Week 1: Statistical Inference

a.
The width of the interval is proportional to $1/\sqrt{n}$, and thus increasing n will make the interval shorter.
Heuristic explanation: larger sample = more information = larger precision = shorter interval.
b.
Decreasing $100(1 -\alpha)$ (increasing alpha) will decrease the corresponding quantile of the t-distribution (we move clo
ser to the center from the tail). As the width of the interval is proportional to the quantile, decreasing the confidence le
vel will make the interval shorter.
Heuristic explanation: lower confidence level = we are satisfied with smaller probability to capture the true value = shor
ter interval.
c.
Increasing the population variance will most likely also increase the sample variance s^2, and consequently the sample s
tandard deviation. As the width of the interval is proportional to s, increasing the variance will make the interval wide
r.
Heuristic explanation: larger variance = data is less accurate = capturing the true value is more difficult = need larger
interval
d. Increasing the expected value will only affect the location of the interval in the x-axis. The width will stay the sam
e

d.
Mean = the average level around which the data fluctuates
SD & Var = the average size of these fluctuations
Min = the lowest point of the curve
Max = the lowest point of the curve
Median = the average level around which the data fluctuates (robust)
MAD = the average size of these fluctuations (robust)
(Mode = difficult to see...)
(Skewness = difficult to see...)
(Kurtosis = difficult to see...)

The data is paired (and the pairs are not independent), making paired t-test a good choice.
Note that using the two-sample t-test is not justified as it assumes the independence of the two samples

F test to compare two variances
Welch Two Sample t-test

In each case we want to use a test which makes the strictest assumptions (such that they are still satisfied). This gives
us maximal power (lowest Type 2 error rate), as we "use more information" about the data. See the lecture examples of week
3.
The assumptions the three tests make besides iid data are:
t-test: normality <- strictest
signed rank test: symmetric continuous distribution <- less strict
sign test: continuous distribution <- even less strict
a.
Exponential distributions are not symmetric -> sign test.
b.
t-test
c.
Laplace distributions are not normal but they are symmetric -> signed rank test.
d.
Poisson distribution is neither continuous nor symmetric so, being strict, none of the tests apply. However, sign test is
regularly applied to discrete data as well (e.g. using the conventions of slide 3.7).

With so few observations it is difficult to say whether the data comes from a normal, or even a symmetric, distribution. S
ign test seems like the safest choice.

p-value = 1
The highest possible p-value -> no evidence against H0 -> the drug 1 is no better than placebo.

Most points are below the y=x -line, meaning that the salary of the man in a pair is more often larger than that of the wo
man

b.
The data is paired (and the pairs are not independent), making paired sign test and paired signed rank test appropriate ch
oices.
Note that using a two-sample rank test is not justified as it assumes the independence of the two samples

d.
Paired sign test does not reject the null but the paired signed rank test does.
e.
Both tests assume that the differences d1, d2, ... ,d8 form an iid. sample from some particular continuous distribution. P
aired signed rank test furthermore assumes that this distribution is symmetric. It is difficult to say whether the underlyin
g distribution is symmetric based on a so small sample so the paired sign test might be a better choice -> no difference in
salaries.

a.
The test assumes that the female and male samples are mutually independent iid samples from the continuous distributions F
x and Fy, respectively. Moreover, the distributions Fx and Fy are assumed to be equal up to location shift ("same-shaped hil
ls").
The null hypothesis is that the medians of Fx and Fy (and consequently the distributions itself) are equal (and the altern
ative hypothesis is the opposite of that).
b.
As the samples where chosen randomly, it is plausible that the samples are independent and iid. Also, googling "female vs.
male height distribution" shows that the male distribution of heights is slightly wider than for females. We assume that thi
s difference in scales is small enough that the test can still be used.

##
Wilcoxon rank sum test
##
data: female and male
W = 25, p-value = 0.0825
alternative hypothesis: true location shift is not equal to 0

p-value = 0.0825 -> not enough evidence to reject H0 on significance level 5% -> no difference in medians. As we "know" that there should be a difference, then either the sample size was too small, the result was caused by randomness or the assumptions weren't justified.

i. The half-width of the conservative 95% interval is 1.96*0.5/sqrt(n). This equals 0.01*a if
1.96*0.5/sqrt(n) = 0.01*a <=> 1.96*0.5/(0.01*a) = sqrt(n) <=> n = 9604/a^2.
That is, the required sample sizes are n = 9604, 2401, 1068.
#
# ii. The half-width of the standard 95% interval for \hat p = 0.05 is 1.96*sqrt(0.05*0.95)/sqrt(n). As in part i, we obtain
n = 1824.76/a^2 and the true required sample sizes are 81% (= 1 - 1824.76/9604) smaller than those approximated in part i.

