Statistical Methods
(summer term 2024)

Frequentist Hypothesis Testing

(based on original lectures by Prof. Dr. N. Christlieb and Dr. Hans-G. Ludwig)

Dr Yiannis Tsapras

ZAH – Heidelberg

Overview

- Z-Test for One Sample
- Student's T-Test (One-Sample and Two-Sample Cases)
- Testing for Correlation (Spearman Rank Test)
- Chi-Squared Test
- Hypergeometric Distribution and Fisher's Exact Test
- Kolmogorov-Smirnov (KS) Test
- Wald-Wolfowitz Runs Test
- Pros and Cons of Classical Hypothesis Testing

(Gaussian) Z-test, one sample

- Tests whether the mean of a Gaussian population deviates from the reference value μ_0
  - uses a sample of n data points x_i
  - assumes the variance σ^2 of the underlying population is known
  - Hypothesis H_0: x_i are drawn from a normal distribution with variance σ^2 and mean μ_0
- The test statistic is
  $$Z = \frac{\bar{x} - \mu_0}{s}$$
  where $s = \sigma/\sqrt{n}$ is the standard error of the sample mean $\bar{x}$
- Under the null hypothesis Z is normally distributed: $Z \sim N(0,1)$
- Applicable whenever the test statistic can be approximated by a normal distribution (or when it is exactly normal)
  - approximation often valid for large sample sizes due to the central limit theorem
  - when the variance σ^2 is unknown, other tests are more appropriate

Example: Z-test (one-sided)

- A random variable X is normally distributed with variance $\sigma^2 = 9$ and unknown mean μ. You have $n = 10$ samples available, drawn from the distribution of X

- Test the hypothesis $\mu = \mu_0 = 24$ against the alternative hypothesis $\mu > \mu_0$ with a significance level $\alpha = 0.05$
  - Null hypothesis: $H_0 : \mu = 24$
  - Alternative hypothesis: $H_1 : \mu > 24$

- The Z-test statistic is given by:

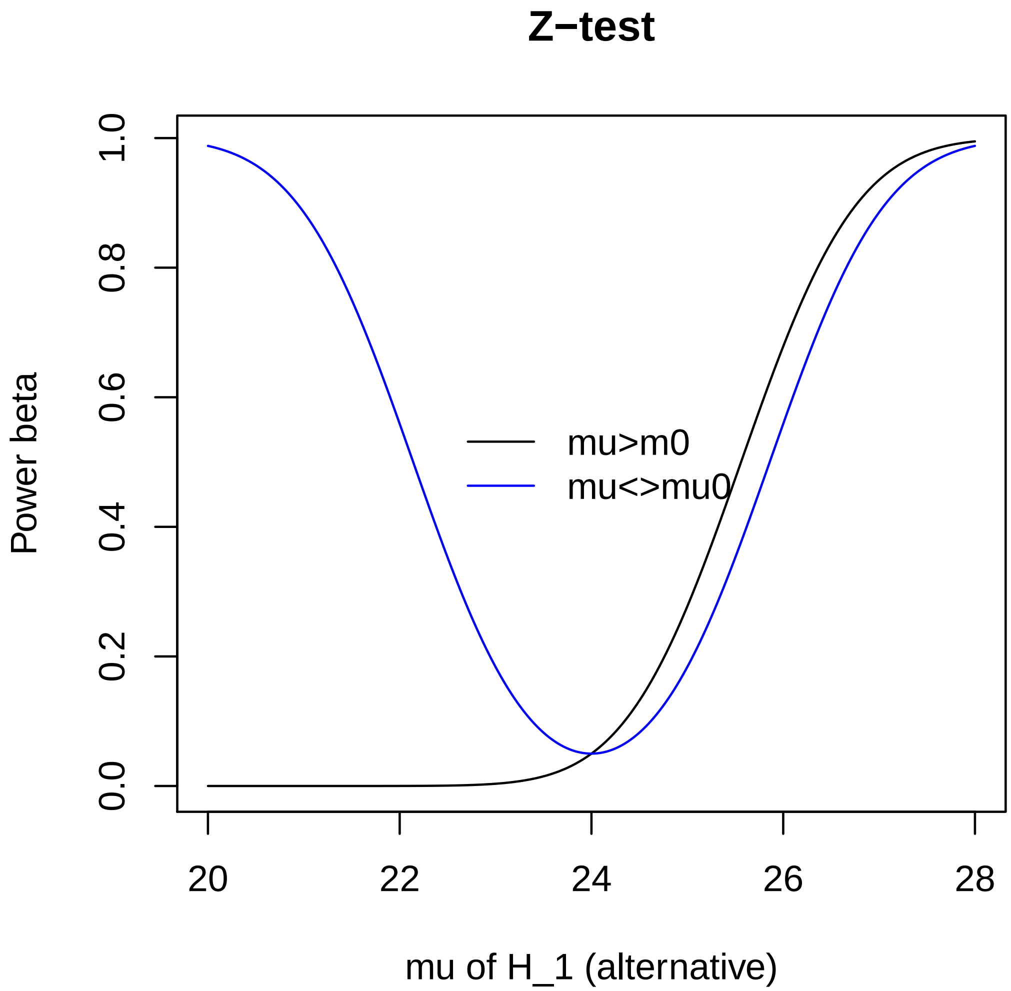
$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- The goal is to determine whether the sample data provides enough evidence to reject H_0 in favor of H_1

