Course 02402 Introduction to Statistics Lecture 6:

Two-sample comparisons and power/sample size

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
   - Precision requirements
   - Power and sample size - one-sample
   - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

Oversigt

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
    - Precision requirements
    - Power and sample size - one-sample
    - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

Motivating example - nutrition study

Nutrition study

In a nutrition study the aim is to investigate if there is a difference in the energy usage for two different types of (moderately physically demanding) work. In the study, the energy usage of 9 nurses from hospital A and 9 (other) nurses from hospital B have been measured. The measurements are given in the following table in mega Joule (MJ):

Sample from each hospital, $n_1 = n_2 = 9$:

Hospital A	Hospital B
7.53	9.21
7.48	11.51
8.08	12.79
8.09	11.85
10.15	9.97
8.40	8.79
10.88	9.69
6.13	9.68
7.90	9.19

Example - nutrition study

The hypothesis of no difference is in focus:

$$H_0: \ \mu_1 = \mu_2$$

Sample means and standard deviations:

$\hat{\mu}_A = \bar{x}_A = 8.293, \ (s_A = 1.428)$

$\hat{\mu}_B = \bar{x}_B = 10.298, \ (s_B = 1.398)$

Is data in accordance with the null hypothesis H_0?

Data: $\bar{x}_B - \bar{x}_A = 2.005$

Null hypothesis: $H_0: \ \mu_B - \mu_A = 0$

NEW:p-**value for difference:**

$$p\text{-value} = 0.0083$$

(Found in the scenario that H_0 is true)

NYT:**Confidence interval for difference**:

$$2.005 \pm 1.412 = [0.59; \ 3.42]$$

Oversigt

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
   - Precision requirements
   - Power and sample size - one-sample
   - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

The definition of hypothesis test and significance (Repetition)

Definition 3.23. Hypothesis test:

We say that we carry out a hypothesis test when we decide against a null hypothesis or not using the data.

A null hypothesis is *rejected* if the p-value, calculated after the data has been observed, is less than some α, that is if the p-value $< \alpha$, where α is some pre-specifed (so-called) *significance level*. And if not, then the null hypothesis is said to be *accepted*.

Definition 3.28. Statistical significance:

An *effect* is said to be *(statistically) significant* if the p-value is less than the significance level α.
(OFTEN we use $\alpha = 0.05$)

Steps by hypothesis tests - an overview (Repetition)

Generelly a hypothesis test consists of the foloowing steps:

1. Formulate the hypotheses and choose the level of significance α (choose the "risk-level")
2. Calculate, using the data, the value of the test statistic
3. Calculate the p-value using the test statistic and the relevant sampling distribution, and compare the p-value and the significance level α and make a conclusion

   OR:

   Alternatively, make a conclusion based on the relevant critical value(s)

The definition and interpretation of the p-value (Repetition)

The p-value expresses the *evidence* against the null hypothesis – Table 3.1:

$p < 0.001$	Very strong evidence against H_0
$0.001 \leq p < 0.01$	Strong evidence against H_0
$0.01 \leq p < 0.05$	Some evidence against H_0
$0.05 \leq p < 0.1$	Weak evidence against H_0
$p \geq 0.1$	Little or no evidence against H_0

Definition 3.21 of the p-value:

The p-value is the probability of obtaining a test statistic that is at least as extreme as the test statistic that was actually observed. This probability is calculated under the assumption that the null hypothesis is true.

Critical value, confidence interval and hypothesis test (Repetition)

Theorem 3.32: Critical value method = Confidence interval method

We consider a $(1-\alpha)\cdot 100\%$ confidence interval for μ:

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

The confidence interval corresponds to the acceptance region for H_0 when testing the (non-directional) hypothesis

$$H_0: \quad \mu = \mu_0$$

(New) interpretation of the confidence interval:

The confidence interval covers those values of the parameter that we believe in given the data.
Those values that we accept by the corresponding hypothesis test.

Oversigt

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
    - Precision requirements
    - Power and sample size - one-sample
    - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

Method 3.48: Two-sample t-test

Computing the test statistic

When considering the null hypothesis about the difference between the means of two *independent* samples:

$$\delta = \mu_2 - \mu_1$$

$$H_0: \ \delta = \delta_0$$

the (Welch) two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Theorem 3.49: The distribution of the (Welch) t-test statistic

Welch t-test statistic is t-distributed

The (Welch) two-sample statistic seen as a random variable:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

approximately, under the null hypothesis, follows a t-distribution with ν degrees of freedom, where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

if the two population distributions are normal or if the two sample sizes are large enough.

