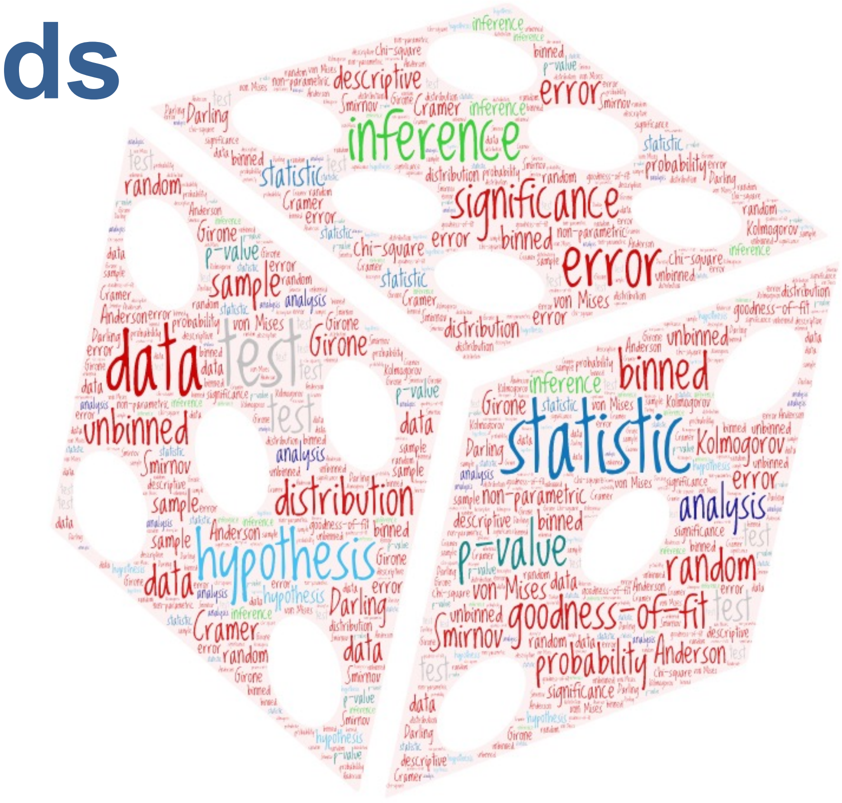
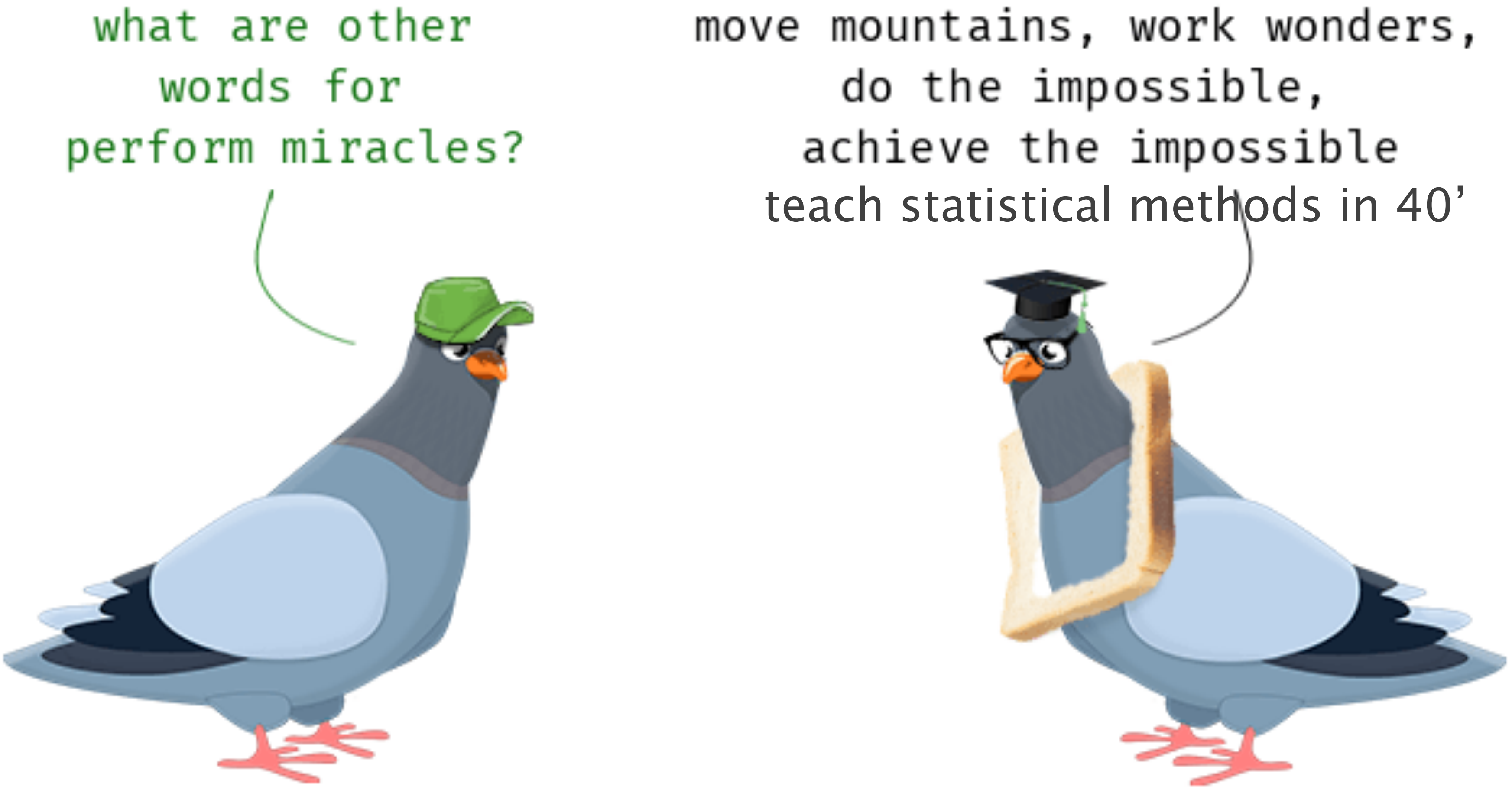
Joint ICTP-IAEA Advanced School/Workshop on
Computational Nuclear Science and Engineering
ICTP, Trieste, 23-27 May 2022

Statistical methods for data analysis

Maria Grazia Pia
INFN Genova, Italy



Acknowledgements: Stefano Bonnini (*Univ. Ferrara),* Elisabetta Ronchieri *(INFN CNAF)*


what are other words for perform miracles?
move mountains, work wonders, do the impossible, achieve the impossible
teach statistical methods in 40’

Statistics?

- Descriptive statistics: characterize the data *often complemented by graphics*
- Statistical inference: draw conclusions from the data

Point estimation, regression, correlation, parameter fitting, hypothesis testing, reliability testing, signal processing, Monte Carlo calculations…

Basic concepts, methods and tools
to compare data distributions

- Validate simulation/calculations w.r.t. experiment
- Evaluate compatibility between experimental data sets
- etc.

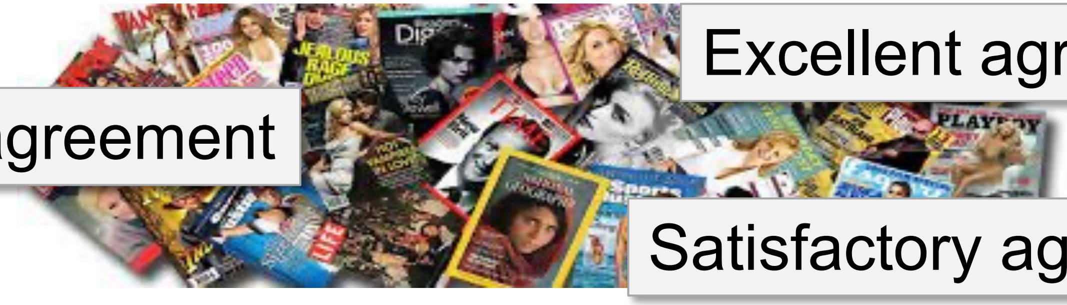
Practical examples and exercises
Suggestions for further reading

nuclear technology

Comparison of data distributions in the literature mainly rests on

- qualitative visual appraisal of figures
- indicators (%) devoid of statistical relevance