- The significance level $\alpha = 0.05$ means that there is a 5% risk of rejecting the null hypothesis when it is actually true

→ blackboard

Example: Z-test (two-sided)



- The black curve is the power of the Z-test for the given example, the blue curve for the two-sided test for $\mu \neq \mu_0$
- Exercise:
  - reproduce the result for the power of the two-sided test!
  - power is the probability of correctly rejecting H_0 when it is false, i.e. $1 - \beta$
  - what is the range of acceptance of $\bar{X}$?

Student's t-test

William Sealy Gosset (1876–1937)

- Worked as chemist with the Guiness brewery in Dublin; quality control, improvement of quality, growing barley

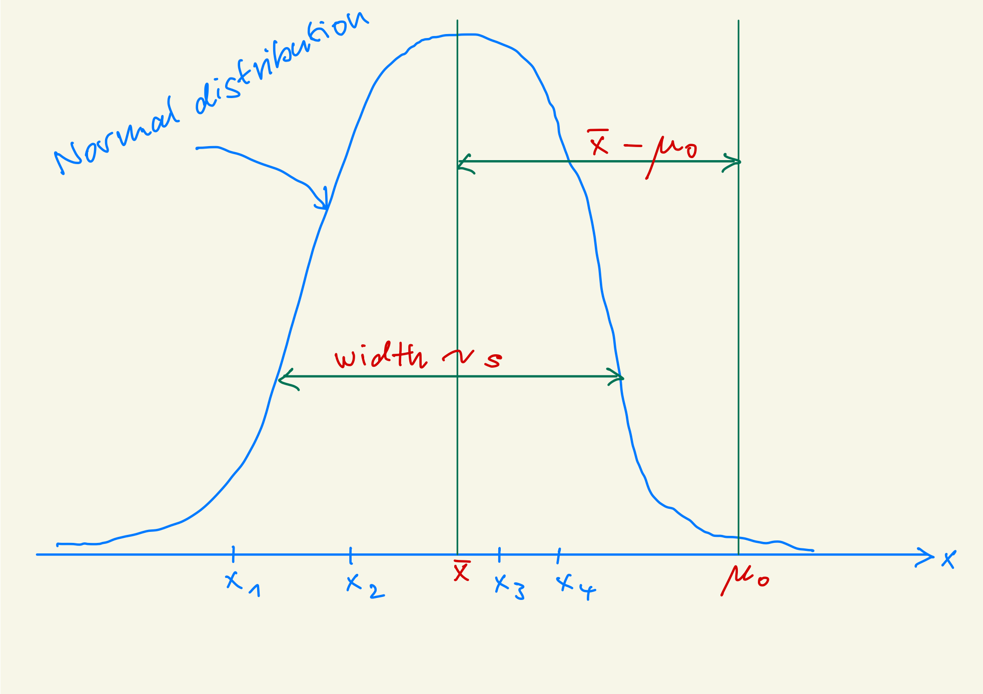
- Published under pseudonym "Student"

- t-test designed having small sample sizes in mind

Student's t-test: one-sample case

- Tests whether the mean of a Gaussian population deviates from reference value μ_0
  - Given a sample of n data points x_i
  - H_0: x_i drawn from a normal distribution of unknown variance σ^2 and mean μ_0
  - estimate of variance derived from the sample

- Test statistic $t = \frac{\bar{x}-\mu_0}{s}$, where $s = \hat{\sigma}/\sqrt{n}$ is the standard error of the sample mean



