Data Analysis and Statistical Methods
Statistics 651

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

Lecture 16 (MWF) One sided and two sided tests, rejection regions and p-values

Suhasini Subba Rao

Review of previous lecture

- In a statistical test we have an idea (conjecture).

  To prove the conjecture we need to disprove its complement.

- We always state the idea we wish to test as the alternative.

  The alternative H_A is the idea or conjecture.

  The null H_0 is the negation (complement) of the conjecture.

- The hypotheses are not on an equal footing. The emphasis is always on the null and checking the viability of the null based on the data. If the null appears unviable, then we have proven the alternative. If the null is viable, we cannot say anything about whether the null or alternative is true.

Example 1 (A one-sided, left tailed test)

- Suppose that you believe that a chocolate company is making chocolates lighter than the 50 grams stated on the chocolate bar, how would you investigate this?

Example 1: Collecting the data

- If chocolate bars are getting lighter, it would be reasonable to suppose that the mean weight of a chocolate bar is less than the 50 grams stated on the packet. Since you are investigating the possibility that the mean weight has decreased you state this as your alternative.

  Let μ denote the mean weight of a chocolate bar. The hypothesis are the following. The alternative is H_A: $\mu < 50$ grams (the hypothesis you wish to investigate). The null is H_0: $\mu \geq 50$ grams.

- Of course you do not know the mean weight of a chocolate bar, but it can be estimated from a sample of chocolate bars.

- A simple random sample of 16 is drawn:

  `http://www.stat.tamu.edu/~suhasini/teaching651/chocolate.dat`

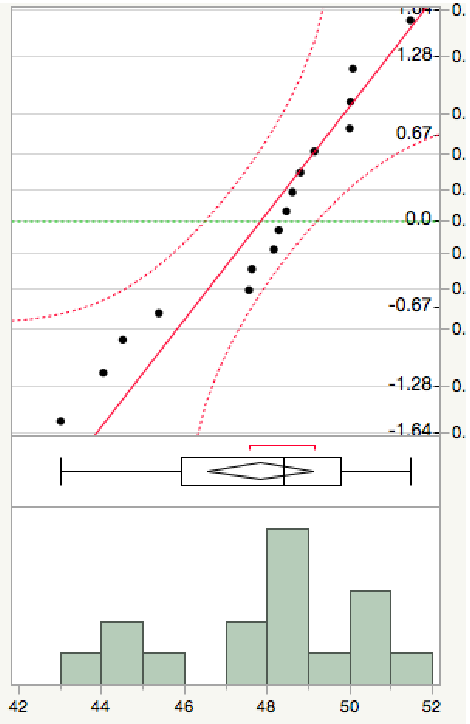
50.09	48.48	48.63	43.03	47.58	44.06	50.03	50.01
48.30	45.40	49.16	44.53	48.82	51.48	47.65	48.18

- The sample mean of the above data is $\bar{X} = 47.8$, the sample standard deviation is $s = 2.4$.

- Since the null is $H_0 : \mu \geq 50$, any sample mean *greater* than 50 is compatable with the null being true (we cannot reject the null).

- We can only reject the null if the sample mean is *far below* 50. Then, the alternative hypothesis seems a better explanation for the observed data.

Solution 1: Distribution of average under the null



- The QQplot of the data shows that it does not deviate hugely from normal. Thus the sample mean (based on 16) is normally distributioned.

- The **estimated** standard error is $2.4/\sqrt{16} = 0.6$, this together with normality of the sample mean, implies using a t-distribution with **15**.

We now analyze the data using (i) Rejection regions (ii) Rejection regions together with the t-transform and (iii) P-values. All methods are identical.

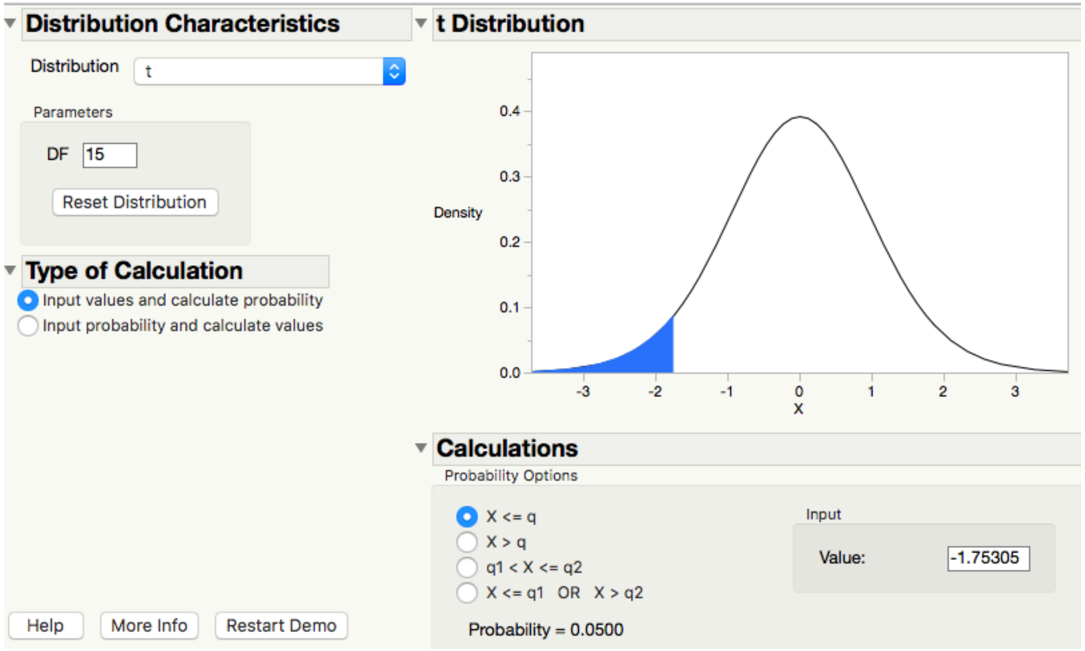
Solution 1: $H_0 : \mu \geq 50$ vs $H_A : \mu < 50$

- Using rejection regions To construct the non-rejection region, we recall that anything to the far left is considered unlikely if the null hypothesis is true. Based on this, the non-rejection region is

  $$[50 - 1.75305 \times 0.6, \infty) = [48.94, \infty)$$

  where -1.75305 corresponds to 5% of the t-dist with 15 df. The plot is below. Compare 47.8 with the plot.