The result is often expressed as the authors' personal opinion

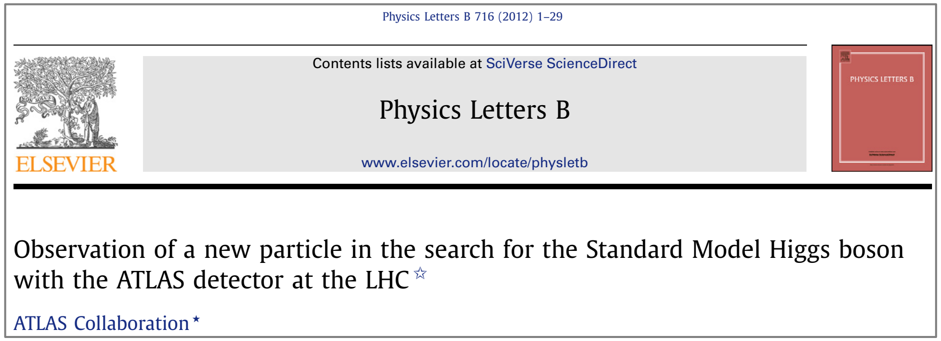


Good agreement

Excellent agreement

Satisfactory agreement

…unlike in fundamental physics literature



Physics Letters B 716 (2012) 1–29

Contents lists available at SciVerse ScienceDirect

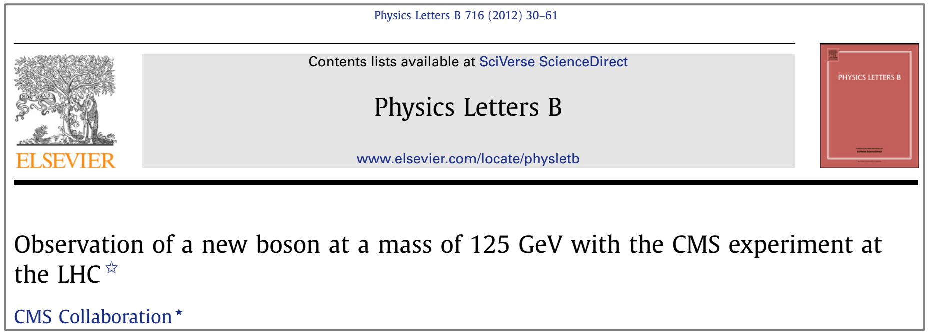
Physics Letters B

www.elsevier.com/locate/physletb

ELSEVIER

Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC ☆

ATLAS Collaboration *



Physics Letters B 716 (2012) 30–61

Contents lists available at SciVerse ScienceDirect

Physics Letters B

www.elsevier.com/locate/physletb

ELSEVIER

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC ☆

CMS Collaboration *

Outline of the lecture

Concepts

- Refresher of basic statistical definitions
- Hypothesis testing: concepts
- Goodness-of-fit testing
- Location-scale tests
- Testing categorical data
- Suggested reading

Practical

- Overview of statistical software packages
- Brief introduction to R
- Application examples (demo)

- If you install R on your PC/Mac, you can run the examples yourself

Hypothesis testing

- Statistical hypothesis testing is the conceptual and mathematical framework that supports data comparison
- It is a domain of **inferential statistics**

Statistical hypothesis

A claim or assertion about the probability function of one or more random variables

or

A statement about the populations from which one or more samples are drawn

e.g. its shape or parameter values

Random variable refresher

Random variable

Rolling a die: since prior to a throw, its outcome can not be predicted with complete certainty, the number of observed dots is called a **random variable**

Sample space

The collection of integer numbers between 1 and 6

Discrete random variable

Can only take on a **finite number of values**

Each possible outcome x_i of the experiment has a probability P_i: $\mathbf{P(X = x_i) = P_i}$

The sum of all P_i for all conceivable outcomes must be equal to one, $\sum P_i = 1$

Continuous random variable

Can have a **continuum of values** within any finite interval

$P\ (x \leq X \leq x+dx)$ is the probability of getting a value in the interval [x, x+dx]

Probability density function

for the continuous random variable X: $f(x)\ dx = P\ (x \leq X \leq x+dx)$

$\int_{\Omega} f(x)\, dx = 1$ where the integration goes over all possible outcomes x defining the sample space Ω

Null and alternative hypothesis

Null hypothesis H_0

The hypothesis under test

Alternative hypothesis H_1

The conclusion reached if the null hypothesis is rejected

A **test of a statistical hypothesis** is a rule that enables one to make a decision whether or not **H_0 should be rejected** on the basis of the observed value of a **test statistic**

The conceptual framework of hypothesis testing is based on **rejecting the null hypothesis**

The null hypothesis is not "accepted": it is either ***rejected*** *or* ***not rejected***

Not rejected could mean that there is not sufficient evidence for rejecting the hypothesis

The probability distribution of the test statistic when H_0 holds is referred to as the **null distribution of the test statistic**

Suppose, for instance, that we have measured the proper decay time of a sample of Ξ^0 hyperons and from these measurements estimated the Ξ^0 mean lifetime. Does this estimate agree with the prediction that the Ξ^0 lives twice as long as the Ξ^-? In other words, do the observations disprove the validity of the physics model?

Let τ_0: the mean lifetime of the Ξ^0 as implied by the model

τ: the mean lifetime indicated by our experiment

We want to test if τ is equal to τ_0 within the experimental errors:

H_0: $\tau = \tau_0$ is the **null hypothesis**

H_1: $\tau \neq \tau_0$ is the **alternative hypothesis**

Simple and composite hypothesis

Simple if the statement completely specifies the population

Composite otherwise

Suppose that we have two hypotheses completely specified
by two different values of a parameter θ entering a p.d.f. f(x|θ)
(x can be a directly observable quantity or a statistic)

$H_0: \theta = \theta_0$
$H_1: \theta = \theta_1$ Completely specified → **Simple** hypothesis

$H_1: \theta > \theta_1$ Not completely specified → **Composite** hypothesis

Statistic

Statistic

A function of one or more random variables that does not depend on any unknown parameter

Function of a sample

The function itself is **independent** of the sample distribution, i.e. the function can be stated before realization of the data

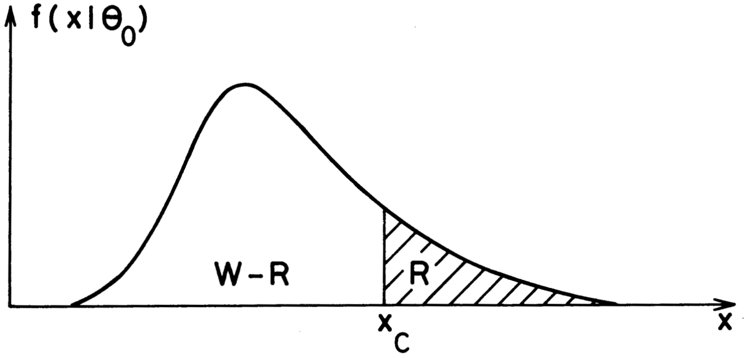
Statistic

The function

The value of the function on a given sample

A statistic is an observable random variable

Critical region



Assuming the null hypothesis H_0 to be true, we can find a region **R** in the sample space W for the observation x such that the probability that x belongs to R is equal to any preassigned numerical value

The region R is called the region of rejection or the **critical region** for H_0

If the observed value x_{obs} of the test statistic falls in R (i.e. x_{obs} exceeds x_c), we shall reject H_0

The preassigned probability α that the observation x will belong to the region R is called the **significance** of the test

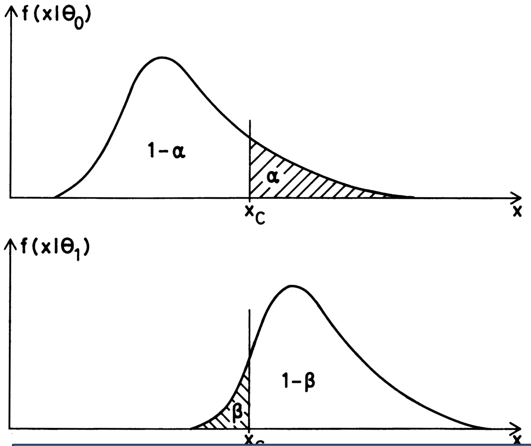
A result is "significant" if the probability that it could have arisen by chance from the null hypothesis is small

The **significance level** is defined as **100 α %**

Errors

There is a probability α that the observed value x_{obs} will fall into R when H_0 is true.
Therefore, in 100α% of all decisions H_0 will be rejected when it should not.
The mistake we do by rejecting H_0 when it is true is called a **Type I error**.

Another error can occur, that we do not reject H_0 as true when it is false. This is called a **Type II error**.
The probability of its occurrence β depends on the alternative hypothesis H_1.

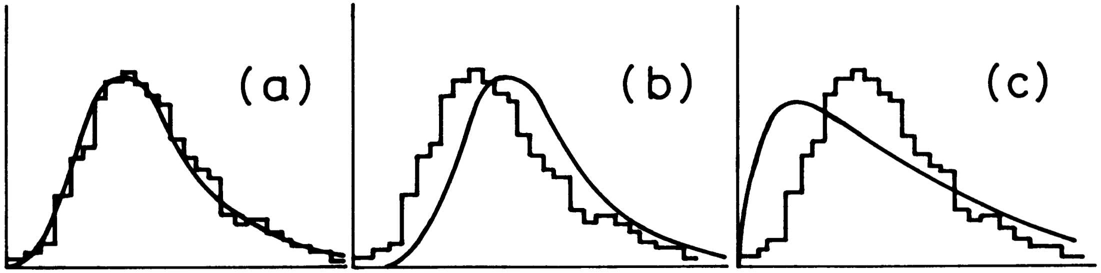


The *best* test is one that makes both α and β as small as possible.
Such a test can be found if – and only if – the hypothesis and its alternative are simple.

Rare in experimental physics!