Student’s t-test: one-sample case

- Note the difference to Z-test: $\hat{\sigma}$ is the sample standard deviation

- the test statistic t follows a t-distribution with $\nu = n - 1$ degrees of freedom. Its PDF is given by:

$$P(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

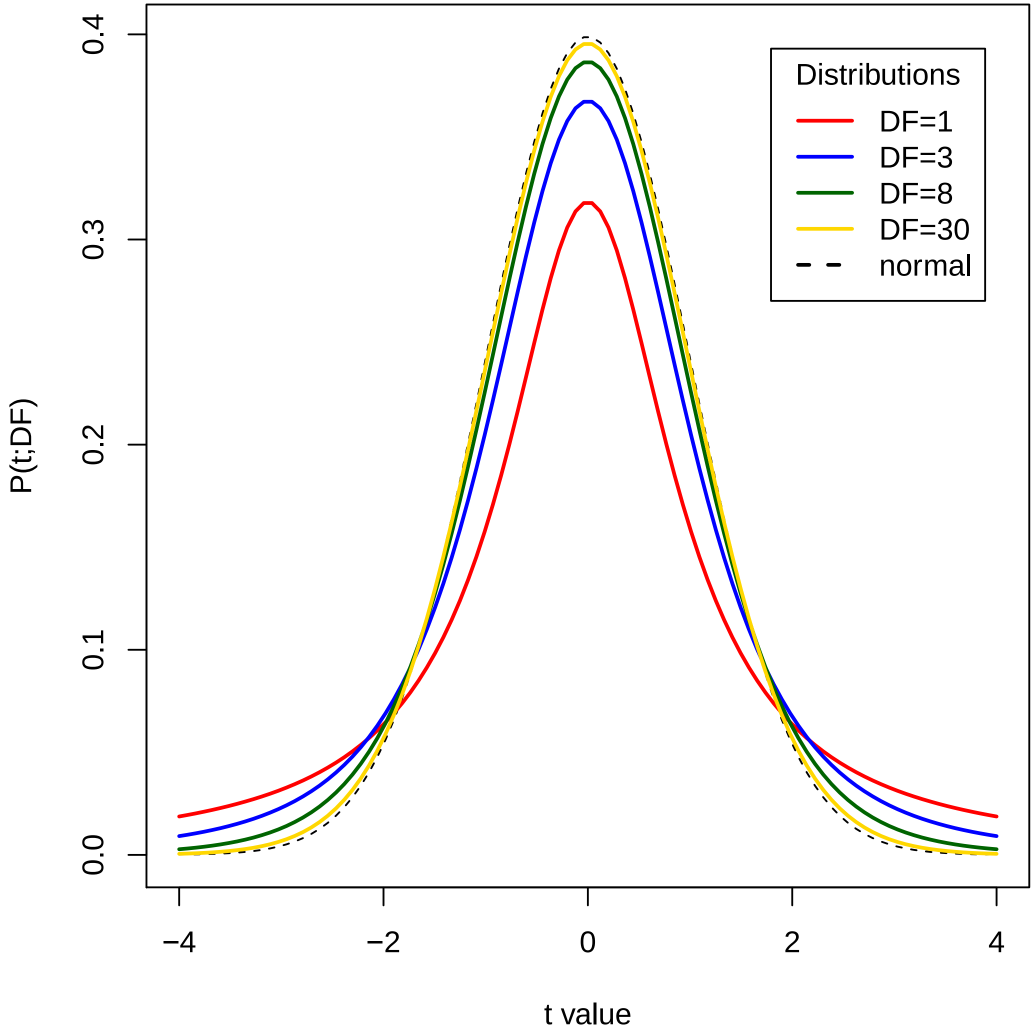
  - where Γ is the gamma function

- for large n (or ν), $P(t;\nu)$ approximates the normal distribution

- its expectation value is always zero

- the variance is $\frac{\nu}{\nu-2}$ for $\nu > 2$. For $\nu \leq 2$, the variance is infinite or undefined, reflecting the heavier tails of the t-distribution compared to the standard Gaussian

Student's t-distributions for various degrees of freedom (DF)



- Note "wings" wider than for normal distribution for small number of DFs
  - additional uncertainty coming from the estimation of the unknown variance

