statistical tests

statistics and data analysis (chapter 6)

Björn Malte Schäfer

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

May 29, 2016

outline

repetition

- description of random distributions
- Gaussians
- weak non-Gaussianities: Gram-Charlier series
- linear regression, polynomial models
- central limit theorem

numerical exercise

fit a linear model $y = ax^2 + bx + c$ to data (y_i, x_i) with noise superimposed. generate new noise realisations and derive the distribution of the parameters a, b and c. what's the shape of the distributions and how do they depend on the amount of noise?

theory of statistical tests

- aim: statistical tests are used to find out whether hypotheses are to be rejected or can be accepted. In this context, hypotheses need to be mathematically formulated
- a hypothesis makes a statement about a statistically measurable quantity, concerning a moment, or a functional shape
- the hypothesis to be tested is referred to as the **null**-hypothesis H_0, H_0 is accepted or rejected in comparison to an alternative hypothesis H_1
- we assign a **likelihood** to data, if the model behind the data is explained by a hypothesis H

please notice

that we try to find the best (true) explanation for already existing data. the random experiment has already been carried out. we **suspect** the true random experiment to be the one with the highest probability of having produced the data on average

classification: (non-)parametric, simple/composite

- a parametric hypothesis makes a quantitative statement for a given model, and a non-parametric hypothesis investigates the truth of a general class of models
- a simple hypothesis makes a statement about the complete parameter set of a random process, a composite hypothesis makes a simultaneous statement about a subset

question: classify the following hypotheses

- each day of the year is equally likely as a birthday among students
- photon arrival times in a detector follow a Poisson distribution
- heights of people in this lecture hall are drawn from a Gaussian distribution with mean 175cm
- heights of people in this lecture hall are drawn from a Gaussian with mean 175 cm and standard deviation 10cm

acceptance regions and critical regions

- x is a random variable with a certain property, described by the hypothesis H_0
- for x, a number of samples are available, so that the property can be **estimated**
- a subset ω of Ω is defined such that
  - events $x \in \Omega \setminus \omega$ provide support for accepting the hypothesis H_0
  - events $x \in \omega$ provide support for rejecting the hypothesis H_0
- test problem: find **for a given value** α a suitable ω such that

$$P(x \in \omega | H_0) = \alpha$$

α is called **size** and is the probability of rejecting H_0 even though it is correct. $1 - \alpha$ is called the level of confidence

acceptance and critical region

ω is called the critical region, $\Omega \setminus \omega$ is the acceptance region

types of error

- 2 types of error are usually interesting
  1. H_0 is rejected but in reality it is true (type 1-error: false negative)
  2. H_1 is accepted (instead of H_0) but in reality it is false (type 2-error: false positive)
- probability of error type 1 is α
- probability of error type 2 is β, **depends** on alternative hypothesis
- power of the test: if $P(x \in \Omega \setminus \omega | H_1) = \beta$, then

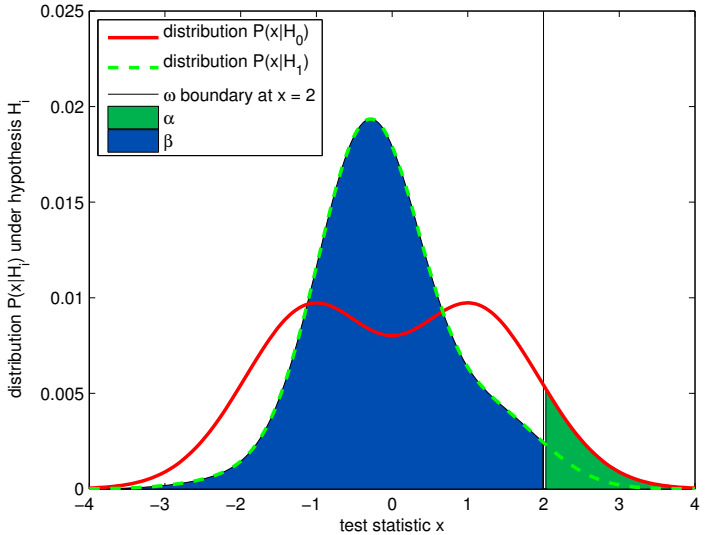
$$P(x \in \omega | H_1) = 1 - \beta$$

is called the power of the test for H_0 **against** the hypothesis H_1

don't you find it weird, that

we are **not** trying to find out, which hypothesis is true (or at least "more true", in the sense of being a better description of Nature)?

