Null Hypothesis Significance Testing

Klaus Sembritzki

Pattern Recognition Lab (CS 5)

18.11.2013

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Table of Contents

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Motivating Examples

1. One-sample significance test

School class **A**, consisting of 15 pupils has an average grade of 2.5 (sample mean) and a sample standard deviation of 1.1 in their final exam. Did the pupils perform significantly different than the **nation-wide** average of 2.0?

2. Two-sample significance test

Did school class **B** consisting of 10 pupils with an average grade of 1.5 perform significantly different from school class **A** (the combined sample standard deviation of **A** and **B** is 1.1)?

Significance Testing

What we want to know

1. Given normally distributed observations $y = \{y_i | i = 1..n\}$, is the population mean nonzero?
2. Given normally distributed observations $y_1 = \{y_{1,i} | i = 1..n_1\}$, $y_2 = \{y_i | i = 1..n_{2,i}\}$ with equal variance, do the population means differ?

Null Hypothesis Significance Testing (NHST)

Reformulation as a NHST problem (not equivalent to original question)

1. Is it reasonably likely to draw samples as extreme as the observed $y = \{y_i | i = 1..n\}$ from a normal distribution with zero mean and unknown variance?
2. Is it reasonably likely to draw samples as different as the observed $y_1 = \{y_{1,i} | i = 1..n_1\}$, $y_2 = \{y_{2,i} | i = 1..n_{2,i}\}$ from the same normal distribution?

Terminology

Null hypothesis

The “boring”, default hypothesis

- The drug has no effects
- Male participants do not differ from female participants

Alternative hypothesis

The exciting, unexpected result that is reported in a scientific publication (a practice which is sometimes criticized)

- Opposite of the null hypothesis
- No one will believe it unless it is proven

Statistical inference

- The process of proving or disproving something with statistical methods
- In NHST, we prove the alternative hypothesis by disproving the null hypothesis

Control of the Type I Error Rate

NHST controls the false positive rate (FPR) of a two-class problem

- Class "n": Samples for which the null hypothesis is fulfilled
- Class "a": Samples for which the alternative hypothesis is fulfilled

Given samples y_i with class labels $y_i \in \{n, a\}$ and a classifier that predicts a class label $\tilde{y}_i$ for each sample, the FPR is defined as the probability $P(\tilde{y}_i = a | y_i = n)$

- We observe exactly one sample in total and therefore perform only one test (classification) in total
- Features are not vector valued in this talk, but just scalars
- FPR is typically used to describe the behaviour of a classifier
- The term *type I error rate* is used instead of FPR in NHST and assigned the letter α
- Freeze the type I error rate at a small value, typically $\alpha = 5\,\%$
- If our sample is classified as "a" despite this small chance of committing a type I error, we **infer** that the null hypothesis does not hold (and that the alternative hypothesis must therefore be correct)

Student's t-Distribution

Chi-squared distribution

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

- $Z_i \sim N(0,1)$ and independent
- n is called "degrees of freedom" (reason for this naming is explained later)
- n times the probability distribution of the sample variance (sample size n) of a standard normal random variable with known mean

Student's t-distribution (signal-to-noise ratio)

$$T = \frac{Z}{\sqrt{V/n}}$$

- $Z \sim N(0,1)$ (Z for z-value),
- $V \sim \chi_n^2$ (V for variance),
- Z and V are independent.

Model Equation

Estimate mean and variance of k normal distributions with equal variance $\{Y_l \sim N(\mu_l, \sigma) | l = 1..k\}$ based on a sample of size $n = n_1 + n_2 + ...n_k$, $y \in \mathbf{R}^n$ with n_l samples from class l.

