Lecture 14 (MWF) The t-distribution

Suhasini Subba Rao
Review of previous lecture

- Often the precision of an estimator is stated in terms of its 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 $[\bar{X} - E, \bar{X} + E]$.

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

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

- 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.
How estimating the standard deviation affects our results

- Underlying the work so far, we have assumed that the standard deviation $\sigma$ is known.

- This is sometimes a plausible assumption. For example if we want to compare the distances travelled last year with those of this year. Last year the mean distance travelled by a person was 2000 km and the standard deviation was 500 km. This year based on a sample of 50 people, the sample mean distance travelled was 2100 km. It would be reasonable to assume that the variance has not changed, only the mean may or may not have changed.

- However in general we will not have apriori knowledge of $\sigma$. $\sigma$ will be unknown.
• Given a data set $X_1, \ldots, X_n$ (say the 9 data values, 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.

• Recall we can estimate the variance using the sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

• **Constructing confidence intervals** It seems reasonable that we then replace $\sigma$ with $s$, to make $z$-transform or a 95% CI:

\[
\text{z-transform} \quad \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \rightarrow \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}}
\]

\[
\text{95% CI} \quad [\bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}] \rightarrow [\bar{X} \pm 1.96 \times \frac{s}{\sqrt{n}}]
\]
Is this valid?

Go to the applet [http://onlinestatbook.com/stat_sim/sampling_dist/index.html](http://onlinestatbook.com/stat_sim/sampling_dist/index.html) and draw a sample from the normal distribution. How well does the sample standard deviation estimate the true standard deviation for small samples?
The effect of estimating the standard deviation

- Recall $s$ is random it varies from sample to sample.

- If the sample size is relatively small it can often under estimate the true standard deviation. This can cause substantial problems when we evaluate the $z$-transform. Recall, the $z$-transform is the number of standard deviations between the mean and the sample mean. If the standard deviation has been underestimated, then the $z$-transform will be artificially larger than what it is suppose to be

\[
    z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \quad \text{(smaller)}
\]
In other words this estimated z-transform, which we call the t-transform

\[ t = \frac{\bar{X} - \mu}{s} \frac{s}{\sqrt{n}} \]

will have tend to have more extreme values than the standard normal (it has thicker tails).

• 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

\[ [\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}] \rightarrow [\bar{X} - 1.96 \times \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{s}{\sqrt{n}}]. \]
If $s$ is smaller than $\sigma$ than the interval will be too narrow for it correspond to 95% confidence.

- Both these arguments support the view than when we use the estimated standard deviation we need to correct for the fact that $s$ tends to underestimate the true standard deviation $\sigma$.

- Indeed, it is very simple to make the correction. All we need to do is change the distribution, go from a normal distribution to a t-distribution.

- Below we give some historical background and what this means.
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 $s^2 = \frac{1}{9} \sum_{i=1}^{10} (X_i - \bar{X})^2$. 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 true variance $\sigma^2$, he replaced it with the sample variance
\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. \]

For each sample of size 10 is constructed the CI he constructed the 95% CI:

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

He counted the number of times the true mean \( \mu \) was in the size 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.

- Hence the true mean is inside the interval less than 95% of the time.
This suggests that the interval is not long enough, and we need to use a longer interval for the 95% CI to be accurate.
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!
Introducing the t-distribution

• He investigated this further, he showed that by standardizing with the sample standard deviation, the transformation $\frac{\bar{X} - \mu}{(s/\sqrt{n})}$ did not have a standard normal distribution.

• Basically because the standard deviation has to be estimated it adds more randomness (uncertainty) into the system.

• More uncertainty means that the z-transform $\frac{\bar{X} - \mu}{(s/\sqrt{n})}$ is more likely to take large values (since the sample variance $s^2$ is random and tends to underestimate the true variance when the sample size is small).
An illustration: sample means and standard deviation

Here we draw a sample of size 10 from a normal distribution. For each sample we evaluate the sample mean and sample standard deviation.

You see that the sample mean is close to normal, but the sample standard deviation also has a distribution. Often the sample standard deviation underestimate the true variance. Therefore, when constructing a confidence interval we need to take the uncertainty associated with the sample standard deviation into account.
The $t$-distribution

- The standardisation with the **sample standard deviation** has the $t$-distribution

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

- $n$ is the number of observations used to estimate $\sigma^2$, eg $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

- Unlike the case where $\sigma^2$ is known and $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ has a standard normal distribution the distribution of $\frac{\bar{X} - \mu}{s / \sqrt{n}}$ depends on the sample size.

