T-test

Anders Stockmarr
Course developers: Anders Stockmarr, Helle Rootzen, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science
Section for Statistics and Data Analysis
Technical University of Denmark
anst@dtu.dk

January 6th, 2024

DTU Compute
Department of Applied Mathematics and Computer Science

Plan for this week

- **Monday** Statistical inference, and the t-test
- **Tuesday** Simple and Multiple regression
- **Wednesday** ANOVA, ANCOVA, and linear models
- **Thursday** Categorical data, statistical report writing, logistic Regression
- **Friday** Introduction to repeated measures , Principal Component Analysis

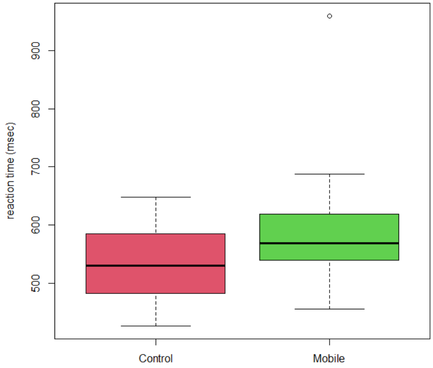
Outline

Mobile Phone Example

- A study to investigate whether mobile phone use impairs drivers' reaction times
- 64 students randomly assigned to two groups (mobile phone or control).
- in a simulated driving situation, the participants were instructed to press the "brake" when they saw a red light flash.
- The mobile phone group were having a conversation, while the control group listened to radio.
- We want to investigate whether the reaction differs between the two groups.

```
Mobile.phone <- read.delim("Data/Mobiltelefon.txt")
```

Mobile Phone Example

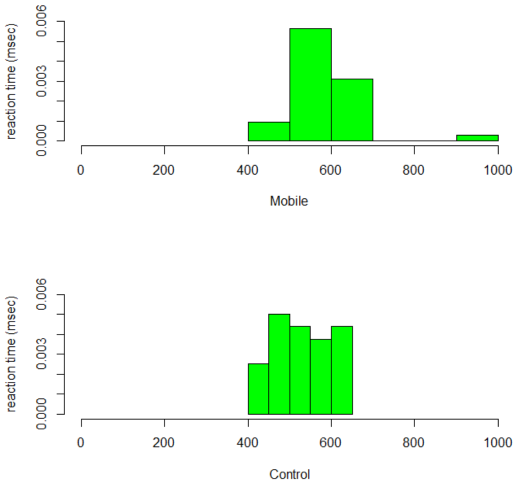


- What does this box plot show?

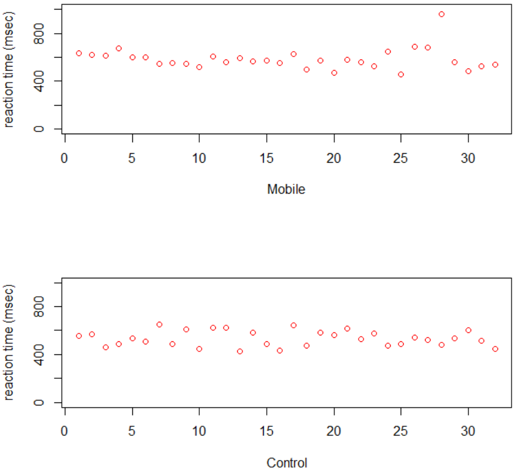
How To Set Up the Analysis

- Draw
    - Histogram
    - Box plot
    - Scatter plot
- Descriptive Statistics
    - Tables
    - Summary Statistics
- Analyses
    - Select model
    - Estimation
    - Test

Mobile Phone Example - Histogram



Mobile Phone Example - Scatter plot



Mobile Phone Example - Summary Statistics

- We have 64 observations of two variables: Time and Group (Mobile/Control).

```
by(Mobile$Time,Mobile$Group,summary)
Mobile$Group: Control
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  426.0   483.5   530.0   533.6   585.2   648.0
------------------------------------------
Mobile$Group: Mobile
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  456.0   540.5   569.0   585.2   617.0   960.0

by(Mobile$Time,Mobile$Group,sd)
Mobile$Group: Control
[1] 65.35998
------------------------------------------
Mobile$Group: Mobile
[1] 89.64606
```

Statistical Model - Two Groups

Model:
Two groups with (possibly) different normal distributions of reaction times:

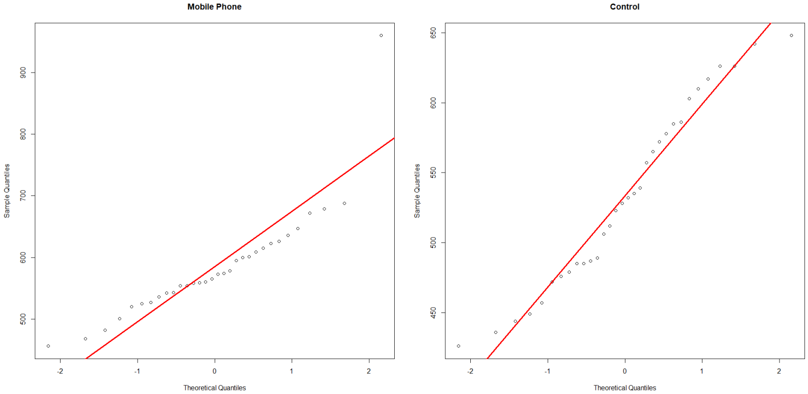
Mobile phone group: $Y_{1i} \sim N\left(\mu_1, \sigma_1^2\right), \quad i = 1, \ldots, 31$
Control group: $Y_{2i} \sim N\left(\mu_2, \sigma_2^2\right), \quad i = 1, \ldots, 32$

Assumptions:

- Normality as described - how could this be violated?
- Independence: all observations are independent - how could this be violated?
- Representativity: students represent a random sample - how could this be violated?

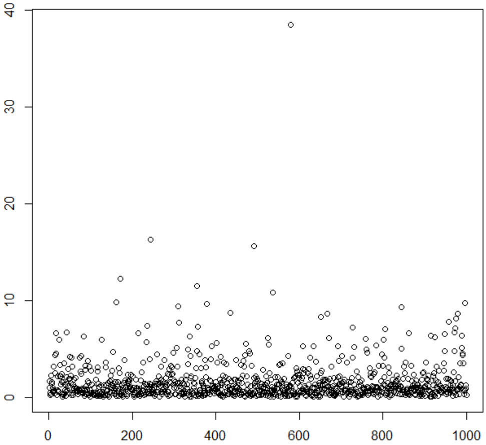
Hypotheses: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$

Normality assumption



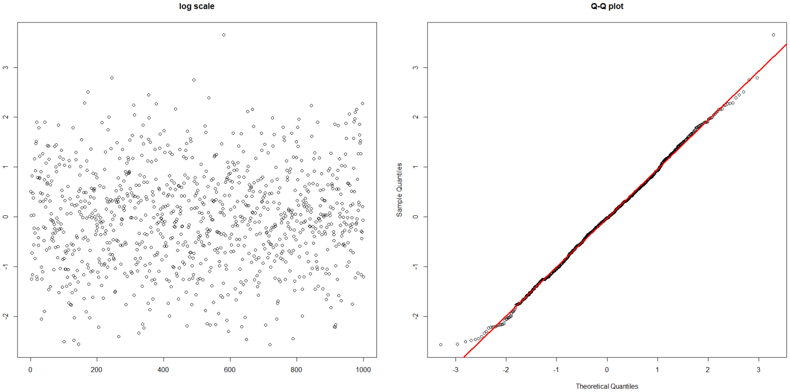
Normality assumption - Outliers

Is there an outlier here?



Normality assumption - Outliers

No outlier - normality on the log scale. Probability of reaching max is 12.2%, not cause for dismissal.



Normality assumption - Outliers

Not so for the Mobile data:

```
Y<-Mobile$Time[Mobile$Group=="Mobile"]
max(Y)
[1] 960
2*(1-pnorm(960-mean(Y),sd=sd(Y))^length(Y))
[1] 0.0009284201
```

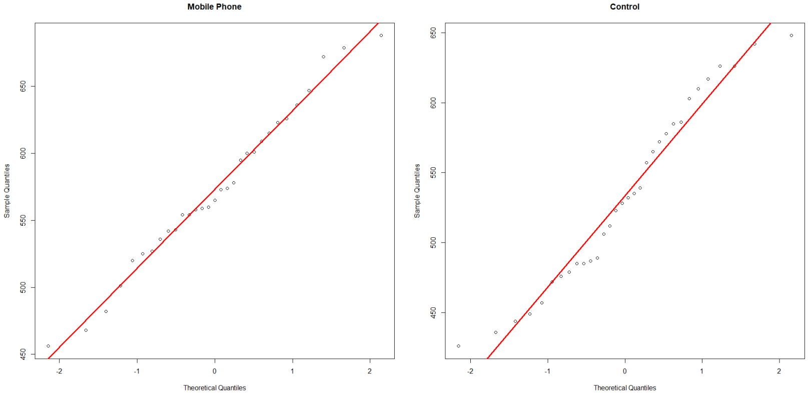
Without the variance-inflating observation:

```
Y<-Y[which(Y<900)]
2*(1-pnorm(960-mean(Y),sd=sd(Y))^length(Y))
[1] 1.584403e-09
```

