TAMU, STAT 415

Mathematical Statistics II: Demystifying statistics

Suhasini Subba Rao

Spring, 2021

Contents

1 Introduction and Review

1.1 Why we spoil statistics with maths

To understand the aims and objectives of this course, we start with a motivating example. Weather stations around the world are constantly collecting data (temperature, air pressure and ozone levels to name but a few). To publish this data in a coherent fashion, it is often summarized, in such a way that one easily finds pertinent trends. For example, if you search for temperature data on the web, you will usually find that the monthly average temperature (at a particular longitude and latitude) is published.

Over the past thirty years or so, scientists have been monitoring the temperatures in Antarctica. The concern is that rises in temperatures in this region will lead to a melting of glaciers and rises in ocean levels. Therefore, we focus our attention on the temperatures collected at Faraday station (later called the Vernadsky research base: a brief history can be found here https://www.bas.ac.uk/about/about-bas/history/british-research-stations-and-refuges/faraday-f/), which has a long history of climatic research (dating back to 1947). From 1951-2004, the monthly temperatures have been published by the British Antarctic Survey. What makes their data set quite unique is that they collect the monthly extremes (max and mins) not just the averages. A plot of the monthly extremes is given in Figure 1.1. As our aim is to understand if the temperatures are rising, we regress the minimum and maximum temperatures against time. However, as this is monthly data, it will have a clear seasonality and that too must be included in the regression. This often done by including a seasonal component in the regression model such as a sine and cosine function with a 12 month period. Thus we fit the following linear regression model to both the maximum and minimum temperatures

$$y_t = \beta_0 + \beta_S \sin\left(\frac{2\pi t}{12}\right) + \beta_C \cos\left(\frac{2\pi t}{12}\right) + \beta_2 t + \varepsilon_t \tag{1.1}$$

to the data.

Remark (Modelling periodicities, such as temperatures)**.** *Monthly temperatures are likely to be periodic with a period of 12 months. A periodic sequence, with a period of 12 months is a sequence* $\{d_{12}(t)\}$ *where* $d_{12}(t) = d_{12}(t+12) = d_{12}(t+24) = \ldots$ *for all* t*. We can model all period sequences with a period of 12 months*

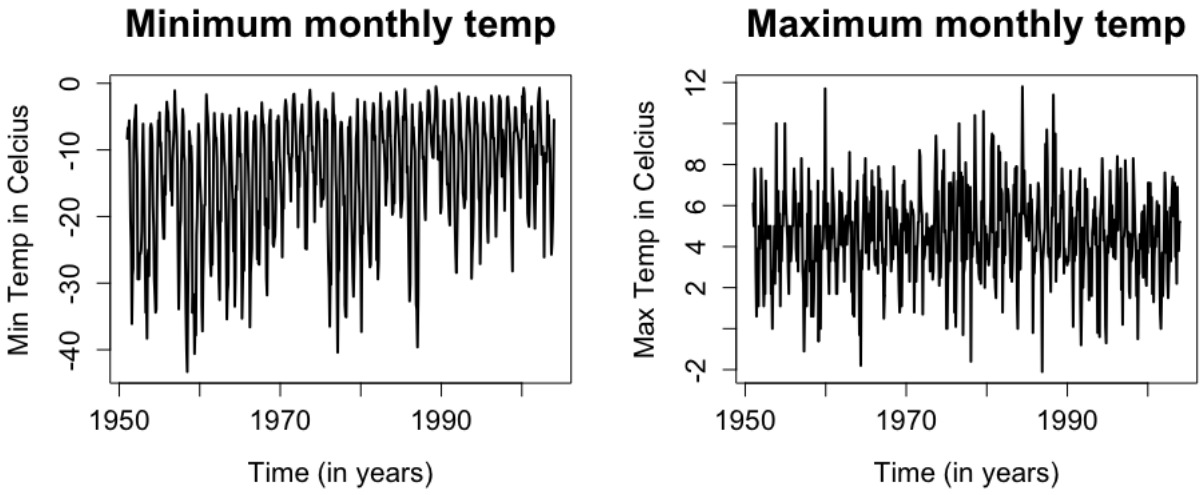


Figure 1.1: Monthly minimum (left) and monthly maximum (right).

using sin and cosine functions;

$$d_{12}(t) = \sum_{s=1}^{6} \left(a_s \sin\left(\frac{2\pi t}{12} s\right) + b_s \cos\left(\frac{2\pi t}{12} s\right) \right).$$

A simple calculation shows that the $d_{12}(t)$ based on this construction is periodic with period 12. As there can be 12 different coefficients for $\{d_{12}(t)\}$ there are 12 different coefficients $\{a_s, b_s\}_{s=1}^{6}$. Thus the above construction is able to model any 12-period sequence. However, this can lead to too many parameters to estimate (lack of parsimony). Thus we often focus on the first sin and cosine regressors $a_1 \sin\left(\frac{2\pi t}{12}\right) + b_1 \cos\left(\frac{2\pi t}{12}\right)$, because in general this contains "most" of the information in $\{d_{12}(s)\}$.

The estimation of the parameters in (1.1) is done in R using the `lm` command, which fits the model using the method of least squares. The R output is given below for the maximum temperatures.

```
> summary(testMax)

Call:
lm(formula = Ymax ~ S + C + time)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6652 -0.9374 -0.0154  1.0011  5.3660

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.739396   8.341244  -0.209    0.835
S            1.417213   0.091405  15.505   <2e-16 ***
C            1.401824   0.091396  15.338   <2e-16 ***
time         0.003205   0.004218   0.760    0.448
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.645 on 644 degrees of freedom
Multiple R-squared:  0.425,     Adjusted R-squared:  0.4223
F-statistic: 158.7 on 3 and 644 DF,  p-value: < 2.2e-16
```

```
> anova(testMax)
Analysis of Variance Table

Response: Ymax
           Df  Sum Sq Mean Sq  F value Pr(>F)
S           1  649.85  649.85 240.1148 <2e-16 ***
C           1  636.93  636.93 235.3418 <2e-16 ***
time        1    1.56    1.56   0.5772 0.4477
Residuals 644 1742.94    2.71
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The left output gives the coefficient estimates, together with with their standard errors. You may recall, from previous statistics classes (STAT212, for instance), that standard errors are instrumental to any data analysis; they quantify the uncertainty we associate with an estimator (we define this precisely below). And to determine if a coefficient is statistically significant we calculate the ratio

$$\text{t value} = \frac{\text{Estimate}}{\text{Std. Error}},$$

where significance of each individual coefficient can be determined by using a t-test (this is the p-value given in $\Pr(>|t|)$). The analysis of the variance on the right measures the reduction in the squared error as each variable is added to the model. The bigger and more dramatic the reduction the smaller the p-value as given by $\Pr(>F)$. Studying the p-values we observe that the seasonal terms (sin and cosine) are statistically significant. However, despite the time coefficient being positive, there is no significant evidence of an increase in the temperatures over time (observe the large p-value corresponding to `time` in both outputs. Recall when there is no statistical evidence of an increase there are two possible explanations (a) there really is no increase or (b) the noise in the data is too large, that it is masks an increase.

We conduct a similar analysis using the minimum temperatures. The output is given below

```
> summary(testMin)

Call:
lm(formula = Ymin ~ S + C + time)

Residuals:
    Min      1Q  Median      3Q     Max
-22.727  -2.652   0.291   3.078  14.364

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -330.68626   26.21982  -12.61   <2e-16 ***
S              8.98994    0.28732   31.29   <2e-16 ***
C              5.48906    0.28729   19.11   <2e-16 ***
time           0.16026    0.01326   12.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.171 on 644 degrees of freedom
Multiple R-squared:  0.697,     Adjusted R-squared:  0.6956
F-statistic: 493.8 on 3 and 644 DF,  p-value: < 2.2e-16
```

```
> anova(testMin)
Analysis of Variance Table

Response: Ymin
           Df  Sum Sq Mean Sq F value    Pr(>F)
S           1 25900.7 25900.7  968.54 < 2.2e-16 ***
C           1  9808.8  9808.8  366.79 < 2.2e-16 ***
time        1  3907.1  3907.1  146.10 < 2.2e-16 ***
Residuals 644 17221.8    26.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Studying the p-values in this output we observe that the seasonal component is statistically significant and also the linear increase in temperatures over time. The data suggests that there is an increase in the monthly minimum temperatures over time. This is very interesting. To summarize, there is no evidence of an increase in the monthly maximum temperatures, but there is evidence of an increase in the minimum temperatures. But before we make any such conclusions, we have to decide if the analysis is valid. This means taking a step back and understanding where all the standard errors, p-values etc come from. If they have been calculated incorrectly, then the conclusions of the analysis may be wrong.

First we make a plot of a histogram of the residuals

$$\widehat{\varepsilon_t} = y_t - \widehat{\beta}_0 - \widehat{\beta}_{S,1} \sin\left(\frac{2\pi t}{12}\right) - \widehat{\beta}_{C,1} \cos\left(\frac{2\pi t}{12}\right) - \widehat{\beta}_2 t.$$

This is given in Figure 1.2. We observe that the distribution of the residuals of the maximum temperatures look symmetric, but the distribution of the residuals of the minimum temperatures appear left skewed.

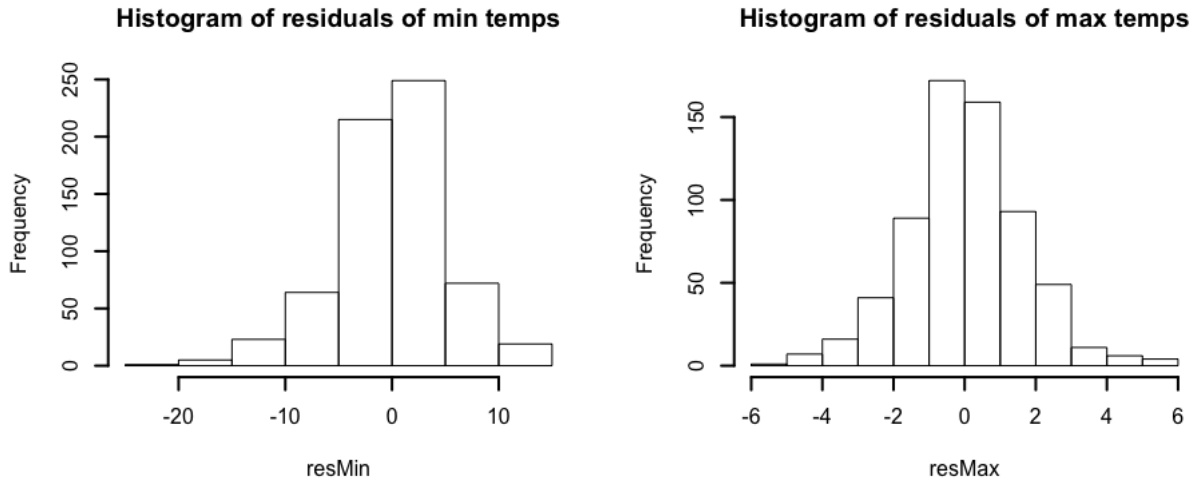


Figure 1.2: Histogram of residuals: Minimum of left and maximum of right.

(1) Clearly the residuals for the minimum temperatures are not normal (the normal/Gaussian distribution is symmetric). Does that matter, do we require normality for the analysis?

(2) What about the standard errors, how are they calculated? How, was the data collected? The data has been collected over time. Often data that has been collected over time, are dependent. The independence assumption probably does not hold. To demonstrate that there is possible dependence, the estimated autocorrelation (ACF) plot of the residuals is given in Figure 1.3. We observe what appears to be dependence (this is beyond this course, and will be discussed in a time series course). Does dependence influence the standard errors? If it does, where will it effect the conclusions of the study?

Though we cannot answer, all the above questions in this course; time series data tends to be extremely complex. In this course, we will hopefully understand why things work. And if we understand why they work, we can also understand when procedures may not work, and how it can influence the conclusions that we draw.

1.2 Joint distributions of random variables

Please Review STAT 414, and pay particular attention to the joint distribution of discrete and continuous random variables. We give a quick summary of what you need to know. Suppose that $\underline{X} = (X_1, \ldots, X_n)$ is a

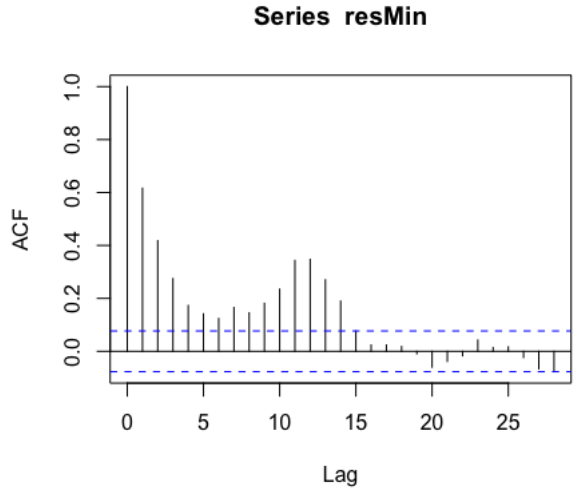


Figure 1.3: The estimated autocorrelation plot of the minimum residuals.

discrete random vector. Then their joint probability mass function is

$$p_{\underline{X}}(x_1, x_2, \ldots, x_n) = P_{\underline{X}}(X_1 = x_1, \ldots, X_n = x_n).$$

Suppose that $(X_1, \ldots, X_n)$ is a vector of continuous random variables then the joint density is a piecewise continuous function $f_{\underline{X}}(x_1, \ldots, x_n)$ where

$$P_{\underline{X}}((X_1, X_2, \ldots, X_n) \in A) = \int_A f_{\underline{X}}(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

Definition 1.1 (Independence)**.** *The random variables $X_1, X_2, \ldots, X_n$ are said to be independent if their joint distribution (or density) function can be written as the product of their marginal distributions*

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \ldots F_{X_n}(x_n) \qquad \text{for all } x_1, \ldots, x_n.$$

Below, we generalize the above notion to independence between vectors.

Definition 1.2. *Suppose $\underline{Y} = (Y_1, \ldots, Y_p)$ and $\underline{X} = (X_1, \ldots, X_q)$ are random vectors. $\underline{X}$ and $\underline{Y}$ are said to be independent if joint cumalative distribution function of $\underline{X}$ and $\underline{Y}$ can be written as the product of the joint distributions*

$$F_{\underline{X},\underline{Y}}(\underline{x}, \underline{y}) = F_{\underline{X}}(\underline{x}) F_{\underline{Y}}(\underline{y}).$$

Definition 1.3 (iid)**.** *The random variables $\{X_i\}_{i=1}^n$ are said to be independent and identically distributed (iid for short) if $\{X_i\}_{i=1}^n$ are independent* and *the marginal distribution is the same for all the random variables.*

Remark (Conditional probabilities)**.** *We recall that the conditional probability of event A given B is*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

The two events A and B are statistically independent if $P(A|B) = P(A)$ (occurence of event B has no impact on the probability of event A). Suppose A is an event corresponding to random variables X (technically we say that A belongs to a sigma-algebra generated by X, but this is not a technical course) and B corresponds to the random variable Y. If X and Y are independent random variables as defined above, then $P(A|B) = P(A)$ and A and B are independent events.

1.3 Euclidean space and matrix multiplication

In this section we review some results from linear algebra, focusing on this simple case of finite dimension Euclidean space. In statistics we store data as vectors (or matrices), therefore either implicitly or explictly we are manipulating vectors. A solid understanding of linear algebra will help in both the mathematical proofs but also writing good and fast pieces of code. For example, the empirical correlation can be understood in terms of projections of one vector onto another. Thus rather than a code the correlation as an n-loop (which takes time) we can simply write it as dot/scalar/inner product which is computationally faster.

Reminder: if α is a scalar (a number) and $\underline{x}$ is a vector, then

$$\alpha \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_d \end{pmatrix}.$$

1.3.1 Inner (scalar) products and projections

We define the inner product (also called the scalar product) between the vector $\underline{x} \in \mathbb{R}^d$ and $\underline{y} \in \mathbb{R}^d$ as

$$\langle \underline{x}, \underline{y} \rangle = \sum_{i=1}^{d} x_i y_i.$$

Observe that by definition $\langle \underline{x}, \underline{y} \rangle = \langle \underline{y}, \underline{x} \rangle$ (the inner product is symmetric). Inner products satisfy some basic properties. The one we will use is

$$\langle \alpha \underline{x} + \beta \underline{z}, \underline{y} \rangle = \alpha \langle \underline{x}, \underline{y} \rangle + \beta \langle \underline{z}, \underline{y} \rangle$$

where $\alpha, \beta \in \mathbb{R}$, which is easily verified. The Euclidean distance is the length of the vector $\underline{x}$, and is $\|\underline{x}\| = \sqrt{\langle \underline{x}, \underline{x} \rangle} = \sqrt{\sum_{i=1}^{d} x_i^2}$ (this is easily understood by considering vectors on $\mathbb{R}^2$ and measuring the distance from the origin $(0, 0)$ to the vector). Often it is useful to deal with the the standardized vector

$$\frac{\underline{x}}{\|\underline{x}\|}.$$

The standardisation means that the length of this vector (its Euclidean distance) is one. In Euclidean space, inner products have a useful geometric properties. We state these below.

Remark (Interpreting a zero inner product). *If $\langle \underline{x}, \underline{y} \rangle = 0$, then $\underline{x}$ and $\underline{y}$ are orthogonal. This means the angle between the two vectors is 90 degrees. In $\mathbb{R}^2$ there will only be one vector (up to a scalar constant) $\underline{y}$ that is orthogonal to $\underline{x}$. In $\mathbb{R}^3$ a plane is orthogonal to $\underline{x}$.*

Definition 1.4 (Projection). *The projection of $\underline{y}$ onto $\underline{x}$ is $\alpha\underline{x}$, where the value α such that the innerproduct between $\underline{y} - \alpha\underline{x}$ and $\underline{x}$ is zero. The value α is obtained by solving*

$$\langle \underbrace{\underline{y} - \alpha\underline{x}}_{\text{red dashed line}}, \underline{x} \rangle = 0$$

$$= \langle \underline{y}, \underline{x} \rangle - \alpha \langle \underline{x}, \underline{x} \rangle = 0 \Rightarrow \alpha = \frac{\langle y, x \rangle}{\langle x, x \rangle} = \frac{\langle \underline{y}, \underline{x} \rangle}{\|\underline{x}\|^2}.$$

Relationship to least squares The projection of $\underline{y}$ onto $\underline{x}$ is the same as finding the α which minimises the least squares criterion

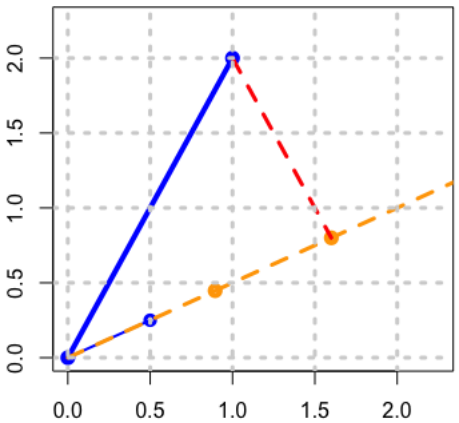
$$L(\alpha) = \sum_{j=1}^{n} (y_j - \alpha x_j) \qquad (= \langle \underline{y} - \alpha\underline{x}, \underline{y} - \alpha\underline{x} \rangle).$$

Differentiating $L(\alpha)$ with respect to α and setting to zero gives

$$\alpha = \frac{\langle y, x \rangle}{\langle x, x \rangle} = \frac{\langle \underline{y}, \underline{x} \rangle}{\|\underline{x}\|^2}.$$

We give an example for dimension $d = 2$. Consider the vectors $\underline{x}$ and $\underline{y}$ where

$$\underline{y} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \text{and} \quad \underline{x} = \begin{pmatrix} 1/4 \\ 1/2 \end{pmatrix}.$$



The vectors $\underline{y}$ and $\underline{x}$ are the blue line and dashed blue line on the plot. The standardized vector $\frac{\underline{x}}{\|\underline{x}\|} = 5^{-1/2}(1, 2)'$ is the yellow point on the plot. The projection of $\underline{y}$ onto the line $\{\alpha\underline{x}; \alpha \in (-\infty, \infty)\}$ is the red point on the yellow dashed line which is orthogonal with the red dashed line in the plot. This point is $\alpha\underline{x}$ where

$$\alpha = \frac{\langle \underline{y}, \underline{x} \rangle}{\|\underline{x}\|^2} = \frac{16}{5} = 3.2.$$

Observe that $\langle \underline{y}, \underline{x} \rangle$ can be viewed as a measure of similarity between the vectors $\underline{x}$ and $\underline{y}$. If $\langle \underline{y}, \underline{x} \rangle = 0$, the vectors are orthogonal and there is no similarity between the vectors.

In the next section we describe the orthogonal representation of vectors. We have already come across this, in the above example. The yellow projection vector $(4/5, 8/5)'$ and the red vector $(6/5, -3/5)$ are orthogonal (observe that they are at right angles). They form the building blocks of $\underline{y}$:

$$\underline{y} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 4/5 \\ 8/5 \end{pmatrix}}_{=\alpha \underline{x}} + \begin{pmatrix} 6/5 \\ -3/5 \end{pmatrix}.$$

1.3.2 Orthonormal basis expansion

The vectors $\underline{e}_1 = (1, 0, 0)$, $\underline{e}_2 = (0, 1, 0)$ and $\underline{e}_3 = (0, 0, 1)$ form what is called an orthonormal basis (defined below) of $\mathbb{R}^3$. It is clear that any $\underline{y} = (y_1, y_2, y_3) \in \mathbb{R}^3$ can writen as $\underline{y} = y_1\underline{e}_1 + y_2\underline{e}_2 + y_3\underline{e}_3$ (a linear combination of the basis). However, this basis is far from unique. We can rewrite $\underline{y}$ in terms of any orthonormal basis. Below we show how this possible using the idea of projections defined in the previous section.

The d-vectors $\{\underline{e}_j\}_{j=1}^d$ is an orthonormal basis of $\mathbb{R}^d$ if for all $1 \leq i, j \leq d$ we have

$$\langle \underline{e}_i, \underline{e}_j \rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

Note from above, the length of the vectors are $\langle \underline{e}_i, \underline{e}_i \rangle = \|\underline{e}_i\|^2 = 1$. Orthonormal basis are not unique (there are an uncountable number of different orthonormal basis on $\mathbb{R}^d$).

For a given orthonormal basis $\{\underline{e}_j\}_{j=1}^d$ in $\mathbb{R}^d$, we now show that we can decompose any vector $\underline{y} \in \mathbb{R}^d$ as a weighted sum of the orthogonal basis. This is easily seen by using the projection argument given in the previous section. Projecting $\underline{y}$ onto $\underline{e}_1$ gives

$$\alpha_1 = \frac{\langle \underline{y}, \underline{e}_1 \rangle}{\|\underline{e}_1\|^2} = \langle \underline{y}, \underline{e}_1 \rangle.$$

The remainder (residual) is $\underline{y} - \langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1$ (which is orthogonal to $\underline{e}_1$). Next we project $\underline{y} - \alpha_1 \underline{e}_1$ onto $\underline{e}_2$. This gives the coefficient

$$\alpha_2 = \frac{\langle \underline{y} - \alpha_1 \underline{e}_1, \underline{e}_2 \rangle}{\|\underline{e}_2\|^2} = \frac{\langle \underline{y}, \underline{e}_2 \rangle - \alpha_1 \langle \underline{e}_1, \underline{e}_2 \rangle}{\|\underline{e}_2\|^2} = \langle \underline{y}, \underline{e}_2 \rangle.$$

Thus the residual after projecting on $\underline{e}_1$ and $\underline{e}_2$ is

$$\underline{y} - \alpha_1 \underline{e}_1 - \alpha_2 \underline{e}_2.$$

Iterating this we obtain the orthonormal basis expansion

$$\underline{y} = \sum_{j=1}^{d} \langle \underline{y}, \underline{e}_j \rangle \underline{e}_j.$$

Thus we have decomposed $\underline{y}$ into vectors into orthogonal building block. Each coefficient $\alpha_j \langle \underline{y}, \underline{e}_j \rangle$ describes the how much of the vector $\underline{e}_j$ is required to build $\underline{y}$. Rewriting $\underline{y}$ in terms of a basis $\{\underline{e}_j\}_{j=1}^{d}$ is the same we writing $\underline{y}$ in terms of a rotation of the usually axis.

Remark (Projection onto planes in $\mathbb{R}^d$). *Suppose that $\{\underline{e}_j\}_{j=1}^{d}$ is an orthonormal basis of $\mathbb{R}^d$, and define the plane, Π, as all linear combination of the vectors $\{\underline{e}_j\}_{j=1}^{r}$ (where $r < d$) i.e.*

$$\Pi = \left\{ \sum_{s=1}^{r} \alpha_s \underline{e}_s; \quad \alpha_1, \ldots, \alpha_r \in \mathbb{R} \right\}.$$

Since $\{\underline{e}_j\}_{j=1}^{s}$ are orthogonal, the projection of $\underline{y} \in \mathbb{R}^d$ onto Π can be done sequentially by projecting onto each $\underline{e}_s$, this gives the projection

$$P_{\Pi}(\underline{y}) = \sum_{s=1}^{r} \langle \underline{e}_s, \underline{y} \rangle \underline{e}_s.$$

Example 1.1 (Examples of orthonormal basis on R^2). *(i) The simplest example is*

$$\underline{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \underline{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Clearly for any $\underline{x}_2 \in \mathbb{R}^2$ we have

$$\underline{x}_2 = x_1 \underline{e}_1 + x_2 \underline{e}_2.$$

(ii) Another orthonormal basis is

$$\underline{e}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \underline{e}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

It is easy to show that $\langle \underline{e}_1, \underline{e}_2 \rangle = (1-1)/2 = 0$. For the purpose of visualisation, it is useful to plot the basis.

Using the above basis, any vector $\underline{x}' = (x_1, x_2) \in \mathbb{R}^2$ can be written as

$$\begin{aligned} \underline{x} &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \langle \underline{x}, \underline{e}_1 \rangle \underline{e}_1 + \langle \underline{x}, \underline{e}_2 \rangle \underline{e}_2 \\ &= \left(\frac{(x_1 + x_2)}{\sqrt{2}} \right) \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \left(\frac{(x_1 - x_2)}{\sqrt{2}} \right) \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \end{aligned}$$

Example 1.2 (Examples of orthonormal basis on R^3). *(i) The simplest example is*

$$\underline{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \underline{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \underline{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

(ii) An alternative orthonormal basis is

$$\underline{e}_1 = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \underline{e}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \qquad \underline{e}_3 = \frac{1}{\sqrt{6}}\begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}$$

Again you can show that $\langle \underline{e}_1, \underline{e}_2 \rangle = \langle \underline{e}_1, \underline{e}_3 \rangle = \langle \underline{e}_2, \underline{e}_3 \rangle = 0$ *(please do this).*

Using the above basis, any vector $\underline{x} \in \mathbb{R}^3$ *can be written as*

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \\ x_2 \end{pmatrix} &= \langle \underline{x}, \underline{e}_1 \rangle \underline{e}_1 + \langle \underline{x}, \underline{e}_2 \rangle \underline{e}_2 + \langle \underline{x}, \underline{e}_3 \rangle \underline{e}_3 \\ &= \left(\frac{x_1 + x_2 + x_3}{\sqrt{3}} \right) \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \left(\frac{x_2 + x_3}{\sqrt{2}} \right) \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} + \left(\frac{-2x_1 + x_2 + x_3}{\sqrt{6}} \right) \frac{1}{\sqrt{6}} \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}. \end{aligned}$$

One way to construct the above basis is to use properties of sins and cosines;

$$\underline{e}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} \cos(2\pi \times 0/3) \\ \cos(2\pi \times 1/3) \\ \cos(2\pi \times 2/3) \end{pmatrix} \approx \frac{1}{1.5} \begin{pmatrix} 1 \\ -0.5 \\ -0.5 \end{pmatrix} \text{ and } \underline{e}_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} \sin(2\pi \times 0/3) \\ \sin(2\pi \times 1/3) \\ \sin(2\pi \times 2/3) \end{pmatrix} \approx \frac{1}{\sqrt{1.5}} \begin{pmatrix} 0 \\ 0.866 \\ -0.866 \end{pmatrix}.$$

1.3.3 Parseval's identity and other L2-norm identities

The purpose of this section is to state some useful identities. In order to derive the t-distribution, MLE for normal random variables, two sample t-test and the ANOVA one needs to establish several identities. One method for establishing these identities is to use cumbersome and not very enlightening chug and plug methods. Another, is to use some powerful results from linear algebra. The approach in this class is to use the latter. I summarize the most pertinent results from linear algebra below.

Parseval's identity

We will use the representation

$$\underline{y} = \sum_{j=1}^{d} \langle \underline{y}, \underline{e}_j \rangle \underline{e}_j,$$

to write the Euclidean distance interms of $\underline{y}$ in terms of $\langle \underline{y}, \underline{e}_j \rangle$. First observe that if the basis is $\underline{e}_1 = (1, 0, 0, \ldots, 0)$, $\underline{e}_2 = (0, 1, 0, \ldots, 0)$,. . ., $\underline{e}_d = (0, 0, 0, \ldots, 1)$, then it is clear that $\sum_{j=1}^d y_j^2 = \sum_{j=1}^n \langle \underline{y}, \underline{e}_j \rangle^2$ (since $\langle \underline{y}, \underline{e}_j \rangle^2 = y_j^2$). We show below that this is true for any orthonormal basis:

$$\begin{aligned} \sum_{j=1}^d y_j^2 &= \langle \underline{y}, \underline{y} \rangle \\ &= \sum_{j_1, j_2 = 1}^d \langle \underline{y}, \underline{e}_{j_1} \rangle \langle \underline{y}, \underline{e}_{j_1} \rangle \langle \underline{e}_{j_1}, \underline{e}_{j_2} \rangle \\ &= \sum_{j=1}^d \langle \underline{y}, \underline{e}_j \rangle^2. \end{aligned} \tag{1.2}$$

This is called Parseval's identity.

A useful extension to the above equality is the following result

$$\| \underline{y} - \langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1 \|^2 = \sum_{j=2}^d \langle \underline{y}, \underline{e}_j \rangle^2.$$

The proof follows the same argument as that given above[1].

We now give a useful application of this result. Suppose $\underline{e}_1 = d^{-1/2}(1, \ldots, 1)$, then $\langle \underline{y}, \underline{e}_1 \rangle = d^{-1/2} \sum_{j=1}^d y_j = d^{1/2} \bar{y}$ (where $\bar{y}$ is the average of the elements in $\underline{y}$). Thus

$$\langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1 = d^{1/2} \bar{y} d^{-1/2} (1, \ldots, 1) = \bar{y}(1, \ldots, 1).$$

Therefore

$$\underline{y} - \langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1 = (y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_d - \bar{y}).$$

Suppose $\{\underline{e}_j\}_{j=2}^d$ are orthonormal to $\underline{e}_1$, then we have

$$\| \underline{y} - \langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1 \|^2 = \sum_{j=1}^d (y_j - \bar{y})^2 = \sum_{j=2}^d \langle \underline{y}, \underline{e}_j \rangle^2. \tag{1.3}$$

[1]The precise proof is that since $\underline{y} - \langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1 = \sum_{j=2}^d \langle \underline{y}, \underline{e}_j \rangle \underline{e}_j$, then

$$\begin{aligned} \| \underline{y} - \langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1 \|^2 &= \langle \underline{y} - \langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1, \underline{y} - \langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1 \rangle \\ &= \langle \sum_{j=2}^d \langle \underline{y}, \underline{e}_j \rangle \underline{e}_j, \sum_{j=2}^d \langle \underline{y}, \underline{e}_j \rangle \underline{e}_j \rangle = \sum_{j=1}^d \langle \underline{y}, \underline{e}_j \rangle^2. \end{aligned}$$

Or alternatively simply use that $\underline{y} - \langle \underline{y}, \underline{e}_1 \rangle \underline{e}_1$ is the projection of $\underline{y}$ onto the subspace spanned by $\{\underline{e}_j\}_{j=2}^d$, and the result immediately follows from Parseval's identity.

L2 norm of orthogonal transformations and their determinant

So now we state one more useful result. Suppose that $\{\underline{e}_j\}_{j=1}^d$ are orthonormal (row) vectors in $\mathbb{R}^d$. Define the orthogonal transformation matrix E where

$$E' = \begin{pmatrix} \underline{e}_1' & \underline{e}_2' & \dots & \underline{e}_d' \end{pmatrix}.$$

Then consider the transformation of the column vector $\underline{y} = E\underline{x}$. It can be shown that

$$\left\|E\underline{x}\right\|_2^2 = \left\|\underline{x}\right\|_2^2. \tag{1.4}$$

PROOF. This result is straightfoward to show, using that $E'X = I_d$ we have

$$\left\|E\underline{x}\right\|_2^2 = \langle \|E\underline{x}, E\underline{x}\rangle = \underline{x}'E'E'\underline{x} = \underline{x}'I_d\underline{x} = \left\|\underline{x}\right\|_2^2.$$

□

Finally, another useful identity. A well known result is that $\det(AB) = \det(A)\det(B)$. This implies that

$$\det(E)^2 = 1$$

since $1 = \det(I_d) = \det(E'E) = \det(E)\det(E') = \det(E)^2$. This implies

$$\det(E'\Sigma E) = \det(E')\det(\Sigma)\det(E) = \pm\det(\Sigma). \tag{1.5}$$

1.3.4 Orthonormal vectors and random coefficients

We now connect orthogonal vector to random vectors. Define the two orthogonal vectors $\underline{e}_1' = (1, 0)$ and $\underline{e}_2' = (0, 1)$ and the random vector

$$\underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = Z_1\underline{e}_1 + Z_2\underline{e}_2$$

where Z_i are iid random variables (assume normal). $\underline{e}_1$ gives the coordinate of the vector and Z_1 the corresponding length. Thus $\underline{X}$ is a combination of these two orthogonal vectors. If you simulated $\underline{Y}$ you would get a symmetric dust cloud. We but we need not stick to orthogonal vectors on the x and y-axis. Also the lengths on each axis need not be the same. Consider the random vector

$$\underline{Y} = 2Z_1\underline{e}_1 + 0.3Z_2\underline{e}_2,$$

where Z_1 and Z_2 are iid standard normal random variables and $\underline{e}_1$ and $\underline{e}_2$ are orthonormal vectors. Here the constribution from $\underline{e}_1$ tends to be more than $\underline{e}_2$. A dust cloud of $\underline{Y}$ will be an ellipse. You will simulate $\underline{Y}$ in HW1.

Any random vector can be decomposed in terms of orthogonal vectors, jus like above. This forms the basis of principal component analysis (PCA).

1.3.5 Matrix multiplication

In this section we review how matrices are multiplied. We start with the simple case A is an $m \times n$ matrix and $\underline{x}$ is a n-dimension column vector. Then we have

$$A\underline{x} = \begin{pmatrix} \underline{a}_1 \\ \underline{a}_2 \\ \vdots \\ \underline{a}_m \end{pmatrix} \underline{x} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \end{pmatrix} = \begin{pmatrix} \langle \underline{a}_1, \underline{x} \rangle \\ \langle \underline{a}_2, \underline{x} \rangle \\ \vdots \\ \langle \underline{a}_m, \underline{x} \rangle \end{pmatrix}$$

Note that $\underline{a}_j$ is row vector and $\underline{x}$ is a column vector, but $\langle \underline{a}_j, \underline{x} \rangle$ treats them as all having the same alignment (which is techically not quite right, but hardly matters). Therefore the $A\underline{x}$ is simply the innerproduct between each row vectors of A with the column vector $\underline{x}$. It can be considered as a linear transformation of the vector $\underline{x}$. Notice that the number of columns of A must match the number of rows in $\underline{x}$, else the innerproduct and $A\underline{x}$ in not well defined.

Definition 1.5 (Transpose)**.** *The transposition A' switches the column vectors into a row vectors and row vectors into column vectors i.e. using the above definitions*

$$\begin{aligned} \underline{x}' &= \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \\ A' &= \begin{pmatrix} a_{11} & a_{12} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix} = \begin{pmatrix} \underline{a}'_1 & \underline{a}'_2 & \dots & \underline{a}'_m \end{pmatrix}. \end{aligned}$$

Suppose A is an $n \times n$ matrix, this is called a squared matrix. If $A' = A$ it is called a symmetric matrix. Symmetric matrices are very important in statistics, as all variance matrices (defined below) are symmetric.

Based on the above definition we observe

$$(A\underline{x})' = \underline{x}'A' = \begin{pmatrix} \langle \underline{a}_1, \underline{x} \rangle & \langle \underline{a}_2, \underline{x} \rangle & \dots & \langle \underline{a}_m, \underline{x} \rangle \end{pmatrix}.$$

Further if $\underline{y}$ is a m-dimension column vector, then

$$\underline{y}'A\underline{x} = \sum_{i=1}^{m} y_i \langle \underline{a}_i, \underline{x} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{i,j} y_i x_j.$$

Example 1.3. *Suppose $\underline{x} = (x_1, x_2, \ldots, x_n)'$.*

(i) Then $\underline{x}\underline{x}'$ is an $n \times n$ matrix:

$$\underline{x}\underline{x}' = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} = \begin{pmatrix} x_1^2 & x_1x_2 & \dots & x_1x_n \\ x_2x_1 & x_2^2 & \dots & x_2x_n \\ x_3x_1 & x_3x_2 & \dots & x_3x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \dots & x_n^2 \end{pmatrix}$$

(ii) Then $\underline{x}'\underline{x}$ is a scalar:

$$\underline{x}'\underline{x} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1^2 + x_2^2 + \ldots + x_n^2 = \langle \underline{x}, \underline{x} \rangle.$$

We can generalize this notion to the product of the $m \times n$ matrix A and $n \times p$ matrix B as follows

$$\begin{aligned} AB &= \begin{pmatrix} \underline{a}_1 \\ \underline{a}_2 \\ \vdots \\ \underline{a}_m \end{pmatrix} \left(\underline{b}_1, \underline{b}_2, \dots, \underline{b}_p \right) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{np} \end{pmatrix} \\ &= \begin{pmatrix} \langle \underline{a}_1, \underline{b}_1 \rangle & \langle \underline{a}_1, \underline{b}_2 \rangle & \dots & \langle \underline{a}_1, \underline{b}_p \rangle \\ \langle \underline{a}_2, \underline{b}_1 \rangle & \langle \underline{a}_2, \underline{b}_2 \rangle & \dots & \langle \underline{a}_2, \underline{b}_p \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{a}_m, \underline{b}_1 \rangle & \langle \underline{a}_m, \underline{b}_2 \rangle & \dots & \langle \underline{a}_m, \underline{b}_p \rangle \end{pmatrix} \end{aligned}$$

Thus AB is comprised of mp innerproducts. The number of columns in A must match the number of columns in B, else AB is not well defined.

Clearly AB is not commutative (in general, you cannot change the order: $AB \neq BA$)

Transpose: Observe the general identity:

$$(AB)' = B'A'.$$

1.4 Expectation, variance and covariance

1.4.1 Expectation

Statisticians almost always deal with averages of a sample. This is because the averages (in most situations) converge to its corresponding expectation. We start with the definition of the average and then define the expectation (we completely avoid the use of measures, sigma algebras etc).

Suppose the random variable X is a discrete valued random variable[2] taking values $\{k_i\}$ and distribution $p(k_i)$. Suppose we observe multiple realisations $\{x_i\}$, then the averages

$$\frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

will (almost surely) limit to the following "expectation":

$$\mathrm{E}[X] = \sum_{i=0}^{\infty} k_i p(k_i) \text{ and in general } \mathrm{E}[g(X)] = \sum_{i=0}^{\infty} g(k_i)p(k_i).$$

respectively, where g is any function. This is called the expectation of the random variable X and $g(X)$. If X is a continuous value random[3] variable with density $f(x)$ then the expectation of X and $g(X)$ is

$$\mathrm{E}[X] = \int_{\mathbb{R}} x f(x) dx \text{ and in general } \mathrm{E}[g(X)] = \int g(x) f(x) dx.$$

The expection can easily generalised to a vector by taking the expectation entrywise. Let $\underline{X} = (X_1, \ldots, X_d)$ be a random row vector, then

$$\mathrm{E}(\underline{X}) = \begin{pmatrix} \mathrm{E}(X_1) & \mathrm{E}(X_2) & \ldots & \mathrm{E}(X_d) \end{pmatrix} = \begin{pmatrix} \mu_1 & \mu_2 & \ldots & \mu_d \end{pmatrix} = \underline{\mu}$$

The joint expectation, $\mathrm{E}(XY)$, generalizes the above definition, and is taken over the joint distribution of (X, Y).

Lemma 1.1 (Expectation of products of independent random variables)**.** *Suppose that X and Y are independent random variables. Then*

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y).$$

Example 1.4 (Expectations of mixtures)**.** *In statistics often modelling data with mixtures of distributions is useful. Let X, Y and U be independent random variables. Let U be a Bernoulli random variable $U \in \{0, 1\}$, where $P(U = 0) = p$ and $P(U = 1) = 1 - p$. Define the new random variable*

$$Z = UX + (1 - U)Y.$$

The expectation

$$\begin{aligned} \mathrm{E}[Z] &= \mathrm{E}[UX + (1-U)Y | U = 0]P(U = 0) + \mathrm{E}[UX + (1-U)Y | U = 1]P(U = 1) \\ &= \mathrm{E}[(1-U)Y | U = 0]P(U = 0) + \mathrm{E}[UX | U = 1]P(U = 1) \\ &= p\mathrm{E}[Y] + (1-p)\mathrm{E}[X]. \end{aligned}$$

By a similar argument we have $Z^2 = (UX + (1-U)Y)^2 = (U^2X^2 + (1-U)^2Y^2 + 2U(1-U)XY$

$$\begin{aligned} \mathrm{E}[Z^2] &= \mathrm{E}[(Z^2 | U = 0]P(U = 0) + \mathrm{E}[Z^2 | U = 1]P(U = 1) \\ &= \mathrm{E}[(1-U)^2Y^2 | U = 0]P(U = 0) + \mathrm{E}[U^2X^2 | U = 1]P(U = 1) \\ &= p\mathrm{E}[Y^2] + (1-p)\mathrm{E}[X^2]. \end{aligned}$$

[2] For example, the binomial or Poisson distribution.

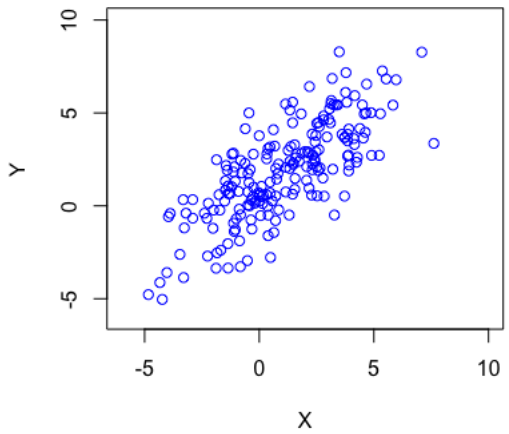
[3] For example, the normal distribution

1.4.2 Example: Interpreting the covariance of bivariate data

Suppose we observe the bivariate data

$$\begin{pmatrix} -0.26 \\ 2.78 \end{pmatrix}, \begin{pmatrix} -1.51 \\ 0.25 \end{pmatrix}, \begin{pmatrix} -0.86 \\ -1.89 \end{pmatrix}, \ldots, \begin{pmatrix} 3.78 \\ 7.17 \end{pmatrix},$$

In total we observe 200 vectors.



On the right we have made a scatter plot of the above data set. Clearly, there appears to be some sort of "linear" dependence between the the two variables. How to measure this? One solution is to use the similarity measure (inner product) described in the previous section.

To do this, we rewrite the bivariate vectors as two 200-dimension vectors

$$\begin{aligned} \underline{x} &= (x_1, x_2, \ldots, x_{200}) = \begin{pmatrix} -0.26, & -1.51, & -0.86, & \ldots & 3.78 \end{pmatrix} \\ \underline{y} &= (y_1, y_2, \ldots, y_{200}) = \begin{pmatrix} 2.78, & 0.25, & -1.89, & \ldots & 7.17 \end{pmatrix}, \end{aligned}$$

If the vectors are "similar" they will almost lie on the same line and have a "large" inner product. We first centralize the vectors by subtracting the average for each vector and then evaluate the (average) inner product

$$n^{-1}\langle \underline{x} - \bar{x}\underline{1}, \underline{y} - \bar{y}\underline{1} \rangle = \frac{1}{n}\sum_{i=1}^{200}(x_i - \bar{x})(y_i - \bar{y})$$

which in this example is 4.7. The average squared spread of the data is

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ and } \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

which for this example is 5.9 and 6.7 respectively. Thus a summary of the inner products is

$$\frac{1}{n}X'X = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{200}(x_i - \bar{x})^2 & \frac{1}{n}\sum_{i=1}^{200}(x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{n}\sum_{i=1}^{200}(x_i - \bar{x})(y_i - \bar{y}) & \frac{1}{n}\sum_{i=1}^{200}(y_i - \bar{y})^2 \end{pmatrix} = \begin{pmatrix} 5.9 & 4.7 \\ 4.7 & 6.7 \end{pmatrix}.$$

If the pairs (X_i, Y_i) are independent realisation (over i) from a distribution. As we increase the number of realisation (see Section 1.4.1) the above estimates the following expection

$$\frac{1}{n}\sum_{i=1}^{200}(x_i - \bar{x})(y_i - \bar{y}) \longrightarrow \mathrm{E}\left[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right].$$

This is called the covariance between the random variables X and Y.

Taking the limit of the above over all possible realisations we have

$$\mathrm{var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \text{ and } \mathrm{var}(Y) = \mathrm{E}\left[(Y - \mathrm{E}(Y))^2\right],$$

which is called the variance of the random variables X and Y respectively[4].

We recall that in Section 1.3.1 we discuss standardized vectors. Where we divide a vector by the distance to ensure it has a length one. We do this to our data vectors $\underline{x}$ and $\underline{y}$:

$$\frac{(x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \qquad \frac{(y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Then the projection of one vector onto the other (it does not matter which way round it is) is

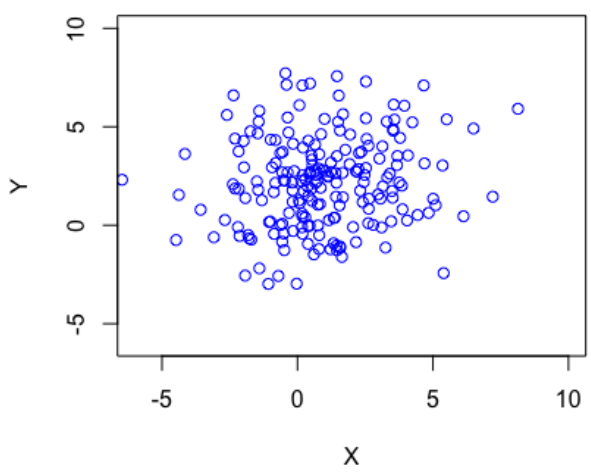
$$\frac{\sum_{i=1}^n (x_1 - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{n^{-1}\sum_{i=1}^n (x_1 - \bar{x})(y_i - \bar{y})}{\sqrt{n^{-1}\sum_{i=1}^n (x_i - \bar{x})^2 n^{-1}\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

This measures the linear dependence between the two vectors, after taking into account their length. Both the numerator and denominator are averages, thus as the sample size grows, the above limits to the correlation between (X, Y), which is defined as

$$\frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}.$$

Below we give an example of a data set where there is little or no correlation.

[4] Often we use the notation $\mathrm{E}[(X - \mu)^2]$ and $\mathrm{E}(X - \mu)^2$ interchangably. When you see $\mathrm{E}(X - \mu)^2$ remember it means $(X - \mu)^2$ is evaluated first, then the expectation of $(X^2 - 2X\mu + \mu^2)$ evaluated.



A scatter plot of a different data set. There does not appear to be any linear dependence between them. The inner product is

$$n^{-1}\langle \underline{x} - \bar{x}\underline{1}, \underline{y} - \bar{y}\underline{1}\rangle = \frac{1}{200}\sum_{i=1}^{200}(x_i - \bar{x})(y_i - \bar{y}) = 0.621$$

and

$$\frac{\sum_{i=1}^{n}(x_1 - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = 0.111.$$

The spread of the bivariate data and its linear dependence can be succinctly summarized in terms of a matrix:

$$\begin{aligned}\frac{1}{n}X'X &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 & \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) & \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \end{pmatrix} \\ &\quad \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} (x_i - \bar{x}) \\ (y_i - \bar{y}) \end{pmatrix}\begin{pmatrix} (x_i - \bar{x}), & (y_i - \bar{y}) \end{pmatrix} \\ &= \begin{pmatrix} 4.55 & 0.62 \\ 0.62 & 6.77 \end{pmatrix}.\end{aligned}$$

This limits to the variance matrix

$$\operatorname{var}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \operatorname{var}(X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(Y,X) & \operatorname{var}(Y) \end{pmatrix},$$

which is an important tool in statistics. We formalize the above to higher dimensions below.

1.4.3 The variance matrix

The covariance between the random variables (X, Y) is a combination of its joint moments. Using the rules of expectations we have

$$\operatorname{cov}(X,Y) = \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

This represents a covariance in terms of its expectations. However, often it is easier to stick with the covariance and to deal with the covariance rather than turning it into an expectation (see Example 1.6).

As mentioned in the previous section, the covariance is a measure of linear dependence between X and Y. If two random variables are independent, then their covariance is zero. But the converse is not necessarily true.

Example 1.5. *Suppose X, Y, Z are zero mean independent random variables. Define the variables $U_1 = ZX$ and $U_2 = ZY$, clearly they are dependent. However,*

$$\operatorname{cov}(U_1, U_2) = 0.$$

See HW1.

The variance is a special case of the covariance

$$\operatorname{cov}(X, X) = \operatorname{var}(X) = \mathrm{E}(X - \mathrm{E}(X))^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2.$$

Generalising the above to matrices, the variance of the vector $\underline{X} = (X_1, \ldots, X_d)$ is the pairwise covariance between each element in the vector

$$\operatorname{var}(\underline{X}) = \mathrm{E}\left((\underline{X} - \mathrm{E}(\underline{X}))(\underline{X} - \mathrm{E}(\underline{X}))'\right) = \begin{pmatrix} \operatorname{var}(X_1) & \operatorname{cov}(X_1, X_2) & \ldots & \operatorname{cov}(X_1, X_d) \\ \operatorname{cov}(X_2, X_1) & \operatorname{var}(X_2) & \ldots & \operatorname{cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_d, X_1) & \operatorname{cov}(X_d, X_2) & \ldots & \operatorname{var}(X_d) \end{pmatrix}.$$

Clearly $\operatorname{var}(\underline{X})$ is a square symmetric matrix: since

$$\operatorname{cov}(X_i, X_j) = \mathrm{E}((X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))) = \operatorname{cov}(X_j, X_i).$$

In general we can define the covariance between two vectors as follows. Suppose $\underline{X} = (X_1, \ldots, X_p)$ and $\underline{Y} = (Y_1, \ldots, Y_q)$. Then $\operatorname{cov}(\underline{X}, \underline{Y})$ is a $(p \times q)$-matrix where

$$\operatorname{cov}(\underline{X}, \underline{Y}) = \begin{pmatrix} \operatorname{cov}(X_1, Y_1) & \operatorname{cov}(X_1, Y_2) & \ldots & \operatorname{cov}(X_1, Y_q) \\ \operatorname{cov}(X_2, Y_1) & \operatorname{cov}(X_2, Y_2) & \ldots & \operatorname{cov}(X_2, Y_q) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p, Y_1) & \operatorname{cov}(X_p, Y_2) & \ldots & \operatorname{cov}(X_p, Y_q) \end{pmatrix}.$$

Example 1.6. *(i) Suppose U, V, W are iid random variables with mean zero and variance one. Define the random vector*

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 2U + V \\ V + W \\ 3W \end{pmatrix}.$$

The expectation of $\underline{X}$ *is* $\mathrm{E}[\underline{X}] = (0,0,0)'$ *and pairwise covariance is*

$$\begin{aligned}
\mathrm{var}(X_1) &= \mathrm{cov}(2U+V, 2U+V) = \mathrm{cov}(2U,2U) + \mathrm{cov}(2U,V) + \mathrm{cov}(V,2U) + \mathrm{cov}(V,V) \\
&= 4\mathrm{var}(U) + 4\mathrm{cov}(U,V) + \mathrm{var}(U) = 4 + 4\times 0 + 1 = 5 \\
\mathrm{cov}(X_1,X_2) &= \mathrm{cov}(2U+V, V+W) = \mathrm{cov}(2U,V) + \mathrm{cov}(2U,W) + \mathrm{cov}(V,W) + \mathrm{cov}(V,V) = \mathrm{var}(V) = 1.
\end{aligned}$$

Applying the above technique to all combinations gives the variance matrix

$$var[\underline{X}] = \mathrm{var}\begin{pmatrix} 2U+V \\ V+W \\ 3W \end{pmatrix} = \begin{pmatrix} 5 & 1 & 0 \\ 1 & 2 & 3 \\ 0 & 3 & 9 \end{pmatrix}.$$

Observe the zero entries in the matrix. This tells us there is no correlation between X_1 *and* X_3*; which is clear as they do not share a common random variable. The pairwise dependence structure can be illustrated using the following network plot:*

(ii) Suppose U, V, W, Z *are iid random variables with mean zero and variance one. Define the random vector*

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} 2U+V \\ V+W \\ 3W+Z \\ Z+U \end{pmatrix}.$$

The expectation of $\underline{X}$ *is* $\mathrm{E}[\underline{X}] = (0,0,0,0)'$ *and variance*

$$var[\underline{X}] = \mathrm{var}\begin{pmatrix} 2U+V \\ V+W \\ 3W+Z \\ Z+U \end{pmatrix} = \begin{pmatrix} 5 & 1 & 0 & 2 \\ 1 & 2 & 3 & 0 \\ 0 & 3 & 10 & 1 \\ 2 & 0 & 1 & 2 \end{pmatrix}.$$

The pairwise dependence structure can be illustrated using the following network plot:

1.4.4 Properties of the variance

We summarize all properties.

Example 1.7. *Suppose* $\mathrm{E}((X_1, X_2)) = (\mu_1, \mu_2)$ *and*

$$\mathrm{var}((X_1, X_2)) = \begin{pmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_2, X_1) & \mathrm{cov}(X_2, X_2) \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Define the new random variable $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$, *then*

$$\mathrm{E}(Y) = \alpha_0 + \alpha_1 \mu_1 + \alpha_2 \mu_2$$

and

$$\begin{aligned} \mathrm{var}(Y) &= \alpha_1^2 \sigma_{1,1} + \alpha_1 \alpha_2 \sigma_{12} + \alpha_1 \alpha_2 \sigma_{21} + \alpha_2^2 \sigma_{22} = \alpha_1^2 \sigma_{11} + 2\alpha_1 \alpha_2 \sigma_{12} + \alpha_2^2 \sigma_{22} \\ &= (\alpha_1, \alpha_2) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}. \end{aligned}$$

We summarize below some properties of expectations and covariances of vectors which we will use:

1. Suppose $Y = aX + b$. Then $\mathrm{E}[Y] = a\mathrm{E}[X] + b$ and $\mathrm{var}(aX + b) = a^2 \mathrm{var}(X)$.

   Remember A shift of b has an impact on the mean but not on the variance (spread) of the transformed random variable.

2. Suppose $Y = \sum_{i=1}^n a_i X_i$ (where X_i are random an a_is are constant). Then $\mathrm{E}[Y] = \sum_{i=1}^n a_i \mathrm{E}[X_i]$ and

$$\mathrm{var}[Y] = \sum_{i,j=1}^{n} a_i a_j \mathrm{cov}(X_i, X_j). \tag{1.6}$$

3. Suppose that $\underline{X}' = (X_1, \ldots, X_d)$, $\underline{b} = (b_1, \ldots, b_n)'$ and A is a $n \times d$ matrix ($\underline{X}$ is a random vector, $\underline{b}$ and A is a constant vector and matrix).

   Let $\underline{Y} = (Y_1, \ldots, Y_n)' = A\underline{X} + \underline{b}$. Then we have $\mathrm{E}(\underline{Y}) = A\mathrm{E}(\underline{X}) + \underline{b}$ and

$$\mathrm{var}(\underline{Y}) = \mathrm{var}\left(A\underline{X}\right) = A\mathrm{var}(\underline{X})A'. \tag{1.7}$$

   To understand why the above holds, assume $\mathrm{E}[\underline{X}] = 0$. Then by using Example 1.3 we have

$$\begin{aligned} \mathrm{var}(\underline{Y}) &= \mathrm{E}\left(\underline{Y}\underline{Y}'\right) = \mathrm{E}\left[\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \begin{pmatrix} Y_1 & Y_2 & \ldots & Y_n \end{pmatrix} \right] = \mathrm{E}\begin{pmatrix} Y_1^2 & Y_1 Y_2 & \ldots & Y_1 Y_n \\ Y_2 Y_1 & Y_2^2 & \ldots & Y_2 Y_n \\ \vdots & \vdots & \ddots & \vdots \\ Y_n Y_1 & Y_n Y_1 & \ldots & Y_n^2 \end{pmatrix} \\ &= \mathrm{E}[A\underline{X}\underline{X}'A'] = A\mathrm{E}[\underline{X}\underline{X}']A' = A\mathrm{var}(\underline{X})A'. \end{aligned}$$

4. Define the random vectors $\underline{X}' = (X_1, \ldots, X_p)$ and $\underline{Y}' = (Y_1, \ldots, Y_q)$. Suppose A and B are two constant matrices. Then

$$\mathrm{cov}\left(A\underline{X}, B\underline{Y}\right) = A\mathrm{cov}(\underline{X}, \underline{Y})B'.$$

Example 1.8. *[The sample mean]*

Define

$$\bar{X}_n = n^{-1}(X_1 + \ldots + X_n).$$

Suppose that $\{X_i\}$ *are iid random variables with mean* μ *and variance* σ^2*. Then we have*

$$\mathrm{E}(\bar{X}_n) = n^{-1}[\mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n)] = \mu.$$

And the variance is

$$\begin{aligned} \mathrm{var}(\bar{X}_n) &= \mathrm{cov}\left(n^{-1}\sum_{i=1}^{n} X_i, n^{-1}\sum_{i=1}^{n} X_i\right) \\ &= n^{-2}\sum_{i_1,i_2=1}^{n} \underbrace{\mathrm{cov}(X_{i_1}, X_{i_2})}_{=0 \text{ if } i_1 \neq i_2} = n^{-2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{\sigma^2}{n}. \end{aligned}$$

Note the above can be shown by using rules of variances of sums (1.6). Alternatively for those who like matrices we observe that

$$\bar{X}_n = (1/n, 1/n, \ldots, 1/n)\underline{X}_n = n^{-1}\underline{1}\underline{X}_n,$$

where $\underline{X}_n' = (X_1, \ldots, X_n)$*. Then by using (1.7) we have*

$$\mathrm{var}(\bar{X}_n) = (1/n, 1/n, \ldots, 1/n)\mathrm{var}(X_n)\begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix} = n^{-2}\underline{1}\mathrm{var}(X_n)\underline{1}'.$$

By expanding out we can see that $\underline{1}\mathrm{var}(\underline{X}_n)\underline{1}'$ *is the sum of all entries in matrix* $\mathrm{var}(\underline{X}_n)$

$$\underline{1}\mathrm{var}(\underline{X}_n)\underline{1}' = \sum_{i_1,i_2=1}^{n}\mathrm{cov}(X_{i_1}, X_{i_2}) = \sum_{i=1}^{n}\mathrm{var}(X_i).$$

This gives an alternative derivation for the same result. They are both the same, choose the method which suits you best.

Observe where we required the assumption of independence. We required the assumption of independence (or at least no correlation) to set $\mathrm{cov}(X_{i_1}, X_{i_2}) = 0$ *when* $i_1 \neq i_2$*. If there is dependence, this may not hold (see HW1)!*

1.5 Modes of convergence

Convergence of the sample mean is a complex idea, and there are several ways of measuring it. In this section we review some standard measures.

But why do we care? Convergence matters, because we never (rarely; an exception is the 2020 census) observe the population, we can never observe the population parameter. We can only come up with some estimator of it based on a sample. But how do we know this estimator is any good? That it even gets "close" to the true parameter for a very large sample size. To answer this question we need to study different modes of convergence. In this section, we focus on the sample mean (as you would have studied it in previous classes). In subsequent chapters we consider more sophisticated estimators, but the basic ideas are the same (indeed most estimator can be written as a type of sample mean).

First recall that sample mean is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where we assume that $\{X_i\}$ are iid random variables with mean μ and variance σ^2. It can be shown with the exception of the most extraordinary situations (by extraordinary we mean a set which has measure zero, it rarely happens), $\bar{X}_n \to \mu$, this is called almost sure convergence. This means as the sample size grow it will get closer and closer to the population mean μ. Though useful, it is not very informative.

To illustrate the ideas in this section we consider a running example. We simulate from an chi-square distribution (we define this formally in the next chapter, however, the type of distribution does not impact the discussion in this section) with one degree of freedom, this means $\mathrm{E}(X) = \mu = 1$ and $\mathrm{var}(X) = \sigma^2 = 2$. For the sample sizes $n = 1, \ldots, 500$ we evaluate the sample mean. Thus for each realisation, we have a trajectory of the sample means from $n = 1$ to $n = 500$. In Figure 1.4 we give a plot of five trajectories; each coloured line corresponds to a specific j where

$$\bar{x}_{j,n} = \frac{1}{n}\sum_{i=1}^{n} x_{j,i} \qquad n = 1, \ldots, 500$$

We observe that each trajectory appears to approach one as n grows. Keep in mind, for a given data set $\{x_{j,i}\}_{i=1}^{n}$, we can only observe one $\bar{x}_{j,n} = n^{-1}\sum_{i=1}^{n} x_{j,i}$ (for example, for the sample size $n = 20$, it may be the red curve at $n = 20$). Since μ is unknown), we do not know how close $\bar{x}_{j,n}$ is μ. In practice we will never know the difference $(\bar{x}_{j,n} - \mu)$. But there are various ways of measuring the "typical" behaviour of $\bar{X}_n$. We describe them in the sections below.

1.5.1 The mean squared error

One measure is the mean squared distance. By mean squared error we mean the average squared distance between each trajectory (for a fixed n) and the mean μ, where the average is taken over all possible realisations. In Figure 1.5 we give the trajectories of 100 realisations (from $n = 1, \ldots, 500$), we denote each realisation as $\bar{x}_{j,n} = n^{-1}\sum_{i=1}^{n} x_{i,j}$. An estimate the mean squared error at sample size n we calculate:

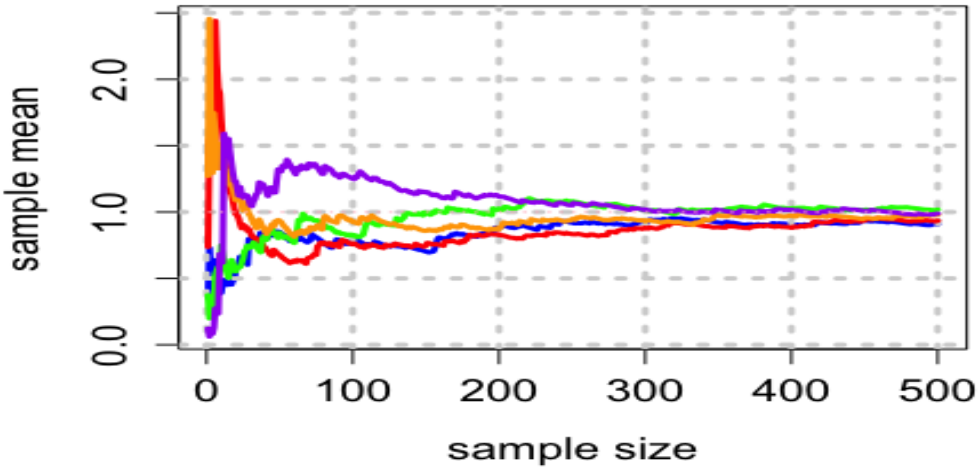


Figure 1.4: Trajectories of five different sample means for sample sizes n=1,...,500.

