Statistics

Hypothesis Testing

Alexandra Posekany
alexandra.posekany@wu.ac.at

Hypothesis Tests

Hypothesis testing is a method of making decisions using data.

Hypothesis tests

hypothesis: assumption about the underlying structure of the population and distribution from which the sample is drawn

- **null hypothesis (H_0)**: basic structure and distribution assumed for the data, if nothing interesting occurs ('standard')
- **alternative hypothesis (H_A)**: interesting case being reasonably different from the standard case ('something happens')

test statistic: value/estimate caluculated from the sample under the distribution assumption of H_0

Watch out with ML based hypothesis testing!

You cannot prove anything, except for maths!

You cannot accept the null hypothesis, only reject it!

Caveat: Null hypothesis and alternative hypothesis are not balanced!

Failure to reject the null hypothesis does not necessarily mean that it is true! Possible wordings in the final conclusion:

- *There is sufficient evidence to warrant rejection of the claim that ...*
- *There is not sufficient evidence to warrant rejection of the claim that ...*

Significance

Significance

statistical significance: a result is significant if it is unlikely that the observed outcome occurred by chance under a null hypothesis given a threshold of significance (**significance level**)
significance level: When assuming the standard structure under H_0, we obtain probabilities for every observed sample estimate (independent of the alternative hypothesis). If the probability of this estimate falls beneath a certain level (=signifcance level), the null hypothesis is rejected.
p-value: the probability of observing the sample estimate (=test statistics) or a more extreme value, if drawing randomly from the distribution defined by the null hypothesis

Statistical Errors

	H_0 TRUE	H_0 FALSE
keep H_0	✓	Type II error
reject H_0	Type I error	✓

- **False positive rate** $=$ significance level (α): probability of committing Type I error, incorrectly rejecting H_0 (positive outcome)
- **False negative rate** (β): probability to commit Type II error, incorrectly accepting H_0 (negative outcome)
- **Power** $(1-\beta)$: probability of correctly rejecting H_0

Regions

Testing regions

- **Region of acceptance**: set of value where the null hypothesis remains valid (cannot be rejected)
- **Region of rejection**: set of values where the null hypothesis is rejected

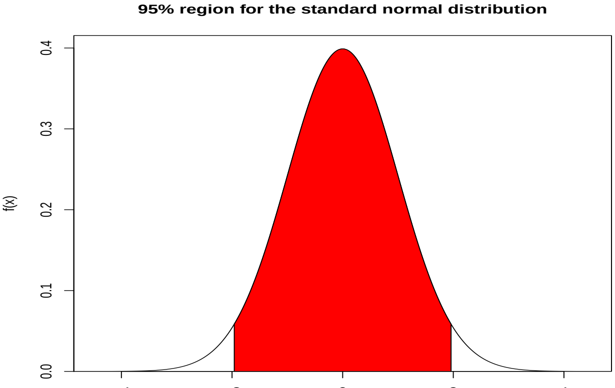
Equivalence to Confidence Regions

Under distribution assumptions identical to the null hypothesis of the corresponding test the confidence interval is identical with the region of acceptance of the test.

Basics of hypothesis testing: Two-tailed test

Using a significance level of $\alpha = 0.05$, the critical values for the alternative hypothesis $\mathrm{H}_1 : p \neq 0.5$ are

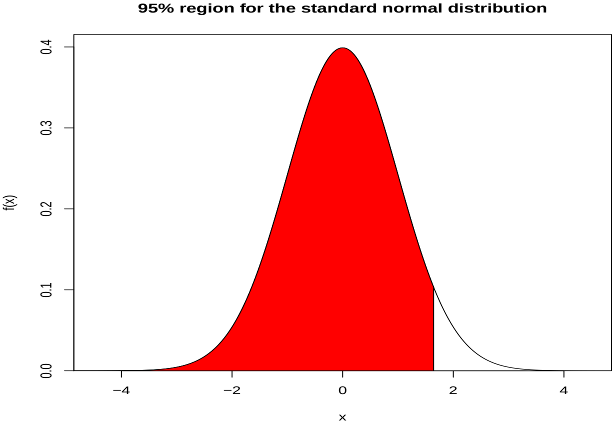
- $crit_L = q_{0.025}, = -1.96$ under normal distribution assumptions,
- $crit_R = q_{0.975}, = 1.96$ under normal distribution assumptions.



Basics of hypothesis testing: Right-tailed test

Using a significance level of $\alpha = 0.05$, the critical value for the alternative hypothesis $\mathrm{H}_1 : p > 0.5$ is

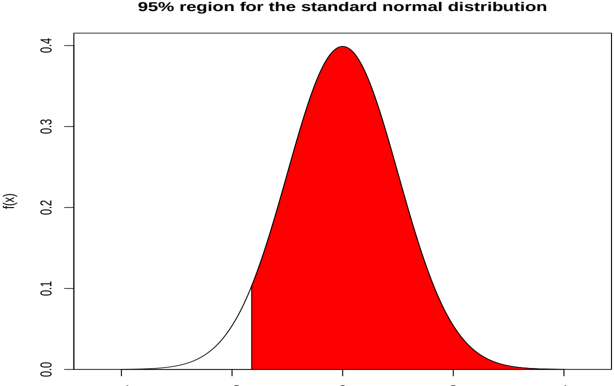
- $crit_R = q_{0.95}, = 1.645$ under normal distribution assumptions,
- $crit_L = q_0 = -\infty$ (The interval is always open on left side.)



Basics of hypothesis testing: Left-tailed test

Using a significance level of $\alpha = 0.05$, the critical value for the alternative hypothesis $\mathrm{H}_1 : p < 0.5$ is

- $crit_L = q_{0.05}, = -1.645$, under normal distribution assumptions
- $crit_R = q_1 = \infty$ (The interval is always open on right side.)



Critics of classical hypothesis testing

"The problem is simple, the researchers are disproving always false null hypotheses and taking this disproof as near proof that their theory is correct." [1]

Multiple testing leads to finding too many significant results.

Overpowered studies lead to finding too many significant results.

Reporting only significant findings leads to systematic bias in research.

[1] (anonymous post from http://andrewgelman.com/)

Confidence regions

Confidence region

A confidence region is a set of points which cover the parameter to estimate with a certain probability

Confidence regions require quantiles and thus distributions
these distributions can come from

- the sample itself (Bootstrap, sampling distribution),
- asymptotic assumptions (CLT) or
- distribution assumptions.

parametric tests

parametric tests assume a specific parametric distribution of the data which leads to a specific parametric distribution of the test statistic

e. g.: **student's t** test

- assumption: data follow (approximately) a normal distribution
- test statistics: $t = \frac{\overline{x} - \mu_0}{\sigma}$
- resulting distribution of the test statistic t is a **student's t** distribution

only for a few parametric distributions such test statistics with predefined distributions exist

most frequently used tests (Central limit theorem)

optimal power, if distribution assumptions are fulfilled

Central Limit Theorem

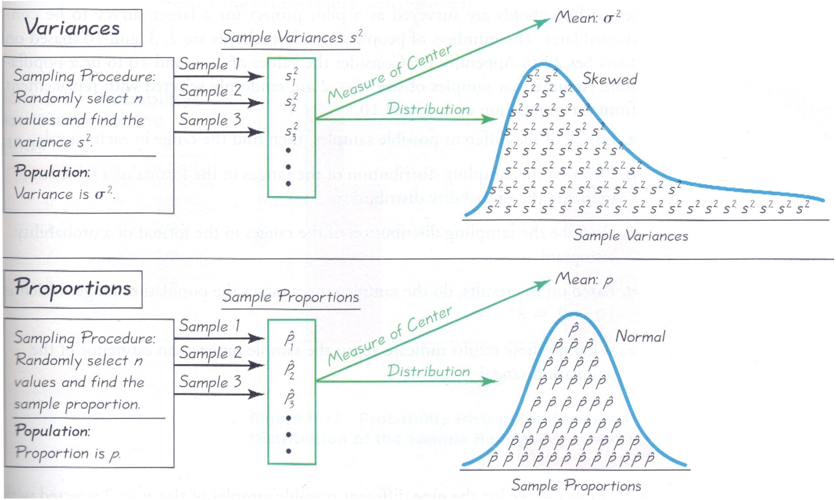
Central Limit Theorem

Let $X1, X2, \ldots$ be a sequence of i. i. d. random variables with expectation $\mathbb{E}[X_i] = \mu$ and $Var[Xi] = \sigma^2 < \infty$. Then as n grows towards infinity, the random variables $\sqrt{n}(\overline{X} - \mu)$ converge in distribution to a normal distribution $N(0, \sigma^2)$.

$$\overline{X} \sim^P N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow$$

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim^P N\left(0, 1\right) \tag{1}$$

Sampling distributions



Non-parametric tests

non-parametric tests do not assume any parametric distribution
however they require:

- a minimum sample size of 100 observations (depends on the scenario, if this minimum is higher)
- a unimodal distribution

typically based on order and rank statistics, such as rank sums etc.

