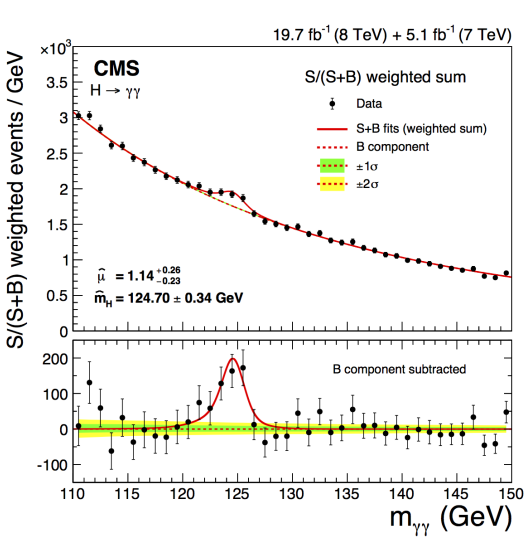
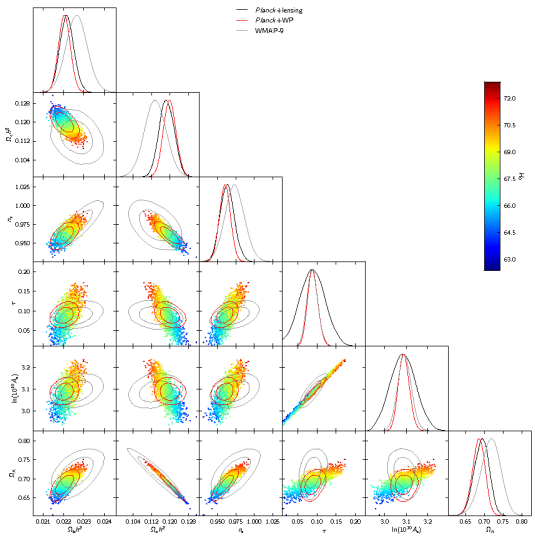
Advanced Statistics 2017/18

University of Amsterdam

Christoph Weniger

January 8, 2019





Everything marked with an asterisk (‘$*$’) is not relevant for the exam.

1 Introduction

Statistical inference is the process of deducing the values of model parameters, statements about the validity of models, or more generally any underlying probability distribution, from data. There are two main approaches to statistical inference. These are connected to two distinct ways of defining the concept of *probability*. One is Frequentist inference, the other Bayesian inference. Both have their advantages and disadvantages, and can be used to do good science or to mislead yourself and others. In this course you will learn the basic ideas behind both approaches, and when and how they work.

1.1 What is probability?

1.1.1 Frequency

The concept of probability that underlies Frequentist inference is, perhaps not surprisingly, that probabilities are *frequencies* (in German one can use the work 'Häufigkeit', which is distinct from 'Frequenz'). Probabilities refer to the *frequency* of **event** X in repeated identical experiments. The probability (or frequency) of event X is then defined by

$$p_x = \lim_{N\to\infty} \frac{n_X}{N} \,,$$

where N denotes the total number of experiments, and n_X the number of outcomes X. Note that n_X is a random number, but that the above expression converges in the large-sample limit, $N \to \infty$.

The **central question** that is asked in a Frequentist analysis is: "Given some hypothesis H, how improbable is my measurement?". The main technique for Frequentist analysis is *Hypothesis testing*.

1.1.2 Belief

In Bayesian inference, the word probability refers to the *degree of belief* in, or the *plausibility* of a **proposition**.[1] The interpretation of specific numerical values of p_x are in principle somewhat arbitrary, but it is convenient to define

$$p_x = 1: \quad \text{Proposition } X \text{ is certainly true}$$

$$p_x = 0: \quad \text{Proposition } X \text{ is certainly false}$$

$$p_x > p_y: \quad \text{Proposition } X \text{ is more plausible/likely/probable than proposition } Y$$

One can now do arithmetics with Bayesian probabilities, since propositions can be logically combined. For instance $p_{X+Y} = p_X + p_Y$, if X and Y are mutually exclusive (otherwise we have $p_X + p_Y \geq p_{X+Y} \geq \max(p_X, p_Y)$).

The **central questions** in Bayesian inference is now *How does a given piece of data change my prior belief in proposition X?* The method of choice is here *Bayes' theorem*.

1.2 Common probability distributions

A number of common probability distribution functions (PDFs) will play repeatedly a role during this course. For reference, I summarize here their main properties.

[1] Note that this can, as a special case, include Frequencies, since a possible proposation can be "Event X has happened."

1.2.1 Binomial distribution

In the case of N independent experiments with the outcome yes/no, the probability mass function (PMF) for obtaining k times a 'yes' is called 'binomial distribution'. It is given by

$$f(k) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k} ,$$

where p is the probability to obtain 'yes' in one experiment. For $N = 1$, we trivially obtain $p(0) = 1 - p$ and $p(1) = p$ (this is the 'Bernoulli distribution'). The mean is given by $\langle k \rangle = Np$, the variance by $\mathrm{var}(k) = Np(1-p)$.

Example: Drawing N green/blue balls from a jar with replacement (or from a jar with an infinite number of balls).

1.2.2 Poisson distribution

The Poisson distribution follows from the binomial distribution in the limit $N \to \infty$ while keeping $\lambda \equiv \langle k \rangle = Np$ constant. The PMF reads

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!} .$$

The mean and variance are given by $\langle k \rangle = \lambda$ and $\mathrm{var}(k) = \lambda$ respectively.

Example: Number of radioactive decays in a given time interval.

1.2.3 Normal distribution

This is the by far most common PDf, which is a consequence of the central limit theorem (more below). It reads

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} .$$

The mean and variance are here given by $\langle x \rangle = \mu$ and $\mathrm{var}(x) = \sigma^2$. We will often talk about the *standard normal distribution*, which corresponds to a normal distribution with $\mu = 0$ and $\sigma = 1$.

Examples: Many distributions approximate asymptotically (in the 'large sample limit') a normal distribution, which is a consequence of the CLT. For instance the (discrete) Poisson distribution approaches a (quasi-continuous) normal distribution in the limit $\lambda \gg 1$, with $\mu = \sigma^2 = \lambda$. An example for an *exact* normal distribution in nature is however harder to find. One such example is the PDF of a ground state of a quantum harmonic oscillator.

1.2.4 The χ^2 distribution

Measurement errors are often approximately normal distributed. If one measures the size of these errors in units of 'standard deviations', and sums their squares, one obtains the χ^2 distribution. Although the χ^2-distribution does not have apparent direct analogues in nature (let me know if you can come up with an example!), it is an indispensable element of statistical inference.

In order to construct a χ^2_k distributed random variable, let us define

$$X = \sum_{i=1}^{k} x_i^2 .$$

If all x_i are independent standard normal distributed random variables, then X is χ^2_k distributed with k 'degrees of freedom'. The corresponding PDF reads

$$f(x) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)} ,$$

where $\Gamma(\cdot)$ is the Gamma function. The mean and variance are given by $\langle x \rangle = k$ and $\mathrm{var}(x) = 2k$.

1.3 Central Limit Theorem

The central limit theorem (CLT) states that the sampling distribution of the mean of any set of independent random variables will be nearly normal distributed, provided the sampling size (number of random variables) is 'large enough'.

More specifically, the CLT states that if we consider the arithmetic mean of random variables $y_1, \ldots, y_N$, given by

$$X = \frac{1}{N}(y_1 + \cdots + y_N)$$

the X follows (provided $N \gg 1$) approximately a normal distribution. The mean is given by

$$\langle X \rangle = \mu = \frac{1}{N}(\mu_1 + \cdots + \mu_N) ,$$

and the variance is given by

$$\mathrm{var}(X) = \sigma^2 = \frac{1}{N^2}\sum_i \sigma_i^2 .$$

Here, μ_i and σ_i are the mean and the variance of random variable y_i.

What exactly constitutes a 'large enough' number of contributing variables depends somewhat on the specific context.[2] I mention here just two simple rules of thumb.

1. If the individual y_i are already close to normal distributed, smaller values of N are sufficient.

2. In general, the central part of the PDF of X approaches the normal distribution faster than the tails of the distribution.

2 Hypothesis test

Statistical inference (both Frequentist and Bayesian) attempts to provide a systematic approach for reasoning in the presence of noise and uncertainties. Bayesian inference deals with this by extending binary logic beyond the values 'true' and 'false', by introducing intermediate truth values. This will be discussed in later chapters.

Frequentist inference, on the other hand, keeps the logic binary at its core. Propositions are always qualified as either 'true' or 'false'. In the presence of noisy data this means, however, that sometimes errors are made.

[2] The conditions can be defined exactlly, by considering generating moments of the distributions.

The main goal of Frequentist inference is to *keep the **frequency** of false conclusions* under control.

The word 'frequency' is here crucial. It is connected to the idea that if we were to repeat the *same* experiment over and over again, we would draw wrong conclusions only in a certain fraction of cases. That has two important consequences.

1. Experiments or measurements that qualify for a Frequentist treatment have to be *repeatable* (at least in principle; we could for instance build another LHC to find the Higgs boson, or we just do the measurement with the same detector twice).

2. The specific steps of the adopted statistical inference strategy have to be the *same* for each of these hypothetical repeated measurements. We *cannot change analysis strategy* depending on the outcome of a specific measurement (this approach is often called 'blind analysis').

2.1 The general strategy

A hypothesis test involves the following steps.

1. Define the null hypothesis, H_0. Your goal is to find observational evidence for H_0 being wrong. In addition, one usually defines an alternative hypothesis, H_A, for which one wants to find supporting evidence, and which guides the design of the test.

2. Define an experiment that has the potential to discriminate between the two. It delivers data $\mathcal{D}$. *But do not measure (or look at) the data yet.*

3. Define a test statistic (TS), which is a real-valued function of the data, $TS = TS(\mathcal{D})$. Like the data itself, TS is a random variable. The TS should have small values if the data supports the null hypothesis, and large values if the data supports the alternative hypothesis.

4. Obtain (by using analytic approximations or Monte Carlo simulations) the *sampling distribution of the TS* under the null hypothesis, $P_{\text{null}}(\text{TS})$.

5. Select the desired 'significance' level of the test, $\alpha \in [0, 1]$. The smaller α, the stronger the test. Find the corresponding threshold value for TS, t, by solving the implicit integral equation
$$\int_0^{t_\alpha} P_{\text{null}}(\text{TS}) d\text{TS} = \alpha$$

6. Make measurement and obtain data $\mathcal{D}$. Calculate measured TS = TS($\mathcal{D}$).

7. Reject null hypothesis if (and only if) TS $> t_\alpha$.

If you had defined an alternative hypothesis, rejecting the null hypothesis is equivalent with 'accepting the alternative'. Note however that there is an asymmetry between the null and the alterntive hypothesis that is intrinsic to the entire process.

Finally, if you now were to repeat steps 6&7 in a loop, *and* if the null hypothesis were true, you would – by construction – falsely reject the null hypothesis just with the (small) frequency α. This is the promise that Frequency inference makes, and it does fulfil it if you stick to the above rules.

