Lecture 10: $t$-Test

S. Massa, Department of Statistics, University of Oxford

20 January 2016
Motivation

In lecture 7 the heights of 198 British men were analysed.

\[ X_1, \ldots, X_n \sim N(\mu, \sigma^2). \]

From the data

\[ \bar{x} = 1732\text{mm} \quad s = 68.8\text{mm}. \]

We want to construct a confidence interval for the true average height \( \mu \).

Recall the \( z \)-statistic computed from the sample

\[ Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \]

If you pay close attention then you may notice that we do not know the true population variance \( \sigma \).
Motivation

We do not know the population variance $\sigma$, we simply replaced it with the sample variance $s$.

Notice that the sample variance of a sample $X_1, \cdots, X_n$ is given by

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

This is only an estimate of $\sigma$: due to chance variation it sometimes is too low or too high.

So the quantity that we computed for $Z$, is not really $Z$ but something else

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$
Motivation

\[ Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, \quad T := \frac{\bar{X} - \mu}{S/\sqrt{n}}. \]

- On average the two are not too different.
- But the numerator of \( T \) is a random variable: the distribution of \( T \) will be more dispersed than that of \( Z \).
- This implies that you underestimate probabilities of extreme observations, or that you reject a true null hypothesis more often than dictated by the confidence level.
- As a result any confidence intervals you compute will be too narrow.
Motivation

\[ T := \frac{\bar{X} - \mu}{S/\sqrt{n}}. \]

- So if we don’t know the population variance \( \sigma \), we cannot compute \( Z \) but instead we compute \( T \).
- Is this a problem?
- \( T \) is also computed from the data, so it is a statistic.
- But for \( T \) to be an useful statistic, we need to know its distribution.
Student’s t-Distribution

- William Gossett computed the distribution of the $t$-statistic while working for the Guinness brewery, trying to choose the best yielding barley variety—he was concerned with small sample sizes.
- He published it under the pseudonym Student, as it was deemed confidential information by the brewery.
- The $t$-distribution has a single parameter called the number of degrees of freedom—this is equal to the sample size minus 1.
- For large samples, typically more than 50, the sample variance is very accurate.
- In this situation, for all practical reasons, the $t$-statistic behaves identically to the $z$-statistic.

**Assumption:** $X_1, \ldots, X_n$ are independent samples from $N(\mu, \sigma^2)$ and hence $\bar{X}$ is approximately normal.
Probability density function

Figure: Probability density functions of $t$-distributions with 1, 2, 5, 10 and 50 degrees of freedom and standard normal.
Table of $t$

**Table:** Critical values at the 95% level, of the $t$-statistic. For very large number of degrees of freedom, they are identical to the $Z$-statistic.

<table>
<thead>
<tr>
<th>degrees of freedom</th>
<th>critical value</th>
<th>degrees of freedom</th>
<th>critical value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.31</td>
<td>1</td>
<td>12.7</td>
</tr>
<tr>
<td>2</td>
<td>2.92</td>
<td>2</td>
<td>4.30</td>
</tr>
<tr>
<td>4</td>
<td>2.13</td>
<td>4</td>
<td>2.78</td>
</tr>
<tr>
<td>10</td>
<td>1.81</td>
<td>10</td>
<td>2.23</td>
</tr>
<tr>
<td>50</td>
<td>1.68</td>
<td>50</td>
<td>2.01</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.64</td>
<td>$\infty$</td>
<td>1.96</td>
</tr>
</tbody>
</table>

(a) one-tailed test  
(b) two-tailed test

Do the bottom values of the table look familiar?
Confidence Intervals: One-Sample

- Suppose you have observations: $x_1, \ldots, x_n$ from a Normal distribution $N(\mu, \sigma^2)$.
- Both $\mu$ and $\sigma^2$ are unknown and you want to compute a confidence interval for $\mu$.
- Look up the $t$-table and find the appropriate number of degrees of freedom.
- For two-sided confidence interval $t$ is such that $P(T < t) = .975$.
- The confidence interval is then given by:

$$
(\bar{x} - t \times \frac{s}{\sqrt{n}}, \bar{x} + t \times \frac{s}{\sqrt{n}}).
$$
Example: Heights of British Men

