LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

USM

University
Observatory
Munich

Joachim Puls, USM

Statistical Methods – an introduction

(Master of Science in Astrophysics, P2.2)

Stochastics: probability theory and statistics

- **Stochastics** [from ancient Greek: στόχαστικὴ τέχνη, (stochastike techne), Latin: ars coniectandi, i.e., the art of conjecture, “die Kunst des Mutmaßens”) comprises, as a generic term, the fields of **probability theory** and **statistics.**

- A **stochastic** (or random) **process** is one whose behavior is non-deterministic in that a system's subsequent state is determined both by the process's predictable actions and by a random element.

- **Statistics** is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. Statisticians improve the quality of data with the design of experiments and survey sampling. Statistics also provides tools for prediction and forecasting using data and statistical models. Statistics is applicable to a wide variety of academic disciplines, e.g., natural and social sciences and business.

- The word statistics can either be singular or plural. In its singular form, statistics refers to the mathematical science. In its plural form, statistics is the plural of the word statistic, which refers to a quantity (such as a mean) calculated from a set of data (see Chap. 7)

Content

Fundamental terms of statistics and data analysis, with examples from physics and astrophysics

1. Description of data
   - data types, binning, bar charts, histograms; averages, variance and standard deviation of a data-set
2. Probabilities
   - axioms, approaches, Bayes-theorem
3. Probability distributions functions – one random variable
   - discrete and continuous distributions, expectation value, central moments, variance, Chebychev's inequality, moments, characteristic function, cumulants, variable transformation
4. Distributions of several random variables –multivariate p.d.f.s
   - marginal distributions, convolution, moments, covariance and correlation, variable transformation, variable reduction, distributions with more than two variables
5. Important distributions and the CLT
   - binomial, multinomial, Poisson, uniform, normal, binormal, chi-squared, log-normal, central limit theorem (CLT)
6. Errors
   - measurement errors, error propagation, systematic errors
7. Estimation
   - random sampling, estimators (bias, consistency, efficiency), basic estimators, stratified sampling, finite population, likelihood (quotient and function), maximum likelihood (ML-) estimators, information inequality and minimum variance bound, minimum variance estimators, asymptotic properties of ML-estimators, errors on ML-estimators, covariances
8. Least squares
   - chi-squared minimization, fitting to a straight line ("linear regression"), variances and correlation, binned data, goodness of fit, errors on x and y, outlook on arbitrary linear and non-linear models
9. MCMC (Markov Chain Monte Carlo) – sampling the posterior
   - how to obtain the distribution and errors of model parameters given a set of measurements, and a corresponding example
10. Confidence intervals and hypothesis testing
    - confidence intervals (classical and based on likelihood function), errors of first and 2nd kind, significance, power, F-test, Student t-test, likelihood-ratio test, χ^2-test, Kolmogorov-Smirnov-test, Spearman-rank test, Wilcoxon rank sum test (Mann-Whitney U-test)

Literature

- R.J. Barlow, *Statistics*, John Wiley & Sons, 1989
- S. Brandt, *Statistical and Computational Methods in Data Analysis,* North-Holland, 1976 (2nd ed.)
- G. Bohm & G. Zech, *Einführung in Statistik und Messwertanalyse für Physiker,* Springer, 2007
- G. Bohm & G. Zech, *Einführung in Statistik und Messwertanalyse für Physiker,* Vorlesungsskript, DESY *http://www-library.desy.de/preparch/books/vstatmp.pdf*
- William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing,* Cambridge University Press, 2007

1. Description of data

Data types

- quantitative or numeric: can be written down as numbers
  - discrete (integers*; example*: numbers of seats in a car) or
  - continuous (cannot be recorded exactly, affected by rounding; *example*: length and weight of a car)
- qualitative or non-numeric: cannot be written down as number (*example*: color of a car;

  but many qualitative data can be transformed to numeric data, e.g., the RGB values of a

  color, or ‘red=1,green=2,blue=3,etc.’)

Binning

- to display information in a clear and concise way
- easy for discrete data;

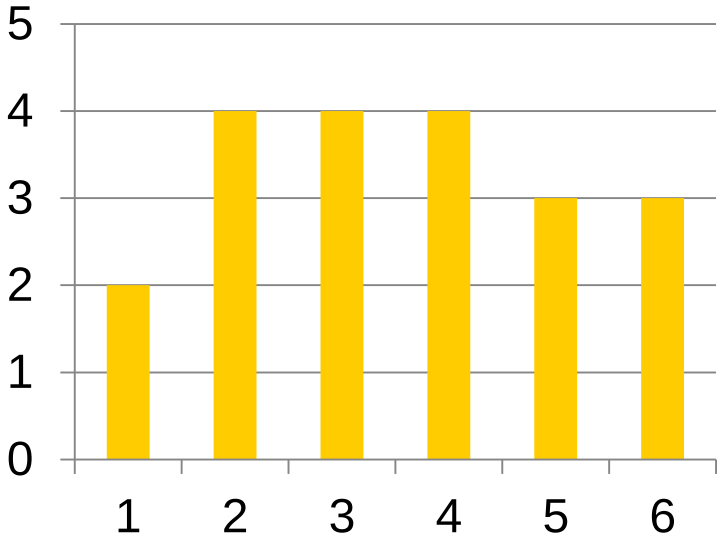
  example: results of throwing 20 times a die (see also next page)

  {4,3,6,1,3,4,5,2,6,2,4,3,1,2,6,5,2,4,3,5} →{2x1,4x2,4x3,4x4,3x5,3x6}
- more difficult for continuous data:
  - need to group adjacent numbers using a range of values to define a bin
  - further rounding and loss of precision
  - usually, uniform bin size, but sometimes also different sizes

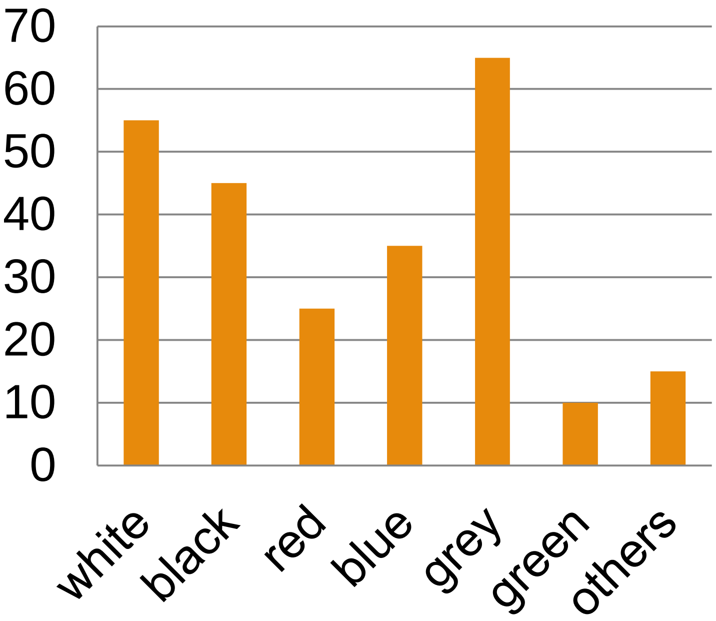
Bar charts …

- bar char: displayed number proportional to *length* of bar
  - can be used also for qualitative (non-numeric) data

results of 20 times throwing a dice

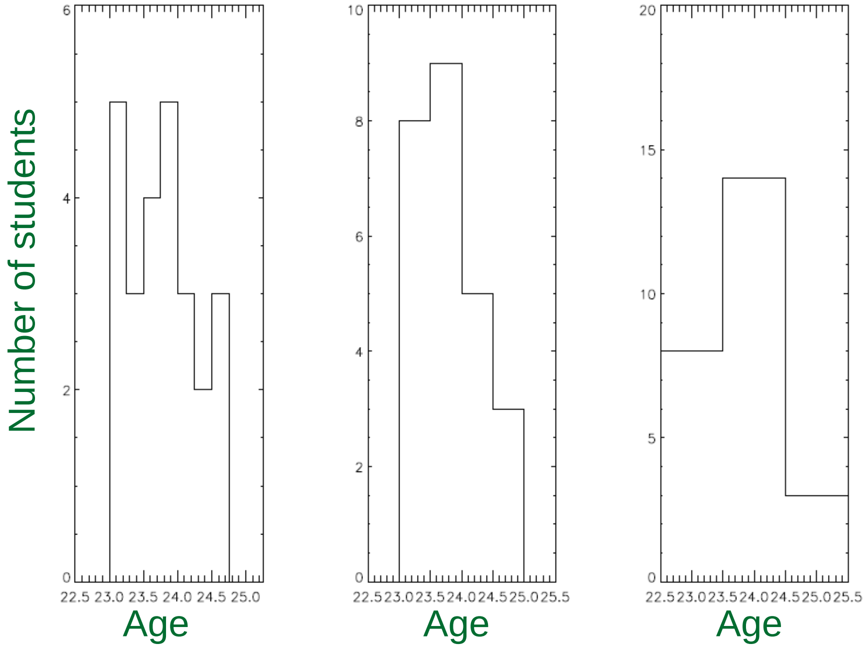


color distribution of a sample of 250 cars



… and histograms

- histogram: displayed number proportional to *area* of bin
  - important if bin size not uniform
  - can be used only for quantitative data, since bin-size has a (numerical) meaning
- bin size
  - if bins too narrow, very few events per bin, and distribution dominated by fluctuations
  - if bins too wide, real details become obscured
  - ideally,
    - at least 5 to10 events per bin
    - difference between contents of adjacent bins should be small



Example:
Age distribution (in years) within a group of students in their 2nd student year. Note the effect of choosing different bin sizes.
left: 0.25 yr, middle: 0.5 yr, right: 1.0 yr

Averages

- The arithmetic mean
  - best and most useful way to describe data by just one number (more later…)

If there is a set of data with N elements, then the mean value of x is

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

Analogously, the mean of any function $f(x)$ is given by

$$\overline{f} = \frac{1}{N}\sum_{i=1}^{N} f(x_i).$$

Don't confuse $\overline{f} = \overline{f(x)}$ (the mean of the function) with $f(\overline{x})$ (the function of the mean).

[see problem set 3]

If the data have been binned, and bin j corresponds to a value x_j (where x_j should be *centered*) and contains n_j data, the means can be written as

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{j\max} n_j x_j \quad \text{and} \quad \overline{f} = \frac{1}{N}\sum_{j=1}^{j\max} n_j f(x_j).$$

Note that in the upper case one summs over elements i, and in the lower over bins j!

Note also that there will be differences between means calculated from unbinned and binned data, mostly because of a finite bin size (different x_i, in the range $x_j \pm \Delta_j$, for the individual elements, and one x_j for all data in bin j).

Other means

- *geometric* mean: $\bar{x}_{geom} = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$ $\quad (x_i \geq 0)$

  inequality of arithmetic and geometrical mean (proof next slide):

  $\bar{x}_{geom} \leq \bar{x}_{arithm}$, equality only for $x_1 = x_2 = \ldots = x_N$

- *harmonic* mean: $\dfrac{N}{1/x_1 + 1/x_2 + \ldots + 1/x_N}$

  astrophys. example:
  Rosseland optical depth

- *root mean square* (rms): $\sqrt{\dfrac{x_1^2 + x_2^2 + \ldots + x_N^2}{N}}$

- *mode*: most popular value in the data set (or ‘highest’ bin)

- *median*: point with half of the data elements below and the other half above it. Certain subtleties involved in precise definition (see Chap. 3)

Geometric mean

- *geometric* mean: $\overline{x}_{geom} = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N} \qquad (x_i \geq 0)$

inequality of arithmetic and geometrical mean:

$\overline{x}_{geom} \leq \overline{x}_{arithm}$, equality only for $x_1 = x_2 = \ldots = x_N$

Proof (George Polya):

For any $x \geq 0$, $\exp(x) \geq 1 + x$ (Inequality for exp-function). Thus,

for arbitrary x_i : $\exp(\frac{x_i}{\overline{x}_{arithm}} - 1) \geq \frac{x_i}{\overline{x}_{arithm}}$. Apply for product involving all x_i,

$$\prod_{i=1}^{N} \exp\left(\frac{x_i}{\overline{x}_{arithm}} - 1\right) \geq \prod_{i=1}^{N} \frac{x_i}{\overline{x}_{arithm}}$$

$$\Rightarrow \prod_{i=1}^{N} \exp\left(\frac{x_i}{\overline{x}_{arithm}} - 1\right) = \exp\left(\frac{\sum_{i=1}^{N} x_i}{\overline{x}_{arithm}} - N\right) = \exp(N - N) = 1 \geq \prod_{i=1}^{N} \frac{x_i}{\overline{x}_{arithm}} = \left(\frac{\overline{x}_{geom}}{\overline{x}_{arithm}}\right)^N$$

$\Rightarrow \overline{x}_{arithm} \geq \overline{x}_{geom}$, q.e.d. $\quad$ (Proof for equality trivial, requires x=0 or $x_i = \overline{x}_{arithm} \ \forall i$)

Variance and standard deviation

- arithmetic mean describes data with one number
- need to describe the width or range of the distribution as well, i.e., the dispersion of the data about the mean
- average deviation from the mean not useful, since

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) = \bar{x} - \frac{N\bar{x}}{N} = 0$$

- force deviations to be always positive by squaring → *variance*

$$Var(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) =$$

$$= \overline{x^2} - 2\bar{x}\cdot\bar{x} + \frac{N\bar{x}^2}{N} = \overline{x^2} - \bar{x}^2, \text{ i.e., alternatively,}$$

$$Var(x) = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2; \quad \text{since by definition } Var(x) \geq 0 \quad \rightarrow \quad \overline{x^2} \geq \bar{x}^2$$

- $Var(f) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - \overline{f})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(f(x_i)\right)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} f(x_i)\right)^2$

and for binned data

$$Var(x) = \frac{1}{N}\sum_{j=1}^{j\max} n_j (x_j - \overline{x})^2 = \frac{1}{N}\sum_{i=1}^{j\max} n_j x_j^2 - \left(\frac{1}{N}\sum_{i=1}^{j\max} n_j x_j\right)^2$$

$$Var(f) = \frac{1}{N}\sum_{j=1}^{j\max} n_j (f(x_j) - \overline{f})^2 = \frac{1}{N}\sum_{i=1}^{j\max} n_j f^2(x_j) - \left(\frac{1}{N}\sum_{i=1}^{j\max} n_j f(x_j)\right)^2$$

Another way to force the deviations to be always positive is by using their absolute values

→ mean absolute deviation $= \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \overline{x}\right|$

rarely used, since difficult to treat (e.g., derivatives are 'horrible')

The standard deviation

- is the root mean squared deviation
- is the square root of the variance
- has similar units as x

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

For individual data sets (e.g., the age distribution of a group of students, see previous example)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2}$$

If the data set is a sample drawn from a 'parent distribution', then an *unbiased estimate* for the standard deviation *of the parent* is given by

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \sqrt{\frac{N}{N-1}\left(\overline{x^2} - \bar{x}^2\right)},$$

with $\bar{x}$ and $\overline{x^2}$ the 'conventional' arithmetic mean(s).

The origin of the denominator $(N-1)$ will be explained in Chap. 7

2. Probabilities

- Majority of predictions affected by uncertainties ("the only certain things in life are taxes and death").
  Thus, dealing with probabilities and statistics is sensible for everybody.
  Inevitable for experimental and empirical sciences.
  - accuracy of experiments restricted by precision of used devices
  - underlying processes often stochastic
  - estimates for measurement quantities and their accuracy required
  - estimates with errors enable to check hypotheses. Results can be improved subsequently, by adding new measurements and suitable averaging prescriptions.
  - statistics yields mathematical algorithms to conclude, from a certain sample, on the properties of the underlying parent population.

- Difference between observation and measurement:
  An observation (event) is the element of a sample (with one or more elements). Measurement is a *parameter estimate*, attributed with an *(in)accuracy*.
  - Example: Decay times for 10 pion decays (observations).
    The estimate of the decay rate is a measurement.
  - Fit to a straight line: observations are data points, slope is measurement.

- **example 1:** Polls allow to predict distribution of parliament seats. Parent population (Grundgesamtheit) is the entity of voters, the sample is a representative selection of them. Important to know the accuracy of the prediction.

- **example 2:** Determine the mean life time of an unstable nucleus, from the observation of 100 decays. Randomness induced by quantum mechanical effects. Sample representative for the entity of all possible decays, if experimental device able to measure all decay times (between zero and infinity) with sufficient precision.

- **example 3:** Determine the frequency of a pendulum, from 10 observations. The estimate for the actual frequency and its uncertainty are determined by suitable averages. It is assumed that the frequency can be determined with arbitrary precision for an infinite number of observations, and that a finite accuracy is the result of a restricted number of observations. Actual observations are a sample collected from an infinite number of possible observations.

- **example 4:** Test whether two experimental devices work similar. Compare samples from both devices. Test whether these samples originate from the same parent population.

Axioms of probability

- let S={E_1, E_2, E_3, …} be the set of possible results of an experiment = events.
- events are said to be *mutually exclusive* if it is impossible that both of them occur in one result.
- For every event E_i there is a probability P(E_i) which is a real number satisfying the axioms of probability (Kolmogorov 1950):

I. $P(E_i) \geq 0$

II. $P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)$ if E_1 and E_2 are mutually exclusive

III. $\sum_{\mathrm{i}} P(E_i) = 1,$ where the sum is over all mutually exclusive events

(simplified version of Kolmogorov's axioms)

- random events → probabilities

  event E → complementary event Ē (not E)

  from axiom III: $P(\bar{E}) = 1 - P(E)$

  and thus $P(E) \leq 1$

- A+B means A *or* B
- A·B means A *and* B;
  - if P(A·B) = 0, then A and B mutually exclusive
- random events can be described by random variables = variates
- a realization of a variate is an observation (event)

Empirical (classical) probabilities

- Frequency definition (frequentists' interpretation):
  In a large number N of experiments the event A is observed to occur n times. Then

$$P(A) = \lim_{N\to\infty} \frac{n}{N}$$

- The set of all N cases (N repetitions of the same experiment or N simultaneous identical experiments) is called the *collective* or *ensemble*
- In this case, the probability is not only a property of the experiment alone, but the joint property of experiment and ensemble
  - example (von Mises, 1957): German insurance companies found that the fraction of their male clients dying at the age of 40 is 1.1%
  - but this is not the probability that a particular man dies at this age. If data had been collected from other samples (all Germans, German hang-glider pilots,…), the outcome would have been different. Thus, the probability depends on the collective from which it has been taken.
- as well: experiments must be repeatable, under identical conditions.
  "What is the probability that is will rain tomorrow?"
  "Will the General Motors shares raise tomorrow?"
- and (old example):
  are we allowed to speak about the probability that, e.g., the mass of the Higgs particle lies in the range of 100 to 200 GeV/c^2

Objective probabilities

- Peirce (1910): probability is a property of device/ experiment, e.g. a die
- resurrected by Popper (in connection with quantum mechanics): objective probability or *propensity* (in German: Hang, Neigung)
- seems reasonable when, e.g., considering equally likely cases, e.g., due to symmetry (coin, die etc.)
- breaks down for continuous variables (transformation can make uniform, symmetric distribution non-uniform, and there is no natural choice for the "best" variable)

Subjective probability – Bayesian statistics/inference

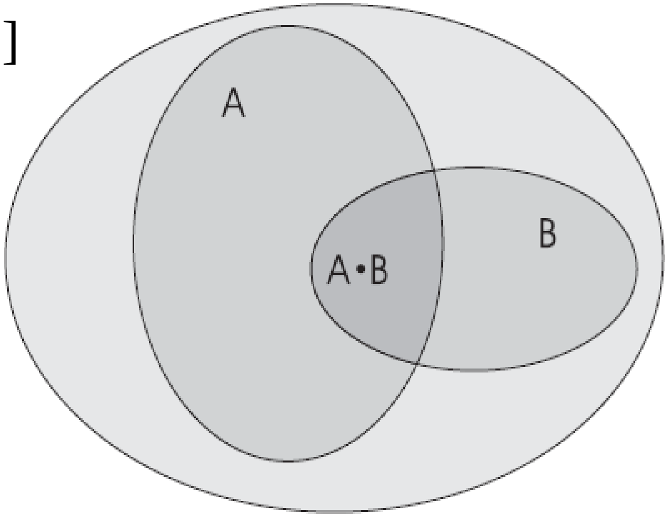
- *definition:* **conditional probability** P(A|B) is the probability of A *given B is true*

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

- implies: P(A·B) = P(A|B) P(B), reasonable
- *definition:* If P(A·B)=P(A) P(B), then the probabilities are **independent** of each other: in this case, P(A|B)=P(A)!
- **Bayes' theorem** (published posthumously by R. Price1763), undisputed:

$$\boxed{P(A|B)P(B)=P(B|A)P(A)}\quad [=P(A\cdot B)]$$

"Venn-diagram"



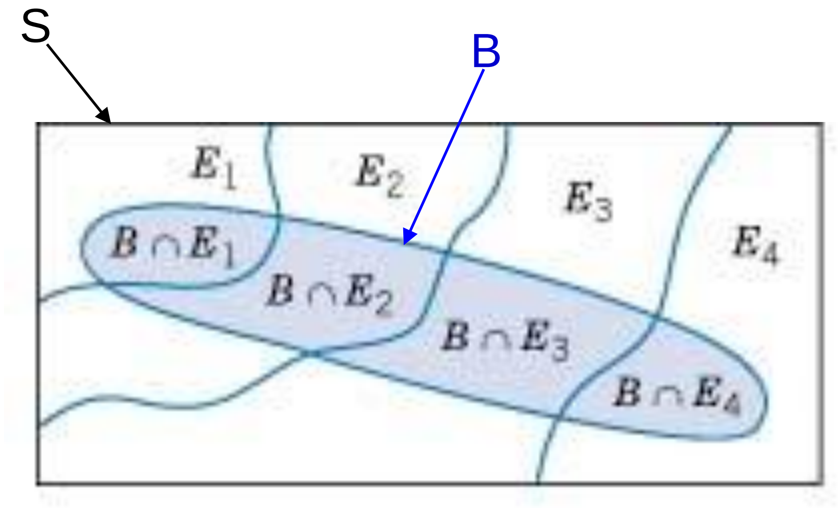
and also

$$P(A+B)=P(A)+P(B)-P(A\cdot B)$$

(generalization of axiom II for non mutually exclusive events)

Rule of total probability

- A collection of (sub-)sets or events E_1, E_2, …, E_k such that $E_1 \cup E_2 \cup E_3 \cup \ldots \cup E_k = S$ (combined, ‘union’) is said to be **exhaustive**



$$B = (B \cap E_1) \cup (B \cap E_2) \cup (B \cap E_3) \cup (B \cap E_4)$$

- Assume E_1, E_2, .., E_k are k mutually exclusive and exhaustive sets. Then

$$
\begin{aligned}
&P(B)\\
&= P(B \cap E_1) + P(B \cap E_2) + \ldots + P(B \cap E_k) \qquad \text{(‘intersection’)}\\
&= P(B \cdot E_1) + P(B \cdot E_2) + \ldots + P(B \cdot E_k)\\
&= P(B|E_1)P(E_1) + P(B|E_2)P(E_2) + \ldots +\\
&= \sum P(B|E_k)P(E_k).
\end{aligned}
$$

- Thus, if A_i mutual exclusive and exhaustive events

$$\boxed{P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{P(B)} = \frac{P(B \mid A_j)P(A_j)}{\sum_i P(B \mid A_i)P(A_i)}}$$

and / or

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})}$$

with $P(\bar{A}) = 1 - P(A)$

Examples

- **example 1:** probabilities for drawing certain cards from a well-shuffled card game with 32 cards
  P(Queen): 4/32 =1/8
  P(spade): 1/4
  P(spade|queen): 1/4 = P(spade) (independent events)
  P(Queen of spade): 1/8*1/4=1/32 (spade and queen)
  P(Queen or spade): 1/8+1/4-1/32=11/32 (not mutually exclusive)

- **example 2:** Calculate the fraction of female students, from the fraction of women and students in the population, and from the fraction of students among the female population

- P(A)=0.05 fraction of students in population
  P(B)=0.52 fraction of women in population
  P(A|B) = 0.07 fraction of students among female population

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.07 \cdot 0.52}{0.05} = 0.728$$

Examples (cont'd)

- **example 3**: **Infected?** (from Gigerenzer 2002, updated for actual numbers) see also very interesting video on the general problem, https://www.youtube.com/watch?v=lG4VkPoG3ko
- HIV-screening for persons without risky behavior
- positive test-result (D) with respect to two modern tests (ELISA, Western-Blot-Test)
- in Germany: H1: one of 10000 men HIV-infected (non risk-group)
- *P(D* **|** *H*1) = 0.999 that positive test (*D*) if man infected ("sensitivity")
- *P(D* **|** *H*2) = 0.0001 that positive test if *not* infected ("false positive")
- **Problem:** even if latter prob. is low, the total number of "false positives" can be large since majority of people is not infected, and can become comparable or even larger than total number of "true positives" (drawn from a much smaller sub-sample)
- Calculate P(*H*1 |*D*) that there is an actual infection if a man (non risk-group) tests positive

$$P(H1|D) = \frac{P(D|H1)P(H1)}{P(D)} = \frac{P(D|H1)P(H1)}{P(D|H1)P(H1)+P(D|H2)P(H2)}$$

$$= \frac{0.999\cdot 10^{-4}}{0.999\cdot 10^{-4}+0.0001\cdot(1-10^{-4})} = 0.4998!$$

$$\text{approximation: } P(D|H1)\approx 1 \approx P(H2) \quad \Rightarrow \quad P(H1|D) \approx \frac{1}{1+\dfrac{P(D|H2)}{P(H1)}};$$

$$P(D|H2) \ll P(H1): \text{test "OK"}$$

$$P(D|H2) = P(H1): \text{ prob. that actually infected = 0.5}$$

$$P(D|H2) \gg P(H1): \text{prob. that actually infected very low, } P(H1|D) \approx \frac{P(H1)}{P(D|H2)}$$

Bayesian statistics/inference

- so far, so good … (if all probabilities known, not disputed)
- but: applied also to statements which are regarded as ‘unscientific’ in the frequency definition.
- probability of a theory (it will rain tomorrow, parity is not violated…) is considered as a subjective ‘degree of belief’. Subsequent experimental evidence then modifies this initial degree of belief.
- expressed as

$$\underset{\text{‘posterior’}}{P(\text{theory}|\text{result})} = \frac{\overset{\text{‘likelihood’ (Chap. 8)}}{P(\text{result}|\text{theory})}}{\underset{\text{‘evidence’}}{P(\text{result})}}\,\underset{\text{‘prior’}}{P(\text{theory})}$$

$$\boxed{\text{posterior} = \frac{\text{likelihood}}{\text{evidence}}\ \text{prior}}$$