The Z-value of the test is proportional to the square root of the sample size n. Thus increasing the sample
size increases the Z-value and consequently pushes it towards the tail of the distribution, decreasing the
p-value. Thus the p-value for n = 200 is smaller
#
An intuitive reasoning for the result is that the difference between 0.06 and 0.09 is "proportionally"
larger for n = 200 than for n = 100 (as larger n implies increased accuracy) and as such also more
deviating.

b.
Assumptions:
The sample is iid from Bernoulli with parameter value p where p is the proportion of people living in
Finland with last name starting with a vowel.
(that is, everyone has their last name beginning with a vowel with equal proability and independently
of each other)

d.
Substitute conclusions here. The conclusions can most likely not be used to draw inference on the

proportion of people in the whole Finland as the session participants make a poor *random* sample of this population.
At best, the participants could be considered a random sample of all Aalto students in
particular programmes.

#
2-sample test for equality of proportions without continuity
correction
##
data: c(x_1, x_2) out of c(n_1, n_2)
X-squared = 9.5406, df = 1, p-value = 0.001005
alternative hypothesis: greater
95 percent confidence interval:
0.0335168 1.0000000
sample estimates:
prop 1 prop 2
0.5576324 0.4861265
c.
The one-sided p-value ~= 0.001 < 0.05 -> we reject the null hypothesis in favor of the alternative.
That is, the support has decreased.
d.
(The assumptions are stated above in the answer to b.) To ensure that the samples are independent and iid
and representative of the nationwide support level, the pollmaker should draw the samples perfectly randomly
from amongst all eligible voters

i. Straight line: sample quantiles depend approximately linearly on normal quantiles
ii. U-shaped curve: the high values are too high (long right tail) and the low values
are too high (short left tail).
iii. Inverse U-shaped curve: for the opposite reasons as in ii.
iv. S-shape near the middle of the plot: the points left of median are too small and the points right of median are too la
rge (too little mass near median)
v. S-shape in the tails: the low values are too high (the left tail is too short) and the high values
are too low (the right tail is too short).
vi. Something resembling a cubic function: the low values are too low (the left tail is too long) and

```
# the high values are too high (the right tail is too long).
# Sample plots:
par(mfrow = c(3, 2))
x <- rnorm(1000)
qqnorm(x)
qqline(x)
x <- rexp(1000, 1)
qqnorm(x)
qqline(x)
x <- -1*rexp(1000, 1)
qqnorm(x)
qqline(x)
b <- rbinom(1000, 1, 1/2) == 1
x <- c(rnorm(1000, 3)[b], rnorm(1000, -3)[!b])
qqnorm(x)
qqline(x)
x <- runif(1000)
qqnorm(x)
qqline(x)
x <- rt(1000, 3)
qqnorm(x)
qqline(x)
```

b. Recall the differences between the interpretations of the χ^2 homogeneity test and χ^2 test for independence. C
ome up with a practical situation where the collected data can be expressed as a 2-by-2 table and a related research questio
n for which the correct interpretation is through
i. the χ^2 homogeneity test,
ii. the χ^2 test for independence.

```
# The key difference between the two tests is in how the data is sampled, i.e., are the margins fixed or not.
# E.g. assume we're interested in studying whether sex (female/male) has an effect on the voting preference
# (democrat/republican) in the US and for this we interview n people in the street.
# These data can be collected into a two-by-two table such that the row variable is sex and the column
# variable is voting preference.
# i. If we choose beforehands that we will interview n1 females and n2 males, then studying the independence of the two vari
```

ables will be questionable (since sex is not fully random anymore with its marginal frequencies fixed). The correct interpre
tation is through the homogeneity test which compares two populations, in this case female and male, in their voting prefere
nces.
ii. If we do not choose beforehands the marginal numbers of females and males, sex is a random variable and we can measure
its independence with the voting behavior. The correct interpretation is now through the test for independence.

d.
Both tests in c. reject their null hypotheses of normality. Based on all the previous evidence, the data can not be deemed
normal enough to rely on normality assumptions in any further analyses.
(Note that it is a different matter whether the next analysis steps involve methods that allow the normality assumption to
be "covered" by large enough sample size (by the central limit theorem).)
e.
See the help file of the dataset: The sample is not iid. as, to obtain the 48 measurements, first 12 "core samples" were o
btained (randomly?) and then from each of these 4 observations were taken to yield the final 48 observations. Thus, the sets
of 4 observations come from a same core sample and as such are not independent, even if the different core samples were.

Mean

Different means

Let $x_1, x_2, \ldots, x_n$ be independent and identically distributed (i.i.d.) observations of a random variable x.

- The sample **mean**,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

  The sample mean estimates the expected value $\mu = \mathrm{E}(x)$ of the random variable x.

- The α-**trimmed mean** is the mean of the sample after discarding the proportion α of both smallest and largest observations,
- **Weighted means** give variable weights for different observations,

$$\sum_{i=1}^{n} w_i x_i,$$

  such that $\sum_{i=1}^{n} w_i = 1$.

Median

Median

Let $y_1 < y_2 < \ldots < y_n$ be ordered values of the data.