- Using rejection regions with the t-distribution The rejection region (at the 5% level) is the blue area in



Distribution: t

DF: 15

● Input values and calculate probability
○ Input probability and calculate values

● X <= q
○ X > q
○ q1 < X <= q2
○ X <= q1 OR X > q2

Value: -1.75305

Probability = 0.0500

- The t-transform for the data set is $\frac{47.8-50}{0.6} = -3.66$. This is in the rejection region, thus we reject the null at the 5% level. The p-value is less than 5%.

- Using p-values Because the alternative is pointing left the p-value is

$$P\left(t_{15} \leq \frac{47.8 - 50}{0.6}\right) = P(t_{15} \leq -3.66) = 0.0011.$$

  Note the exact p-value was found using statistical software. However, you can also find upper and lower bounds using the t-tables.

- Since 0.11% is substantially below 5% there is (strong) evidence to reject the null.

- Regardless of the method used, based on our data there is evidence to suggest that the mean weight of chocolate bars has decreased.

Example 2 (A two-sided test)

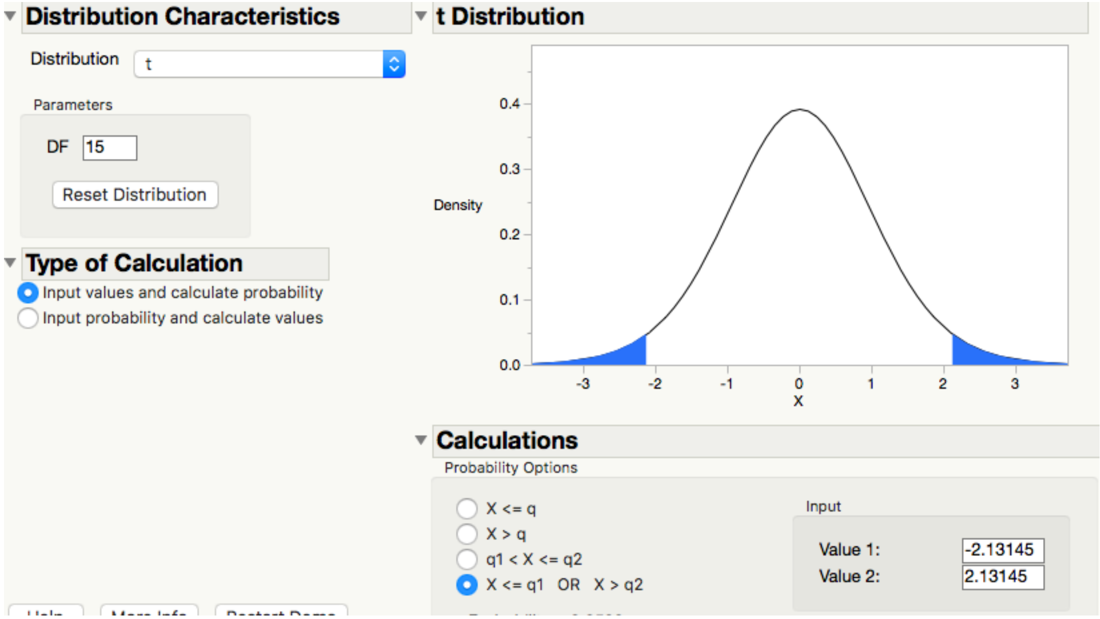
Suppose you want to test that the mean weight is not 50 grams. State your null and alternative. Using the same data as before do the test at the 5% level.

Solution 2

- The competing hypotheses are $H_0 : \mu = 50$ against $H_A : \mu \neq 50$.
- We use the same data set as in Example 1, and use the t-distribution (since the standard deviation is estimated from the data).
- Using rejection regions

$$[50 - 2.13 \times 0.6, 50 + 2.13 \times 0.6] = [48.72, 51.28].$$

- Using t-transforms The t-transform $t = (\bar{X} - 50)/0.6$ is $[-2.13, 2.13]$ (at the 5% level) and the critical region is the blue area in



Distribution Characteristics

Distribution: t

Parameters

DF: 15

Type of Calculation

● Input values and calculate probability

○ Input probability and calculate values

Calculations

Probability Options

○ X <= q

○ X > q

○ q1 < X <= q2

● X <= q1 OR X > q2

Input

Value 1: -2.13145

Value 2: 2.13145

- Since the t-value is $\frac{47.8-50}{0.6} = -3.66$ is in the critical region, we reject the null (the p-value is less than 5%).

- Using p-value Calculate two times the smallest area:

  $$P\left(|t_{15}| \geq \left|\frac{47.8-50}{0.6}\right|\right) = 2 \times P\left(t_{15} \leq -3.66\right) = 2 \times 0.0011 = 0.0022.$$

  The p-value is 0.22%. This is smaller than 5%, there is evidence to suggest that the mean has changed.

Observe it is harder to reject the null for a two-sided test than a one-sided test.

Example 3 (A one sided right-sided test)

The manufacturer lawyers start to intervene. They claim that the mean weight is larger than 50 grams. Using the same data as before we test this at the 5% level.