- What is the probability of a theory???
- if complete ignorance, uniform distribution assumed … (see example below, “The first night in paradise”)
  - otherwise, suitable choice due to symmetry arguments, laws of nature, empirical knowledge, experts opinion…
- But: with respect to which parameter? (example: mass or mass^2 give different priors)

- example: assume you toss a coin 3 times and obtain always “head”. Calculate probability that coin is a phoney, i.e., has a head on each side.

$$P(\text{phoney}|\text{3 heads}) = \frac{P(\text{3 heads}|\text{phoney})}{P(\text{3 heads}|\text{phoney})P(\text{phoney}) + P(\text{3 heads}|\text{ not phoney})(1 - P(\text{phoney}))} P(\text{phoney})$$

$$P(\text{3 heads}|\text{phoney}) = 1$$

$$P(\text{3 heads}|\text{ not phoney}) = \left(\frac{1}{2}\right)^3 = 0.125$$

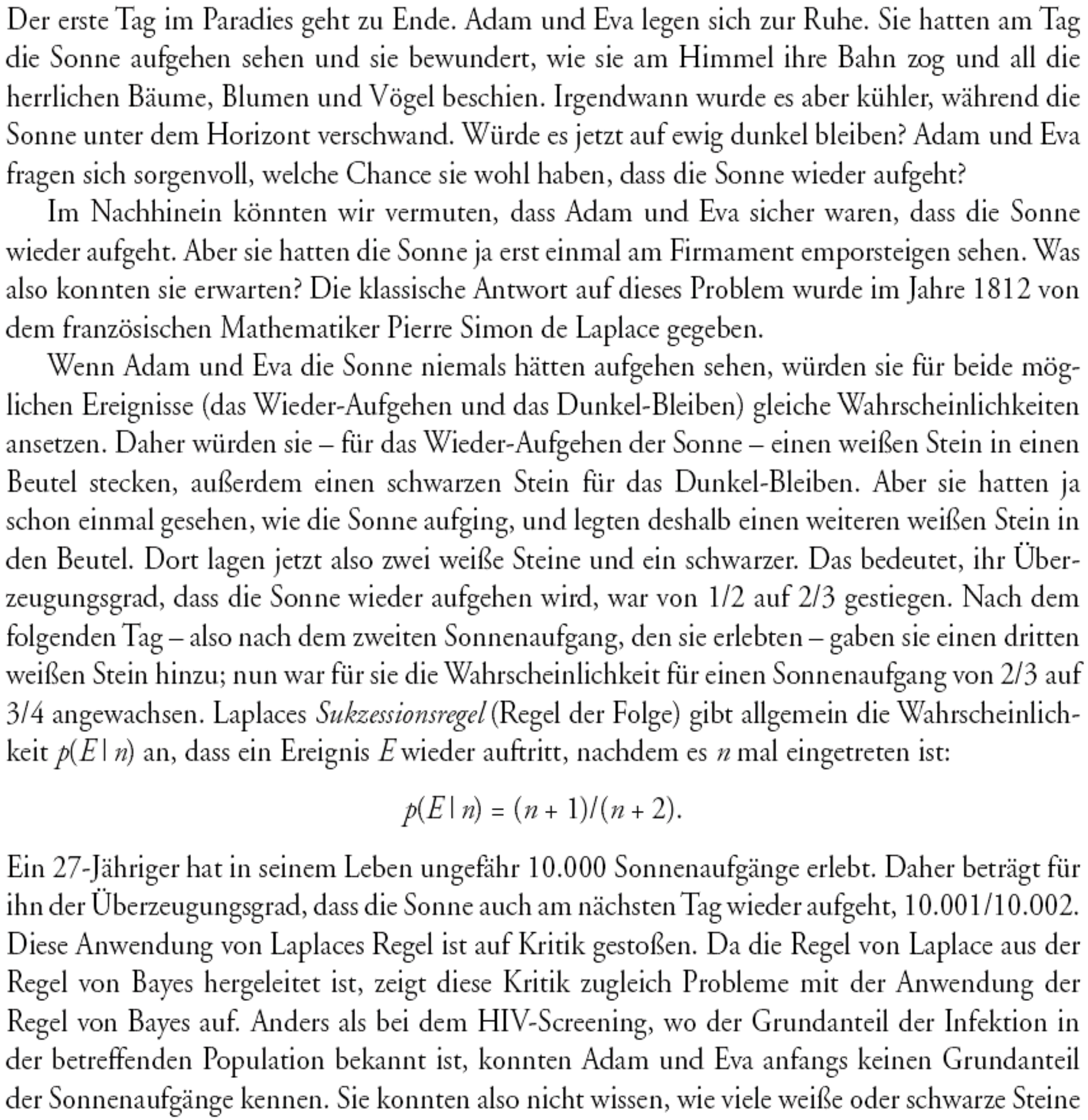
$$\text{prior}: P(\text{phoney}) = ???$$

- If you have drawn the coin from your pocket, the prior should be very small. Let P(phoney)=10^{-6}.Then
P(phoney|3 heads)=8· 10^{-6}, i.e., reasonably small

- Now assume that you have played against the car salesman Honest Eddi for a beer, and that Honest Eddi has given you the coin. In this case, the *a priori probability* that the coin is a phoney might be higher, you estimate 5%, and one finds P(phoney|3 heads)=0.3, which is a considerable chance.

The first night in paradise

From G. Gigerenzer 2004, “The evolution of statistic thinking”, Unterrichtswissenschaft, 32



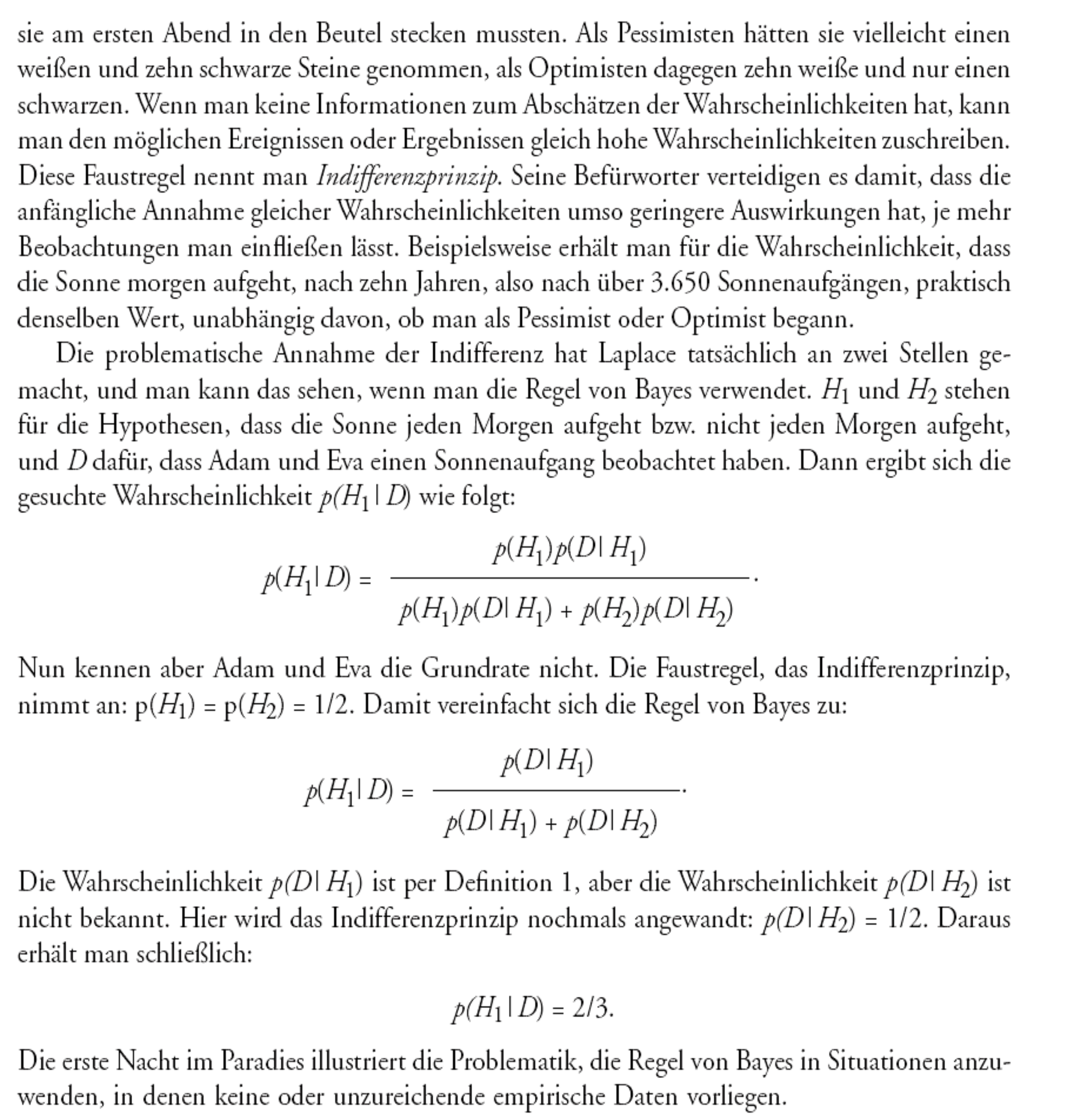
Der erste Tag im Paradies geht zu Ende. Adam und Eva legen sich zur Ruhe. Sie hatten am Tag die Sonne aufgehen sehen und sie bewundert, wie sie am Himmel ihre Bahn zog und all die herrlichen Bäume, Blumen und Vögel beschien. Irgendwann wurde es aber kühler, während die Sonne unter dem Horizont verschwand. Würde es jetzt auf ewig dunkel bleiben? Adam und Eva fragen sich sorgenvoll, welche Chance sie wohl haben, dass die Sonne wieder aufgeht?

Im Nachhinein könnten wir vermuten, dass Adam und Eva sicher waren, dass die Sonne wieder aufgeht. Aber sie hatten die Sonne ja erst einmal am Firmament emporsteigen sehen. Was also konnten sie erwarten? Die klassische Antwort auf dieses Problem wurde im Jahre 1812 von dem französischen Mathematiker Pierre Simon de Laplace gegeben.

Wenn Adam und Eva die Sonne niemals hätten aufgehen sehen, würden sie für beide möglichen Ereignisse (das Wieder-Aufgehen und das Dunkel-Bleiben) gleiche Wahrscheinlichkeiten ansetzen. Daher würden sie – für das Wieder-Aufgehen der Sonne – einen weißen Stein in einen Beutel stecken, außerdem einen schwarzen Stein für das Dunkel-Bleiben. Aber sie hatten ja schon einmal gesehen, wie die Sonne aufging, und legten deshalb einen weiteren weißen Stein in den Beutel. Dort lagen jetzt also zwei weiße Steine und ein schwarzer. Das bedeutet, ihr Überzeugungsgrad, dass die Sonne wieder aufgehen wird, war von 1/2 auf 2/3 gestiegen. Nach dem folgenden Tag – also nach dem zweiten Sonnenaufgang, den sie erlebten – gaben sie einen dritten weißen Stein hinzu; nun war für sie die Wahrscheinlichkeit für einen Sonnenaufgang von 2/3 auf 3/4 angewachsen. Laplaces *Sukzessionsregel* (Regel der Folge) gibt allgemein die Wahrscheinlichkeit $p(E \mid n)$ an, dass ein Ereignis E wieder auftritt, nachdem es n mal eingetreten ist:

$$p(E \mid n) = (n + 1)/(n + 2).$$

Ein 27-Jähriger hat in seinem Leben ungefähr 10.000 Sonnenaufgänge erlebt. Daher beträgt für ihn der Überzeugungsgrad, dass die Sonne auch am nächsten Tag wieder aufgeht, 10.001/10.002. Diese Anwendung von Laplaces Regel ist auf Kritik gestoßen. Da die Regel von Laplace aus der Regel von Bayes hergeleitet ist, zeigt diese Kritik zugleich Probleme mit der Anwendung der Regel von Bayes auf. Anders als bei dem HIV-Screening, wo der Grundanteil der Infektion in der betreffenden Population bekannt ist, konnten Adam und Eva anfangs keinen Grundanteil der Sonnenaufgänge kennen. Sie konnten also nicht wissen, wie viele weiße oder schwarze Steine



sie am ersten Abend in den Beutel stecken mussten. Als Pessimisten hätten sie vielleicht einen weißen und zehn schwarze Steine genommen, als Optimisten dagegen zehn weiße und nur einen schwarzen. Wenn man keine Informationen zum Abschätzen der Wahrscheinlichkeiten hat, kann man den möglichen Ereignissen oder Ergebnissen gleich hohe Wahrscheinlichkeiten zuschreiben. Diese Faustregel nennt man *Indifferenzprinzip*. Seine Befürworter verteidigen es damit, dass die anfängliche Annahme gleicher Wahrscheinlichkeiten umso geringere Auswirkungen hat, je mehr Beobachtungen man einfließen lässt. Beispielsweise erhält man für die Wahrscheinlichkeit, dass die Sonne morgen aufgeht, nach zehn Jahren, also nach über 3.650 Sonnenaufgängen, praktisch denselben Wert, unabhängig davon, ob man als Pessimist oder Optimist begann.

Die problematische Annahme der Indifferenz hat Laplace tatsächlich an zwei Stellen gemacht, und man kann das sehen, wenn man die Regel von Bayes verwendet. H_1 und H_2 stehen für die Hypothesen, dass die Sonne jeden Morgen aufgeht bzw. nicht jeden Morgen aufgeht, und D dafür, dass Adam und Eva einen Sonnenaufgang beobachtet haben. Dann ergibt sich die gesuchte Wahrscheinlichkeit $p(H_1 | D)$ wie folgt:

$$p(H_1 | D) = \frac{p(H_1)p(D | H_1)}{p(H_1)p(D | H_1) + p(H_2)p(D | H_2)}.$$

Nun kennen aber Adam und Eva die Grundrate nicht. Die Faustregel, das Indifferenzprinzip, nimmt an: $\mathrm{p}(H_1) = \mathrm{p}(H_2) = 1/2$. Damit vereinfacht sich die Regel von Bayes zu:

$$p(H_1 | D) = \frac{p(D | H_1)}{p(D | H_1) + p(D | H_2)}.$$

Die Wahrscheinlichkeit $p(D | H_1)$ ist per Definition 1, aber die Wahrscheinlichkeit $p(D | H_2)$ ist nicht bekannt. Hier wird das Indifferenzprinzip nochmals angewandt: $p(D | H_2) = 1/2$. Daraus erhält man schließlich:

$$p(H_1 | D) = 2/3.$$

Die erste Nacht im Paradies illustriert die Problematik, die Regel von Bayes in Situationen anzuwenden, in denen keine oder unzureichende empirische Daten vorliegen.

The game show problem

Adapted from Christian Rieck
www.spieltheorie.de/Spieltheorie_Anwendungen/ziegenproblem.htm

See also http://en.wikipedia.org/wiki/Monty_Hall_problem

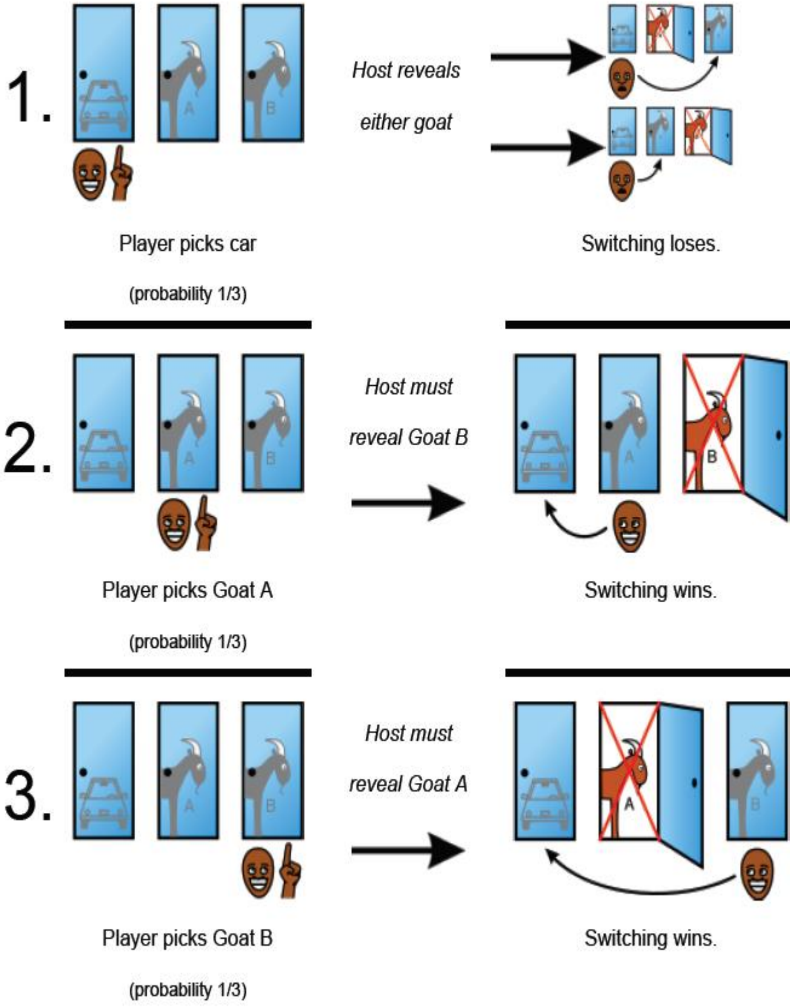
- Das Ziegenproblem (auch game-show problem, Monty-Hall problem) ist eines der Probleme, das die Gemüter lange Zeit erhitzt hat und ganze Scharen von Mathematikern an den Rand der Verzweiflung gebracht hat (insbesondere, weil sie von ihrer Intuition irregeführt wurden und lange gebraucht haben, das zu bemerken). Es gibt wohl keinen Spieltheoretiker, der Ende der 1980er Jahre nicht in irgendeiner Form über dieses Problem nachgedacht hat.

- In einer amerikanischen Quizsendung steht eine Kandidatin vor drei verschlossenen Türen, hinter denen in einem Fall ein Auto steht und in zwei Fällen eine Ziege. Die Kandidatin darf jetzt eine der Türen wählen; anschließend öffnet der Showmaster eine der verbleibenden zwei Türen, und zwar immer so, dass auf jeden Fall eine Tür mit Ziege geöffnet wird, sodass das Auto also hinter einer der noch verschlossenen Türen sein muss. Er bietet der Kandidatin dann an, jetzt noch einmal die Türe zu wechseln oder bei der zuerst gewählten Tür zu bleiben, bevor sie geöffnet wird. Die Kandidatin bekommt dann das, was hinter der von ihr endgültig gewählten Tür steht (wobei wir hier davon ausgehen wollen, dass sie das Auto der Ziege vorzieht).

- In einer Kolumne von Marilyn vos Savant (www.marilynvossavant.com/articles/gameshow.html) stellte jemand die Frage, ob es in dieser Situation besser sei zu wechseln oder bei der ursprünglichen Wahl zu bleiben. Die meisten Menschen dachten damals, dass es egal sein müsse. Marilyn vos Savant, die den höchsten jemals gemessenen IQ hat und daher als der intelligenteste Mensch der Welt gilt, antwortete allerdings lapidar mit "wechseln ist besser" und löste damit die Diskussion aus, in der es Wochen dauerte, bis sich die Menschheit auf die bis heute akzeptierte Lösung einigen konnte. Davor bekam sie allerdings so nette Zuschriften wie: "Sie sind die Ziege!", oder: "Sie haben einen Fehler gemacht. ... Wenn sich all diese Doktoren irren würden, dann wäre unser Land in ernsthaften Schwierigkeiten." Aber wenigstens ist der intelligenteste Mensch der Welt dadurch berühmt geworden.

The game show problem (cont'd)

Popular solution
The player has an equal chance of initially selecting the car, Goat A, or Goat B. Switching results in a win 2/3 of the time.



From http://en.wikipedia.org/wiki/Monty_Hall_problem

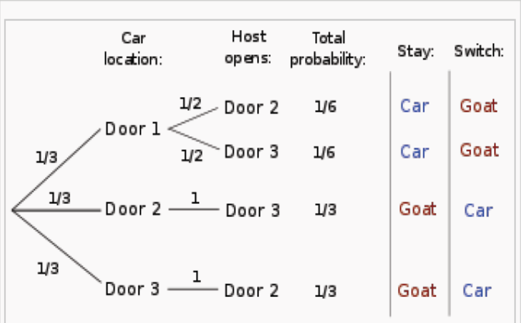
The game show problem (cont'd)

- Dabei lässt sich das Problem durch Anwendung des Satzes von Bayes lösen. Darin ist A_i das unbeobachtbare Ereignis (wo steht das Auto?) und Q_j ist die Beobachtung (welche Tür öffnet der Quizmaster?).
  - Ai: das Auto steht hinter Tür i
  - Qj: Quizmaster öffnet Tür j
  - O.E.d.A. nehmen wir an, dass der Kandidat Tür 1 wählt, und der Quizmaster Tür 2 öffet (-> Q2)
  - wir wollen die Wahrscheinlichkeit für den Gewinn mit Türwechsel berechnen, also P(A3|Q2), kennen aber nur die bedingten Wahrscheinlichkeiten P(Qj|Ai).
    Mit Hilfe des Bayesschen Theorems und der „rule of total probability“ (A1,A2,A3 sind erschöpfend und sich gegenseitig ausschließend) finden wir

$$P(A3\,|\,Q2)=\frac{P(Q2\,|\,A3)P(A3)}{P(Q2\,|\,A3)P(A3)+P(Q2\,|\,A1)P(A1)+P(Q2\,|\,A2)P(A2)}$$

  - P(Ai)=1/3
  - P(Q2|A3)=1 (der Quizmaster muss Tür 2 öffnen)
  - P(Q2|A1)=0.5 (er könnte auch Tür 3 öffnen)
  - P(Q2|A2)=0 (das ist „verboten“)
  - insgesamt resultiert also

$$P(A3\,|\,Q2)=\frac{1\cdot\frac{1}{3}}{1\cdot\frac{1}{3}+0.5\cdot\frac{1}{3}+0\cdot\frac{1}{3}}=\frac{2}{3}$$



Tree showing the probability of every possible outcome if the player initially picks Door 1

From http://en.wikipedia.org/wiki/Monty_Hall_problem

The game show problem (cont’d)

- Die Wahrscheinlichkeit, dass das Auto hinter der Tür steht, die der Quizmaster geschlossen lässt, beträgt somit 2/3, wogegen sie hinter der ursprünglichen Tür nur 1/3 beträgt. Somit ist klar, dass man seine Chancen auf das Auto verdoppelt, wenn man wechselt. Vos Savant hatte also Recht.

- Beachte: der Quizmaster verhält sich wie ein rein ausführender Algorithmus, der nur eine Tür öffnet, hinter der mit Sicherheit kein Auto steht.

- Falls der Quizmaster nicht wüsste, wo das Auto steht, wäre $P(Q2|Ai) =1/3$ für alle $i=1,3$, und damit $P(A3|Q2)=P(A2|Q2)=1/3$. In diesem Fall würde also Wechseln nichts bringen!

3. Probability distributions – one random variable

- random events are characterized by random variables
- Probability distribution functions associate random variables with corresponding probabilities
- discrete and continuous random variables (r.v.)
- in the following, probabilities refer to *one* r.v. x, i.e., one property which can be quantified.

definition: **(cumulative) distribution function** (c.d.f) F(*t*) defines the probability of finding a random variable x being smaller than *t*,

$$F(t) = P(\mathrm{x} < t) \quad \text{with} \quad -\infty < t < \infty$$

  - here and in the following we denote a random variable by non-italics and "ordinary" variables in italics **whenever a confusion might be possible** (otherwise, we provide no distinction).

