Lecture 19  Small samples and t tests

In previous lecture, either

- population standard deviation was known or

- the sample size was sufficiently large ($n \geq 20$)
  $\sigma$ could be replaced by its large sample estimate, $\hat{\sigma} = s$

- Central Limit Theorem indicates that the distribution of means and differences of means is reliably approximated by the normal distribution.

In many cases the population standard deviation is unknown and the sample size is $< 20$.
The sample standard deviation ($s$) is used but the distribution of the mean or differences has to take into account that this estimate has itself a component of uncertainty.

**t-distribution**

The t-distribution is similar to the standardized normal distribution, $Z \sim N(0, 1)$ but differs because

(i) the variance is $> 1$,

(ii) the shape of the distribution depends upon the sample size ($n$) or more specifically the degrees of freedom ($df$) for the variance ($n - 1$).

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

The t test

$t_{\alpha, n-1}$, read as “$t$ sub alpha on 2 with $n - 1$ degrees of freedom”
Example 19.1

An accepted belief is that the mean of a population is \( \mu_0 = 6.5 \). A small sample \((n = 10)\) is taken. The sample mean and standard deviation are \( \bar{X} = 3.15 \) and \( s = 1.90 \). Does the sample support the belief?

Calculate,

\[
T = \frac{\mu_0 - \bar{X}}{s/\sqrt{n}} = \frac{6.5 - 3.15}{1.9/\sqrt{10}} = 5.6
\]

The \( 2\% \) and \( 97\% \)’iles are found by:-

\[
> \text{qt}(p=c(0.025,0.975),df=9)
\]
\[
[1] \quad -2.262157 \quad 2.262157
\]

\( P(t > T) \) is found by:-

\[
> \text{pt}(q=5.6,df=9,lower.tail=F)
\]
\[
1] \quad 0.00018
\]

In Rcmdr use the menu of the t-distribution for calculating quantiles from probabilities or use the tail probabilities menu.

A union official would like to establish that the average salary is less than \$35,500.
A random sample of 15 has a mean salary of \$34,886 with a standard deviation of \$845.
Test to establish whether the claim is correct at 5% significance.
The hypothesis is that there is a pseudo-population with \( \mu = 35500 \) and the population in the study will have a \( \mu_0 \) (estimated by \( \bar{X} \)) which is significantly less than this.
The interest is only if it be less so the test is one-tailed.

\[
T_x = \frac{34886 - 35500}{845/\sqrt{15}} = -2.8
\]

From R, \( t_{0.05,14} = -1.76 \)
( qt(\( p=0.05, df=14 \) ) )
An official claims that average commuting distance is 36km. The group disagree and a random sample of 15 yielded a mean of 35km with sd=5km. Test the claim at the 5% level of significance. This is a 2-tailed test since because the disagreement is not whether is more or less but whether it is not equal; could be more or could be less.

\[ T_x = \frac{36 - 35}{\frac{5}{\sqrt{15}}} = 0.77 \]

This is not as extreme as \( t_{0.025,14} \) quantile so conclude there is no evidence against the claim. Quick R calculations are:-

```r
Tx <- (36-35)/(5/sqrt(15))
> qt(p=c(0.025,0.975),df=14)
[1] -2.144787  2.144787
> print(pt(q=Tx,df=0.975,lower.tail=F))
[1] 0.2915507
```

A summary of the t-test for a population mean

Assumptions - (1) Normal population or large sample, (2) \( \sigma \) unknown.

1. The null hypothesis is whether the mean of a sample (\( \mu \)) could have arisen from a population with mean \( \mu_0 \).
   This is written as \( H_0: \mu = \mu_0 \).
   The true mean of the sample is unknown and is replaced by the sample mean \( \bar{X} \).
   The question is whether the observed difference \( \mu_0 - \bar{X} \) is a random effect or is due to systematic differences between the populations.
   We find the probability that the observed difference could be due to random effects using the \( t \)-distribution.
   The alternate hypothesis is one of

   \[ H_a: \mu \neq \mu_0 \quad H_a: \mu < \mu_0 \quad H_a: \mu > \mu_0 \]  
   (2 tailed) (left tailed) (right tailed)

   Decide on the significance \( \alpha \). Suggested is \( \alpha = 0.05 \).
   The critical values are:-
Compute the test statistic,

\[ T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \]

If the value of \( T \) falls in the rejection region, reject \( H_0 \); otherwise do not reject \( H_0 \).