- For the heights of British men example we used the distribution of the $Z$-statistic.
- We should have used the $t$-distribution with $198-1=197$ degrees of freedom.
- Due to the large sample size, the critical values are almost identical to those of the normal distribution, so the error in our confidence interval was not very large.
Example 1

The level of phosphate, in mg/dl in the blood of a patient undergoing dialysis treatment was measured on six consecutive visits.

\[ 5.6, 5.1, 4.6, 4.8, 5.7, 6.4. \]

Construct a symmetric 99% confidence interval.

The sample size is \( n = 6 \). We can compute the sample mean and sample variance as follows:

\[
\bar{x} = \frac{1}{6}(5.6 + 5.1 + 4.6 + 4.8 + 5.7 + 6.4) = 5.4 \text{mg/dl}
\]

\[
s^2 = \frac{1}{5}(5.6 - 5.4)^2 + (5.1 - 5.4)^2 + (4.6 - 5.4)^2
\]

\[
+ (4.8 - 5.4)^2 + (5.7 - 5.4)^2 + (6.4 - 5.4)^2
\]

\[
= (0.67 \text{mg/dl})^2.
\]
The number of degrees of freedom is \( n - 1 = 5 \).
Thus, the symmetric confidence interval will be

\[
\left( 5.4 - t \frac{0.67}{\sqrt{6}}, 5.4 + t \frac{0.67}{\sqrt{6}} \right) \text{ mg/dl},
\]

where \( t \) is chosen so that the \( T \) variable with 5 degrees of freedom has probability 0.01 of being bigger than \( t \).
We want to find $t$ so that $P(-t \leq T \leq t) = 0.99$.

We look at the official table and locate the row corresponding to 5 degrees of freedom.

The critical value is in the final column and is $t = 4.03$.

Using the $t$-Table

![Diagram showing the probability $P$ of lying outside $\pm t$ with critical value $t = 4.03$.]

<table>
<thead>
<tr>
<th>d.f.</th>
<th>$P=0.10$</th>
<th>$P=0.05$</th>
<th>$P=0.02$</th>
<th>$P=0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.31</td>
<td>12.71</td>
<td>31.82</td>
<td>63.7</td>
</tr>
<tr>
<td>2</td>
<td>2.92</td>
<td>4.30</td>
<td>6.96</td>
<td>9.93</td>
</tr>
<tr>
<td>3</td>
<td>2.35</td>
<td>3.18</td>
<td>4.54</td>
<td>5.84</td>
</tr>
<tr>
<td>4</td>
<td>2.13</td>
<td>2.78</td>
<td>3.75</td>
<td>4.60</td>
</tr>
<tr>
<td>5</td>
<td>2.02</td>
<td>2.57</td>
<td>3.36</td>
<td>4.03</td>
</tr>
<tr>
<td>6</td>
<td>1.94</td>
<td>2.45</td>
<td>3.14</td>
<td>3.71</td>
</tr>
</tbody>
</table>
Example: Kidney Dialysis I

- Having identified the critical value to be 4.03 we compute the confidence interval as

  \[(4.3 \text{ mg/dl}, 6.5 \text{ mg/dl})\].

- If we repeat the experiment 100 times then the true mean would be in this interval 99 times.

- If we had used the \(Z\)-distribution, the critical value would have been 2.6 resulting in the much narrower interval

  \[(4.7 \text{ mg/dl}, 6.1 \text{ mg/dl})\].

- Here you can immediately see that in small samples the difference between \(Z\) and \(T\) is significant.
Continuing with the kidney dialysis example, suppose that 4 mg/dl is a dangerously low phosphate level. We want to test this hypothesis at the 0.99 significance level.