Related definitions

Before doing any measurement, one can calculate

- α: The 'significance level' (aka 'false positive rate')
- $(1-\alpha)\cdot 100\%$: 'Confidence level' of the test.
- t_α: The corresponding value of the test statistics TS, connected via the expression
$$\alpha = \int_{t_\alpha}^{\infty} P_{\text{null}}(t)dt$$
- β: The 'false negative rate', which depends on the selected significance level α, and the PDF of the TS value assuming that the *alternative* hypothesis is true
$$\beta = \int_{-\infty}^{t_\alpha} P_{\text{alt}}(t)dt \ .$$
- $1-\beta$: This is called 'statistical power' of the test. It depends on the adopted significance level α, and on the shape of the test statistic PDFs P_{null} and P_{all}. Experimental design typically aims at increasing the statistical power of the experiment (and the subsequent statistical analysis), while keeping the significance level fixed to some specific value. This is achieved by reducing the overlap between null and alternative hypothesis PDFs as much as possible.

After the measurement, one can then calculate

- t_m: The value of TS inferred from the measured data.
- p: The corresponding p-value, defined as
$$p = \int_{t_m}^{\infty} P_{\text{null}}(t) \ .$$
- Outcome of hypothesis test: Reject null hypothesis if $p < \alpha$ (or equivalently $t_m > t_\alpha$).

As mentioned above, in the presence of uncertain data it is unavoidable that sometimes during hypothesis testing false conclusions are drawn. A Frequentist cares about keeping the frequency of these errors (again, in repeated experiments) under control. There are two types of errors.

- 'type I error': Incorrect rejection of a true null hypothesis, or 'false positive'. The expected rate in repeated experiments is α.
- 'type II error': Failure to reject a wrong null hypothesis, or 'false negative'. If the alternative hypothesis is true, the expected rate is β.

2.2 Goodness-of-fit test

A 'goodness-of-fit test' is one of the simplest hypothesis tests. It does not require the definition of an alternative hypothesis, which means that it can be rather universally used. However, it also means that *it is not the most powerful test if the alternative hypothesis can be formulated* (this will be discussed in later sections).

We will here discuss the Pearson's chi-squared test, which is applicable in a large number of cases (another common goodness-of-fit test is the Kolmogorov-Smirnov test).

Pearson's chi-squared test

We illustrate the Pearson's chi-squared test with a few simple examples.

Simple hypothesis. (= no free model parameters). Consider the temperature curve given by

$$T(t) = (T_1 - T_2)e^{-t/\tau} + T_0$$

here t denotes time, $T_{0,1,2}$ are reference temperatures, and τ the decay rate. Let's suppose we measure the temperature T_i at multiple times t_i, with Gaussian errors with a standard deviation of ΔT. We can now define the quantity

$$\chi^2 \equiv \sum_{i=1}^{k} \frac{((T(t_i) - T_i)^2}{\Delta T^2}$$

as our test statistic (but we call it here χ^2, for reasons that become clear shortly). If we now make the variable substitution

$$T_i \mapsto Y_i = \frac{T_i - T(t_i)}{\Delta T}$$

we find that χ^2 can be expressed in a much simpler way, namely

$$\chi^2 = \sum_{i=1}^{k} Y_i^2 \, .$$

This looks suspiciously close to the definition of a random variable that is χ^2_k distributed (see discussion of the χ^2_k distribution above). And indeed, if the T_i are random variables with mean $T(t_i)$ and variance ΔT^2 (this is precisely what is meant with 'errors are Gaussian'), then Y_i are standard normal distributed. And then χ^2 is a random variable that is distributed like a χ^2_k distribution with k degrees of freedom (beware of the sometimes confusing double-use of the notation 'χ^2', once as a random variable connected to the actual measurement T_i, and once as the name of a specific probability distribution function). In short, we can say that $\chi^2 \sim \chi^2_k$.

The goodness-of-fit can now be estimated from the 'reduced chi-squared value',

$$\chi^2_{\text{red}} \equiv \frac{\chi^2}{k}$$

If the null hypothesis is true, then $\langle \chi^2_{\text{red}} \rangle = 1$ and $\text{var}(\chi^2_{\text{red}}) = 2/k$. If the null hypothesis gives a bad fit, then usually $\chi^2_{\text{red}} - 1 \gg \sqrt{2/k}$.

Composite hypothesis (= free model parameters)

Assume that in the above model T_1 is unknown and determined from the measurement. One way to do this is to consider the 'maximum likelihood value', which corresponds to the value of T_1 that minimizes χ^2 (we will see below what this means).

Consider

$$\chi^2_{\text{min}} = \min_{T_1} \chi^2 = \min_{T_1} \sum_{i=1}^{k} \frac{(T(t_i, T1) - T_i)^2}{\Delta T^2}$$

Now, one can show that

$$\chi^2_{\text{min}} \sim \chi^2_{k-1}$$

i.e. χ^2_{min} follows a chi-squared distribution with $k-1$ degrees of freedom. In general

$$k = \#\text{data points} - \#(\text{free}) \text{ model parameters}$$

2.3 Signal detection

If the alternative hypothesis is known and can be formulated, one can derive statistical tests that are much more powerful than goodness-of-fit tests. Like above, we are interested in defining a test statistics TS, which takes into account information about the alternative hypothesis. This can be done in various ways.

2.3.1 Simple Poisson process

Consider the measurement of photons, from for instance an astronomical object, or more generally the number of events in a counting experiment. We expect on average a number of b background events, and are interested in whether the existence of signal events above this background can be claimed.

The corresponding expectation values of the null and alternative hypothesis are given by

- Null hypothesis: $\mu_{\text{null}} = b$
- Alternative hypothesis: $\mu_{\text{alt}} = b + s$

where s is the (unknown) signal rate. As test statistic, we can now simply take TS $= k_m$ (k_m is the number of measured events). In fact, we have here not much choice, since we only measured one number (namely k_m); all test statistics would be here equivalent.

The corresponding p-value is given by

$$p = \sum_{k \geq k_m} f(k|\mu_{\text{null}})$$

It can be used to decide what the 'background-only' hypothesis should be rejected in favor of a signal+background hypothesis or not.

2.3.2 $\Delta\chi^2$-method

This test is somewhat similar to Pearson's goodness-of-fit test, but also accounts for possible extra information that we have about the alternative hypothesis. This makes the test statistically (much) more powerful. An important requirement is here that *the null and the alternative hypothesis must be nested.* Two statistical models are 'nested' if the first model can be transformed into the second by imposing constraints on the parameters of the first model.

As a simple example, consider a random variables

$$X_i = \mu_i(\vec{\theta}) + \epsilon_i \ , \quad \epsilon_i \sim \mathcal{N}(0|\sigma_i^2)$$

where $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)^T$ are the model parameters of the d-dim alternative hypothesis, μ_i the mean value and ϵ_i some Gaussian noise. A possible $d-k$-dim null hypothesis is now obtained by fixing $k < d$ parameters to zero, $\theta_1 = \theta_2 = \cdots = \theta_k = 0$. This is the example that we will typically consider throughout the lecture, although the arguments apply to (almost) any constraints onto a $d-k$-dim sub-manifold $\mathcal{M}_0$ of the original parameter space.

We can now define the $\Delta\chi^2$ test statistic as

$$\Delta\chi^2 \equiv \chi^2_{\text{null}} - \chi^2_{\text{alt}}$$

with

$$\chi^2_{\text{null}} \equiv \min_{\vec{\theta}\in\mathcal{M}_0} \chi^2 \,, \quad \chi^2_{\text{alt}} \equiv \min_{\vec{\theta}} \chi^2 \,, \quad \text{and} \quad \chi^2 \equiv \sum_i \frac{(\mu_i(\vec{\theta}) - x_i)^2}{\sigma_i^2} \,.$$

From the above definitions, it follows that $\Delta\chi^2 \geq 0$.

Let us now assume the null hypothesis is correct. That means that there is a parameter $\vec{\theta}^* \in \mathcal{M}_0$ such that $X_i = \mu_i(\vec{\theta}^*) + \epsilon_i$. An important property of the $\Delta\chi^2$ statistic is then that it follows a (guess what) χ^2_k distribution with k degrees of freedom,

$$\Delta\chi^2 \sim \chi^2_k \quad \text{(if null hypothesis is true)},$$

where the number of degrees of freedom is given by the difference in the number of free model parameters

$$k = d_{\text{alt}} - d_{\text{null}} \,.$$

*Outline of proof.** Consider n data points, a simple null hypothesis and an alternative with k parameters. Without loss of generality, we can set $\sigma_i^2 = 1$, and the $\mu_i = 0$ for the null hypothesis (otherwise, use a linear transform that will leave the $\Delta\chi^2$ statistic invariant). In that case, we find that

$$\Delta\chi^2 = \chi^2_{\text{null}} - \chi^2_{\text{alt}} = \sum_{i=1}^{n} x_i^2 - \left(x_i - \mu_i(\hat{\vec{\theta}})\right)^2 \,,$$

where we defined the minimum chi-square estimator

$$\hat{\vec{\theta}} \equiv \operatorname*{argmin}_{\vec{\theta}} \left(x_i - \mu_i(\vec{\theta})\right)^2 \,.$$

Let us now make the linear approximation $\mu_i(\vec{\theta}) \simeq \sum_j A_{ij}\theta_j$, where A is a $(n \times k)$ matrix. The matrix A spans the k-dim vector space $V \in \mathbb{R}^n$.

You can now convince yourself (by defining an appropriate basis for V and minimizing analytically) that μ_i evaluated at the minimum chi-square estimator equals the projection of $\vec{x}$ onto the subspace V,

$$\vec{\mu}(\hat{\vec{\theta}}) = P_V \vec{x} \,.$$

From this, we find that

$$\Delta\chi^2 = \sum_i x_i^2 - (x_i - (P_V\vec{x})_i)^2 = \|\vec{x}\|^2 - \|P_{V_\perp}\vec{x}\|^2 = \|P_V\vec{x}\|^2$$

where $V_\perp$ is the orthogonal complement of V, and the last step is a high-dimensional version of the Pythagorean theorem.

Now remember that, per definition, $x_i \sim \mathcal{N}(0,1)$. However, only k of these standard normal distributed variables contribute effectively (after projection P_V) to $\Delta\chi^2$. Hence, we find that indeed, under the assumptions made above, the $\Delta\chi^2$ statistic is distributed like a chi-squared distribution with k degrees of freedom.

Q.E.D.

3 Estimators

3.1 Basic definitions

A 'point estimator' is a rule for calculating an estimate of a variable of interest, based on observed (or simulated) data. Since the data is random, the estimator is a random variable. The word 'point' refers to the fact that the inference result is a single point in the model parameter space (in contrast to, for instance, confidence regions, which we will discuss below and which are subsets of the model parameter space with specific sampling properties).

Let us introduce some definitions.