Solution 3

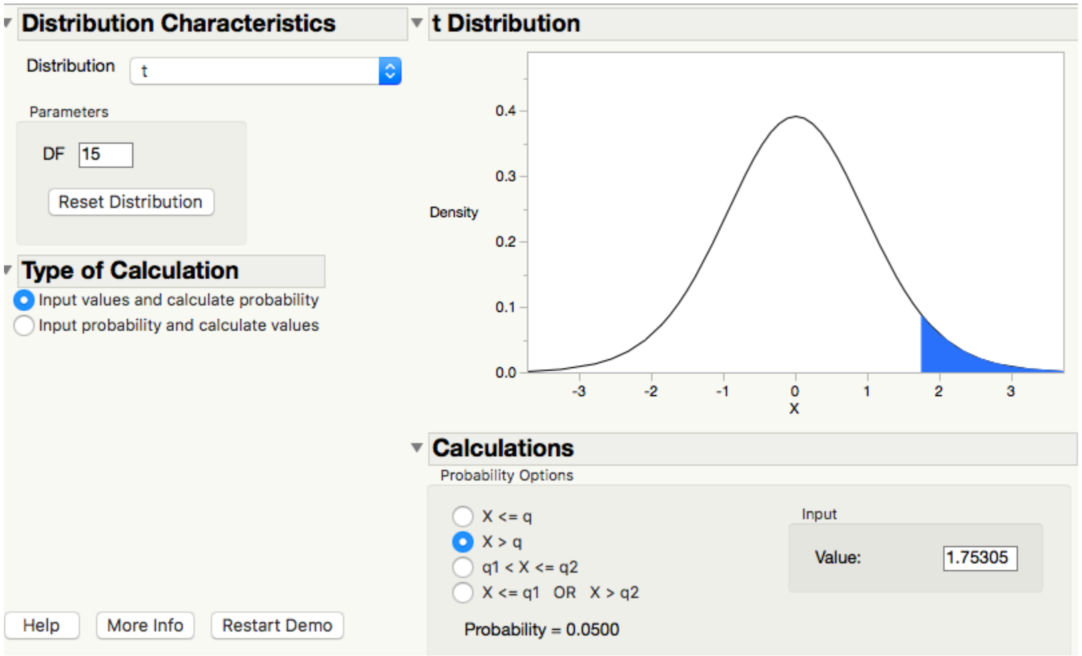
- The hypotheses are $H_0 : \mu \leq 50$ against $H_A : \mu > 50$.

- All statistical tests are based on disproving the null; can the data observed be easily generated if the null were true? The sample mean $\bar{X} = 47.8$, can easily be generated under the null being true (though we still cannot say the null is true). We cannot reject the null.

- Using rejection regions The non-rejection region is

$$(-\infty, 50 + 1.75305 \times 0.6] = (-\infty, 51.05].$$

- Using t-transform The critical region is the blue area in



Distribution Characteristics

Distribution: t

Parameters

DF: 15

Type of Calculation

● Input values and calculate probability

○ Input probability and calculate values

t Distribution

Calculations

Probability Options

○ X <= q

● X > q

○ q1 < X <= q2

○ X <= q1 OR X > q2

Input

Value: 1.75305

Probability = 0.0500

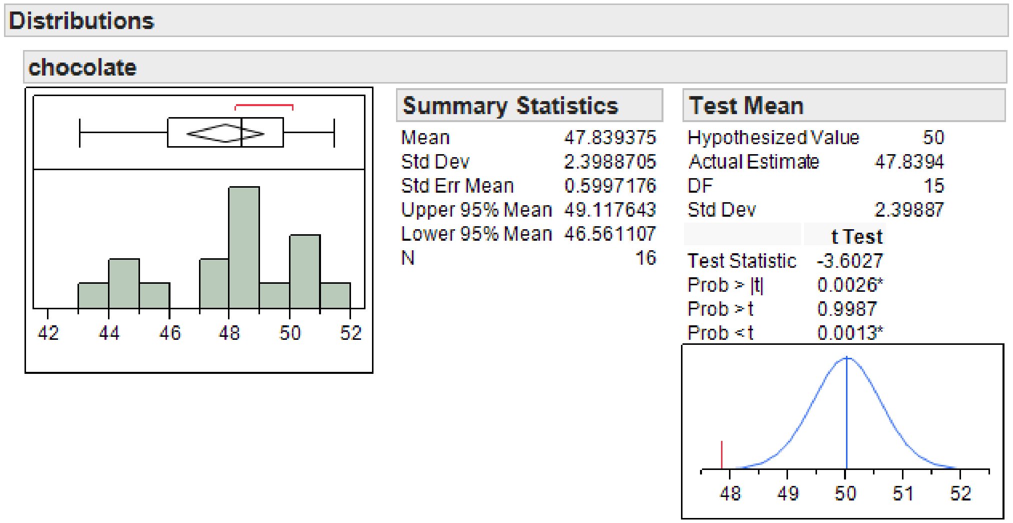
- Since $t = \frac{47.8-50}{0.6} = -3.66$ is in the white non-rejection region, we cannot reject the null.

- P-value calculation The p-value it is the area to the right of 47.8:

$$P\left(t_{15} \geq \frac{47.8-50}{0.6}\right) = P(t_{15} \geq -3.66) = 1 - 0.0011$$

- The p-value is close to 99.9%, there is no evidence to reject the null. There is no evidence to suggest the mean weight of the chocolate bar has increased.

