descriptive statistics

statistics and data analysis (chapter 4)

Björn Malte Schäfer

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

24.Oct.2011

outline: lecture 3 - descriptive statistics

1. characteristic function
2. Gaussians
3. histograms
4. Edgeworth
5. central limit theorem
6. regression
7. summary

repetition

- random distributions
- Bernoulli-, Poisson- and Gauss-distribution
- relations between these distributions
- characterisation of a distribution
- multivariate Gaussians, covariance and correlation coefficient
- conditions on Gaussians: Schur-complement

numerical exercise

generate the sum of n arbitrarily distributed random numbers. show that the higher-order cumulants κ_k tend to zero $\propto n^{(2-k)/2}$

characteristic function $\phi(t)$

- characteristic function $\phi(t)$: **Fourier-transform** of $p(x)\mathrm{d}x$:

$$\phi(t) = \int \mathrm{d}x\, p(x) \exp(-\mathrm{i}tx) \leftrightarrow p(x) = \int \frac{\mathrm{d}t}{2\pi}\, \phi(t) \exp(+\mathrm{i}tx)$$

- relation to moments: Taylor-expand the exponential:

$$\phi(t) = \int \mathrm{d}x\, p(x) \sum_n \frac{(-\mathrm{i}tx)^n}{n!} = \sum_n \langle x^n \rangle \frac{(-\mathrm{i}t)^n}{n!}, \quad \langle x^n \rangle = \int \mathrm{d}x\, x^n p(x)$$

- in analogy: moment generating function $m(t)$

$$m(t) = \langle \exp(-tx) \rangle = \int \mathrm{d}x\, p(x) \exp(-tx)$$

it's a matter of taste to use either the Fourier- or Laplace-transform, with either sign

question

symmetric distribution have vanishing odd-numbered moments

cumulants and the cumulant generating function

- cumulants: expand the **logarithm** of the moment-generating function:

$$K(t) = \ln m(t) = \sum_n \kappa_n \frac{t^n}{n!} \quad \rightarrow \quad \kappa_n = \frac{\partial^n}{\partial t^n} K(t)|_t = 0 \tag{1}$$

  $K(t)$ is called the cumulant generating function
- naturally, the moment generating function is given by

$$m(t) = \exp(K(t)) \tag{2}$$

- cumulant-generating function of a Gaussian is a second-order polynomial
- there are only two nonzero cumulants in a Gaussian: mean and variance
- with cumulants you can quantify how close a distribution is to a Gaussian

Gaussian distribution - why is it so special?

- all moments exist and are finite
- $(2n)$th moment is $\propto$ variance^n: $\langle x^{2n}\rangle = (2n-1)!! \times \langle x^2\rangle^n$
- $\phi(t)$ and $m(t)$ are Gaussians again

question

show directly by induction (and partial integration) that $\langle x^{2n}\rangle \propto \langle x^2\rangle^n$

question

compute $\langle x^{2n}\rangle$ from $m(t)$ for a Gaussian pdf!

question

show that $\langle x^{2n}\rangle = (2n-1)!! \times \langle x^2\rangle^n$ for a Gaussian pdf!

sum of Gaussians - the ideal central limit theorem

- sum of Gaussian distributed **uncorrelated** random numbers is exactly Gaussian distributed → ideal case of the **central limit theorem**
- look at the characteristic function $\phi_x(t)$ and $\phi_y(t)$ of two Gaussian distributed random numbers x and y

$$\phi_{x+y}(t) = \langle\exp(\mathrm{i}t(x+y))\rangle = \langle\exp(\mathrm{i}tx)\exp(\mathrm{i}ty)\rangle$$

- use independency

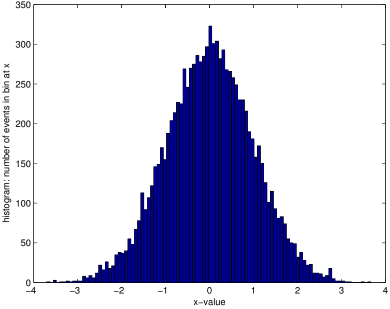
$$\ldots = \langle\exp(\mathrm{i}tx)\rangle\langle\exp(\mathrm{i}ty)\rangle = \phi_x(t)\phi_y(t)$$

- characteristic function of a Gaussian is a Gaussian again:

$$\ldots\exp\left(-\frac{\sigma_x^2 t^2}{2}\right)\exp\left(-\frac{\sigma_y^2 t^2}{2}\right) = \exp\left(-\frac{(\sigma_x^2+\sigma_y^2)t^2}{2}\right)$$