- θ: Parameter of interest (we just consider the 1-dim case here, which is straightforward to generalize to multiple dimensions).
- $\hat{\theta}$: Estimator, a function of data $\mathcal{D}$, $\hat{\theta} = \hat{\theta}(\mathcal{D})$.
- $b = \langle\hat{\theta}\rangle - \theta$: Bias of the estimator. Averages $\langle\cdot\rangle$ are taken over data $\mathcal{D}$ assuming model parameter θ is correct.
- $\mathrm{var}(\hat{\theta}) = \langle(\hat{\theta} - \langle\hat{\theta}\rangle)^2\rangle$: Variance of the estimator (the spread of the estimator under multiple measurements).
- $\mathrm{MSE}(\hat{\theta}) = \langle(\hat{\theta} - \theta)^2\rangle$: Mean square error (the typical deviation of the estimator from the true value).
- An estimator is 'unbiased' if $b = 0$. Estimators with small bias are 'accurate' (but not necessarily precise).
- An estimator is 'minimum variance' if any other estimator has equal or larger variance. Estimators with low variance are 'precise' (but not necessarily accurate).

Optimal estimators are minimum-variance unbiased estimators (MVUE). We will see examples for such estimators below.

3.2 Fisher information

The Fisher information is a way to quantify the amount of information that a measurement carries about a specific variable of interest.[3] It is a property of the **likelihood function** $\mathcal{L}(\vec{\theta}|\vec{x})$, which is related the parametric PDF for data $\vec{x}$ via $\mathcal{L}(\vec{\theta}|\vec{x}) \propto P(\vec{x}|\vec{\theta})$.[4]

We will here first concentrate on models with a single parameter. The 'score' of the likelihood function is defined as $s(\theta|\vec{x}) \equiv \partial/\partial\theta \ \ln\mathcal{L}(\theta|\vec{x})$. One can show (see box below) that the expectation value of the score is in fact zero, $\langle s\rangle = 0$, *if* averaged over

[3] Note that there are various very definitions of the word 'information'. In particular 'information entropy' is a very different concept, that is only loosely connected to what we discuss here.

[4] Note that in the context of Frequentist inference, the overall normalization of the likelihood function does not play a role and drops out of all typical calculations. This is *not* the case in context of Bayesian inference, where the properties of the likelihood function matter both as function of the model parameters and as function of the data variables.

data generated for a model with parameter θ. The variance of the score is the Fisher information,

$$\mathcal{I}(\theta) \equiv \left\langle \left(\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta|\vec{x}) \right)^2 \right\rangle \ .$$

The average is here, again, taken over data sampled from $P(\vec{x}|\theta)$. One can show (under weak regularity conditions on the likelihood function, see box below) that the Fisher information can be also written in terms of the *second* derivative of the log-likelihood function,

$$\mathcal{I}(\theta) = - \left\langle \frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta|\vec{x}) \right\rangle \ .$$

This definition is somewhat easier to interpret than the previous one. It becomes evident that the Fisher information characterizes the *curvature* of the likelihood function. The width of the peak in $\ln \mathcal{L}(\theta|\vec{x})$ (as function of θ) gives hence information about the amount of information that a measurement of $\vec{x}$ can provide us about the parameter θ. A narrower peak corresponds to increased information.

Proof. Let us first show that the mean value of the score is zero. To this end (I use here f for the likelihood function and the PDF) we just have to expand the derivative of $\ln f$, and then exploit that the PDF is normalized to one for all values of θ,

$$\langle s \rangle = \int d^n x f(\vec{x}|\theta) \frac{\partial}{\partial \theta} \ln f(\vec{x}|\theta) = \int d^n x \frac{\partial}{\partial \theta} f(\vec{x}|\theta) = \frac{\partial}{\partial \theta} 1 = 0 \ .$$

Second, we show that the covariance between s and *any unbiased* estimator $\hat{\theta} = \hat{\theta}(\vec{x})$ equals one. Since the expectation value of s is zero, the covariance is simply given by

$$\mathrm{cov}(s, \hat{\theta}) \equiv \langle s\hat{\theta} \rangle - \langle s \rangle \langle \hat{\theta} \rangle = \langle s\hat{\theta} \rangle \ .$$

However, this equals one, because

$$\langle s\hat{\theta} \rangle = \int d^n x f(\vec{x}|\theta) \hat{\theta} \frac{\partial}{\partial \theta} \ln f(\vec{x}|\theta) = \frac{\partial}{\partial \theta} \int d^n x \hat{\theta} f(\vec{x}|\theta) = \frac{\partial}{\partial \theta} \theta = 1 \ ,$$

where in the last steps we used that $\langle \hat{\theta} \rangle = \theta$ for an unbiased estimator (and we had to use that this is true for all values of θ). Armed with this information, we can use the general property of the covariance,

$$\mathrm{cov}(s, \hat{\theta})^2 \leq \mathrm{var}(\hat{\theta}) \mathrm{var}(s) \ ,$$

from which follows that

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta)} \ .$$

This is the famous Cramér-Rao bound that states that the variance of *any* unbiased estimator cannot be smaller than the inverse Fisher information.

Q.E.D.

Note: The above argument can be extended to multiple parameters. In that case, the Fisher information would become a symmetric positive-definite matrix.

The size dimensionality of the matrix corresponds to the number of parameters. We will not discuss this further in the present course.

3.3 Maximum likelihood estimator

We introduced above the *likelihood function*, which equals the probability distribution function of a parametric model, interpreted as function of the model parameters. In the Frequentist context, the normalization of the likelihood function usually does not matter and drops out of all calculations.

The *maximum likelihood estimator* (MLE) is now a specific estimator that happens to maximize the likelihood function for a given measurement $\vec{x}$,

$$\hat{\theta}_{\text{MLE}} = \operatorname*{argmax}_{\theta} \mathcal{L}(\theta|\vec{x}) \ .$$

The MLE maximizes the probability of the observed data (not the 'probability of the model being true' etc). Remember that.[5]

The MLE has a number of useful properties. Although it is not in general unbiased and minimum variance (in the sense of saturating the Cramér-Rao bound), it acquires these properties in the large sample limit. In general, the MLE is

- 'consistent': Unbiased in the large sample limit.
- 'efficient': Saturates the Cramér-Rao bound (is 'minimum variance') in the large sample limit
- 'asymptotically normal': The distribution of the MLE tends to a normal distribution in the large sample limit.

Some cautionary remark: The MLE depends on the parameterization of the model. This means that if one replaces the model parameter θ with η, using some arbitrary bijection $\theta \mapsto \eta(\theta)$, then in general we have that $\hat{\eta} \neq \eta(\hat{\theta})$. In particular this implies that it is always possible to start with a perfectly well behaved unbiased MLE, and use some wicked model parameter transformation to obtain a MLE with very bad behaviour (very biased etc). Problems of this kind evaporate, usually, in the large sample regime. This is illustrated in the following example.

Example. In order to illustrate the consistency of the MLE, we consider multiple measurements of a variable that follows the exponential distribution. For a single measurement, the PDF of the exponential distribution reads

$$P(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} \ , & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \ .$$

One can show that, for a single measurement, the MLE of λ is $1/x$. Hence, the MLE estimator is here clearly biased (the mean value diverges). However, the MLE is not parametrization invariant. If we were to replace $\lambda \mapsto \beta = 1/\lambda$, we

[5] We cannot talk about the probability of model parameters, since the likelihood function is not a probability distribution for θ. It is not normalized to one, and integrating it over θ does in general not give a dimensionless number (think of the case where the data has the dimension of length, in which case $P(\vec{x}|\theta)$ would have the dimension Length^n, where n is the length of vector $\vec{x}$).

would find that $\langle\hat{\beta}\rangle = \beta$, and hence have an unbiased MLE. But for the sake of the argument, let's keep the parameter λ instead.

However, if we consider N measurements instead, the combined likelihood function is given by the product of the PDFs of the individual independent measurements

$$\mathcal{L}(\lambda|\vec{x}) = \prod_{i=1}^{N} \lambda e^{-\lambda x_i} \ .$$

The log-likelihood function acquires then the form

$$\ln \mathcal{L} = N \ln \lambda - \lambda \sum_{i=1}^{N} x_i \ .$$

From this it follows (just differentiate w.r.t. λ and set the result to zero) that the MLE is given by

$$\hat{\lambda}(\vec{x}) = \frac{N}{\sum_{i=1}^{N} x_i}$$

and one can show that the expectation value of this MLE is given by

$$\langle\hat{\lambda}\rangle = \lambda \frac{N}{N-1} \ .$$

Hence, the MLE is biased, unless we are in the limit $N \gg 1$.

4 Maximum likelihood ratio test

In the previous sections we introduced various statistical concepts: Hypothesis testing, the $\Delta\chi^2$ method, likelihood functions, estimators, and the maximum likelihood estimator. Here, we will bring these concepts together and discuss the arguably most common test statistic that you will use and encounter when performing Frequentist inference: The 'maximum likelihood ratio'.

4.1 Basic definitions

Consider the PDF $P(\vec{x}|\vec{\theta})$, where $\vec{\theta} \in \mathbb{R}^n$ are n model parameters. We define two nested composite hypotheses,

- null hypothesis: First k parameters equal to $\theta_1^*, \ldots, \theta_k^*$.
- alternative hypothesis: $\vec{\theta}$ can be anything.

Profile likelihood. We can now define the *profile likelihoods* that correspond to the null and the alternative hypothesis,[6]

$$\mathcal{L}_{\text{null}}(\theta_1^*, \ldots, \theta_k^*|\vec{x}) \equiv \max_{\theta_{k+1},\ldots,\theta_n} \mathcal{L}(\vec{\theta}|\vec{x}) = \mathcal{L}(\theta_1^*, \ldots, \theta_k^*, \hat{\theta}_{k+1}, \ldots, \hat{\theta}_n|\vec{x})$$

[6]The word 'profiling' refers in this context to the process of maximizing the function of interest w.r.t. specific parameters. Do not confuse this with 'marginalization', which we will discuss below in the context of Bayesian inference.

and

$$\mathcal{L}_{\text{alt}}(\vec{x}) = \max_{\theta_1,\ldots,\theta_n} \mathcal{L}(\vec{\theta}|\vec{x}) = \mathcal{L}(\hat{\vec{\theta}}'|\vec{x}) \; .$$

Here, $\hat{\theta}$ refers to the maximum likelihood estimation. Note that in general $\hat{\theta}_i \neq \hat{\theta}'_i$, since in one case all parameters are maximized, in the other k of the parameters are kept fixed.

Profile likelihood ratio. Finally, the profile likelihood ratio test statistic is defined as

$$\text{TS}(\theta_1^*, \ldots, \theta_k^*) = -2 \ln \frac{\mathcal{L}_{\text{null}}(\theta_1^*, \ldots, \theta_k^*|\vec{x})}{\mathcal{L}_{\text{alt}}(\vec{x})} \; .$$

Since in general and always $\mathcal{L}_{\text{null}} \leq \mathcal{L}_{\text{alt}}$ by construction, TS ≥ 0. TS is minimized (equal zero) when the maximum likelihood estimators happen to equal the fixed values, $\theta_i^* = \hat{\theta}'_i$ for $i = 1, \ldots, k$.