Bootstrap distribution

The bootstrap estimator simulates the sampling distribution of an estimator by sampling with replacement from the original sample and calculating the estimator repeatedly for this ‘new’ sample. This means that 'artificial' additional samples are drawn from the distribution defined by the ECDF in order to approximate the underlying estimators distribution in cases when the CLT does not kick in.

Permutation Distribution

The permutation distribution is

- the exact distribution of any reasonably constructed estimator for a small enough sample, after calculating all possible values of the test statistic under permutations of the data points' labels
- or the approximation of the exact distribution of any reasonably constructed estimator for a sample too large to caluculate all possible permutations

Confidence Interval

Confidence Interval

A set of values which contains the estimate of an unknown population parameter with a certain percentage (confidence level 1-α), if many samples were drawn repeatedly. Formally, the confidence interval at confidence level $1-\alpha$ for a random sample X with probability distribution depending on parameter(s) θ is an interval with random endpoints $(\ell(X), u(X))$ which fulfills

$$\mathbb{P}\left[\ell(X) \leq \theta \leq u(X)\right] = 1-\alpha$$

additional assumptions for $\ell(X)$ and $u(X)$ can be:

- symmetric around the "center" of the interval (mean or median)
- the same amount of the distribution $\frac{\alpha}{2}$ lies below $\ell(X)$ and above $u(X)$

Interpretation of Confidence intervals

in terms of:

- **(repeated) samples**
  "If this procedure was repeated on multiple samples, the calculated confidence interval (differing for each sample) would contain the true parameter 95% of the cases."
- **single sample**
  "With 95% probability the confidence interval contains the true value of the population parameter."
  probability statement about confidence interval, not parameter!
- **acceptance region of hypothesis test**
  "The confidence interval contains possible values of the parameter for which the difference between the parameter and the observed estimate is not significant at the 5% level."

Overview of tests by structure

- Classical **parametric tests**
  Assume a certain distribution of the data characterised by parameters
  calculate a sufficient statistics which follows a different type of distribution
- **non-parametric tests**
  Do not assume a certain distribution
  work with general properties like ranks and rank statistics
- **resampling** methods
  sample an exact or approximate distribution of an estimator based on 'newly drawn samples' out of the original data sample

Overview of tests for the mean

- Classical parametric tests
  - Gauss test for 1 or 2 samples
    special case: test for proportions (assuming Binomial distribution for which the porportion is estimated)
  - Student's t test (exact) for 1 or 2 samples
  - Welch's test (approximate student's t test) for 1 or 2 samples
  - ANOVA for 2 or more samples
- Bayesian test for the mean (and proportions) (for 1 or more samples)
- non-parametric tests
  - Wilcoxon-Mann-Whitney test (for 2 samples)
  - Kruskal-Wallis test (for 2 or more samples)
  - resampling methods (bootstrapping, permutation test) (for 1 or more samples)

Variance comparison

- parametric tests
  - F test
    comparing two variances under normal distribution assumption
  - ANOVA (Analysis of Variance)
    special case of the F test for comparing means of different samples and models
- Bayesian test for variance and Bayesian ANOVA
- non-parametric tests based on permutation or bootstrapping

Distribution comparison

- χ^2-test for homogeneity
  comparison of frequency distributions
- Kolmogorow-Smirnow test, Cramer-von-Mises test
  whether sample data stem from a predetermined distribution
- Shapiro-Wilks test
  whether data comes from a normal distribution

Proportions test

parametric test

assumed distribution for the data
Binomial distribution $B(n, p)$

parameter p is the proportion of 'successes' out of n trials

assumed distribution for the parameter
normal distribution

Testing proportions

- Task: test the observed relative frequency p based on a sample against an assumed frequency p_0
- Null hypothesis: the frequency of an event is p_0
  $H_0 : p = p_0$ or $H_0 : p \geq p_0$ or $H_0 : p \leq p_0$
- Alternative hypothesis: the frequency of an event differs from p_0
  $H_A : p \neq p_0$ or $H_A : p < p_0$ or $H_A : p > p_0$
- Significance level: often $1 - \alpha = 0.95$ ($\Rightarrow \alpha = 0.05$)

Estimating a population proportion

- Task: estimating an unknown relative frequency p based on a sample
- Relative frequency in population: $p = \frac{m}{N}$, where m denotes the amount of items with a certain property
- Point estimate for p: relative frequency $\hat{p}$ in a simple random sample
- Interval estimate for p: confidence interval around $\hat{p}$ that will contain p with a probability of $1 - \alpha$
- Confidence level: often $1 - \alpha = 0.95$ ($\Rightarrow \alpha = 0.05$)

Central Limit Theorem

Central Limit Theorem:

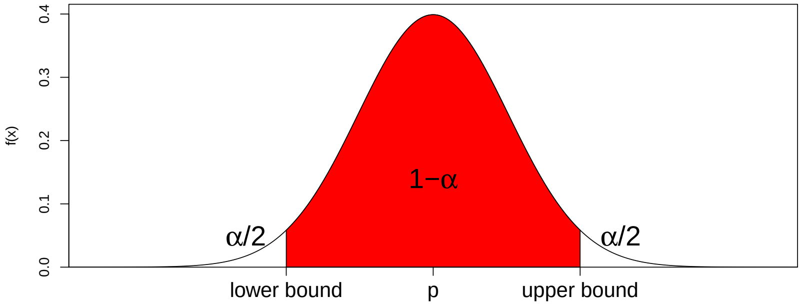
For large sample size n (rule of thumb: $n \geq 100$), the relative frequencies $\hat{p}$ are approximately normally distributed

- with expected value p and
- variance $\frac{p(1-p)}{n}$.

$\Rightarrow$ **We continue using the normal distribution!**

Step 1: Symmetrical interval for $\hat{p}$

- We know the parameter p and sample size n
- We look for a statement about the sample proportion $\hat{p}$
- We find the (approximate) symmetrical interval $[\hat{p}_l, \hat{p}_u]$, in which possible sampling results $\hat{p}$ lie with a certain fixed probability $1 - \alpha$
- $\mathbb{P}(\hat{p}_l \leq \hat{p} \leq \hat{p}_u) = 1 - \alpha$



Step 1: Symmetrical interval for $\hat{p}$

The "standard machine" yields the following procedure:

1. Find the $(1-\frac{\alpha}{2})$ quantile $q_{1-\frac{\alpha}{2}}$, often $q_{0.975}$ or $q_{0.995}$
2. Do the (inverse) z-transformation $x = \mu + z\sigma$

Upper Bound

$$\hat{p}_u = p + q_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

Lower Bound

$$\hat{p}_l = p - q_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

- Statement about probabilities of sample proportions $\hat{p}$ this interval is used for 'prognosis' under a **known** value p or to see for which values the null hypothesis of the corresponding test cannot be rejected

Step 2: Acceptance/Confidence interval for p

If we do not know p we use the estimate $\hat{p}$

Acceptance/Confidence interval for p with confidence level $1-\alpha$

$$p_u = \hat{p} + q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2}$$

$$p_l = \hat{p} - q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{3}$$

where

- $\hat{p} \ldots$ relative sample frequency
- $n \ldots$ sample size
- $q_{1-\frac{\alpha}{2}} \ldots (1-\frac{\alpha}{2})$ quantile of the standard normal distribution

We say: "The confidence interval contains p with probability $1-\alpha$."