General linear model

$$Xb = \mu = y - \varepsilon$$

ε : Statistical error

$$X\hat{b} = \hat{\mu} = y - \hat{\varepsilon}$$

$\hat{\varepsilon}$: Statistical residual

Least squares solution

$$\hat{\varepsilon} = y - X\hat{b} = \mu + \varepsilon - XX^+(\mu + \varepsilon)$$
$$= (I - XX^+)\varepsilon$$

$\hat{\varepsilon}$ is a projective mapping of ε onto an N-k dimensional subspace

$$\hat{\varepsilon} \perp X$$

$\hat{\varepsilon}$ independent of model parameters

Example for k=2

$$X = \begin{matrix} 1 & -1 \\ 1 & 1 \\ 1 & 1 \end{matrix} \qquad y = \begin{matrix} 1.1 \\ 1.9 \\ 2.1 \end{matrix} \qquad \mu = \begin{matrix} 1 \\ 2 \\ 2 \end{matrix}$$

The first column of the design matrix X models the intercept.
The second column is -1 for class 1 and 1 for class 2.

Basu's Theorem

Reminder: Student's t-distribution

$$T = \frac{Z}{\sqrt{V/k}}$$

- $Z \sim N(0,1)$,
- $V \sim \chi^2_n$,
- Z and V are independent.

Idea: Calculate a t-statistic from $\hat{b}$ and $\hat{\sigma}$

- Convert $\hat{b}$ into something with a standard normal distribution
- Convert $\hat{\sigma}$ into something with a chi-squared distribution

$\hat{b}$ and $\hat{\sigma}$ are independent due to Basu's theorem: "Any *boundedly complete sufficient* statistic is independent of any *ancillary* statistic."

- *Boundedly complete sufficient*: fulfilled, but this talk will not go into detail
- *Ancillary*: fulfilled, because $\hat{\varepsilon} \perp X$ and $\hat{\varepsilon}$ does thus not depend on the model parameters μ_l

In other words, the sample mean and the sample standard deviation of a normally distributed random variable are statistically independent (the normal distribution is the only probability distribution with this property)

Variance Estimation and t-Statistic

$\widehat{\Sigma}$, the probability distribution $\hat{\sigma}$ is drawn from

- $\hat{\varepsilon}$, drawn from $\widehat{E}$, results from a projective mapping of ε onto an N-k dimensional subspace $\hat{\varepsilon} = (I - XX^{+})\varepsilon$
- Probability density point symmetric around origin $\Rightarrow$ rotate euclidian basis $y_1, .., y_k$ to new basis $y'_1, .., y'_k$, aligned to image and nullspace of the mapping (see figure)
- Coefficients in new basis still i.i.d.
- $\Rightarrow \sum_i^n \widehat{E}_i^2 = (N-k) \cdot \widehat{\Sigma}^2 \sim \chi^2_{N-k} \cdot \sigma^2$

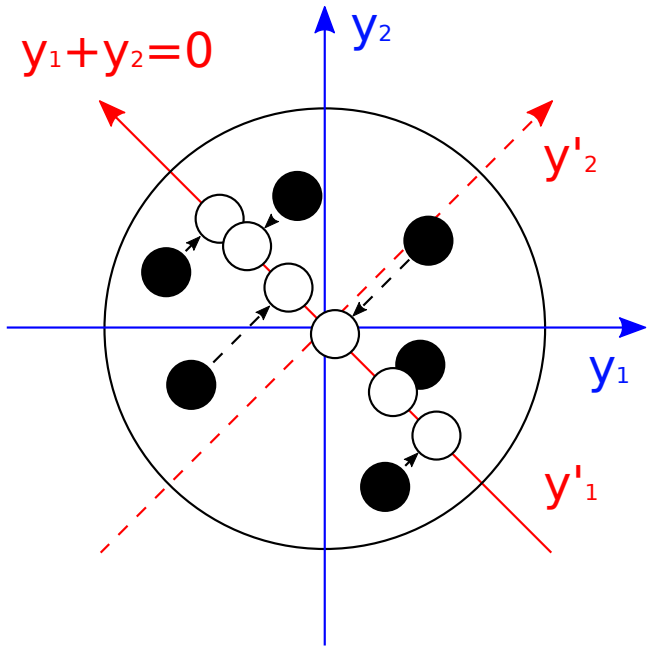
$\widehat{B}$, the probability distribution $\hat{b}$ is drawn from