visualisation of α and β



$\omega = [2, \ldots, \infty]$, and therefore $\Omega \setminus \omega = [-\infty, \ldots, 2]$

- $P(x \in \omega|H_0) = \int_\omega \mathrm{d}x \mathcal{L}(x|H_0) = \alpha$ determines ω for a given α
- $P(x \in \Omega \setminus \omega|H_1) = \int_{\Omega\setminus\omega} \mathrm{d}x \mathcal{L}(x|H_1) = \beta$ for the alternative hypothesis

chain of definitions

choose size $\alpha \rightarrow$ rejection $\omega \rightarrow$ acceptance $\Omega \setminus \omega \rightarrow$ power $1 - \beta$

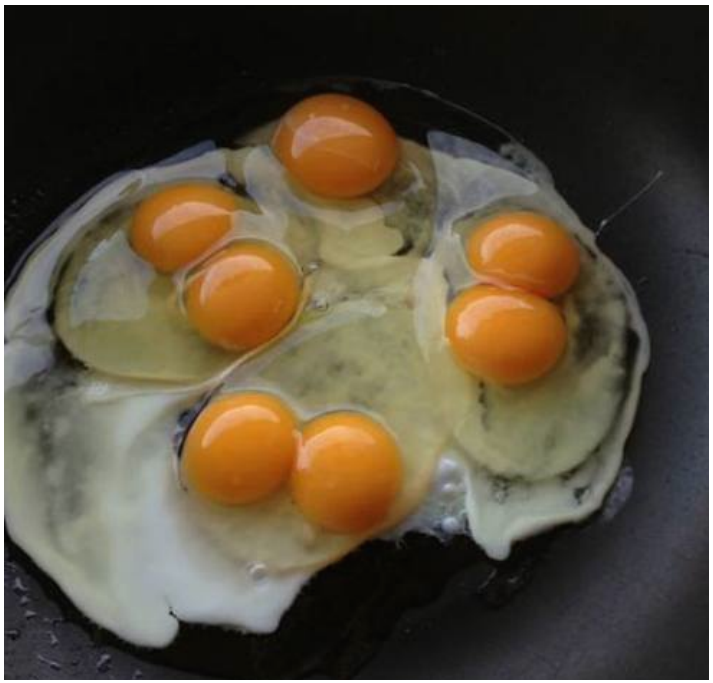
likelihoods

- if samples x_i for a random variable x are available, and if the properties of the random variable are described by a hypothesis H_0, it makes sense to ask if it would have been likely to obtain these x_i
- **likelihood** is the probability that a set of samples has been obtained
- if each sample originates independently from a random distribution $p(x)\mathrm{d}x$ described by a hypothesis H_0, then the likelihood for a set x_i

$$\mathcal{L}(x_i|H_0) = \prod_i p(x_i)$$

- the likelihood would be different if one assumes a different hypothesis, even for the **same** data! $\mathcal{L}(x_i|H_0) \neq \mathcal{L}(x_i|H_1)$
- likelihood: model is variable, but data is fixed

difference between likelihood and probability???



improbable or unlikely?

- both are probabilities in the mathematical sense, obeying the Kolmogorov-axioms
- probability: **fixed model**, makes a statement about the future
- likelihood: data is there, interpret data with **different models**

likelihood ratios

- the likelihood ratio is the quotient between the likelihood of H_0 explaining data x_i and the likelihood of the alternative hypothesis H_1

$$q = \frac{\mathcal{L}(x_i|H_1)}{\mathcal{L}(x_i|H_0)}$$

- $\ln q > 0$ prefers the explanation H_1 over H_0 for explaining the data x_i, $\ln q < 0$ vice versa
- the test problem is then reduced to finding a number c such that the hypothesis H_0 is accepted if $q < c$ and rejected if $q > c$
- we will discuss 3 specific types of tests
  - t-test: tests for the mean of an assumed Gaussian distribution
  - F-test: tests whether the mean values obtained in different types of experiment can be considered equal
  - Kolmogorov-Smirnov test: tests whether two samples can be considered as drawn from the same distribution

