UNIVERSITÄT BERN

http://www.math.unibe.ch/continuing_education/cas_applied_data_science/index_eng.html

CAS Applied Data Science - Module 2 – Day 3

Statistical Inference for Data Science

Prof. Dr. Géraldine Conti, PD Dr. Sigve Haug

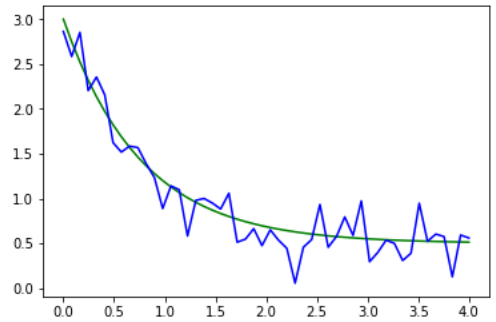
Bern, 2019-08-29

Discussion Session

- Review of Notebook 3
- Questions from the chat
- Summary
- The online book covering most of Module 2 and more : **Think Stats 2e**

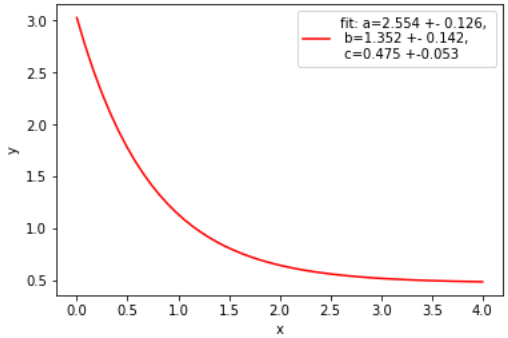
Review of Notebook 3

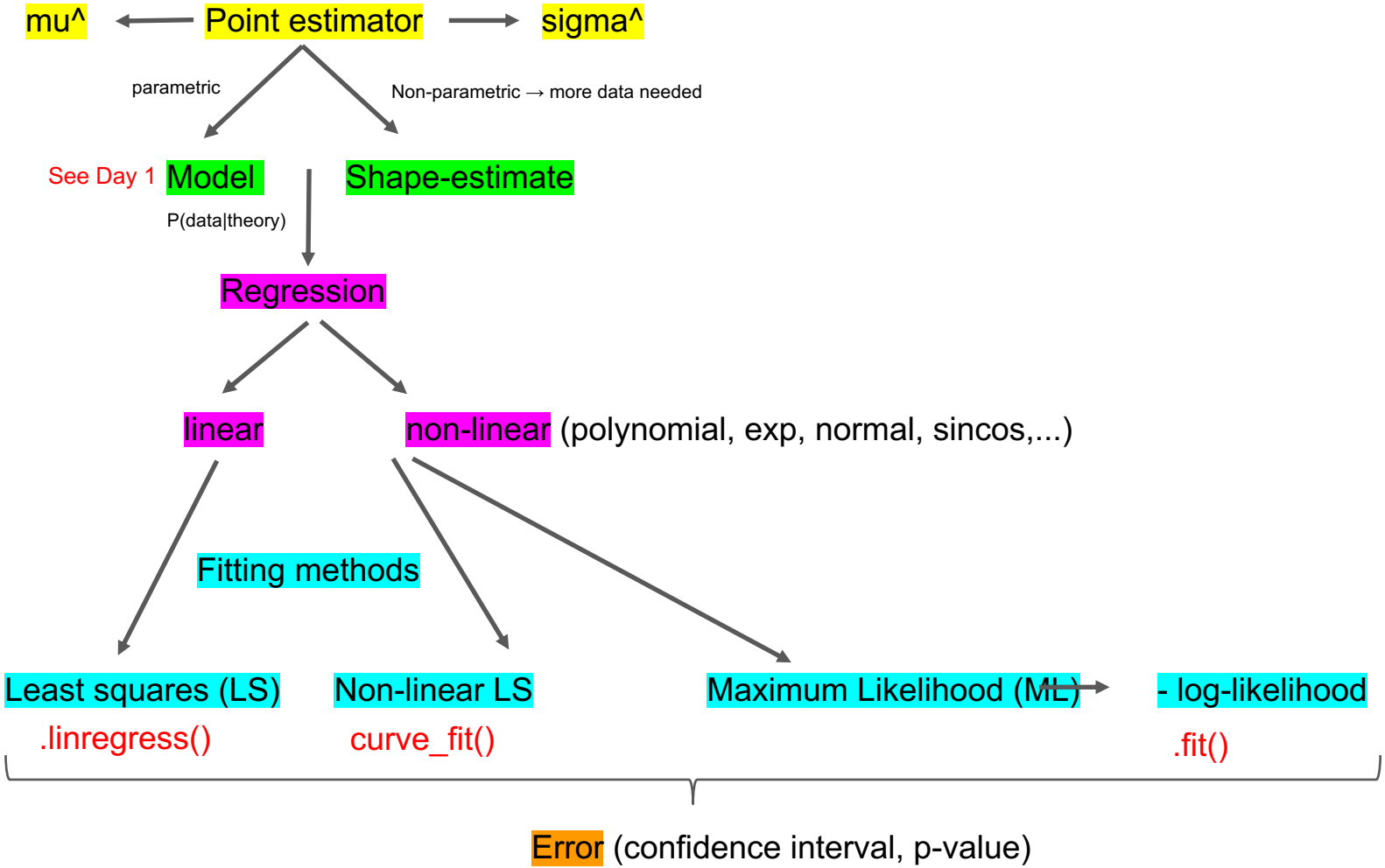
In [15]:
```
from scipy.optimize import curve_fit

def func(x, a, b, c):
     return a * np.exp(-b * x) + c

xdata = np.linspace(0, 4, 50) #
y = func(xdata, 2.5, 1.3, 0.5)
plt.plot(xdata, y, 'g-', label='Generated data')
np.random.seed(1729)
y_noise = 0.2 * np.random.normal(size=xdata.size)
ydata = y + y_noise
plt.plot(xdata, ydata, 'b-', label='Generated data with noise')
plt.show()
```