The test in JMP



```
Distributions
chocolate
Summary Statistics                 Test Mean
Mean              47.839375        Hypothesized Value      50
Std Dev           2.3988705        Actual Estimate    47.8394
Std Err Mean      0.5997176        DF                      15
Upper 95% Mean    49.117643        Std Dev            2.39887
Lower 95% Mean    46.561107                      t Test
N                        16        Test Statistic   -3.6027
                                   Prob > |t|       0.0026*
                                   Prob > t         0.9987
                                   Prob < t         0.0013*
```

- Test Statistic gives the p-value. Prob$>|t|$ the p-value for $H_0 : \mu = 50$ vs $H_A : \mu \neq 50$. Prob$> t$ the p-value for $H_0 : \mu \leq 50$ vs $H_A : \mu > 50$. Prob$< t$ the p-value for $H_0 : \mu \geq 50$ vs $H_A : \mu < 50$.

- The sum of the p-values for the two one-sided tests, $0.13\% + 99.87\% = 100\%$. This will always be the case.

- The p-value for two-sided test is always double the smallest one-sided test p-value $2 \times 0.13\% = 0.26\%$.

Types of test

There are three main types of tests.

- Two sided tests $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$.
- One sided tests (case I) $H_0 : \mu \geq \mu_0$ and $H_A : \mu < \mu_0$.
- One sided tests (case II) $H_0 : \mu \leq \mu_0$ and $H_A : \mu > \mu_0$.

Look at `handout_lecture16.pdf` pages 12-14.

- Remember, the null should always contain an equal sign in it, whether it is $H_0 : \mu = \mu_0$, $H_0 : \mu \geq \mu_0$ or $H_0 : \mu \leq \mu_0$.

Rejection regions verses p-values

- You will have seen in the handout that I did the tests in what seemed to be two different ways:
  - Rejection regions.
  - p-values.

- Both methods are the same, they lead to identical conclusions. The p-value method is what you will see when you look at output. On the other hand, the rejection region method is easier to understand, and is required for "power" calculations.

Testing at other significance levels

- So far we have stated everything for 5% level. The discussion above can easily be adapted for other levels 10% and 1% too.

- All this means is that in the decision rule we compare the p-value with 10% or 1%.

- Often in research articles the p-value in the test is quoted, this allows the reader to choose their own significance level.

Example 4

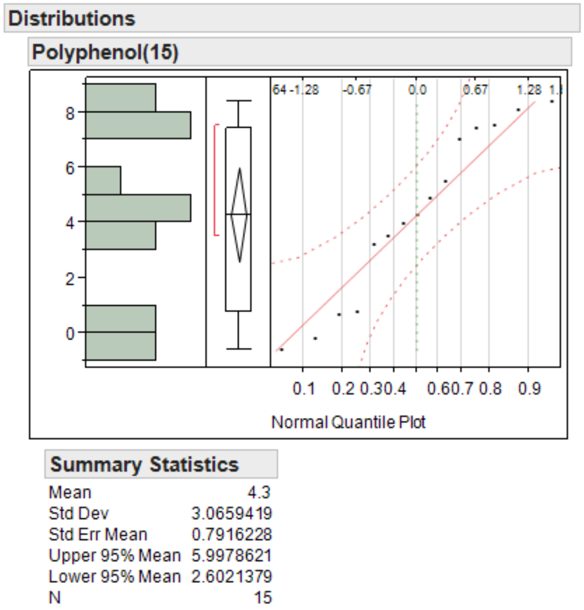
Let us return to the example on increasing polyphenol levels by drinking red wine.

- We are interested in whether moderate read wine consumption leads, on average, drinking to an increase polyphenol levels.
- There for if we let μ denote the mean change in polyphenol levels after drinking read wine hypotheses are $H_0 : \mu \leq 0$ against $H_A : \mu > 0$.

To investigate this, 15 healthy males were included in the study and for each male the increase (or decrease) in polyphenols is measured (after drinking red wine):

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A-B	0.7	3.5	4	4.9	5.5	7	7.4	8.1	8.4	3.2	0.8	4.3	-0.2	-0.6	7.5

The sample mean is 4.3 and the sample standard deviation is 3.06. Can this data set have been realized if the population mean μ is zero or less. If this is highly unlikely (p-value is small) we have disproved the null.



Summary Statistics

Mean	4.3
Std Dev	3.0659419
Std Err Mean	0.7916228
Upper 95% Mean	5.9978621
Lower 95% Mean	2.6021379
N	15

Solution 4

- We want to test $H_0 : \mu \leq 0$ against $H_A : \mu > 0$.

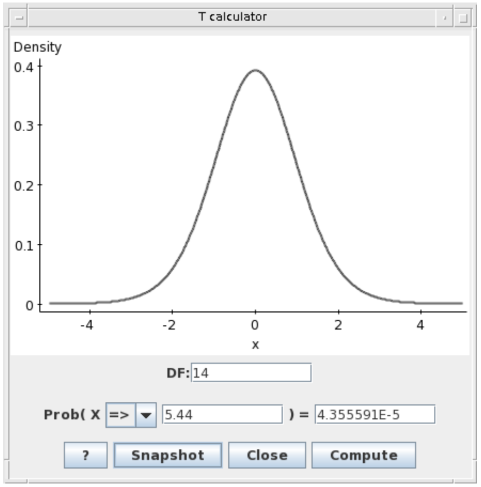
- The data does not appear to deviate massively from normality (see the plot above), and the sample size of 15 seems large enough for us to assume normality of the sample mean.

- The sample standard error for the data is s.e$= 3.06/\sqrt{15} = 0.79$. Since we have estimated the the standard deviation from the data we need to use the t-distribution.