Testing the proportion

Testing for proportions

We reject the null hypothesis, if our test statistic

$$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

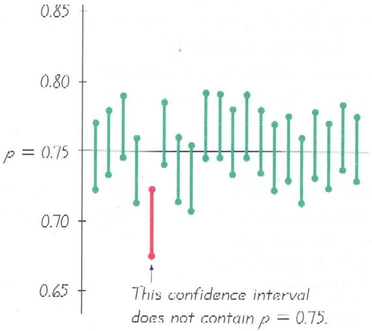
exceeds the $\alpha/2$ quantile $q_{1-\frac{\alpha}{2}}$ (two-sided) or α quantile $q_{1-\alpha}$ (one-sided) of the normal distribution, which is valid for p under H_0.

R

prop.test(x, n, p = NULL, alternative = c(two.sided", less", "greater"), conf.level = 0.95, correct = TRUE)

Step 2: Confidence interval for p

- We know the sample proportion $\hat{p}$ and the sample size n
- We look for a statement about the (true) proportion p within the population
- Question: If the symmetric interval $[\hat{p}_l, \hat{p}_u]$ around p contains the sample proportion $\hat{p}$ with a probability of $1 - \alpha$, what's the probability that an interval built around $\hat{p}$ will contain the "true" parameter p?



Proportion of Adults Believing in Global Warming

In a Pew Research Center poll, 1051 of 1501 randomly selected adults in the United States believe in global warming, so the sample proportion is $\hat{p} = 0.70$. Is this significantly different from random guessing?

```
binom.test(c(1051, 450), p = 0.5)
prop.test(1051, 1501, p = 0.5)
```

Proportion of Adults Believing in Global Warming

In a Pew Research Center poll, 1051 of 1501 randomly selected adults in the United States believe in global warming, so the sample proportion is $\hat{p} = 0.70$. Is this significantly different from random guessing?

Acceptance/Confidence interval with $1 - \alpha = 0.95$ confidence level:

$$
\begin{aligned}
p_u &= \hat{p} + q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.7 + 1.96\sqrt{\frac{0.7(1-0.7)}{1501}} \approx 0.723 \\
p_l &= \hat{p} - q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.7 - 1.96\sqrt{\frac{0.7(1-0.7)}{1501}} \approx 0.677
\end{aligned}
$$

Interpretation:

With a probability of 95%, the interval $[0.677, 0.723]$ contains the relative frequency of believers in global warming in the US. As the 0.5 mark is not contained in this interval, we may safely say that the majority of adults in the US believe in global warming. Thus, people have a strong belief which is significantly different from random guessing ($p = 0.5$).

A statement summarizing the results

70% of United States adults believe that the earth is getting warmer. That percentage is based on a Pew Research Center poll of 1501 randomly selected adults in the United States. In theory, in 95% of such polls, the percentage should differ by no more than 2.3 percentage points in either direction from the percentage that would be found by interviewing all adults in the United States.

Note:

- Sample should be a simple random sample.
- Confidence level should be provided.
- Sample size should be provided.
- Except for rare cases, the quality of the poll results depends on the sampling method and the size of the sample; the size of the population is usually not a factor.

Required sample size

What sample size is needed?

- Sample size depends on the required accuracy (formula for $\mathbb{V}[\hat{p}]$ has n in the denominator, thus the standard deviation of $\hat{p}$ decreases proportionally to $\sqrt{n}$).
- Calculation can be done by solving the corresponding formula

$$n = \frac{q^2_{1-\frac{\alpha}{2}} \hat{p}(1-\hat{p})}{(\underbrace{p_u - \hat{p}}_{E})^2},$$

where $E = p_u - \hat{p} = \hat{p} - p_l$ denotes the desired margin of error.

- If no estimate $\hat{p}$ is known, we simply assume $\hat{p} = 0.5$, yielding

$$n = \frac{q^2_{1-\frac{\alpha}{2}}}{4E^2}.$$

Inference about two proportions

We compare two population proportions, and call

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

the pooled sample proportion, where x_i denotes the number of successes and n_i the size of sample $i \in \{1, 2\}$, and $p_i = \frac{x_i}{n_i}$.

Requirements

1. Sample proportions are from two simple random samples that are independent (i.e. sample values are not related or paired).
2. The normal approximation applies (i.e. $x_i = n_i p_i \geq 5$ and $n_i - x_i = n_i(1 - p_i) \geq 5$ for each sample i).

Inference about two proportions

Under these assumptions, it can easily be shown (see e.g. Triola 2011, p. 455) that

Test Statistic for Two Proportions

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$$

($p_1 - p_2$ is fixed in H_0, usually $p_1 = p_2$ implying $p_1 - p_2 = 0$)

asymptotically follows a standard normal distribution. This means that we can look up P-values and critical values in the tables, and the confidence interval estimate of $p_1 - p_2$ is given by:

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$$

where $E = q_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.

Example: Inference about two proportions

Do Airbags Save Lives?

The table below lists results from a simple random sample of front-seat occupants involved in car crashes (based on data from "Who Wants Airbags?" by Meyer and Finney, *Chance*, Vol. 18, No. 2). Use a 0.05 significance level to test the claim that the fatality rate of occupants is lower for those in cars equipped with airbags.

	Airbag	No Airbag
Occupant Fatalities	41	52
Total Number of Occupants	11,541	9,853

Example: Inference about two proportions

Hypotheses:

$$\begin{aligned} \mathrm{H}_0 : &\quad p_1 \geq p_2 \\ \mathrm{H}_1 : &\quad p_1 < p_2 \end{aligned}$$

Requirements:

1. Two independent simple random samples $\Rightarrow$ OK!
2. $41 > 5$ and $11500 > 5$ and $52 > 5$ and $9801 > 5 \Rightarrow$ OK!

R

```
airbag<-c(41,52)
total<-c(11541,9853)
prop.test(airbag,total,alternative=less")
```

Testing a population mean (variance σ^2 known)

parametric test

assumed distribution for the data
normal distribution $N(\mu, \sigma^2)$

parameter μ is the mean of the data

assumed distribution for the parameter
normal t distribution, only if σ is known

Testing a population mean (variance σ^2 unknown)

parametric test

assumed distribution for the data
normal distribution $N(\mu, \sigma^2)$

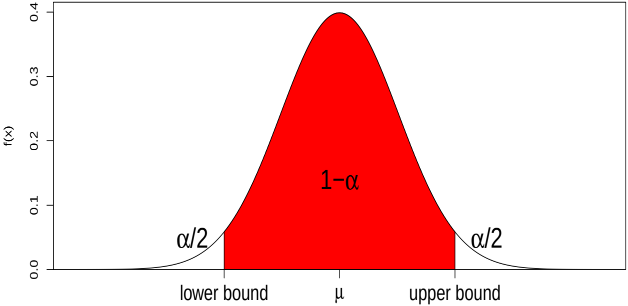
parameter μ is the mean of the data

assumed distribution for the parameter
student's t distribution, if variance σ^2 is unknown (almost always the case)
$\Rightarrow$ **student's t test**

Estimating a population mean (σ known)

Using the Central Limit Theorem we can construct intervals for testing and estimation Requirements:

- We have a simple random sample.
- The value of the population standard deviation σ is **known**.
- The population is normally distributed *or* approximately normally distributed (rule fo thumb: $n > 100$, symmetric and without heavy tails or outliers).