Neyman-Pearson lemma

- theoretical justification for the definition of a likelihood ratio
- Neyman-Pearson lemma: the likelihood ratio,

$$q = \frac{\mathcal{L}(x_i|H_1)}{\mathcal{L}(x_i|H_0)} \leq c \quad \text{with} \quad P(q < c|H_0) = \alpha$$

is the **most powerful** test with **size** α at the threshold c

proof of the Neyman-Pearson lemma: idea

compare the likelihood ratio q with another test between two hypotheses H_0 and H_1 and show that the **rejection region of another test is smaller**, meaning that the probability of getting a q inside the rejection region is smaller and the test weaker.

proof of the Neyman-Pearson lemma

- rejection region for H_0 under the likelihood ratio

$$R_{NP} = \left\{ x : q = \frac{\mathcal{L}(x|H_0)}{\mathcal{L}(x|H_1)} \leq c \right\}$$

  while any other test has the rejection region R_A

- probability of data falling in R under the hypothesis H

$$P(R, H) = \int_R \mathcal{L}(x|H)$$

- assume: both tests have the same size α: $P(R_{NP}|H_0) = P(R_A|H_0) = \alpha$
- decompose

$$\begin{aligned} P(R_{NP} \cap R_A|H) + P(R_{NP} \cap CR_A|H) &= P(R_{NP}|H) \\ P(R_{NP} \cap R_A|H) + P(CR_{NP} \cap R_A|H) &= P(R_A|H) \end{aligned}$$

proof of the Neyman-Pearson lemma

- equate: equal probabilities of accepting H_0

$$P(R_{NP} \cap CR_A|H_0) = P(CR_{NP} \cap R_A|H_0)$$

- advantage of likelihood ratio over any other test for H_1

$$P(R_{NP}|H_1) \geq P(R_A|H_1) \quad \text{if} \quad P(R_{NP} \cap CR_A|H_1) \geq P(CR_{NP} \cap R_A|H_1)$$

- because

$$\begin{aligned}
P(R_{NP} \cap CR_A|H_1) &= \int_{R_{NP} \cap CR_A} \mathcal{L}(x|H_1) \geq \frac{1}{c} \int_{R_{NP} \cap CR_A} \mathcal{L}(x|H_0) \\
&= \frac{1}{c} P(R_{NP} \cap CR_A|H_0) = \frac{1}{c} P(CR_{NP} \cap R_A|H_0) \\
&= \frac{1}{c} \int_{CR_{NP} \cap R_A|H_0} \mathcal{L}(x|H_0) \geq \int_{CR_{NP} \cap R_A|H_0} \mathcal{L}(x|H_1) \\
&= P(CR_{NP} \cap R_A|H_1)
\end{aligned}$$

statistical testing

- we check, if a hypothesis is true
- this needs the formulation of a null-hypothesis H_0 and an alternative hypothesis H_1
- we define a quantity, the **test statistic**, which can be measured from data
- then, we figure out the distribution of this test statistic, for being able to quantify, how probable a value for the test statistic is
- for a given probability α we can then determine the acceptance region for the hypothesis H_0: these would be acceptable values for the test statistic, with total probability α
- in a **right sided test**, we could then simply state a threshold number, such that the integrated probability for the test statistic to be **above the threshold** is equal to α

t-test: idea

motivation:

A pasta-producing company checks the length of their spaghetti and wants to see if they conform to the specifications. are spaghetti in a package described by a specified mean length?