- sum is Gaussian distributed, with new variance $\sigma^2 = \sigma_x^2 + \sigma_y^2$

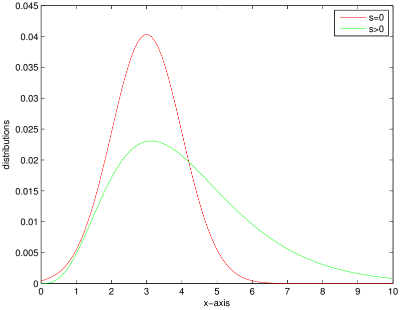
histograms



histogram of 10^4 draws from a Gaussian distribution

- histogram: count number of **events** falling inside a given **bin** $\rightarrow$ discrete approximation to the probability density
- typical error in each bin: Poisson statistics, $\sqrt{n_i}$ for n_i events
- rule of thumb: $\sqrt{n}$ bins for n events

skewness



Gaussian ($s = 0$) and Planck ($s > 0$)-distribution

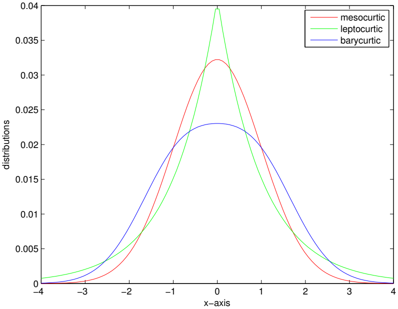
- skewness $s = \langle x^3\rangle / \langle x^2\rangle^{3/2}$: **asymmetry** of a distribution $p(x)\mathrm{d}x$

  $s > 0$ skewed to right

  $s = 0$ symmetric distribution

  $s < 0$ skewed to left

kurtosis



Gaussian distribution, and distributions with kurtosis $\neq 3$

- kurtosis $k = \langle x^4 \rangle / \langle x^2 \rangle^2$: **curvature** of a distribution $p(x)\mathrm{d}x$

$k > 3$	flat	Table Mountain	barycurtic
$k = 3$	Gaussian	Mont Blanc	mesocurtic
$k < 3$	peaked	Matterhorn	leptocurtic

weak non-Gaussianity: Edgeworth-expansion

- describe approximatively a probability density $g(x)\mathrm{d}x$ close to a Gaussian $p(x)\mathrm{d}x$ with measured skewness and kurtosis
  - $g(x)$ has cumulants κ_n and characteristic function $g(t)$
  - likewise, $p(x)$ has cumulants γ_n and the characteristic function $p(t)$
- consider characteristic function, and its expansion into cumulants:

$$g(t) = \exp\left[\sum_n (\kappa_n - \gamma_n)\frac{(\mathrm{i}t)^n}{n!}\right]p(t)$$

- $(\mathrm{i}t)^n p(t)$ is the Fourier transform of $(-\frac{\mathrm{d}}{\mathrm{d}x})^n p(x)$
- transformed back into real space:

$$g(x) = \exp\left[\sum_n (\kappa_n - \gamma_n)\frac{(-1)^n}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\right]p(x)$$

weak non-Gaussianity: Edgeworth-expansion

- $p(x)$ =Gaussian, chosen such that $\mu = \kappa_1$ and $\sigma^2 = \kappa_2$ (remember that a Gaussian has only two non-zero cumulants)
- approximate $g(x)$ with a Gaussian + correction terms

$$g(x) = \exp\left[\sum_{n=3} \kappa_r \frac{(-1)^n}{n!} \frac{\mathrm{d}^n}{\mathrm{d}x^n}\right] \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- this approximation series is in general case called **Gram-Charlier A-series**, if $p(x)$ is chosen as Gaussian, one refers to the expansion as **Edgeworth** expansion
- carrying out the derivatives yields a series in Hermite-polynomials

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\left[1 + \frac{\kappa_3}{3!\sigma^3} H_3\left(\frac{x-\mu}{\sigma}\right) + \frac{\kappa_4}{4!\sigma^4} H_4\left(\frac{x-\mu}{\sigma}\right)\right]$$

truncating the series after the 4th order

weak non-Gaussianity: Edgeworth-expansion

- carrying out the derivatives yields a series in Hermite-polynomials

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\left[1 + \frac{\kappa_3}{3!\sigma^3}H_3\left(\frac{x-\mu}{\sigma}\right) + \frac{\kappa_4}{4!\sigma^4}H_4\left(\frac{x-\mu}{\sigma}\right)\right]$$

- $H_n(x)$ are the Hermite polynomials

$$H_n(x)$$

  beware of the two differing definitions in the literature!

