STATS201.stats 201 30 Statistical inference

Anna Helga Jónsdóttir
Sigrún Helga Lund

December 14, 2012

Contents

1 Statistical inference

1.1 Estimators

Estimator
An **estimator** is a statistic that estimates parameters of probability distributions.

- Estimators for parameters of normal distribution, Poisson distribution and binomial distribution.
- μ, σ, λ and p.
- The outcome of the estimators are called estimates
- They are denoted with $\hat{\mu}$, $\hat{\sigma}$, $\hat{\lambda}$ and $\hat{p}$.

1.2 Estimator for the mean of a random variable

Metill á meðaltal slembistærðar
The estimator used for the mean of a random variable is

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

where n is the total number of measurements.

It is believed that the number of children that break a leg every day in Iceland follows a Poisson distribution. A doctor at the emergency room wants to estimate how many children, on average, break a leg per day. He has some data describing the number of breaks the past ten days: 2, 0, 1, 7, 3, 3, 6, 4, 4, 1. What is the doctors estimate of the average number of leg breaks per day?

The λ parameter in the Poisson distribution represents the average number of breaks per day. The estimator is X, which is simply the mean of the measurements:

$$\hat{\lambda} = \bar{x} = \frac{2+0+1+7+3+3+6+4+4+1}{10} = \frac{31}{10} = 3.1.$$

1.3 Estimator for the variance of a random variable

> **Estimator for the variance of a random variable**
> The estimator used for the variance of a random variable is
>
> $$s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$
>
> where $\bar{X}$ is the estimator for the mean of the measurements and n is the total number of measurements. mælinga.

Helga believes that womens shoe sizes follow a normal distribution. She wants to open up a shoe shop so she is interested in finding out what the variance of womens shoe size is to find out how many pairs she needs to buy of each number. She measures the shoe size of eight women and gets: 40, 36, 37, 39, 38, 39, 40, 38. What is the estimated variance?

1. The mean is $\frac{40+36+37+39+38+39+40+38}{8} = 38.375$.
2. The deviation from the mean is: 1.625, -2.375, -1.375, 0.625, -0.375, 0.625, 1.625, -0.375.
3. The numbers squared are: 2.641, 5.641, 1.891, 0.391, 0.141, 0.391, 2.641, 0.141.
4. The sum of the squared number is 13.878.
5. $\frac{13.878}{8-1} = \frac{13.878}{7} = 1.983$

So: $\hat{\sigma}^2 = s^2 = 1.983$.

1.4 Estimator for the ratio of a random variable

> **Estimator for the ratio of a random variable**
> The estimator used for the ratio of a random variable is
>
> $$P = \frac{X}{n}$$
>
> where X is the number of successful confidence intervals and n is the total number of confidence intervals.

Let us assume that the numbers of rotten apples in a box of 20 follows a binomial distribution. Anna wants to insure that she buys enough if good apples so she wants to estimate the proportion of rotten apples per box. She buys a box of 20 apples and finds 2 rotten apples. What is the estimated ratio of rotten apples?

We have $n = 20$ and $x = 2$ so the estimated proportion is:

$$\hat{p} = \frac{x}{n} = \frac{2}{20} = 0.1.$$

1.5 Confidence level

Usually there is no probability that our estimate is exactly the true value of the parameter.

Confidence intervals
1 - α **confidence interval** is a numerical interval that contains the true value with the confidence level 1 - α.

Confidence level
Confidence level is the ratio of cases when the confidence interval contains the true value of the parameter, when the experiment is repeated very often.

1.6 Confidence limits

Confidence limits
Confidence limits are the endpoints of the confidence interval. The upper confidence limit is the upper endpoint of the interval (the highest value in the interval), but the lower confidence limit is the lower endpoint (the smallest value in the interval).

Type I error
Type I error denoted α, is the ratio of cases where the confidence interval contains the true value of the parameter, if the experiment is repeated very often.

1.7 The ideology behind hypothesis tests

The ideology behind hypothesis tests

A hypothesis is found that describes what we want to demonstrate and another that describes a neutral case.

A statistic is found that has a known probability distribution in the neutral case. This statistic is our test statistic.

It is defined what values of the test statistic are "improbable" according to the probability distribution in the neutral case.

If the retrieved estimate classifies as "improbable" the hypothesis for the neutral stage is rejected and the hypothesis we want to demonstrate is claimed.

If the estimate is not "improbable" no claims are made.

1.8 Hypothesis

Null hypothesis
Null hypothesis is a hypothesis that can be rejected with observed data. It can never we be claimed. It is usually denoted with H_0.

Alternative hypothesis
Alternative hypothesis is the hypothesis we wish confirm with the experiment. It can only be claimed but not rejected. It is either denoted with H_1 or H_a.

1.9 Directions of hypothesis tests

Two-sided tests
If the data allows, a **two-sided test** claims that one or more parameters of the population or populations are **not equal** to each other or a certain value.

One-sided tests
There are two types of **one-sided tests**:
Those who claim that one parameter of the probability distribution is **larger** then another parameter or a certain value, if the measurements allow.
Those who claim that one parameter of the probability distribution is **smaler** then another parameter or a certain value, if the measurements allow.

1.10 Test statistics

Test statistic
A **test statistic** is a statistic that can be used to reject a null hypothesis if the measurements allow.

Null hypothesis rejected
A null hypothesis is **rejected** if the test statistic receives a improbable value compared to the probability distribution it should have if the null hypothesis would be true.

1.11 Rejection areas and α-levels

α-level
The α **level** of a hypothesis test is the highest acceptable probability that we receive an improbable value when the null hypothesis is true.

Rejection areas of hypothesis tests
Rejection areas of hypothesis tests are the intervals that contain **all** of the improbable values and **only** those values.
If the test statistics falls within the rejection interval of the hypothesis test, we reject the null hypothesis.
If it does not fall within the rejection interval of the hypothesis test, we make no claims

1.12 Rejection areas and α-levels

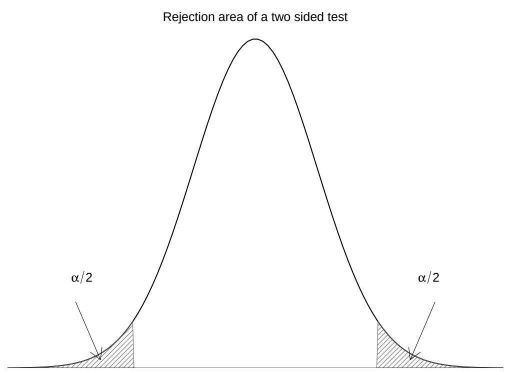
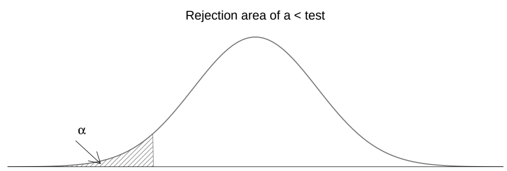
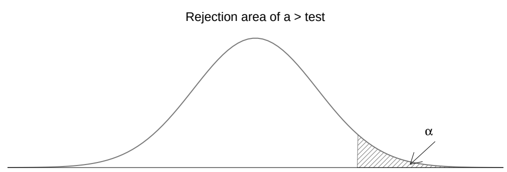


Figure 1: Rejection areas of two-sided tests

1.13 Rejection areas and α-levels





1.14 Rejection areas and α-levels

The probability that a test statistic falls within the rejection are when the null hypothesis is true is exactly the α-level of the hypothesis test.

In order to define rejection ares one needs to decide:

- What is the direction of the test? (one- or two-sided test)
- What is an acceptable α-level for the test.

1.15 p-values

p-values
A **p-value** is the probability of receiving as improbable value or an value even more improbable as the one received with the measurements if the null hypothesis is true. The H_0 shall be rejected if the p-value is less then α. If the p-value is greater then α the null hypothesis cannot be rejected.

Power
The **power** of a hypothesis test is the probability of rejecting a null hypothesis that is not true. It is denoted with $1-\beta$.

1.16 Errors of type I and II

Type I error
Type I error is the error of rejecting a null hypothesis that was true. The probability of a type I error is the α-level of the hypothesis test.