- t-transform under the null is

$$t = \frac{4.3 - 0}{0.79} = 5.44$$

Using the t-distribution with 14df gives:

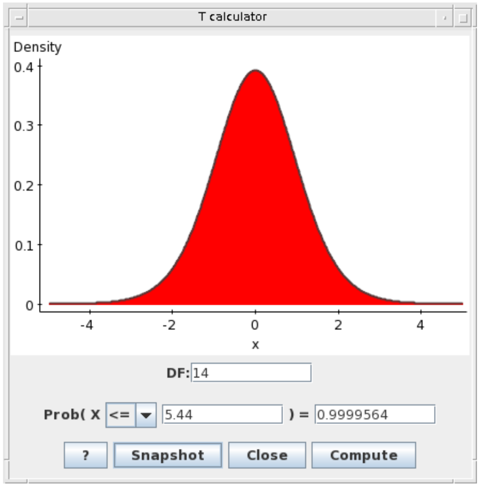


DF: 14

Prob(X => 5.44) = 4.355591E-5

- As the alternative is pointing to the right the p-value is the area to the right of 5.44 which is 0.0044%. 0.0044% is substantially smaller than 5% that we can reject the null.

- Further discussion Suppose we want to test whether drinking wine *reduced* polyphenol levels $H_0 : \mu \geq 0$ against $H_A : \mu < 0$. It is clear that there is no evidence in the data of a decrease since the sample mean is $\bar{x} = 4.3$, and this corresponds to a p-value which is 0.9999564 (almost one), which is far greater than 5%. The p-value is given below:

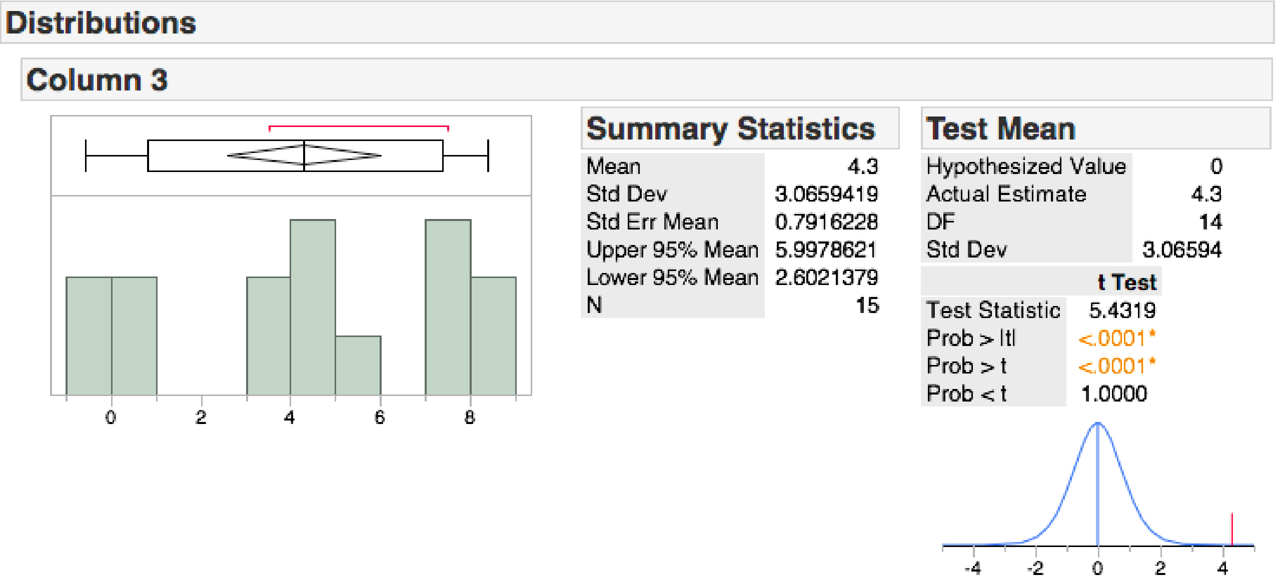


DF: 14

Prob(X <= 5.44) = 0.9999564

The p-value for this test is the area to the left of 5.4.

Solution 4: The test in JMP



Distributions

Column 3

Summary Statistics

Mean	4.3
Std Dev	3.0659419
Std Err Mean	0.7916228
Upper 95% Mean	5.9978621
Lower 95% Mean	2.6021379
N	15

Test Mean

Hypothesized Value	0
Actual Estimate	4.3
DF	14
Std Dev	3.06594

	t Test
Test Statistic	5.4319
Prob > |t|	<.0001*
Prob > t	<.0001*
Prob < t	1.0000

Observe from the output that the result of the test $H_0 : \mu \leq 0$ vs $H_A : \mu > 0$ leads to a p-value which is very small. Thus there is evidence in the data to suggest that drinking red wine leads to an increase in polyphenol levels in the blood.