**Two-sided test**

\( H_0 : \mu = 4.0 \) mg/dl,
\( H_1 : \mu \neq 4.0 \) mg/dl,

Under the null hypothesis \( \mu = 4.0 \), and therefore

\[
T := \frac{\bar{X} - 4.0}{S/\sqrt{n}} \sim t_5, \hspace{1em} \text{(under the null hypothesis)}
\]

has the \( t \)-distribution with \( n - 1 = 5 \) degrees of freedom.
We compute

\[
    t_{\text{obs}} = \frac{\bar{x} - 4.0}{s/\sqrt{n}} = \frac{5.4 - 4.0}{0.67/\sqrt{6}} = 5.12.
\]

At the 99% level, we already know the theoretical critical value to be 4.03. Since the observed value is \( t_{\text{obs}} = 5.12 > 4.03 \) we reject the null hypothesis.
One-sided test

$H_0 : \mu = 4.0 \text{ mg/dl}$,
$H_1 : \mu > 4.0 \text{ mg/dl}$,

In this case we look for the one-sided critical value. The table gives critical values for the two-sided test, so we need the value of $t$ such that $P(-t < T < t) = 1 - 2 \times 1\% = 0.98$.

The theoretical value is $t = 3.36$. From the structure of our alternative hypothesis, we reject the null if the observed value $t_{obs}$ is more than 3.36. With the one-sided test we would have been more likely to reject the null hypothesis.
For testing and constructing a confidence interval for the true mean of a population:

- **Know $\sigma$?**
  - No: Estimating a proportion?
  - Yes: Large sample?

- Estimating a proportion?
  - No: Use $Z$
  - Yes: Large sample?
    - Yes: Use $Z$
    - No: Use $T$

**Figure:** Flowchart for deciding whether to use $Z$ or $T$. 
Paired $t$-Test

The percentage of aggregated blood platelets in the blood of 11 individuals were measured before and after they smoked a cigarette.

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td>27</td>
<td>37</td>
<td>10</td>
</tr>
<tr>
<td>44</td>
<td>56</td>
<td>12</td>
</tr>
<tr>
<td>30</td>
<td>46</td>
<td>16</td>
</tr>
<tr>
<td>67</td>
<td>82</td>
<td>15</td>
</tr>
<tr>
<td>53</td>
<td>57</td>
<td>4</td>
</tr>
<tr>
<td>53</td>
<td>80</td>
<td>27</td>
</tr>
<tr>
<td>52</td>
<td>61</td>
<td>9</td>
</tr>
<tr>
<td>60</td>
<td>59</td>
<td>-1</td>
</tr>
<tr>
<td>28</td>
<td>43</td>
<td>15</td>
</tr>
</tbody>
</table>

It seems there was more clotting after the cigarette. Can we test this hypothesis?
Paired $t$-test: Standard Error

Suppose that a random individual has a normally distributed “Before” score $X_i$. Smoking a cigarette adds an independent random, normally distributed effect $D_i$. We want to know if $D_i$ tends to be positive on average. So we want to compare the sample mean of the $D_i$s to their standard error. Notice that the standard error we should use is that of the $D_i$’s not of the $X_i$’s. That would be much greater since

$$\text{var}(X + D) = \text{var}(X) + \text{var}(D) > \text{var}(D).$$
Paired sample $t$-test: Standard Error

$H_0 : \mu_d = 0$

$H_1 : \mu_d > 0$

Sample mean and sample standard deviation of the differences are

$\bar{d} = 10.3, \quad s_d = 7.98.$

The observed value of the $T$-statistic is

$t_{obs} = \frac{10.3 - 0}{7.98/\sqrt{11}} = 4.28.$

We are doing a one-sided test.

We want $P(T > t) = .05$. Since the table is giving us $t$ so that $P(|T| > t) = \alpha$, we want the $t$ that corresponds to $2 \times 0.05 = 0.1.$
Paired $t$-Test

Now there are 11 measurements, so the are 10 degrees of freedom. So look at first column.

