Data Analysis and Statistical Methods
Statistics 651

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

Lecture 14 (MWF) The t-distribution

Suhasini Subba Rao
Review of previous lecture

- Often the precision of an estimator is stated in terms of it’s margin of error.

- For example, the proportion of Americans that are happy is 40% with a margin of error 2.5%.

- We now know that margin of error corresponds to the plus/minus part in a confidence interval

\[ [\bar{X} - E, \bar{X} + E] = [\bar{X} - 1.96 \times \sqrt{\frac{\sigma^2}{n}}, \bar{X} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}]. \]

- The margin of error does not mean that the proportion of Americans that
are happy is definitely in the interval $[37.5, 42.5]\%$ (this is the difference between knowing for certain and a confidence interval).

Technically, the margin of error means that for every 100 sample mean drawn about 95% of them will lie inside the interval $[ar{X} - E, \bar{X} + E]$.

- We can use the margin of error to determine the ‘ideal’ sample size using the formula

$$n = \left( \frac{z\alpha/2}{E} \right)^2$$

- To calculate the margin of error we had to assume the standard deviation is known. If it is not known we need to come up with an intelligent guess for an upper bound.
Terminology: Standard deviations and errors

- The **standard deviation** is a measure of variation/spread of a variable (in the population). This is typically denoted as $\sigma$. See Lecture 4.

- The **standard error** is a measure of variation/spread of the sample mean. The standard error of the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is

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

  See Lecture 12.

- Usually, $\sigma$ is unknown. To get some idea of the spread, we estimate it
from the sample \( \{X_i\}_{i=1}^{n} \) using the formula

\[
s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}.
\]

We call \( s \) the sample standard deviation. It is an estimator of the standard deviation \( \sigma \). Usually \( s \neq \sigma \). Often \( s < \sigma \) (especially, when the sample size is not large).

- Since \( \sigma \) is usually unknown, the standard error \( \sigma/\sqrt{n} \) is usually unknown. Instead we estimate it using the sample standard error is

\[
\frac{s}{\sqrt{n}}.
\]
Motivation

• We take a SRS of 5 students and record their heights 61, 63, 65, 66, 72. The sample mean/average is 65.4.

• Our objective is to construct a 95% confidence interval for the population mean of students. Putting numbers into the formula gives

$$\left[ 65.4 - 1.96 \times \frac{\sigma}{\sqrt{5}}, 65.4 + 1.96 \times \frac{\sigma}{\sqrt{5}} \right]$$

• But the population standard deviation, $\sigma$, is unknown.

• We can estimate it from the data 61, 63, 65, 66, 72, using the sample standard deviation which is

$$s = \sqrt{\frac{1}{4-1} (61 - 65.4)^2 + (63 - 65.4)^2 + (65 - 65.4)^2 + (66 - 65.4)^2 + (72 - 65.4)^2} = 4.16$$
and put 4.16 into the above confidence interval.

- What we want to know is whether it changes anything. In fact it turns out that the population standard deviation is $\sigma = 4.3$...what does this tells us about the interval?
How estimating the standard deviation effects our results

- So far we have assumed that the standard deviation $\sigma$ is known.

- This is sometimes a plausible assumption. There are situations when one may know the standard deviation but not the population mean.

- However in general we will not know $\sigma$. $\sigma$ is unknown and has to be estimated from the data.

- Given a data set $X_1, \ldots, X_n$ (say the 9 observations 0.025, 0.025, 0.057, 0.064, 0.054, 0.035, 0.047, 0.059, 0.045, used in the lecture 13) we can estimate it.
• We can estimate the variance using the sample standard deviation

\[ s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}. \]

• **Constructing confidence intervals** In this case it seems reasonable to replace \( \sigma \) with \( s \) when evaluations a \( z \)-transform or a 95% CI:

\[
\begin{align*}
\text{z-transform} & \quad \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \rightarrow \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \\
\text{t-transform} & \quad \text{95% CI} \quad \left[ \bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}} \right] \rightarrow \left[ \bar{X} \pm 1.96 \times \frac{s}{\sqrt{n}} \right]
\end{align*}
\]

• But have we lost anything in replacing \( \sigma \) with \( s \)?
The effect of estimating the standard deviation

- In the discussion below we are assuming that the observations \( \{X_i\} \) are independent random variables from a normal distribution with mean \( \mu \) and standard deviation \( \sigma \).