In [16]:
```
popt, pcov = curve_fit(func, xdata, ydata)
print(popt)
perr = np.sqrt(np.diag(pcov)) # Standard deviation = square root of the varianc
plt.plot(xdata, func(xdata, *popt), 'r-',label= \
         'fit: a=%5.3f +- %5.3f, \n b=%5.3f +- %5.3f, \n c=%5.3f +-%5.3f' % \
         (popt[0],perr[0],popt[1],perr[1],popt[2],perr[2]))
plt.xlabel('x')
plt.ylabel('y')
plt.legend()
plt.show()
perr = np.sqrt(np.diag(pcov)) # Standard deviation = square root of the varianc
perr
```

```
[2.55423706 1.35190947 0.47450618]
```




mu^
Point estimator
sigma^
parametric
Non-parametric → more data needed
See Day 1
Model
Shape-estimate
P(data|theory)
Regression
linear
non-linear (polynomial, exp, normal, sincos,...)
Fitting methods
Least squares (LS)
.linregress()
Non-linear LS
curve_fit()
Maximum Likelihood (ML)
- log-likelihood
.fit()
Error (confidence interval, p-value)

3rd day : Hypothesis Testing

Introduction

- Hypotheses and tests
- Error types

Frequent tests

- Normality tests
- One-sample test
- Comparing two samples

Take home

Lorem ipsum dolor sit amet, consectetur adipiscing elit

Hypothesis Testing

Topics in Statistics edit · view

General topics	Probability	Descriptive statistics	Inferential statistics	Specialized topics
• Levels of measurement • Sampling • Statistical survey • Design of experiments • Data analysis • Statistical graphics • History of statistics	• Probability theory • Random variable • Probability distribution • Independence • Expected value • Variance, covariance • Central limit theorem	• Averages • Statistical dispersion • Summary statistics • Skewness • Correlation • Frequency distribution • Contingency table	• Hypothesis testing • Estimator • Maximum likelihood • Bayesian inference • Non-parametric statistics • Analysis of variance • Regression models	• Computational statistics • Decision theory • Multilevel models • Multivariate statistics • Statistical process control • Survival analysis • Time series analysis

Introduction

Traumatic brain injury causes millions of neurons to become hyperactive and this damages the neurons. A drug company invents a medication that suppresses this process. Will the drug work as an effective treatment for Traumatic brain injury ?