Method 3.50: Two-sample t-test

The level α test is

1. Compute t_{obs} and ν as given above.
2. Compute the evidence against the *null hypothesis*[a] $H_0 : \mu_1 - \mu_2 = \delta_0$ vs. the *alternative hypothesis* $H_1 : \mu_1 - \mu_2 \neq \delta_0$ by the

   $$p\text{–value} = 2 \cdot P(T > |t_{\text{obs}}|)$$

   where the t-distribution with ν degrees of freedom is used.
3. If p–value $< \alpha$: We $\underline{\text{reject } H_0}$, otherwise we $\underline{\text{accept } H_0}$.

   OR

   The rejection/acceptance conclusion could alternatively, but equivalently, be made based on the critical value(s) $\pm t_{1-\alpha/2}$:

   If $|t_{\text{obs}}| > t_{1-\alpha/2}$ we $\underline{\text{reject } H_0}$, otherwise we $\underline{\text{accept } H_0}$.

[a] We are often interested in the test where $\delta_0 = 0$

Example - nutrition study

The hypothesis of no difference is in focus:

$$H_0: \ \delta = \mu_B - \mu_A = 0$$

versus the non-directional($=$ two-sided) alternative:

$$H_0: \ \delta = \mu_B - \mu_A \neq 0$$

First the computations of t_{obs} and ν:

$$t_{\text{obs}} = \frac{10.298 - 8.293}{\sqrt{2.0394/9 + 1.954/9}} = 3.01$$

and

$$\nu = \frac{\left(\frac{2.0394}{9} + \frac{1.954}{9}\right)^2}{\frac{(2.0394/9)^2}{8} + \frac{(1.954/9)^2}{8}} = 15.99$$

Example - nutrition study

Next the p-value is found:

$$p\text{–value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2P(T > 3.01) = 2 \cdot 0.00415 = 0.0083$$

```
## p-value for nutrition study example
1 - pt(3.01, df = 15.99)

## [1] 0.004154
```

Evaluate the evidence (Table 3.1):

There is strong evidence AGAINST the null hypothesis.

Conclude based on $\alpha = 0.05$:

We reject the null hypothesis, as there is a significant difference of the two groups - nurses on Hospital B can be said to have a larger (mean) energy usage than nurses on Hospital A.

Oversigt

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
    - Precision requirements
    - Power and sample size - one-sample
    - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

Method 3.46: Confidence interval for $\mu_1 - \mu_2$

The Confidence interval for the mean difference:

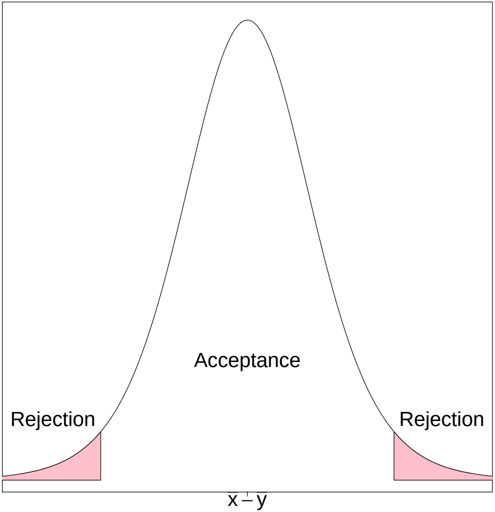
For two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$-quantile from the t-distribution with ν degrees of freedom given from Theorem 3.49 (as above)

The Confidence interval and hypothesis test (Repetition)

The acceptance region is the potential values for $\mu_1 - \mu_2$ that are not too far away from the data:



Example - nutrition study - everything in R:

Let us find the 95% confidence interval for $\mu_B - \mu_A$. Since the relevant t-quantile is, using $\nu = 15.99$,

$$t_{0.975} = 2.120$$

the confidence interval becomes:

$$10.298 - 8.293 \pm 2.120 \cdot \sqrt{\frac{2.0394}{9} + \frac{1.954}{9}}$$

which then gives the result as also seen above:

$$[0.59;\ 3.42]$$

Example - nutrition study - everything in R:

```
## Read the two-sample in R
xA=c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4, 10.88, 6.13, 7.9)
xB=c(9.21, 11.51, 12.79, 11.85, 9.97, 8.79, 9.69, 9.68, 9.19)
## A two sample Welch t-test
t.test(xB, xA)

##
##  Welch Two Sample t-test
##
## data:  xB and xA
## t = 3, df = 16, p-value = 0.008
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5923 3.4166
## sample estimates:
## mean of x mean of y
##    10.298     8.293
```