- the t-test is a test of **identical mean**, either of a data set with a specified number, or between two data sets
- we treat a special case of Gaussian distributions of the random numbers, with a certain (unknown) mean and variance
- formulation of hypotheses:
  - null-hypothesis: equal means, $\mu_1 = \mu_2$
  - alternative hypothesis: unequal means, $\mu_1 \neq \mu_2$
- both means and variances are estimated from each of the data sets

t-test

- naive (but not stupid!): $\bar{x} = \sum_i x_i / n$ is Gaussian-distributed with mean μ_1 and the standard deviation $\sigma_1 / \sqrt{n} \rightarrow$ any definition if the acceptance region for a given α is easy
- but: we don't know the true σ^2, which needs to be estimated from data as well
- define the test statistic:

$$t = \sqrt{n}\frac{\bar{x} - \mu_1}{S} \quad \text{with} \quad S^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \tag{1}$$

- t follows the student-t distribution
- advantage: the definition of t is a common test statistic with a distribution which you could derive for a Gaussian with zero mean and unit variance, and it maps all problems with an underlying Gaussian distribution on the same t-distribution of the test statistic

student-t distribution

- $X = \bar{x}$ is Gaussian distributed, and $Y = s(x)$ is χ^2_n-distributed (we'll do a detailed derivation next lecture in a different context)
- what's the distribution of $t = X/\sqrt{Y/k}$?
- assume independent random variables X and Y from p_x and p_y
- cumulative distribution of Z:

$$P(Z \leq z) = \int \int_{x/y \leq z} \mathrm{d}x\mathrm{d}y \; p_x(x) p_y(y)$$

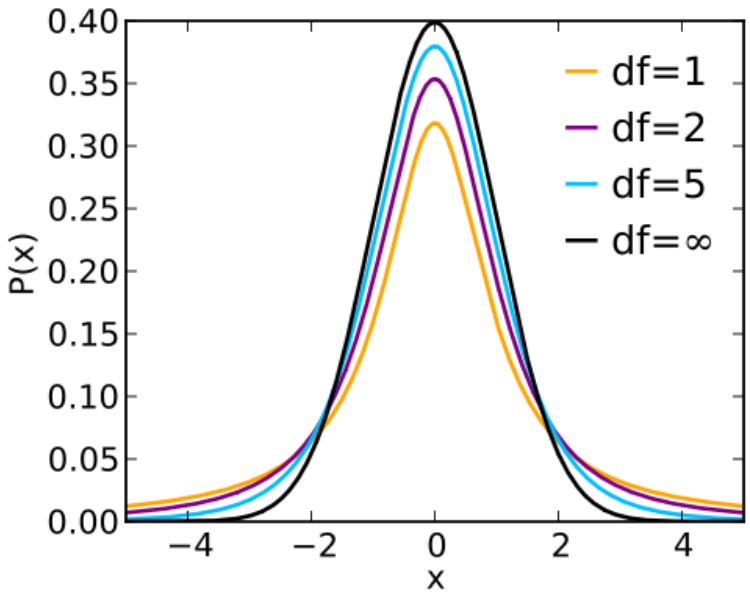
- define $t = x/y$, $\mathrm{d}x = y\mathrm{d}t$

$$P(Z \leq z) = \int_{-\infty}^{z} \mathrm{d}t \int_{-\infty}^{+\infty} \mathrm{d}y \; y p_x(ty) p_y(y)$$

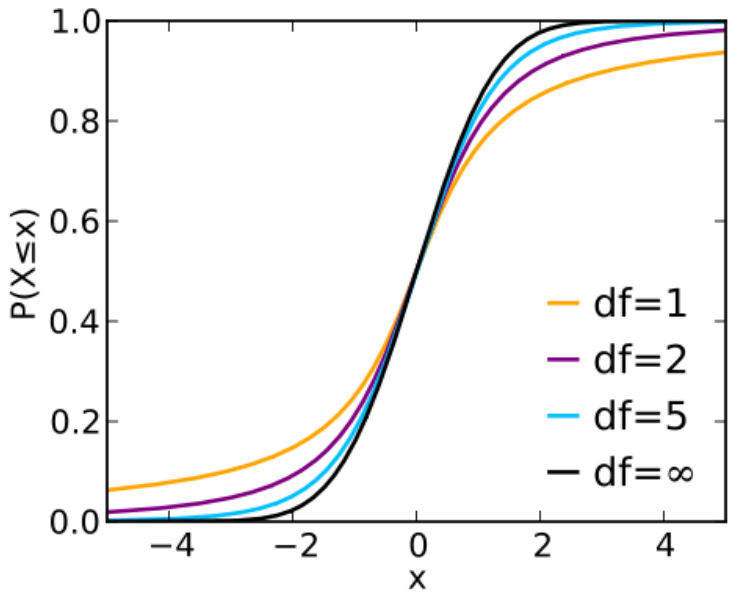
- with $p(z) = \partial P(Z \leq z)/\partial z$: student-$t$ distribution with k degrees of freedom