- The sample **median** is the middle value of the ordered values.
- If the number of observations is even, the sample median is the average of the two middle observations.
- The sample median estimates the population median m_x, the value with the following property

$$\mathrm{P}(x < m_x) \leq \frac{1}{2} \quad \text{and} \quad \mathrm{P}(x \leq m_x) \geq \frac{1}{2}.$$

Quantiles

Quantiles

- The sample β-**quantile**, $0 < \beta < 1$, is the data point y_k, where $k = \lceil \beta n \rceil$ and n is number of observations.
- The sample β-quantile estimates the population β-quantile β_x, defined as
$$\mathrm{P}(x < \beta_x) \leq \beta \quad \text{and} \, \mathrm{P}(x \leq \beta_x) \geq \beta.$$
- 0.25- and 0.75-quantiles are called first and third **quartiles**, and
- the **mid-hinge** is their average

$$\frac{y_{\lceil 0.25n \rceil} + y_{\lceil 0.75n \rceil}}{2}$$

Mean absolute deviation

Median absolute deviation

The **median absolute deviation**, MAD, is the median of the sample $|x_1 - m_x|, |x_2 - m_x|, \ldots, |x_n - m_x|$.

MAD is often multiplied with by the constant 1.4826 to make it a consistent estimator of the standard deviation in a normal model.

Variance

Variance

- The sample **variance**

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

  The sample variance estimates the population variance $\sigma^2 = \mathrm{E}[(x-\mu)^2]$.
- The sample **standard deviation**,

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

  is often preferred over variance as it is measured in the same units as the data.

Correlation

Correlation

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y).

- The **sample covariance**

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

  estimates the population covariance $\sigma_{xy} = \mathrm{E}[(x - \mathrm{E}[x])(y - \mathrm{E}[y])]$.
- The **sample correlation**

$$\hat{\rho}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  estimates the population correlation $\rho(x, y) = \sigma_{xy}/(\sigma_x \sigma_y)$.

Correlation measures the linear dependence between two random variables. The coefficient is always in the interval $[-1, 1]$

Range

Range

- The **sample range** is the interval $[x_{min}, x_{max}]$ and its length is

$$x_{max} - x_{mix}.$$

- The **interquartile range**, IQR, is the distance between the first and third quartile,

$$y_{\lceil 0.75n \rceil} - y_{\lceil 0.25n \rceil}$$

  IQR is often multiplied with by the constant 0.7413 to make it a consistent estimator of the standard deviation in a normal model.

Skewness and Kurtosis

Measures of skewness and kurtosis

Thus far, roughly:

- First moment = measures of location
- Second moment = measures of spread

If we continue onward, the next two moments give us measures of *skewness* and *kurtosis*

Skewness describes the deviation of the data from symmetry.

Kurtosis describes the heaviness of the tails of the data.

The following slides list various measures of skewness and kurtosis.

Skewness

The sample **skewness** is

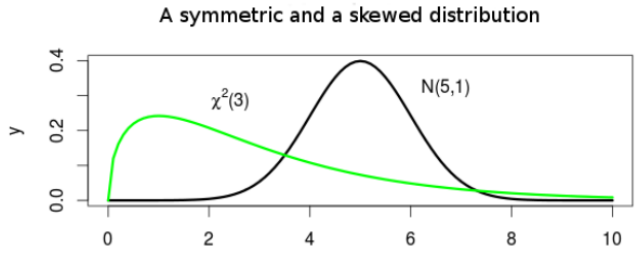
$$\hat{\gamma} = \frac{m_3}{s^3},$$

where

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

Sample skewness coefficient estimates the population skewness,

$$\gamma = \mathrm{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right].$$