Both numbers point towards an outlier.

Normality assumption

Leaving out the outlier in data:



Test of hypothesis H_0 vs. H_1

We use the *Welch t-test*, accounting for possibly unequal variances, and leaving out the outlier from the Mobile Phone group:

$$T = \frac{\overline{Y}_1 - \overline{Y}_2}{\widehat{se}(\overline{Y}_1 - \overline{Y}_2)} = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Satterthwaites approximation to the number of degrees of freedom ν:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$

Test of hypothesis H_0 vs. H_1

- We have observed $T = 39.50/15.67 = 2.52$
- The approximate degrees of freedom are found as $\nu = 60.69052$.
- Values critical for H_0 are numerically large values. The p-value is the probability of observing something more critical than the actual observation of T.
- calculating the p-value in R:

```
2*(1-pt(T,df=nu))
[1] 0.01432928
```

- The p-vaue is thus below the standard test level of $\alpha = 0.05$. At the 0.05 test level, the data do not support that the control group and the Mobile Phone group have similar reaction times (p=0.01).

Estimated Difference

- We estimate the difference in reaction times as follows:

$$\hat{\mu}_1 - \hat{\mu}_2 = \overline{Y}_1 - \overline{Y}_2 = 573.0968 - 533.5938 = 39.5030 \text{ msec.}$$

- What is the uncertainty of this estimate?

$$\widehat{se}(\overline{Y}_1 - \overline{Y}_2) = \sqrt{\widehat{var}(\overline{Y}_1) + \widehat{var}(\overline{Y}_2)} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = 15.6667$$

Confidence Interval for Estimated Difference

Standard confidence interval:

$$\hat{\mu} \pm q_{0.975} \times sd(\hat{\mu})$$

where $q_{0.975}$ is the 97.5% percentile in the relevant t-distribution. In our case, with $\nu = 60.69$ which gives $q_{0.975} = 1.9998$:

$$CI(\mu_1 - \mu_2) = [39.50 - 2 * 15.67; 39.50 + 2 * 15.67] = [8.17; 70.83]$$

Compare with the tighter approximative interval, where we use normal uncertainty of 1.96 rather than the t_ν uncertainty of 2:

$$[39.50 - 1.96 * 15.67; 39.50 + 1.96 * 15.67] = [8.80; 70.21]$$

A fairly good approximation here.

t-test in R

```
t.test(Y1,Y2)

        Welch Two Sample t-test

data:  Y1 and Y2
t = 2.5215, df = 60.691, p-value = 0.01433
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.172203 70.833845
sample estimates:
mean of x mean of y
 573.0968  533.5938
```

Similar Standard Deviation

- In many situations it makes sense to have an extra model assumption:

$$\sigma_1^2 = \sigma_2^2$$

ie. the variation in the two groups are identical. The model in this case is thus

Mobile phone group:	$Y_{1i} \sim N\left(\mu_1, \sigma^2\right),$	$i = 1, \ldots, 32$
Control group:	$Y_{2i} \sim N\left(\mu_2, \sigma^2\right),$	$i = 1, \ldots, 32$

Similar Standard Deviation

- The assumption of similar standard deviation model leads to a different test statistic, where the empirical variances are pooled:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 3877.745$$

$$T = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = 2.52$$

Similar Standard Deviation

- This case of equal variances is much simpler, and no approximations to the number of degrees of freedom for the t-test is needed: It is $n_1 + n_2 - 2 = 61$. The test provides a higher power than the Welch t-test, because the model has one less parameter to estimate.

- However, if the difference in variance is considerable, the similar variance t-test med be misleading. Without thorough investigations, the Welch version of the t-test should be used. In particular, for small sample sizes, it may be difficult to detect differences in variation with sufficient strength.