- from the probability axioms, we obtain the following properties for F(*t*)
  - F(*t*) increases monotonically with t
  - F(-∞)=0
  - F(∞) =1
- There are discrete and continuous distributions

Discrete distributions

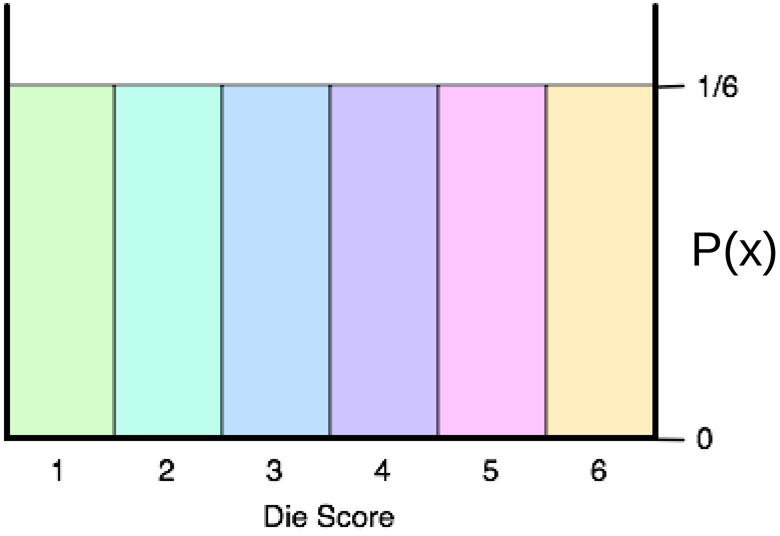
$$P(x{=}x_i) = F(x_i + \varepsilon) - F(x_i - \varepsilon)$$

- describe probabilities for the occurrence of N discrete, different events, with

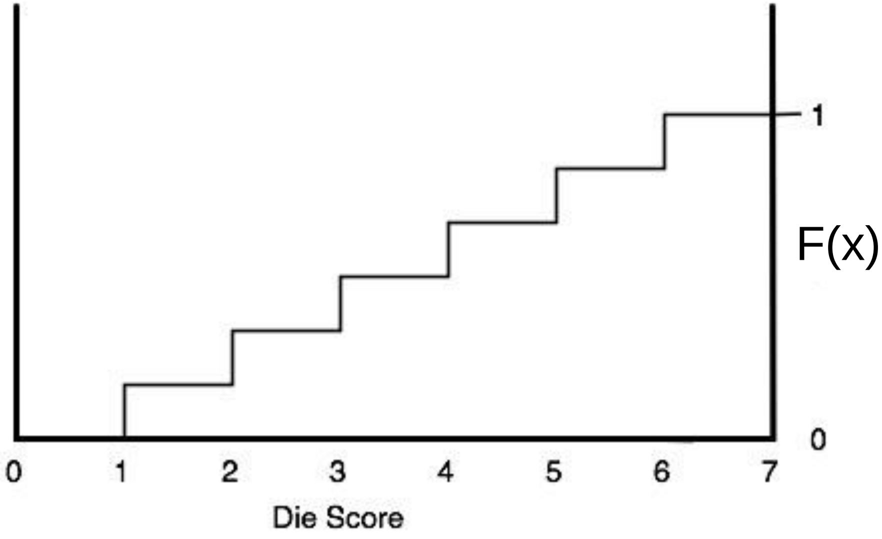
$$\sum_i P(x_i) = 1$$

- **example**: die; the probability to dice a certain number x_i is $P(x_i) = 1/6$, $x_i = i$ for $i = 1,6$

- discrete distributions can be treated as continuous distributions, via the Dirac δ-function



probability distribution



distribution function

Continuous distributions

- instead of probability distribution, define **probability density** *f(x)* (p.d.f. = prob. density function) with

$$f(x) = \frac{dF(x)}{dx}$$

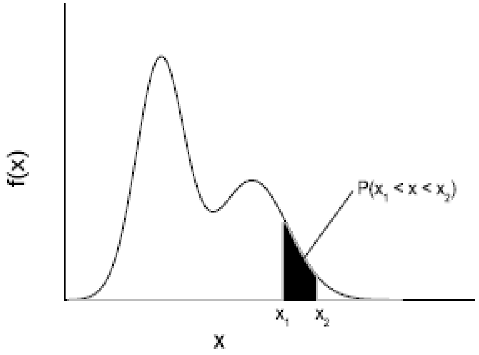
and properties

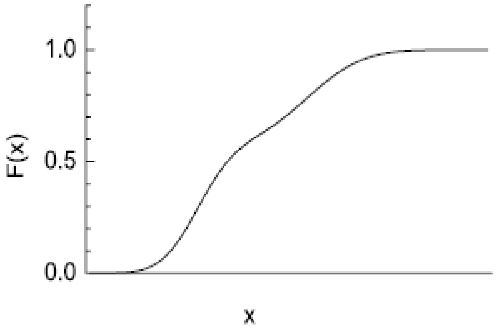
$$f(-\infty) = f(+\infty) = 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1; \qquad \text{thus,}$$

$$P(a \le \mathrm{x} < b) = F(b) - F(a) = \int_a^b f(x)dx \quad \text{and}$$

$f(x)dx$ is probability that x in the interval $[x, \mathrm{x} + dx]$





- example: life-times of instable particles follow an exponential distribution

$$f(t) = \frac{\exp(-t/\tau)}{\tau} \quad \text{for t} \geq 0 \text{ and with mean life-time } \tau$$

$$\Rightarrow$$

$$F(t) = \int_{-\infty}^{t} f(t')dt' \rightarrow \int_{0}^{t} f(t')dt' = 1 - \exp(-t/\tau),$$

and the probability that the particle lives longer than τ is

$$P(\text{t} > \tau) = F(\infty) - F(\tau) = \exp(-1)$$

Expectation value

- note: if x is a r.v., than any function u(x) is a r.v. as well
- distributions have characteristic parameters such as expectation value, width and asymmetry

- the *expectation value* or *mean of a r.v.* x results from averaging over x according to its distribution,

$$E(\mathrm{x}) = <\mathrm{x}> = \mu = \begin{cases} \sum_i x_i P(\mathrm{x} = x_i) & \text{discrete dist.} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{continuous dist.} \end{cases}$$

$$E(\mathrm{u}(\mathrm{x})) = <\mathrm{u}> = \begin{cases} \sum_i u(x_i) P(\mathrm{x} = x_i) & \text{discrete dist.} \\ \int_{-\infty}^{\infty} u(x) f(x) dx & \text{continuous dist.} \end{cases}$$

- Note that the expectation value (and similar functions) is **not** a random variable but clearly defined

- calculation rules: let α,β be constants and u and v functions of x

$$E(\alpha)=\alpha; \quad E(E(\mathrm{u}))=E(\mathrm{u})$$

$$E(\alpha\mathrm{u}+\beta\mathrm{v})=\alpha E(\mathrm{u})+\beta E(\mathrm{v});$$

the expection value is a linear operator!

if x,y are independent r.v., then

E(u(x)v(y))=E(u)E(v) (see Chap. 4)

- the expectation value is the centre of gravity of the distribution

Central moments of a r.v.

- Let's choose especially

$\mathrm{u(x)} = (\mathrm{x} - \mu)^n$ with $E(\mathrm{u(x)}) =: \mu'_n = E\{(\mathrm{x} - \mu)^n\}$

which is called the n-th central moment or

the n-th moment about the mean (the latter fact denoted by the prime).

Lowest order central moments are

$\mu'_0 = 1$ and $\mu'_1 = 0$

The quantity

$\mu'_2 = Var(\mathrm{x}) = \sigma^2(\mathrm{x}) = E\{(\mathrm{x} - \mu)^2\}$

is the lowest central moment which contains information about the average deviation of x from the mean.

It's called the variance of x, and σ is the standard deviation

Variance

- the variance measures the mean quadratic deviation from the mean.
- the standard deviation σ=√ Var has the same units as x, will be identified with the errors of measurements
- the mechanical analogue to the variance is the moment of inertia
- calculation rules

$$Var(\alpha) = 0, \quad \boxed{Var(\alpha \mathrm{x}) = \alpha^2 Var(\mathrm{x})}$$

$$Var(\alpha \mathrm{x} + \beta \mathrm{y}) = \alpha^2 Var(\mathrm{x}) + \beta^2 Var(\mathrm{y}) \quad \text{if x,y are independent} \quad \text{(see Chap. 4)}$$

- different representation

$$Var(\mathrm{x}) = E\{(\mathrm{x}-\mu)^2\} = E(\mathrm{x}^2 - 2\mathrm{x}\mu + \mu^2) =$$

$$= E(\mathrm{x}^2) - 2\mu^2 + \mu^2 =$$

$$= E(\mathrm{x}^2) - \mu^2 \quad \text{or} \quad <\mathrm{x}^2> - <\mathrm{x}>^2$$

The variance (and all other central moments)

is invariant to translations of the r.v.!!!

Variance of a convolution

- We measure a quantity x with pdf g, and the measurement is ‘smeared out’ according to a pdf h (→ convolution, see Chap. 4).
- We look for the variance of the (‘combined’) measurement x’.
- Alternative interpretation: x’ is the sum of two r.v., x’=x+u, where x is distributed according to g(x), and u according to h(u).
- Then, x’ is distributed according to (see page 73)

$$f(x') = \int g(x)h(x'-x)dx.$$

with (see also problem set 2)

$$\left.\begin{array}{l} E(\mathrm{x}') = < \mathrm{x}' > = < \mathrm{x} >_g + < \mathrm{u} >_h \\ \text{and} \\ Var(\mathrm{x}') = Var_g(\mathrm{x}) + Var_h(\mathrm{u}) \end{array}\right\} \left(\begin{array}{l} \text{consistent with "calculation rules"} \\ \text{for independent variates} \end{array}\right)$$

- The variance of x’ is the sum of the variances of the distributions g and h. For sequential measurements of a quantity the individual errors add quadratically (see Chap. 6)

Skewness

- measures the asymmetry of a distribution

$$\gamma_1 = \frac{\mu'_3}{\sigma^3} = E\{(\mathrm{x}-\mu)^3\}/\sigma^3 = \ldots$$

$$= \frac{\mathrm{E}(\mathrm{x}^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

skewness is invariant to translations and elongations

sometimes one finds $\beta_1 = (\gamma_1)^2$

a positive skew describes a distribution with a tail which extends to the right.

Curtosis/Kurtosis

- measures how pronounced the tails of the distribution are

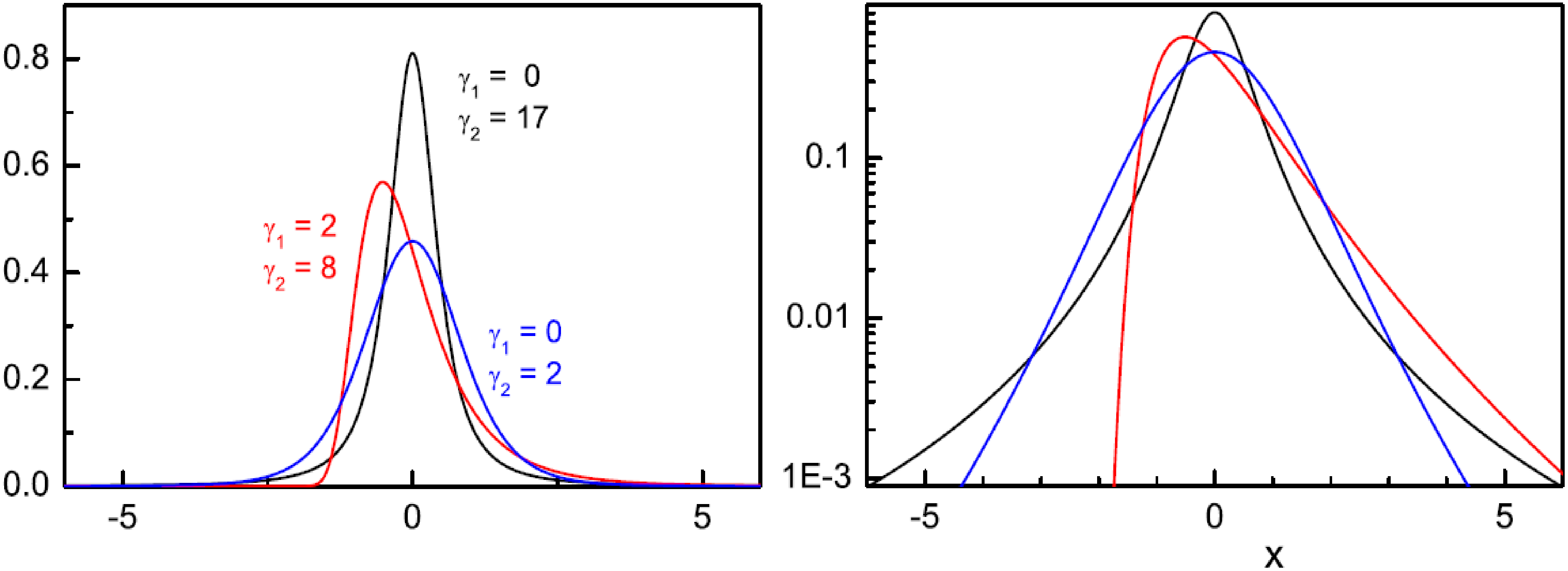
$$\beta_2 = \frac{\mu'_4}{\sigma^4} = E\{(\mathrm{x}-\mu)^4\}/\sigma^4 = \ldots$$

$$= \frac{E(\mathrm{x}^4) - 4E(\mathrm{x}^3)\mu + 6E(\mathrm{x}^2)\mu^2 - 3\mu^4}{\sigma^4}$$

$\gamma_2 = \beta_2 - 3$ is defined in such a way as to be zero for a normal=Gaussian distribution

positive γ_2 implies a relatively higher, narrower peak and wider wings than the normal distribution with same mean and σ, and vice versa (wider peak, shorter wings) for negative γ_2.

Examples



3 different p.d.f, all with zero mean and unit variance, but different skewness and curtosis.
Left: linear scale; right: logarithmic scale.

Let $\mathrm{u}(\mathrm{x}) = \frac{\mathrm{x} - \mu}{\sigma}$. Then $E(\mathrm{u}) = 0$ and $Var(\mathrm{u}) = \frac{1}{\sigma^2} Var(\mathrm{x} - \mu) = \frac{\sigma^2}{\sigma^2} = 1$

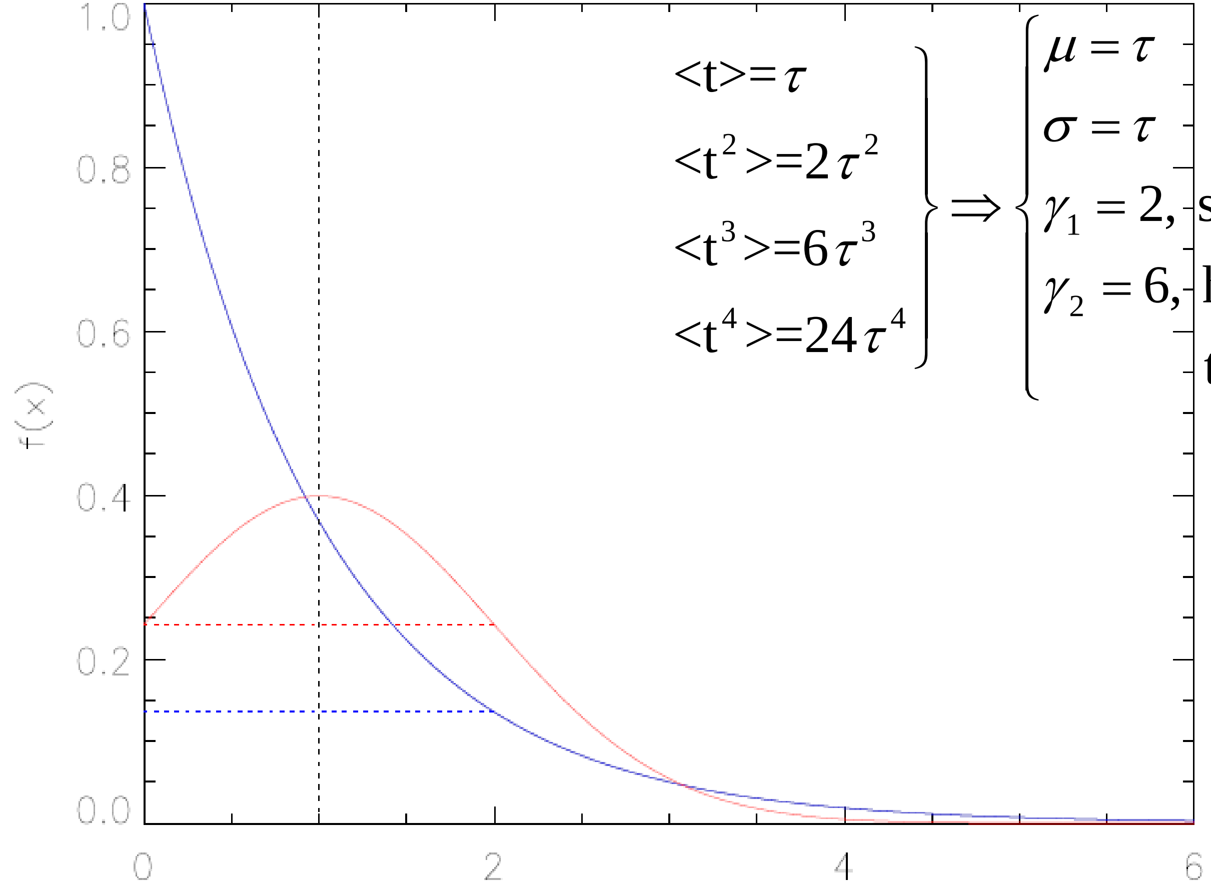
The r.v. u has particularly simple properties,

and is called a reduced (normalized) variable

Examples (cont'd)

life-time (exponential) distribution

$$<\mathrm{t}^{\mathrm{n}}> = \int_0^\infty \frac{t^n}{\tau}\exp(-t/\tau)dt = n!\tau^n$$



$$\left.\begin{array}{l} <\mathrm{t}>=\tau \\ <\mathrm{t}^2>=2\tau^2 \\ <\mathrm{t}^3>=6\tau^3 \\ <\mathrm{t}^4>=24\tau^4 \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \mu=\tau \\ \sigma=\tau \\ \gamma_1=2,\ \text{skewed, with tail to the right} \\ \gamma_2=6,\ \text{higher peak and wider wings} \\ \qquad\quad \text{than normal dist.} \end{array}\right.$$

life-time (blue, τ=1) and normal (red, μ=σ=1) distribution,
both distributions have identical mean and variance (indicated by dotted lines)

Other parameters of a distribution

mode x_m : $P(\mathrm{x} = x_m) = \max$

if distribution has a differentiable probability density, then the mode is determined via

$$\frac{d}{dx} f(x) = 0, \quad \frac{d^2}{dx^2} f(x) < 0$$

if one maximum, distr. unimodal, otherwise multimodal

median $x_{0.5}$:

a) continuous data: $\mathrm{F}(x_{0.5}) = P(\mathrm{x} < x_{0.5}) = 0.5,$ i.e., $\int_{-\infty}^{x_{0.5}} f(x)dx = 0.5$

For a continous distribution, the median divides the total range of x into two regions of equal probability.

Example : half-live $t_{0.5}$ of a radio-active nucleus, which yields the time after which 50% of the nuclei have decayed. With respect to the exponential distribution, $t_{0.5} = \tau \ln 2$

b) discrete data: The value $x_{0.5}$ is the median of a data set if at most half of the events have a value $< x_{0.5}$, and at most half of the events a value $> x_{0.5}$.

After sorting the data $\{x_1, x_2, ..., x_N\}$

$$x_{0.5} = \begin{cases} x_{\frac{N+1}{2}} & \text{if N odd} \\ \frac{1}{2}\left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}\right) & \text{if N even (sometimes also differently defined)} \end{cases}$$

Other parameters of a distribution (cont'd)

The median minimizes the mean absolute deviation (Chap. 1), i.e.,

$$\frac{1}{N}\sum_{i=1}^{N}\left|x_i - x_{0.5}\right| \quad \leq \quad \frac{1}{N}\sum_{i=1}^{N}\left|x_i - x\right|$$

Big advantage of the median:

The median preserves its meaning even for (strictly) monotonic non-linear transformations of the r.v. (later in this chapter). E.g., if $x_{0.5}$ is the median with respect to the distribution of x, then

$$y_{0.5} = y(\, x_{0.5})$$

is the median of the distribution of y(x).

This property is usually *not* valid for the mean [see problem set 3] or the mode!

Median from a histogram by linear interpolation:

$$x_{0.5} = x_{j,low} + \frac{N/2 - \sum_{k=1}^{j-1} n_k}{n_j} \cdot \Delta_j$$

where N is the total number of events in the histogram, j is the index of the bin which corresponds to $N/2$, $x_{j,low}$ is the x-value of the lower boundary of bin j (usually x_j-Δ_j/2 when the x-values of the bins are centered), n_j the number of events in bin j, and Δ_j its size.

Other parameters of a distribution (cont'd)

Example: Age distribution (in years) within a group of students in their 2nd student year, with bin sizes 0.25 yr and 0.5 yr (see histogram in Chap. 1)

Original, sorted, unbinned data: N=25, $x_{0.5}$=23.8
23.0, 23.1, 23.1, 23.2, 23.2, 23.3, 23.4, 23.4, 23.5, 23.5, 23.7, 23.7, **23.8**, 23.8, 23.8, 23.9, 23.9, 24.0, 24.1, 24.2, 24.4, 24.4, 24.5, 24.5, 24.7

bin	1	2	3	4	5	6	7
$x_{j,low}$	23.0	23.25	23.5	23.75	24.0	24.25	24.5
n_j	5	3	4	5	3	2	3

bin	1	2	3	4
$x_{j,low}$	23.0	23.5	24.0	24.5
n_j	8	9	5	3

$$x_{0.5} = x_{j,low} + \frac{N/2 - \sum_{k=1}^{j-1} n_k}{n_j} \cdot \Delta_j$$

$$\Delta = 0.25 : N/2 = \frac{5+3+4+5+3+2+3}{2} = 12.5$$

$$j = 4$$

$$x_{0.5} = 23.75 + \frac{12.5-12}{5} \cdot 0.25 = \mathbf{23.775}$$

$$\Delta = 0.50 : N/2 = \frac{8+9+5+3}{2} = 12.5$$

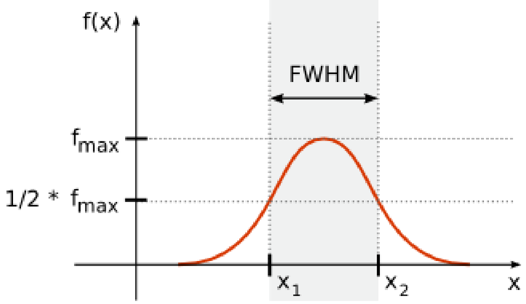
$$j = 2$$

$$x_{0.5} = 23.5 + \frac{12.5-8}{9} \cdot 0.5 = \mathbf{23.75}$$

lower and upper quartiles: $\mathrm{F}(x_{0.25}) = 0.25; \quad \mathrm{F}(x_{0.75}) = 0.75$

full width of half maximum (FWHM) is independent of the tails;

for a Gaussian distribution, $FWHM = 2.35\sigma$



Chebychev’s inequality

The values of a r.v. are somewhere in the neighbourhood of the mean μ. Deviations from the mean are less probable the larger they are compared with σ. This fact is expressed by Chebychev’s inequality (which is generally very weak):

$$\boxed{P(|\,\mathrm{x}-\mu\,|\geq k\sigma)\leq\frac{1}{k^2},\quad k\geq 1}$$

"The probability of x being k and more standard deviations away from the mean is lower/equal than k^{-2}"

Proof for a continuous r.v.

$$P = P(|\,\mathrm{x}-\mu\,|\geq k\sigma) = P((\mathrm{x}-\mu)^2 \geq k^2\sigma^2)$$

$P = \int\limits_{k^2\sigma^2}^{\infty} g(t)dt$ with g the p.d.f. of $t=(x-\mu)^2$

$$\sigma^2 = E\{(\mathrm{x}-\mu)^2\} = E(\mathrm{t}) = \int\limits_{-\infty}^{\infty} tg(t)dt = \int\limits_{0}^{k^2\sigma^2} tg(t)dt + \int\limits_{k^2\sigma^2}^{\infty} tg(t)dt$$

Since integration over positive values only and g(t) positiv definite (p.d.f),

the integral can be approximated $\left(\int\limits_a^b tg(t)dt \geq a\int\limits_a^b g(t)dt\right)$ as

$\sigma^2 \geq 0 + k^2\sigma^2 \int\limits_{k^2\sigma^2}^{\infty} g(t)dt = k^2\sigma^2 P$, i.e., $P\leq\dfrac{1}{k^2}$ q.e.d.

Alternative formulations and examples

$$P(|\,\mathrm{x}-\mu\,|\geq \lambda)\leq \frac{\sigma^2}{\lambda^2}, \quad \lambda>0$$

$$P(|\,\mathrm{x}-\mu\,|< \lambda)\geq 1-\frac{\sigma^2}{\lambda^2}, \quad \lambda>0$$

Example 1: Let‘s assume that an A&A (Astronomy & Astrophysics) article has an average length of 10 pages, with a standard deviation of 2 pages. By means of Chebychev’s inequality, we find a lower limit for the probability that an A&A article has a length between 6 and 14 pages, which is 75%:

$$P(|\,\mathrm{x}-10\,|< 4)\geq 1-\frac{2^2}{4^2}=0.75$$

Example 2: Another implication is that for any probability distribution with expectation value μ and finite standard deviation σ at least half of the values are located within the interval

$$\mu-\sqrt{2}\sigma < x < \mu+\sqrt{2}\sigma \quad \left(\text{since } 1-\frac{1}{k^2}=\frac{1}{2}\right)$$

Examples (continued)

Example 3
When applied to the arithmetic mean of N independent, identically distributed r.v.'s x_1 … x_N, we find one of the *weak* **laws of big numbers**.

Later, we will see that the expectation value of such an arithmetic mean is just the actual expectation value of the distribution, and that the variance of the arithmetic mean is the variance of the distribution, divided by N, i.e.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad E(\bar{x}) = E(x) = \mu, \quad Var(\bar{x}) = Var(x)/N. \text{ Thus, } \boxed{P(|\bar{x}-\mu| \geq \lambda) \leq \frac{Var(x)}{N\lambda^2}}$$

and the r.h.s. becomes arbitrarily small when N→∞. In other words, the arithmetic mean *converges stochastically* (i.e., w.r.t. probability) towards the expectation value of the distribution.

Moments of a distribution

- remember central moments (of r.v. or distribution)

$$\mu'_{\mathrm{n}} = E\{(\mathrm{x}-\mu)^n\} = \int_{-\infty}^{\infty} (x-\mu)^n f(x)dx$$

- analogue definition: moments of distribution (without prime)

$$\mu_{\mathrm{n}} = E(\mathrm{x}^n) = \int_{-\infty}^{\infty} x^n f(x)dx \quad \text{or} \quad \mu_{\mathrm{n}} = E(\mathrm{x}^n) = \sum_{k=1}^{\infty} \mathrm{x}_k^n P(\mathrm{x}=x_k)$$

$$\mu_1 := \mu =< \mathrm{x} >= E(\mathrm{x})$$

- remember as well

$$\mu'_1 = 0$$

$$\mu'_2 = \sigma^2 = Var(\mathrm{x})$$

$$\mu'_3 = \gamma_1\sigma^3$$

$$\mu'_4 = (\gamma_2 + 3)\sigma^4$$

- the probability density function is uniquely defined by its moments, as we will show now

Characteristic function

- *definition:* The **characteristic function** of a p.d.f. *f(x)* is

$$\phi(t) = E(e^{itx}) = \int_{-\infty}^{\infty} e^{itx} f(x)dx \text{ or } \sum_{\text{k=1}}^{\infty} e^{itx_k} P(\text{x} = x_k) \quad \text{Note: the lower summation}$$

index might be also 0

- for a continuous distribution, the characteristic function is the Fourier transform of *f(x)* (Note the (missing) normalization). Thus, the transform is invertible

$$\begin{aligned}
\phi(t) &= \int_{-\infty}^{\infty} e^{itx'} f(x')\, dx' \\
\int \phi(t) e^{-itx} dt &= \int e^{-itx} dt \int_{-\infty}^{\infty} e^{itx'} f(x')\, dx' \\
&= \int_{-\infty}^{\infty} f(x') \left(\int e^{it(x'-x')} dt \right) dx' \\
&= 2\pi \int_{-\infty}^{\infty} f(x')\, \delta(x'-x) dx' \\
&= 2\pi f(x)
\end{aligned}$$

$$\boxed{f(x) = \frac{1}{2\pi} \int \phi(t) e^{-itx} dt}$$

- … and the characteristic function defines the p.d.f.

Characteristic function *and* moments

The n-th derivative of the characteristic function is

$$\frac{d^n\phi(t)}{dt^n} = \int_{-\infty}^{\infty} (ix)^n e^{itx} f(x)dx.$$

At $t = 0$ one obtains

$$\left.\frac{d^n\phi(t)}{dt^n}\right|_{t=0} = \int_{-\infty}^{\infty} (ix)^n f(x)dx, \quad \text{i.e.,} \qquad \left.\frac{d^n\phi(t)}{dt^n}\right|_{t=0} = i^n \mu_n$$

Thus, the Taylor expansion of $\phi(t)$ around $t = 0$,

$$\phi(t) = \sum_{n=0}^{\infty} \frac{1}{n!} t^n \left.\frac{d^n\phi(t)}{dt^n}\right|_{t=0} = \sum_{n=0}^{\infty} \frac{1}{n!}(it)^n \mu_n = 1 + \sum_{n=1}^{\infty} \frac{1}{n!}(it)^n \mu_n$$

depends on the moments alone. Since the Fourier transform can be uniquely inverted and the Taylor expansion of the characteristic function consists of the moments, we conclude that indeed the moments define the p.d.f., as stated above.