What we discuss below has nothing to do with correcting for normality of the observations. It is about estimation of the population standard deviation \( \sigma \).

- The sample standard deviation \( s \) is random it varies from sample to sample.

- If the sample size is relatively small it can often underestimate the true standard deviation. This can cause substantial problems.
• The z-transform is the number of standard errors that can “fit between” the mean and the sample mean. If the standard deviation has been underestimated, then the z-transform will be larger than what it is suppose to be

\[
z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \begin{cases} \text{larger} \end{cases}
\]

- smaller

• There is a change in terminology (when we replace the population standard error with the sample standard error) we call it

\[
t\text{-transform} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}.
\]
• Equivalently, if we use the estimated standard deviation to construct the confidence interval, an underestimated standard deviation will result in a confidence interval that is too narrow. Consider the 95% confidence interval

\[
\left[ \bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}} \right] \rightarrow \left[ \bar{X} - 1.96 \times \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{s}{\sqrt{n}} \right].
\]

If \( s \) is smaller than \( \sigma \), then the interval will be too narrow for it to be a 95% confidence interval.

• We need to correct for the fact that \( s \) tends to underestimate the population standard deviation \( \sigma \).

• Indeed, it is very simple to make the correction. All we need to do is change the distribution from a normal distribution to a t-distribution.
Gossett’s experiment

• We find that when we estimate the variance (rather than use the true variance) we need to increase the size of the confidence interval to account for the greater variation in the $Z$-transform.

• This fact was discovered by William Gossett, who was a chemist, working for Guinness the brewery (in Ireland) and had to judge the quality of several brews.

• He was working with a small sample size $X_1, \ldots, X_{10}$ (sample size is 10), and estimated the standard deviation from this $s = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (X_i - \bar{X})^2}$. From previous experiments he knew that the true mean was $\mu = 4$.

• He wanted to construct 95% CIs for the mean. But, rather than use the population standard deviation $\sigma$, he replaced it with the sample standard
deviation $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$. For each sample of size 10 he constructed the 95% CI:

$$\left[ \bar{X} - 1.96 \frac{s}{\sqrt{10}}, \bar{X} + 1.96 \frac{s}{\sqrt{10}} \right]$$

He counted the number of times the true mean $\mu$ was in the interval.

- You would expect that about 5% of the time the true mean should be outside the interval (since it is a 95% CI).

- What William Gosset noticed was that the true mean was outside the interval more than 5% of time.

- This interval is not a 95% confidence interval.
An illustration: Confidence intervals

We draw a sample of size 10, from a normal distribution, and estimate both the sample mean and standard deviation and construct a 95% CI using $z = 1.96$. Observe only 91 of the 100 confidence intervals contain the mean. We have less confidence in this interval than the stated 95% level!
The t-distribution

- The transform (which we formally called the z-transform)

\[ t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n - 1), \]

has a \( t \)-distribution with \( (n - 1) \)-degrees of freedom where \( n \) is the number of observations used to estimate \( \sigma \) and \( \mu \).

- Since \( t \) can usually be larger than \( z \) (since the sample standard error tends to be smaller than the population standard error), the distribution of \( t \) has “thicker” tails than a normal distribution. This means it can have extreme values or outliers. This is reflected in the critical values which are given 3 slides on.
• The term degrees of freedom is a word commonly used in statistics. It refers to the “effective” sample used to estimate the population standard deviation. The \( (n - 1) \)- comes into play because once the sample mean is estimated the effective sample size is \( (n - 1) \) and not \( n \).

• The distribution of \( \frac{\bar{X} - \mu}{s/\sqrt{n}} \) depends on the sample size.