Power of a test ➡ the probability of rejecting a hypothesis when it is false

Goodness-of-fit testing



a) comparable in shape and location
b) differing in **location**
c) differing in **shape**

A typical goodness-of-fit problem:

$x_1, x_2, \ldots, x_n$ sample values for a random variable x whose true probability function f (x), continuous or discrete, is not known

$f_0(x)$ some particular specified distribution

$H_0 : f(x) = f_0(x)$ hypothesis to be tested on the basis of the sample values

We need a **test statistic** whose distribution, assuming H_0 true, defines a critical region with **probabilities** α (rejection) and **1 − α** (non-rejection)

We may not formulate an alternative hypothesis H_1, since H_1 can be the ensemble of all conceivable hypotheses different from H_0

H_1 is often left unspecified, and the power of the test not taken into account

p-value

Express goodness-of-fit by giving the **p-value** for H_0

It is the probability that, if the null hypothesis is true, the test statistic assumes a value at least as extreme as the one observed

It is **NOT** the probability of the hypothesis being true, given the data

They are not the same: for example, the probability of being pregnant, assuming you are female, is not the same as the probability of being female, given that you are pregnant

If the null hypothesis is true,
the p-value is **flat** between 0 and 1

Common error

Comparing distribution A and distribution B, distribution A and distribution C, $\alpha = 0.01$

$p_{AB} = 0.992$

$p_{AC} = 0.357$

does not authorize one to conclude that the agreement between A and C is worse than the agreement between A and B

GoF recipe

Ingredients:

- data
- H_0
- α
- test *(test statistic)*

1. We define the null hypothesis (and the alternative hypothesis)
2. We set the significance level of the test (α)
   - A priori, as appropriate to the experimental scenario
3. We calculate the value of the test statistic
   - According to the mathematical formulation of the test we choose
4. We compare the value of the test statistic with **tabulated critical values**

OR

4. we calculate the **p-value** corresponding to the test statistic and compare it with the pre-set α

p-value < α → We **reject** the null hypothesis

p-value ≥ α → We **do not reject** the null hypothesis

Statistical software systems

Open source, free

ADMB	A software for non-linear statistical modeling in C++
DAP	A free replacement for SAS
Fityk	Nonlinear regression software
OpenEpi	A web-based, open source series of programs for use in epidemiology and statistics
SciPy	Regressing, plotting, GLM, time series analysis, Non-parametric statistics, ANOVA etc.
PSPP	A free software alternative to IBM SPSS Statistics
R	A free implementation of the S language

Proprietary

…and more

GraphPad InStat	Simple functionality
GraphPad Prism	Biostatistics and nonlinear regression
MATLAB	Programming language with statistical features
SAS	Comprehensive statistical package
SPSS	Statistical package for the social sciences
StatsDirect	Statistics packages designed for biomedical and health science
SPSS Statistics	Comprehensive statistics package

R https://www.r-project.org/ RStudio https://www.rstudio.com/

An integrated development environment (IDE) for R

R is the **lingua franca** for statistical algorithms

- good for statistical programming and data analysis tasks
- used in industry and academia
- free and open-source software
- available on different operating systems: Linux, Windows and macOS
- online communities and resources
- easy at reading and writing data
- equipped with many packages for statistics and graphics

but

- R is slow (it is an interpreted language)
- all the objects are in memory
- vector programming with R is hard to learn

nevertheless

- Rcpp addresses some of these problems

Goodness-of-fit tests

One-sample tests

Compare a data distribution w.r.t. a function
(e.g. gaussian, exponential etc.)
Most common: tests for normality

Two-sample tests

Compare two data distributions

k-sample tests

Most common tests

- χ^2 test
- Kolmogorov-Smirnov test
- Cramer-von Mises test
- Anderson-Darling test

No time to review them in detail
Suggested books and references in

B. Mascialino, A. Pfeiffer, M. G. Pia, A. Ribon, P. Viarengo, "New developments of the Goodness-of-Fit Statistical Toolkit", *IEEE Trans. Nucl. Sci., v*ol. 51, no. 5, pp. 2056-2063, 2004

Pay attention to the characteristics of your data!

Binned/unbinned distributions

Some tests are pertinent to **binned** *(histogram)* or unbinned **distributions** only, or to **continuous** or **discrete** distributions
e.g. the Kolmogorov-Smirnov test is applicable to continuous distributions

Some tests are applicable both to binned and unbinned distributions, but the test statistic has a slightly different formulation in either case

Parametric/non-parametric tests

Non-parametric = distribution-free

In parametric tests some assumptions are made regarding some characteristics of the data distributions
e.g. normality, homoschedasticity etc.

These **assumptions must be verified** for the test to be applicable

χ^2 test

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

n independent variables x_i are each normally distributed with mean μ_i and variance σ_i^2

becomes a matrix equation if errors are correlated

$$P(\chi^2; n) = \frac{2^{-n/2}}{\Gamma(n/2)} \chi^{n-2} e^{-\chi^2/2}$$

χ^2 distribution
with n dof

mean n
variance $2n$

p-value = integral from χ^2 to ∞

If the data are binned (i.e. in a histogram), Pearson's χ^2 applies:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

O = observed
E = expected

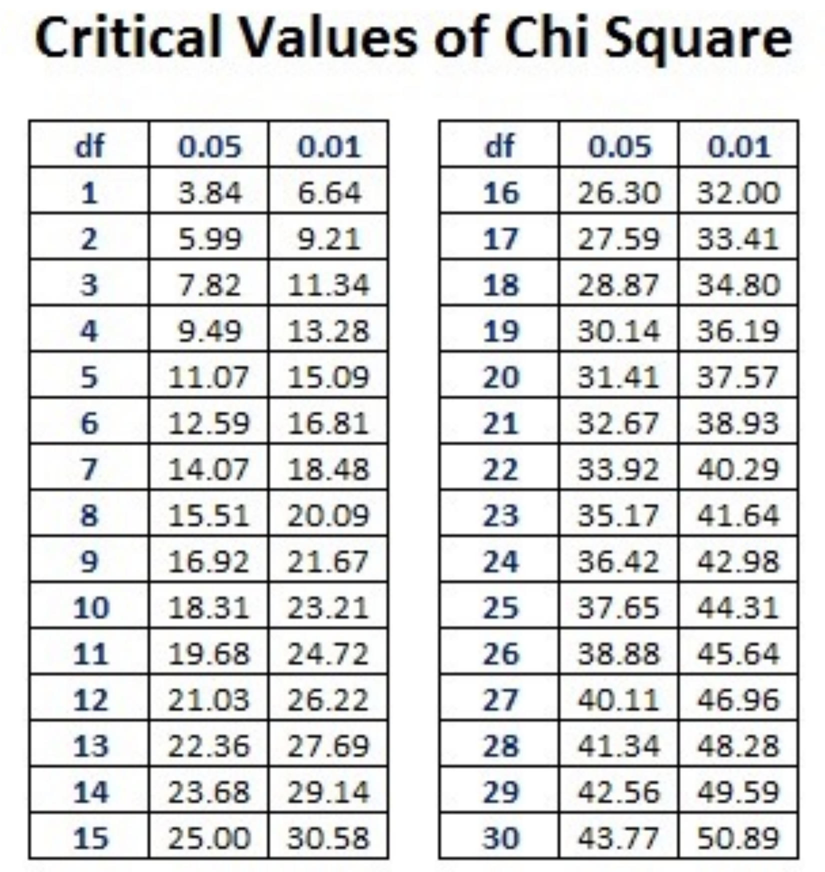
(based on Poisson distribution)

χ^2/N

χ^2 distribution
with n dof

mean n
variance $2n$

Rule of thumb: $\chi^2/n \approx 1$



Critical Values of Chi Square

df	0.05	0.01
1	3.84	6.64
2	5.99	9.21
3	7.82	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58

df	0.05	0.01
16	26.30	32.00
17	27.59	33.41
18	28.87	34.80
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.92	40.29
23	35.17	41.64
24	36.42	42.98
25	37.65	44.31
26	38.88	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89

OK to get a rough idea of the outcome of your test,
inappropriate to report a scientific result

```
pvChi2 <- pchisq(chi2,n,lower.tail=FALSE)
```

Experimental uncertainties

The χ^2 statistic involves experimental uncertainties explicitly

- Beware of biased test outcomes because of unrealistic estimates of the experimental uncertainties

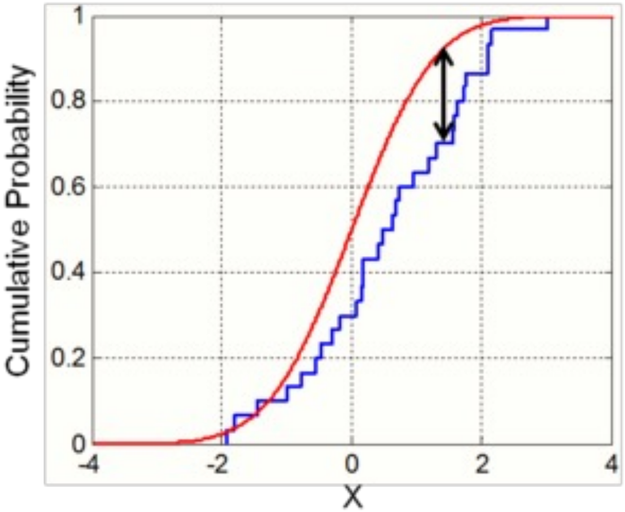
- The null hypothesis is rejected
  - Your model does not truly fit the data
  - Or perhaps your experimental uncertainties are **underestimated**
  - *You may also want to check if you have any outliers*

- The p-value is suspiciously ≈ 1
  - You were lucky…
  - Or perhaps your experimental uncertainties are **overestimated**

Kolmogorov-Smirnov test

It is a test for **unbinned** *(continuous)* distributions

The test statistic is formed as the **maximum difference** between the two **cumulative distributions** being compared



$$\rho(F, G) \equiv \sup_x |F(x) - G(x)|$$

cdf corresponding to the null hypothesis

empirical cdf of our data set

Cramer-von Mises test (AKA Fitz-Cramer-von Mises test)

Formulations exist for both **binned** and **unbinned** distributions
Use the appropriate one!