For the central moments, we find in analogy

$$\phi'(t) = E(e^{it(\mathrm{x}-\mu)}) = \int_{-\infty}^{\infty} e^{it(x-\mu)} f(x)dx \quad \rightarrow \sum_{n=0}^{\infty} \frac{1}{n!}(it)^n \mu'_n$$

Note in particular that $\mu'_2 = \boxed{\sigma^2(\mathrm{x}) = -\left.\frac{d^2\phi'(t)}{dt^2}\right|_{t=0}}$

Example

- characteristic function and moments of the exponential distribution.

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0 \quad (\text{e.g., } \lambda = \frac{1}{\tau} \text{ for the life-time distribution})$$

$$\phi(t) = \int_0^\infty e^{itx} \lambda e^{-\lambda x} dx = \frac{\lambda}{-\lambda + it} e^{(-\lambda + it)x} \Big|_0^\infty = \frac{\lambda}{\lambda - it}$$

By differentiation,

$$\frac{d\phi(t)}{dt} = \frac{i\lambda}{(\lambda - it)^2}$$

$$\frac{d^n\phi(t)}{dt^n} = \frac{n!i^n\lambda}{(\lambda - it)^{n+1}}$$

$$\frac{d^n\phi(t)}{dt^n}\Big|_{t=0} = \frac{n!i^n}{\lambda^n}, \text{ we obtain the moments}$$

$\mu_{\text{n}} = n!\lambda^{-n} = n!\tau^n$ (compare Fig. page 45)

without explicitly calculating the integrals defining the expectation values!

Sum of two independent r.v.

Let z = x + y with independent r.v. x,y and corresponding p.d.f.s $f(x),\ g(y)$.
Calculate the distribution $h(z)$.

$$\phi_h(t) = E(e^{it(\mathrm{x+y})}) = E(e^{it\mathrm{x}}e^{it\mathrm{y}}) \overset{\text{x,y independent}}{=} E(e^{it\mathrm{x}})E(e^{it\mathrm{y}}),\ \text{i.e.,}$$

$$\boxed{\phi_h(t) = \phi_f(t)\phi_g(t)}$$

and thus

$$h(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itz}\phi_h(t)dt$$

The pdf of the sum of two independent r.v. is the (inverse) Fourier-transform of the product of the two corresponding characteristic functions!

Example

Calculate the distribution $h(z)$ of the sum $\mathrm{z} = \sum_{i=1}^{n} \mathrm{x}_i$,

with x_i uniformly distributed (see Chapt. 5) in the interval $\left[0,1\right]$:

$$f(x_i) = \begin{cases} 1 \ \text{ for } 0 \le x_i \le 1 \\ 0 \ \text{ else} \end{cases}$$

(for other solution methods, see Chap. 4)

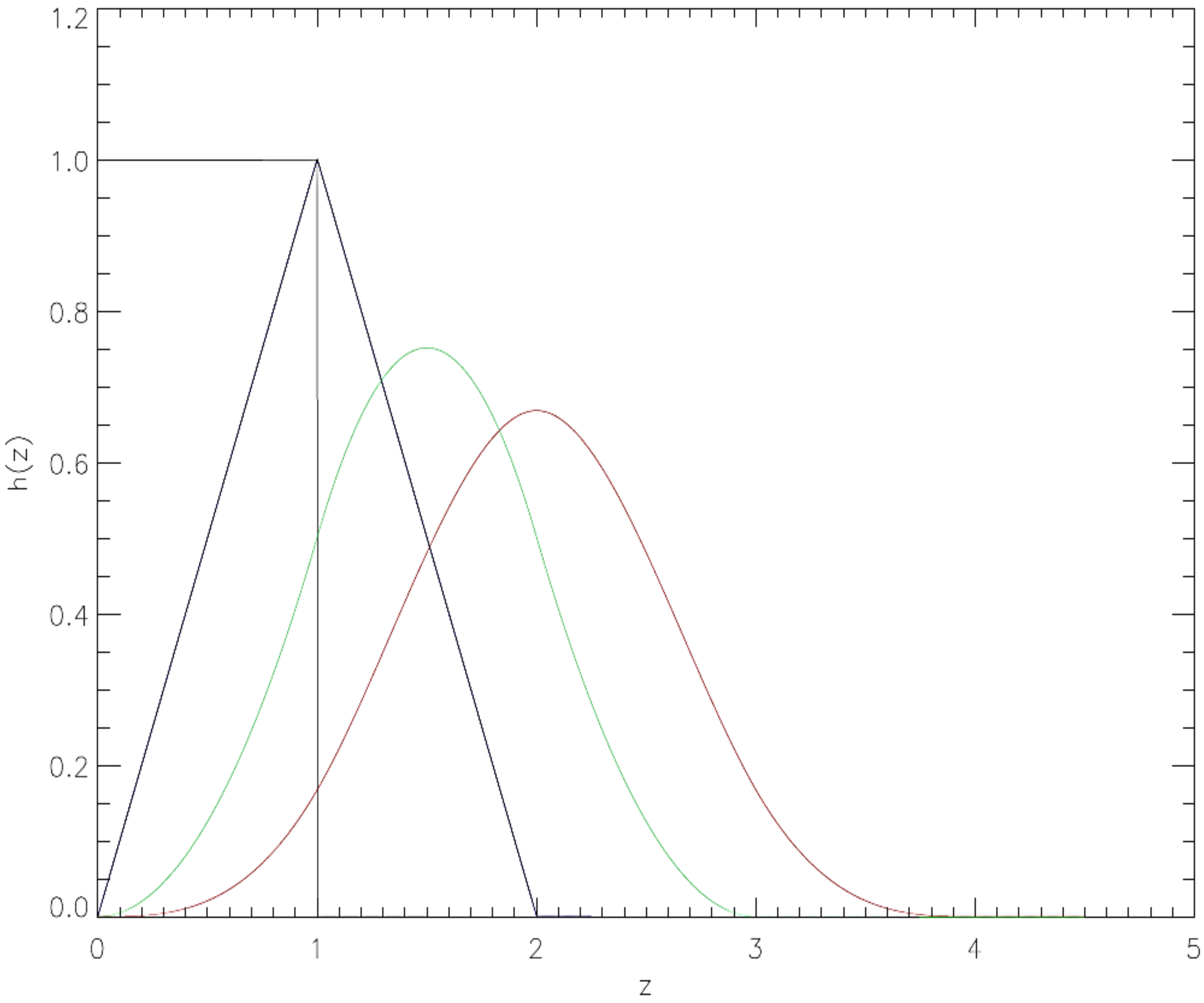
$$\phi_f(t) = E\left(\exp(it\mathrm{x})\right) = \int_{-\infty}^{\infty} f(x)\exp(itx)dx = \int_0^1 1\cdot\exp(itx)dx = \frac{i}{t}\left(1-\exp(it)\right)$$

$$\phi_h(t) = \left(\phi_f(t)\right)^n$$

$$h(z) = \frac{i^n}{2\pi}\int_{-\infty}^{\infty}\left(\frac{1-\exp(it)}{t}\right)^n \exp(-itz)dt = \text{ (mathematica, maple...)}$$

$$= \begin{cases} n = 2: \ \frac{1}{2}\left(|z-2| - 2|z-1| + |z|\right) \ \triangleq \ z \text{ for } 0 \le z < 1 \text{ and } 2 - z \text{ for } 1 \le z < 2 \\ n = 3: \ \frac{1}{2}z^2 \text{ for } 0 \le z < 1, \ \frac{1}{2}(-2z^2 + 6z - 3) \text{ for } 1 \le z < 2 \text{ and } \frac{1}{2}(z-3)^2 \text{ for } 2 \le z < 3 \\ n = 4: \ \frac{1}{12}\left(|z-4|^3 - 4|z-3|^3 + 6|z-2|^3 - 4|z-1|^3 + |z^3|\right) \end{cases}$$

Example (cont'd)



Distribution *h(z)* (with z the sum of n uniformly distributed numbers within [0,1]), for different n:

black: n=1 (=$f(x_i)$)
blue: n=2
green: n=3
red: n=4

The distributions have been calculated via Fast Fourier transform

Cumulants

Characteristic function defined by moments, $\phi(t) = 1 + \sum_{n=1}^{\infty} \frac{(it)^n}{n!}\mu_n$, with $\phi(0) = 1$

Often, $\phi(t)$ has exponential character, e.g., for Poisson (p. 101) and *normal* distribution (p. 108).

Thus, expand $\ln\phi(t)$:

$$\ln\phi(t) = \sum_{n=1}^{\infty} \frac{(it)^n}{n!}\kappa_n \quad \text{with } \kappa_n = \frac{1}{i^n} \frac{d^n \ln\phi(t)}{dt^n}\bigg|_{t=0} \quad (n = 0 \text{ term vanishes since } \ln\phi(0) = \ln(1) = 0)$$

κ_n 'cumulants' of the distribution, can be expressed in terms of (central) moments.

First four cumulants:

$$\kappa_1 = \mu_1; \quad \kappa_2 = \mu_2 - \mu_1^2 = \mu'_2; \quad \kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 = \mu'_3; \quad \kappa_4 = \mu'_4 - 3\left(\mu'_2\right)^2$$

Can be shown:

If x is a random variable, $\kappa_n(\mathrm{x})$ the corresponding cumulant (w.r.t. to $\ln\phi_x(t)$), and $c \in \mathbb{R}$, then

$\kappa_1(\mathrm{x}+c) = \kappa_1(\mathrm{x}) + c$; $\quad \kappa_n(\mathrm{x}+c) = \kappa_n(\mathrm{x})$ for $n \geq 2$ ('almost' translation-invariant)

$\kappa_n(c\mathrm{x}) = c^n\kappa_n(\mathrm{x})$ $\left(\text{homogeneous of degree } n\right)$ $[\kappa_n(c\mathrm{x})$ is the cumulant for y=cx w.r.t. to $\ln\phi_y(t)]$

If x and y independent random variables and z = x + y, then

$$\boxed{\kappa_n(\mathrm{z}) = \kappa_n(\mathrm{x}) + \kappa_n(\mathrm{y})}$$

Note: For *normal* distributions, only the first two cumulants are different from zero!!!

Transformation of variables

- given a p.d.f. *f(x),* we'd like to know the p.d.f. *g(u),* when *u* is a (invertible) function of *x, u(x)*
- example: given a distribution of velocities *f(v),* we want to calculate the distribution of energies, $\frac{1}{2}mv^2$
- for discrete distributions, this is trivial. The probability for the event $u(x_k)$ (where *u* is a function of *x*) is the same as for the event x_k itself,

$$P(u(x_k))=P(x_k)$$

- for continuous distributions, we have to invoke calculus

Calculation of the transformed p.d.f.

Given: pdf $f(x)$ and uniquely invertible + differentiable (implies strictly monotonic) function $u(x)$.

Calculate g(u).

$P(x_1 < \mathrm{x} < x_2) = P(u_1 < \mathrm{u} < u_2)$ or $P(u_2 < \mathrm{u} < u_1)$ with $u_1 = u(x_1)$ and $u_2 = u(x_2)$.

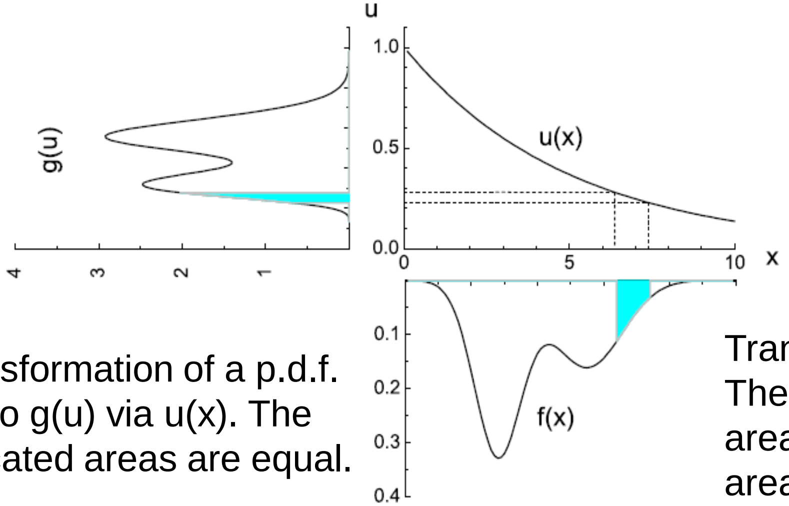
$$P = \int_{x_1}^{x_2} f(x)dx = \left| \int_{u_1}^{u_2} g(u)du \right| \Rightarrow |g(u)du| = f(x)dx \text{ and thus } \qquad \boxed{g(u) = f(x)\left|\frac{dx}{du}\right|}$$

The absolute sign garantuees that the pdf is positive. Integrating this equation yields

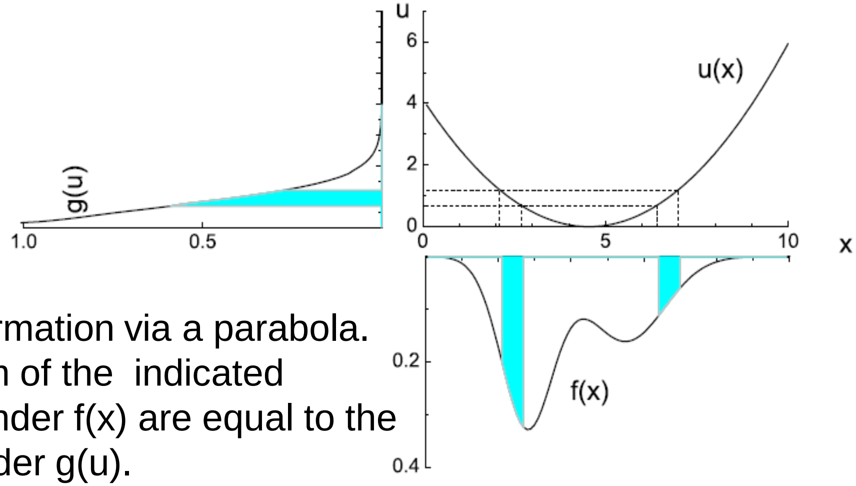
$$F(x) = \begin{cases} G(u) & \text{for } u(x_{\min}) < u(x_{\max}), \text{ i.e., } u(x) \text{ monotonically increasing} \\ 1 - G(u) & \text{for } u(x_{\min}) > u(x_{\max}), \text{ i.e., } u(x) \text{ monotonically decreasing} \end{cases}$$

If $u(x)$ is invertible, but no longer uniquely, and thus $x(u)$ is multi-valued, one has to sum over all contributing branches (within the branches, the derivative must not change its sign).

$$g(u) = \left\{ f(x)\left|\frac{dx}{du}\right| \right\}_{\text{branch 1}} + \left\{ f(x)\left|\frac{dx}{du}\right| \right\}_{\text{branch 2}} + \ldots$$



Transformation of a p.d.f. f(x) to g(u) via u(x). The indicated areas are equal.



Transformation via a parabola. The sum of the indicated areas under f(x) are equal to the area under g(u).

Examples

example 1: calculate the pdf for the area of a circle from a uniform (see Chap. 5) distribution of radii between 0 and r$_m$.

$$\text{p.d.f. for } r: f(r) = \frac{1}{r_m - 0} \text{ for } 0 < r < r_m; \quad f(r) = 0 \text{ else.}$$

$$g(A) = f(r)\left|\frac{dr}{dA}\right| \text{ with } A = r^2\pi \quad \left[r(A) \text{ single-valued, since } r > 0\right]$$

$$\frac{dA}{dr} = 2\pi r; \quad g(A) = \frac{1}{r_m 2\pi r}$$

$$g(A) = \frac{1}{2}\sqrt{\frac{1}{A_m A}}; \text{ Test:} \int_0^{A_m} g(A)dA = \frac{1}{2\sqrt{A_m}} \int_0^{A_m} A^{-\frac{1}{2}} dA = 1!$$

example 2: Calculate the distribution for the square of a reduced r.v. where the original r.v. should be normally distributed.

$$u = \left[\frac{(\mathrm{x} - \mu)}{\sigma}\right]^2 \text{ and } f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{(see Chap. 5)}$$

The function $x(u)$ has two branches [since $(x - \mu)$ positive or negative]!

Examples (cont'd)

$$\frac{dx}{du} = \pm\frac{\sigma}{2\sqrt{u}}; \quad g(u) = \left\{\frac{\sigma}{2\sqrt{u}}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{u}{2}}\right\}_{\text{branch 1}} + \left\{\left|-\frac{\sigma}{2\sqrt{u}}\right|\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{u}{2}}\right\}_{\text{branch2}}$$

Since the contributions from branch 1 and 2 are identical, we obtain

$$g(u) = \frac{1}{\sqrt{2\pi u}}e^{-u/2},$$

which is the so-called χ^2-distribution for one degree of freedom (see Chap. 5)

example 3: kinetic energy for a 1-D ideal gas. The pdf of the velocity of a particle into direction *x* is

$f(v) = \sqrt{\frac{m}{2\pi kT}}e^{-\frac{mv^2}{2kT}}$. Calculate the corresponding energy distribution.

As above [two branches, since *v* positive or negative]

$\frac{dv}{dE} = \pm\frac{1}{\sqrt{2mE}}$; both branches have similar contributions, thus

$$g(E) = \frac{2}{\sqrt{2mE}}\sqrt{\frac{m}{2\pi kT}}e^{-E/kT} = \frac{1}{\sqrt{\pi kTE}}e^{-E/kT}$$

Calculation of the transformation

- Now, the original and the transformed p.d.f., *f(x)* and *g(u),* are given, and the transformation *u(x)* needs to be calculated.
- This situation is frequently met in Monte-Carlo simulations. Random number generators usually create uniformly distributed r.v., and we look for the transformation law which transforms these uniformly distributed r.v. into others which are distributed following a given p.d.f. (defined by the process to be investigated).

$$\int_{-\infty}^{x} f(x')dx' = \int_{-\infty}^{u} g(u')du'; \text{ integration yields the c.d.f.s}$$

Note 1: no problem with abs-value here, since f and g positive by definition (pdf)

$$F(x) = G(u) \text{ and thus } u(x) = G^{-1}(F(x))$$

Note 2: If $g(u)$ has been derived from a transformation as described before, via $u'(x)$, and $x(u')$ was multi-valued, the transformation law $u(x)$ can be different from $u'(x)$, but delivers the same transformed distribution $[\rightarrow$ problem sheet 4$]$

- The problem can be solved analytically only if both p.d.f.s f and g can be integrated analytically, and if the inverse of G can be calculated.
- in other cases (which are the majority), numerical methods have to be applied. Most powerful is the rejection method by von Neumann (see, e.g., “Numerical Recipes” and http://www.usm.uni-muenchen.de/people/puls/lessons/numpraktnew/montecarlo/mc_manual.pdf

In the former case of f being a uniform distribution over the unit interval, i.e.,

$f(x) = 1$ for $0 \le x \le 1$ and $f(x) = 0$ else [random number generator],

we obtain $F(x) = x$ and thus $G(u) = x$; $\boxed{u = G^{-1}(x).}$

In this case, only G needs to be calculable and invertible.

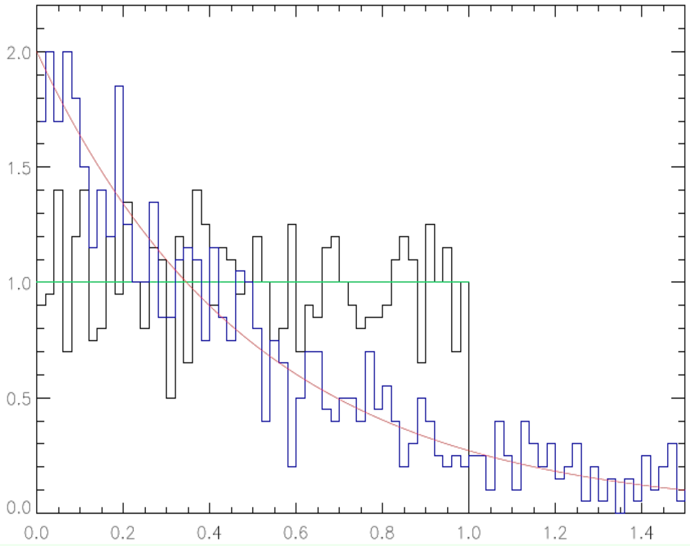
example: Create exponentially distributed r.v. from a uniform distribution.

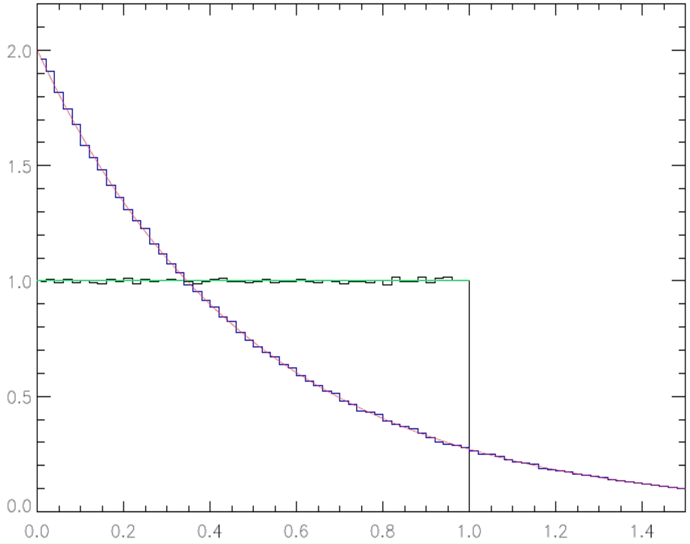
$$g(u) = \lambda e^{-\lambda u};$$

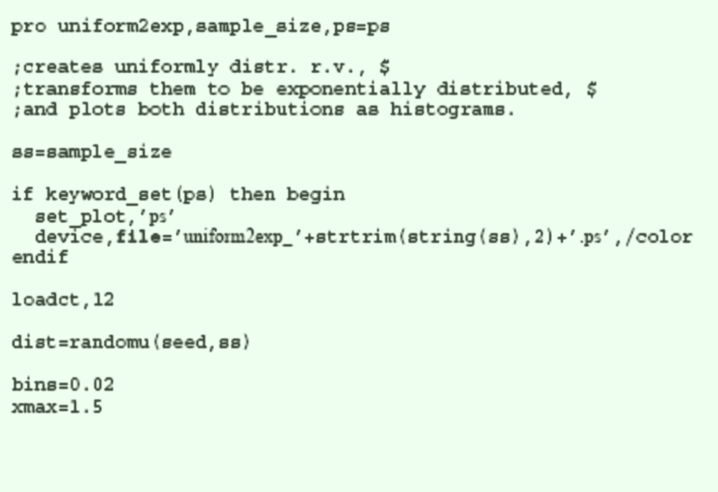
$$G(u) = \int_0^u \lambda e^{-\lambda u'} du' =: F(x) = x;$$

$$1 - e^{-\lambda u} = x; \quad u(x) = -\ln(1-x)/\lambda \quad \overset{\text{uniformly distr. x in unit interval}}{\triangleq} \quad -\ln(x)/\lambda$$

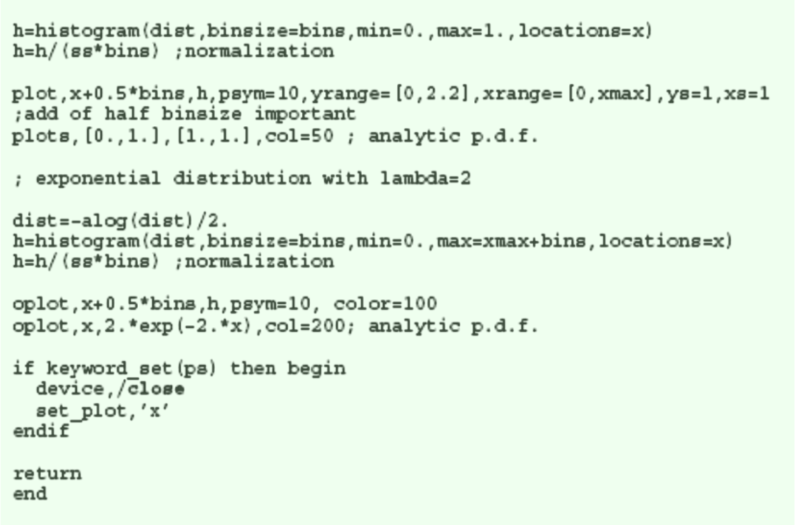
P.d.f.s for a uniform distribution (black), generated by a random number generator from N=10^3 (left) and N=10^6 subsequent numbers. The corresponding exponential distribution (λ=2, blue) has been created from these numbers using the transformation method as described above. Displayed are (normalized) histograms with bin size 0.02. Analytical p.d.f.s in green and red. IDL (interactive data language) code below.







```
pro uniform2exp,sample_size,ps=ps

;creates uniformly distr. r.v., $
;transforms them to be exponentially distributed, $
;and plots both distributions as histograms.

ss=sample_size

if keyword_set(ps) then begin
  set_plot,'ps'
  device,file='uniform2exp_'+strtrim(string(ss),2)+'.ps',/color
endif

loadct,12

dist=randomu(seed,ss)

bins=0.02
xmax=1.5
```



```
h=histogram(dist,binsize=bins,min=0.,max=1.,locations=x)
h=h/(ss*bins) ;normalization

plot,x+0.5*bins,h,psym=10,yrange=[0,2.2],xrange=[0,xmax],ys=1,xs=1
;add of half binsize important
plots,[0.,1.],[1.,1.],col=50 ; analytic p.d.f.

; exponential distribution with lambda=2

dist=-alog(dist)/2.
h=histogram(dist,binsize=bins,min=0.,max=xmax+bins,locations=x)
h=h/(ss*bins) ;normalization

oplot,x+0.5*bins,h,psym=10, color=100
oplot,x,2.*exp(-2.*x),col=200; analytic p.d.f.

if keyword_set(ps) then begin
  device,/close
  set_plot,'x'
endif

return
end
```

4. Distributions of several random variables – multivariate p.d.f.s

- until now, *univariate* distributions: one r.v.
- generalization to several r.v. “easy”: multivariate (also: more-dimensional) distributions
- in the following, only continuous distributions

- *definition* of prob. distribution for two r.v., x,y:

$$F(x,y) = P(\mathrm{x} < x, \mathrm{y} < y) \text{ with}$$

$$F(\text{-}\infty,\text{-}\infty) = 0, \quad F(\infty,\infty) = 1$$

- corresponding *joint* p.d.f.

