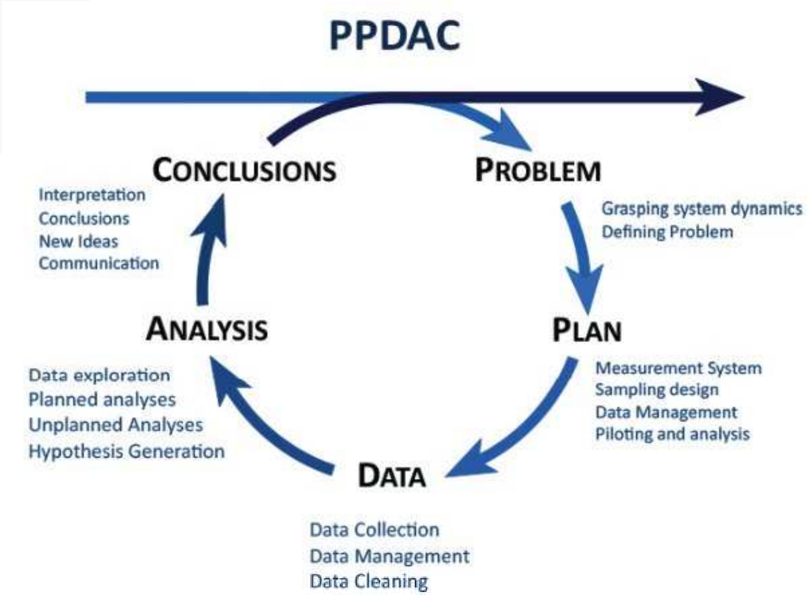
CEITEC
Central European Institute of Technology
BRNO | CZECH REPUBLIC

Introduction to Bioinformatics (LF:DSIB01)

Week 8 : Statistics

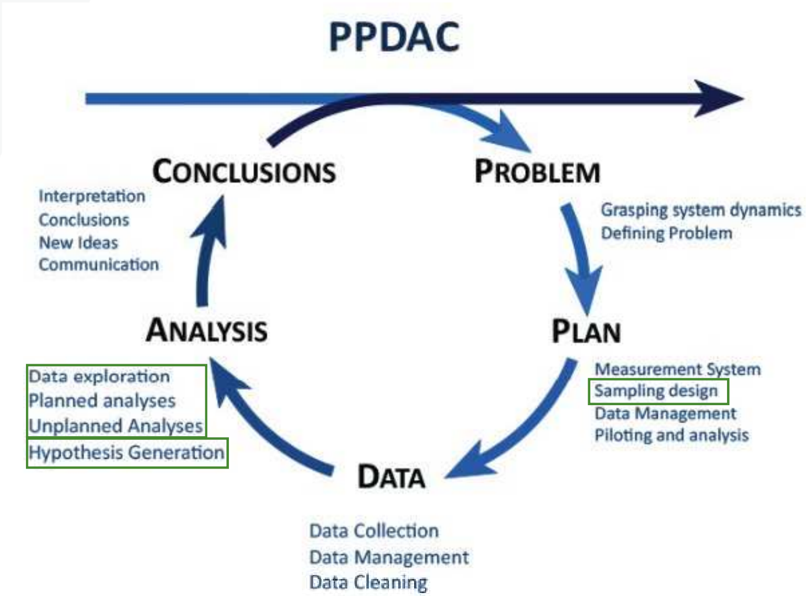
What is statistics?

- Statistics is the science of learning from **data**, and of measuring, controlling and communicating **uncertainty**. - American Statistical Association (ASA)



What is statistics?

- Statistics is the science of learning from **data**, and of measuring, controlling and communicating **uncertainty**. - American Statistical Association (ASA)



Sampling Design

- Randomization
  - Sampling designs should be as random as possible
- Overrepresentation
  - Preferentially select units where the dispersion is larger
  - Sample is not necessarily a “scale copy” of population
  - It makes sense to increase depth in categories that are more variable
- Restriction
  - Should restrict or exclude problematic samples such as samples with empty categories
  - Stratification: fixing the sample size in categories of the population
  - Is not in contrast with Randomization as long as there are enough possible samples

Statistical Hypothesis Testing

To ask questions on data

we use statistical methods that provide

a confidence or likelihood about the answers.