The theoretical value is 1.81. The observed value is 4.28 ($> 1.81$) therefore we reject the null hypothesis.

<table>
<thead>
<tr>
<th>d.f.</th>
<th>$P=0.10$</th>
<th>$P=0.05$</th>
<th>$P=0.02$</th>
<th>$P=0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.31</td>
<td>12.71</td>
<td>31.82</td>
<td>63.7</td>
</tr>
<tr>
<td>2</td>
<td>2.92</td>
<td>4.30</td>
<td>6.96</td>
<td>9.93</td>
</tr>
<tr>
<td>3</td>
<td>2.35</td>
<td>3.18</td>
<td>4.54</td>
<td>5.84</td>
</tr>
<tr>
<td>4</td>
<td>2.13</td>
<td>2.78</td>
<td>3.75</td>
<td>4.60</td>
</tr>
<tr>
<td>5</td>
<td>2.02</td>
<td>2.57</td>
<td>3.36</td>
<td>4.03</td>
</tr>
<tr>
<td>6</td>
<td>1.94</td>
<td>2.45</td>
<td>3.14</td>
<td>3.71</td>
</tr>
<tr>
<td>7</td>
<td>1.90</td>
<td>2.37</td>
<td>3.00</td>
<td>3.50</td>
</tr>
<tr>
<td>8</td>
<td>1.86</td>
<td>2.31</td>
<td>2.90</td>
<td>3.36</td>
</tr>
<tr>
<td>9</td>
<td>1.83</td>
<td>2.26</td>
<td>2.82</td>
<td>3.25</td>
</tr>
<tr>
<td>10</td>
<td>1.81</td>
<td>2.23</td>
<td>2.76</td>
<td>3.17</td>
</tr>
</tbody>
</table>

Conclusion: There is a significant increase in blood clotting after smoking a cigarette.
The critical region
Paired $t$-Test: Discussion

- Essentially we compared the sample means of two samples.
- Our goal was to understand if the true mean of the first sample was greater than the true mean of the second.
- In the next lecture we will see more about comparing the means and distributions of two samples.
- In the paired test: the data is structured in pairs.
- This will not always be the case.
- We will see that this experiment design results in more effective hypothesis testing (power of a test).
Does Population Size Matter?

- In our analysis of the heights of 198 married men we ignored the population size.
  
  *What if these 198 men were all the men in the UK?*

- In that case there would be no sampling error at all.

- What if the total population were 300 or 400? There should be less error than if there were 20 million.

- Indeed the sampling error does depend on the size of the population, but this effect decays very quickly as the population grows.
Suppose that we sample $n$ from a population of $N$.

So far we have used the following formula for the standard error:

$$SE = \text{var}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$ 

This is based on the premise that we are sampling from an infinite population.

Usually sampling is performed from a finite population and without replacement.

In this case, if a significant proportion of the population $> 5\%$ is sampled we need to use the correction factor

$$\text{correction factor} = \sqrt{\frac{N - n}{N - 1}},$$

where $N$ is the population size and $n$ the sample size.
Correction Factor

So the standard error becomes

$$SE = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}.$$ 

Let’s have a look at how the correction factor behaves with a fixed sample size of 200.

<table>
<thead>
<tr>
<th>N</th>
<th>200</th>
<th>300</th>
<th>1000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corr. Factor</td>
<td>0</td>
<td>0.5783</td>
<td>0.8949</td>
<td>0.9899</td>
</tr>
</tbody>
</table>
Confidence Intervals from Small Populations

- How does this affect our confidence intervals?
- In the heights of British men example, we calculated
  \[ \bar{x} = 1732 \text{m.m.} \quad s = 68.8 \text{m.m.} \]

- Assuming a large population we computed the confidence interval as
  \[ 1732 \pm 1.96 \times \frac{68.8}{\sqrt{198}} = (1722, 1742) \text{mm} \]

- Let’s see how this changes if we now assume that these were sampled from a village with total population 300
  \[ 1732 \pm 1.96 \times \frac{68.8}{\sqrt{198}} \sqrt{\frac{102}{299}} = (1726, 1738) \text{mm}, \]
  a narrower confidence interval.
Measurement Bias vs Random error

- All measurements are prone to error.
- We can split this error to two types
  1. measurement bias, or systematic error, and
  2. random error.