- if the non-Gaussianities κ_n, $n \geq 3$, become too large, $p(x)$ might get negative in violation of the Kolmogorov axioms $\rightarrow$ the Gram-Charlier-series can only be approximative

question

please verify by integration that the cumulants of $g(x)$ are in fact μ, σ and $\kappa_{3,4}$

adding and scaling random distributions

- adding random numbers ≡ multiply the characteristic functions

$$\phi_{x+y}(t) = \langle\exp(\mathrm{i}t(x+y))\rangle = \langle\exp(\mathrm{i}tx)\exp(\mathrm{i}ty)\rangle$$

- use independency

$$\ldots = \langle\exp(\mathrm{i}tx)\rangle\langle\exp(\mathrm{i}ty)\rangle = \phi_x(t)\phi_y(t)$$

- consequently, cumulants $\kappa_n \propto \ln\phi$ add : $\kappa_n(x+y) = \kappa_n(x) + \kappa_n(y)$
- scaling random numbers

$$\phi_{cx}(t) = \langle\exp(\mathrm{i}t(cx))\rangle = \langle\exp(\mathrm{i}tcx)\rangle$$

- cumulant is a homogeneous function of order n

$$\kappa_n(cx) = c^n\kappa_n(x), \quad \text{because} \quad \frac{\partial^n}{\partial t^n}\phi_{cx}(t) = c^k\frac{\partial^n}{\partial(ct)^n}\phi_{cx}(t) = c^k\frac{\partial^n}{\partial t^n}\phi_x(t)$$

central limit theorem

central limit theorem

the sum of a large number of random numbers is approximately Gaussian distributed, if the numbers originate from **independent** random processes with **finite variance**

- CLT is the reason why Gaussian distributions are so ubiquitous
- define auxiliary variable y

$$y = \frac{1}{\sqrt{n}} \sum_{i}^{n} x_i$$

- notice similarity to the law of large numbers!

derivation of the central limit theorem

- assume (without loss of generality) that the x_i originiate from the **same** underlying distribution
- consider additivity of x_i in the definition of y:

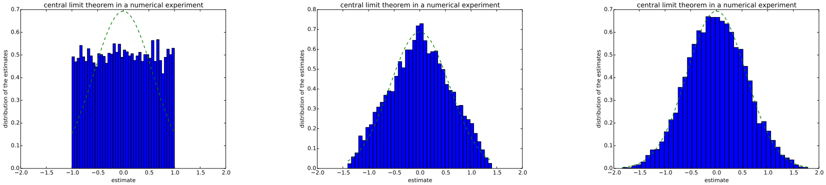
$$\kappa_k(y) = \sum_i^n \kappa_k\left(\frac{x_i}{\sqrt{n}}\right) = n^{-k/2} \sum_i^n \kappa_k(x_i)$$

- cumulants $\kappa_1 < a$ and $\kappa_2 < b$ are finite, with two numbers a, b:

$$\kappa_1(y) \leq n^{-1/2} na = \sqrt{n}a \quad \text{and} \quad \kappa_2(y) \leq n^{-1} nb = b$$

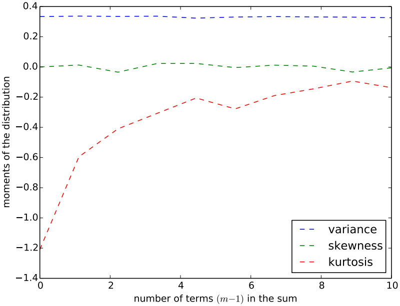
- cumulants with $k \geq 3$ are suppressed and approximate zero, because their proportionality $\propto n^{(2-k)/2}$
- in the limit $n \rightarrow \infty$, only two cumulants remain: Gaussian
- y is Gaussian distributed with $\mu = \sqrt{n}\kappa_1(x_i)$ and $\sigma^2 = \kappa_2(x_i)$

central limit theorem: convolution



distributions of the sum of 1,2,4 uniform distributed random numbers

central limit theorem: convergence



convergence of the moments towards the Gaussian values

- start with a uniform distribution and build up $x = \sum_i^m x_i / \sqrt{m}$
- measure the moments of x
- if m is large, the moments approximate their Gaussian values