It is a variation on the Kolmogorov–Smirnov approach, where one replaces the supremum distance function with another common measure of distance:
the **average squared deviation**

$$\boxed{C_N^2(F) = \int_{-\infty}^{\infty} [F_N(y) - F(y)]^2 dF(y)}$$

test statistic

Anderson-Darling test

Formulations exist for both **binned** and **unbinned** distributions
Use the appropriate one!

It is designed to give more weight to the tails of the distribution

→

It is thus particularly powerful in tests involving **deviations in the tails**

Some distributions may be approximately normal in the central region, but the tails may be significantly non-Gaussian: the Anderson–Darling test is known to be especially powerful in detecting such cases

$$\boxed{A^2_N(\mathbf{x}) = N \int_{-\infty}^{\infty} \frac{[F_N(y) - F(y)]^2}{F(y)[1 - F(y)]} \, dF(y)}$$

The test statistic modifies the Cramer–von Mises statistic with the addition of a weighting term, which gives **preferential weight to the tails** of the distribution

More exotic tests

- Generalised **Girone** test
- **Watson** test *(another variation of the Cramer-von Mises test)*
- **Weighted** Kolmogorov-Smirnov and Cramer-von Mises tests
- Approximated calculations
  - Useful when computational performance is an issue
  - **Goodman** and **Tiku** tests: approximated versions of Kolmogorov-Smirnov and Cramer-von Mises tests

B. Mascialino, A. Pfeiffer, M. G. Pia, A. Ribon, P. Viarengo, "New developments of the Goodness-of-Fit Statistical Toolkit", *IEEE Trans. Nucl. Sci.,* vol. 53, no. 6, pp. 3834-3841, 2006 *and references therein*

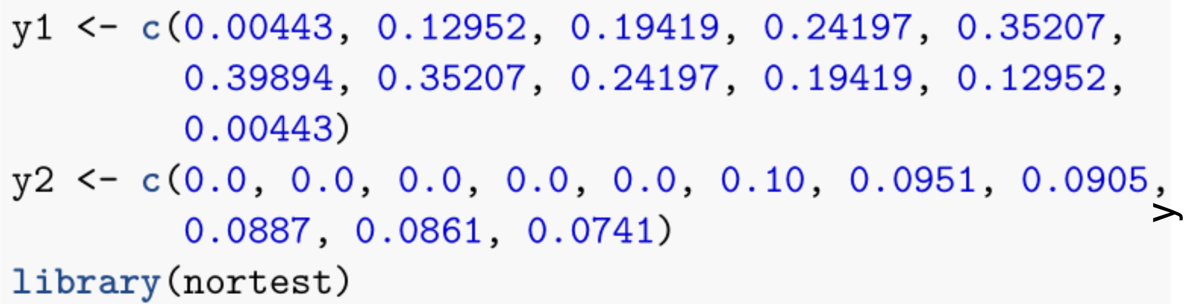
Example: test normality

Given a data distribution, assess if it is normally distributed

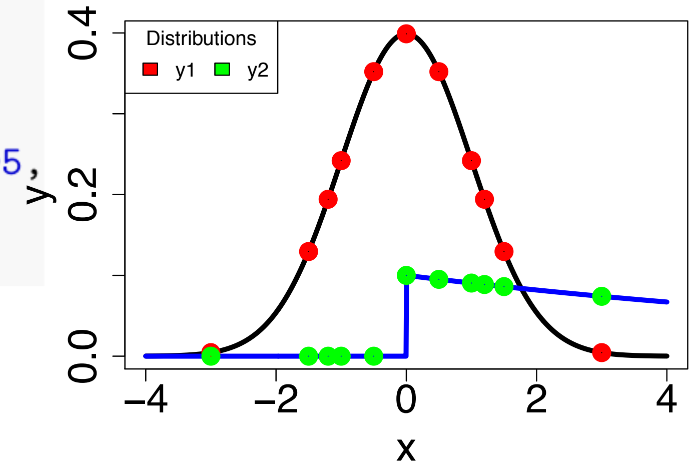
Null hypothesis: the data are normally distributed

Significance level: let’s set $\alpha = 0.01$

`nortest` package provides a set tests for **normality**, including Pearson chi-square test, Cramer-von Mises test and Anderson-Darling test



```
y1 <- c(0.00443, 0.12952, 0.19419, 0.24197, 0.35207,
        0.39894, 0.35207, 0.24197, 0.19419, 0.12952,
        0.00443)
y2 <- c(0.0, 0.0, 0.0, 0.0, 0.0, 0.10, 0.0951, 0.0905,
        0.0887, 0.0861, 0.0741)
library(nortest)
```



pt1 <- **cvm.test**(y1)
pt2 <- **cvm.test**(y2)

Cramer-von Mises test

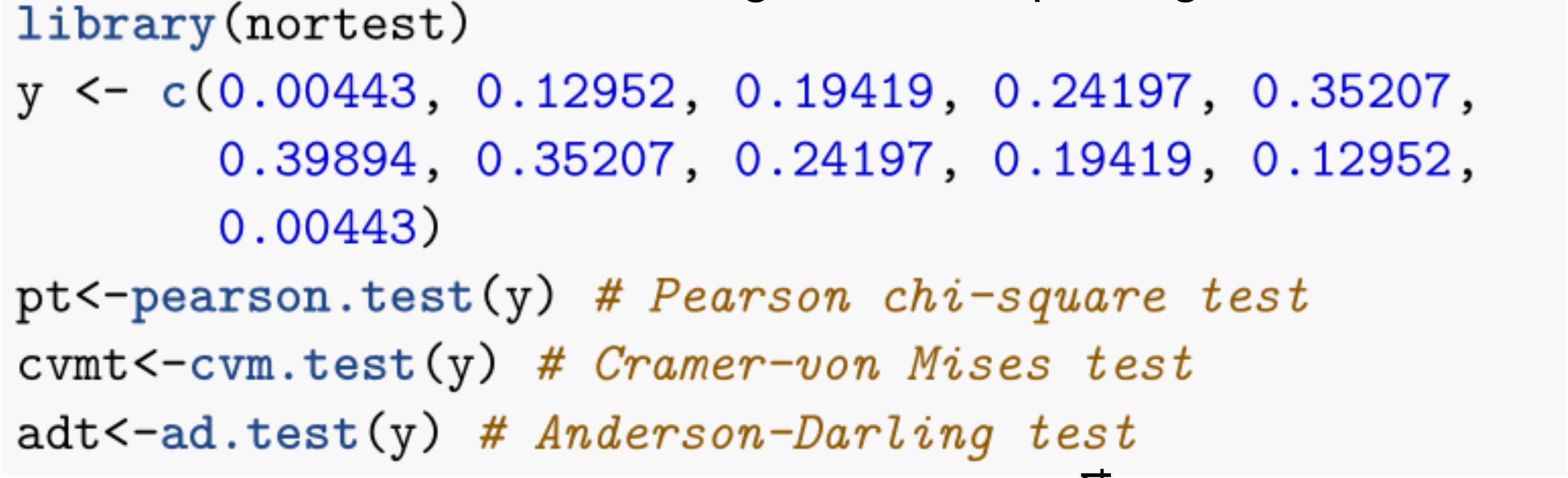
For y1: p-value = 0.7037954

For y2: p-value = 0.0018620

Example: test normality

Calculate p-values for different normality tests.

using R **nortest** package

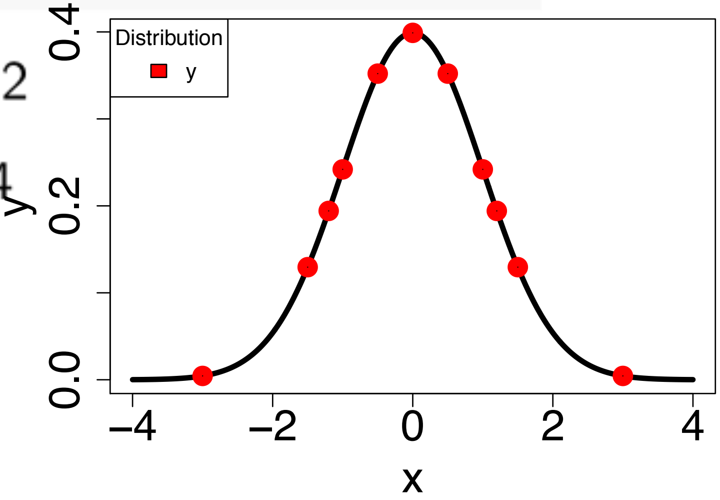


```
library(nortest)
y <- c(0.00443, 0.12952, 0.19419, 0.24197, 0.35207,
       0.39894, 0.35207, 0.24197, 0.19419, 0.12952,
       0.00443)
pt<-pearson.test(y) # Pearson chi-square test
cvmt<-cvm.test(y) # Cramer-von Mises test
adt<-ad.test(y) # Anderson-Darling test
```