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \quad \Rightarrow \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1 \text{ and}$$

$$P(a \le \mathrm{x} < b, c \le \mathrm{y} < d) = \int_a^b \int_c^d f(x,y)dxdy$$

Marginal distributions

- following problem: sometimes the c.d.f. *F(x,y)* is approximately determined (by many measurements), but only the probability distribution of *x* (irrespective of *y*) is of interest.
- **example:** the appearance of a certain disease is known as a function of location and date. For a certain investigation, the dependence on date is without interest.
- In this case, we *marginalize* the distribution, i.e., we integrate over the whole range in *y*

$$P(a \leq \mathrm{x} < b, -\infty < \mathrm{y} < \infty) = \int_a^b \left[\int_{-\infty}^{\infty} f(x,y)dy \right] dx = \int_a^b g(x)dx$$

$g(x) = \int\limits_{-\infty}^{\infty} f(x, y)dy$ is a p.d.f. of x, called the marginal distribution of x.

The corresponding distribution of y is

$$h(y) = \int\limits_{-\infty}^{\infty} f(x, y)dx$$

Marginal distributions are "projections" of the joint p.d.f. onto the axes.

Two r.v. x,y are independent if

$$f(x, y) = g(x)h(y)$$

Now, we can define the conditional probability for y given that x is known:

$P(y \le \mathrm{y} < y + dy \mid x \le \mathrm{x} < x + dx).$

The corresponding p.d.f. is given by

$$f(y \mid x) = \frac{f(x, y)}{g(x)},$$

and the above probability results as $f(y \mid x)dy$.

Note: conditional probabilities as defined above are normalized!

The rule of total probability

$$\mathrm{P(B)} = \sum_{\mathrm{k}} \mathrm{P(B \cdot E_k)} = \sum_{\mathrm{k}} \mathrm{P(B|E_k)P(E_k)}$$

(see Chap. 2) is then expressed by

$$h(y) = \int_{-\infty}^{\infty} f(x,y)dx \;= \int_{-\infty}^{\infty} f(y\,|\,x)g(x)dx\ .$$

If the variables are independent, then

$$f(y\,|\,x) = \frac{f(x,y)}{g(x)} = \frac{g(x)h(y)}{g(x)} = h(y)$$

Any constraint on one variable cannot contribute information about the other, if the variables are independent!

Bayes theorem for two-dimensional distributions:

$$f(x\,|\,y)h(y) = f(y\,|\,x)g(x) = f(x,y)$$

NOTE: often, this theorem is used to construct $f(x,y)$ from the conditional and the marginal probabilities.

Example

- superposition of two normal-distributions, with corresponding marginal and conditional p.d.f.s

$$f(x,y) = \frac{1}{2\pi}\left[0.6\exp\left(-\frac{x^2}{2}-\frac{y^2}{2}\right)+\frac{0.4}{\sqrt{3}}\exp\left(-\frac{(x-2)^2}{3}-\frac{(y-2.5)^2}{4}\right)\right]$$

$$g(x) = \frac{1}{\sqrt{2\pi}}\left[0.6\exp\left(-\frac{x^2}{2}\right)+\frac{0.4}{\sqrt{1.5}}\exp\left(-\frac{(x-2)^2}{3}\right)\right] = \int f(x,y)dy$$

$$h(y) = \frac{1}{\sqrt{2\pi}}\left[0.6\exp\left(-\frac{y^2}{2}\right)+\frac{0.4}{\sqrt{2}}\exp\left(-\frac{(y-2.5)^2}{4}\right)\right] = \int f(x,y)dx$$

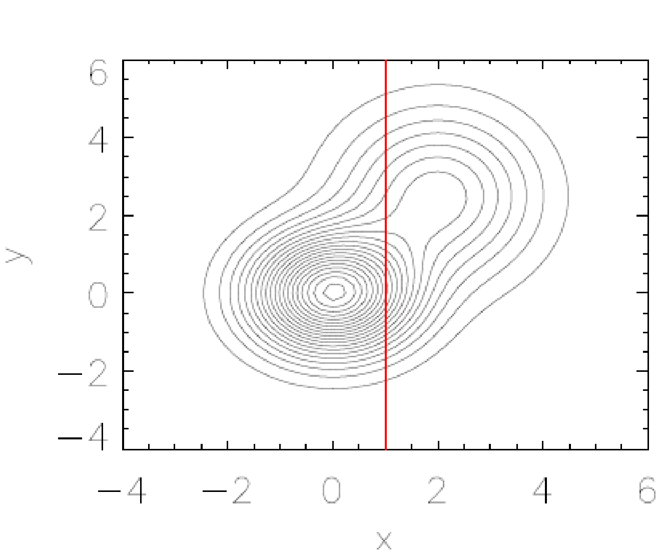
$$f(y,x = 1) = \frac{1}{2\pi}\left[0.6\exp\left(-\frac{1}{2}-\frac{y^2}{2}\right)+\frac{0.4}{\sqrt{3}}\exp\left(-\frac{1}{3}-\frac{(y-2.5)^2}{4}\right)\right]$$

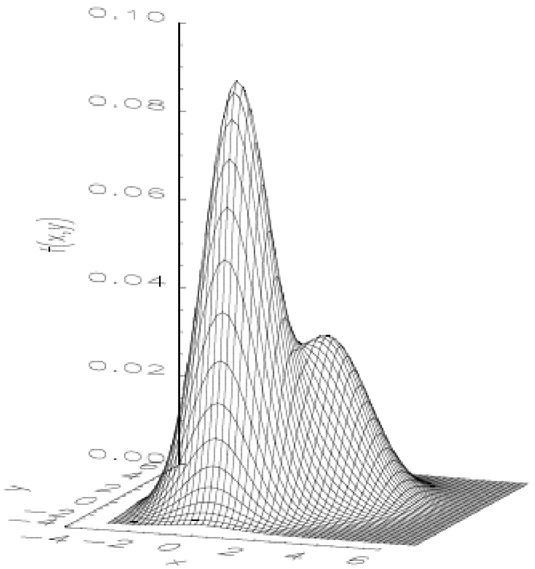
$$f(y|x = 1) = 0.6672\left[0.6\exp\left(-\frac{1}{2}-\frac{y^2}{2}\right)+\frac{0.4}{\sqrt{3}}\exp\left(-\frac{1}{3}-\frac{(y-2.5)^2}{4}\right)\right]$$

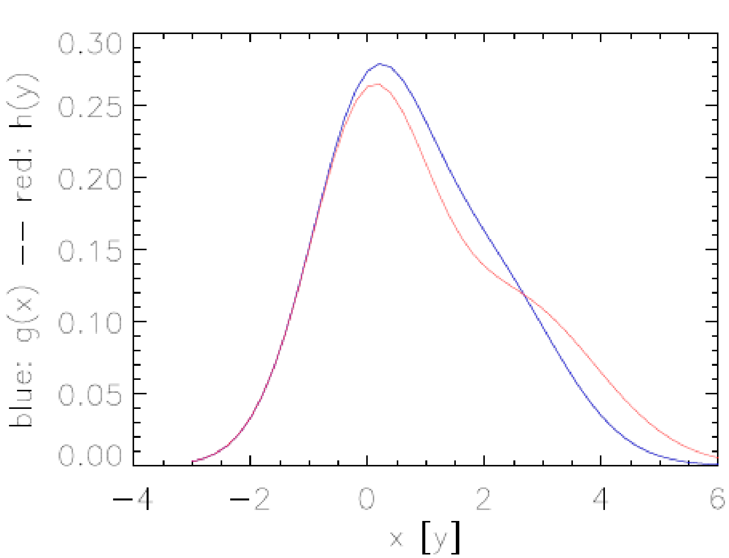
$= f(y,x=1)/g(1);$ Remember that this conditional pdf is normalized, i.e., $\int f(y\,|\,x=1)dy = 1$

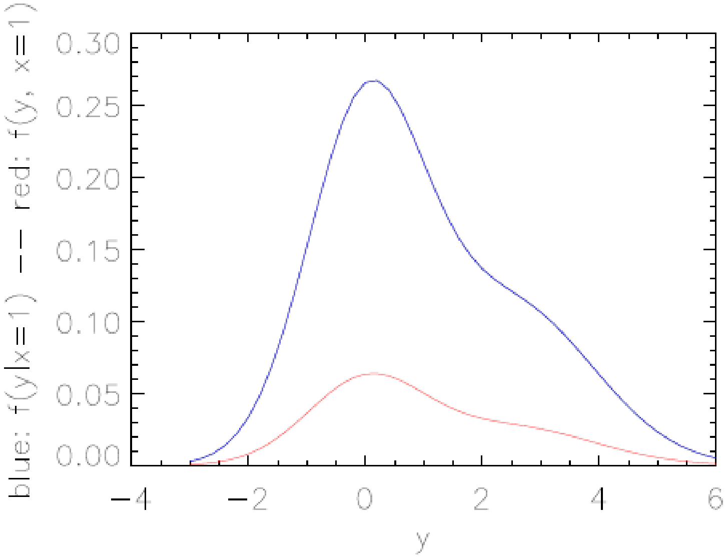
Note: x and y not independent, since $f(y\,|\,x)$ depends on x!

Example (cont'd)









Convolution

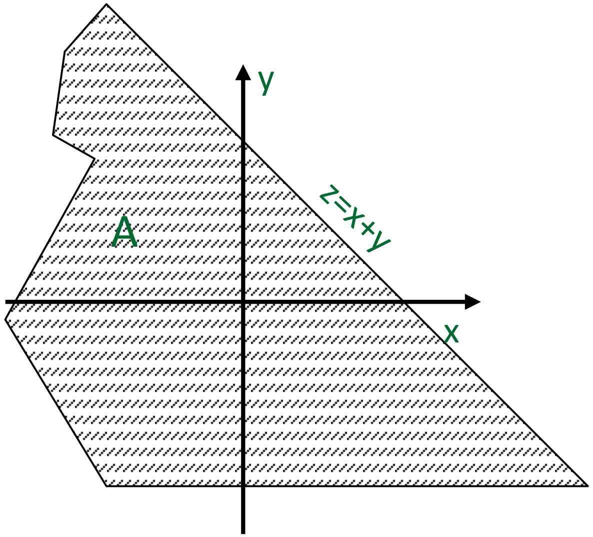
Let z=x+y the sum of two independent r.v. x,y, with a joint pdf

$$f(x,y) = f_x(x)\,f_y(y)$$

The (cumulative) distribution function of z is

$$F(z) = P(z < z) = P(x+y < z)$$

and given by the integration of the joint pdf over the shaded area in the right hand figure.



$$F(z) = \iint_A f_x(x)\,f_y(y)dxdy = \int_{-\infty}^{\infty} f_x(x)dx \int_{-\infty}^{z-x} f_y(y)dy = \int_{-\infty}^{\infty} f_y(y)dy \int_{-\infty}^{z-y} f_x(x)dx$$

The probability density is found by differentiating $F(z)$,

$$f(z) = \frac{dF(z)}{dz} = \int_{-\infty}^{\infty} f_x(x)\,f_y(z-x)dx = \int_{-\infty}^{\infty} f_y(y)\,f_x(z-y)dy$$

This is a convolution of the two pdf's [see also Chap. 3, p. 41].

- In Chap. 3, we *alternatively* proved that the p.d.f. of the sum of two independent r.v. is the (inverse) Fourier-transform of the product of the two corresponding characteristic functions, where the characteristic functions themselves are the Fourier transforms of the individual p.d.f.'s (except for a different normalization)
- As we will show now, this is nothing else than a **convolution,** in agreement with the above result

Convolution

There are many situations where we have a signal $S(x')$ and a so-called "response function" $R(x, x')$. R decribes the probability that a certain fraction of the signal within "channel" x' emerges within channel x. Thus, the signal will be washed out by the response function.

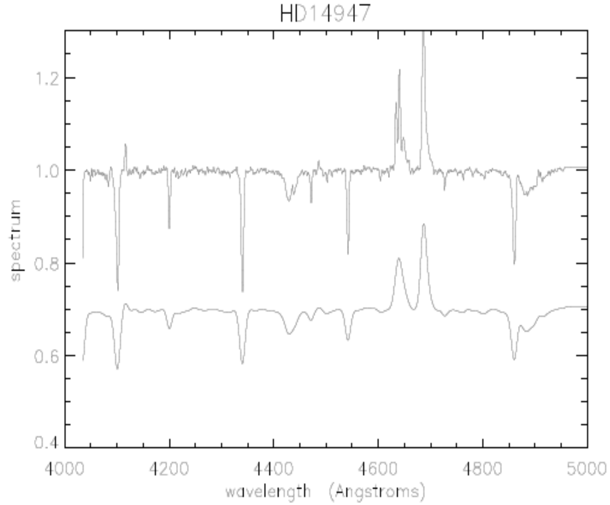


Figure 1.1: Top: optical spectrum of the hot supergiant HD 14947, observed with high resolution. Bottom: same spectrum, observed with 8-times lower resolution (shifted by −0.3): the signal has been washed out.

1.1 Example (Resolution of a detector). A stellar spectrum $S(\nu')$ (with frequency ν') should be collected by a detector (spectrograph) with finite resolution $\Delta\nu$.

In this case, the response function can be approximated by

$$R(\nu, \nu') = R(\nu - \nu') = \frac{1}{\sqrt{\pi}\Delta\nu} e^{-\left(\frac{\nu-\nu'}{\Delta\nu}\right)^2},$$

with $\Delta\nu$ the spectrograph's resolution. (see Fig. 1.2). If $\Delta\nu$ is larger than the width of the individual spectral lines, one observes a spectrum which has been washed out,

$$S'(\nu) = \int\limits_{-\infty}^{\infty} S(\nu')R(\nu - \nu')\,\mathrm{d}\nu' = \int\limits_{-\infty}^{\infty} S(\nu - \nu')R(\nu')\,\mathrm{d}\nu' \qquad \text{“convolution”}.$$

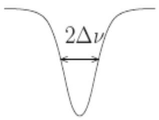


Figure 1.2: Resolution of a detector: the "response function"

The signal has been *convolved* with the response function. Due to the finite resolution, contributions from different signals $S(\nu')$ are summed up at frequency ν, primarily from the range $\nu' = \nu \pm \Delta\nu$ (Fig. 1.1, bottom).

If the detector has a very high resolution ($\lambda/\Delta\lambda \gtrsim 50,000$), the response function might be approximated by a δ-function,

$$R(\nu - \nu') \approx \delta(\nu - \nu') \qquad \Rightarrow \qquad S'(\nu) \simeq S(\nu),$$

and the original signal will be hardly distorted (Fig. 1.1, top).

If one likes to compare the observation $S'(\nu)$ with theoretical simulations of the spectrum, this problem needs to be considered. Usually, the theoretical signal is convolved with the response function of the detector. If the noise of the observed signal is (very) low, one might also "deconvolve" the observed spectrum ($S' \to S$).

If $S(\nu')$ comprises N frequency points and $R(\nu-\nu')$ M frequency points in the range $R(\nu-\nu') \neq 0$, at first glance this convolution would require $N \cdot M$ essential operations, accounting for the formulation as given above,

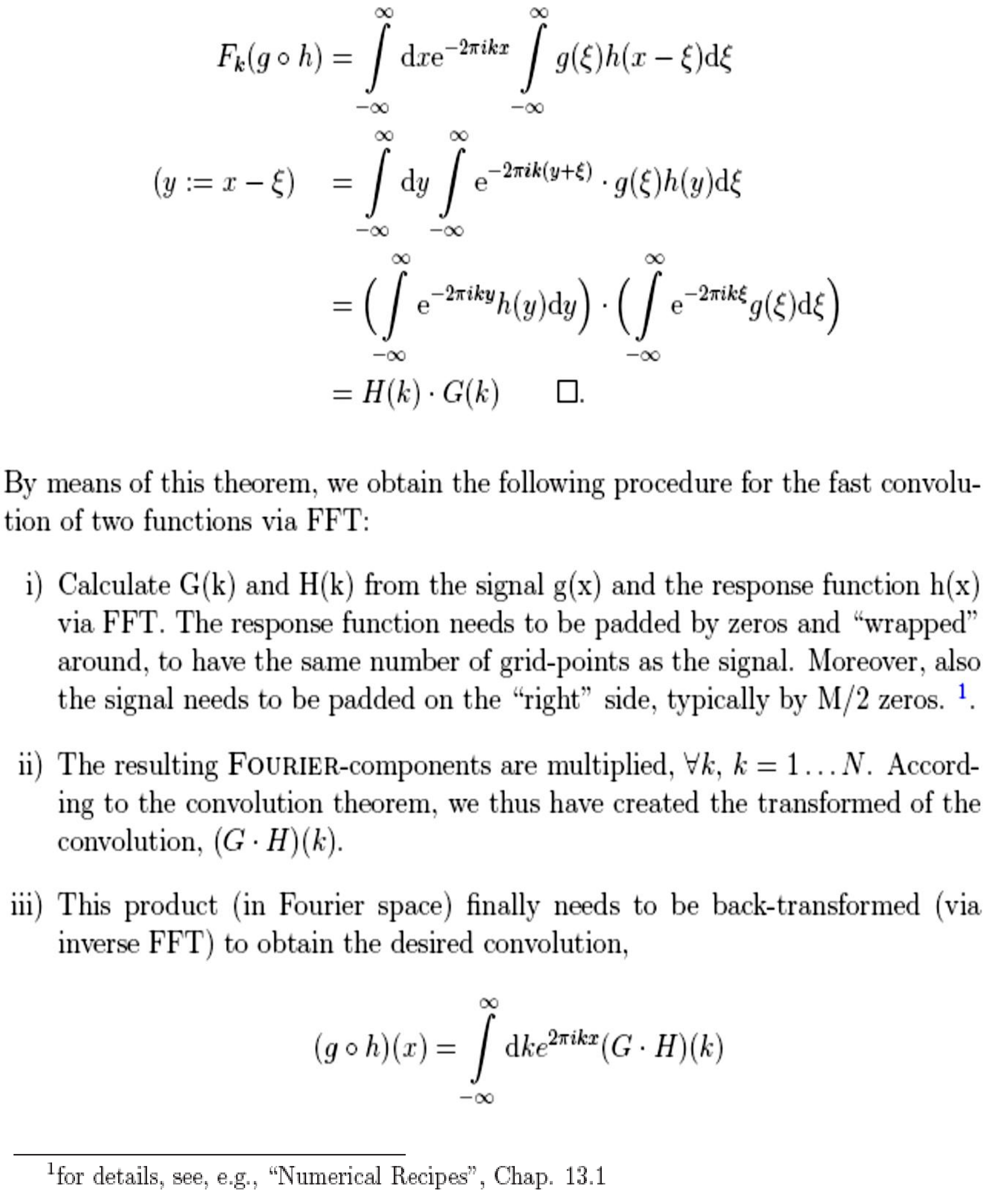
By using the "Fast Fourier Transformation (FFT)", the computational effort can be significantly reduced. To understand this, we need the following

1.2 Theorem (convolution theorem). *The* FOURIER*-transform of a convolution of two functions $(g \circ h)$ is the product of the* FOURIER*-transforms of the individual functions, $G \cdot H$.*

1.3 Proof. *(using a somewhat different normalization than for the characteristic function and $k = 2\pi t$)*

$$(g \circ h)\,(x) = \int\limits_{-\infty}^{\infty} g(\xi)h(x - \xi)\mathrm{d}\xi$$

Convolution



$$
\begin{aligned}
F_k(g \circ h) &= \int_{-\infty}^{\infty} \mathrm{d}x e^{-2\pi ikx} \int_{-\infty}^{\infty} g(\xi)h(x-\xi)\mathrm{d}\xi \\
(y := x-\xi) \quad &= \int_{-\infty}^{\infty} \mathrm{d}y \int_{-\infty}^{\infty} \mathrm{e}^{-2\pi ik(y+\xi)} \cdot g(\xi)h(y)\mathrm{d}\xi \\
&= \Big(\int_{-\infty}^{\infty} \mathrm{e}^{-2\pi iky} h(y)\mathrm{d}y\Big) \cdot \Big(\int_{-\infty}^{\infty} \mathrm{e}^{-2\pi ik\xi} g(\xi)\mathrm{d}\xi\Big) \\
&= H(k) \cdot G(k) \qquad \square.
\end{aligned}
$$

By means of this theorem, we obtain the following procedure for the fast convolution of two functions via FFT:

i) Calculate G(k) and H(k) from the signal g(x) and the response function h(x) via FFT. The response function needs to be padded by zeros and "wrapped" around, to have the same number of grid-points as the signal. Moreover, also the signal needs to be padded on the "right" side, typically by M/2 zeros. [1].

ii) The resulting FOURIER-components are multiplied, $\forall k,\ k = 1 \ldots N$. According to the convolution theorem, we thus have created the transformed of the convolution, $(G \cdot H)(k)$.

iii) This product (in Fourier space) finally needs to be back-transformed (via inverse FFT) to obtain the desired convolution,

$$(g \circ h)(x) = \int_{-\infty}^{\infty} \mathrm{d}k e^{2\pi ikx} (G \cdot H)(k)$$

[1] for details, see, e.g., "Numerical Recipes", Chap. 13.1

Since the Fourier-transformation via FFT costs roughly N $\log_2$ N operations, the total number of essential operations is of the order of

$N \cdot (1 + 3 \log_2 N)$.

The first part results from the multiplications $G_k \cdot H_k$, and the second part from two forward and one backward transformations.

Thus, if $(1 + 3 \log_2 N) < M$ (with M the original number of grid points for the response function), the convolution by FFT is faster than the "simple" method by conventional integration.

If, e.g., N = 1000 frequency points, for all $M > (1 + 3 \log_2 N) \approx 30$ there would be a gain in computation time.

Moments

- in analogy to univariate distributions, we define

$$
\begin{aligned}
\mu_x &= E(x) \\
\mu_y &= E(y) \\
\sigma_x^2 &= E\left[(x-\mu_x)^2\right] && = \mu'_{20} \\
\sigma_y^2 &= E\left[(y-\mu_y)^2\right] && = \mu'_{02} \\
\sigma_{xy} &= E\left[(x-\mu_x)(y-\mu_y)\right] = \mu'_{11} = \\
&= E(\mathrm{xy}) - \mu_x\mu_y = \\
&= \mathrm{cov(x,y)} \quad \text{"covariance"}
\end{aligned}
$$

$$
\begin{aligned}
\mu_x &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\,dx\,dy = \int_{-\infty}^{\infty} x g(x)\,dx \\
\mu_y &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y)\,dx\,dy = \int_{-\infty}^{\infty} y h(y)\,dy \\
\mu'_x &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x) f(x,y)\,dx\,dy = \int_{-\infty}^{\infty} (x-\mu_x) g(x)\,dx \\
\mu'_y &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y-\mu_y) f(x,y)\,dx\,dy = \int_{-\infty}^{\infty} (y-\mu_y) h(y)\,dy
\end{aligned}
$$

$$
\begin{aligned}
\mu_{lm} &= E(x^l y^m) \\
\mu'_{lm} &= E\left[(x-\mu_x)^l (y-\mu_y)^m\right]
\end{aligned}
$$

$$
\begin{aligned}
\mu_{lm} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^l y^m f(x,y)\,dx\,dy \\
\mu'_{lm} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)^l (y-\mu_y)^m f(x,y)\,dx\,dy
\end{aligned}
$$

- similarly, we define

$$E(\mathrm{u(x,y)}) = \iint u(x,y)\,f(x,y)\,dxdy$$

$$\sigma^2(\mathrm{u(x,y)}) = E\left\{\left[\mathrm{u(x,y)} - E(\mathrm{u(x,y)})\right]^2\right\} = E\left\{\mathrm{u}^2\mathrm{(x,y)}\right\} - \left(E\left\{\mathrm{u(x,y)}\right\}\right)^2$$

- **examples**

$$\mathrm{u(x,y)} = a\mathrm{x} + b\mathrm{y} \;\Rightarrow\; E(a\mathrm{x} + b\mathrm{y}) = aE(\mathrm{x}) + bE(\mathrm{y})$$

$$\sigma^2(a\mathrm{x} + b\mathrm{y}) = E\left[\left((a\mathrm{x} + b\mathrm{y}) - E(a\mathrm{x} + b\mathrm{y})\right)^2\right] =$$

$$= E\left[\left(a(\mathrm{x} - \mu_x) + b(\mathrm{y} - \mu_y)\right)^2\right] =$$

$$= E\left[a^2(\mathrm{x} - \mu_x)^2 + b^2(\mathrm{y} - \mu_y)^2 + 2ab(\mathrm{x} - \mu_x)(\mathrm{y} - \mu_y)\right] =$$

$$= a^2\sigma^2(\mathrm{x}) + b^2\sigma^2(\mathrm{y}) + 2ab\,\mathrm{cov}(x,y) \quad \text{(cf. Chap. 3)}$$

$$\mathrm{u(x,y)} = \mathrm{xy} \text{ and x,y independent, i.e., } f(x,y) = g(x)h(y) \Rightarrow$$

$$\text{i) } E(\mathrm{xy}) = \iint xyf(x,y)\,dxdy = \iint xyg(x)h(y)\,dxdy = \left(\int xg(x)\,dx\right)\left(\int yh(y)\,dy\right) =$$

$$= E(\mathrm{x})E(\mathrm{y}) \quad \text{(cf. Chap. 3)}$$

$$\text{ii) } \mathrm{cov(x,y)} = \iint (x - \mu_x)(y - \mu_y)g(x)h(y)\,dxdy = 0 \text{ !!!}$$

Covariance, correlation coefficient

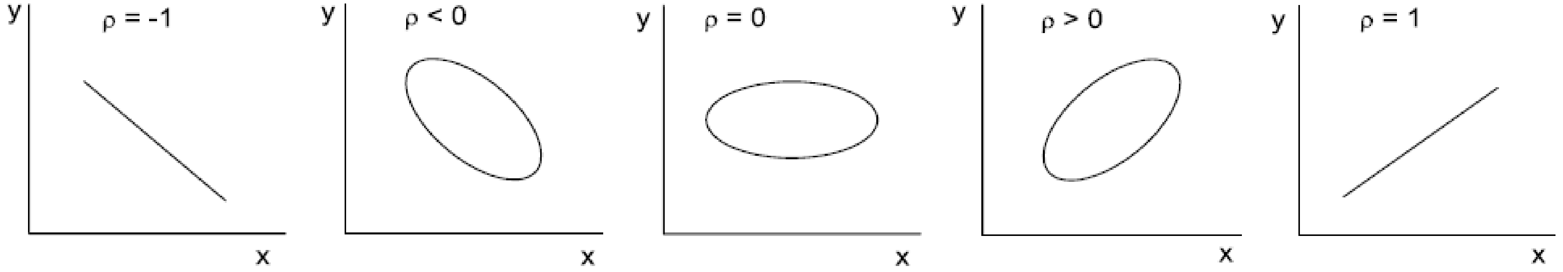
- from definition of covariance, we see that
  - cov(x,y) is positive if values $x>\mu_x$ ($x<\mu_x$) appear preferentially together with values $y>\mu_y$ ($y<\mu_y$).
  - cov(x,y) is negative if values $x>\mu_x$ ($x<\mu_x$) appear preferentially together with values $y<\mu_y$ ($y>\mu_y$).
  - if the knowledge of x does not give information about the probable position of y, the covariance vanishes (see Fig. below)
- if cov(x,y) ≠0, the variables x,y are called correlated.
- correlation is quantified by the dimensionless correlation coefficient

$$\rho(\mathrm{x,y}) = \frac{\mathrm{cov(x,y)}}{\sigma(\mathrm{x})\sigma(\mathrm{y})}, \quad -1 \le \rho(\mathrm{x,y}) \le 1;$$

the limiting values are reached when $\mathrm{y} = a + b\mathrm{x}$ and $b>0$ ($\rho = 1$) or $b<0$ ($\rho = -1$)

proof: calculate

$$\left.\begin{aligned} \mathrm{cov(x,y)} &= E(\mathrm{xy}) - E(\mathrm{x})E(\mathrm{y}) \\ \text{or } \mathrm{cov(x,y)} &= \frac{1}{2}\left[\sigma^2(\mathrm{x}+\mathrm{y}) - \sigma^2(\mathrm{x}) - \sigma^2(\mathrm{y})\right] \end{aligned}\right\} \text{with } \mathrm{y} = a + b\mathrm{x} \text{ and then } \rho \to \frac{b}{\sqrt{b^2}} = \pm 1$$



f(x,y)=const for different correlation coefficients (linear dependence: f(x,y)=f(y|x)f(x) with f(y|x)=δ(y-(a+bx))

- Note: for independent (uncorrelated) variables → cov(x,y)=0
- But: cov(x,y)=0 does not necessarily imply that x,y are independent, *since covariance detects only linear dependencies*.
- Example: let x be uniformly distributed between [-1,1], and $y=x^2$
  - Then: y depends on x, but $cov(x,y)=E(x^3)-E(x)E(x^2)=0$, since expectation values of odd quantities=0 for symmetric intervalls!
- In other words: there are cases when cov(x,y)=0, but the conditional p.d.f. *f(y|x)* depends on *x*.
- Independence is only warranted if *f(y|x) = f(y)* [or, equivalently, *f(x,y)=g(x)h(y)*]!