• We call \( t(n - 1) \) the Student \( t \)-distribution with \( (n - 1) \)-degrees of freedom. We use the name Student, in honor of William Gosset (he wrote all his papers under the pseudonym Student).
How does this change things?

- We do almost everything as we did before, but when we estimate the standard deviation we use the \( t \)-distribution instead of the standard normal.

- The \( t \)-values are larger than the \( z \)-values to compensate for the underestimation of standard deviation.

- Rather than use the normal tables we use the \( t \)-tables which are very easy to use and can be found on my website.

- Most statistical software (such as JMP)
**Reading t-tables (Table 2)**

Percentage points of Student’s t-distribution

<table>
<thead>
<tr>
<th>d.f.</th>
<th>.005</th>
<th>.01</th>
<th>.025</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>1.000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>318.309</td>
<td>636.619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4.303</td>
<td>6.965</td>
<td>9.925</td>
<td>22.327</td>
<td>31.599</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
<td>10.215</td>
<td>12.924</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
<td>7.173</td>
<td>8.610</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
<td>5.893</td>
<td>6.869</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Confidence intervals using the t-distribution

• When the standard deviation $\sigma$ is \textbf{known}. The $(1 - \alpha)100\%$ CI is

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

• When the standard deviation $\sigma$ is \textbf{unknown}, we estimate it from the data

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

and use the CI

$$
\left[ \bar{X} - |t_{\alpha/2}(n-1)| \frac{s}{\sqrt{n}}, \bar{X} + |t_{\alpha/2}(n-1)| \frac{s}{\sqrt{n}} \right].
$$
**An illustration: Confidence intervals**

We draw a sample of size 10, from a normal distribution, and estimate **both** the sample mean and standard deviation and construct a 95% CI using $t_{0.025}(9) = 2.262$ (compare with $z = 1.96$). By using the t-distribution we have 95% confidence the interval contains the mean.
Example 1: Red Wine and polyphenols

It has been suggested that drinking red wine in moderation may protect against heart attacks. This is because red wine contains polyphenols which act on blood cholesterol.

To see if moderate wine consumption does increase polyphenols, a group of nine random selected males were assigned to drink half a bottle of red wine daily for two weeks. The percentage change in their blood levels are

$0.7, 3.5, 4, 4.9, 5.5, 7, 7.4, 8.1, 8.4$

Here’s the data: [http://www.stat.tamu.edu/~suhasini/teaching651/red_wine_polyphenol.txt](http://www.stat.tamu.edu/~suhasini/teaching651/red_wine_polyphenol.txt). The sample mean is $\bar{x} = 5.5$ and sample standard deviation is 2.517. Construct a 95% confidence interval and discuss what your results possibly imply.
The 95% confidence interval constructed by default in JMP is [3.56, 7.43]. We discuss what this means below.
Solution 1: Red Wine

• The sample size is small, therefore to construct a reliable confidence interval we need that the distribution of the blood samples is does not deviate much from a normally distributed.

Discussion of the polyphenol data set When the sample size is so small it is hard to tell from the 9 points on the QQplot whether the data has come from a normal distribution. However, the these points do not deviate too much from the line for us to believe it is skewed. Furthermore, a blood samples tend to come from a biological experiment. Based on these two facts, it seems plausible that the data does not come from a distribution severe skew or heavy tails. If this is the case, the distribution of the data is unlikely to deviate hugely from normality. Thus, the sample mean based on 9 is likely to be close to normal.

• We do not know the standard deviation and JMP estimates it from the
data. Therefore the 95% confidence interval constructed in JMP uses the t-distribution and not the normal distribution.

The exact calculation:

• Use the t-tables with 8df (sample size, 9, minus one) and 2.5%. This gives the critical value \( 2.306 \). Based on this the 95% CI for the mean is

\[
\left[ 5.5 \pm 2.306 \times \frac{2.517}{\sqrt{9}} \right] = [3.57, 7.43],
\]

which are exactly the numbers given in the JMP output.
Example 2: Red Wine II

- We return to the same question but in order to get a smaller margin of error we include 6 extra males in our study. [http://www.stat.tamu.edu/~suhasini/teaching651/red_wine_polyphenol.txt](http://www.stat.tamu.edu/~suhasini/teaching651/red_wine_polyphenol.txt).