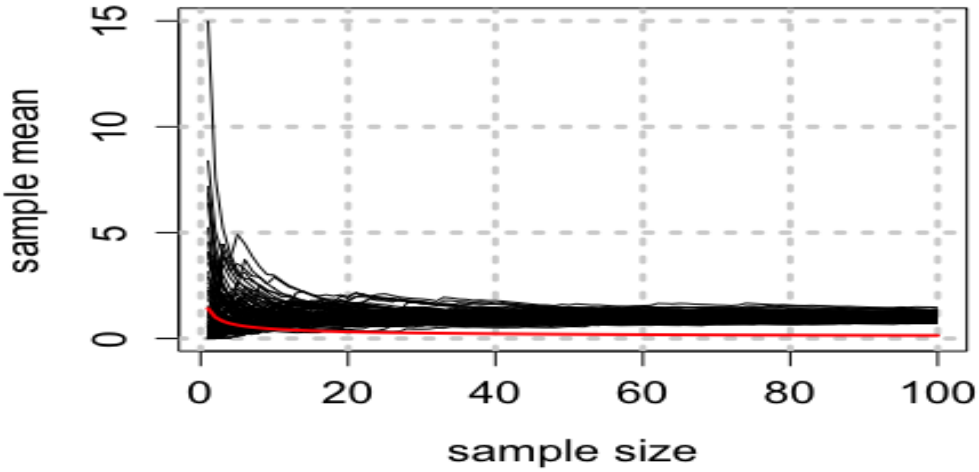


Figure 1.5: 100 different trajectories for sample sizes n=1,...,100. In red is the standard error $\sqrt{2/n}$ of the sample mean.

$$\frac{1}{100}\sum_{i=1}^{100}(\bar{x}_{j,n}-\mu)^2,$$

where $\bar{x}_{i,n}$ denotes the jth trajectory of the plot at sample size n and $\mu = 1$. A plot of these average squared error is given in Figure 1.6 However, the true mean squared error should be over all realisations. We recall,

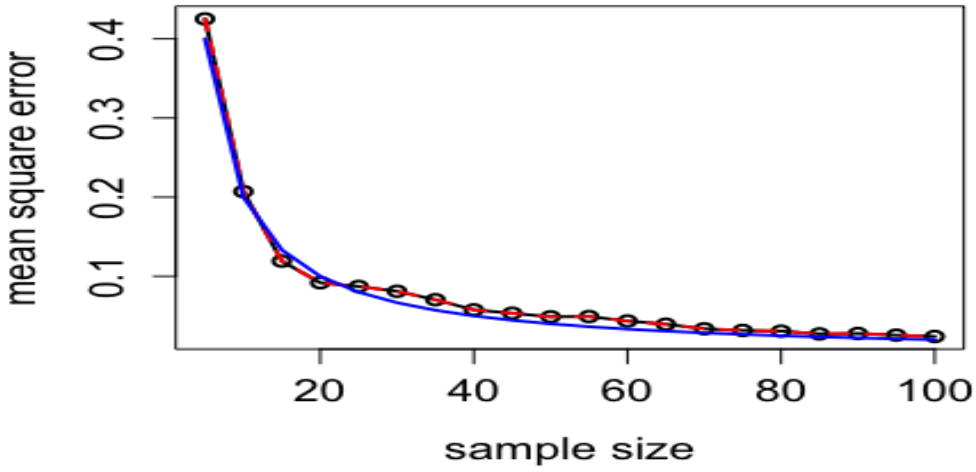


Figure 1.6: The average squared error evaluated over 100 different realisations (dots and red) and the mean squared error 2/100 (blue).

that this is simply the expectation of the squared difference $(\bar{X}_n - \mu)$:

$$\mathrm{E}\left(\bar{X}_n - \mu\right)^2.$$

By simply expanding the expectation it can be shown that

$$\begin{aligned}
\mathrm{E}\left(\bar{X}_n - \mu\right)^2 &= \mathrm{E}\left(\bar{X}_n - \mathrm{E}(\bar{X}_n) + \mathrm{E}(\bar{X}_n) - \mu\right)^2 \\
&= \mathrm{E}\left(\bar{X}_n - \mathrm{E}(\bar{X}_n)\right)^2 + 2\underbrace{\mathrm{E}\left(\bar{X}_n - \mathrm{E}(\bar{X}_n)\right)}_{=0}\mathrm{E}\left(\mathrm{E}(\bar{X}_n) - \mu\right) + \mathrm{E}\left(\mathrm{E}(\bar{X}_n) - \mu\right)^2 \\
&= \mathrm{E}\left(\bar{X}_n - \mathrm{E}(\bar{X}_n)\right)^2 + \left(\mathrm{E}(\bar{X}_n) - \mu\right)^2 = \underbrace{\mathrm{var}(\bar{X}_n)}_{\text{variance}} + \underbrace{\left(\mathrm{E}(\bar{X}_n) - \mu\right)^2}_{\text{bias squared}}.
\end{aligned} \tag{1.8}$$

This is the "classical" decomposition of the mean squared error of an estimator in terms of its variance and its bias squared.

Definition 1.6 (Bias and standard error)**.** *Suppose $\widehat{\theta}_n$ is an estimator of a parameter θ. The bias of $\widehat{\theta}_n$ is defined as*

$$B_\theta(\widehat{\theta}_n) = (\mathrm{E}[\widehat{\theta}_n] - \theta)$$

and the standard error defined as the square root of the variance of the estimator:

$$s.e(\widehat{\theta}_n) = \sqrt{\mathrm{var}(\widehat{\theta}_n)}.$$

The mean squared error is

$$\mathrm{E}\left(\widehat{\theta}_n - \theta\right)^2 = \mathrm{var}(\widehat{\theta}_n) + B_\theta(\widehat{\theta}_n)^2.$$

We proved the above mean square error decomposition in (1.8).

Example 1.9. *Suppose $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are two unbiased estimators of θ. Show that for all $0 \leq a \leq 1$, $\widehat{\theta}_3 = a\widehat{\theta}_1 + (1-a)\widehat{\theta}_2$ is an unbiased estimator of θ. Assume that $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent, with $\mathrm{var}(\hat{\theta}_1) = \sigma_1^2$ and $\mathrm{var}(\hat{\theta}_2) = \sigma_2^2$. How should the constant a be chosen in order to minimize the variance of $\hat{\theta}_3$?*

Solution Just by taking expectations we have

$$\mathrm{E}[\widehat{\theta}_3] = a\mathrm{E}[\widehat{\theta}_1] + (1-a)\mathrm{E}[\widehat{\theta}_2] = a\theta + (1-a)\theta = \theta.$$

Under the assumption of independence we have

$$\mathrm{var}[\widehat{\theta}_3] = a^2\sigma_1^2 + (1-a)^2\sigma_2^2$$

To minimise the above take derivatives wrt a and set to zero

$$\frac{df(a)}{a} = 2a\sigma_1^2 - 2(1-a)\sigma_2^2 = 0.$$

This gives $a = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}$. It needs to be checked this is the minimum (and not maximum).

Definition 1.7 (A means squared consistent estimator)**.** *An estimator, $\widehat{\theta}_n$ is mean square consistent estimator of θ, if*

$$\mathrm{E}\left[\widehat{\theta}_n - \theta\right]^2 \to 0$$

as $n \to \infty$.

For our example, where we consider the sample mean $\mathrm{E}(\bar{X}_n) = \mu$ (see Example 1.8), the bias is zero for all sample sizes and $\mathrm{E}\left(\bar{X}_n - \mu\right)^2 = \mathrm{var}(\bar{X}_n) = \sigma^2/n$ (also from Example 1.8). The the average squared errors given in Figure 1.6 is a good approximation of $2/n$ (since $\sigma^2 = 2$); compare the red and blue lines. To summarise, the mean squared error gives the average squared distance from the estimator to the population parameter. It is not an asymptotic result, i.e. it holds for any sample size. However, it does not give any guarantees on the proportion of realisations which are within, say two standard errors (square root of the MSE if the bias is zero) of the population parameter. For that we need to know the distribution of the estimator. But first we describe convergence in probability.

1.5.2 Convergence in probability

Convergence in probability is evaluated at every n. Roughly speaking it is the "proportion" of trajectories $\bar{X}_n$, at sample size n, which deviates from μ by more than ε. If this proportion converges to zero for every ε as $n \to \infty$, then the estimator converges in probability to μ. Formally, if for every $\varepsilon > 0$ we have

$$P(|\bar{X}_n - \mu| > \varepsilon) \to 0$$

as $n \to \infty$. Then we say $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$ as $n \to \infty$.

Convergence in probability is a weaker form of convergence than almost sure convergence and convergence in mean square. This means almost sure convergence and convergence in mean square imply convergence in probability, but the converse is not true. Though this is very important, it is something that we do not worry too much about in this class. But we should keep in mind that there exists strange examples, where convergence in probability can occur but not almost sure convergence. That is, examples where, as n grows, collectively the group of trajectories become tightly gathered about the μ (increasingly bunched together). But individually, many trajectory on an infinite number of occasions deviates far from μ. Thus, individually these trajectories do not converge to μ (so no almost sure convergence). It is difficult to illustrate. But we make an attempt in Figure 1.7; we observe that in general all the trajectories are congregating about the mean, $\mu = 1$, but there are excursions; each individually trajectory is not converging to $\mu = 1$.

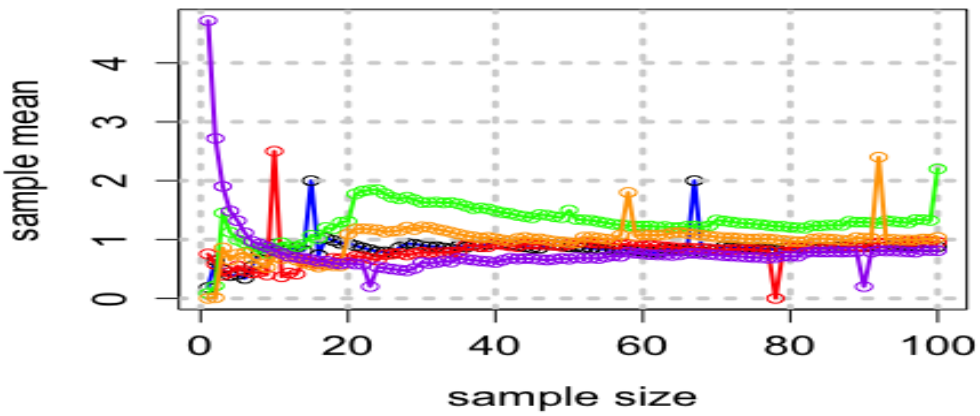


Figure 1.7: Trajectories of five sample means (observe the individual excursions away from one).

1.5.3 Sampling distributions and the central limit theorem

We return to the trajectories in Figure 1.5. For sample sizes $n = 1, 5, 10$ and 100 we make a histogram of

$$\bar{X} \quad \text{and} \quad \sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma}$$

together with a QQplot of $\sqrt{n}\frac{(\bar{X}_n-\mu)}{\sigma}$ against the quantiles of a standard normal distribution. The reason we consider the "z-transform" $\sqrt{n}\frac{(\bar{X}_n-\mu)}{\sigma}$ (recall we make this transform when looking up the z-tables) is that it mean zero and variance one (it standardizes $\bar{X}_n$). This is is similar to taking a cross section across Figure 1.5 $n = 1, 5, 10$ and 100 and studying the distribution of the trajectories at each of these intersections. The plots are given in Figures 1.8 - 1.11.

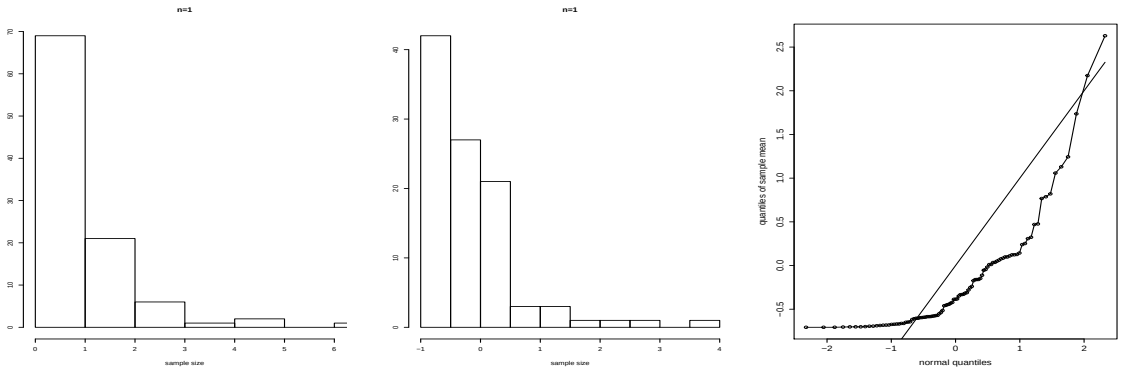


Figure 1.8: Sample size $n = 1$. Histogram of X, $(X-1)/\sqrt{2}$ and the QQplot against a standard normal distribution. Using 100 replications.

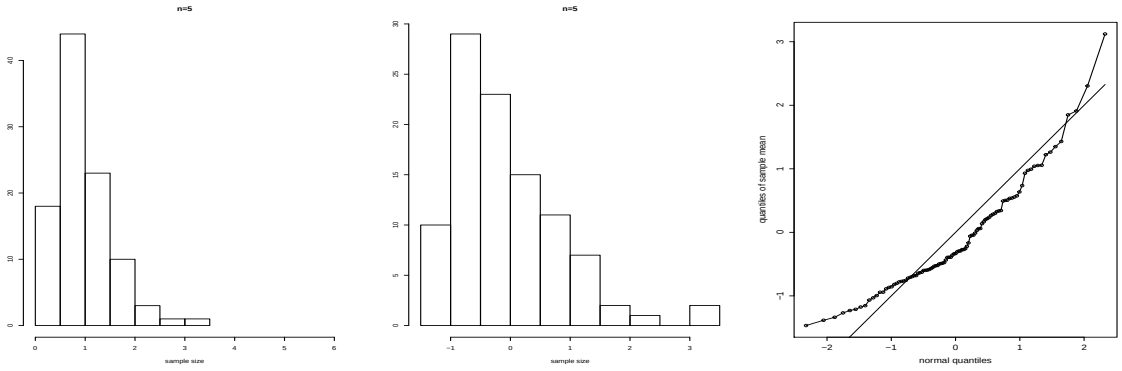


Figure 1.9: Sample size $n = 5$. Histogram of $\bar{X}_5$, $\sqrt{5}(\bar{X}_5 - 1)/\sqrt{2}$ and the QQplot against a standard normal distribution. Using 100 replications.

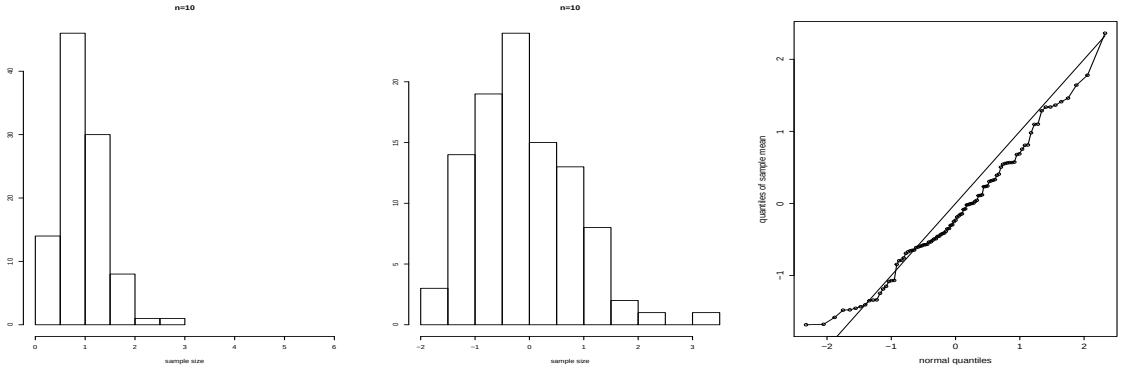


Figure 1.10: Sample size $n = 10$. Histogram of $\bar{X}_{10}$, $\sqrt{10}(\bar{X}_{10} - 1)/\sqrt{2}$ and the QQplot against a standard normal distribution. Using 100 replications.

Definition 1.8 (Sampling distribution of an estimator)**.** *The distribution of an estimator is called the sampling distribution of the estimator. For the $\bar{X}_n$ described in our running sample, the sampling distribution of $\bar{X}_n$ are*

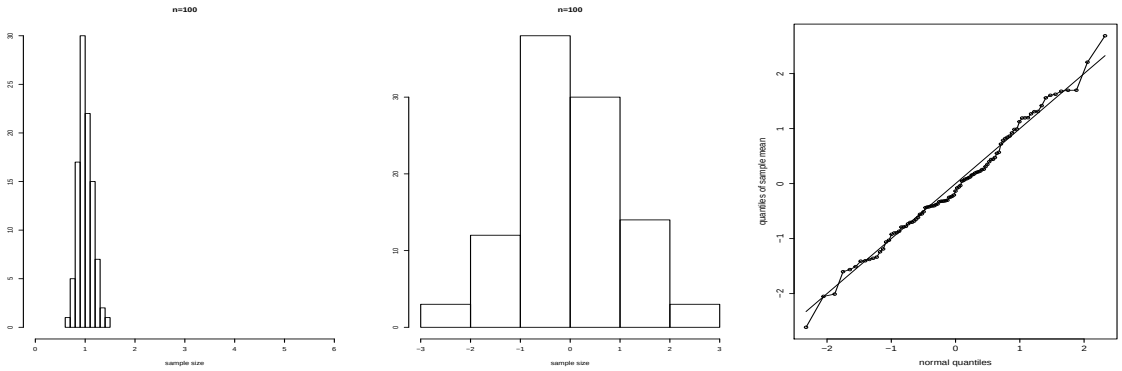


Figure 1.11: Sample size $n = 100$. Histogram of $\bar{X}_{100}$, $\sqrt{100}(\bar{X}_{100} - 1)/\sqrt{2}$ and the QQplot against a standard normal distribution. Using 100 replications.

the histograms given in Figures 1.8-1.11.

Definition 1.9 (Quantile Quantile plot)**.** *A Quantile Quantile plot (QQplot for short) is a useful method for checking if the data can plausibly come from a conjectured distribution. It plots the ordered data against the corresponding quantiles of the corresponding distribution. In Figure 1.8-1.11, we have plotted ordered sample means (for a given sample size) against the quantiles of a standard normal distribution.*

More precisely, suppose we observe the data $Y_1, Y_2, \ldots, Y_n$ (these could be raw data or several averages, as given in this example). We order the data from the smallest number to the largest, often denoted $Y_{(1,n)}, Y_{(2,n)}, \ldots, Y_{(n,n)}$. If $\{Y_i\}$ came from a standard normal distribution we would expect the median of the data, $Y_{(n/2,n)}$ to closely match the 50% (median) quantile in the standard normal distribution (which is zero). Similarly, we would expect $Y_{(n/4,n)}$ to close match the first quartile of a standard normal (which is -0.674, you can get these numbers from the z-tables) and $Y_{(3n/4,n)}$ to close match the third quartile of a standard normal (which is 0.674).

Extending this argument, we would expect that $Y_{(i,n)}$ roughly matches the quantile corresponding the probability i/n in the normal distribution. Based on this argument, we define the n quantiles in a standard normal distribution. Let $z_{i,n}$ be such that

$$P(Z \leq z_{i,n}) = \frac{(i - 0.5)}{n},$$

where Z is a standard normal distribution (mean zero and variance one). We subtract use $(i - 0.5)/n$ rather than i/n to avoid the case $P(Z \leq z_{n,n}) = 1$ (when $i = n$). $\{z_{i,n}\}$ are called the standard normal quantiles (you can find them in the z-tables). A standard normal QQplot is a plot of $\{(z_{i,n}, Y_{(i,n)})\}_{i=1}^{n}$. The line is usually (but not always) the line corresponding to $(z_{i,n}, z_{i,n})$. If the data is normal, it will the QQplot will lie close to the line.

In R, the function qq*plot plots the data against the quantiles of the normal distribution whose mean and variance are the sample mean and variance calculated from the data. The function* qq*lines makes a line which goes through the 25th and 75th quantiles of the standard normal distribution and corresponding data.*

Studying the plots from $n = 1, 5, 10$ and 100 we first observe that the histogram of the first plot becomes narrower as the sample size grows (this corresponds to the trajectories in Figure 1.5 getting increasingly bunched together). This is because the standard error $\sigma/\sqrt{n}$ gets smaller as the sample size grows. Further, the histograms tend to resemble a normal distribution as the sample size grow. Observe further, as the sample size grow the quantiles of standardized sample mean match well the quantiles of the standard normal distribution. This is called the central limit theorem, and we state this formally below. We mention that in our example, the distribution of the original data $\{X_i\}$ is skewed (see the plot in Figure 1.8, which is the histogram of X_i, since the sample size is $n = 1$). The skew in the original distribution means that it takes a larger sample size for the samling distribution of the sample mean to be close to normal. Observe that when $n = 10$, evidence of a skew is still seen in the QQplot, but it is no longer so evident for $n = 100$.

We state the central limit theorem in its simplest form.

Theorem 1.1

Suppose that $\{X_i\}$ are iid random variables with mean μ and variance σ^2 (a finite variance is a necessary condition). Let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Using the results in Section 1.4.1 we have

$$\mathrm{E}(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) = \mu \text{ and } \mathrm{var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Then

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) \qquad n \to \infty.$$

This means that the density (or histogram) associated with the random variable $\sqrt{n}\left(\bar{X}_n - \mu\right)$ gets more and more standard normal looking as the number of X_is used to construct the $\bar{X}_n$ increase. By dividing by σ and using the results in Section 1.4.1 we have

$$\mathrm{E}\left[\sqrt{n}\frac{\left(\bar{X}_n - \mu\right)}{\sigma}\right] = 0 \qquad var\left[\sqrt{n}\frac{\left(\bar{X}_n - \mu\right)}{\sigma}\right] = 1$$

and

$$\sqrt{n}\frac{\left(\bar{X}_n - \mu\right)}{\sigma} \xrightarrow{\mathcal{D}} N(0, 1)$$

as $n \to \infty$. Alternatively, if we want to apply the above results, then we can write the above result as

$$\bar{X}_n \xrightarrow{\mathcal{D}} N\left(\mu, \frac{\sigma^2}{n}\right).$$

This way of writing the result corresponds to the left hand side histogram in Figures 1.8-1.11.

However, the actual sample size required for the normal approximation to hold well depends on the characteristics of the distribution of X_i. The factor that plays the largest role is the skewness (asymmetry) of the original distribution. The greater the level of asymmetry in the density or pmf of X_i, the larger the sample size required for the normal approximation to hold. This effect is easily seen in simulations. Further, as can be seen from the QQplots, the deviance between the distribution of the sample mean and normal distribution differs greatest in the tails. The asymmetric is measured using skewness, which we define in the section below.

Definition 1.10 (Summary of different modes of convergence)**.** *(i) Almost sure convergence. This is where all the trajectories (except for the really weird and exceptional ones) converge to the mean, μ. In some sense this is the easiest to understand. We often denote this as $\bar{X}_n \overset{a.s.}{\to} \mu$.*

(ii) Mean squared convergence. This is essentially the average square distance between the trajectory (at sample size n) and μ. If the estimator converges in mean square then $\mathrm{E}(\bar{X}_n - \mu)^2 \to 0$ as $n \to \infty$.

(iii) Convergence in probability: If for every $\varepsilon > 0$ we have

$$P(|\bar{X}_n - \mu| > \varepsilon) \to 0$$

as $n \to \infty$. Then we say $\bar{X}_n \overset{\mathcal{P}}{\to} \mu$ as $n \to \infty$.

(iv) Convergence in distribution: we may write

$$\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} \overset{\mathcal{D}}{\to} \mathcal{N}(0,1)$$

or $\sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma \overset{\mathcal{D}}{\to} Z$, where Z is a standard normal random variable.

Skewness: Measure of asymmetry

Skewness is usually defined using the third moment

$$S_3 = \frac{\mathrm{E}(X_i - \mu)^3}{\sigma^3}$$

To understand why this measure asymmetry assume $\mu = 0$ (without loss of generality). It is easily seen if the the distribution is symmetric about the mean, then $S_3 = 0$;

$$\begin{aligned}
\mathrm{E}(X^3) &= \int_{-\infty}^{\infty} x^3 f(x)dx = \int_0^{\infty} x^3 f(x)dx + \int_{-\infty}^{0} x^3 f(x)dx \\
&= \int_{-\infty}^{\infty} x^3 f(x)dx = \int_0^{\infty} x^3 f(x)dx + \int_{-\infty}^{0} x^3 f(-x)dx \qquad \text{(change variables } x = -y\text{)} \\
&= \int_{-\infty}^{\infty} x^3 f(x)dx = \int_0^{\infty} x^3 f(x)dx - \int_0^{\infty} y^3 f(y)dy = 0,
\end{aligned}$$

essentially the positives in $(X_i - \mu)^3$ cancel with the negatives (this is best seen with a picture)[5] If the distribution is not symmetric then this cancellation may not be possible and S_3 may not be zero. It can be shown that the size of $|S_3|$ together with the sample size effects the quality of the normal approximation of the distribution of the sample mean. For the χ^2 distribution with m-df (we define it formally in the next chapter), $S_3 = \sqrt{8/m}$. Observe that the level of skewness decreases as m grows.

Suppose we observe the iid random variables $\{X_i\}_{i=1}^n$, we can estimate $\mathrm{E}(X - \mu)^3$ with the centralized average

$$\hat{\mu}_3 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^3$$

and the variance σ^2 with

$$\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

This gives an estimator of S_3

$$\widehat{S_3} = \frac{n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^3}{\widehat{\sigma}_n^3}.$$

1.5.4 Functions of sample means

In statistics many estimators we encounter are averages or functions of averages. Suppose $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$ and $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$, what happens to $\bar{X}_n^2$? By the continuous mapping theorem

$$\bar{X}_n^2 \xrightarrow{\mathcal{P}} \mu^2.$$

Further, if $\mu \neq 0$, then

$$\sqrt{n}(\bar{X}_n^2 - \mu^2) \xrightarrow{\mathcal{D}} N(0, 4\mu^2\sigma^2)$$

as $n \to \infty$.

In general, we have the following result (usually called the continuous mapping theorem).

Lemma 1.2

Suppose $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$ as $n \to \infty$ and g is a continuous function, then

$$g(\bar{X}_n) \xrightarrow{\mathcal{P}} g(\mu),$$

as $n \to \infty$. Furthermore, if $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$ and $g'(\mu) \neq 0$ then we have

$$\sqrt{n}[g(\bar{X}_n) - g(\mu)] \xrightarrow{\mathcal{D}} N(0, [g'(\mu)]^2\sigma^2) \tag{1.9}$$

[5] A simpler proof uses that $x^3 f(x)$ is an odd function and this integrates to zero.

as $n \to \infty$.

PROOF. The proof is beyond this course, but a rough outline of the normality result is given below. The second order mean value theorem of $g(\bar{X}_n)$ about $g(\mu)$ gives

$$g(\bar{X}_n) - g(\mu) = (\bar{X}_n - \mu)g'(\mu) + \frac{1}{2}(\bar{X}_n - \mu)^2 g''(\alpha\mu + (1-\alpha)\bar{X}_n).$$

Since $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$, $(\bar{X}_n - \mu)^2 << ((\bar{X}_n - \mu))$; thus the first term of the RHS of the above "dominates" the second term and we have

$$g(\bar{X}_n) - g(\mu) \approx (\bar{X}_n - \mu)g'(\mu).$$

Note this is why we require $g'(\mu) \neq 0$. Since we have asymptotic normality of $(\bar{X}_n - \mu)$ and $g'(\mu)$ is a constant, we have the result. □

This result turns out to be very useful in many of the methods we discuss in the subsequent chapters.

To illustrate the result, in Figures 1.12 and 1.13 we give a plot of the histogram of $\bar{X}_n^2$ (for two different sample sizes $n = 200$ and 1000, conducted over 1000 replications) in the case that $\mu = 0$ and $\mu = 0.5$. We recall that $g'(\mu) = 2\mu$) and that for (1.9) to hold, we require $g'(\mu) \neq 0$. If this condition is violated, as is the case that $\mu = 0$, then this result does not hold. This is clearly seen in Figure 1.12, for both the sample sizes $n = 200$ and $n = 1000$, the distributions is clearly not normal. On the other hand, when $\mu = 0.5$, then

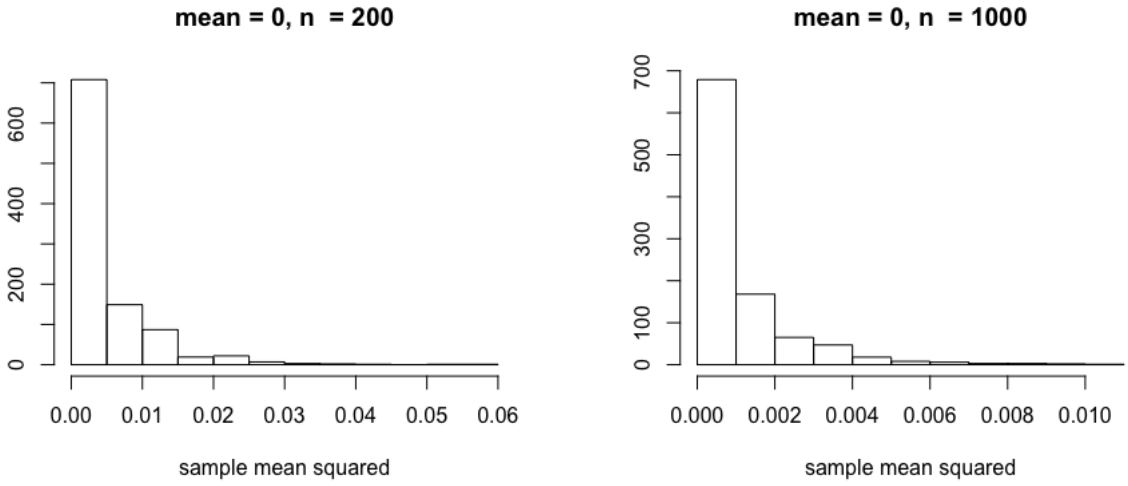


Figure 1.12: The histogram of $\bar{X}_n^2$ when $\mu = 0$: Left $n = 200$, Right: $n = 1000$.

$g'(0.5) \neq 0$, thus (1.9) should hold when n is large. In Figure 1.13 we see that this is indeed the case. For $n = 200$, there appears to be a small right skew, which appears to have almost diminished when $n = 1000$. This essentially illustrates the power of (1.9); a transformation of the sample mean is also asymptotically normal, so long as $g'(\mu) \neq 0$.

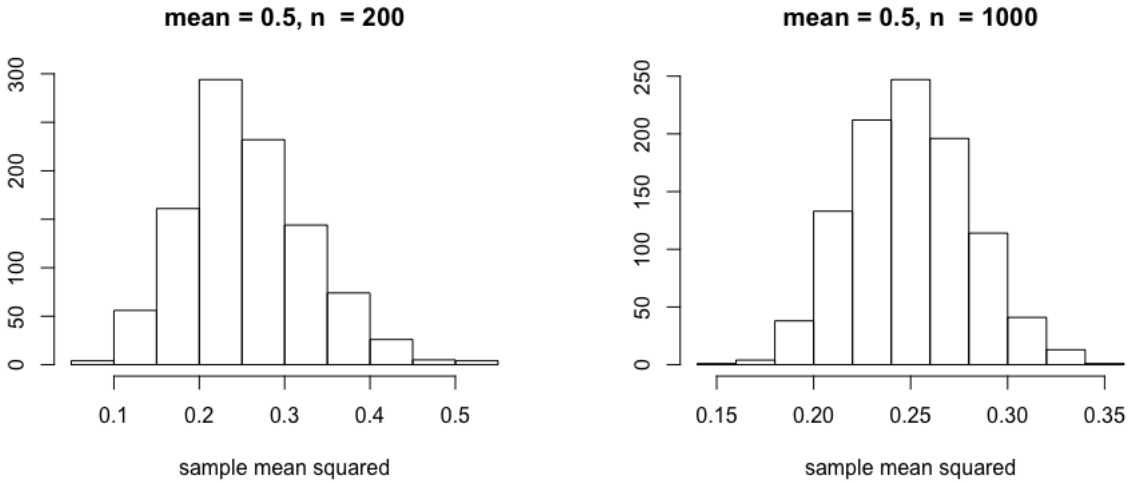


Figure 1.13: The histogram of $\bar{X}_n^2$ when $\mu = 0.5$: Left $n = 200$, Right: $n = 1000$.

Example 1.10 (Transformations to reduce skewness)**.** *In a simulation, I draw from a Poisson with $\lambda = 0.5$, 5 times and evaluate the sample mean (this is done several times). The histogram for the distribution of $\bar{X}_5$ is the top plot in Figure 1.14. Observe the huge skew. We now take a power transform; $\bar{X}_5^{\alpha}$ using $\alpha = 1/2$. The histogram of the lower plot in Figure 1.14. Observe that after taking the square root of the sample mean, it is normal looking. I found that taking a square root is a lot better than taking a cube root. This is a neat trick for making estimators more normal in their distribution. Note that from Lemma 1.2 we have*

$$\sqrt{n}(\bar{X}_n^{1/2} - \lambda^{1/2}) \xrightarrow{\mathcal{D}} N\left(0, \lambda \times [\lambda^{-1/2}/2]^2\right).$$

The above can be generalized to functions of several averages. Suppose $\{X_i\}$ and $\{Y_i\}$ are iid random variables with mean μ_X and μ_Y (respectively) and variance

$$\Sigma = \begin{pmatrix} var(X) & \text{cov}(X,Y) \\ \text{cov}(Y,X) & \text{var}(Y) \end{pmatrix}.$$

Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$. Suppose the multivariate CLT holds (see STAT414, for details) such that

$$\sqrt{n}\begin{pmatrix} \bar{X}_n - \mu_X \\ \bar{Y}_n - \mu_Y \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma).$$

Let $Z_n = g(\bar{X}_n, \bar{Y}_n)$, where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$. If

$$\underline{g}(\mu_1, \mu_2) = \left(\frac{\partial g(x,y)}{\partial x}, \frac{\partial g(x,y)}{\partial y}\right)\rfloor_{(x=\mu_x, y=\mu_y)} \neq 0,$$

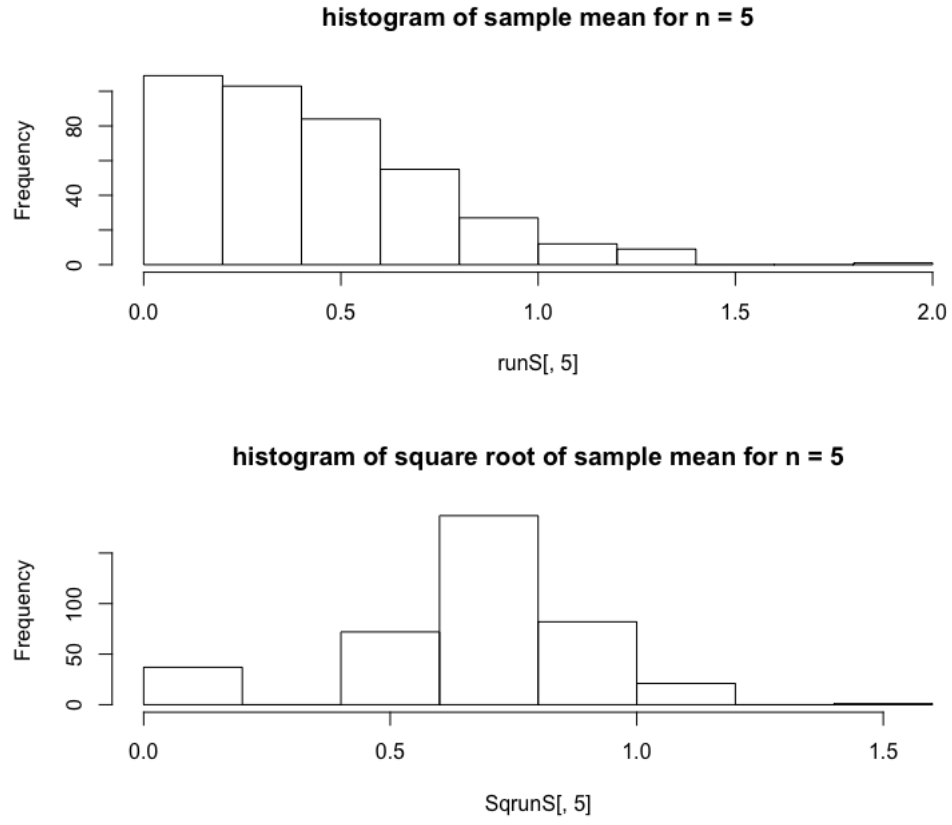


Figure 1.14: Top plot histogram of $\bar{X}_5$. Lower plot histogram of $\bar{X}_n^{1/2}$.

then we have

$$\sqrt{n}\left(g(\bar{X}_n, \bar{Y}_n) - g(\mu_X, \mu_Y)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \underline{g}(\mu_X, \mu_Y)\Sigma\underline{g}(\mu_X, \mu_Y)'\right).$$

Research 1. *Run some simulations for different functions of averages. Plot the histogram and calculate the standard deviations in the simulations. Do they match the results given above for sufficiently large n?*

1.6 A historical perspective

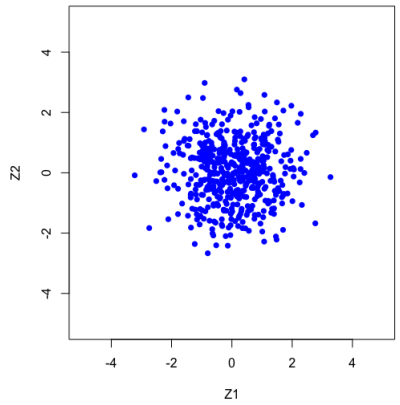
The Central Limit Theorem dates back to Laplace in 1810. In 1824, Poisson gave a more rigourous proof of the result. However, the first rigourous version of the proof was established by Lyapunouv in 1901. In 1922 Lindeburg, established the result under the weaker condition that only the first and second moments of the random variable are finite (this turns out to be sufficient and necessary condition).

2 Classical distributions and the first foray into sampling distributions

2.1 The Multivariate Gaussian distribution

2.1.1 Motivation through the bivariate Gaussian

Suppose that Z_1 and Z_2 are iid normal random variables with mean zero and variance one. Below is a plot of Z_1 against Z_2 over 500 replications.



If we made a histogram of the above, it would resemble the joint density of Z_1 and Z_2. Because Z_1 and Z_2 are independent, their joint density is the product of the marginals. The joint density is

$$\begin{aligned} f_{Z_1,Z_2}(z_1,z_2) &= f_{Z_1}(z_1)f_{Z_2}(z_2) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z_1^2\right)\times\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z_2^2\right) = \frac{1}{(2\pi)}\exp\left(-\frac{1}{2}(z_1^2+z_2^2)\right) \\ &= \frac{1}{(2\pi)}\exp\left(-\frac{1}{2}\begin{pmatrix} z_1 & z_2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right) \\ &= \frac{1}{(2\pi)}\exp\left(-\frac{1}{2}\begin{pmatrix} z_1 & z_2 \end{pmatrix} I_2^{-1}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right) \end{aligned} \tag{2.1}$$

The variance matrix of $\underline{Z} = (Z_1, Z_2)'$ is

$$\text{var}\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_2.$$

The appearance of $\text{var}[\underline{Z}] = I_2$ in f_{Z_1,Z_2} is not a coincidence. A perspective plot and contour plot is given in Figure 2.1. The perspective plot gives you an idea of which regions of (z_1, z_2) in $\mathbb{R}^2$ are more likely to arise. The contour plot is a "birds eye" view of the density (like a contour map); each line corresponds to where f_{Z_1,Z_2} is the same. From the perspective and contour may we observe that the density is completely symmetric.

Let λ_1 and λ_2 denote two constants (for the examples below we set $\lambda_1 = \sqrt{0.9}$ and $\lambda_2 = \sqrt{0.1}$). We transform (Z_1, Z_2) using the transformation

$$\begin{aligned} \underline{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &= \lambda_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} Z_1 + \lambda_1 \begin{pmatrix} 1 \\ -1 \end{pmatrix} Z_2 = \begin{pmatrix} \lambda_1 & \lambda_2 \\ \lambda_1 & -\lambda_2 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \\ &= A\underline{Z}. \end{aligned}$$

The variance (matrix) of $\underline{X}$ is

$$\text{var}[\underline{X}] = \begin{pmatrix} \lambda_1^2 + \lambda_2^2 & \lambda_1^2 - \lambda_2^2 \\ \lambda_1^2 - \lambda_2^2 & \lambda_1^2 + \lambda_2^2 \end{pmatrix} = AA'.$$

The correlation between (X_1, X_2) is

$$\text{cor}(X_1, X_2) = \frac{\lambda_1^2 - \lambda_2^2}{\lambda_1^2 + \lambda_2^2}.$$

A plot of X_1 against X_2 is given in Figure 2.2. We observe the alignment of the points have changed dramatically and appear to lie on an ellipse. To evaluate the bivariate density of $\underline{X}' = (X_1, X_2)$ we note that the iid random variables $\underline{Z}' = (Z_1, Z_2)$ can be rewritten in terms of $\underline{X}' = (X_1, X_2)$:

$$\underline{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & \lambda_2 \\ \lambda_1 & -\lambda_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = A^{-1}\underline{X}.$$

Using this, (2.1) and the change variables for multiple integrals (see your Calculus III book, Chapter 13.11) we can derive the joint density of $\underline{X}' = (X_1, X_2)$ (we do not give the details here, but it is straightforward). The joint density of (X_1, X_2) is

$$\begin{aligned} f_{X_1,X_2}(x_1, x_2) &= \frac{1}{|\det(A)|2\pi} \exp\left(-\frac{1}{2}\begin{pmatrix} X_1 & X_2 \end{pmatrix}\begin{pmatrix} \lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix}^{-1}\begin{pmatrix} \lambda_1 & \lambda_2 \\ \lambda_1 & -\lambda_2 \end{pmatrix}^{-1}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right) \\ &= \frac{1}{\sqrt{(2\pi)^2 \det(\Sigma)}} \exp\left(-\frac{1}{2}\underline{X}'\Sigma^{-1}\underline{X}\right), \end{aligned}$$

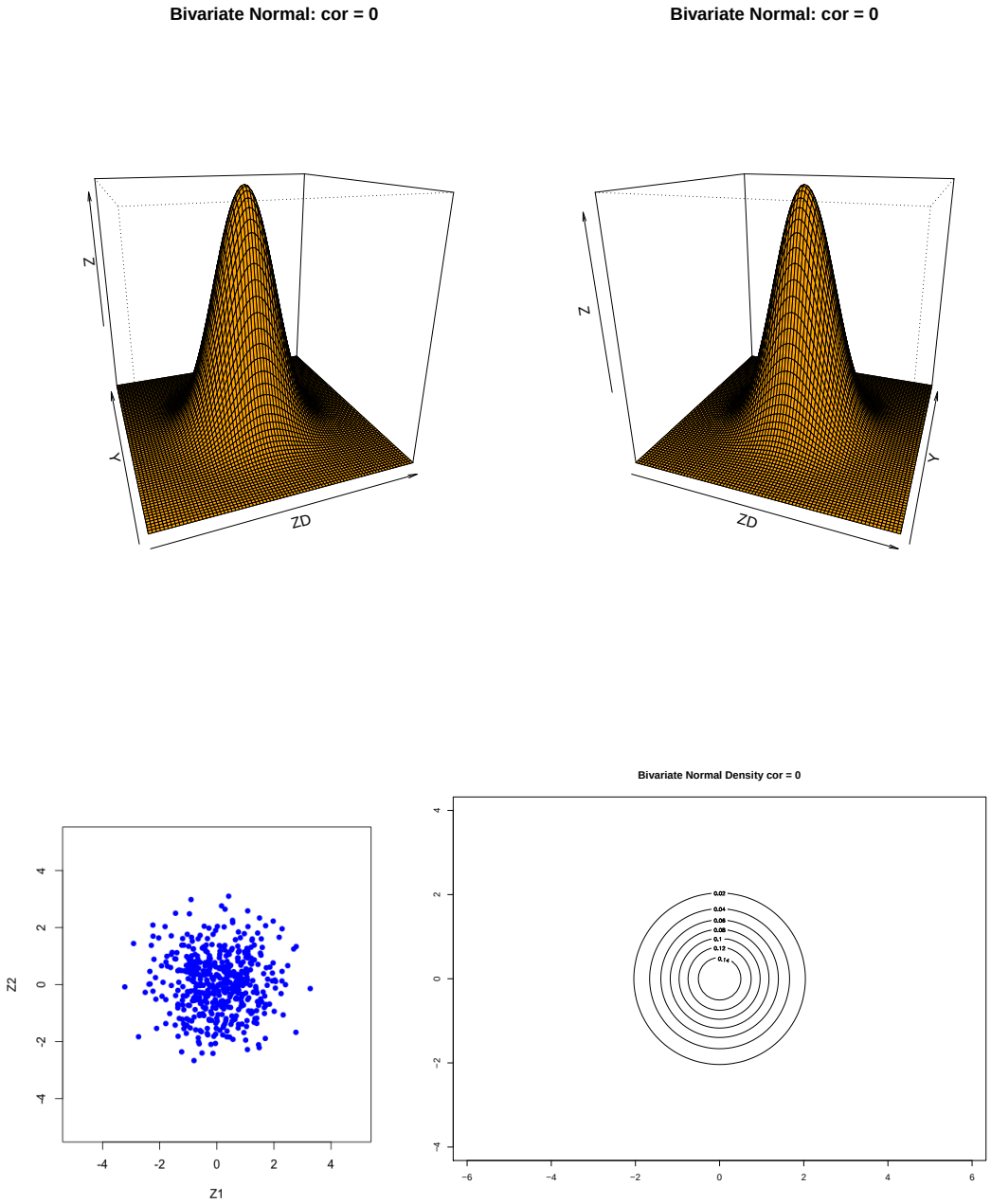


Figure 2.1: Perspective plot (at different angles) and contour map of the bivariate iid random variables.

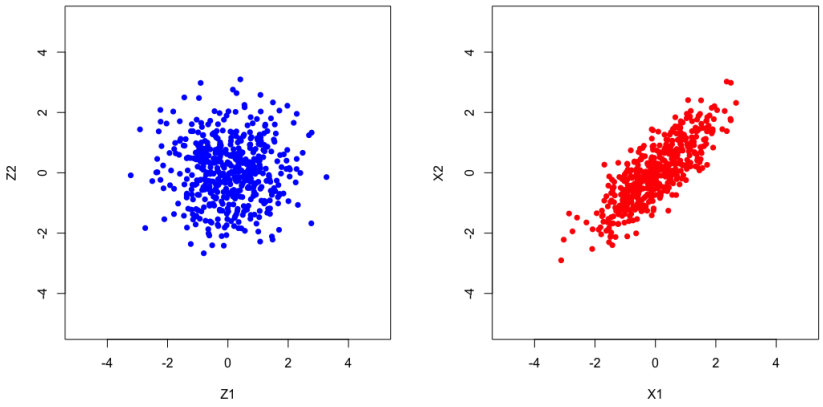


Figure 2.2: Left: Plot of (Z_1, Z_2) and Right: Plot of transformed $\underline{X} = A\underline{Z}$ (with $\lambda_1 = \sqrt{0.9}$ and $\lambda_2 = \sqrt{0.1}$).

since $\Sigma = AA'$. A perspective plot and contour plot for $\lambda_1 = \sqrt{0.9}$ and $\lambda_2 = \sqrt{0.1}$ is given in Figure 2.3. From the perspective and contour may we observe that the density is *not* symmetric at all rotations. The main spread is along the $X_1 = X_2$ line.

The plots in Figure 2.3 are for the case $\lambda_1 = \sqrt{0.9}$ and $\lambda_2 = \sqrt{0.1}$. In Figure 2.4 we give the analogous plots for $\lambda_1 = \sqrt{0.1}$ and $\lambda_2 = \sqrt{0.9}$. In general, any jointly bivariate Gaussian random variable has the joint density

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \det(\Sigma)}} \exp\left(-\frac{1}{2}\underline{X}'\Sigma^{-1}\underline{X}\right).$$

This derivations is based on the fact that any jointly normal random variable can be expressed as $\underline{X} = A\underline{Z}$ (where $\underline{Z} = (Z_1, Z_2)$ are iid standard normal random variables).

In the following section we generalize the joint density from (X_1, X_2) to the joint density of a d-dimensional multivariate Gaussian.

2.1.2 The general multivariate Gaussian

The random vector $\underline{X}_d' = (X_1, \ldots, X_d)$ is jointly normal (also called Gaussian, I switch between the two) with mean $\underline{\mu}'$ and variance Σ if the joint density of $\underline{X}_d$ (assuming Σ is invertible, that is there exists a unique matrix Σ^{-1} where $\Sigma\Sigma^{-1} = I_d$) is

$$f_{\underline{X}_d}(\underline{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})'\Sigma^{-1}(\underline{x} - \underline{\mu})\right)$$

where $\det(\Sigma)$ denotes the determinant of the matrix Σ (you will need to look it up, to find out what it exactly is, we rarely use it, and when we need it I will give you the pertinent results). I tend to be a sloppy and will

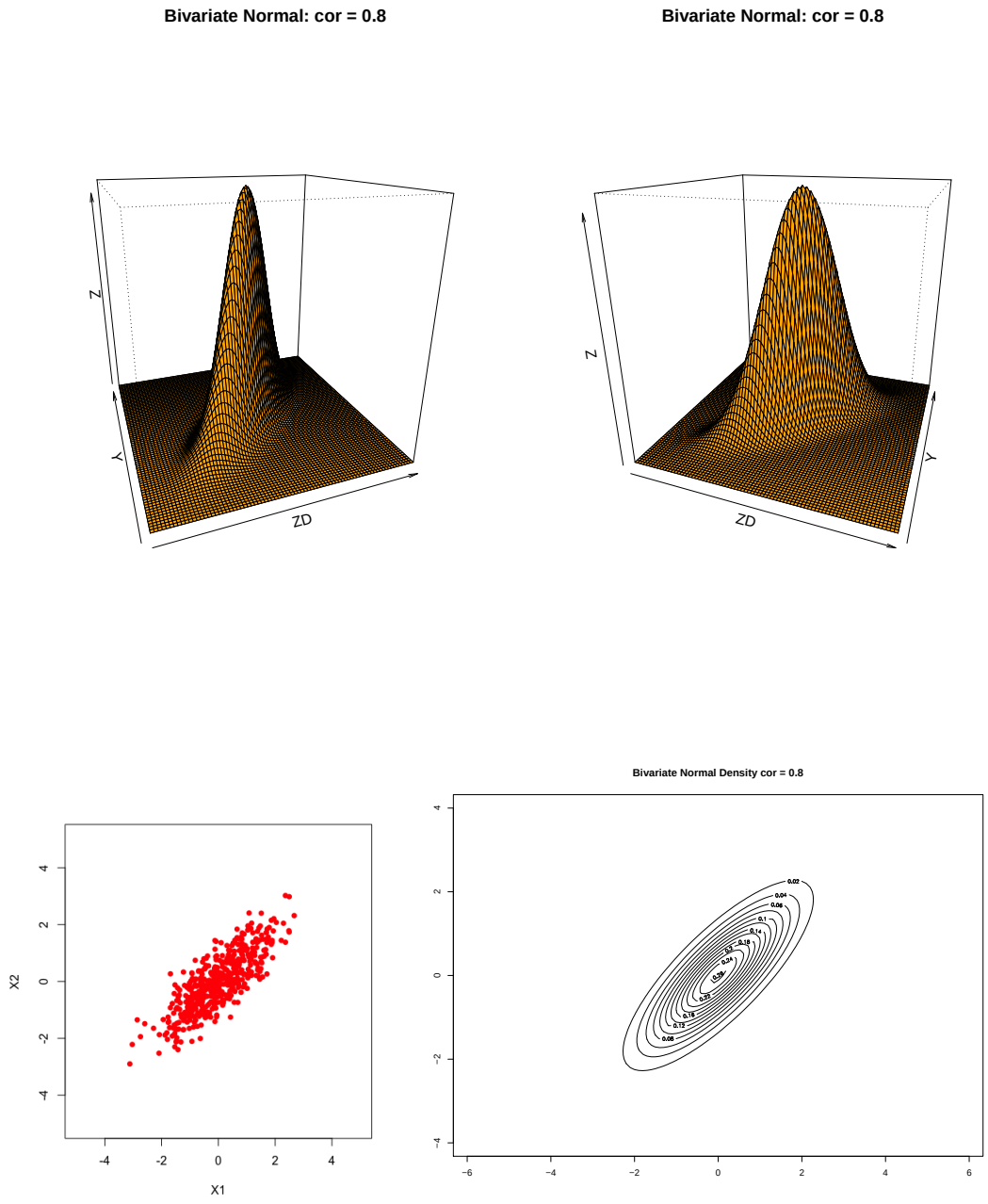


Figure 2.3: Perspective plot (at different angles) and contour map of the bivariate dependent random variables ($\lambda_1 = \sqrt{0.9}$ and $\lambda_2 = \sqrt{0.1}$).

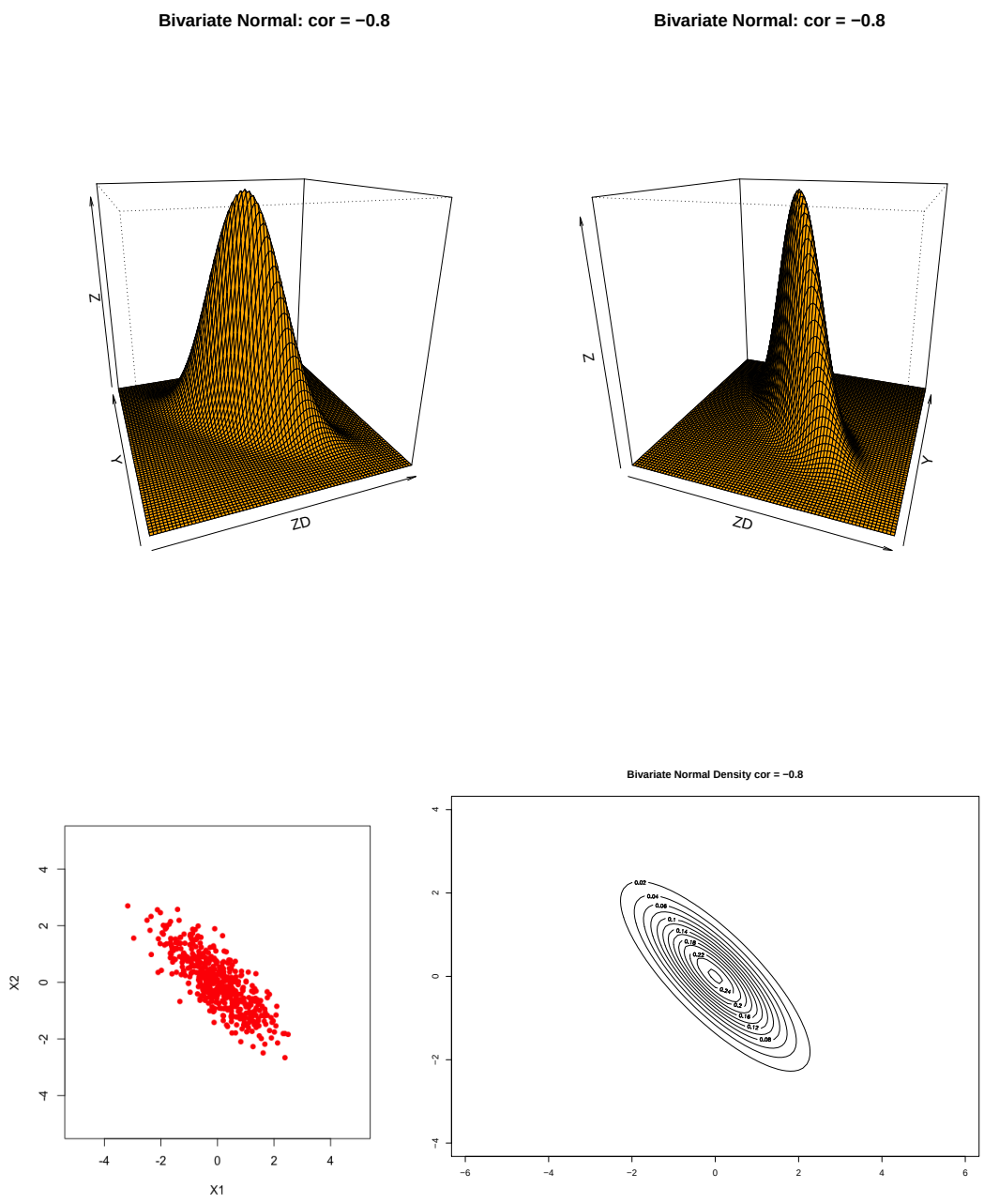


Figure 2.4: Perspective plot (at different angles) and contour map of the bivariate dependent random variables ($\lambda_1 = \sqrt{0.1}$ and $\lambda_2 = \sqrt{0.9}$).

switch between the notation $\det(\Sigma)$ and $|\Sigma|$ for determinant of a matrix. The inverse Σ^{-1} is a matrix where $\Sigma^{-1}\Sigma = \Sigma\Sigma^{-1} = I_d$ (generalisation of the inverse of a real number a). Often we denote the distribution of $\underline{X}_d$ as

$$\underline{X}_d \sim \mathcal{N}_d(\mu, \Sigma).$$

In general the density looks awful. In this course, you will never have to explicitly evaluate the above density in the case the X_is are dependent. But the joint density nicely simplifies if the random variables are uncorrelated. We show why below.

If $\{X_i\}_{i=1}^d$ are Gaussian random variables with mean μ_i, variance σ^2 and $\text{cov}(X_{j_1}, X_{j_2}) = 0$ for $j_1 \neq j_2$. Then

$$\text{var}(\underline{X}_d) = \Sigma = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \sigma^2 I_d.$$

This is a diagonal matrix with inverse

$$\Sigma^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \sigma^{-2} I_d.$$

Using algebra the joint density is

$$f_{\underline{X}_d}(\underline{x}) = \frac{1}{\sqrt{(2\pi)^d |\sigma^2 I_d|}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})(\sigma^2 I_d)^{-1}(\underline{x} - \underline{\mu})'\right) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma^2}\right) = \prod_{i=1}^d f_{X_i}(x_i).$$

Observe that the joint density is the product of the marginals. Thus if the Gaussian random variables are uncorrelated, then they are independent.

The multivariate Gaussian distribution is a "work horse" in statistics. It drives many modern statistical methods (often it is assumed in the background). It has many useful properties. One of these is that any linear combination/transformation of jointly Gaussian random variables is also Gaussian. This means that all one has to do is evaluate the mean and variance of the linear transformation (see Section 1.4) and transformation will be Gaussian with the new evaluated mean and variance. Below, we give some examples (return to Section 1.4 and practice these).

1. Suppose $X \sim N(\mu, \sigma^2)$ and let $Y = aX + b$. Then by using the results in Sections 1.4.1 and 1.4.3 we have

$$\mathrm{E}(Y) = a\mu + b \qquad \mathrm{var}(Y) = a^2\sigma^2.$$

Thus

$$Y \sim \mathcal{N}_d\left(a\mu + b, a^2\sigma^2\right).$$

A commonly used transformation is the z-transform:

$$\sigma^{-1}(X - \mu) \sim \mathcal{N}_d(0, 1).$$

2. Suppose Y_i are independent, normal random variables with $\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i$ and $\mathrm{var}[Y_i] = \sigma^2$ (independence means $\mathrm{cov}(Y_i, Y_j) = 0$ if $i \neq j$). Then

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_d \end{pmatrix} \sim \mathcal{N}_d\left(\begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_d \end{pmatrix}, \sigma^2 I_d\right).$$

To do Write out the density of $\underline{Y}$.

3. If $\underline{X}_d \sim \mathcal{N}_d(\mu, \Sigma)$ where $\Sigma_{j_1,j_2} = \sigma_{j_1 j_2}$. Then any linear combination of $\underline{X}_d$; $Y = \sum_{j=1}^d \alpha_j X_j$ is normal with mean $\sum_{j=1}^d \alpha_j \mu_j$ and variance

$$\mathrm{var}(\sum_{j=1}^d \alpha_j X_j) = \underline{\alpha}\Sigma\underline{\alpha}' = \sum_{j_1,j_2=1}^d \alpha_{j_1}\alpha_{j_2}\mathrm{cov}(X_{j_1}, X_{j_2}) = \sum_{j_1,j_2=1}^d \alpha_{j_1}\alpha_{j_2}\sigma_{j_1 j_2}$$

where $\underline{\alpha} = (\alpha_1, \ldots, \alpha_d)$ and the (j_1, j_2) entry of Σ is $\sigma_{j_1 j_2}$. In summary

$$\begin{aligned} \sum_{j=1}^d \alpha_j X_j &\sim \mathcal{N}\left(\sum_{j=1}^d \alpha_j \mu_j, \sum_{j_1,j_2=1}^d \alpha_{j_1}\alpha_{j_2}\sigma_{j_1 j_2}\right) \\ \text{same as saying } \underline{\alpha}\underline{X}_d &\sim \mathcal{N}\left(\underline{\alpha}\underline{\mu}_d, \underline{\alpha}\Sigma\underline{\alpha}'\right) \end{aligned}$$

4. Generalising the above if $Y_1 = \sum_{j=1}^d \alpha_j X_j$ and $Y_2 = \sum_{j=1}^d \beta_j X_j$, then

$$\begin{aligned} \begin{pmatrix} \underline{\alpha}\underline{X}_d \\ \underline{\beta}\underline{X}_d \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} \underline{\alpha} \\ \underline{\beta} \end{pmatrix}\underline{\mu}_d, \begin{pmatrix} \underline{\alpha} \\ \underline{\beta} \end{pmatrix}\Sigma(\underline{\alpha}', \underline{\beta}')\right) \\ &= \mathcal{N}\left(\begin{pmatrix} \underline{\alpha}\underline{\mu} \\ \underline{\beta}\underline{\mu} \end{pmatrix}, \begin{bmatrix} \underline{\alpha}\Sigma\underline{\alpha}' & \underline{\alpha}\Sigma\underline{\beta}' \\ \underline{\beta}\Sigma\underline{\alpha}' & \underline{\beta}\Sigma\underline{\beta}' \end{bmatrix}\right) \end{aligned}$$

5. If $\underline{X}_d \sim \mathcal{N}_d(\mu, \Sigma)$ then

$$\left(A\underline{X}_d + \underline{b}\right) \sim \mathcal{N}_d\left(A\underline{\mu} + \underline{b}, A\Sigma A'\right).$$

A commonly used transformation is $\underline{Y} = \Sigma^{-1/2}(\underline{X} - \underline{\mu})$. This gives

$$\mathrm{E}[\underline{Y}] = \Sigma^{-1/2}\mathrm{E}[\underline{X} - \underline{\mu}] = 0$$

and

$$\begin{aligned}\mathrm{var}[\underline{Y}] &= \mathrm{var}\left[\Sigma^{-1/2}\underline{X}\right] = \Sigma^{-1/2}\mathrm{var}(\underline{X})\Sigma^{-1/2} \\ &= \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma^{1/2}\Sigma^{1/2}\Sigma^{-1/2} = I_d.\end{aligned}$$

Therefore

$$\underline{Y} = \Sigma^{-1/2}\left(\underline{X}_d - \underline{\mu}\right) \sim \mathcal{N}_d\left(0, I_d\right).$$

6. If $\underline{X}$ and $\underline{Y}$ are p and q dimensional random vectors which are jointly normal and $\mathrm{cov}(\underline{X}, \underline{Y}) = 0$ (a matrix of zeroes), then

$$var\left[\begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix}\right] = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$$

where $\mathrm{var}(\underline{X}) = A$ and $\mathrm{var}(\underline{Y}) = B$. The above is known as a block diagonal matrix. Further, it can be shown that $\underline{X}$ and $\underline{Y}$ are independent of each other (your HW).

7. Suppose E is an orthonormal transformation matrix as defined Section 1.3.3. If this sounds scary thing of a system of orthonomal vectors such as $\{(1,2)/\sqrt{5}, (-2,1)/\sqrt{5}\}$ of $\{(1,1)/\sqrt{2}, (1,-1)/\sqrt{2}\}$ which give the 2×2 matrix

$$E = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} \text{ or } E = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

Let $\underline{X} \sim N(\underline{\mu}, \Sigma)$ and define the linear transform $\underline{Y} = E\underline{X}$. Then by using equations (1.4) and (1.5) the density of $f_{\underline{Y}}$ is

$$\frac{1}{\sqrt{(2\pi)^d \det(E'\Sigma E)}} \exp\left(-\frac{1}{2}(\underline{y} - E\underline{\mu})'(E'\Sigma E)^{-1}(\underline{y} - E\underline{\mu})\right) \tag{2.2}$$

In the special case $\Sigma = I_d$ we have $E'I_dE = E' = E = I_d$ the density of $E\underline{X}$ is

$$\frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}(\underline{y} - E\underline{\mu})'(\underline{y} - E\underline{\mu})\right)$$

Definition 2.1. *We say Z has a standard normal distribution if Z has a normal distribution with mean zero and variance one.*

Research 2. *The inverse of the variance matrix Σ is called the precision matrix. It has many interesting properties (related to linear regression, conditional independence and graphical models). You may want to read up on it.*

2.2 Relatives of the Gaussian distribution

2.2.1 The chi-square distribution

Suppose $Z_1, \ldots, Z_n$ are iid standard normal random variables. Then the distribution of

$$Z_1^2 + \ldots + Z_n^2 \sim \chi_n^2,$$

where χ_n^2 denotes a (central) chi-square distribution with n degrees of freedom (df for short). For example, the distribution of Z^2 is a chi-square with one df. In general if $X_1, \ldots, X_n$ are iid normal random variables with mean μ and variance σ^2, then

$$\left(\frac{X_1 - \mu}{\sigma}\right)^2 + \ldots + \left(\frac{X_n - \mu}{\sigma}\right)^2 \sim \chi_n^2.$$

An analytic form for the density exists. But the plots of the densities are more illuminating. A plot of the density of a chi-square distribution for various degrees of freedom is given in Figure 2.5. Observe, that as

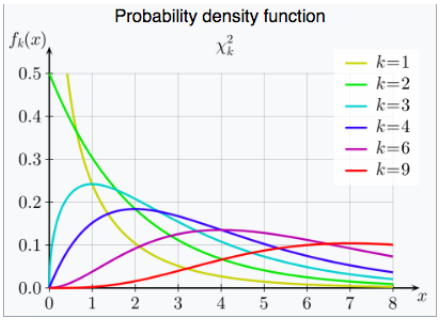


Figure 2.5: Plot of different chi-squared densities. Stolen from Wiki

n moves away from one the hump of the chi-square distribution moves along the x-axis (the mode is at $\max(0, n-2)$). This makes sense, as n grows, more and more positive random variables are being summed together (without any standardisation), so $\sum_{i=1}^n X_i^2$ grows. This impacts the mean and variance, as we can see in the following lemma.

Lemma 2.1

If Y_n has a chi-square distribution with n df. Then $\mathrm{E}[Y_n] = n$, $\mathrm{var}[Y_n] = 2n$ and skewness is $S_3 = \sqrt{8/n}$.

Using the above result and the rules of variances given in Section 1.4.3 we have: if Y_n has a chi-square distribution with n degrees of freedom, then

$$\mathrm{var}(n^{-1}Y_n) = n^{-2}(2n) = 2n^{-1}.$$

2.2.2 The t-distribution

Suppose that $Z_0, \ldots, Z_n$ are iid standard normally distributed random variables with a standard normal distribution. Then the ratio

$$T_n = \frac{Z_0}{\sqrt{n^{-1}\sum_{i=1}^{n} Z_i^2}}$$

is said to have a t-distribution with n-df. Often we write it as $Z = Z_0$ and $U_n = \sum_{i=1}^{n} Z_i^2$, then

$$T_n = \frac{Z}{\sqrt{n^{-1}U_n}},$$

has a t-distribution with n df. The density of a t-distribution with n df is given by

$$f_n(x) = \frac{\Gamma(n+1)/2}{\sqrt{n\pi}\Gamma(n/2)}\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} \qquad x \in \mathbb{R},$$

where Γ denotes gamma function. From the formula we observe that for large x (ignoring some constants)

$$f_n(x) \sim \frac{1}{|x|^{-(n+1)}}.$$

Further, $f_n(x)$ is a proper density for all $n > 0$ (and not just the integers). The non-integer case has useful applications too (you may have used it in the independent two sample t-test when the population variances are assumed different). A plot of the density for different integer values is given in Figure 2.6.