Test for the population mean

Test for the mean, (approximate) normal distribution, σ known

$$
\begin{aligned}
H_0: \quad \mu &= \mu_0 \\
H_A: \quad \mu &= \mu_A \text{ or } \neq \mu_0 \text{ or } \geq \mu_0 \text{ or } \leq \mu_0
\end{aligned}
$$

The corresponding test statistic is

$$
Z = \frac{\mu - \hat{x}}{\sigma/\sqrt{n}}
$$

Interval for the population mean (σ known)

Acceptance/Confidence interval for μ with confidence level $1-\alpha$

$$\mu_u = \bar{x} + q_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{4}$$

$$\mu_l = \bar{x} - q_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{5}$$

where

- $\bar{x} \dots$ sample mean
- $\sigma \dots$ population standard deviation
- $n \dots$ sample size
- $q_{1-\frac{\alpha}{2}} \dots (1-\frac{\alpha}{2})$ quantile of the standard normal distribution

We say: "The confidence interval contains μ with probability $1-\alpha$."

Choosing the appropriate distribution

Is the population normally distributed?	←*YES*	Is σ known?	→*NO*	Is the population normally distributed?
↓*NO*				↓*NO*
symmetric and no haevy tails and no outliers				symmetric and no heavy tails and no outliers
↓*YES*		NO		↓*YES*
Use normal z distribution		Use non-parametric or resampling methods		Use student's t distribution

Intervals for the population mean (σ unknown)

Stylized facts:

- Up to know, we "magically" knew the population standard deviation σ, enabling us to find confidence intervals for the population mean μ by applying the Central Limit Theorem.
- More realistic: we have to *estimate* σ from our data – how do the results change?
- In one sentence: Replace the standard normal quantiles by so called Student t quantiles.
- More precisely: Given that a population is normally distributed, the quantity

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad (\text{"}z\text{-transformation for } \bar{x} \text{ with unknown } \sigma\text{"})$$

follows a Student t distribution with degrees of freedom $n-1$.

Estimating a population mean (σ unknown)

- Objective: Construct a confidence interval used to estimate a population mean.
- Requirements:
  1. The sample is a simple random sample.
  2. Either the sample is from a normally distributed population (symmetric and without heavy tails or outliers).
- Confidence interval:

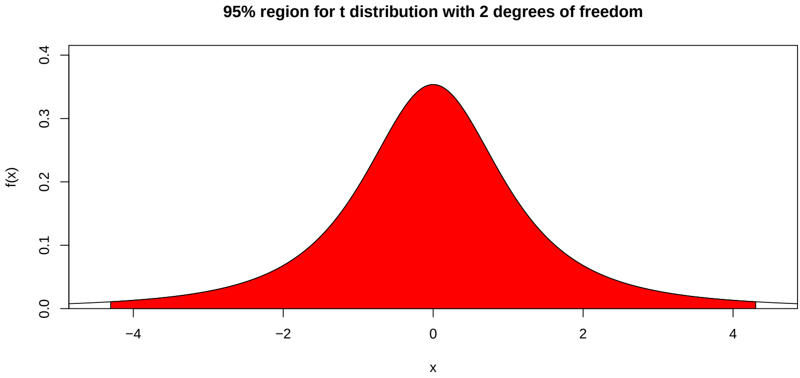
$$\bar{x} - E < \mu < \bar{x} + E, \quad \text{where } E = t_{\alpha/2}\frac{s}{\sqrt{n}} \text{ and } df = n - 1.$$

The Student t distribution

Finding a critical t value

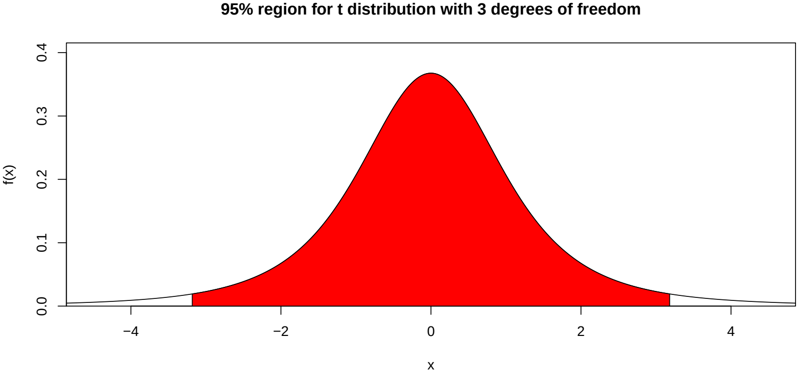
A simple random sample of size n is taken from a normally distributed population. Find the critical value $t_{\alpha/2}$ corresponding to a 95% confidence level (i.e. $t_{0.025}$ or the 0.975 quantile).

The Student t distribution



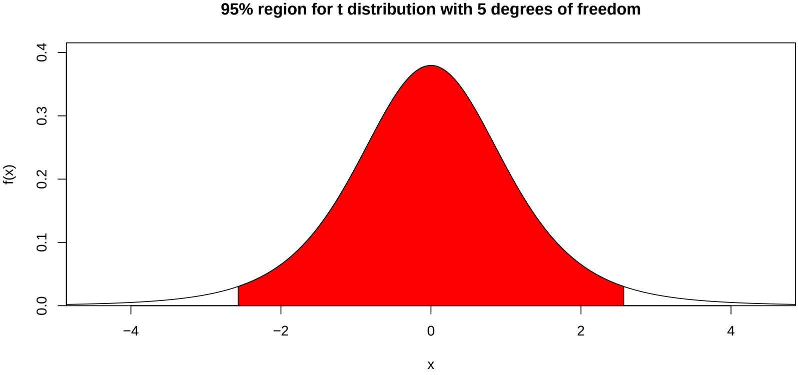
$$df = 2 \text{ implies } t_{0.025} \approx 4.30$$

The Student t distribution



$$df = 3 \text{ implies } t_{0.025} \approx 3.18$$

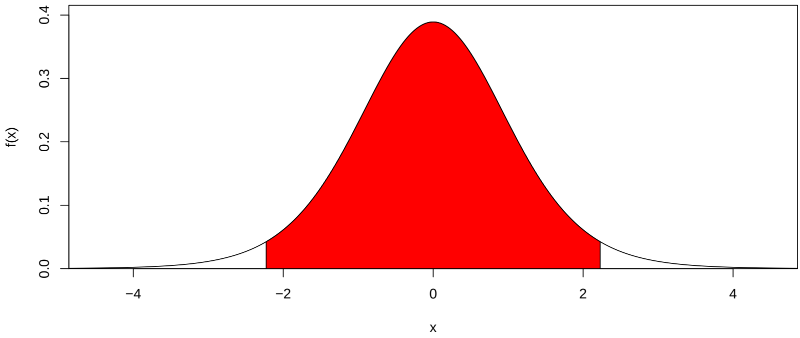
The Student t distribution



$$df = 5 \text{ implies } t_{0.025} \approx 2.57$$

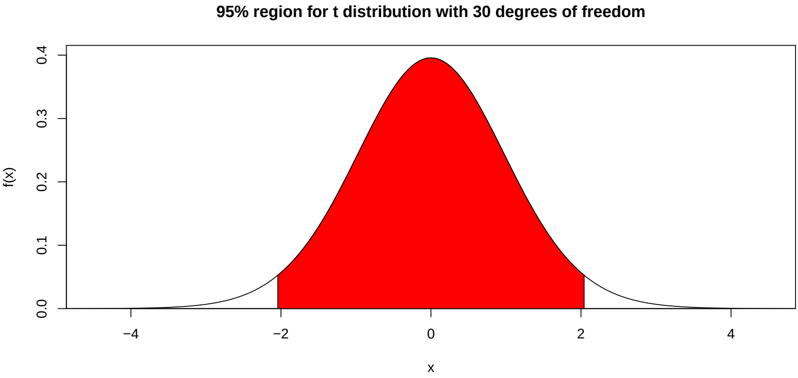
The Student t distribution

95% region for t distribution with 10 degrees of freedom



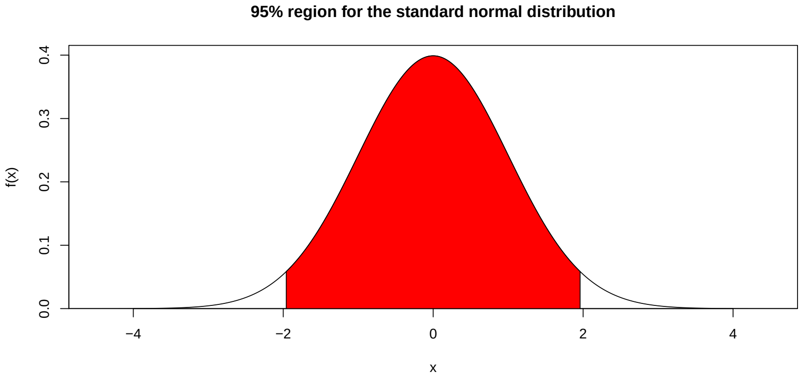
$$df = 10 \text{ implies } t_{0.025} \approx 2.23$$