Pearson chi-square test: p-value = 0.4511502

Cramer-von Mises test: p-value = 0.7037954

Anderson-Darling test: p-value = 0.5589221



Example: Anderson-Darling test

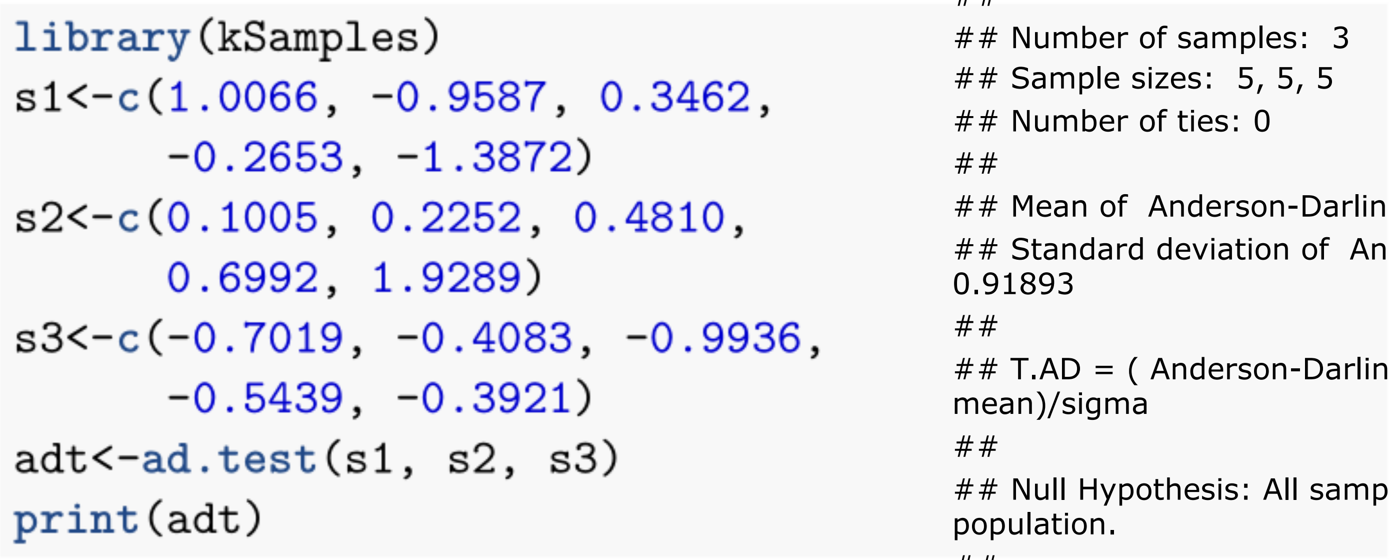
Given three data distributions, assess if they arise from a common (unspecified) distribution function

`kSamples` package: compare *k* samples using the Anderson-Darling test

$\alpha = 0.01$

version 1: Continuous population
version 2: Discrete parent population

Null hyphothesis: The samples have a common distribution function.

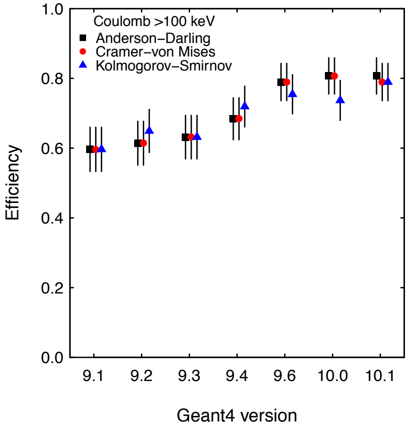


```
library(kSamples)
s1<-c(1.0066, -0.9587, 0.3462,
      -0.2653, -1.3872)
s2<-c(0.1005, 0.2252, 0.4810,
      0.6992, 1.9289)
s3<-c(-0.7019, -0.4083, -0.9936,
      -0.5439, -0.3921)
adt<-ad.test(s1, s2, s3)
print(adt)
```

```
## Anderson-Darling k-sample test.
##
## Number of samples:  3
## Sample sizes:  5, 5, 5
## Number of ties: 0
##
## Mean of  Anderson-Darling  Criterion: 2
## Standard deviation of  Anderson-Darling  Criterion:
0.91893
##
## T.AD = ( Anderson-Darling  Criterion -
mean)/sigma
##
## Null Hypothesis: All samples come from a common
population.
##
##                AD  T.AD  asympt. P-value
## version 1: 4.08   2.26          0.0355
## version 2: 4.08   2.27          0.0354
```

Which GoF test to use?

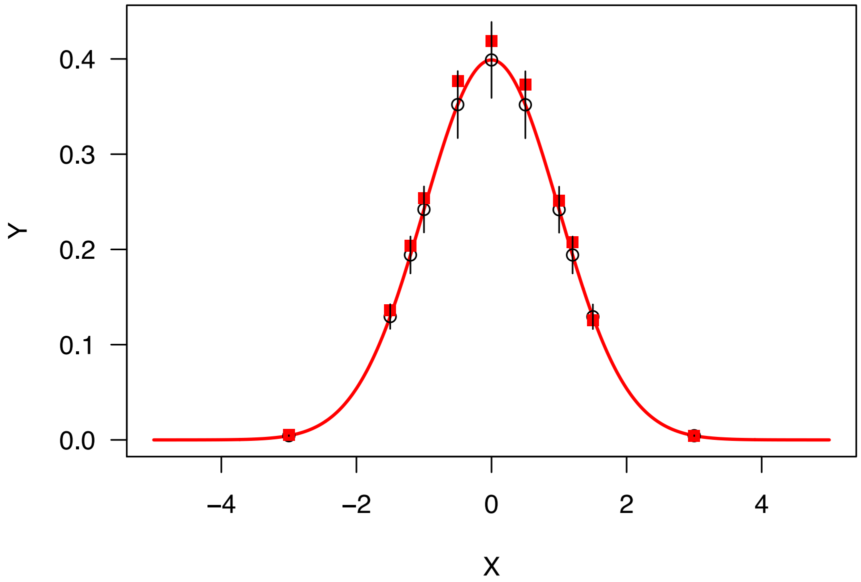
- The test should be appropriate to the data and to the experimental scenario which it is applied to
  - Binned/unbinned data
  - Known (reliable) experimental uncertainties
  - Characteristics of the data: fat tails, smooth, cyclic etc.



- Good practice: apply several tests to the data
  - Is the outcome (rejection/non-rejection) consistent?
  - What could be the source of inconsistencies?
    - Unrealistic experimental uncertainties, intrinsic characteristics of the data…

Tests of randomness

Is it sufficient to perform goodness-of-fit tests to conclude that the data are consistent with the model?



Goodness-of-fit tests do not tell us anything about the possible presence of **systematic effects**

Tests of randomness

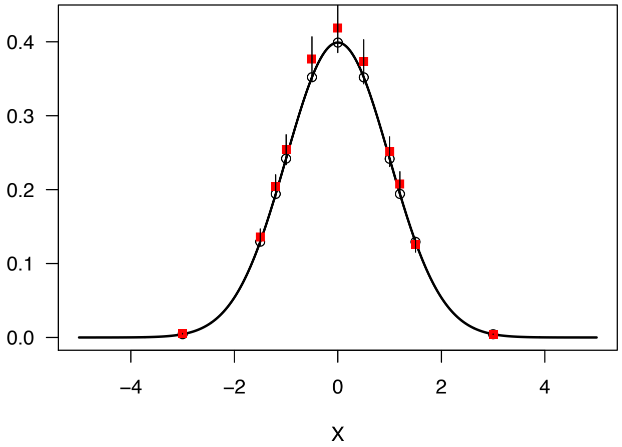
are useful to identify systematic effects

Runs test

$\alpha = 0.01$

χ^2 test → p-value = **0.848**

K-S test → p-value = **0.997**



Experimental data appear systematically larger than reference data

The sequence of red data is AboveAAAAAAAABelowB = 2 runs

randtests package

runs.test

```
result<-runs.test(diff, "two.sided", 0.)
```

performs the **Wald-Wolfowitz Runs Test**

runs test → p-value = **0.008**

A. Wald, J. Wolfowitz, On a test whether two samples are from the same population, *Ann. Math. Stat.*, vol. 11, pp. 147–162, 1940