Null Hypothesis: H_0: The default position that there is nothing new happening

How can we test our confidence in the Null Hypothesis?

Statistical Hypothesis Testing

Usual goal:

Reject Null Hypothesis with some confidence (0.05)

Confirm statistically significant effect.

Example

You can’t confirm null Null Hypothesis!

Frequentist (classical) vs. Bayesian statistics

A Probability value can be thought of in several ways:

1. Long-term frequency
2. Degree of belief
3. Degree of logical support

Frequentist Statistics works with (1) while Bayesian Statistics with (2 and 3)

Frequentist Statistics

- Only repeatable random events (like the result of flipping a coin) have probabilities.
- These probabilities are equal to the long-term frequency of occurrence of the events in question.
- Cannot apply probabilities to hypotheses or to any fixed but unknown values in general

Bayesian Statistics

- Probabilities can represent the uncertainty in any event or hypothesis
- Newly collected data narrows down the probability distribution over the parameter.

Example:

Doctor knows that 20% of the population has Disease
Test that shows + for 90% of Disease individual
but also shows + for 30% of Healthy individual

A patient comes in the Doctor's office and takes the test.

What is the probability that the patient has Disease?

The patient takes Test and the test comes back +

What is the probability that the patient has Disease?
(given that the Test showed +)

Frequentist Statistics

- Only repeatable random events (like the result of flipping a coin) have probabilities.
- These probabilities are equal to the long-term frequency of occurrence of the events in question.
- Cannot apply probabilities to hypotheses or to any fixed but unknown values in general

Bayesian Statistics

- Probabilities can represent the uncertainty in any event or hypothesis
- Newly collected data narrows down the probability distribution over the parameter.

Example:

Doctor knows that 20% of the population has Disease
Test that shows + for 90% of Disease individual
but also shows + for 30% of Healthy individual

A patient comes in the Doctor's office and takes the test.

What is the probability that the patient has Disease?

20% (1:4)

The patient takes Test and the test comes back +

What is the probability that the patient has Disease?
(given that the Test showed +)

43% (3:7)

How?

Answered in Lecture 10: Bayesian Inference

$$\mathbb{P}(X \mid Y) := \frac{\mathbb{P}(X \wedge Y)}{\mathbb{P}(Y)}$$

Response vs. explanatory variable

- Dependent vs. measured variable
- Example

- What if variables are independent.
  - Correlation
- What if it is a chicken egg scenario.
  - select it to fit your model/test