4.1.1 Some examples

Consider the case where the PDF is given by a multivariate Gaussian distribution. More specifically, we assume that $P(\vec{x}|\vec{\theta})$ is given by

$$P(\vec{x}|\vec{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\frac{(x_i - \mu(\vec{\theta})_i)^2}{\sigma_i^2}} \; . \tag{1}$$

In that case, one finds that (-2 times) the log-likelihood ratio is given by

$$-2 \ln \frac{\mathcal{L}_{\text{null}}}{\mathcal{L}_{\text{alt}}} = \Delta\chi^2 = \chi^2_{\text{null}} - \chi^2_{\text{alt}} \; .$$

Hence, if errors are normal distributed, we recover the $\Delta\chi^2$ formalism that we discussed in the previous section.

However, the log-likelihood ratio can also be calculated for very different PDFs. One common example is the Poisson likelihood function, where the combined PDF takes the form

$$P(\vec{x}|\vec{\theta}) = \prod_{i=1}^{N} \text{Pois}(x_i|\lambda_i(\vec{\theta})) \; . \tag{2}$$

In this case, the log-likelihood ratio becomes instead

$$-2 \ln \frac{\mathcal{L}_{\text{null}}}{\mathcal{L}_{\text{alt}}} = 2 \sum_{i=1}^{N} \left(\lambda_i^{\text{null}} - \lambda_i^{\text{alt}}\right) + x_i \ln \frac{\lambda_i^{\text{alt}}}{\lambda_i^{\text{null}}} \; .$$

For the case of the normal distribution, we discussed already in context of the $\Delta\chi^2$ test that the test statistic is under the null hypothesis distributed like a chi-squared distribution. I very important observation is now that this result also applies to the above likelihood ratio, in a large number of cases (and in general in the large sample limit). This will be discussed in the next subsection.

4.2 Wilks' theorem

Wilks' theorem is a statement about the asymptotic distribution of the log-likelihood ratio in the large sample limit. It goes as follows.

If

$$\mathrm{TS}(\theta_1^*, \ldots, \theta_k^*) = -2 \ln \frac{\mathcal{L}_{\mathrm{null}}}{\mathcal{L}_{\mathrm{alt}}} \, .$$

with

$$\mathcal{L}_{\mathrm{null}}(\vec{\theta}_1^*, ;\vec{\theta}_k^*) = \max_{\theta_{k+1},\ldots,\theta_n} \mathcal{L}(\vec{\theta}|\vec{x}) = \mathcal{L}(\theta_1^*, \ldots, \theta_k^*, \hat{\theta}_{k+1}, \ldots, \hat{\theta}_n|\vec{x})$$

and

$$\mathcal{L}_{\mathrm{alt}} = \max_{\theta_1,\ldots,\theta_n} \mathcal{L}(\vec{\theta}|\vec{x}) = \mathcal{L}(\hat{\vec{\theta}}|\vec{x})$$

and if $\vec{x}$ are random variables following the PDF corresponding to $\mathcal{L}(\theta_1^*, \ldots, \theta_k^*, \tilde{\theta}_{k+1}, \ldots, \tilde{\theta}_n|\vec{x})$, where $\tilde{\theta}_{k+1}, \ldots, \tilde{\theta}_n$ are the true values, then

$$\mathrm{TS}(\theta_1^*, \ldots, \theta_k^*|\vec{x}) \sim \chi_k^2 \, .$$

The TS follows the a chi-squared distribution with k degrees of freedom.

4.3 Limitations of Wilks' theorem

Due to its enormous convenience, Wilks' theorem plays a very central role in many aspects of Frequentist inference. However, one should not forget that it is only an approximate statement that holds under specific circumstances. If these are not fulfilled, the sampling distribution of the test statistic (whether it is the log-likelihood ratio or something else) has to be in general obtained by time-consuming Monte Carlo simulations. It is therefore important to understand, at least qualitatively, the caveats of Wilks' theorem. The three caveats that we discuss in this subsection are directly related to the various steps of the above heuristic derivation of Wilks' theorem.

4.3.1 Parameter boundaries

One of the steps in the above proof is to assume that the MLEs are normal distributed. This is obviously not possible if the parameter range over which the MLEs are allowed to vary is bounded from above or below. Since many physical quantities are required to be non-negative, this is actually a quite common situation. We show here with one specific example with a boundary that the resulting sampling distribution is not χ_k^2 distributed, even not in the large sample limit.

Consider the above example with a normal distribution, Eq. (1), assuming $N = 1$ for simplicity. Furthermore, we assume that the true value is $\tilde{\theta}_1 = 0$, and that there is a non-negativity bound, $\theta_1 \geq 0$. In that case, we can find the sampling distribution of the log-likelihood ratio, or the TS, as follows. If there were no boundary, then the MLE $\hat{\theta}_1$ would be in 50% of the cases larger than zero, and in 50% smaller than zero. In the presence the boundary however, $\hat{\theta}_1 \geq 0$ in 50% of the cases and $\hat{\theta}_1 = 0$ in the other 50%. This means that the PDF of the TS has the form

$$\frac{1}{2} P_{\chi^2_{k=1}}(t) + \frac{1}{2}\delta(t) \, ,$$

where $\delta(\cdot)$ is here the Dirac delta function. This result holds true irrespectively of the sampling size, and hence is a clear exception to Wilks' theorem.

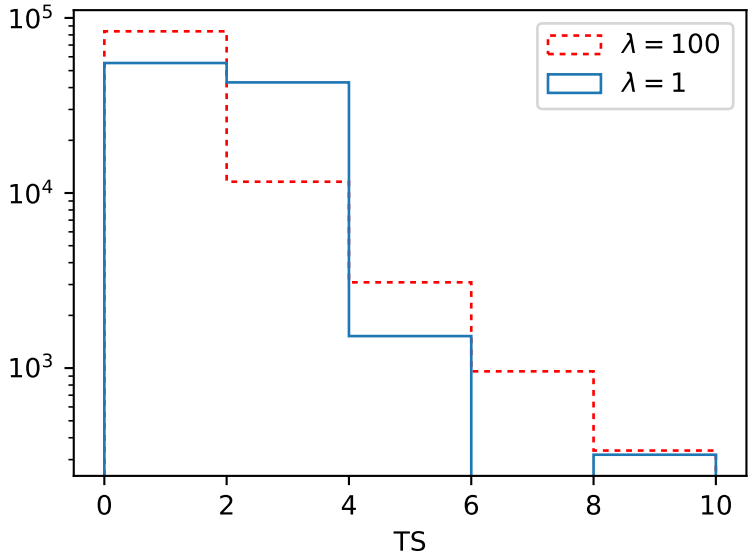


Figure 1: Histogram of TS values, using the log-likelihood ratio test statistic and a single bin Poisson likelihood, for two different values of the background-only expectation rate λ. In the case of $\lambda = 100$, the distribution becomes similar to a χ^2_k with $k = 1$, but for $\lambda = 1$ strong deviations can be observed. Due to the discreteness of the Poisson distribution, these deviations are not continuous and can introduce seemingly random fluctuations in the various bins of the histogram.

4.3.2 Sample size

One example that illustrates the impact of a small sample size is a single Poisson distribution, Eq. (2) with $N = 1$. In this case, the true sampling distribution of TS can be derived with a Monte Carlo simulation and compared with the $\chi^2_{k=1}$ distribution, which we do in Fig. 1.

4.3.3 Parameters that are irrelevant under the null hypothesis

This situation is a bit harder to understand, but not at all uncommon. Consider the case where you are searching for a peak-like signal in some data, where the exact position of the signal is unknown (think about the position of a source in the sky, or the energy of photons produced in the decay of the Higgs boson). In that case, the alternative (background + signal) hypothesis has two more parameter than the background only hypothesis. One is the signal strength/normalization, let's call this A, and one is the signal position, let's call this E_0. Now, in the null hypothesis (no signal), we have the constraint $A = 0$. *However*, in that limit the value of E_0 becomes completely irrelevant. It make no difference whether we fix E_0 or not. The difference in the number of degrees of freedom between the null and the alternative hypotheses is for that reason not clearly defined.

The above example suggests that in the presence of irrelevant parameters in the null hypothesis, the difference in the degrees of freedom is not well defined, and this should have impact on the applicability of Wilks' theorem. Indeed, the presence of irrelevant parameters is equivalent to a degenerate matrix XYZ in XYZ, which breaks

one of the important steps in the derivation of Wilks' result.

Lastly, one can conform with Monte Carlo simulations that the log-likelihood ratio in the above situation is not exactly χ^2_k distributed for any value of k. One has to resort to Monte Carlos.

4.4 Intermezzo: Trials factors

The concept of trials factor is easiest to understand with this XKCD comic, https://xkcd.com/882/. If you repeat a measurement often enough, you are doomed to find occasionally a significant result (and you know the rate at which this happens, it is given by α, compare this with the number of tests in the comic). This is a problem in cases where one searches for potential signals in many different parts of the data, or repeats a large number of similar tests. Again, a typical example would be the search for new astronomical sources, where the position is unknown.

Example: Imagine we looked at N different sky regions, and found in one of them a 3σ detection of (or hint for) a source. The corresponding p-value, the frequency of a chance observation of a 3σ excess in a given specific sky region, is $p = 1.3 \times 10^{-3}$ (since this p-value is calculated for a single isolated sky region, it is often called 'local' p-value).

What is the *overall* probability to observe a 3σ excess in *any* of the N sky regions, p_G? It turns out this is simply to calculate. The (one minus) probability p_G is given by

$$1 - p_G = (1 - p_{\text{loc}})^N \Rightarrow p_G \approx N p_{\text{loc}}$$

Example: $N = 8, p_{\text{loc}} = 1.3 \times 10^{-3} \Rightarrow p_G = 1.1 \times 10^{-2}$, which is just a 2.3σ excess.

5 Confidence regions

5.1 Motivation from hypothesis testing

Since estimators are random variables, it is a natural question to ask 'what is the probability that the estimator is correct?' In the Frequentist context we would actually ask for the frequency of the correct result. For discrete variables it is clear that the answer will be usually zero (the probability that a estimator equals *exactly* the true value is usually infinitesimally small). The way around of this problem is to consider not '*point*-estimators' but 'confidence *regions*'. It is then a reasonable question to ask 'what is the frequency with which the confidence region *covers* the true value?'.[7] This frequency is often called 'coverage'. A 95% confidence level (CL) interval should, for instance, cover the true value in 95% of the cases, etc. See Fig. 2 for an illustration.

5.2 General definition

A confidence interval (1-dim) or region (n-dim), $\mathcal{R}(\mathcal{D})$, is a 'interval estimator' that covers the true parameter value a predefined fraction of times (under repeated experiments). This fraction is called the 'confidence level' of the interval / region. Formally, this means that

$$P(\vec{\theta}_t \in \mathcal{R}(\vec{x})) = 1 - \alpha \,,$$

where $\vec{\theta}_t$ is the true parameter value, and $1 - \alpha$ the confidence level of the interval.

[7] In the Frequentist world, the word 'cover' is used to mean 'include'.

