operations on random numbers

statistics and data analysis - (chapter 2)

Björn Malte Schäfer

Graduate School for Fundamental Physics
Fakultät für Physik und Astronomie, Universität Heidelberg

April 24, 2016

contents

repetition

- operations on sets
- probability
- expectation value, variance, covariance
- estimates and the law of large numbers
- Chebyshev-inequality

Cauchy-Schwarz inequality (geometric proof)

- imagine two vectors $\vec{x}$ and $\vec{y}$: can you find a solution for $\vec{x} + \lambda\vec{y} = 0$?
  - yes, if they're parallel
  - no, if they're not parallel
- the norm of a vector is positive definite: if the norm is zero, the vectors are zero

$$\left|\vec{x} + \lambda\vec{y}\right| = 0 \quad \leftrightarrow \quad \vec{x} + \lambda\vec{y} = 0 \tag{1}$$

- compute $\left|\vec{x} - \lambda\vec{y}\right|^2 = \vec{x}^2 + 2\lambda\vec{x}\vec{y} + \lambda^2\vec{y}^2 = 0$
- find the solutions to the resulting quadratic equation

$$\lambda_\pm = \frac{-2\vec{x}\vec{y} \pm \sqrt{4(\vec{x}\vec{y})^2 - 4\vec{x}^2\vec{y}^2}}{2\vec{y}^2} \tag{2}$$

where there can be only 1 or zero solutions, corresponding to the cases $\vec{x} \parallel \vec{y}$ and $\vec{x} \nparallel \vec{y}$

Cauchy-Schwarz inequality (continuous distributions)

- the number of solutions for λ is determined by the square root:
  - $(\vec{x}\vec{y})^2 = \vec{x}^2\vec{y}^2$: vectors parallel, one solution
  - $(\vec{x}\vec{y})^2 \leq \vec{x}^2\vec{y}^2$: vectors not parallel, no solution
- combine both cases: Cauchy-Schwarz inequality

$$\left|\vec{x}\vec{y}\right| \leq \sqrt{\vec{x}^2\vec{y}^2} = \left|\vec{x}\right|\,\left|\vec{y}\right| \tag{3}$$

  using the fact that the root is monotonic
- discrete distributions:

$$\sigma_x^2 = \vec{x}^2 = \sum_i x_i^2 p(x_i) \quad \text{and} \quad \mathrm{cov}_{x,y} = \vec{x}\vec{y} = \sum_i \sum_j x_i y_j p(x_i, y_j) \tag{4}$$

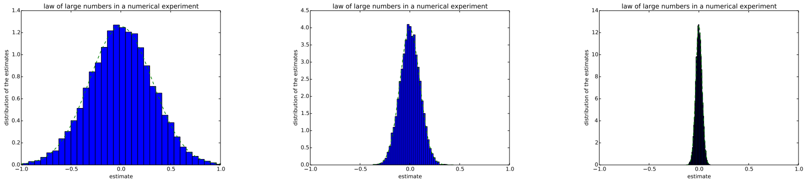
- continuous distributions:

$$\sigma_x^2 = \vec{x}^2 = \int \mathrm{d}x\, x^2 p(x) \quad \text{and} \quad \mathrm{cov}_{x,y} = \vec{x}\vec{y} = \int \mathrm{d}x \int \mathrm{d}y\, xy\, p(x, y) \tag{5}$$

law of large numbers

- we saw that the mean as as estimate of the expectation value can be expected to be close to the expectation value in the limit of large sample sizes n
- typical behaviour:

$$\bar{x} = \frac{1}{n}\sum_i x_i \quad \rightarrow \quad \sigma_{\bar{x}} \propto \frac{1}{\sqrt{n}} \tag{6}$$



$n = 10$ $n = 100$ $n = 1000$

- still probabilistic, very large $\bar{x} \gg 0$ can (and will) occur!