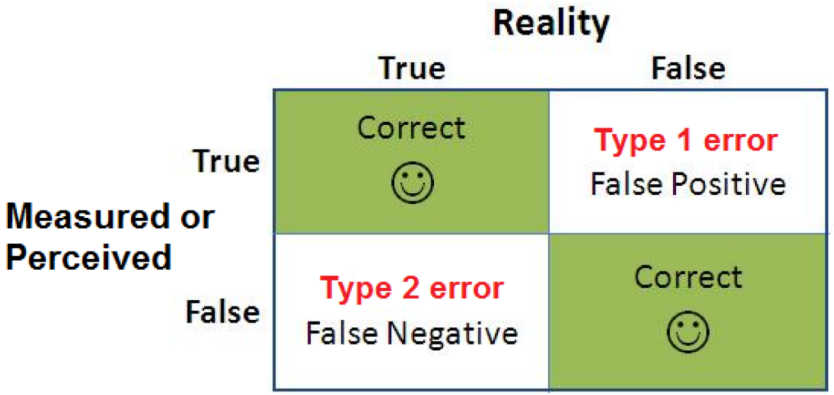
- The drug is not effective (Null hypothesis (H0)) ← *hypothesis to be tested*
- The drug reduces the damage and the patient will have a better recovery. The drug is effective (Alternative hypothesis (H1)).

Statistical test : test statistic of a sample is calculated and the p-value is given under the H0 assumption.

Probability of obtaining such a sample if H0 is true

Errors Types

- Type 1 : Reject the null hypothesis due to a fluctuation (*false positive*)
- Type 2 : Keep the null hypothesis by interpreting a real effect as a fluctuation (*false negative*)
- **Example** : Guilty vs Innocent, jailed/set free



An innocent person is set free	An innocent person is jailed
A guilty person is set free	A guilty person is jailed

Data to be tested

1 sample

- Normal
- Symmetric
- ...

Tests can take advantage of relations (cancellations in divisions)

≥2 samples (columns/sets/samples)

Paired

- Dependent
- Repeated measurements on the *same object/individual*

Unpaired

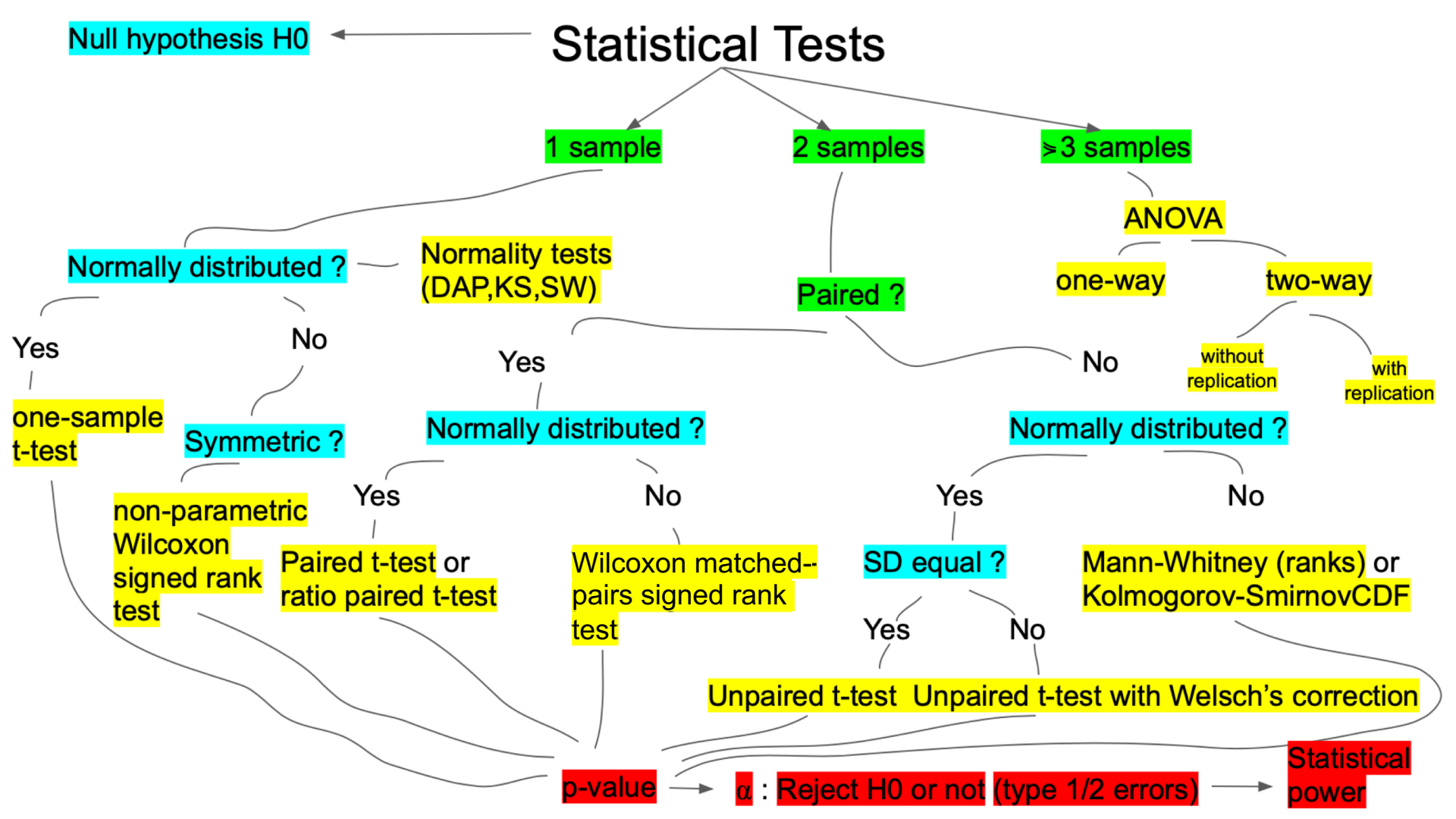
- Independent
- From *separate individuals*