Type II error
Type II error is the error of not rejecting a null hypothesis that was not true. The probability of a type II error is β, where $1-\beta$ is the power of the hypothesis test.

	H_0 is true	H_0 is false
Reject H_0	Type I error Probability: α	Right decision Probability: $1-\beta$
Not reject H_0	Right decision Probability: 1-α	Type II error Probability: β

1.17 Not rejecting a null hypothesis

There can be various reasons behind one not rejecting a null hypothesis:

- The number of measurements was to small and therefore the hypothesis test had little power.
- The null hypothesis is true.
- Our model does not fit the measurements - the assumptions we made about the measurements do not hold.

We may never claim which one of the following cases was the reason!

But we may make arguments for one reason being the most plausible.

1.18 Conducting hypothesis tests

Conducting hypothesis tests

1 Decide which hypothesis test is appropriate for our measurements.

2 Decide the α-level.

3 Propose a null hypothesis and decide the direction of the test (one- or two-sided).

4 Calculate the test statistic for the hypothesis test.

5a See whether the test statistic falls within the rejection interval.

5b Look at the p-value of the test statistic.

6 Draw conclusions.

1.19 The relationship between confidence intervals and hypothesis tests

If the α-level is the same for both the confidence interval and the hypothesis test, the following are equivalent:

- We **reject** the null hypothesis that a particular statistic has a certain value.
- The confidence interval calculated does **not** contain that value.

If we conduct an hypothesis test with the α-level 5% and calculate a 95% confidence interval:

- We reject the null hypothesis that the statistic is equal to the number 1 if the number 1 is not within the confidence interval.
- The number 1 is not within the confidence interval if we reject the null hypothesis that the statistic is equal to the number 1.

2 Inference on the mean of a population

2.1 Hypothesis tests for μ

- In this lecture we will discuss hypothesis tests and confidence intervals that apply when making inference on the mean of a population, μ.
- We will use them for example to test the hypothesis that mean precipitation in Reykjavik in june is less then 50 mm, that the average heart rate of men over fifty years old is greater then 99 beats p/minute, that the average number of nights slept in hotels and hostels in june differs from 100000 and so and so forth.
- All hypothesis test that will be discussed have the same null hypothesis, that the mean of the population is equal to a certain value that is called μ_0.

2.2 Hypothesis tests for μ

The null hypothesis is written:

$$H_0 : \mu = \mu_0$$

.

- It depends on the direction of the hypothesis test, what conclusions are made if we reject the null hypothesis.
- If the hypothesis test is two sided, we can conclude that the mean of the population, μ, differs from μ_0.
- If it is one sided we can only conclude that it is greater in one case or less in the other case then μ_0 depending on the case.

2.3 Hypothesis tests for μ

- Circumstances can be very different when we make inference on the mean of a population and we categorize them into four different cases, but each case is treated differently.
- The decision tree on slide **??** shows which case corresponds to which circumstance , but in order to select the appropriate case we need to answer three questions that are shown on slide **??**.
- The question are about the probability distribution of the population, whether it's variance, σ^2 is known and the size of the population, n. Each case is discussed separately in the following slides

2.4 Decision tree

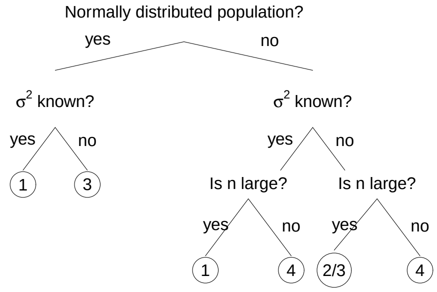


Figure 2: Decision tree for μ

2.5 Decision tree

These three questions are answered in correct order and the answers decide how we trace us down the decision tree.

1. Is the population normally distributed
   This need to be based on prior experience or by looking at the distribution of the sample and conclude from that. It can though be douptful if the sample is small.

2. Is the variance of the **population**, σ^2, known?
   Notice that this is rarely the case, although it may happen that such detailed prior investigations have been made that we can assume that the variance is known.

3. Is the sample large?
   We use the rule of thumb that n is large if $n > 30$. This is not a universal rule though.

2.6 Conducting hypothesis tests

Conducting hypothesis tests

1 Decide which hypothesis test is appropriate for our measurements.

2 Decide the α-level.

3 Propose a null hypothesis and decide the direction of the test (one- or two-sided).

4 Calculate the test statistic for the hypothesis test.

5a See whether the test statistic falls within the rejection interval.

5b Look at the p-value of the test statistic.

6 Draw conclusions.

2.7 Case 1

Case 1 corresponds to:

- When one can assume that the population follows a normal distribution and the variance (σ^2) of the distribution is known.
- when n is large and σ^2 is known, although the population is not normally distributed.

2.8 Confidence interval for μ - case 1

2.9 Hypothesis test for μ - case 1

John is producing fish to export. The packages he exports follow a normal distribution with known variance, σ^2 = 0.8 kg. He wants to make a 95% confidence interval for μ and test the hypothesis that μ is different from 50 kg. He therefore takes a random sample of $n = 12$ and calculates the sample average to be 50.84 kg. Use $\alpha = 0.05$.

We have: $\bar{x} = 50.84$, $n = 12$, $\sigma^2 = 0.8$. We use the normal table to find $z_{1-\alpha/2}$. $z_{1-\alpha/2} = z_{1-0.05/2} = z_{0.975} = 1.96$. The lower limit is:

$$\bar{x} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 50.84 - 1.96 \cdot \frac{\sqrt{0.8}}{\sqrt{12}} = 50.84 - 0.506 = 50.33$$

and the upper limit:

$$\bar{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 50.84 + 0.506 = 51.35.$$

The confidence interval is:

$$50.33 < \mu < 51.35$$

We test the hypothesis using six steps:

1. We are testing a hypothesis concerning μ and the variance of the distribution is known.
2. Við fengum uppgefið að nota $\alpha = 0.05$.
3. The hypotheses are
$$\begin{aligned} H_0 &: \mu = 50 \\ H_1 &: \mu \neq 50 \end{aligned}$$
4. We have: $\bar{x} = 50.84$, $n = 12$, $\sigma^2 = 0.8$. The value of the test statistic is
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{50.84 - 50}{\sqrt{0.8}/\sqrt{12}} = 3.25.$$
5. $z_{1-\alpha/2} = z_{0.975} = 1.96$. We reject the null hypothesis if $z < -1.96$ or $z > 1.96$. We see that $z > 1.96$.
6. We reject the null hypothesis and conclude that the boxes are heavier than 50 kg.

2.10 μ - case 2

Case 2 corresponds to:

- when the sample is large and we do not know the variance of the population. We do not need to assume that the population is normally distributed.

Be careful! One can always calculate the variance of the **sample** but the variance of the **population** is rarely known!

2.11 μ - case 2

As the variance of the population is not known, we use the variance of the sample to estimate the variance of the population with

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

In order to find the standard deviation of the sample, we take the square root of the variance

$$s = \sqrt{s^2}.$$

2.12 Confidence interval for μ - case 2

2.13 Hypothesis test for μ - case 2

Hypothesis test for μ - case 2
The null hypothesis is:

$$H_0 : \mu = \mu_0$$

The test statistic is:

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

If the null hypothesis is true, the test statistic follows the standardized normal distribution, or $Z \sim N(0,1)$.
The rejection areas are:

Alternative hypothesis	Reject H_0 if:
$H_1 : \mu < \mu_0$	$Z < -z_{1-\alpha}$
$H_1 : \mu > \mu_0$	$Z > z_{1-\alpha}$
$H_1 : \mu \neq \mu_0$	$Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$

Gugu is the CEO in a car company. She claims that the cars the company produces can drive 20 km. pr. liter. Ingibjorg is working for a consumer board and she has had many complains that the cars cannot drive 20 km. pr. liter. Therefore she decided to conduct an experiment where she wanted to test the hypothesis that the average number of kilometers is fewer than 20. A random sample of 40 cars was taken and the average number of kilometers calculated to be 19.2 and standard deviation of 1.7. Test the hypothesis. Use $\alpha = 0.05$.