law of large numbers in dice rolls

- let's build some intuition about the law of large numbers
- imagine rolling 2 dice simultaneously and look at the sum of points: in how many ways can you achieve a specific number of points:

sum of points	2	3	4	5	6	7	8	9	10	11	12
combinations	1	2	3	4	5	6	5	4	3	2	1

- if you increase the number of dice, the distribution will fall off more rapidly: for 3 dice you've got one possibility of getting 3 points, but already 3 ways of getting 4 points.
- in comparison to the many realisations to roll a number of points close to the expectation value there are fewer possibilities to roll an extremely large or small number of points
- again, this is not excluded, but only gets unlikely with increasing n

distributions of random variables

- a random variable assigns a **value** to the random event which occurs at a certain probability
- it makes sense to quote directly the probability for a random variable
- the set of discrete values $p(x_i)$ or of the continuous values $p(x)$ is called **distribution**
- we will assume that distributions are smooth functions
- the random variable value x with the largest probability is the most probable value

question

find the most probable value x of the distributions $p(x) \propto x^n \exp(-x)$ and $p(x) \propto x^n/(\exp(x) \pm 1)$ with integer n and positive x

discrete and continuous random processes

quote probability for:

- discrete random process: only a finite (countable) number of possible values for the random variable

$$p(x_i) \tag{7}$$

- continuous random process: random variabe lies within an interval (only integral statements would make sense)

$$p(x_a \leq x \leq x_b) = \int_{x_a}^{x_b} \mathrm{d}x\, p(x) \tag{8}$$

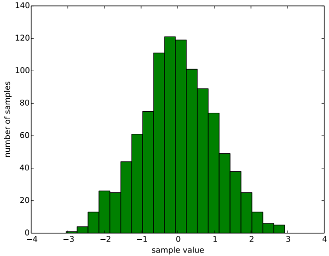
  with the **probability density** $p(x)$

- be careful with distributions: often you'll see $p(x)\mathrm{d}x$ as the probability of x to be within an infinitesimal interval $\mathrm{d}x$ around x

question

what's the unit of the probability density $p(x)$ if the random variable x is not dimensionless?

histograms



- construct an idea of the distribution from a list of data points
- Laplacian approximation of the probability p_i or $p(x_a \leq x \leq x_b)$ by number n_i of counts in the bin $x_a \leq x \leq x_b$ in $n = \sum_i n_i$ trials
- one gets an **estimate** of the expectation value, not the value itself
- for understanding this, we need **Poisson**-statistics (see lecture 4!)

interpretation of histograms for continuous distributions

- be careful in interpreting histograms: the numbers of the y-axis make only sense for a given number of repetitions n and for a bin size Δx
- even though the ratio n_i/n is smaller than one, it is **not** the probability $p(x_a \leq x \leq x_b)$
- how many events in a bin $x_a \leq x \leq x_b$ do you expect?

$$n_i = n \int_{x_a}^{x_b} \mathrm{d}x\, p(x) \simeq n\,(x_b - x_a)p(x) \quad \rightarrow \quad p(x) \simeq \frac{1}{x_b - x_a}\frac{n_i}{n} \tag{9}$$

if the binning is chosen finely enough such that variations of $p(x)$ don't matter and the approximation holds

rejection sampling

- distribution can be used for designing a random process that yields x-values at the probability $p(x)$
- computers provide **uniformly distributed** random numbers
- rejection sampling:
  1. draw a proposal value x
  2. decide by a random experiment if you want to keep it:
     - a value x should occur with a probability $p(x)$
     - draw a second value a from an interval $0 \leq a \leq a_{\max}$ and keep it if $a \leq p(x)$, reject it if $p(x) < a < a_{\max}$, $a_{\max} = \max\left[p(x)\right]$

question

can you optimise rejection sampling such that the number of valid samples is large?

