# Inference on the means of two populations

## (STATS201.stats 201 30: Statistical inference)

Anna Helga Jónsdóttir
Sigrún Helga Lund

December 14, 2012

# 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 $\mu_1$ and $\mu_2$ 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.

# 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.

# Conducting hypothesis tests

## Conducting hypothesis tests

1 Decide which hypothesis test is appropriate for our measurements.

2 Decide the $\alpha$-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.

# 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 $\delta$.
- 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 $\delta$.
- If it is one sided we can only conclude that the difference is greater in one case or less in the other case then $\delta$ depending on the case.

# 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 $\mu_1$, $\sigma_1^2$ and $n_1$ but the other with $\mu_2$, $\sigma_2^2$ and $n_2$.

# Decision tree - independent measurements

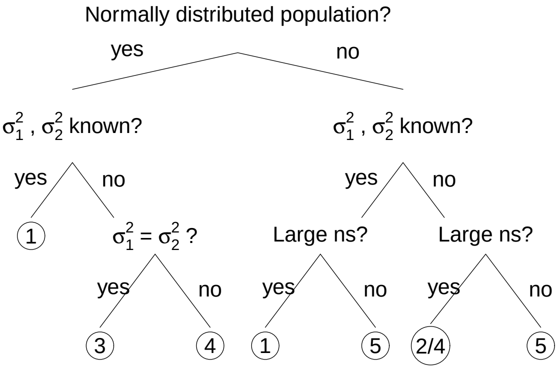
Normally distributed population?
yes
no
$\sigma_1^2$ , $\sigma_2^2$ known?
$\sigma_1^2$ , $\sigma_2^2$ known?
yes
no
yes
no
1
$\sigma_1^2 = \sigma_2^2$ ?
Large ns?
Large ns?
yes
no
yes
no
yes
no
3
4
1
5
2/4
5

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

# 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**, $\sigma_1^2, \sigma_2^2$, 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 universal rule though.
4. When the variances are unknown, can we yet assume that they are equal for the two populations, that is if $\sigma_1^2 = \sigma_2^2$?
   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.

# $\mu_1 - \mu_2$ - case 1

Case one applies when:

- it can be assumed that the populations are normally distributed and the variances of the populations, ($\sigma_1^2$ and $\sigma_2^2$) are known.
- when $n_1$ and $n_2$ are large and $\sigma_1^2$ and $\sigma_2^2$ are known, although the populations are not normally distributed.

# 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:

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

# 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 $\sigma_1^2$, $\sigma_2^2$ are the population variances. $z_{1-\alpha/2}$ is found in the standardized normal distribution table.

# 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: |
|---|---|
| $\mathrm{H}_1 : \mu_1 - \mu_2 < \delta$ | $Z < -z_{1-\alpha}$ |
| $\mathrm{H}_1 : \mu_1 - \mu_2 > \delta$ | $Z > z_{1-\alpha}$ |
| $\mathrm{H}_1 : \mu_1 - \mu_2 \neq \delta$ | $Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$ |

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

## $\mu_1 - \mu_2$ - case 2

Case 2 applies when:

- we do not know the population variances ($\sigma_1^2$ and $\sigma_2^2$) 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

# 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:

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

# 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.

# 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: |
|---|---|
| $\mathrm{H}_1 : \mu_1 - \mu_2 < \delta$ | $Z < -z_{1-\alpha}$ |
| $\mathrm{H}_1 : \mu_1 - \mu_2 > \delta$ | $Z > z_{1-\alpha}$ |

# $\mu_1 - \mu_2$ - case 3

Case 3 applies when:

- One can assume that the populations are normally distributed, the variances ($\sigma_1^2$ and $\sigma_2^2$) 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.

# $\mu_1 - \mu_2$ - 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.

# 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.

# 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: |
|---|---|
| $\mathrm{H}_1 : \mu_1 - \mu_2 < \delta$ | $T < -t_{1-\alpha,(n_1+n_2-2)}$ |
| $\mathrm{H}_1 : \mu_1 - \mu_2 > \delta$ | $T > t_{1-\alpha,(n_1+n_2-2)}$ |
| $\mathrm{H}_1 : \mu_1 - \mu_2 \neq \delta$ | $T < -t$ or $T > t$ |

## $\mu_1 - \mu_2$ - case 4

Case 4 applies when:

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

# $\mu_1 - \mu_2$ - 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 $\nu$ 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.

# 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:

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

# 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. $\nu$ is the number of degrees of freedom.

# 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 $\nu$ frígráðum or $T \sim t(\nu)$ where $\nu$ is calculated as shown in slide **??**

| Alternative hypothesis | Reject $H_0$ if: |
|---|---|
| $\mathrm{H}_1 : \mu_1 - \mu_2 < \delta$ | $T < -t_{1-\alpha,(\nu)}$ |
| $\mathrm{H}_1 : \mu_1 - \mu_2 > \delta$ | $T > t_{1-\alpha,(\nu)}$ |

## $\mu_1 - \mu_2$ - 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.

# Paired measurements

- Assume that we have $n$ pair of measurements $(X_i, Y_i)$, $i = 1, 2, 3...n$.
- 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 $\mu_D$.
- The hypothesis tests make inference on $\mu_D$.

# 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.

# 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, $\mu_D$, 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.

# 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, $\mu_D$, 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.

# 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 $\mu_D$.
- 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 $\mu_D$ 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).

# 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: |
|---|---|
| $\mathrm{H}_1 : \mu_D < \mu_{D,0}$ | $Z < -z_{1-\alpha}$ |
| $\mathrm{H}_1 : \mu_D > \mu_{D,0}$ | $Z > z_{1-\alpha}$ |
| $\mathrm{H}_1 : \mu_D \neq \mu_{D,0}$ | $Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$ |

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

# 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 $T \sim t_{(n-1)}$.

# 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: |
|---|---|
| $\mathrm{H}_1 : \mu_D < \mu_{D,0}$ | $T < -t_{1-\alpha,(n-1)}$ |
| $\mathrm{H}_1 : \mu_D > \mu_{D,0}$ | $T > t_{1-\alpha,(n-1)}$ |
| $\mathrm{H}_1 : \mu_D \neq \mu_{D,0}$ | $T < -t_{1-\alpha/2,(n-1)}$ or $T > t_{1-\alpha/2,(n-1)}$ |

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