Kurtosis

The sample **kurtosis coefficient** is

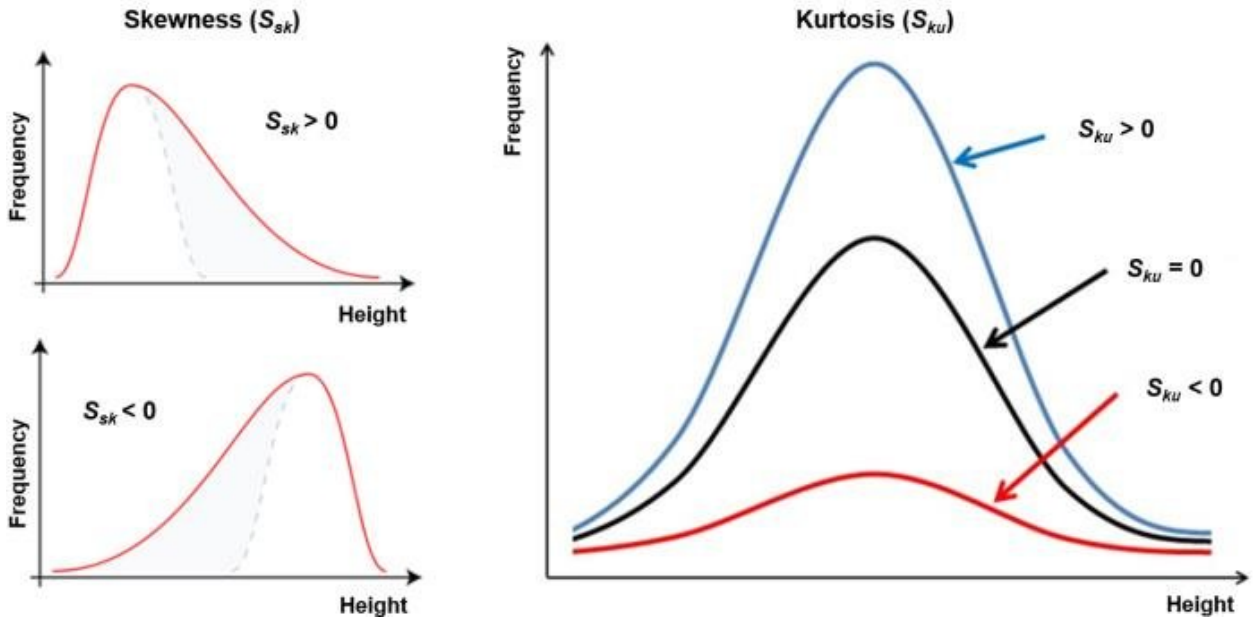
$$\hat{\kappa} = \frac{m_4}{s^4} - 3,$$

where

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

The sample kurtosis coefficient estimates the population kurtosis

$$\kappa = \mathrm{E}\left[\left(\frac{x-\mu}{\sigma}\right)^4\right] - 3.$$



Four moments of distribution: mean, variance, skewness, kurtosis

Moment	Uncentered	Centered
1st	$E(X) = \mu$	
2nd	$E(X^2)$	$E((X-\mu)^2)$
3rd	$E(X^3)$	$E((X-\mu)^3)$
4th	$E(X^4)$	$E((X-\mu)^4)$

Mean(X)	=	$E(X)$
Var(X)	=	$E((X-\mu)^2) = \sigma^2$
Skewness(X)	=	$E((X-\mu)^3) / \sigma^3$
Kurtosis(X)	=	$E((X-\mu)^4) / \sigma^4$

Week 2: Confidence intervals and hypothesis testing

Confidence Interval

Confidence interval

A confidence interval gives an estimated range of values which is likely to contain the value of an unknown population parameter.

The confidence level of a confidence interval determines the probability that the confidence interval produced (interpreted as a random interval) will contain the true parameter value.

E.g. if 95% confidence intervals for an unknown parameter are computed from 100 independent samples, approximately 95 of the these will contain the true parameter value — but we do not know which!

Note that any particular realized confidence interval either contains the true value or not; the 95% frequency concens the probability *in the sampling process*

Bootstrap

The bootstrap

The standard formulas for confidence intervals either make heavy parametric assumptions or work only for parameters estimable by means (CLT).

The standard non-parametric procedure for estimating confidence intervals is known as the bootstrap.

Bootstrap creates pseudo-samples by drawing n observations *from the data*, *with replacement*, *repeating* the procedure for a large number of times.

If n is large enough, the pseudo-sampling approximates true sampling from the population.

Bootstrap confidence intervals

1. Select n data points randomly with replacement from the original sample $\{x_1, x_2, ..., x_n\}$. Each data point can be selected once, multiple times, or not at all. (Note that the sample size of the new sample is the same as the sample size of the original sample.)
2. Use this new sample to calculate a new estimate for the parameter θ.
3. Repeat the previous steps B times.
4. After the replications, order the B estimates from the smallest to the largest.
5. A $100(1-\alpha)\%$ confidence interval is now obtained by choosing the $\lfloor B \times (\alpha/2) \rfloor$ ordered estimate as the lower endpoint and the $\lfloor B \times (1-\alpha/2) \rfloor$ ordered estimate as the upper endpoint.

Exact confidence interval

Exact confidence intervals, normal distribution

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$ where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown.

A level $100(1-\alpha)\%$ confidence interval for μ is obtained as,

$$\left(\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right),$$

where $t_{n-1,\alpha/2}$ is the $(1-\alpha/2)$-quantile of the Student's t-distribution with $n-1$ degrees of freedom.

For large values of n, the Student's t-distribution with $n-1$ degrees of freedom approaches the standard normal distribution and its corresponding quantile can be substituted in place of $t_{n-1,\alpha/2}$.

A level $100(1-\alpha)\%$ confidence interval for σ^2 is obtained as,

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \right),$$

where $\chi^2_{n-1,\alpha/2}$ is the $(1-\alpha/2)$-quantile of the χ^2_{n-1}-distribution and $\chi^2_{n-1,1-\alpha/2}$ is the $(\alpha/2)$-quantile of the χ^2_{n-1}-distribution.

Hypothesis testing

Statistical hypothesis testing is based on

1. Selecting a statistical model/assumptions and
2. Setting a null hypothesis and often also an alternative hypothesis,
3. Choosing a suitable test statistic, the value of which is calculated from a sample of observations.

The result of statistical hypothesis testing is a p-value, based upon which the conclusions are drawn.

Statistical model

- Statistical model/assumptions casts the problem in a mathematical context and defines the rules of probability governing it.
- Statistical models are usually of the form:

  "Let $x_1, \ldots, x_n$ be an i.i.d. sample from the distribution F with the unknown parameter θ".
- The validity of the model can, and should, be tested separately.