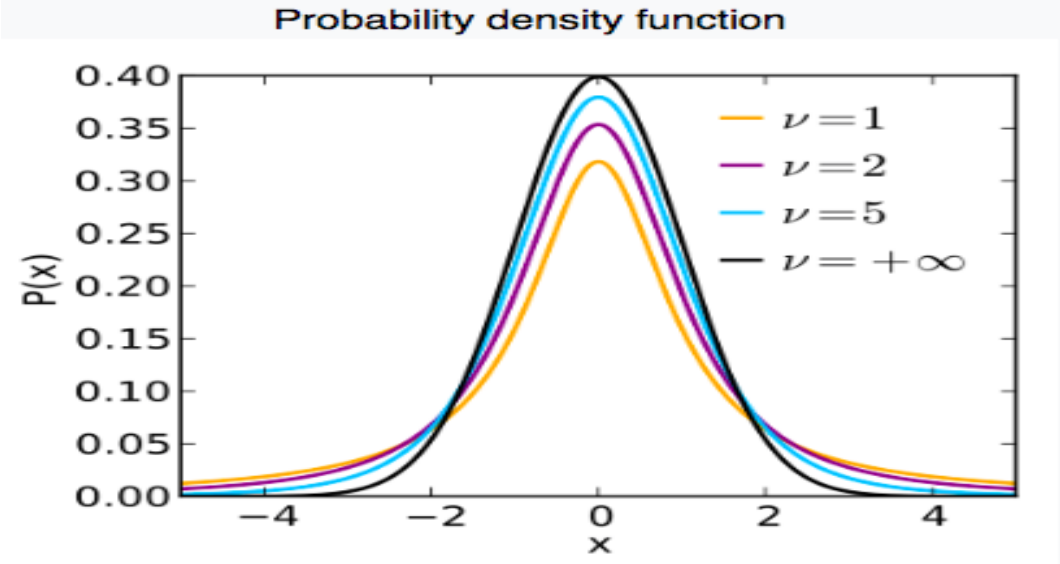


Figure 2.6: Plot of different t-densities. Black curve corresponds to a normal distribution. Stolen from Wiki

Lemma 2.2

If T_n has t-distribution with n df. Then for $n > 1$ $\mathrm{E}[T_n] = 0$ and for $n > 2$ $\mathrm{var}[T_n] = n/(n-2)$.

If the df is too low, the moments do not exist. This is because the tails of the t-distribution for low df are very "thick". Thick tails mean that extremes are likely to happen with a large probability. This means functions can happen with a "large" probability, such that $\mathrm{E}(T_n^2)$ is not finite.

Example 2.1 (What does a moment being undefined or non-existent mean?)**.** *Let us look at the case $n = 1$ (this is also called the Cauchy distribution). The density of the t-distribution reduces to*

$$\begin{aligned} f_1(x) &= \frac{\Gamma(1)}{\sqrt{\pi}\Gamma(1/2)}\frac{1}{(1+x^2)} \quad x \in \mathbb{R} \\ &= \frac{1}{\pi(1+x^2)}. \end{aligned}$$

This is a distribution which is symmetric about zero. We would expect it to have a mean that it is zero. But this turns out not to be the case because the expectation is not defined. We now explain why:

$$\begin{aligned} \mathrm{E}(T_1) &= \int_{-\infty}^{\infty} x f_1(x)dx = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}dx \\ &= \int_{-\infty}^{0} \frac{x}{\pi(1+x^2)}dx + \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}dx. \end{aligned}$$

Let us consider the last term in the above integral and partition it into two sums with M being "large"

$$\begin{aligned} \int_0^{\infty} \frac{x}{\pi(1+x^2)}dx &= \underbrace{\int_0^{M} \frac{x}{\pi(1+x^2)}dx}_{\text{some number}} + \int_M^{\infty} \frac{x}{\pi(1+x^2)}dx \\ &= \underbrace{\int_0^{M} \frac{x}{\pi(1+x^2)}dx}_{\text{some number}} + \int_M^{\infty} \underbrace{\frac{x}{\pi(1+x^2)}}_{\approx x^{-1}} dx \end{aligned}$$

To see why in the second term, for large x, $\frac{x}{(1+x^2)} \approx 1/x$ divide the numerator and denominator by x:

$$\frac{x}{1+x^2} = \frac{1}{1/x + x^2/x} = \frac{1}{1/x + x} \approx \frac{1}{x}.$$

Recall elementary calculus;

$$\int_M^{y} \frac{1}{x}dx = \log y - \log M.$$

Thus letting $y \to \infty$ gives $\int_M^{\infty} \frac{1}{x}dx = \infty$. Therefore, the above gives

$$\int_0^{\infty} \frac{x}{\pi(1+x^2)}dx = \infty.$$

Using a similar set of argument we have $\int_{-\infty}^{0} \frac{x}{\pi(1+x^2)} dx = -\infty$. *Thus* $\mathrm{E}[X]$ *is not well defined (since* $-\infty + \infty$ *has no logical meaning).*

On the other hand, by the same argument, we can show that $\mathrm{E}[X^2] = \infty$. *Thus for the t-distribution with 1df, the second moment does not exist.*

Remark. *Note if* Z *and* U_n *are not independent, then* $T_n = Z/(U_n/n)$ *does not a have t-distribution with n df.*

2.2.3 The F-distribution

Suppose that U_p and V_q are two independent random variables with a chi-square distribition with p and q degrees of freedom. Then the ratio

$$\frac{U_p/p}{V_q/q}$$

has an F-distribution with (p, q) degrees of freedom.

Example 2.2. *Suppose that* T_n *has t distribution with n degrees of freedom. Then* T_n^2 *has an F-distribution with* $(1, n)$ *degrees of freedom.*

2.3 The exponential class of distributions

There exists a general, algebraic expression that characterises the Gaussian distribution, chi-squared distribution, binomial distribution and many other distributions to boot.

If X comes from the exponential class of distribution, then it has a "parametric" density/distribution with the form

$$\begin{aligned} f(x;\theta) &= \exp\left[s(x)T(\theta) + b(\theta) + c(x)\right] \qquad x \in A, \\ \Rightarrow \log f(x;\theta) &= s(x)T(\theta) + b(\theta) + c(x), \end{aligned} \tag{2.3}$$

where the functions s, T, b and c are all known. The only unknown is the parameter θ, which though unknown, it is known to belong to the parameter space Θ. Usually, Θ is the set of all parameters where $f(x;\theta)$ is a proper density/distribution (integrates to one and is positive). It is important to note that the set A does not depend on the parameter θ. If A is a fuction of θ, then $f(x;\theta)$ does not belong to the exponential family. This is very important and is the reason that the exponential family does not include the uniform distribution whose support is a function of the parameter θ. The exponential family also excludes other important distributions such as the t-distribution and the Weibull distribution.

The above is for the single parameter case. For multiple parameters, the generalisation is

$$f(x;\underline{\theta}) = \exp\left[\sum_{i=1}^{K} s_i(x)T_i(\underline{\theta}) + b(\underline{\theta}) + c(x)\right] \qquad x \in A, \tag{2.4}$$

where A does not involve the unknown parameter $\underline{\theta}$ and $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ (usually the number of terms $T_i(\underline{\theta})$ match the dimension of $\underline{\theta}$; the number of parameters). We give below some examples of distribution families which can be written in the above form.

This large class of distributions (we give examples below) is mainly considered because it is has interesting mathematical properties and it is also practically very useful. We will not dwell too much on the actual properties of the class of distributions, but will occasionally discuss it. The main point of interest for this course, is that distributions which belong to the exponential family have similar type of estimation and inferential properties (see Chapter 3). Furthermore, the likelihood corresponding to distributions from the exponential family (see Chapter 3 for the definition of a likelihood) are easy to maximise because they are concave functions (if you are interested it is worth investigating this further).

Example 2.3. *(i) The exponential distribution is for positive continuous response random variables. It has the pdf is $f(x;\lambda) = \lambda\exp(-\lambda x)$, which can be written as*

$$\log f(x;\lambda) = (-x\lambda + \log\lambda) \qquad x \geq 0.$$

The parameter space is $\Theta = (0,\infty)$. Therefore $s(x) = -x$ and $b(\lambda) = \log\lambda$.

(ii) The binomial distribution is for positive discrete response random variables with outcomes $\{0,1,\ldots,n\}$. The probability mass function is

$$P(X = k;\pi) = \binom{n}{k}\pi^k(1-\pi)^{n-k}, \qquad k = 0,\ldots,n.$$

The parameter space is $\pi \in [0,1]$. Sometimes, have a parameter space which is restricted is not convenient for estimation. So we "reparameterize" let $\theta = \log(\frac{\pi}{1-\pi})$, then the log of the pmf is

$$\log f(x;\theta) = \log f(y;\log\frac{\pi}{1-\pi}) = x\theta - n\log\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right) + \log\binom{n}{x}.$$

Since $\log(\frac{\pi}{1-\pi})$ is bijective (we can go from θ to π and back again), the equivalent parameter space of θ is $\Theta \in (-\infty,\infty)$.

(iii) The normal distribution is for continuous response random variables. It is characterised by its mean and variance, μ and σ^2 respectively. Its logarithm is

$$\begin{aligned}\log f(x;\mu,\sigma^2) &= \left(-\frac{(x-\mu)^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2 + \frac{1}{2}\log(2\pi)\right)\\ &= \left(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} + \frac{\mu^2/2}{\sigma^2} + \frac{1}{2}\log(2\pi)\right) \qquad x \in (-\infty,\infty).\end{aligned}$$

The parameter space is

$$\Theta = \{\mu \in (-\infty, \infty), \sigma^2 \in [0, \infty)\}.$$

(iv) The Poisson is a positive discrete response random variable whose set of possible outcomes is $\{0, 1, 2, \ldots\}$. *The probability mass function is*

$$P(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!} \qquad k \geq 0.$$

The parameter space is $\Theta \in (0, \infty)$. *The log distribution can be written as*

$$\log f(x; \lambda) = x \log \lambda - \lambda + \log x!.$$

An alternative parameterisation is to let $\lambda = \log \mu$.

(v) The Gamma distribution distribution is for continuous response positive random variables. It has density

$$f(x; \lambda, \alpha) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x) \qquad x \geq 0.$$

The parameter space is

$$\Theta = \{\alpha \in (0, \infty), \beta \in (0, \infty)\}.$$

The exponential distribution is a special case of the Gamma. A plot of several different Gamma distributions is given in Figure 2.7

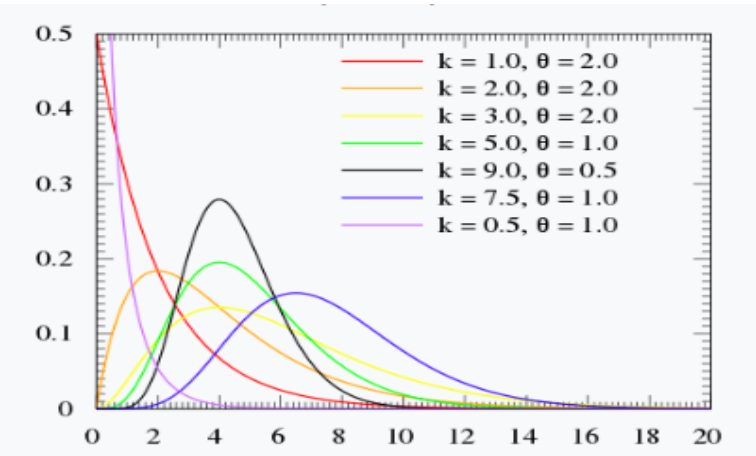


Figure 2.7: Plot of different gamma distributions. Stolen from Wiki

(vi) Other members in this family include the beta, Multinomial and inverse Gaussian to name but a few.

Example 2.4 (Example of distributions that do not belong to the exponential family). *(i) The uniform distribution where the support of the distribution is the unknown parameter (HW problem).*

- *The Weibull distribution (which is usually used to model failure times):*

$$f(x;\lambda,k) = \left(\frac{k}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{k-1}\exp\left(-\left[\frac{x}{\lambda}\right]^{k}\right) \qquad x \geq 0$$

where $\Theta = \{(k,\lambda), k \in (0,\infty), \lambda \in (0,\infty)\}$. *Then*

$$\log f(x;\lambda,k) = -\left(\frac{x}{\lambda}\right)^{k} + (k-1)\log x - (k-1)\log\lambda + \log k - \log\lambda.$$

Observe that in the case $k \neq 1$, *the parameter* $(x/\lambda)^k = f(x,\lambda,k)$ *cannot be separated into* $T(\lambda,k)s(x)$. *Because of this, the Weibull distribution does not belong to exponential family. So does not inherit some of their nice properties.*

A plot of several different Weibull distributions is given in Figure 2.8.

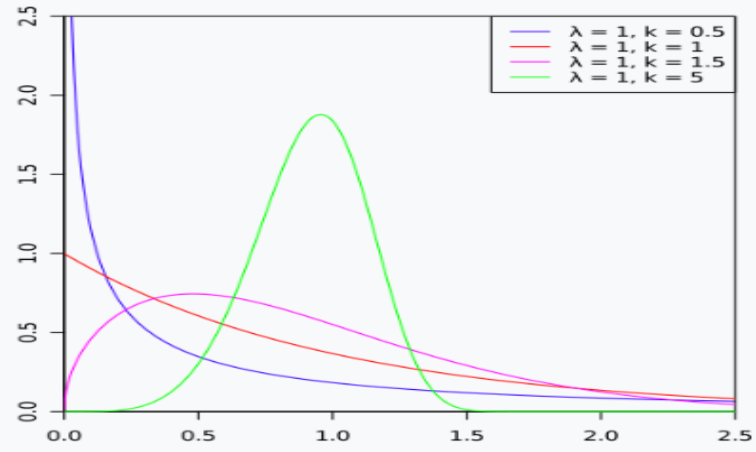


Figure 2.8: Plot of different Weibull distributions. Stolen from Wiki

So far the reparameterisation of a distribution n terms of the exponential family seems to be an algebraic exercise. However, the exponential family has been widely studied in statistics. This is because

- Under certain conditions, if the data comes from a distribution which belongs to the exponential family then is extremely simple to estimate the parameters. We cover this later, but maximisation of the likelihood (which is often quite fiendish) is straightforward for the exponential family.
- The exponential family has useful properties. For example all the information about the parameters in the distribution can be described in terms of their so called sufficient statistics. Loosely speaking, this is a function of the observed data.

2.4 The sample mean and variance: Sampling distributions

2.4.1 The sample mean

Suppose $\{X_i\}_{i=1}^n$ are iid random variables with mean μ and variance σ^2. The most obvious estimator of μ is the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

By using the results in Example 1.8 we have

$$\mathrm{E}(\bar{X}_n) = n^{-1}[\mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n)] = \mu$$

and

$$\mathrm{var}(\bar{X}_n) = n^{-2}\sum_{i_1,i_2=1}^{n}\mathrm{cov}(X_{i_1}, X_{i_2}) = n^{-2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{\sigma^2}{n}.$$

If $\{X_i\}_{i=1}^n$ are normally distributed, then $\bar{X}_n$ is normal, regardless of the sample size. On the other hand, if $\{X_i\}_{i=1}^n$ is not normally distrbute then only for "sufficiently" large sample size is $\bar{X}_n$ close to normal (this is the central limit theorem coming into play; see Section 1.5).

Usually $\sqrt{\mathrm{var}(\bar{X}_n)}$ is called the standard error of the sample mean i.e.

$$s.e. = \frac{\sigma}{\sqrt{n}},$$

which you would have encountered in an elementary statistics. Recall from Example 1.8 that we need uncorrelatedness of $\{X_i\}$ for this to be the correct standard error. If there is correlation between the X_is the standard error will be different. Look back at the calculation and see what happens to the standard error.

2.4.2 The sample variance

Suppose we observe $X_1, \ldots, X_n$ which are iid random variables with mean μ and variance σ^2. Suppose μ is known and σ^2 is unknown (this rarely ever happens). Then an estimator of σ^2 is

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

If $\{X_i\}$ are iid normal, then $(X_i - \mu)/\sigma$ are iid standard normal random variables it immediately follows from Section 2.2.1 that

$$ns_n^2/\sigma^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

should follow a χ^2 distribution with n degrees of freedom. Therefore

$$s_n^2 \sim \frac{\sigma^2}{n}\chi_n^2$$

i.e. the distribution of s_n^2 is the same as a χ_n^2 random variable multipled by σ^2/n.

A more realistic situation is that both μ and σ^2 are unknown. We know that a reasonable estimator of μ is $\bar{X}$. An estimator of σ^2 is based on a similar idea. First let us return to the iid random variables $\{X_i\}$ where $X_i \sim N(\mu, \sigma^2)$. An alternative but equivalent representation of X_i is to write X_i as an equation:

$$X_i = \mu + \varepsilon_i$$

where ε_i is called the residual with $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{var}[\varepsilon_i] = \mathrm{E}[\varepsilon_i^2] = \sigma^2$. Since $\mathrm{E}[\varepsilon_i^2] = \sigma^2$, one can use the average of $\{\varepsilon_i^2\}_{i=1}^n$ as an estimator of σ^2. However, $\varepsilon_i = X_i - \mu$ is not observed. Thus $n^{-1}\sum_i \varepsilon_i^2$ cannot be used as an estimator. Instead we can replace the residual with its estimate. We recall that $\bar{X}$ is an estimator of μ thus an estimator of ε_i is the estimated residual

$$\widehat{\varepsilon_i} = X_i - \bar{X}.$$

This then leads to the potential estimator

$$\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon_i}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

However, the classical estimator of the sample variance is not exactly the above, but something very close

$$s_n^2 = \frac{1}{(n-1)}\sum_{i=1}^n (X_i - \bar{X}_n)^2. \tag{2.5}$$

Why we divide by $(n-1)$ rather than n will become clear in Theorem 2.3 (and its subsequent proof).

Remark. *In the extreme case, we only observe X_1. We can estimate the mean with $\bar{X}_1 = X_1$. But it is impossible to estimate the variance, since the data contains no information about the spread of the data. Therefore s_n^2 is only meaningful when $n > 1$.*

We now show that the distribution of s_n^2 with an appropriate standardisation follows a χ^2-distribution with $(n-1)$ degrees of freedom (df for short).

Theorem 2.3

Suppose $X_1, \ldots, X_n$ are iid random variables with mean μ and variance σ^2. Let s_n^2 be defined as in (2.5). Then

$$\mathrm{E}[s_n^2] = \sigma^2$$

If, in addition, $X_i \sim N(\mu, \sigma^2)$, then

$$(n-1)\frac{s_n^2}{\sigma^2} \sim \chi^2_{n-1},$$

and s_n^2 and $\bar{X}_n$ are independent random variables.

To prove the result, we focus on the case that $X_i \sim N(\mu, \sigma^2)$. Though seemingly complex, the proof below simply relies on properties of Gaussian random variables and basic results from linear algebra such as projections and orthonormal basis (which we reviewed in the previous chapter).

Remark. *Like most important results there are several different ways to prove it. One proof uses that the sample mean and sample variance $\bar{X}_n$ and s_n^2, are independent. This result can be shown by evaluating the covariance between $\bar{X}_n$ and $(X_i - \bar{X}_n)$:*

$$\begin{aligned}
\text{cov}\left(\bar{X}_n, X_i - \bar{X}_n\right) &= \text{cov}\left(X_i, \bar{X}_n\right) - \text{cov}(\bar{X}_n, \bar{X}_n) \\
&= \text{cov}\left(X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\right) - var(\bar{X}_n) \\
&= \frac{1}{n}\sum_{j=1}^{n} \text{cov}(X_i, X_j) - \frac{\sigma^2}{n} = \frac{1}{n}\text{cov}(X_i, X_i) - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0 \qquad \textit{for all} \quad i.
\end{aligned}$$

Since $\bar{X}_n$ and $\{X_i - \bar{X}_n\}_{i=1}^n$ are uncorrelated and jointly normal they are independent. Using this, the moment generating function of $\sum_{i=1}^n (X_i - \mu)^2$ is written in terms of the moment generating function of $n(\bar{X}_n - \mu)^2$ times the moment generating function of s_n^2. From which one can deduce that the moment generating function corresponding to s_n^2 is a χ^2 with $(n-1)$-df.

The proof we give is based on linear transformation of the observations and projections. These ideas can easily be generalised to the independent two-sample t-test, ANOVA and linear regression. But it does require the results from Section 1.3.2.

We prove Theorem 2.3 in the case $n = 2$ and 3. The proof below will help answer the following questions:

- Why divide by $(n-1)$ and not n?

  We show below that by dividing by $(n-1)$ rather than n, the resulting estimator is unbiased. This effect is most pronounced in small sample sizes (when $n = 2$ it is obvious). For relatively large sample sizes the difference between $(n-1)$ and n is not so large.

- Why $(n-1)$ df? What does a df mean anyway?

  We show below that one interpretation of df is the number of independent random variables used to construct the estimator.

- Does the estimator get better as the sample size grows?

  As with most estimators as the sample size grows the estimator improves. The improvement in the estimator is measured with the standard error which is the standard deviation of the estimator.

The proof hinges on an appropriate transformation of the observed data vector

$$\underline{X}' = (X_1, X_2, \ldots, X_n).$$

This is how we store the data, so making meaningful transformation of it is a very natural thing to do.

Important preliminaries

An important property we will use is that for normally distributed random variables, no correlation implies independence (this is not true for other distributions). We consider some simple examples below.

Example 2.5. *Suppose that* $(X_1, X_2)' \sim N((\mu, \mu)', \sigma^2 I_2)$. *Then* X_1 *and* X_2 *are independent.*

(i) Show that the two transformations; $Y = (X_1 + X_2)$ *and* $U = (X_1 - X_2)^2$ *are independent too.* U *is not normal, but* $W = (X_1 - X_2)$ *is normal (as it is a linear combination of jointly normal random variables). Since* $U = W^2$*, independence of* Y *and* W *will imply independence of* Y *and* $U = W^2$*. We now show that* $\text{cov}(Y, W) = 0$*:*

$$\begin{aligned} \text{cov}(Y, W) = \text{cov}(X_1 + X_2, X_1 - X_2) &= \text{cov}(X_1, X_1) - \text{cov}(X_1, X_2) + \text{cov}(X_2, X_1) - cov(X_2, X_2) \\ &= \sigma^2 - 0 + 0 - \sigma^2 = 0. \end{aligned}$$

Thus, by joint normality of (Y, W)*,* Y *and* W *are independent. This immediately implies any transformation of* Y *and* W *are independent. Thus* Y *and* $U = W^2$ *are independent.*

(ii) Show that the two transformations; $Y = (X_1 + X_2)$ *and* $U = (X_1 - 2X_2)^2$ *are dependent. As above we show that* Y *and* $W = (X_2 - 2X_2)$ *are dependent, which will usually mean that* Y *and* $U = W^2$ *are dependent too. We now show that* $\text{cov}(Y, W) \neq 0$*:*

$$\begin{aligned} \text{cov}(Y, W) = \text{cov}(X_1 + X_2, X_1 - 2X_2) &= \text{cov}(X_1, X_1) - 2\text{cov}(X_1, X_2) + \text{cov}(X_2, X_1) - 2cov(X_2, X_2) \\ &= \sigma^2 - 0 + 0 - 2\sigma^2 = -\sigma^2. \end{aligned}$$

Thus we have shown Y *and* W *are dependent (if* $\sigma^2 > 0$*). Thus,* Y *and* $U = W^2$ *are dependent.*[1]

[1] There can arise very strange examples where two random variables are dependent but their squares are not. For example if X, Y and δ are independent random variables where $\delta = \{-1, 1\}$. Then $U_1 = \delta X$ and $U_2 = \delta Y$ are dependent, but $U_1^2 = X^2$ and $U_2 = Y^2$ are indepedendent. But you can ignore such cases in this course (the technical reason for this is due to their sigma algebras, but do not even think about this).

Proof of Theorem 2.3 in the case the sample size is $n = 2$

Let

$$s_2^2 = \frac{1}{2-1}\left[(X_1 - \bar{X}_2)^2 + (X_2 - \bar{X}_2)^2\right]$$

and

$$\bar{X}_2 = \frac{1}{2}(X_1 + X_2).$$

We will show that $\mathrm{E}[s_2^2] = \sigma^2$ and $\frac{s_2^2}{\sigma^2} \sim \chi_1^2$. This is the same as saying that the density of $\frac{s_2^2}{\sigma^2}$ has the shape of a χ^2 with one df.

We observe that s_2^2 contains the sum of squares of normal random variables. Therefore it seems reasonable that the distribution of s_2^2 is a chi-square "type" distribution. Moreover, since s_2^2 is the sum of two random variables, it would, on first appearances, appear that s_2^2 should follow a χ_2^2 with two df and not one as stated above. However, a more careful study shows that $s_2^2 = Y_1^2 + Y_2^2$, where

$$Y_1 = (X_1 - \bar{X}_2) = \frac{1}{2}(X_1 - X_2) \qquad \text{and} \qquad Y_2 = (X_2 - \bar{X}_2) = \frac{1}{2}(X_2 - X_1).$$

Clearly $Y_1 = -Y_2$, thus besides the sign change Y_1 and Y_2 are the same (Y_2 does not convey any additional information), they are certainly *not* independent. Therefore

$$s_2^2 = \frac{1}{2-1}\left[(X_1 - \bar{X}_2)^2 + (X_2 - \bar{X}_2)^2\right] = \left(\frac{1}{\sqrt{2}}[X_1 - X_2]\right)^2 = Z^2, \tag{2.6}$$

where $Z = \frac{1}{\sqrt{2}}[X_1 - X_2]$. Since Z is the sum of two random variable whose joint distribution is normal, then by the properties of a normal distribution (see Section 2.1), Z must also be normal, characterized by $\mathrm{E}[Z]$ and $\mathrm{var}(Z)$. Using the rules of expectations and variances given in Section 1.4.1 we have

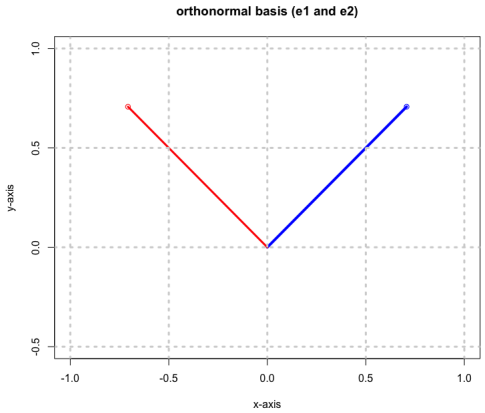
$$\begin{aligned}
\mathrm{E}(Z) &= \frac{1}{\sqrt{2}}[\mathrm{E}(X_1) - \mathrm{E}(X_2)] = \frac{1}{\sqrt{2}}[\mu - \mu] = 0 \\
\mathrm{var}(Z) &= \frac{1}{2}\left[\mathrm{var}(X_1) + \mathrm{var}(X_2) - \underbrace{2\mathrm{cov}(X_1, X_2)}_{0}\right] = \sigma^2.
\end{aligned}$$

Thus $Z \sim N(0, \sigma^2)$. Therefore

$$\mathrm{E}[s_2^2] = \mathrm{E}[Z^2] = \mathrm{var}(Z) + \underbrace{\mathrm{E}[Z]^2}_{=0} = \sigma^2.$$

Observe that in the definition of s_2^2 we divide by one and not two. The reason for this becomes apparent from the calculation above, this is to ensure that s_2^2 is an unbiased estimator of σ^2. Further, since $s_2^2 = Z^2$ = normal squared. Then $Z/\sigma \sim N(0,1)$. Thus $\frac{s_s^2}{\sigma^2}$ is a standard normal squared, leading to $\frac{s_s^2}{\sigma^2} \sim \chi_1^2$ (a chi-square distribution with one df). This proves the first part of Theorem 2.3 in the case $n = 2$.

We now discuss the relationship of the sample mean $\bar{X}_2 = 2^{-1}(X_1+X_2)$ and sample variance s_2^2. We will show that they are independent of each other. Since $s_2^2 = Z^2$, we need only show that $\bar{X}_2$ and Z are independent of each other[2] . Further, because $\bar{X}_2$ and Z are a linear combination of jointly normal random variables, they must be normal too[3]. We show that the vector $(\bar{X}, Z)$ is a linear transformation of (X_1, X_2) (importantly, it is an orthogonal transformation, based on the orthogonormal vectors $\underline{e}_1$ and $\underline{e}_2$ (see Example 1.1));



Place the vector $\underline{X}_2 = (X_1, X_2)'$ on the plot on the right. The projection of $\underline{X}_2$ onto the blue line gives the coefficient $\langle \underline{e}_1, \underline{X}_2 \rangle = 2^{-1/2}(X_1 + X_2) = 2^{1/2}\bar{X}$. The projection of $\underline{X}_2$ onto the red line gives the coefficient $\langle \underline{e}_2, \underline{X}_2 \rangle = 2^{-1/2}(X_2 - X_2) = Z$;

$$\underbrace{\begin{pmatrix} \bar{X} \\ Z \end{pmatrix}}_{=\underline{Y}_2} = \begin{pmatrix} 2^{-1/2}\langle \underline{e}_1, \underline{X}_2 \rangle \\ \langle \underline{e}_2, \underline{X}_2 \rangle \end{pmatrix} = \underbrace{\begin{pmatrix} 1/2 & 1/2 \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}}_{E_2} \underbrace{\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}}_{=\underline{X}_2}.$$

The transformed vector $\underline{Y}_2' = (\bar{X}, Z)$ is normal with mean $(\mu, 0)'$ and variance

$$\begin{aligned} \text{var}(\underline{Y}_2) &= \text{var}(E_2\underline{X}_2) = E_2\text{var}(\underline{X}_2)E_n' \\ &= \begin{pmatrix} 1/2 & 1/2 \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1/2 & 1/\sqrt{2} \\ 1/2 & -1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned} \tag{2.7}$$

An alternative argument for proving (2.7) uses that $\underline{e}_1$ and $\underline{e}_2$ forms and orthonormal basis, and has the advantage that it simply uses properties without the need for brute force calculations. We briefly summarize it now. Using the notation from Example 1.1 we have

$$\begin{pmatrix} 1/2 & 1/2 \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 2^{-1/2}\underline{e}_1 \\ \underline{e}_2 \end{pmatrix}$$

where $\underline{e}_1 = (1/\sqrt{2}, 1/\sqrt{2})$ and $\underline{e}_2 = (1/\sqrt{2}, -1/\sqrt{2})$ are an orthonormal basis of $\mathbb{R}^2$. E.g. using the orthonormality (that is $\underline{e}_1\underline{e}_1' = \underline{e}_2\underline{e}_2' = 1$ and $\underline{e}_1\underline{e}_2'$=0) property of the vectors and that $\text{var}(\underline{X}_2) = \sigma^2 I_2$ we have

$$\text{var}(\underline{Y}_2) = \begin{pmatrix} 2^{-1}\underline{e}_1\underline{e}_1' & 2^{-1/2}\underline{e}_1\underline{e}_2' \\ 2^{-1/2}\underline{e}_2\underline{e}_1' & \underline{e}_2\underline{e}_2' \end{pmatrix} = \sigma^2 \begin{pmatrix} 2^{-1}\langle \underline{e}_1, \underline{e}_1 \rangle & 2^{-1/2}\langle \underline{e}_1, \underline{e}_2 \rangle \\ 2^{-1/2}\langle \underline{e}_2, \underline{e}_1 \rangle & \langle \underline{e}_2\underline{e}_2 \rangle \end{pmatrix} = \sigma^2 \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}.$$

[2]Random variables are independent if their distributions are a product of their marginals. If X and Y are independent, then $g(X)$ and $h(Y)$ are also independent (for any function g and h) .

[3]Brute force calculations give

$$\begin{aligned} \text{cov}\left(2^{-1}(X_1 + X_2), \frac{1}{\sqrt{2}}(X_1 - X_2)\right) &= \frac{1}{2^{3/2}}\left(\text{var}(X_1) - \text{cov}(X_1, X_2) + \text{cov}(X_1, X_2) - \text{var}(X_2)\right) \\ &= \frac{1}{2^{3/2}}\left(\sigma^2 - 0 + 0 - \sigma^2\right) = 0. \end{aligned}$$

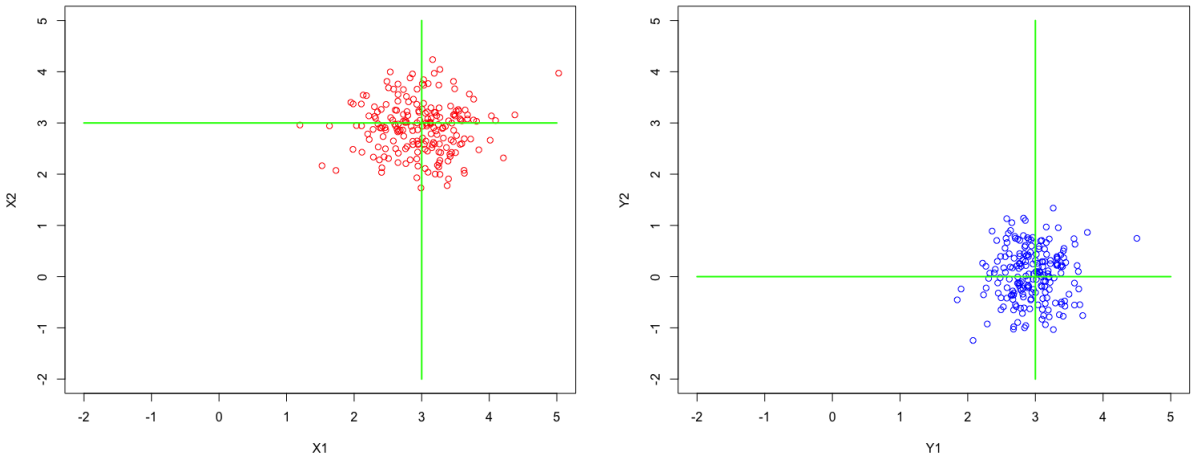


Figure 2.9: Left: Bivariate data (X_1, X_2) from a iid normal $N(3, 0.25)$. Right: Data transformed onto new axis independent normal (but not identically distributed). With $Y_1 = \bar{X}_2$ axis: $N(3, 0.25/2)$ and $Y_2 = Z$ axis: $N(0, 0.25)$. The green lines indicate the means. Observe that $\mathrm{E}[Y_1] = 3$, thus contains information on the mean and $\mathrm{E}[Y_2^2] = 0.25$, thus contains information on the variance.

In summary,

$$\begin{pmatrix} \bar{X} \\ Z \end{pmatrix} \sim MNV_2\left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

See Figure 2.9 for the transformation from $\underline{X}$ to $\underline{Y}$. Observe that the mean of the data cloud gets shifted from $(3, 3)$ to $(3, 0)$ and the spread on Y_1 is slightly less than the spread on Y_2.

Since the variance matrix of $(\bar{X}, Z)$ is diagonal (and the distribution is normal), the distribution of $(\bar{X}, Z)$ are independent random variables.

Remark. *An alternative way to view the above, is as the coefficients of the orthogonal expansion*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = 2^{1/2}\bar{X}_2\underline{e}_1 + Z\underline{e}_2,$$

where $2^{1/2}\bar{X}_2 = \langle \underline{e}_1, \underline{X}_2 \rangle = 2^{-1/2}(X_1 + X_2)$ *and* $Z = \langle \underline{e}_2, \underline{X}_2 \rangle = 2^{-1/2}(X_1 - X_2)$. *Since* $\text{var}[(X_1, X_2)] = \sigma^2 I_2$, *then* $\text{var}[(2^{1/2}\bar{X}_2, Z)] = \sigma^2 I_2$[4].

Rearranging the above expansion gives

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} - 2^{1/2}\bar{X}_2\underline{e}_1 = Z\underline{e}_2 \qquad \Rightarrow \qquad \begin{pmatrix} X_1 - \bar{X}_2 \\ X_2 - \bar{X}_2 \end{pmatrix} = Z\underline{e}_2.$$

Thus by using (1.3) we have

$$(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = Z^2.$$

This gives an alternative proof exactly to the identity (2.5). This method is how we tackle the case $n = 3$ *(and will later be used to prove the ANOVA expressions).*

To summarize, we have shown that $\bar{X}$ and $Z = 2^{-1/2}(X_1 - X_2)$ are independent. This proving that $\bar{X}$ and $s_2^2 = Z^2$ are independent. This proves the last part of Theorem 2.3.

Remark. *The above proof allows us to understand when Theorem 2.3 does not exactly hold.*

(i) Suppose we drop the assumption of Gaussianity and assume that $\{X_i\}$ *are iid random variables with mean* μ *and variance* σ^2 *(but not Gaussian). We still have that*

$$\underbrace{\begin{pmatrix} \bar{X} \\ Z \end{pmatrix}}_{=\underline{Y}_2} = \underbrace{\begin{pmatrix} 1/2 & 1/2 \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}}_{E_2} \underbrace{\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}}_{=\underline{X}_2},$$

where $\text{E}(\underline{Y}_2') = \text{E}(\bar{X}, Z) = (\mu, 0)$ *with variance*

$$\text{var}(\underline{Y}_2) = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}.$$

However, in the Gaussian case $(\bar{X}, Z)$ *are independent because* (X_1, X_2) *are uncorrelated and Gaussian. In the non-Gaussian case* $(\bar{X}, Z)$ *is an uncorrelated vector but not independent. For non-Gaussian random variables orthogonal transformations can induce dependence even though the orthogonal transforms are uncorrelated. This result means that* $\bar{X}_2$ *and* σ_2^2 *are not independent of each other. Further,* Z^2 *is not a* χ_1^2.

(ii) In the case that (X_1, X_2) *are correlated (but could still be Gaussian), the vector* $(\bar{X}, Z)$ *is no longer uncorrelated. Thus even in the case of Gaussianity,* $\bar{X}_2$ *and* s_2^2 *are dependent.*

Before we move on to the case $n = 3$, we briefly discuss the meaning of degrees of freedom and also the relationship of Z with $\bar{X}_2$.

[4]The proof is given previously, but try to do it yourself.

Degrees of freedom

The degree of freedom is ubiquitious in statistics and usually refers to the number of *independent* random variables used to build an estimator. In this example, it is the number used to build the sample variance. When $n = 2$ "it looks like two", but because have to estimate the mean (so one piece of information has already been used), it turns out to be only one. Often df is considered a measure of the "effective" sample size, that is the number of independent random variables required to build an estimator.

Remark (Effective sample size (small digression))**.** *As an extreme consider the sample $X_1, \ldots, X_n$ where $X_i = Z$ for all i (the same random variable for the entire sample). The sample size is n, but the effective sample size (the amount of independent pieces of information the sample contains) is only one.*

Proof of Theorem 2.3 in the case sample size is $n = 3$

Though the idea of the proof for $n = 3$ is similar to that of $n = 2$, the proof is a little more complicated. We give the proof below, pointing out the similarities and the differences. Let

$$s_3^2 = \frac{1}{3-1}\left[(X_1 - \bar{X}_3)^2 + (X_2 - \bar{X}_3)^2 + (X_3 - \bar{X}_3)^2\right]$$

and

$$\bar{X}_3 = \frac{1}{3}(X_1 + X_2 + X_3)\,.$$

Again let

$$\begin{aligned} Y_1 &= (X_1 - \bar{X}_3) = \frac{1}{3}(2X_1 - X_2 - X_3), \qquad Y_2 = (X_2 - \bar{X}_3) = \frac{1}{3}(2X_2 - X_1 - X_3) \\ Y_3 &= (X_3 - \bar{X}_3) = \frac{1}{3}(2X_3 - X_1 - X_2). \end{aligned}$$

It is easily seen that

$$Y_3 = -(Y_1 + Y_2).$$

Since Y_3 is just a linear combination of Y_1 and Y_2, it contains no additional information. Thus the effective sample size of Y_1, Y_2, Y_3 is two (and not three) and

$$s_3^2 = \frac{1}{3-1}\left(Y_1^2 + Y_2^2 + (Y_1 + Y_2)^2\right).$$

This is analogous to the case $n = 2$ (however Y_1 and Y_2 are still dependent).

We have shown that s_3^2 really involves two dependent random variables (and not three), this is a start, but not enough. Our next objective is to show the stronger result, that s_3^2 can be written as the sum of squares two independent normal random variables, each with mean zero and variance σ^2. To do this we first define

the vector $\underline{e}_1 = 3^{-1/2}(1,1,1)$. We project $\underline{X}_3$ onto this vector, because the coefficient of this vector is the sample mean $\bar{X}$: precisely the vector

$$\langle \underline{X}_3, \underline{e}_1 \rangle \underline{e}_1 = 3^{1/2} \bar{X}_3 \underline{e}_1$$

is the projection of the vector $\underline{X}_3 = (X_1, X_2, X_3)$ onto the line in $\mathbb{R}^3$ defined by the vector $\underline{e}_1$. The remainder after the projection onto $\underline{e}_1$ is

$$\underline{X}_3 - 3^{1/2} \bar{X}_3 \underline{e}_1 = \begin{pmatrix} X_1 - \bar{X}_3, & X_2 - \bar{X}_3, & X_3 - \bar{X}_3 \end{pmatrix}.$$

The vector $\underline{X}_3 - 3^{1/2} \bar{X}_3 \underline{e}_1$ is the residuals vector (after removing the sample mean from $\underline{X}_n$)[5]. The residuals vector lies on a plane in $\mathbb{R}^3$ (called a subspace) that is orthogonal to the line defined by the vector $\underline{e}_1$. The residuals vector is commonly studied in statistics, and will usually lie on a subspace of $\mathbb{R}^n$.

The plane in $\mathbb{R}^3$ which is orthogonal to $\underline{e}_1$ (need to draw a picture here) can be described by two orthonormal vectors (which are orthogonal to $\underline{e}_1$). These two vectors are not unique, but for the purpose of illustration we use Example 1.2 and set

$$\underline{e}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}, \qquad \underline{e}_3 = \frac{1}{\sqrt{6}} \begin{pmatrix} -2 & 1 & 1 \end{pmatrix}.$$

Since $\underline{e}_1$, $\underline{e}_2$ and $\underline{e}_3$ are orthonormal, by using Section 1.3.2 we have the representation

$$\begin{aligned} \underline{X}_3 &= \langle \underline{X}_3, \underline{e}_1 \rangle \underline{e}_1 + \langle \underline{X}_3, \underline{e}_2 \rangle \underline{e}_2 + \langle \underline{X}_3, \underline{e}_3 \rangle \underline{e}_3 \\ &= 3^{1/2} \bar{X}_3 \underline{e}_1 + \underbrace{\langle \underline{X}_3, \underline{e}_2 \rangle}_{=Z_2} \underline{e}_2 + \underbrace{\langle \underline{X}_3, \underline{e}_3 \rangle}_{=Z_3} \underline{e}_3. \end{aligned}$$

Thus

$$\underline{X}_3 - 3^{1/2} \bar{X}_3 \underline{e}_1 = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} - 3^{1/2} \bar{X} 3^{-1/2} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ X_3 - \bar{X} \end{pmatrix} = Z_2 \underline{e}_2 + Z_3 \underline{e}_3.$$

Now the interesting part: we show that $Z_1 = \sqrt{3}\bar{X}_3$, $Z_2 = \langle \underline{X}_3, \underline{e}_2 \rangle$ and $Z_3 = \langle \underline{X}_3, \underline{e}_3 \rangle$ are uncorrelated random variables. This result, decomposes $(X_1 - \bar{X}_3, X_2 - \bar{X}_3, X_3 - \bar{X}_3)$ into the sum of <u>two</u> orthogonal vectors, where the coefficients of the vectors are independent. With this we show that $s_3^2 = 2^{-1}(Z_2^2 + Z_3^2)$.

But first we show that Z_1, Z_2 and Z_3 are normally distributed independent random variables. From the definitions of Z_1, Z_2 and Z_3 we have

$$\begin{aligned} \begin{pmatrix} \sqrt{3}\bar{X} \\ \langle \underline{X}_3, \underline{e}_2 \rangle \\ \langle \underline{X}_3, \underline{e}_3 \rangle \end{pmatrix} &= \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ -2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \\ &= E_3 \underline{X}_3. \end{aligned}$$

[5] We observe that $\underline{X}_3 - 3^{1/2} \bar{X}_3 \underline{e}_1$ contains the building blocks of the sample variance s_3^2

Since $\underline{Y}_3' = (Z_1, Z_2, Z_3)$ is a linear combination of Gaussian random variables it is multivariate Gaussian with mean

$$\begin{pmatrix} \mathrm{E}(Z_1) \\ \mathrm{E}(Z_2) \\ \mathrm{E}(Z_3) \end{pmatrix} = \begin{pmatrix} \mu/\sqrt{3} + \mu/\sqrt{3} + \mu/\sqrt{3} \\ 0 + \mu/\sqrt{2} - \mu/\sqrt{2} \\ -2\mu/\sqrt{6} + \mu/\sqrt{6} + \mu/\sqrt{6} \end{pmatrix} = \begin{pmatrix} 3^{1/2}\mu \\ 0 \\ 0 \end{pmatrix}$$

and variance

$$\begin{aligned} \mathrm{var}(\underline{Y}_3) &= \mathrm{var}(E_3 \underline{X}_3) \\ &= \underbrace{E_3 \, \mathrm{var}(\underline{X}_3)}_{=\sigma^2 I_3} E_3^* = \sigma^2 E_3 E_3^*. \end{aligned}$$

Using that $\underline{e}_1, \underline{e}_2$ and $\underline{e}_3$ are orthonormal vectors we have

$$\mathrm{var}(\underline{Y}_3) = \sigma^2 \begin{pmatrix} \underline{e}_1\underline{e}_1' & \underline{e}_1\underline{e}_2' & \underline{e}_1\underline{e}_3' \\ \underline{e}_2\underline{e}_1' & \underline{e}_2\underline{e}_2' & \underline{e}_2\underline{e}_3' \\ \underline{e}_3\underline{e}_1' & \underline{e}_3\underline{e}_2' & \underline{e}_3\underline{e}_3' \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$$

Altogether this gives

$$\begin{pmatrix} \sqrt{3}\bar{X}_3 \\ Z_2 \\ Z_3 \end{pmatrix} \sim \mathcal{N}_3 \left(\begin{pmatrix} 3^{1/2}\mu \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \right)$$

the joint density of $\underline{Y}_3$ is the product of its normal marginal densities. Therefore Z_1, Z_2 and Z_3 are mutually independent. We discuss the role of $Z_1 = \bar{X}_3$ at the end of this subsection and return to

$$\begin{pmatrix} X_1 - \bar{X}_3, & X_2 - \bar{X}_3, & X_3 - \bar{X}_3 \end{pmatrix} = Z_2\underline{e}_2 + Z_3\underline{e}_3,$$

and obtain an expression for the sample variance using the above. By exploiting the orthonormality of the vectors we have

$$\begin{aligned} \sum_{j=1}^{3} (X_j - \bar{X}_3)^2 &= \langle (Z_2\underline{e}_2 + Z_3\underline{e}_3), (Z_2\underline{e}_2 + Z_3\underline{e}_3) \rangle \\ &= \sum_{j_1, j_2 = 2}^{3} Z_{j_1} Z_{j_2} \langle \underline{e}_{j_1}, \underline{e}_{j_2} \rangle \\ &= Z_1^2 \langle \underline{e}_1, \underline{e}_1 \rangle + Z_1 Z_2 \langle \underline{e}_1, \underline{e}_2 \rangle + Z_2 Z_1 \langle \underline{e}_2, \underline{e}_1 \rangle + Z_2^2 \langle \underline{e}_2, \underline{e}_2 \rangle \\ &= Z_2^2 + Z_3^2, \end{aligned}$$

note that above is Parseval's identity given in (1.2). Using this we can easily evaluate the expectation of s_3^2

$$\mathrm{E}[s_3^2] = \frac{1}{3-1} \mathrm{E} \left[\sum_{j=1}^{3} (X_j - \bar{X}_3)^2 \right] = \frac{1}{2} \mathrm{E} \left(Z_2^2 + Z_3^2 \right) = \sigma^2.$$

Since Z_2, Z_3 are iid random variables with mean μ and variance σ^2 we have

$$\frac{2}{\sigma^2}s_3^2 = \frac{1}{\sigma^2}\left(Z_2^2 + Z_3^2\right) = \left(\left[\frac{Z_2}{\sigma}\right]^2 + \left[\frac{Z_3}{\sigma}\right]^2\right) \sim \chi_2^2.$$

This proves Theorem 2.3 for the case $n = 3$. The proof for general n is similar.

Let us summarize the main ingredients of the proof. We have shown that the residuals $(X_1 - \bar{X}_3, X_2 - \bar{X}_3, X_3 - \bar{X}_3)$ lie on a two dimensional plane in $\mathbb{R}^3$. That the residuals can be decomposed as the sum of two random variables Z_2 and Z_3. That $\bar{X} = Z_1$, Z_2 and Z_3 are independent. And that $s_2^2 = 2^{-1}(Z_2^2 + Z_3^2)$. Thus proving Theorem 2.3.

However, it is worth taking a step back on reflecting on what we have shown. The main ingredient is that the observed vector $\underline{X}_3$, whose distribution is

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}_3\left(\begin{pmatrix} \mu \\ \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}\right).$$

is linear transformed such that the result

$$\begin{pmatrix} \sqrt{3}\bar{X}_3 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} \underline{e}_1' \\ \underline{e}_2' \\ \underline{e}_3' \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}_3\left(\begin{pmatrix} 3^{1/2}\mu \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}\right)$$

are independent random variables. But most importantly not identically distributed. Each variable contains different pieces of information. The first variable, $\bar{X}_3$ contains information on the mean, μ. However, Z_2 and Z_3 are identically distributed containing information of the variance but nothing on the mean.

Remark (Connections to Principle Component Analysis)**.** *From the above above you should have noticed that*

$$\underline{X}_3 \quad = \quad 3^{1/2}Z_1\underline{e}_1 + Z_2\underline{e}_2 + Z_3\underline{e}_3.$$

This is a decomposition of $\underline{X}_3$ into three orthonormal vectors whose coefficients are uncorrelated random variables. If $\text{var}(\underline{X}_3) = \sigma^2 I_3$ *($\{X_i\}_{i=1}^3$ are independent random variables), the basis and representation is not unique. However, in the general case that* $\text{var}(\underline{X}_3)$ *is general matrix, then this representation is usually unique and the orthogonal basis $\underline{e}_1$, $\underline{e}_2$ and $\underline{e}_3$ conveys interesting information about "patterns" in the data.*

For the proof in the case $n > 3$ we use the same ideas. We define n-orthonormal vectors $\{\underline{e}_j\}_{j=1}^n$ where the first vector is $\underline{e}_1 = n^{-1/2}(1, 1, 1 \ldots, 1)$ which transform $\underline{X}_n$ to a different basis. The coefficients of $\underline{X}_n$ on this new basis are:

$$\underline{Y}_n = E_n\underline{X}_n = \begin{pmatrix} \underline{e}_1 \\ \underline{e}_2 \\ \vdots \\ \underline{e}_n \end{pmatrix}\underline{X}_n = \begin{pmatrix} \sqrt{n}\bar{X} \\ \langle \underline{e}_2, \underline{X}_n \rangle \\ \vdots \\ \langle \underline{e}_n, \underline{X}_n \rangle \end{pmatrix}\underline{X}_n$$

Since $\{X_i\}$ are iid normal, the joint distribution of $\underline{Y}_n$ is normal with mean $\mathrm{E}[E_n\underline{X}_n] = (\mu, 0 \ldots, 0)'$ and variance $\mathrm{var}(E_n\underline{X}_n) = E_n\mathrm{var}(\underline{X}_n)E_n^* = \sigma^2 E_n E_n^* = \sigma^2 I_n$;

$$\underline{Y}_n \sim \mathcal{N}\left(\begin{pmatrix} \sqrt{n}\mu \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}\right).$$

We use the above construction to prove the result.

The sampling distribution of the sample variance s_n^2

Using Lemma 2.1 and Theorem 2.3 we can obtain the sampling properties of the sample variance s_n^2 defined in (2.5).

We observe from the proof of Theorem 2.3 that there exists $(n-1)$ iid standard normal random variables where

$$\frac{s_n^2}{\sigma^2} = \frac{1}{(n-1)} \underbrace{\left(Z_1^2 + \ldots + Z_{n-1}^2\right)}_{\chi^2_{n-1}}.$$

Thus by using Lemma 2.3 (or properties of expectations) we have $s_n^2 = \sigma^2$ (which we already know) and

$$\mathrm{var}\left(\frac{s_n^2}{\sigma^2}\right) = \frac{1}{(n-1)^2}\mathrm{var}(\chi^2_{n-1}) = \frac{1}{(n-1)^2}2(n-1) = \frac{2}{n-1},$$

where $\mathrm{var}(\chi^2_{n-1}) = 2(n-1)$. Since $\sigma^{-4}\mathrm{var}(s_n^2) = 2/(n-1)$ we have $\mathrm{var}(s_{n-1}^2) = 2\sigma^4/(n-1)$. The standard error associated with $\mathrm{var}(s_{n-1}^2)$ is

$$\sqrt{\frac{2}{n-1}}\sigma^2.$$

Thus as $n \longrightarrow \infty$ the standard error of the variance estimator goes to zero (the estimator improves as the sample size grows). Furthermore, we observe that the distribution of s_n^2 has a $\sigma^2\chi^2_{n-1}/(n-1)$. This is the exact distribution of s_{n-1}^2 (no approximation involved). A plot of the distributions for $n = 11, 26$ and 51 (with $\sigma^2 = 1/2$) in given in Figure 2.10.

We observe from Figure 2.10 that as n grows the distribution s_n^2 coalescing about $\sigma^2 = 1/2$. This fits with exactly how we understand averages to behave. We recall that

$$s_n^2 = \frac{\sigma^2}{(n-1)}\left(Z_1^2 + \ldots + Z_{n-1}^2\right)$$

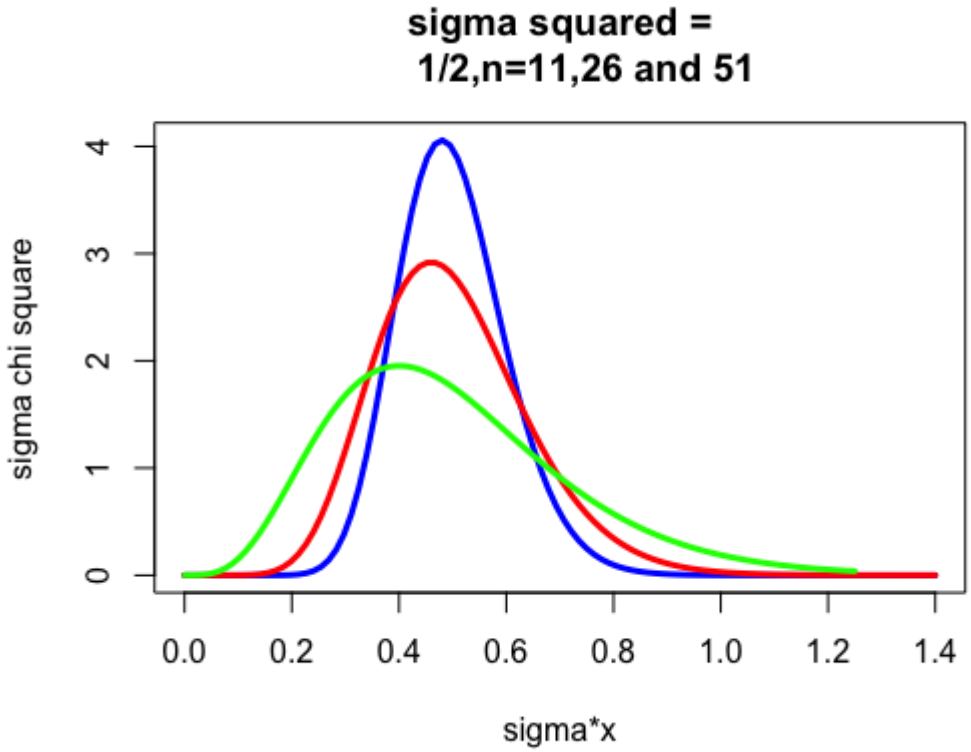


Figure 2.10: Distribution of s_n^2 for $n = 11, 26$ and 51 with $\sigma^2 = 1/2$. Green = 11, Red = 26 and Blue = 51.

is an average of iid random variables (where $\text{var}(Z_i^2) = 2$). Thus its standard error decreases as n grows (this explains the coalescing). Further, from the central limit theorem (Section 1.5), the average of iid random variables are close to normal for large sample sizes. Thus s_n^2 should also be close to normal when n is large. However, Z_i^2 is skewed; the skewness is $S_3 = \sqrt{8}$, which is quite large. This means a relatively large sample size n, is required for s_n^2 to be close to normal. Refer to the example and plots considered at the start of Section 1.5 and Figures 1.8 - 1.11, where the random variables were sampled from a chi-squared distribution with one degree of freedom (these averages have the same distribution as a sample variance of the form s_{n+1}^2).

2.4.3 The t-statistic

In an introductory statistics class you would have studied the t-test and constructed confidence intervals, both of which involve the t-statistic. We will return to these procedures later in the course. But we start by deriving the sampling distribution of the t-statistic.

Suppose $X_1, \ldots, X_n$ are iid, normally distributed random variables with mean μ and standard deviation σ^2. Let $\bar{X}_n$ denote the sample mean, which we have shown is an unbiased estimator of μ. Further, since $\{X_i\}$

are iid normal by using Section 2.1 we can show that

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Since linear transformations preserve normality we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1). \tag{2.8}$$

We now turn our attention to

$$Z = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}. \tag{2.9}$$

Recall that $\sigma/\sqrt{n}$ is the *standard error* which measures the uncertainty we associate with the estimator $\bar{X}$. Z can be treated as a "measure" of distance, between $\bar{X}$ and μ relative to the standard error. Transformations of the type Z are crucial in statistical inference. If Z is too large, then μ is an implausible candidate for the mean. To determine if the distance Z is large or not we require the distribution of Z. Under the normal assumptions on $\{X_i\}$, (2.8) shows that Z is a standard normal free of any parameters (in statistics this is called a pivotal quantity). Observe that if $n = 2$, then $\text{var}[\bar{X}_2] = \sigma^2/2$ and

$$\frac{\sqrt{2}(\bar{X}_2 - \mu)}{\sigma} \sim N(0,1). \tag{2.10}$$

Thus even with $n = 2$, 95% of transformations $\frac{\sqrt{2}(\bar{X}_2-\mu)}{\sigma}$ with lie with $[-1.96, 1.96]$.

However, σ is usually unknown. Thus we replace σ in Z, with the sample standard deviation s_n, defined in (2.5);

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

For example when $n = 2$

$$s_2^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = 2(X_1 - X_2)^2.$$

It is clear that in general this must be a pretty crude (and usually awful) estimator of σ^2 and this will be reflected in the estimation scheme. Using this we define the t-statistic

$$T_n \quad = \quad \sqrt{n}\frac{(\bar{X}_n - \mu)}{s_n}. \tag{2.11}$$

Whereas the distribution of Z is a standard normal, the distribution of t is *not* normal despite the numerator $\sqrt{n}(\bar{X} - \mu)$ being normal. This is because t involves the *ratio* of two random variables; the numerator is a random variable which is normally distributed and the denominator involves a random variables, which we have shown in Theorem 2.3 is the square root of a chi. It is useful to understand the impact of dividing $(\bar{X}_n - \mu)$ by s_n rather than σ.

Example 2.6 (The case $n = 2$). *Suppose X_1 and X_2 are iid normal random variables with mean μ and variance σ^2. Our aim is compare*

$$z = \frac{\sqrt{2}(\bar{X} - \mu)}{\sigma} \quad \text{with} \quad T_2 = \frac{\sqrt{2}(\bar{X} - \mu)}{s_2}$$

where s_2 is the sample variance based on X_1 and X_2. We recall from Theorem 2.3 that $s_2^2 = \sigma^2 \chi_1^2$ (σ^2 times chi-square with one df). We observe that

$$P(s_2 < \sigma) = P\left(s_2^2 < \sigma^2\right) = P\left(\sigma^2 \chi_1^2 < \sigma^2\right) = P(\chi_1^2 < 1) = 0.682 \text{ using tables or R.}$$

and

$$P\left(s_2 < \frac{1}{2}\sigma\right) = P\left(s_2^2 < \frac{1}{4}\sigma^2\right) = P\left(\sigma^2 \chi_1^2 < \frac{1}{4}\sigma^2\right) = P(\chi_1^2 < 1/4) = 0.382 \text{ using tables or R.}$$

Thus means there is a 68.2% chance s_2 underestimates σ and a 38.2% chance s_2 underestimates σ by more than a factor two. This has severe consequences on the t-transform;

$$P(T_2 > z) = P\left(T_2 = \frac{\sqrt{2}(\bar{X} - \mu)}{s_2} > \frac{\sqrt{2}(\bar{X} - \mu)}{\sigma}\right) = P\left(s_2 < \sigma\right) = 0.682$$

and

$$P(T_2 > 2z) = P\left(T_2 = \frac{\sqrt{2}(\bar{X} - \mu)}{s_2} > 2\frac{\sqrt{2}(\bar{X} - \mu)}{\sigma}\right) = P\left(s_2 < \frac{1}{2}\sigma\right) = 0.382.$$

- For small samples, $X_1, \ldots, X_n$ may underestimate the spread, thus s_n^2 under estimates σ.
- Simulate T_2, T_3 and T_{10} and compare this to a standard normal (with histograms and QQplots).

By using Theorem 2.3, under the assumption that $\{X_i\}$ are iid normal, $\bar{X}_n$ and s_n^2 are independent and s_n^2 is the sum of squares of $(n-1)$ independent normal random variables. Thus by dividing the numerator and denominator in T_n by σ we have

$$\begin{aligned} T_n &= \sqrt{n}\frac{(\bar{X} - \mu)/\sigma}{s_n/\sigma} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{s_n/\sigma} \\ &= \frac{Z_0}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1} Z_i^2}} \end{aligned}$$

where $\{Z_i\}_{i=0}^n$ are iid standard normal random variables. Using the definition of the t-distribution in Section 2.2.2, we that $T_n \sim t_{n-1}$. We state this result in the following theorem.

Theorem 2.4

Suppose $\{X_i\}_{i=1}^n$ are iid normal random variables with mean μ and variance σ^2. Let $\bar{X}_n$ and s_n^2 denote

the sample mean and variance respectively. Then we have

$$T_n = \sqrt{n}\frac{(\bar{X}_n - \mu)}{s_n} \sim t_{n-1} \tag{2.12}$$

Example 2.7 (The influence of different sample sizes). *When $n = 2$, then $T_2 \sim t_1$. The sample standard deviation is $s_2^2 = 2(X_1 - X_2)^2$, which could be "too close" to zero. This means that there is a 5% chance that the ratio*

$$T_n = \sqrt{2}\frac{(\bar{X}_2 - \mu)}{\sqrt{2(X_1 - X_2)^2}}$$

will be larger than 12.71 (look up the t-tables); this exactly fits with your calculations in HW4, Q1, where you showed that the tails are so thick, the second moment is infinite! Compare this with (2.10), where the same data is used but the population standard deviation is used in the denominator. It will be extremely rare that $\frac{\sqrt{2}(\bar{X}_2-\mu)}{\sigma}$ is more than 12.71.

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869

However, observe, that as soon as one uses three observations to calculate the standard deviation, the corresponding t-distribution has 2df. The thickness of the tail reduces dramatically from 12.71 *to* 4.30.

Remark. *The t-distribution result is the exact distribution of the t-statistic. If the data is not iid normal, then the result does not hold. However, simulations show that for using the t-distribution as an approximation of the distribution of non-normal data is relatively robust (if the data does not deviate much from normality). See HW4. However, one must careful. Once the assumption of normality of the data is dropped, the there can be huge deviations from the t-distribution. For example, if the data is discrete (for example from a Binomial or Poisson distribution), then there is positive chance that all the data is the same. Resulting in a zero sample standard deviation. Of course, in this case there is a positive probability the t-transform is ∞.*

When the sample size, n, is small the random variable T_n has a larger number of outliers/extremes than the standard normal distribution. You can see this from the t-tables, where the critical value at the 2.5% level for a t-distribution with one df is 12.7 as compared with 1.96 which is the corresponding critical value of a standard normal distribution. However, for large sample sizes the T_n-statistic has a t_{n-1}-dustribution which is close to normality. One can measure how "extremal" a distribution is as compared with the normal distribution using kurtosis, which is defined below.

A QQplot of a t-statistic (generated with iid normal random variables and $n = 3$, replicated 1000 times) against a standard normal distribution is given in 2.11). Observe the signature S shape and that the t-statistic

can take large values (a few of the really extreme t-statistic, larger than 30 were removed to make the plot clearer). To check if it is t-distribution, we plot the same values against a t-distribution with $df = 2$. We observe a close match.

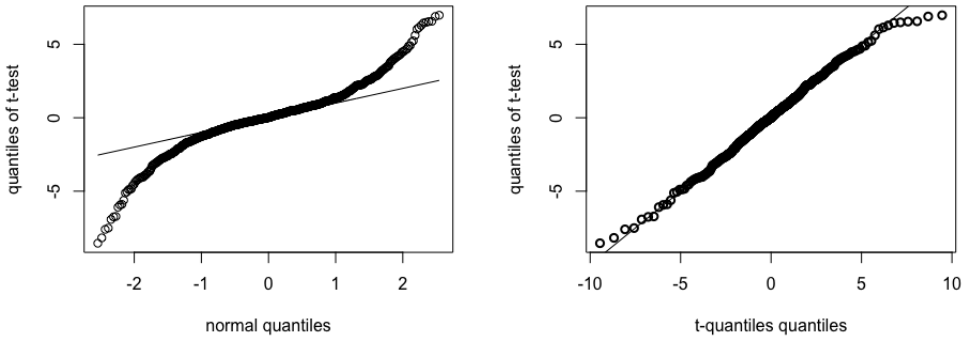


Figure 2.11: The quantiles of the t-statisic, $n = 3$, against a standard normal (left) and t-distribution with df=2 (right).