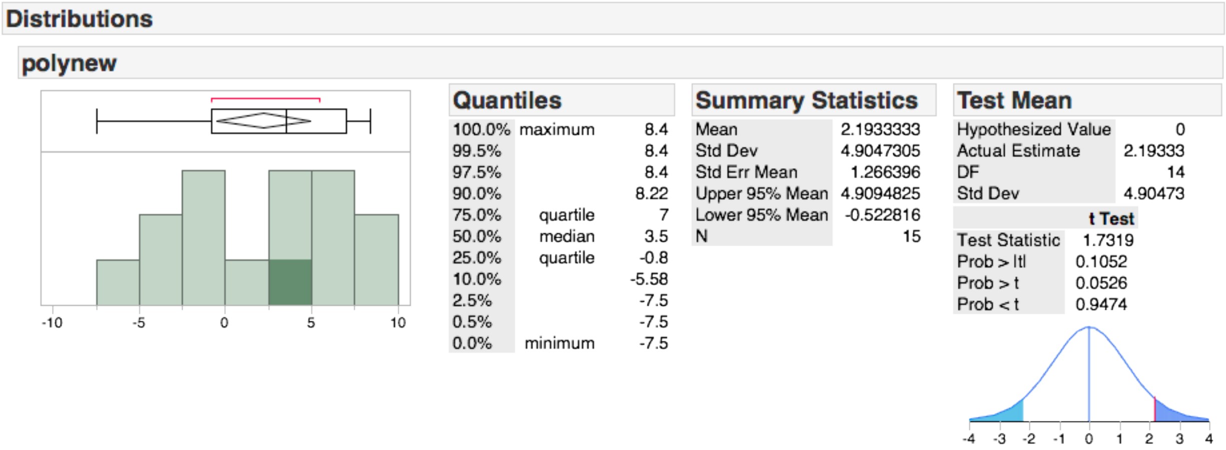
Question 5

- In the previous example, it was clear from the data that the result of the test $H_0 : \mu \leq 0$ vs $H_A : \mu > 0$ would lead to the rejection of the null. Most of the data was positive, the standard error was small, this meant the p-value would be very small and we would reject the null.
- Consider the less clear example that the the change in polyphenol level is

  $$0.7, 3.5, 4, 4.9, 5.5, 7, 7.4, 8.1, 8.4, -3.2, -0.8, -4.3, -0.2, -0.6, -7.5.$$

  By inspective the data it is unclear whether we can reject the null or not.

Solution 5



Distributions

polynew

Quantiles

100.0%	maximum	8.4
99.5%		8.4
97.5%		8.4
90.0%		8.22
75.0%	quartile	7
50.0%	median	3.5
25.0%	quartile	-0.8
10.0%		-5.58
2.5%		-7.5
0.5%		-7.5
0.0%	minimum	-7.5

Summary Statistics

Mean	2.1933333
Std Dev	4.9047305
Std Err Mean	1.266396
Upper 95% Mean	4.9094825
Lower 95% Mean	-0.522816
N	15

Test Mean

Hypothesized Value	0
Actual Estimate	2.19333
DF	14
Std Dev	4.90473

	t Test
Test Statistic	1.7319
Prob > |t|	0.1052
Prob > t	0.0526
Prob < t	0.9474

- We see that the p-value is 5.26%. This means the chance of observing the data under the null is 5.26%. This is a relatively small chance, but not overwhelmingly so. It is definitely not enough to convincingly disprove the null.

Conclusion There is not enough evidence in the data to disprove the null.

Example 6

A patient is classified as having low potassium if her mean potassium level is below 3.5. The potassium level in each blood sample taken will vary from sample to sample, but it is known that the standard deviation 0.4 ($\sigma = 0.4$). 4 patients are being examined, for each patient 20 (!!) blood samples are taken and the sample mean calculated, the data is summarised below.

(i) Patient 1: $\bar{x} = 3.3$.

(ii) Patient 2: $\bar{x} = 3.4$.

(iii) Patient 3: $\bar{x} = 3.6$.

(iv) Patient 4: $\bar{x} = 3.9$.

Is there any evidence at the 5% level of low potassium in each of these patients, compare your answers to the confidence intervals constructed in Lecture 12. Note the rejection point (where we say that we believe the patient has low potassium) is any sample mean less that $3.5 - 1.64 \times 0.4/\sqrt{20} = 3.35$.

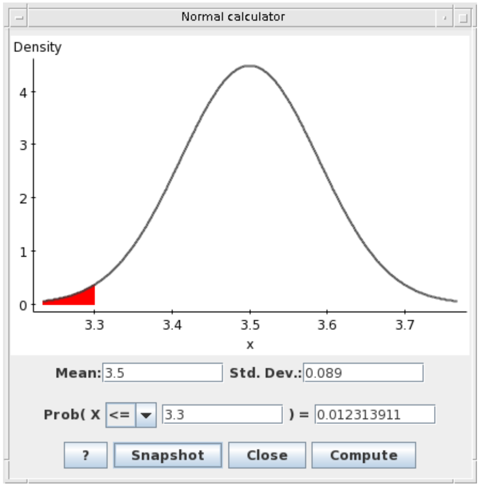
Solution 6

- The sample size is relatively large, so we can assume normality of the sample mean (it is unlikely the distribution of the blood samples deviate massively from normality).

- The standard error is s.e.$= 0.4/\sqrt{20} = 0.089$. Since we have not estimated the standard deviation from the data, we can use the normal distribution (not the t-distribution).

- As we want to see whether there is any evidence of low potassium our hypotheses are $H_0 : \mu \geq 3.5$ (patient does not have low potassium) against $H_A : \mu < 3.5$ (patient has low potassium).

Solution 6: Patient 1

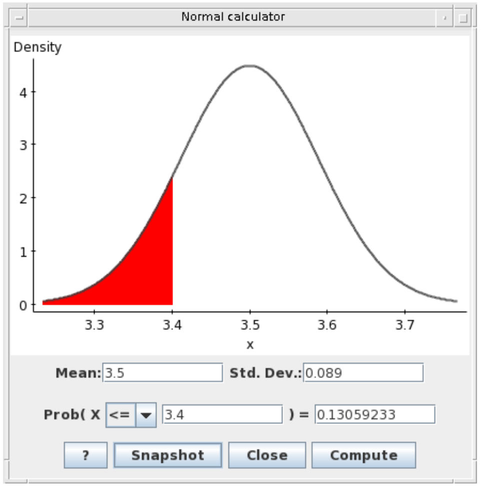


Mean: 3.5 Std. Dev.: 0.089

Prob(X <= 3.3) = 0.012313911

The p-value is 1.2%, as this is less than 5%, there is evidence to suggest the mean level is less than 3.5 (that the patient may have low potassium).