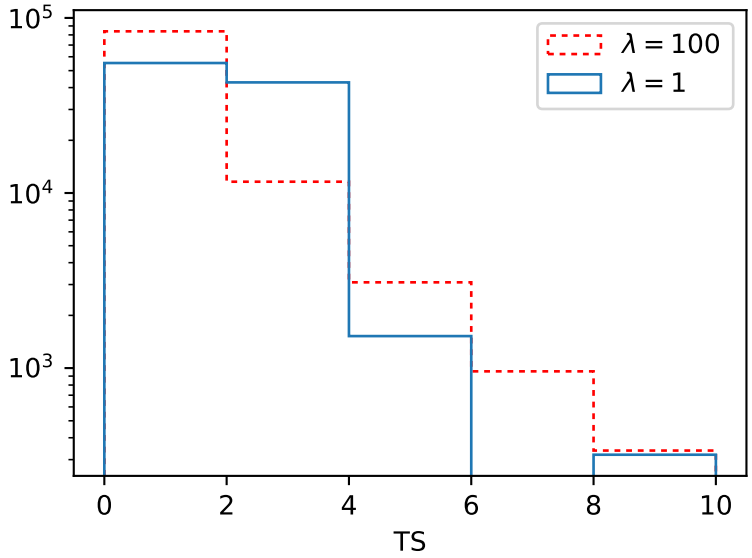


Figure 2: Histogram of TS values generated for the maximum likelihood ratio using mock data sampled from a Poisson distribution with the indicated expectation values. The larger the number of counts, the closer the distribution becomes to a $\chi^2_{k=1}$.

The above definition makes intuitively sense. However, it also includes a few pathological cases. One example would be where $\mathcal{R} = \mathbb{R}^d$ in $1-\alpha$ of the cases (we assume a d-dimensions model parameter space), and $\mathcal{R} = \emptyset$ in α of the cases. This is a proper $1-\alpha$ confidence level interval, but obviously quite useless. When defining strategies for the derivation of confidence intervals, it is hence not only important to obtain good coverage, but also to ensure that the intervals have a reasonable and useful extend in for all possible measurements.

5.3 Construction from the MLE and the MLR

Consider a model with two model parameters, θ_1 and θ_2. In the null hypothesis, we keep both parameters fixes, and we keep both parameters free in the alternative. The maximum-likelihood ratio is then given by

$$\mathrm{TS}(\theta_1, \theta_2) = -2 \ln \frac{\mathcal{L}(\theta_1, \theta_2 | \vec{x})}{\mathcal{L}(\hat{\theta}_1, \hat{\theta}_2 | \vec{x})} \,,$$

where $\hat{\theta}_i$ are the MLEs, and the θ_i the fixed values. Thanks to Wilks' theorem, we know that $\mathrm{TS} \sim \chi^2_{k=2}$, provided that θ_1 and θ_2 denote the correct values (from which the data $\vec{x}$ is sampled).

We can now define confidence regions as

$$\mathcal{R}_\alpha = \{\theta_1, \theta_2 \in \mathcal{R} | \mathrm{TS}(\theta_1, \theta_2) < 2.30\} \quad (68.3\%\mathrm{CL})$$

$$\mathcal{R}_\alpha = \{\theta_1, \theta_2 \in \mathcal{R} | \mathrm{TS}(\theta_1, \theta_2) < 6.18\} \quad (95.4\%\mathrm{CL})$$

In words, the regions include all parameter points for which the TS value is smaller than the indicated threshold. If Wilks' theorem applies, this happens for the true parameter

values exactly the predefined number of cases. For this reason, *the above regions have automatically the correct coverage, per definition.*

5.3.1 Interval constrution in practice (1-dim case)

The above definition of the confidence intervals in terms of the MLR is somewhat abstract. However, one can define a few simple clear steps to obtain the boundaries of the intervals. We will discuss this here in $d = 1$, in 1 dimensions.

1. Find MLE for θ, *i.e.*, the parameter that minimizes the quantity $\Lambda \equiv -2 \ln \mathcal{L}(\theta|\vec{x})$.[8]
2. Increase/decrease θ from the MLE, $\hat{\theta}$, until Λ changes by 1. In this way, you can obtain the upper and lower boundaries of a 68.3%CL interval for θ, $\mathcal{R} = [\theta_L, \theta_R]$.

Question: How much does Λ have to change to obtain a 95.4%CL interval?

5.3.2 Upper limits

The confidence interval from the previous example is *two-sided*, since it is constrained from two sides. However, sometimes it is interesting to construct *one-sided* intervals, where the interval extents for instance to $-\infty$ (or just 0 in the case where the parameter cannot become negative) on the left side. In that case, the right end of the interval corresponds to an *upper limit* on the parameter of interest. A typical situation in which this would be used is the null-result for searches for new particles in a particle detector. In that case, one just wants to quite an upper limit on the possible signal flux (or expected number of signal photons).

The construction of such an interval is similar to above, namely we would define a 95%CL interval as

$$\mathcal{R}_\alpha = \{\theta \geq 0 | \widetilde{\mathrm{TS}}(\theta) < 2.71\} \quad (95\%\mathrm{CL}) \,, \tag{3}$$

where we use the modified test statistic

$$\widetilde{\mathrm{TS}}(\theta) = \begin{cases} \mathrm{TS}(\theta) \,, & \text{if } \theta \geq \hat{\theta} \\ 0 \,, & \text{otherwise} \end{cases} .$$

The test statistic and the associated confidence interval are here defined such that values $\theta < \hat{\theta}$ are always included in the interval. Hence, the interval always extends all the way down to zero.

One can show that, for true values of θ that are far enough from the zero boundary, the PDF of the modified test statistic is given by $P_{\widetilde{\mathrm{TS}}}(x) = \frac{1}{2} P_{\chi^2_{k=1}}(x) + \frac{1}{2}\delta(x)$, and hence 50% a regular chi-squared, and 50% identically zero. Using this distribution, one can derive the 2.71 threshold value that is quote above in the definition of the interval Eq. (3).

[8] Note that $\chi^2 = \Lambda$ in the case of normal distributions.

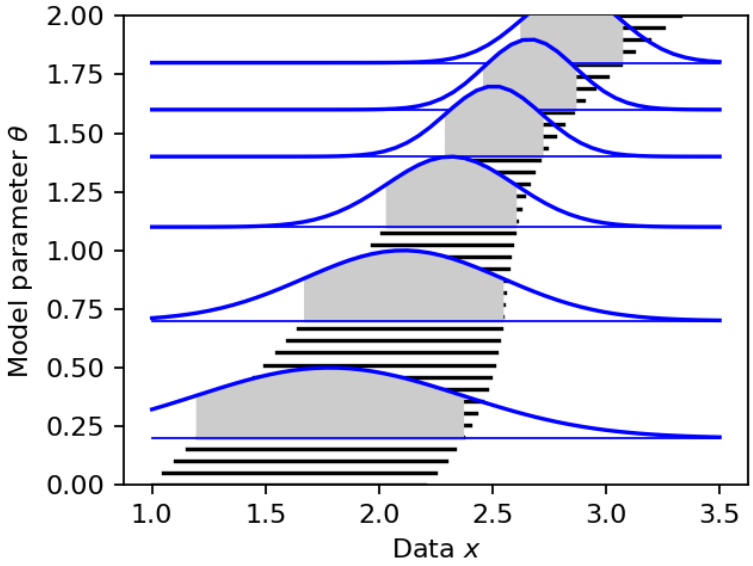


Figure 3: Illustration of the parametric PDF $P(x|\theta)$ as function of the data x, for different values of the model parameter θ. The gray regions indicate 68.3% containment.

5.4 The Neyman belt construction

The above construction breaks down when Wilks' theorem does not apply, and/or when the MLE is not available. In such cases, it is useful to resort to the *most general* construction of confidence regions. This is based on the so-called 'Neyman belt'. We will here only discuss the one-dimensional case, since it allows to visualize the problem in an intuitive way. However, the construction can be immediately applied to multivariate data and high-dimensional models as well.

Consider an arbitrary parametric PDF, $P(x|\theta)$. The entire parametric family can be illustrated like shown in Fig. 3. Based on this PDF, one can construct the Neyman belt with the following steps.

1. For each θ, select an interval $\mathcal{B}(\theta) \subset \mathbb{R}$, such that

$$\int_{\mathcal{B}(\theta)} dx\, P(x|\theta) = 0.683 \quad \text{(for 68.3\% CL)} \tag{4}$$

   For a different confidence level, $1-\alpha$, the RHS would be changed accordingly. The constructed regions are again illustrated in Fig. 3. It is evident that this procedure leads, in the present case, to a belt with a width in x direction and length in θ direction. The shape of the belt is largely arbitrary, as long as the above integral condition holds. However, usually one would start 'filling up the integral' by adding regions in x where the PDF has the largest values.[9]

2. The previous step already constructed the Neyman belt. Step two is now to simply read off the resulting confidence interval, given some measurement x. The procedure is illustrated in Fig. 4, where the blue line indicates the measured value

[9] An exception is the construction of upper or lower limits, where one would start the regions from $x = \infty$ or $x = -\infty$, respectively. This will not be further discussed here.

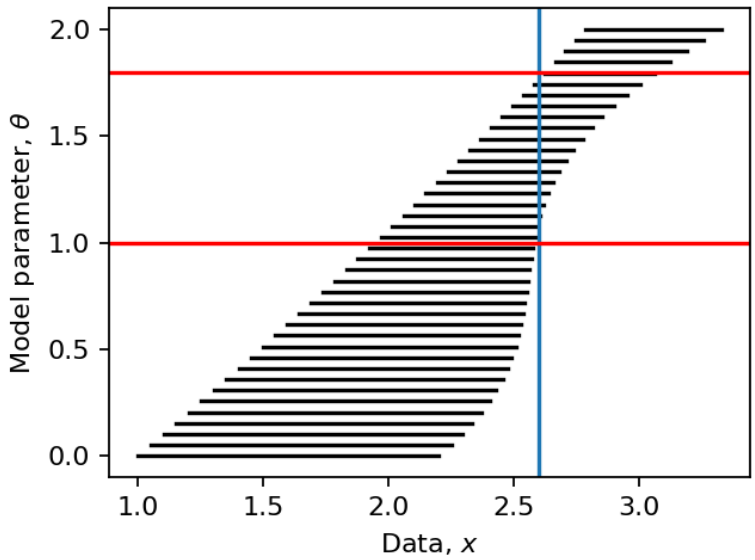


Figure 4: Illustration of Neyman belt construction. The blue vertical line indicates the measured value. The red horizontal lines indicate the resulting boundaries of the confidence interval.

of x. The confidence interval associated with this value x (shown by the red lines) is now given by all values of θ where the blue line overlaps with the Neyman belt.

Following the above example, the confidence region corresponding to the measured value x is defined as

$$\mathcal{R}(x) = \{\theta \in \mathbb{R} | x \in \mathcal{B}(\theta)\} \ ,$$

it contains all values of θ for which the corresponding region $\mathcal{B}(\theta)$ includes the value x. That's it.

The crucial point is now, although it might not be obvious on first sight, that *per definition* this interval has exactly the correct coverage, $P(\theta_t \in \mathcal{R}(x)) = 0.683$. To understand why this is the case, remember that coverage of a confidence interval means that the true value is included in the interval in a certain fraction of cases. Let's assume the true value is $\theta = 1$. In this case, x will lie in the range $\mathcal{B}(1) = [1.9, 2.6]$ in about 68.3% of the cases. Each time x is within $\mathcal{B}(1)$, $\theta = 1$ is within $\mathcal{R}(x)$ (per construction). Hence, the coverage for $\theta = 1$ is exactly 68.3%. The same argument can be made for any true value of θ, which means that confidence intervals constructed via the Neyman belt have proper coverage behavior.