Transformation of variables

- analogous to 1-D (univariate) case
  - given f(x,y) and u(x,y), v(x,y)

$$\text{Then: } g(u,v)dudv = f(x,y)dxdy \Rightarrow g(u,v) = f(x,y)\cdot \underbrace{\left|\frac{\partial x,\partial y}{\partial u,\partial v}\right|}_{\substack{\text{absolute value of}\\ \text{Jacobi-determinante}}}$$

- **example**

Transform 2-D *normal* distribution $\frac{1}{2\pi}e^{-(x^2+y^2)/2}$ into polar coordinates $x = r\cos\varphi, y = r\sin\varphi$

$$\left|\frac{\partial x,\partial y}{\partial r,\partial\varphi}\right| = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial\varphi} & \frac{\partial y}{\partial\varphi} \end{vmatrix} = \begin{vmatrix} \cos\varphi & \sin\varphi \\ -r\sin\varphi & r\cos\varphi \end{vmatrix} = r$$

$$\Rightarrow g(r,\varphi) = \frac{1}{2\pi} re^{-r^2/2}, \text{with marginal distributions}$$

$$g_r = \int_0^{2\pi} g(r,\varphi)d\varphi = re^{-r^2/2} \text{ and } g_\varphi = \int_0^{\infty} g(r,\varphi)dr = \frac{1}{2\pi}, \text{ i.e.,}$$

the distribution factorizes into the marginal distributions (independent variates!)

Reduction of variables

- problem: we have *f(x,y),* and need *g(u)* with *u(x,y).*
- solution: use standard transformation, by introducing a 2nd variable *v(x,y)* (usually, choose *v*=*x*)

$$f(x,y) \to h(u,v)$$

and marginalize with respect to u

$$g(u) = \int h(u,v)dv$$

- example: given 2-D uniform distribution

$$f(x,y) = \begin{cases} \frac{1}{\Delta^2} & \text{if } x \in [0,\Delta] \text{ and } y \in [0,\Delta] \\ 0 & \text{else} \end{cases}$$

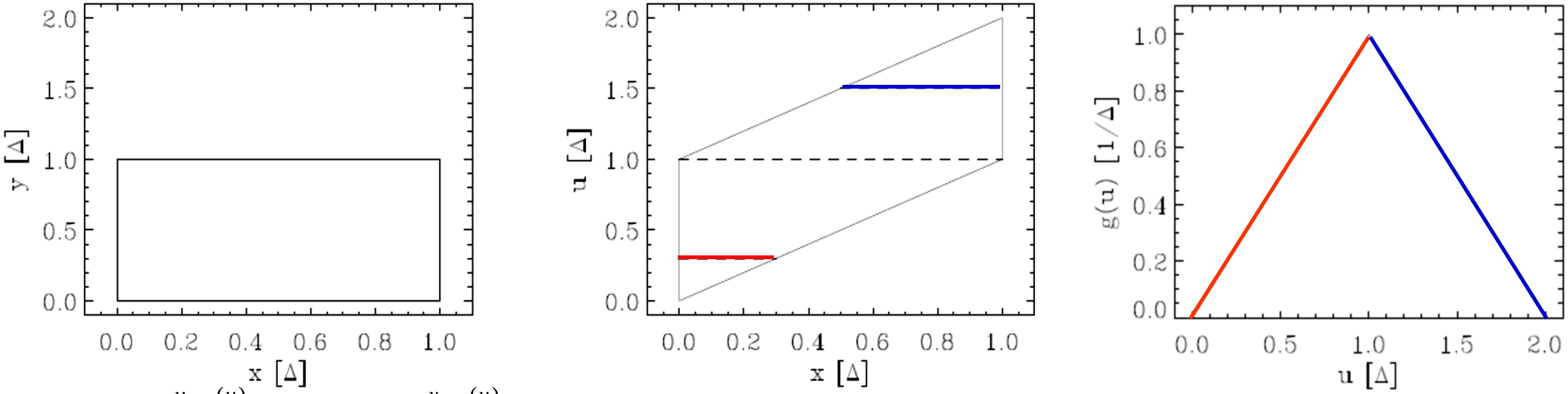
see next page, left figure

Calculate $g(x+y)$! (see also Chap. 3, page 57/58 and Chap. 4, page 73)

Note: $f(x,y)$ already normalized

$$u = x+y \qquad v = x$$

$$\left|\frac{\partial x,\partial y}{\partial u,\partial v}\right| = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 0 \end{vmatrix} = 1 \Rightarrow h(u,v) = f(x,y) = \frac{1}{\Delta^2}$$



$$g(u) = \int_{v_{\min}(u)}^{v_{\max}(u)} h(u,v)dv = \int_{x_{\min}(u)}^{x_{\max}(u)} h(u,x)dx = \frac{1}{\Delta^2}(x_{\max}(u) - x_{\min}(u))$$

From above figure (middle): $u(x) = (x + y) \in [x, x+\Delta]$, since $y \in [0, \Delta]$

$u < \Delta : x(u) \in [0, u]$ (slope=1) $\quad \Rightarrow \quad g(u) = \frac{1}{\Delta^2}(u - 0) = \frac{1}{\Delta^2} u$

$u > \Delta : x(u) \in [u - \Delta, \Delta] \quad \Rightarrow \quad g(u) = \frac{1}{\Delta^2}(\Delta - (u - \Delta)) = \frac{1}{\Delta^2}(2\Delta - u)$

$$g_{\max} = g(\Delta) = \frac{1}{\Delta}$$

The distribution of the sum of two uniformly distributed quantities is triangular-shaped, see above figure (right). See also Chap. 3, page 57/58

Note: the distribution of $x - y$ looks similar, when the abscissa is shifted by -Δ

Calculation of the transformation

- as in the 1-D case: integration and inversion of primitive function
- **important example**: Box-Muller algorithm to create *normally* distributed variates from uniform distribution (random number generator)

remember 2-D *normal* distribution in polar coordinates

$$g(r,\varphi)drd\varphi = \frac{1}{2\pi} re^{-r^2/2} drd\varphi \;\; \text{(factorized in } r \text{ and } \varphi\text{)}$$

distribution in r:

$$G(r) = \int_0^r r' e^{-r'^2/2} dr' = F(x_1) = x_1 \;\; \text{(uniform distribution w.r.t. } [0,1]\text{)}$$

$$G(r) = -e^{-r'^2/2} |_0^r = x_1; \quad r = \sqrt{-2\ln(1-x_1)}$$

distribution in φ:

$$H(\varphi) = \int_0^\varphi \frac{1}{2\pi} d\varphi' = F(x_2) = x_2 \;\; \text{(uniform distribution } F\text{)}$$

$$H(\varphi) = \frac{\varphi}{2\pi} = x_2; \quad \varphi = 2\pi x_2$$

in Carthesian coordinates:

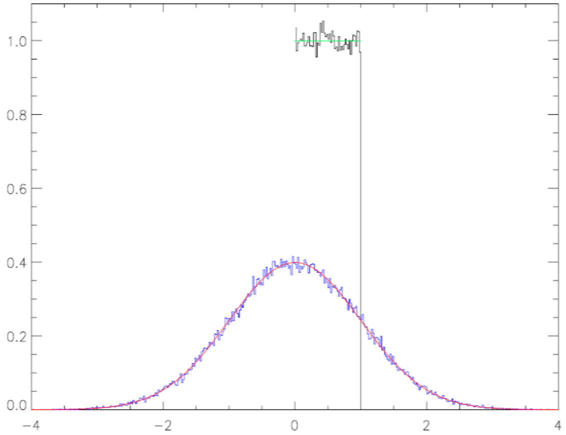
$$x = r\cos\varphi = \sqrt{-2\ln(1-x_1)}\cos(2\pi x_2) \triangleq \sqrt{-2\ln(x_1)}\cos(2\pi x_2)$$

$$y = r\sin\varphi = \sqrt{-2\ln(1-x_1)}\sin(2\pi x_2) \triangleq \sqrt{-2\ln(x_1)}\sin(2\pi x_2)$$

These variables are independent and *normally* distributed with expectation value zero and unit variance.

$$f(x,y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

Thus: two uniformly distributed variates x_1, x_2 $\Rightarrow$ two *normally* distributed variates x,y



P.d.f.s for a uniform distribution (black), generated by a random number generator from N=10^5 subsequent numbers. The corresponding normal distribution (blue) has been created from these numbers using the Box-Muller algorithm. Displayed are histograms with bin size 0.02. Analytical p.d.f.s in green and red.

Distributions with more than two variables

probability density

$f(x_1, x_2, x_3, ...x_N) = f(\mathbf{x})$ (in vector notation)

expectation value

$$E(\mathrm{u}) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} u(\mathbf{x}) f(\mathbf{x}) \prod_{i=1}^{N} dx_i$$

particularly important is covariance matrix C,

$C_{ij} = \mathrm{cov}(\mathrm{x_i}, \mathrm{x_j}) = E\left\{(\mathrm{x_i} - \mu_i)(\mathrm{x_j} - \mu_j)\right\}$ (see also Chap. 6)

The covariance matrix is symmetric, and the diagonal elements are the variances: $C_{ii} = Var(\mathrm{x_i}) = \sigma^2(\mathrm{x_i})$

Matrix notation: with $\mathbf{x}^\mathrm{T} = (x_1, x_2, x_3, ...x_N)$ and $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_N \end{pmatrix}$,

$$\mathbf{C} = E\left\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathrm{T}\right\}$$

transformation of variables with Jacobi determinant

$$\mathrm{g}(\mathbf{y}) = f(\mathbf{x}) \left| \frac{\partial x_1 \ldots \partial x_N}{\partial y_1 \ldots \partial y_N} \right|$$

independent, identically distributed (i.i.d.) variables

(u.i.v. = unabhängig, identisch verteilt)

For parameter estimates, sample of N independent measurements might be used. The joint p.d.f. $\tilde{f}$ for N independent variables which are identically distributed according to $f(x)$ is given by

$$\tilde{f}(\mathbf{x}) = \prod_{i=1}^{N} f(x_i)$$

Binomial distribution

- experiment with two mutually exclusive outcomes, i.e.,

  $S = A + \bar{A}$ with $P(A) = p$ and $P(\bar{A}) = 1 - p = q$

- calculate the probability that *n* experiments have *k* times the outcome *A*.
  - What is the probability to obtain (exactly!) 4 times the six when rolling the die 10 times? Answer: ≈0.054

    $n = 10, \quad k = 4, \quad p(A) = \frac{1}{6} \qquad p(\bar{A}) = \frac{5}{6}$

  - What is the probability to toss “number” only one time in 20 trials? Answer: ≈1.91 $\cdot 10^{-5}$

    $n = 20, \quad k = 1, \quad p(A) = \frac{1}{2} \qquad p(\bar{A}) = \frac{1}{2}$

let's assign the random variable x_i to the outcome of experiment *i*.

$\mathrm{x}_i = 1$ if the result *A* occurs, and $\mathrm{x}_i = 0$ if $\bar{A}$ occurs. Our above question can be rephrased then to the question regarding the probability distribution of the random number

$$\mathrm{x} = \sum_{i=1}^{n} \mathrm{x}_i,$$

and, particularly, to the probability $\mathrm{P}(\mathrm{x} = k)$

- answer depends on two factors

i) What is the probability to obtain the result A in the *first* k experiments and to obtain $\bar{A}$ in the remaining $n-k$?

Since the experiments are independent, this probability is given by the product of the probabilities of the individual events, i.e.,

$$p^k(1-p)^{n-k}$$

ii) How many possibilities for the event "k times result A in n experiments" do exist? This is given by the binomial coefficients,

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

Thus, the probability $P(\mathrm{x}=k)$ is given by

$$B_p^n(k) = \frac{n!}{k!(n-k)!} p^k(1-p)^{n-k}$$

- expectation value and variance

expectation value and variance of *single* experiment

$$E(\mathrm{x_i}) = 1\cdot p + 0\cdot(1-p) = p$$

$$Var(\mathrm{x_i}) = E(\mathrm{x_i}^2) - \left(E(\mathrm{x_i})\right)^2 = \left[1^2\cdot p + 0^2\cdot(1-p)\right] - p^2 = p(1-p) = pq$$

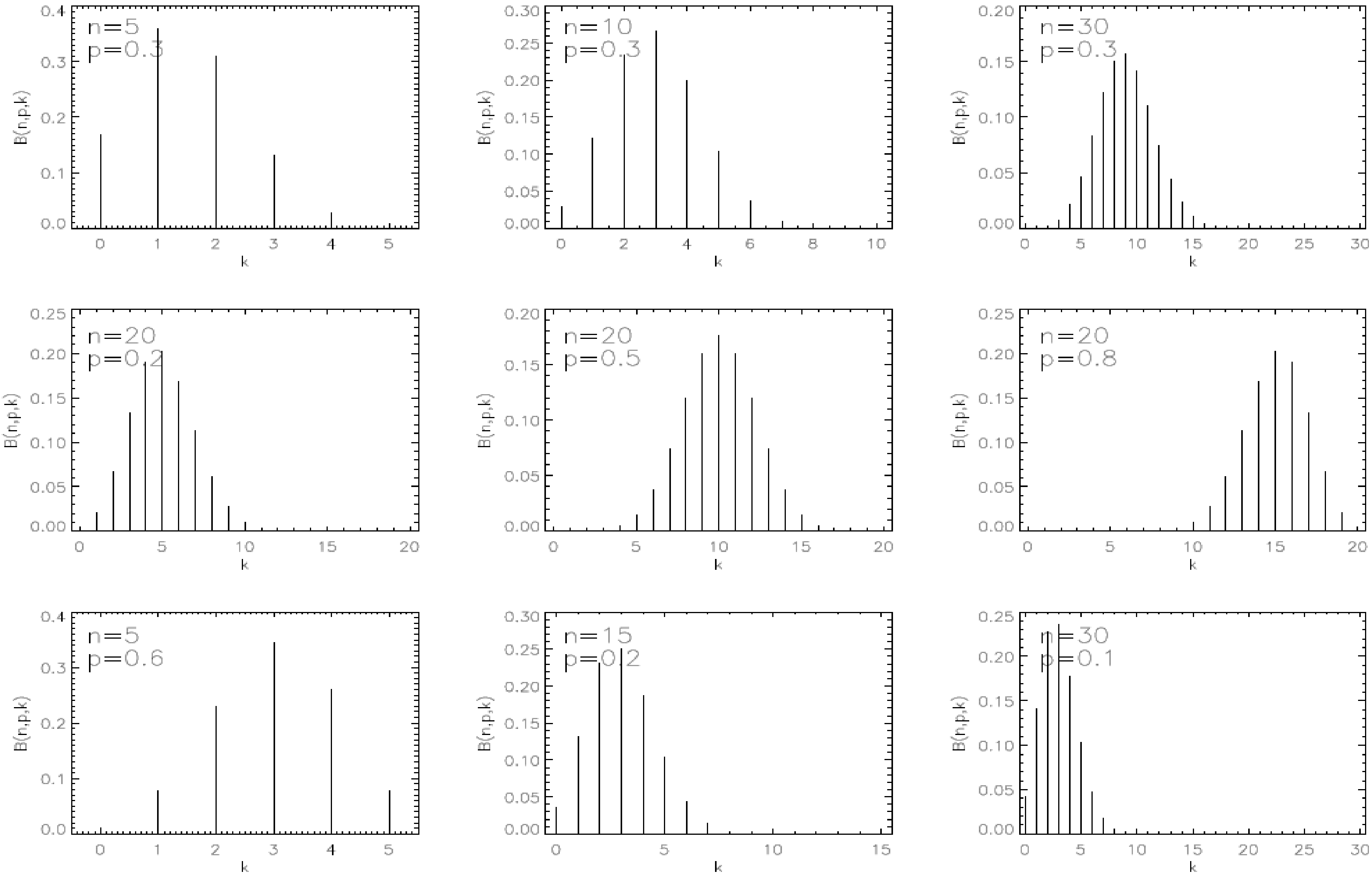
The corresponding values for the random variable $\mathrm{x}=\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{x}_i$ are

(exploiting the calculation rules for independent variates)

$$\boxed{\begin{array}{l} E(\mathrm{x}) = \langle \mathrm{k} \rangle = np \qquad \text{"mean number of successes"} \\ Var(\mathrm{x}) = \sigma^2(\mathrm{x}) = np(1-p) = npq \end{array}}$$

(cumulative) distribution function

$$F(k) = F(\mathrm{k} < k) = \sum_{k'=0}^{k-1} B_p^n(k') = \sum_{k'=0}^{k-1} \frac{n!}{k'!\left(n-k'\right)!}\, p^{k'}(1-p)^{n-k'}$$



Binomial distribution, $B_p^n(k)$, as a function of *k*. Top panel: fixed *p*, different *n*; middle: fixed *n*, different *p*; bottom: different values of *n* and *p*, but *np*=const

- Example: Detector efficiency
  - spark chambers (95% efficient) are used to measure the tracks of cosmic rays. At least three points are needed to define a track. How efficient is a stack of three chambers? Would using 4 or 5 chambers give significant improvement?

The probability of three hits from three chambers is

$$P(3;3,0.95) = B^{3}_{0.95}(3) = \frac{3!}{3!0!}\,p^{3}(1-p)^{0} = 0.95^{3} = 0.857$$

For four chambers, the probability of three or four hits is

P(3;4,0.95)+P(4;4,0.95)=0.171+0.815=0.986

For five chambers, the probability of three, four or five hits is

P(3;5,0.95)+P(4;5,0.95)+P(5;5,0.95)=0.021+0.204+0.774=0.999!

A weak law of big numbers

- Application of Chebychev’s inequality (page 49/50) to the binomial distribution results in a *weak* law of big numbers:
  The expectation value for the *frequency* of an event A (*n* experiments with *k* times event A occuring) is given by *E(k/n)*= *p*, with variance *Var(k/n)=1/n*^{2} *Var(k)=p(1-p)/n*. Thus,

  $$P\left(\left|\frac{k}{n}-p\right|\geq\varepsilon\right)\leq\frac{p(1-p)}{n\varepsilon^2}\leq\frac{1}{4n\varepsilon^2}$$

  $$\left[\begin{array}{l}\text{applying the inequality of arithmetic and geometric mean (Chap. 1) for n=2}\\ \sqrt{p(1-p)}\leq\frac{p+(1-p)}{2}=\frac{1}{2}\end{array}\right],$$

  and the frequency converges stochastically towards its expectation value.

Multinomial distribution

- binomial distribution: 2 different outcomes
- multinomial: more than 2 different outcomes, mutually exclusive!

$S = A_1 + A_2 + A_3 + \ldots + A_l$ with $P(A_j) = p_j$ and $\sum_{j=1}^{l} p_j = 1$

When n experiments are performed, the probability of finding k_j events A_j is given by

$$M^{n}_{p_1,p_2,p_3,\ldots,p_l}(k_1,k_2,k_2,\ldots,k_l) = \frac{n!}{\prod_{j=1}^{l} k_j!} \prod_{j=1}^{l} p_j^{\,k_j}$$

We define $\mathrm{x}_{ij} = 1$ if experiment i yields A_j, and 0 otherwise. Then

$\mathrm{x}_j = \sum_{i=1}^{n} \mathrm{x}_{ij}$ and

$E(\mathrm{x}_\mathrm{j}) = np_j$, with covariance matrix

$C_{ij} = np_i(\delta_{ij} - p_j)$ (δ_{ij} Kronecker δ), i.e.,

$C_{ii} = np_i(1 - p_i)$ as before, but nonvanishing, negative covariance $C_{ij} = -np_i p_j$

That there *is* a correlation was to be expected, since the x_j are not independent due to the constraint $\sum p_j = 1$. I.e., if there are more successes for class i than expected ($E(\mathrm{x}_\mathrm{i})$), the values of x_j for all other classes j are smaller than $E(\mathrm{x}_\mathrm{j})$

$\Rightarrow$ negative correlation!

Frequency; law of big numbers

- probabilities, e.g., p_j in case of the multinomial distribution, are usually not known a priori but have to be obtained from experiments. The *frequency* of event A_j in n experiments is given by

$$\mathrm{h} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{ij} = \frac{1}{n}\mathrm{x}_{j}$$

- This frequency is a random number, since it depends on the results of the particular n experiments.

$$E(\mathrm{h}) = E\left(\frac{\mathrm{x}_\mathrm{j}}{n}\right) = \frac{1}{n}E(\mathrm{x}_\mathrm{j}) = p_j,$$

i.e., the expectation value of the frequency of an event is the corresponding probability , and

$$Var(\mathrm{h}) = Var\left(\frac{\mathrm{x}_\mathrm{j}}{n}\right) = \frac{1}{n^2}Var(\mathrm{x}_\mathrm{j}) = \frac{1}{n}p_j(1-p_j) \Rightarrow \sigma(\mathrm{h}) \propto \frac{1}{\sqrt{n}}$$

This is the law of big numbers! For large *n*, the standard deviation of the frequency vanishes below any given limit, which "justifies" the frequency definition of probability (cf. Chap. 2, p. 18).

Poisson distribution

- The study of the lower panel of the last figure (binomial distribution) suggests that this distribution approaches a fixed distribution if *n* tends to infinity but the product (the expectation value) *np*=*λ* is kept constant.
- Indeed,

$$P(k;n,p) = P(k;n,\lambda/n) = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$\frac{n!}{(n-k)!} = n(n-1)(n-2)\cdots(n-k+1) \to n^k \text{ for } n \to \infty$$

$$\left(1-\frac{\lambda}{n}\right)^{n-k} \to \left(1-\frac{\lambda}{n}\right)^{n} \to e^{-\lambda} \text{ for } n \to \infty \quad \text{(definition of the exp function)}$$

Thus,

$$P(k;n,\lambda/n) \xrightarrow{n\to\infty} P(k,\lambda) = \frac{n^k}{k!}\left(\frac{\lambda}{n}\right)^k e^{-\lambda} = \frac{e^{-\lambda}\lambda^k}{k!}$$ which is the Poisson-distribution

and describes the probability of obtaining k events if the expected number is λ

calculation: start with P(0)=$e^{-\lambda}$, and then successively multiply by λ

and divide by 1,2,3,4,.... to obtain P(1), P(2) etc.

Interpretation

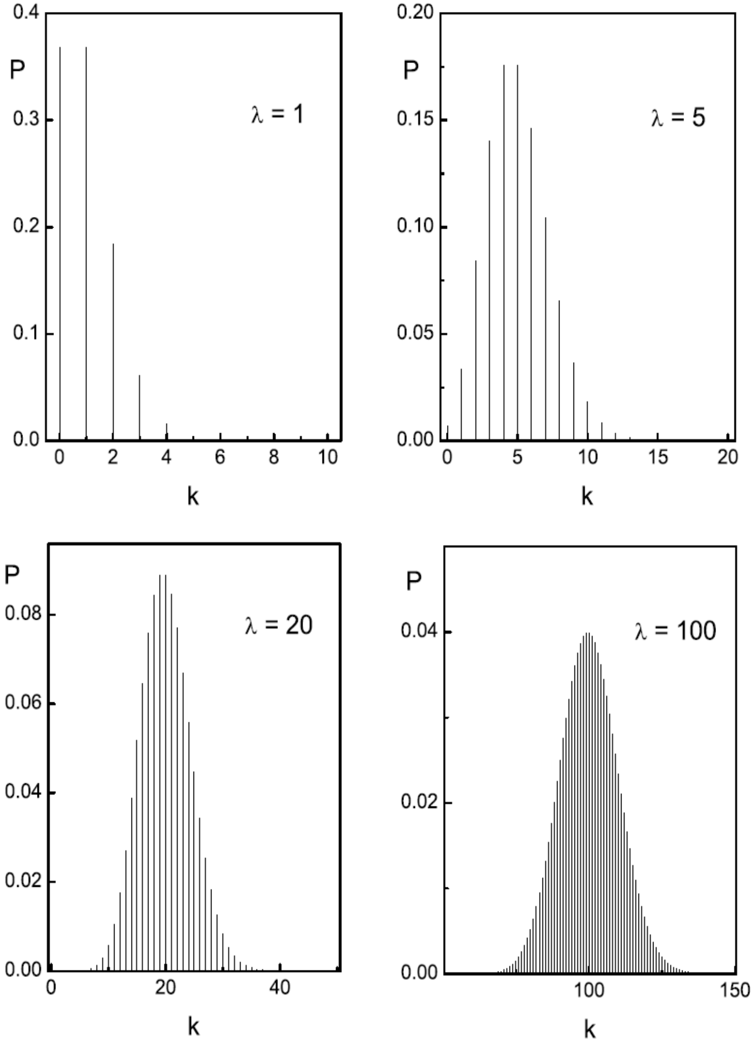
- Suppose λ events are *expected* to occur in some interval. Split up this interval into n very small sections, so that the chance to find two events in one section is negligible. The probability that one section contains one event is then $p=\lambda/n$.

