

# Tools for Physicists: Statistics

Wolfgang Gradl  
Peter Weidenkaff

Institut für Kernphysik

Summer semester 2019



JOHANNES GUTENBERG  
UNIVERSITÄT MAINZ

# The scientific method: how we create ‘knowledge’

## Theory / model

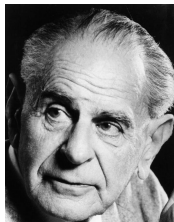
- usually mathematical
- self-consistent
- simple explanations, few (arbitrary) parameters
- testable predictions / hypotheses

Advance of scientific knowledge is *evolutionary* process with occasional revolutions

Statistical methods are important part of this process

## Experiment

- modify or even reject theory in case of disagreement with data
- if theory requires too many adjustments it becomes unattractive
- generate surprises



Karl Popper  
(1902–1994)

# Statistics in science

Statistics is needed to:

- characterise and summarise experimental results (impractical to always deal with raw data)
- quantify uncertainty of a measurement
- assess whether two measurements of the same quantity are compatible, combine measurements
- estimate parameters of an underlying model or theory
- test hypotheses:  
determine whether a model is compatible with data
- ...

# Aims of this mini-series

- **Statistical inference:** from data to knowledge
  - ▶ Should we believe a physics claim?
  - ▶ Develop intuition
  - ▶ Know (some) pitfalls: avoid making mistakes others have already made
- **Understand statistical concepts**
  - ▶ Ability to understand physics papers
  - ▶ Know some methods / standard statistical toolbox
- **Use tools**
  - ▶ Hands-on part with Python / Jupyter
  - ▶ Application to your own work



# Practical information

Three sessions:

1. Basics, introduction, statistical distributions
2. Parameter estimation
3. Confidence intervals, hypothesis testing

About 60 minutes of lecture, then  $\geq 30$  minutes hands-on tutorial

I hope this will be useful for you,  
but keep in mind that there is much more  
to statistics than can be covered  
in three brief hours.



# Two quick questions

<https://pingo.coactum.de/529916>



- What is your (main) area of research / interest?
- Which programming language(s) do you speak?

# Useful reading material

## Books:

- G. Cowan, Statistical Data Analysis
- R. Barlow, Statistics: A guide to the use of statistical methods in the physical sciences
- L. Lyons, Statistics for Nuclear and Particle Physicists
- A. J. Bevan, Statistical data analysis for the physical sciences
- G. Bohm, G. Zech, Introduction to Statistics and Data Analysis for Physicists (available [online](#))

## Lectures on the web:

- G. Cowan, Royal Holloway University London: Statistical Data Analysis
- K. Reygers, U Heidelberg, Stat. Methods in Particle Physics

# Dealing with uncertainty

- Underlying theory is probabilistic (quantum mechanics / QFT)  
source of **true** randomness
- Limited knowledge about measurement process  
even without QM  
random measurement errors
- Things we could know in principle, but don't  
e.g. from limitations of cost, time, ...

Quantify uncertainty using **probability**

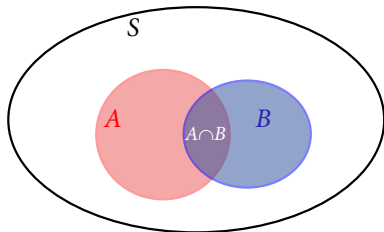
# Mathematical definition of probability

Kolmogorov axioms:

Consider a set  $S$  (the **sample space**) with subsets  $A, B, \dots$  (**events**).