Comment on t-distribution

- The t-distribution is the distribution of the random variable

$$T = \frac{X}{\sqrt{Y/\nu}}$$

  where X and Y are independent random variables. ν is a positive integer – the number of degrees of freedom (NB: difference to n in definition of t)

- $X \sim N(0,1)$ and $Y \sim \chi^2_\nu$

- The ratio T gives a measure of the deviation of the sample mean from the hypothesized mean in terms of the standard error, considering the variability in the sample

- shows that the t-distribution accounts for additional uncertainty due to the estimation of the sample variance, especially in small samples

- For a sample drawn from a normal distribution mean and variance are independent

Comment on t-distribution

- For a sample drawn from a normal distribution mean and variance are independent
  - generate a large number of random samples, calculate the sample means and variances, and then check for correlation between them:

```
# quick check in R for correlation which is a necessary
# while not sufficient condition for independence
a <- matrix(rnorm(10*10000), nrow=10, ncol=10000)
m <- apply(a,2,mean)
v <- apply(a,2,var)
cor.test(m,v)
```

Comment on t-distribution

- From that one may get an idea how Gosset came up with the t-distribution:
  - under H_0, the distance $\overline{x} - \mu_0$ is a normally distributed random variable
  - the natural unit to use when measuring this difference is the standard deviation, which represents the width of the normal distribution
  - knowing that the estimated distance and the width are independent random variables allows one to derive the t-distribution

- As a side note: the distribution of the ratio of two random variables is not just the ratio of their two distributions...
  - it depends on how these two variables interact, especially around their means, variances, and the relationship between them
  - more specifically, the distribution of the ratio is obtained by changing variables, which (as we previously saw) involves integrating the joint probability distribution of X and Y over the appropriate range
    - ⋆ this requires evaluating convolutions and is not as simple as taking the ratio of the two PDFs

Exercise: testing a batch of batteries during production

A manufacturer of batteries wants to test whether a certain batch – say – of 1000 batteries complies to quality standards. The batteries are supposed to hold a charge of 200 mAh. The batch is considered not to quality standards if the mean charge of the batteries likely deviate too much from the nominal value. Since the batteries are emptied during the testing only 10 batteries can be tested. Their charges x were measured to $x \in \{177, 194, 209, 228, 229, 235, 241, 244, 244, 287\}$ mAh. Conduct a two-sided t-test to decide whether the batch is likely following quality standards or needs to be rejected on a significance level $\alpha = 0.05$.

- Program the the t-test yourself (you are allowed to use the `R`-functions for the t-distribution `_t()`)!
- What is the t- and p-value of the measured sample?
- What are the intervals of non-rejection in x and t?
- Is the batch acceptable or not?
- Compare with the result of the `R`-function `t.test()`!

Student's t-test: two-sample case

- n data x_i and m data y_i are from two different Gaussians, however, with $\sigma_x = \sigma_y$
- Test statistic $t = \frac{\bar{x}-\bar{y}}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$
- Here s_p is the pooled standard deviation

$$s_p \equiv \sqrt{\frac{\sum(x_i-\bar{x})^2 + \sum(y_i-\bar{y})^2}{n+m-2}}$$

- Note the difference to the sample standard deviation $\hat{\sigma}$
- t follows a t-distribution with $n+m-2$ degrees of freedom (-2, because we calculate two means from the data)
- There is a generalistation for the case $\sigma_x \neq \sigma_y \rightarrow$ Welch test
- In R, `t.test(...)` implements it all (one- or two-sample, paired-data, one- or two-tailed, equal or unequal variances)

Testing for correlation: Spearman rank test

- Correlations are not necessarily linear (assumption implicit to the Pearson product moment coefficient)

- The Spearman rank correlation coefficient is a measure of correlation, measuring how well a monotonuous function can decribe the relation among two variables

- Coefficient derived from relative ordering of pairs of N observations (x_i, y_i)

- The Spearman rank correlation coefficient is defined as

$$r_s = 1 - 6\frac{\sum_{i=1}^{N}(\text{rank}(x_i) - \text{rank}(y_i))^2}{N^3 - N}$$

- The `rank` is the position in a sorted list of values, see `rank()` in R

Testing for correlation: Spearman rank test

- For not too small samples ($N > 30$, or so), the test statistic

  $$t_r = r_s \sqrt{\frac{N-2}{1-r_s^2}}$$

  follows a Student's t-distribution with $\nu = N - 2$ degrees of freedom

  ```
  # R Example:
  # Sample data
  x <- c(10, 20, 30, 40, 50)
  y <- c(15, 25, 35, 45, 55)

  # Using cor() directly with method = "spearman"
  spearman_corr <- cor(x, y, method = "spearman")
  print(spearman_corr)
  1
  ```