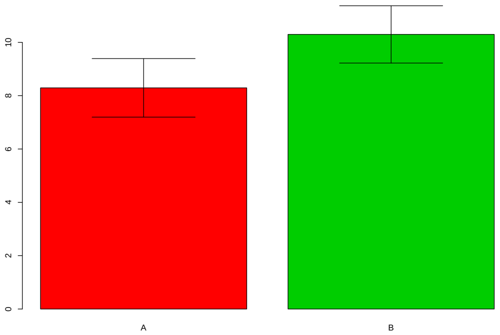
Oversigt

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
   - Precision requirements
   - Power and sample size - one-sample
   - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

Example - nutrition study - presentation of result

Barplot with *error bars* are often seen

A grouped barplot with some "error bars" - below the 95%-confidence intervals for each group is shown:



Be careful about using "overlapping confidence intervals"

The approach actually is using an incorrect variation for evaluation of the difference:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} \neq \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

$$\mathrm{Var}(\bar{X}_A - \bar{X}_B) = \mathrm{Var}(\bar{X}_A) + \mathrm{Var}(\bar{X}_B)$$

Assume that the two standard-errors are 3 and 4: The sum is 7, but $\sqrt{3^2 + 4^2} = 5$

The correct relation between the two hence is:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} < \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

Be careful about using "overlapping confidence intervals"

Remark 3.58. Rule for using "overlapping confidence intervals":

When two CIs do NOT overlap: The two groups are significantly different

When two CIs DO overlap: We do not know what the conclusion is

Oversigt

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
   - Precision requirements
   - Power and sample size - one-sample
   - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

Motivating example - sleeping medicine

Difference of sleeping medicines?

In a study the aim is to compare two kinds of sleeping medicine A and B. 10 test persons tried both kinds of medicine and the following results are obtained, given in prolonged sleep length (in hours) for each medicine type:

Sample, $n = 10$:

Person	A	B	$D = B - A$
1	$+0.7$	$+1.9$	$+1.2$
2	-1.6	$+0.8$	$+2.4$
3	-0.2	$+1.1$	$+1.3$
4	-1.2	$+0.1$	$+1.3$
5	-1.0	-0.1	$+0.9$
6	$+3.4$	$+4.4$	$+1.0$
7	$+3.7$	$+5.5$	$+1.8$
8	$+0.8$	$+1.6$	$+0.8$
9	0.0	$+4.6$	$+4.6$
10	$+2.0$	$+3.4$	$+1.4$

The paired setup and analysis = one-sample analysis

```
## Read the two samples
x1=c(.7,-1.6,-.2,-1.2,-1,3.4,3.7,.8,0,2)
x2=c(1.9,.8,1.1,.1,-.1,4.4,5.5,1.6,4.6,3.4)
## Take the difference to get a paired t-test
dif=x2-x1
## Calculate the test and results
t.test(dif)

##
##  One Sample t-test
##
## data:  dif
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of x
##      1.67
```

The paired setup and analysis = one-sample analysis

```
## Another way to calculate the paired setup
t.test(x2, x1, paired=TRUE)

##
##  Paired t-test
##
## data:  x2 and x1
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of the differences
##                    1.67
```

Paired versus independent experiment

Completely Randomized (independent samples)

20 patients are used and completely at random allocated to one of the two treatments (but usually making sure to have 10 patients in each group). So: different persons in the different groups.

Paired (dependent samples)

10 patients are used, and each of them tests both of the treatments. Usually this will involve some time in between treatments to make sure that it becomes meaningful, and also one would typically make sure that some patients do A before B and others B before A. (and doing this allocation at random). So: the same persons in the different groups.

Example - Sleeping medicine - WRONG analysis

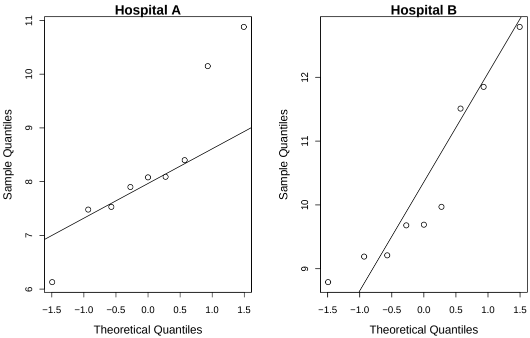
```
t.test(x1,x2)

##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -1.9, df = 18, p-value = 0.07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.4854  0.1454
## sample estimates:
## mean of x mean of y
##      0.66      2.33
```