State the conclusion in words.

Suppose that in Example 19.1, the data were \( 1.88 \ 2.54 \ 6.12 \ 3.14 \ 3.26 \ 6.43 \ 3.92 \ 0.47 \ 1.63 \ 2.11 \).

A 2 tailed test of whether the mean of this sample differs from 3 is found using \texttt{t.test()} as follows.

```r
#------- ttest.r ------
options(digits=2)
library(ctest)
x <- c(1.88,2.54,6.12,3.14,3.26,6.43,3.92,0.47,1.63,2.11)
print(t.test(x,mu=6.5,alternative="two.sided"))
```

```r
> source("ttest.r")

One Sample t-test
data:  x
t = -5.6, df = 9, p-value = 0.0003546
alternative hypothesis: true mean is not equal to 6.5
95 percent confidence interval:
1.8 4.5
sample estimates:
mean of x
3.1
```

In Rcmdr,

- **Data → New data set**. Name this data set as \texttt{tdata1} (say).

- Enter the data. Change the variable name to \texttt{X} (say). [Quit] when all data type in.

- **Statistics → Means → Single-sample t test**. For this examples, the default values \( \mu_0 = 0 \) and Confidence Level = 0.95 are what is required.
#________ ttest2.r ______
options(digits=2)
library(ctest)
v <- c(33694,36762, 33682, 35067, 35450, 33942, 34282, 34870,
      35906, 34995, 34664, 34283, 35543, 35203, 34949)
print(t.test(v,mu=35500,alternative="less"))

> source("ttest2.r")

One Sample t-test
data: v
t = -2.8, df = 14, p-value = 0.006906
alternative hypothesis: true mean is less than 35500
95 percent confidence interval:
   -Inf 35270
sample estimates:
  mean of x
  34886

In Rcmdr, proceed as in Example ?? . Check the option Population mean < µ₀.

Confidence Intervals

The sample confidence interval for the population mean µ is:-

\[
\bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}
\]

A random sample of 16 has a mean salary of $43,000 with a standard deviation of $1,000. Construct a 95% confidence interval for the true mean salary.

> Q <- qt(p=c(0.025,0.975),df=15)
> Q
[1] -2.131450 2.131450
> CI <- 43 + round(Q*1/4,2)
> print(CI)
[1] 42.47 43.53
Lecture 20  

Differences between 2 independent sample means

The $t$ distribution is used to make statistical inferences regarding whether the means from 2 random samples are significantly different or not.

The null and alternate hypotheses are:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

The $t$ test for this is:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\text{s.e.}(\bar{X}_1 - \bar{X}_2)}$$  \hspace{1cm} (20.1)

Example from an industrial experiment.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample</th>
<th>(X)</th>
<th>(s^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1</td>
<td>89.7</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>81.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>84.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>84.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>87.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>79.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>85.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>81.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>83.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>84.5</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>84.7</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>86.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>83.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>91.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>86.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>79.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>82.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>89.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>83.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>88.5</td>
<td></td>
</tr>
</tbody>
</table>

Calculation of \(s(X_1 - X_2)\)

- \(\text{var}(\bar{X}) = \frac{\sigma^2}{n}\)
- \(\text{var}(X_1 - X_2) = \text{var}(X_1) + \text{var}(X_2)\) if \(X_1\) and \(X_2\) are independent.

If the 2 samples are independent, \(\bar{X}_1\) and \(\bar{X}_2\) are independent and

$$\text{var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$  \hspace{1cm} (20.2)

In equation (20.2), \(\sigma_1^2\), \(\sigma_2^2\) would be replaced by their sample estimates \(s_1^2\) and \(s_2^2\).

If \(s_1^2 \approx s_2^2\),

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$  \hspace{1cm} (20.3)

$$s^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1,i} - \bar{X}_1)^2$$
the numerator of (20.3) is

$$\sum_{i=1}^{n_1} (X_{1,i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2,i} - \bar{X}_2)^2$$
$n_1 - 1$ and $n_2 - 1$ degrees of freedom from samples 1 and 2 respectively.