Define a function  $P : \mathfrak{P}(S) \mapsto [0, 1]$  with

1.  $P(A) \geq 0$  for all  $A \in S$
2.  $P(S) = 1$
3.  $P(A \cup B) = P(A) + P(B)$  if  $A \cap B = \emptyset$ ,  
i.e.  $A$  and  $B$  are exclusive



From these we can derive further properties:

- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup \bar{A}) = 1$
- $P(\emptyset) = 0$
- If  $A \in B$ , then  $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

for the mathematically inclined: proper treatment will use *measure theory*

# Interpretations

## ■ Classical definition

- ▶ Assign equal probabilities based on symmetry of problem, e.g. rolling ideal dice:  $P(6) = 1/6$
- ▶ difficult to generalise, sounds somewhat circular

## ■ Frequentist: relative frequency

- ▶  $A, B, \dots$  outcomes of a repeatable experiment

$$P(A) = \lim_{n \rightarrow \infty} \frac{\text{times outcome is } A}{n}$$

## ■ Bayesian: subjective probability

- ▶  $A, B, \dots$  are hypotheses (statements that are either true or false)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

...all three definitions consistent with Kolmogorov's axioms

# Conditional probability, independent events

Conditional probability for two events  $A$  and  $B$ :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example: rolling dice

$$P(n < 3 | n \text{ even}) = \frac{P((n < 3) \cap (n \text{ even}))}{P(n \text{ even})} = \frac{1/6}{1/2} = 1/3$$

Events  $A$  and  $B$  independent  $\iff P(A \cap B) = P(A) \cdot P(B)$

$A$  is independent of  $B$  if  $P(A|B) = P(A)$

# Bayes' theorem

Definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

But obviously  $P(A \cap B) = P(B \cap A)$ , so:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Allows to 'invert' statements about probability:

of great interest to us. Want to infer  $P(\text{theory}|\text{data})$  from  $P(\text{data}|\text{theory})$

Often these two are confused, knowingly or unknowingly  
(advertising, political campaigns, ...)



# Example for Bayes' theorem: Rare disease

Base probability (for anyone) to have a disease  $D$ :

$$P(D) = 0.001$$

$$P(\text{no } D) = 0.999$$

# Example for Bayes' theorem: Rare disease

Base probability (for anyone) to have a disease  $D$ :

$$P(D) = 0.001$$

$$P(\text{no } D) = 0.999$$

Consider a test for  $D$ : result is positive or negative (+ or -):

$$P(+|D) = 0.98$$

$$P(+|\text{no } D) = 0.03$$

$$P(-|D) = 0.02$$

$$P(-|\text{no } D) = 0.97$$

# Example for Bayes' theorem: Rare disease

Base probability (for anyone) to have a disease  $D$ :

$$P(D) = 0.001$$

$$P(\text{no } D) = 0.999$$

Consider a test for  $D$ : result is positive or negative (+ or -):

$$P(+|D) = 0.98 \qquad P(+|\text{no } D) = 0.03$$

$$P(-|D) = 0.02 \qquad P(-|\text{no } D) = 0.97$$

Suppose your result is +; should you be worried?

# Example for Bayes' theorem: Rare disease

Base probability (for anyone) to have a disease  $D$ :

$$P(D) = 0.001$$

$$P(\text{no } D) = 0.999$$

Consider a test for  $D$ : result is positive or negative (+ or -):

$$P(+|D) = 0.98 \qquad P(+|\text{no } D) = 0.03$$

$$P(-|D) = 0.02 \qquad P(-|\text{no } D) = 0.97$$

Suppose your result is +; should you be worried?

$$\begin{aligned} P(D|+) &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\text{no } D)P(\text{no } D)} \\ &= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} = 0.032 \end{aligned}$$

Probability that you have disease is **3.2%**, i.e. you're probably ok

# Bayes' theorem: degree of belief in a theory

$$P(\text{theory}|\text{data}) = \frac{P(\text{data}|\text{theory})P(\text{theory})}{P(\text{data})}$$

likelihood

prior (before seeing the data, subjective)

posterior probability, i.e., after seeing the data

normalization

# Criticisms — Frequentists vs. Bayesians

## ■ Criticisms of the frequentist interpretation

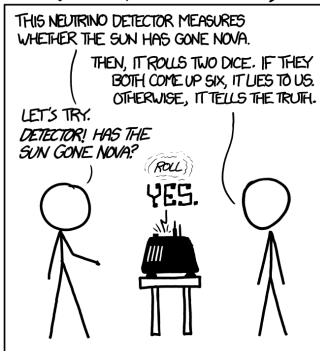
- ▶  $n \rightarrow \infty$  can never be achieved in practice. When is  $n$  large enough?
- ▶ Want to talk about probabilities of events that are not repeatable
  - ▶  $P(\text{rain tomorrow})$  — but there's only one tomorrow
  - ▶  $P(\text{Universe started with a big bang})$  — only one universe available
- ▶  $P$  is not an intrinsic property of  $A$ , but depends on how the ensemble of possible outcomes was constructed
  - ▶  $P(\text{person I talk to is a physicist})$  strongly depends on whether I am at a conference or at the beach