$$\mathrm{Var}(\widehat{B}) = \mathrm{Var}(X^{+}(\mu + E)) = \mathrm{diag}(X^{+}(X^{+})^{T}) \cdot \sigma^2$$

$$\widehat{B} \sim N(0, \sqrt{\mathrm{diag}(X^{+}(X^{+})^{T})} \cdot \sigma)$$

$$t = \frac{\hat{b}}{\sqrt{\mathrm{diag}(X^{+}(X^{+})^{T})} \cdot \hat{\sigma}}$$

t is sampled from a Student's t distribution with N-k degrees of freedom



Special Design Matrices

One-sample t-test

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}}$$
$$\text{df} = n - 1$$

Two-sample t-test

$$t = \frac{\overline{y}_1 - \overline{y}_2}{s_{y_1 y_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
$$s_{y_1 y_2} = \sqrt{\frac{(n_1 - 1)s_{y_1}^2 + (n_2 - 1)s_{y_2}^2}{n_1 + n_2 - 2}}$$
$$\text{df} = n_1 + n_2 - 2$$

Unequal variances

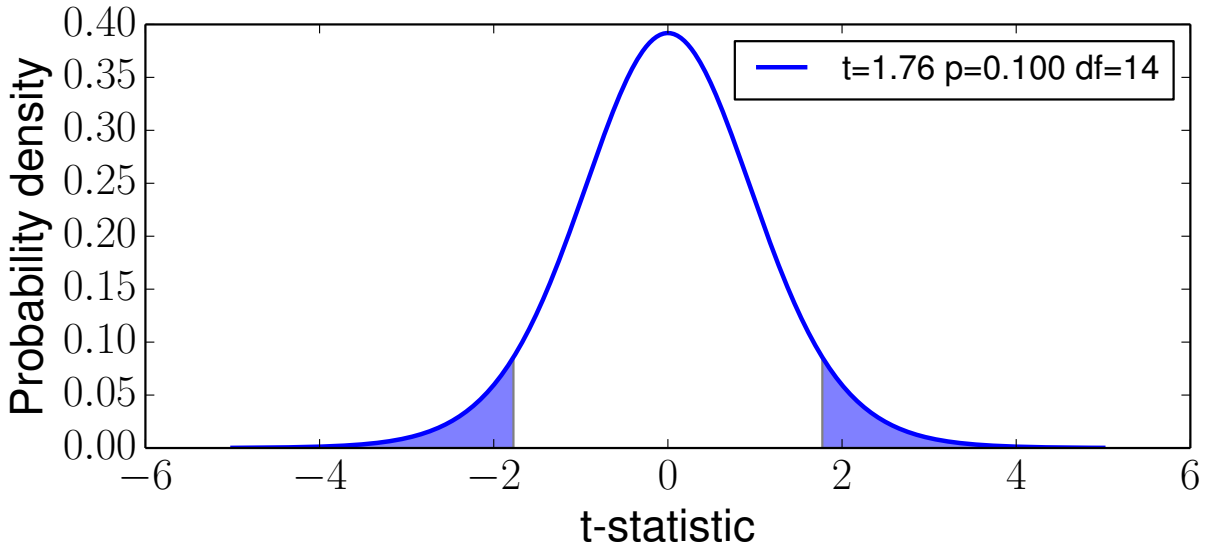
Not considered in this talk

p-Value

1. Two-tailed one-sample t-test

School class **A**, consisting of 15 pupils, has an average grade of 2.5 (sample mean) and a sample standard deviation of 1.1 in their final exam. Did the pupils perform significantly different than the **nation-wide** average of 2.0?