$$p_k(x) = \frac{\Gamma\left(\frac{1+k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{1}{2}\right)} \frac{1}{\sqrt{k}} \left(1 + \frac{x^2}{k}\right)^{-(1+k)/2}$$

student-t distribution: visualisation (from wikipedia)



student-t probability density



student-t cumulative distribution

student-t distribution

- the test statistic $T(x)$ is distributed according to the **student-t distribution** with $n - 1$ degrees of freedom
- for some α we can now figure out the rejection and acceptance regions ω and $\Omega \setminus \omega$ of the test statistic
- if a measured t falls within this acceptance region, $t \in \Omega \setminus \omega$, we would conclude that the hypothesis H_0, which states that $\bar{x} = \mu_0$, is true
- only in a fraction α of all cases we would have rejected the hypothesis $\bar{x} = \mu_0$ wrongly, meaning that we would (by chance) obtain $t \in \omega$ when in reality the means coincide given the variances of the random experiment
- we can not quantify the type-2 error β because for the hypothesis $\bar{x} \neq \mu$ one can not write down a likelihood. this would be different if the alternative hypothesis was $\bar{x} = \mu_1$
- then, we would have accepted the hypothesis H_1 wrongly in a fraction β of all cases

variant:

- two sets of random numbers from two random experiments with unknown means and variances
- test for equal mean, irrespective of the variance
- difference to the previous case: both variances and both means are estimated
- formulate hypotheses
  - null-hypothesis: $\mu_x = \mu_y$
  - alternative hypothesis: $\mu_x \neq \mu_y$
- define the test statistic

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{2}$$

- variance

$$S^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n + m - 2} \tag{3}$$

F-test

motivation:

one has survey data for studying physics at different universities: Two universities have equal mean study times, but do the students finish with the same variance?

- the F-test is a test of **identical variance** of a number of data sets
- some of the argumentation will remind you of the CLT
- we treat a special case where the random numbers follow a Gaussian distribution, of which we try to decide, if the mean values are statistically equal
- formulation of hypotheses:
  - null-hypothesis: equal variances, $\sigma_1^2 = \sigma_2^2$
  - alternative hypothesis: unequal variances, $\sigma_1^2 \neq \sigma_2^2$
- where the variances and the means are estimated from each of the sets

F-test

- define the test statistic:

$$F = \frac{\frac{1}{n_2-1}\sum_i^{n_2}(x_{2,i} - \bar{x}_2)^2}{\frac{1}{n_1-1}\sum_i^{n_1}(x_{1,i} - \bar{x}_1)^2} \tag{4}$$

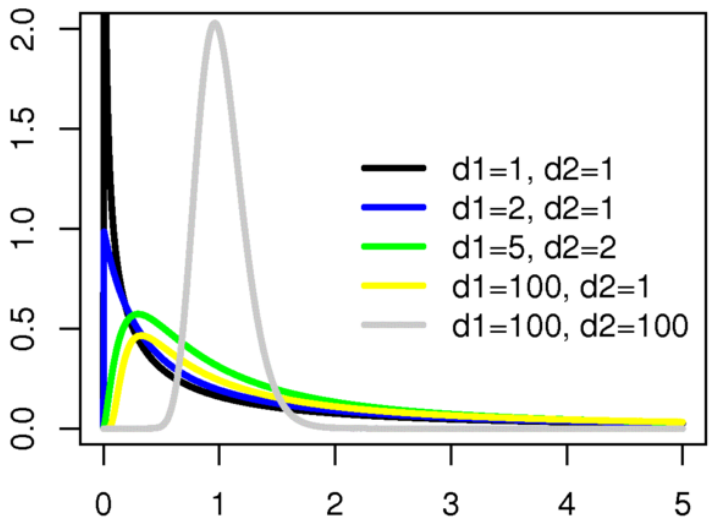
- means $\bar{x}_1$ and $\bar{x}_2$ are estimated from the data $x_{1,i}$ and $x_{2,i}$
- if the value of F is too large, the idea that both variances are equal is rejected
- F is distributed according to the F-distribution $F(n_2 - 1, n_1 - 1)$, with