## ■ Criticisms of the subjective interpretation

- ▶ 'Subjective' estimate has no place in science
- ▶ How to quantify the prior state of our knowledge?

'Bayesians address the questions everyone is interested in by using assumptions that no one believes, while Frequentists use impeccable logic to deal with an issue that is of no interest to anyone' — Louis Lyons

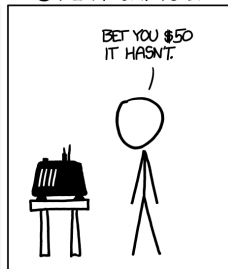
# DID THE SUN JUST EXPLODE? (IT'S NIGHT, SO WE'RE NOT SURE.)



FREQUENTIST STATISTICIAN:



BAYESIAN STATISTICIAN:



## **Describing data**



# Random variables and probability density functions

Random variable:

- Variable whose possible values are numerical outcomes of a random phenomenon

Probability density function (pdf) of a continuous variable:

$$P(X \text{ found in } [x, x + dx]) = f(x)dx$$

Normalisation:

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \quad x \text{ must be somewhere}$$

# Histograms

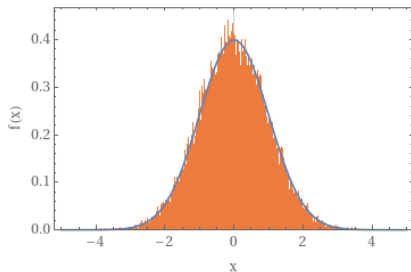
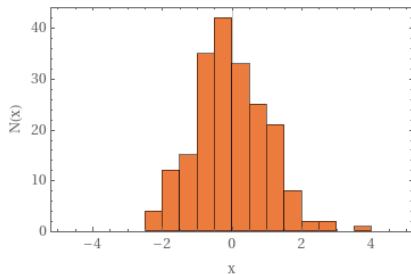
## Histogram

- representation of the frequencies of numerical outcome of a random phenomenon

pdf = histogram for

- infinite data sample
- zero bin width
- normalised to unit area

$$P(x) = \lim_{\Delta x \rightarrow 0} \frac{N(x)}{N\Delta x}$$



# Median, mean, and mode

Arithmetic **mean** of a data sample  
(‘sample mean’):

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

Mean of a pdf:

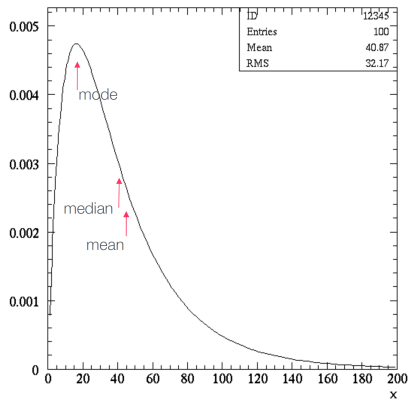
$$\begin{aligned} \mu &\equiv \langle x \rangle \equiv \int x f(x) dx \\ &\equiv \text{expectation value } E[x] \end{aligned}$$

## Median:

point with 50% probability above and  
50% prob. below

## Mode:

most likely value



not necessarily the same, for skewed  
distributions

# Variance, standard deviation

Variance of a distribution:

$$V(x) = \int dx P(x) (x - \mu)^2 = E[(x - \mu)^2]$$

Variance of a **data sample**

$$V(x) = \frac{1}{N} \sum_i (x_i - \mu)^2 = \overline{x^2} - \mu^2$$

Requires knowledge of *true* mean  $\mu$ . Replacing  $\mu$  by sample mean  $\bar{x}$  results in underestimated variance!