normalisation

- distributions are normalised as a consequence of the Kolmogorov axioms:
  - discrete distribution:
$$\sum_i p(x_i) = 1 \tag{10}$$
  - continuous distribution:
$$\int \mathrm{d}x\, p(x) = 1 \tag{11}$$
- but sometimes, distributions are normalised to a certain physical value: for example, the Planck-spectrum $S(\nu)\mathrm{d}\nu$ is normalised to yield the total power emitted by a black body

question

normalise the distributions $p(x) \propto x^n \exp(-x)$ and $p(x) \propto x^n/(\exp(x) - 1)$ with integer n and positive x

transformation of random variables

- suppose you know the distribution $p(x)\mathrm{d}x$ of a random distribution
- can you write down the distribution of a function $y(x)$?
- look at probability of each interval, which should be conserved by the mapping:
$$\int \mathrm{d}y\, p(y) = \int \mathrm{d}x\, p(y(x))\frac{\mathrm{d}y}{\mathrm{d}x} \tag{12}$$
using **integration by parts**
- consequently: $p(x)\mathrm{d}x = p(y)\mathrm{d}y$ if the above holds for any interval

question

what properties does the remapping $x \to y$ need to have?

cumulative distribution

- from every probability density $p(x)\mathrm{d}x$ one can construct the **cumulative distribution** $P(x)$:

$$P(x) = \int_{-\infty}^{x} \mathrm{d}x' p(x') \tag{13}$$

- interpretation: probability of the random variable to be smaller than x

question

why is $P(x)$ always monotonically increasing?

complementary cumulative distribution

- the **complementary cumulative distribution** $Q(x)$ is defined as the opposite,

$$Q(x) = \int_x^{+\infty} \mathrm{d}x'\, p(x') \tag{14}$$

which gives the probability of the random variable to be at least as large as x

- obviously,

$$P(x) + Q(x) = 1 \tag{15}$$

if correctly normalised

question

design an algorithm for computing the cumulative distribution for a list of random numbers without histogramming them first!

quartiles and percentiles

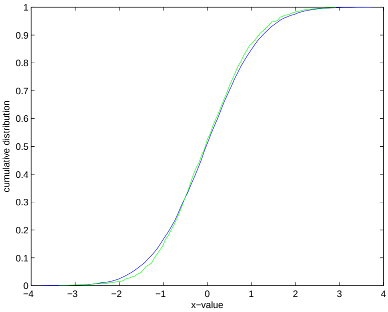
- instead of $p(x)\mathrm{d}x$ or $P(x)$ one often quotes percentiles:

$$P(x_a) = a\% \tag{16}$$

  x_a is called the ath percentile

- it is customary to give quartiles, where $a = 0.25, 0.5, 0.75$
- or $n\sigma$-intervals, in particular for symmetric distributions, containing 0.68, 0.95 or 0.99 of the total normalisation

cumulative distribution

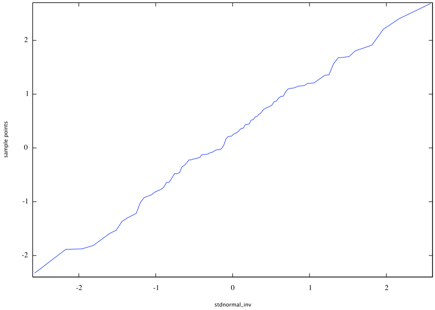


cumulative, normalised distribution of 10^3, 10^4 draws from a Gaussian

question

how would you generate a cumulative distribution from data?

qq-plots and percentiles



qqplot for 10^3 draws from a Gaussian

- with weak statistics, it is difficult to see the distribution due to the large Poisson noise in each bin entry $\rightarrow$ cumulative distribution works better
- some tools can plot the cumulative distribution with the y-axis rescaled such that the Gaussian distribution gives a straight line

Gaussian probability density

- Gaussian probability density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{17}$$

  with variance σ^2

- many processes in Nature follow a Gaussian distribution
- reason: central limit theorem