- Null hypothesis H_0: “The pupils’ grades have the same mean as the nation’s”
- Calculate the one-sample t-statistic of “grade - 2.0” with 14 degrees of freedom
- Reject H_0 if the p-value $p < \alpha$, with $p = P(|T| > |t|)$ and significance level $\alpha = 5\%$

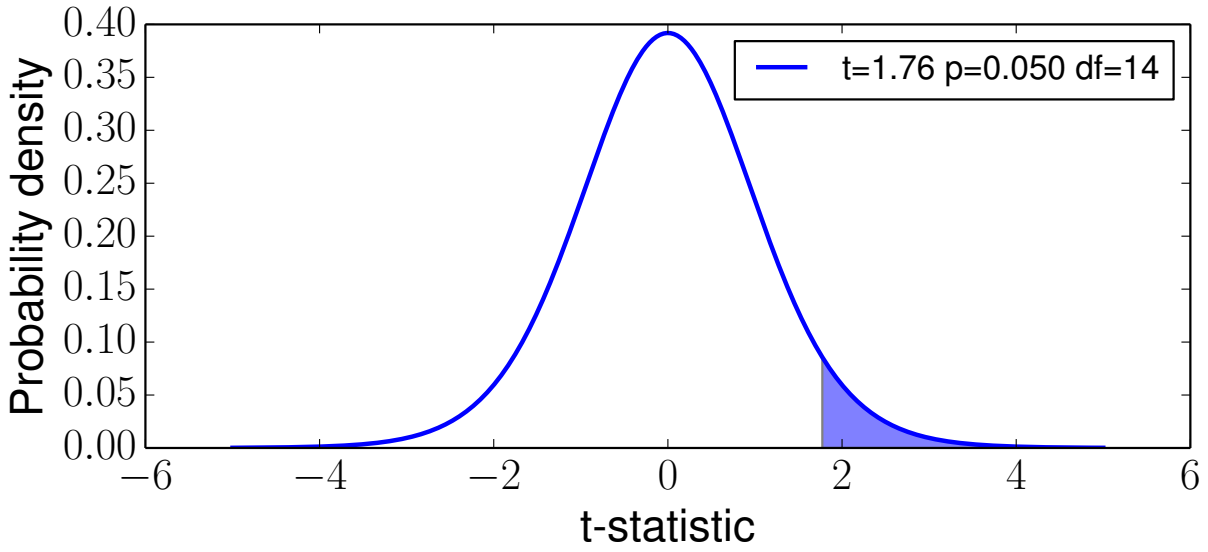


p-Value

1. Right-tailed one-sample t-test

School class **A**, consisting of 15 pupils, has an average grade of 2.5 (sample mean) and a sample standard deviation of 1.1 in their final exam. Did the pupils perform significantly different than the **nation-wide** average of 2.0?

- Null hypothesis H_0: “The pupils’ grades are lower than the nation’s”
- Calculate the one-sample t-statistic of “grade - 2.0” with 14 degrees of freedom
- Reject H_0 if the p-value $p < \alpha$, with $p = P(T > t)$ and significance level $\alpha = 5\%$