- Notice some of the new guys actually had a drop in their polyphenol levels!
• The **sample mean** is 4.3 and the **sample standard deviation** is 3.06.

• **Solution** We now use a t-distribution with 14 degrees of freedom and the 95% CI for the mean level after drinking wine (for two weeks) is

\[
[4.3 \pm 2.145 \times \frac{3.06}{\sqrt{15}}] = [2.1, 6].
\]

• The factor **2.145** has decreased from the **2.306** given in the previous example. This is because, the sample standard deviation based on \( n = 15 \) tends to be closer to the population standard deviation.
Comparing Example 1 and 2

The difference between Example 1 and Example 2 is the sample size has grown from 9 to 15. We compare the two samples below:

- We see that the smaller sample size contains less ‘extreme’ values (the people whose polyphenol level went down with wine consumption). Less spread in the smaller sample size means that the corresponding estimated standard deviation will be less than the second sample (look at the output below and compare for $n = 9$, $s = 2.5$ whereas for $n = 15$, $s = 3.1$). We
see that for smaller sample sizes the estimated standard deviation tends to underestimate the true population standard deviation.
**Extreme example: n= 3**

Consider the data set 4, 5.5, 6. The sample mean and standard deviation

\[ \bar{x} = 5.17 \]

\[ s = \sqrt{\frac{1}{2}[(4 - 5.17)^2 + (5.5 - 5.17)^2 + (6 - 5.17)^2]} = 1.04 \]

With just three observations it is highly likely the sample standard deviation is anywhere close to the population standard deviation.

- The 95% confidence interval for the population mean is

\[ \left[ 5.17 - 4.3 \times \frac{1.04}{\sqrt{3}}, 5.17 + 4.3 \times \frac{1.04}{\sqrt{3}} \right] = [2.6, 7.8] \]
• Observe that the factor 4.3 is used instead of 1.96, since we have estimated the standard deviation using just 3 observations.
Sample size and the sample standard deviation

As the sample size grows, the standard error of the sample mean gets smaller (see green plot) and the sample standard deviation concentrates about the population mean (see blue plot). Below are plots of the distribution of sample means and standard deviation when \( n = 10 \) and \( n = 40 \), see the spread reduces as \( n \) gets larger.
Example: 95% Confidence intervals

If $\sigma$ is known. The 95% CI is

$$\left[ \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right]$$

Below are the CIs using the sample standard deviation

- $n = 3, \ n - 1 = 2, \ t_{0.025}(2) = 4.303.$
  $$\left[ \bar{X} - 4.303 \frac{s}{\sqrt{3}}, \bar{X} + 4.303 \frac{s}{\sqrt{3}} \right].$$

- $n = 10, \ n - 1 = 9, \ t_{0.025}(9) = 2.262.$
  $$\left[ \bar{X} - 2.262 \frac{s}{\sqrt{10}}, \bar{X} + 2.262 \frac{s}{\sqrt{10}} \right].$$
\( n = 121, \ n - 1 = 120, \ t_{0.025}(120) = 1.98. \)

\[
[\bar{X} - 1.98 \frac{s}{\sqrt{121}}, \bar{X} + 1.98 \frac{s}{\sqrt{121}}].
\]

As the sample size grows the critical values in the t-distribution get closer to the critical values of the normal distribution (in this case 1.96).
Common misunderstandings

As the sample size gets large two completely different things happen:

- The distribution of the sample mean gets close to the normal distribution (lecture 11 and 12). This is called the central limit theorem.

- The sample standard error tends to get closer to the population standard deviation. This means the critical values of the t-distribution converge to those of a normal distribution.

The t-distribution and the fact that the critical values of a t-distribution get closer to those of a normal distribution has nothing to do with the central limit theorem.
Conditions for using a t-distribution

• Observations are from a Simple Random Sample.