1. We want to test a hypothesis regarding μ. We do not know the probability distribution not the variance of that distribution but the sample is large.
2. $\alpha = 0.05$.
3. The hypotheses are
$$\begin{aligned} H_0 &: \mu = 20 \\ H_1 &: \mu < 20 \end{aligned}$$
4. We have $\bar{x} = 19.2,\ s = 1.7,\ n = 40$. The test statistic is:
$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{19.2 - 20}{1.7/\sqrt{40}} = -2.98.$$
5. $-z_{1-\alpha} = -z_{0.95} = -1.64$. We reject the null hypothesis if $z < -1.64$. We see that $z < -1.64$.
6. We reject the null hypothesis and conclude that the average consumption of gas is greater than 20 km per liter.

2.14 μ - case 3

Case 3 corresponds to two cases:
In both cases, the variance (σ^2) of the **population**, to which the sample belongs, unknown. On the other hand, we either need to assume that:

- the population is normally distributed
- or that we have many measurements in our sample (then the population does not have to be normally distributed). This is the same as case 2.

When calculating confidence intervals and conducting hypothesis test in this case, one uses the t-distribution.

2.15 Of the overlap of case 2 and 3

- Notice that when case 2 (which uses z-test) is valid, one can successfully use case 3 instead (which uses t-test). This is because when the number of degrees of freedom is large, the t-distribution is similar to the normal distribution.

- T-test, unlike z-tests, are built in most statistical software and therefore more used.

- If we are doing calculations by hand it is often better to use z-tests because then we can easily calculate p-values.

2.16 Confidence interval for μ - case 3

2.17 Hypothesis test for μ - case 3

A cigarette producer states that the a certain types of cigarettes has on average 14 mg of nicotine per cigarette. Health authorities wanted to investigate if is was in fact higher. Therefore they made an experiment to test the hypothesis that the average nicotine level is higher than 14 gr./cigarette. A random sample of size 12 was taken and the average level found to be 14.3 and the standard deviation of 0.9. Test the hypothesis using $\alpha = 0.05$. We can assume that the nicotine level follows a normal distribution.

1. We want to test an hypothesis regarding μ we can assume that the population follows a normal distribution, σ^2 is unknown.
2. $\alpha = 0.05$.
3. The hypothesis are:
$$\begin{aligned} H_0 &: \mu = 14 \\ H_1 &: \mu > 14 \end{aligned}$$
4. We get: $n = 12$, $\bar{x} = 14.3$, $s = 0.9$. The value of the test statistic is:
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{14.3 - 14}{0.9/\sqrt{12}} = 1.15.$$
5. We have $n-1 = 11$ degrees of freedom: $t_{1-\alpha,(n-1)} = t_{0.95,(11)}$ = 1.796. We reject the null hypothesis if $t > 1.796$. We see that $t < 1.796$.
6. We cannot reject the null hypothesis so we cannot conclude that the average level is higher than 14 mg/cigarette.

2.18 μ - case 4

In case 4 one can neither use z-test nor t-test unless further approximations are used. In these cases one can do one of the following:

- Transform the data
- Use nonparametric tests
- Check whether the population follows some other known distribution and use tests that are applicable for them.

3 Inference on the means of two populations

3.1 Inference on the mean of two populations

- In this lecture we will discuss hypothesis tests and confidence intervals that apply when making inference on the mean of two populations.
- The means are named μ_1 and μ_2 and we wish to make inference on their difference, $\mu_1 - \mu_2$.
- The tests applied can broadly be divided into two groups:
  - Tests for independent measurements.
  - Tests for paired measurements.

3.2 Independent or paired?

The first question we need to ask is if the measurements are independent or paired. Examples of tests for independent measurements:

- Height of 50 men and 50 women used to test the hypothesis that men are on average taller then women.
- The heart rate of 30 women in the age 41-50 and 30 women in the age 51-60 measured to test the hypothesis that there is a difference in the heart rate of women in these two age groups.

Examples of tests for paired measurements:

- The weight of 30 men before they undergo an intensive workout program. The weight is measure again after the program to test the hypothesis that the workout is successful for loosing weight.
- The age of 40 men and their wives is noted to test the hypothesis that in marriages of men and women are the men on average older then the women.

3.3 Conducting hypothesis tests

Conducting hypothesis tests

1 Decide which hypothesis test is appropriate for our measurements.

2 Decide the α-level.

3 Propose a null hypothesis and decide the direction of the test (one- or two-sided).

4 Calculate the test statistic for the hypothesis test.

5a See whether the test statistic falls within the rejection interval.

5b Look at the p-value of the test statistic.

6 Draw conclusions.

3.4 Independent measurements

- All hypothesis tests in this lecture test the same null hypothesis, whether the difference of the two means is equal to a certain value that we call δ.
- The null hypothesis is $H_0 : \mu_1 - \mu_2 = \delta$.
- It depends on the direction of the hypothesis test, what conclusions are made if we reject the null hypothesis.
- If the hypothesis test is two sided, we can conclude that the difference of the means, $\mu_1 - \mu_2$, differs from δ.
- If it is one sided we can only conclude that the difference is greater in one case or less in the other case then δ depending on the case.

3.5 Independent measurements

- As with one mean, we use different tests for different circumstances.
- The circumstances are categorized into five cases:
- The decision tree shows which case corresponds to which circumstance , but in order to select the appropriate case we need to answer four questions that are shown on an upcoming slide.
- We note the mean, variance and sample size of one population with μ_1, σ_1^2 and n_1 but the other with μ_2, σ_2^2 and n_2.

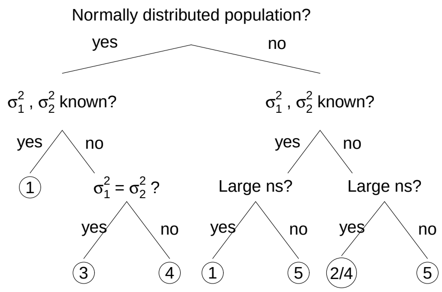


Figure 3: Decision tree for $\mu_1 - \mu_2$

3.6 Decision tree - independent measurements

3.7 Decision tree - independent measurements

1. Are the populations normally distributed?
   This need to be based on prior experience or by looking at the distributions of the samples and conclude from that. It can though be douptful if the samples are small.

2. Is the variance of the **populations**, σ_1^2, σ_2^2, known?
   Notice that this is rarely the case, although it may happen that such detailed prior investigations have been made that we can assume that the variance is known.

3. Are the samples large?
   We use the rule of thumb that the samples are large if $n_1 > 30$ and $n_2 > 30$. This is not a

3.8 $\mu_1 - \mu_2$ - case 1

Case one applies when:

- it can be assumed that the populations are normally distributed and the variances of the populations, (σ_1^2 and σ_2^2) are known.
- when n_1 and n_2 are large and σ_1^2 and σ_2^2 are known, although the populations are not normally distributed.

3.9 Confidence interval for the difference of two means - case 1

Confidence interval for the difference of two means - case 1
Lower bound of $1-\alpha$ confidence interval is:

$$\bar{x}_1 - \bar{x}_2 - z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Upper bound of $1-\alpha$ confidence interval is:

$$\bar{x}_1 - \bar{x}_2 + z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

3.10 Confidence interval for the difference of two means - case 1

Confidence interval for the difference of two means - case 1
The confidence interval is:

$$\bar{x}_1 - \bar{x}_2 - z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $\bar{x}_1$, $\bar{x}_2$ are the sample means and σ_1^2, σ_2^2 are the population variances. $z_{1-\alpha/2}$ is found in the standardized normal distribution table.

3.11 Hypothesis test for the difference of two means - case 1

Hypothesis test for the difference of two means - case 1
The null hypothesis is:

$$H_0 : \mu_1 - \mu_2 = \delta$$

The test statistic is:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

If the null hypothesis is true, the test statistic follows the standardized normal distribution, or $Z \sim N(0,1)$.

Alternative hypothesis	Reject H_0 if:
$H_1 : \mu_1 - \mu_2 < \delta$	$Z < -z_{1-\alpha}$
$H_1 : \mu_1 - \mu_2 > \delta$	$Z > z_{1-\alpha}$
$H_1 : \mu_1 - \mu_2 \neq \delta$	$Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$

Notice that δ can be any number at all, but in most cases $\delta = 0$.