Replacing $\sigma_1^2$ and $\sigma_2^2$ in (20.2) by the pooled variance gives

$$\text{var}(\bar{X}_1 - \bar{X}_2) = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

and

$$\text{s.e.}(\bar{X}_1 - \bar{X}_2) = s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (20.4)$$

If $n_1 = n_2 = n$, (20.3) becomes $s_p^2 = \frac{s_1^2 + s_2^2}{2}$ and (20.5) simplifies to

$$\text{s.e.}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2 + s_2^2}{2}} \times \sqrt{\frac{2}{n}}$$

As the ratio is < 1 meaning that the difference is less than 1 standard error of the difference, it is obvious that there is insufficient information to conclude a difference between the means exists.

From R, $P(t > 0.88) = 0.195$ and by symmetry $P(t < -0.88) = 0.195$. Hence there is a $p = 0.39$ probability that the observed difference could have happened by chance.

$> \text{pt}(q=0.88,\text{df}=18,\text{lower.tail}=.F)$

[1] 0.1952

industrial <- expand.grid(batch=1:10,proc=c("S","M"))
industrial$yields <- c(89.7,81.4,84.5,84.8,87.3,79.7,85.1,81.8,83.7,84.5,
                      84.7,86.1,83.2,91.9,86.3,79.3,82.6,89.1,83.7,88.5)
boxplot(yields ~ proc,data=industrial,las=1)

Alternatively, you could make a data file called industrial.txt and use Rcmdr to Import the data followed by the boxplot. The data file would resemble this:-

<table>
<thead>
<tr>
<th>batch</th>
<th>proc</th>
<th>yields</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S</td>
<td>89.7</td>
</tr>
<tr>
<td>2</td>
<td>S</td>
<td>81.4</td>
</tr>
<tr>
<td>3</td>
<td>S</td>
<td>84.5</td>
</tr>
<tr>
<td>4</td>
<td>S</td>
<td>84.8</td>
</tr>
<tr>
<td>5</td>
<td>S</td>
<td>87.3</td>
</tr>
<tr>
<td>6</td>
<td>S</td>
<td>79.7</td>
</tr>
<tr>
<td>7</td>
<td>S</td>
<td>85.1</td>
</tr>
<tr>
<td>8</td>
<td>S</td>
<td>81.7</td>
</tr>
<tr>
<td>9</td>
<td>S</td>
<td>83.7</td>
</tr>
<tr>
<td>10</td>
<td>S</td>
<td>84.5</td>
</tr>
<tr>
<td>1</td>
<td>M</td>
<td>84.7</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>86.1</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>83.2</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>81.9</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>86.3</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>79.3</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
<td>82.6</td>
</tr>
<tr>
<td>8</td>
<td>M</td>
<td>89.1</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
<td>83.7</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>88.5</td>
</tr>
</tbody>
</table>
> source("ttestdiff.r")

Welch Two Sample t-test
data: S and M
t = -0.8816, df = 17.129, p-value = 0.3902
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-4.409417 1.809417
sample estimates:
mean of x  mean of y
84.24  85.54

If you use Rcmdr,
• Make \texttt{industrial} the active data set.
• \texttt{Statistics} \rightarrow \texttt{Means} \rightarrow \texttt{Independent samples t-test}.
• Select \texttt{yields}
• Check Assume equal variances

![Diagram of two bell curves](image)

1. If \( \mu_1 = \mu_2 \), then \( T = (\bar{X}_1 - \bar{X}_2)/\text{se}(\bar{X}_1 - \bar{X}_2) \) follows a \( t \) distribution.
2. \( T \) will vary as \( \bar{X}_1 \) and \( \bar{X}_2 \) vary through sampling but its frequency will be a \( t \)-distribution
3. If observed \( T \approx 0 \), \( \Rightarrow \mu_1 \approx \mu_2 \).
4. If \(|T|\) is large, \( \Rightarrow \mu_1 \neq \mu_2 \).
5. We assess the risk in assuming that \( \mu_1 = \mu_2 \) by the probability of the \( t \) statistic being at least as large in magnitude as the observed \( T \).