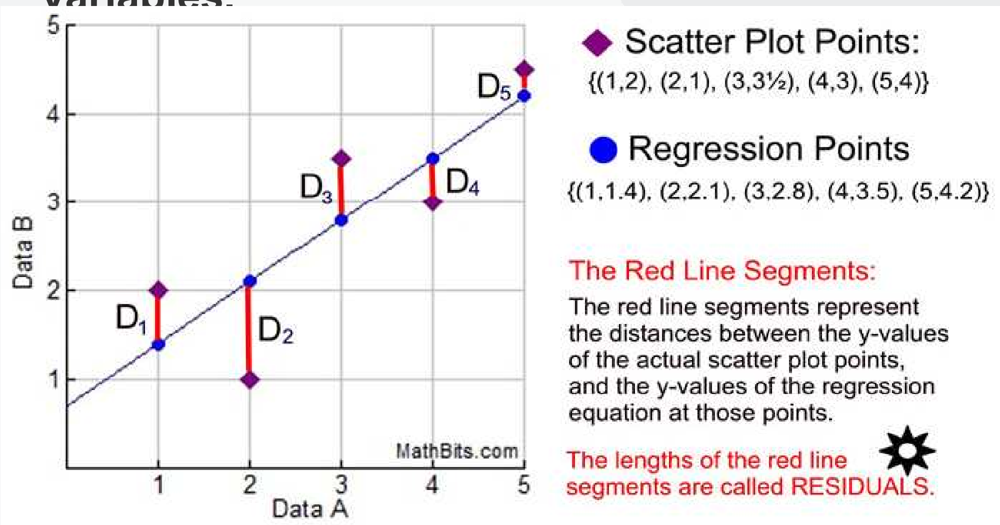
- Why do we care?
  - Formula in R

Response vs. explanatory variable

Response variable type	Explanatory variable type	Example tst type
Categorical	Categorical	Fisher test
Categorical (two groups)	Continuous	t-test
Categorical (multiple groups)	Continuous	ANOVA
Continuous	Continuous	Linear regression
Continuous	Categorical (two groups)	Logistic regression

Linear regression

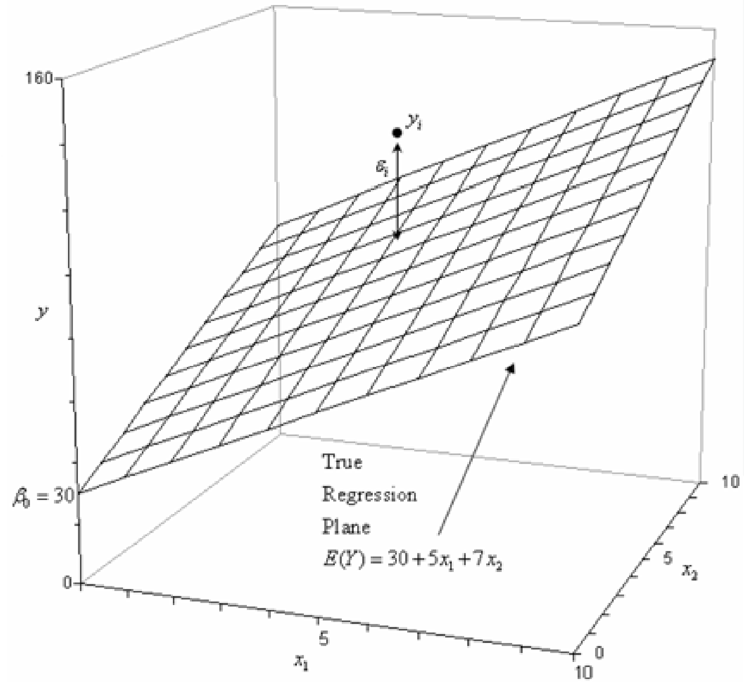
Simple linear regression is used to model the relationship between two continuous variables.



$$Y = \beta_0 + \beta_1 X_1$$

Multiple linear regression is used when we we have multiple explanatory variables.

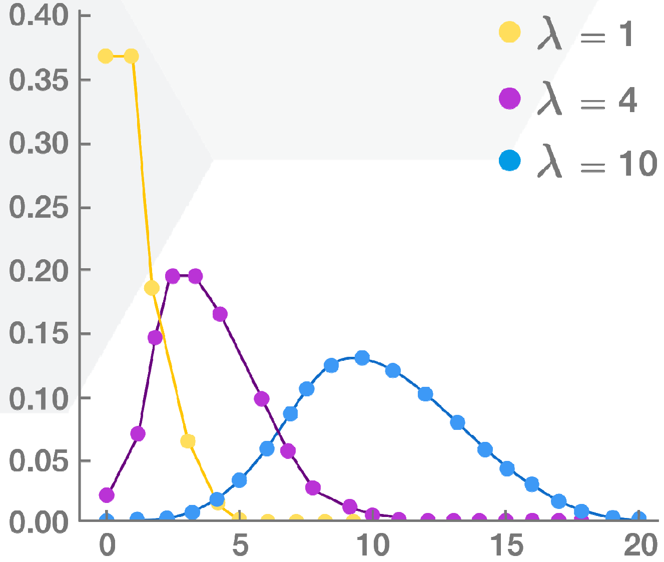
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