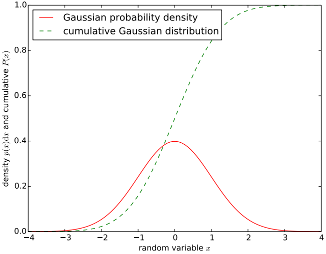
question

verify the normalisation $\sqrt{2\pi\sigma^2}$ of the Gaussian probability density

question

show that the width of the Gaussian at half height is $2\sqrt{2\ln(2)}\sigma$

Gaussian probability density



Gaussian density $p(x)\mathrm{d}x$ and cumulative $P(x)$, variance $\sigma^2 = 1$

error function and Φ-function

- cumulative distribution $P(x) = \Phi(x)$ of a **unit Gaussian** with $\sigma^2 = 1$:

$$\Phi(x) = \int_{-\infty}^{x} \mathrm{d}x'\, p(x') = \frac{1}{2}\left(1 + \mathrm{erf}(x/\sqrt{2})\right) \tag{18}$$

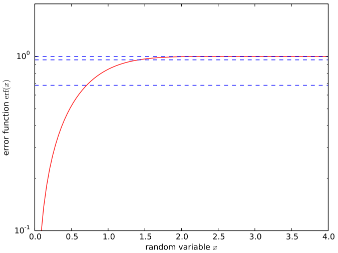
- the **error function** $\mathrm{erf}(x)$ is defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{d}t\, \exp(-t^2) \tag{19}$$

- the error function is a convenient way to express $n\sigma$-intervals:

$$\int_{-n\sigma}^{+n\sigma} \mathrm{d}x'\, p(x') = \mathrm{erf}(n/\sqrt{2}) \tag{20}$$

error function



Gaussian density $p(x)\mathrm{d}x$ and cumulative $P(x)$, variance $\sigma^2 = 1$

- $\mathrm{erf}(n/\sqrt{2})$ are integrals of the Gaussian from $-n\sigma$ to $+n\sigma$

inversion sampling

- idea: map samples y from a known distribution to samples x if the relationship $x(y)$ is known
- generate a random number y from the uniform unit interval
- map y onto $x = P^{-1}(y)$
- x is distributed according to $p(x)\mathrm{d}x$:

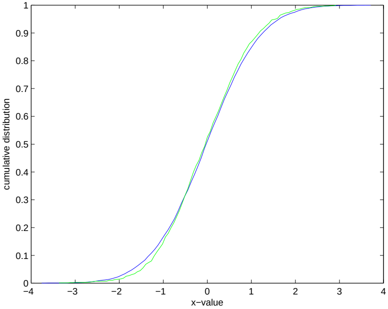
$$x = P^{-1}(y) \rightarrow y = P(x) \rightarrow \frac{\mathrm{d}y}{\mathrm{d}x} = p(x) \rightarrow p(x)\mathrm{d}x = 1 \times \mathrm{d}y \tag{21}$$

  meaning that $x(y)$ is $p(x)\mathrm{d}x$-distributed if y is uniformly distributed
- using that $\mathrm{d}P/\mathrm{d}x = p(x)$ due to the fundamental theorem of calculus
  - advantage: every sample is valid
  - disadvantage: inversion might be difficult

question

why does the inversion $x = P^{-1}(y)$ always have a solution for x?

inversion sampling: intuition



cumulative distribution of a Gaussian probability density

- every interval $\mathrm{d}y$ gets squeezed or stretched into an interval $\mathrm{d}x$
- amount of squeezing or stretching is proportional to $p(x)$

independent random processes

- drawing two random number independently means that the probabilities of the two draws can be multiplied

$$p(x,y) = p(x) \times p(y) \tag{22}$$

  for the **joint distribution** $p(x,y)$ from the individual distributions for x and y
- this is called a **Markovian** process of length zero
- we will encounter correlated random variables and Markovian processes with a long memory

conditional random processes

- often, the outcome of a random experiment depends on a previous outcome: then, the two probabilties can **not** be multiplied
- in this case, the correlation coefficient is **not zero** and
- the distribution does **not** factorise
- instead:

$$p(x, y) = p(x|y) \tag{23}$$

  which defines a Markovian process of length one
- what about Bayes' law? it does not matter if $p(x, y) = p(y, x)$

sum distribution

- let's combine two independent, indentically distributed random numbers x and y into a sum
- a certain fixed value $s = x + y$ for the product can result from the entire range of x and y
- accumulate the total probability p_s for getting s: if x is the first number, the second number needs to be $y = s - x$ so that the sum is s:

$$p_s(s) = \int \mathrm{d}x \int \mathrm{d}y\, p(x)p(y)\, \delta_D(s - x + y) = \int \mathrm{d}x\, p(x)p(s - x) \quad (24)$$