- In the present case we have estimates $s_1^2 = 3470.424$ and $s_2^2 = 4271.926$. The data does not support that these values should be different (p=0.31). More on this on Thursday.

Similar Standard Deviation

The t-test with assuming equal variances:

```
t.test(Y1,Y2,var.equal=TRUE)

	Two Sample t-test

data:  Y1 and Y2
t = 2.5172, df = 61, p-value = 0.01447
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.123071 70.882978
sample estimates:
mean of x mean of y
 573.0968  533.5938
```

Sensitivity Analysis

- We found an outlier in the Mobile Phone group, with a reaction time of 960 msec - nearly a second!
- We removed the outlier, as it clearly wasn't comparable to the remaining data - likely a student that wasn't up for the task and was thinking about something else.
- Without the removal of the outlier, we would be violating the normality assumption, and the t-test would no longer be valid.

Sensitivity Analysis

- **Problem**: Are we testing on an idealized population, without the proper association to reality?
- To supplement our analysis, we will investigate how to include the outlier in an analysis, to see if the presence affect our conclusion.
- The interpretation is that we allow for a fraction not being observant at all, not being "up for the task".

Sensitivity Analysis

- Test statistic: We still use the t-test statistic, but this time we INCLUDE the outlier in the mobile phone group T_1:

$$T = \frac{\overline{Y}_1 - \overline{Y}_2}{\widehat{se}(\overline{Y}_1 - \overline{Y}_2)} = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}} = 2.63$$

- : Estimated group difference: 51.59. Much bigger than without the outlier, but the variance is also increased.

- The assumption of normality is seriously violated however, and we have to resort to other means to evaluate the statistic T.

Sensitivity Analysis - the permutation test

- Consider the Nul hypothesis that we wish to test:

$$H_0 : \mu_1 = \mu_2$$

- Under H_0, the mean in the two groups is identical.
- Thus, if we resample our reaction times in two new groups, the two groups will still theoretically have the same mean.
- We use this technique to investigate the sampling variation may be the cause of the difference in means.

Sensitivity Analysis - the permutation test

- Strategy:
  - resample the 64 data points in two new random groups.
  - calculate the test statistic T for the two new groups.
  - Compare the new T statistic with the original, to see if it is bigger.
  - Repeat the above a large number of times.
  - Use the fraction of T statistics bigger than the original as the p-value, as this simulates the probability of getting a more extreme result than our original due to sampling variation.

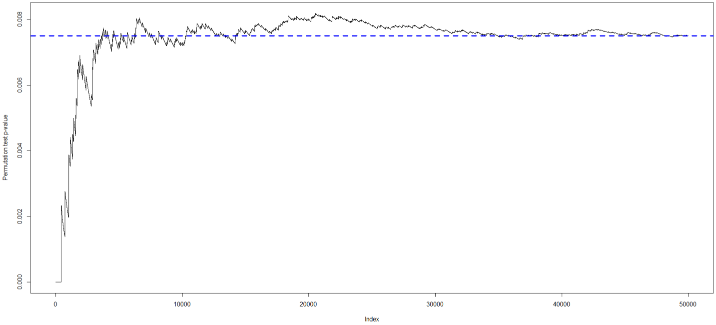
Sensitivity Analysis - the permutation test

R code:

```
my.reaction.times<-Mobile$Time
my.t.statistics<-numeric(50000)
for(i in 1:50000){
  index<-sample(1:64,32)
  Y1.temp<-my.reaction.times[index]
  Y2.temp<-my.reaction.times[-index]
  my.t.statistics[i]<-t.test(Y1.temp,Y2.temp)$statistic
  }
my.p.value<-length(my.t.statistics[abs(my.t.statistics)>T])/
           50000
my.p.value
[1] 0.0075
```

Sensitivity Analysis - the permutation test

- The permutation test supports the previous conclusions - but have we performed enough simulations?



Exercise: MTcars Example

We want to compare miles per gallon for cars with and without manual transmission.

- Access the builtin data set mtcars with the command

  ```
  data(mtcars)
  ```

- Plot the Miles per Gallon for the two groups (am=0 or 1).

- Formulate the relevant hypothesis to test, and the alternative.

- Are the underlying assumptions for the t-test fulfilled?

- What is the estimated difference in mpg, and the corresponding 95% confidence interval?

- What can we conclude about H_0?

Paired t-test: The Glucose Data

- The *Glucose12* data set features data from two different methods to measure blood glucose.
- 73 subjects have had their levels of blood glucose measured with both metods.
- We would like to know if the two methods measure the same?