- Also through `cor.test(..., method="spearman")`

χ^2-test

- The χ^2-test is often used as a goodness-of-fit estimator

- Idea: x_i are normally distributed, independent random variables with means μ_i and variances σ_i^2

- The test statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

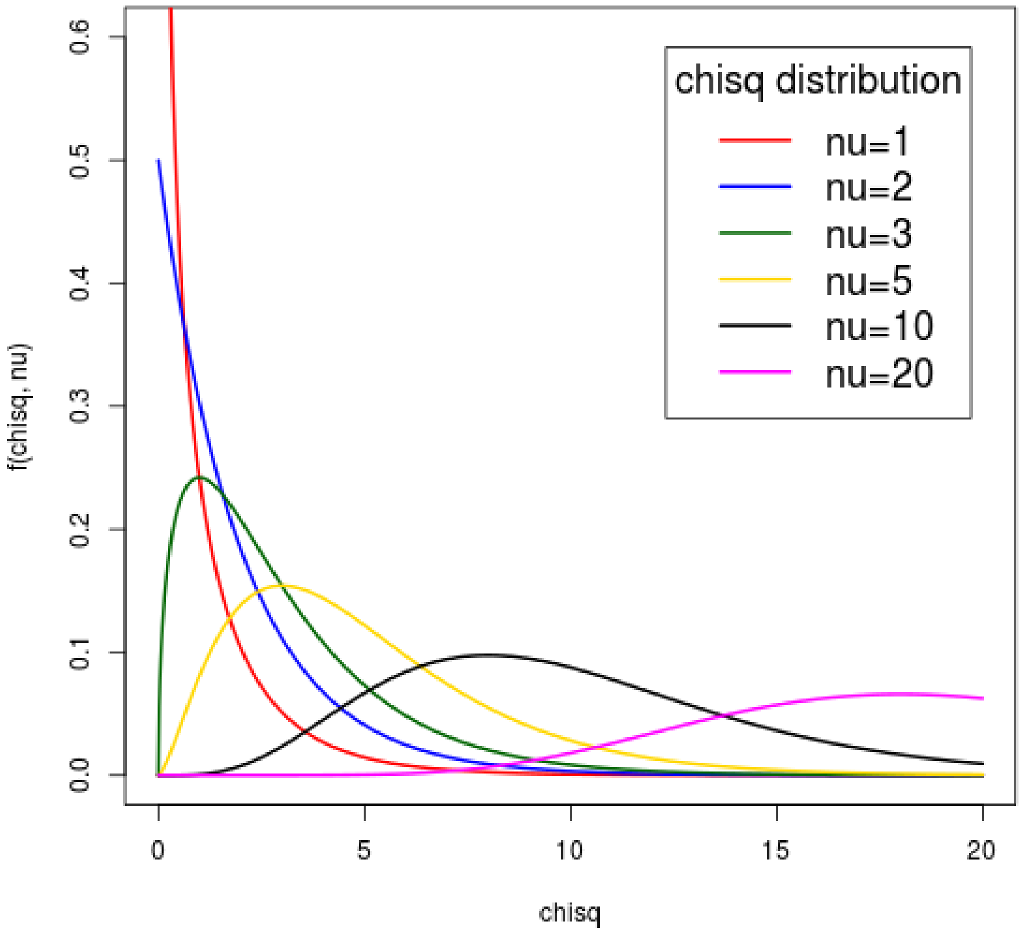
follows a χ^2 distribution

$$f(x = \chi^2, \nu) = \frac{x^{(\nu/2-1)} e^{-x/2}}{2^{\nu/2}\Gamma(\frac{\nu}{2})},$$

  - The χ^2 distribution is only defined for $x > 0$ and is 0 otherwise
  - ν is the number of degrees of freedom, Γ is the Gamma function

- For large ν, the χ^2 distribution approaches a normal distribution with $\mu = \nu$ and $\sigma^2 = 2\nu$

χ^2-distribution



- $\nu =$ number of data points $-$ number of parameters fitted
  - this holds if the model is linear in the parameters

χ^2-test in practice: is a histogram compatible with assumed underlying PDF?