Measure blood pressure from a group of patients before/after giving a medicine	Measure blood pressure from a group of patients who were given a medicine and from another group not given it

Normality tests

Quantify how close a sample of values are to the normal distribution

- Answers with a p-value to the question *“If you randomly sample from a Gaussian population, what is the probability of obtaining a sample that deviates from a Gaussian distribution as much (or more so) as this sample does?”*
- May not work very well for small samples (<10-20)


Null hypothesis H0
Statistical Tests
1 sample
2 samples
⩾3 samples
Normally distributed ?
Normality tests (DAP,KS,SW)
Paired ?
ANOVA
one-way
two-way
Yes
No
Yes
No
without replication
with replication
one-sample t-test
Symmetric ?
Normally distributed ?
Normally distributed ?
Yes
No
Yes
No
Yes
No
non-parametric Wilcoxon signed rank test
Paired t-test or ratio paired t-test
Wilcoxon matched-pairs signed rank test
SD equal ?
Mann-Whitney (ranks) or Kolmogorov-SmirnovCDF
Yes
No
Unpaired t-test
Unpaired t-test with Welsch's correction
p-value
α : Reject H0 or not (type 1/2 errors)
Statistical power

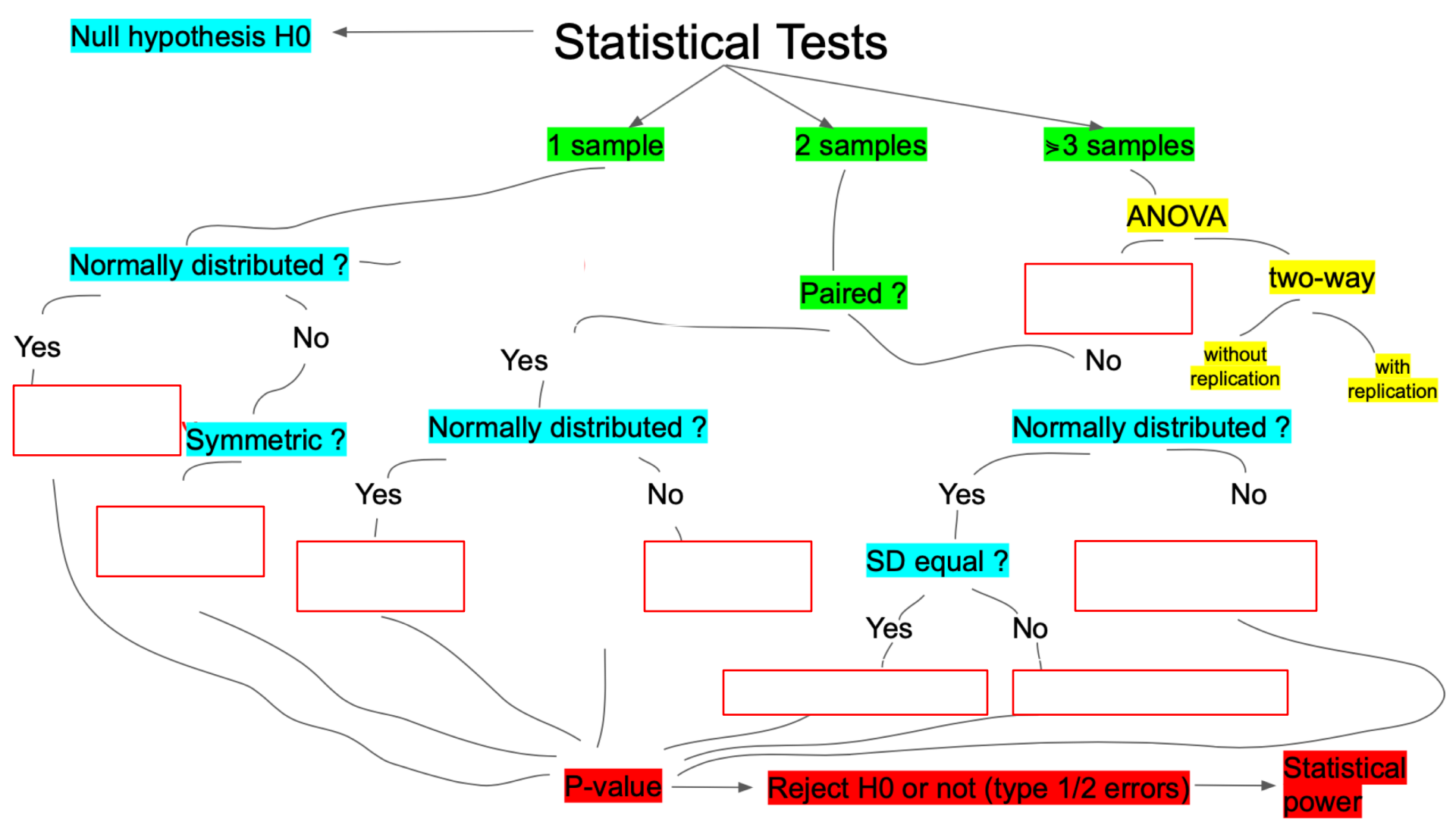
Exercise