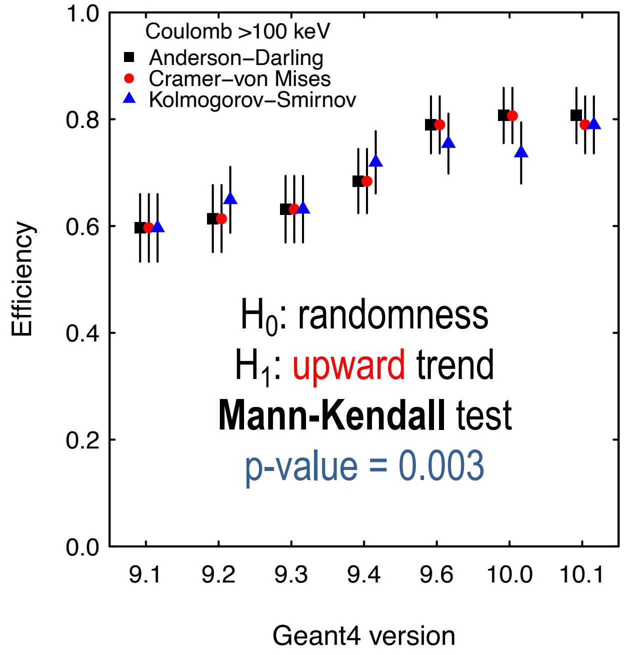
Trend tests

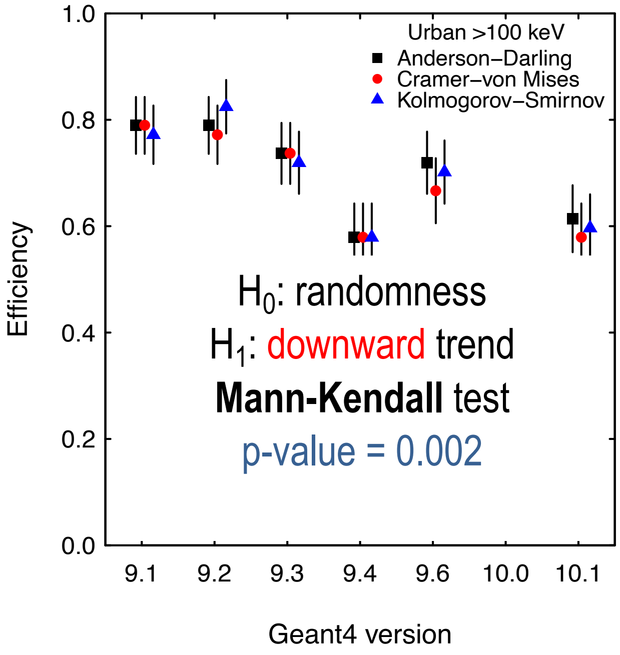
Mann-Kendall test

Cox-Stuart test

Can detect the presence of a trend in the data

e.g. “upward”, “downward” *(one-sided)*
or just “trend” *(two-sided)* w.r.t. randomness



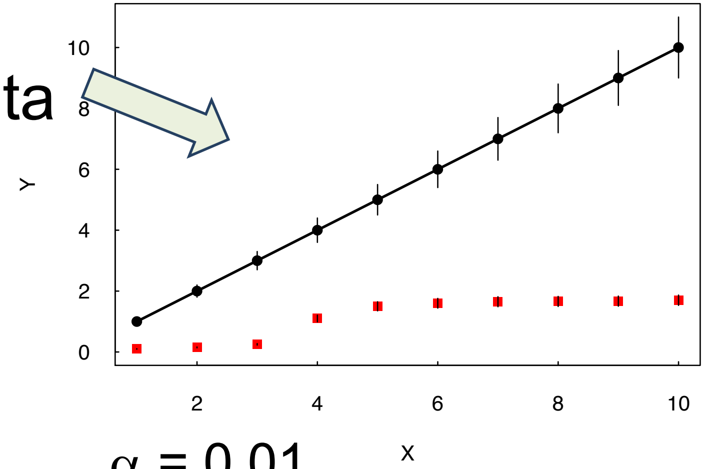


Correlation and GoF

Common error: confusing correlation with goodness-of-fit testing

A correlation coefficient provides a **description** of the data,
it does not support the **inference** that two data samples
are drawn from the same parent distribution

Correlation between red and black data



linear

Pearson correlation coefficient = **0.896**

Kendall τ = **1**

Spearman ρ = **1**

χ^2 test → p-value **< 0.0001**

K-S test → p-value **= 0.0002**

$\alpha = 0.01$

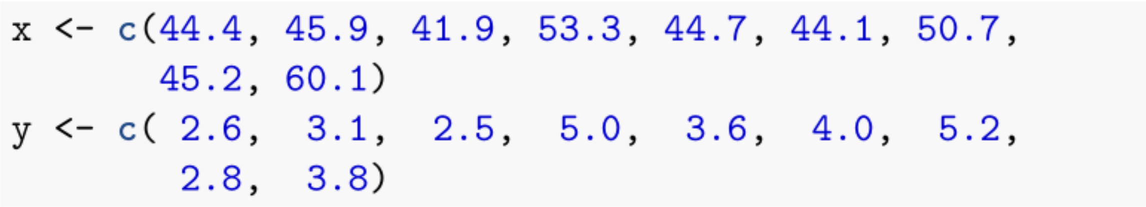
GoF tests reject the hypothesis of compatibility between red and black data

Example

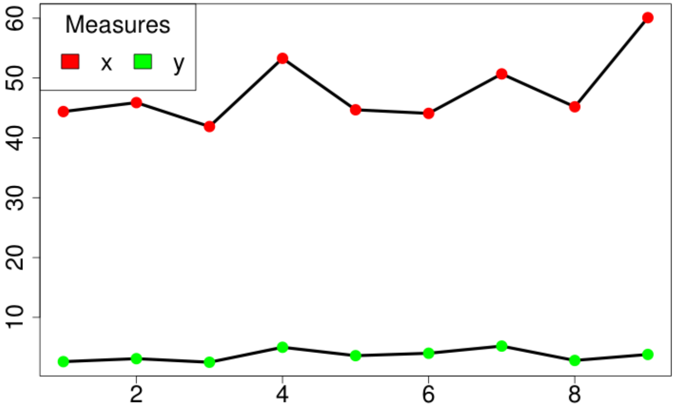
`stats` **package**

Correlation using Pearson's correlation coefficient, Kendall's τ or Spearman's ρ

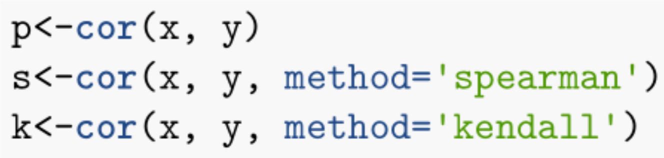
Let's consider two data sets and assess their relationship



```
x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7,
       45.2, 60.1)
y <- c( 2.6,  3.1,  2.5,  5.0,  3.6,  4.0,  5.2,
        2.8,  3.8)
```



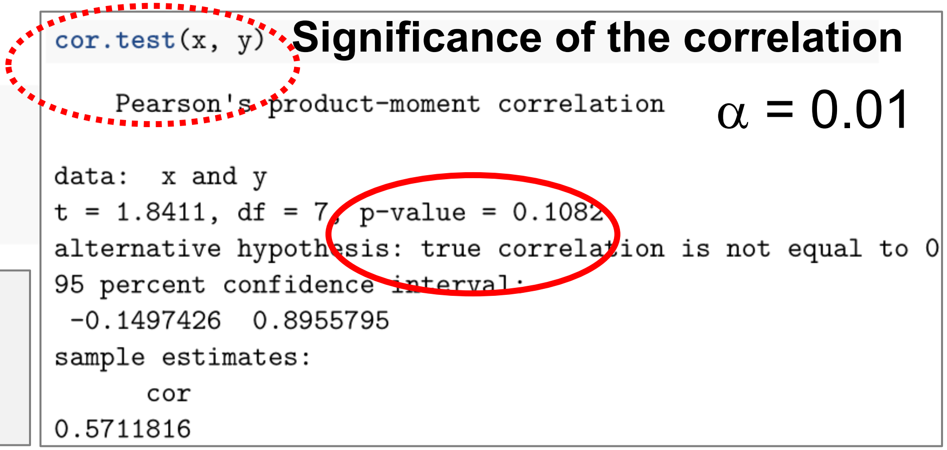
Get correlation coefficient.



```
p<-cor(x, y)
s<-cor(x, y, method='spearman')
k<-cor(x, y, method='kendall')
```

Use the *method* argument to change the way the association is measured

Pearson correlation coefficient:	0.5711816
Spearman ρ:	0.6
Kendall τ:	0.4444444



Significance of the correlation

α = 0.01

```
cor.test(x, y)

	Pearson's product-moment correlation

data:  x and y
t = 1.8411, df = 7, p-value = 0.1082
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1497426  0.8955795
sample estimates:
      cor
0.5711816
```

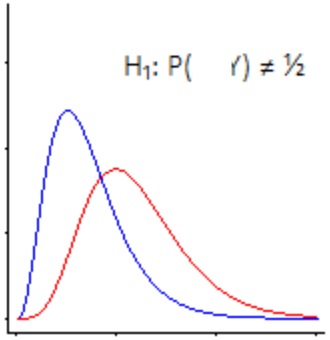
Do not confuse correlation with GoF!