Instead, use this:

$$\hat{V}(x) = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

Standard deviation:

$$\sigma = \sqrt{V(x)}$$

# Multivariate distributions

Outcome of an experiment  
characterised by tuple  $(x_1, \dots, x_n)$

$$P(A \cap B) = \int f(x, y) dx dy$$

with  $f(x, y)$  the 'joint pdf'

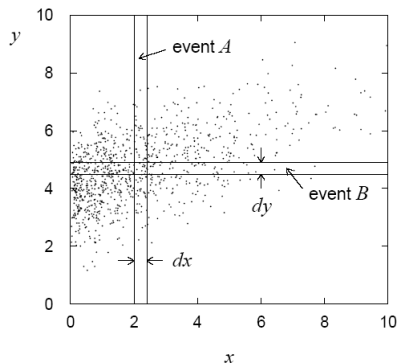
Normalisation

$$\int \dots \int f(x_1, \dots, x_n) dx_1 \dots dx_n = 1$$

Sometimes, only the pdf of one  
component is wanted:

$$f_1(x_1) = \int \dots \int f(x_1, \dots, x_n) dx_2 \dots dx_n$$

$\approx$  projection of joint pdf onto individual  
axis



# Covariance and correlation

Covariance:

$$\text{cov}[x, y] = E[(x - \mu_x)(y - \mu_y)]$$

Correlation coefficient:

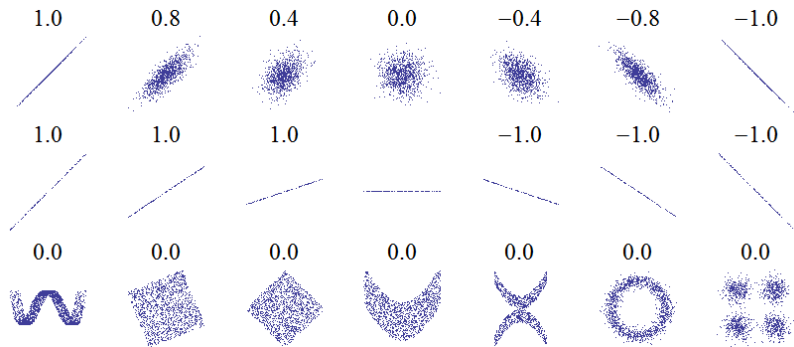
$$\rho_{xy} = \frac{\text{cov}[x, y]}{\sigma_x \sigma_y}$$

If  $x, y$  independent:

$$E[(x - \mu_x)(y - \mu_y)] = \int (x - \mu_x) f_x(x) dx \int (y - \mu_y) f_y(y) dy = 0$$

Note: converse not necessarily true

# Covariance and correlation



# Linear combinations of random variables

Consider two random variables  $x$  and  $y$  with known covariance  $\text{cov}[x, y]$

$$\langle x + y \rangle = \langle x \rangle + \langle y \rangle$$

$$\langle ax \rangle = a \langle x \rangle$$

$$V[ax] = a^2 V[x]$$

$$V[x + y] = V[x] + V[y] + 2 \text{cov}[x, y]$$

For uncorrelated variables, simply add variances.

How about combination of  $N$  independent measurements (estimates) of a quantity,  $x_i \pm \sigma$ , all drawn from the same underlying distribution?

$$\bar{x} = \frac{1}{N} \sum x_i \quad \text{best estimate}$$

$$V[N\bar{x}] = N^2 \sigma$$

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{N}} \sigma$$



# Combination of measurements: weighted mean

Suppose we have  $N$  independent measurements of the same quantity, but each with a different uncertainty:  $x_i \pm \delta_i$