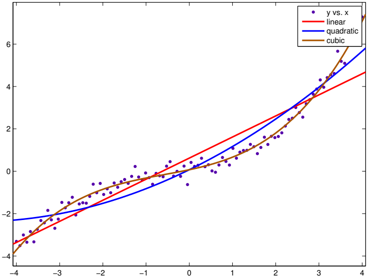
central limit theorem: visualisation

- adding random numbers means multiplying their characteristic functions
- transfrom back to real space: multiplications in Fourier-space are convolutions in real space
- adding random numbers: convolve their probability density
- convolution forces the pdfs to become Gaussian
- final state: convolution of two Gaussians is a Gaussian again

very curious...

the self-convolution of a Cauchy-distribution is the Cauchy-distribution again

fun with moments: linear regression



data points (x_i, y_i), polynomial models $y(x)$

- data (x_i, y_i) with errors σ_i, polynomial model $y(x)$
- best model?
  $\rightarrow$ linear inversion problem formulated with the moments!

fitting of a straight line

- Gauß' idea: minimise squared distance between model and data

$$\chi^2 = \sum_{i=1}^{N} |y(x_i) - y_i|^2 = \sum_{i=1}^{N} |mx_i + b - y_i|^2 \geq 0$$

- minimisation: partial derivatives of χ^2 wrt model parameters vanish

$$\frac{\partial\chi^2}{\partial m} = 0 \quad\rightarrow\quad m\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i x_i \tag{3}$$

$$\frac{\partial\chi^2}{\partial b} = 0 \quad\rightarrow\quad m\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} 1 = \sum_{i=1}^{N} x_i \tag{4}$$

- write as a matrix equation (after division with N)

$$\underbrace{\begin{pmatrix} \langle x_i^2\rangle & \langle x_i\rangle \\ \langle x_i\rangle & 1 \end{pmatrix}}_{=Q} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \langle y_i x_i\rangle \\ \langle y_i\rangle \end{pmatrix}$$

fitting of a straight line

- matrix equation can be solved, if $\det(Q) \neq 0$, so that Q^{-1} exists
- no numerical extremisation is necessary, and the fitting is mathematically exact
- normalisation by N affects χ^2, not the χ^2 we're going to treat in the lecture about likelihoods, but convenient because the moments turn out correctly normalised
- fit can be extended to polynomials, but the inversion of the matrix becomes difficult
- overfitting of data is possible - a polynomial of order $m = N$ will go through all data points exactly

question

derive the fit of a horizontal line, i.e. of the model $y(x) = b$ to data (x_i, y_i). show that $b = \langle y_i \rangle$ (as one would expect)!

fitting of a polynomial

- fit can be extended to a polynomial model of order m

$$y(x) = \sum_{j=0}^{m} p_j x^j$$

- which gives a linear system of equations of the type

$$\left(\begin{array}{ccc} \langle x_i^{2m}\rangle & \ldots & \langle x_i^m\rangle \\ \vdots & \ddots & \vdots \\ \langle x_i^m\rangle & \ldots & 1 \end{array}\right)\left(\begin{array}{c} p_m \\ \vdots \\ p_0 \end{array}\right) = \left(\begin{array}{c} \langle y_i x_i^m\rangle \\ \vdots \\ \langle y_i\rangle \end{array}\right)$$

which can be inverted for the parameters $p_0 \ldots p_m$

question

it is desirable to introduce a weighting $\propto 1/\sigma_i$ if σ_i are the individual errors in y_i. why σ_i^{-1}? and how would you incorporate it?

fitting of a horizontal line

- fit a **very simple** model:

$$y(x) = b$$

- which gives a single equation

$$\frac{\partial\chi^2}{\partial b} = 0 = 2\sum_i (y_i - b) \rightarrow b = \frac{1}{N}\sum_i y_i$$

question

is this a surprising result?

question

what happens if there are more parameters p_j than data points x_i?

summary

- Gaussian has **amazing** properties
- characteristic function gives a way of adding probability distributions
- distributions close to a Gaussian can be approximated with the Edgeworth expansion
- inference of a probability density from data is difficult: only a finite number of moments is measurable
- fitting of polynomials to data can be formulated as a linear problem using the moments of the data
- central limit theorem shows why most random process are approximately Gaussian

now:

we know everything to derive a theory of fitting of arbitrary models!