3.12 $\mu_1 - \mu_2$ - case 2

Case 2 applies when:

- we do not know the population variances (σ_1^2 and σ_2^2) but the samples are large. We do not need to assume that the populations are normally distributed.
- In this case one can successfully use case 4, but that is built in most statistical software (such as R). When calculating in hands case 2 is easier though.

As the variance of the population is not known, we use the variance of the sample to estimate the variance of the population with

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

In order to find the standard deviation of the sample, we take the square root of the variance

$$s = \sqrt{s^2}.$$

These values are calculated for each sample separately and named s_1 and s_2 as appropriate.

3.13 Confidence interval for the difference of the mean of two populations - case 2

Confidence interval for the difference of the mean of two populations - case 2
Lower bound of $1-\alpha$ confidence interval is:

$$\bar{x}_1 - \bar{x}_2 - z_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Upper bound of $1-\alpha$ confidence interval is:

$$\bar{x}_1 - \bar{x}_2 + z_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

3.14 Confidence interval for the difference of the mean of two populations - case 2

Confidence interval for the difference of the mean of two populations - case 2
The confidence interval is:

$$\bar{x}_1 - \bar{x}_2 - z_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + z_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $\bar{x}_1$, $\bar{x}_2$ are the sample means and s_1^2, s_2^2 are the sample variances. $z_{1-\alpha/2}$ is found in the standardized normal distribution table.

3.15 Hypothesis test for the difference of the means of two populations - case 2

Hypothesis test for the difference of the means of two populations - case 2
The null hypothesis is:

$$H_0 : \mu_1 - \mu_2 = \delta$$

The test statistic is:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

If the null hypothesis is true, the test statistic follows the standardized normal distribution, or $Z \sim N(0,1)$.

Alternative hypothesis	Reject H_0 if:
$H_1 : \mu_1 - \mu_2 < \delta$	$Z < -z_{1-\alpha}$
$H_1 : \mu_1 - \mu_2 > \delta$	$Z > z_{1-\alpha}$
$H_1 : \mu_1 - \mu_2 \neq \delta$	$Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$

Notice that δ can be any number at all, but in most cases $\delta = 0$.

3.16 $\mu_1 - \mu_2$ - case 3

Case 3 applies when:

- One can assume that the populations are normally distributed, the variances (σ_1^2 and σ_2^2) of the populations are unknown, but we assume that $\sigma_1^2 = \sigma_2^2$.

Later on we will see how to test this hypothesis formally, but until then we will use the rule of thumb that if one sample variance is more then four times greater then the other, we cannot assume that the variances of the populations are equal.
The t-distribution is used for calculating confidence intervals and hypothesis testing in this case.

3.17 $\mu_1 - \mu_2$ - case 3

Before we can calculate confidence intervals and conduct hypothesis tests we need to calculate the **pooled variance** of the samples, which is denoted s_p^2.

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

where s_1^2 and s_2^2 are calculated in the same way as earlier.

3.18 Confidence interval for the difference of the mean of two populations - case 3

Confidence interval for the difference of the mean of two populations - case 3
Lower bound of $1-\alpha$ confidence interval is:

$$\bar{x}_1 - \bar{x}_2 - t_{1-\alpha/2,(n_1+n_2-2)} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Upper bound of $1-\alpha$ confidence interval is:

$$\bar{x}_1 - \bar{x}_2 + t_{1-\alpha/2,(n_1+n_2-2)} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $\bar{x}_1$, $\bar{x}_2$ are the sample means and s_1^2, s_2^2 are the sample variances . $t_{1-\alpha/2,(n_1+n_2-2)}$ is found in the t-distribution table.

3.19 Hypothesis test for the difference of the means of two populations - case 3

Hypothesis test for the difference of the means of two populations - case 3
The null hypothesis is:

$$H_0 : \mu_1 - \mu_2 = \delta$$

The test statistic is:

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

If the null hypothesis is true, $T \sim t_{(n_1+n_2-2)}$.

Alternative hypothesis	Reject H_0 if:
$H_1 : \mu_1 - \mu_2 < \delta$	$T < -t_{1-\alpha,(n_1+n_2-2)}$
$H_1 : \mu_1 - \mu_2 > \delta$	$T > t_{1-\alpha,(n_1+n_2-2)}$
$H_1 : \mu_1 - \mu_2 \neq \delta$	$T < -t_{1-\alpha/2,(n_1+n_2-2)}$ or $T > t_{1-\alpha/2,(n_1+n_2-2)}$

Notice that δ can be any number at all, but in most cases $\delta = 0$.

Ingunn og Arni are interested in investigating if there is a difference in mean salaries of males and females working in fisheries in Iceland. Random samples were taken, 20 males and 20 females. Mean and standard deviation of male salaries were 245163 kr and 22814. Mean and standard deviation in the female sample were 218634 and 18312. Test the hypothesis that there is a difference in average salaries between males and females. Use $\alpha = 0.05$. We can assume that the salaries follow a normal distribution.

1. We want to compare the means of two populations. The samples are independent. We can assume that the salaries follow a normal distribution, we dont know the variances but we can assume that they are the same.

2. $\alpha = 0.05$.

3. The hypotheses are:

$$H_0 \quad : \quad \mu_1 - \mu_2 = 0$$
$$H_1 \quad : \quad \mu_1 - \mu_2 \neq 0.$$

4. The value of the test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

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

We have that $n_1 = 20$, $n_2 = 20$, $\bar{x_1}$ = 245163, s_1 = 22814,
$\bar{x_2}$ = 218634, s_2 = 18312 og δ = 0.

$$s_p = \sqrt{\frac{(20-1)\cdot 22814^2 + (20-1)\cdot 18312^2}{20+20-2}} = 20685.84$$

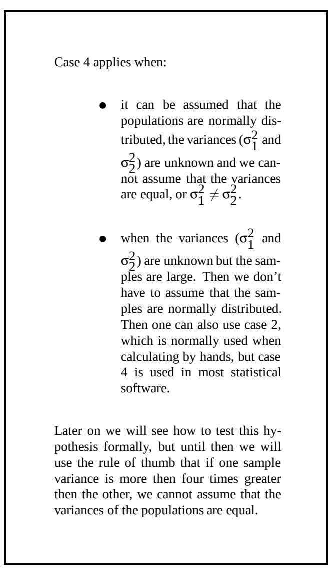
and

$$t = \frac{245163 - 218634 - 0}{20685.84\sqrt{\frac{1}{20}+\frac{1}{20}}} = 4.06.$$

5. We look up after: $n_1 + n_2 - 2 = 38$. $t_{1-\alpha/2,(n_1+n_2-2)} = t_{0.975,(38)}$ = 2.024, so we reject the null-hypothesis if $t > 2.024$ or $t < -2.024$. We see that $t > 2.024$.

6. We reject the null hypothesis and conclude that there is a difference between males and females in the average salaries in fisheries in Iceland. jöfn.

3.20 $\mu_1 - \mu_2$ - case 4



Case 4 applies when:

- it can be assumed that the populations are normally distributed, the variances (σ_1^2 and σ_2^2) are unknown and we cannot assume that the variances are equal, or $\sigma_1^2 \neq \sigma_2^2$.

- when the variances (σ_1^2 and σ_2^2) are unknown but the samples are large. Then we don't have to assume that the samples are normally distributed. Then one can also use case 2, which is normally used when calculating by hands, but case 4 is used in most statistical software.

Later on we will see how to test this hypothesis formally, but until then we will use the rule of thumb that if one sample variance is more then four times greater then the other, we cannot assume that the variances of the populations are equal.

3.21 $\mu_1 - \mu_2$ - case 4

In this case the confidence interval and the test statistic resembles the one in case 2 but here it follows the t-distribution. The number of degrees of freedom in this t-distribution is denoted with ν and calculated by

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

where s_1^2 and s_2^2 are calculated by same methods as earlier. This hypothesis test is rarely done by hands but a statistical software used for the calculations.