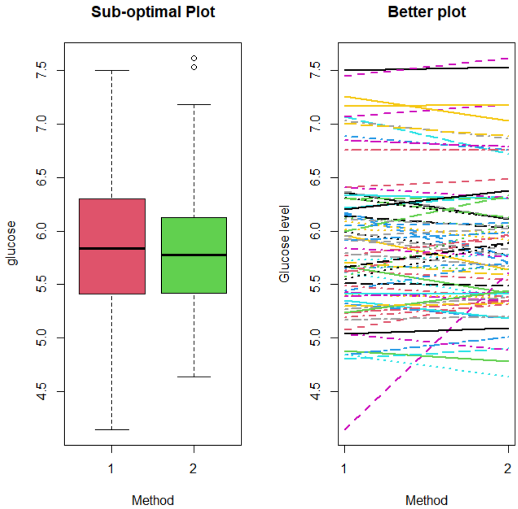
```
head(glucose12)
  subject Glucose1 Glucose2
1       1     6.36     6.11
2       2     5.08     5.35
3       3     6.12     6.04
4       4     5.65     5.69
5       5     7.07     6.72
6       6     5.43     5.34
```

Paired t-test: The Glucose Data

- For paired data, each subject act as its own control.
- Greatly reduces person-to-person variation, and may give a much more powerful test.
- We will look for differences between the two types of measurements.
  - Are the differences independent of the size of the measurements?
  - Do we need to look at relative differences (log-transform data)?
- Overall, we wish to investigate of the difference between the two types of measurements is 0.

Paired t-test: The Glucose Data

Why is the first plot not optimal?



Model for Paired Data

Data:
Method 1: $X_i, i = 1, \ldots, 73$
Method 2: $Y_i, i = 1, \ldots, 73$
Difference: $D_i = X_i - Y_i, i = 1, \ldots, 73$

Model: Differences are assumed independent and identically distributed with $D_i \sim N(\mu, \sigma^2)$-

Assumptions

- Assumptions on differences as above;
- NO further assumptions on X and Y.

Hypotheses

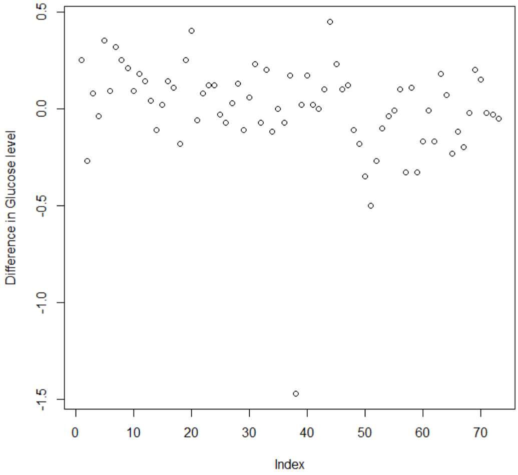
$$H_0 \, : \, \mu = 0 \qquad H_1 \, : \, \mu \neq 0$$

Assumptions for the Paired t-test

- Independence - consider circumstances.
- Same variances - look for patterns in the scatter plot of differences. IF data are normally distributed: Look at mean of methods vs. differences in methods.
- Normality - consider the qq-plot vs. the normal distribution.

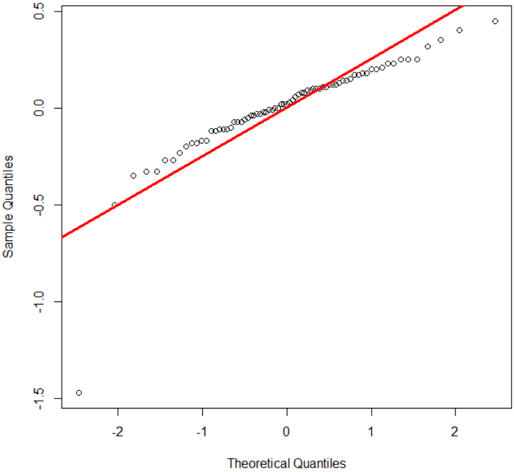
Assumptions for the Paired t-test

Same variances:



Assumptions for the Paired t-test

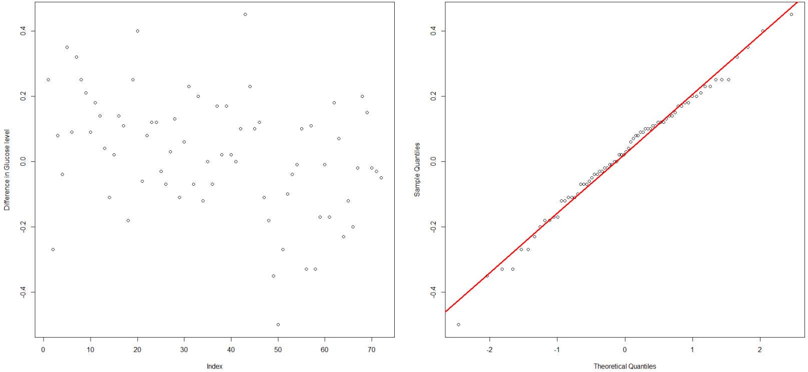
Normality:



Assumptions for the Paired t-test

Continuing without the outlier:

```
glucose12.new<-glucose12[glucose12$Glucose1-glucose12$Glucose2>-1,]
glucose12.new$D<-glucose12.new$Glucose1-glucose12.new$Glucose2
```



Estimation of Method Difference

We wish to estimate the mean difference, ie. the parameter μ. This puts us back to a one-sample problem. We have 72 measurements of the same normally distributed variable D:

- The estimate of the mean difference is $\hat{\mu} = \overline{D} = 0.0238$.
- The standard deviation of D: 0.1822
- The standard error of $\hat{\mu}$: $SEM = \frac{sd(D)}{\sqrt{n}} = \frac{0.1822}{\sqrt{72}} = 0.0215$

Confidence Interval

The 95% confidence interval for μ;

$$\begin{aligned}
&\overline{D} \pm t_{97.5\%}(72) \times SEM \\
=&0.0238 \pm 1.9935 \times 0.0215 \\
=&[-0.0191; 0.0666]
\end{aligned}$$

Compare with the standard confidence interval from normal errors:

$$\begin{aligned}
&\overline{D} \pm 1.96 \times SEM \\
=&[-0.0183; 0.0658]
\end{aligned}$$

Confidence Interval

We found a confidence interval of

$$[-0.0191; 0.0666]$$

- This interval includes 0; a t-test will show that the data do not support a systematic bias at the 5% test level.
- We will expect the mean of the differences in a similar experiment with 72 subjects to be within this interval, with 95% probability.
- In our study, there could still be a systematic bias less than 0.019, but the t-test from the study will not have enough power to detect it.

Test of No Bias

Let us test the hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu \neq 0$:

$$T = \frac{\hat{\mu} - 0}{SEM} = \frac{0.0238 - 0}{0.0215} = 1.1059 \sim t(71)$$

The p-value in a t-distribution with 71 degrees of freedom is $p = 0.27$, so the hypothesis is accepted; the data do not support a systematic difference between the two methods at the 5% test level.

Note the correspondence between test and confidence interval:

- If the CI contains 0, the t-test will be statistically insignificant;
- If the CI does not contain 0, the t-test will be statistically significant.

Paired t-test in R

- We don't have to calculate this by hand but we can use **R**:

```
t.test(glucose12.new$D)

        One Sample t-test

data:  glucose12.new$D
t = 1.1059, df = 71, p-value = 0.2725
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.01906955  0.06656955
sample estimates:
mean of x
  0.02375
```

Paired t-test in R - Alternative Formulation

```
t.test(glucose12.new$Glucose1,glucose12.new$Glucose2,paired=TRUE)

        Paired t-test

data:  glucose12.new$Glucose1 and glucose12.new$Glucose2
t = 1.1059, df = 71, p-value = 0.2725
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01906955  0.06656955
sample estimates:
mean of the differences
                0.02375
```

Exercise - Mobile Phones 1

Recall the study of reaction times when driving. In this exercise we have results from a paired study design, where each subject performs both the 'Mobile' and the 'Control' experiment.

- Load the data Mobile_Matched.txt.
- Make relevants plots of the data, and formulate the hypotheses to test the method difference.
- Evaluate the model control, and perform the test.

Exercise - Mobile Phones 2

Repeat the analysis from the previous exercise, but this time transform the original reaction times with any log transform.

- See if the check for normality check went better. Comment and compare to the previous exercise.

- Present you results both on the chosen log-scale and back-transformed to the original scale. Is the conclusion altered compared to the non-transformed data?