- H_0: The assumed PDF $F(x)$ correctly describes the population from which the sample $x_1, ..., x_n$ was drawn

- Step 1: generate a histogram where each bin j contains at least $b_j \geq 5$ counts, so the approximation is valid (b stands for bin here)
  - if a bin contains fewer than 5 counts, it should be combined with a neighboring bin to satisfy this requirement
  - denote the total number of bins as K

- Step 2: For each bin j, calculate the expected number of counts $e_j = np_j$ based on the assumed PDF $F(x)$
  - p_j is the probability that an observation falls within bin j under the assumed PDF

- Step 3: Compute the χ^2 statistic given the observed and expeced counts:

$$X_0^2 = \sum_{j=1}^{K} \frac{(b_j - e_j)^2}{e_j}$$

χ^2-test in practice: is a histogram compatible with assumed underlying PDF?

- Step 4: choose a significance level α (0.05, 0.01, or similar), for rejecting or not rejecting H_0

- Step 5: calculate the critical value c by solving $c = P^{-1}(1-\alpha)$ using the inverse cumulative χ^2-distribution with $K-1$ degrees of freedom;
  in R use: `qchisq(1-alpha, df=K-1)`

- If $X_0^2 \leq c$ accept H_0, otherwise reject

- Quantification of the statement that the mean squared deviation should not exceed a given level

Exercise: 1) generate some test data by drawing 1000 samples from a standard normal distribution and use the χ^2-test to check the hypothesis that they were drawn from a normal PDF. Plot the histogram and overplot the expected Normal distribution (use $\alpha = 0.05$ and 10 bins). 2) Repeat the test, but this time draw the samples from a t-distribution with a) 8 and b) 24 degrees of freedom →`chi2test_1.ipynb`

The hypergeometric distribution

- Discrete probability distribution
- models the scenario of drawing objects from a finite population $without$ replacement
  - unlike the binomial distribution where each draw is independent (due to replacement)
- in this example we consider only 2 possible types of objects, black or white marbles
- N objects ("marbles") in total, K of one kind ("white marble") $N - K$ of the other kind ("black marble")
- Hypergeometric distribution gives the probability to draw k white marbles in n draws without replacement

$$\mathrm{P}(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

The hypergeometric distribution

$$\mathrm{P}(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

- $\binom{K}{k}$ is a binomial coefficient, also called "K choose k," representing the number of ways to choose k objects from the K objects of the first type

- $\binom{N-K}{n-k}$ represents the number of ways to choose the remaining $n-k$ objects from the $N-K$ objects of the second type

- $\binom{N}{n}$ total number of ways to choose n objects from all N objects, regardless of type

Interlude: counting raccoons on Königstuhl mountain (Muira 2011)

Racoons have been brought from North America to Germany, and happily spread out. A question is: how many live in Germany? Perhaps a bit less ambitiously, we ask how many live in the forest up Königstuhl mountain?

For counting animals in a habitat biologists developed the so-called capture – recapture technique: a number of racoons is randomly captured, labeled and released. Afterwards, racoons are randomly captured again.

Question: $n_1 = 25$ racoons got caught, labeled, and released. Afterwards, $n_2 = 25$ were re-captured, of which $n_\mathrm{x} = 7$ turned out to be labeled. What is the maximum likelihood estimate $\tilde{N}$ of the number of racoons living on Königstuhl mountain?

→ `racoons.ipynb`

Fisher's exact test – example

Assume that psychologists conducted the following experiment: 48 male bank supervisors had to decide on the promotion of employees in their bank. By random selection, 24 supervisors were given personnel files labeled as "female", 24 labeled as "male". In fact, except for the label the files were identical. The following table (called contingency table or truth table) summarizes the outcome:

	male	female	$\sum$
promote	21	14	35
not promote	3	10	13
$\sum$	24	24	48

- Looking at the numbers the question arises: Is there a gender bias?
  - 48 supervisors choosed to promote 35 employees of which only 14 are female
  - put differently: is there an association between gender and probability of being promoted?
- Can the imbalance between the promotion of females and males be understood as statistical fluctuation?