There are a number of cases where the above construction is used in practice, and I will list here a few.

- Construction of confidence regions when Poisson noise is dominating (only few events etc). In that case, the integral in Eq. (4) is replaced by a sum over the number of measured events. That means that in general no region can be defined that *exactly* integrates to the target confidence level. Instead a region is selected that has at least the target confidence level. This leads to confidence intervals

that slightly over-cover (have a too large coverage), which can however not be avoided for discrete distributions.

- In scenarios where no MLE exists (e.g. because the PDF is flat in the relevant region as function of θ), or where the MLE is strongly biased, or close to boundaries. In that case, a construction can be still made, and would lead to the correct coverage behavior.

6 Bayesian Inference

After discussing some central aspects of Frequentist inference in the previous sections, we will now turn by 90° and introduce the very basics of Bayesian inference. Both statistics schools have their own pros and cons, and it can be somewhat confusing to keep the concepts apart in the beginning. In particular, many terms are used in both schools, but with slightly different meaning. It is hence useful to first think about Bayesian inference and Frequentist inference completely separately. Once you got the hang if them and some intuition, you can start asking where are the contact points and how they compare.

6.1 Bayes' theorem

The center of Bayesian inference is *Bayes' theorem*. Bayes' theorem is nothing else the the rule of conditional probability, however with a specific interpretation of the prepositions. It reads

$$P(H|\mathcal{D}, I) = \frac{P(\mathcal{D}|H, I)P(H, I)}{P(\mathcal{D}, \mathcal{I})} \,. \tag{5}$$

Here, I indicates any external knowledge, and it reminds us that there are always a large number of implicit and explicit assumptions made when defining an hypothesis. We will drop I in the following to not clutter the notation.

The meaning of the individual factors in Eq. (5) is as follows.

- $P(H|\mathcal{D})$: Posterior probability of proposition H, given data $\mathcal{D}$.
- $P(\mathcal{D}|H)$: Likelihood function, probability of measuring data $\mathcal{D}$ given hypothesis H.
- $P(H)$: Prior probability of hypothesis H, *before* the data $\mathcal{D}$ is taken into account.
- $P(\mathcal{D}) \equiv \int \mathrm{d}H'\, P(\mathcal{D}|H')\, P(H')$: Global likelihood / model evidence. The main effect of this factor is to ensure the proper normalization of the posterior probability distribution, which has to sum to one if summed over hypotheses.

We will illustrate the application of Bayes' theorem in two simple examples.

Example I

Question: What is the temperature outside?

1. Starting point: somebody told you an hour ago that it is 'around zero'. You trust the ability of that person to feel the temperature right up to, say, 3°C. In that case, a justifiable prior belief in the temperature is[10]

$$P(T) = \mathcal{N}(T|0, \sigma = 3°\text{C})$$

2. New data: You see through the window that it rains outside. Let's suppose that it only can rain at a temperature above 0°C. In that case, you improper likelihood function is given by

$$P(\text{rains}|T) = \begin{cases} \alpha\,, & \text{if } T > 0 \\ 0 & \text{otherwise} \end{cases} = \alpha\theta_H(T) \tag{6}$$

Here, θ_H is the Heaviside step function, and α is some (here) arbitrary value larger than zero and smaller than one, which indicates the probability that it rains given a specific temperature. We will see that it drops out of the calculation.

Given the above likelihood function, we can now update our initial belief for the temperature, using Bayes' theorem. We have to calculate

$$P(T|\text{rain}) = \frac{\alpha\theta_H(T)\mathcal{N}(T|0, 3°\text{C})}{\int dT\alpha\theta_H(T)\mathcal{N}(T|0, 3°\text{C})} = 2\theta_H(T)\mathcal{N}(T|0, 3°\text{C})\,. \tag{7}$$

The final answer after taking the rain into account is hence a posterior distribution with is zero at $T < 0$, and follows the initial Guassian for $T > 0$. The factor of two makes sure that this truncated PDF is still correctly normalized to one.

3. Assume now you find a thermometer outside. It shows that the temperature is $T = 1.3 \pm 0.1°\text{C}$. It is instructive to consider the impact of this new information. The corresponding likelihood function can be assumed to have the form $\mathcal{N}(T|1.3°\text{C}, 0.1°\text{C})$.

$$\begin{aligned} P(T|\text{rain}, \text{therm}) &= \frac{\mathcal{N}(T|1.3°\text{C}, 0.1°\text{C}) \cdot 2\theta_H(T)\mathcal{N}(T|0, 3°\text{C})}{\int dT\mathcal{N}(T|1.3°\text{C}, 0.1°\text{C}) \cdot 2\theta_H(T)\mathcal{N}(T|0, 3°\text{C})} \\ &\simeq \frac{\mathcal{N}(T|1.3°\text{C}, 0.1°\text{C}) \cdot \mathcal{N}(1.3°\text{C}|0, 3°\text{C})}{\mathcal{N}(1.3°\text{C}|0, 3°\text{C})} = \mathcal{N}(T|1.3°\text{C}, 0.1°\text{C})\,. \end{aligned} \tag{8}$$

Here, we used the fact that the likelihood function that corresponds to the newly added information is significantly more peaked than the prior that we obtained (as posterior) during the previous step. This can be used to approximately evaluate the integral. The final result is that the new very precise information overrides our previous knowledge about the outside temperature.

Exercise: Think about what would have happened if the thermometer would have shown a negative temperature. How does this change if you allow for a tiny probability of rain when it is below zero degree?

[10] Note that would have absolutely no clue what temperature is outside, you would probably want to use a flat prior instead that extends over a large range. Flat priors will be further discussed below.

6.1.1 Example II & Bayesian model comparison

The following example is motivated by a classical example for Frequentist inference that you can find online.[11] Imagine your aunt claims that she can recognize whether you put first milk and then sugar in her tea, or first sugar and second milk. You do not belief her and propose an experiment. Let's assume you test her 20 times, and she guesses right 15 times. What does this imply for the credibility of her claim?

As a Frequentist, you are bound to do a hypothesis test. What you hence would do is to calculate the p-value for guessing 15 out of 20 cups correctly by chance. The p-value is given by

$$p = \sum_{k=15}^{2} 0\mathrm{Bin}(k|n = 20, p = 0.5) = 2.1\% \ .$$

You hence would conclude that you can reject the null hypothesis (aunt is delusional) at $> 95\%$CL.

On the other hand, as a Bayesian, you would calculate the posterior probability.

$$P(\mathrm{T}|k = 15) = \frac{P(k = 15|\mathrm{T})P(\mathrm{T})}{P(k = 15|\mathrm{T})P(\mathrm{T}) + P(k = 15|\mathrm{F})P(\mathrm{F})} \ ,$$

where $P(\mathrm{T})$ is your prior belief that your aunt is right, and $P(\mathrm{F})$ that your aunt is wrong. In order to make a decision about the statement from your aunt, it is now conventional to look at the so-called *Bayes factor*, which is defined as the ratio of likelihoods corresponding to the hypotheses of interest. This is called *Bayesian model comparison* and the Bayesian alternative to classical hypothesis testing. In the present case, the Bayes factor is is given by

$$K = \frac{P(k = 15|\mathrm{T})}{P(k = 15|\mathrm{F})}$$

Note that the Bayes factor corresponds to the posterior ratio if the prior probabilities of the two alternative hypotheses are equal.

In the aunt example, we would find $K \simeq 13$. How should this be interpreted? In general, a value of $K > 1$ means that data supports the hypothesis T. Harold Jefferys proposed an interpretation of the Bayesian evidence. Note that this is empirically calibrated, and should be always interpreted in the context of the specific problem at hand.

Jeffreys' scale is as follows

K	Evidence
< 1	negative
1–3	barley worth mentioning
3–10	substantial
10–30	strong
30–100	very strong
> 100	decisive

In the present example, the support for the claims of your aunt are hence 'strong'. However, if you take into account that your prior belief in your aunt's claims is (probably) not very high, say with an odds ratio of $1 : 9$, the resulting posterior ratio would be around one, leaving the test inconclusive.

[11] https://en.wikipedia.org/wiki/Lady_tasting_tea

Comparison

Let us briefly summarize some of the main differences between the Bayesian and the Frequentist approach to statistical inference.

Bayesian	**Frequentist**
Works for all definitions of probability.	Works for frequencies only.
Priors must be specified.	No concept of priors.
Results are always prior dependent, although the dependence can be very weak if the data is strong enough.	No prior-dependence of results.
No estimate of the rate (or probability) of type I error.	Frequency of type I errors (false positives) is known by construction.

6.2 Ockham's razor

One of the many interesting aspect of Bayesian model comparison is that it automatically penalizes additional (fine-tuned, we will see below what that means) model parameters. If you have two models that you compare with the data, and one has more parameters, the more extended model tends to give a better fit to the data. Bringing this to the extreme, you could introduce as many model parameters as you have data points, and then completely over-fit the data. In Freuqentist hypothesis testing, this problem is not taken into account (at least not out of the box, see the Akaike information criterion). In the Bayesian framework, it is automatically. We will see this in the following example.

The entire mechanism is sometimes referred to as 'Ockham's razor', since it implements in some way a problem solving strategy that is attributed to William of Ockham (ca. 1287–1347, English Franciscan). One formulation of this statement is that 'Entities must not be multiplied beyond necessity.' More complex models should be disfavored, unless they are strongly supported by data.

Assume you compare hypothesis $\mathcal{M}_0$ ($\theta = 0$) with hypothesis $\mathcal{M}_1$ ($\theta \geq 0$). Then the posterior probability for model 1 is

$$P(\mathcal{M}_1|\vec{x}) = \frac{\int_0^\infty d\theta\, P(\vec{x}|\theta)P(\theta|\mathcal{M}_1)P(\mathcal{M}_1)}{P(\vec{x})} \,,$$

where we have used the likelihood $P(\vec{x}|\theta)$, the prior for the model parameter θ in case of model 1, $P(\theta|\mathcal{M}_1)$, and the prior belief for model 1, $P(\mathcal{M}_1)$. Furthermore, the posterior probability for model 0 is

$$P(\mathcal{M}_0|\vec{x}) = \frac{P(\vec{x}|0)P(\mathcal{M}_0)}{P(\vec{x})} \,,$$

where the individual factors are defined analogously to model 1.