3.22 Confidence interval for the difference of the mean of two populations - case 4

Confidence interval for the difference of the mean of two populations - case 4
Lower bound of $1-\alpha$ confidence interval is:

$$\bar{x}_1 - \bar{x}_2 - t_{1-\alpha/2,(\nu)} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Upper bound of $1-\alpha$ confidence interval is:

$$\bar{x}_1 - \bar{x}_2 + t_{1-\alpha/2,(\nu)} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

3.23 Confidence interval for the difference of the mean of two populations - case 4

Confidence interval for the difference of the mean of two populations - case 4
The confidence interval is:

$$\bar{x}_1 - \bar{x}_2 - t_{1-\alpha/2,(\nu)} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + t_{1-\alpha/2,(\nu)} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $\bar{x}_1$, $\bar{x}_2$ are the sample means and s_1^2, s_2^2 are the sample variances . $t_{1-\alpha/2,(\nu)}$ is found in the t-table. ν is the number of degrees of freedom.

3.24 Hypothesis test for the difference of the means of two populations - case 4

Hypothesis test for the difference of the means of two populations - case 4
The null hypothesis is:

$$H_0 : \mu_1 - \mu_2 = \delta$$

The test statistic is:

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

If the null hypothesis is true, the test statistic t-dreifingu með ν frígráðum or $T \sim t(\nu)$ where ν is calculated as shown in slide **??**

Alternative hypothesis	Reject H_0 if:
$H_1 : \mu_1 - \mu_2 < \delta$	$T < -t_{1-\alpha,(\nu)}$
$H_1 : \mu_1 - \mu_2 > \delta$	$T > t_{1-\alpha,(\nu)}$
$H_1 : \mu_1 - \mu_2 \neq \delta$	$T < -t_{1-\alpha/2,(\nu)}$ or $T > t_{1-\alpha/2,(\nu)}$

Notice that δ can be any number at all, but in most cases $\delta = 0$.

3.25 $\mu_1 - \mu_2$ - case 5

In case 5 one can neither use z-test nor t-test unless further approximations are made. In these cases one of the following can be done:

- Transform the data
- Use resampling methods
- Use nonparametric tests
- Test whether the measurements follow any known distributions and look at tests that apply to them.

3.26 Paired measurements

- Assume that we have n pair of measurements (X_i, Y_i), $i = 1, 2, 3 \dots n$.
- We need to find the differences of these pairs:

$$D_i = X_i - Y_i.$$

D_i is a random variable of size n from a population with mean μ_D.

- The hypothesis tests make inference on μ_D.

3.27 Paired measurements

Before conducting hypothesis tests, the following statistics need to be calculated:

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

which is the mean of the differences and

$$s_D{}^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}$$

is the standard deviation of the differences.

3.28 Paired measurements

- We test the null hypothesis that the mean of the differences is equal to a certain value that is denoted $\mu_{D,0}$.
- The null hypothesis is $H_0 : \mu_D = \mu_{D,0}$.
- It depends on the direction of the hypothesis test, what conclusions are made if we reject the null hypothesis.
- If the hypothesis test is two sided, we can conclude that the difference of the means, μ_D, differs from $\mu_{D,0}$.
- If it is one sided we can only conclude that the difference is greater in one case or less in the other case then $\mu_{D,0}$ depending on the case.

3.29 Paired measurements

- We test the null hypothesis that the mean of the differences is equal to a certain value that is denoted $\mu_{D,0}$.
- The null hypothesis is $H_0 : \mu_D = \mu_{D,0}$.
- It depends on the direction of the hypothesis test, what conclusions are made if we reject the null hypothesis.
- If the hypothesis test is two sided, we can conclude that the difference of the means, μ_D, differs from $\mu_{D,0}$.
- If it is one sided we can only conclude that the difference is greater in one case or less in the other case then $\mu_{D,0}$ depending on the case.

3.30 Paired measurements

- It depends on how many pairs of measurements we have en whether it can be assumed that the difference of the measurements is normally distributed if we use a z-test or a t-test to make inference on μ_D.
- If n is large, which here denotes the number of pairs, we can always use the z-test.
- The t-test can be used if μ_D is normally distributed and/or the sample is large.
- When we use a statistical software the t-test is preferred to the z-test when both tests are valid (when n is large).

3.31 Inference on paired measurements, n large

Inference on paired measurements, n large
The null hypothesis is:

$$H_0 : \mu_D = \mu_{D,0}$$

The test statistic is:

$$Z = \frac{\overline{D} - \mu_{D,0}}{S_D/\sqrt{n}}$$

If the null hypothesis is true, the test statistic follows the standardized normal distribution, or $Z \sim N(0,1)$.

Alternative hypothesis	Reject H_0 if:
$H_1 : \mu_D < \mu_{D,0}$	$Z < -z_{1-\alpha}$
$H_1 : \mu_D > \mu_{D,0}$	$Z > z_{1-\alpha}$
$H_1 : \mu_D \neq \mu_{D,0}$	$Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$

$z_{1-\alpha/2}$ is found in the standardized normal distribution table.

3.32 Inference on paired measurements, normally distributed differences and/or large n

When n is small the difference of the measurements need to be normally distributed.

Inference on paired measurements, normally distributed differences and/or large n
The null hypothesis is:

$$H_0 : \mu_D = \mu_{D,0}$$

The test statistic is:

$$T = \frac{\overline{D} - \mu_{D,0}}{S_D/\sqrt{n}}$$

If the null hypothesis is true, the test statistic is t-distributed with $(n-1)$ degrees of freedom, or $T \sim t_{(n-1)}$.

3.33 Inference on paired measurements, normally distributed differences and/or large n

Inference on paired measurements, normally distributed differences and/or large n

Alternative hypothesis	Reject H_0 if:
$H_1 : \mu_D < \mu_{D,0}$	$T < -t_{1-\alpha,(n-1)}$
$H_1 : \mu_D > \mu_{D,0}$	$T > t_{1-\alpha,(n-1)}$
$H_1 : \mu_D \neq \mu_{D,0}$	$T < -t_{1-\alpha/2,(n-1)}$ or $T > t_{1-\alpha/2,(n-1)}$

$t_{1-\alpha/2,(n-1)}$ is found in t-table

An experiments has been performed to asses the hypothesis that over weighted males could lose weight by exercising in 30 minutes per day for two months. A random sample of 6 over weighted males was taken and the males weighted before and after the two months of exercise. The results can be seen below. Test the hypothesis using $\alpha = 0.05$. It can be assumed that weight follows normal distribution.

Individual	Befire [kg]	After [kg]
1	123	120
2	112	108
3	107	106
4	101	99
5	112	112
6	116	114

We need D_i, $\overline{D}$ and S_D.

Individual	Before (x_i)	After (y_i)	$d_i = x_i - y_i$
1	123	120	3
2	112	108	4
3	107	106	1
4	101	99	2
5	112	112	0
6	116	114	2

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{3+4+1+2+0+2}{6} = 2$$

$${s_d}^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1} = \frac{(3-2)^2 + (4-2)^2 + ... + (2-2)^2}{5} = 2.$$

1. We have paired measurements.

2. $\alpha = 0.05$.

3. The hypotheses are:

$$H_0 \quad : \quad \mu_D = 0$$
$$H_1 \quad : \quad \mu_D > 0$$

4. The value of the test statistic is:

$$t = \frac{\overline{d} - \mu_{D,0}}{s_D/\sqrt{n}} = \frac{2-0}{\sqrt{2/6}} = 3.46.$$

5. We look up after $n-1=5$ degrees of freedom: $t_{1-\alpha} = t_{0.95,(n-1)}$ = 2.015, and we reject the null-hypothesis if $t > 2.015$. We see that $t > 2.015$.

6. We reject the null hypothesis and conclude that males lose weight on average after doing exercises for 30 minutes/day for two months.

4 Analysis of variance (ANOVA)

4.1 Introduction

- In lecture 110 we discussed inference on the mean of a population (μ).
- In the former part of lecture 120 we discussed inference on the difference of the mean of two populations ($\mu_1 - \mu_2$).
- In the latter part of lecture 120 we discussed inference on paired measurements (μ_D).
- Now we will discuss a method that we can apply to compare the means of two or more populations. The method is called analysis of variance, or ANOVA.