The Student t distribution



$$df = 30 \text{ implies } t_{0.025} \approx 2.04$$

The Student t distribution $\rightarrow$ Normal distribution



$$df = \infty \text{ implies } t_{0.025} = z_{0.025} \approx 1.96$$

Example: Testing a claim about a mean with σ unknown

Boat Safety

Data set 1 from Triola contains the following information about body weights of people on a boat: $n = 40$, $\bar{x} = 172.55$ lb, $s = 26.33$ lb. Do not assume that the value of σ is known. Use these results to test the claim that men have a mean weight greater than 166.3 lb, which was the weight in the National Transportation and Safety Board's recommendation M-04-04. Use a 0.05 significance level.

Requirement check: (1) simple random sample, (2) σ is not known, (3) $n > 30$ or the population is normally distributed. $\Rightarrow$ OK!

Example: Testing a claim about a mean with σ unknown

- Traditional method:

$$\begin{aligned} H_0: &\quad \mu \leq 166.3 \\ H_1: &\quad \mu > 166.3 \end{aligned}$$

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{s/\sqrt{n}} = 1.501 < 1.685$$

- P-value method: Find area to the right of the test statistic $t = 1.501$: P-value is 0.0707.
- Confidence interval method: 90% confidence interval: 165.54 lb $< \mu <$ 179.59 lb, which contains the assumed $\mu = 166.3$.

Because we fail to reject the null, we conclude that there is not sufficient evidence to support a conclusion that the population mean is greater than 166.3 lb, as in the National Transportation and Safety Board's recommendation.

Required sample size

What sample size is needed?

- Sample size depends on the required accuracy (formula for $\mathbb{V}[\bar{x}]$ has n in the denominator, thus the standard deviation of $\bar{x}$ decreases proportionally to $\sqrt{n}$).
- Calculation can be done by solving formula (4) or (5) for

$$n = \left(\frac{q_{1-\frac{\alpha}{2}} \sigma}{E} \right)^2 .$$

where again $E = \mu_u - \bar{x} = \bar{x} - \mu_l$ denotes the desired margin of error.

Example: Required sample size

IQ Scores of Statistics Students

Assume that we want to estimate the mean IQ score for the population of statistics students. How many must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean?

For a 95% confidence interval, we have $\alpha = 0.05$, so $q_{1-\frac{\alpha}{2}} = 1.96$. Because we want the sample mean to be within 3 IQ points of μ, the margin of error is $E = 3$. Also, $\sigma = 15$. We get:

$$n = \left(\frac{q_{1-\frac{\alpha}{2}}\sigma}{E}\right)^2 = \left(\frac{1.96 \times 15}{3}\right)^2 = 96.04 \approx 97 \quad \text{(rounded up).}$$

Example: Required sample size

IQ Scores of Statistics Students

Assume that we want to estimate the mean IQ score for the population of statistics students. How many must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean?

Interpretation:

Among the thousands of statistics students, we need to obtain a simple random sample of at least 97 students. Then we need to get their IQ scores. With a simple random sample of only 97 stats students, we will be 95% confident that the sample mean $\bar{x}$ is within 3 IQ points of the true population mean μ.

Example: Finding a CI for μ (σ unknown)

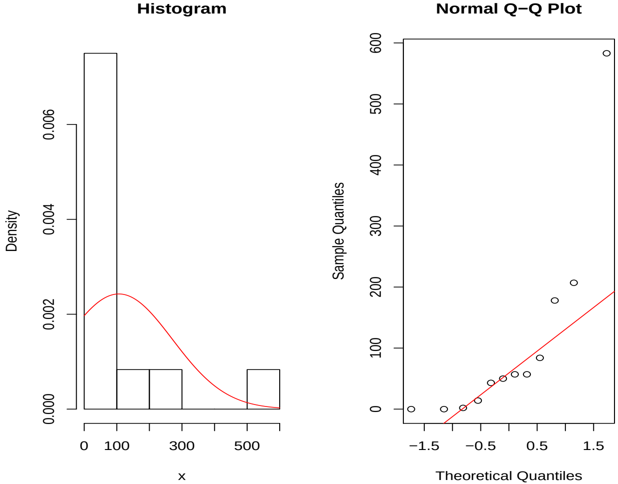
Confidence Interval for Alcohol in Video Games

Twelve different video games showing substance use were observed. The duration times (in seconds) of alcohol were recorded, with the times listed below (based on data from "Content and Ratings of Teen-Rated Video Games," by Haninger and Thompson, *Journal of the American Medical Association*, Vol. 291, No. 7). The design of the study justifies the assumption that the sample can be treated as a simple random sample. Use the sample data to construct a 95% confidence interval estimate of μ, the mean duration time that the video showed the use of alcohol.

84 14 583 50 0 57 207 43 178 0 2 57

Example: Finding a CI for μ (σ unknown)

Caveat: $n = 12 < 100$, thus we must determine whether the data appear to be from a normal population.



Example: Finding a CI for μ (σ unknown)

The requirements are not satisfied $\Rightarrow$ STOP!

Let's continue anyway:

- $n = 12$
- $\bar{x} = 106.25$
- $s \approx 164.33$
- $\alpha = 0.05$, $df = n - 1 = 11 \Rightarrow t_{\alpha/2} \approx 2.20$
- $E = t_{\alpha/2} \frac{s}{\sqrt{n}} \approx 104.42$
- $\bar{x} - E < \mu < \bar{x} + E \Rightarrow 1.84 < \mu < 210.66$

This result is highly questionable because it assumes incorrectly that the requirements are satisfied! Other methods such as nonparametric estimation or bootstrap resampling are needed, the latter yielding a confidence interval of $35.3 < \mu < 205.6$ (Triola).

Comparing 2 samples

- variances are equal → two-sample t test
- variances are not equal → no exact solution exists !!!
  Approximation: Welch's t test

Caveat: when comparing 2 samples, we need to assure equal variances or use the Welch approximation

R

t.test(x, y = NULL, alternative = c(two.sided", less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)

var.test(x, y, ratio = 1, alternative = c(two.sided", less", "greater"), conf.level = 0.95, ...)

Non-parametric test for comparing means

A non-parametric test assumes no parametric distribution of the population, e.g. normal

- Wilcoxon Rank Sum test (for 1 sample), Mann Whitney U test (for two samples)
- Wilcoxon Signed Rank test (for the difference between 2 samples)

R

wilcox.test(x, y = NULL, alternative = c(two.sided", less", "greater"), mu = 0, paired = FALSE, exact = NULL, correct = TRUE, conf.int = FALSE, conf.level = 0.95, ...)

Inference about two means: Independent samples

Independent and Dependent Samples

- Independent samples: The sample values from one population are not related to or somehow naturally paired or matched with the sample values from the other population. Example: Average number of words spoken per day, measured amongst 123 men and 134 women.
- Dependent samples: The sample values are *paired*, i.e. from the same subject (before/after) or from matched pairs (such as husband/wife). Example: "Freshman 15".

Remark

We cover the case where σ_1 and σ_2 are unknown and not assumed to be equal, which is the most common case. Nevertheless, these assumptions can easily be modified (c.f. Triola 9-3).