Definition 2 (Random error)

This is the part of the error that is due to chance fluctuations. Random error is on average zero, and as the sample size increases its contribution vanishes.
Measurement Bias vs Random error

Example 3
Suppose that you are measuring length with a metal stick. The stick’s length fluctuates due to changes in temperature. But this should on average cancel out, as sometimes the temperature will be too high, and sometimes too low. Also purely by chance, sometimes your measurements will be too high, and sometimes too low.
Definition 4 (Bias)

Bias, or systematic error, is not due to chance alone, but rather due to the measuring procedure itself. Bias cannot be removed by increasing the sample size, but only by improving the measurement procedure.

Example 5

Example: suppose that you are measuring length with a stick which is 99cm long, but you think it’s 1m long. All your measurements will be off by 1cm /m before adding random variation to it.

Although random errors ”tend” to be normally distributed, and are fairly well understood, bias is application specific.
Statisticians often use surveys to gather data about the general population. These suffer from the following types of bias:

**Selection bias:** You are not sampling from a truly random subset of the population.
How do you pick a random sample of 1000 people from the 64 million people in the UK?

**Ascertainment bias:** You are not sampling from a representative subset of the population.
In medical sciences, subjects can not be identified if they are not diagnosed. Sometimes mild cases of diseases are not diagnosed and therefore are systematically under-represented in surveys.
Bias in surveys II

Non-response bias: *The subjects who choose to respond to the survey may not represent the general population.* Example: when polling about some type of illegal behaviour, people engaging in it could refrain from responding.

Response bias: *Subjects may choose to give answers they perceive as more acceptable.* Example: how many racists will declare themselves so in a poll?
Recap

- When estimating a population mean, if we don’t know $\sigma$ and estimate it using $S$ we use the $t$-statistic,

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

- This has the $t$-distribution: a family of distributions parameterised by number of degrees of freedom

$$d.f. = n - 1, \quad n \text{ is the sample size}.$$

- Use $t$-table for $(1 - \alpha) \times 100\%$ confidence intervals

$$\overline{x} \pm t_{1-\alpha/2} \frac{s}{\sqrt{n}}, \quad \text{where} \quad P(|T| < t_{1-\alpha/2}) = 1 - \alpha.$$
Recap: $Z$ or $T$?

For testing and constructing a confidence interval for the true mean of a population:

- **know $\sigma$?**
  - no → **estimating a proportion?**
    - no → **large sample?**
      - yes → **Use $T$**
      - no → **Use $Z$**
    - yes → **Use $Z$**
  - yes → **Use $Z$**

*Figure: Flowchart for deciding whether to use $Z$ or $T$.***
Recap

- When estimating a population mean, if we don’t know \( \sigma \) and estimate it using \( S \) we use the \( t \)-statistic,

\[
T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.
\]

- This has the \( t \)-distribution: a family of distributions parameterised by number of degrees of freedom

\[
d.f. = n - 1, \quad n \text{ is the sample size.}
\]

- Use \( t \)-table for \((1 - \alpha) \times 100\%\) confidence intervals

\[
\bar{x} \pm t_{1-\alpha/2} \frac{s}{\sqrt{n}}, \quad \text{where } P(|T| < t_{1-\alpha/2}) = 1 - \alpha.
\]
Recap

- In the paired $t$-test, the data are naturally structured in pairs.
- Bias is the systematic error in our measurements that cannot be removed without changing the procedure.
- Surveys suffer from:
  - Selection/Ascertainment bias;
  - Non-response bias;
  - Response bias.