Solution 6: Patient 2

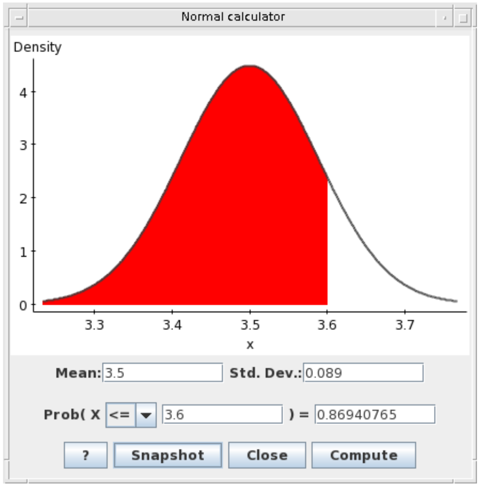


Mean: 3.5 Std. Dev.: 0.089

Prob(X <= 3.4) = 0.13059233

The p-value is 13%, as this is greater than 5%, there is not enough evidence in the data to suggest the mean level is less than 3.5 (we do not know whether the patient has low potassium or not).

Solution 6: Patient 3

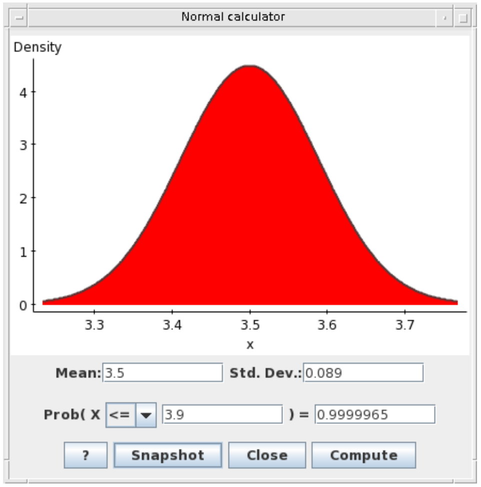


Mean: 3.5 Std. Dev.: 0.089

Prob(X <= 3.6) = 0.86940765

The p-value is 87%, as this is a lot greater than 5%, there is not enough evidence in the data to suggest the mean level is less than 3.5 (we do not know whether the patient has low potassium or not).

Solution 6: Patient 4



Mean: 3.5 Std. Dev.: 0.089

Prob(X <= 3.9) = 0.9999965

The p-value is 99.99965%, as this is far greater than 5%, there is no evidence in the data to suggest the mean level is less than 3.5. In fact, if test $H_0 : \mu \leq 3.5$ against $H_A : \mu > 3.5$ the p-value would be $100 - 99.99965 = 0.00035\%$, which is so small that there is evidence that the patient has healthy potassium levels!

Discussion of the above results

- The solutions above explain what the test is actually doing. It is testing the viability of the null hypothesis given the data (it is not testing the viability of the alternative).

- If we test $H_0 : \mu \geq 3.5$ against $H_A : \mu < 3.5$, the sample mean is $\bar{X} = 3.4$, the p-value is 13%. This is sufficiently large for the null to be viable, even though $\bar{X} = 3.4 < 3.5$.

- On the other hand, if the patient's sample mean is $\bar{X} = 3.3$, then the likelihood of getting this reading when the patient has a healthy potassium level is 1.2%. This is relatively small, as this is below our decision criterion of 5%, we can say there is evidence against the null and suggest that the patient has low potassium.

- It is important to note, that if 20 healthy patients were to have blood samples take, one out of 20 may be falsey diagnosed with low potassiums (since the test is done at the 5% level). Falsely rejecting the null is often called a false positive.

Example 7

An airline wants to evaluate the depth perception of pilots over the age of fifty. A random sample of pilots over the age of fifty are used. The sample data of the pilots error is listed below.

2.7	2.4	1.9	2.6	2.4	1.9	2.3
2.2	2.5	2.3	1.8	2.5	2.0	2.2

(a) Construct a 95% CI for the mean error.

(b) The mean error for ‘young’ plots is 2, test the hypothesis that the mean error is larger for ‘older’ pilots.

Solution 7

(a) The average is 2.26 and the sample standard deviation is 0.27 sample standard error is $0.27/\sqrt{14} = 0.0778$.

To evaluate the sample variance we use $\frac{1}{14-1}\sum_{i=1}^{14}(X_i - \bar{X})^2$. There are 14 observations hence the z-transform is a t-distribution with 13 degrees of freedom.

By looking up in the tables we see that $t_{0.025}(13) = 2.16$ (we use this instead of $z_{0.025} = 1.96$). The 95% CI is

$$\left[2.26 - 2.16 \times \frac{0.27}{\sqrt{14}}, 2.26 + 2.16 \times \frac{0.27}{\sqrt{14}}\right].$$

(b) The hypotheses are $H_0 : \mu \leq 2$ against $H_A : \mu > 2$. We do a one-sided test so we need $t_{0.05}(13) = 1.771$ The non-rejection region is:

$$\left(-\infty, 2 + 1.771 \times \frac{0.27}{\sqrt{14}}\right] = (-\infty, 2.13].$$

Therefore the rejection region is

$$\left(2 + 1.771 \times \frac{0.27}{\sqrt{14}}, \infty\right) = (2.13, \infty).$$

Since the sample average $\bar{X} = 2.26$ lies in the rejection region there is enough evidence to reject the null and accept the alternative. That is there is evidence to suggest that ‘older’ pilots tend to make a larger error.