Inference about two means: Independent samples

A Word of Caution

Before conducting a hypothesis test, consider the context of the data, the source of the data, the sampling method, and explore the data with graphs and descriptive statistics. Be sure to verify that the requirements are satisfied.

Requirements

1. σ_1 and σ_2 are unknown and not necessarily equal.
2. The two samples are independent.
3. Both samples are simple random samples.
4. The two sample both come from populations with normal distributions (symmetric and without heavy tails or outliers).

Inference about two means: Independent samples

It can easily be shown (see e.g. Triola 2011, p. 466) that

Test Statistic for Two Means: Independent Samples

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

($\mu_1 - \mu_2$ is fixed in H_0, usually $\mu_1 = \mu_2$ implying $\mu_1 - \mu_2 = 0$)

approx. follows a Student-t distribution with $df = \min(n_1, n_2) - 1$. This means that we can look up P-values and critical values in the tables, and the confidence interval estimate of $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

where $E = t_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and $df = \min(n_1, n_2) - 1$.

"Are Women Really More Talkative Than Men?" by Mehl et al., *Science, Vol. 317, No. 5834*

Number of Words Spoken in a Day

Men			Women		
n_1	$=$	186.0	n_2	$=$	210.0
$\bar{x}_1$	$=$	15,668.5	$\bar{x}_2$	$=$	16,215.0
s_1	$=$	8632.5	s_2	$=$	7301.2

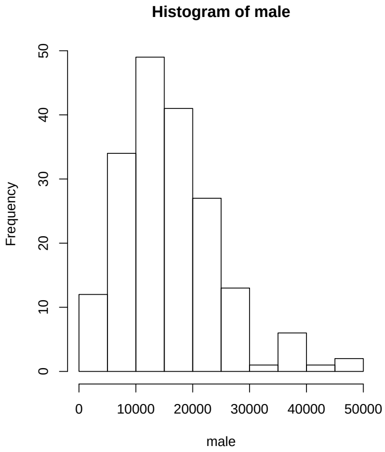
Use a 0.05 significance level to test the claim that men and women speak the same mean number of words in a day.

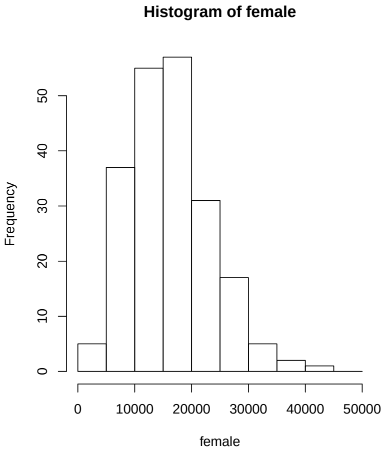
Requirement check:

1. Population standard deviations are not known $\Rightarrow$ OK!
2. The samples are independent $\Rightarrow$ OK!
3. We assume that we have simple random samples $\Rightarrow$ OK!
4. Both samples are large $\Rightarrow$ OK!

Inference about two means: Independent samples

Even though samples are large, let's look at our data:





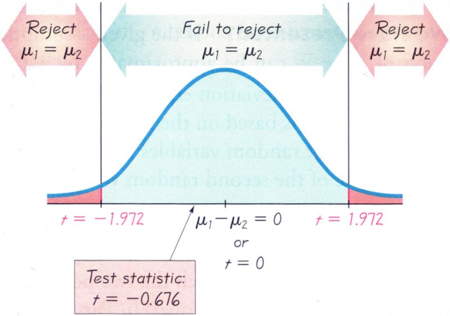
Inference about two means: Independent samples

Hypotheses:

$$H_0: \quad \mu_1 = \mu_2$$
$$H_1: \quad \mu_1 \neq \mu_2$$

Calculations:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \overbrace{(\mu_1 - \mu_2)}^{0}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -0.676, \quad df = \min(n_1, n_2) - 1 = 185$$



Inference about two means: Dependent samples

We have paired data, thus we can talk about individual differences d between the two values in a single matched pair. Examples: weight gain, age difference between husband and wife, etc.

Requirements

1. The sample data are dependent/paired.
2. Both samples are simple random samples.
3. The pairs of values have differences that are from a population with a normal distribution (symmetric and without heavy tails or outliers).

Inference about two means: Dependent samples

The quantity

Test Statistic for Two Means: Dependent Samples

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

(μ_d is fixed in H_0, usually $\mu_d = 0$)

approximately follows a Student-t distribution with $df = n - 1$. This means that we can look up P-values and critical values in the tables, and the confidence interval estimate of μ_d is given by:

$$\bar{d} - E < \mu_d < \bar{d} + E$$

where $E = t_{\alpha/2} \times \frac{s_d}{\sqrt{n}}$ and $df = n - 1$.

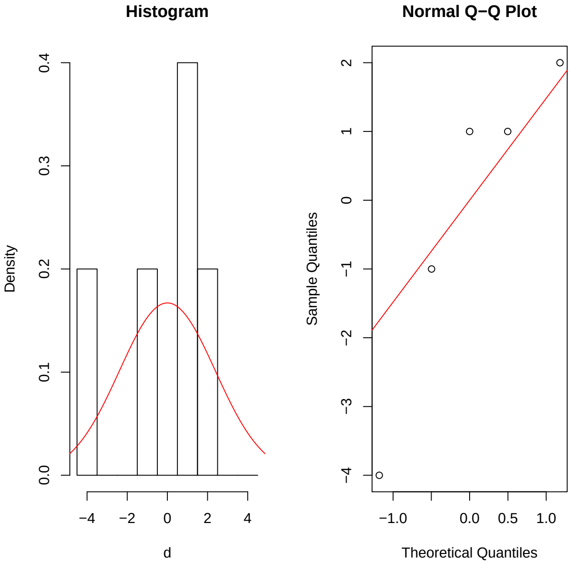
Inference about two means: Dependent samples

Baby Data Set: "Freshman 15"

April weight *(after)*	66	52	68	69	71
September weight *(before)*	67	53	64	71	70
Difference d *(gain)*	-1	-1	4	-2	1

1. Dependent samples $\Rightarrow$ OK!
2. Voluntary response, not simple random sample $\Rightarrow$ FAIL! Let's continue anyway but be careful with interpreting the results
3. Sample size is small $\Rightarrow$ Normality?

Inference about two means: Dependent samples



Inference about two means: Dependent samples

Hypotheses:

$$H_0: \quad \mu_d = 0$$
$$H_1: \quad \mu_d \neq 0$$

Calculations:

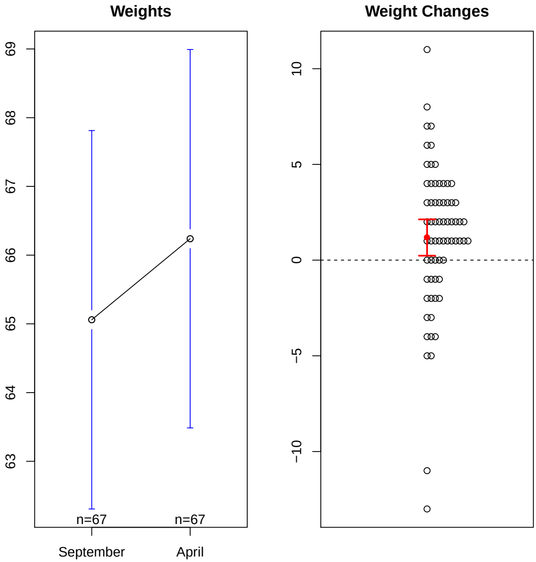
$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{0.2}{\frac{\sqrt{5.7}}{\sqrt{5}}} \approx 0.187, \quad df = n - 1 = 4.$$

Confidence interval:

$$E = t_{\alpha/2}\frac{s_d}{\sqrt{n}} = 2.776 \times \frac{\sqrt{5.7}}{\sqrt{5}} \approx 2.964 \Rightarrow -2.764 < \mu_d < 3.164$$

Not sufficient evidence to warrant rejection of the claim that the mean change in weight from September to April is equal to 0kg. Based on our sample, there does not appear to be a significant weight gain. Limitations: (1) only Rutgers students, (2) potential self-selection bias!