Kurtosis: Measuring extremes

Kurtosis is a measure of frequency of extremes or outliers in a distribution. The tails of a normal distribution decay very fast, so the chance of outliers is quite slim. We recall that for normally distributed data, "most" observations are with three standard deviations of the mean. For the distributions which have more mass in the tails, this happens more often. One measure for the extremal events is to consider the fourth moment of a distribution (the second moment measures the variance). If a random variable is normally distrubuted with mean μ and variance σ^2 it can be shown that

$$\mathrm{E}[X - \mu]^4 = 3\sigma^4$$

Based on the above, Karl Pearson defined the notion of kurtosis as

$$K_4 = \frac{\mathrm{E}[X - \mu]^4}{\sigma^4}.$$

For a normal distribution $K_4 = 3$. For a distribution with more mass in the tails, K_4 will be larger. Thus often $K_4 - 3$, is called the excess kurtosis. It is used to measure the extremal behaviour of a distribution as compared with the normal distribution. If $K_4 - 3$ is larger than zero, than it has more extremes than a normal distribution (or thicker tails). For a t-distribution with n-df the excess kurtosis is

$$K_4 - 3 = \frac{6}{n-4} \qquad \text{for } n > 4.$$

For $2 \leq n \leq 4$, K_4 does not exist because the tail of the corresponding t-distribution decreases so slow, that large values of X can happen frequently. So frequently that $\mathrm{E}[X^4] = \infty$. To estimate K_4 from the data we replace expectations with their sample means to give the estimator

$$\widehat{K}_4 = \frac{n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^4}{s_n^4}.$$

2.5 Confidence intervals

A $(1-\alpha)100\%$ confidence interval for the parameter θ based on the data $\underline{X} = (X_1, \ldots, X_n)$, is an interval $C_\alpha(\underline{X})$, where

$$P\left(\theta \in C_\alpha(\underline{X})\right) = 1 - \alpha.$$

Note that the interval $C_\alpha(\underline{X})$ is not unique. However, our main requirement is that it for a stated level of confidence it is as narrow as possible.

2.5.1 Confidence interval for the mean

Suppose $\{X_i\}_{i=1}^n$ are iid normally distributed random variables with mean μ and variance σ^2. We have shown that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Thus based on the above (given that σ^2 is known), then

$$\begin{aligned} & P\left(\mu \in \left[\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) = 1-\alpha \\ = \quad & P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \end{aligned}$$

where $z_{\alpha/2}$ is such that $z_{\alpha/2} = P(Z \geq z_{\alpha/2})$ with $Z \sim N(0,1)$. Thus a $(1-\alpha)100\%$ CI for μ is

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

This the classical 95% interval that we see in an intro statistics class. But the interval in not unique. By using the same argument can easily construct (non-symmetric) intervals which have the same level of confidence. The interesting aspect of this interval is that it is the narrowest.

2.5.2 Confidence interval for the variance

Suppose $\{X_i\}_{i=1}^n$ are iid normally distributed random variables with mean μ and variance σ^2. Let $s_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})$. We have shown in Theorem 2.3 that

$$\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Based on the above we have

$$P\left(\frac{(n-1)s_n^2}{\sigma^2} \in \left[\chi^2_{n-1}(1-\alpha/2), \chi^2_{n-1}(\alpha/2)\right]\right) = \alpha$$

where $P(\chi^2_{n-1} \geq \chi^2_{n-1}(\alpha/2)) = \alpha/2$. Rearranging the above gives

$$\begin{aligned}
&= P\left(\chi^2_{n-1}(1-\alpha/2) \leq \frac{(n-1)s_n^2}{\sigma^2} \leq \chi^2_{n-1}(\alpha/2)\right) \\
&= P\left(\frac{\chi^2_{n-1}(1-\alpha/2)}{(n-1)s_n^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{n-1}(\alpha/2)}{(n-1)s_n^2}\right) \\
&= P\left(\frac{\chi^2_{n-1}(1-\alpha/2)}{(n-1)s_n^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{n-1}(\alpha/2)}{(n-1)s_n^2}\right) \\
&= P\left(\frac{(n-1)s_n^2}{\chi^2_{n-1}(\alpha/2)} \leq \sigma^2 \leq \frac{(n-1)s_n^2}{\chi^2_{n-1}(1-\alpha/2)}\right) = \alpha.
\end{aligned}$$

Thus a $(1-\alpha)100\%$ CI for σ^2 is

$$\left[\frac{(n-1)s_n^2}{\chi^2_{n-1}(\alpha/2)}, \frac{(n-1)s_n^2}{\chi^2_{n-1}(1-\alpha/2)}\right].$$

Observe that unlike the interval for the mean, this interval is not symmetric.

2.6 A historical perspective

3 Parameter Estimation

3.1 Introduction

Over the past 50 years, statisticians, computer scientists and engineers have developed an amazing array of algorithms for extracting interesting features from data. The current vogue name for this huge array of algorithms is machine learning. However, it is worth bearing in mind, that the data is collected through "experiments" (either physical experiment, sample surveys etc). If we redo the experiment and collect a new set of data, the numbers in the new data set are unlikely to match the numbers in the old data set. Further the estimates from both data sets are unlikely to be the same. So what exactly are the numbers that we have extracted from the data?

Given that multiple experments give rise to different values, we can treat our observations as random and the estimator as a random variable (Rice, nicely describes this on page 257), this is the same as the multiple trajectories for the sample mean seen in Figure 1.4. Once we understand that for each sample, we obtain an estimator, and these are random variables with a sampling distribution (see Definition 1.8). Then we can start to understand what we are estimating. The estimator is an estimate of a parameter in the sampling distribution, usually its mean. But this is not very informative, without understanding what underpins this distribution. To do this, we make assumptions about how the data is collected. In this course, we will usually assume that the data $\{X_i\}_{i=1}^n$ are iid random variables from a certain distribution, that is a function of an unknown parameter (this is often called the the data generating process). The sampling distribution and the data generating process are closely related. And the estimator is an estimate of a parameter in the data generating process. We often call this the population parameter and treat $\{X_i\}_{i=1}^n$ as iid sample from the population.

Definition 3.1 (Statistical inference)**.** *Statistical inference is the mapping of what we estimate from the data onto the entire population, which is unobserved. Formally, we say we drawing conclusions or making inference, about the underlying population based on the observed sample.*

Recall, that in Section 1.5 we made inference on the population mean based on the sample mean.

In summary, point estimation involves two main steps:

- Find methods and algorithms which allow us to evaluate features in the data. This is called a point

estimate.

- Underlying all algorithms is a model and an unobserved population from which the data is collected. The task of a statistician is to make inference about parameters in the population based on the estimator at hand.

Often to construct an estimator we assume that the observations come from a certain family of distributions. In this chapter, we will assume that the $\{X_i\}$ is an iid sample from a known family of distributions $\{f(x;\theta)\}$ where $\theta = (\theta_1, \ldots, \theta_p)$ are a small number of unknown parameters, which we aim to estimate. The family of distributions is determined either from scientific evidence (how the data was collected) or empirical evidence (for example making a histogram of the data).

In most real life situation, the true distribution is unlikely to be known. We can only make intelligent guesses on what it should be. Keep in mind the famous quotes made by various statistician over the past hundred years: "All models are wrong but some models are useful." Therefore, it is also important to understand what the estimator is actually estimating when the model has not been correctly specified (for example, the data comes from a beta-distribution but we estimate the parameters as if it came from an exponential). However, this analysis is beyond this course.

Let us recall the set-up. We assume that we observe the iid random variables $\{X_i\}_{i=1}^n$ which come from the known family of distributions $\{f(x;\theta)\}$ where $\theta = (\theta_1, \ldots, \theta_p)$ and $\theta \in \Theta$. Θ is called the parameter space, it is the set of all parameters where $f(x;\theta)$ makes sense as a density (it is positive and integrates to one). Our objective is to estimate θ, this estimator is called a point estimator.

3.2 Estimation: Method of moments

The method of moments was developed over a hundred twenty years ago, and is a precursor of the maximum likelihood estimator, described in a later section. It may not be the most efficient estimator (defined in Section 3.8) and is prone to finite sample bias. But it is extremely simple to evaluate and is conceptionally easy to understand. We start with a few motivating examples. A nice description is given in Section 8.4 of Rice.

3.2.1 Motivation

Often use the notation

$$\mu_r = \mathrm{E}(X_i^r)$$

for the rth moment.

(i) Suppose the random variable X is exponentially distributed with density $f(x;\lambda) = \lambda \exp(-\lambda x)$ $(x \geq 0)$, then

$$\mathrm{E}(X) = \frac{1}{\lambda}.$$

Thus

$$\lambda = \frac{1}{\mu_1}.$$

(ii) Suppose the random variables X has a Poisson distribution with probability mass function $f(k;\lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$. Then

$$\mathrm{E}(X) = \lambda.$$

Thus $\lambda = \mu_1$

(iii) Suppose the random variables X has a normal distribution with density $f(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp(-(x-\mu)^2/\sigma^2)$. Then

$$\mathrm{E}(X) = \mu \quad \text{and} \quad \mathrm{E}(X^2) = \sigma^2 + \mu^2.$$

Thus $\mu = \mu_1$ and $\sigma^2 = \mu_2 - \mu_1^2$

(iv) Suppose the random variable X has Gamma distribution distribution with density $f(x;\lambda,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$. Then

$$\mathrm{E}(X) = \frac{\alpha}{\beta} \quad \text{and} \quad \mathrm{E}(X^2) = \frac{\alpha(\alpha+1)}{\beta^2}.$$

Thus

$$\mu_1 = \frac{\alpha}{\beta} \quad \text{and} \quad \mu_2 = \mu_1^2 + \frac{\mu_1}{\beta}.$$

Hence

$$\beta = \frac{\mu_1}{\mu_2 - \mu_1^2} \qquad \alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

Thus the parameters in the distribution are embedded within the moments of the estimators. And we can rewrite the parameters as a function of the moments. Thus by estimating the moments, we can also estimate the parameters by simply substituting the moment estimators into the formula for parameters in terms of the moment. Moments are like means, they can easily be estimated by taking the average. For example, an estimator of $\mu_r = \mathrm{E}[X_i^r]$ based on the iid random variables $\{X_i\}$ is

$$\hat{\mu}_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r.$$

Therefore if $\theta = g(\mu_1, \ldots, \mu_K)$, then the method of moments estimator θ is

$$\widehat{\theta}_n = g(\widehat{\mu}_1, \ldots, \widehat{\mu}_K).$$

3.2.2 Examples

Example 3.1. *(i) Suppose the random variable X is exponentially distributed with density $f(x;\lambda) = \lambda\exp(-\lambda x)$, then a moments estimator of λ is*

$$\widehat{\lambda}_n = \frac{1}{\widehat{\mu}_1}.$$

(ii) Suppose the random variables X has a Poisson distribution with probability mass function $f(k;\lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$. Then $\widehat{\lambda}_n = \widehat{\mu}_1$.

(iii) Suppose the random variables X has a normal distribution with density $f(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp(-(x-\mu)^2/\sigma^2)$. Then $\widehat{\mu} = \widehat{\mu}_1$ and $\widehat{\sigma}^2 = \widehat{\mu}_2 - \widehat{\mu}_1^2$

(iv) Suppose the random variable X has Gamma distribution distribution with density $f(x;\beta,\alpha) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)$. Then

$$\widehat{\beta}_n = \frac{\widehat{\mu}_1}{\widehat{\mu}_2 - \widehat{\mu}_1^2} \qquad \widehat{\alpha}_n = \frac{\widehat{\mu}_1^2}{\widehat{\mu}_2 - \widehat{\mu}_1^2}.$$

Method of moment estimators are usually not unique, an example is given below.

Example 3.2. *Consider the exponential distribution $f(x;\lambda) = \lambda\exp(-\lambda x)$. Basic algebra gives*

$$\begin{aligned}
\mathrm{E}(X^r) &= \lambda\int_0^\infty x^r\exp(-\lambda x)dx \\
&= \frac{\lambda}{\lambda^r}\int_0^\infty (\lambda x)^r\exp(-\lambda x)dx \text{ (change variables } y = \lambda x) \\
&= \frac{1}{\lambda^r}\int_0^\infty y^r\exp(-y)dy \\
&= \frac{1}{\lambda^r}\Gamma(r+1),
\end{aligned}$$

where $\Gamma(r+1)$ is the gamma function (and does not depend on λ). Thus for all $r \geq 1$ we have

$$\lambda = \left(\frac{\Gamma(r+1)}{\mu}\right)^{1/r}$$

Based on the above, for any r we can use

$$\widehat{\lambda}_{r,n} = \left(\frac{\Gamma(r+1)}{\widehat{\mu}_r}\right)^{1/r}$$

as an estimator of λ. Question: which moments estimator should one use?

The take home message from the above example is that moments estimator do not have to depend on the first few moments. They can depend on higher order moments, the can also depend on moments of the transformed data, for example even $\{\log X_i\}$.

3.2.3 Sampling properties of method of moments estimators

In this section we derive the sampling properties of the method of moments estimators. We start by considering the estimator of the rate λ in the exponential distribution. From the previous section we observe that the moments estimator is

$$\widehat{\lambda}_n = \frac{1}{\bar{X}_n}.$$

We simulate the estimator in the case that $\lambda = 1.5$ and a plot of 5 trajectories ($n = 1, \ldots, 100$) is given in Figure 3.1. We observe that they all appear to converge to the truth $\lambda = 1.5$ as $n \to \infty$. Indeed this should be the case. Since $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$, by using Lemma 1.2 we have that $\widehat{\lambda}_n \xrightarrow{\mathcal{P}} 1/\mu$. This means the method of moments estimator is asymptotically consistent. In general it will converge to the true parameter as the sample size grows. This is true for all method of moment estimators.

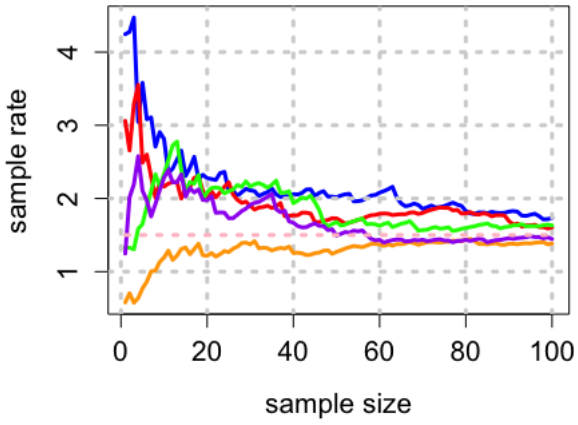


Figure 3.1: 5 trajectories of $\widehat{\lambda}_n$. The pink dashed line is the truth = 1.5.

To obtain the sampling distribution and variance of $\widehat{\lambda}_n$ we use the results in Section 1.5.4. We observe that $\widehat{\lambda}_n = g(\bar{X}_n)$, thus by using (1.9), on transformations of sample means we can obtain the asymptotic distribution of $\widehat{\lambda}_n$. Since $g'(\mu) = -1/\mu^2 \neq 0$ we have

$$\sqrt{n}(\widehat{\lambda}_n - \lambda) \xrightarrow{\mathcal{D}} N\left(0, \frac{\text{var}(X)}{\mu^4}\right) = N(0, \lambda^2) \tag{3.1}$$

as $n \to \infty$. Thus the asymptotic sampling variance of the estimator is

$$\sigma^2_{\widehat{\lambda}_n} = \frac{\lambda^2}{n}.$$

The corresponding standard error for $\widehat{\lambda}_n$

$$\frac{\lambda}{\sqrt{n}}.$$

Of course this standard error is quite useless if we want to use it to construct confidence intervals, since λ is unknown. However, we overcome this issue by replacing λ with its estimator $\widehat{\lambda}_n$ to yield the estimated (asymptotic) sampling variance

$$\widehat{\sigma}^2_{\widehat{\lambda}_n} = \frac{\widehat{\lambda}^2_n}{n}.$$

Since we know by Lemma 1.2 that $\widehat{\lambda}_n \xrightarrow{\mathcal{P}} \lambda$, then we have $\widehat{\lambda}^2_n \xrightarrow{\mathcal{P}} \lambda^2$, thus for sufficiently large n, $\widehat{\sigma}^2_{\widehat{\lambda}_n}$ is a good approximation of $\sigma^2_{\widehat{\lambda}_n}$.

To see how good the above approximations of the sampling distribution of $\widehat{\lambda}_n$ and its true finite sample variance we conduct some simulations. For $n = 3, \ldots, 100$ we simulate from an exponential with $\lambda = 1.5$ and evaluate $\widehat{\lambda}_n$ we do this 500 times.

Comparing the asymptotic and empirical variance

We evaluate the empirical variance:

$$\widehat{\sigma}^2_n = \frac{1}{500}\sum_{i=1}^{500}(\widehat{\lambda}_{i,n} - \bar{\lambda}_n)^2 \quad \text{with} \quad \bar{\lambda}_n = \frac{1}{500}\sum_{i=1}^{500}\widehat{\lambda}_{i,n} \tag{3.2}$$

this should be close to the true variance (since it was done over 500 replications). In Figure 3.2 we plot $n\widehat{\sigma}^2_n$ against n. Recall that asymptotic sampling variance is λ^2/n. It is not the "true" finite sample variance (which can usually only be evaluated through simulations). However, we would expect that $n\widehat{\sigma}^2_n$ (defined in (3.2)) to be "close" to λ^2 for "large" n. We do see that this is the case, for $n > 40$ they are closely aligned, but for smaller n the match is not so close.

Comparing the asymptotic and finite sample sampling distributions

We observe a similar effect for the sampling distribution of the estimators. We recall that the result in (3.1) is an asymptotic approximation of the true finite sample distribution $\widehat{\lambda}_n$. In other words, for a sufficiently large n

$$\widehat{\lambda}_n \sim N\left(\lambda, \frac{\lambda^2}{n}\right).$$

To see how close this approximation is to the truth, in Figures 3.3-3.5 we make a histogram of the estimates (evaluated over 500 replication) together with a histogram of the corresponding normal approximation. We observe that for sample size $n = 5$ the true sampling distribution of $\widehat{\lambda}_5$ is right skewed and normal approximation is really quite bad. For sample size $n = 20$, the true sampling distribution has less of a right skew, and the sampling distribution is better. And for sample size $n = 100$ there appears to be a close match in the true sampling distribution and the normal approximation. In summary, for relatively large sample sizes the asymptotic variance λ^2/n and normal approximation are quite good approximations of the true

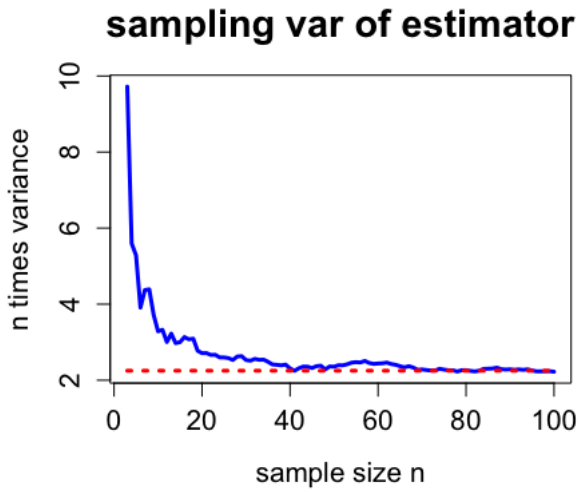


Figure 3.2: A plot of the sampling variance times the sample size for n=3,...,100. The blue line is $n\widehat{\sigma}_n^2$, the red line is the $\lambda^2 = 1.5^2 = 2.25$.

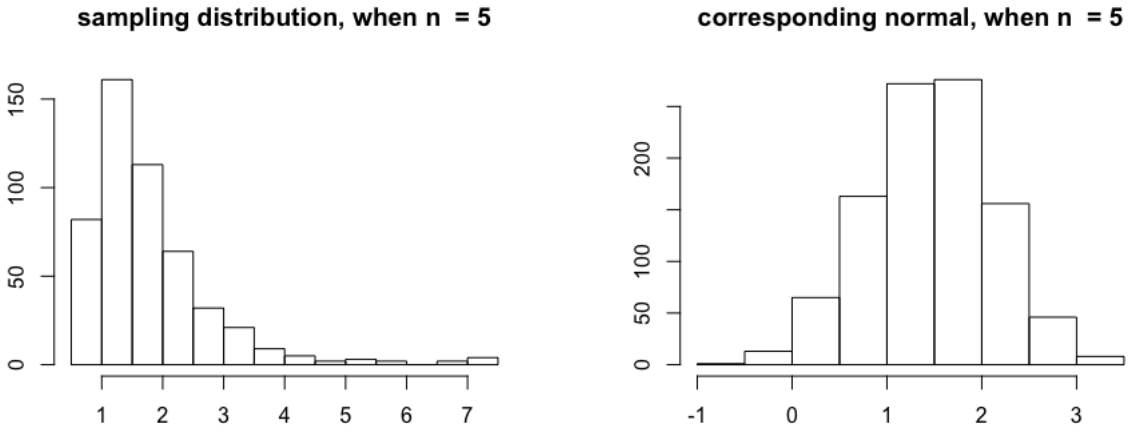


Figure 3.3: Expnential distribution. Left: True sampling distribution when n = 5. Asymptotic normal distribution $N(1.5, 2.25/5)$

variance and sampling distribution of the estimator. But for small sample sizes some caution is required when applying these approximations (to constructing confidence intervals and testing). As can be clearly seen the quality of the normal approximation as asymptotic variance is not so good when the sample size is small.

The above results concern the method of moments estimator corresponding to the exponential distribution. However, similar results also hold for general method of moments estimators. Some examples be given in

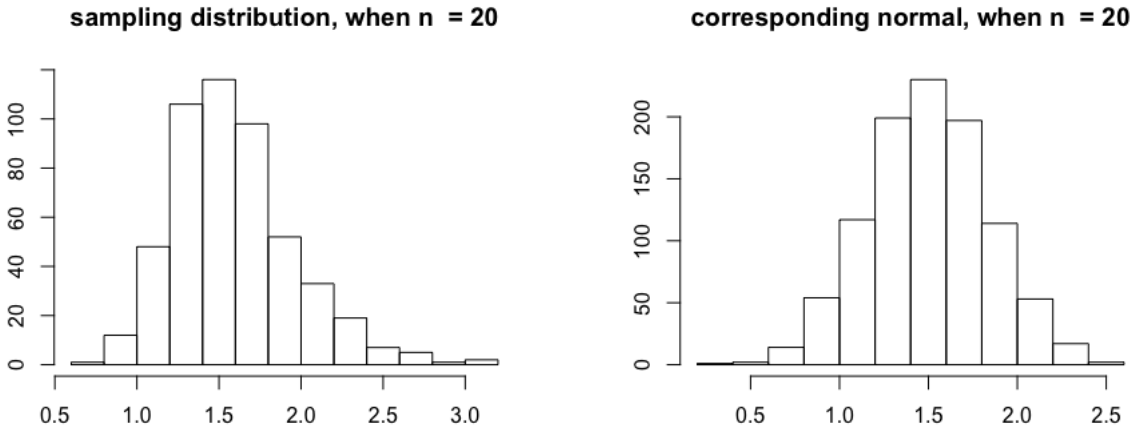


Figure 3.4: Exponential distribution. Left: True sampling distribution when n = 20. Asymptotic normal distribution $N(1.5, 2.25/20)$

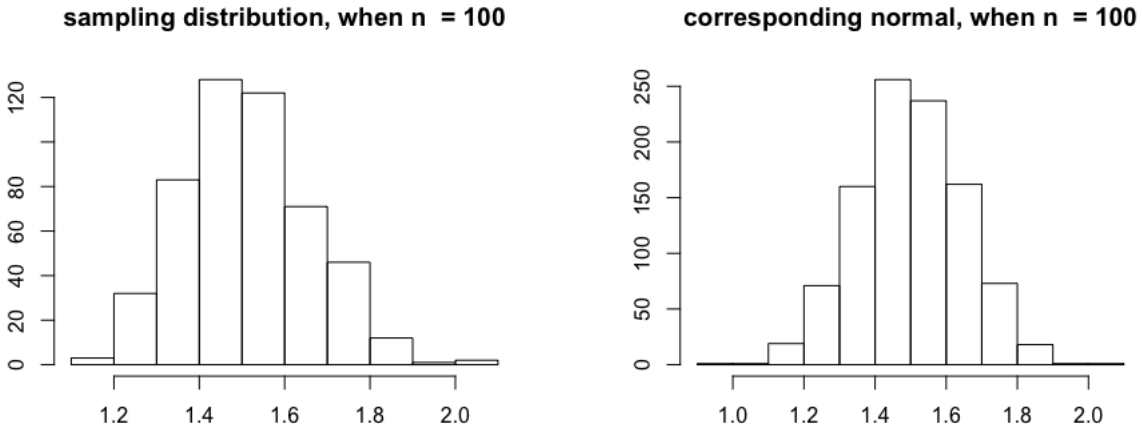


Figure 3.5: Exponential distribution. Left: True sampling distribution when n = 100. Asymptotic normal distribution $N(1.5, 2.25/100)$

your homework. However, to get more practice, in the section below we consider the asymptotic sampling properties of the moments estimator of the gamma distribution.

Sampling properties of the Gamma distribution

Suppose that $\{X_i\}$ are iid random variables with Gamma distribution with density $f(x; \beta, \alpha) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$. Suppose (for simplicity) that β is known. Then the method of moments estimator can simply be constructed

using the first moment since

$$\mathrm{E}[X] = \frac{\alpha}{\beta} \qquad \alpha = \beta \mathrm{E}[X].$$

This yields the moment estimator

$$\widehat{\alpha}_n = \beta \bar{X}_n$$

of α. Using Theorem 1.1 we have

$$\sqrt{n}\left(\widehat{\alpha}_n - \alpha\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \beta^2 \mathrm{var}(X)\right), \tag{3.3}$$

where we note that $\mathrm{var}(X) = \alpha/\beta^2$. Thus the asymptotic variance of $\widehat{\alpha}_n$ is α.

3.2.4 Application of asymptotic results to the construction of confidence intervals

From a practical perspective the reason the asymptotic sampling properties of the moments estimators are of interest is to construct confidence intervals for the parameter we are estimating. We recall from elementary statistics that if $\{X_i\}$ are iid random variables with mean μ and variance σ^2, then the sample mean $\bar{X}_n$ satisfies the following distributional results

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2).$$

Thus for sufficiently large n we roughly have

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

This result is used to construct the $(1-\alpha)100\%$ confidence interval for the mean μ

$$\left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

The idea being that if $\bar{X}$ is normally distributed, then for every 100 intervals constructed on average $(1-\alpha)100\%$ of them would contain the population mean μ. To understand this, in Figure 3.6, left plot, we construct 100 95% confidence intervals (for $n = 3$ based on iid normal observations). We observe that in this simulation, 3 out of 100 do not contain the population mean. If the population variance is unknown we replace σ with the estimator s_n, this induces additionally variability. We observe in Figure 3.6 (right plot) that by simply replacing σ with s_3 (but still using iid normal observations) and still using the normal distribution to construct the CI means that the CI has less confidence than the 95% that is stated (18 confidence intervals out of 100 do not contain the mean). Since we know that for normal data $T_n = \sqrt{n}(\bar{X} - \mu)/s_n \sim t_{n-1}$, when constructing the CI we replace the normal distribution with the t-distribution with $(n-1)$-df:

$$\left[\bar{X}_n - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}\right].$$

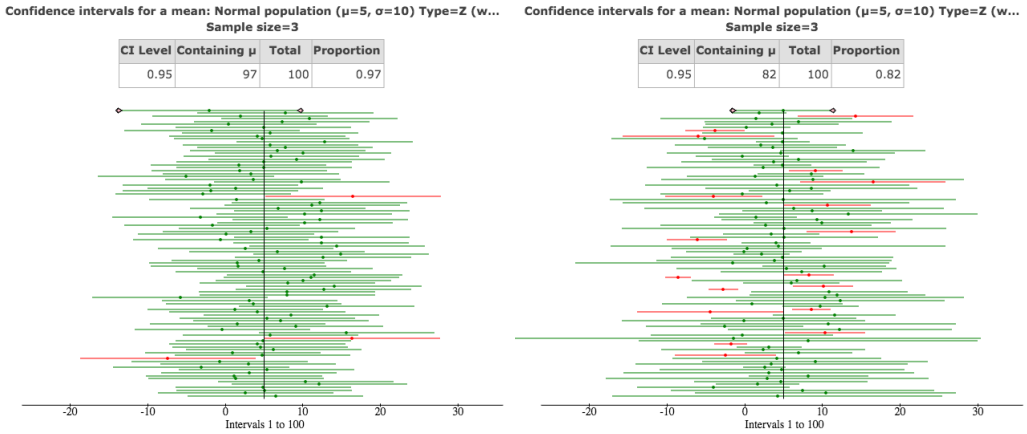
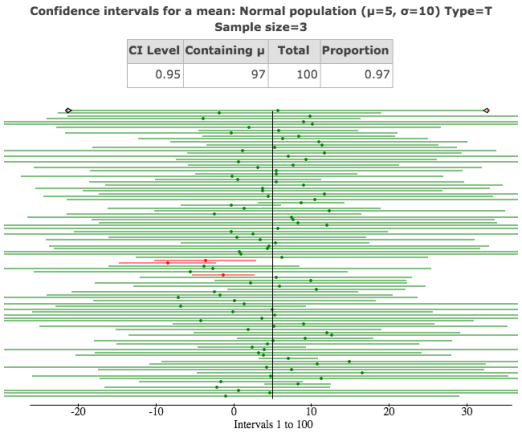


Figure 3.6: Confidence intervals μ using $\bar{X}$, where $X_i \sim N(\mu = 5, \sigma^2 = 10^2)$ and $n = 3$ (over 100 replications). Left: Confidence interval constructed using z-values and $\sigma = 10$ is used. Right: Left: Confidence interval constructed using z-values and s_3 is used.



To see how effective using the t-distribution is, in the plot on the right, we simulate 100 confidence intervals (using iid normal observations, sample size $n = 3$). We use s_n and the t-distribution to construct the CI. We observe that out 3 out of 100 do not contain the confidence interval. Hence the t-distribution does improve the "coverage" of the confidence interval.

For non-normal data and small sample sizes neither

$$\left[\bar{X}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s_n}{\sqrt{n}}\right] \quad \text{nor} \quad \left[\bar{X}_n - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}\right]$$

is truely a $(1-\alpha)100\%$ for the mean μ.

We now apply the same ideas described above to obtain confidence intervals parameter estimators. We observe the iid random variables $\{X_i\}_{i=1}^n$ which has density $f(x;\theta)$. Suppose the method of moments estimator of θ is $\widehat{\theta}_n$ and it can be shown that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} N\left(0, \sigma_\theta^2\right).$$

For a sufficiently large n we roughly have

$$\widehat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{\sigma_\theta^2}{n}\right).$$

This result is used to construct the $(1-\alpha)100\%$ confidence interval for θ;

$$\left[\widehat{\theta}_n - z_{\alpha/2}\frac{\sigma_\theta}{\sqrt{n}}, \widehat{\theta} + z_{\alpha/2}\frac{\sigma_\theta}{\sqrt{n}}\right].$$

Note that often we have to estimate $\widehat{\sigma}_\theta$, in which case we replace σ_θ with its estimator $\widehat{\sigma}_\theta$ to yield the interval

$$\left[\widehat{\theta}_n - z_{\alpha/2}\frac{\widehat{\sigma}_\theta}{\sqrt{n}}, \widehat{\theta} + z_{\alpha/2}\frac{\widehat{\sigma}_\theta}{\sqrt{n}}\right].$$

Observe that we did not replace $z_{\alpha/2}$ with the corresponding t-distribution, because it is not really clear if the t-distribution is able to model correctly the additional uncertainty caused by replacing σ_θ with its estimator $\widehat{\sigma}_\theta$.

3.3 Monte Carlo methods and correcting for the lack of normality

We now investigate how to estimate the finite sample distribution of estimators. We focus on the method of moment estimators, but the methods described below also apply to many other estimators.

3.3.1 The parametric Bootstrap

For many estimators asymptotic normality can be shown. But as demonstrated in Figures 3.3-3.5 (and the variance estimator in Figure 3.2) this approximation is not very reliable when the sample size is small. The plots given on the left hand side of Figures 3.3-3.5 are the histograms corresponding to the true sampling distribution and the blue line in Figure 3.2 is the true variance. In an ideal world we would use this distribution and variances for inference (constructing confidence intervals etc, we cover this in a later chapter). But the simulations are based on the simulating the exponential distribution with true underlying parameter (look at the R code in `Chapter3Rcode`). In reality the true parameter is unknown (else we would not be estimating it).

To get round this problem we can estimate the finite sampling distribution by simulating from the distribution using the estimated parameter. To demonstrate what we mean by this, we return to the exponential distribution.

- The idea is to sample from an exponential distribution with sample size $n = 20$.
- Suppose we observe

$$\begin{aligned} \text{demo} \quad = \quad & 0.5475, 0.0089, 1.1269, 0.7519, 0.5628, 0.8547, 1.9941, 0.7383, 0.0529, 0.0243 \\ & 0.5029, 0.4628, 0.3951, 0.7799, 2.3609, 1.3204, 0.1754, 1.5920, 1.0729, 0.7659, \end{aligned}$$

which are drawn from an exponential distribution (with $\lambda = 1.5$, which is treated as unknown). The sample mean is 0.805 and the method od moments estimator of λ is $\widehat{\lambda}_{20} = \bar{X}^{-1} = 1.243$.

- The bootstrap step is based on simulating from an exponential distribution with $\lambda = 1.243$ for $n = 20$ and calculate $\widehat{\lambda}^*_{20}$ from these simulated values. We use the $*$ notation to denote the fact that $\widehat{\lambda}^*_{20}$ is estimated from simulated or the bootstrap sample.

  We repeat this several times (say 500 times), storing $\widehat{\lambda}^*_{20}$ for each replication.

- A plot of the estimated histogram together with the true distribution is given in Figure 3.7. Further the QQplot of the estimated distribution against true distribution (calculated based on the true parameter $\lambda = 1.5$) is given in Figure 3.8. We observe a relatively close match, though there is a slight shift.

- Based on this simulation, the estimated variance is 0.115, whereas the true sampling variance of $\widehat{\lambda}_{20}$ is 0.135.

- For every sample, the estimated distribution will change a little.

From the demonstration above and Figures 3.7 and 3.8 we observe that the Monte Carlo method appears to capture the right skew in the sampling distribution of $\widehat{\lambda}_{20}$, which the normal approximation given in Figure 3.4 clearly does not.

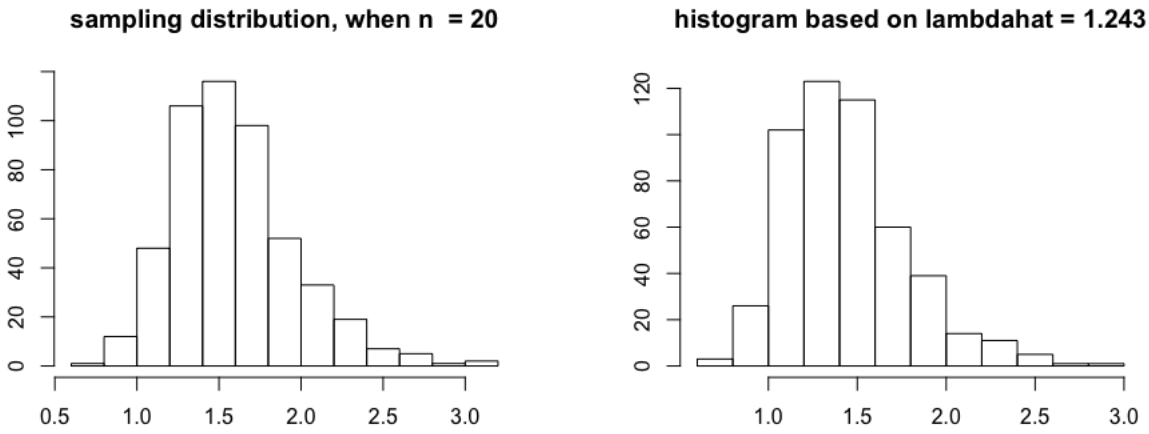


Figure 3.7: Left: True sampling distribution when n = 20. Right: Distribution based on sampling from estimated exponential with $\widehat{\lambda}_n = 1.243$.

The Monte Carlo method described above can be generalised to many distributions. The "take home" message is that we replace the unknown parameter with the estimated parameter, when conducting the replications. Monte Carlo methods are also nicely described in Rice, Section 8.4, page 264-265.

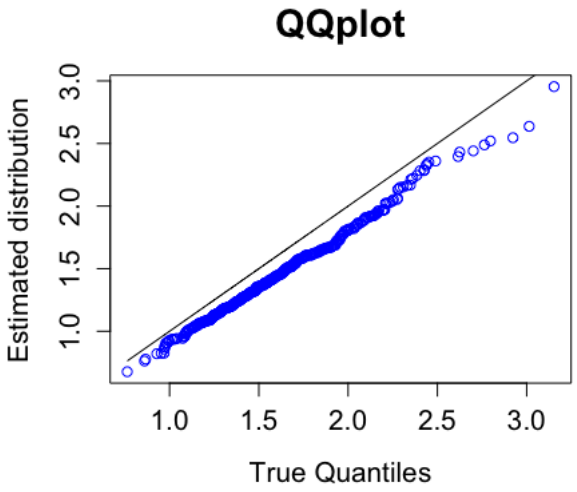


Figure 3.8: QQplot of quantiles of true distribution of $\widehat{\lambda}_n$ (calculated through simulations) against the estimated distribution of $\widehat{\lambda}_n$ based on the estimated parameter.

3.3.2 The nonparametric Bootstrap

Monte Carlo methods, as described, above are very useful. However, by sampling from an exponential distribution we make the assumption the observed data has really been drawn from an exponential distribution. If this is not the case, then this method will not give a good approximation of the sampling distribution of the estimator. It can be completely wrong.

There are more general methods, which allow for what we called misspecification of the distribution. These are nonparametric methods (where the distribution is not assumed known). This entails sampling not from the conjectured distribution (such as the exponential) but sampling (with replacement) from the data itself. This is often called the nonparametric bootstrap. It is robust to misspecification of the distribution. A bootstrap sample in R can be obtained using the command `sample(data,replace = T)`.

We outline the nonparametric bootstrap for the exponential example described in the previous section.

- Start with the data

$$\begin{aligned} \text{demo} \quad = \quad & 0.5475, 0.0089, 1.1269, 0.7519, 0.5628, 0.8547, 1.9941, 0.7383, 0.0529, 0.0243 \\ & 0.5029, 0.4628, 0.3951, 0.7799, 2.3609, 1.3204, 0.1754, 1.5920, 1.0729, 0.7659, \end{aligned}$$

- Sample from demo with replacement. For example, one bootstrap sample is

```
> temp = sample(demo,replace = T)
> temp
 [1] 0.3480 0.4494 0.3645 3.5009 0.9539 0.1917 0.7354 0.8323 0.2294 0.6017 0.8323 2.2859
```

```
[13] 0.1917 0.7354 0.4627 0.2294 0.1917 0.0348 0.1917 0.4627
```

The estimator corresponding to the above is $\widehat{\lambda}^*_{20} = 1.446$.

- We repeat the bootstrap procedure described above. For each sample we evaluate $\widehat{\lambda}^*_{20}$ and store it. We do this many times (I did it 500, but the more the better).
- A histogram of the bootstrap estimates of $\widehat{\lambda}_{20}$ is given in Figure 3.9. A QQplot using the bootstrap quantiles against the true quantiles is given in Figure 3.10.
- The bootstrap estimated variance of $\widehat{\lambda}_{20}$ is 0.169, whereas the true sampling variance of $\widehat{\lambda}_{20}$ is 0.135 (calculated using simulation).

  A possible reason for the over estimation in the bootstrap standard error is that there is slightly more spread in the original data set than the spread of the true, underlying exponential distribution. The nonparametric bootstrap mimics the properties of the data from which it samples from.
- For every sample demo, the estimated bootstrap distribution will change (as this distribution is random and depends on the original sample).

Based on this one sample, the nonparametric bootstrap seems to estimate the true sampling distribution quite well. It is comparable to the parametric bootstrap described in the previous section.

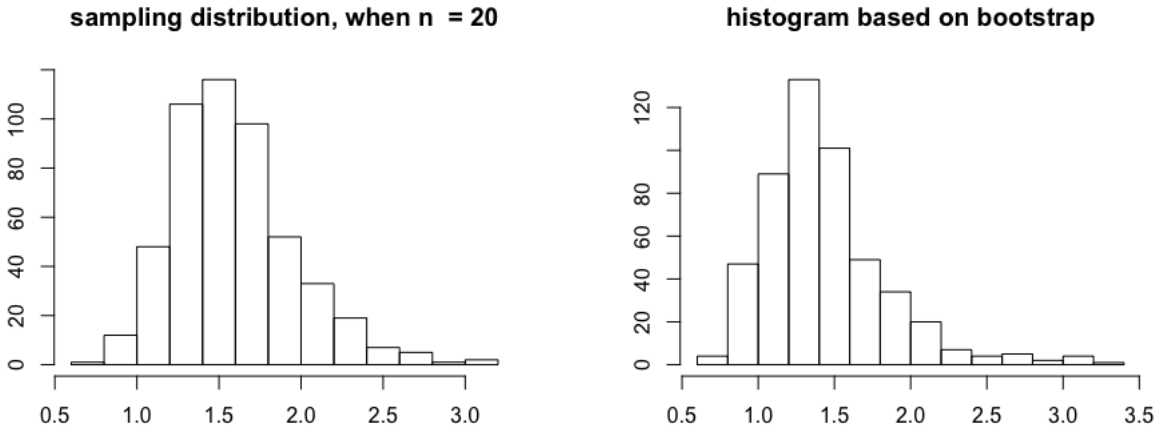


Figure 3.9: Left: True sampling distribution when n = 20. Right: Distribution based on the boostrap samples.

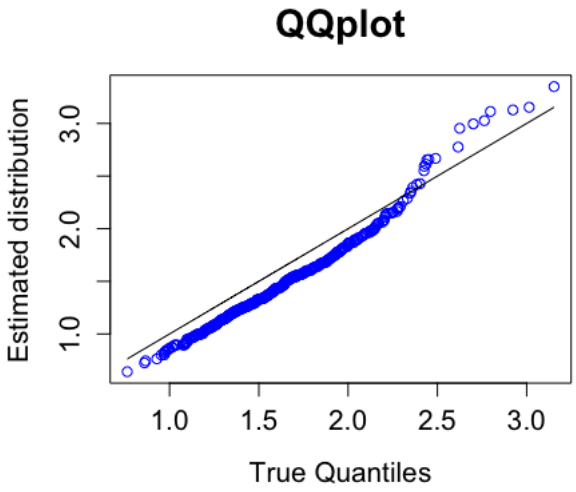


Figure 3.10: QQplot of quantiles of true distribution of $\widehat{\lambda}_n$ (calculated through simulations) against the estimated distribution of $\widehat{\lambda}_n$ based on the estimated parameter.

Using the bootstrap to construct confidence intervals

We use the above idea to estimate the distribution of the z-transform

$$z = \frac{\sqrt{n}(\widehat{\lambda}_n - \lambda)}{\widehat{\lambda}_n}, \tag{3.4}$$

which can be used to estimate the critical values in a 95% CI. The above z-transform is often called a *pivotal* statistic.

- The data:

$$\begin{aligned} \text{demo} \quad = \quad & 0.5475, 0.0089, 1.1269, 0.7519, 0.5628, 0.8547, 1.9941, 0.7383, 0.0529, 0.0243 \\ & 0.5029, 0.4628, 0.3951, 0.7799, 2.3609, 1.3204, 0.1754, 1.5920, 1.0729, 0.7659, \end{aligned}$$

- Use the data to estimate λ; $\widehat{\lambda} = 1/\bar{x} = 1.242$.
- Next we sample from demo with replacement. For example, one bootstrap sample is

```
> temp = sample(demo,replace = T)
> temp
 [1] 0.3480 0.4494 0.3645 3.5009 0.9539 0.1917 0.7354 0.8323 0.2294 0.6017 0.8323 2.2859
[13] 0.1917 0.7354 0.4627 0.2294 0.1917 0.0348 0.1917 0.4627
```

The estimator corresponding to the above is $\widehat{\lambda}^*_{20} = 1.446$. The corresponding bootstrap z-transform

corresponding to (3.4) replaces λ with the estimate from the data $\widehat{\lambda} = 1/\bar{x} = 1.242$ and $\widehat{\lambda}$ with $\widehat{\lambda}^*_{20}$:

$$z^* = \frac{\sqrt{n}(\widehat{\lambda}^*_{20} - \widehat{\lambda})}{\widehat{\lambda}^*_{20}}.$$

- We repeat the bootstrap procedure described above. For each sample we evaluate z^* and store it. We do this many times (I did it 500, but the more the better). Using these samples we have the distribution function of z^*, $\widehat{F}^*(u)$.
- A histogram of the bootstrap estimates of z^* is given in Figure 3.11 (in the case the true $\lambda = 1.5$.).
- Next calculate the 2.5% and 97.5% quantiles using the quantile function in R. For this example it is -1.63 and 1.39 respectively. We use this to construct the 95% confidence interval for λ with

$$\left[\widehat{\lambda}_{20} - 1.63 \times \frac{\widehat{\lambda}_{20}}{\sqrt{20}}, \widehat{\lambda}_{20} + 1.39 \times \frac{\widehat{\lambda}_{20}}{\sqrt{20}}\right]$$

  with $\widehat{\lambda}_{20} = 1.24$.

Warnings It is tempting to believe that the bootstrap is the true distribution and not an estimator of it. We emphasize that the bootstrap distribution of z^* is only an *estimator* of the distribution of the standardized statistic in (3.4). Indeed in the example in Figure 3.11 the bootstrap seems to underestimate the spread. The bootstrap quantile of 2.5% and 97.5% is -1.63 and 1.39 whereas the "truth" is -2.29 and 1.70 respectively.

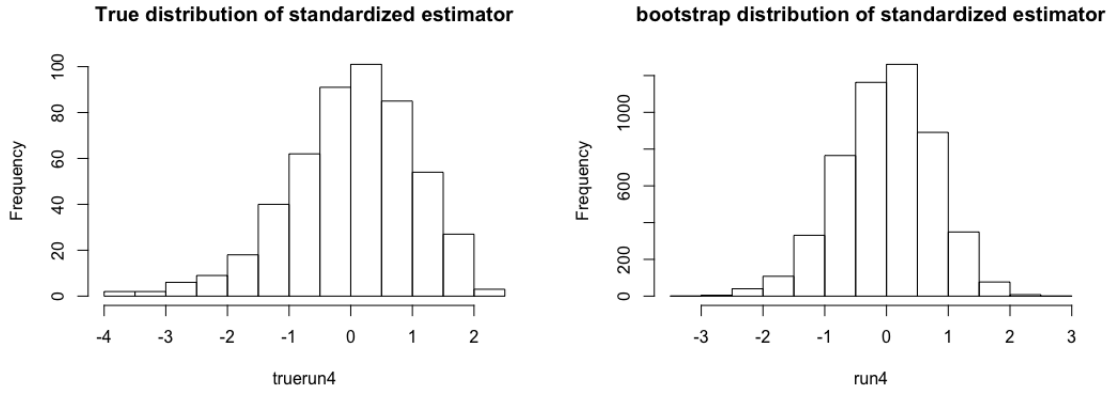


Figure 3.11: Left: Distribution of (3.4) for $\lambda = 1.5$ and $n = 20$ (over a few thousand replications). Right: Bootstrap estimate of the distribution of (3.4) for $\lambda = 1.5$ (over 2000 bootstrap replications) based on the data demo.

The bootstrap is conceptionally simple to understand. It can be shown that the bootstrap is able to capture the skewness of the (finite sampling) distribution of the estimator but this is an asymptotic result. The actually details are quite delicate (using Edgeworth expansion) and is beyond this class.

3.3.3 The power transform approach

As mentioned above the bootstrap is able to capture in the skewness in the sampling distribution of the estimator (which is useful in inference). An alternative approach, is transform the estimator in such a way to remove the skewness in the sampling distribution.

We recall from HW2, Question 3, that for skewed random variables, the sampling distribution of the sample mean tends to be skewed for small sample sizes. As seen from the simulations above, bootstrap methods tend to capture the skewness. An alternative approach is to make a power transform, for the form $\widehat{\theta}_n^\alpha$, similiar to that described in HW2, question 3. By using Lemma 1.2 we have

$$\sqrt{n}\left(\widehat{\theta}_n^\alpha - \theta^\alpha\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, [\alpha\theta^{\alpha-1}]^2\sigma_\theta^2\right).$$

But we also recall from HW2, question 3, that for $\alpha < 1$ the power transform tends to reduce the skewness of the parameter and obtain a better approximate the normal distribution. Thus we can construct a 95% confidence interval for θ^α using

$$\left[\widehat{\theta}_n^\alpha - 1.96 \times \frac{[\alpha\theta^{\alpha-1}]^2\sigma_\theta^2}{n^{1/2}}, \widehat{\theta}_n^\alpha + 1.96 \times \frac{[\alpha\theta^{\alpha-1}]^2\sigma_\theta^2}{n^{1/2}}\right] = [L, U].$$

Since $\theta = (\theta^\alpha)^{1/\alpha}$, a 95% confidence interval for θ is $[L^{1/\alpha}, U^{1/\alpha}]$.

Of course, it is necessary to select the optimal α, Chen and Deo (2004) propose a method for selecting the "best" α based on minimising the skewness of the sampling distribution (beyond this course).

3.4 Estimation: Maximum likelihood (MLE)

3.4.1 Motivation

Empirical evidence suggests that the life time of an incandescent light bulb follows an exponential distribution with density $f(x;\lambda) = \lambda\exp(-\lambda x)$ with mean lifetime λ^{-1}. It is well known that the light bulb manufactoring industry did their level best to ensure that light bulbs did not last "too long", if it lasted too long it would reduce profits (sounds rather like what manufacturers today and goes contrary to the circular economy, but I digress). An interesting summary is give in wiki and here.

Suppose a manufacturer has three options on their machine for producing light bulbs with a mean life time of 750 hours, 1000 hours or 1250 hours. However, the settings on the machine are stuck and the writing has worn away so noone knows the setting it is stuck at. It can be either 750 hours, 1000 hours or 1250 hours. To figure it out, 20 light bulbs are produced in the stuck setting machine and their lifetimes measure. The

data is summarized below.

$$\begin{aligned}\underline{x} = \; & 202.5, \quad 962.4, \quad 342.4, \quad 596.1, \quad 1331.8, \quad 902.8, \quad 501.7, \quad 1393.1, \quad 620.8, \quad 1604.2, \quad 241.0 \\ & 372.6, \quad 143.0, \quad 1420.5, \quad 74.7, \quad 2342.4, \quad 1072.3, \quad 1309.2, \quad 1650.7, \quad 163.7\end{aligned}$$

The sample mean is 862.4, which is slightly closer to 1000 than 750. But simply comparing the sample means is really not enough. It makes sense to consider the joint distribution of the data, which we believe comes from an exponential distribution. Under the assumption that the lifetime of the light bulbs are independent of each other the joint density is the product of the marginals

$$\begin{aligned} f_{\underline{X}}(\underline{x};\lambda) &= \prod_{i=1}^{20} \lambda \exp\left(-\lambda x_i\right) = \lambda^{20} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) = \lambda^{20} \exp\left(-\lambda 20 \bar{x}_{20}\right) \\ &= \lambda^{20} \exp\left(-\lambda \times 20 \times 862.4\right). \end{aligned}$$

But unlike applications in probability, we are not using the joint density to calculate the probability of an event. The data is observed, it cannot change. Our aim is to select the λ based on the observed data; the λ that most likely to give the observed data. The idea in statistics is that a "typical" sample is most likely to be drawn from the maximum of the density (close to the peak). Thus from a statistical perspective $f_{\underline{X}}(\underline{x};\lambda)$ is a likelihood function, and is a function of λ rather than $\underline{x}$. Often to emphasis the dependence on λ we rewrite the $f_{\underline{X}}(\underline{x};\lambda)$ as

$$L(\lambda;\underline{x}) \quad = \quad = \lambda^{20} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) = \lambda^{20} \exp\left(-\lambda \times 20 \times 862.4\right).$$

A plot of $L(\lambda;\underline{x})$ is given in Figure 3.12 (see the left hand plot). However, often to make the plot easier to read the logarithm of the likelihood is used

$$\mathcal{L}(\lambda;\underline{x}) \quad = \quad = 20 \log \lambda - \lambda \sum_{i=1}^{n} x_i = 20 \log \lambda - 862.4\lambda.$$

Since log is a monotonic transform, it does not change the the characterisics in $L(\lambda;\underline{x})$. The log-likelihood is given on the right hand side of Figure 3.12. As the potential candidates are mean= $\lambda^{-1} = 750, 1000, 1250$, these correspond to the red vertical lines in the plots. From the plot we observe that $\lambda^{-1} = 1250$ is unlikely given the observed data. Visually it is difficult to tell the difference between 750, 1000, but the actual log-likelihoods at these values are given in the table below.

λ^{-1}	750	100	1250
log-likelihood	-155.399	-155.403	-156.416

We observe that just by a very small margin, $\lambda^{-1} = 750$ maximises the likelihood over $\lambda^{-1} = 1000$. Thus based on the observed data set, using the likelihood criterion we make the decision that the setting is stuck at 750 hours. In fact in this case, we have made the correct decision, the sample was generated from an exponential with mean $\lambda^{-1} = 750$. But we do observe from the plots in Figure 3.12 that the likelihood

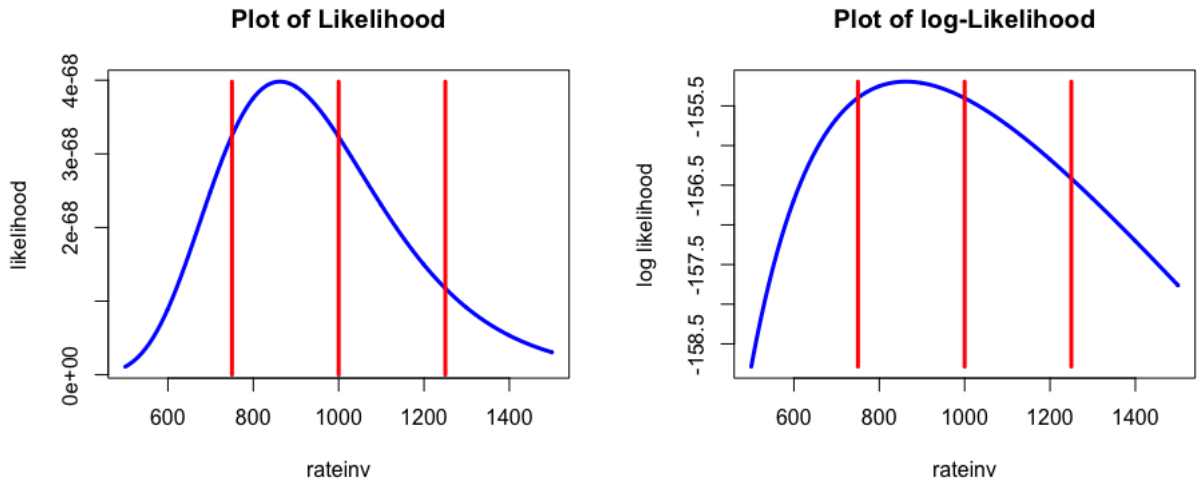


Figure 3.12: Left: Likelihood. Right: log-likelihood.

(and equivalently log-likelihood) is maximised at $\lambda^{-1} = 862.4$ (the sample mean). Thus if any value of λ^{-1} were a potential candidate for the mean, using $\lambda^{-1} = 862.4$ or equivalently $\lambda = 862.4^{-1}$ appears to be the most likely. This is exactly the maximum likelihood estimator of λ based on the exponential distribution (interestingly it is the same as the method of moments estimator too).

We now formally define the likelihood for iid random variables. Let us suppose that $\{X_i\}_{i=1}^n$ are iid random variables with density $f(x;\theta)$, the likelihood is defined as

$$L_n(\theta;\underline{X}) = \prod_{i=1}^{n} f(X_i;\theta)$$

and the log-likelihood is the logarithm of the likelihood

$$\mathcal{L}_n(\theta;\underline{X}) = \sum_{i=1}^{n} \log f(X_i;\theta).$$

The maximum likelihood estimator (denoted from now on as MLE) is defined as parameter which maximises the likelihood. If you like maths notation we say

$$\widehat{\theta}_{mle} = \arg\max_{\theta\in\Theta} L_n(\theta;\underline{X}) = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta;\underline{X}),$$

where Θ denotes the parameter space (see Section 2.3, recall it is all parameters where $f(x,\theta)$ is a density or probability mass function). The MLE is said to be the estimator that is most consistent with the observed data (it is most likely given the data).

Remember the data vector $\underline{X}$ is observed, so we are maximising over the unknown parameter θ.

Often it is easier reparametrize a distribution. For example, for the exponential distribution $f(x;\lambda) = \lambda\exp(-\lambda x)$ we require that $\lambda > 0$, but if we rewrite $\lambda = \exp(\gamma)$, then $\gamma \in (-\infty,\infty)$ (the parameter space of γ

is not restricted). In this reparametrized world we can write the exponential distribution as

$$g(x;\gamma) = \exp(\gamma)\exp\left[-\exp(\gamma)x\right].$$

In the following lemma we show that if the mapping $\lambda \implies \exp(\gamma)$ is one-to-one and onto (a bijection/invertible), then the MLE of γ and λ yield (after transformation) the same values.

Lemma 3.1 (The invariance property)**.** *The invariance property of the MLE says that it makes no difference which parameterization we use for inding the MLE. If $\widehat{\theta}$ is the MLE of θ and $g(\theta)$ is a one-to-one (and onto) function (invertible), then $g(\widehat{\theta})$ is the MLE of $g(\theta)$.*

The proof is straightforward. We may not use this property so much in this course, but is very useful in applied statistics. For example, in generalized linear models etc.

3.4.2 Examples

We now give some examples of distributions and their MLE. We note that in general to maximise the likelihood

$$L(\theta;\underline{X}) = \prod_{i=1}^{n} f(X_i;\theta) \text{ equivalently } \mathcal{L}_n(\theta;\underline{X}) = \sum_{i=1}^{n} \log f(X_i;\theta),$$

over the parameter space Θ, we differentiate $L(\theta;\underline{X})$ with respect to θ and set the derivative to zero and solve for θ (there are some issues which arise if the maximum happens at the boundary of the parameter space, which is beyond this course).

We also evaluate the second derivative at the solution to ensure it is the (local maximum, local because there could be a few maximums). If $\frac{d^2L(\theta;\underline{X})}{d\theta^2}$ is negative (or negative definite), then we can be sure we have "caught" the (local) maximum. Often it will not be possible to obtain an explicit expression for the maximum, and we discuss strategies on dealing with this in a later section.

The Normal/Gaussian distribution

Suppose $\{X_i\}$ are iid random variables, then the Gaussian likelihood is

$$L(\theta;\underline{X}_n) = \prod_{i=1}^{n}\left\{\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)\right\} = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}\right].$$

Often (especially for the exponential family of distributions) it is easier to take the logarithm and maximise that

$$\mathcal{L}_n(\theta;\underline{X}_n) = \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{irrelevant}} - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}.$$

Evaluating the partial derivative of $\mathcal{L}_n(\theta; \underline{X})$ with respect to μ and σ^2 and setting to zero gives

$$\begin{aligned}\frac{\partial \mathcal{L}_n(\theta; \underline{X}_n)}{\partial \mu} &= \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma^2} = 0 \\ \frac{\partial \mathcal{L}_n(\theta; \underline{X})}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^4}.\end{aligned}$$

Solving the above gives

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i^2 \qquad \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2,$$

which is the same as the method of moments estimator.

An alternative derivation of the MLE (which on first appearances does not appear to have any advantages) is based on transforming the data vector $\underline{X}_n = (X_1, \ldots, X_n)$. It is uses the transformation proof described in Section 2.4.2 used to prove Theorem 2.3 together with identity (2.2) (orthonormal transformations of a normally distributed vector). We recall, we had defined n-orthonormal vectors $\{\underline{e}_j\}_{j=1}^n$ where the first vector is $\underline{e}_1 = n^{-1/2}(1, 1, 1 \ldots, 1)$ which transform $\underline{X}_n$ to a different basis. The coefficients of $\underline{X}_n$ on this new basis are:

$$\underline{Y}_n = E_n \underline{X}_n = \begin{pmatrix} \underline{e}_1 \\ \underline{e}_2 \\ \vdots \\ \underline{e}_n \end{pmatrix} \underline{X}_n = \begin{pmatrix} \sqrt{n}\bar{X} \\ \langle \underline{e}_2, \underline{X}_n \rangle \\ \vdots \\ \langle \underline{e}_n, \underline{X}_n \rangle \end{pmatrix}.$$

Since $\{X_i\}$ are iid normal, the joint distribution of $\underline{Y}_n$ is normal with mean $\mathrm{E}[E_n \underline{X}_n] = (n^{1/2}\mu, 0 \ldots, 0)'$ and variance $\mathrm{var}(E_n \underline{X}_n) = E_n \mathrm{var}(\underline{X}_n) E_n^* = \sigma^2 E_n E_n^* = \sigma^2 I_n$;

$$\underline{Y}_n \sim \mathcal{N}\left(\begin{pmatrix} \sqrt{n}\mu \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \right)$$

Since $\{Y_j\}$ are independent the joint density of $\{Y_j\}$ is the product of the marginals:

$$\begin{aligned} L_n(\mu, \sigma^2; \underline{Y}_n) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Y_1 - \sqrt{n}\mu)^2 \right) \prod_{j=2}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} Y_j^2 \right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\sqrt{n}\bar{X} - \sqrt{n}\mu)^2 \right) \prod_{j=2}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} Y_j^2 \right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{n}{2\sigma^2}(\bar{X} - \mu)^2 \right) \prod_{j=2}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} Y_j^2 \right). \end{aligned}$$

This leads to the log-likelihood of the transformed data

$$\mathcal{L}_n(\theta; \underline{Y}_n) = \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{irrelevant}} - \frac{n}{2}\log\sigma^2 - \frac{n(\bar{X}_n - \mu)^2}{2\sigma^2} - \frac{\sum_{i=2}^n Y_i^2}{2\sigma^2}. \tag{3.5}$$

Further, by using (2.2) we have that the transformed data $\underline{Y} = E\underline{X}$ and the original data $\underline{X}$ have exactly the same likelihood;

$$\begin{aligned}\mathcal{L}_n(\theta; \underline{Y}_n) &= \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{irrelevant}} - \frac{n}{2}\log\sigma^2 - \frac{n(\bar{X}_n - \mu)^2}{2\sigma^2} - \frac{\sum_{i=2}^n Y_i^2}{2\sigma^2} \\ &= \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{irrelevant}} - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2} = \mathcal{L}_n(\theta; \underline{X}_n)\end{aligned} \tag{3.6}$$

Thus we can obtain the MLE of $\underline{Y}_n$ without any loss in information. Observe the first entry of $\underline{Y}_n$ contains information on the mean, but the rest do not. Differentiating this likelihood with respect to μ and σ^2 gives

$$\begin{aligned}\frac{\partial \mathcal{L}_n(\theta; \underline{Y}_n)}{\partial \mu} &= \frac{n(\bar{X}_n - \mu)}{\sigma^2} = 0 \\ \frac{\partial \mathcal{L}_n(\theta; \underline{X})}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{(\bar{X}_n - \mu)^2}{2\sigma^4} + \frac{\sum_{i=2}^n Y_i^2}{2\sigma^4}.\end{aligned}$$

Solving the above gives the MLE

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i^2 \qquad \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=2}^n Y_i^2.$$

Now on first impression it would appear that the estimators are different, since $\frac{1}{n}\sum_{i=2}^n Y_i^2$ looks different to $n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2$. But using equation (1.3) in Section 1.3.2 we have that

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=2}^n Y_i^2 = n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2.$$

Thus the MLE estimators for both transformations of the data are the same. Which it will always be. An advantage of the transformed data is that the mean information is contained in only coefficient, whereas the variance information is encrypted in the other coefficients of the vector.

Take home message: So long as the matrix E_n is known (and non-singular) no information is lost or gained in the transform $E_n\underline{X}_n$. But in terms of estimation sometimes it is easier to deal with the transformed data.

Note that the MLE estimator of σ^2 has a bias (which goes away for large n). We know from Theorem 2.3 that the unbiased estimator of σ^2 is

$$s_n^2 = \frac{1}{(n-1)}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2.$$

Observe that the MLE coincides with the moments estimator for the normal distribution.

Exponential distribution

We now derive the MLE of the exponential distribution. Suppose that $\{X_i\}$ are iid exponential, then the likelihood is

$$L(\theta; \underline{X}_n) = \prod_{i=1}^{n} [\lambda \exp(-\lambda X_i)] = \lambda^n \exp(-\lambda \sum_{i=1}^{n} X_i).$$

The log likelihood is

$$\mathcal{L}_n(\theta; \underline{X}_n) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i. \tag{3.7}$$

Differentiating the above wrt λ gives

$$\frac{d\mathcal{L}_n(\theta; \underline{X}_n)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0.$$

Thus the MLE is

$$\widehat{\lambda}_n = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n}.$$

Again, the MLE coincides with the moments estimator for the normal distribution.

The Poisson distribution

We now derive the MLE of the Poisson distribution. Suppose that $\{X_i\}$ are iid poisson, then the likelihood is

$$L(\theta; \underline{X}_n) = \prod_{i=1}^{n} \frac{\lambda^{X_i} \exp(-\lambda)}{X_i!}.$$

The log likelihood is

$$\mathcal{L}_n(\theta; \underline{X}_n) = \sum_{i=1}^{n} [X_i \log \lambda - \lambda - \log X_i] = 0.$$

Differentiating the above wrt λ gives

$$\frac{d\mathcal{L}_n(\theta; \underline{X}_n)}{d\lambda} = \lambda^{-1} \sum_{i=1}^{n} X_i - n.$$

This gives the MLE is

$$\widehat{\lambda}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n.$$

Again, the MLE coincides with the moments estimator for the normal distribution.

The Gamma distribution

We now derive the MLE of the gamma distribution. Suppose that $\{X_i\}$ are iid gamma, then the likelihood is

$$L(\theta; \underline{X}_n) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} X_i^{\alpha-1} \exp(-\beta X_i)$$

The log-likelihood is

$$\mathcal{L}_n(\theta; \underline{X}_n) = \sum_{i=1}^{n} \left[(\alpha - 1)\log X_i - \beta X_i + \alpha \log \beta - \log \Gamma(\alpha)\right].$$

Differentiating the above wrt β and α gives

$$\begin{aligned} \frac{\partial \mathcal{L}_n}{\partial \beta} &= \sum_{i=1}^{n} \left[-X_i + \frac{\alpha}{\beta}\right] = 0 \\ \frac{\partial \mathcal{L}_n}{\partial \alpha} &= \sum_{i=1}^{n} \left[\log X_i + \log \beta - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}\right] = 0. \end{aligned}$$

Thus, solving the above we can rewrite the solution of β in terms of α (this is often called profiling)

$$\widehat{\beta}(\alpha) = \frac{\alpha}{\bar{X}_n}.$$

Substituting this into the next derivative gives

$$\sum_{i=1}^{n} \left[\log X_i + \log \frac{\alpha}{\bar{X}_n} - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}\right] = 0.$$

But to solve the above we need to use a numerical routine.