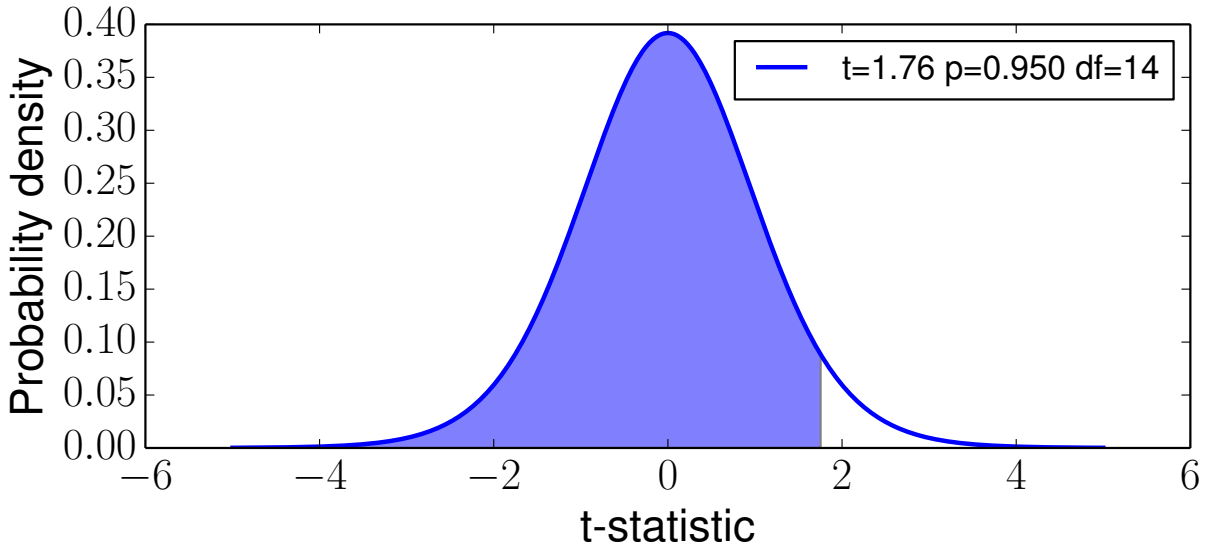


p-Value

1. Left-tailed one-sample t-test

School class **A**, consisting of 15 pupils, has an average grade of 2.5 (sample mean) and a sample standard deviation of 1.1 in their final exam. Did the pupils perform significantly different than the **nation-wide** average of 2.0?

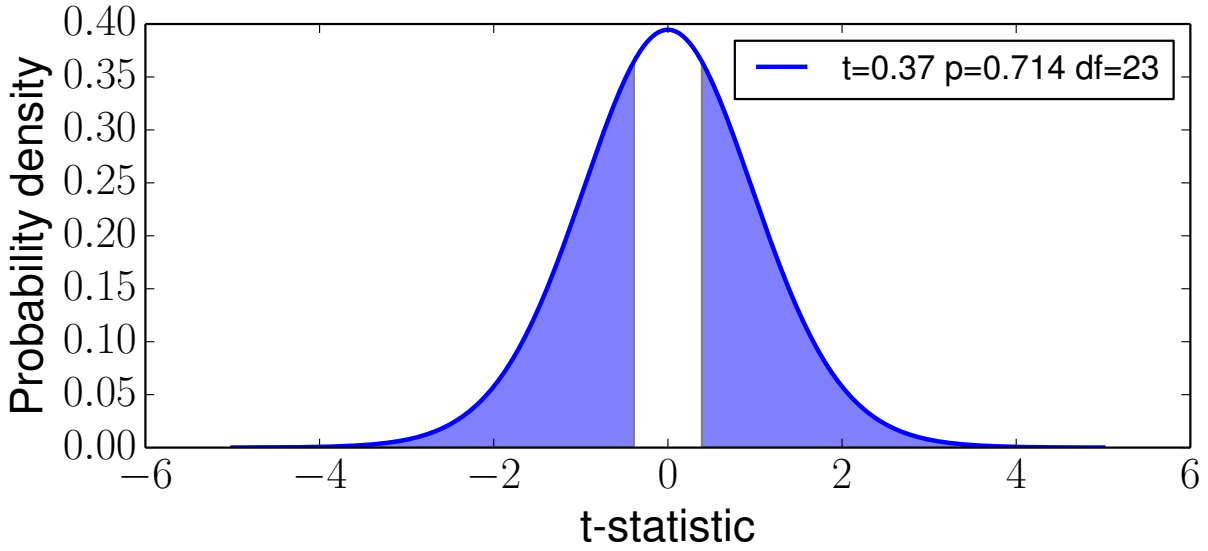
- Null hypothesis H_0: “The pupils’ grades are higher than the nation’s”
- Calculate the one-sample t-statistic of “grade - 2.0” with 14 degrees of freedom
- Reject H_0 if the p-value $p < \alpha$, with $p = P(T < t)$ and significance level $\alpha = 5\%$



p-Value

2. Two-tailed two-sample significance test

Did school class **A**, consisting of 15 pupils with an average grade of 2.5, perform significantly different from school class **B**, consisting of 10 pupils with an average grade of 1.5 (the combined sample standard deviation of **A** and **B** is 1.1)?