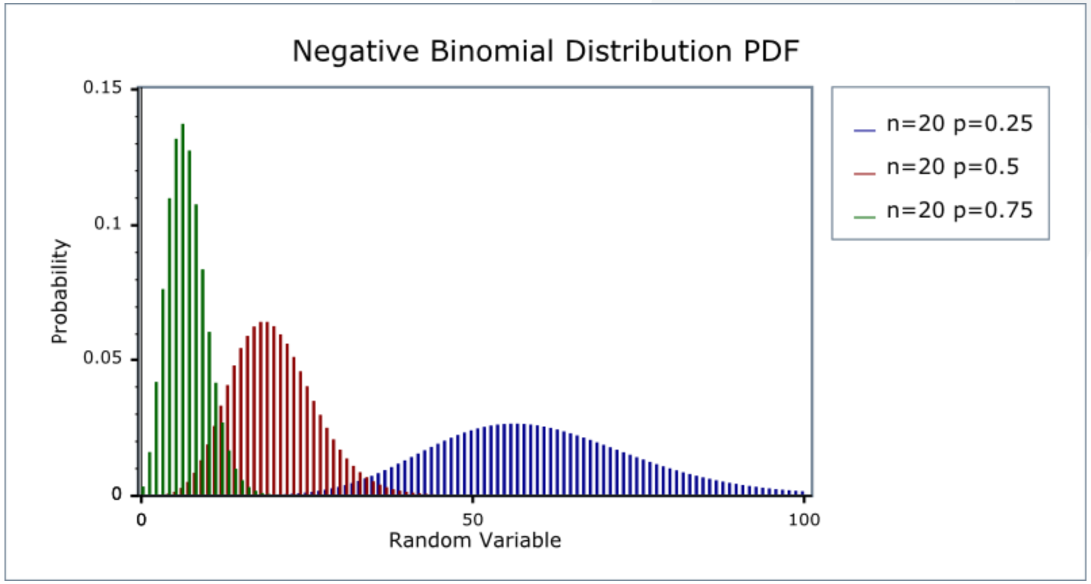


Beware - overfitting!

Nonlinear regression

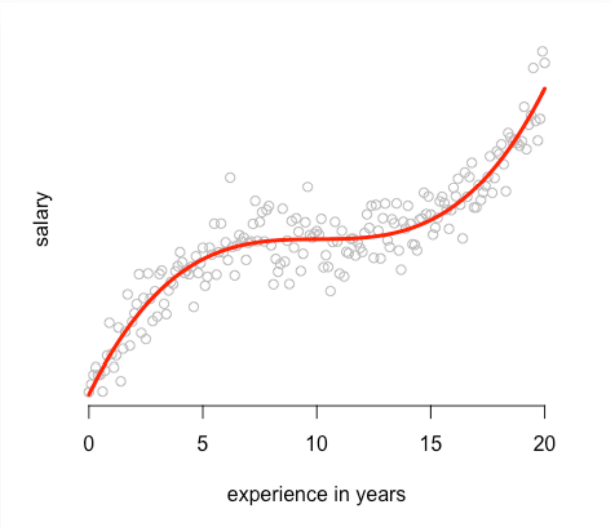
- Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear. ;-)
- Polynomial regression
- Poisson distribution