We note that the above MLE does not coincide (in an obvious way) to the method of moments estimator described in Section 3.2.1.

Research 3. *In many of the examples above we showed that the MLE in the exponential family coincides with the method of moments estimators (the exception is the Gamma). This is not a coincidence, and there are reasons for it.*

The curious amongst you may want to investigate why. Hint: As a start you may want to evaluate the expectation of X and $\log X$ for the Gamma distribution. Then study the properties of moments of the exponential (natural) family.

The Uniform distribution

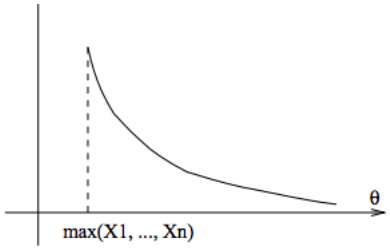
Let us suppose $\{X_i\}$are iid random variables with a uniform distribution with density

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta} \qquad x \in [0,\theta] \\ f(x;\theta) &= 0 \qquad x \notin [0,\theta]. \end{aligned}$$

We can write the density as $f(x,\theta) = \theta^{-1} I_{[0,\theta]}(x)$, where $I_{[0,\theta]}(x)$ is an indicator function which is one between $[0,\theta]$ and zero elsewhere (please draw it). Based on this the likelihood is

$$L(\theta, \underline{X}) = \left[\prod_{i=1}^{n} \theta^{-1} I_{[0,\theta]}(X_i)\right].$$

Now recall that in a likelihood we treat the data as fixed and the parameter θ as variable. Thus $L(\theta, \underline{X}_i)$ is a function of θ. The log-likelihood has no meaning. Instead, it is easier to just directly maximise the likelihood. Since the data is kept fixed and $L(\theta, \underline{X})$ is a function of θ, we observe (by making a plot) that $L(\theta, \underline{X})$ is zero for any $X_i \leq \theta \leq X_{i+1}$. Thus $L(\theta, \underline{X})$ is zero for $\theta \leq \max_i X_i$ (see below).



Since the data is kept fixed and $L(\theta, \underline{X})$ is a function of θ, we observe (see plot on the left) that $L(\theta, \underline{X})$ is zero for any $X_i \leq \theta \leq X_{i+1}$. Thus $L(\theta, \underline{X})$ is zero for $\theta \leq \max_i X_i$. But for $\theta > \max_i(X_i)$, it drops at the rate θ^{-n}. The way to visualize this is to think of a simple data set $X_1 = 2$, $X_2 = 2.5$ and $X_3 = 3.2$. $L(\theta, \underline{X}) = 0$ if $\theta < 3.2$ (since this event cannot happen), thus θ must be 3.2 or greater. And the MLE if 3.2.

Thus, $L(\theta, \underline{X})$ is maximised at $\theta = \max_i(X_i)$, and the maximum likelihood estimator for the uniform distribution is

$$\widehat{\theta}_n = \max_{1 \leq i \leq n} X_i.$$

Inflated Poisson distribution

The Poisson distribution is often used to model count data. A plot of a Poisson mass function for different values of λ is given in Figure 3.13. We observe that when λ is small, there is a large mass at zero, then it rapidly drops. On the other hand if λ is large, then the mass tends to be very small at zero. This gives a dichotomy in modelling. Either the probabilities will be large at zero and small elsewhere, or small at zero and large elsewhere. Such data can arise, but often we require more flexibility in the distribution.

There arises many real life situations where the chance of observing zero outcomes is very high, but also the chance of observing 5, 6, 7, can also be high. The regular Poisson cannot model both these behaviour simultaneuously. But by mixing two distributions more flexible characteristics can be achieved. For example, suppose

$$\begin{aligned} U &\in \{0,1\} \text{ a Bernoulli random variable} \\ V &\in \{0,1,2,\ldots\} \text{ a Poisson random variable} \end{aligned}$$

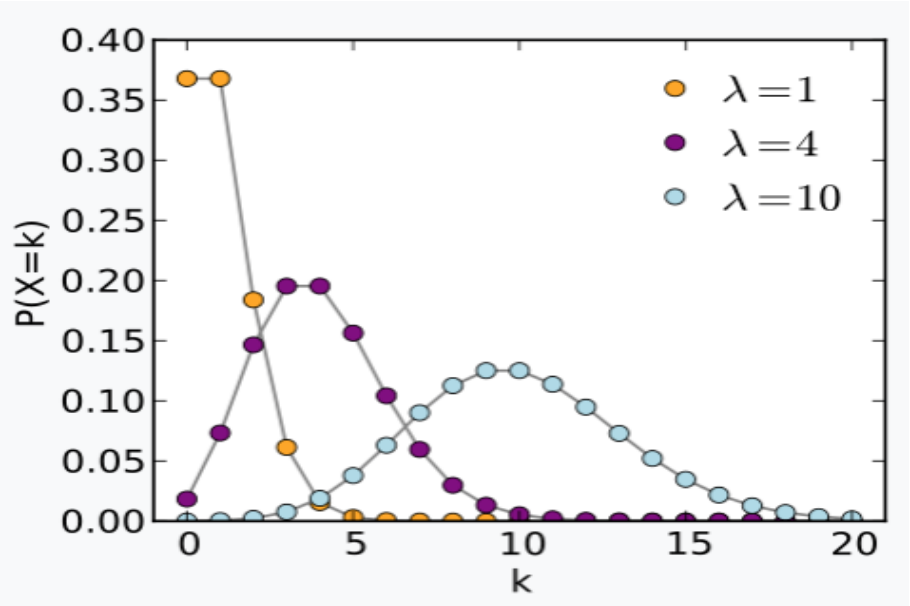


Figure 3.13: Left: Mass function of the Poisson distribution (stolen from Wiki)

where

$$\begin{aligned} P(U=0) &= p \qquad P(U=1) = 1-p \\ P(V=k) &= \frac{\lambda^k \exp(-\lambda)}{k!} \end{aligned}$$

We define a new random variables which mixes the Bernoulli and the Poisson:

$$X = UV.$$

We observe that

$$\begin{aligned} P(X=0) &= P(X=0|U=0)P(U=0) + P(X=0|U=1)P(U=1) \\ &= p + \exp(-\lambda)(1-p) \\ P(X=k) &= \underbrace{P(X=k|U=0)}_{=0} P(U=0) + P(X=k|U=1)P(U=1) \\ &= \frac{\lambda^k \exp(-\lambda)(1-p)}{k!} \qquad k = 1, 2, \ldots \end{aligned}$$

Thus combining the two sets of probabilities we can write

$$P(X=k) = [p + \exp(-\lambda)(1-p)]^{I(k=0)} \left[\frac{\lambda^k \exp(-\lambda)(1-p)}{k!}\right]^{1-I(k=0)} \qquad k = 0, 1, 2, \ldots$$

where $I(\cdot)$ is an indicator variable: $I(k=0)=1$ of $k=0$ and $I(k=0)=0$ if $k\neq 0$ (it is like the `if` and `else` function when we code). A plot for different λ is given in Figure 3.14. Observe that even for large λ, the probability at zero can be large as well as "far" from zero.

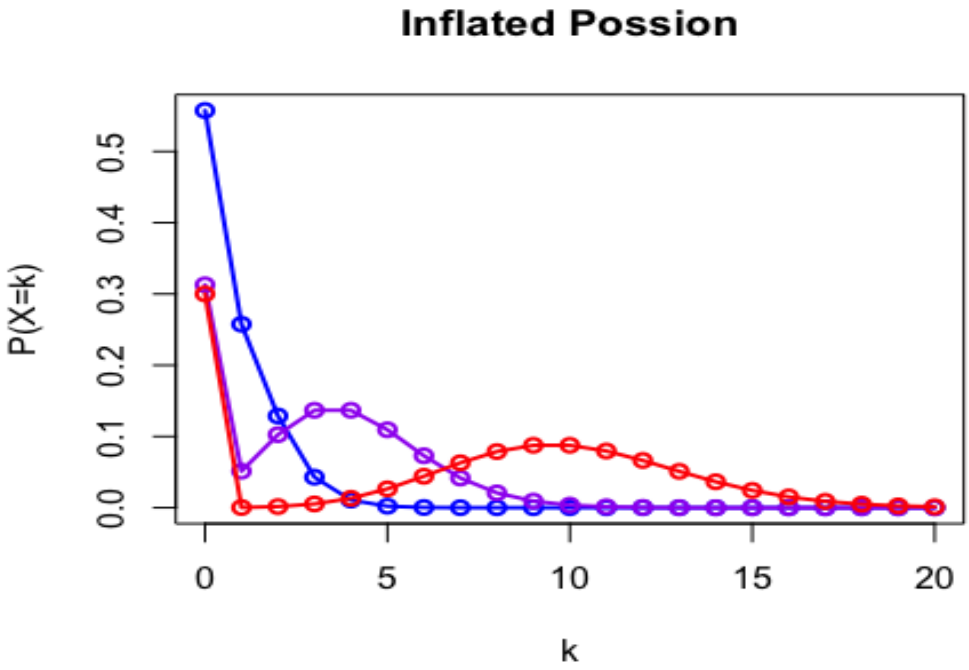


Figure 3.14: Inflated Poisson distribution for $p=0.3$, $\lambda=1$ (red), $\lambda=4$ (purple) and $\lambda=10$ (red).

The above model is called the Inflated zero Poisson model and it was first proposed by Diane Lambert for modelling manufacturing defects. Observe that the inclusion of the Bernoulli random variables allows one to model a large number of zeros without the need to use the Poisson distribution. This distribution does not belong to the exponential family.

Given the iid observations $\{X_i\}$ the likelihood function is

$$L(\theta,\underline{X}_n)=\prod_{i=1}^{n}\left([p+\exp(-\lambda)(1-p)]^{I(X_i=0)}\left[\frac{\lambda^{X_i}\exp(-\lambda)(1-p)}{X_i!}\right]^{1-I(X_i=0)}\right)$$

and the log-likelihood is

$$\mathcal{L}_n(\theta,\underline{X}_n)=\sum_{i=1}^{n}I(X_i=0)\log\left[p+\exp(-\lambda)(1-p)\right]+\sum_{i=1}^{n}[1-I(X_i=0)]\left[X_i\log\lambda-\lambda-\log X_i!+\log(1-p)\right].$$

The derivatives are

$$\begin{aligned}\frac{\partial \mathcal{L}_n(\theta, \underline{X}_n)}{\partial p} &= \sum_{i=1}^{n} I(X_i = 0)\frac{1-\exp(-\lambda)}{p+\exp(-\lambda)(1-p)} - \sum_{i=1}^{n}[1 - I(X_i = 0)]\frac{1}{1-p} = 0 \\ \frac{\partial \mathcal{L}_n(\theta, \underline{X}_n)}{\partial \lambda} &= \sum_{i=1}^{n} I(X_i = 0)\frac{-\exp(-\lambda)(1-p)}{p+\exp(-\lambda)(1-p)} + \sum_{i=1}^{n}[1 - I(X_i = 0)]\left[\frac{X_i}{\lambda} - 1\right] = 0.\end{aligned}$$

An analytic solution does not exist for the above and a numerical routine needs to be used (note that a moment type of estimator is possible).

Using constraints: Lagrange multipliers

This section goes beyond the syllabus of this course, but it is important when constructing likelihoods. In some examples, one needs to place constraints on the parameters. The simplest case is the multinomial distribution. The multinomial distribution is a generalisation of the binomial distribution. To construct a binomial distribution, we use that for each trial, X_s, there are two options "success" and "failure", where the probability of a success is π. However, for many data sets, each trial may have more than two options, for examples in a survey, one may give a person several fruit K different options and ask which is their favourite fruit. To simplify notation, we label each fruit r for $1 \leq r \leq K$. The person can only answer one fruit, where the probability of fruit i being selected is π_r. In this set-up, suppose N people are random selected (with replacement) asked the fruit question. Let $\underline{M} = (M_1, \ldots, M_K)$ denote total number of responses for each fruit. For example M_r denotes the number of people out of N who like fruit r. The sample space is $\{(m_1, \ldots, m_K); 0 \leq m_i \leq N, \sum_{r=1}^{K} m_r = K\}$. It can be shown that

$$P(M_1 = m_1, M_2 = m_2, \ldots, M_K = m_k) = \binom{N}{m_1, \ldots, m_K}\pi_1^{m_1} \ldots \pi_K^{m_K}.$$

This is called a multinomial distribution. Suppose that n surveys are conducted and in each survey N_i people are asked which was their favourite fruit, we observe the random vector $\{\underline{M} = (M_{i,1}, \ldots, M_{i,K})\}_{i=1}^{n}$. The log-likelihood is

$$\mathcal{L}_n(\pi_1, \ldots, \pi_K) = \sum_{i=1}^{n}\sum_{r=1}^{K} M_{i,r}\log\pi_r + \sum_{i=1}^{n}\log\binom{N_i}{M_{i,1}, \ldots, M_{i,K}}.$$

However, we observe that we have an additional condition on the parameters, that is the probabilities $\sum_{r=1}^{K}\pi_r = 1$ (thus in the end we only estimate $(K-1)$ parameters not K, since the $\pi_K = 1 - \pi_1 - \ldots - \pi_{K-1}$). We can either place this condition into the likelihood itself:

$$\mathcal{L}_n(\pi_1, \ldots, \pi_{K-1}) = \sum_{i=1}^{n}\left[\sum_{r=1}^{K-1} M_{i,r}\log\pi_r + M_{i,K}\log(1 - \pi_1 - \ldots - \pi_{K-1})\right] \sum_{i=1}^{n}\log\binom{N_i}{M_{i,1}, \ldots, M_{i,K}}$$

or we include an Lagrange muliplier. This is done by including an extra "dummy" variable into the likelihood

$$\mathcal{L}_n(\pi_1, \ldots, \pi_K, \lambda) = \sum_{i=1}^{n}\sum_{r=1}^{K} M_{i,r}\log\pi_r + \sum_{i=1}^{n}\log\binom{N_i}{M_{i,1}, \ldots, M_{i,K}} - \lambda\left(\sum_{r=1}^{K}\pi_r - 1\right),$$

the last term forces the $\sum_{r=1}^{K} \pi_r = 1$. This can be seen when we differentiate $\mathcal{L}_n(\pi_1, \ldots, \pi_K, \lambda)$ with respect to $\{\pi_r\}_{r=1}^{K}$ and λ:

$$\begin{aligned} \frac{\partial}{\partial \pi_s} \mathcal{L}_n(\pi_1, \ldots, \pi_K, \lambda) &= \sum_{i=1}^{n} \frac{M_{i,s}}{\pi_s} - \lambda \qquad 1 \leq s \leq K \\ \frac{\partial}{\partial \lambda} \mathcal{L}_n(\pi_1, \ldots, \pi_K, \lambda) &= \sum_{r=1}^{K} \pi_r - 1. \end{aligned}$$

Thus when we solve the $(K + 1)$ equations above (by setting them to zero), the very last one constrains $\sum_{r=1}^{K} \pi_r = 1$. In this example, the there is not much advantage of adding the additional term $\lambda \left(\sum_{r=1}^{K} \pi_r - 1 \right)$. But often placing restrictions to the parameters in a likelihood using Lagrange multipliers can be extremely useful.

3.4.3 Evaluation of the MLE for more complicated distributions

Most methods are designed for minimisation of a criterion. Therefore to use these methods we define the negative of the likelihood (which simply turns the likelihood upside down). Therefore, the parameter which maximises the likelihood is the same as the parameter which minimises the negative likelihood. Brute force or minimisation of a function (such as a likelihood) in R can be achieved using the function `optim` or `nlm`. The type of algorithm that one uses is very important. If the likelihood function is concave (equivalent to the negative of the likelihood being convex), the maximising can be achieved using the Newton-Raphson algorithm or related convex optimisation schemes.

We illustrate the above routine in R for the MLE of the gamma distribution.

```
# Simulate from a gamma function alpha = 9, beta = 2
# often alpha is called the shape and beta the rate.

xdata = rgamma(n=100,shape =9,rate =2)

# Below is the negative-likelihood based on
# gamma distribution using the generated data.

LikeGamma = function(par){
  n = length(xdata) # xdata is the data vector we input into likelihood
  alpha = par[1]; beta = par[2]
 loglik = (alpha-1)*sum(log(xdata))- beta*sum(xdata) + n*alpha*log(beta) - n*log(gamma(alpha))
  nloglik= -loglik
  return(nloglik)
```

```
}

## "minimization" using optim function
## par: initial value; fn = function to minimize
# As initial value we give the vector alpha = 3 and
# beta = 3.
# A better initial value is to give the Moments estimator
# this would speed up the algorithm and if the
# likelihood was not concave (which for this example it is)
# it is more likely to converge to its global minimum.

fit.optim = optim(par= c(3,3), fn=LikeGamma, hessian = T)

fit.optim$par
fit.optim$convergence
solve(fit.optim$hessian) #inverse hessian
# solve(fit.optim$hessian) gives the asymptotic variance
# of the estimator. See Section 3.5.2 below.
```

Many of the algorithms require an initial value for the estimator $\widehat{\theta}$, to start the algorithm off. If you can provide an initial values which is based on a crude estimator of θ (such as a method of moments estimator), this would be great.

Optimisation for a concave distribution is relatively straightfoward. Given any initial value you are usually gauranteed to get to the global maximum of the distribution. Under quite general conditions most of the distributions in the exponential are concave, so are ideal for maximising. However, if the distribution is not concave (which often happens in the case of mixtures of distributions as studied in HW2 or the inflated Poisson), then the routine can run into problems. For example, it may converge to a local maximum rather than a global maximum. Typically, runs the routine with several initial values in the hope of finding a global maximum, but this is far from a simple task.

Remark (The EM-algorithm)**.** *The Expectation-Maximisation algorithm is an algorithm designed for specifically maximising the likelihood. It works by constructing a likelihood based on the observed data and unobserved data. By using a "clever" choice of unobserved data the combined likelihood (often complete likelihood) has a simple form. The EM-algorithm is based on maximisation of the conditional expectation of this complete likelihood. The precise details are beyond this course.*

Remark (Profiling)**.** *So called profiling the likelihood is another strategy for maximising the likelihood.*

3.5 Sampling properties of the MLE

3.5.1 Consistency

It can be shown that the MLE (under certain conditions) is asymptotically consistent. This means that if the distribution is correctly specified then $\widehat{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$ (where θ_0 is the true parameter in the distribution) as the sample size grows. The proof of this result is quite technical and beyond this class, but we give a heuristic (some ideas) as to why the Maximum likelihood estimator actually works. Rice discusses this in Section 8.5.2.

Let us suppose that X_i are iid random variables with density $f(x;\theta_0)$ the true parameter θ_0 but it is known it belongs to the parameter space Θ. The log-likelihood

$$\mathcal{L}_n(\theta;\underline{X}) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta).$$

We observe that we divide by n in the above definition to turn the sum into an average (which we like), in terms of estimation it does not change anything. Since $\mathcal{L}_n(\theta;\underline{X})$ is an average of iid random variables (in this case $\{\log f(X_i,\theta)\}$) as n grows large it limits to its expectation (see Section 1.4.1)

$$\mathcal{L}(\theta) = \mathrm{E}\left[\log f(X;\theta)\right].$$

We recall that the expectation is defined as

$$\mathrm{E}\left[\log f(X;\theta)\right] = \int [\log f(x,\theta)] f(x,\theta_0) dx,$$

observe this is a function of θ.

Idea of heuristic proof We need to show that in the ideal situation that we have an infinite sample size and thus the observed likelihood is $\mathrm{E}\left[\log f(X;\theta)\right]$. That the θ that maximises this likelihod is the true population parameter θ_0 (that generates the density $f(x;\theta_0)$). If this did not hold. Then the MLE estimator cannot work. This is the bare minumum requirement for the method to work. Such result form the basics in any statistical method. If it does not work for an infinite data set it cannot work for any data set!

Returning to the proof To show that $\mathcal{L}(\theta) = \mathrm{E}\left[\log f(X;\theta)\right]$ is maximised at θ_0 we take the derivative with respect to θ and show that the derivative is zero at θ_0. Thus

$$\frac{d\mathcal{L}(\theta)}{d\theta} = \frac{d}{d\theta}\int [\log f(x,\theta)] f(x,\theta_0) dx.$$

We make the assumption that the derivative can be put inside the integral (which for many distributions does hold, but not all):

$$\frac{d\mathcal{L}(\theta)}{d\theta} = \int \frac{d}{d\theta}[\log f(x,\theta)] f(x,\theta_0) dx = \int \frac{df(x,\theta)}{d\theta}\frac{1}{f(x,\theta)} f(x,\theta_0) dx.$$

Thus at θ_0 we have

$$\begin{aligned}\frac{d\mathcal{L}(\theta)}{d\theta}\rfloor_{\theta=\theta_0} &= \int \frac{df(x,\theta)}{d\theta}\rfloor_{\theta=\theta_0}\frac{1}{f(x,\theta_0)}f(x,\theta_0)dx \\ &= \int \frac{df(x,\theta)}{d\theta}\rfloor_{\theta=\theta_0}dx = \frac{d}{d\theta}\underbrace{\int f(x,\theta)dx}_{=1}\rfloor_{\theta=\theta} = \frac{d1}{d\theta} = 0.\end{aligned}$$

The take home message in the above proof is that the density of pmf integrates to one, it does not depend on a parameter.

Using the above result as a starting point, we obtain the following result.

Lemma 3.1

Suppose that $\{X_i\}_{i=1}^n$ are iid random variables with density $f(x;\theta_0)$. Let $\widehat{\theta}_n$ be the MLE of θ based on $\{X_i\}_{i=1}^n$. Then

$$\widehat{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$$

we $n \to \infty$.

3.5.2 The distributional properties of the MLE

The aim in this section is to study the sampling properties of the MLE. These results will be used in later chapter for testing and constructing confidence intervals. As we mentioned above many MLE estimators are the same as the method of moments estimator. And we we showed in Section 3.2.3 that the methods of moments estimator is usually asymptotically normal in distribution. In this section we show that (under certain conditions) all MLE estimators are asymptotically normal.

We first illustrate this result with an example. We consider the MLE of the parameters in the Gamma distribution (which is not, in an obvious way, a method of moments estimator). We simulate from a Gamma distribution with $\alpha = 9$ and $\beta = 2$. For simplicity, we treat $\beta = 2$ as known and only estimate α[1]. The simulations were done over 400 realisations for $n = 1, \ldots, 100$. We estimate α using the routine outlined above and denote the MLE for sample size n as $\widehat{\alpha}_n$.

[1]Keep in mind the distribution will change a little if we estimate both. Further the variances will be different if we estimate only one parameters rather than both.

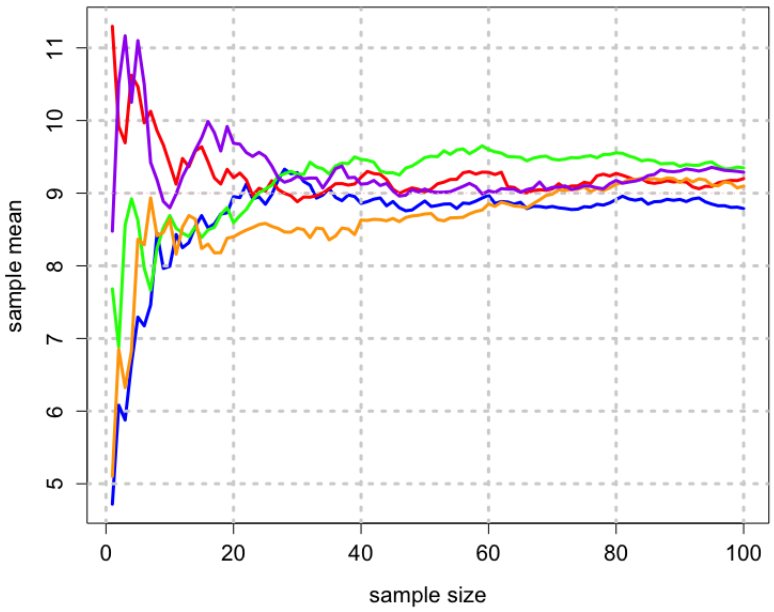


Figure 3.15: Left: trajectories of Gamma estimator $\alpha = 9$

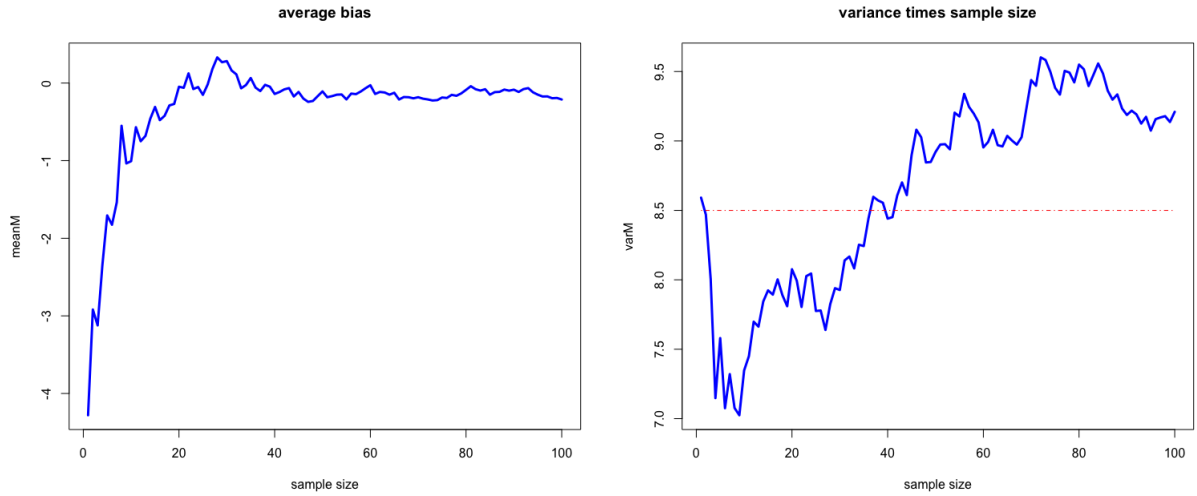


Figure 3.16: Left: Estimated bias. Right: variance × sample size when $\alpha = 9$. The red dotted line is the asymptotic time samples size $(I(\alpha)^{-1})$.

In Figure 3.15 we observe that each realisation does appear to "converge" to the true parameter as the sample size grows. In Figure 3.16 we observe that the estimator has a "finite" sample bias, but the bias decreases as the sample size grows. The n times the variance fluctuates quite a lot but it around 8.5 – 9 (8.5 is the asymptotic variance times n) for a n larger than 30. Finally, in Figure 3.17 we make a plot of the

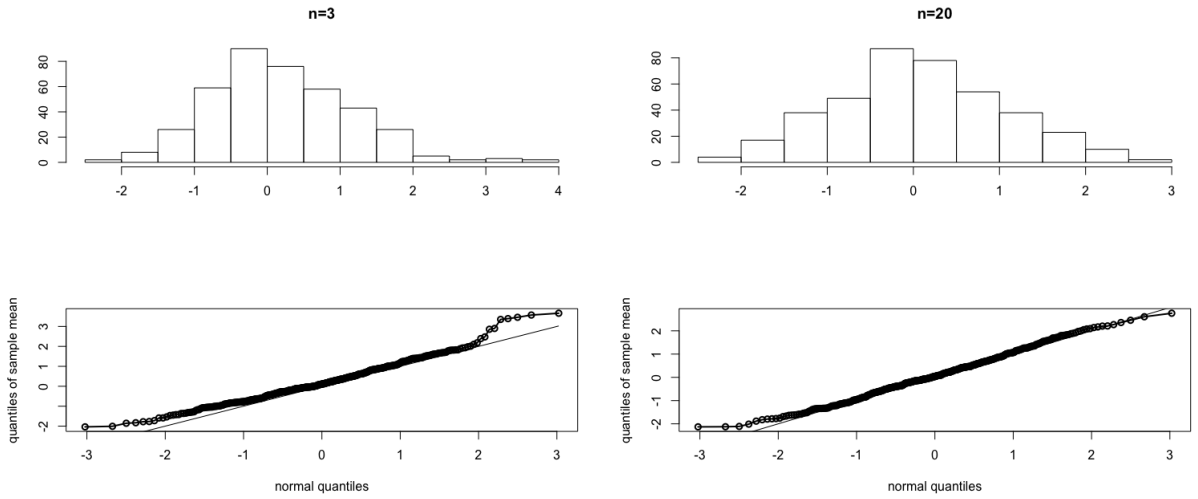


Figure 3.17: Top: Distribution of the standardized sample mean $Z_n = \sqrt{nI(\alpha)}(\widehat{\alpha}_n - 9)$ ($n = 3$ and $n = 20$). Bottom: QQplot against standard normal quantiles.

histogram and QQplot against the standard normal of

$$Z_n = \sqrt{nI(\alpha)}(\widehat{\alpha}_n - 9),$$

we discuss what $I(\alpha)$ is below. But we observe that for $n = 3$, the estimator appears to have a small right skew, which is reduced $n = 20$.

These results allude to the result that as the sample size grows, like the method of moment estimators, the distribution of the maximum likelihood estimator asymptotically becomes normal.

Theorem 3.2

Suppose $\{X_i\}$ are iid random variables with parametric distribution $f(x, \theta_0)$, where θ_0 is unknown. The density $f(x, \theta)$ is smooth over θ (such that we can evaluate its derivative) and the support of the density does not depend on the parameter θ. Then asymptotically the MLE is

$$\sqrt{nI(\theta_0)}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

as $n \to \infty$, where

$$I(\theta_0) = -\int_{-\infty}^{\infty} \left(\frac{d^2 \log f(x;\theta)}{d\theta^2}\right) \rfloor_{\theta=\theta_0} f(x;\theta_0) dx$$

From the above result for a large enough n we say that

$$\widehat{\theta}_n \sim \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right).$$

Important The Fisher information matrix is always evaluated at the true parameter. Thus

$$I(\theta) = -\int_{-\infty}^{\infty} \left(\frac{d^2 \log f(x;\theta)}{d\theta^2}\right) f(x;\theta) dx$$

the θ in the $\frac{d^2 \log f(x;\theta)}{d\theta^2}$ must match the θ in the density $f(x, \theta)$.

The above thoerem is stated for univariate MLE estimators. However, a similar results hold for the MLE of several parameters (for example the MLE of both α and β in the geometric distribution). In this case we replace $I(\theta)$ with a matrix, which is the Hessian of the log-likelihood.

Application We apply this result to construct an approximate 95% confidence interval for θ_0. Suppose we evaluate the MLE $\widehat{\theta}_n$, then the approximate 95% CI for θ_0 is

$$\left[\widehat{\theta}_n - 1.96 \times \frac{1}{\sqrt{nI(\theta_0)}}, \widehat{\theta}_n + 1.96 \times \frac{1}{\sqrt{nI(\theta_0)}}\right].$$

Of course θ_0 is unknown to calculate $I(\theta_0)$ instead we replace with its estimator $I(\widehat{\theta}_n)$ to give the approximate 95% CI for θ

$$\left[\widehat{\theta}_n - 1.96 \times \frac{1}{\sqrt{nI(\widehat{\theta}_n)}}, \widehat{\theta}_n + 1.96 \times \frac{1}{\sqrt{nI(\widehat{\theta}_n)}}\right].$$

Outline of proof of Theorem 3.2

The precise proof of Theorem 3.2 is quite technical. But we give the basic ideas here. In many respects the proof resembles the proof of Lemma 1.2. Even though $\widehat{\theta}_n$ does not appear to be an average, at the estimation method is an average; the log-likelihood which is the sum (or average) of $\log f(X_i, \theta)$. As the log-likelihood is an average it will asymptotically be normal, which implies (in a way described below) that $\widehat{\theta}_n$ can also be written as an average and will be asymptotically normal.

We first recall that since

$$\widehat{\theta}_n = \arg\max \mathcal{L}_n(\theta),$$

then in general (if $\widehat{\theta}_n$ lies inside the parameter space and not on the boundary), $\widehat{\theta}_n$ is the solution of

$$\frac{d\mathcal{L}_n(\theta)}{d\theta}\rfloor_{\theta=\widehat{\theta}_n} = 0.$$

We recall that $\frac{d\mathcal{L}_n(\theta)}{d\theta}$ is

$$\frac{d\mathcal{L}_n(\theta)}{d\theta} = n^{-1}\sum_{i=1}^{n}\frac{d\log f(X_i;\theta)}{d\theta},$$

we divide by n to turn the derivative of the likelihood into an average, which does not change the estimator. Now by using the mean value theorem and expanding $\frac{d\mathcal{L}_n(\theta)}{d\theta}\rfloor_{\theta=\widehat{\theta}_n}$ about θ_0 (the true parameter) gives

$$\frac{d\mathcal{L}_n(\theta)}{d\theta}\rfloor_{\theta=\widehat{\theta}_n} \approx \frac{d\mathcal{L}_n(\theta)}{d\theta}\rfloor_{\theta=\theta_0} + (\widehat{\theta}_n - \theta_0)\frac{d^2\mathcal{L}_n(\theta)}{d\theta^2}\rfloor_{\theta=\widehat{\theta}_n},$$

see the illustration in Figure 3.18 to understand why. Since $\frac{d\mathcal{L}_n(\theta)}{d\theta}\rfloor_{\theta=\widehat{\theta}_n}$ this gives

$$\frac{d\mathcal{L}_n(\theta)}{d\theta}\rfloor_{\theta=\theta_0} \approx -(\widehat{\theta}_n - \theta_0)\frac{d^2\mathcal{L}_n(\theta)}{d\theta^2}\rfloor_{\theta=\widehat{\theta}_n},$$

observe that the above involves the Fisher information before taking expectation (often called the observed Fisher information). Going back to the definitions of the likelihood we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{df(X_i,\theta_0)}{d\theta} \approx (\widehat{\theta}_n - \theta_0)\frac{-1}{n}\sum_{i=1}^{n}\frac{df(X_i,\theta_0)}{d\theta^2}.$$

Since $\frac{-1}{n}\sum_{i=1}^{n}\frac{df(X_i,\theta_0)}{d\theta^2}$ is an average we will replace it by its expectation, which is the Fisher information $I(\theta_0)$. This gives

$$\frac{1}{n}\sum_{i=1}^{n}\frac{df(X_i,\theta_0)}{d\theta} \approx (\widehat{\theta}_n - \theta_0)nI(\theta_0).$$

Thus

$$(\widehat{\theta}_n - \theta_0) \approx (nI(\theta_0))^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{df(X_i,\theta_0)}{d\theta}.$$

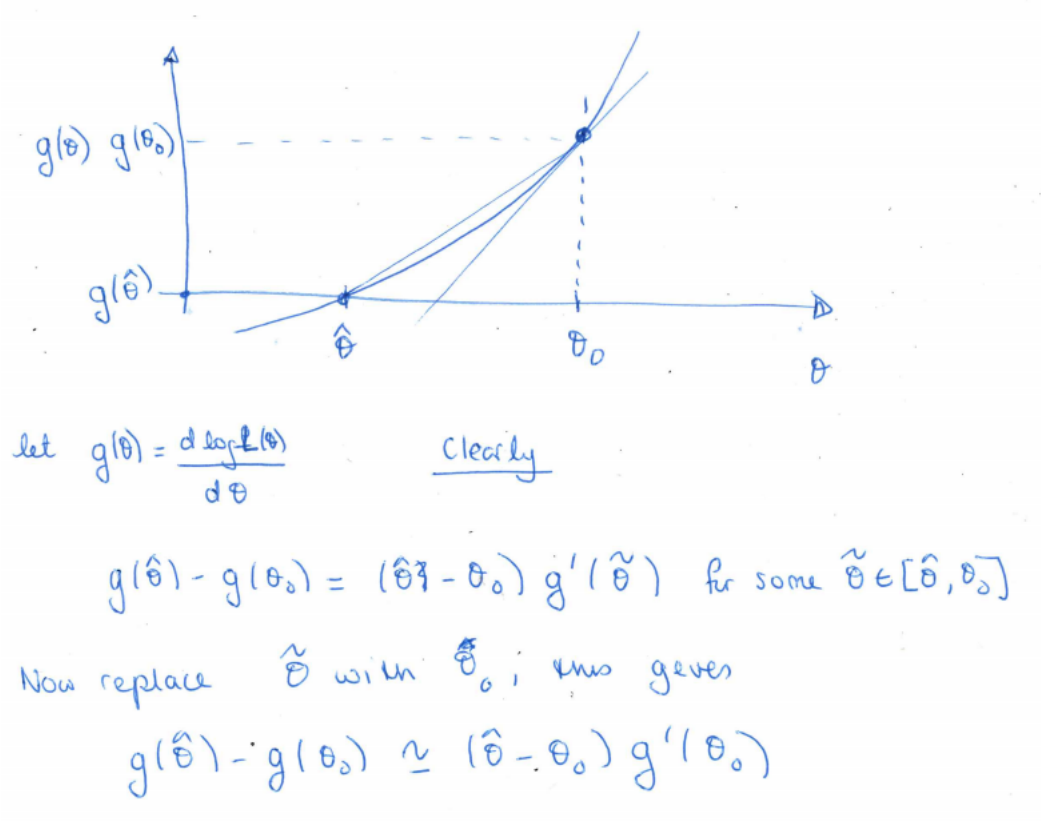


Figure 3.18: Illustration of mean value theorem.

Observe that $I(\theta_0)^{-1}$ is a constant, so the limiting distribution of $(\widehat{\theta}_n - \theta_0)$ is determined by

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{df(X_i, \theta_0)}{d\theta},$$

which is an average of iid random variables. Further, we have shown in Section 3.5.1, that $\mathrm{E}[\log f(X;\theta)]$ is maximised at $\theta = \theta_0$, which is equivalent (under certain conditions) to $d\mathrm{E}[\log f(X;\theta)]/d\theta\rfloor_{\theta=\theta_0} = 0$. Thus $\mathrm{E}[\bar{Y}_n] = 0$. Now by the CLT in Theorem 1.1 we have that

$$\sqrt{nJ(\theta_0)}\bar{Y}_n \xrightarrow{\mathcal{D}} N(0,1) \text{ and } \sqrt{n}\bar{Y}_n \xrightarrow{\mathcal{D}} N(0, J(\theta_0))$$

where

$$J(\theta_0) = \mathrm{var}\left(\frac{df(X_i,\theta_0)}{d\theta}\rfloor_{\theta=\theta_0}\right).$$

Thus altogether we have

$$(\widehat{\theta}_n - \theta_0) \approx (nI(\theta_0))^{-1}\bar{Y}_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I(\theta_0)^{-1}J(\theta_0)I(\theta_0)^{-1}\right) \qquad n \to \infty.$$

Finally, in Lemma 3.3 we prove that $J(\theta_0) = I(\theta_0)$. Thus the above reduces to

$$(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, (nI(\theta_0))^{-1}\right) \qquad n \to \infty,$$

which is what we want to show.

Example: Gamma distribution

Consider the Gamma example considered above where $\alpha = 2$ in the distribution is assumed known. The log-likelihood is

$$\mathcal{L}_n(\theta; \underline{X}_n) = \sum_{i=1}^{n} \left[(\alpha - 1)\log X_i - \beta X_i + \alpha \log \beta - \log \Gamma(\alpha)\right].$$

Differentiating the above wrt α gives

$$\frac{\partial \mathcal{L}_n(\theta; \underline{X}_n)}{\partial \alpha} = \sum_{i=1}^{n} \left[\log X_i + \log \beta - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}\right].$$

So

$$\frac{\partial \log f(X_i, \theta)}{\partial \alpha} = \left[\log X_i + \log \beta - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}\right]$$

and

$$-\frac{\partial^2 \log f(X_i, \theta)}{\partial \alpha^2} = \left[\frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2}\right].$$

Notice that the second derivative does not depend on the data. This is quite common for distributions in the exponential family (under certain parameterisations of the parameter), but is not the rule for all distributions. Often the second derivative will depend on the observed data. Based on the above, the Fisher information is

$$\begin{aligned} I(\alpha) &= -\mathrm{E}\left[\frac{\partial^2 \log f(X_i, \theta)}{\partial \alpha^2}\right] \\ &= \left[\frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2}\right]. \end{aligned}$$

Therefore the limiting distribution of α is

$$\sqrt{n\left(\frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2}\right)}(\widehat{\alpha}_n - \alpha) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1). \tag{3.8}$$

This is exactly what we observed in Figures 3.16 and 3.17; the asymptotic variance is $1/(nI(\alpha))$ and the standardized estimator is close to a standard normal for a large enough sample size.

Given the MLE $\widehat{\alpha}_n$ the approximate 95% confidence interval for α is

$$\left[\widehat{\alpha}_n - 1.96 \times \left(n\left(\frac{\Gamma''(\widehat{\alpha}_n)}{\Gamma(\widehat{\alpha}_n)} - \frac{\Gamma'(\widehat{\alpha}_n)^2}{\Gamma(\widehat{\alpha}_n)^2}\right)\right)^{-1/2}, \widehat{\alpha}_n + 1.96 \times \left(n\left(\frac{\Gamma''(\widehat{\alpha}_n)}{\Gamma(\widehat{\alpha}_n)} - \frac{\Gamma'(\widehat{\alpha}_n)^2}{\Gamma(\widehat{\alpha}_n)^2}\right)\right)^{-1/2}\right].$$

Example: Exponential distribution

The exponential data the log-likelihood is

$$\mathcal{L}_n(\theta; \underline{X}_n) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

Differentiating the above wrt λ gives

$$\frac{\partial \mathcal{L}_n(\theta; \underline{X}_n)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i.$$

So

$$\frac{\partial \log f(X_{i,\lambda})}{\partial \lambda} = \frac{1}{\lambda} - X_i.$$

and the second derivative is

$$\frac{d^2 \log f(X_{i,\lambda})}{d\lambda^2} = -\frac{1}{\lambda^2}.$$

Thus the Fisher information is

$$I(\lambda) = -\mathrm{E}\left(\frac{d^2 \log f(X_{i,\lambda})}{d\lambda^2}\right) = \frac{1}{\lambda^2}.$$

Therefore the limiting distribution of α is

$$\sqrt{n\frac{1}{\lambda^2}}\left(\widehat{\lambda}_n - \lambda\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

Based on the above the approximate 95% CI for λ is

$$\left[\widehat{\lambda}_n - 1.96 \times \frac{\widehat{\lambda}_n}{\sqrt{n}}, \widehat{\lambda}_n + 1.96 \times \frac{\widehat{\lambda}_n}{\sqrt{n}}\right].$$

3.6 The Fisher information matrix

We recall that for iid random variables $\{X_i\}$ with mean μ and the variance σ^2, the sample mean (for large n) is approximately

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

And for Theorem 3.2, the MLE (for large n) is approximately

$$\widehat{\theta}_n \sim \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right).$$

Comparing the two results we observe that σ^2 and $I_n(\theta_0)^{-1}$ play similar roles. It is clear σ^2 is small when the variance of X_i is small. What we want to do is understand when $I_n(\theta_0)^{-1}$ is small or equivalently, the Fisher information $I_n(\theta_0)$ is large. However, why should the (asymptotic) variance of the MLE involve the second derivative of the likelihood and what it actually means can be quite difficult to understand. We make an attempt in this section.

The Fisher information matrix $I(\theta)$ is really quite a cryptic object, even its name appears nonsensical! The word Fisher comes from Ronald Fisher, so no need to explain that part. But the word information appears strange; roughly speaking $nI(\theta)$ describes how much information the data contains about the true parameter. Below, we try to explain why.

Background: First we recall from our calculus days, that the size of the second derivative at the maximum of a function corresponds to how peaky or curvey it is at the maximum. The "larger" the second derivative the more the curvature (peaky) the likelihood is a the maximum. The "smaller" the second derivative the flatter the function is about the maximum. Recall, that when the second derivative is zero, we have a saddle point. These insights are useful in the discussion below.

3.6.1 Example: Exponential distribution

For simplicity we focus on the exponential distribution (though the discussion below applies to most distributions). We recall from equation (3.7) that the exponential density is $f(x;\theta)$ and the log-density is

$$\log f(X,\lambda) = \log\lambda - \lambda X.$$

The second derivative is

$$\frac{d^2}{d\lambda^2}\log f(X,\lambda) = -\frac{1}{\lambda^2}.$$

Thus the information matrix is

$$I(\lambda) = -\mathrm{E}\left[\frac{d^2}{d\lambda^2}\log f(X,\lambda)\right] = \frac{1}{\lambda^2}.$$

Thus for a the data set $\{X_i\}_{i=1}^n$ the information matrix is

$$nI(\lambda) = \frac{n}{\lambda^2}.$$

Observe that for the exponential distribution $\frac{d^2}{d\lambda^2}\log f(X,\lambda)$ does not depend on the data. So the curvature of the likelihood is the same for all data sets. However, for other distributions $\frac{d^2}{d\lambda^2}\log f(X,\lambda)$ will depend on the data, which is why we consider the mean/expected curvature of $\log f(X,\lambda)$. The larger $nI(\lambda)$, the greater the information in the data about λ. This is easily understood, by studying the likelihood. Since $nI(\lambda)$ corresponds to the (expected) negative second derivative of the likelihood, it measures the steepness

of the likelihood about the true, population parameter. The large $nI(\lambda)$ the more pronounced the likelihood about the peak, and the "easier" it is to find the maximum. To see this, we simulate from two different exponential distributions, with $\lambda = 1$ and $\lambda = 100$ and sample size $n = 10$. For $\lambda = 1$, we generate the numbers:

$$2.0546 \quad 0.9861 \quad 0.3977 \quad 1.9480 \quad 0.8082 \quad 0.8082 \quad 0.0491 \quad 2.5444 \quad 0.4528 \quad 0.9950.$$

The sample mean is 1.101 and the log-likelihood is given in Figure 3.19. For $\lambda = 100$, we generate the numbers:

$$0.0008424 \quad 0.0321545 \quad 0.0009847 \quad 0.0070063 \quad 0.0044617$$
$$0.0470954 \quad 0.0076897 \quad 0.0055562 \quad 0.0000079 \quad 0.0050735.$$

The sample mean is 0.01108 and the likelihood and log-likelihood is given in Figure 3.20. What we observe

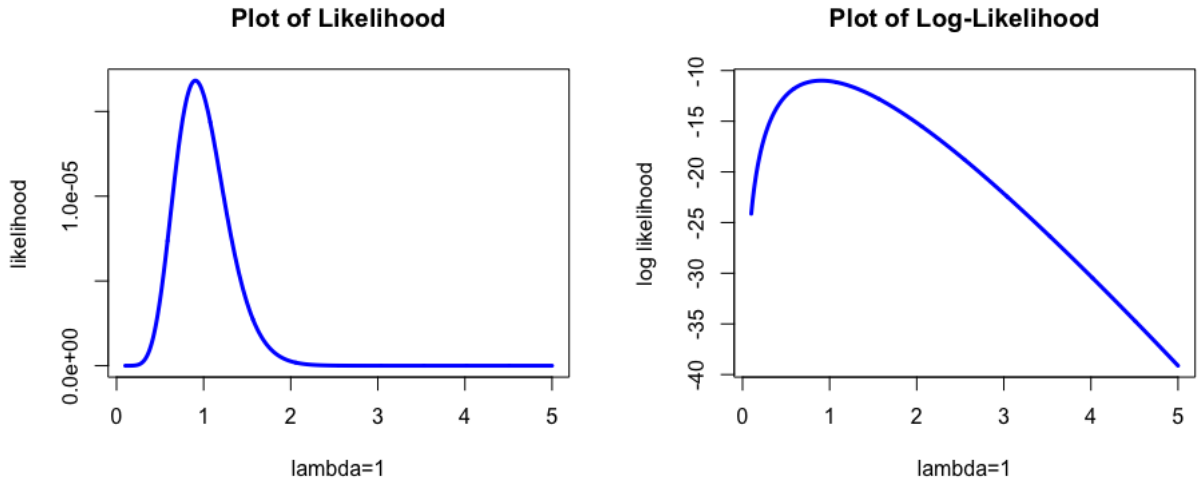


Figure 3.19: The likelihood (left) and corresponding log likelihood (right) for one simulation for an exponential distribution with $\lambda = 1$ and $n = 10$. $nI(\lambda) = 10/\lambda^2 = 10$.

is when λ is small, the likelihood is more peaked about the maximum. We are better able to find the maximum; this corresponds to a smaller variance and "more" information in the data about the underlying parameter (thus $nI_n(\lambda)$ is large or equivalently the asymptotic variance $(nI_n(\lambda))^{-1}$ is small). On the other hand, when λ is large, the likelihood is more "flat" about the maximum. Making it much harder to "find" the maximum. It is difficult to distinguish the maximum from other values in the neighbourhood. This means there is "less" information in the data about the underlying parameter (thus $nI_n(\lambda)$ is small or equivalently the asymptotic variance $(nI_n(\lambda))^{-1}$ is large). If you run the same simulation, but with a larger n, you will see that the likelihood gets increasingly steeper about its maximum.

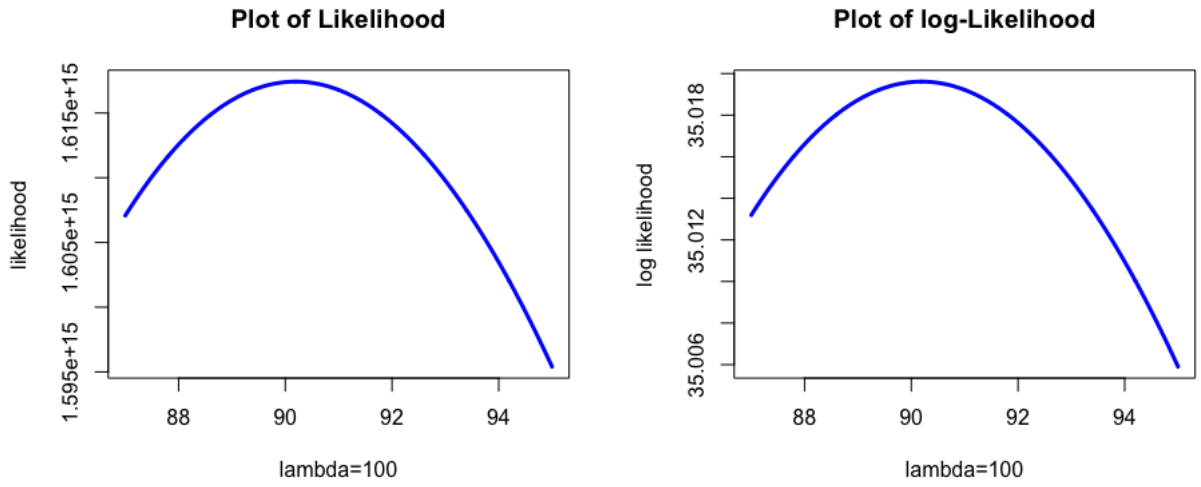


Figure 3.20: The likelihood and corresponding log-likelihood for one simulation from an exponential distribution with $\lambda = 100$ and $n = 10$. $nI(\lambda) = 10/\lambda^2 = 10/100^2$. Observe that the true $\lambda = 100$ is far from the maximum in the data set and how shallow this curve is.

In summary The data holds more information about λ, if the λ in the exponential density is small. We observe that small λ, corresponds to a flatter density (see Figure 3.21).

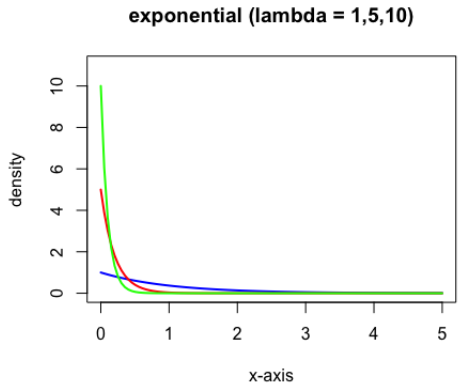


Figure 3.21: Density of exponential for different λ. When λ is small, Fisher information is large.

Example 3.3 (An alternative parameterisation of the exponential distribution)**.** *An alternative but commonly used parameterisation for the exponential distribution is*

$$f(x;\mu) = \frac{1}{\mu}\exp(-x/\mu) \qquad x \geq 0$$

for any $\mu > 0$. Comparing μ with θ above we observe that $\mu = \theta^{-1}$. Using this parametrisation, a random

variable with density $f(x;\mu)$ has expectation $\mathrm{E}[X] = \mu$. The log density is

$$\log f(x;\mu) = -\log \mu - \frac{x}{\mu}$$

and its derivatives are

$$\begin{aligned} \frac{d \log f(x;\mu)}{d\mu} &= -\frac{1}{\mu} + \frac{x}{\mu^2} \\ \frac{d^2 \log f(x;\mu)}{d\mu^2} &= \frac{1}{\mu^2} - \frac{2x}{\mu^3} \end{aligned}$$

Based on the above the Fisher information matrix of μ is

$$\begin{aligned} I(\mu) &= -\mathrm{E}\left(\frac{d^2 \log f(X;\mu)}{d\mu^2}\right) = -\frac{1}{\mu^2} + \frac{2\mathrm{E}[X]}{\mu^3} \\ &= -\frac{1}{\mu^2} + \frac{2\mu}{\mu^3} = \frac{1}{\mu^2}. \end{aligned}$$

Thus we observe that the smaller μ is the larger the Fisher information.

In summary The data holds more information about μ, if the μ in the exponential density is small. We make a plot for this parametisation of the exponential density in Figure 3.22). We observe that small μ, corresponds to a density concentrated about zero (see Figure 3.22).

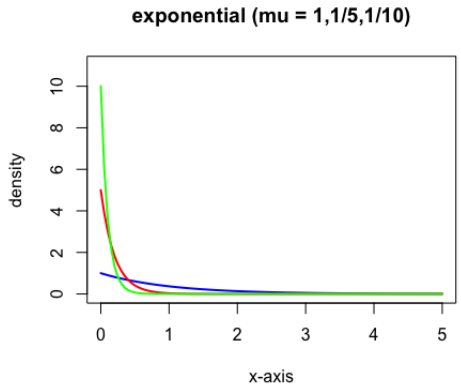


Figure 3.22: Density of exponential for different μ. When μ is small, Fisher information is large.

3.6.2 Example: Poisson distribution

In now obtain the Fisher information matrix for the Poisson distribution. We recall the probability mass function for the Poisson distribution is

$$p(x;\lambda) = \frac{\lambda^x \exp(-\lambda x)}{x!} \qquad x \geq 0.$$

The log pmf is

$$\log p(x;\lambda) = x\log\lambda - \lambda x - \log x!$$

and its derivatives are

$$\begin{aligned}\frac{d\log p(x;\lambda)}{d\lambda} &= \frac{x}{\lambda} - \lambda \\ \frac{d^2\log p(x;\lambda)}{d\lambda^2} &= -\frac{x}{\lambda^2}.\end{aligned}$$

Thus the Fisher information matrix is

$$I(\lambda) = -\mathrm{E}\left[\frac{d^2\log p(x;\lambda)}{d\lambda^2}\right] = \frac{\mathrm{E}[X]}{\lambda^2} = \frac{1}{\lambda}.$$

Thus we observe that Fisher information matrix is large for very small λ. This means the the data holds more information about λ, if the λ small. Note from Figure 3.23 that small λ corresponds to a pmf that is concentrated close to zero.

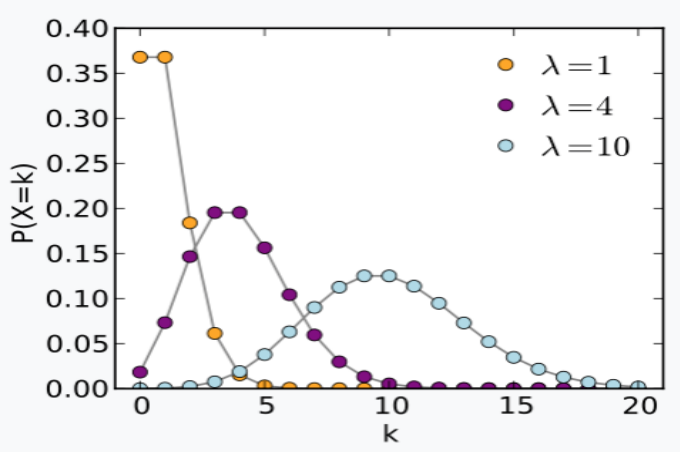


Figure 3.23: Left: Mass function of the Poisson distribution. Small λ corresponds to a mass concentrated close to zero.

We recall (from Section 3.4.2) that the MLE for λ is $\widehat{\lambda}_n = \bar{X}$ and by using Theorem 1.1 we have for large n

$$\widehat{\lambda}_n \sim N\left(\lambda, \frac{\mathrm{var}[X]}{n}\right) = N\left(\lambda, \frac{\lambda}{n}\right).$$

But this exactly the same answer as the limiting distribution result in Theorem 3.2;

$$\widehat{\lambda}_n \sim N\left(\lambda, \frac{1}{nI(\lambda)}\right) = N\left(\lambda, \frac{\lambda}{n}\right).$$

3.6.3 A useful identity

We now give a useful identity which links the variance of $\frac{d \log L_n(\theta;\underline{X})}{d\theta}\rfloor_{\theta=\theta_0}$ to $\mathrm{E}\left(\frac{d^2 \log L_n(\theta;\underline{X})}{d\theta^2}\rfloor_{\theta=\theta_0}\right)$. This is useful in obtaining "nice" expressions for the Fisher information.

Lemma 3.3

$$\mathrm{E}\left(\frac{d \log L_n(\theta;\underline{X})}{d\theta}\rfloor_{\theta=\theta_0}\right)^2 = -\mathrm{E}\left(\frac{d^2 \log L_n(\theta;\underline{X})}{d\theta^2}\rfloor_{\theta=\theta_0}\right).$$

PROOF. To prove this result we use the fact that the likelihood L_n is a density/distribution, thus it integrates to one:

$$\int L_n(\theta,\underline{x})d\underline{x} = 1.$$

Now by differentiating the above with respect to θ gives

$$\frac{\partial}{\partial\theta}\int L_n(\theta,\underline{x})d\underline{x} = 0.$$

Thus

$$\int \frac{\partial L_n(\theta,\underline{x})}{\partial\theta}d\underline{x} = 0 \Rightarrow \quad \int \frac{\partial \log L_n(\theta,\underline{x})}{\partial\theta}L_n(\theta,\underline{x})d\underline{x} = 0$$

Differentiating again with respect to θ and taking the derivative inside gives

$$\begin{aligned}
&\int \frac{\partial^2 \log L_n(\theta,\underline{x})}{\partial\theta^2}L_n(\theta,\underline{x})d\underline{x} + \int \frac{\partial \log L_n(\theta,\underline{x})}{\partial\theta}\frac{\partial L_n(\theta,\underline{x})}{\partial\theta}d\underline{x} = 0\\
\Rightarrow\quad &\int \frac{\partial^2 \log L_n(\theta,\underline{x})}{\partial\theta^2}L_n(\theta,\underline{x})d\underline{x} + \int \frac{\partial \log L_n(\theta,\underline{x})}{\partial\theta}\frac{1}{L_n(\theta,\underline{x})}\frac{\partial L_n(\theta,\underline{x})}{\partial\theta}L_n(\theta,\underline{x})d\underline{x} = 0\\
\Rightarrow\quad &\int \frac{\partial^2 \log L_n(\theta,\underline{x})}{\partial\theta^2}L_n(\theta,\underline{x})d\underline{x} + \int \left(\frac{\partial \log L_n(\theta,\underline{x})}{\partial\theta}\right)^2 L_n(\theta,\underline{x})d\underline{x} = 0
\end{aligned}$$

Thus

$$-\mathrm{E}\left(\frac{\partial^2 \log L_n(\theta,\underline{X})}{\partial\theta^2}\right) = \mathrm{E}\left(\frac{\partial \log L_n(\theta,\underline{X})}{\partial\theta}\right)^2.$$

Note in all the derivations we are evaluating the second derivative of the likelihood at the *true parameter* θ of the underlying distribution. □

3.7 The curious case of the uniform distribution

We recall that if $\{X_i\}$ are iid random variables with uniform density $f(x,\theta) = \theta^{-1}I_{[0,\theta]}(x)$. The uniform distribution is one of those distributions that do not satisfy the so called "regularity" conditions mentioned

above (this is because the support of the distribution involves the parameter, thus Lemma 3.3 does not hold). We do not maximise the likelihood (or log-likelihood) by differentiating it. We would expect that its sampling properties are also strange, we now show that they are.

We recall from Section 3.4.2, that the MLE is

$$\widehat{\theta}_n = \max_{1 \le i \le n} (X_i).$$

The sampling distribution of $\widehat{\theta}_n$ is actually quite straightforward to derive. It turns on that for the uniform distribution its MLE not asymptotically normal. We start with the distribution function:

$$\begin{aligned} P_\theta\left(\widehat{\theta}_n \le x\right) &= P\left(\max_{1 \le i \le n} (X_i) \le x\right) \\ &= P\left(X_1 \le x \text{ and } X_2 \le x \text{ and } \ldots \text{ and } X_n \le x\right) = \underbrace{\prod_{i=1}^{n} P(X_i \le x)}_{\text{by independence}} = \frac{x^n}{\theta^n}. \end{aligned}$$

This gives the distribution function. The density is the derivative of the distribution function with respect to x

$$f_{\widehat{\theta}_n}(x) = \frac{dP_\theta\left(\widehat{\theta}_n \le x\right)}{dx} = n\frac{x^{n-1}}{\theta^n} \qquad x \in [0, \theta]$$

Thus the distribution is the MLE is $n\frac{x^{n-1}}{\theta^n}$, which is <u>not</u> anywhere near normal, even for large n. This is one strange distribution. Now we obtain a very strange variance. We start with the expectation:

$$\mathrm{E}[\widehat{\theta}_n] = n\int_0^\theta x\frac{x^{n-1}}{\theta^n}dx = \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_{x=0}^{\theta} = \frac{n}{n+1}\theta.$$

Thus there is a finite sample bias:

$$\mathrm{E}[\widehat{\theta}_n] - \theta = \frac{n}{n+1}\theta - \theta = \frac{-1}{n}\theta,$$

which tends to zero as $n \to \infty$. Next we consider the variance. We start with the second moment:

$$\mathrm{E}[\widehat{\theta}_n^2] = n\int_0^\theta x^2\frac{x^{n-1}}{\theta^n}dx = \frac{n}{\theta}\left[\frac{x^{n+2}}{n+2}\right]_{x=0}^{\theta} = \frac{n}{n+2}\theta^2.$$

Thus

$$\begin{aligned} \mathrm{var}[\widehat{\theta}_n] &= = \mathrm{E}[\widehat{\theta}_n^2] - (\mathrm{E}[\widehat{\theta}_n])^2 = \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 \\ &= n\left(\frac{1}{n+2} - \frac{n}{(n+1)^2}\right)\theta^2 \\ &= n\left(\frac{n+1-n}{(n+2)(n+1)^2}\right)\theta^2 = \frac{n}{(n+1)^2(n+2)}\theta^2. \end{aligned}$$

Now observe what happens to the variance as n gets large

$$\text{var}[\widehat{\theta}_n] = \frac{n}{(n+1)^2(n+2)}\theta^2 = \frac{1}{(n+1)^2(n/n+2/n)}\theta^2 \sim \frac{1}{(n+1)^2}\theta^2.$$

The standard error for the MLE is

$$\sqrt{\text{var}[\widehat{\theta}_n]} = \frac{1}{(n+1)\sqrt{(n/n+2/n)}}\theta \sim \frac{1}{(n+1)}\theta.$$

This is (much, much) faster than the usual standard error of the sample mean which is $\sigma/\sqrt{n}$. Thus the variance of the MLE for the uniform distribution is in general "very good" for small sample sizes (indeed the bias and the variance are of the same "size"). But we do observe that the standard error for the MLE depends on the parameter θ. Thus the larger θ, the more spread out the data and the larger that standard error of the MLE.

Figure 3.24 offers an intuitive explanation as to why the MLE for the uniform distribution is so "good".

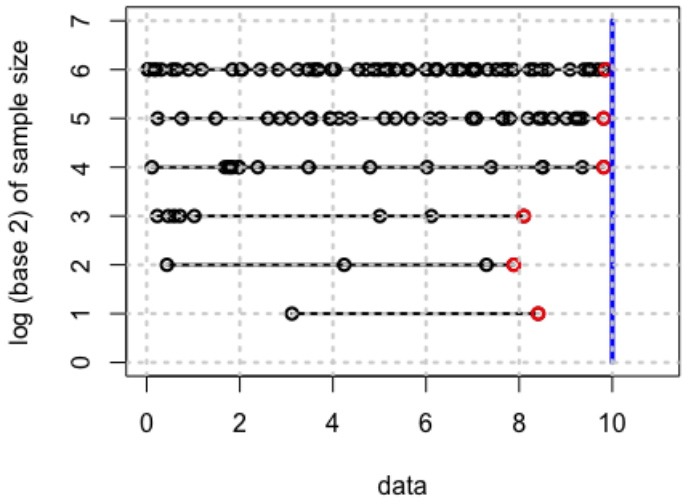


Figure 3.24: A "typical" set of realisations from a uniform distribution with $\theta = 10$. Sample sizes $n = 2, 2^2 = 4, 2^3 = 8, 2^4 = 16, 2^5 = 32, 2^6 = 64$. Red point is the MLE.

3.8 What is the best estimator?

Remember We measure the quality of an estimator by its mean squared error, or variance (if it is close to unbiased). If the estimator is close to unbiased and its true variance very difficult to evaluate, then studying its asymptotic variance is the simplest thing to do. For example, the variance of $1/\bar{X}_n$ is usually very difficult to derive but we can use the asymptotic variance given in Lemma 1.2.

3.8.1 Measuring efficiency

For many different procedures there is no unique estimation method. For example, in the previous sections we considered the method of moments estimator of the Gamma distribution and the MLE for the Gamma distribution. Given the array of estimation methods how would one choose a particular method. It seems sensible to use the estimator which has the greatest chance of "concentrating" about true parameter. There are various ways to measure this, but one method is the mean squared error $\mathrm{E}[\widehat{\theta}_n - \theta]^2$, and to select the method with the smallest MSE for a given n. We recall (see Section 1.5.1) that

$$\mathrm{E}[\widehat{\theta}_n - \theta]^2 = \mathrm{var}(\widehat{\theta}_n) + \underbrace{\left(\mathrm{E}[\widehat{\theta}_n] - \theta\right)^2}_{\text{bias squared}}.$$

Assuming that the estimator in unbiased (or nearly unbiased), then one would compare the variances of the estimator. Often this is not easy to derive (especially for finite sample sizes), but the asymptotic variance can often be derived.