Null hypothesis

- The statement of interest about a model parameter is called the null hypothesis H_0.
- H_0 is assumed to be true unless there is strong evidence that indicates otherwise, in which case it is rejected.
- In simple statistical tests the null hypothesis can often be stated as an *equality*, $H_0 : \theta = \theta_0$, where θ is the parameter being tested and and θ_0 is a fixed value of the parameter.
- The null hypothesis is often conceptually of the form "*is the same*" or "*there is no difference*".

Alternative Hypothesis

Alternative hypothesis

- The null hypothesis is usually accompanied by an alternative hypothesis H_1, which is often the logical opposite of H_0 (though not always).
- If H_0 is rejected, then H_1 is accepted.
- The alternative hypothesis is often conceptually of the form "*is not the same*" or "*there is a difference*".

Examples of alternative hypotheses

- $H_1 : p \neq 0.5$.
- $H_1 : \mu_1 > \mu_2$.

Most tests in these lecture slides are for simplicity formulated using *two-tailed alternative hypotheses*,

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0.$$

Test statistic is a random variable that is calculated from sample data and used in a hypothesis test. You can use test statistics to determine whether to reject the null hypothesis. The test statistic compares your data with what is expected under the null hypothesis.

Test statistic

- Test statistic measures deviation of the observed sample from the null hypothesis H_0.
- A test statistic is a random variable and its value depends on the observations.
- The distribution of the test statistic under the null hypothesis H_0 must be known for assessing the compatibility of the observations and the null hypothesis H_0.

Examples of test statistics

- The proportion of correct guesses out of the total n.
- $\bar{x} - \bar{y}$

p-value

p-value

- The *p*-value of a statistical test is the probability of observing at least as deviating value towards H_1 as the observed value of the test statistic under the null hypothesis H_0.
- Note that what is considered as "deviating" depends on the form of the hypotheses.
- If the *p*-value is *too small* (the observation is too strange to have happened under H_0), we reject H_0 in favor of H_1.
- Note that we can never accept H_0 based on the test, only "continue to believe in it".

significance level

Significance level and critical values

- Significance level α is used to make a cut-off between small and large *p*-values.
  - If $p < \alpha$ we reject H_0.
  - If $p \geq \alpha$ we do not reject H_0.
- Commonly used significance levels are $\alpha = 0.05, 0.1, 0.01, 0.001.$ and it should be set before the study.
- The set of values of the test statistic for which the null hypothesis is rejected (i.e. the values that yield a *p*-value smaller than α) is called the critical region.
- The threshold values delimiting the regions of non-rejection and rejection for the test statistic are called the critical values.

Type I and type II error

Errors in statistical hypothesis testing

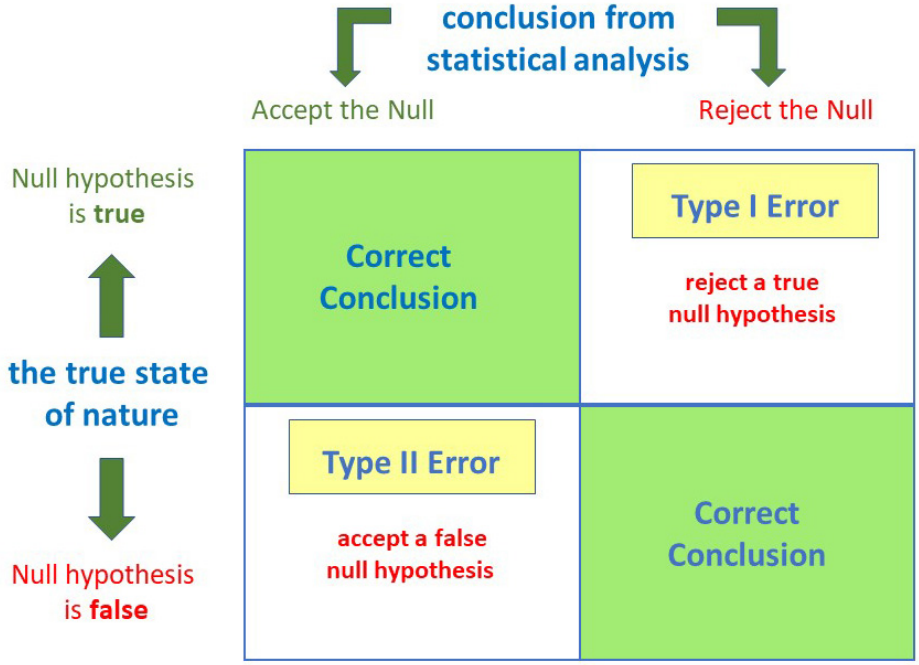
The are two kinds of errors related to the rejection of the null hypothesis H_0.

- Type 1 error: True null hypothesis is rejected.
- Type 2 error: False null hypothesis is not rejected.

The type 1 error rate is the probability of rejecting a true H_0. It is at most α.

The type 2 error rate is the probability of not rejecting a false H_0. Type 2 error rate is more difficult to control as it is usually a function of the possible distributions under H_1.

Power of a test is equal to $1-$ "type 2 error rate". The larger the power, the better the test detects false null hypotheses.



One-sample t-test

One-sample t-test compares the expected value of a distribution to a given constant.