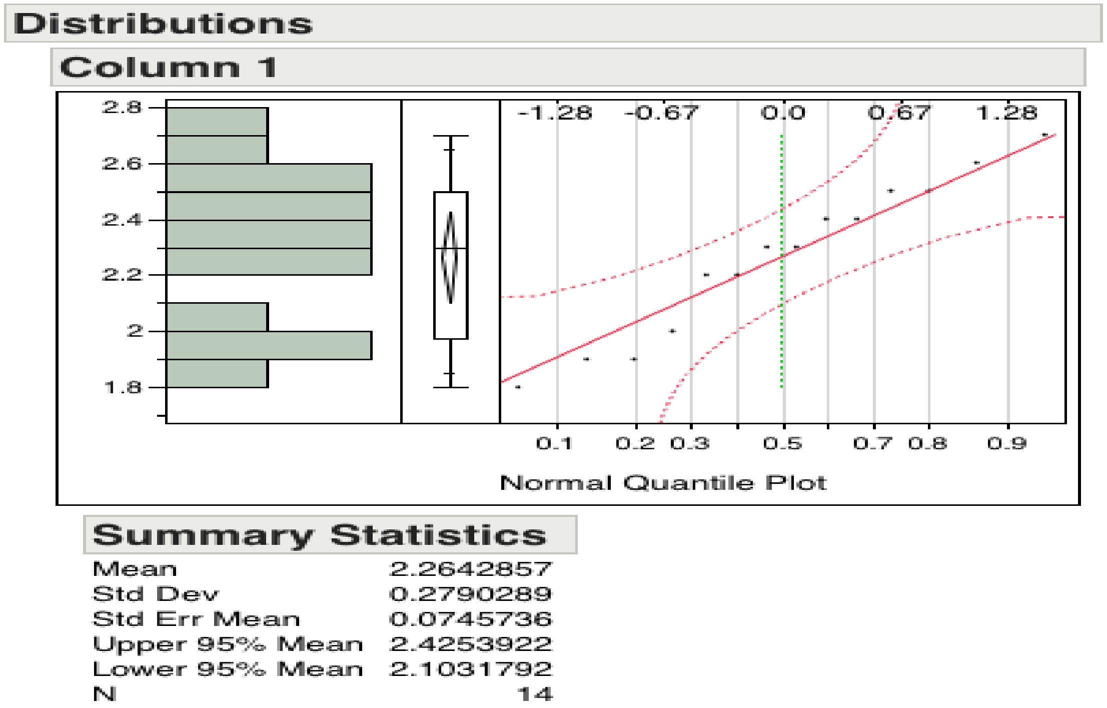
- Alternatively we can obtain the p-value

$$t = \frac{\bar{X} - 2}{\frac{0.27}{\sqrt{14}}} = \frac{2.26 - 2}{\frac{0.27}{\sqrt{14}}} = 3.48.$$

If we used Statcrunch of JMP we could find that the area to the right of 3.48 (since the alternative is pointing right) is 0.02%. If we only have tables available, since the test is done at the 5% level first look up $t_{0.05}(13) = 1.771$ in the tables. Now compare the t-transform with $t_{0.05}(13) = 1.771$, since $3.48 > 1.771$, the p-value must be less than 5% (easiest seen with a plot). Therefore, there is evidence at the 5% level to reject the null.

- All these methods are equivalent.

Solution using JMP output



Distributions

Column 1

```
Summary Statistics
Mean              2.2642857
Std Dev           0.2790289
Std Err Mean      0.0745736
Upper 95% Mean    2.4253922
Lower 95% Mean    2.1031792
N                        14
```

Everything we need to do a test or construct a CI (of any type) can be found in the above output.

- The data does not deviate much from the x=y axis on the QQplot. This suggests that the distribution of pilot perception does not deviate much from normality. Therefore, even though the sample size is quite small, both the 95% confidence intervals and the tests will be accurate (ie. we really will be 95% confident about the location of the mean and the p-values are really those percentages).

- Using this information we can construct, say, a 99% confident interval for the mean (by looking up t-tables with 13df and 0.5%)

$$[2.26 \pm 3.01 \times 0.075].$$

- We can use this information to do the test $H_0 : \mu \leq 2$ against $H_A : \mu > 2$.

- For the purposes of an exam, I may only give half the information in the summary statistics and you need to deduce the rest.

Type II errors

- So far we have addressed the issue of a Type I error. This is the significant level and is the probability of falsely rejecting the null when in fact it is true.

- There is of course another error we could make when making a decision based on a statistical test. That is the probability of not rejecting null when in fact the alternative is true (in other words, not detecting the alternative hypothesis in the test).

- This is called a Type II error, formally it is defined as

$$P(\text{Not enough evidence to reject } H_0 | H_A \text{ is true})$$

Example

$$P(\text{Unable to reject null: mean price of wheat changed}|\text{mean price has change}$$

- In plain terms, the Type II error is finding no difference when there is a difference.

- In a criminial trial it is the probability of finding no evidence of guilt when there is guilt.

- The Type II error cannot be controlled when the Type I error is controlled.

- Indeed the smaller we make the Type I error, the larger the Type II error will be (there is a trade off between the two).

- In a criminal trial, usually we want to control the number of innocent people we send to prison, this is our type I error (the amount of evidence that it required to convict someone).

A summary of the decision process and the possible mistakes

This is a two-way decision process, which we can write as:

		Truth	
		H_0	H_A
decision	reject H_0	$P(\text{reject} \quad H_0 \mid H_0)$ α (Type I error)	$P(\text{reject} \quad H_0 \mid H_A)$ $\underbrace{1-\beta}_{=\text{power}}$ (correct outcome)
made	cannot reject H_0	$P(\text{cannot reject} \quad H_0 \mid H_0)$ $1-\alpha$ (correct outcome)	$P(\text{cannot reject} \quad H_0 \mid H_A)$ β (Type II error)

Notice if the probability of rejecting the null (H_0) when it is true, the probability of not rejecting the null when it is true is $1-\alpha$.