Fisher's exact test – example

	male	female	$\sum$
promote	21	14	35
not promote	3	10	13
$\sum$	24	24	48

- Null hypothesis: there is no gender bias. Moreover, table margins are fixed:
  - there are always 24 female and 24 male candidates for promotion
  - since we do not know any better the total number of promotions (or no-promotions) is also fixed
    - ⋆ one may consider the supervisors as "promoters" and "no-promoters" who either always promote or don't promote irrespective of file content
    - ⋆ could be different when the probability of promotion were known from additional sources of information (but then another test would have to be made)
- if no gender bias, we would expect a 50-50 split of promotions
- like in the kangaroo problem the table has only one degree of freedom
- Singling (arbitrarily) out the upper left corner of male promotions Fisher found that the number of male promotions follows a hypergeometric distribution which becomes the test statistic

Fisher's exact test – test statistic

$$
\begin{array}{cc|c}
N_{11} & N_{12} & n_{1.} \\
N_{21} & N_{22} & n_{2.} \\
\hline
n_{.1} & n_{.2} & n_{..}
\end{array}
$$

(only variables with capital names are random variables)

- Writing the contingency as above (→ partial sums) the test statistic can be written

$$P(N_{11}) = \frac{\binom{n_{1.}}{N_{11}}\binom{n_{2.}}{N_{21}}}{\binom{n_{..}}{n_{.1}}} = \text{dhyper}(N_{11},\ \text{promote},\ \text{not\_promote},\ \text{draws})$$

- For the example one obtains the following table for $P(N_{11})$

N_{11}	11	12	13	14	15	16	17
$P(N_{11})$	.000	.000	.004	.021	.072	.162	.241
N_{11}	18	19	20	21	22	23	24
$P(N_{11})$	.241	.162	???	.021	.004	.000	.000

Confirm the table, and have a look at the documentation of `dhyper()`!

→`Fisher_test.ipynb`

Fisher's exact test – result

N_{11}	11	12	13	14	15	16	17
$P(N_{11})$	.000	.000	.004	.021	.072	.162	.241
N_{11}	18	19	20	21	22	23	24
$P(N_{11})$	.241	.162	???	.021	.004	.000	.000

- From the table the rejection region for the two-sided test with $\alpha = 0.05$ is $N_{11} \in \{11, 12, 13, 14,\ 21, 22, 23, 24\}$.
  - we need to sum the probabilities from the most extreme values (both small and large N_{11}) until the cumulative probability reaches or exceeds $\alpha/2 =$ 0.025 in each tail

- The p-value for the observed $N_{11} = 21$ evaluates to

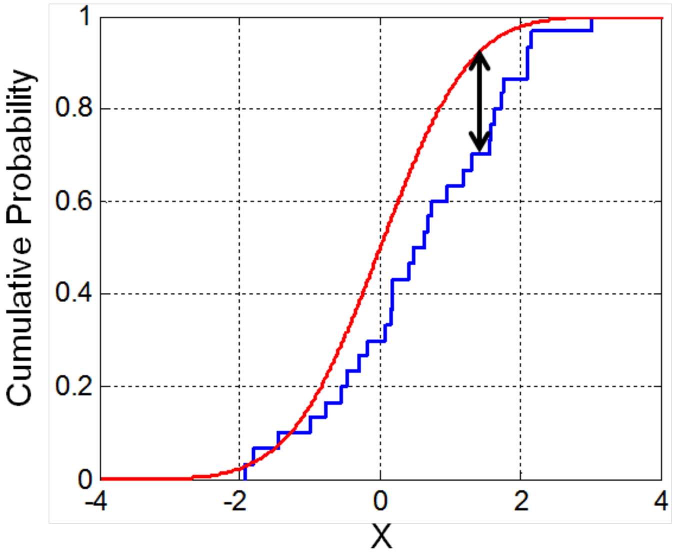
$$\sum_{N_{11} \in \{11,12,13,14,\ 21,22,23,24\}} P(N_{11}) = 0.049\,.$$

- Hence, the test would reject the null hypothesis at significance level 0.05

Kolmogorov-Smirnov (KS) test

- Test for continuous probability distributions

- Tests whether a given sample x_i was drawn from a particular continuous PDF
  - non-parametric test
  - works with discrete cumulative distribution function → no binning necessary
  - suitable for small sample sizes



- Test statistic: maximum absolute deviation between cumulative distribution functions:

$$D = \max \left| \tilde{F}(X) - F(X) \right|$$

- Interestingly, distribution of test statistic does not depend on particular $F(X)$!