Oversigt

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
    - Precision requirements
    - Power and sample size - one-sample
    - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

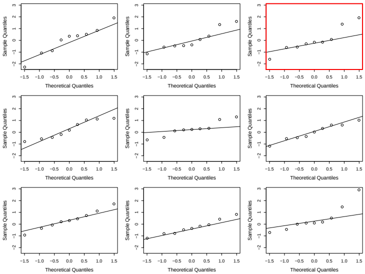
Example - Q-Q plot for EACH sample:

```
## Q-Q plot for each sample
par(mfrow=c(1,2))
qqnorm(xA, main="Hospital A")
qqline(xA)
qqnorm(xB, main="Hospital B")
qqline(xB)
```



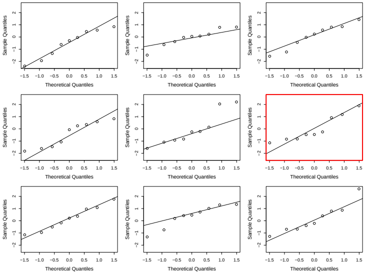
Example - Comparing with simulated, A

```
require(MESS)
fit1 <- lm(xA ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...)}
wallyplot(fit1, FUN=qqnorm.wally, main="")
```



Example - Comparing with simulated, B

```
## Multiple (simulated) Q-Q plots for each sample
fit1 <- lm(xB ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...)}
wallyplot(fit1, FUN=qqnorm.wally, main="")
```



Oversigt

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
    - Precision requirements
    - Power and sample size - one-sample
    - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

Planning of study with requirements to the precision

Method 3.62: The one-sample CI sample size formula:

When σ is known or guessed at some value, we can calculate the sample size n needed to achieve a given margin of error, ME, with probability $1-\alpha$ as:

$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME}\right)^2$$

Example, height data again

Sample mean og standard deviation:

$$\bar{x} = 178$$
$$s = 12.21$$

Estimate the population mean and standard deviation:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

If we want that $ME = 3$cm with 95% confidence, how large should n then be?

$$n = \left(\frac{1.96 \cdot 12.21}{3}\right)^2 = 63.64$$

Planning, Power

What is the power of a future study/experiment:

- The probability of detecting an (assumed) effect
- $P(\text{Reject } H_0)$ when H_1 is true
- Probability of correct rejection of H_0
- Challenge: The null hypothesis can be wrong in many ways!
- Practically: Scenario based approach
  - E.g. "What if $\mu = 86$, how good will my study be to detect this?"
  - E.g. "What if $\mu = 84$, how good will my study be to detect this?"
  - etc

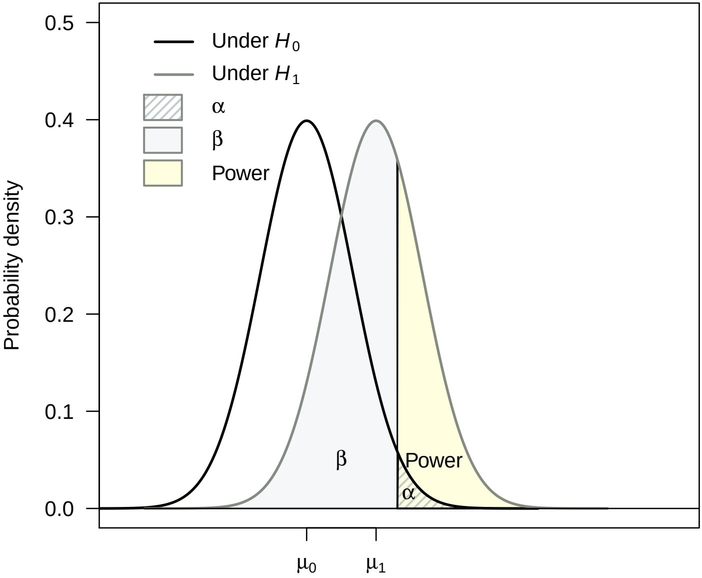
Planning and power

When the test to use has been set:

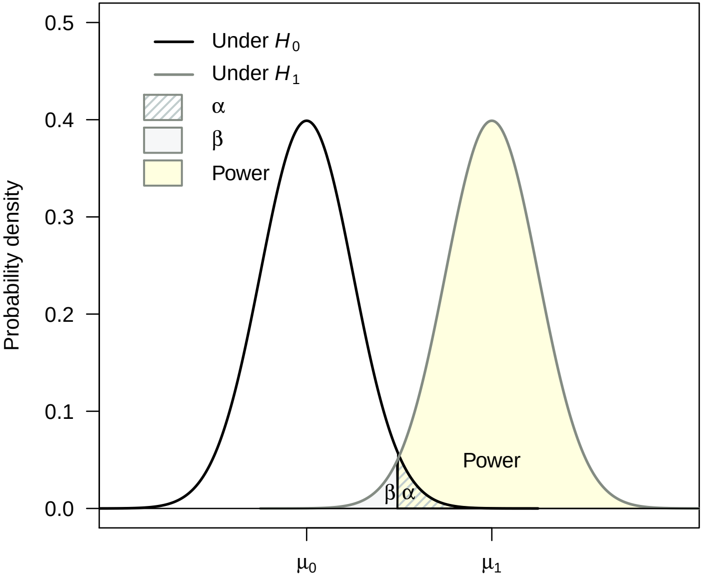
If you know (or set/guess) four out of the following five pieces of information, you can find the fifth:

- The sample size n
- Significance level α of the test.
- A change in mean that you would want to detect (effect size) $\mu_0 - \mu_1$.
- The population standard deviation, σ.
- The power $(1-\beta)$.

Low power example



High power example



Planning, Sample size n

The big practical question: What should n be?

The experiment should be large enough to detect a relevant effect with high power (usually at least 80%):

Metode 3.64: The one-sample sample size formula:

For the one-sample t-test for given α, β and σ:

$$n = \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{(\mu_0 - \mu_1)} \right)^2$$

Where $\mu_0 - \mu_1$ is the change in means that we would want to detect and $z_{1-\beta}$, $z_{1-\alpha/2}$ are quantiles of the standard normal distribution.

Example - The power for $n = 40$

```
power.t.test(n = 40, delta = 4, sd = 12.21,
             type = "one.sample")

##
##      One-sample t test power calculation
##
##               n = 40
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.5242
##     alternative = two.sided
```

Example - The sample size for power$= 0.80$

```
power.t.test(power = .80, delta = 4, sd = 12.21,
             type = "one.sample")

##
##      One-sample t test power calculation
##
##               n = 75.08
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Power and sample size - two-sample

Finding the power of detecting a group difference of 2 with $\sigma = 1$ for $n = 10$:

```
## Power calculation for two-sample
power.t.test(n = 10, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.9882
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Power and sample size - two-sample

Finding the sample size for detecting a group difference of 2 with $\sigma = 1$ and power$= 0.9$:

```
## Sample size calculation for two-sample
power.t.test(power = 0.90, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 6.387
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Power and sample size - two-sample

Finding the detectable effect size (delta) with $\sigma = 1$, $n = 10$ and power$= 0.9$:

```
################################
## Detectable effect size calculation for two-sample
power.t.test(power = 0.90, n = 10, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 1.534
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Oversigt

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
    - Precision requirements
    - Power and sample size - one-sample
    - Power and sample size - two-sample
9. The pooled t-test - a possible alternative

The pooled two-sample t-test statistic

The *pooled* estimate of variance (assuming $\sigma_1^2 = \sigma_2^2$)

Method 3.51

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

The pooled test statistic, Method 3.52

When considering the null hypothesis about the difference between the means of two *independent* samples:

$$\delta = \mu_2 - \mu_1$$

$$H_0: \ \delta = \delta_0$$

the pooled two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

Theorem 3.53: The distribution of the pooled test-statistic

is a *t*-distribution:

The pooled two-sample statistic seen as a random variable:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}$$

follows, under the null hypothesis and under the assumption that $\sigma_1^2 = \sigma_2^2$, a t-distribution with $n_1 + n_2 - 2$ degrees of freedom if the two population distributions are normal.

We always use the "Welch' version

Almost (fool)proof to use the Welch-version always:

- if $s_1^2 = s_2^2$ the Welch and the Pooled test statistics are the same.
- Only when the two variances become really different the two test-statistics may differ in any important way, and if this is the case, we would not tend to favour the pooled version, since the assumption of equal variances appears questionable then.
- Only for cases with a small sample sizes in at least one of the two groups the pooled approach may provide slightly higher power if you believe in the equal variance assumption. And for these cases the Welch approach is then a somewhat cautious approach.

Agenda

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the difference
5. Overlapping confidence intervals?
6. The paired setup
7. Checking the normality assumptions
8. Planning for wanted precision or power
   - Precision requirements
   - Power and sample size - one-sample
   - Power and sample size - two-sample
9. The pooled t-test - a possible alternative