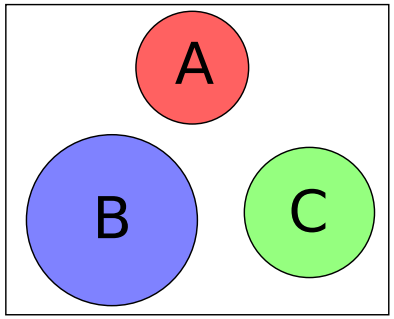


Confidence Intervals

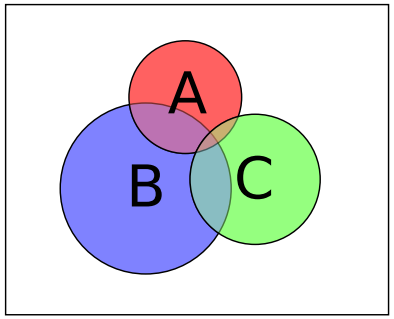
Lower (c_l) and upper (c_u) value of confidence interval

$$c_l = \min_c \{c | P(T > t(x - c)) < \alpha/2\}$$
$$c_u = \max_c \{c | P(T < t(x - c)) < \alpha/2\}$$

Divide significance levels by 2 because the upper and lower type I error events are disjoint.



$P(A \cup B \cup C) = P(A) + P(B) + P(C)$
Disjoint events



$P(A \cup B \cup C) < P(A) + P(B) + P(C)$
Overlapping events

Venn diagram of the sample space and associated events

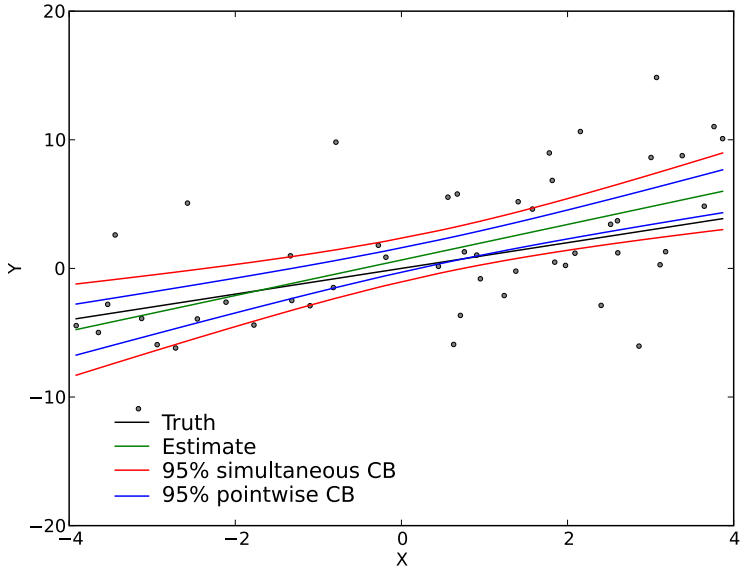
Confidence Intervals

Lower (c_l) and upper (c_u) value of confidence interval

$$c_l = \min_c \{c | P(T > t(x - c)) < \alpha/2\}$$

$$c_u = \max_c \{c | P(T < t(x - c)) < \alpha/2\}$$

Extension: simultaneous and pointwise confidence bands



Source: Wikipedia “Confidence and prediction bands”

Probability of a Type I Error

- The **two-tailed t-test** and the **confidence intervals** have a probability of **exactly** α to do a type I error under H_0
- The **one-tailed** test has a probability of **at most** α to do a type I error under H_0. The maximum type I error is reached if $\mu = 0$.

Paired Test

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

If observations are paired, e.g. each patient was measured before and after treatment, then those observations are no more statistically independent and a two-sample t-test can not be performed. However, in such a situation a paired test can be done.

Paired test by pairwise subtraction

Let $y_{b,i}$ be the measurement of the i-th subject before the treatment and $y_{a,i}$ be the measurement of the same subject after the treatment. Then perform a one-sample t-test on the differences $y_{a,i} - y_{b,i}$.

Nonparametric Tests

If the data is not normally distributed, nonparametric tests can be used

- Mann–Whitney U test
- Kruskal-Wallis non-parametric ANOVA
- Permutation tests
- Wilcoxon signed-rank test (paired test)

Multiple Comparison Procedures

Consider the following null hypothesis test.

- H_0: It is not Christmas today.
- Test statistic: Roll two dice. If both come up six, it is Christmas.