One-sample t-test, assumptions

Let $x_1, x_2, \ldots, x_n$ be and i.i.d. sample from $\mathcal{N}(\mu, \sigma^2)$.

One-sample t-test, hypotheses

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0.$$

One-sample t-test

One-sample t-test, test statistic

- The t-test statistic,
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$
- follows Student's t-distribution with $n - 1$ degrees of freedom under H_0.
- The expected value of t under the null hypothesis H_0 is 0 and if the value of t has **large absolute value**, evidence against the null hypothesis H_0 is found.

If the sample size is large, then the one-sample t-test is not very sensitive to moderate deviations from normality.

Two-sample t-test

Two-sample t-test compares the expected values of two distributions.

Two-sample t-test, assumptions

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from $\mathcal{N}(\mu_x, \sigma_x^2)$ and let $y_1, y_2, \ldots, y_m$ be an i.i.d. sample from $\mathcal{N}(\mu_y, \sigma_y^2)$. Furthermore, let the two samples be independent.

Two-sample t-test, hypotheses

$$H_0 : \mu_x = \mu_y \qquad H_1 : \mu_x \neq \mu_y.$$

Two-sample t-test

Two-sample t-test, test statistic

- The t-test statistic,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m}},$$

follows approximately the Student's t-distribution with

$$\frac{(s_x^2/n + s_y^2/m)^2}{((s_x^2/n)^2/(n-1)) + ((s_y^2/m)^2/(m-1))}.$$

degrees of freedom under H_0.

- The expected value of t under H_0 is 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis H_0 is found.

If the sample size is large, then the two-sample t-test is not very sensitive to moderate deviations from normality.

pair t-test

Paired t-test

Paired t-test, assumptions

Observations consist of an i.i.d. sample of pairs (x_{i1}, x_{i2}), $i = 1, 2, ..., n$ (the values **within** a pair need not be independent). The differences $d_i = x_{i1} - x_{i2}$ have the normal distribution $\mathcal{N}(\mu_d, \sigma_d^2)$.

Paired t-test, hypotheses

$$H_0 : \mu_d = 0 \qquad H_1 : \mu_d \neq 0.$$

The recipe is now simple: Apply the *one-sample* t-test to the differences d_i, to test whether their expected value is zero (whether there is no systematic difference between the values in a pair).
(Stop for a moment to think why this works.)

Variance test

Variance test

The variance test compares the variance of a distribution to a given constant.

Variance test, assumptions

Let $x_1, x_2, \ldots, x_n$ be and i.i.d. sample from $\mathcal{N}(\mu, \sigma^2)$.

Variance test, hypotheses

The null hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \qquad H_1 : \sigma^2 \neq \sigma_0^2.$$

Variance test

Variance test, test statistic

- The χ^2-test statistic,
$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$
follows χ^2-distribution with $n-1$ degrees of freedom under H_0.
- The expected value of the test statistic under H_0 is $n-1$ and both **large** and **small** values of the test statistic suggest that the null hypothesis H_0 is false.

The variance test is sensitive to deviations from normality and does not work, even for large samples, if the underlying distribution is skewed.

Variance comparison test

The variance comparison test compares the variances of two distributions.

Variance comparison test, assumptions

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from $\mathcal{N}(\mu_x, \sigma_x^2)$ and let $y_1, y_2, \ldots, y_m$ be an i.i.d. sample from $\mathcal{N}(\mu_y, \sigma_y^2)$. Furthermore, let the two samples be independent.

Variance comparison test, hypotheses

$$H_0 : \sigma_x^2 = \sigma_y^2 \qquad H_1 : \sigma_x^2 \neq \sigma_y^2.$$

Variance comparison test

Variance comparison test, test statistic

- The F-test statistic,

$$F = \frac{s_x^2}{s_y^2},$$

  follows the F-distribution with $n-1$ and $m-1$ degrees of freedom under H_0.
- The expected value of the test statistic under H_0 is ≈ 1 and both **large** and **small** values of the test statistic suggest that the null hypothesis H_0 is false.

Also the variance comparison test is sensitive to deviations from normality and does not work, even for large samples, if the underlying distribution is skewed.

Week 3: Non-parametric test

A model is parametric if it is fully defined by a set of parameters $\theta_1, \ldots, \theta_K$ (we assume a certain "family" of distributions, e.g. "some normal" or "some exponential").

A model is non-parametric (roughly) if it does not assume a specific family of distributions.

Stricter assumptions $\leftrightsquigarrow$ *more powerful results* $\leftrightsquigarrow$ *more narrow area of application.*

One-sample sign test

One-sample sign test is applied in similar testing problems as the one-sample t-test. However, it makes much milder distributional assumptions.

One-sample sign test, assumptions

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from a continuous distribution with the median m.