- The probability of finding k events in the n sections is given by the binomial distribution,

  $P(k;n,p=\lambda/n)$

  which approaches the Poisson distribution for large n.

- Note: the Poisson distribution is defined only for integer values of k!



Poisson distribution for different expectation values

Total probability

$$\sum_{k=0}^{\infty} P(k,\lambda) = 1$$

expectation value and variance

$$E(\mathrm{k}) = \lambda$$

$$Var(\mathrm{k}) = \sigma^2(\mathrm{k}) = \lambda$$

[this is consistent with the binomial distribution:

$$E(\mathrm{k}) = np = n\frac{\lambda}{n} = \lambda \text{ and}$$

$$\mathrm{Var}(\mathrm{k}) = np(1-p) = n\frac{\lambda}{n}(1-\frac{\lambda}{n}) \rightarrow \lambda \text{ for } n \rightarrow \infty]$$

all cumulants are identical (=λ)!

skewness

$$\mu'_3 = \lambda \text{ (third central moment)} \rightarrow$$

$$\gamma = \frac{\mu'_3}{\sigma^3} = \frac{\lambda}{\lambda^{3/2}} = \lambda^{-1/2},$$

i.e., the distribution becomes increasingly symmetric for increasing λ

- application:

  Poisson distribution describes asymptotic behavior of binomial distribution with constant *$\lambda=np$*, i.e., with a (very) low probability for the individual process. Thus, it should be applied when there are many trials but only few successes, and the mean (expectation value) is known. Since one has no idea on the number of trials (only that there are many), it describes the cases of *sharp* events occurring in a continuum.

- examples:
  - the number of flashes of lightning in a thunderstorm (it is meaningless to ask how often there is no flash) when the mean is known
  - the number of clicks in a Geiger counter (meaningless to ask about “non-clicks”) when the mean is known
  - the actual number of photons from an average signal/background

A historical example

- Statistics on the numbers of Prussian soldiers kicked to death by horses. In the 19th century is was reported that there were 122 deaths in ten different army corps over twenty years, i.e., the mean number of deaths per corps and per year is λ=122/200=0.61.
- The probability of, e.g., *no* death is then P(0,0.61)=0.5434 per year and corps.
- In twenty years and ten corps, there should be 108.7 cases where no death should have happened. Actually, 109 such events have been reported.

Number of deaths per year and corps	actual number reported for 20 years and 10 corps	predictions from Poisson statistics
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6

- (1*65+2*22+3*3+4*1=122)

Supernova 1987A

- The following table gives the numbers of neutrino events detected in 10 s intervals by the Irvine-Michigan-Brookhaven experiment on Feb. 23rd 1987 (around which time SN1987A has been firstly seen)

no. of events	0	1	2	3	4	5	6	7	8	9
no. of intervals	1042	860	307	78	15	3	0	0	0	1
prediction	1064	823	318	82	16	2	0.3	0.03	0.003	0.0003

- The average number of events per interval (ignoring the interval with 9 events) is 0.77
- The Poisson predictions agree well with the data, except for the interval with the 9 events. Thus, the background due to random events is Poisson and well understood, and the nine events cannot be due to fluctuations, but must have come from a different event (the supernova).

Two Poisson distributions

- If there are two separate types of Poisson distributed events, and we do not distinguish between the two, then the probability of $k=k_1+k_2$ events is also Poisson, with mean equal to the sum of the two individual means.

$$P(k)=\sum_{k_1=0}^{k} P(k_1,\lambda_1)P(k-k_1,\lambda_2)=P(k,\lambda_1+\lambda_2)$$

Proof via characteristic function of Poisson distribution

$$\phi_{\mathrm{P}}(t)=\sum_{k=0}^{\infty} e^{itk}P(k,\lambda)=\sum_{k=0}^{\infty} e^{itk}\frac{\lambda^k e^{-\lambda}}{k!}=e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{it})^k}{k!}=e^{-\lambda}\exp(\lambda e^{it})=\exp\left[\lambda(e^{it}-1)\right]$$

Remember: characteristic function of sum of independent variables is product of their characteristic functions (Chap. 3) $\Rightarrow$

$$\phi_{\mathrm{sum}}(t)=\phi_{\mathrm{P}(\lambda_1)}(t)\phi_{\mathrm{P}(\lambda_2)}(t)=\exp\left[\lambda_1(e^{it}-1)\right]\exp\left[\lambda_2(e^{it}-1)\right]=$$

$$=\exp\left[(\lambda_1+\lambda_2)(e^{it}-1)\right]=\phi_{\mathrm{P}(\lambda_1+\lambda_2)}(t).$$

Thus, the sum of two independent, Poisson distributed variables is Poisson-distributed as well, with $\lambda=\lambda_1+\lambda_2$

- can be generalized to any number of Poisson processes

- example: signal with background
  - expected are S signals with an average background B. The average fluctuation (standard dev.) of the observed number of events k is thus
    $\sigma(\mathrm{S+B}) = \sqrt{S+B}$
  - If we subtract the average background from the signal, this fluctuation remains conserved, of course.

    For an expected signal $S=100$ and background $B=50$ we observe on average 150 events with a standard deviation of $\sqrt{150}$. After subtracting the background, the average signal is $S=100 \pm \sqrt{150}$
  - If the exact expectation value of the background is not known, the uncertainty is even larger (error propagation)

Uniform distribution

- so far, only distributions of one or more *discrete* variables discussed
- will now turn to continuous distribution functions
- most simple case: the uniform distribution (already mentioned before):

constant probability density in a certain interval, elsewhere 0.

$$f(x) = c \qquad a \leq x < b$$

$$f(x) = 0 \qquad x < a,\ x \geq b$$

From the normalization, $\int_{-\infty}^{\infty} f(x)dx = 1,$ we obtain

$$c = \frac{1}{b-a},$$

and the distribution function becomes

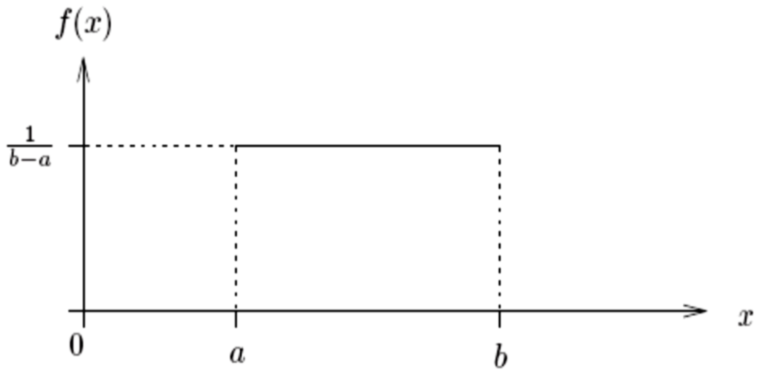
$$F(x) = \int_a^x \frac{1}{b\text{-}a}dx = \frac{x-a}{b-a} \qquad a \leq x < b$$

$$F(x) = 0 \qquad x < a$$

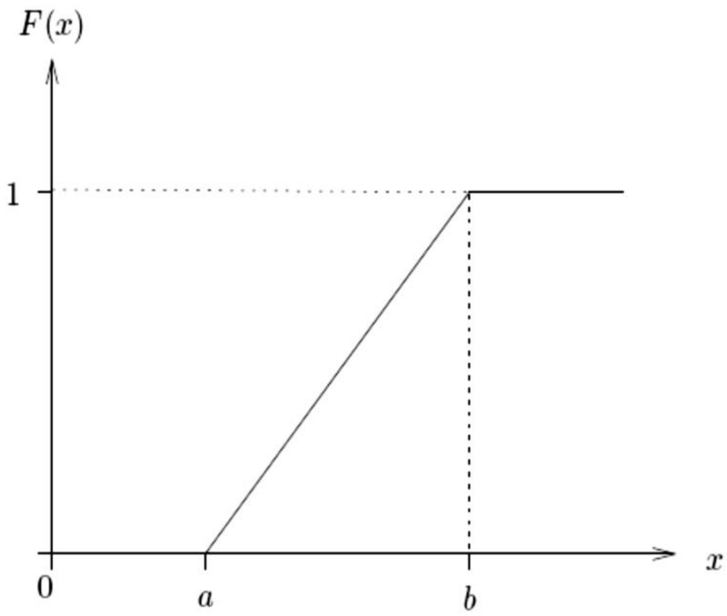
$$F(x) = 1 \qquad x \geq b$$

$$E(x) = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{2}(a+b)$$

$$Var(x) = \frac{1}{12}(b-a)^2$$



Probability density function (pdf), $f(x)$, of a *uniform* distribution within the interval [a,b].



Corresponding cumulative probability distribution, $F(x)$.

- uniform distributions with a=0, b=1, i.e., f(x)=1 and F(x)=x, created by random number generators (RNGs).
  Note: in many RNGs, “0” not included, i.e., lowermost value =ε (machine dependent)
- important for Monte Carlo methods
- different distributions obtained from transformation methods (see Chap. 3/4)

Gaussian (or normal) distribution

- assume binomial distribution with random variable k

$$P(k;n,p)=\frac{n!}{k!\left(n-k\right)!}p^{k}(1-p)^{n-k}$$

characteristic function (see exercises):

$$\phi(t)=\sum_{k=0}^{n}e^{itk}P(k;n,p)=\left[\exp(it)p+(1-p)\right]^{n}$$

consistent with charact. function of Poission distribution: $p=\frac{\lambda}{n}\rightarrow\phi(t)=\left(1-\frac{\lambda}{n}+\exp(it)\frac{\lambda}{n}\right)^{n}=$

$$\left(1+\frac{\lambda(\mathrm{e}^{it}-1)}{n}\right)^{n}\overset{n\rightarrow\infty}{\rightarrow}\exp\left(\lambda(\mathrm{e}^{it}-1)\right),$$ cf. page 101

use reduced variable

$$\mathrm{u}=\frac{\mathrm{k}-\left\langle\mathrm{k}\right\rangle}{\sigma}=\frac{\mathrm{k}-np}{\sigma}$$

$$\Rightarrow\phi_{u}(t)=\exp\left(-\frac{itnp}{\sigma}\right)\left[\exp\left(\frac{it}{\sigma}\right)p+(1-p)\right]^{n}$$ (see exercises)

Gaussian (or normal) distribution

$$\phi_u(t) = \exp\left(-\frac{itnp}{\sigma}\right)\left[\exp\left(\frac{it}{\sigma}\right)p + (1-p)\right]^n$$

$$\ln\phi_u(t) = -\frac{itnp}{\sigma} + n\ln\left[1 + p\left(\exp\left(\frac{it}{\sigma}\right) - 1\right)\right]$$

Expand in t/σ $\left(\propto t/\sqrt{np(1-p)},\ \text{small quantity for large } n\right)$

$$\ln\phi_u(t) = -\frac{itnp}{\sigma} + n\ln\left[1 + p\left(\frac{it}{\sigma} - \frac{1}{2}\left(\frac{t}{\sigma}\right)^2 + \dots\right)\right]$$

$$\left[\ln(1+\varepsilon) = \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3} - \dots\right]$$

$$\Rightarrow \ln\phi_u(t) = -\frac{itnp}{\sigma} + n\left(p\left(\frac{it}{\sigma} - \frac{1}{2}\left(\frac{t}{\sigma}\right)^2 + \dots\right) - \frac{1}{2}p^2\left(\frac{it}{\sigma} - \frac{1}{2}\left(\frac{t}{\sigma}\right)^2 + \dots\right)^2 + \dots\right) =$$

$$= -\frac{itnp}{\sigma} + n\left(\left(\frac{itp}{\sigma} - \frac{1}{2}p\left(\frac{t}{\sigma}\right)^2\right) + \frac{1}{2}p^2\left(\frac{t}{\sigma}\right)^2 + O(\sigma^{-3})\right)$$

$$= -\frac{itnp}{\sigma} + n\left(\frac{itp}{\sigma} - \frac{1}{2}\left(\frac{t}{\sigma}\right)^2\left(p - p^2\right) + O(\sigma^{-3})\right) =$$

$$= -\frac{1}{2}t^2\frac{np(1-p)}{\sigma^2} + n\cdot O(\sigma^{-3})$$

Thus, accounting for $\sigma^2 = np(1-p)$ and in the limit of n $\to \infty$, we find

$$\phi_u(t) = \exp\left(-\frac{1}{2}t^2\right)$$

This is the characteristic function of a binomial distribution, using a reduced random variable, in the limit of large n [no assumption regarding p, contrasted to Poisson]
Back-transformation yields the corresponding p.d.f.,

$$\boxed{f(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)}$$

which is called the Gaussian or normal distribution.

Since u is a reduced variable, $E(\mathrm{u})$ should be 0 and $Var(\mathrm{u})$ should be 1.
Test:

$$E(\mathrm{u}) = \int_{-\infty}^{\infty} u\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right) = 0$$

$$Var(\mathrm{u}) = -\left.\frac{d^2\phi'(t)}{dt^2}\right|_{t=0} \overset{E(\mathrm{u})=0}{=} -\left.\frac{d^2\phi(t)}{dt^2}\right|_{t=0} = 1, \qquad \text{q.e.d.}$$

A more general form of the normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi b}} \exp\left(-\frac{(x-a)^2}{2b^2} \right).$$

Since $E(\mathrm{x}) = a$ and $Var(\mathrm{x}) = b^2$, the conventional representation is

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

The inflection points of this distribution (zero curvature) are located at $\mu \pm \sigma$. Once again, this is the limit of a binomial distribution with the above expectation value and variance, in the limit $n \to \infty$.

The corresponding characteristic function is (with $\mathrm{x}=\sigma\mathrm{u} + \mu$)

$$\phi_x(t) = \int e^{itx} f(x)dx = \exp(it\mu)\exp\left(-\frac{1}{2}\sigma^2 t^2 \right) \quad (\text{see exercise})$$

Theorem: The characteristic function of a normal distribution with zero mean is itself a normal distribution with zero mean. The product of the variances of both distributions is one.

The characteristic function transformed to y = x − μ is

$$\phi'(t) = \exp\left(-\frac{1}{2}\sigma^2 t^2\right)$$

With $\mu'_\mathrm{n} = \frac{1}{i^n}\frac{d^n\phi'(t)}{dt^n}\Big|_{t=0}$ (Chap. 3), we find the central moments

$\mu'_1 = 0, \quad \mu'_2 = \sigma^2, \quad \mu'_3 = 0, \quad \mu'_4 = 3\sigma^4$ (remember curtosis, Chap. 3),

and

$$\mu'_{2k+1} = 0,\ k = 0,1,2,3,...$$

$$\mu'_{2k} = \frac{(2k)!}{2^k k!}\sigma^{2k}.$$

Cumulants: $\kappa_1 = 0$ (here; otherwise = μ),

$\kappa_2 = \sigma^2,\ \kappa_\mathrm{n} = 0$ for n ≥ 3

Corresponding cumulative distribution functions are

$$\psi_o(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{1}{2}x^2\right)dx$$

$$\psi(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(x-\mu)/\sigma}\exp\left(-\frac{1}{2}u^2\right)du = \psi_o\left(\frac{x-\mu}{\sigma}\right)$$

The probability of observing t within a band width $2x$ around the expectation value zero is

$$P\left(|\mathrm{t}| \leq x\right) = \int_{-x}^{x} f(u)du = \int_{-x}^{0} f(u)du + \int_{0}^{x} f(u)du = \quad \left[f(u) = f(-u)\right]$$

$$= 2\int_{0}^{x} f(u)du = 2\int_{-\infty}^{x} f(u)du - 2\int_{-\infty}^{0} f(u)du = 2\psi_o\left(x\right) - 1$$

and

the probability of a random variable being observed within an integer multiple of the standard deviation from the mean

$$P\left(|\mathrm{x}-\mu| \leq n\sigma\right) = 2\psi_o\left(\frac{n\sigma}{\sigma}\right) - 1 = 2\psi_o\left(n\right) - 1$$

		from Chebychev inequality (Chap. 3)
$P\left(\|\mathrm{x}-\mu\| \leq \sigma\right) = 0.682$	$P\left(\|\mathrm{x}-\mu\| \geq \sigma\right) = 0.318$	$P\left(\|\mathrm{x}-\mu\| \geq \sigma\right) \leq 1.0$
$P\left(\|\mathrm{x}-\mu\| \leq 2\sigma\right) = 0.954$	$P\left(\|\mathrm{x}-\mu\| \geq 2\sigma\right) = 0.046$	$P\left(\|\mathrm{x}-\mu\| \geq 2\sigma\right) \leq 0.25$
$P\left(\|\mathrm{x}-\mu\| \leq 3\sigma\right) = 0.998$	$P\left(\|\mathrm{x}-\mu\| \geq 3\sigma\right) = 0.002$	$P\left(\|\mathrm{x}-\mu\| \geq 3\sigma\right) \leq 0.11$

“3σ-error“

Multivariate normal distribution

The joint normal distribution of *n* variables $(\mathrm{x}_1,\mathrm{x}_2,...,\mathrm{x}_n) = \mathbf{x}^{\mathrm{T}}$ is defined as

$$\phi(\mathbf{x}) = k\exp\left\{-\frac{1}{2}(\mathbf{x - a})^{\mathrm{T}}\mathbf{B}(\mathbf{x - a})\right\}$$

with **B** a symmetric, $n \times n$ Matrix. Since $\phi(\mathbf{x})$ symmetric about **a**,

$$\int_{-\infty}^{\infty}(\mathbf{x}-\mathbf{a})\phi(\mathbf{x})\mathrm{d}\mathbf{x} = \mathbf{0},\ \text{i.e.,}\quad \left[\ \text{i.e.,}\ \iint...\int\begin{pmatrix} x_1 - a_1 \\ ... \\ x_n - a_n\end{pmatrix}\phi(x_1, x_2, ..., x_n)\mathrm{d}x_1\mathrm{d}x_2...\mathrm{d}x_n = \begin{pmatrix} 0 \\ ... \\ 0\end{pmatrix}\right]$$

$E(\mathbf{x}) = \mathbf{a} = \boldsymbol{\mu}.$

Differentiating w.r.t. **a** (=0), we find for the *i*th component

$$\frac{\partial}{\partial a_i}\int_{-\infty}^{\infty}(\mathbf{x}-\mathbf{a})\phi(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{-\infty}^{\infty}\left(-\phi(\mathbf{x})\mathbf{e}_i + (\mathbf{x}-\mathbf{a})\phi(\mathbf{x})\left(-\frac{1}{2}\right)(-2)\sum B_{ik}(x-a)_k\right)\mathrm{d}\mathbf{x} = \overset{\text{vector}}{\mathbf{0}}\ ,$$

and for all components (*n* columns → $n \times n$ matrix)

$$\left(\frac{\partial}{\partial a_1},\frac{\partial}{\partial a_2},...,\frac{\partial}{\partial a_n}\right)\int_{-\infty}^{\infty}(\mathbf{x}-\mathbf{a})\phi(\mathbf{x})\mathrm{d}\mathbf{x} = -\int_{-\infty}^{\infty}\left(\mathbf{I} - (\mathbf{x - a})\left[\mathbf{B}(\mathbf{x - a})\right]^{\mathrm{T}}\right)\phi(\mathbf{x})\mathrm{d}\mathbf{x} = \overset{n\times n\ \text{matrix}}{\mathbf{0}}\ ,$$ which implies that

$E\left((\mathbf{x - a})(\mathbf{x - a})^{\mathrm{T}}\right)\mathbf{B} = \mathbf{I}$ (**B** symmetric) and thus

$$\mathbf{C} = E\left((\mathbf{x - a})(\mathbf{x - a})^{\mathrm{T}}\right) = \mathbf{B}^{-1}$$

The Matrix **B** in the exponent of $\phi(\mathbf{x})$ is just the inverse of the covariance matrix, and the vector **a** the vector formed by the expectation values.

Binormal (or bivariate normal) distribution

With $\mathbf{C} = \mathbf{B}^{-1} = \begin{pmatrix} \sigma_1{}^2 & \mathrm{cov}(\mathrm{x}_1,\mathrm{x}_2) \\ \mathrm{cov}(\mathrm{x}_1,\mathrm{x}_2) & \sigma_2{}^2 \end{pmatrix}$, we obtain

$$\mathbf{B} = \frac{1}{\sigma_1{}^2\sigma_2{}^2 - \mathrm{cov}^2(\mathrm{x}_1,\mathrm{x}_2)} \begin{pmatrix} \sigma_2{}^2 & -\mathrm{cov}(\mathrm{x}_1,\mathrm{x}_2) \\ -\mathrm{cov}(\mathrm{x}_1,\mathrm{x}_2) & \sigma_1{}^2 \end{pmatrix}$$

Case 1: independent variables →

$$\mathbf{B} = \begin{pmatrix} \frac{1}{\sigma_1{}^2} & 0 \\ 0 & \frac{1}{\sigma_2{}^2} \end{pmatrix} \to \phi(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1{}^2}\right) \exp\left(-\frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_2{}^2}\right), \text{ i.e.,}$$

ϕ becomes the product of two normal distributions (the leading factor from normalization)

(for n variables with vanishing or non-vanishing covariance, one obtains $k = \dfrac{1}{(2\pi)^{n/2}\sqrt{\det(\mathbf{C})}} = \dfrac{\sqrt{\det(\mathbf{B})}}{(2\pi)^{n/2}}$

Case 2: dependent variables

$$\phi(x_1,x_2) = \frac{1}{(2\pi\sigma_1\sigma_2)\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1{}^2} - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2{}^2}\right)\right]$$

Let's use reduced variables, $\mathrm{u}_\mathrm{i} = \dfrac{\mathrm{x}_\mathrm{i} - \mu_i}{\sigma_i}$, $i = 1,2$ and correlation coefficients $\rho = \dfrac{\mathrm{cov}(\mathrm{x}_1,\mathrm{x}_2)}{\sigma_1\sigma_2} = \mathrm{cov}(\mathrm{u}_1,\mathrm{u}_2) \to$

$$\phi(u_1,u_2) = \frac{\sqrt{\det \mathbf{B}}}{2\pi}\exp(-\frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}), \quad \text{with}$$

$$\mathbf{B} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

Lines of constant probability density result from constant exponent

$$\left(\frac{1}{2}\right)\frac{1}{1-\rho^2}(u_1{}^2 + u_2{}^2 - 2u_1u_2\rho) = \text{const}\left(\frac{1}{2}\right)$$

Let const=1, i.e., the prob. density has decreased by a factor of

$\exp(-1/2)=1/\sqrt{\mathrm{e}}$ from the maximum, $\phi(0,0)$.

(This corresponds to the 1-D case where at $u = \pm 1$ (i.e., $\pm (x - \mu) = \sigma$)

the prob. density has decreased by the same factor.)

In the original variables, we then have

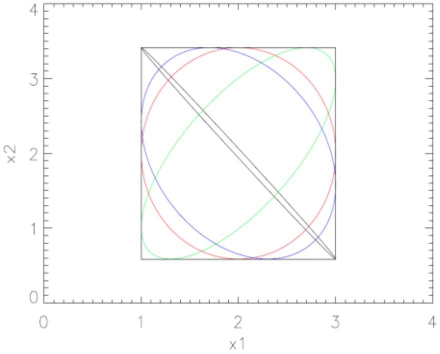
$$\frac{(x_1-\mu_1)^2}{\sigma_1{}^2} - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2{}^2} = 1-\rho^2,$$

which is the equation of an ellipse with center at (μ_1, μ_2) and is called the ellipse of covariance (Fehlerellipse).

The extreme values of x_1 and x_2 are located at

$\mu_1 \pm \sigma_1$ and $\mu_2 \pm \sigma_2$ (independent of ρ!),

ellipses of covariance for various binormal distributions, with different ρ (all other para-meters identical)



i.e., the ellipse fits exactly into the rectangular box between these limits.
The total probability of observing a pair of x_1 and x_2 inside the ellipse is $1-\exp(-1/2)$, see below.

By a simple rotation, the correlation can be put to zero (diagonalization by orthogonal transformation).
The corresponding transformation for $\mu_1 = \mu_2 = 0$ (in case, correct for non-vanishing expectation values) is

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \text{ with}$$

$$\tan 2\theta = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2},$$

and new semi-major and semi-minor axes (corresponding to the variances of the uncorrelated variables x_1' and x_2')

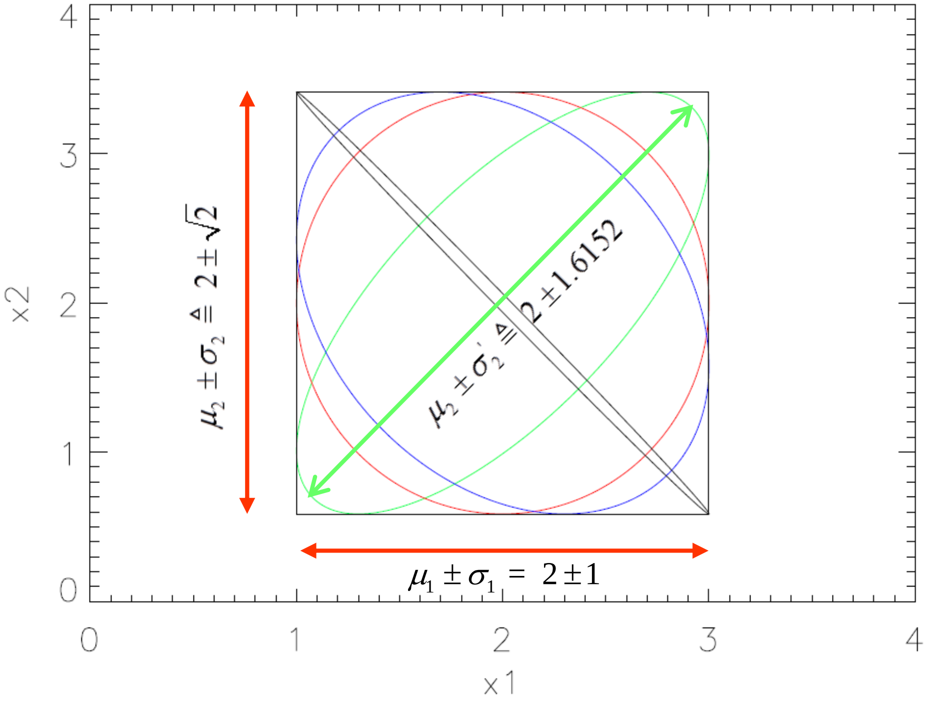
$$\sigma_1'^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2\sin^2\theta + \sigma_2^2\cos^2\theta - 2\rho\sigma_1\sigma_2\sin\theta\cos\theta}$$

$$\sigma_2'^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2\cos^2\theta + \sigma_2^2\sin^2\theta + 2\rho\sigma_1\sigma_2\sin\theta\cos\theta}$$

In the rotated coordinate system, the distribution has the simple form

$$\phi(x_1', x_2') = \frac{1}{2\pi\sigma_1'\sigma_2'} \exp\left\{ -\frac{1}{2}\left(\frac{x_1'^2}{\sigma_1'^2} + \frac{x_2'^2}{\sigma_2'^2} \right) \right\},$$

and is also centered at $\mu_1' = \mu_2' = 0$.



covariance ellipses centered at (2,2), with $\sigma_1 = 1, \ \sigma_2 = \sqrt{2}$, and

ρ=0.7 (green) $\rightarrow \theta = -31.60^\circ, \ \sigma_1' = 0.6252, \ \sigma_2' = 1.6152$

ρ=0.0 (red) $\rightarrow \theta = 0.00^\circ, \ \sigma_1' = 1.0000, \ \sigma_2' = \sqrt{2}$

$\rho = -0.3$ (blue) $\rightarrow \theta = 20.16^\circ, \ \sigma_1' = 0.9188, \ \sigma_2' = 1.4683$

$\rho = -0.999$ (black) $\rightarrow \theta = 35.26^\circ, \ \sigma_1' = 0.0365, \ \sigma_2' = 1.7316$

All ellipses fit into the rectangular box $\mu_1 \pm \sigma_1$ and $\mu_2 \pm \sigma_2$!