- $t(n - 1)$ is a distribution like the standard normal distribution. The main difference is for different $n$ we have a **different** distribution.
• We call $t(n - 1)$ the Student $t$-distribution with $(n - 1)$-degrees of freedom. We use Student, in honor of William Gosset (he wrote all his papers under the pseudonym Student).
What does this means for us?

- We can pretty much do everything was we did before, but when we estimate the variance we need to 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).

- Replace every standard normal with the t-distribution!

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

T-distribution with 5 degrees of freedom
Confidence intervals using the t-distribution

- We know when the variance $\sigma^2$ is known. The $(1 - \alpha)100\%$ CI is

$$\bar{X} - |z_{\alpha/2}| \sqrt{\frac{\sigma^2}{n}}, \bar{X} + |z_{\alpha/2}| \sqrt{\frac{\sigma^2}{n}}.$$  

- When the variance $\sigma^2$ is unknown, estimate it from the data $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ and use the CI

$$\bar{X} - |t_{\alpha/2}(n-1)| \sqrt{\frac{s^2}{n}}, \bar{X} + |t_{\alpha/2}(n-1)| \sqrt{\frac{s^2}{n}}.$$
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.
Solution 1: in JMP

Summary Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.5</td>
</tr>
<tr>
<td>Std Dev</td>
<td>2.5169426</td>
</tr>
<tr>
<td>Std Err Mean</td>
<td>0.8309809</td>
</tr>
<tr>
<td>Upper 95% Mean</td>
<td>7.4546933</td>
</tr>
<tr>
<td>Lower 95% Mean</td>
<td>3.5553067</td>
</tr>
<tr>
<td>N</td>
<td>9</td>
</tr>
</tbody>
</table>
**Solution 1: in JMP**

When the sample size is so small it is very hard to tell from the QQplot whether the data has come from a normal distribution. Instead we need to rely on our knowledge of how the data was collected. As a blood sample is a biological experiment, it seems plausible that its distribution does not have a 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. Under this assumption we can proceed with the analysis.
Solution 1: Red Wine

- There are two problems when constructing a confidence interval for the above data:
  
  - The sample size is small, therefore to construct a reliable confidence interval we need that the distribution of the blood samples is not too far from normally distributed (see the discussion on the previous page).
  
  - We do not know the standard deviation and have to estimate it from the data. We demonstrated using the applet that when the sample size is small, we are likely to under estimate the standard deviation. This means that we have under estimated the margin of error in the confidence interval. In turn this means that our 95% confidence using the normal (1.96) will not be 95% and the confidence level is less than 95% (our interval is not 95% reliable).
• To address the second issue we use the t distribution instead of the normal distribution.

• The calculation: t-tables with 8df (sample size, 9, minus one) we get a 95% CI for the mean which is \([5.5 \pm 2.306 \times \frac{2.517}{\sqrt{9}}] = [3.57, 7.43]\).
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

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

• Observe that the factor 2.145 has decreased from the 2.306 given in the previous example. This is because, the sample standard deviation based on \( n = 15 \) will tend to be closer to the population standard deviation, hence we don’t require such a large interval to be 95% confident the CI captures the mean.
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). The less spread in the smaller sample size means that the estimated standard deviation using the smaller sample will be smaller than the second sample (look at the output below). We see that for smaller sample sizes the
estimated standard deviation tends to underestimate the true population standard deviation.

- What this means, is that if we were to WRONGLY USE 1.96 (z-value at 2.5% level using normal tables) to construct a 95% confidence interval, the interval would be much narrower than the desired 95% confidence level.
• We correct for this problem by using the t-distribution instead of the normal distribution.

• However, for large sample sizes, the estimated standard deviation is likely to be closer to the true populations standard deviation, therefore we do not need to correct so much for the underestimation of the standard deviation (this is why we use the factor 2.145 (2.5% from t-tables with 14df) is used rather than 2.306 (2.5% from t-tables with 8df)).
Comparing t and normal distributions

Two times the area to the RIGHT of 1.96, tells is the exact confidence we have in the interval when we use 1.96 instead of the correct t-value/
Toy example: using the wrong distribution

Consider the data set 4, 5.5, 6, my estimate of the mean is \( \bar{x} = 5.17 \) and the estimated standard deviation is \( s = \sqrt{\frac{1}{2}[(4 - 5.17)^2 + (5.5 - 5.17)^2 + (6 - 5.17)^2]} = 1.04 \). With just three points it is highly likely the true standard deviation has been substantially underestimated.