Weighted sum:

$$x = w_1 x_1 + w_2 x_2$$
$$\delta^2 = w_1^2 \delta_1^2 + w_2^2 \delta_2^2$$

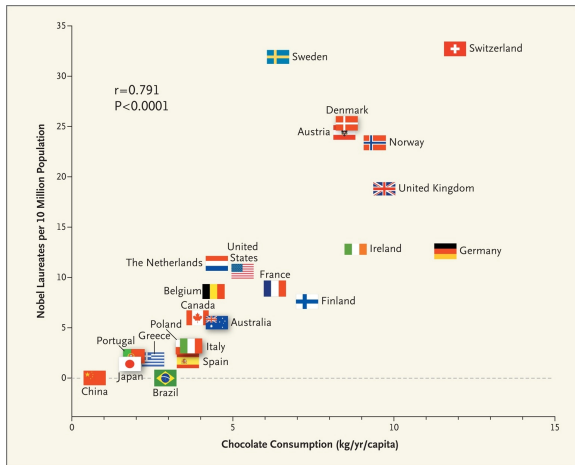
Determine weights  $w_1, w_2$  under constraint  $w_1 + w_2 = 1$  such that  $\delta^2$  is minimised:

$$w_i = \frac{1/\delta_i^2}{1/\delta_1^2 + 1/\delta_2^2}$$

If original raw data of the two measurements are available, can improve this estimate by combining raw data

alternatively, use log-likelihood curves to combine measurements

# Correlation $\neq$ causation



F. Messerli, N Engl J Med 2012; 367:1562

Correlation coefficient:  
0.791

significant correlation  
( $p < 0.0001$ )

0.4 kg/year/capita to  
produce one additional  
Nobel laureate

improved cognitive  
function associated  
with regular intake of  
dietary flavonoids?

## Some important distributions

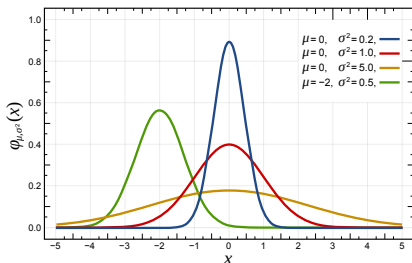
# Gaussian

A.k.a. normal distribution

$$g(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Mean:  $E[x] = \mu$

Variance:  $V[x] = \sigma^2$



Standard normal distribution:  $\mu = 0, \sigma = 1$

Cumulative distribution related to error function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz = \frac{1}{2} \left[ \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) + 1 \right]$$

In Python: `scipy.stats.norm(loc, scale)`

# $p$ -value

Probability for a Gaussian distribution corresponding to  $[\mu - Z\sigma, \mu + Z\sigma]$ :

$$P(Z\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-Z}^{+Z} e^{-\frac{x^2}{2}} = \Phi(Z) - \Phi(-Z) = \operatorname{erf}\left(\frac{Z}{\sqrt{2}}\right)$$

68.27% of area within  $\pm 1\sigma$

95.45% of area within  $\pm 2\sigma$

99.73% of area within  $\pm 3\sigma$

90% of area within  $\pm 1.645\sigma$

95% of area within  $\pm 1.960\sigma$

99% of area within  $\pm 2.576\sigma$

$p$ -value:

probability that random process  
(fluctuation) produces a measurement  
at least this far from the true mean

$$p\text{-value} := 1 - P(Z\sigma)$$

Available in ROOT: `TMath::Prob(Z*Z)`

and Python: `2*stats.norm.sf(Z)`

Deviation	$p$ -value (%)
$1\sigma$	31.73
$2\sigma$	4.55
$3\sigma$	0.270
$4\sigma$	0.006 33
$5\sigma$	0.000 057 3

# Why are Gaussians so useful?

Central limit theorem: sum of  $n$  random variables approaches Gaussian distribution, for large  $n$

True, if fluctuation of sum is not dominated by the fluctuation of one (or a few) terms

- **Good example:** velocity component  $v_x$  of air molecules
- **So-so example:** total deflection due to multiple Coulomb scattering.  
Rare large angle deflections give non-Gaussian tail
- **Bad example:** energy loss of charged particles traversing thin gas layer.  
Rare collisions make up large fraction of energy loss → Landau PDF