Inference about two means: Dependent samples



Testing the variance - χ^2 test

parametric test

distribution assumption for the data
normal distribution

distribution assumption for the test statistic
χ^2 distribution

Estimating a population variance

Given:

- a normally distributed population with variance σ^2
- independent samples of size n with sample variance s^2 for each sample

Then:

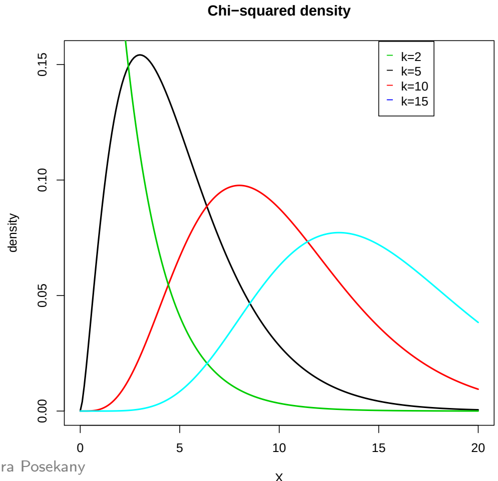
- The sample statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

  has a sampling distribution called the chi-square distribution with $n-1$ degrees of freedom.

The χ^2-distribution

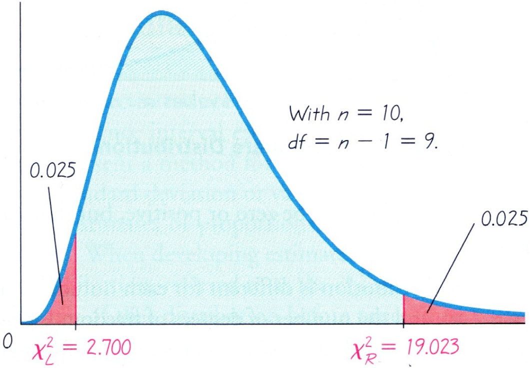
Density functions of a χ^2-distribution with $df = k$ degrees of freedom:



- only nonnegative values
- not symmetric
- moves "to the right" and approaches a normal distribution as $df \to \infty$

Finding Critical Values of χ^2

A simple random sample is obtained. Construction of a confidence interval for the population variance σ^2 requires the left and right critical values of χ^2 corresponding to a confidence level of 95% and a sample size of $n = 10$. Thus, we are looking for a critical value χ^2_L separating an area of 0.025 in the left tail, and a critical value χ^2_R separating an area of 0.025 in the right tail.



Estimating a Population Standard Deviation or Variance

- Objective: Construct a confidence interval used to estimate a population standard deviation or variance.
- Requirements:
  1. The sample is a simple random sample.
  2. The population must have normally distributed values (even if the sample is large).
- Confidence interval for σ^2:

$$\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$$

- Confidence interval for σ:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_R}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_L}}$$

Example: Finding a CI for σ

Confidence Interval for Bottle Fillings

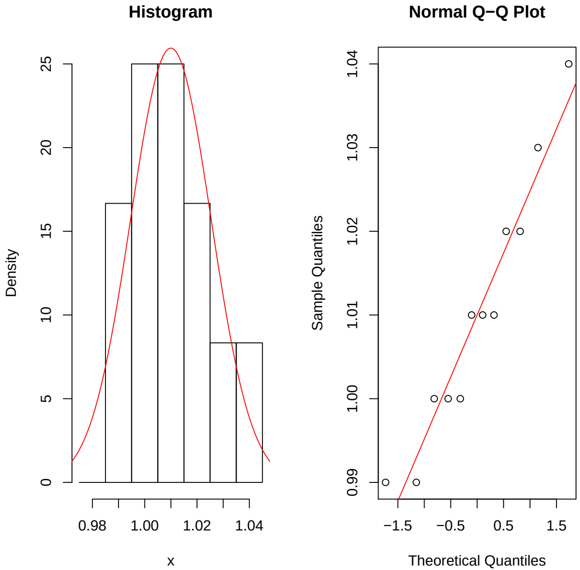
Market conditions require fillings of bottles that do not vary much. Listed below are twelve fillings (in liters). Use the sample data to construct a 95% confidence interval estimate of the standard deviation of all fillings.

0.99 1.01 1.00 1.01 1.03 1.01 1.02 0.99 1.00 1.02 1.00 1.04

Requirement check:

1. Simple random sample $\Rightarrow$ OK
2. Normality?

Example: Finding a CI for σ



Example: Finding a CI for σ

- $s = 0.015374$
- $n = 12$, $df = 11 \Rightarrow \chi^2_L = 3.816$, $\chi^2_R = 21.920$
- $$\begin{aligned} \sqrt{\frac{(n-1)s^2}{\chi^2_R}} &< \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_L}} \\ 0.011 &< \sigma < 0.026 \end{aligned}$$

Based on this result, we have 95% confidence that the limits of 0.011 liter and 0.026 liter contain the true value of σ. The confidence interval can also be expressed as $(0.011, 0.026)$ or $[0.011, 0.026]$, but the format of $s \pm E$ *cannot* be used because the confidence interval does not have s at its center.

Comparing 2 variances - F test

parametric test

distribution assumption for the data
normal distribution

distribution assumption for the test statistic
F distribution

Statistics for nominal data

A (test-)statistic for the dependence of nominal data:

χ^2 (chi-squared)

- Applicable for 2 nominal variables. Also for one nominal and a second variable with few levels or intervals.
- Dependence, if the conditional distributions of one variable given the "subpopulation" of the other variable are not equal.
- Goals:
  - Measuring strength of statistical dependence
  - Testing for that dependence
  - Testing for homogeneity

Example: Measuring dependence of two nominal variables

Distribution of 500 freshman students at JKU Linz (Faculty of Social Sciences, Economics and Business)

Observed frequencies O_{ij}:

Field of study

Sex	BWL	Soz	VWL	SoWi	Stat	**Sum**
female	**110**	120	20	30	20	300
male	90	60	30	10	10	200
Sum	200	180	50	40	30	**500**

Observed relative frequencies o_{ij}:

Field of study

Sex	BWL	Soz	VWL	SoWi	Stat	**Sum**
female	**0.22**	0.24	0.04	0.06	0.04	0.60
male	0.18	0.12	0.06	0.02	0.02	0.40
Sum	0.40	0.36	0.10	0.08	0.06	1

o_{11} = relative frequency of female BWL-students = $110/500 = \mathbf{0.22}$

Example: Measuring dependence of two nominal variables

Conditional distribution

If there were **no statistical relationship** between *sex* and *field of study*, the **conditional distribution** of *field of study* **would be equal** amongst males and females!

Conditional distribution:

	Field of study					
Sex	BWL	Soz	VWL	SoWi	Stat	**Sum**
female	**0.367**	0.400	0.067	0.100	0.067	1
male	**0.450**	0.300	0.150	0.050	0.050	1

Example: Measuring dependence of two nominal variables

Conditional distribution (given independence)

The distribution amongst males and females equals the marginal distribution of *field of study*.

Conditional distribution given independence:

Field of study

Sex	BWL	Soz	VWL	SoWi	Stat	**Sum**
female	**0.40**	0.36	0.10	0.08	0.06	1
male	**0.40**	0.36	0.10	0.08	0.06	1
marginal	**0.40**	0.36	0.10	0.08	0.06	1

What would the **expected (relative) frequencies** look like, if there were **no dependence** between males/females?

Example: Measuring dependence of two nominal variables

Expected frequencies (given independence)

Field of study

Sex	BWL	Soz	VWL	SoWi	Stat	**Sum**
female	**120**	108	30	24	18	**300**
male	80	72	20	16	12	200
Sum	200	180	50	40	30	500
rel. freq.	**0.40**	0.36	0.10	0.08	0.06	1

Given independence, we expect 40% of the 200 male (300 female) students to study BWL, 36% Soz, ...