4.2 Analysis of variance

- Analysis of variance is one of the most commonly used statistical methods. There are several variants of it that can be used in a vast number of various different cases.
- We will only look at one variant of the method that is called *one-sided ANOVA*.
- It is applied to data that contain samples from two or more populations and it is common to speak of groups when discussing the samples.
- The method compares the variability of the measurements within the groups on one hand and between them on the other hand.
- ANOVA assumes that the samples are random samples, that they are sampled from populations with a normal distribution and that the variance is the same in all populations.

4.3 Conducting hypothesis tests

Conducting hypothesis tests

1 Decide which hypothesis test is appropriate for our measurements.

2 Decide the α-level.

3 Propose a null hypothesis and decide the direction of the test (one- or two-sided).

4 Calculate the test statistic for the hypothesis test.

5a See whether the test statistic falls within the rejection interval.

5b Look at the p-value of the test statistic.

6 Draw conclusions.

4.4 One sided ANOVA - example of application

A pharmaceutical company is testing new blood pressure medicine and conducts a little experiment. Eighteen individuals participated in the experiment and they were randomly allocated to three groups. Group one got drug 1, group two drug 2 and group three drug 3. The blood pressure was measured before and after the intake of the drug. The variable of interest is the difference in blood pressure before and after the drug intake. The mean difference blood pressure in the three groups was calculated. In all cases the blood pressure had decreased on average.

Average change group 1: $\bar{y}_{1.} = 8.14$
Average change group 2: $\bar{y}_{2.} = 6.28$
Average change group 3: $\bar{y}_{3.} = 13.01$

The question is, do the drug decrease the blood pressure equally or not?

4.5 The data

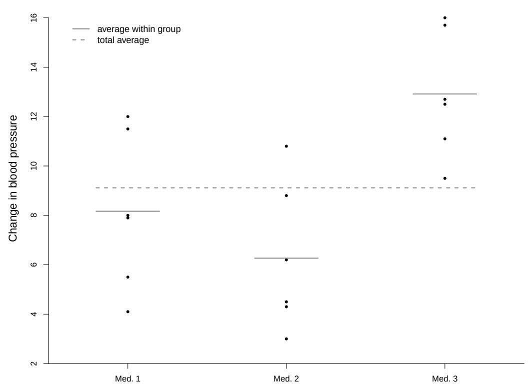


Figure 4: Data for ANOVA

4.6 Syntax

4.7 Sums of squares

- We need to calculate three sums of squares, and are they denoted with SS_T, SS_{Tr} and SS_E.
- SS_T is the total sums of squares and is a measure of the total variation of the measurements.
- SS_{Tr} is a measure of the variation between groups (or treatments), that is, how much to the means of the groups vary.
- SS_E is a measure of the variability within groups (or treatments) and is therefore a measure of the error. It shows how much the measurements deviate from the mean of the group.

4.8 Sums of squares

4.9 Sums of squares - graphically

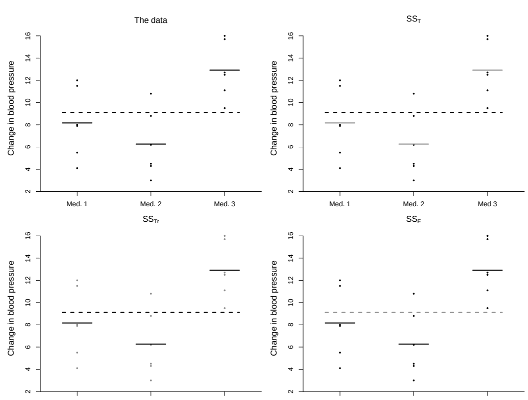


Figure 5: Sums of squares

4.10 ANOVA table

- It is common to visualize the sums of squares in a so-called *ANOVA table*.
- The table consist of three columns and three lines.
- The first column contains the sums of squares, the next one contains the number of *degrees of freedom*. The first column contains the sums of squares, the second one contains the number of *degrees of freedom* for each sum of squares and the third column contains so-called mean sum of squares.
- Mean sum of squares is calculated by dividing the corresponding sum of squares with the number of corresponding degrees of freedom (in the same line).

4.11 ANOVA table

Sums of squares	Degrees of freedom	Mean sum of squares
SS_{Tr}	$a-1$	$MS_{Tr} = \frac{SS_{Tr}}{a-1}$
SS_E	$N-a$	$MS_E = \frac{SS_E}{N-a}$
SS_T	$N-1$	

4.12 Hypothesis testing with ANOVA

4.13 Hypothesis testing with one-sided ANOVA

- The alternative hypothesis is that at least one of the means differs from the others, it is therefore then only information we receive if the null hypothesis is rejected.
- We do not know which of the means differs from the others or if they are potentially all different.
- Further analysis needs to be done in order to find that out. A common test i Tukey's test, but they will not be covered in this lecture.

Look at the example from the lecture about the blood pressure medicine. The data are the following:

Medicine 1	Medicine 2	Medicine 3
4.29	10.32	12.89
11.28	3.23	15.68
5.37	4.51	16.03
7.89	4.57	9.43
8.10	8.85	12.86
11.93	6.23	11.15

1. We would like to compare three mean values, the samples are independent, the variance is similar so we use analysis of variance.
2. $\alpha = 0.05$ að venju.
3. The hypotheses are
$$H_0 : \mu_1 = \mu_2 = \mu_3$$
og
$$H_1 : \text{at least one mean different from the others.}$$
4. We need to calculate the sums of squares.

   We have three groups so $a = 3$. We have six measurements per group so $n_1 = n_2 = n_3 = 6$ and $N = 6+6+6 = 18$. The grand mean is:
$$\bar{y}_{..} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}}{N} = \frac{4.29+10.32+12.98+11.28+\ldots+11.15}{18} = 9.15$$
   and the averages within the groups:
$$\bar{y}_{1.} = \frac{\sum_{j=1}^{n_1} y_{1j}}{n_1} = \frac{4.29+11.28+\ldots+11.93}{6} = 8.14,$$
$$\bar{y}_{2.} = \frac{\sum_{j=1}^{n_2} y_{2j}}{n_2} = \frac{10.32+3.23+\ldots+6.23}{6} = 6.29,$$
$$\bar{y}_{3.} = \frac{\sum_{j=1}^{n_3} y_{3j}}{n_3} = \frac{12.89+15.68+\ldots+11.15}{6} = 13.01.$$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{..})^2$$
$$= (4.29 - 9.15)^2 + (11.28 - 9.15)^2 + ... + (11.15 - 9.15)^2 = 262.16.$$

$$SS_{Tr} = \sum_{i=1}^{a} n_i(\overline{y}_{i.} - \overline{y}_{..})^2$$
$$= 6 \cdot (8.14 - 9.15)^2 + 6 \cdot (6.29 - 9.15)^2 + 6 \cdot (13.01 - 9.15)^2 = 144.53.$$

$$SS_E = SS_T - SS_{Tr} = 117.63.$$

Lets make a SS table:

SS	DF	MS
SS_{Tr} = 144.53	$a - 1$ = 2	$MS_{Tr} = 72.27$
SS_E = 117.63	$N - a$ = 15	$MS_E = 7.84$
SS_T = 262.16	$N - 1$ = 17	

The value of the test statistic:

$$F = \frac{MS_{Tr}}{MS_E} = \frac{72.27}{7.84} = 9.21.$$

5. We look up for $a - 1$ = 2 og $N - a$ = 15 degrees of freedom. $F_{1-\alpha,((a-1),(N-a))}$ = $F_{0.95,(2,15)}$ = 3.68. We see that $F > 3.68$.

6. We reject the null hypothesis and conclude that at leas one of the mean values is different from the others.

5 Inference on variances

5.1 Introduction

- In this lecture we will discuss inference on the variance of a normally distributed population and how to compare the variances in two normally distributed populations.
- First we will discuss confidence intervals and hypothesis test for the variance of a normally distributed population.
- Then we explore hypothesis tests that can be used when comparing the variances of two normally distributed populations.

5.2 Conducting hypothesis tests

Conducting hypothesis tests

1 Decide which hypothesis test is appropriate for our measurements.

2 Decide the α-level.

3 Propose a null hypothesis and decide the direction of the test (one- or two-sided).

4 Calculate the test statistic for the hypothesis test.

5a See whether the test statistic falls within the rejection interval.