Nonlinear regression

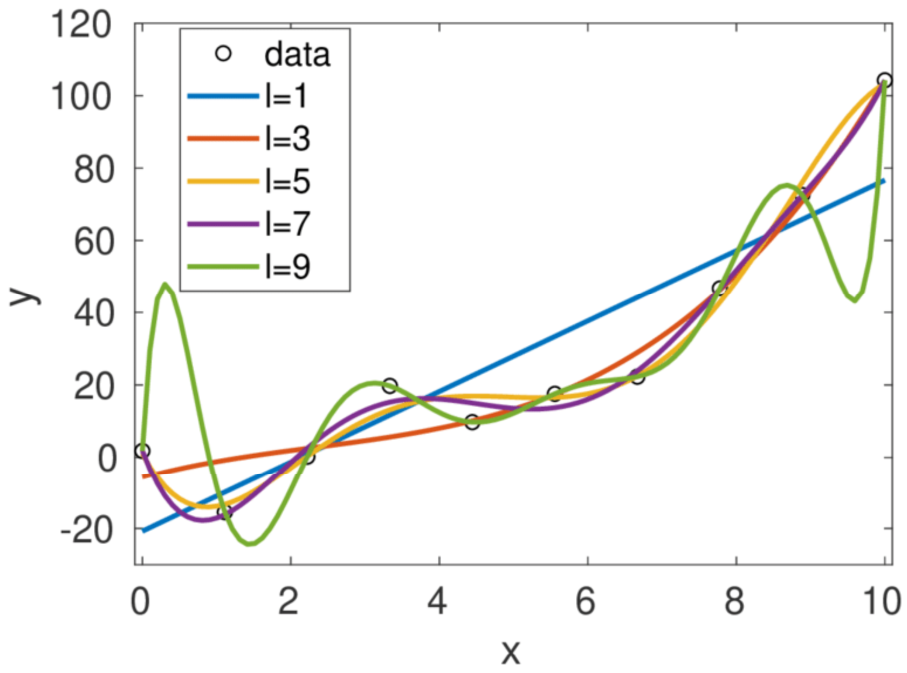
- Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear. ;-)
- Polynomial regression



Simple Linear Regression	$y = b_0 + b_1 x_1$
Multiple Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$
Polynomial Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_1^2 + \dots + b_n x_1^n$

Polynomial regression

- Overfitting



Parametric vs. non-parametric tests

Parametric tests assume underlying statistical distributions in the data. Therefore, several conditions of validity must be met so that the result of a parametric test is reliable. For example, Student's t-test for two independent samples is reliable only if each sample follows a normal distribution and if sample variances are homogeneous.

Nonparametric tests do not rely on any distribution. They can thus be applied even if parametric conditions of validity are not met. They are generally weaker.

Parametric test	Non-Parametric equivalent
Paired t-test	Wilcoxon Rank sum Test
Unpaired t-test	Mann-Whitney U test
Pearson correlation	Spearman correlation
One way Analysis of variance	Kruskal Wallis Test

Wilcoxon rank-sum test (Mann–Whitney U test)

- Comparing medians of two population
- No assumptions about the populations – but data must be ordinal
  - Beware of ties
- One population can be sub-sample of the other
  - Does selected genes have a higher expression?

	id	gender	ad1
1	1	Female	94
2	4	Female	92
3	5	Female	93
4	2	Male	92
5	6	Male	49
6	7	Male	53

MANN-WHITNEY TEST

Population Mean Ranks Equal?

1 metric / ordinal outcome variable
2 groups of cases

Kruskal–Wallis test

	group	outcome
1	1	7
2	1	2
3	1	3
4	2	4
5	2	8
6	2	6
7	3	1
8	3	9
9	3	5

KRUSKAL-WALLIS TEST

Population Mean Ranks Equal?

1 metric / ordinal outcome variable on 3(+) groups

Kruskal-Wallis Test: Overall Satisfaction versus Customer Type

Descriptive Statistics

Customer Type	N	Median	Mean Rank	Z-Value
1	31	3.56	36.0	-3.34
2	42	4.34	65.9	4.53
3	27	3.51	43.1	-1.56
Overall	100		50.5	

Test

Null hypothesis H_0: All medians are equal

Alternative hypothesis H_1: At least one median is different

Method	DF	H-Value	P-Value
Not adjusted for ties	2	21.36	0.000
Adjusted for ties	2	21.36	0.000

Wilcoxon signed-rank test

- Paired data points

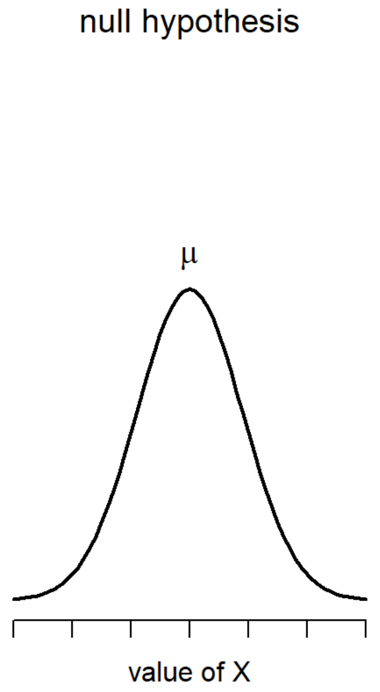
Population Distributions Equal?