In the rotated coordinate system (by θ), x_1' and x_2' are uncorrelated!

The probability enclosed by the covariance-ellipse can be calculated as follows:
Consider the rotated coordinate system, and work in **reduced** variables. In this case, the p.d.f. reads

$$\phi(u_1^{'}, u_2^{'}) = \frac{1}{2\pi} \; \exp(-\frac{1}{2}(u_1^{'2} + u_2^{'\,2}),$$

and the total probability inside the covariance-ellipse (which in the transformed variables is the unit circle) can be calculated from

$$\iint_{\text{circle}} \phi(u_1^{'}, u_2^{'}) du_1^{'} du_2^{'} = \frac{1}{2\pi} \int_0^{2\pi} d\varphi \int_0^1 r \; \exp(-r^2/2) dr = -\exp(-r^2/2)\Big|_0^1 = 1 - \exp(-1/2) = 0.393$$

This is the probability that any (x_1, x_2) pair is located within the covariance-ellipse,
and applies for all binormal distributions, independent of their specific correlation term
(distribution in transformed coordinate system independent of correlation).

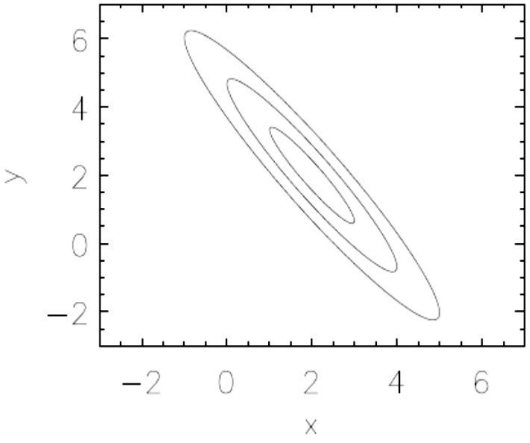
The area inside the covariance ellipse is called the "1-σ confidence region", since it comprises the region where the p.d.f. has decreased from the maximum by a factor of $\exp(-1/2)$, in analogy to the 1-D case (independent of correlation and the specific $\sigma_{1,2}$).

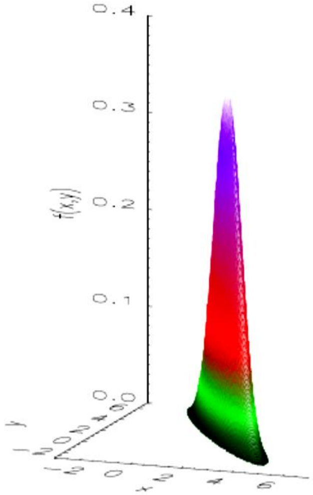
Similarly, one can calculate the 2-σ confidence region (where the probability density has decreased by a factor of $\exp(-(2\sigma)^2/2\sigma^2) = \exp(-4/2)$, with a total probability inside the corresponding ellipse of
$1 - \exp(-4/2) = 0.865$
(in the above integral, replace the upper limit by r=2), and so on for the n-σ interval.

Finally, one can generalize this consideration to arbitrary dimensions.

Generally, the 1-σ confidence interval denotes the region where the probability density has decreased by the factor of $\exp(-1/2)$

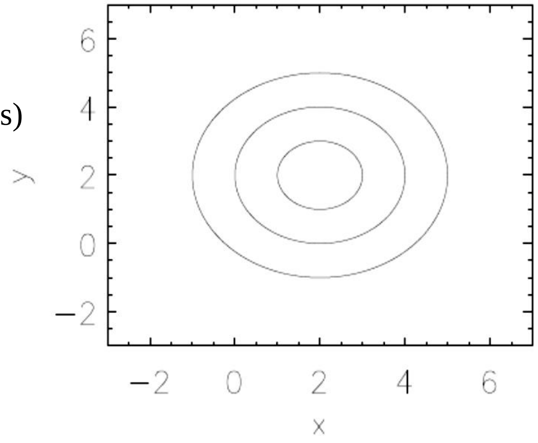
binormal distribution as before,
with $\rho = -0.9$, and contour plots
for the 1-,2- and 3-σ covariance ellipses

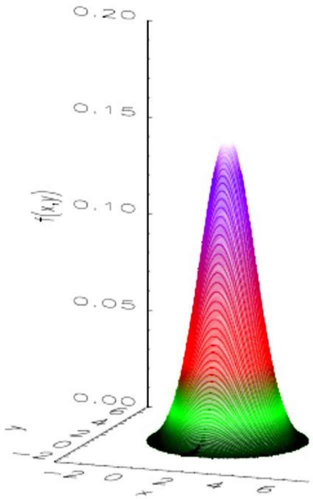


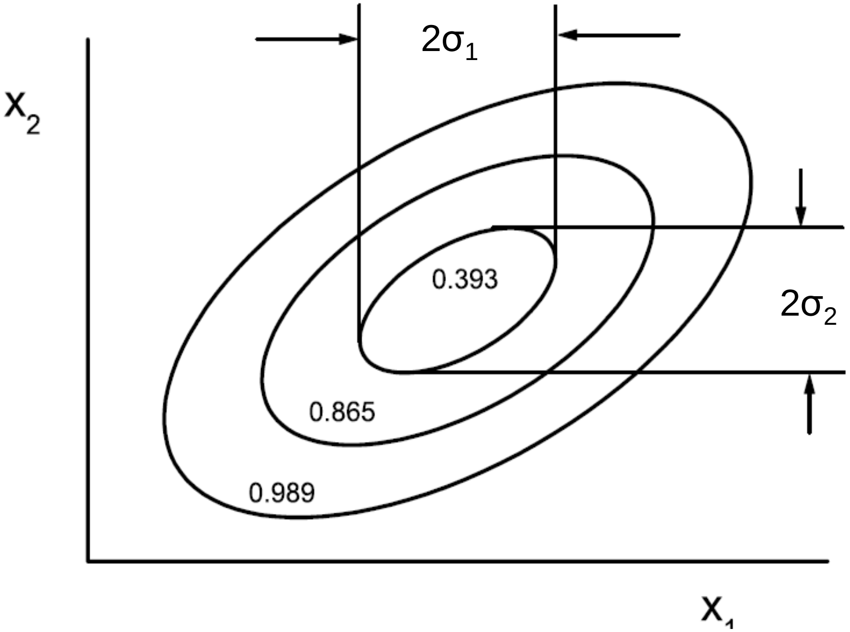


In the lower panel, the coordinate system
has been transformed (rotated, streched)
and displays the transformed binormal
distribution (with unit variances and ρ=0)
and corresponding covariance "ellipses"
for σ=1,2,3

Note that the volume (corresponding to
the total probability inside the contour levels)
remains preserved under the transformation
(e.g., for thin ellipses with large $|\rho|$
the probability densities are larger)







covariance ellipses for σ=1,2,3, corresponding probabilities and standard-deviations with respect to the two directions

deviation	Dimension 1	2	3	4
1σ	0.683	0.393	0.199	0.090
2σ	0.954	0.865	0.739	0.594
3σ	0.997	0.989	0.971	0.939
4σ	≈1.	≈1.	0.999	0.997

confidence-level [%]	Dimension 1	2	3	4
0.50	0.67	1.18	1.54	1.83
0.90	1.65	2.14	2.50	2.79
0.95	1.96	2.45	2.79	3.08
0.99	2.57	3.03	3.37	3.64

left: probability inside n-σ confidence region; right: interval limits in units of σ, for a given confidence level (probability)

χ^2-distribution

Remember example 2 from Chap. 3, "calculation of the transformed p.d.f.":

Calculate the distribution for the square of a reduced r.v.
which itself should be normally distributed.

$$\mathrm{u} = \left[\frac{(\mathrm{x} - \mu)}{\sigma}\right]^2 \quad \text{and} \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\Rightarrow g(u) = \frac{1}{\sqrt{2\pi u}} e^{-u/2},$$

which is the so-called χ^2-distribution for one degree of freedom.

For convenience, we denote χ^2 by u in the following.

$$E(\mathrm{u}) = 1, \quad Var(\mathrm{u}) = 2$$

Now, let's add f squares of independent, normally distributed and reduced random variables

$$\chi^2 = \mathrm{u} = \sum_{\mathrm{i=1}}^{f} \frac{(\mathrm{x_i} - \mu_i)^2}{\sigma_i^2}$$

This results in the χ^2-distribution for f degrees of freedom, and plays an important role in the comparison of measurements and theoretical predictions (e.g., linear regressions). In this case (e.g., via the charact. function)

$$g(u) = \frac{1}{\Gamma(f/2)2^{f/2}} u^{f/2-1} e^{-u/2},$$ with Gamma-function Γ and

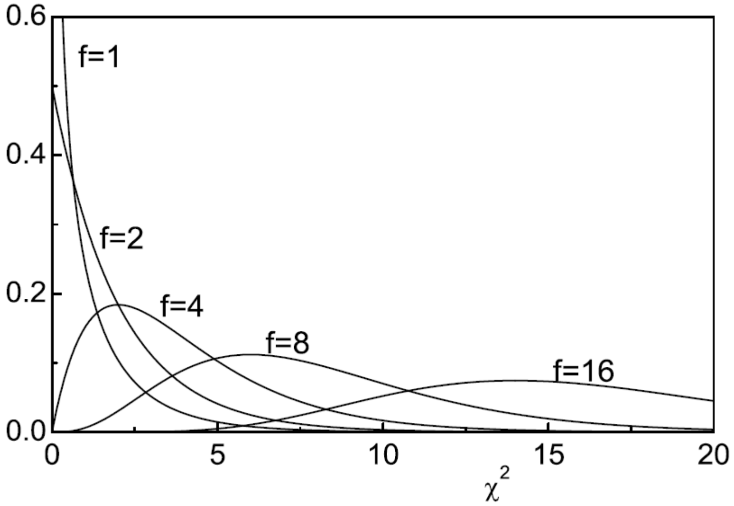
$E(\mathrm{u}) = f, \quad Var(\mathrm{u}) = 2f$ (from the definition and using the calculation rules for expectation value and variance)

Maximum (mode) of χ^2-distribution for $f > 2$ at $u_{\max} = f - 2$.

For $f = 2$, we obtain an exponential distribution

For large f, χ^2-distribution approaches normal distribution.

The role of the degrees of freedom will be discussed in Chap. 8



The central limit theorem (CLT)

Remember: Normal distribution was derived as the asymptotic distribution for

$$\mathrm{x} = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{x}_\mathrm{i}$$

when x_i describes the outcome of an experiment with two possible results, $\mathrm{x}_\mathrm{i} = \{0,1\}$.

Let's now investigate more general sums of this type.

"Classical" theorem:

We assume that the x_i are independent r.v. and originate from the same, arbitrary distribution with well-defined mean μ and variance σ^2. The characteristic function of this distribution (for $\mathrm{x}_\mathrm{i}^{'} = \mathrm{x}_\mathrm{i} - \mu$) is

$$\phi_{x_i^{'}}(t) = E(e^{it(\mathrm{x}_\mathrm{i}-\mu)}), \text{ with } \left.\frac{d\phi(t)}{dt}\right|_{t=0} = 0 \text{ and } \left.\frac{d^2\phi(t)}{dt^2}\right|_{t=0} = -\sigma^2$$

Thus, the Taylor expansion is given by $[\phi(0) = 1]$

$$\phi_{x_i^{'}}(t) = 1 - \frac{1}{2}\sigma^2 t^2 + O(t^3)$$

We introduce now a new variable

$\mathrm{u_i} = \frac{\mathrm{x_i'}}{\sigma\sqrt{n}} = \frac{\mathrm{x_i} - \mu}{\sigma\sqrt{n}}$, which simply contracts the scale. The corresponding charact. function is

$\phi_{u_i}(t) = E(e^{it\mathrm{u_i}}) = E(\exp(it\frac{\mathrm{x_i} - \mu}{\sigma\sqrt{n}})) = \phi_{x_i'}(\frac{t}{\sigma\sqrt{n}}),$ and therefore

$\phi_{u_i}(t) = 1 - \frac{t^2}{2n} + ...$ with higher terms at most of order $O(n^{-3/2})$

Making use of the fact that the characteristic function of the sum of n independent r.v. is given by the product of the individual charact. functions, and going to the limit $n \to \infty$, we find for

$\mathrm{u} = \lim_{n\to\infty}\sum_{i=1}^{n}\mathrm{u_i} = \lim_{n\to\infty}\sum_{i=1}^{n}\frac{\mathrm{x_i} - \mu}{\sigma\sqrt{n}}$ that

$$\phi_u(t) = \lim_{n\to\infty}\left(\phi_{u_i}(t)\right)^n = \lim_{n\to\infty}\left(1 - \frac{t^2}{2n} + ...\right)^n$$

$$\phi_u(t) = \exp(-\frac{1}{2}t^2),$$

which is just the charact. function of the standardized normal distribution, with expectation value 0 and variance 1.

In terms of the arithmetic mean of the original variables x_i then,

$$\overline{\mathrm{x}}=\frac{\mathrm{x}}{n}=\lim_{n\to\infty}\frac{1}{n}\left[\sum_{i=1}^{n}(\sigma\sqrt{n}\mathrm{u}_i+\mu)\right]=\lim_{n\to\infty}\frac{\sigma}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{u}_i+\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\mu}{n}=\frac{\sigma}{\sqrt{n}}\mathrm{u}+\mu,$$

$$E(\overline{\mathrm{x}})=\mu,\quad Var(\overline{\mathrm{x}})=\sigma^2/n$$

the back-transformed distribution is normal, with mean μ and standard deviation $\sigma/\sqrt{n}$.

For a corresponding (easier) proof using the properties of *cumulants*, see exercises.

Thus, the "classical" **central limit theorem** reads:

If the x_i are a set of independent r.v. each distributed with (existing) mean μ and variance σ^2, then in the limit of $n\to\infty$ their arithmetic mean

$$\overline{\mathrm{x}}=\frac{\mathrm{x}}{n}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_\mathrm{i}$$

is normally distributed with mean μ and variance σ^2/n.

theorem not valid for “pathologic“ distributions with undefined/non-existing mean or variance, e.g., the Cauchy (Lorentz) distribution

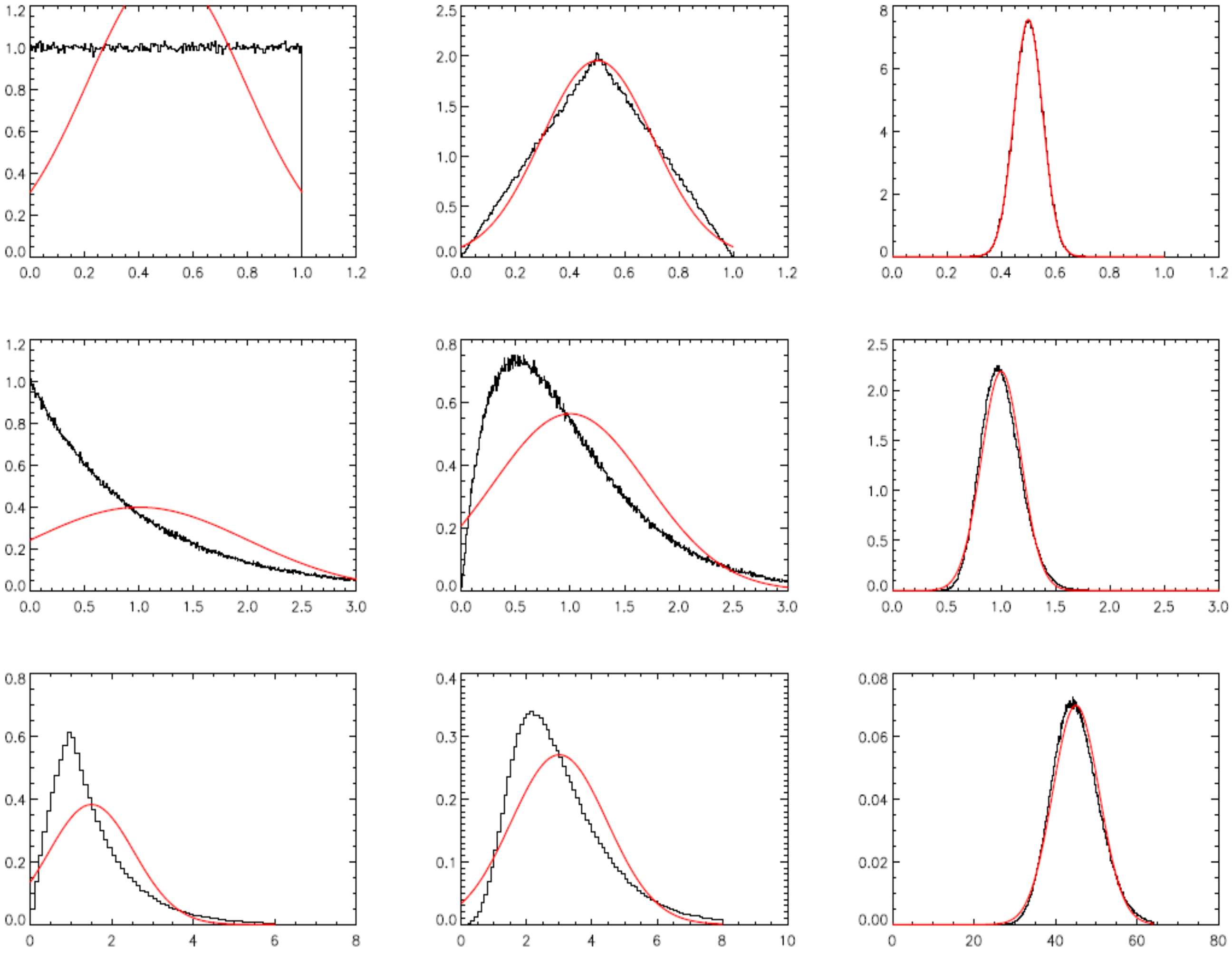
Under certain assumptions [see, e.g., Wikipedia: the Lyapunov criterium ("weak" asymmetry) or the even weaker Lindeberg condition], a "generalized" central limit theorem can be formulated. If these conditions apply, the sum of arbitrary (i.e., not identical) distributed r.v converges to a normal distribution, with mean $\sum_{i=1}^{n}\mu_i$ and variance $\sum_{i=1}^{n}\sigma_i^2$.

Examples for the CLT

CLT for several cases:
upper panel: arithmetic mean of n=1, 2, 30 uniformly distributed r.v. Overplotted is the corresponding Gaussian with μ= 0.5 and variance =1/(12*n)

middle panel: arithmetic mean of n=1, 2, 30 exponentially (λ=1) distributed r.v. Overplotted is the corresponding Gaussian with μ= 1 and variance =1/n

lower panel: sum of n=1, 2, 30 exponentially (λ=1) plus n=1,2, 30 uniformly distributed r.v. overplotted is the corresponding Gaussian with μ= n*1+n*0.5 and variance =n*1+n/12.



sample size =1e6, bin size=0.005

- The CLT in its generalized form is the base of assuming experimental errors as being normally distributed:

- each measurement error is assumed to consist of an accumulation of small individual errors (with unknown distribution), whereas their sum (the measured error) can be described by a Gaussian.

log-normal distribution

- single-tailed probability distribution of a random variable whose logarithm is normally distributed. If y is a random variable with a normal distribution, then x = exp(y) has a log-normal distribution
- likewise, if x is log-normally distributed, then log(x) is normally distributed. (The base of the logarithmic function does not matter)
- a variable might be modeled as log-normal if it can be thought of as the product of many independent factors which are positive and close to 1. (see figure next page)
- log (x) = log of product = sum of log's -> CLT -> log (x) normally distributed
- plays an important role in, e.g., economy, biology, mechanics and astrophysics

$$f(x,\mu,\sigma)dx = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)dx \qquad \text{or}$$

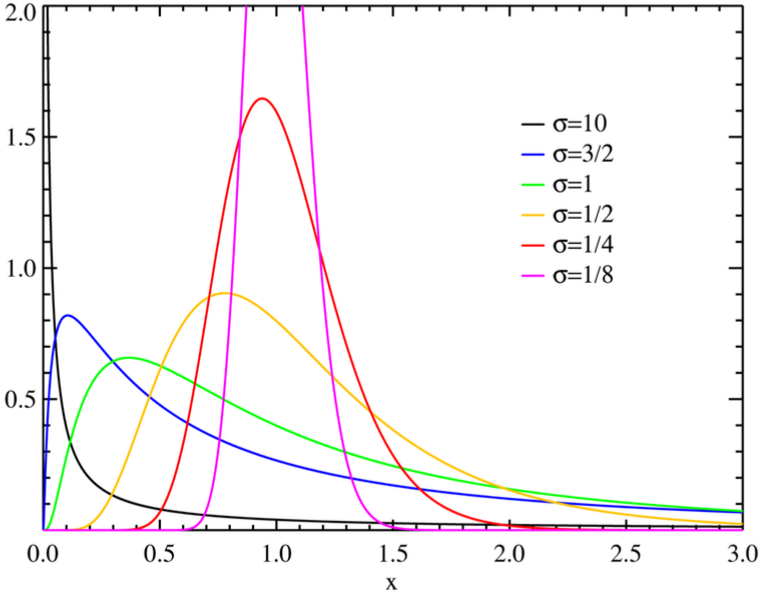
$$f(\ln x,\mu,\sigma)d\ln x = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)d\ln x \quad \rightarrow \quad \ln x \text{ is normally distributed}$$

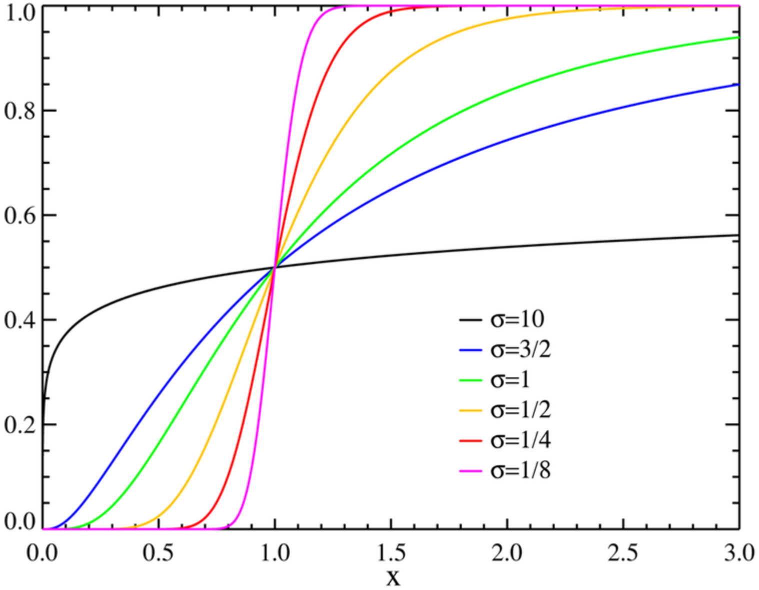
(of course, w.r.t. $d\ln x$)

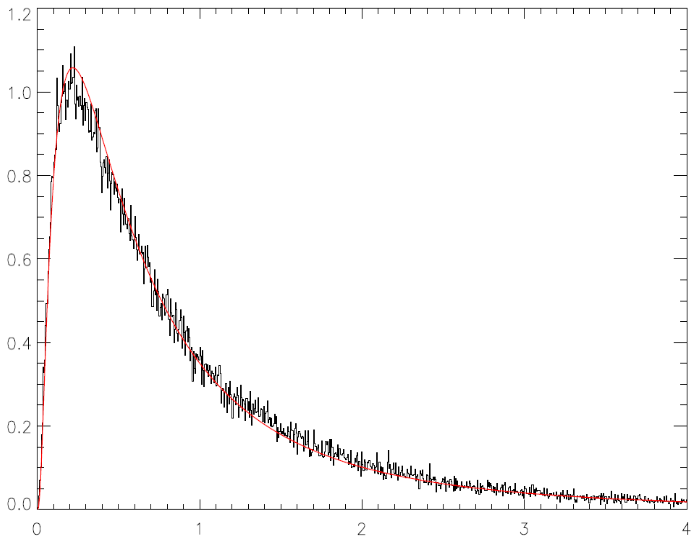
$$E(\mathrm{x}) = e^{\mu+\sigma^2/2}$$

$$Var(\mathrm{x}) = (e^{\sigma^2}-1)e^{2\mu+\sigma^2}$$

pdf (left) and cumulative distribution function (right) for a log-normal distribution with μ=0 and different σ as a function of x (linear scale)







Left: simulation of a log-normal distribution from a sample of 10^5 r.v. which are distributed according to

$$x=\prod_{i=1}^{7} x_i \quad \text{with independent } x_i,$$

where the x_i are uniformly distributed within the interval $[0.4,1.6]$.

The estimators (Chap. 7) for μ and σ are

$\hat{\mu}$=-0.47 and $\hat{\sigma}$=1.01

Overplotted is a theoretical log-normal distribution with these parameters