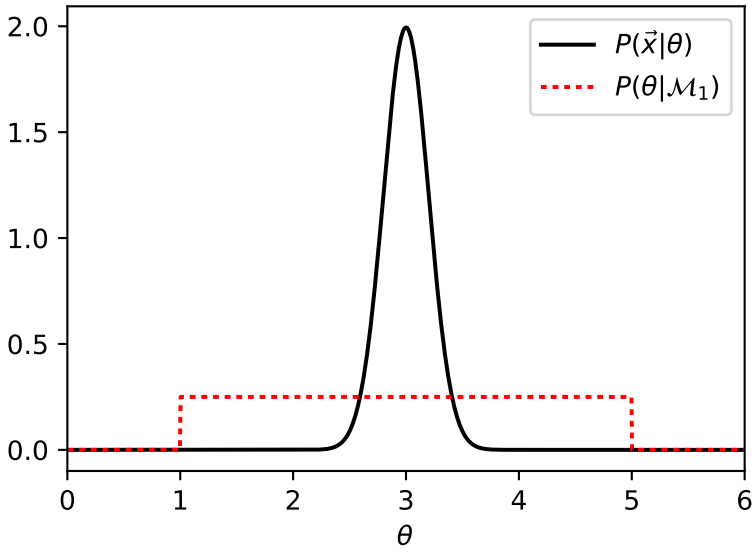


Figure 5: Illustration of the likelihood function and the prior for the model parameter θ. The prior range is here $\Delta\theta = 4$, and the width of the likelihood function approximately $\delta\theta = \frac{1}{2}$ (the peak height is $1/\delta\theta = 2$). Integrating over the product of both functions gives $\delta\theta/\Delta\theta$. Remember that the likelihood function as function of θ is *not* normalized to one.

We can now calculate the Bayes factor, which is simply given by

$$K = \frac{\int_0^\infty d\theta\, P(\vec{x}|\theta) P(\theta|\mathcal{M}_1)}{P(\vec{x}|0)} \,.$$

Assume now that the prior for θ is flat, and spans a range $\Delta\theta$. Since the prior must be normalized to one, this means that either $P(\theta|\mathcal{M}_1) = 0$ outside the prior range, or $P(\theta|\mathcal{M}_1) = 1/\Delta\theta$ inside the prior range. Furthermore, we assume that the measurement of θ is reasonably precise, so that the width of the likelihood function $P(\vec{x}|\theta)$ is approximately $\delta\theta$. The peak height of the likelihood function is in this case of the order $1/\delta\theta$. This situation is illustrated in Fig. 5.

Using the above assumptions, we can now evaluate the integral in K approximately, which leads to

$$K \sim \frac{P(\vec{x}|\hat{\theta})}{P(\vec{x}|0)} \frac{\delta\theta}{\Delta\theta} \,,$$

where $\hat{\theta}$ is the MLE. (Note that this is really only an order of magnitude estimate; the exact prefactor will depend on the shape of the likelihood function.) We find that the Bayesian evidence is now of the order of the classical maximum likelihood ratio, times a factor that becomes small in the limit $\delta\theta \ll \Delta\theta$. If you now assume that the likelihood ratio is only marginally larger than one, the additional factor can significantly reduce the Bayes factor, favoring the model with less parameters. This effect is particularly strong if the additional parameter is 'find-tuned', i.e. has to acquire a very specific value to fit the data, although the prior range of the parameter is large. The penalty factor can only be overcome if the data strongly prefers the additional parameter.

7 Credible intervals and typical priors

7.1 Credible intervals

The result of Bayes' theorem is the posterior distribution function, $P(\vec{\theta}|\vec{x})$. By construction, it is normalized to one,

$$\int d^n\theta \, P(\vec{\theta}|\vec{x}) = 1 \; .$$

where n is the dimensionality of the model parameter space. The full posterior distribution function contains all the information that were obtained as result of the application of Bayes' theorem. Nevertheless, the result of Bayesian inference is often summarized in terms of a few quantities that can be derived from the posterior. We will here discuss a few of them.

First, if there is only one model parameter of interest, say θ_i, we can consider the *marginal posterior*, which is given by

$$P(\theta_i|\vec{x}) = \int \prod_{j \neq i} d\theta_j P(\vec{\theta}|\vec{x}) \; .$$

This is now a one-dimensional posterior distribution. A 68.3% credible interval is now given by any interval $\mathcal{R}$ that satisfies

$$\int d\theta_i P(\theta_i|\vec{x}) = 0.683 \; .$$

This works analogously in the multivariate case (in that case we do not obtain intervals but regions). The form of $\mathcal{R}$ is arbitrary, but a conventional approach is to define the credible interval as the region with the highest posterior density,

$$\mathcal{R} = \{\theta_i \in \mathbb{R} | P(\theta_i|\vec{x}) > P_{\rm cut}\}$$

where $P_{\rm cut}$ is adjusted such that that the credible interval includes the correct amount of probability.

7.2 Dimensional analysis

A very useful way to test the correctness of equations is to check that the units work out. We will now do this for Bayes theorem, using a specific example.

Consider the observation of a number of photons c from a distant source with luminosity L (units $\mathrm{ph\,s^{-1}}$) and distance D (units cm). The expected number of measured photons is then given by

$$\lambda = \frac{L}{4\pi D^2} A_{\rm eff} T_{\rm obs} \; ,$$

where $A_{\rm eff}$ is the effective area (units $\mathrm{cm^2}$) of the detector and $T_{\rm obs}$ the observation time (units s). The expected number of events, λ, is unit-less.

Let's now suppose we are looking at a 'standard candle', and we hence know the value of L for the particular source of interest. We are interested in inferring the distance D. The posterior distribution of D is then given by

$$P(D|k) = \frac{P(k|D)P(D)}{P(k)}$$

where k is the number of measured events, and $P(D)$ the prior for the source distance. Now, the posterior and the prior for D have units m^{-1}, whereas the likelihood function and the marginal likelihood, $P(k)$, are dimension-less.

Already the various units of the functions inform you know what you can do with the function and what not. For instance, the posterior and priors for D can be integrated over distance, yielding dimensionless number. On the other hand, integrating over the likelihood function would *not* give a dimensionless number (instead you obtain unit m), which obviously means that a probabilistic interpretation of the outcome of this integral would not make much sense.

7.3 Upper and lower limits

Above, we discussed the maximum posterior density credible interval. In a similar way, one can also construct intervals that correspond to upper and lower limits on a model parameter. The construction is simple: Fix one of the integration boundaries to $-\infty$ (or $+\infty$ in the case of a lower limit), and adjust the other boundary such that the required credible level is reproduced. We will here discuss this in the context of a Poisson process, and contrast it with the results that we would obtain from a Frequentist analysis.

Consider the posterior for the expectation value of a Poisson process, given some number of measured events k. It is given by

$$P(\lambda|k) = \frac{P(k|\lambda)P(\lambda)}{P(k)} \ ,$$

where $P(k) = \int_0^\infty d\lambda P(k|\lambda)P(\lambda)$ is the marginal likelihood, $P(\lambda)$ the prior for the expectation value λ and $P(k|\lambda)$ the PDF of a Poisson process. The 95% credible interval upper limit, $\lambda_{\mathrm{UL,B}}$, is now implicitly given by

$$\int_0^{\lambda_{\mathrm{UL,B}}} d\lambda P(\lambda|k) = 0.95 \ .$$

On the other hand, Frequentist upper limit (95%CL), would be given by

$$\sum_{k'=0}^{k} P(k|\lambda_{\mathrm{UL,F}}) = 0.05 \ ,$$

namely the largest value of λ for which the probability to measure k events equals 5%.[12]

In Fig. 6, you find upper limits based on the Frequentist and the Bayesian construction compared. The Bayesian upper limits depend, not suprisingly, on the assumed prior for the expectation value λ. The Frequentist upper limit is completely independent of any priors (you can convince yourself that it is also completely invariant under any reasonable re-parametrization of the problem). Interestingly, however, and *for the particular case at hand*, we find that the Bayesian upper limits corresponding to flat

[12] If you find this surprising, please think about how to draw the Neyman belt in this particular case. Remember that for discrete distributions, exact coverage is in general not possible, and you will end up with intervals that for most values of λ slightly over-cover. In the present case, since you are interested in upper limits on λ, the intervals in data space should range from some threshold value $k_{\mathrm{th}}(\lambda)$ to $k = \infty$.

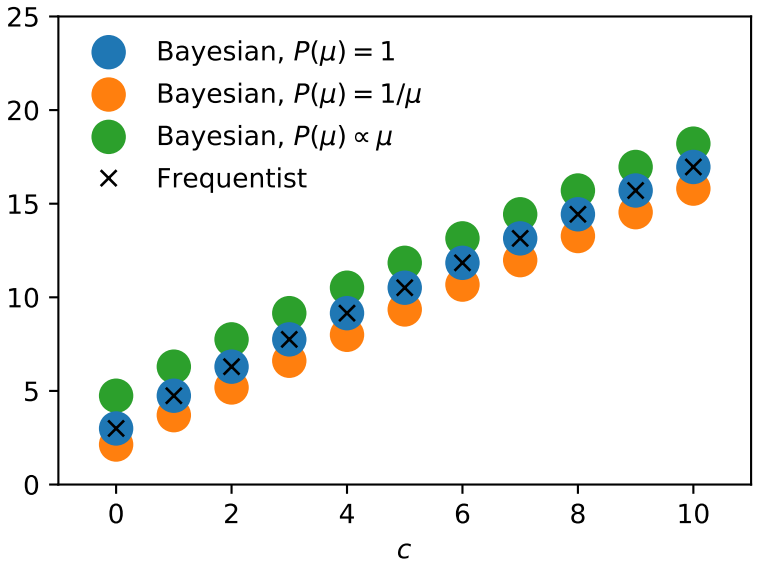


Figure 6: Upper limits on the expectation value of a Poisson process, μ, derived using the Frequentist method (build around the Neyman belt construction), or using credible intervals assuming different priors. Limits are shown as function of the number of measured events, c.

priors exactly coincides with the Frequentist upper limit. This is a property of the specific form of the Poisson distribution. However, we also see that the Bayesian limits actually *do* depend on the adopted prior, as they become stronger (weaker) when more weight is given to lower (higher) values of λ by adopting different priors.

7.4 Priors

7.4.1 Informative priors

First a few words about informative priors. These apply to cases where indeed prior information about a quantity of interest is available. Sometimes the information is specific enough to directly imply a functional form for the prior (for instance in cases where you have a measurement and an associated error bar, and reasons to belief that the error is approximately Gaussian distributed). In many other cases, you need to make some educated guess about the most plausible functional form of the prior (e.g., a flat prior if you know that the parameter must be in a specific reasonably narrow range). However, in the vast majority of cases, you will have at least some parameters for which no informative prior exists (otherwise, you would likely not discover anything new with your data). Being able to select reasonable *non-informative* priors is hence an important and non-trivial task.

7.4.2 Non-informative priors

Non-informative priors apply to situations where you are interested in measuring parameters on which you have little to know prior knowledge. What functional form for the prior should be used in such a case? Any prior that you could chose makes un-

avoidably some statement about which parameter values you find more plausible and which less. The goal must hence be to select a prior that (a) accounts best for our ignorance (under some measure), and (b) that minimizes the prior-dependence of the inference result. We will here concentrate on point (a).

Let us start with a simple example, a set of discrete hypotheses, H_i. If their number is finite, say $i = 1, \ldots, N$, then an obvious non-informative prior (motivated by what is sometimes called *principle of indifference*, where one strives at giving each element the same weight) is to assign a prior probability of $p(H_i) = 1/N$ to each of the hypotheses.

This plausible approach finds its limitations when considering a discrete but infinite number of possibilities, since the probability per item approaches zero. This hindrance can be circumvented if one extends the definition of priors to so-called *improper priors*, where the normalization requirement is dropped. Such improper priors are often still very useful in practice, since they lead – after application of Bayes' theorem – to properly normalized posteriors, provided the likelihood function is constraining enough and leads to convergence when calculating the marginal likelihood.