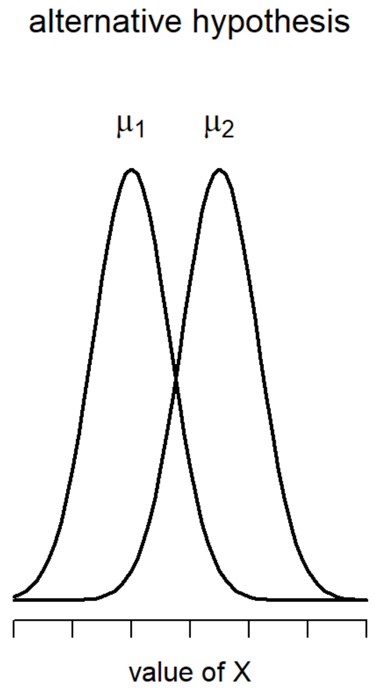
WILCOXON SIGNED-RANKS TEST

	id	score_1	score_2	var	var	var	var	var	va
1	1	8	10						
2	2	5	4						
3	3	7	6						
4	4	9	5						
5	5	8	8						

2 metric / ordinal outcome variables
1 group of cases

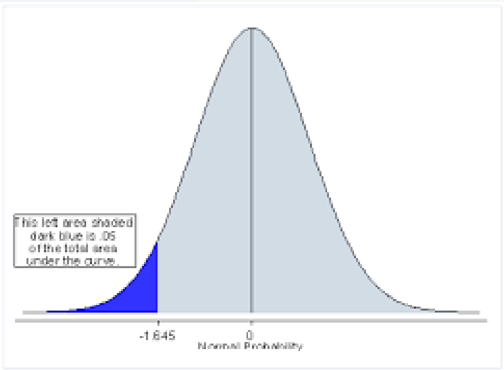
Student's t-test

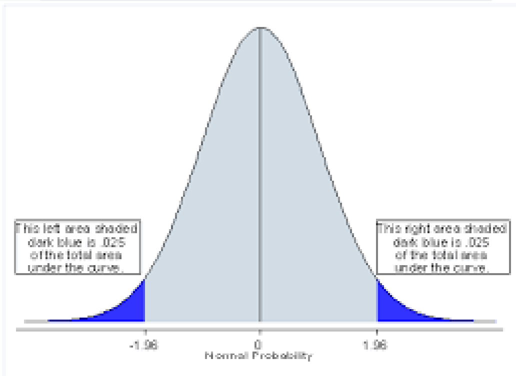




Student's t-test

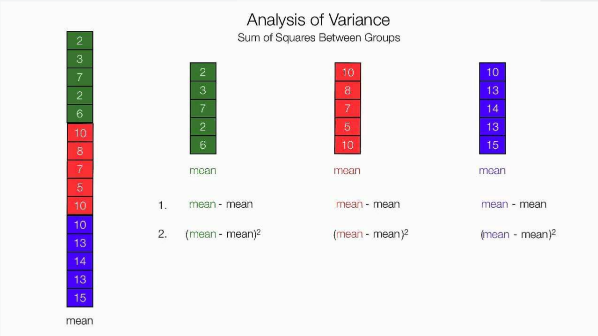
- One sided vs. two sided





- Also paired version

ANOVA



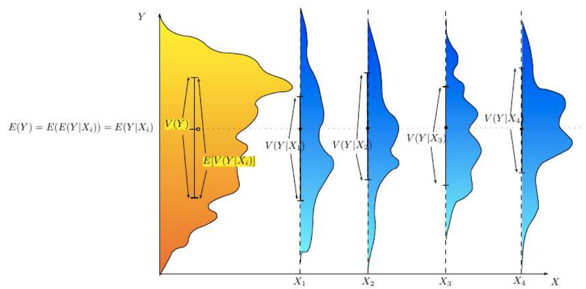


Figure 2: ANOVA : No fit

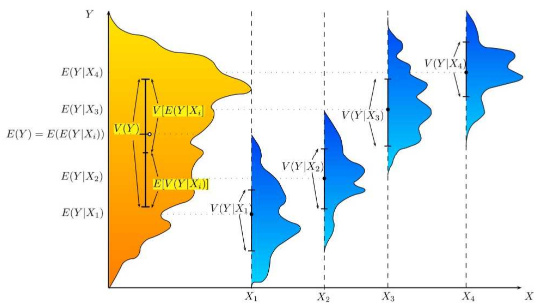


Figure 1: ANOVA : Fair fit

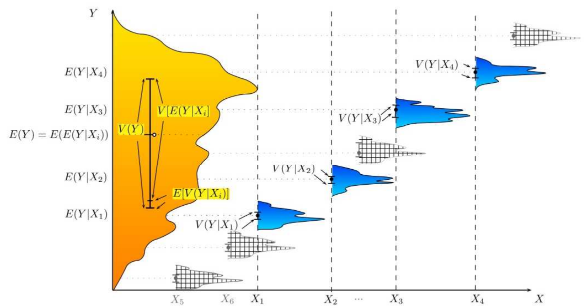


Figure 3: ANOVA : very good fit

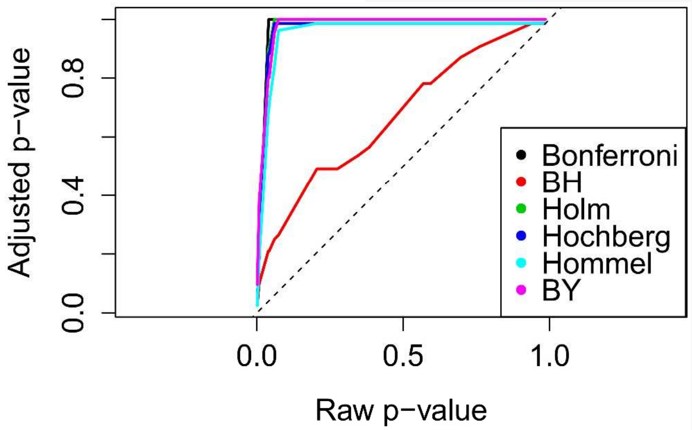
p-value

- *Many researchers in various areas use standard routines in statistical software in the expectation that the software can condense their research into a single summary (most often a p-value) that ‘objectifies’ their results. This idea of objectivity is in stark contrast with the realization by many of these researchers at some point that depending on individual inventiveness there are many ways to arrive at such a number.”*

p-value adjustment

- Multiple testing problem
  - Minimize false positive error rate

i	p(i)	(K-i+1)	(K-i+1)p(i)	Holm adjustment max(p*(1),p*(2),... ,p*(i))	Bonferroni adjustment 10*p(i)
1	0.0002	10	0.0020	0.0020	0.0020
2	0.0011	9	0.0099	0.0099	0.0110
3	0.0012	8	0.0096	0.0099	0.0120
4	0.0015	7	0.0105	0.0105	0.0150
5	0.0022	6	0.0132	0.0132	0.0220
6	0.0091	5	0.0455	0.0455	0.0910
7	0.0131	4	0.0524	0.0524	0.1310
8	0.0152	3	0.0456	0.0524	0.1520
9	0.0311	2	0.0622	0.0622	0.3110
10	0.1986	1	0.1986	0.1986	1.0000



CEITEC

@CEITEC_Brno

Thank you for your attention!
60 minutes lunch break.

www.ceitec.eu