Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 11 (MWF) The rationale behind confidence intervals

Suhasini Subba Rao

Review of previous lecture

- So far we have used the normal distribution to calculate probabilities.
- We have also discussed the QQplot, which is a method for checking whether the data actually comes from a normal distribution (once we know the data is normal then we can use the normal distribution to calculate percentiles etc).
- However, the percentile calculations were based on the assumption that the mean of the population is known.
- Typically, this will not be the case. Our *aim* is to use the data to locate the mean.
- The first thing we do when given the data is to take the sample mean of the observations, this is our best estimator of the population mean.

- The problem is that the sample mean, usually denoted $\bar{X}$, is unlikely to be exactly equal to the population mean μ.

- Instead we can construct an interval about the mean and hope the mean lies in there.

  How wide should this interval be?

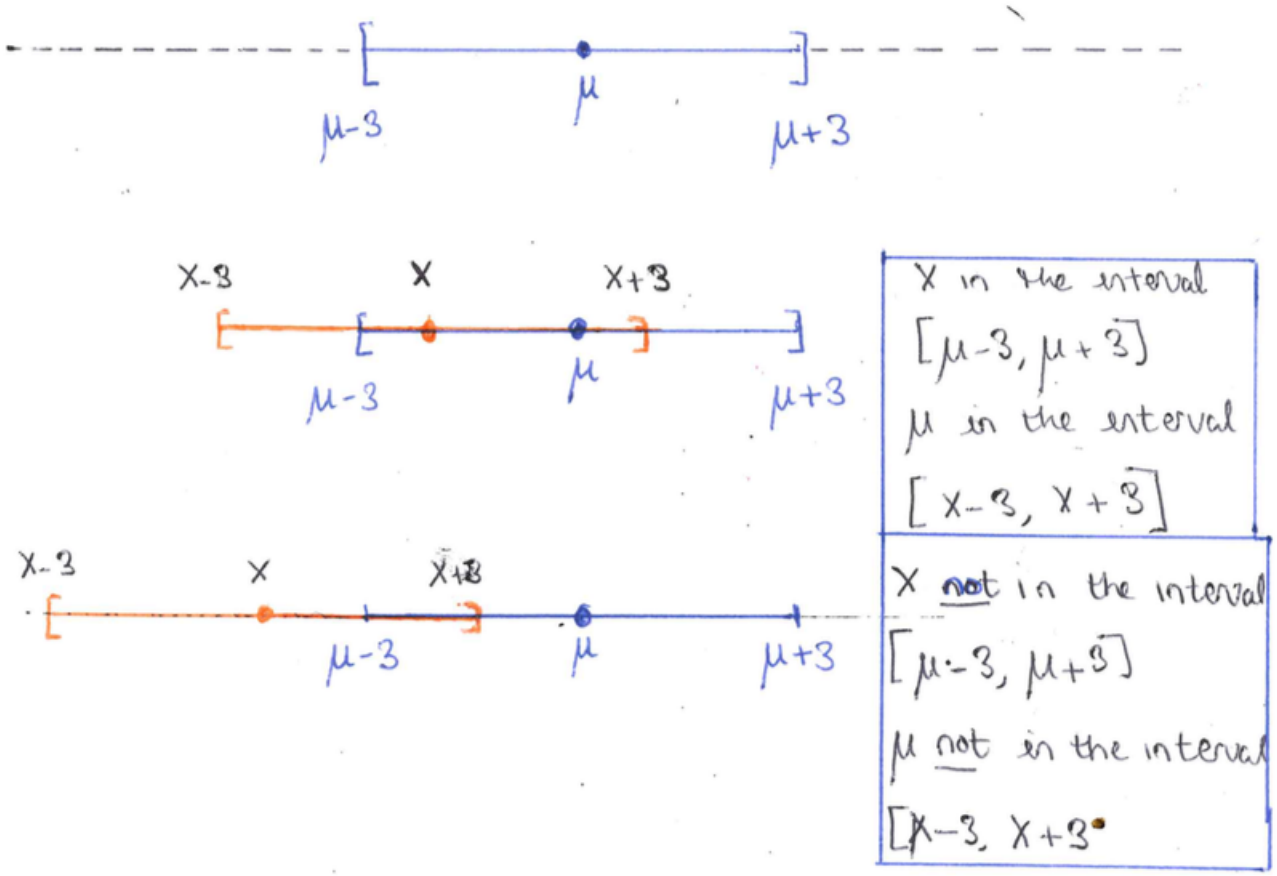
- If an interval is too wide, it won't be very informative. Example: We know that human adult weights can be anything from 70 pounds going up to 400 pounds. If I want to know the mean human weight, I can be 100% confident that the population mean is in the interval [70,400]. However this interval is $uninformative$ about the location of the population mean!

- If the mean is too narrow, then we can miss the mean. This interval is $unlikely\ to\ contain\ the\ population\ mean$

- We need to find a balance.

- We should be able to expression some level of confidence that the mean lies within the stated interval.