Example 3.4 (Gamma distribution). *The asymptotic variance of the method of moments estimator based on the first moment (as given in (3.3)) is approx*

$$\left(\widehat{\alpha}_{MoM,n} - \alpha\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\alpha}{n}\right).$$

Similarly, by using the MLE estimator (see (3.8)) we approximately have

$$\left(\widehat{\alpha}_{MLE,n} - \alpha\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, n^{-1}\left(\frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2}\right)^{-1}\right).$$

Thus the asymptotically most efficient estimator is the one with the smallest variance

$$\frac{1}{n}\left(\frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2}\right)^{-1} \quad vs \quad \frac{\alpha}{n}. \tag{3.9}$$

Definition 3.2 (Efficiency). *Given the two estimators $\widehat{\theta}_{1,n}$ and $\widehat{\theta}_{2,n}$ which are unbiased (or it is very small), we measure the efficiency of $\widehat{\theta}_{1,n}$ relative to $\widehat{\theta}_{2,n}$ using the following measure:*

$$eff(\widehat{\theta}_{1,n}, \widehat{\theta}_{2,n}) = \frac{\mathrm{var}(\widehat{\theta}_{1,n})}{\mathrm{var}(\widehat{\theta}_{2,n})}.$$

Example 3.5 (Gamma distribution (cont)). *For the gamma distribution example, the (asymptotic) relative efficiency (since we do not have an analyic expression for the variance of $\widehat{\alpha}_{mle,n}$) based on (3.9)*

$$\begin{aligned} asyeff(\widehat{\theta}_{MoM,n}, \widehat{\theta}_{MLE,n}) &= \frac{\alpha}{n} \times \left(\frac{1}{n}\left(\frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2}\right)^{-1}\right)^{-1} \\ &= \alpha\left(\frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2}\right). \end{aligned}$$

A plot of asyeff$(\widehat{\theta}_{MoM,n}, \widehat{\theta}_{MLE,n})$ is given in Figure 3.25. We observe that asyeff$(\widehat{\theta}_{MoM,n}, \widehat{\theta}_{MLE,n}) > 1$, thus (asymptotically) the MLE estimator tends to be more efficient than the method of moments estimator (an explanation as to why is given in the next section). However, as α grows, the relative efficiency converges to one. This means that for large α (asymptotically) there is very little difference in performance of the estimators.

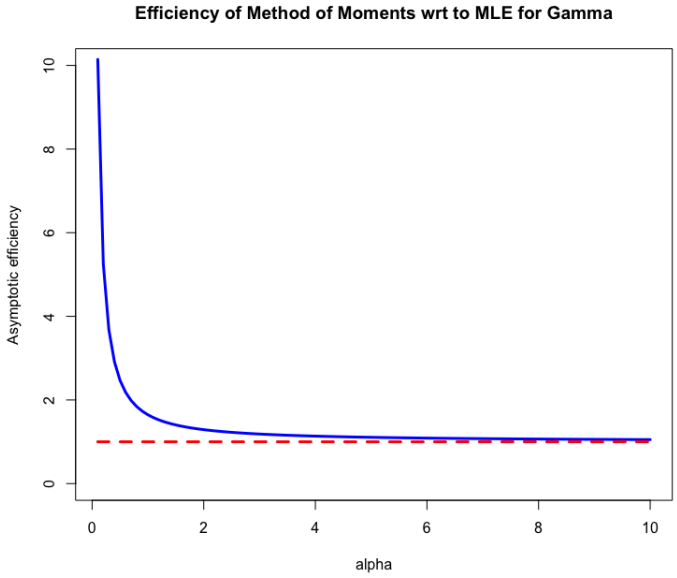


Figure 3.25: A plot of asyeff$(\widehat{\theta}_{MoM,n}, \widehat{\theta}_{MLE,n})$ for the Gamma distribution.

In the following two sections we explain why the MLE (under uncertain regularity conditions) performs so well.

Example 3.6 (The method of moments estimator and MLE for the uniform distribution parameter)**.**

HW Question: Compare different method of moment estimators of exponential distribution.

3.8.2 The Cramer-Rao Bound

As we mentioned in the previous section, different estimators may have different mean squared errors. Any one family of distributions may have an infinite number of different estimator to select. How, do we know if we have the best estimator or if it can be improved. If there existed a lower bound for mean squared error, we can use this as a bench mark to compare the estimator with. If the variance of our estimator attains the lower bound or it is only slightly higher than it, then there is no need to search for a better estimator.

We now state a classical result where such a lower bound is obtained. This result only applies to the class of estimators which are unbiased, nevertheless it is extremely powerful and useful. It is interesting to note that was independently derived by C.R. Rao when he was only 24 years old and Harold Cramer.

Theorem 3.4

Suppose $\{X_i\}_{i=1}^n$ are iid random variables with density $f(x;\theta)$. Let $T(X_1,\ldots,X_n)$ be an unbiased estimator of θ i.e. $\mathrm{E}[T(X_1,\ldots,X_n)] = \theta$. Then under certain regularity conditions on the density we have

$$\mathrm{var}[T(X_1,\ldots,X_n)] \geq \frac{1}{nI(\theta)},$$

where $I(\theta)$ denotes the Fisher information matrix corresponding X.

This result gives a lower bound on how small the variance of an unbiased estimator can actually be. Interestingly, we previously showed that asymptotic variance of the MLE is $(nI(\theta))^{-1}$. This means that for finite samples the MLE may not be best estimator of θ, but at least asymptotically (this means for large samples), using the MLE will usually yield an estimator that is close to the best. This observation justifies the frequent use of the MLE in estimation.

3.9 Sufficiency

Suppose you are an astronaut/cosmonaut exploring Pluto (the unfortunate x-planet). You are collecting rock samples and want to transmit their weights to earth. However, it takes several minutes to transmit one piece of information (the weight of one rock) and you have 1000 rocks. It would take several days to transmit all the information about the rocks to earth. The people on earth are getting quite impatient and want the information as fast as possible. What do you do? Is it really necessary to give the people on earth the exact data (data specific information). May be it is sufficient to produce a synthetic data (based on the information you send then), which reproduces the main features in the true data as closely as possible. If such information were sufficient, we should transmit the maximum information about the distribution of the data.

We now make this precise, by defining the notion of a sufficient statistic.

Definition 3.3 (Sufficiency)**.** *Suppose the joint distribution of $\underline{X} = (X_1,\ldots,X_n)$ depends on the unknown parameter θ. Let $T(\underline{X})$ be a function of the data $\underline{X}$. Then $T(\underline{X})$ is called a sufficient statistic for the parameter θ, if the conditional distribution of $\underline{X}$ given $T(\underline{X})$ does not depend on the parameter θ. Formally we write this as*

$$F(\underline{X}|T(\underline{X})) = g(\underline{X}).$$

In other words, the conditional distribution of F conditioned on the sufficient statistic $T(\underline{X})$ does not depend on the parameter θ. The distribution function $g(\underline{X})$ does not depend on θ.

The above above definition tells us the following. Suppose the sample mean $\bar{X}_n$ is a sufficient statistic for the population mean μ (for some distribution). If we observe the the sample mean $\bar{X}_n = 3$. Then the conditional distribution of $X_1, \ldots, X_n$ conditioned on $n^{-1}\sum_{i=1}^n X_i = 3$ does not depend on the population mean μ. Returning to our the example of you collecting rocks samples from Pluto. If it is known that the distribution of the weight of rocks comes form a known parametric family, with unknown parameter θ, and $T(\underline{X})$ is a sufficient statistic for θ. Rather than transmitting the weights of 1000 rocks to earth you can simply evaluate and send $t = T(\underline{X})$ to earth (which only takes a few minutes). And the people back on earth can produce a synthetic data set by drawing samples from numbers from the conditional density

$$f(X_1 = x_1, \ldots, X_n = x_n | T(\underline{X}) = t),$$

which is known (since the parametric density is known and the above conditional density does not depend on unknown parameters). This procedure does not give the entire data set, but gives a reconstruction of the data set based on some vital information about it.

The above is a rather artifical example. But a sufficient statistic contains all the important ingredients/information about the unknown parameter. And "most" good estimators will be a function of the sufficient statistic. We show in Section 3.9.1, that a sufficient statistic can be used to improve an estimator (so it is closely related to estimation).

But we start by giving an example of a sufficient statistic and then state necessary and sufficient conditions for sufficiency. This leads to a simple method for obtaining a sufficient statistic from a distribution and data.

Example 3.7 (Bernoulli random variables)**.** *Suppose $\{X_i\}_{i=1}^n$ are iid Bernoulli random variables where $P(X_i = 0) = 1 - \pi$ and $P(X_i = 1) = \pi$. We now show that $T_n(\underline{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for π.*

Note that the joint distribution of $X_1, \ldots, X_n$ is

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n (1-\pi)^{1-x_i}\pi^{x_i},$$

which clearly depends on π. We will show that $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T_n(\underline{X}) = t)$ does not depend on π. There are two ways this can be shown.

We first use the brute force method. Using the classical $P(A|B) = P(A \cap B)/P(B)$ we have

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T_n(\underline{X}) = t) = \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T_n(\underline{X}) = t)}{P(T_n(\underline{X}) = t)}.$$

To evaluate the joint probability $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T_n(\underline{X}) = t)$ we first consider a some simple example $n = 2$ and $T_n(\underline{X}) = 1$, then

$$P(X_1 = 1, X_2 = 1, T_n(\underline{X}) = 1) = 0 \text{ but } P(X_1 = 1, X_2 = 0, T_n(\underline{X}) = 1) = P(X_1 = 1, X_2 = 0) = \pi(1-\pi).$$

Using the above, in general we observe that

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T_n(\underline{X}) = t) = \begin{cases} \pi^t(1-\pi)^{n-t} & \sum_{i=1}^n x_i = t \\ 0 & \sum_{i=1}^n x_i \neq t \end{cases}$$

This deals with the numerator. Next, we consider the denominator: since $T_n = \sum_{i=1}^n X_i \sim Bin(n, p)$*, then* $P(T_n = t) = \binom{n}{t}\pi^t(1-\pi)^{n-t}$*. Putting the numerator and denominator together gives the conditional probability*

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T_n(\underline{X}) = t) &= \begin{cases} \frac{\pi^t(1-\pi)^{n-t}}{\binom{n}{t}\pi^t(1-\pi)^{n-t}} & \sum_{i=1}^n x_i = t \\ \frac{0}{\binom{n}{t}\pi^t(1-\pi)^{n-t}} & \sum_{i=1}^n x_i \neq t \end{cases} \\ &= \begin{cases} \frac{1}{\binom{n}{t}} & \sum_{i=1}^n x_i = t \\ 0 & \sum_{i=1}^n x_i \neq t \end{cases} \end{aligned}$$

In other words, if $\sum_{i=1}^n x_i = t$ *then*

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T_n(\underline{X}) = t) = \frac{1}{\binom{n}{t}}.$$

A simpler method would be to use basic combinatorical arguments.

To summarize, what we observe is that if the total number of successes is known, than $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T_n(\underline{X}) = t)$ *does not depend on the underlying parameter* π*. Thus* $T_n(\underline{X}) = \sum_{i=1}^n X_i$ *is a sufficient statistic for* π*.*

In general it is difficult to "come up" with sufficient statistics. The following result gives necessary and sufficient conditions for a sufficient statistic.

Theorem 3.5

A necessary and sufficient condition for $T(\underline{X})$ to be a sufficient statistic for θ is that the joint probability function of $\underline{X}$ (either density or probability mass function) can be factorised as follows

$$f_{\underline{X}}(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n) g\left[T(x_1, \ldots, x_n); \theta\right].$$

Remark: The important aspect of this result, is that the function $h(\cdot)$ does not depend on θ. All the information on the function θ is contained in the function $g(\cdot; \theta)$.

It is often (but not always) easier to consider the log of the joint probability function. By the factorisation theorem the log probability is

$$\log f_{\underline{X}}(x_1, \ldots, x_n; \theta) = \log h(x_1, \ldots, x_n) + \log g\left[T(x_1, \ldots, x_n); \theta\right].$$

Example: The exponential distribution

Suppose $\{X_i\}$ are iid random variables with exponential density $f(x;\lambda) = \lambda \exp(-\lambda x)$. Then the log of the joint density is

$$\begin{aligned} \log f_{\underline{X}}(x_1,\ldots,x_n;\lambda) &= \sum_{i=1}^{n} \log f(x;\lambda) = n\log\lambda - \lambda \underbrace{\sum_{i=1}^{n} x_i}_{=T(x_1,\ldots,x_n)} \\ &= \log g\left(T(X_1,\ldots,X_n),\lambda\right). \end{aligned}$$

All the data is summarized in $T(\underline{x}) = \sum_{i=1}^{n} x_i$. Thus by the factorisation theorem $T(\underline{x})$ is a sufficient statistic for λ.

Example: The normal distribution

We use the log-expansion given in Example 2.3. The log of the normal density is

$$\log f_{\underline{X}}(x_1,\ldots,x_n;\lambda) = \sum_{i=1}^{n} \log f(x;\lambda) = \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{n\mu^2/2}{\sigma^2} + \frac{n}{2}\log(2\pi).$$

Again we see that the data is "clustered together" in two terms $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$. Thus we observe that $T_1(\underline{x}) = \sum_{i=1}^{n} x_i$ and $T_2(\underline{x}) = \sum_{i=1}^{n} x_i^2$ are together sufficient statistics for μ and σ^2.

Example: All distributions in the exponential family

The above two examples both come from the exponential family, with good reason. All distributions which come from the exponential family can be summarized in a few sufficient statistics (that do not depend on the sample size). To understand why, we recall from Section 2.3 that a family of distributions belong to the exponential family if they can be written as

$$f(x;\underline{\theta}) = \exp\left[\sum_{j=1}^{K} s_j(x)T_j(\underline{\theta}) + b(\underline{\theta}) + c(x)\right] \qquad x \in A,$$

where A does not depend on the parameter $\underline{\theta}$ and $\underline{\theta} = (\theta_1,\ldots,\theta_K)$. Thus if $\{X_i\}$ are iid random variables with distribution $f(x;\theta)$, then their joint density is

$$\begin{aligned} \log f_{\underline{X}}(x_1,\ldots,x_n;\lambda) &= \sum_{i=1}^{n}\left[\sum_{j=1}^{K} s_j(x_i)T_j(\underline{\theta}) + b(\underline{\theta}) + c(x_i)\right] \\ &= \underbrace{\sum_{j=1}^{K}\sum_{i=1}^{n} s_j(x_i)T_j(\underline{\theta}) + nb(\underline{\theta})}_{g[T_1(\underline{x}),\ldots,T_K(\underline{x});\underline{\theta}]} + \underbrace{\sum_{i=1}^{n} c(x_i)}_{\log h(x_1,\ldots,x_n)}. \end{aligned}$$

Thus $T_1(\underline{x}),\ldots,T_K(\underline{x})$ are collectively sufficient statistics for $\underline{\theta}$. Observe that since $T_1(\underline{x}),\ldots,T_K(\underline{x})$ does not depend on the sample size, we are able to summarize important aspects of the data in just K terms.

Example: The uniform distribution (not in exponential family)

The uniform distribution is an example of a distribution that does not belong to the exponential family but whose number of sufficient statistics does not depend on the sample size. We recall that if $\{X_i\}$ are iid random variables with uniform density, then their joint density (see Section 3.4.2) is

$$f_{\underline{X}}(x_1, \ldots, x_n; \theta) = \theta^{-n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i),$$

where $I_{[0,\theta]}(x)$ is the indicator variable, which is one for $x \in [0, \theta]$ and zero elsewhere. With some thought we observe that this can be written as

$$f_{\underline{X}}(x_1, \ldots, x_n; \theta) = \theta^{-n} I_{[0,\theta]}(\max_i x_i).$$

Thus $T(\underline{x}) = \max_i x_i$ is a sufficient statistic for θ.

Example: The Weibull distribution; sufficient statistic cannot be reduced

We now show that the sufficient statistics of the Weibull distribution cannot be reduced to just a few (which do not depend on the sample size).

Suppose that $\{X_i\}$ are iid random variables with the Weibull distribution;

$$f(x; \lambda, k) = \left(\frac{k}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left[\frac{x}{\lambda}\right]^k\right) \qquad x \geq 0.$$

The log of the joint density is

$$\log f_{\underline{X}}(x_1, \ldots, x_n; \lambda, k) = -\frac{1}{\lambda^k} \sum_{i=1}^{n} x_i^k + (k-1) \sum_{i=1}^{n} \log x_i - n(k-1) \log \lambda + n \log k - n \log \lambda.$$

Now suppose that k is assumed known. Then $\sum_{i=1}^{n} x_i^k$ is a sufficient statistic for λ; since

$$\log f_{\underline{X}}(x_1, \ldots, x_n; \lambda, k) = \underbrace{-\frac{1}{\lambda^k} \sum_{i=1}^{n} x_i^k - n(k-1) \log \lambda - n \log \lambda}_{g(T(\underline{x}); \lambda)} + \underbrace{(k-1) \sum_{i=1}^{n} \log x_i}_{h(\underline{x})}.$$

However, no such sufficient statistic exists for k (except for the original data).

3.9.1 Application of sufficiency to estimation: Rao-Blackwellisation

Sufficiency has various applications in statistcis. But we conclude this section with a very elegant application to estimation. The theorem below is called the Rao-Blackwell theorem.

Theorem 3.6

Suppose that $\widehat{\theta} = \widehat{\theta}(\underline{X})$ is an estimator of θ based on the random variables $\underline{X}$ (and $\mathrm{E}[\widehat{\theta}^2] < \infty$) and $T(\underline{X})$ is a sufficient statistic for θ. Define the "new estimator" $\widetilde{\theta} = \mathrm{E}[\widehat{\theta}(\underline{X})|T(\underline{X})]$.

Then

$$\mathrm{E}\left[\widetilde{\theta} - \theta\right]^2 \leq \mathrm{E}\left[\widehat{\theta} - \theta\right]^2.$$

PROOF. The proof is straightforward (but will not be tested). We first note that by iterated expectation we have

$$\mathrm{E}\left(\widehat{\theta}\right) = \mathrm{E}\left(\mathrm{E}\left[\widehat{\theta}|T(\underline{X})\right]\right) = \mathrm{E}\left(\widetilde{\theta}\right)$$

Thus the bias of both $\widehat{\theta}$ and $\widetilde{\theta}$ are the same. Next we focus on the variance. By using the well known conditonal variance identity (see HW2) we have

$$\mathrm{var}[\widehat{\theta}] = \mathrm{var}[\mathrm{E}[\widehat{\theta}|T(\underline{X})]] + \mathrm{E}\left(\mathrm{var}[\widehat{\theta}|T(\underline{X})]\right).$$

Since $\mathrm{E}\left(\mathrm{var}[\widehat{\theta}|T(\underline{X})]\right) \geq 0$, this immediately implies that

$$\mathrm{var}[\widehat{\theta}] \geq \mathrm{var}[\mathrm{E}[\widehat{\theta}|T(\underline{X})]] = \mathrm{var}[\widetilde{\theta}].$$

Thus giving the required result. □

The Rao-Blackwell theorem allows one to improve an estimator by conditioning on a sufficient statistic. It also makes is understand that "good estimators" should be functions of the sufficient statistic. If you return to the MLE, in particular the MLE of the exponential family, you will observe that the maximum likelihood estimator is a function of the sufficient statistics (this will take some algebraic manipulation). This gives credence to the claim that the MLE yield good estimators.

Under stronger conditions on the sufficient statistic, one can show that if an unbiased estimator is a function of the sufficient statistic it cannot be improved (this result is beyond this class but is related to minimal sufficiency and completeness). It has the smallest/minimal variance.

Remark. *The MLE will be a function of the sufficient statistic. Under the additional condition of minimal sufficiency and that the estimator is unbiased, then it will have the smallest variance that an estimator can have.*

3.10 What happens if we get the assumptions wrong

NEED TO DO.

3.11 A historical perspective

Mention Jackknife (bias reduction) proposed by Maurice Quenouille. Discuss Blackwell one of the early Afro-American statistician who made fundamental contributions to statistics.

Also discuss C.R.Rao.

4 Hypothesis testings

The aim of this chapter is to develop an understanding of a statistical test. In particular, why we use certain types of tests. We start by reviewing some of the testing methods you have previously encountered.

4.1 A short review

Let us start with the simplest example you may have encountered. Suppose $\{X_i\}$ are iid normal random variables with mean μ (which is unknown) and variance σ^2, which for now is assumed known. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. From Section 2.1 we have

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Our aim is to use the results above to make a decision between two hypotheses, the null and the alternative. We recall that the null and alternative are not treated as "equals". A statistical test is a decision process, where we either reject the null in favour of the alternative, or do not reject the null. As it is a decision process, we can alway make a mistake. We recall that if we falsely reject the null, when the null is true, this is called a type I error. If on the other hand, we do not reject the null when the alternative is true, this is called a type II error. The emphasis is always on having full control of the null hypothesis and the type I error. The classical example is a criminal trial where H_0: innocent against H_A: guilty. Usually our aim is to ensure that the number of innocent people that are convicted of a crime is minimal and no more than a certain percentage. Thus we are controlling the the type I error.

4.1.1 The simple hypothesis

A simple hypothesis is when the distribution under both the null and alternative are completely specified. Let us return to the iid normal example and test $H_0 : \mu = \mu_0$ vs $H_A : \mu = \mu_1$. The test is conducted at the α%-level. This means the type I error is α%. In most situations a simple hypothesis is a highly unrealistic situation. However, it does allow us to get a handle on what a statistical test can actually do. We will show later that for simple hypothesis tests one can construct a test which is optimal (is best at detecting the alternative); this is called the Neyman-Pearson Lemma. Once such a test can be constructed we can use it to

compare it with testing procedures in the more realistic set up that the hypotheses is not simple. This will allow us to decide if these testing procedures are as good as the optimal testing procedures. Interestingly, there will be situations where they are.

To fix ideas, let us consider the example where $H_0 : \mu = 1$ against $H_A : \mu = 4$ (and $\sigma^2 = 1$ is assumed known) and $n = 1$. The test is conducted at the 5% level. We reject the null if the sample mean $\bar{X} \geq 1 + 1.64 \times \frac{\sigma}{\sqrt{n}}$ (where in this case $\sigma = 1$ and $n = 1$). A plot of the two densities and the rejection region is given in Figure 4.1. The area to the right of the blue/red is the rejection region. And if the null hypothesisis true, there is a 5% chance this can happen. We often formally write the above decision as follows. Let C denote the

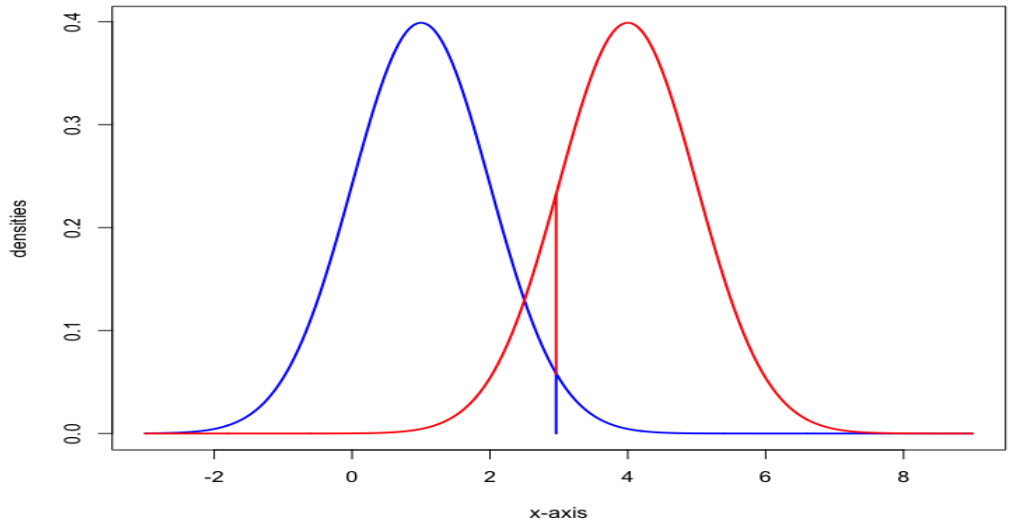


Figure 4.1: $H_0 : N(1, 1)$ vs $H_A : N(4, 1)$ with $n = 1$. Area to the right of blue/red line and below blue curve is 5% (under H_0).

rejection region. Then given the data $\underline{X}$ we define the variable

$$\delta(\underline{X}) = \begin{cases} 1 & \underline{X} \in C \\ 0 & \underline{C} \notin C \end{cases}$$

If $\delta(\underline{X}) = 1$ we reject the null, and:

(i) $P(\delta(\underline{X}) = 1|H_0)$ is the Type I error.

(ii) $P(\delta(\underline{X}) = 1|H_A)$ is the power of the test. The Type II error is

$$P(\delta(\underline{X}) = 0|H_A) = 1 - P(\delta(\underline{X}) = 1|H_A).$$

Returning to the example, suppose $C = [2.64, \infty)$, then

$$P\left(\delta(\underline{X}) = 1|H_0\right) = P\left(X \geq 2.64|H_0\right) = P\left(\frac{X-1}{1} \geq \frac{2.64-1}{1}\right) = P(Z \geq 1.64) = 0.05, \tag{4.1}$$

since $Z \sim N(0,1)$.

But why use this rejection region?. We could easily find other regions where the $P(X \in R|H_0) = 5\%$. For example, if we let $R_1 = (-\infty, -0.64)$, then using the same calculation as above, we can show that $P(X \in R_1|H_0) = 0.05$. Why not use a rejection region which is the union of disjoint intervals.

There is a good reason to use $C = [2.64, \infty)$ It transpires that the region $C = [2.64, \infty)$ is the best region to use, because it keeps the type I error at 5%, but maximises the probability (power) of detecting the alternative $H_A : \mu = 4$. The power can easily be calculated for this rejection region C by calculating the area below the red curve (to the right of the blue/red line)

$$P\left(\delta(\underline{X}) = 1|H_A\right) = P\left(X \geq 2.64|H_A\right) = P\left(\frac{X-4}{1} \geq \frac{2.64-4}{1}\right) = P(Z \geq -1.36) = 0.913.$$

In constrast, by the same argument, the rejection region $R_1(-\infty, -0.64)$ would have led to a power close to zero ($P(X \in R_1|H_a) = P(X \leq -0.64|H_A) = P(Z \leq -4.64) \approx 0$). Visually, it is clear from Figure 4.1 that C is the rejection region which maximises the power since it is in the direction of the alternative. Interestingly the actual value of θ_1 plays no role in the test, only its direction with respect to θ_0 (which determines the location of the rejection region; either left or right of θ_0).

To summarize:

- If we reduce the level α (the type I error), then the rejection interval gets pushed further to the right. This reduces the power of test (the probability of detecting the alternative).
- If we increase the sample size, but keep the level at, say 5%, then the rejection region becomes

  $$C = \left[1 + 1.64 \times \frac{1}{\sqrt{n}}, \infty\right).$$

  And the power of the test is

  $$\begin{aligned} P\left(\delta(\underline{X}) = 1|H_A\right) &= P\left(\bar{X} \geq 1 + 1.64n^{-1/2}|H_A\right) = P\left(\frac{\bar{X}-4}{n^{-1/2}} \geq \frac{1 + 1.64n^{-1/2} - 4}{n^{-1/2}}\right) \\ &= P(Z \geq -3n^{1/2} + 1.64). \end{aligned}$$

  Clearly, the power grows with the sample size (but the type I error remains the same).
- For any alternative $\mu_1 > \mu_0$, the general rejection region is

  $$C = \left[\mu_0 + z_\alpha \times \frac{\sigma}{\sqrt{n}}, \infty\right).$$
- Again, you should ask yourself why use the critical region

  $$C = \left\{\underline{X}_n; \bar{X} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right\}$$

  why not another region $R \subset \mathrm{R}^n$ where $P(\underline{X} \in R|H_0) = \alpha$?

Example 4.1 (General means). *Let $H_0 : \mu = \mu_0$ vs $H_A : \mu = \mu_1$ (where $\mu_1 > \mu_0$). To test this hypothesis, we observe the iid normal random variables $\{X_i\}_{i=1}^n$ with mean μ and variance σ^2. We use the z-test at the α significance level, our objective is to calculate the power of the test.*

The rejection region is $C = [\mu_0 + z_\alpha \sigma/\sqrt{n}, \infty)$. Thus the power of the test is

$$\begin{aligned} P\left(\bar{X} \geq \mu_0 + z_\alpha \sigma/\sqrt{n} | \mu = \mu_1\right) &= P\left(\frac{\sqrt{n}(\bar{X} - \mu_1)}{\sigma} \geq \frac{\sqrt{n}(\mu_0 + z_\alpha \sigma/\sqrt{n} - \mu_1)}{\sigma} | H_0\right) \\ &= P\left(Z \geq \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma} + z_\alpha | H_0\right). \end{aligned}$$

The power should always be illustrated with a plot.

4.1.2 Composite hypothesis

The case of the sample hypothesis is the simplest setting for a hypothesis test. In most real life settings it is not realistic to assume that the distribution is fully specified by the the null and alternative. When this is not the case, we are in composite hypothesis framework. Composite hypothesis can arise in several different ways. For example, the alternative hypothesis may not be completely specified, examples include $H_A : \mu > \mu_0$ (unlike previously where $H_A : \mu = \mu_1$) or $H_A : \mu < \mu_0$ or $H_A : \mu \neq \mu_0$. Alternatively, it could be that neither the null or alternatively are completely specified. For example, the hypothesis may concern the mean μ, but the variance σ^2 is unknown. We consider some examples below.

One sided tests

Suppose we observe the iid normal random variables $\{X_i\}$ with mean μ and variance σ^2. Once again the variance is assumed known. We consider tests of the form $H_0 : \mu = \mu_0$ against $H_A : \mu > \mu_0$. Observe this is a composite hypothesis in the alternative (as the distribution null hypothesis is completely specified, it is $X_i \sim N(\mu_0, \sigma^2)$). Again the aim is to control the Type I error, but in such a way that the power is maximised. Since the rejection region for the simple hypothesis does not depend on the alternative μ_1 (on its direction with respect to μ_0), exactly the same rejection region described in the simple hypothesis test applies to the one-sided test. Thus at the 5% level, the rejection region is

$$C = \left[\mu_0 + 1.64 \times \frac{\sigma}{\sqrt{n}}, \infty\right).$$

And, in general, at α% level the rejection region

$$C = \left[\mu_0 + z_\alpha \times \frac{\sigma}{\sqrt{n}}, \infty\right).$$

If tests is the form $H_0 : \mu = \mu_0$ against $H_A : \mu < \mu_0$, then at the 5% level the rejection region is

$$C = \left[\mu_0 - 1.64 \times \frac{\sigma}{\sqrt{n}}, \infty\right).$$

Two sided tests

How to perform the two-sided test where $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ is not so clear cut. The intervals described above for one-sided tests do not apply to the two sided. For example, if we use the rejection region $[\mu_0 + 1.64 \times \frac{\sigma}{\sqrt{n}}, \infty)$, but for the alternative $\mu < \mu_0$, then the test has no power. The standard compromise is to evenly "distribute" the rejection region on both sides of μ_0. Such that the rejection region at the 5% level is

$$C_- \cup C_+ \text{ where } C_- = \left(-\infty, \mu_0 - 1.96 \times \frac{\sigma}{\sqrt{n}}, \infty\right] \text{ and } C_+ = \left[\mu_0 + 1.96 \times \frac{\sigma}{\sqrt{n}}, \infty\right).$$

Further, at any α level the rejection region is

$$C_- \cup C_+ \text{ where } C_- = \left(-\infty, \mu_0 - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \infty\right] \text{ and } C_+ = \left[\mu_0 + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \infty\right).$$

A compositive null hypothesis (when the variance is unknown): The t-test

In the previous tests we have assumed the variance is known and thus the distribution under the null was fully specified. As we have shown in Section 2.4.3 this is usually not a realistic assumption. In this case we replace σ^2 with its estimator. Suppose $\{X_i\}$ are iid random variables with mean μ and variance σ^2. Our aim is to test $H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0$. Then the t-statistic is

$$T_n = \frac{(\bar{X} - \mu_0)}{s_n/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s_n},$$

which measures the standardized distance between μ_0 and $\bar{X}$. We know that under the null hypothesis that $T_n \sim t_{n-1}$. Thus rejection region for the test (at the α significance level) is

$$C = \left[\mu_0 + t_{n-1,\alpha} \times \frac{s_n}{\sqrt{n}}, \infty\right).$$

4.2 The likelihood ratio test

By the mid 1920s there was a whole bunch of different tests for various different hypothesis (the t-test, independent sample t-test, the chi-square test for independence to name but a few). However, at that time it still was not clear what exactly linked all these tests together and what test, if any, was better than another. Then in a series of seminal papers Neyman and Pearson introduced the principle of likelihood[1]. This method linked many of the tests already in use and introduced new testing methods. Further, in 1933, Neyman and Pearson showed that the principle of likelihood (what is now called the likelihood ratio test)

[1] To understand the Neyman and Pearson thought process, here are their 1928 and 1933 papers, the latter gives the Neyman-Pearson Lemma.

in fact was the most efficient (or powerful testing method). This is now referred to as the Neyman-Pearson Lemma (and will be discussed below). We start with this section by introducing the likelihood ratio test (in a subsequent section we show that a variant of this test includes the t-test and various other tests as a special case).

We focus first on the case of a simple hypothesis. Suppose the random vector $\underline{X} = (x_1, \ldots, x_n)$ has either the density $f(\underline{x}; \theta_0)$ or $f(\underline{x}; \theta_1)$ (it two distributions could also be f_0 and f_1, it does not matter)). We do not know which, but we place importance on the null hypothesis (i.e. we only reject the null if the evidence points away from it) and pose the null and alternative as the simple hypothesis $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_1$. The Likelihood ratio test is based on the ratio

$$\frac{f(\underline{X}; \theta_1)}{f(\underline{X}; \underline{\theta}_0)} = \frac{L_n(\theta_1; \underline{X})}{L_n(\underline{\theta}_0; \underline{X})}$$

(note that without any loss of generality we can switch the ratio around $\frac{f(\underline{x};\theta_0)}{f(\underline{x};\underline{\theta}_1)}$). Keep in mind the joint density and the likelihood are the same. The rationale for considering this ratio is similiar to the likelihood estimator; in general data tends to lie where the probability mass function is largest. Thus if we are choosing between two density functions, we select the density which dominates the other for a given data set. We return to the ratio

$$LR(\underline{X}) = \frac{f(\underline{X}; \theta_1)}{f(\underline{X}; \underline{\theta}_0)} = \frac{L_n(\theta_1; \underline{X})}{L_n(\underline{\theta}_0; \underline{X})}.$$

Treating $f(\underline{x}; \underline{\theta}_0)$ and $f(\underline{x}; \underline{\theta}_0)$ as probabilities if $LR(\underline{x}) > 1$ then the probability of $\underline{x}$ under the alternative is greater than under the null. Conversely, if $LR(\underline{x}) < 1$ then the probability of $\underline{x}$ under the null is greater than under the alternative. This would make a sensible method for deciding between the null and the alternative. But in statistical test both distributions are not given the same weighting. The emphasis is always on disproving the null. Thus to err on the cautious side, rather than using the threshold 1 for making the decision we select a K_α that is larger than one. How large depends on the choice of the type I error.

More precisely, we reject the null hypothesis if $LR(\underline{x})$ is sufficiently "large" (the situation that the null seems implausible). We select a threshold K_α and the corresponding rejection region C_α where

$$C_\alpha = \left\{ \underline{x}; \frac{f(\underline{x}; \theta_1)}{f(\underline{x}; \theta_0)} \geq K_\alpha \right\} = \left\{ \underline{x}; \frac{L_n(\theta_1; \underline{x})}{L_n(\underline{\theta}_0; \underline{x})} \geq K_\alpha \right\}$$

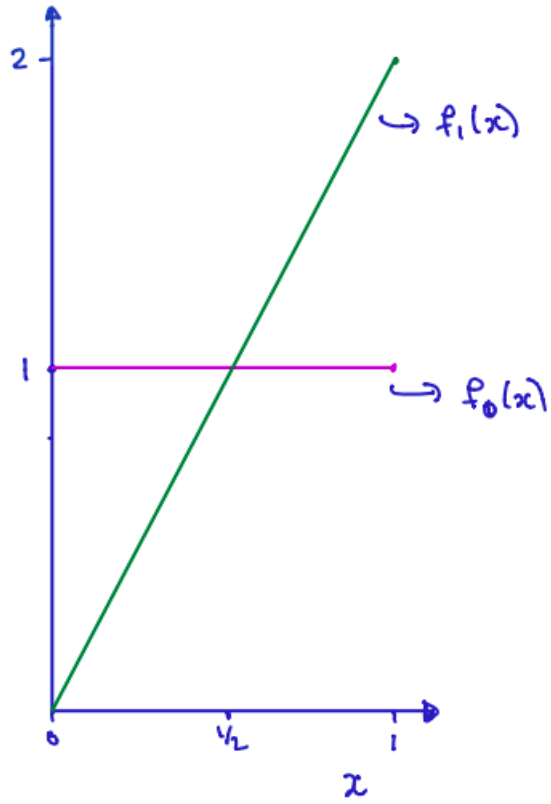
and K_α is chosen such that $P(\underline{X} \in C_\alpha | H_0) = \alpha$. I.e. keeping in mind that $LR(\underline{X})$ is a random variable we choose K such that

$$P\left(LR(\underline{X}) \geq K_\alpha \middle| H_0\right) = \alpha.$$

On first impression it is not clear how the likelihood ratio test is related to the z-test described above. Below we show that they are in fact equivalent (the same).

4.2.1 Toy Example

We start with a rather artificial example to understand what the likelihood ratio test is actually doing. Suppose that X is a continuous random variable defined on $[0, 1]$ with density f.



There are two potential candidates for the density of X, $f_0(x) = 1$ for $x \in [0, 1]$ and zero elsewhere or $f_1(x) = 2x$ for $x \in [0, 1]$ and zero elsewhere. The plot on the left gives both densities. Remember, the height of the density indicates the likelihood of that x value happening.
Observe that if $x = 0.5$ each density is equally likely. For x close to zero, f_0 is more likely, whereas for x close to one f_1 is more likely.

We use the likelihood ratio test to test $H_0 : f(x) = f_0(x)$ versus $H_A : f(x) = f_1(x)$. We use the sample size $n = 1$ (very unrealistic in practice, but at least it presents us with the idea of the method). The likelihood ratio is

$$LR(X) = \frac{2X}{1}.$$

Now we need to isolate X to construct a rejection region for X under the null hypothesis being true;

$$C_\alpha = \{x; LR(x) \geq K_\alpha\}$$

where K_α is such that

$$P(X \in C_\alpha | H_0) = P\left(LR(X) \geq K_\alpha | H_0 : f = f_0\right) = \alpha.$$

We now evaluate C_α

$$\begin{aligned} P\left(LR(X) \geq K_\alpha | H_0 : f = f_0\right) &= P\left(2X \geq K_\alpha | H_0 : f = f_0\right) = P\left(X \geq \frac{K_\alpha}{2} | H_0 : f = f_0\right) \\ &= \int_{K_\alpha/2}^{1} dx = 1 - \frac{K_\alpha}{2} = \alpha. \end{aligned}$$

Thus $K_\alpha = 2(1-\alpha)$ and the rejection region is

$$C_\alpha = \{x \geq 1-\alpha\}.$$

Thus for $x > 1-\alpha$ we reject the null. Observe for a very small α, this means that the rejection point is very close to one.

Example Suppose $\alpha = 0.1$. We observe $x = 0.95$, since $0.95 \in C_{0.1} = \{x \geq 1-0.1\}$, we reject the null. In other words, given the data there is evidence at the 10% level that the alternative is true.

Power calculation We recall that the power is the ability of the test to reject the null, that is the probability an observation lies in C_α when H_A is true. For this example it is

$$\begin{aligned} P(LR(X) \geq K_\alpha | H_A : f = f_1) &= P(X \in C_\alpha K_\alpha | H_A : f = f_1) \\ &= \int_{K_\alpha/2}^{1} 2x dx = 1 - (1-\alpha)^2 = \alpha(2-\alpha). \end{aligned}$$

Observe that since $\alpha \in (0,1)$, then $P(LR(X) \geq K_\alpha | H_A) = \alpha(2-\alpha) > \alpha$.

For $n > 1$, calculation of the rejection region becomes more difficult (but the power of the test increases for a given α). Below we consider more realistic examples, where calculation of the rejection region is straightforward even for $n > 1$.

Keep in mind the basic recipe. Calculate the rejection region by isolating the random variables in the likelihood ratio.

4.2.2 Example: The normal distribution

Below we will go through a series of (sometimes awful) algebraic manipulations. This is because often the likelihood ratio is a complicated function of the Xs. Thus figuring out that distribution of $LR(\underline{X})$ under the null is not feasible. However, by its very construction the distribution of the random variable X is specified under the null. Therefore we will transform the likelihood ratio through algebraic manipulation to isolate the Xs (or sum of Xs). Often this is done by taking the log of the likelihoods, but not always.

Now you may ask what allows us to do this. The reason is the rejection region. The rejection region consists of the values of X where this ratio is bigger than a threshold. The probability of X lying in this set under the null is α (say 5%). The algebraic manipulations *do not change* this set. It is like saying for what values of x is $x^2 - 2 > 11$, this is exactly the same as the set of x where $x > 3$ or $x < -3$. They are equivalent sets; different formulas, but the same set.

Simple hypothesis

Let us suppose that $\{X_i\}$ are iid normal random variables. We want to test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_1$ with σ known.

The case $n = 1$ (when the likelihood and the marginal density are the same)

We consider the normal example considered in Section 4.1.1, with $H_0 : \mu = 1$ against $H_A : \mu = 4$ (and $\sigma^2 = 1$ is assumed known) and $n = 1$. Since log is a monotonic function, analysis of the ratio is equivalent to the analysis of $\log \frac{f(\underline{x};\theta_1)}{f(\underline{x};\underline{\theta_0})}$. In this example we have

$$LR(x) = \frac{\exp\left(-(x-4)^2/2\right)}{\exp\left(-(x-1)^2/2\right)},$$

isolating the x is difficult. Thus we take logarithms (which does not change anything but simplifies the calculation of the rejection region)

$$\log LR(x) = \log \frac{f(x;\mu_1 = 4)}{f(x;\mu_0 = 1)} = -\frac{1}{2}(x-4)^2 + \frac{1}{2}(x-1)^2 = x(4-1) - \frac{1}{2}(4^2 - 1^2), \tag{4.2}$$

a plot is given in Figure 4.2. Observe that the ratio $\log LR(x)$, monotonically grows the further to the right x is from $\theta_0 = 1$ (this can be interpretated as stronger evidence against the null hypothesis). Our aim is to find the region C where

$$\begin{aligned} P(X \in C|\mu = 1) &= P(LR(X) \geq K|\mu = 1) = P(\log LR(X) \geq \log K|\mu = 1) \\ &= P\left(X(4-1) - \frac{1}{2}(4^2 - 1^2) \geq \log K|H_0\right) \\ &= P\left(X \geq 3^{-1}\log K + 3^{-1}\frac{1}{2}(4^2 - 1^2)|H_0\right) = \alpha. \end{aligned}$$

Thus setting $\widetilde{K} = 3^{-1}\log K + 3^{-1}\frac{1}{2}(4^2 - 1^2)$ and $\alpha = 5\%$ we find that

$$P\left(X \geq \widetilde{K}|\mu = 1\right) = 0.05 \quad \Rightarrow P\left(\underbrace{\frac{X-1}{1}}_{=Z} \geq \underbrace{\frac{\widetilde{K}-1}{1}}_{=1.64}\middle|\mu = 1\right) = 0.05.$$

Though in practice this is not necessary, if you really wanted to find the threshold K then solve

$$\widetilde{K} = 3^{-1}\log K + 3^{-1}\frac{1}{2}(4^2 - 1^2) = 1 + 1.64.$$

This gives

$$\log K = 3 \times (1 + 1.64) - 15/2 = -0.708 \qquad K = 0.49.$$

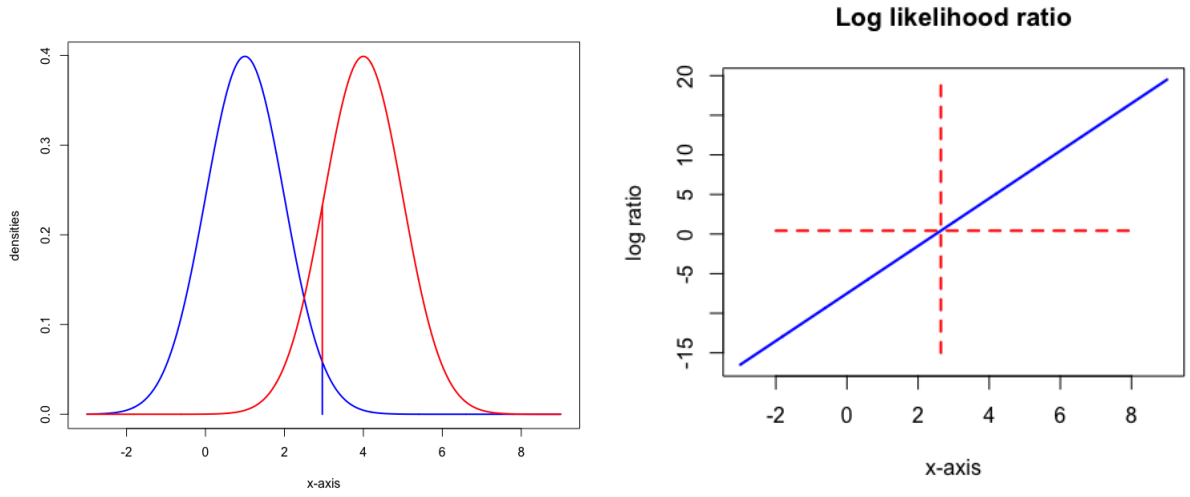


Figure 4.2: $H_0 : N(1,1)$ vs $H_A : N(4,1)$ with $n = 1$ log. The left hand plot is of the two densities under the null and alternative hypothesis. The red/blue line is the rejection region at the 5% level using the z-test. The Right hand plot is equation $\log LR(x)$ in (4.2). For normal observations the LRT(x) is linear. The red horizontal line corresponds the rejection value $\log K$ on the y-axis (it is not zero). If $\log LR(x)$ is greater than this value we reject the null (at the 5% level). This is for all x to the right of the vertical red line. Observe the log-likelihood ratio is monotonically increasing with x.

$\log K = -0.708$ corresponds to the red horizonal dotted line on the right hand plot in Figure 4.2 (though it is not so clear). Thus we have the rejection region

$$C = \left\{x; \frac{f(x;\mu_1 = 4)}{f(x;\mu_0 = 1)} \geq K\right\} = \left\{\frac{X-1}{1} \geq 1.64\right\} = \{X \geq 2.64\},$$

where $P(C|\mu = 1)P(X \geq 2.64|\mu = 1) = 0.05$. The interesting observation is that this rejection region is the same as the one constructed using the z-test (see equation (4.1). Further, more for any alternative $H_A : \mu = \mu_1$ (where $\mu_1 > 1$) we have that

$$\log LR = \log \frac{f(x;\mu_1)}{f(x;\mu_0 = 1)} = -\frac{1}{2}(x-1)^2 + \frac{1}{2}(x-\mu_1)^2 = x(\mu_1 - 1) - \frac{1}{2}(\mu_1^2 - 1^2).$$

Thus we observe that $\log LR$ is a monotonic function in x, and following the same calculations as above if we set $\alpha = 0.05$ the rejection region is

$$C = \left\{x; \frac{f(x;\mu_1)}{f(x;\mu_0 = 1)} \geq K = 0.49\right\} = \{x \geq 1 + 1.64\}$$

where $P(C|\mu = 1) = 0.05$. Thus the rejection region is (a) the same as in the z-test (see equation (4.1)) and (b) it is the same for any $\mu_1 > 1$ (we later learn that this means for the normal distribution, this test is

uniformly most powerful for any alternative $\mu_1 > 1$). The above construction is for $n = 1$, but the argument is easy to generalise to any n.

General sample size n. The likelihood is the joint density

Suppose $\underline{x} = (x_1, \ldots, x_n)$, it is straightforward to show that for n iid random variables we have

$$\begin{aligned} \log LR(\underline{x}) = \log \frac{f(\underline{x};\mu_1)}{f(\underline{x};\mu_0=1)} &= \mathcal{L}_n(\mu_1;\underline{x}) - \mathcal{L}_n(\mu_0=1;\underline{x}) \\ &= -\frac{n}{2}(\bar{x}-1)^2 + \frac{n}{2}(\bar{x}-\mu_1)^2 = n\bar{x}(\mu_1-1) - \frac{n}{2}(\mu_1^2-1^2). \end{aligned}$$

Using the same argument as in the case $n = 1$, this corresponds to the rejection region

$$\begin{aligned} C &= \left\{\underline{x}; \frac{f(\underline{x};\mu_1)}{f(\underline{x};\mu_0=1)} \geq K\right\} = \left\{\underline{x}; \frac{L_n(\mu_1;\underline{x})}{L_n(\mu_0=1;\underline{x})} \geq K\right\} \\ &= \{\bar{x} \geq 1 + 1.64n^{-1/2}\} \end{aligned}$$

where $P(C|\mu = 1) = 0.05$. We observe that the value of μ_1 plays no role in the construction of this region.

Power Calculation The power of the test is identical to the power in the z-test (since both the LR-test and the z-test are identical). As a reminder the Power is calculated as

$$\begin{aligned} P\left(\bar{X} \geq 1 + 1.64n^{-1/2}|H_A\right) &= P\left(\bar{X} \geq 1 + 1.64n^{-1/2}|\bar{X} \sim N(\mu_1, 1/n)\right) \\ &= P\left(\frac{\bar{X}-\mu_1}{n^{-1/2}} \geq \frac{1-\mu_1+1.64n^{-1/2}}{n^{-1/2}}|H_A\right) \\ &= P\left(Z \geq \frac{1-\mu_1+1.64n^{-1/2}}{n^{-1/2}}\right), \end{aligned}$$

where $Z \sim N(0,1)$.

Composite hypothesis (one-sided test): Normal distribution (known variance)

We observe that for the normal distribution the rejection region is the same for all $\mu_1 > \mu_0$ (it does not depend on μ_1).

Thus, suppose we observe n iid normal random variables with mean μ and variance σ^2. The hypothesis $H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0$ leads to the the rejection region

$$C = \left\{\bar{x} \geq \mu_0 + z_\alpha \frac{\sigma}{n^{1/2}}\right\}$$

where $P(C|\mu_0) = P(\bar{X} \geq \mu_0 + z_\alpha \frac{\sigma}{n^{1/2}}|\mu_0) = \alpha$. Comparing C, with the rejection region in the z-test, we observe that the likelihood ratio test and the z-test are equivalent for the one-sided tests (and the variance is known) of iid normal random variables.

Composite hypothesis (one-sided test): Normal distribution (unknown variance)

It can be shown that when the variance is unknown the log-likelihood ratio test and the t-test are equivalent for the one-sided tests for iid normal random variables (see Cox and Hinkley (1974), Example 5.5, page 142). This is not covered in the syllabus.

4.2.3 Example: The Binomial distribution

Simple hypothesis

We consider the LR test for the binomial distribution where $X = Bin(m = 10, p)$ (where we simply observe one $X \sim Bin(m = 10, p)$). The hypothesis is $H_0 : p = p_0 = 0.3$ vs $H_A : p = p_1 = 0.6$. The log likelihood ratio is

$$\begin{aligned}
\log LR(X) &= \log \binom{10}{X} p_1^X (1-p_1)^{10-X} - \log \binom{10}{X} p_0^X (1-p_0)^{10-X} \\
&= X \log p_1 + (10 - X) \log(1 - p_1) - X \log p_0 - (10 - X) \log(1 - p_0) \\
&= X \left[\log \left(\frac{p_1}{1-p_1} \right) - \log \left(\frac{p_0}{1-p_0} \right) \right] + 10 \left(\log(1-p_1) - \log(1-p_0) \right). \qquad (4.3)
\end{aligned}$$

The table for the probabilities at the $\log LR$ is given below:

Outcome	0	1	2	3	4	5	6	7	8	9	10
$H_0 : p = 0.3$	0.0282	0.1211	0.2335	0.2668	0.2001	0.1029	0.0368	0.0090	0.0014	0.0001	0.0000
$H_A : p = 0.6$	0.0001	0.0016	0.0106	0.0425	0.1115	0.2007	0.2508	0.2150	0.1209	0.0403	0.0060
$\log LR$	-5.5962	-4.3434	-3.0906	-1.8379	-0.5851	0.6677	1.9204	3.1732	4.4259	5.6787	6.9315
$P(X \geq k \mid p = 0.3)$	1.0000	0.9718	0.8507	0.6172	0.3504	0.1503	0.0473	0.0106	0.0016	0.0001	0.0000
$P(X \geq k \mid p = 0.6)$	1.0000	0.9999	0.9983	0.9877	0.9452	0.8338	0.6331	0.3823	0.1673	0.0464	0.0060

Remark (Comparing the log LR(x) of normal and binomial). *The log $LR(X)$ for the binomial and normal look very different. But we observe that if $p_1 > p_0$ then the $\log LR(X)$ of the binomial can be written as*

$$\log LR(X) = \alpha X + \beta,$$

where α is positive. Thus as X increases, $\log LR(X)$ increases in X. Thus

$$\begin{aligned}
& P\left(\frac{f(X, p = 0.6)}{f(X, p = 0.3)} \geq K \mid X \sim Bin(10, p = 0.3) \right) = 0.05 \\
= \; & P\left(\log \frac{f(X, p = 0.6)}{f(X, p = 0.3)} \geq \log K \mid X \sim Bin(10, p = 0.3) \right) \\
= \; & P\left(\alpha X + \beta \geq \log K \mid X \sim Bin(10, p = 0.3) \right) \\
= \; & P\left(X \geq \underbrace{\frac{\log K - \beta}{\alpha}}_{\widetilde{K}} \Big| X \sim Bin(10, p = 0.3) \right) = 0.05.
\end{aligned}$$

If we set $\widetilde{K} = 6$, the we have

$$\begin{aligned} & P\left(\frac{f(X, p = 0.6)}{f(X, p = 0.3)} \geq K | X \sim Bin(10, p = 0.3)\right) = 0.05 \\ = \quad & P(X = 6) + P(X = 7) + \ldots + P(X = 10) = 0.0473. \end{aligned}$$

We write this formally below.

We observe that the log $LR(X)$ increases with X, which is what we would expect. A plot of both distributions (under null and alternative) and log $LR(X)$ against X is given in Figure 4.3.

As x gets larger $H_0 : p = 0.3$ becomes less likely and $H_A : p = 0.6$ becomes more likely. The likelihood ratio test rejects the null when the ratio exceeds the threshold K, where

$$P\left(\frac{f(X, p = 0.6)}{f(X, p = 0.3)} \geq K | X \sim Bin(10, p = 0.3)\right) = 0.05.$$

From the table we observe the probabilities are discrete, so finding an x, where we have exactly 0.05 is not possible. Eric Lehmann found ways to deal with this case. But instead we simply look for the nearest approximation. Since $LR(X)$ increases with X, we just select the smallest X, where we get a probability under the null less than 0.05. From the table above we observe that

$$P(X \geq 6 | X \sim Bin(10, p = 0.3)) = 0.0473.$$

Thus we have the rejection region:

$$C = \left\{x; \frac{f(x, p = 0.6)}{f(x, p = 0.3)} \geq \exp(1.9204)\right\} = \{x \geq 6\}$$

where $P(C|p = 0.3) = 0.0473$.

Power Calculation, under alternative $p = 0.6$ The power calculation is the probability X lies in the rejection region under the alternative

$$\begin{aligned} P(X \in C | p = 0.6) \quad &= \quad P(X = \{6, 7, \ldots, 10\} | p = 0.6) = \sum_{k=6}^{10} P(X = k | p = 0.6) \\ &= \quad 0.6331, \end{aligned}$$

which can be read from the above table and is the sum of the red heights for the Binomial in Figure 4.3.

The composite hypothesis

To motivate where we are going in the case of composite hypothesis, we start by considering a different alternative hypothesis (but same null as in the previous example). We test $H_0 : p = p_0 = 0.3$ vs $H_A : p = p_1 = 0.8$. The likelihood ratio table is given below.

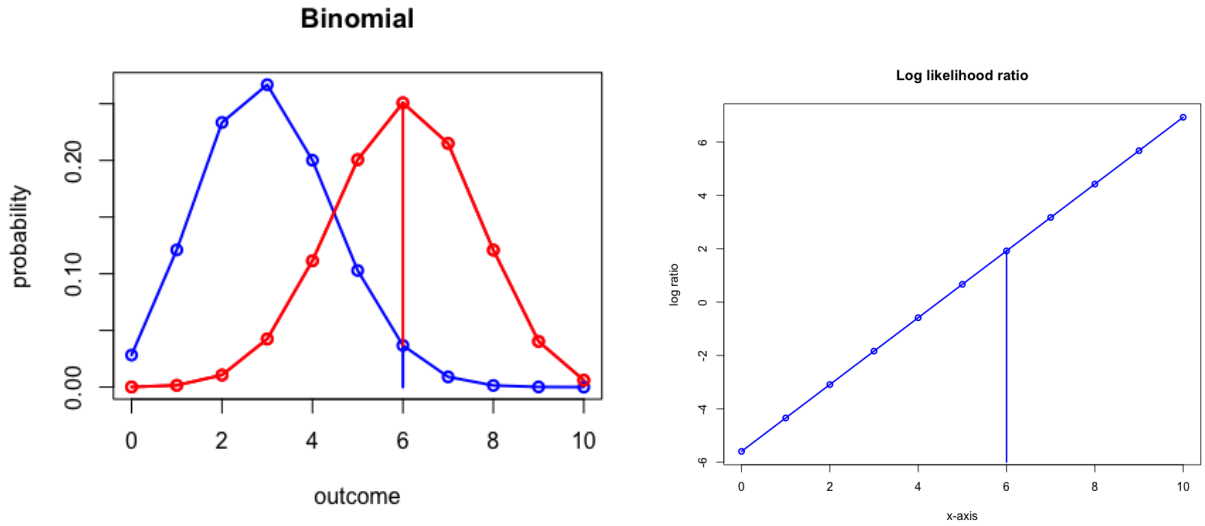


Figure 4.3: $H_0 : p = 0.3$ vs $H_A : p = 0.6$. Left hand side: Plot of both probability mass functions under the null and the alternative hypothesis. Right hand side: Plot of $\log LR(x)$ against x given in equation (4.3). Probability of a realisation x being on the right of the vertical blue line is under 5% under the null. Observe that the log-likelihood ratio $\log LR(x)$ is increasing with x.

Outcome	0	1	2	3	4	5	6	7	8	9	10
$H_0 : p = 0.3$	0.0282	0.1211	0.2335	0.2668	0.2001	0.1029	0.0368	0.0090	0.0014	0.0001	0.0000
$H_A : p = 0.8$	0.0000	0.0000	0.0001	0.0008	0.0055	0.0264	0.0881	0.2013	0.3020	0.2684	0.1074
$\log LR$	-12.5276	-10.2940	-8.0604	-5.8269	-3.5933	-1.3597	0.8739	3.1075	5.3411	7.5747	9.8083
$P(X \geq k \vert p = 0.3)$	1.0000	0.9718	0.8507	0.6172	0.3504	0.1503	0.0473	0.0106	0.0016	0.0001	0.0000
$P(X \geq k \vert p = 0.8)$	1.0000	1.0000	1.0000	0.9999	0.9991	0.9936	0.9672	0.8791	0.6778	0.3758	0.1074

Analogous to the alternative $H_A = 0.6$, we observe $\log LR(x)$ grows with x (it is a monotonic function in x). For this reason the rejection region is the same as in the previous case where $H_A : p = 0.6$. Returning to the new alternative $H_A : p = 0.8$ we have

$$C = \left\{ x; \frac{f(x, p = 0.8)}{f(x, p = 0.3)} \geq \exp(0.8739) \right\} = \{S_{10} \geq 6\} \tag{4.4}$$

where $P(C|p = 0.3) = 0.0473$.

Power Calculation, under alternative $p = 0.8$ The power calculation is the probability X lies in the rejection region under the alternative

$$\begin{aligned} P(X \in C|p = 0.8) &= P(X = \{6, 7, \ldots, 10\}|p = 0.8) = \sum_{k=6}^{10} P(X = k|p = 0.8) \\ &= 0.9672, \end{aligned}$$

which can be read from the above table. Observe that the power for the test $H_0 : p = 0.3$ vs $H_A : p = 0.8$ is

0.9672, which is quite a lot greater than the power for the test $H_0 : p = 0.3$ vs $H_A : p = 0.6$ (the power is 0.6331) derived in the previous section.

We deduce from these two examples and the general expression (4.3), that for any alternative $p > 0.3$ the rejection region is the same C. Thus the rejection region of the likelihood ratio test corresponding to the one-sided hypothesis $H_0 : p = 0.3$ vs $H_A : p > 0.3$ is the same C defined in (4.4)[2].

Recall from elementary statistics, when testing $H_0 : p = p_0$ against $H_A : p > p_0$, to obtain the rejection region one used the binomial distribution under the null (see the plot in Figure 4.4). This is exactly what was done above. Thus the likelihood ratio is equivalent to the binomial one-sided test. In Section 4.3 we show that there is no other test that will give better power at the same level. This gives a strong justification for using these testing methods.

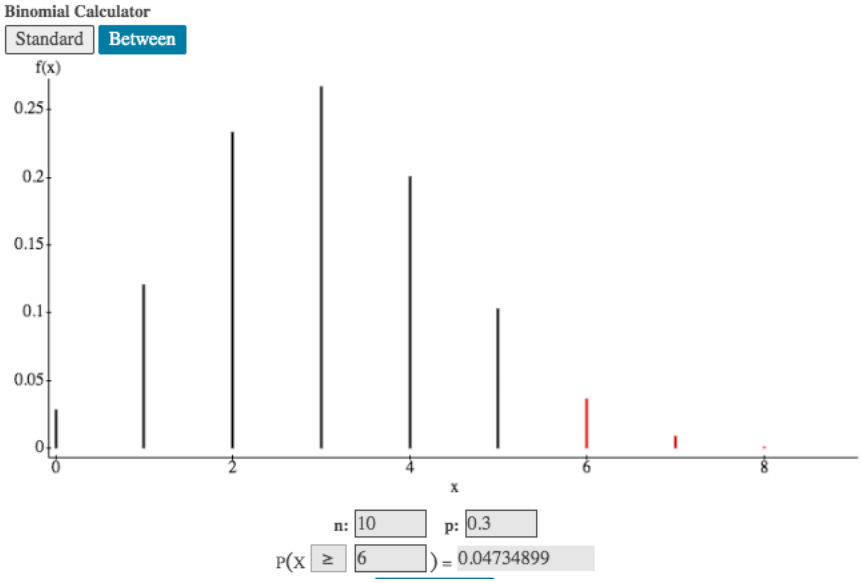


Figure 4.4: The red region is the rejection region for $H_0 : p = 0.3$ vs $H_A : p > 0.3$.

Review of normal and binomial examples We observe that for both the binomial and normal distribution the rejection region for the LRT with the composite the hypothesis is the same as that in for simple hypothesis tests.

Remark (One sided tests "pointing" right and left)**.** *In all the above examples, the alternative was "pointing" to the right of the null $H_A : \mu > \mu_0$ or $H_A : p > p_0$. More precisely, $\mu_1 - \mu_0 > 0$ (for the mean in the normal) or $p_1 - p_0 > 0$ (for the probability in the Binomial). This immediately gave a rejection region which was on the right hand tail of the distribution. In a homework, you will consider similar examples, but where the alternative is such that $H_A : \mu < \mu_0$ or $H_A : p < p_0$ (i.e. $\mu_1 - \mu_0 < 0$ or $p_1 - p_0 < 0$). This will lead to a change in the*

[2]Note that $P(C|H_0) = 0.0473$, thus the test is not conducted at precisely the 5% level. The boundary of the rejection region is between 5 and 6. To overcome the boundary issue Lehmann (1958) proposes using a random boundary of 5 and 6.

direction of the rejection region. The rejection region will be on the left hand tail of the distribution.

You would have learnt this in introductory statistics classes. The reason these tests are used (and the obsession as to whether the rejection region lies) is because they are log-likelihood ratio tests. We will show in Section 4.3 have maximum power and cannot be bettered.

4.2.4 Exponential family (with one parameter)

This section will not be tested. An interesting feature that binds the normal and the binomial simple hypothesis test is that for any $\theta > \theta_0$ (single parameter under any alternative greater than the null) the likelihood ratio $LR = f(x:\theta)/f(x:\theta_0)$ is a monotonically non-decreasing function in x. This gives rise to a rejection region which is one connected interval that does not depend on the alternative value θ (besides its direction with respect to θ_0). This result is does not hold true for all distributions. But it does hold for the single parameter distributions in the exponential family (with a natural parameterisation). To see why, suppose X_i are iid random variables, with density $f(x,\theta)$, which can be written as

$$\begin{aligned} f(x;\theta) &= \exp\left[s(x)\theta + b(\theta) + c(x)\right] \qquad x \in A, \\ \Rightarrow \log f(x;\theta) &= s(x)\theta + b(\theta) + c(x), \end{aligned}$$

where θ is a univariate parameter. Observe that both the normal distribution (with known variance) and binomial distribution have this representation.

Suppose we test $H_0 : \theta = \theta_0$ vs $H_A : \theta = \theta_1$ (where $\theta_1 > \theta_0$), then the log-likelihood ratio is

$$\log LR(\underline{x}) = (\theta_1 - \theta_0)\sum_{i=1}^{n} s(x_i) + b(\theta_1) - b(\theta_0). \tag{4.5}$$

Example 4.2. *We demonstrate that the normal and binomial can be written as (4.5).*

(i) Normal observations:

$$\log LR(\underline{x}) = \log\frac{f(\underline{x};\mu_1)}{f(\underline{x};\mu_0)} = \sum_{i=1}^{n}\left[-\frac{1}{2\sigma^2}(x_i-\mu_1)^2 + \frac{1}{2\sigma^2}(x_i-\mu_0)^2\right] = \frac{\sum_{i=1}^{n} x_i}{\sigma^2}(\mu_1-\mu_0) - \frac{1}{2\sigma^2}(\mu_1^2-\mu_0^2),$$

(ii) Binomial observations:

$$\log LR(\underline{x}) = \log\frac{p(\underline{x},p_1,m)}{p(\underline{x},p_0,m)} = \sum_{i=1}^{n} x_i\left[\log\left(\frac{p_0}{1-p_0}\right) - \log\left(\frac{p_1}{1-p_1}\right)\right] + nm\left(\log(1-p_0) - \log(1-p_1)\right).$$

Thus to determine the rejection region we use that

$$\begin{aligned} P(LR(\underline{x}) \geq K) &= P\left((\theta_1-\theta_0)\sum_{i=1}^{n} s(x_i) + b(\theta_1) - b(\theta_0) \geq K | H_0\right) \\ &= P\left((\theta_1-\theta_0)\sum_{i=1}^{n} s(x_i) \geq \widetilde{K} | H_0\right). \end{aligned}$$

This corresponds to a region $C = \{y; y = \sum_{i=1}^{n} s(x_i) \geq k\}$ where k is such that $P(\sum_{i=1}^{n} s(X_i) \geq k|H_0) = \alpha$.

