# Statistics

## Section: Tests and Confidence Intervals

## Exercises

1. Hypothesis and Null Hypothesis ........ 2
  1.1 How to derive the null-hypothesis ........ 2
  1.2 null-hypothesis and p-value ........ 2
  1.3 Conclusions ........ 3
2 Statistical significance and practical significance (relevance) ........ 4
3 Confidence Interval ........ 5
  3.1 Definition ........ 5
  3.2 Sample size ........ 5

# 1. Hypothesis and Null Hypothesis

### 1.1 How to derive the null-hypothesis

- It is __ to first look at the data and then frame $H_0$ to fit what the data show.

| | |
|---|---|
| 1 | mathematically correct |
| 2 | a normal procedure |
| 3 | quite OK |
| 4 | natural |
| 5 | cheating |

### 1.2 null-hypothesis and p-value

- Mark correct and false statements (Yes/No)

| | |
|---|---|
| | A null hypothesis is a statement about a population. |
| | A null hypothesis is a statement about the sample. |
| | $H_0$: There is no difference in the true means. |
| | $H_0$: There is no difference in the sample means. |
| | $H_0$: $\mu = 0$ |
| | $H_0$: $\bar{x} = 0$ |
| | The p-value is the probability of getting an outcome as extreme as or more extreme than the actually observed outcome (sample) under the null hypothesis. |
| | Usually the null hypothesis is a statement of "no effect", "no difference" or "=0" and we are eager to find evidence against it. |
| | p-value = Prob(Sample \| $H_0$) |
| | p-value = Prob($H_0$ \| Sample) |
| | Beware of searching for significance |
| | Legitimately you cannot test a hypothesis on those data that first suggested that hypothesis. |
| | When large samples are available, even tiny deviations from the null hypothesis will be significant. |

### 1.3 Conclusions

There are two populations, e.g. two groups of farmers (traditional and organic farming technique)

$H_0$: Mean $Income_1$ = Mean $Income_2$

You draw a random sample of size 30 from each population.

Test-result: p = 0.96

Mark correct and false conclusions (Yes/No)

| | |
|---|---|
| The mean income of the two farmer samples is equal. | |
| The mean income in the two farmer populations is equal. | |
| Traditional farmers have the same mean income as organic farmers. | |
| With probability 0.96, traditional farmers have the same mean income as organic farmers. | |
| With probability 0.04, traditional and organic farmers have the same mean income. | |
| The sample is not compatible (p=0.04) with the null hypothesis: the mean income in the two farmer populations is equal. | |
| We could not find a significant (at level 0.05) difference in mean income between the two farmer populations. | |
| Data did not show a significant (p=0.96) difference in mean income between traditional and organic farmers. | |
| Data are compatible (p=0.96) with the null hypothesis that traditional and organic farmers have the same mean income. | |

Test-result: p = 0.04

Mark correct and false conclusions (Yes/No)

| | |
|---|---|
| Mean income in the two farmer samples differs significantly (p=0.04). | |
| Mean income in the two farmer populations differs significantly (p=0.04). | |
| Traditional farmers have the same mean income as organic farmers (p=0.04). | |
| With probability 0.04, traditional and organic farmers have the same mean income. | |
| The null hypothesis, that the mean income of traditional and organic farmers are equal, is rejected at significance level α=0.05. | |
| The sample is not compatible (p=0.04) with the null hypothesis that the mean income of traditional and organic farmers is equal. | |
| Data are not compatible (p=0.04) with equal mean income gained by organic and traditional farming. | |
| Organic and traditional farmers have different mean income (p=0.04). | |

## 2 Statistical significance and practical significance (relevance)

- A study comparing wages (each 5 male and female) reported
  mean difference (male-female) = -1.5, p = 0.10, CI95% = [-2; 1]

- A study comparing wages (each 5 male and female) reported
  mean difference (male-female) = 260, p = 0.10 CI95% = [-20; 500]

- A study comparing wages (each 500 male and female) reported
  mean difference (male-female) = -1.5, p = 0.0001 CI95% = [-2; -1]

- A study comparing wages (each 500 male and female) reported
  mean difference (male-female) = 475, p = 0.0001 CI95% = [450; 500]

- A randomized trial of interventions for reducing transmission of HIV-1 reported[1]:

  $H_0$ : incident rate[2]$_{intervention}$ / incident rate$_{control}$ = 1 (i.e. incident rate$_{intervention}$ = incident rate$_{control}$)

  The test was not significant at 0.05

  95% confidence interval was reported as [0.63 ; 1.58][3].

  Conclusion: As the test was not significant at 0.05 we conclude that intervention has no influence on the infection rate of HIV-1.

  What do you think?

[1] Example from Moore (2006), p. 425
[2] Incident rate = rate of new infections
[3] Intervention might decrease the infection by 37% or increase the infection by 58%.

## 3 Confidence Interval

### 3.1 Definition

- Mark correct and false statements (Yes/No)

| | |
|---|---|
| | A confidence interval always covers the true parameter. |
| | A confidence interval covers the true parameter with a given probability. |
| | A confidence interval covers the statistic with a given probability. |
| | In the long run, 95% of your confidence intervals will contain the true parameter |

A confidence interval (CI) is determined by:
mean, standard deviation (Std), sample size (n) and the confidence level (C)

- Circle correct statements

All else being equal

n=100 CI is smaller / wider than n=10 CI

Std=10 CI is smaller / wider than Std = 100 CI

a 80% CI is smaller / wider than a 95% CI

a 95% CI is smaller / wider than a 99% CI

### 3.2 Sample size

- A pilot study reveals that the real value μ is around 100 and σ = 50

Determine the sample size so that a confidence interval derived from a sample of this size covers the real μ with C = 80%; 90%; 95%; 99.7% confidence level and has a width of less than 10% of the true parameter.

$$CI : \bar{x} \pm z \cdot SEM$$

| C | rule of thumb z | exact z |
|---|---|---|
| 80% | 1.3 | 1.28 |
| 90% | 1.6 | 1.64 |
| 95% | 2 | 1.96 |
| 99% | 2.6 | 2.56 |
| 99.7 % | 3 | 2.97 |

$$\Pr ob(\bar{x} - z * SEM \leq \mu \leq \bar{x} + z * SEM) \geq C \Leftrightarrow$$

$$\Pr ob(|\mu - \bar{x}| \leq z * \frac{s}{\sqrt{n}}) \geq C$$