One-sample sign test, hypotheses

$$H_0 : m = m_0 \quad H_1 : m \neq m_0.$$

One-sample sign test, test statistic

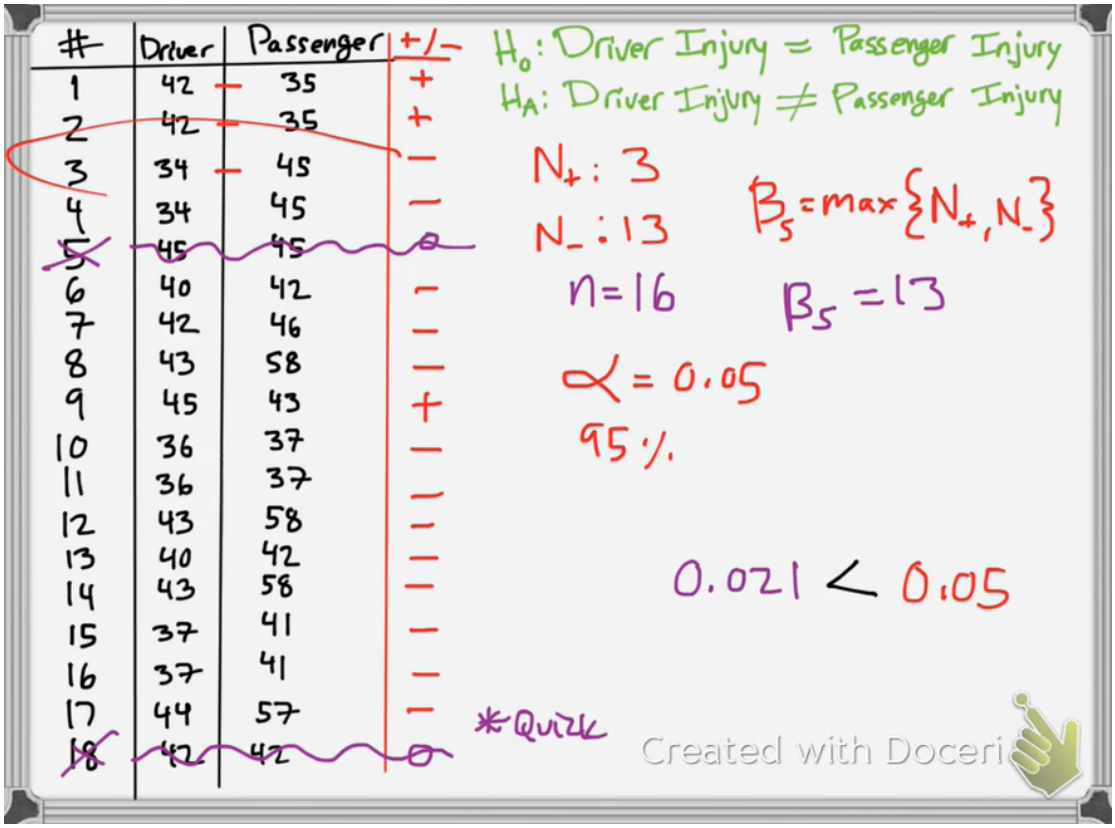
- The test statistic S equals the number of cases with $x_i > m_0$ (alternatively, the number of the cases with $x_i < m_0$) and follows binomial distribution with the parameters n and $1/2$ under H_0.
- Under H_0, the expected value of the test statistic is $\frac{1}{2}n$ (its variance is $\frac{1}{4}n$) and both **large** and **small** values of the test statistic suggest that the null hypothesis H_0 is false.

The distribution of the test statistic S is tabulated and statistical software gives exact p-values of the test.

- If the sample size is large, then the standardized test statistic,

$$Z = \frac{S - n/2}{\sqrt{n/4}},$$

  follows approximately the standard normal distribution under H_0.
- For large n, the p-value of the test can thus be retrieved from a normal table.
- The approximation is usually good enough if $n > 20$. For smaller samples, it is better to rely on the exact distribution of the test statistic S (which is the binomial distribution).



Paired sign test

Paired sign test is a non-parametric version of the paired t-test.

Paired sign test, assumptions

Observations consist of an i.i.d. sample of pairs (x_{i1}, x_{i2}), $i = 1, 2, \ldots, n$ (the values **within** a pair need not be independent). The distribution of the differences $d_i = x_{i2} - x_{i1}$ is continuous (denote the median of this distribution by m_d).

Paired sign test, hypotheses

$$H_0 : m_d = 0 \quad H_1 : m_d \neq 0.$$

The test is conducted by applying the one-sample sign test to the differences d_i.

One-sample signed rank test/Wilcoxon test

More sophisticated versions of the sign tests are given by signed rank tests, which consider not only the signs of the observations but also their relative order (and thus use more information).

One-sample signed rank test, assumptions

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from continuous, symmetric distribution with the median m.

One-sample signed rank test, hypotheses

$$H_0 : m = m_0 \quad H_1 : m \neq m_0.$$

Example:

You have measured the reaction time of a small group of people once in the morning and once in the evening and you want to know if there is a difference.

Case	morning	evening	diff (evening - morning)	Rank from |diff|	Calculation of the rank sums
1	34	45	11	7	
2	33	36	3	2	
3	41	35	-6	5 (-)	$T^+ = 7 + 2 + 3 + 4 + 6 = 22$
4	39	43	4	3	
5	44	42	-2	1 (-)	$T^- = 5 + 1 = 6$
6	37	42	5	4	
7	39	46	7	6	

Null hypothesis

Both rank sums are the same.