H_0 is rejected with a probability of $1/36 \approx 2.8\% < 5\%$.

The test seems useless for two reasons.

Multiple Comparison Procedures

Consider the following null hypothesis test.

- H_0: It is not Christmas today.
- Test statistic: Roll two dice. If both come up six, it is Christmas.

H_0 is rejected with a probability of $1/36 \approx 2.8\% < 5\%$.

The test seems useless for two reasons.

- It has a low statistical power (probability that the test will reject the null hypothesis when the alternative hypothesis is true).
  This talk will not discuss statistical power.

Multiple Comparison Procedures

Consider the following null hypothesis test.

- H_0: It is not Christmas today.
- Test statistic: Roll two dice. If both come up six, it is Christmas.

H_0 is rejected with a probability of $1/36 \approx 2.8\% < 5\%$.

The test seems useless for two reasons.

- It has a low statistical power (probability that the test will reject the null hypothesis when the alternative hypothesis is true).
  This talk will not discuss statistical power.
- One thinks “So can I just roll the dice multiple times until they show up 6, and then it is Christmas?”.
  The answer to this question is “no”, which is addressed by multiple comparison procedures.

Multiple Comparison Procedures

Real-world examples

- Which out of 130 regions of interest inside the brain show activation in an fMRI experiment using 30 subjects?
- Which out of 100,000 voxels show activation in an fMRI experiment using 30 subjects?
- Which out of 10,000 genes differ in their expression level in 100 different persons?

Bonferroni Correction

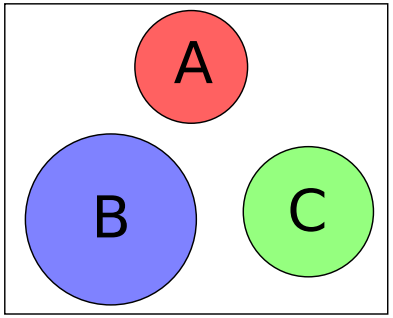
We perform m null hypothesis tests, V is the number of false positives (type I errors) of those m tests. The familiywise error rate (FWER) is defined as

$$\text{FWER} = \Pr(V \geq 1)$$

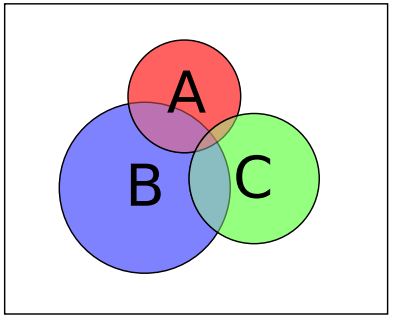
- Assumption: m possible type I errors mutually exclusive (worst case, see the figure)
- $\Rightarrow$ FWER, the probability of at least one type I error out of the m observations, is $m \cdot \alpha$
- $\Rightarrow$ Threshold not at α, but at α/m

Drawbacks

- Overestimates FWER because it assumes mutually exclusive type I errors
- FWER might not be the desired error measure (see false discovery rate on next slide)



$P(A \cup B \cup C) = P(A) + P(B) + P(C)$
Disjoint events



$P(A \cup B \cup C) < P(A) + P(B) + P(C)$
Overlapping events

Venn diagram of the sample space and associated events

False Discovery Rate

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Definition

$$\text{FDR} = Q_e = \mathrm{E}[Q] = \mathrm{E}\left[\frac{V}{R}\right]$$

where

- $\frac{V}{R}$ is defined to be 0 when R = 0
- V is the number of false positives (type I errors, also called “false discoveries”)
- R is the number of rejected null hypotheses (also called “discoveries”)

FDR control using the Benjamini-Hochberg procedure (BH step-up procedure)

- Order the p-values in increasing order and denote them by $P_1 \ldots P_m$
- For a given significance level α, find the largest k such that $P_k \leq \frac{k}{m}\alpha$
- Reject (i.e. declare positive discoveries) all $H_{(i)}$ for $i = 1, \ldots, k$

The BH procedure is valid when the m tests are independent, and also in various scenarios of dependence.

Thank you for your attention.