5b Look at the p-value of the test statistic.

6 Draw conclusions.

5.3 Inference on the variance of a population

- In this section we discuss hypothesis tests and confidence intervals that apply when making inference on the variance of a normally distributed population, σ^2.
- When calculating confidence intervals and testing hypothesis for the variance of a population, the χ^2-distribution is used.
- The null hypothesis in this section is that the variance of the population equals some specific value that we denote σ_0^2.
- The null hypothesis is written $H_0 : \sigma^2 = \sigma_0^2$.
- It depends on the direction of the hypothesis test what conclusion are drawn if the null hypothesis is rejected.
- If the hypothesis test is two-sided we conclude that the variance of the population, σ^2, differs from σ_0^2 but if it is one-sided we can only conclude that the variance is greater or less than σ_0 depending on the case.

5.4 Confidence interval for the variance of a population

5.5 Inference on the variance of a population

A consumer group is investigating whether there is to little soda in cans from a certain soda factory. It is important that the filling process in the factory is stable, that is the variance is not higher than 10 ml^2, so that there is not many bottles with too much or too little soda. To investigate this an experiment was performed where a random sample of 30 bottles was taken and the standard deviation calculated to, $s = 3.5$. It can be assumed that the soda level follows a normal distribution. Test the hypothesis that the variance is higher than 10 ml^2. Use $\alpha = 0.05$.

1. We would like to test a hypothesis regarding the variance of a normal distribution.
2. $\alpha = 0.05$.
3. The hypotheses are:
$$H_0 \ : \ \sigma^2 = 10$$
$$H_1 \ : \ \sigma^2 > 10.$$
4. The test statistic is:
$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}.$$
$n - 1 = 30 - 1 = 29$, $\sigma_0 = 10$.
$$\chi^2 = \frac{29 \cdot 12.25}{10} = 35.53.$$
5. $\chi^2_{0.95,(29)} = 42.56$. We reject the null hypothesis if $\chi^2 > 42.56$. We see that $\chi^2 < 42.56$.
6. We cannot reject the null hypothesis so we cannot conclude that the variance is larger than 10 ml^2.

5.6 Inference on the variance of two populations

- The hypothesis tests that we discuss in this section are used to compare the variance of two populations that both are normally distributed.
- Tests of this kind are often conducted before hypothesis tests where the means of two populations are compared and the variance of the populations is unknown and the samples are not large.
- The null hypothesis i this section is that the variance of the two populations is equal, written $H_0 : \sigma_1^2 = \sigma_2^2$.
- If the hypothesis test is two-sided we can draw the conclusion that the variances are unequal, but if it is one-sided we can only draw the conclusion that the variance in one sample is greater then the variance in the other sample.

5.7 Hypothesis tests for the variances of two populations

Let us go back to an example form the lecture on difference between two mean values where we compared the average salaries fro males and females working in fisheries in Iceland. There we assumed that the variance in the two populations is the same. We are going to check that assumption know using the appropriate hypothesis test. Random samples were taken from both populations of size 20. The average and standard deviation in the make sample was 245163 kr and 22814. The average and standard deviation in the female sample was 218634 og 18312. Use $\alpha = 0.05$.

1. We are going to test whether the variances in two normally distributed populations is the same.
2. $\alpha = 0.05$.
3. The hypotheses are:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_0 &: \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

4. The test statistic is:

$$F = \frac{S_M^2}{S_m^2}.$$

$$F = \frac{22814^2}{18312^2} = 1.55.$$

5. $F_{1-\alpha/2,(n_M-1,n_m-1)} = F_{0.975,(19,19)}$. we use that value that is next, $F_{0.975,(20,19)} = 2.506$, so we reject the null-hypothesis if $F > 2.506$. We see that $F < 2.506$.
6. We cannot reject the null-hypothesis so we cannot conclude that the variances are different.

6 Inference on ratios and contingency tables

6.1 Estimate of the ratio of a population

In this section we will discuss confidence intervals and hypothesis tests for one ratio p that describes the ratio of subjects within a population that have a particular value of a categorical variable.

6.2 Bernoulli trial and the binomial distribution

Bernoulli trial
Every trial in a group of repeated trials is classified as a **Bernoulli trial** if the following holds:

1. Every trial has only two possible outcomes (positive and negative).

2. The probability of a positive outcome are the same in every trial.

3. The outcomes are indipendant.

The number of positive outcomes in n Bernoulli trials follows the **binomial distribution** with the parameters n and p, written $X \sim B(n, p)$, where p is the probability of a positive outcome.

6.3 Estimate of the ratio of a population

The ratio of the population, denoted p, is estimated with the sample proportion:

$$\widehat{p} = \frac{x}{n}$$

where x is the number of measurements that receive the corresponding outcome and n is the size of the sample.

6.4 Normal approximation

- When certain criteria is met, the binomial distribution is similar to the normal distribution.

- Then we can use methods that assume the characteristics of the normal distribution to make inference on random variables that in deed are binomially distributed.

- That is called to apply a **normal approximation**

When can one use normal approximation?
If $n\hat{p}$ and $n(1-\hat{p})$ are greater then 15, the normal approximation can be used to make inference on the proportion of a binomial distribution.

6.5 Confidence interval

> **Confidence interval for the ratio of a population**
> If the criteria for using the normal approximation is met, the lower bound for p can be calculated with:
>
> $$\widehat{p} - z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$
>
> and the upper bound with:
>
> $$\widehat{p} + z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$
>
> where $\widehat{p} = \frac{x}{n}$ and $z_{1-\alpha/2}$ is in the standardized normal distribution table

A company decided to make a pole to investigate whether over half of the population is in favour of the government. Out of the 8750 that were asked, 4530 said yes and 4220 said no. Find a 95% confidence interval for p, the ratio of those that are in favour of the government.

We start by finding $\widehat{p}$:

$$\widehat{p} = \frac{x}{n} = \frac{4530}{8750} = 0.5177.$$

The conditions for using the normal approximation are met since $n\hat{p}$ and $n(1-\hat{p})$ are both larger than 15.

The lower limit is:

$$\widehat{p} - z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0.5177 - 1.96 \cdot \sqrt{\frac{0.5177(1-0.5177)}{8750}} = 0.5072$$

and the upper limit:

$$\widehat{p} + z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0.5177 + 1.96 \cdot \sqrt{\frac{0.5177(1-0.5177)}{8750}} = 0.5282.$$

The confidence interval is:

$$0.5072 < p < 0.5282.$$

6.6 The null hypothesis

- The null hypothesis in this section tests the hypothesis that the ratio of the sample, p is equal to a certain value that we call p_0.
- The null hypothesis is written $H_0 : p = p_0$.
- If the test is two-sided we can conclude that the ratio p differs from p_0.
- If it is one sided we can only conclude that p is either greater or less than p_0, depending on the case.

6.7 Hypothesis test for the ratio of a population

Hypothesis test for the ratio of a population
If the criteria for using the normal approximation are met, the following hypothesis test can be used The null hypothesis is

$$\mathrm{H}_0 : p = p_0$$

The test statistic is

$$Z = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$

where X is the number of successful experiments and n is the size of the sample.
If the null hypothesis is true,

the test statistic follows the standardized normal distribution, or $Z \sim N(0, 1)$.

6.8 Alternative hypothesis for the ratio of a sample.

Alternative hypothesis for the ratio of a sample.
The alternative hypothesis along with the rejection areas are shown below.

Alternative hypothesis	Reject H_0 if:
$H_1 : p < p_0$	$Z < -z_{1-\alpha}$
$H_1 : p > p_0$	$Z > z_{1-\alpha}$
$H_1 : p \neq p_0$	$Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$

6.9 Inference on the ratio of two populations

We often want to compare the ratios of a certain value of a categorical variable in two populations.

We denote the ratios in the two populations with p_1 and p_2 and estimate them with

$$\widehat{p}_1 = \frac{x_1}{n_1}, \qquad \widehat{p}_2 = \frac{x_2}{n_2}$$

where x_1 and x_2 are the number of successful outcomes in the two samples.