Confidence Intervals for \( \mu_1 - \mu_2 \)
A \( 100(1-\alpha)\% \) CI for \((\mu_1 - \mu_2)\) is

\[
(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2,df} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
\]  

(20.6)

The 95\% confidence interval for the difference of means in the Example is

\[
1.3 \pm 2.1 \times \sqrt{\frac{2 \times 10.9}{10}} = [(1.3 - 2.1 \times 1.5), (1.3 - 2.1 \times 1.5)] = [-1.85, 4.45]
\]
We observe that \texttt{t.test()} evaluates this CI

If it is not tenable to assume that $s_1^2 \approx s_2^2$ the pooled variance should not be used and directly from (20.2)

$$\text{se}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(20.7)

This variant is known as \textit{Welch’s 2 sample t test} and is the default for \texttt{t.test()}. 
Lecture 21  Paired $t$ tests

- the differences amongst the boys are greater than the differences between materials.
- If a $t$ test was performed using

$$T = \frac{X_1 - X_2}{s.e.(X_1 - X_2)}$$

and

$$s.e.(X_1 - X_2) = s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

, the inference would be that there was no difference between materials because that se($X_1 - X_2$) would also include the variability amongst boys.

The statistical model:-

$$y_{ij} = \mu_i + \beta_j + \epsilon_{ij} \quad i = 1, 2 \quad j = 1, \ldots, 10$$

where

- $y_{ij}$ is the wear of the $i$th material on shoes worn by the $j$th boy,
- $\mu_i$ is the mean of the $i$th material,
- $\beta_j$ is the effect of the $j$th boy.
- $\epsilon_{ij} \sim N(0, \sigma^2)$

Calculating B-A for each boy gives

$$D_j = y_{2j} - y_{1j} = \tau_2 - \tau_1 + \epsilon_j^* \quad j = 1, \ldots, n$$
• $\epsilon^*_j \sim N(0, \sigma^2_d)$

• $s^2_d = \frac{1}{n-1} \sum_{j=1}^n (D_J - \bar{D})^2$, $\bar{D} = \frac{1}{n} \sum_{j=1}^n D_j$

<table>
<thead>
<tr>
<th>material</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>14.0</td>
<td>8.8</td>
<td>11.2</td>
<td>14.2</td>
<td>11.8</td>
<td>6.4</td>
<td>9.8</td>
<td>11.3</td>
<td>9.3</td>
<td>13.6</td>
</tr>
<tr>
<td>A</td>
<td>13.2</td>
<td>8.2</td>
<td>10.9</td>
<td>14.3</td>
<td>10.7</td>
<td>6.6</td>
<td>9.5</td>
<td>10.8</td>
<td>8.8</td>
<td>13.3</td>
</tr>
<tr>
<td>B-A</td>
<td>0.8</td>
<td>0.6</td>
<td>0.3</td>
<td>-0.1</td>
<td>1.1</td>
<td>-0.2</td>
<td>0.3</td>
<td>0.5</td>
<td>0.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

The mean difference is 0.41 and $s^2_d = 0.149$ giving

$$T = \frac{\bar{D}}{s_d/\sqrt{n}}$$

$$= \frac{0.41}{0.386/\sqrt{10}}$$

$$= 3.4$$

To do this problem we extend our repertoire in R by introducing the function `expand.grid()`

```r
Shoes <- expand.grid(material=LETTERS[1:2],boy=1:10)
t.test(paired=T)
```

With the data organised in a data frame, `t.test` can use a formula to test material differences on wear

- `paired=T`
- `alternative="two.sided"`
- If we were only interested in whether $y_B > y_A$, `alternative="more"`

```r
#_______ pairt.r ___________
Shoes <- expand.grid(material=LETTERS[1:2],boy=1:10)
Shoes$wear <- c(13.2,14, 8.2,8.8, 10.9,11.2,
14.3,14.2, 10.7,11.8,
6.6,6.4, 9.5,9.8, 10.8,11.3,
8.8,9.3, 13.3,13.6)
pairt <- t.test(wear ~ material,data=Shoes,
paired=T,alternative="two.sided")
print(pairt)
> source("pairt.r")
```

Paired t-test

data:  wear by material
t = -3.3489, df = 9, p-value = 0.008539
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
-0.6869539 -0.1330461
sample estimates:
mean of the differences
-0.41

The Rcmdr menu for a paired t-test requires that the 2 measurements be paired in columns, e.g. Shoes.txt.
Import the data into a data frame called *Shoes*.

Statistics $\rightarrow$ Means $\rightarrow$ Paired t-test. Leave the options as 2-sided and the Confidence Level as 0.95.