- 3 slides (<team-nr>.pdf) to be uploaded to Ilias by 4pm today:
  - 1 slide : Question that the test tries to answer, assumptions to be able to use the test, other details
  - 1 slide : example from “real life” (*provide reference*)
  - 1 slide : your conclusions from the Notebook on this test
  - 1 question to another group
- Will be presented at tomorrow’s discussion session

	Statistical Test to study	Question to team :	Team (by random break out rooms)
1	One-sample t-test	4	
2	Non-param Wilcoxon signed rank test	3	
3	Paired t-test	7	
4	Wilcoxon matched-pairs signed rank test	5	
5	Unpaired t-test	8	
6	Unpaired t-test with Welsch’s correction	2	
7	Mann-Whitney rank test	1	
8	One-way ANOVA	6	

Summary

Statistical tests


Null hypothesis H0
Statistical Tests
1 sample
2 samples
⩾3 samples
Normally distributed ?
Paired ?
ANOVA
two-way
Yes
No
Yes
No
without replication
with replication
Symmetric ?
Normally distributed ?
Normally distributed ?
Yes
No
Yes
No
SD equal ?
Yes
No
P-value
Reject H0 or not (type 1/2 errors)
Statistical power

Next self study with Notebooks

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit
2. Sed do eiusmod tempor incididunt ut labore
3. Ut enim ad minim veniam, quis nostrud exercitation