See practical part of today's lecture

# Binomial distribution

$N$  independent experiments

- Outcome of each is either 'success' or 'failure'
- Probability for success is  $p$

$$f(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k} \quad E[k] = Np \quad V[k] = Np(1-p)$$

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

binomial coefficient: number of permutations to have  $k$  successes in  $N$  tries

Use binomial distribution to model processes with two outcomes

Example: detection efficiency = #(particles seen) / #(all particles)

In the limit  $N \rightarrow \infty, p \rightarrow 0, Np = \nu = \text{const}$ , binomial distribution can be approximated by a Poisson distribution

# Poisson distribution

$$p(k; \nu) = \frac{\nu^k}{k!} e^{-\nu}$$

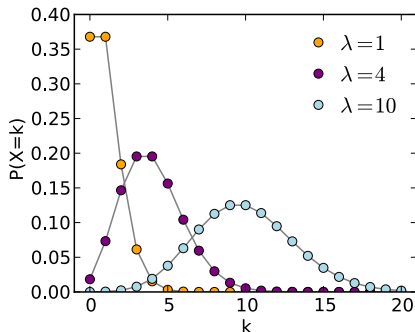
$$E[k] = \nu; \quad V[k] = \nu$$

Properties:

- If  $n_1, n_2$  follow Poisson distribution, then also  $n_1 + n_2$
- Can be approximated by Gaussian for large  $\nu$

Examples:

- Clicks of a Geiger counter in a given time interval
- Cars arriving at a traffic light in one minute



- Number of Prussian cavalymen killed by horse-kicks

Number of deaths in 1 corps in 1 year	Actual number of such cases	Poisson prediction
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6



# Uniform distribution

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

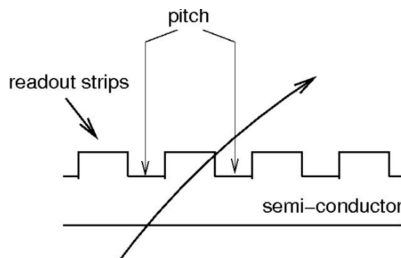
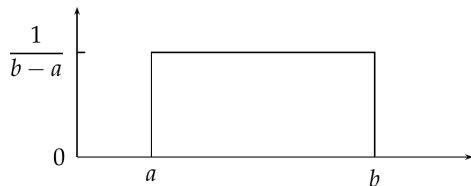
Properties:

$$E[x] = \frac{1}{2}(a + b)$$

$$V[x] = \frac{1}{12}(a + b)^2$$

Example:

- Strip detector:  
resolution for one-strip clusters:  
pitch /  $\sqrt{12}$



# Exponential distribution

$$f(x; \xi) = \begin{cases} \frac{1}{\xi} e^{-x/\xi} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E[k] = \xi; \quad V[k] = \xi^2$$

Example:

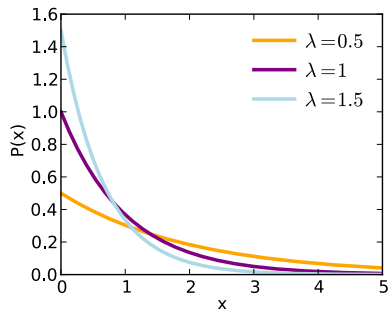
- Decay time of an unstable particle at rest

$$f(t; \tau) = \frac{1}{\tau} e^{-t/\tau}$$

$\tau$  = mean lifetime

Lack of memory (unique to exponential):  $f(t - t_0 | t \geq t_0) = f(t)$

Probability for an unstable nucleus to decay in the next minute is independent of whether the nucleus was just created or has already existed for a million years.