* Important

When to use sign tests and signed rank tests

Both families of tests are non-parametric counterparts of the t-test and worthy alternatives when normality cannot be assumed.

Both types of tests are appropriate in similar problems:

- one sample — comparison of the location to a constant,
- paired samples — comparison of the locations.

The assumptions of the two types of tests differ:

- sign test — continuous distribution,
- signed rank test — continuous symmetric distribution.

If normality can be assumed, use t-tests. If symmetry (but no normality) can be assumed, use signed rank tests. Otherwise, use sign tests.

Week 4: Inference for binary data

Binary observations

In many applications the observations are binary.

- Something is true/false.
- Something happened/did not happen.
- Someone belongs/does not belong to a group.

In such a case the observations are most conveniently coded as 0/1.

Recall that if we have a iid sample of binary observations, their distribution is necessarily the *Bernoulli distribution*.

Bernoulli distribution

The random variable x is said to obey the Bernoulli distribution with the probability of success p if,

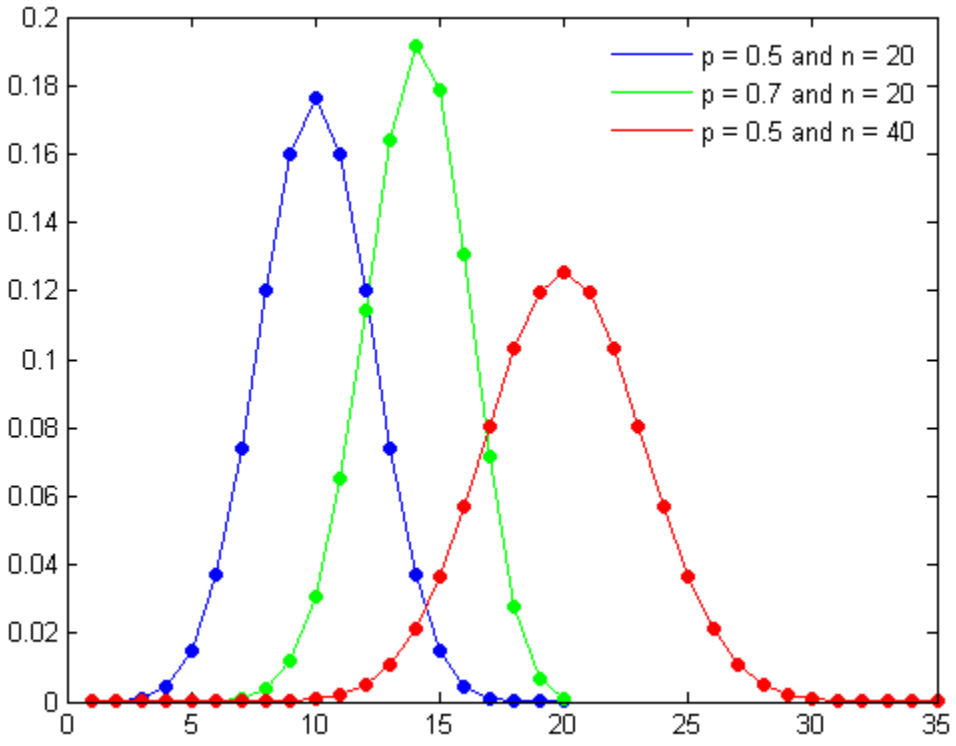
$$\mathbb{P}(x = 1) = p \quad \text{and} \quad \mathbb{P}(x = 0) = 1 - p.$$

The expected value and variance of x are,

$$\mathbb{E}(x) = p$$
$$\mathrm{Var}(x) = p(1 - p).$$

That is, the Bernoulli distribution has only a single parameter to estimate.

The sum of n i.i.d. Bernoulli random variables with the success probability p has the binomial distribution with the parameters n and p.



Approximate confidence interval

Central limit theorem can be used to obtain a confidence interval for the success probability p of a Bernoulli distribution.

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from the Bernoulli distribution with the success probability/expected value p.

For large n, a level $100(1-\alpha)\%$ confidence interval for the success probability p is obtained as

$$\left(\hat{p} - z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right),$$

where $\hat{p}$ is the observed proportion of successes and $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

One-sample proportion test

To test whether the success probability of a Bernoulli distribution equals some pre-specified value, we employ one-sample proportion test.

One-sample proportion test, assumptions

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from a Bernoulli distribution with the success probability p.

One-sample proportion test, hypotheses

$$H_0 : p = p_0 \quad H_1 : p \neq p_0.$$

One-sample proportion test, test statistic

- The test statistic,

$$C = \sum_{i=1}^{n} x_i,$$

  follows the binomial distribution with parameters n and p_0 under H0.
- Under H_0, the test statistic has $\mathrm{E}[C] = np_0$ and $\mathrm{Var}(C) = np_0(1 - p_0)$ and both large and both **large** and **small** values of the test statistic suggest that the null hypothesis H_0 is false.

The distribution of the test statistic C is tabulated and statistical software calculate exact p-values of the test.