Criteria for normal approximation
A normal approximation can be used if $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$ and $n_2(1-\hat{p}_2)$ are all greater than 15

6.10 Confidence interval for the ratio of two populations

Confidence interval for the ratio of two populations
If the criteria for using the normal approximation are met, the lower bound for the difference p_1 and p_2 can be calculated with:

$$\widehat{p_1} - \widehat{p_2} - z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

and the upper bound with:

$$\widehat{p_1} - \widehat{p_2} + z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

where $\widehat{p_1} = \frac{x_1}{n_1}$, $\widehat{p_2} = \frac{x_2}{n_2}$ and $z_{1-\alpha/2}$ is in the standardized normal distribuiton table

6.11 The null hypothesis

- The hypothesis test in this section tests the null hypothesis that the ratios in the two populations are equal.
- The null hypothesis is written $H_0 : p_1 = p_2$.
- If the hypothesis test is two sided we draw the conclusion that the ratios are different if we reject the null hypothesis.
- It is one sided we can only conclude that one ratio is greater than the other or vice verse, depending on the case.

6.12 Hypothesis test for the ratio of two populations

Hypothesis test for the ratio of two populations
If the criteria for using the normal approximation are met, the following hypothesis test can be used:

The null hypothesis is:

$$H_0 : p_1 = p_2$$

The test statistic is:

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\widehat{P}(1-\widehat{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \text{ where } \widehat{P} = \frac{X_1+X_2}{n_1+n_2}$$

If the null hypothesis is true the test statistic follows the standardized normal distribution, or $Z \sim N(0,1)$.

Let us look again at the example where governmental support was measured. Now we get the additional information that in fact two samples were taken, 4375 females and 4375 males.

The result was that 4530 said yes in total and 4220 said no. Out of the 4530 that said yes, 2337 were females. Find a 95% confidence interval for the difference in ratio between females and mails that support the government and test the hypothesis that there is a difference in the ratio between the females and the mails that support the government. Use $\alpha = 0.05$.

The conditions for using a normal approximation are fulfilled since $n_1\widehat{p_1}$, $n_1(1-\widehat{p_1})$, $n_2\widehat{p_2}$ and $n_2(1-\widehat{p_2})$ are all larger than 15.

We need to find $\widehat{p_1}$ and $\widehat{p_2}$. We have that: $n_1 = n_2 = 4375$. We also know that the number of females supporting the government is 2337 and the number of males then: 4530-2337 = 2193, so $x_1 = 2337$ og $x_2 = 2193$.

$$\widehat{p_1} = \frac{x_1}{n_1} = \frac{2337}{4375} = 0.5342 \text{ og } \widehat{p_2} = \frac{x_2}{n_2} = \frac{2193}{4375} = 0.5013.$$

The lover limit is:

$$\widehat{p_1} - \widehat{p_2} - z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}} =$$

$$0.5342 - 0.5013 - 1.96\sqrt{\frac{0.5342(1-0.5342)}{4375} + \frac{0.5013(1-0.5013)}{4375}} = 0.0119$$

and the upper limit:

$$\widehat{p_1} - \widehat{p_2} + z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}} =$$

$$0.5342 - 0.5013 + 1.96\sqrt{\frac{0.5342(1-0.5342)}{4375} + \frac{0.5013(1-0.5013)}{4375}} = 0.0537.$$

The confidence interval is:

$$0.0119 < p_1 - p_2 < 0.0537.$$

1. We want to make a test regarding the difference between two ratios. We use a normal approximation.

2. $\alpha = 0.05$.

3. The hypotheses are:

$$H_0 \quad : \quad p_1 = p_2$$
$$H_1 \quad : \quad p_1 \neq p_2$$

4. We know that $\widehat{p}_1 = 0.5342$ and $\widehat{p}_2 = 0.5013$. We need to find $\widehat{p}$:

$$\widehat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{4530}{8750} = 0.5177.$$

The test statistic is:

$$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.5342 - 0.5013}{\sqrt{0.5177(1-0.5177)\left(\frac{1}{4375} + \frac{1}{4375}\right)}} = 3.08.$$

5. $z_{1-\alpha/2}$ = $z_{0.975}$ = 1.96. We reject the null-hypothesis if $z > 1.96$ OR IF $z < -1.96$. We see that $z > 1.96$.

6. We reject the null hypothesis and conclude that there is a difference in the ratio between males and females.

6.13 The alternative hypothesis

The alternative hypothesis
The alternative hypothesis along with their rejection areas are shown below:

Alternative hypothesis	Reject H_0 if:
$H_1 : p_1 < p_2$	$Z < -z_{1-\alpha}$
$H_1 : p_1 > p_2$	$Z > z_{1-\alpha}$
$H_1 : p_1 \neq p_2$	$Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$

6.14 Chi squared test

- The hypothesis in last section can be generalized such that it compares the ratio of more than two populations.
- Then one cannot use methods based on normal proximation, but so called chi-squared tests are used (χ^2-test).
- The method can also be used when comparing the ratios of two populations, but only if the alternative hypothesis is two-sided.
- Then the Chi-squared test statistic and the Z-statistic be the same.

6.15 The null hypothesis

- The hypothesis test in this section tests whether the ratios of c populations are all equal.
- It is written $H_0 : p_1 = p_2 = \dots = p_c$.
- If it is rejected we can conclude that the ratio are not all equal.
- That does not mean that they are all different!
- The hypothesis test does not say which of the ratios differ from the other.
- More evolved methods are used to do so, which are not taught in this lecture.

6.16 Tables for chi-squared próf

Töflur fyrir chi-squared próf
Þegar framkvæma á chi-squared test er gott að búa til þrjár töflur:

- Table 1: Contains the observed frequency in the investigation, denoted with o.
- Table 2: Containes the expected frequency in the investigation, denoted with e. The values are calculated by multiplying the sums for the corresponding column and row and divide by the total number of measurements. All values in this table need to be greater than 5 for the test to be valid.
- Table 3: Contains the tribute to the test statistic, calculated with $\frac{(o-e)^2}{e}$. Finally all the values in Table 3 are added together to calculate the value of the test statistic (see next slide).

6.17 Chi-squared test for ratios

Chi-squared test for ratios
The hypothesis are:

$$H_0 : p_1 = p_2 = ... = p_c$$

H_1 : the ratios are not all equal

The test statistic is:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where r is the number of rows, c is the number of columns, o is the observed frequency and e is the expected frequency.

If the null hypothesis is true, the test statistic follows the χ^2-distribution with $(r - 1) \cdot (c - 1)$ degrees of freedom. The null hypothesis is rejected if $\chi^2 > \chi^2_{1-\alpha,((r-1)\cdot(c-1))}$.

The following data are the results from an experiment where employees in three governmental departments were asked if the were in favour of their pension plan.

	Department 1	Department 2	Department 3
In favour	66	85	108
Not in favour	34	65	42

We need to start make the three tables, that observed frequencies, the expected frequencies and the contribution to the test statistic.

Table 1 - *o*	Department 1	Department 2	Department 3	Total
In favour	66	85	108	259
Not in favour	34	65	42	141
Total	100	150	150	400

We get the values in Table 2 by multiplying the totals in the corresponding line and column from Table 1 and divide with the total number.

Table 2 - *e*	Dep. 1	Dep. 2	Dep. 3
In favour	$\frac{100 \cdot 259}{400} = 64.75$	$\frac{150 \cdot 259}{400} = 97.13$	$\frac{150 \cdot 259}{400} = 97.13$
Not in favour	$\frac{100 \cdot 141}{400} = 35.25$	$\frac{150 \cdot 141}{400} = 52.88$	$\frac{150 \cdot 141}{400} = 52.88$

The values in Table three we get by: $\frac{(o-e)^2}{e}$

Table 3	Gov. 1	Gov. 2	Gov. 3
In favour	$\frac{(66-64.75)^2}{64.75} = 0.02$	$\frac{(85-97.13)^2}{97.13} = 1.51$	$\frac{(108-97.13)^2}{97.13} = 1.22$
Not in favour	$\frac{(34-35.25)^2}{35.25} = 0.04$	$\frac{(65-52.88)^2}{52.88} = 2.78$	$\frac{(42-52.88)^2}{52.88} = 2.24$