• The sample mean is close to normally distributed.
Example: Comparing the mean number of M&Ms in a bag

We now analyse the M&M data to see whether the mean number of M&Ms in a bag vary according to the type of M&M. The data can be found here:

http://www.stat.tamu.edu/~suhasini/teaching651/MandMs_2013.csv

There is a proper formal method called ANOVA, which we cover in lecture 24, where we can check to see whether all three have the same mean or not. However, a crude method is to simply check their confidence intervals.
Lecture 14 (MWF) The t-distribution
Solution: Analysis and interpretation

- As the sample sizes used to construct each confidence interval are large (over 30 in each case), even though the distribution of M&Ms is not normal (they are integer valued!), it is safe to assume that the average is close to normal, therefore these 95% confidence intervals are reliably 95%.

- A summary of the output is given below:
  - Plain: sample mean = 17.2, standard error = 0.31, CI = [16.67, 17.92].
  - Peanut: sample mean = 8.6, standard error = 0.49, CI = [7.67, 9.76].
  - Peanut butter: sample mean = 10.9, standard error = 0.26, CI = [10.37, 11.45].

- As none of the confidence intervals (recalling that in this interval we believe the mean for each case should like) intersect our crude analysis suggests that the means are all different.
• In lecture 19 we will make the above precise (by constructing a confidence interval for the differences in the means).
Statistics in articles

This is a snapshot from the article on the influence of CO2 on diet by Eweis et al. (2017). Below are the glucose and cholesterol levels in rats after drinking only regular water, a sugar soda, diet soda and decarbonated sugar soda (for 6 months).

The table gives the [sample mean± sample standard deviation] for each group. In each group there are 4 rats. From these numbers, we can calculate the 95% confidence intervals for the population mean under each treatment.

<table>
<thead>
<tr>
<th>Group</th>
<th>Glucose (mg/dl)</th>
<th>Cholesterol (mg/dl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>157 ± 22</td>
<td>127 ± 3</td>
</tr>
<tr>
<td>RCB</td>
<td>187 ± 0.4</td>
<td>135 ± 3</td>
</tr>
<tr>
<td>DCB</td>
<td>172 ± 21</td>
<td>135 ± 5</td>
</tr>
<tr>
<td>DgCB</td>
<td>192 ± 5</td>
<td>138 ± 3</td>
</tr>
</tbody>
</table>
• When reading an article it is important to check if the ± is the margin of error (in which case the authors have given the confidence interval) of the sample standard deviation (in which case you need to construct the CI).

• In the article above the 95% confidence intervals for the mean level of water and RCB (regular soda) are

\[
\begin{align*}
157 & \pm 3.18 \times \frac{22}{\sqrt{4}} = [121, 192] \\
187 & \pm 3.18 \times \frac{0.4}{\sqrt{4}} = [186.3, 187.6].
\end{align*}
\]

• The intervals intersect, which means we have to cautious about saying that they have different treatment groups have different means.
• However, the variation between the two data sets is very different (22 vs 0.4), which suggests that there are differences in the populations. But we need to keep in mind that these are estimated using very small sample sizes.

• Warning: Comparing the confidence intervals of several treatment groups can lead to “false positives”. This is one reason we do ANOVA, which is a method for collectively comparing the means across groups. We cover this later on in the course.
IMPORTANT!!

- A common mistake that students make is that the t-distribution is used to correct for the non-normality of sample mean (for example when the sample size is not large enough).

- NOOOOOOOOOOOOOOOOOOOOOOOOOOOO

- In order to use the t-distribution we require that the sample mean is close to normal.

- THE ONLY REASON WE USE THE T-DISTRIBUTION is because the true population standard deviation is unknown and us estimated from the data. The t-distribution is used to correct for the error in the estimated standard deviation.
The t-distribution cannot correct for non-normality of the data

Here we draw a sample of size 10 from a right-skewed distribution and use the t-distribution to construct a confidence interval for the mean. We see that only 87% of the confidence intervals contain the mean. Using the