Observe once again that this rejection region does not depend on θ_1. Thus in general for the test $H_0 : \theta = \theta_0$ vs $H_A : \theta > \theta_0$ the rejection region is

$$C = \left\{ y; y = \sum_{i=1}^{n} s(x_i) \geq k \right\} \qquad P\left(\sum_{i=1}^{n} s(X_i) \geq k|H_0 \right) = \alpha.$$

4.2.5 Non-monotonic likelihood ratios

In the examples above, the likelihood ratio is monotonic in x, this gives rise to a rejection region which is one continuous interval. This is not always the case. The purpose of this example is to show that for many distributions the rejection region using the LR-test can be very difficult to analytically calculate.

Example 1

Suppose the random variable X is a discrete random variables with $X \sim \{0, 1, 2, 3, 4, 5\}$ there are two competing distributions for X:

k	0	1	2	3	4	5
P_0	$P_0(X = 0) = 0.1$	$P_0(X = 1) = 0.2$	$P_0(X = 2) = 0.3$	$P_0(X = 3) = 0.1$	$P_0(X = 4) = 0.1$	$P_0(X = 5) = 0.2$
P_1	$P_1(X = 0) = 0.3$	$P_1(X = 1) = 0.1$	$P_1(X = 2) = 0.1$	$P_1(X = 3) = 0.2$	$P_1(X = 4) = 0.1$	$P_1(X = 5) = 0.2$

Aim Using just one random variable test $H_0 : X \sim P_0$ vs $H_A : X \sim P_1$. We do the test at the 20% level using the likelihood ratio test. The derivation is given below:

k	0	1	2	3	4	5
P_0	$P_0(X = 0) = 0.1$	$P_0(X = 1) = 0.2$	$P_0(X = 2) = 0.3$	$P_0(X = 3) = 0.1$	$P_0(X = 4) = 0.1$	$P_0(X = 5) = 0.2$
P_1	$P_1(X = 0) = 0.3$	$P_1(X = 1) = 0.1$	$P_1(X = 2) = 0.1$	$P_1(X = 3) = 0.2$	$P_1(X = 4) = 0.1$	$P_1(X = 5) = 0.2$
$\frac{P_1}{P_0}$	$\frac{0.3}{0.1} = 3$	$\frac{0.1}{0.2} = 0.5$	$\frac{0.1}{0.3} = 0.333$	$\frac{0.2}{0.1} = 2$	$\frac{0.1}{0.1} = 1$	$\frac{0.2}{0.2} = 1$

Using the above we observe

$$\begin{aligned}
P\left(\frac{P_1(X)}{P_0(X)} \geq 0.33|H_0\right) &= 1 \\
P\left(\frac{P_1(X)}{P_0(X)} \geq 0.5|H_0\right) &= 1 - P(X=2|H_0) = 1 - 0.3 = 0.7 \\
P\left(\frac{P_1(X)}{P_0(X)} \geq 1|H_0\right) &= 1 - P(X=2|H_0) - P(X=1|H_0) = 1 - 0.3 - 0.2 = 0.5 \\
P\left(\frac{P_1(X)}{P_0(X)} \geq 2|H_0\right) &= P(X=0|H_0) + P(X=3|H_0) = 1 - 0.3 - 0.2 - 0.1 - 0.2 = 0.2 \\
P\left(\frac{P_1(X)}{P_0(X)} \geq 3|H_0\right) &= P(X=3|H_0) = 1 - 0.3 - 0.2 - 0.1 - 0.2 - 0.1 = 0.1
\end{aligned}$$

Using the above we can obtain the threshold, since

$$P\left(\frac{P_1(X)}{P_0(X)} \geq 2|H_0\right) = P(X=2|H_0) + P(X=3|H_0) = 0.2,$$

the rejection region is any X where $\frac{P_1(X)}{P_0(X)} \geq 2$. We see this corresponds to the set $R = \{0, 3\}$. Thus if our observed data x is in the set $R = \{0, 3\}$ we reject the null at the 20% level.

The power of the test This is the probability X lies in $R = \{0, 3\}$ if the true distribution is f_1;

$$P(X \in \{0, 3\}|H_A) = P(X = 0|H_A) + P(X = 3|H_A) = 0.3 + 0.2 = 0.5.$$

Example 2

We start with a very simple example, that gives rise to a very complicated rejection region under the LR-test. Suppose that X is normally distributed (to keep things simple we stick with sample size $n = 1$). As the null hypothesis is $H_0 : X \sim N(\mu_0, \sigma_0^2)$ vs $H_A : X \sim N(\mu_1, \sigma_1^2)$ (where μ_0, μ_1, σ_0^2 and σ_1^2) are all known numbers. The difference between this example and the previous example is that we do not assume the variance is the same under the null and the alternative. Our aim is to construct the rejection region (which is always calculated under the null that $X \sim N(\mu_0, \sigma_0^2)$) and then to calculate the power (which is calculated under the alternative).

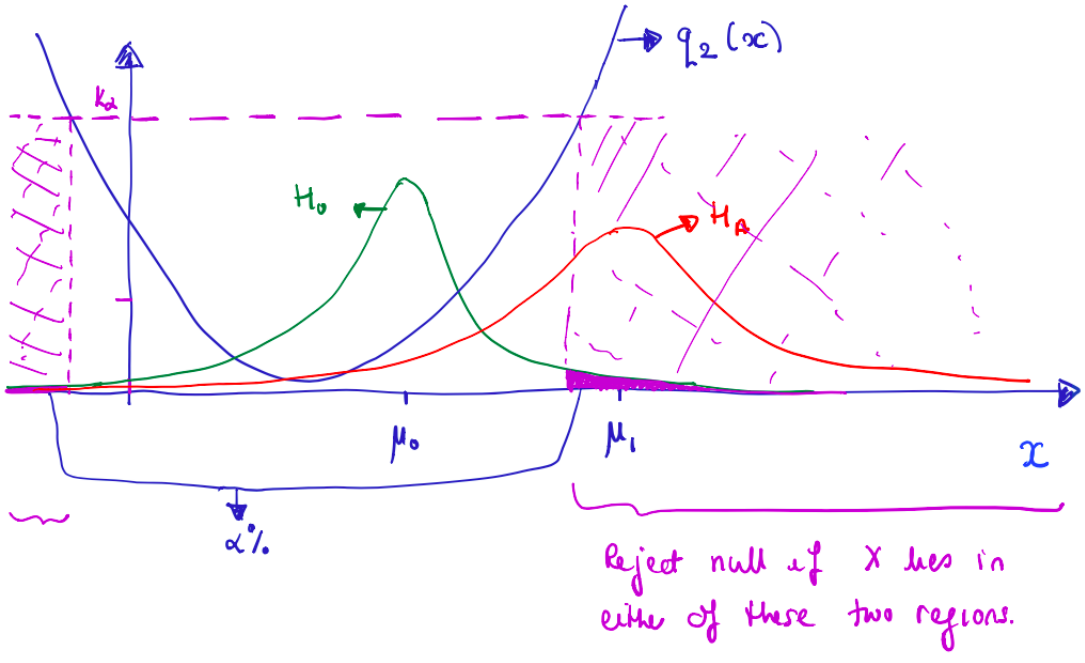
Calculating the rejection region for X at the $\alpha\%$ level. The log-likelihood ratio

$$\begin{aligned}\log\frac{f_0(X)}{f_1(X)} &= \log\frac{(2\pi\sigma_1^2)^{-1/2}\exp(-(2\sigma_1^2)^{-1}(X-\mu_1)^2)}{(2\pi\sigma_0^2)^{-1/2}\exp(-(2\sigma_0)^{-2}(X-\mu_0)^2)}\\ &= -\frac{1}{2\sigma_1^2}(X-\mu_1)^2+\frac{1}{2\sigma_0^2}(X-\mu_0)^2-\frac{1}{2}\log\sigma_0^2+\frac{1}{2}\log\sigma_1^2\\ &= \left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_1^2}\right)X^2-2X\left(\frac{\mu_1}{2\sigma_1^2}-\frac{\mu_0}{2\sigma_0^2}\right)+\left(\frac{\mu_0^2}{2\sigma_0^2}-\frac{\mu_1^2}{2\sigma_1^2}\right)-\frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2}\\ &= q_2(X,\mu_0,\mu_1,\sigma_0^2,\sigma_1^2)-\frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2}.\end{aligned}$$

To simply notation we set $q_2(X) = q_2(X,\mu_0,\mu_1,\sigma_0^2,\sigma_1^2)$, but keep in mind that the coefficients of this quadratic polynomial depend on the parameters. Thus using the above to obtain the rejection region we need to calculate

$$\begin{aligned}P\left(\log\frac{f_0(X)}{f_1(X)}\geq\log K\middle|H_0\right) &= P\left(q_2(X)-\frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2}\geq\log K\middle|H_0\right)\\ &= P\left(q_2(X,\mu_0,\mu_1,\sigma_0^2,\sigma_1^2)\geq\widetilde{K}\middle|H_0\right).\end{aligned}$$

where $\widetilde{K}=\frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2}+\log K$. The rejection region is difficult to construct. But we give an illustration below (where we assume that $\sigma_0^2\leq\sigma_1^2$ and $\mu_0<\mu_1$). A plot of the normal distribution under the null and the alternative is given, together with the quadratic polynomial.



The vertical dotted like corresponds to the threshold K_α. If $q_2(X)\geq K_\alpha$ then we reject the null at the α level. This corresponds to the purple checked area, which is the rejection region. The $\alpha\%$ level is the purple area below the green normal curve under the null. This example illustrates that with a small change of the

null and alternative, the rejection region starts to depend on the alternative distribution and can become highly complex.

Power Calculation The power of the test (not shown). Is the area of below under red curve in the checked purple area.

Example 3

Here we consider an even more complicated example. Consider the following null and alternative

$$\begin{aligned} H_0: \quad f_0(x) &= N(2.5, 1) \\ H_A: \quad f_1(x) &= \frac{1}{2}\left(N(1,1) + N(4,1)\right) \end{aligned}$$

A plot of both densities is given in Figure 4.5 (left hand plot). The log ratio

$$LR(x) = \frac{f_1(x)}{f_0(x)} = \frac{0.5\exp\left(-2^{-1}(x-1)^2\right) + 0.5\exp\left(-2^{-1}(x-4)^2\right)}{\exp\left(-2^{-1}(x-2.5)^2\right)}$$

is given in the right plot of Figure 4.5. Recall we reject the null for sufficiently large values of LR, but this happens over two disconnected regions. Observe that $\log LR(x)$ looks linear in x, though it is not. However, it is symmetric about $x = 2.5$. Suppose we only observe one random variable X, which comes from either

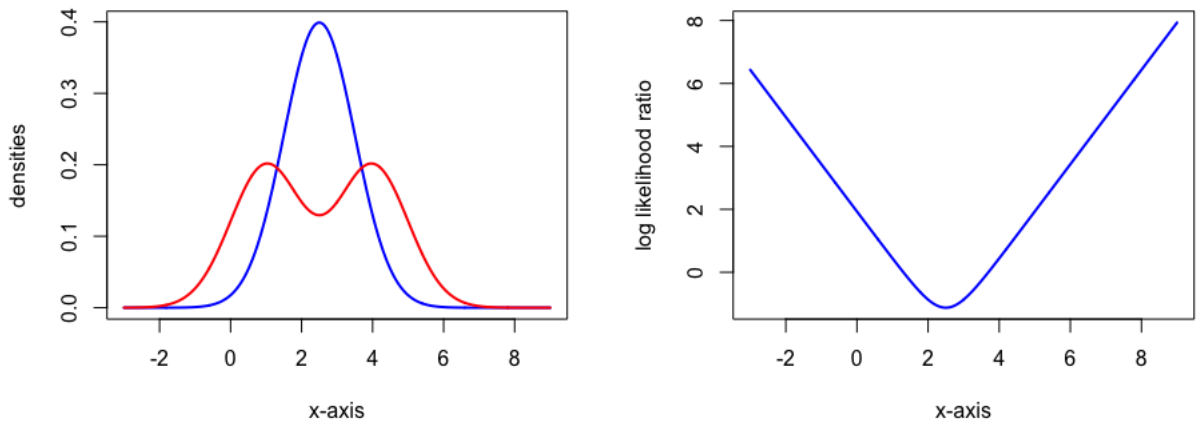


Figure 4.5: Left: Blue density (null), red density (alternative). Right: the non-monotonic likelihood ratio $LR(x)$.

f_0 or f_1. Even in this simple setting, calculation of the rejection region is not simple since

$$\begin{aligned} & P\left(\log LR(X) \geq \log K | X \sim f_0\right) \\ = \; & P\left(\log\left[0.5\exp\left(-2^{-1}(X-1)^2\right) + 0.5\exp\left(-2^{-1}(X-4)^2\right)\right] - \frac{1}{2}(X-2.5)^2 \geq \log K | X \sim N(2.5,1)\right). \end{aligned}$$

Observe that unlike the normal and binomial distribution considered in the previous section it is difficult to isolate X on its own. Further, the critical region of X will depend on the null and the alternative (unlike the binomial and normal). Since $\log LR(x)$ is symmetric about $x = 2.5$ the rejection region will be symmetric about 2.5. But I suspect that an analytic expression for the rejection region cannot be evaluate.

Power Calculation The power calculation is the probability X lies in the rejection region under the alternative. But gosh this is difficult to calculate!

4.3 The most powerful test: The Neyman-Pearson Lemma

In the examples above we have shown that the both the one-sided z-test, the one-sided binomial test are examples of the Likelihood Ratio test. There are several other examples of tests which can be written within the likelihood framework. Further, the generalized likelihood ratio (we describe this in Section 4.4, below) includes an even wider array of tests. During the late 1920s and early 1930s when Neyman and Pearson were developing the framework they realized that this could not be coincidence, that there must be a reason that the likelihood framework was so powerful. This lead to the Neyman-Pearson lemma, which we state below.

Before we state the result, we recall that every statistical test comes with a rejection region (they are

synonyms), where the probability the data lies in that region, under the null hypothesis is the Type I error. The Power of the test is the probability that the data lies in the rejection region under the alternative being true. Given a data set $\underline{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ (this can be $\{0, 1, \ldots, m\}^n$ in the case of n iid Binomial with m trials). There are many (sometimes infinite) ways of dividing $\mathbb{R}^n$ into a rejection region and non-rejection region. The Neyman-Pearson Lemma shows that the LR-test gives the "best" region.

Theorem 4.1

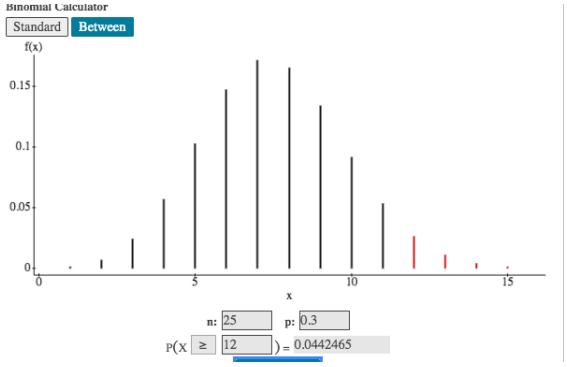
Suppose H_0 and H_A are two simple hypotheses. Consider the test that rejects the null whenever the likelihood ratio is greater than a threshold K, such that the probability this occurs under H_0 is α. Then any other test with significant level α or under, has power less than or equal to the likelihood ratio test.

The above result tells us that for two simple hypotheses, the most powerful test is the likelihood ratio test. No other test can do a better job of detecting the alternative (under the constraint the significance level is α or less). Therefore, for normal data (with known variance) the z-test is the best, and for binomial data the binomial test has the best power (for simple hypothesis).

Example 4.3. *Let us consider the example* $X \sim Bin(N = 25, p = 0.3)$. *We test the hypothesis*

$$H_0 : p = 0.3 \text{ vs } H_A : p > 0.3.$$

The distribution under the null is given in the plot below.



The LR-test at the 5% *level is the red region*

$$C_{0.05} = \{X = 12, 13, \ldots, 25\}.$$

Note that $P(C_{0.05}|p = 0.3) = 0.044$. *Observe that for* $p = 0.6$ *the power of this test is*

$$P(C_{0.05}|p = 0.6) = P(X \geq 12|p = 0.6) = \sum_{j=12}^{25} P(X = j|p = 0.6) = 0.922.$$

The power is 92.2% for this alternative. The Neyman-Pearson lemma states this rejection region (at the 5% level) has the greatest power for any alternative $p > 0.3$. To see check if this case, consider a different test with a different decision rule. Consider the rejection region $A_{0.05} = \{X = 4\}$. Now it is easily seen that

$$P(A_{0.05}|p = 0.3) = P(X = 4|p = 0.3) = 0.057,$$

which is relatively close to 5%. Suppose we rejected the null if $X = 4$, would this decision rule have much power for the alternative $p > 0.3$? Let is calculate the power with the alternative $p = 0.6$. That is calculating the chance of on observation being equal to 4 under the alternative $p = 0.6$:

$$P(A_{0.05}|p = 0.6) = P(X = 4|p = 0.6) = \binom{25}{4} 0.6^4 (1 - 0.6)^{25-4} = 0.00000721.$$

This means there is a 0.000721% of detecting the alternative under this decision rule. This is awful! This test has barely any power. Clearly, this test is much worse than the LR-test!

The Neyman-Pearson lemma concerns simple hypothesis. But in some situations it can be extended to composite hypothesis. We give a classical example below.

Example 4.4 (Normal distribution and one-parameter exponential family: One sided test)**.** *Recall that for the normal one-sided tests considered in the previous sections. The construction of the rejection region is completely specified by the null hypothesis (sample size and significance level) but it does not depend on the value of the alternative. Thus for every simple alternative the same test has maximum power. Thus because the test is the most powerful and is the same for every alternative, it is said to be uniformly most powerful.*

Indeed for any one-sided test in the natural (one-parameter) exponential class, the Likelihood Ratio test is uniformly most power (this result also holds for other distributions too) for the hypothesis $H_0 : \theta = \theta_0$ vs $H_A : \theta > \theta_0$

But we show below, that by a simple change of the test, and uniformity may no longer hold.

Example 4.5 (Normal distribution: Two sided test (uniformity does not hold))**.** *Suppose that we test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$. Consider the alternative $\mu_1 > \mu_0$, then the test with the most power for this alternative yields the rejection region*

$$C = \left\{ \bar{x}; \bar{x} \geq \mu_0 + z_\alpha \frac{\sigma}{n^{1/2}} \right\}$$

at the α level. However, such a test has no power against the alternative $\mu < \mu_0$. This means, if $\mu < \mu_0$ the chance that $\bar{x} \in C$ is extremely small (less than α%!), this is what we mean by no power.

Indeed if $\mu < \mu_0$ then the test with the best power has rejection region

$$C = \left\{ \bar{x}; \bar{x} \geq \mu_0 - z_\alpha \frac{\sigma}{n^{1/2}} \right\},$$

which results in the same problem, such a test has no power for $\mu > \mu_0$. Thus for the two sided test, there is no single test which has the highest power for all the possible alternatives.

Note that the two-sided z-test does not give optimal power. Since the rejection regions on both side of the null are narrower than the one-sided rejection region.

In summary, the Neyman-Pearson lemma and its extension to uniformly most power test, give credence to the likelihood ratio test.

The proof of the Neyman-Pearson lemma is extremely simple, but it does not appeal to my intution as to why it works. In the following subsection, I try to reason why the LRT is the most powerful test (this is not part of the syllabus).

4.3.1 My heuristic understanding of the LRT and the Neyman Pearson Lemma

Consider the two simple hypothesis $H_0 : \theta = \theta_0$ and $H_A : \theta = \theta_1$. Our aim is to show that the LRT is the most powerful test for this alternative. We recall that a test is a decision process where

$$\delta(\underline{x}) = \begin{cases} 1 \text{ (Reject Null)} & \underline{x} \in C \\ 0 \text{ (Do not reject Null)} & \underline{x} \notin C \end{cases}$$

We start with the conditions that are required. The region C should be such that

$$P\left(\delta(\underline{X})|\theta = \theta_0\right) = \int_{\mathbb{R}} \delta(\underline{x}) f(\underline{x};\theta_0) d\underline{x} = \int_C f(\underline{x};\theta_0) d\underline{x} = \alpha,$$

where α is the type I error (typically 5%). But the region C should lead to the maximum power, that is $P(C)$, where

$$P\left(\delta(\underline{X})|\theta = \theta_1\right) = \int_{\mathbb{R}} \delta(\underline{x}) f(\underline{x};\theta_1) d\underline{x} = \int_C f(\underline{x};\theta_1) d\underline{x} = P(C)$$

should be maximum. Our aim is to show that the LRT leads to the test with maximal power. In particular, find the region C which maximises power under the constraint that $P\left(\delta(\underline{X})|\theta = \theta_0\right) = \alpha$. To solve this we use the method of Calculus of Variations (Lagrange multipliers). We want to find the region C which maximises;

$$L(C,\lambda) = \int_C f(\underline{x};\theta_1) d\underline{x} - \lambda \left(\int_C f(\underline{x};\theta_0) d\underline{x} - \alpha \right).$$

The solution is found by differentiating the above wrt C and λ (differentiating wrt λ forces the constraint $f(\underline{x};\theta_0)d\underline{x} = \alpha$). Differentiating over C is a little tricky (for my brain). So I start with the case $n = 1$ i.e. $\underline{x} = x$. In this case C reduces to a sequence $\{C_{\ell,1}, C_{\ell,2}\}$ and we have

$$\begin{aligned} L(C,\lambda) &= \int_C f(x;\theta_1) dx - \lambda \left(\int_C f(x;\theta_0) dx - \alpha \right) \\ &= \sum_\ell \int_{C_{\ell,1}}^{C_{\ell,2}} f(x;\theta_1) dx - \lambda \left(\sum_\ell \int_{C_{\ell,1}}^{C_{\ell,2}} f(x;\theta_0) dx - \alpha \right). \end{aligned}$$

Differentiating wrt to $C_{\ell,1}, C_{\ell,2}$ and λ gives

$$\begin{aligned}
\frac{\partial}{\partial C_{\ell,2}} L(C,\lambda) &= \left[f(C_{\ell,2};\theta_1) - \lambda f(C_{\ell,2};\theta_0)\right] = 0\\
\frac{\partial}{\partial C_{\ell,1}} L(C,\lambda) &= -\left[f(C_{\ell,1};\theta_1) - \lambda f(C_{\ell,1};\theta_0)\right] = 0\\
\frac{\partial}{\partial \lambda} L(C,\lambda) &= \int_C f(\underline{x};\theta_0)d\underline{x} - \alpha = 0.
\end{aligned}$$

Solving the above, we observe that the boundary of the regions are defined by the points where

$$\frac{f(x;\theta_1)}{f(x;\theta_0)} = \lambda.$$

And within each region $[C_{\ell,1}, C_{\ell,2}]$, we have $f(c;\theta_1) - \lambda f(c;\theta_0) \geq 0$. To satisfy condition (4.6), we choose λ such that the set

$$C = \left\{x; \frac{f(x;\theta_1)}{f_0(x;\theta_0)} \geq \lambda\right\}$$

satisfies $P(C|\theta = \theta_0) = \alpha$. To summarize in the case $n = 1$, the log-likelihood ratio test is most optimal.

The proof is probably more illuminating for the case $n \geq 2$. In this case the optimising equation is

$$L(C,\lambda) = \int_C f(\underline{x};\theta_1)d\underline{x} - \lambda\left(\int_C f(\underline{x};\theta_0)d\underline{x} - \alpha\right).$$

Note that C is a region or multiple regions. Then, roughly speaking (using dodgy calculus), taking derivatives along C gives

$$\begin{aligned}
\frac{\partial}{\partial C} L(C,\lambda) &= \int_{\partial C}\left[f(\underline{x};\theta_1)d\underline{x} - \lambda f(\underline{x};\theta_0)d\underline{x}\right] = 0\\
\frac{\partial}{\partial \lambda} L(C,\lambda) &= \int_C f(\underline{x};\theta_0)d\underline{x} - \alpha = 0.
\end{aligned}$$

Note that ∂C is the boundary of the region, and (4.6) gives conditions on the boundary. It shows that the boundary should be defined by

$$\frac{f(\underline{x};\theta_1)}{f(\underline{x};\theta_0)} = \lambda.$$

In other words, $\partial C(\underline{x})$ defines the contour on $\mathbb{R}^n$ where the function $g(\underline{x}) = \frac{f(\underline{x};\theta_1)}{f(\underline{x};\theta_0)} = \lambda$. The derivative of function $f(\underline{x};\theta_1)d\underline{x} - \lambda f(\underline{x};\theta_0)$ along $\partial C(\underline{x})$ is zero (just like when we find parameters which maximise a function). Inside this region $f(\underline{x};\theta_1)d\underline{x} - \lambda f(\underline{x};\theta_0) > 0$. Thus we choose the λ such that the set

$$C = \left\{\underline{x}; \frac{f(\underline{x};\theta_1)}{f_0(\underline{x};\theta_0)} \geq \lambda\right\}$$

satisfies the condition $P(C|\theta = \theta_0) = \alpha$. Thus gives the most efficient (powerful) test for the simple hypothesis.

4.4 Generalized Likelihood Ratio Test

The likelihood ratio test is remarkable in the sense that it is the test with the greatest power for simple hyothesis. And in certain situations it also has greatest power for some composite hypothesis. However, in most composite hypothesis test the likelihood ratio test cannot be evaluated (since the parameters are unknown). In this section we discuss a generalisation of the likelihood ratio test, which may not have the optimality properties of the likelihood ratio test but nevertheless performs relatively well. We recall that the likelihood ratio test is based on evaluating the ratio or log ratio under the two hypothesis

$$\frac{f(\underline{x};\theta_1)}{f(\underline{x};\theta_0)} \quad \text{or} \quad \log f(\underline{x};\theta_1) - \log f(\underline{x};\theta_0),$$

where $f(\underline{x},\theta)$ is the joint density of $\underline{x}$. Alternatively (and equivalently) we can view the above as likelihoods and log-likelihoods

$$\frac{L_n(\theta_1;\underline{X})}{L_n(\theta_0;\underline{X})} \quad \text{or} \quad \mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_0).$$

If this quantity is sufficiently large, the null is deemed implausible. However, the above can only be evaluated for simple hypothesis (since the distribution under the null and alternative are fully specified). Let us consider two (possibly composite) hypothesis Ω_0 and Ω_1. For the generalized likelihood ratio test, one finds the parameter which maximises likelihood in both the parameter spaces and computes the subsequent likelihood. Namely, the generalized likelihood is

$$\Lambda(\underline{x}) = \frac{\sup_{\theta_1\in\Omega_1} f(\underline{x};\theta_1)}{\sup_{\theta_0\in\Omega_0} f(\underline{x};\theta_0)} = \frac{\sup_{\theta_1\in\Omega_1} \mathcal{L}_n(\theta_1)}{\sup_{\theta_0\in\Omega_0} \mathcal{L}_n(\theta_0)}.$$

Just as in the likelihood ratio test, we find the threshold K, such $P\left(\Lambda(\underline{x}) \geq K|H_0\right) = \alpha$. If $\Lambda(\underline{x}) \geq K$ we reject the null hypothesis at the α significance level.

4.4.1 Example: Normal data (variance known), two-sided test

Suppose $\{X_i\}$ are iid normal random variables with mean μ and variance σ^2 (we assume the variance σ^2 is known). We test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$. To prove the result we use the transformation

$$\underline{Y}_n = E_n\underline{X}_n = \begin{pmatrix} \underline{e}_1 \\ \underline{e}_2 \\ \vdots \\ \underline{e}_n \end{pmatrix} \underline{X}_n = \begin{pmatrix} \sqrt{n}\bar{X} \\ \langle \underline{e}_2, \underline{X}_n \rangle \\ \vdots \\ \langle \underline{e}_n, \underline{X}_n \rangle \end{pmatrix}.$$

where $\{\underline{e}_j\}_{j=1}^n$ are orthonormal vectors and $\underline{e}_1 = n^{-1/2}(1,1,1\ldots,1)$. Which was exactly the same transformation used the derivation of the MLE for the normal distribution. By using (3.6) we have

$$\begin{aligned}\mathcal{L}_n(\mu;\underline{X}_n) &= -\frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2} \\ &= -\frac{n}{2}\log\sigma^2 - \frac{n(\bar{X}_n-\mu)^2}{2\sigma^2} - \frac{\sum_{i=2}^n Y_i^2}{2\sigma^2} = \mathcal{L}_n(\mu;\underline{Y}_n).\end{aligned}$$

Thus simplicity of calculation we use the (log)-likelihood of the transformed data given in equation (3.5)

$$\mathcal{L}_n(\mu;\underline{Y}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{n}{2\sigma^2}(\bar{X}-\mu)^2 - \frac{1}{2\sigma^2}\sum_{i=2}^n Y_i^2$$

Under the null hypothesis the (log)-likelihood is completely specified:

$$\mathcal{L}_n(\mu_0;\underline{Y}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{n}{2\sigma^2}(\bar{X}-\mu_0)^2 - \frac{1}{2\sigma^2}\sum_{i=2}^n Y_i^2.$$

Whereas under the alternative we have

$$\begin{aligned}\max_{\mu\in R/\{\mu_0\}} \mathcal{L}_n(\mu;\underline{Y}) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{n}{2}(\bar{X}-\bar{X})^2 - \frac{1}{2\sigma^2}\sum_{i=2}^n Y_i^2 \\ &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=2}^n Y_i^2.\end{aligned}$$

The log-likelihood ratio Thus the (log) likelihood ratio is

$$\log\Lambda(\underline{x}) = \log\frac{\sup_{\theta_1\in\Omega_1} f(\underline{x};\theta_1)}{\sup_{\theta_0\in\Omega} f(\underline{x};\theta_0)} = \frac{n}{2\sigma^2}(\bar{X}-\mu_0)^2.$$

Multiplying by two gives

$$2\log\Lambda(\underline{X}) = n\left(\frac{\bar{X}-\mu_0}{\sigma}\right)^2.$$

Under the null hypothesis

$$2\log\Lambda(\underline{X}) = n\left(\frac{\bar{X}-\mu_0}{\sigma}\right)^2 = \left(\frac{(\bar{X}-\mu_0)}{\sigma/\sqrt{n}}\right)^2 \sim \chi_1^2.$$

From the above, we observe that generalized log-likelihood ratio test, at the α-level is the K such that

$$P\left(\log\Lambda(\underline{x}) \geq K|H_0\right) = P\left(n\left(\frac{\bar{X}-\mu_0}{\sigma}\right)^2 \geq K|H_0\right) = P(\chi_1^2 \geq K) = \alpha. \tag{4.6}$$

Thus we use the threshold $K = \chi_1^2(\alpha)$. Or equivalently we use that

$$\begin{aligned}P\left(\log\Lambda(\underline{x}) \geq K|H_0\right) &= P\left(n\left(\frac{\bar{X}-\mu_0}{\sigma}\right)^2 \geq K|H_0\right) \\ &= P\left(\frac{|\bar{X}-\mu_0|}{\sigma/\sqrt{n}} \geq K^{1/2}|Z\sim N(0,1)\right) = P(|Z|\geq z_{\alpha/2}) = \alpha.\end{aligned}$$

This corresponds to the usual regular region for the two-sided test (using the z-test);

$$C_- \cup C_+ \text{ where } C_- = \left(-\infty, \mu_0 - 1.96 \times \frac{\sigma}{\sqrt{n}}, \infty\right] \text{ and } C_+ = \left[\mu_0 + 1.96 \times \frac{\sigma}{\sqrt{n}}, \infty\right).$$

Thus the generalised likelihood ratio test and the two-sided z-test are equivalent (the same) testing procedures.

In statistical software if a two-sided z-test is conducted always the p-value quoted is the smallest area times two. The reason behind this is given in (4.6). The p-value is

$$\begin{aligned} & P\left(\underbrace{\log \Lambda(\underline{X})}_{\text{Random variable}} \geq \underbrace{n\left(\frac{\bar{x}-\mu_0}{\sigma}\right)^2}_{\text{observed data}} | H_0 \right) \\ = \; & P\left(\chi_1^2 \geq n\left(\frac{\bar{x}-\mu_0}{\sigma}\right)^2\right) = 2 \times P\left(Z \geq \left|\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}\right|\right). \end{aligned}$$

From above we observe that the p-value is the two times the smallest area in the normal distribution.

4.4.2 Example: Normal data (variance unknown), two-sided test

We test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ (but the variance is assumed unknown). We will show that the generalised likelihood ratio test when the variance is unknown is equivalent to the two-sided t-test.

Suppose that $\{X_i\}$ are iid normal random variables with mean μ and variance σ^2. Again for simplicity of calculation we use the likelihood of the transformed data and (3.6)

$$\begin{aligned} \log f(\underline{y}; \mu, \sigma^2) &= -\frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2} \\ &= -\frac{n}{2}\log\sigma^2 - \frac{n(\bar{x}_n-\mu)^2}{2\sigma^2} - \frac{\sum_{i=2}^n y_i^2}{2\sigma^2} = \log f(\underline{x}; \mu, \sigma^2) \end{aligned}$$

By using that $\sum_{i=2}^n y_i^2 = (n-1)s_n^2$ (recall Section 2.4.2) we have

$$\log f(\underline{y}; \mu, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{n}{\sigma^2}(\bar{x}-\mu)^2 - \underbrace{\frac{(n-1)}{\sigma^2}s_n^2}_{=\sigma^{-2}\sum_{i=2}^n y_i^2}.$$

Thus under the null hypothesis $\mu = \mu_0$ the MLE of σ^2 is

$$\widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu_0)^2 = \frac{1}{n}\left((n-1)s_n^2 + n(\bar{x}-\mu_0)^2\right),$$

where $s_n^2 = (n-1)^{-1}\sum_{i=1}^n (x_i-\bar{x})^2$. To see why the equivalence in (4.7) is true use linear algebra (you do not have to). This results in the maximum log-likelihood (under null)

$$\log \widehat{f}_0(\underline{x}; \mu_0, \widehat{\sigma}_0^2) = -\frac{n}{2}\log(\widehat{\sigma}_0^2) - \frac{n}{2}.$$

Under the alternative, the MLE of μ and σ^2 is

$$\widehat{\mu}_1 = \bar{x} \qquad \widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(n-1)}{n}s_n^2.$$

This results in the maximum the log-likelihood (under the alternative)

$$\log \widehat{f}_1(\underline{x}; \widehat{\mu}, \widehat{\sigma}_1^2) = -\frac{n}{2}\log(\widehat{\sigma}_1^2) - \frac{n}{2}$$

The log-likelihood ratio Thus the (log) likelihood ratio is

$$\begin{aligned}
\log \Lambda(\underline{x}) = \log \widehat{f}_1(\underline{x}; \widehat{\mu}, \widehat{\sigma}_1^2) - \log \widehat{f}_0(\underline{x}; \mu_0, \widehat{\sigma}_0^2) &= -\frac{n}{2}\log\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_0^2} = \frac{n}{2}\log\left(\frac{\frac{1}{n}((n-1)s_n^2 + n(\bar{x}-\mu_0)^2)}{\frac{n-1}{n}s_n^2}\right) \\
&= \frac{n}{2}\log\left(1 + \frac{(\bar{x}-\mu_0)^2)}{\frac{n-1}{n}s_n^2}\right) \\
&= \frac{n}{2}\log\left(1 + \frac{n(\bar{x}-\mu_0)^2)}{(n-1)s_n^2}\right)
\end{aligned}$$

Thus we want to find the region where

$$P\left(\log \Lambda(\underline{x}) \geq e^K | H_0\right) = P\left(\log \widehat{f}_1(\underline{X}; \widehat{\mu}, \widehat{\sigma}_1^2) - \log \widehat{f}_0(\underline{X}; \mu_0, \widehat{\sigma}_0^2) \geq K | H_0\right) = \alpha.$$

This is equivalent to finding a region where

$$\begin{aligned}
P\left(\log \Lambda(\underline{x}) \geq e^K | H_0\right) &= P\left(\log \widehat{f}_1(\underline{X}; \widehat{\mu}, \widehat{\sigma}_1^2) - \log \widehat{f}_0(\underline{X}; \mu_0, \widehat{\sigma}_0^2) \geq K | H_0\right) \\
&= P\left(1 + \frac{n(\bar{X}-\mu_0)^2}{(n-1)s_n^2} \geq e^{2K/n} | H_0\right) \\
&= P\left(\frac{n(\bar{X}-\mu_0)^2}{s_n^2} \geq (n-1)(e^{2K/n}-1) | H_0\right) = \alpha.
\end{aligned}$$

Setting $R = \frac{n-1}{n}(e^K - 1)$ we have

$$\begin{aligned}
& P\left(\log \widehat{f}_1(\underline{X}; \widehat{\mu}, \widehat{\sigma}_1^2) - \log \widehat{f}_0(\underline{X}; \mu_0, \widehat{\sigma}_0^2) \geq K | H_0\right) \\
= \; & P\left(\underbrace{\frac{n(\bar{X}-\mu_0)^2}{s_n^2}}_{=T_n^2} \geq \widetilde{K} | H_0\right) = \alpha.
\end{aligned}$$

Thus we choose R such that

$$P\left(T_n^2 \geq \widetilde{K}\right) = P\left(|T_n| \geq \widetilde{K}^{1/2}\right) = \alpha,$$

noting that T_n^2 follows a $F_{1,n-1} = t_{n-1}^2$-distribution (under the null). Since $T_n = n^{1/2}(\bar{X}-\mu_0)/s_n$, this is exactly the two-sided t-test. Thus the generalized likelihood ratio test and the two-sided t-test are equivalent tests.

In statistical software if a two-sided t-test is conducted always the p-value quoted is the smallest area times two. This is because the p-value is

$$P\left(\underbrace{\log \Lambda(\underline{X})}_{\text{Random variable}} \geq \underbrace{n\left(\frac{\bar{x}-\mu_0}{s_n}\right)^2}_{\text{observed data}} | H_0 \right)$$
$$= P\left(t_{n-1}^2 \geq n\left(\frac{\bar{x}-\mu_0}{s_n}\right)^2\right) = 2 \times P\left(t_{n-1} \geq \left|\frac{\sqrt{n}(\bar{x}-\mu_0)}{s_n}\right|\right).$$

From above we observe that the p-value is the two times the smallest area in the t-distribution.

4.4.3 Example: Binomial distribution

Suppose that $\{X_i\}_{i=1}^n$ are iid random variables which follow a Binomial distribution, i.e. $X_i \sim Bin(m, p)$.

Examples of such data could be that in each of $n = 30$ cities around the US, $m = 20$ people were sampled. In each sample of 20, they were asked whether they liked orange juice, yes or no. Let X_i be the number of people out of 20 in city i who liked orange juice. We assume that the preference of the orange juice is homogenuous over the entire of the US. This means there is no city where the preference is more than another city. Consequently we can treat $\{X_i\}$ as iid random variables (which are Binomial random variables, as long as the $m = 20$ is far smaller than the population size; we require this assumption because the sampling is done without replacement; see HW1, Q1).

Suppose we test $H_0 : p = p_0$ vs $H_A : p \neq p_0$. We now construct the generalized log-likelihood ratio test for this procedure. The log-likelihood is

$$\begin{aligned}\mathcal{L}_n(p; \underline{X}) &= \sum_{i=1}^n \log \binom{m}{X} p^{X_i}(1-p)^{m-X_i} \\ &= \sum_{i=1}^n \left[\log \binom{m}{X} + X_i \log p + (m - X_i)\log(1-p)\right].\end{aligned}$$

Observe that $Y = \sum_{i=1}^n X_i \sim Bin(nm, p)$, so we could apply the same approach to Y, but for now we ignore this (they are both equivalent). Under the null hypothesis the log-likelihood is

$$\mathcal{L}_n(p_0; \underline{X}) = \sum_{i=1}^n \left[\log \binom{m}{X} + X_i \log p_0 + (m - X_i)\log(1-p_0)\right].$$

Under the alternative hypothesis we need to deduce the maximum likelihood estimator of p and plug it into the log-likelihood. The MLE (deduced by differentiating with respect to p) is

$$\frac{d\mathcal{L}_n(p; \underline{X})}{dp} = \frac{1}{p}\sum_{i=1}^n X_i - \frac{1}{(1-p)}\sum_i (m - X_i) = 0.$$

Basic algebra gives

$$\widehat{p} = \frac{\bar{X}}{m} = \frac{\sum_{i=1}^{n} X_i}{nm}.$$

Thus the maximum of the likelihood under the alternative is

$$\mathcal{L}_n(\widehat{p}; \underline{X}) = \sum_{i=1}^{n} \left[\log \binom{m}{X} + X_i \log \widehat{p} + (m - X_i) \log(1 - \widehat{p}) \right].$$

The log-likelihood ratio Thus the (log) generalized likelihood ratio for the binomial is

$$\begin{aligned} \log \Lambda(\underline{X}) &= \sum_{i=1}^{n} \left[X_i \log \widehat{p} + (m - X_i) \log(1 - \widehat{p}) \right] \\ &\quad - \sum_{i=1}^{n} \left[X_i \log p_0 + (m - X_i) \log(1 - p_0) \right]. \end{aligned}$$

This can be rewritten as

$$\begin{aligned} \log \Lambda(\underline{X}) &= \sum_{i=1}^{n} \left[X_i \log \frac{\widehat{p}}{p_0} + (m - X_i) \log \frac{(1 - \widehat{p})}{1 - p_0} \right] \\ &= mn(1 - \widehat{p}) \log \left(\frac{1 - \widehat{p}}{1 - p_0} \right) + mn\widehat{p} \log \frac{\widehat{p}}{p_0}, \end{aligned} \tag{4.7}$$

which you may observe is the generalized likelihood ratio for the Binomial(nm, p)!

Note, that unlike the normal example considered above, it is unclear what the distribution of $\log \Lambda(\underline{X})$ is under the null hypothesis. This makes construction of the rejection region difficult. An approximation of the distribution of $\log \Lambda(\underline{X})$ is given in the following section.

4.5 Asymptotic sampling properties of the generalized likelihood ratio test under the null hypothesis

The two examples we considered above concern the normal distribution. Thus the exact distribution of the generalized likelihood ratio test under the null hypothesis can be derived. But for the binomial example, the exact distribution of $\Lambda(\underline{x})$ under the null is not simple to derive. This makes obtaining the rejection regions for "determining" when $\Lambda(\underline{x})$ is too large impossible. For general distributions obtaining the finite sample distribution of the test statisic is not possible (though bootstrap methods do exist for obtaining approximations). Instead we resort to asymptotic sampling properties. This means finding the approximate distribution of $\Lambda(\underline{x})$. This approximation is close/good when the sample size is large, but may not be so good when the sample size is small.

To simplify the discussion we consider the simplest case that the null hypothesis, $\theta = \theta_0$ is simple (the finite sampling distribution is completely described by the null hypothesis), but the alternative hypothesis is composite, $\theta \in \Omega$. Further, the null hypothesis is nested in the sense that $\theta_0 \in \Omega$. Note that this seems different to the previous cases where $H_0 : \theta = \theta_0$ and $H_A : \theta \neq \theta_0$ (hence the two hypothesis are disjoint), but they are almost the same since if we complete the set $(-\infty, \theta_0) \cap (\theta_0, \infty)$ we obtain $\mathbb{R}$, which contains θ_0. We consider the null and alternative $H_0 : \theta = \theta_0$ vs $H_A : \theta \in \Theta$ and generalized likelihood ratio test

$$\begin{aligned} \Lambda(\underline{X}) &= \frac{\max_{\theta \in \Omega} f(\underline{x}; \theta)}{f(\underline{x}, \theta_0)} = \frac{L_n(\widehat{\theta}_n)}{L_n(\theta_0)} \\ \text{and } \log \Lambda(\underline{X}) &= \mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0). \end{aligned}$$

where $\widehat{\theta}_n = \arg\max_{\theta \in \Omega} L_n(\theta)$ (see Section 3.4). If it can be shown that the the maximum likelihood estimator (see Section 3.5) is asymptotic normal such that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}} N(0, [I(\theta)]^{-1}), \tag{4.8}$$

where $I(\theta)$ is the Fisher information matrix of the univariate density $f(x, \theta)$. If this holds, we obtain the following result.

Theorem 4.2

Under certain "regularity" conditions we can show that under the null hypothesis ($\theta = \theta_0$) we have

$$2 \log \Lambda(\underline{X}) \xrightarrow{\mathcal{D}} \chi^2_d, \quad \text{as} \quad n \to \infty, \tag{4.9}$$

where d is the dimension of the parameter vector θ.

This and more general versions of the result is often referred to as Wilks' theorem. A heuristic proof of the result is by making a second order Taylor expansion of $\log f(\underline{x}, \theta)$ about $\log f(\underline{x}, \theta_0)$ (which is beyond the syllabus).

Remark. *This result is quite remarkable. It says that that under the null hypothesis, the asymptotic distribution of* $2\Lambda(\underline{x}) \xrightarrow{\mathcal{D}} \chi^2_d$ *does not depend in any way on the distribution under the null. It does not even include parameters which are unknown! Contrast this with the distribution of the MLE (4.8), whose variance is the Fisher information matrix (and depends on parameters we do not know).*

Example 4.6 (Comparing the distribution of the Generalized LR-test for the normal data with Theorem 4.2)**.** *Below we discuss how the exact distributions obtain in Section 4.4.1 and 4.4.2 relate to the asymptotic distribution in Theorem 4.2.*

(i) *Section 4.4.1: Normal data, with σ^2 known*

We showed that

$$2\log\Lambda(\underline{x}) = \left(\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}\right)^2.$$

Under the null hypothesis $2\log\Delta(\underline{x}) \sim \chi^2_1$. *This exactly matches the asymptotic distribution result in Theorem 4.2 (no large sample size required!).*

(ii) *Section 4.4.2: Normal data, with* σ^2 *unknown*

We showed that

$$2\log\Lambda(\underline{X}) \quad = \quad n\log\left(1+\frac{n(\bar{x}-\mu_0)^2}{(n-1)s_n^2}\right)$$

Now it is not immediately clear how this links to the chi. But we can do some approximations. You may recall from your Calc I class, if x *is "small", then* $\log(1+x)\approx x$. *Applying this approximation to the above*[3] *we have*

$$2\log\Lambda(\underline{X}) \quad \approx \quad n\left[\frac{n(\bar{x}-\mu_0)^2)}{(n-1)s_n^2}\right].$$

Note that $n/(n-1)\approx 1$ *for large* n. *Using this gives*

$$2\log\Lambda(\underline{X}) \quad \approx \quad \left[\frac{n(\bar{x}-\mu_0)^2)}{s_n^2}\right] = \left[\frac{\sqrt{n}(\bar{x}-\mu_0))}{s_n}\right]^2,$$

which for large n *is close to a* χ^2_1*-distribution (since for large* n, $s_n\approx\sigma$*). As stated by Theorem 4.2.*

Remark (Power).

- *Unlike the likelihood ratio test. The generalized likelihood ratio test will usually not have optimal power for any given alternative.*

  For example, we know that for normal data with known variance that the one-sided z-test test has optimal power. However, using the generalized t-test for the two-sided hypothesis does not have optimal power (since the one-sided test does!). However, the generalized likelihood ratio test does have good power, which we discuss below.

- *We observe that both the two-sided t and z-test have statistical power (the ability to detect the alternative, with a high probability, when it is true). The same is true for the general generalized LR-test using the chi-square distribution approximation. It can be shown that if the alternative is true, then* $\log\Lambda(\underline{x})$ *tends to be very large (especially for large sample sizes). The reason for this goes back to Theorem 3.2, but the exact details are beyond this class.*

[3] $\frac{n(\bar{x}-\mu_0)^2)}{(n-1)s_n^2}$ will be quite "small" under the null, since $\frac{n(\bar{x}-\mu_0)^2)}{s_n^2}$ has a $t^2_{n-1} = F_{1,n-1}$-distribution.

A heuristic proof of Wilks' Theorem

We observe that $2\log\Lambda(\underline{X})$ asymptotically follows a χ^2-distribution. Which may seem a little surprising, as up to now most of the results we have considered are asymptotically normal. But we recall that

$$\begin{aligned}\log\Lambda(\underline{X}) &= \log f(\underline{X};\widehat{\theta}_n) - \log f(\underline{X},\theta_0)\\ &= \mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0),\end{aligned}$$

where $\mathcal{L}_n(\cdot)$ is the log-likelihood. Since $\widehat{\theta}_n$ is the MLE, it maximises the likelihood. Thus for any other $\theta \in \Theta$ we have $\mathcal{L}_n(\widehat{\theta}_n) \geq \mathcal{L}_n(\theta)$. Since $\log\Lambda(\underline{x})$ is non-negative, the normal distribution is an unlikely candidate. Next, we recall that Wilks' Theorem states that $2\log\Lambda(\underline{X})$ is asymptotic a chi-square. This suggests that $2\log\Lambda(\underline{X})$ can be written as the square of standard normal random variables. Why this should be true, is not immediately obvious. However, the clue is by making a Taylor expansion (mean value theorem) of θ_0 about $\widehat{\theta}$. For simplicity we restrict ourselves to the case that θ is one-dimension (such as the exponential and Poisson distribution given in HW10). Then by making a second order Taylor expansion of $\mathcal{L}_n(\theta_0)$ about $\mathcal{L}_n(\widehat{\theta}_n)$ we have

$$\mathcal{L}_n(\theta_0) \approx \mathcal{L}_n(\widehat{\theta}_n) + (\theta_0 - \widehat{\theta}_n)\frac{d\mathcal{L}_n(\theta)}{d\theta}\rfloor_{\theta=\widehat{\theta}_n} + \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^2\frac{d^2\mathcal{L}_n(\theta)}{d\theta^2}\rfloor_{\theta=\widehat{\theta}_n}.$$

Recall how the likelihood is maximised, we differentiate $\mathcal{L}_n(\theta)$ wrt θ and use the θ which solves $\frac{d\mathcal{L}_n(\theta)}{d\theta} = 0$. This immediately implies that $\frac{d\mathcal{L}_n(\theta)}{d\theta}\rfloor_{\theta=\widehat{\theta}_n} = 0$. This insight reduces the above expansion to

$$\mathcal{L}_n(\theta_0) \approx \mathcal{L}_n(\widehat{\theta}_n) + \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^2\frac{d^2\mathcal{L}_n(\theta)}{d\theta^2}\rfloor_{\theta=\widehat{\theta}_n}.$$

Rearranging the above "equation" gives

$$\begin{aligned}2\left[\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0)\right] &\approx -(\widehat{\theta}_n - \theta_0)^2\frac{d^2\mathcal{L}_n(\theta)}{d\theta^2}\rfloor_{\theta=\widehat{\theta}_n}\\ \Rightarrow 2\log\Lambda(\underline{X}) &\approx -(\widehat{\theta}_n - \theta_0)^2\frac{d^2\mathcal{L}_n(\theta)}{d\theta^2}\rfloor_{\theta=\widehat{\theta}_n}.\end{aligned} \tag{4.10}$$

The beauty of the above approximation is that $2\log\Lambda(\underline{X})$ is the square of $\widehat{\theta}_n - \theta_0$, which we know from Theorem 3.2 is asymptotically normal. Which starts to give us some clues as to where the chi-square comes from. Next let us look at the second derivative. From the definition of the log-likelihood we have

$$\frac{d^2\mathcal{L}_n(\theta)}{d\theta^2}\rfloor_{\theta=\widehat{\theta}_n} = \sum_{i=1}^n \frac{d^2\log f(X_i,\theta)}{d\theta^2}\rfloor_{\theta=\widehat{\theta}_n} \approx \sum_{i=1}^n \frac{d^2\log f(X_i,\theta)}{d\theta^2}\rfloor_{\theta=\theta_0}.$$

The above replacement is because for large n, the MLE $\widehat{\theta}_n$ is "close" to the truth θ_0. Dividing by n to turn the second derivative of the log-likelihood into an average gives

$$\frac{1}{n}\frac{d^2\mathcal{L}_n(\theta)}{d\theta^2}\rfloor_{\theta=\widehat{\theta}_n} \approx \frac{1}{n}\sum_{i=1}^n \frac{d^2\log f(X_i,\theta)}{d\theta^2}\rfloor_{\theta=\theta_0} \approx \mathrm{E}\left(\frac{d^2\log f(X_i,\theta)}{d\theta^2}\rfloor_{\theta=\theta_0}\right) = -nI(\theta_0).$$

Thus for "large" n we have

$$\frac{d^2\mathcal{L}_n(\theta)}{d\theta^2}\rfloor_{\theta=\theta_0} \approx -nI(\theta_0) \tag{4.11}$$

Substituting (4.11) into (4.10) gives

$$2\log\Lambda(\underline{X}) \approx (\widehat{\theta}_n - \theta_0)^2 nI(\theta_0) = \left(\sqrt{nI(\theta)}(\widehat{\theta}_n - \theta_0)\right)^2.$$

But from Theorem 3.2 we know that for large n that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}} N(0, [I(\theta)]^{-1}),$$

This implies that

$$\sqrt{nI(\theta)}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}} N(0,1).$$

Therefore

$$2\log\Lambda(\underline{X}) = \left(\sqrt{nI(\theta)}(\widehat{\theta}_n - \theta_0)\right)^2 = Z^2 \sim \chi^2_1.$$

Thus heuristically proving for the case $d = 1$.

4.5.1 Example: Binomial distribution

Recall the Binomial example, where $\{X_i\}_{i=1}^n$ are iid random variables which follow a Binomial distribution, i.e. $X_i \sim Bin(m, p)$. Suppose we test $H_0 : p = p_0$ vs $H_A : p \neq p_0$. From equation (4.7) we observe that the log-likelihood ratio is

$$\log\Lambda(\underline{X}) = mn(1-\widehat{p})\log\left(\frac{1-\widehat{p}}{1-p_0}\right) + mn\widehat{p}\log\frac{\widehat{p}}{p_0}.$$

Observe that the number of variables being estimated is $d = 1$. Thus by using the (4.9) with $d = 1$, under the null hypothesis we have

$$2\log\Lambda(\underline{X}) = 2mn(1-\widehat{p})\log\left(\frac{1-\widehat{p}}{1-p_0}\right) + 2mn\widehat{p}\log\frac{\widehat{p}}{p_0}.$$

In other words, for a sufficiently large n (remember m is kept fixed it is n that grows) the distribution of $2\Lambda(\underline{X})$ follows a chi-square distribution under the null. From the chi-tables we know that

$$P\left(\chi^2_1 \geq 3.84\right) = 0.05.$$

Thus if we do a test at the 5% level, then the approximate rejection region is

$$C_{0.05} = \left\{\underline{X}; 2\log\Lambda(\underline{X}) \geq 3.84\right\}.$$

Example Suppose we want to test the hypothesis that the proportion of people who like orange juice is $H_0 : p = 0.4$ vs $H_A : p \neq 0.4$.

A polling organisation asks $m = 20$ people (via telephone so we can keep social distancing) in each city, and samples $n = 30$ cities. If they like orange juice yes or no. Let X_i be the number of people out of 20 who like orange juice. The observed data is

$$\underline{x} = (9, 13, 8, 11, 9, 11, 10, 9, 12, 11, 11, 9, 12, 11, 10, 10, 11, 12, 11, 10, 8, 8, 16, 13, 8, 12, 12, 9, 7, 11).$$

Under the null hypothesis $p_0 = 0.4$ but under the alternative we use the MLE estimator which is

$$\widehat{p} = \frac{9 + 13 + \ldots + 11}{20 \times 30} = \frac{314}{20 \times 30} = 0.52.$$

Substituting this into $\Delta(\underline{x})$ gives

$$2 \log \Lambda(\underline{x}) \quad = \quad 2 \times 20 \times 30(1 - 0.52) \log \left(\frac{1 - 0.52}{1 - 0.4} \right) + 2 \times 20 \times 30 \times 0.52 \log \frac{0.52}{0.4} \approx 35.$$

Since $35 >> 3.84$ we reject the null at the 5% level (in fact at most reasonable levels!).

Sanity Check Recall that $Y = \sum_{i=1}^{n} X_i \sim Bin(nm, p)$. Therefore we could apply the normal approximation of the binomial distribution (covered in an introductory statistics class) to $\widehat{p}$. Suppose we test $H_0 : p = -p_0$ against $H_A : p \neq p_0$, then the test statistic you would have used is

$$T = \frac{(\widehat{p} - p_0)}{\sqrt{p(1 - p)/(mn)}}.$$

Under the null, for a large n is $T \xrightarrow{\mathcal{D}} N(0, 1)$ or equivalently $T^2 \sim \chi^2_1$. Given $\widehat{p} = 0.52$ we have

$$T = \frac{(0.52 - 0.4)}{\sqrt{0.6 \times 0.4/(20 \times 30)}} = 6.$$

Thus $T^2 = 36$, which is very close to $2\Lambda(\underline{x}) = 35$. This is not a coincidence, it can be shown that $2\Lambda(\underline{x}) \approx T^2$.

4.5.2 Example: The chi-square goodness of fit test

Suppose the null is $H_0 : \pi_1 = \tilde{\pi}_1, \ldots, \pi_m = \tilde{\pi}_m$ (where $\{\tilde{\pi}_i\}$ are some pre-set probabilities) and H_A : the probabilities are not the given probabilities. Hence we are testing restricted model (where we do not have to estimate anything) against the full model where we estimate the probabilities using $\pi_i = Y_i/n$.

The log-likelihood ratio in this case is

$$W = 2\Big\{ \arg\max_{\pi} \mathcal{L}_n(\pi) - \mathcal{L}_n(\tilde{\pi}) \Big\}.$$

Under the null we know that $W = 2\{\arg\max_\pi \mathcal{L}_n(\pi) - \mathcal{L}_n(\tilde{\pi})\} \xrightarrow{\mathcal{P}} \chi^2_{m-1}$ (because we have to estimate $(m-1)$ parameters). We now derive an expression for W and show that the Pearson-statistic is an approximation of this.

$$\begin{aligned}\frac{1}{2}W &= \sum_{i=1}^{m-1} Y_i \log\big(\frac{Y_i}{n}\big) + Y_m \log\frac{Y_m}{n} - \sum_{i=1}^{m-1} Y_i \log\tilde{\pi}_i - Y_m \log\tilde{\pi}_m \\ &= \sum_{i=1}^{m} Y_i \log\big(\frac{Y_i}{n\tilde{\pi}_i}\big).\end{aligned}$$

Recall that Y_i is often called the observed $Y_i = O_i$ and $n\tilde{\pi}_i$ the expected under the null $E_i = n\tilde{\pi}_i$. Then $W = 2\sum_{i=1}^m O_i \log\big(\frac{O_i}{E_i}\big) \xrightarrow{\mathcal{P}} \chi^2_{m-1}$. By making a Taylor expansion of $x\log(xa^{-1})$ about $x = a$ we have $x\log(xa^{-1}) \approx a\log(aa^{-1}) + (x-a) + \frac{1}{2}(x-a)^2/a$. We let $O = x$ and $E = a$, then assuming the null is true and $E_i \approx O_i$ we have

$$W = 2\sum_{i=1}^{m} Y_i \log\big(\frac{Y_i}{n\tilde{\pi}_i}\big) \approx 2\sum_{i=1}^{m}\big((O_i - E_i) + \frac{1}{2}\frac{(O_i-E_i)^2}{E_i}\big).$$

Now we note that $\sum_{i=1}^m E_i = \sum_{i=1}^m O_i = n$ hence the above reduces to

$$W \approx \sum_{i=1}^{m}\frac{(O_i-E_i)^2}{E_i} \xrightarrow{\mathcal{D}} \chi^2_{m-1}.$$

We recall that the above is the Pearson test statistic. Hence this is one methods for deriving the Pearson chi-squared test for goodness of fit.

Example: The independent two-sample t-test

See Chapter 5.

4.5.3 P-values

4.6 Confidence intervals and hypothesis tests

In this section we demonstrate the duality between certain hypothesis tests and confidence intervals. Previously, you probably saw something similar to this in an elementary statistics class.

We start with a motivating example. Suppose $\{X_i\}$ are iid normal random variables with mean μ and variance σ^2 (we assume the variance σ^2 is known). We test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$. We recall from Section 4.4.1 the generalized (log) likelihood ratio is

$$\log\Lambda(\underline{x}) = \log\frac{\sup_{\theta_1\in\Omega_1} f(\underline{x};\theta_1)}{\sup_{\theta_0\in\Omega} f(\underline{x};\theta_0)} = \frac{n}{2\sigma^2}(\bar{X}-\mu_0)^2.$$

If μ_0 is the true mean, than $2\log\Lambda(\underline{x}) \sim \chi^2_1$. Our objective is to show that the generalized likelihood ratio test can be used to construct a confidence interval for θ. To do so we emphasis the role of μ on the notation $\log\Lambda(\underline{x})$ and let

$$2\log\Lambda(\underline{x};\mu) \quad = \quad n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2.$$

We use as a $(1-\alpha)100\%$ confidence interval (or set) for μ the region

$$C_\alpha(\underline{x}) = \left\{\mu; 2\log\Lambda(\underline{x};\mu) = n\left(\frac{\bar{x}-\mu}{\sigma}\right)^2 < \chi_1(1-\alpha)^2\right\}.$$

We will show that this is a $(1-\alpha)100\%$ confidence interval for the mean and gives rise to the confidence interval as we know it. It is immediately clear that

$$\mu \in C_\alpha(\underline{X}) \text{ if and only if } n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 < \chi_1(1-\alpha)^2.$$

This implies that

$$P\left(\mu \in C_\alpha(\underline{X})|\mu\right) = P\left(n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 < \chi^2_{1-\alpha}|\mu\right) = 1-\alpha.$$

Thus $C_\alpha(\underline{X})$ is a $(1-\alpha)100\%$ confidence interval for μ.

We now show that $C_\alpha(\underline{X})$ is the same set as the classical confidence interval for the mean μ. If

$$\left(\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}\right)^2 \leq \chi^2_1(1-\alpha)$$

(this corresponds to not rejecting the null), then equivalently

$$-\sqrt{\chi^2_1(1-\alpha)} \leq \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \leq \sqrt{\chi^2_1(1-\alpha)}.$$

Multipling by $\sigma/\sqrt{n}$

$$-\sqrt{\chi^2_1(1-\alpha)}\frac{\sigma}{\sqrt{n}} \quad \leq \quad (\bar{X}-\mu) \leq \sqrt{\chi^2_1(1-\alpha)}\frac{\sigma}{\sqrt{n}}$$

or that μ lies in the interval

$$\left[\bar{X} - \sqrt{\chi^2_1(1-\alpha)}\frac{\sigma}{\sqrt{n}}, \bar{X} + \sqrt{\chi^2_1(1-\alpha)}\frac{\sigma}{\sqrt{n}}\right].$$

In other words, we do not reject the null at the $\alpha\%$ level if and only if μ lies in the interval

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

which is exactly the $(1-\alpha)\%$ confidence interval for the mean μ.

Summary There is an equivalence between a statistical test and the confidence interval. I.e. if

$$\left(\frac{(\bar{X}-\mu_0)}{\sigma/\sqrt{n}}\right)^2 \leq \chi_1^2(1-\alpha).$$

then μ_0 lies in the interval

$$\left[\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Equivalently if μ_0 lies in the interval

$$\left[\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right], \text{ then } \left(\frac{(\bar{X}-\mu_0)}{\sigma/\sqrt{n}}\right)^2 \leq \chi_1^2(1-\alpha)$$

and we do not reject the null.

This observation can be generalized to any test on a parameter.

Theorem 4.3

Converting a test into a confidence interval.

We observe the random vector $\underline{X}$. We test $H_0 : \theta = \theta_0$ against $H_A : \theta \neq \theta_0$. Suppose $\underline{X} \in R_\alpha(\theta_0)$, then we not reject the null at the $\alpha\%$ level. Using this we define the interval

$$C_\alpha(\underline{X}) = \left\{\theta; \underline{X} \in R_\alpha(\theta)\right\}.$$

Then $C_\alpha(\underline{X})$ is a $(1-\alpha)100\%$ confidence interval for θ.

Converting a confidence interval into a test.

Suppose $C_\alpha(\underline{X})$ is a $(1-\alpha)100\%$ confidence interval for the θ. Define the interval

$$R_\alpha(\theta) = \left\{\underline{X}; \theta \in C_\alpha(\underline{X})\right\}.$$

Then, if we test $H_0 : \theta = \theta_0$ against $H_0 : \theta \neq \theta_0$ the nonrejection region at the $\alpha\%$ level is $R_\alpha(\theta_0)$.