# $\chi^2$ distribution

Let  $x_1, \dots, x_n$  be  $n$  independent standard normal ( $\mu = 0, \sigma = 1$ ) random variables. Then the sum of their squares

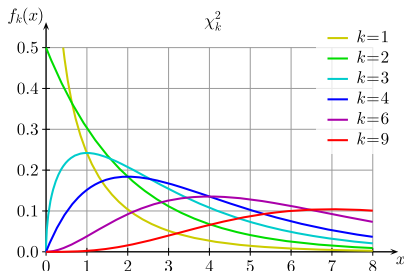
$$z = \sum_{i=1}^n x_i^2 = \sum_i \frac{(x_i' - \mu')^2}{\sigma'^2}$$

follows a  $\chi^2$  distribution with  $n$  degrees of freedom.

$$f(z; n) = \frac{z^{n/2-1}}{2^{n/2} \Gamma(\frac{n}{2})} e^{-z/2}, \quad z \geq 0$$

$$E[z] = n, \quad V[z] = 2n$$

Quantify goodness of fit, compatibility of measurements, ...



# Student's $t$ distribution

Let  $x_1, \dots, x_n$  be distributed as  $N(\mu, \sigma)$ .

Sample mean and  
estimate of variance:

$$\bar{x} = \frac{1}{n} \sum_i x_i, \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

Don't know true  $\mu$ , therefore have to estimate variance by  $\hat{\sigma}$ .

$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$  follows  $N(0, 1)$

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

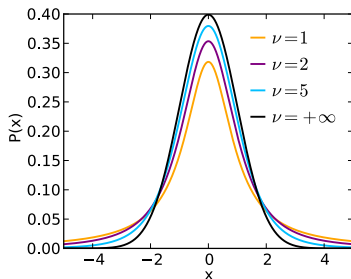
For  $n \rightarrow \infty$ ,  $f(t; n) \rightarrow N(t; 0, 1)$

Applications:

- Hypothesis tests: assess statistical significance between two sample means
- Set confidence intervals (more of that later)

$\frac{\bar{x} - \mu}{\hat{\sigma} / \sqrt{n}}$  not Gaussian.

Student's  $t$ -distribution with  $n - 1$   
d.o.f.



# Landau distribution

Describes energy loss of a (heavy) charged particle in a thin layer of material due to ionisation

tail with large energy loss due to occasional high-energy scattering, e.g. creation of delta rays

$$f(\lambda) = \frac{1}{\pi} \int_0^{\infty} \exp(-u \ln u - \lambda u) \sin(\pi u) du$$

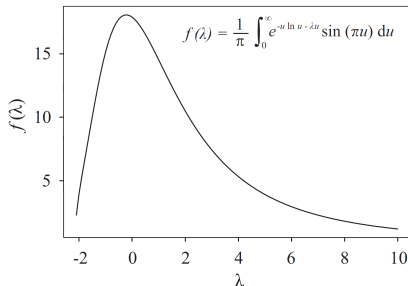
$$\lambda = \frac{\Delta - \Delta_0}{\xi}$$

$\Delta$ : actual energy loss

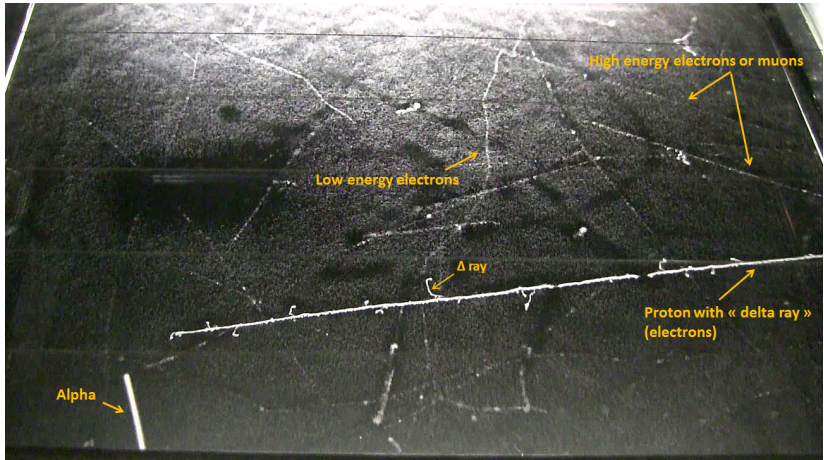
$\Delta_0$ : location parameter

$\xi$ : material property

Unpleasant: mean and variance (all moments, really) are not defined



# Delta rays



Julien SIMON, CC-BY-SA 3.0