Location-scale tests

Less common in a physics validation context

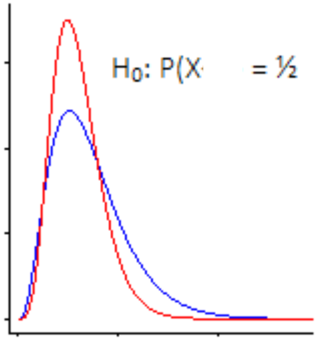
Location

- Two-sample tests for comparison of the means
- Two-sample permutation tests for location
- Multi-sample case



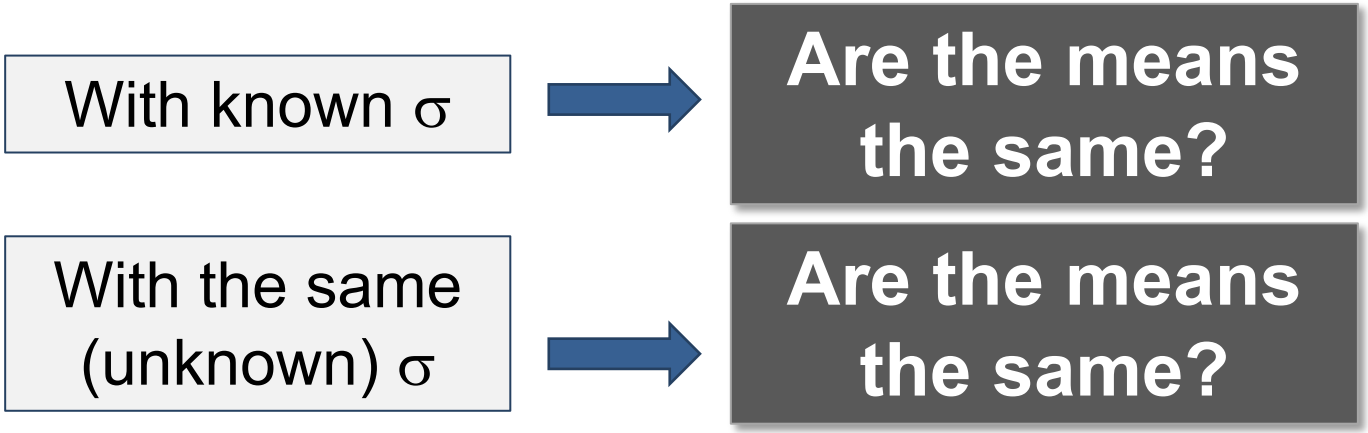
Scale

- Test on the ratio of two variances
- Variability comparison



From a physicist's perspective

Assuming two Gaussian samples:



With known σ → Are the means the same?

With the same (unknown) σ → Are the means the same?

Is σ the same?

Known σ

X and Y are randomly distributed about their true value(s) according to the Gaussian distribution with standard deviations σ_X and σ_Y

Are they "really" the same?

Equivalent to asking whether X-Y is compatible with 0

X-Y has variance **$V(X-Y) = \sigma_X^2 + \sigma_Y^2$**

The question reduces to how many σ correspond to the difference.
A decision at a preset significance level is made according to the table of the integrated Gaussian

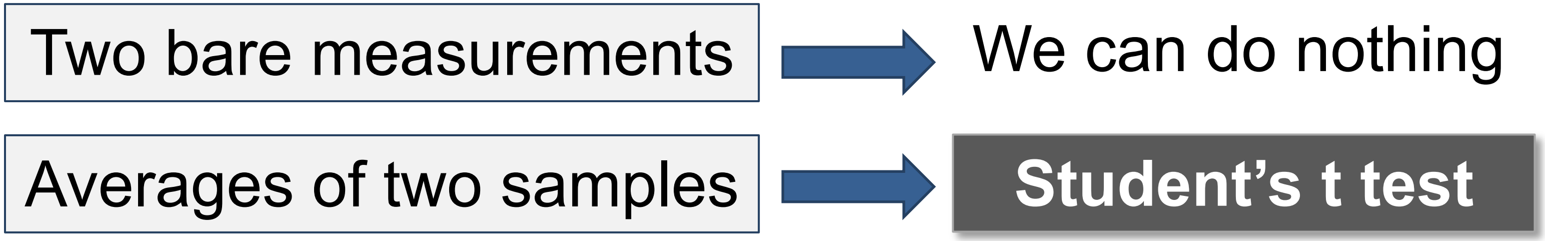
Example: two experimental measurements with detectors of known resolution σ_X and σ_Y

Experiments A and B measure
$E_A = 202 \pm 3$ MeV, $E_B = 209 \pm 4$ MeV

error = sqrt(9+16) MeV = 5 MeV

$E_B - E_A = 7$ MeV → **1.4 σ**

Unknown σ


Two bare measurements → We can do nothing

Averages of two samples → Student's t test

Estimate of σ:

$$\hat{\sigma}_x = s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N_x - 1}} \qquad \hat{\sigma}_y = s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{N_y - 1}}$$

Under the null hypothesis $\mu_1=\mu_2$, and **if $\sigma_1=\sigma_2$** →

$$\frac{\bar{x} - \bar{y}}{S\sqrt{(1/N_x) + (1/N_y)}}$$

Pooled estimate of the variance

$$S^2 = \frac{(N_x - 1)s_x^2 + (N_y - 1)s_y^2}{N_x + N_y - 2}$$

is distributed according to Student's t distribution with N_x+N_y-2 degrees of freedom

Wilcoxon rank sum test (AKA Mann-Whitney test)

Non-parametric location test

Two populations characterized by the <u>same distribution</u>, possibly except for the location

Assumptions

- Continuous, independent random variables
- Identical underlying data distributions within the same sample

Compute the **ranks** of all $n=n_1+n_2$ observations of the pooled sample

Test statistic:

$$W = \sum_{i=1}^{n_1} r(X_{1i})$$

sum of the ranks of the first sample *(if no ties)*

The distribution of W is obtained considering all possible rank assignments and depends only on the sample sizes

Gaussian approximation for large samples and if ties

Kruskal-Wallis test: multi-sample generalization

Scale comparison: F test on the ratio of variances

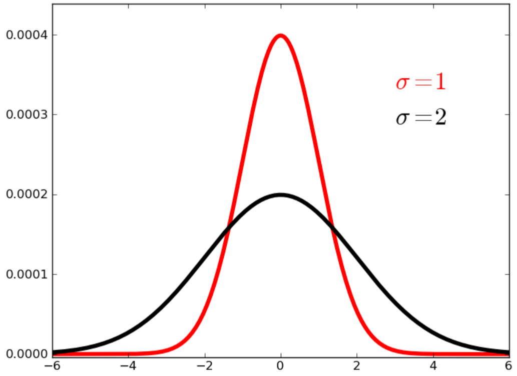
Under the assumption of normality of the data:

- Two independent random samples from a normal distribution
- Mean and variance are unknown for both populations

$H_0: \sigma_1^2 = \sigma_2^2$

$H_1: \sigma_1^2 \neq \sigma_2^2$

Typical scale parameter : variance



Test statistic: ratio of two sample variances

$F = S_1^2 / S_2^2$

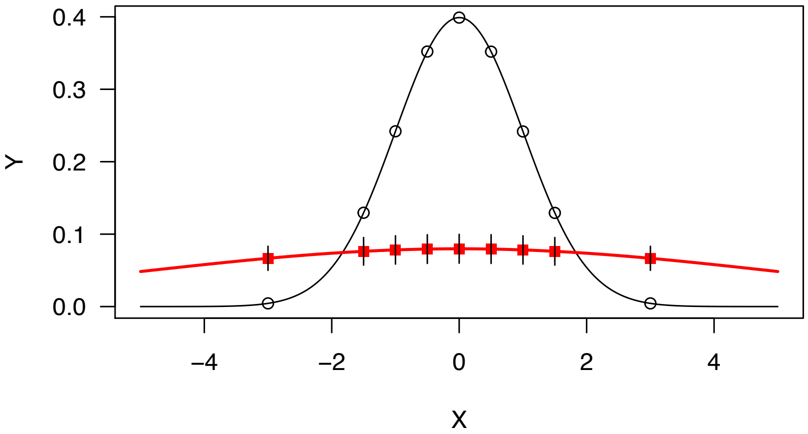
code: `var.test(x1,x2,"two.sided")`

Beware of what you want to test!

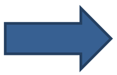
Two data samples, drawn from two gaussians:

G_1: μ = 0, σ = 1

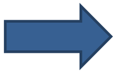
G_2: μ = 0, σ = 5



α = 0.01

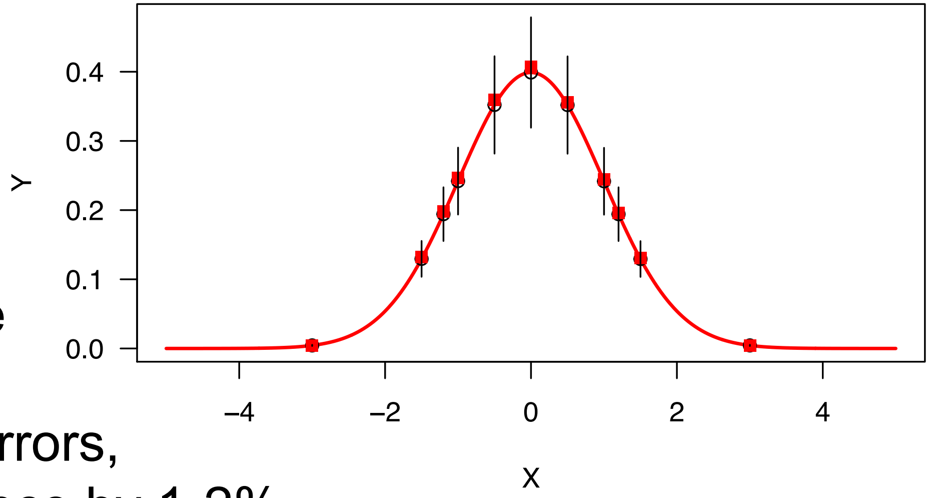


Wilcoxon Signed Rank test → p-value **> 0.01** **location**



χ^2 test
Kolmogorov-Smirnov test → p-value **< 0.01** **general alternative**

Beware of the **assumptions** underlying the tests!