Intervals: inside and outside



Study the plot above:

- Suppose there is a 95% chance an observation lies with 3 cm of the mean (μ). This is the same as the proportion of times we draw an X which is within the blue lines of the plot above.

- If the observation (number) X lies in the interval $[\mu - 3, \mu + 3]$. Then the mean μ lies in the interval $[X - 3, X + 3]$. Look at the middle plot.

- If the observation (number) X does not lie in the interval $[\mu - 3, \mu + 3]$. Then the mean μ does not lies in the interval $[X - 3, X + 3]$. Look at the bottom plot.

- Since, there is a 95% chance X lies in the interval $[\mu - 3, \mu + 3]$ using the above argument, there is a 95% chance that the mean μ lies in the interval $[X - 3, X + 3]$.

The confidence interval

- The interval $[X-3, X+3]$ can be used as a way of locating the mean μ.
- Suppose we draw an X and $X = 33$, and construct the interval $[33-3, 33+3]$.
- What can we say about the mean μ being inside the interval $[30, 36]$?
- Look back at the picture, observe that either the mean μ is in the interval $[30, 36]$ or it is not in the interval. We cannot assign a formal probability to the interval.
- Once we have collected the data, we cannot say there is a 95% chance the mean lies in $[30, 36]$. But this interval is still informative about the mean.

- To avoid using the word "probability", we say that with 95% confidence the mean lies in the interval $[30, 36]$.

- We call $[30, 36]$ a 95% confidence interval for the mean μ.

The distribution of X

- The above definition is a very general definition of a confidence interval. It does not require that the distribution of X is normal.

- However, it does require us to know the distribution of X such that we can say that there is an 95% chance X lies in $[\mu - 3, \mu + 3]$.

- Usually the distribution of X is unknown. Recently, subsampling methods have been developed to estimate the distribution of X.

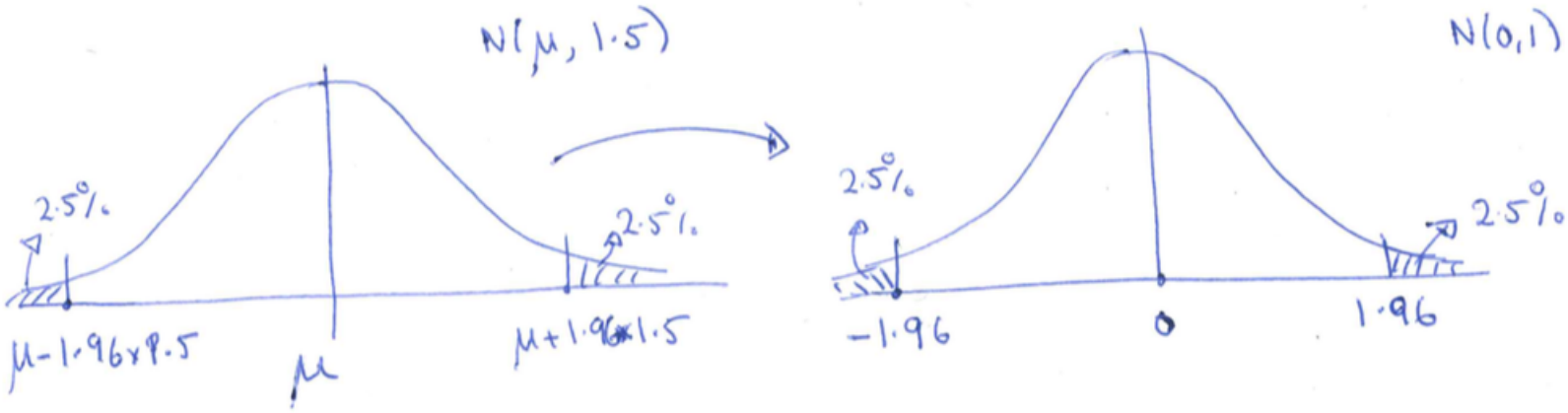
- However, in the remainder of course we will assume that X is normal (we justify this in Lecture 12). We will use subsampling to justify this result.

Example: assuming normality

- **Question** Suppose $X \sim N(\mu, \sigma = 1.5)$ (mean μ and standard deviation 1.5).

- Find the interval centered about the mean, where 95% of the population should lie.

- We did this in Lecture 9:



- Based on normality of the observation X, there is a 95% chance X lies in the interval $[\mu - 1.96 \times 1.5, \mu + 1.96 \times 1.5]$[1]

- Using the argument on the previous slide. Based on normality of X, there is a 95% chance the mean μ lies in the interval

  $[X - 1.96 \times 1.5, X + 1.96 \times 1.5]$.

- For a given number $X = \bar{x}$,

  $[\bar{x} - 1.96 \times 1.5, \bar{x} + 1.96 \times 1.5]$ is a 95% confidence interval for the mean μ.

[1] For a given percentage level, the interval which is **centered** about the mean will always be the smallest (if it is normally distributed). If it is not centered about the mean it will be longer.