• Suppose we ignore the fact that the sample standard deviation has been estimated and use instead the normal factor to construct the confidence interval. Our incorrect 95% CI is

\[
[5.17 - 1.96 \times \frac{1.04}{\sqrt{3}}, 5.17 + 1.96 \times \frac{1.04}{\sqrt{3}}].
\]
• On two slides previous, we showed that if we use 1.96 as the critical value, then by going to the distribution with 2df, the areas to the right of 1.96 is 9.45% (not 2.5%). This means that the interval we have constructed above is not a 95% confidence interval, but a $100 - 2 \times 9.45 = 80.1\%$ confidence interval. There is a lot less confidence in this interval containing the mean then we thought we had.

• We can see from the next slides that if we want to have 95% confidence in the interval containing the mean, we need to use the interval

$$[5.17 - 4.3 \times \frac{1.04}{\sqrt{3}}, 5.17 + 4.3 \times \frac{1.04}{\sqrt{3}}].$$

• This is a lot wider than we than the previous incorrect interval.
Sample size and the sample standard deviation

As the sample size grows, not both the spread of the sample mean and the sample standard deviation decrease (less spread means they are more likely to be close to the population parameters). Below we give the spread of the sample standard deviation when $n = 10$ and $n = 40$, see the spread reduces as $n$ gets larger.
The t-distribution at 2.5% percentile for different $n$

As the spread of the sample variance decreases, we see that the t-values get closer to 1.96 as $n$ grows.
Why the distribution depends on the sample size?

• Consider the following situation. If you were to estimate the variance using 100 observations, you would expect it to be far better than an estimator calculated using 10 observations.

  The idea here is exactly the same as expecting the estimator of the mean involving 100 observation to be far better than an estimator involving 10 observations.

• Therefore when the sample size is small, the sample variance is more likely to underestimate the true population variance than when the sample is large.

• Hence when the sample size is small, it is reasonable to suppose that the distribution of \( \frac{\bar{X}_{10} - \mu}{s_{10}/\sqrt{10}} \) is ‘more’ random than when the sample size is large.
large \( \frac{\bar{X}_{100} - \mu}{\sqrt{\frac{s^2_{100}}{100}}} \). Therefore, the t-distribution for smaller sample sizes will have thicker tails (more likely to obtain extreme values) than for large samples. This manifests in larger t-values at the same level.
Example: 95% Confidence intervals

- We know when the variance $\sigma^2$ is known. The 95% CI is $[\bar{X} - 1.96\sqrt{\frac{\sigma^2}{n}}, \bar{X} + 1.96\sqrt{\frac{\sigma^2}{n}}]$.

- When the variance is estimated, the 95% CI is
  - $n = 3, \ n - 1 = 2, \ t_{0.025}(2) = 4.303$
    The CI is $[\bar{X} - 4.303\sqrt{\frac{s^2}{3}}, \bar{X} + 4.303\sqrt{\frac{s^2}{3}}]$.
  - $n = 10, \ n - 1 = 9, \ t_{0.025}(9) = 2.262$
    The CI is $[\bar{X} - 2.262\sqrt{\frac{s^2}{10}}, \bar{X} + 2.262\sqrt{\frac{s^2}{10}}]$.
  - $n = 30, \ n - 1 = 29, \ t_{0.025}(29) = 2.045$
    The CI is $[\bar{X} - 2.045\sqrt{\frac{s^2}{30}}, \bar{X} + 2.045\sqrt{\frac{s^2}{30}}]$.
  - $n = 121, \ n - 1 = 120, \ t_{0.025}(120) = 1.98$
    The CI is $[\bar{X} - 1.98\sqrt{\frac{s^2}{121}}, \bar{X} + 1.98\sqrt{\frac{s^2}{121}}]$. 
The t-distribution and sample size

- We know that \(|z_{0.025}| = 1.96\). For the t-distribution we have
  - \(n = 3, \ n - 1 = 2, \ t_{0.025}(2) = 4.303\).
  - \(n = 10, \ n - 1 = 9, \ t_{0.025}(9) = 2.262\).
  - \(n = 30, \ n - 1 = 29, \ t_{0.025}(29) = 2.045\).
  - \(n = 121, \ n - 1 = 120, \ t_{0.025}(120) = 1.98\).

- We see as \(n\) gets larger the value of \(t_{0.025}(n)\) gets closer to \(|z_{0.025}| = 1.96\).
  In fact for \(n > 50\) we generally don’t use the t-distribution, but instead approximate this by the normal distribution.
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).
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).

• NOOOOOOOOOOOOOOOOOOOOOOOOOOOOO

• 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