- the sum distribution is the convolution of the two individual distributions

question

going back to the law of large numbers, do you see the convolution there?

product distribution

- let's combine two independent, identically distributed random numbers x and y to a product
- a certain fixed value $q = x \times y$ for the product can result from the entire range of x and y
- accumulate the total probability p_q for getting q: if x is the first number, the second number needs to be $y = q/x$ so that the product is q:

$$p_q(q) = \int \mathrm{d}x \int \mathrm{d}y\, p(x)p(y)\,\delta_D(q - xy) = \int \frac{\mathrm{d}x}{|x|}\, p(x)p(q/x) \tag{25}$$

with the Dirac-δ distribution

question

show that $\int \mathrm{d}x\, \delta_D(\alpha x) = 1/\alpha$, then $\int \mathrm{d}x\, p(x)\delta_D(\alpha x) = p(0)/\alpha$ and then the above relation

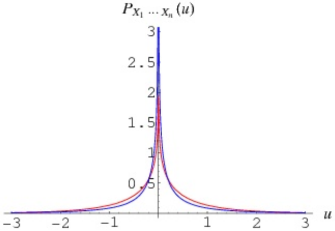
Gaussian product distribution

- combine two independent, identically Gaussian-distributed random numbers x, y to a product $q = xy$

$$p_q(q) = \frac{1}{2\pi\sigma^2}\int \mathrm{d}x \int \mathrm{d}y \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)\delta_D(xy-q) = \frac{K_0(q/\sigma^2)}{\pi\sigma^2} \quad (26)$$

with a Bessel function of the second kind

Gaussian product distribution



Gaussian product distribution (source: mathworld)

ratio distribution

- let's combine two independent, identically distributed random numbers x and y to a ratio $r = x/y$
- in analogy to the product distribution, x as the first number needs to x/r for the second number such that the ratio is r:

$$p_r(r) = \int \mathrm{d}x \int \mathrm{d}y\, p(x)p(y)\, \delta_D(r - x/y) = \int \mathrm{d}x\, |x|\; p(x)p(rx) \qquad (27)$$

Gaussian ratio distribution

- combine two independent, identically Gaussian distributed random numbers x, y to a ratio $r = x/y$

$$p_r(r) = \frac{1}{2\pi\sigma^2}\int \mathrm{d}y\,|y|\,\exp\left(-\frac{y^2}{2\sigma^2}\left[1+r^2\right]\right) \tag{28}$$

- with $\int y\mathrm{d}y\;\exp(-\alpha y^2) = 1/(2\alpha)$ this becomes:

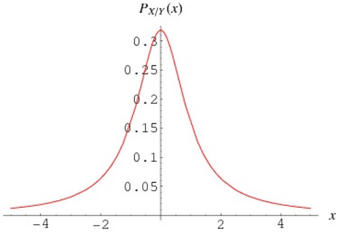
$$p_r(r) = \frac{1}{\pi}\frac{1}{1+r^2} \tag{29}$$

- this distribution is called the **Cauchy-distribution** and is mean!

the Cauchy-distribution

does not have a finite variance, and therefore, you can not apply the Chebyshev-inequality or the law of large numbers

Gaussian ratio distribution



Gaussian ratio distribution (source: mathworld)

summary

- distributions and probability densities
- cumulative distributions
- transformations between random variables
- sampling of random numbers from a distribution
- Gaussian probability density and error function
- sum, product and ratio distribution