$$F_{m,n}(x) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}\, m^{m/2} n^{n/2}\, \frac{x^{m/2-1}}{(n+mx)^{(m+n)/2}} \tag{5}$$

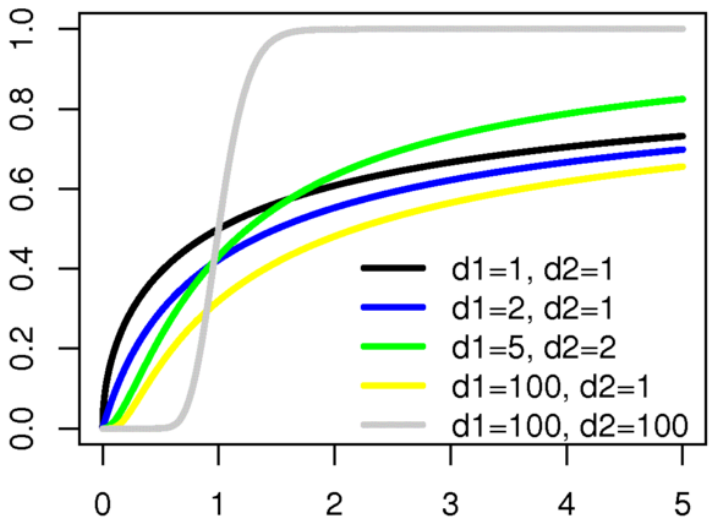
- this is related to the the χ^2-distribution because of

$$F_{m,n}(x) = \frac{\chi^2_m/m}{\chi^2_n/n} \tag{6}$$

F-distribution: visualisation (from wikipedia)



student-t probability density



student-t cumulative distribution

Kolmogorov-Smirnov test

motivation:

2 successful TV-series among students are *Breaking Bad* (chemistry) and *The Big Bang Theory* (physics). a survey gives ages of watchers for a limited number of people. Are the distributions identical?

- KS-test is **nonparametric**: it tests (entire) distributions for equality
- N sample points x_i are given. then, $S_N(x)$ is the fraction of data points with $x_i < x$
- $S_N(x)$ is an estimate of the cumulative distribution, and $0 \leq S_N(x) \leq 1$
- test statistic D: 2 variants
  - test, if a set of events conforms to a given distribution
  - test, if two sets of events are statistically equivalent

Kolmogorov-Smirnov test

- test, if a set of events conforms to a given distribution $P(x)$

$$D = \sup_x |S_N(x) - P(x)|$$

- test, if two sets of events are statistically equivalent

$$D = \sup_x \left|S_{N_1}(x) - S_{N_2}(x)\right|$$

- the distribution of D can be computed: in the limit of sufficiently many data points, $N_e = N$ for one sample and $N_e = N_1 N_2/(N_1 + N_2)$, the probability of D being larger than a threshold c is

$$\lim_{N_e \to \infty} P(D > c/\sqrt{N_e}) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 z^2)$$

summary

- formulation of hypotheses
- critical regions and acceptance regions
- types of error
- likelihood and likelihood ratio for comparing hypotheses
- 3 specific examples
  - t-test
  - F-test
  - Kolmogorov-Smirnov test
- this was the most complicated lecture! ☺

data fitting

is a particular type of statistical test: we will look for a model $y(x)$ which provides the largest likelihood for explaining data y_i and will make statements about the model parameters p_μ such as confidence intervals