- Procedure of calculation: look at all "jump positions" of sample CDF

Kolmogorov-Smirnov (KS) test

- There is also a two-sample version of the KS test, meaning that two empirical distribution functions are compared

- The KS test cannot be used if first parameters of the underlying distribution have been estimated from the sample
  - necessary distributions could be established by Monte Carlo simulations

- The KS test is mainly sensitive to the center of the distribution

- The function `ks.test()` in R implements everything, including the value of the test statistic D
  - for the D.I.Y. aficionados `ecdf()` implements the empirical CDF

- Since KS, more powerful tests have been developed, notably the Anderson-Darlington test. It has proven particularly powerful for testing deviations from normality. In R: `ad.test{nortest}`

Wald-Wolfowitz runs test: Testing for randomness

- This test whether the data are random, in the sense that successive data points are uncorrelated

- Applicable to binary statistic constructed from the data, for instance

  - heads vs. tails
  - random number > 0.5 vs. random number < 0.5
  - positive residual vs. negative residual

- A run is a group of successive data points with the same characteristic, e.g. a sequence with 17 runs . . .
  + + - - - + + + + - + + - - - + + - + + + + - + - + - - - + - +

Wald-Wolfowitz runs test: Testing for randomness

- Test statistic is the number of runs r, having as input
  - $m =$ number of "heads"
  - $n =$ number of "tails"

- The number of runs r is distributed as:
  - if r is even:

$$p(r = 2q) = \frac{2\binom{m-1}{q-1}\binom{n-1}{q-1}}{\binom{m+n}{m}}$$

  - if r is odd:

$$p(r = 2q + 1) = \frac{\binom{m-1}{q}\binom{n-1}{q-1} + \binom{m-1}{q-1}\binom{n-1}{q}}{\binom{m+n}{m}}$$

- Approximately, the statistic $p(r)$ follows a normal distribution $N(\mu, \sigma^2)$ with

$$\mu = \frac{2mn}{m+n} + 1 \qquad \text{and} \qquad \sigma^2 = \frac{2\,m\,n\,(2\,m\,n\, - m - n)}{(m+n)^2\,(m+n-1)}$$

- In R, use `druns(..)` and `runs.test(...)` in package `randtests`

Exercise: analyse human generated random sequences using the Wald-Wolfowitz test

- In the file human-sequences.txt you will find 200 sequences with a length of 50 items each which have been painstakingly created by UKSta students. During creation, the students were told that the mean relative occurence of ones and zeros should be 50/50. Read them in, and split the zeros and ones into vector elements via

  ```
  d <- scan('human-sequences.txt', what='character')
  d <- sapply(strsplit(d, ''), function(x) as.integer(x))
  ```

- Use runs.test() to calculate the distribution of p-values of the 200 sequences.

- Compare the distribution (visually) with the distribution to be expected under the null hypothesis. Conclusions?

- Test your own ability to create random sequences: type strings of zeros and ones and subject them to the Wald-Wolfowitz test!

Classical hypothesis testing: pros and cons

- Strengths
  - Frequentist methods provide a wide array of statistical tests tailored to many common scenarios
  - low p-value can suggest that there is evidence against the null hypothesis, potentially indicating that "something might be going on"
  - classical hypothesis testing doesn't always require a generative model for the alternative hypothesis H_1, making it simpler in certain contexts

- Weaknesses
  - Rejection of H_0 is not proof of H_1
  - one is left with the question of what the actual probability of a hypothesis is, given the observed data
  - many scientific questions are interested in the relative likelihood of H_0 versus H_1 or even H_2, H_3 etc $\rightarrow$ Bayesian model selection
  - the p-value is often misunderstood as the probability of H_0 being true, which it is not. It is actually the probability of obtaining the observed data, or more extreme data, assuming H_0 is true

Go Bayesian or go Frequentist?

- Frequentist methods rely on the long-run frequency of events. The focus is on obtaining the probability of observing the data given that H_0 is true (characterized by the p-value)
  - Uncertainty is expressed through p-values and confidence intervals
  - The p-value does not provide the probability that the null hypothesis is true
  - Thresholds are used to decide whether to reject or not the null hypothesis
  - Require less computational power and are often easier to implement than Bayesian methods
- Bayesian methods rely on priors and can be used to update the probability of a hypothesis given the data. They provide a posterior probability of the hypothesis, which can be directly interpreted as the probability that the hypothesis is true given the observed data
  - Uncertainty is expressed in terms of probabilities that directly quantify belief in a hypothesis after seeing the data
  - Compare the probability of the hypotheses directly by calculating the posterior odds (or Bayes factor)
  - Can be computationally expensive because they often rely on methods like MCMC to compute posteriors