Example: Measuring dependence of two nominal variables

Expected relative frequencies given independence: $\mathbf{e_{ij}}$

Field of study

Sex	BWL	Soz	VWL	SoWi	Stat	**Sum**
female	**0.24**	0.216	0.06	0.048	0.036	**0.60**
male	0.16	0.144	0.04	0.032	0.024	0.40
Sum	**0.40**	0.36	0.10	0.08	0.06	1

e_{11} = expected relative frequency of female BWL-students
$= 0.60 \times 0.40 = \mathbf{0.24}$

Example: Measuring dependence of two nominal variables

About finding one measure of dependence

Idea: Use the differences between observed and expected relative frequencies.

Observed relative frequencies o_{ij}:

	Field of study					
Sex	BWL	Soz	VWL	SoWi	Stat	**Sum**
female	**0.22**	0.24	0.04	0.06	0.04	0.60
male	0.18	0.12	0.06	0.02	0.02	0.40
Sum	0.40	0.36	0.10	0.08	0.06	1

Expected relative frequencies e_{ij} (given independence):

	Field of study					
Sex	BWL	Soz	VWL	SoWi	Stat	**Sum**
female	**0.24**	0.216	0.06	0.048	0.036	0.60
male	0.16	0.144	0.04	0.032	0.024	0.40
Sum	0.40	0.36	0.10	0.08	0.06	1

Measuring dependence of two nominal variables: χ^2

Measure of Dependence: Chi-Squared χ^2

$$\chi^2 = n \times \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

o_{ij}	...	observed relative frequencies
O_{ij}	...	observed (absolute) frequencies
e_{ij}	...	expected relative frequencies (given independence)
E_{ij}	...	expected (absolute) frequencies (given independence)
n	...	sample size

Example: Measuring dependence of two nominal variables

Back to the example: Calculating χ^2

$$\boldsymbol{\chi^2} = n \times \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} =$$

$$= 500 \times \left[\frac{(\mathbf{0.22} - \mathbf{0.24})^2}{\mathbf{0.24}} + \ldots + \frac{(0.02 - 0.024)^2}{0.024} \right] =$$

$$= \mathbf{18.06}$$

A measure of dependence: χ^2

Summarizing the above, we conclude:

- $\chi^2 = 0$, if there is no statistical dependence between the two variables ($o_{ij} = e_{ij}$)
- $\chi^2 > 0$, if there is a dependence between the variables ($o_{ij} \neq e_{ij}$ for some i, j)
- χ^2 will never be negative (why?)

What we know (now)...

We have info, if there is dependence, but we don't know...

- whether this dependence is statistically significant
- how strong the dependence is (scaling necessary!)

Testing for significance

- Objective: Conduct a hypothesis test for dependence between the row variable and column variable in a contingency table.
- Requirements: For every cell, the *expected* absolute frequency is at least 5, i.e. $E_{ij} > 5$ for all i, j. Otherwise, check out *Fisher Exact Test* for 2×2 tables.
- Hypotheses:

  H_0 : The row and column variables are independent.

  H_1 : The row and column variables are dependent.

- Test statistic: $\chi^2 = n \times \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
- Critical values and P-values: These stem from a χ^2-distribution with $(r-1) \times (c-1)$ degrees of freedom.

Example: Testing for dependence of two nominal variables

Requirements: For every cell, the *expected* absolute frequency is at least 5, i.e. $E_{ij} > 5$ for all i, j.

Expected frequencies (given independence)

			Field of study			
Sex	BWL	Soz	VWL	SoWi	Stat	**Sum**
female	120	108	30	24	18	300
male	80	72	20	16	12	200
Sum	200	180	50	40	30	500

OK!

Example: Testing for dependence of two nominal variables

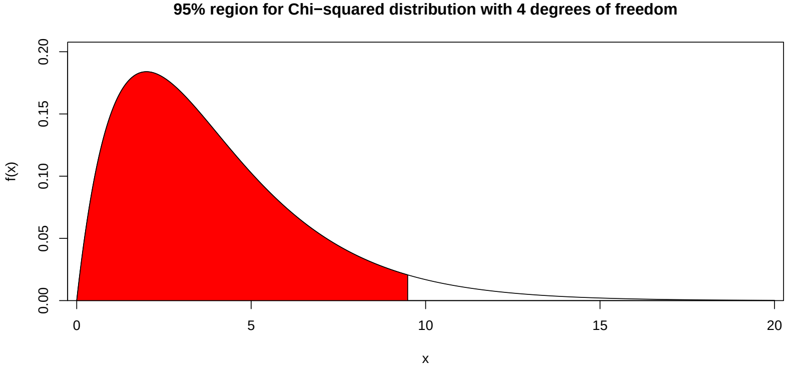
- Hypotheses:

  H_0 : The variables *Sex* and *Field of study* are independent.

  H_1 : The variables *Sex* and *Field of study* are dependent.

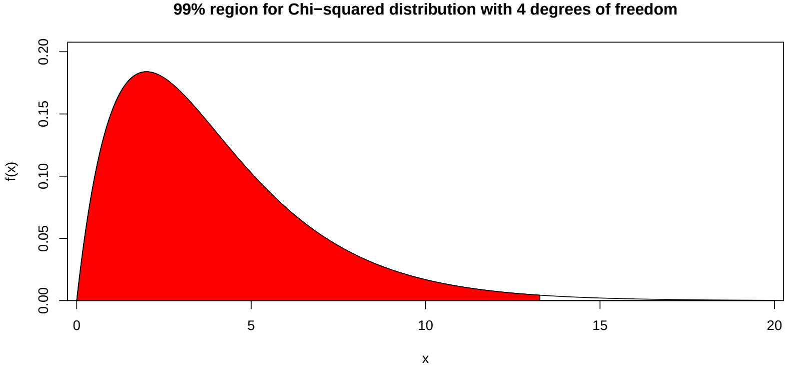
- Test statistic: $\chi^2 = n \times \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 18.06$
- Critical values and P-values: These stem from a χ^2-distribution with $(r-1)(c-1) = (2-1)(5-1) = 4$ degrees of freedom.

Example: Testing for dependence of two nominal variables



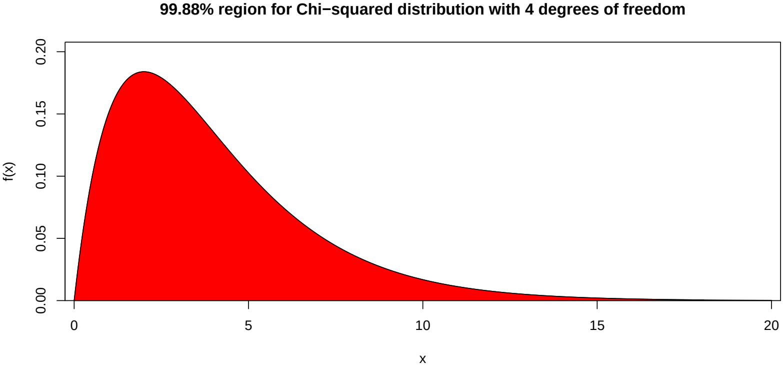
Critical value at 95%-level: $9.488 \ll 18.06 \Rightarrow$ reject H_0!

Example: Testing for dependence of two nominal variables



Critical value at 99%-level: $13.277 < 18.06 \Rightarrow$ reject H_0!

Example: Testing for dependence of two nominal variables



P-value of 18.06 is 0.0012 $\Rightarrow$ reject H_0 at 99.88%-level!

Test of homogeneity

Sometimes we are interested whether two different "populations" (such as males/females) have the same proportion of some characteristics (such as *field of study*).

"New" hypotheses:

H_0 : The proportions of *fields of study* are equal amongst male and female students at WU.

H_1 : The proportions are different.

Good news: Use the same method (it's an equivalent formulation).

Exercise: Is the nurse a serial killer?

Two-way table with deaths when Gilbert was working

	Shifts with a death	Shifts without a death
G. at work	40	217
G. not at work	34	1350