The situation becomes even more problematic when priors for continuous parameters are considered. In this case, the specific functional form of the prior is parametrization dependent.

Examples. - Even though there are all these caveats, here a few exmaples - Jump then to a general principle - Lin prior - Log prior

Maximum entropy principal. A powerful way to derive the form of non-informative priors from first principles is the so-called *maximum entropy prior*. In short, it is the prior with maximum information entropy, given various external constraints.

For discrete distributions, information entropy is defined as

$$S = -\sum_{i=1}^{N} p_i \ln p_i \ .$$

where p_i the the probability for outcome i, and $\sum_{i=1}^{N} p_i = 1$. If you accept $\ln 1/p_i$ as a measure for the self-information or 'surprise' of event i, then S is the average self-information or expected surprise that each drawing from the underlying distribution would provide (there is much more to say about this, but we keep it short here and concentrate on a few practical examples). In particular, if the outcome is completely clear ($p_i = 1$ for one of the i), then the event has no information value, and $S = 0$.

We can now consider what kind of probability distribution functions emerge depending on the constraints that we apply. In the simplest case, the only constraint is that the probabilities should sum up to one,

$$\sum_{i=1}^{N} p_i = 1 \ .$$

We can now use the Lagrange multiplier method to implement these constraints while maximizing S. To this end, we have to consider

$$S(p_i, \beta) = -\sum_{i=1}^{N} p_i \ln p_i + \beta \left(1 - \sum_i p_i\right) \ .$$

Taking the derivative w.r.t. p_i gives

$$\frac{\partial S}{\partial p_i} = -1 - \ln p_i - \beta$$

and w.r.t. β gives

$$\frac{\partial S}{\partial \beta} = 1 - \sum_i p_i \; .$$

Requiring that all derivatives vanish implies then that $p_1 = p_2 = \cdots = p_N = 1/N$, which is the flat distribution motivated by the 'principle of indifference above'.

If we instead require that besides the normalization also the mean value is fixed

$$\sum_{i=1}^{N} p_i = 1 \quad \text{and} \quad \sum_{i=1}^{N} x_i p_i = \mu \; ,$$

(to this end, some values x_i have to be introduced) a similar calculation would find that the corresponding probability distribution function is the discrete version of the exponential distribution

$$p_i = A e^{-\lambda x_i} \; ,$$

where A and λ are chosen to fulfill the above constraints.

Lastly, if we require that besides the normalization also the mean value *and the variance* are constrained, namely

$$\sum_{i=1}^{N} p_i = 1 \; , \quad \sum_{i=1}^{N} x_i p_i = \mu \quad \text{and} \quad \sum_{i=1}^{N} (x_i - \mu)^2 p_i = \sigma^2 \; ,$$

then we obtain back the normal distribution, with appropriately chosen mean and variance.

Jefferys' prior. The last prior that I will mention is a specific prior that is related to *Fisher information*, which we discussed in earlier sections of the lecture in context of minimum-variance estimators. Interestingly, Fisher information can be used to define a parameterization-independent prior. In one dimensions, it is given by

$$P(\theta) \propto \sqrt{\mathcal{I}(\theta)} \; ,$$

namely the square-root of the Fisher information. You can convince yourself that this prior is independent under a redefinition of parameters. It emphasizes regions of the parameter space which can be well tested by the considered experiment.

8 Sampling techniques

8.1 Basics

The goal of Bayesian inference is to derive the posterior PDF. However, only in very few cases this is possible in a closed analytical form. Usually, numerical approximations are required to obtain results for the posterior PDF. There are *many* possible approaches towards estimating the posterior. We will here concentrate on methods that are related

to sampling from the posterior PDF. In order to see that sampling from the posterior is indeed useful, consider the following integral.

$$\mu = \int_{\mathcal{D}} d^n x \, f(\vec{x}) p(\vec{x}) \approx \frac{1}{N} \sum_{i=1}^{N} f(\vec{x}_i) \; .$$

In the last step, we replaced the integral by a sum over random samples from $p(\vec{x})$. Note that the mean value of each of the terms in this sum equals μ, and hence the RHS provides an unbiased estimator for μ.

The sample of points $\vec{x}_i$ can then be used to calculate the maximum posterior density, the mean, mode and median of the distribution, confidence intervals etc. We will here discuss first a few basic sampling techniques, and then one of the most popular sampling algorithms used in Bayesian inference, the Metropolis-Hastings algorithm.

8.2 Simple sampling techniques

8.2.1 Inverse transform sampling

In many simple cases (in particular if the PDF is given by an analytic function of some form) it is possible to derive the cumulative distribution function, CDF. In 1-dim, it is given by

$$\mathrm{CDF}(x) = \int_{-\infty}^{x} dx' \; \mathrm{PDF}(x') \; .$$

If this is possible, sampling from the PDF is simple. First, sample $z \sim \mathrm{Uniform}(0, 1)$ from the the uniform distribution with boundaries $[0, 1]$. Second, map this onto $x = \mathrm{CDF}^{-1}(z)$. Here, we used the inverse of the CDF, which is simple to obtain numerically (the CDF is a monotonic function). The resulting random variable x is is then distributed like the target PDF,

$$x \sim \mathrm{PDF} \; .$$

The huge advantage of this technique is that *any* randomly drawn value of z results in a valid sample x from the PDF. Hence, this algorithm is rather efficient and fast. This is not the case for the below algorithms, that can however be applied more generally.

8.2.2 Rejection sampling

A completely different technique, that works for any PDF that can be point-wise evaluated, is called rejection sampling. The only required input that one needs is the maximum of the PDF, $p_{\max} = \max_x \mathrm{PDF}(x)$, which either has to be guessed or evaluated somehow. Then, the steps are as follow.s

1. Draw $w \sim \mathrm{Uniform}(0, p_{\max})$
2. Draw $\vec{x} \sim \mathrm{Uniform}(\mathcal{D})$, where $\mathcal{D}$ refers to the domain over which the PDF is defined.
3. Evaluate $p = \mathrm{PDF}(\vec{x})$.
4. If $p > w$, accept $\vec{x}$ as a sample from the PDF, otherwise ignore it and start again from one.

The list of samples $\vec{x}$ that are generated in this way are effectively sampling the target PDF. To see how this works, consider a 1-dim example. It is evident that the density of accepted points around the value x will be proportional to $p(x)$. Each of the dots corresponds to a (x, w) pair, however, only the x values that are shown as crosses are accepted as samples from the PDF.

An important aspect for the efficiency of the above algorithm is the acceptance rate. If most proposal points $\vec{x}$ are ultimately rejected, it might take a long time before a large number of samples is generated. Interestingly, in 1-dim, the acceptance rate corresponds simply to the area that is covered by the PDF within the sampled rectangle (spanned by the data domain $\mathcal{D}$ and the height $[0, p_{\max}]$). This area is particularly small if the PDF is very peaked (since then the $p_{\max}$, and with it the area, has to extend to high values). For these and similar functions, the algorithm might be hence quite inefficient. We will discuss in the next subsection an algorithm that can better deal with such situations.

8.3 Intermezzo: Markov Chains

Before we can finally introduce the last sampling algorithm that we will discuss in this course, we have to take a side-tour and discuss the concept of 'Markov Chains'.

A Markov Chain is a sequence of random variables, $X_1, X_2, \ldots, X_N$, with the special property that that each subsequent value is *only* dependent conditioned on the previous value. This means that

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{n+1} = x | X_n = x_n) \; .$$

This is called the 'Markov property'. The next step depends only on the current value, not on the history how the sequence arrived at the current value.

We need one more definition before we can proceed. A Markov Chain is called **reversible**, *if there exists a probability distribution function* $\Pi(x)$ *such that*

$$P(y|x)\Pi(x) = P(x|y)\Pi(y) \; .$$

In this case, $\Pi(x)$ is called the equilibrium distribution. The reason is the following: Imagine that the Markov Chain corresponds to an object that can acquire different states and has a given probability to jump from one state to another. The sequence of states would then (if the Markov property holds) a Markov Chain. Imagine now that there is a large ensemble of these objects, with states distributed – at a given moment in time – like the equilibrium distribution $\Pi(x)$. The dynamics of the Markov Chain will then not alter this equilibrium distribution, and the system will remain in equilibrium (this is referred to as 'detailed balance'). Furthermore, out-of-equilibrium ensembles will be typically driven back to equilibrium. Finally, instead of thinking about an ensemble of a large number of objects, one can also just consider the dynamics of a single object (or Markov Chain) for a sufficiently long time. Its occupation of different states, averaged over a sufficiently long time, will again start to resemble $\Pi(x)$. It will become clear in the next subsection why this is important.

8.4 Markov Chain Monte Carlo

Following the discussion of the previous subsection, our goal is now to write an algorithm for a reversible Markov Chain with an equilibrium distribution that equals the

PDF that we would like to sample. If we manage to do that, the Markov Chain will, if it runs for a sufficiently long time, act as a random sample from the PDF (I use the word *act*, since subsequent elements in a Markov Chain are correlated, which would be not the case in a truly random sample; more below).

One of the most common MCMC algorithms is called Metropolis-Hastings. Its individual steps to sample from the PDF $\pi(\vec{x})$ are as follows.

1. Generate some initial value $\vec{x}_1$. Set the counter to $i = 1$.
2. Randomly pick some new proposal state z according to the symmetric proposal distribution $g(\vec{z}|\vec{x}_i)$.
3. Accept the newly proposed state z with an acceptance probability that is given by
$$A(\vec{x}'|\vec{x}_i) = \begin{cases} 1\,, & \text{if } \pi(\vec{z}) > \pi(\vec{x}_i) \\ \frac{\pi(\vec{z})}{\pi(\vec{x}_i)} & \text{otherwise} \end{cases}$$
4. If the new point is accepted, then $\vec{x}_{i+1} = \vec{z}$, otherwise $\vec{x}_{i+1} = \vec{x}_i$.
5. Increase the counter i by one, and repeat from step two.

The algorithm should stop once a sufficiently long chain is generated. Furthermore, we had to introduce some proposal function g which describes how far subsequent points can jump. It requires some further insight in the dynamics of the chains to understand what choices are here the most reasonable, and lead to the highest acceptance ratio and fastest convergence of the chain. This will be discussed below.

Note that the above algorithm implies that the jump probability is given by

$$P(\vec{x}_{i+1}|\vec{x}_i) = g(\vec{x}_{i+1}|\vec{x}_i)A(\vec{x}_{i+1}|\vec{x}_i) + B(\vec{x}_i)\delta(\vec{x}_{i+1} - \vec{x}_i)\,,$$

where the last term corresponds to cases where the new point is not accepted and ensure that the jump probability is normalized to one. Now, this implies that

$$\frac{P(\vec{x}|\vec{y})}{P(\vec{y}|\vec{x})} = \frac{\pi(\vec{x})}{\pi(\vec{y})}\,.$$

This is however exactly the criterion for a reversible Markov Chain, with $\pi(\vec{x})$ as the equilibrium distribution, which is precisely what we wanted.