Comparing the theorem with the motivating example above, we set

$$R_\alpha(\mu_0) = \left[\mu_0 - \sqrt{\chi_1^2(1-\alpha)}\frac{\sigma}{\sqrt{n}}, \mu_0 + \sqrt{\chi_1^2(1-\alpha)}\frac{\sigma}{\sqrt{n}}\right]$$

and

$$C_\alpha(\underline{X}) = \left[\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

4.6.1 Example: The binomial distribution

The statement of Theorem 4.3 is a little confusing. But is easiest understood with an example. Let us return to the binomial example in Section 4.5.1. Set $m = 1$ and let n be large (or visa-versa). If we test $H_0 : p = p_0$ vs $H_A : p \neq p_0$, then from (4.7) the log-likelihood ratio is

$$2 \log \Lambda(\underline{X}; p_0) \quad = \quad 2n(1 - \widehat{p}) \log \left(\frac{1 - \widehat{p}}{1 - p_0} \right) + 2n\widehat{p} \log \frac{\widehat{p}}{p_0},$$

where $\widehat{p} = X/n$. If the null holds, for sufficiently large m we have $2 \log \Lambda(\underline{X}; p_0) \leq \chi^2_1$ (this follows from Theorem 4.2). Therefore, we do not reject the null if

$$2 \log \Lambda(\underline{X}; p_0) < \chi^2_1(1 - \alpha).$$

We use this result to construct an $(1 - \alpha)\%$ confidence interval for p. We use as the $(1 - \alpha)100\%$ confidence interval p, all p such that

$$C_\alpha(\underline{X}) = \left\{ p; 2 \log \Lambda(\underline{X}; p) < \chi^2_1(1 - \alpha) \right\} \quad = \quad \left\{ p; 2n(1 - \widehat{p}) \log \left(\frac{1 - \widehat{p}}{1 - p} \right) + 2n\widehat{p} \log \frac{\widehat{p}}{p} < \chi^2_1(1 - \alpha) \right\}.$$

And the reason this is a $(1 - \alpha)100\%$ confidence interval is that for any given p

$$P\left(p \in C_\alpha(\underline{X}) | p \right) = P\left(2 \log \Lambda(\underline{X}; p) < \chi^2_1(1 - \alpha) | p \right) \approx 1 - \alpha \text{ for large enough } n.$$

5 Comparing two populations

5.1 Comparing two independent samples

5.1.1 Example: Independent two sample data

It is conjectured that exposure to different light colours may influence the strength egg shells of hens. A team of animal science students decided to investigate this. Hens were randomly assigned to two different light treatments: red light or white light. The hens either spent 6 weeks exposed to only red light or 6 weeks only exposed to white light (both with dark inbetween). After 6 weeks the strength of the eggs for each hen was measured (in terms of the Haugh index) and A snapshot of the data is summarized below, note that each column corresponds to one hen.

Treatment						
White Light	99.85	99.62	104.82	108.36	107.75	108.01
Red Light	102.12	108.37	99.05	98.58	99.52	

We observe in the way that the experiment was designed that there is no dependence between the hens in treatment groups and between the treatment groups. Thus we can assume all the observations are independent. This is a big assumption and can determine the type of procedure that one uses. If this assumption does not hold but procedures assuming independent was done, then the results of the test are not valid. And can give rise to spurious conclusions.

Remark (How data is presented in a spreadsheet)**.** *In most spreadsheets the data won't be presented as above. Usually the data is presented as follows:*

Observation	X_1	X_2	Y_1	Y_2	X_3	$\ldots$	X_n	Y_m
Treatment	A	A	B	B	A	$\ldots$	A	B

One row (or column) contains the treatment an "individual" is given and another row (column) is the response given the treatment. For example, the treatment could be white or red light and the observation is the strength of the egg corresponding to that light treatment.

Our objective is to see if different light treatments give rise to eggs with different shell strengths. Of course, one cannot compare individual eggs. So by different, we mean in the sense of parameters within both

populations. The most common parameter of comparison that one uses (but it is not the only one) is the population means. We now state the model assumptions. But before we start with the modelling, we first plot the data. Comparing the two shapes of the histograms is difficult. They do not "look" alike, but this

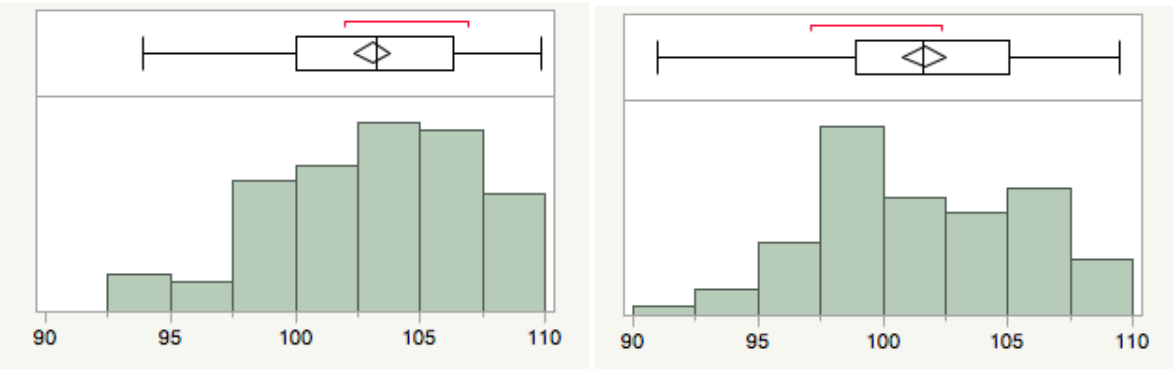


Figure 5.1: Histogram of Haugh index of given different light treatments. Left: Red Treatment (sample mean 103.1). Right: White Treatment (sample mean 101.7) Sample sizes in both groups is 101.

could be due to sampling variation. It is difficult to tell if they are normal (a QQplot would be useful). However, we do observe differences in the sample means. The question we must ask ourselves is if this difference is statistically significant.

5.1.2 Modelling assumptions

We start with some formal definitions. We assume that $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$ are independent random variables (both between and within groups) where

$$X_i \sim N(\mu_1, \sigma^2) \quad \text{and} \quad Y_j \sim N(\mu_2, \sigma^2).$$

Observe that we have made the assumption that the variance is the same for both sets of random variables. In other words, the data is assumed to have come from a normal distribution. Often you may see X_i and Y_j written in the following equivalent way

$$X_i = \mu_1 + \varepsilon_i \quad \text{and} \quad Y_j = \mu_2 + \epsilon_j \tag{5.1}$$

where $\{\varepsilon_i, \epsilon_j\}$ are independent, identically distributed normal random variables with mean zero and variance σ^2. This equivalent representation separates the mean from the noise. It is very similar in flavour to the representation of a linear regression model.

Based on the above model a hypothesis of interest is

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_0 : \mu_1 - \mu_2 \neq 0.$$

This can easily be generalized to test

$$H_0 : \mu_1 - \mu_2 = \delta \text{ vs } H_0 : \mu_1 - \mu_2 \neq \delta.$$

One way of testing this hypothesis is to estimate $\mu_1 - \mu_2$, by estimating it with $\bar{X} - \bar{Y}$. Thus we need to obtain the distributional properties of $\bar{X} - \bar{Y}$. Since $X_i \sim N(\mu_1, \sigma^2)$ and $Y_j \sim N(\mu_2, \sigma^2)$, by the assumption of independence and Section 2.1 we have

$$\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/m \end{pmatrix} \right).$$

Since $\bar{X}$ and $\bar{Y}$ are jointly normal, $\bar{X} - \bar{Y}$ is normal too with

$$\mathrm{E}\left(\bar{X} - \bar{Y}\right) = \mathrm{E}(\bar{X}) - \mathrm{E}(\bar{Y}) = \mu_1 - \mu_2$$

and

$$\begin{aligned} \mathrm{var}\left(\bar{X} - \bar{Y}\right) &= \mathrm{cov}(\bar{X} - \bar{Y}, \bar{X} - \bar{Y}) \\ &= \mathrm{var}[\bar{Y}] + \mathrm{var}[\bar{X}] - 2\underbrace{\mathrm{cov}(\bar{X}, \bar{Y})}_{=0} \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{m}. \end{aligned}$$

Using the above the distribution of $\bar{X} - \bar{Y}$ is

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right).$$

Thus under the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$ we have

$$\bar{X} - \bar{Y} \sim N\left(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right).$$

Therefore to test if $\mu_1 - \mu_2 = 0$ we use the distance $\bar{X} - \bar{Y}$ but take into account its standard error $\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}$. Under the null it is clear that the z-transform is

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim N(0, 1).$$

But if the alternative is true then

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} = \underbrace{\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}}_{=Z \sim N(0,1)} + \underbrace{\frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}}_{=\text{mean}}$$

this gives

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}} \sim N\left(\frac{\mu_1-\mu_2}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}, 1\right). \tag{5.2}$$

It is the "size" of $\frac{\mu_1-\mu_2}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}$ that gives the statistical test power.

Using what we learnt in Section 4.1; under the null $Z = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}$ is a standard normal, then the rejection region at the α level is

$$\left|\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}\right| \geq z_{\alpha/2}.$$

Or equivalently, we reject the null if $\bar{X}-\bar{Y}$ lies in $C_- \cup C_+$ where

$$C_- = \left(-\infty, -z_\alpha\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}\right] \qquad C_- = \left[z_\alpha\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}, \infty\right).$$

Note that $P(\bar{X}-\bar{Y} \in C_- \cup C_+ | H_0) = \alpha$.

Power of the test Of course there is no point to a test, if it does not have power in detecting the alternative. Without loss of generality let us assume that $\mu_1 - \mu_2 = \delta > 0$. The power in the test is

$$P\left(\bar{X}-\bar{Y} \in C_- \cup C_+ | \mu_1-\mu_2=\delta\right) \approx P\left(\bar{X}-\bar{Y} \in C_+ | \mu_1-\mu_2=\delta\right),$$

since if $\delta > 0$ it is unlikely that $\bar{X}-\bar{Y} \leq -z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}$. Thus

$$\begin{aligned} P\left(\bar{X}-\bar{Y} \in C_- \cup C_+ | \mu_1-\mu_2=\delta\right) &\approx P\left(\bar{X}-\bar{Y} \geq z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}} | \mu_1-\mu_2=\delta\right) \\ &= P\left(Z + \frac{\mu_1-\mu_2}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}} \geq z_{\alpha/2}\right) \\ &= P\left(Z \geq -\frac{\delta}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}} + z_{\alpha/2}\right), \end{aligned}$$

where $Z \sim N(0,1)$. Now for a given δ, σ^2, μ_1 and μ_2 this probability can easily be calculated. As is always the case, the larger δ, or m and n the larger the power (if you have doubts try different values and calculate the power).

Observation The testing procedure assumes that the variance σ^2 is known. This will not usually be the case. In the following section we consider a method for estimating σ^2.

5.1.3 Pooling information: The pooled sample variance

Our aim in this section is to estimate σ^2. The method we propose turns out to be MLE estimator of σ^2 under the assumption the the means are different (see Section 5.2).

The following derivation of the sample variance is based on the estimation of residuals. Let us return to "model" in (5.1), where we wrote X_i and Y_j as

$$X_i = \mu_1 + \varepsilon_i \quad \text{and} \quad Y_j = \mu_2 + \epsilon_j$$

where $\{\varepsilon_i, \epsilon_j\}$ are independent, identically distributed normal random variables with $\mathrm{E}[\varepsilon_i] = 0$, $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{var}[\varepsilon_i] = \sigma^2$ and $\mathrm{var}[\epsilon_i] = \sigma^2$. The random variables ε_i and ϵ_j are often called the residuals in the model (the part of the model that cannot be explained by the mean). The are not observed, but they can be estimated using what is called the sample residuals. We estimate μ_1 and μ_2 with $\bar{X}$ and $\bar{Y}$ respectively. Thus the estimated residuals are

$$\widehat{\varepsilon_i} = X_i - \bar{X} \quad \text{and} \quad \widehat{\epsilon_j} = Y_j - \bar{Y}.$$

We can imagine that if the sample size is "large" then $\bar{X}$ and $\bar{Y}$ are "close" to μ_1 and μ_2 respectively. Thus approximately

$$\begin{aligned}
\mathrm{E}[\widehat{\varepsilon_i}] &\approx \mathrm{E}[\varepsilon_i] = 0 \\
\mathrm{E}[\widehat{\varepsilon_i}^2] &\approx \mathrm{E}[\varepsilon_i^2] = \sigma^2 \\
\mathrm{E}[\widehat{\epsilon_i}] &\approx \mathrm{E}[\epsilon_i] = 0 \\
\mathrm{E}[\widehat{\epsilon_i}^2] &\approx \mathrm{E}[\epsilon_i^2] = \sigma^2.
\end{aligned}$$

What the above tells us that the expectation of the square of the estimated residuals is the almost the variance σ^2. As the variance is the same for both $\{X_i\}$ and $\{Y_j\}$ we "pool" the two bits of information to obtain a better estimator.

To pool the information, we recall if the expectation of a random variable is equal to θ, the sample mean of the random variables is a good estimator of θ. By this principle, the average of all the squared residuals should yield a good estimator of the variance:

$$\widehat{\sigma}^2 = \frac{1}{m+n}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right).$$

Observe in the above we have pooled information from both the Xs and Ys.

But the above estimator is biased. We now show why. Recall from Theorem 2.3 that

$$\mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \sigma^2 \quad \text{and} \quad \mathrm{E}\left(\frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2\right) = \sigma^2.$$

Rearranging the above gives

$$\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=(n-1)\sigma^2 \quad \text{and} \quad \mathrm{E}\left(\sum_{j=1}^{m}(Y_j-\bar{Y})^2\right)=(m-1)\sigma^2.$$

Using these expectations we have

$$\begin{aligned}\mathrm{E}[\widehat{\sigma}^2] &= \frac{1}{m+n}\left(\mathrm{E}[\sum_{i=1}^{n}(X_i-\bar{X})^2]+\mathrm{E}[\sum_{j=1}^{m}(Y_j-\bar{Y})^2]\right)\\ &= \frac{1}{m+n}\left(\mathrm{E}[(n-1)\sigma^2+(m-1)\sigma^2\right)\\ &= \frac{m+n-2}{m+n}\sigma^2=\sigma^2-\frac{2}{m+n}\sigma^2.\end{aligned}$$

Thus the estimator $\widehat{\sigma}^2$ has a small bias (the bias is $-\frac{2}{m+n}\sigma^2$). An unbiased estimator of σ^2 is the pooled sample variance which is

$$\begin{aligned}\widehat{s}_p^2 &= \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{j=1}^{m}(Y_j-\bar{Y})^2\right)\\ &= \frac{1}{m+n-2}\left((n-1)s_X^2+(m-1)s_Y^2\right),\end{aligned}$$

where s_X^2 and s_Y^2 are the unbiased sample variance estimators

$$s_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad \text{and } s_Y^2=\frac{1}{m-1}\sum_{j=1}^{m}(Y_j-\bar{Y})^2.$$

We recall from Theorem 2.3 that

$$(n-1)s_X^2=\sum_{i=1}^{n}(X_i-\bar{X})^2\sim\sigma^2\chi_{n-1}^2 \quad \text{and } (m-1)s_Y^2=\sum_{j=1}^{m}(Y_j-\bar{Y})^2\sim\sigma^2\chi_{m-1}^2.$$

Since $\{X_i\}$ and $\{Y_j\}$ are independent of each other, then $(n-1)s_X^2$ and $(m-1)s_Y^2$ are independent. Further if χ_{n-1}^2 and χ_{m-1}^2 are independent chi-square random variables then $\chi_{n-1}^2+\chi_{m-1}^2$ has a χ_{m+n-2}^2-distribution. Thus altogether we have

(i) Unbiased pooled sample variance

$\mathrm{E}[s_p^2]=\sigma^2$.

(ii) Distribution of the pooled sample variance

$$\left(\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{j=1}^{m}(Y_j-\bar{Y})^2\right)=\left((n-1)s_X^2+(m-1)s_Y^2\right)\sim\sigma^2\chi_{n+m-2}^2. \tag{5.3}$$

(iii) Independence from sample mean

Since $\sum_{i=1}^{n}(X_i-\bar{X})^2$ and $\sum_{j=1}^{m}(Y_j-\bar{Y})^2$ are independent of $\bar{X}$ and $\bar{Y}$ (again by Theorem 2.3) we have that $(\bar{X}-\bar{Y})$ are independent of s_p^2.

Very important the above derivations are based in the assumption (a) independence (b) same variance in both groups (extremely important) and (c) normality of the random variables.

Remark (Alternative, but worse estimators of the variance)**.** *You may wonder why we do not use*

$$\frac{1}{2}\left(s_X^2 + s_Y^2\right) \tag{5.4}$$

as an estimator of σ^2. This estimator can also be used. In fact if $m = n$ the pooled sample variance and the above are the same. However, in the case $m \neq n$, the pooled sample variance gives a larger weight to the group with the larger sample size:

$$\widehat{\sigma}^2 \quad = \quad \frac{(n-1)}{m+n-2}s_X^2 + \frac{(m-1)}{n+m-2}s_Y^2,$$

which leads to an estimator with a smaller variance and bias (is better) than (5.4).

5.1.4 The independent two sample t-test

In the previous section we obtained an estimator of σ^2. We recall from (5.2) that if σ^2 is known then

$$\frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1)\,. \tag{5.5}$$

And from (5.3) that the pooled sample variance s_p^2 is independent of $\bar{X}-\bar{Y}$ and $(n+m-2)s_p^2/\sigma^2 \sim \chi^2_{n+m-2}$. Thus replacing σ with s_p yields

$$T = \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}. \tag{5.6}$$

To obtain the distribution of T we divide the numerator and denominator by σ (which does not change t_{n+m-2}) this gives

$$T = \left(\frac{\sigma}{s_p}\right)\frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim \frac{Z}{\chi_{m+m-2}/\sqrt{m+n-2}} \sim t_{n+m-2} \tag{5.7}$$

Returning to the test

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_0 : \mu_1 - \mu_2 \neq 0.$$

We use as the rejection region at the α-level, all $\bar{X}$ where

$$|T| = \left|\frac{(\bar{X}-\bar{Y})}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}\right| \geq t_{n+m-2,\alpha/2}. \tag{5.8}$$

Thus if $|T| \geq t_{n+m-2,\alpha/2}$ we reject the null hypothesis, and there is evidence that $\mu_1 \neq \mu_2$.

Extending the above to the one-sided set-up is straightforward.

5.2 Generalized likelihood ratio test and the independent two sample t-test

Our aim in this section is to show that independent two-sample t-test falls within the generalized log-likelihood ratio t-test. The proof mirrors the proof of showing that the two-sided t-test for one sample was equivalent to the generalized likelihood ratio t-test (described in Section 4.4.2). This section is not in the syllabus but may be of interest to you. We consider the simplest setting that the mean of the two populations under the null are the same:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_0 : \mu_1 - \mu_2 \neq 0.$$

The generalized likelihood ratio test is constructed under the following assumptions:

- The variances in both groups are the same.
- The two groups $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$ are independent with $X_i \sim N(\mu_1, \sigma^2)$ and $Y_j \sim N(\mu_2, \sigma^2)$.

To obtain the likelihoods, we define the concatonated $(n+m)$-dimension vector $\underline{Z}_{n+m}$

$$\underline{Z}'_{n+m} = (X_1, \ldots, X_n, Y_1, \ldots, Y_m),$$

and n-dimension and m-dimension vector $\underline{X}' = (X_1, \ldots, X_n)$ and $\underline{Y}' = (Y_1, \ldots, Y_m)$. Since

$$\underline{Z}_{n+m} \sim N\left(\begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}, \sigma^2 I_{n+m} \right).$$

The log-likelihood for $\underline{Z}$ is

$$\mathcal{L}_{n+m}(\mu_1, \mu_2, \sigma^2; \underline{Z}_{n+m}) = -\frac{n+m}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu_1)^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{m} (Y_i - \mu_2)^2 \tag{5.9}$$

(we have ignored the $\frac{n+m}{2} \log(2\pi)$ term). The aim is to obtain an expression for the generalized log-likelihood ratio statistic:

$$\begin{aligned} \log \Delta(\underline{z}) &= \underbrace{\sup_{\mu_1, \mu_2, \sigma^2} \mathcal{L}_{n+m}(\mu_1, \mu_2, \sigma^2; \underline{Z}_{n+m})}_{\text{log-likelihood under alternative}} - \underbrace{\sup_{\mu_0, \sigma^2} \mathcal{L}_{n+m}(\mu_0, \mu_0, \sigma^2; \underline{Z}_{n+m})}_{\text{log-likelihood under null}} \\ &= \mathcal{L}_{n+m}(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2; \underline{Z}_{n+m}) - \mathcal{L}_{n+m}(\widehat{\mu}_0, \widehat{\mu}_0, \widehat{\sigma}_0^2; \underline{Z}_{n+m}). \end{aligned} \tag{5.10}$$

We will show that

$$\log \Delta(\underline{z}) = \frac{m+n}{2} \log \left(1 + \frac{\frac{nm}{n+m} \left(\bar{X} - \bar{Y} \right)^2}{(n+m-2) s_{n+m-2}^2} \right), \tag{5.11}$$

from which we derive the independent sample t-test.

Deriving (5.11) through brute force is extremely painful (but possible; see most standard text books). Instead, we use results in linear-algebra to ease the pain and make the proof more informative.

The main result we use is stated in Section 2.1, point (7). If one makes an orthonormal transform of the normal random vector $\underline{Z}$, then the likelihood of $\underline{Z}$ and $E\underline{Z}$ are exactly the same. We have used this result frequently namely in Sections 3.4.2, 4.4.1 and 4.4.2. It is so useful, we use it again here. Our objective is to find two orthonormal transforms, one suitable for the null and the other suitable for the alternative hypothesis.

A useful orthonormal transform for the alternative hypothesis

t is clear that if the alternative is true, we do not need to "pool" the sample means together and the MLE for μ_1 and μ_2 should be the sample means $\bar{X}$ and $\bar{Y}$. This we simply use the constructions from Section 2.4.2 and the constructions in Sections 3.4.2, 4.4.1 and 4.4.2.

Define the n-dimension orthonormal vectors $\underline{e}_{1,n}, \ldots, \underline{e}_{n,n}$ where

$$\underline{e}_{1,n} = n^{-1/2}(1, \ldots, 1)$$

and $\{\underline{e}_{j,n}\}_{j=2}^{n}$ are orthonormal to $\underline{e}_{1,n}$ (examples include sines and cosines the discrete Haar transform, it actually does not matter what this basis is, except that it exists). Similarly one can define the m-dimension orthonormal vectors $\underline{f}_{1,m}, \ldots, \underline{f}_{m,m}$ where

$$\underline{f}_{1,m} = m^{-1/2}(1, \ldots, 1)$$

and $\{\underline{f}_{j,m}\}_{j=2}^{m}$ are orthonormal to $\underline{f}_{1,m}$. Using these vectors we can define a $(n+m) \times (n+m)$ matrix which transforms $\underline{Z}_{n+m}$ such that

$$\underline{W}_{n+m} = E\underline{Z}_{n+m} = \begin{pmatrix} \langle \underline{X}, \underline{e}_{1,n} \rangle \\ \langle \underline{Y}, \underline{f}_{1,m} \rangle \\ \langle \underline{X}, \underline{e}_{2,n} \rangle \\ \langle \underline{X}, \underline{e}_{3,n} \rangle \\ \vdots \\ \langle \underline{Y}, \underline{f}_{m,m} \rangle \end{pmatrix} = \begin{pmatrix} \sqrt{n}\bar{X} \\ \sqrt{m}\bar{Y} \\ U_2 \\ U_3 \\ \vdots \\ V_m \end{pmatrix} \sim N\left(\begin{pmatrix} \sqrt{n}\mu_1 \\ \sqrt{m}\mu_2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 I_{n+m} \right).$$

This gives rise to the log-likelihood (see equation (3.5) for an analogous calculation)

$$\begin{aligned} \mathcal{L}_n(\mu_1, \mu_2, \sigma^2; \underline{Z}) &= \mathcal{L}_n(\mu_1, \mu_2, \sigma^2; \underline{W}) \\ &= -\frac{(n+m)}{2} \log \sigma^2 - \frac{n}{2\sigma^2}(\bar{X} - \mu_1)^2 - \frac{m}{2\sigma^2}(\bar{Y} - \mu_1)^2 - \frac{1}{\sigma^2} \sum_{i=2}^{n} U_i^2 - \frac{1}{\sigma^2} \sum_{j=2}^{m} V_j^2. \end{aligned}$$

Differentiating the above wrt μ_1, μ_2 and σ^2 leads to the MLE estimators

$$\begin{aligned}\widehat{\mu}_1 &= \bar{X}, \quad \widehat{\mu}_2 = \bar{Y} \text{ and} \\ \widehat{\sigma}_1^2 &= \frac{1}{n+m}\left(\sum_{i=2}^{n} U_i^2 + \sum_{j=2}^{m} V_j^2\right) = \frac{1}{n+m}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right].\end{aligned}$$

Substituting this into $\mathcal{L}_n(\mu_1, \mu_2, \sigma^2; \underline{Z})$ gives

$$\begin{aligned}\mathcal{L}_n(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2; \underline{Z}) &= \mathcal{L}_n(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2; \underline{W}) \\ &= -\frac{(n+m)}{2}\log\widehat{\sigma}_1^2 - \frac{(n+m)}{2}.\end{aligned}$$

A useful orthonormal transform for the null hypothesis

Finding the transform when the null is true is a little more tricky. Under the null, $\{X_i\}, \{Y_j\}$ are iid normal random variables and the MLE of μ should "pool" the sample means together. Thus the first vector we use is the $(n+m)$-dimension vector

$$\underline{h}_{1,n+m} = (n+m)^{-1}(1, 1, 1, 1, \ldots, 1).$$

But we want to keep the vectors $\{\underline{e}_{i,n}\}_{i=2}^{n}$ and $\{\underline{f}_{j,m}\}_{j=2}^{m}$, to cancel out the $\frac{1}{\sigma^2}\sum_{i=2}^{n} U_i^2$ and $\frac{1}{\sigma^2}\sum_{j=2}^{m} V_j^2$ when we subtract the null from the alternative. Thus we still keep the $(n+m-2)$ $(n+m)$-dimension vectors

$$\underline{e}_i = (\underline{e}_{i,n}, \underline{0}) \qquad i = 2, \ldots, n \text{ and } \underline{f}_j = (\underline{0}, \underline{f}_{j,m}) \qquad j = 2, \ldots, m.$$

Altogether this gives $(n+m-1)$-orthonormal transform vectors $\{\underline{h}_{1,n+m}, \underline{e}_i, \underline{f}_j\}$; with a little thought you can see that $\underline{h}_{1,n+m}$ is orthogonal to $\underline{e}_i$ and $\underline{f}_j$ and further that $\underline{e}_i$ and $\underline{f}_j$ are orthogonal to each other. But we need one more vector to complete the basis. If you get a cup of tea and spends a few moments thinking you start to realize there is only vector that can be orthonormal to the others and it is

$$\underline{h}_{2,n+m} = \sqrt{\frac{nm}{m+n}}\left(\underbrace{(1/n), \ldots, (1/n)}_{n\text{of these}}, \underbrace{(-1/m), \ldots, (-1/m)}_{m\text{of these}}\right).$$

This completes the orthonormal basis over $\mathbb{R}^{n+m}$. Using these vectors we define the orthonormal transform matrix F and the transform $F\underline{Z}_{n+m}$. Under the null the distribution of $F\underline{Z}_{n+m}$ we have

$$\underline{S}_{n+m} = F\underline{Z}_{n+m} = \begin{pmatrix} \langle \underline{Z}, \underline{h}_{1,n}\rangle \\ \langle \underline{Z}, \underline{h}_{2,m}\rangle \\ \langle \underline{X}, \underline{e}_{2,n}\rangle \\ \langle \underline{X}, \underline{e}_{3,n}\rangle \\ \vdots \\ \langle \underline{Y}, \underline{f}_{m,m}\rangle \end{pmatrix} = \begin{pmatrix} \sqrt{n+m}(n\bar{X} + m\bar{Y}) \\ \langle \underline{Z}, \underline{h}_{2,m}\rangle \\ U_2 \\ U_3 \\ \vdots \\ V_m \end{pmatrix} \sim N\left(\begin{pmatrix} \sqrt{n+m}\mu_0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 I_{n+m}\right).$$

This gives rise to the log-likelihood (see equation (3.5) for an analogous calculation)

$$\begin{aligned}\mathcal{L}_n(\mu_0, \mu_0, \sigma^2; \underline{Z}) &= \mathcal{L}_n(\mu_0, \mu_0, \sigma^2; \underline{S}) \\ &= -\frac{(n+m)}{2}\log\sigma^2 - \frac{(n+m)}{2\sigma^2}(\bar{Z}-\mu)^2 - \frac{1}{2\sigma^2}\langle \underline{Z}, \underline{h}_{2,m}\rangle^2 - \frac{1}{\sigma^2}\sum_{i=2}^{n} U_i^2 - \frac{1}{\sigma^2}\sum_{j=2}^{m} V_j^2.\end{aligned}$$

By differentiating the above wrt σ^2 and μ_0 the MLE estimators of σ^2 and μ_0 are

$$\begin{aligned}\widehat{\mu}_0 &= \frac{1}{n+m}\left(\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j\right) = \frac{1}{n+m}\left(n\bar{X} + m\bar{Y}\right) \\ \widehat{\sigma}_0^2 &= \frac{1}{n+m}\left(\langle \underline{Z}, \underline{h}_{2,m}\rangle^2 + \sum_{i=2}^{n} U_i^2 + \sum_{j=2}^{m} V_j^2\right).\end{aligned}$$

Substituting this into $\mathcal{L}_n(\mu_0, \mu_0, \sigma^2; \underline{Z})$ gives

$$\begin{aligned}\mathcal{L}_n(\widehat{\mu}_0, \widehat{\mu}_0, \widehat{\sigma}_0^2; \underline{Z}) &= \mathcal{L}_n(\widehat{\mu}_0, \widehat{\mu}_0, \widehat{\sigma}_0^2; \underline{W}) \\ &= -\frac{(n+m)}{2}\log\widehat{\sigma}_0^2 - \frac{(n+m)}{2}.\end{aligned}$$

The log-generalized likelihood ratio statistic

Substituting the above likelihoods into (5.10) gives

$$\begin{aligned}\log\Delta(\underline{z}) &= \mathcal{L}_{n+m}(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2; \underline{Z}_{n+m}) - \mathcal{L}_{n+m}(\widehat{\mu}_0, \widehat{\mu}_0, \widehat{\sigma}_0^2; \underline{Z}_{n+m}) \\ &= \frac{m+n}{2}\log\left(\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_0^2}\right) \\ &= \frac{m+n}{2}\log\left(\frac{\langle \underline{Z}, \underline{h}_{2,m}\rangle^2 + \sum_{i=2}^{n} U_i^2 + \sum_{j=2}^{m} V_j^2}{\sum_{i=2}^{n} U_i^2 + \sum_{j=2}^{m} V_j^2}\right) \\ &= \frac{m+n}{2}\log\left(1 + \frac{\langle \underline{Z}, \underline{h}_{2,m}\rangle^2}{\sum_{i=2}^{n} U_i^2 + \sum_{j=2}^{m} V_j^2}\right).\end{aligned}$$

We now make some simplications of the above. We recall from Section 2.4.2 that

$$\begin{aligned}\sum_{i=2}^{n} U_i^2 &= \sum_{i=2}^{n}\langle \underline{e}_{i,n}, \underline{X}\rangle^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \\ \sum_{j=2}^{m} V_j^2 &= \sum_{j=2}^{m}\langle \underline{f}_{j,n}, \underline{Y}\rangle^2 = \sum_{j=1}^{m}(Y_i - \bar{Y})^2.\end{aligned}$$

Therefore, from the definition of the pooled sample variance we have

$$s_{n+m}^2 = \frac{1}{n+m-2}\left(\sum_{i=2}^{n} U_i^2 + \sum_{j=2}^{m} V_j^2\right).$$

Further Before we evaluate the generalised log-likelihood ratio we mention that

$$\begin{aligned}\langle \underline{Z}, \underline{h}_{2,m}\rangle^2 &= \langle \underline{Z}, \underline{h}_{1,m}\rangle^2 - n\bar{X}^2 - m\bar{Y}^2 \\ &= \frac{1}{n+m}\left(n\bar{X}+m\bar{Y}\right)^2 - n\bar{X}^2 - m\bar{Y}^2 \\ &= \frac{nm}{n+m}\left(\bar{X}-\bar{Y}\right)^2.\end{aligned}$$

The above can be directly verified or follows immediately from Parseval's identity. These two identities give

$$\log \Delta(\underline{z}) = \frac{m+n}{2}\log\left(1+\frac{\frac{nm}{n+m}\left(\bar{X}-\bar{Y}\right)^2}{(n+m-2)s^2_{n+m-2}}\right),$$

which proves (5.11)

Finally we need to evaluate the rejection region for $\{X_i\}$ and $\{Y_j\}$ under the null such that

$$\begin{aligned}P\left(\log\Lambda(\underline{z}) \geq K|H_0\right) &= P\left(\frac{m+n}{2}\log\left(1+\frac{\frac{nm}{n+m}\left(\bar{X}-\bar{Y}\right)^2}{(n+m-2)s^2_{n+m-2}}\right) \geq K|H_0\right) \\ &= P\left(\frac{\frac{nm}{n+m}\left(\bar{X}-\bar{Y}\right)^2}{s^2_{n+m-2}} \geq \widetilde{K}|H_0\right),\end{aligned}$$

where $\widetilde{K} = (n+m-2)e^{2K/(m+1)} - 1$. Using that

$$\frac{nm}{n+m}\frac{\left(\bar{X}-\bar{Y}\right)^2}{s^2_{n+m-2}} = \left(\frac{1}{n}+\frac{1}{m}\right)^{-1}\frac{(\bar{X}-\bar{Y})^2}{s^2_{n+m-2}}$$

it is clear that under the null and by using (**??**) we have

$$\frac{nm}{n+m}\frac{\left(\bar{X}-\bar{Y}\right)^2}{s^2_{n+m-2}} = \left(\frac{1}{n}+\frac{1}{m}\right)^{-1}\frac{(\bar{X}-\bar{Y})^2}{s^2_{n+m-2}} \sim t^2_{n+m-2} = F_{1,n+m-2}.$$

Thus

$$P\left(\log\Lambda(\underline{z}) \geq K|H_0\right) = P\left(\left(\frac{1}{n}+\frac{1}{m}\right)^{-1}\frac{(\bar{X}-\bar{Y})^2}{s^2_{n+m-2}} \geq \widetilde{K}\middle|H_0\right) = \alpha.$$

Thus under the generalized log-likelihood ratio test we reject the null when

$$\left(\frac{1}{n}+\frac{1}{m}\right)^{-1/2}\frac{|\bar{X}-\bar{Y}|}{s_{n+m-2}} \geq t_{n+m-2,\alpha/2},$$

which is identical to the independent two-sample t-test.

5.3 Matched data

5.3.1 Example: matched data

Biologists want to understand how the weight of mammals changes from birth to a week after birth. In their investigation they focussed on new born calves. They randomly sampled 44 calves at various different

farms and took their weight at birth and then one week later. Plots of the data are given in Figure 5.2. Observe that the data set comes in pairs, where i denotes the individual calf, X_i the weight of the calf at

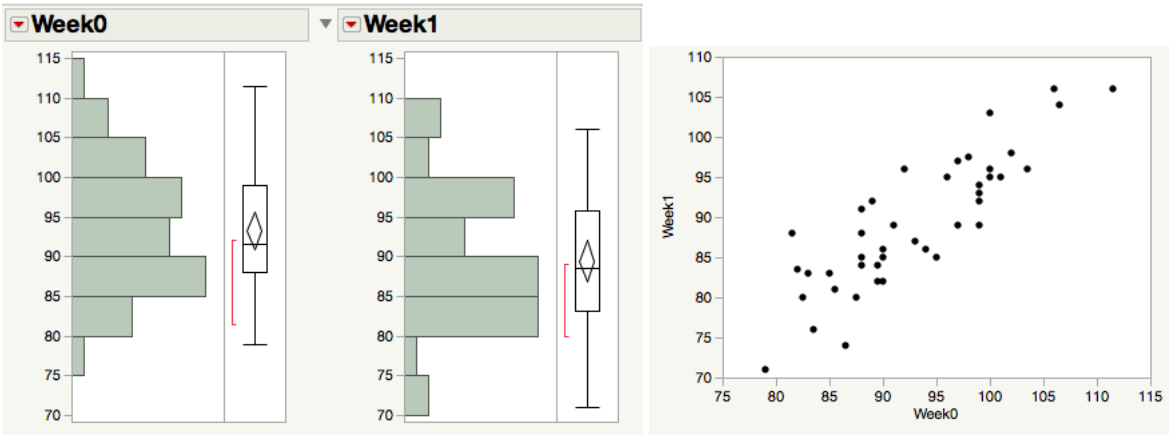


Figure 5.2: Left: The histogram of the 44 calves at week 0 and week 1. Right: A scatterplot of the calf data, Week 1 weight plotted against Birth weight.

birth and Y_i the weight of the calf at week 1. Other examples of matched data include (i) the running times of a person at a high altitude and a low altitude (ii) the response of an individual before and after treatment (iii) even data at is collected at the same time in the year.

In this data set we do observe a drop in the sample mean from birth to Week 1. Our objective is to investigate if this is statistically significant using the test

$$H_0 : \mu_{\text{Birth}} - \mu_{\text{Week1}} = 0 \text{ vs } H_A : \mu_{\text{Birth}} - \mu_{\text{Week1}} \neq 0.$$

The type of testing procedure used depends the relationship between (X_i, Y_i). A scatter plot of the pairs $\{(X_i, Y_i)\}_{i=1}^n$ is on the right plot of Figure 5.2. We observe a clear "dependence" between the two variables. This is a violation of a fundamental assumption in the independent two sample t-test. Below we obtain a model for matched data that allows for dependence.

5.3.2 Model assumptions

For matched data, we usually match each individual pairing as following

$$X_i = \delta_i + \mu_1 + \varepsilon_{X,i} \quad \text{and} \quad Y_i = \delta_i + \mu_2 + \varepsilon_{Y,i} \qquad i = 1, \ldots, n. \tag{5.12}$$

Observe that for each pairing we allow for an "individual effect" δ_i. This individual effect allows for the individual weight of a calf, inate running ability of a person (regardless of the altitude) etc.

It is important to understand the assumptions behind matched data and similarities and dissimilarities to independent sample data. We list the main points below.

We first state some assumptions that are similar to those for the independent sample t-test.

(i) The pairings $\{X_i, Y_i\}_{i=1}^n$ are independent over the individuals.

(i) The data is jointly normal.

However, there are differences between matched and independent two sample data that are important to emphasis:

(i) For a given individual i, the errors $\varepsilon_{X,i}$ and $\varepsilon_{Y,j}$ can be correlated. In other words, $\text{cov}(\varepsilon_{X,i}, \varepsilon_{Y,i}) \neq 0$. However we assume that the pairs $\{\varepsilon_{X,i}, \varepsilon_{Y,i}\}_{i=1}^n$ are identically distributed covariance is the same over all i i.e. $\text{cov}(\varepsilon_{X,i}, \varepsilon_{Y,i}) = \rho$, $\text{var}[X_i] = \sigma_X^2$ and $\text{var}[Y_i] = \sigma_Y^2$.

(ii) Unlike the independent sample t-test each pair can have its own mean. Observe the common mean δ_i for each pair.

Remark. *Allowing for dependence in the errors* $\text{cov}(\varepsilon_{X,i}, \varepsilon_{Y,i})$ *or a mean specific to each individual,* δ_i *depends on the data. For example, in terms of the calf data, my personal preference is to model each calves weight with an individual mean. Where this mean is specific to the, say this calves family. Either way we require a term that models the clear linear dependence seen in Figure 5.2. It does not matter which way this is done.*

Once again we want to compare the means between two different distributions and the hypothesis of interest is

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_0 : \mu_1 - \mu_2 \neq 0.$$

As in the independent two sample t-test the focus is on the difference of the sample means that is

$$\bar{D} = \bar{X} - \bar{Y}.$$

But the variance of $\bar{X} - \bar{Y}$ is different to the variance in the independent two sample t-test. This can make a fundamental difference to the analysis we should do. We discuss this issue in the following section, but first we focus on the variance of $\bar{D}$:

$$\begin{aligned}
\text{var}[\bar{D}] &= \text{var}[n^{-1}\sum_{i=1}^n X_i - n^{-1}\sum_{i=1}^n Y_i] \\
&= \text{var}[n^{-1}\sum_{i=1}^n (X_i - Y_i)] \text{ now we use the independence assumption} \\
&= \frac{1}{n^2}\sum_{i=1}^n \text{var}(X_i - Y_i) = \frac{1}{n^2}\sum_{i=1}^n [\text{var}(X_i) + 2\text{cov}(X_i, Y_i) + \text{var}(Y_i)] \\
&= \frac{1}{n}\left(\sigma_X^2 - 2\rho + \sigma_Y^2\right).
\end{aligned}$$

Thus any method we use based on the difference $\bar{X} - \bar{Y}$ should estimate $\text{var}[\bar{D}]$ correctly. If we do not do this, the standard errors will be incorrect leading to the misleading statistical conclusions. We show in the

following section, that the pooled sample variance described in Section 5.1.3 is estimating a completely different quantity. The implication of this is that applying the independent sample t-test to matched data can lead to spurious results.

5.3.3 Why the independent two sample t-test should not be used for matched data

The problem with applying the independent sample t-test to matched data is that the pooled variance cannot consistently (correctly) estimate the true variance. From the above the standard error of $\bar{X} - \bar{Y}$ is

$$\sqrt{\frac{1}{n}\left(\sigma_X^2 - 2\rho + \sigma_Y^2\right)}. \tag{5.13}$$

But the estimated standard error which ignores the pairing between the data is (with $n = m$) leads to the estimator:

$$\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{n}\right)} \quad = \quad \sqrt{\frac{s_X^2 + s_Y^2}{2}\left(\frac{2}{n}\right)} = \sqrt{\frac{1}{n}\left(s_X^2 + s_Y^2\right)}. \tag{5.14}$$

- Comparing (5.13) and (5.14) we observe that the covariance -2ρ between the pairings is not being estimated. Not estimating this covariance means the standard error is not being correctly estimated. Often the covariance will be positive (see the calf plot). If this is the case than the true standard error is such that

  $$\sqrt{\frac{1}{n}\left(\sigma_X^2 - 2\rho + \sigma_Y^2\right)} < \sqrt{\frac{1}{n}\left(\sigma_X^2 + \sigma_Y^2\right)},$$

  thus the standard error has been overestimated.

- But another (related error) has incurred. It "looks" like s_X^2 and s_Y^2 are estimating the variance σ_X^2 and σ_X^2. Thus turns out not to be the case. We explain why below.

To understand what is happening we recall that

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\widehat{\varepsilon}_{X,i}^2.$$

where $\widehat{\varepsilon}_{X,i}$ are the "estimated residuals"

$$\begin{aligned}
\widehat{\varepsilon}_{X,i} &= X_i - \bar{X} = \delta_i + \mu_1 + \varepsilon_{X,i} - \frac{1}{n}\sum_{j=1}^{n}\left(\delta_j + \mu_1 + \varepsilon_{X,j}\right) \\
&= \varepsilon_{X,i} - \frac{1}{n}\sum_{j=1}^{n}\varepsilon_{X,j} + \delta_i - \frac{1}{n}\sum_{j=1}^{n}\delta_j \\
&\approx \varepsilon_{X,i} + \delta_i - \frac{1}{n}\sum_{j=1}^{n}\delta_j,
\end{aligned}$$

since for "large" n we have $\frac{1}{n}\sum_{j=1}^{n}\varepsilon_{X,j} \approx \mathrm{E}[\varepsilon_{X,j}] = 0$. But observe that that "mean term" $\delta_i - \frac{1}{n}\sum_{j=1}^{n}\delta_j$ remains (which we do not want). This gives

$$\widehat{\varepsilon}^2_{X,i} \approx \varepsilon^2_{X,i} - 2\varepsilon_{X,i}\left(\delta_i - \frac{1}{n}\sum_{j=1}^{n}\delta_j\right) + \left(\delta_i - \frac{1}{n}\sum_{j=1}^{n}\delta_j\right)^2.$$

Therefore for large n

$$s^2_X = \frac{1}{n-1}\sum_{i=1}^{n}\widehat{\varepsilon}^2_{X,i} \approx \frac{1}{n-1}\sum_{i=1}^{n}\varepsilon^2_{X,i} + \frac{2}{n-1}\sum_{i=1}^{n}\varepsilon_i\left(\delta_i - \frac{1}{n}\sum_{j=1}^{n}\delta_j\right) + \frac{1}{n-1}\sum_{i=1}^{n}\left(\delta_i - \frac{1}{n}\sum_{j=1}^{n}\delta_j\right)^2.$$

This is a little beyond this course one can show that the above is approximately

$$s^2_X = \frac{1}{n-1}\sum_{i=1}^{n}\widehat{\varepsilon}^2_{X,i} \approx \sigma^2 + \frac{1}{n-1}\sum_{i=1}^{n}\left(\delta_i - \frac{1}{n}\sum_{j=1}^{n}\delta_j\right)^2.$$

The main point is that by not removing the correct mean, the variance estimator s^2_X completely misses its mark and will over estimate it. Leading to a "sample standard deviation" which is far larger than it should be. Subsequently, this will lead to a standard error that is "too large" making it difficult to reject the null. We observe exactly this effect in the data analysis of the calf data in Section 5.3.5.

To summarize, the variance estimator in the independent two sample t-test does not correctly estimate the true variance of $\bar{X} - \bar{Y}$ because

(i) The variance estimator does not estimate the covariance $\rho = \mathrm{cov}(\mathrm{cov}(X_i, Y_i)$.

(ii) It does not estimate σ^2_X and σ^2_Y. Indeed it tends to over estimate it.

In conclusion the z-transform

$$Z = \frac{n^{1/2}(\bar{X} - \bar{Y})}{\sqrt{\left(s^2_X + s^2_Y\right)}}$$

will not be standard normal or t-distribution under the null hypothesis because the standardisation is completely wrong.

The fix to this issue is very simple, we briefly outline it below.

5.3.4 The matched paired t-test

Clearly when there is a dependence between the pairs, $\{X_i\}$ and $\{Y_i\}$ should not be treated as independent of each other. Instead we overcome the common individual mean and estimation of the covariance by simply taking differences

$$\begin{aligned} D_i = X_i - Y_i &= \delta_i + \mu_1 + \varepsilon_{X,i} - \delta_i - \mu_2 - \varepsilon_{Y,i} \\ &= \underbrace{\mu_1 - \mu_2}_{\mu_d} + \underbrace{\varepsilon_{X,i} - \varepsilon_{Y,i}}_{=\epsilon_i}, \end{aligned}$$

observe we relable the residual $\epsilon_i = \varepsilon_{X,i} - \varepsilon_{Y,i}$ and the mean as $\mu_d = \mu_1 - \mu_2$. Thus we have transformed two samples into one sample. It is simple to show that the sample mean of $\{D_i\}$ is

$$\bar{D} = \bar{X} - \bar{Y},$$

and the sample variance

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})$$

consistently estimates $\sigma_X^2 - 2\rho + \sigma_Y^2$. This allows us to apply the one-sample t-test to the differences:

$$H_0 : \mu_d = 0 \text{ and } H_0 : \mu_d \neq 0.$$

Conclusion If there is any doubt that there may be matching in the data always use a matched t-test. Even if there is no matching in the data, the variance estimator using a matched t-test will still consistently estimate the variance. And the matched t-test will give always valid results even if the data has no matching.

5.3.5 Application to data

The calf data

We conclude this section by analysing the calf data presented at the start of the section. Because I was too lazy to upload the data into R I did it in JMP.

In Figure 5.3 the output in JMP using the independent two sample t-test is given. Recall that due to the dependence between the variables this is a completely inappropriate test. We observe that the standard error is $s.e. = 1.74$ which results in a t-value of $t = -2.17 = -3.78/1.74$). This gives a p-value which is relative small (3%) but not overwhelmingly so. Keep in mind the standard error of 1.74.

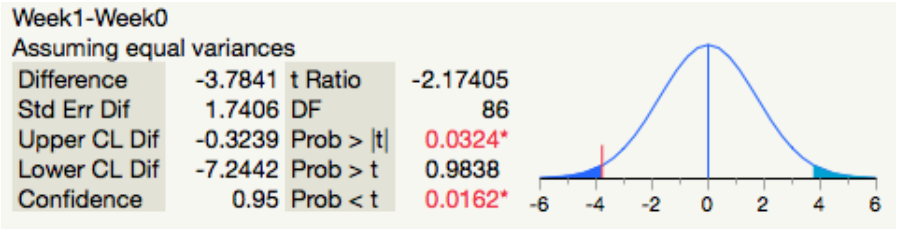


Figure 5.3: The output when applying independent two sample t-test to the calves data (the incorrect test).

Next we apply the matched paired t-test for the same data. Given the linear dependence we observe between the variables, this appears to be the appropriate test. The results of the test are given in Figure 5.4. The

differences in the sample means is -3.78 (which is the same as for the independent two sample t-test, and as expected). But we observe a dramatic decrease in the standard error, which when estimated correctly is $s.e. = 0.63$ (compare this with the standard error for the independent two sample t-test which gives a standard error of 1.74). The resulting t-value is $t = -3.78/0.63 = -5.94$. The t-value is far larger than the t-value corresponding to the independent sample t-test, making the results of the test more significant. Indeed,the p-value for the matched t-test is less than 0.01% and is highly significant.

Week1	89.4318	t-Ratio	-5.94636
Week0	93.2159	DF	43
Mean Difference	-3.7841	Prob > |t|	<.0001*
Std Error	0.63637	Prob > t	1.0000
Upper 95%	-2.5007	Prob < t	<.0001*
Lower 95%	-5.0675		
N	44		
Correlation	0.87004		

Figure 5.4: The output when applying independent two sample t-test to the calves data (the correct test).

The hens and eggs shells data

white-red
Assuming equal variances

Difference	-1.4279	t Ratio	-2.53179
Std Err Dif	0.5640	DF	200
Upper CL Dif	-0.3158	Prob > |t|	0.0121*
Lower CL Dif	-2.5400	Prob > t	0.9939
Confidence	0.95	Prob < t	0.0061*

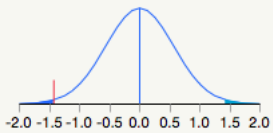


Figure 5.5: The independent two sample t-test applied to light treatment and strength of egg shell data.

6 ANOVA

6.1 Post-hoc analysis

Under the null

$$\frac{\max |\bar{X}_i - \bar{X}_j|}{s_p\sqrt{2/n}}$$

6.1.1 Studentised range distribution

https://en.wikipedia.org/wiki/Studentized_range_distribution

6.2 Proof of one-way ANOVA

We recall that the ANOVA statistics is (in the case of three groups)

Theorem 6.1

Suppose $\{\{X_{s,i}\}_{s=1}^{n_2}\}_{s=1}^{K}$ are iid normal random variables with mean μ and variance σ^2. We define the F-statistic

$$F = \frac{\sum_{s=1}^{M} n_s(\bar{X}_s - \bar{X})^2}{\sum_{s=1}^{K}\sum_{i=1}^{n_s}(X_{s,i} - \bar{X}_s)^2},$$

where $\bar{X}_s$ denotes the group means and $\bar{X}$ the global mean. Then $F \sim F_{s-1,n-s}$.

We prove the results for three groups, with $\{X_i\}_{i=1}^{n_1}$, $\{Y_i\}_{i=1}^{n_2}$ and $\{Z_i\}_{i=1}^{n_3}$ with group sample means $\bar{X}$, $\bar{Y}$ and $\bar{Z}$ respectively. Precisely, we prove that

$$\frac{U}{V} = \frac{n_1(\bar{X} - \bar{W})^2 + n_1(\bar{Y} - \bar{W})^2 + n_3(\bar{Z} - \bar{W})^2}{\sum_{i=1}^{n_1}(X_{1,i} - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_{1,i} - \bar{Y})^2 + \sum_{i=1}^{n_3}(Z_{1,i} - \bar{Z})^2} \sim F_{2,n_1+n_2+n_3-3},$$

where

$$\bar{W} = \frac{1}{n_1 + n_2 + n_3}\left(\sum_{i=1}^{n_1} X_{s,i} + \sum_{i=1}^{n_2} Y_{s,i} + \sum_{i=1}^{n_3} Z_{s,i}\right).$$

The proof uses similar construction to the proof of Theorem 2.3. First by showing the numerator are denominator are independent and then showing that the numerator and denominator follow a chi-square distribution. In line with the proof Theorem 2.3 we prove the result using orthogonal projections.

We start by defining the $(n_1 + n_2 + n_3)$-dimensional vector

$$\underline{W} = \begin{pmatrix} \underline{X} \\ \underline{Y} \\ \underline{Z} \end{pmatrix}$$

Under the stated the assumptions we have $\underline{W} \sim N(\mu\underline{1}, \sigma^2 I_{n_1+n_2+n_2})$. We will write $\underline{W}$ in terms of an appropriate orthonormal basis to prove the result.

Let $\{\underline{\widetilde{e}}_i\}_{i=1}^{n_1}$ denote a basis of $\mathbb{R}^{n_1}$ where $\underline{\widetilde{e}}_1' = n_1^{-1/2}(1, \ldots, 1)$, $\{\underline{\widetilde{f}}_i\}_{i=1}^{n_2}$ denote a basis of $\mathbb{R}^{n_2}$ where $\underline{\widetilde{f}}_1' = n_2^{-1/2}(1, \ldots, 1)$ and $\{\underline{\widetilde{g}}_i\}_{i=1}^{n_3}$ denote a basis of $\mathbb{R}^{n_3}$ where $\underline{\widetilde{g}}_1' = n_3^{-1/2}(1, \ldots, 1)$. For all $1 \leq i \leq n_s$ we define the $(n_1 + n_2 + n_3)$-dimensional vectors

$$\underline{e}_i = \begin{pmatrix} \underline{\widetilde{e}}_i \\ \underline{0} \\ \underline{0} \end{pmatrix}, \quad \underline{f}_i = \begin{pmatrix} \underline{0} \\ \underline{\widetilde{f}}_i \\ \underline{0} \end{pmatrix}, \quad \underline{g}_i = \begin{pmatrix} \underline{0} \\ \underline{0} \\ \underline{\widetilde{g}}_i \end{pmatrix}$$

By construction $\{\{\underline{e}_i\}_{i=1}^{n_1}, \{\underline{f}_i\}_{i=1}^{n_2}, \{\underline{g}_i\}_{i=1}^{n_3}\}$ forms an orthogonal basis of $\mathbb{R}^{n_1+n_2+n_3}$. Thus we can represent $\underline{W}$ as

$$\underline{W} = \langle \underline{e}_1, \underline{W}\rangle \underline{e}_1 + \langle \underline{f}_1, \underline{W}\rangle \underline{f}_1 + \langle \underline{g}_1, \underline{W}\rangle \underline{g}_1 + \sum_{i=2}^{n_1} \langle \underline{e}_i, \underline{W}\rangle \underline{e}_i + \sum_{i=2}^{n_2} \langle \underline{f}_i, \underline{W}\rangle \underline{f}_i + \sum_{i=2}^{n_3} \langle \underline{g}_i, \underline{W}\rangle \underline{g}_i.$$

Using that $\underline{W} \sim N((\mu_1\underline{1}_{n_1}, \mu_2\underline{1}_{n_2}, \mu_1\underline{1}_{n_2})', \sigma^2 I_{n_1+n_2+n_2})$, and the orthonormality of the basis we can show that $\{\langle \underline{e}_i, \underline{W}\rangle\}$, $\{\langle \underline{f}_i, \underline{W}\rangle\}$ and $\{\langle \underline{g}_i, \underline{W}\rangle\}$ are iid normal random variables with variance σ^2, where for $i > 1$, $\mathrm{E}[\langle \underline{e}_i, \underline{W}\rangle] = 0$ (since $\langle \underline{e}_1, \underline{e}_i\rangle = \sum_{s=1}^{n_1} 1 \times e_{i,s} = 0$) further for $i = 1$ we have $\langle \underline{e}_1, \underline{W}\rangle = \sqrt{n_1}\bar{X}$, $\langle \underline{f}_1, \underline{W}\rangle = \sqrt{n_2}\bar{Y}$ and $\langle \underline{g}_1, \underline{W}\rangle = \sqrt{n_3}\bar{Z}$. Thus under the assumption that the group means are all different the vector $\underline{W}$ can be rewritten as

$$\begin{pmatrix} \langle \underline{e}_1, \underline{W}\rangle \\ \langle \underline{f}_1, \underline{W}\rangle \\ \langle \underline{g}_1, \underline{W}\rangle \\ \langle \underline{e}_2, \underline{W}\rangle \\ \vdots \\ \langle \underline{g}_{n_3}, \underline{W}\rangle \end{pmatrix} \sim MVN_{n_1+n_2+n+3}\left(\begin{bmatrix} \sqrt{n_1}\mu_1 \\ \sqrt{n_2}\mu_2 \\ \sqrt{n_3}\mu_3 \\ \underline{0} \end{bmatrix}, \sigma^2 I_{n_1+n_2+n_2}\right). \tag{6.1}$$

Observe that all the information about the population means in each group are encoded in the first three entries of this transformed vector. Therefore, by using the same arguments to those in Theorem 2.3 we have

$$\begin{aligned}\underline{W} - \langle \underline{e}_1, \underline{W}\rangle \underline{e}_1 + \langle \underline{f}_1, \underline{W}\rangle \underline{f}_1 + \langle \underline{g}_1, \underline{W}\rangle \underline{g}_1 &= \sum_{i=2}^{n_1} \langle \underline{e}_i, \underline{W}\rangle \underline{e}_i + \sum_{i=2}^{n_2} \langle \underline{f}_i, \underline{W}\rangle \underline{f}_i + \sum_{i=2}^{n_3} \langle \underline{g}_i, \underline{W}\rangle \underline{g}_i \\ \Rightarrow \begin{pmatrix} \underline{X} - \sqrt{n_1}\bar{X}\underline{e}_1 \\ \underline{Y} - \sqrt{n_2}\bar{Y}\underline{f}_1 \\ \underline{Z} - \sqrt{n_3}\bar{Z}\underline{f}_1 \end{pmatrix} &= \sum_{i=2}^{n_1} \langle \underline{e}_i, \underline{W}\rangle \underline{e}_i + \sum_{i=2}^{n_2} \langle \underline{f}_i, \underline{W}\rangle \underline{f}_i + \sum_{i=2}^{n_3} \langle \underline{g}_i, \underline{W}\rangle \underline{g}_i.\end{aligned}$$

The Euclidean distance of the above vector together with Parseval's equality in (1.2) gives

$$\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2 + \sum_{i=1}^{n_2}(Z_i - \bar{Z})^2 = \sum_{i=2}^{n_1} \langle \underline{e}_i, \underline{W}\rangle^2 + \sum_{i=2}^{n_2} \langle \underline{f}_i, \underline{W}\rangle^2 + \sum_{i=2}^{n_3} \langle \underline{g}_i, \underline{W}\rangle^2. \tag{6.2}$$

Thus

$$V = \sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2 + \sum_{i=1}^{n_2}(Z_i - \bar{Z})^2 \sim \chi^2_{n_1+n_2+n_3-3}.$$

Furthermore, thanks to the orthgonal decomposition, V is independent of

$$\langle \underline{e}_1, \underline{W}\rangle \underline{e}_1 + \langle \underline{f}_1, \underline{W}\rangle \underline{f}_1 + \langle \underline{g}_1, \underline{W}\rangle \underline{g}_1 = \sqrt{n_1}\bar{X}\underline{e}_1 + \sqrt{n_2}\bar{Y}\underline{f}_1 + \sqrt{n_3}\bar{Z}\underline{g}_1.$$

This proves that the numerator, U, and denominator, V, are independent.

The final part of the proof involves rewriting $\sqrt{n_1}\bar{X}\underline{e}_1 + \sqrt{n_2}\bar{Y}\underline{f}_1 + \sqrt{n_3}\bar{Z}\underline{g}_1$ in terms of a different orthnormal basis (which includes the global average $\bar{W}$ as a coefficient). Define the vector $\underline{h}_1 = (n_1 + n_2 + n_3)^{-1/2}(1, 1, \ldots, 1)$, and the orthonormal vectors $\underline{h}_2$ and $\underline{h}_3$ which form a basis for the space spanned by $(\underline{e}_1, \underline{f}_1, \underline{g}_1)$. Thus

$$\sqrt{n_1}\bar{X}\underline{e}_1 + \sqrt{n_2}\bar{Y}\underline{f}_1 + \sqrt{n_3}\bar{Z}\underline{g}_1 = \langle \underline{h}_1, \underline{W}\rangle \underline{h}_1 + \langle \underline{h}_2, \underline{W}\rangle \underline{h}_2 + \langle \underline{h}_3, \underline{W}\rangle \underline{h}_3,$$

where $(n_1 + n_2 + n_3)^{1/2}\bar{W}$. Therefore

$$\sqrt{n_1}\bar{X}\underline{e}_1 + \sqrt{n_2}\bar{Y}\underline{f}_1 + \sqrt{n_3}\bar{Z}\underline{g}_1 - \langle \underline{h}_1, \underline{W}\rangle \underline{h}_1 = \langle \underline{h}_2, \underline{W}\rangle \underline{h}_2 + \langle \underline{h}_3, \underline{W}\rangle \underline{h}_3.$$

Since $\underline{h}_1$ and $\underline{h}_2$ are orthonormal transformation vectors we have $\mathrm{var}(\langle \underline{h}_1, \underline{W}\rangle) = \sigma^2$ and $\langle \underline{h}_2, \underline{W}\rangle) = \sigma^2$. Under the null hypothesis that the population means in all three groups are the same. This gives

$$\begin{aligned}\mathrm{E}(\langle \underline{h}_2, \underline{W}\rangle) &= \sum_{i=1}^{n_1} h_{2,i}\mathrm{E}(X_i) + \sum_{i=1}^{n_2} h_{2,n_1+i}\mathrm{E}(Y_i) + \sum_{i=1}^{n_3} h_{2,n_1+n_2+i}\mathrm{E}(Z_i) \\ &= \mu \sum_{i=1}^{n_1+n_2+n_3} h_{2,i} = \mu\langle \underline{h}_1, \underline{h}_2\rangle = 0.\end{aligned}$$

By the same argument $\mathrm{E}(\langle \underline{h}_3, \underline{W}\rangle) = 0$. Therefore under the null all the group means are the same we have

$$\begin{pmatrix} \langle \underline{h}_1, \underline{W}\rangle \\ \langle \underline{h}_2, \underline{W}\rangle \\ \langle \underline{h}_3, \underline{W}\rangle \\ \langle \underline{e}_2, \underline{W}\rangle \\ \vdots \\ \langle \underline{g}_{n_3}, \underline{W}\rangle \end{pmatrix} \sim MVN_{n_1+n_2+n+3}\left(\begin{bmatrix} \sqrt{m}\mu \\ \underline{0} \end{bmatrix}, \sigma^2 I_{n_1+n_2+n_2}\right) \tag{6.3}$$

where $m = n_1 + n_2 + n_3$. Thus, by using (1.2), and noting that $\underline{e}_1, \underline{f}_1$ and $\underline{g}_1$ are n_1, n_2 and n_3-dimension vectors containing the same quantity (under the null) we have

$$\begin{aligned} & \left\| \sqrt{n_1}\bar{X}\underline{e}_1 + \sqrt{n_2}\bar{Y}\underline{f}_1 + \sqrt{n_3}\bar{Z}\underline{g}_1 - \langle \underline{h}_1, \underline{W}\rangle \underline{h}_1 \right\|_2^2 \\ = \;& n_1(\bar{X} - \bar{W})^2 + n_2(\bar{Y} - \bar{W})^2 + n_3(\bar{Z} - \bar{W})^2 = \langle \underline{h}_2, \underline{W}\rangle^2 + \langle \underline{h}_3, \underline{W}\rangle^2 \sim \chi^2_2, \end{aligned}$$

where $\|\underline{x}\|_2^2 = \sum_{i=1}^d x_i^2$. Altogether, by using (6.2) and the above we obtain the result, in the case $K = 3$. The same proof can be generalized to the case $K > 3$.

In summary $\underline{W} \sim \mathcal{N}(\mu\underline{1}, \sigma^2 I_{n_1+n_2+n_3})$. However, we have shown that, equivalently, $\underline{W}$ can be written in terms of orthnormal vectors whose coefficients are uncorrelated and contain information that can be used to test for differences between the group population means. For precisely we have the decomposition

$$\underline{W} = (n_1 + n_2 + n_3)^{1/2}\bar{W}\underline{h}_1 + \underbrace{\langle \underline{h}_2, \underline{W}\rangle \underline{h}_2 + \langle \underline{h}_3, \underline{W}\rangle \underline{h}_3}_{\sqrt{n_1}\bar{X}\underline{e}_1 + \sqrt{n_2}\bar{Y}\underline{f}_1 + \sqrt{n_3}\bar{Z}\underline{g}_1 - \langle \underline{h}_1, \underline{W}\rangle \underline{h}_1} + \sum_{i=2}^{n_1} \langle \underline{e}_i, \underline{W}\rangle \underline{e}_i + \sum_{i=2}^{n_2} \langle \underline{f}_i, \underline{W}\rangle \underline{f}_i + \sum_{i=2}^{n_3} \langle \underline{g}_i, \underline{W}\rangle \underline{g}_i.$$

This decomposes the $\underline{W}$ into the global mean vector into three independent blocks:

- The global mean vector.
- Differences between each group sample mean and the global mean
- The residuals after removing the group mean.

From this expansion we immediately obtain the classical decomposition of SST

$$\left\| \underline{W} - (n_1 + n_2 + n_3)^{1/2}\bar{W}\underline{h}_1 \right\|_2^2 = \|\langle \underline{h}_2, \underline{W}\rangle \underline{h}_2 + \langle \underline{h}_3, \underline{W}\rangle \underline{h}_3\|_2^2 + \left\| \sum_{i=2}^{n_1} \langle \underline{e}_i, \underline{W}\rangle \underline{e}_i + \sum_{i=2}^{n_2} \langle \underline{f}_i, \underline{W}\rangle \underline{f}_i + \sum_{i=2}^{n_3} \langle \underline{g}_i, \underline{W}\rangle \underline{g}_i \right\|_2^2.$$

Since

$$\begin{aligned} SST &= \left\| \underline{W} - (n_1 + n_2 + n_3)^{1/2}\bar{W}\underline{h}_1 \right\|_2^2 = \sum_{i=1}^{n_1} (X_i - \bar{W})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{W})^2 + \sum_{i=1}^{n_2} (Z_i - \bar{W})^2 \\ SSB &= \|\langle \underline{h}_2, \underline{W}\rangle \underline{h}_2 + \langle \underline{h}_3, \underline{W}\rangle \underline{h}_3\|_2^2 = n_1(\bar{X} - \bar{W})^2 + n_2(\bar{Y} - \bar{W})^2 + n_3(\bar{Z} - \bar{W})^2 \\ SSW &= \left\| \sum_{i=2}^{n_1} \langle \underline{e}_i, \underline{W}\rangle \underline{e}_i + \sum_{i=2}^{n_2} \langle \underline{f}_i, \underline{W}\rangle \underline{f}_i + \sum_{i=2}^{n_3} \langle \underline{g}_i, \underline{W}\rangle \underline{g}_i \right\|_2^2 = \sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 + \sum_{i=1}^{n_2} (Z_i - \bar{Z})^2 \end{aligned}$$

The above expansion yields the well known result $SST = SSB + SSW$. □