General definition: The 95% Confidence interval

- Suppose $X \sim N(\mu, \sigma)$ (where σ is assumed known), then $P(\mu - 1.96 \times \sigma \leq X \leq \mu + 1.96 \times \sigma) = 0.95$. Given a value for X, the 95% confidence interval for the mean is

  $[X - 1.96 \times \sigma, X + 1.96 \times \sigma]$.

- It basically says that if we draw X one thousand times (the number of times does not matter), and for each X drawn, we construct an interval $[X - 1.96 \times \sigma \leq \mu \leq X + 1.96 \times \sigma]$, then about 950 of these intervals will contain the mean μ.

- If the mean μ is unknown we will not know which of the 1000 intervals contain the mean μ. But we do know that about 950 of the intervals will contain it. In this way the interval is informative about the mean.

A 90% Confidence interval for the mean

- Using the above argument, since $P(\mu - 1.64 \times \sigma \leq X \leq \mu + 1.64 \times \sigma). = 0.90$,

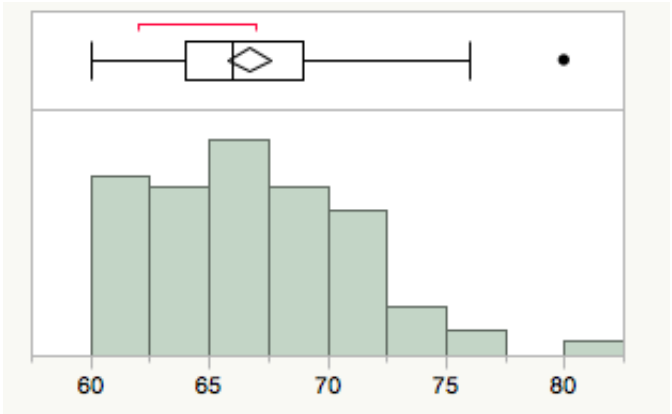
- $[X - 1.64 \times \sigma, X + 1.64 \times \sigma]$ is a 90% confidence interval for the mean μ.

- Remember Once we have number for $X = x$, we cannot associate a probability to

  $[x - 1.64 \times \sigma, x + 1.64 \times \sigma]$.

  Either the mean is in the interval or it is not in the above interval.

The heights of students (sample 1)



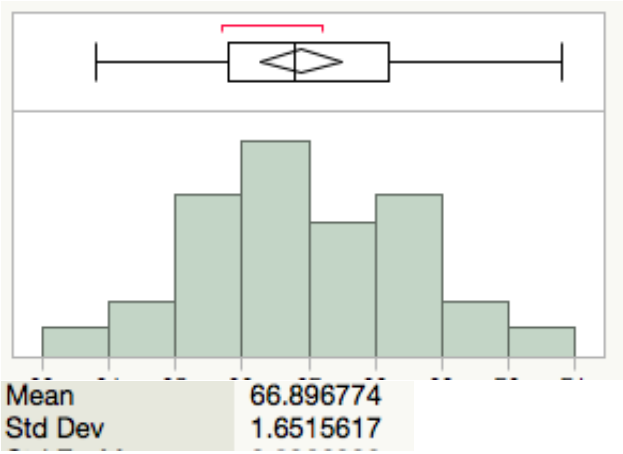
On the left is the distribution of heights. One height is drawn from this distribution, say 62 inches. Using the above prescription a 95% confidence interval for the mean is
$[62 \pm 1.96 \times 4.3] = [53.5, 70.5]$

First, we observe that this is a very wide interval. Further, we use the normal tables to get the 1.96 and the 95% confidence. However, the distribution of heights does not look normal. This means the 95% confidence we assign to the interval is incorrect.

The heights of students (sample 5)

Suppose that 5 heights are drawn, 68, 68, 69, 70, 70 inches. The average of this is 69 inches.

The histogram below are the heights of sample means (averages based on 5 students); this was done by taking several SRS of size 5 and making the plot (in practice this cannot be done).



The average of 69 inches is drawn from this distribution. The standard derivation/standard error is far smaller 1.65 inches. Using the above prescription a 95% confidence interval for the mean is
$[69 \pm 1.96 \times 1.65] = [65.8, 72.2]$.

- This interval is narrower, and gives greater precision in locating the mean.
- Moreover, the distribution looks more normal (see lecture 10 and the QQplots to check this).
- This the 95% confidence that we assign the interval is correct. We really can say we have 95% confidence the mean lies within this interval.

Summary and future lectures

- The examples, above, illustrated some important issues.

- The interval is very wide (and inaccurate) based on just one observation. We need to the class average in order to obtain a narrower interval.

- We require normality of the estimator in order to construct confidence intervals based on the normal z-values.

- Many data sets will not be normal, this includes the combined heights of males and females (recall this is bimodal). However, regardless of the original distribution, the sample mean will have a distribution that is closer to normal.

In lecture 12 we address the issues raised above.