Two data samples drawn from the same gaussian ($\mu = 0, \sigma = 1$)
G_1: experimental data with 20% errors,
G_2: the data differ from the reference by 1-2%

According to physical intuition, these data samples are consistent

$\alpha = 0.01$

Wilcoxon Signed Rank test → p-value **< 0.01**

The assumption we are missing is that the distribution must be symmetric; in this case the differences between the two data sets are not

χ^2 test → p-value **> 0.01**

Comparing categorical data

- A categorical variable has a measurement scale consisting of a set of categories
  - e.g. accommodation could use categories apartment, house, castle, igloo…
- Widely used in **social sciences** and **health sciences**
- Quite uncommon in experimental physics analyses
- Used to compare the ”validity” of different physics models, or to compare simulation configurations

Contingency tables

	Category A	Category B
Obs X	N_{AX}	N_{BX}
Obs Y	N_{AY}	N_{BY}

Category A and B could be, for instance, two physics models, two simulation configurations, two detector prototypes…

Are the observed results of the two categories compatible with chance, or do they exhibit any significant difference?

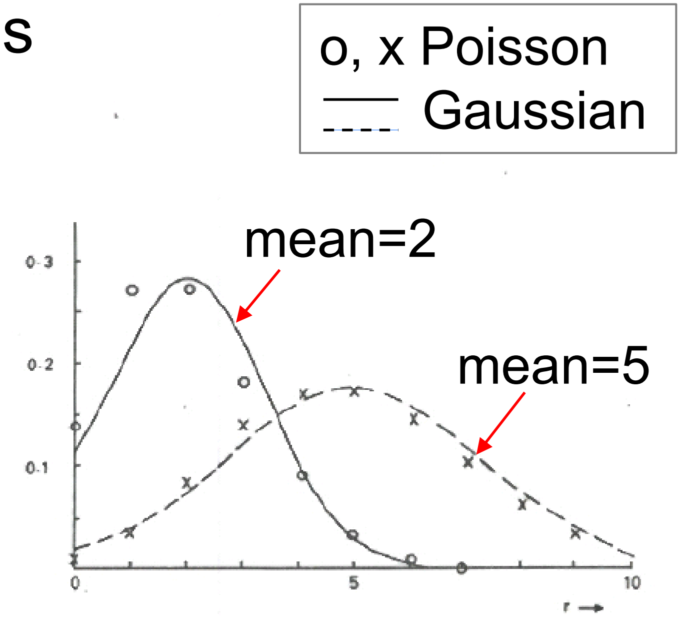
$\boxed{\pi_{ij} = P(X = i, Y = j)}$ probability that (X, Y) falls in the cell in row i and column j

Probabilities $\{\pi_{ij}\}$ form the **joint distribution** of X and Y, $\sum \pi_{ij} = 1$

marginal distributions: the row and column totals of the joint probabilities

Pearson χ^2 test

- Can be applied to contingency tables
  - provided the number of entries in each cell is “sufficiently large”
  - Otherwise: Yates’ continuity correction
- Similar to the application of the test to histograms
  - Observed and expected



Exact tests

Several tests implemented in R packages:
Exact2x2, Exact, Barnard…

Conditional tests

- Conditions on the marginals
- Usually not experimentally realistic
- Tend to be conservative
- Can be used even with small number of entries in cells

- **Fisher exact test**

Unconditional tests

- No conditions on marginals
- Experimentally realistic
- Slow computation

- **Barnard test** and approximations
- **Boschloo test**
- **Suissa-Schuster test**
- …

Example

Function	Package
`fisher.test`	`stats`
`chisq.test`	`stats`
`boschloo`	`exact2x2`
`barnard`	`Barnard`
`exact.test`	`exact`

Let us consider the following contingency table extracted from Table IV in Han *et al.* paper

- Each **model** implements a photoionization cross section
- Do the Biggs and EPDL model exhibit equivalent, or significantly different compatibility with experiment?

Model	Experimental	
	Pass	**Fail**
Biggs	33	9
EPDL	36	6

- p-values from tests applied to the contingency table

Fisher	χ^2	*Boschloo*	*Barnard*	*Z-pooled*
0.570	0.393	0.427	0.942	0.427

Same analysis for the original Biggs model and an "improved" version implemented in Geant4

Model	Experimental	
	Pass	Fail
Biggs	10	4
BiggsG4	1	13

Fisher	χ^2	*Boschloo*	*Barnard*	*Z-pooled*
0.0014	not applicable	0.0004	0.0004	0.0004

M. C. Han, H. S. Kim, M. G. Pia, T. Basaglia, M. Batič, G. Hoff, C. H. Kim, and P. Saracco, "Validation of Cross Sections for Monte Carlo Simulation of the Photoelectric Effect", *IEEE Trans. Nucl. Sci,* vol. 63, no. 2, pp. 1117-1146, 2016

McNemar test for related data

	Category A success	Category A failure
Category B success	$N_{succ.A,\ succ.B}$	$N_{fail.A,\ succ.B}$
Category B failure	$N_{succ.A,\ fail.B}$	$N_{fail.A,\ fail.B}$

Table especially filled for **McNemar test**

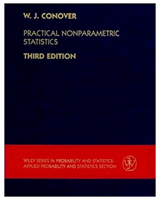
Focuses on the significance of the discordant results

Null hypothesis: the proportion of discordant results is the same in the two cells corresponding to *"success-fail"* or *"fail-success"* associated with the two categories subject to test

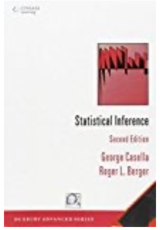
Application examples in: S. H. Kim et al.,Validation Test of Geant4 Simulation of Electron Backscattering, IEEE Trans. Nucl. Sci., vol. 62, no. 2, pp. 451-479, 2015

Suggested reading

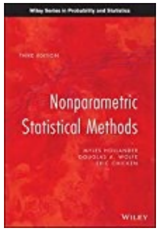
Classical statistics books



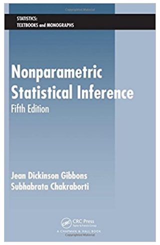
W. J. Conover,
Practical Nonparametric Statistics,
Wiley



G. Casella, R. Berger,
Statistical Inference,
Duxbury Press



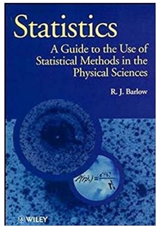
M. Hollander,
Nonparametric Statistical Methods,
Wiley



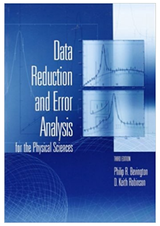
J. D. Gibbons, S. Chakraborti,
Nonparametric Statistical Inference,
Chapman and Hall/CRC

Suggested reading

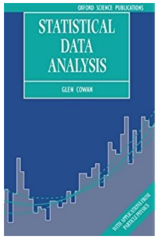
Books written by physicists



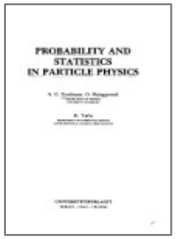
R. J. Barlow, **Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences,** Wiley



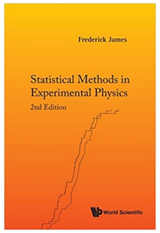
P. Bevington and D. K. Robinson, **Data Reduction and Error Analysis for the Physical Sciences**, McGraw-Hill



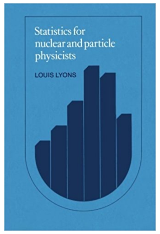
G. Cowan, **Statistical Data Analysis,** Oxford Science Publications



A. G. Frodesen, O. Skjeggestad, H. Tofte, **Probability and Statistics in Particle Physics**, Universitetsforlaget



F. James, **Statistical Methods in Experimental Physics**, World Scientific



L. Lyons, **Statistics for Nuclear and Particle Physicists**, CUP

Suggested reading

Statistical software tools

https://cran.r-project.org/manuals.html

R documentation

(or the documentation of any other software system you may want to use)

Includes "Introduction to R",
which is a recommended reading for anybody wishing to learn R

https://www.r-project.org/doc/bib/R-books.html

>150 books!

Many introductory books, pick your favourite

P. Dalgaard, **Introductory Statistics with R**, Springer

Plenty of online R material; *be careful about the authoritativeness of the source…*

Summary

- Hypothesis testing
  - Conceptual framework
- Goodness-of-fit
  - One sample
    (compare with function)
  - Two or more samples
- Location
- Scale
- Categorical data

- Statistical software
  - Overview of available systems
  - R
- Application examples
- Suggested reading

…within 40 minutes?

Conclusion

Statistical methods let us compare data distributions **objectively** and **quantitatively**, consistent with the **scientific method**

Computational tools are available to facilitate their use, but beware of the **assumptions** and **conditions** for their applicability

Examples and exercises in a GitHub repository
https://doi.org/10.5281/zenodo.6567236

Question, feedback, discussions are welcome
Feel free to contact MGP

A longer hands-on course at IAEA/ICTP?