Statistics for little penguins

C.A. Argüelles[1] and G.H. Collin[1]

[1]Massachusetts Institute of Technology, Cambridge, MA 02139, USA

1 Introduction and objective of this note

This note is not meant to be an exhaustive review on the topic, for that please see the referred literature at the end of this note. This note is intended to be minimal, but useful. This attempts to cover three basic skills: parameter estimation, model selection, and goodness-of-fit in the bayesian and frequentist frameworks.

2 What is probability?

2.1 Definition

Consider a coin. The probability of landing on heads or tails is 50% each. If we give C a value that depends on the coin toss, then C is called a random variable. Then we can write the probability of heads as

$$\mathbb{P}(C = \text{heads}) = 0.5 \tag{1}$$

and for tails

$$\mathbb{P}(C = \text{tails}) = 0.5 \tag{2}$$

C is said to belong to a coin toss distribution. If the coin were weighted, then the probabilites of heads and tails would differ. C would then belong to a weighted coin toss distribution. We can denote this as

$$C \sim T(w) \tag{3}$$

where $T(w)$ represents the weighted coin distribution. This distribution is parameterized by w, which gives the amount of bias that the coin has towards heads or tails.

2.2 Basic discrete distributions

This coin toss distribution we introduced goes by another name in the statistics literature. It is called a Bernoulli distribution, and represents the probability of a random variable X being either 0 or 1 (which can be thought of as heads or tails).

$$X \sim \text{Ber}(\beta) \tag{4}$$

Here β gives the probability of X being 1:

$$\mathbb{P}(X = 1) = \beta \tag{5}$$

and thus is follows that

$$\mathbb{P}(X = 0) = 1 - \beta \tag{6}$$

Probabilities must be between 0 and 1, so β must also be a real number between 0 and 1.

We rarely see a physical process that follows a Bernoulli distribution. Instead, a more common distribution in physics is called the Poisson distribution.

$$X \sim \text{Pois}(\lambda) \tag{7}$$

This distribution gives the probability of seeing X events in a given period of time

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{8}$$

Here λ is the average number of events we expected to see. If we were looking in a time interval t, and the rate of events was r, then $\lambda = rt$.

All probability distributions must be normalized. For discrete distributions, this means that the sum of the probabilities for every possible outcome must be equal to 1. For example, for the Poisson distribution:

$$\sum_{k=0}^{\infty} \mathbb{P}(X = k) = 1 \tag{9}$$

The mean of a distribution is also known as the average. It is usually denoted as μ and is found by summing over all outcomes of the distribution:

$$\mu = \sum X\mathbb{P}(X) \tag{10}$$

This act of summing over all outcomes is called the expected value, or E:

$$\mu = E[X] \tag{11}$$

The poisson distribution is defined such that

$$\mu = \sum_{k=0}^{\infty} k\mathbb{P}(X = k) = \lambda \tag{12}$$

The variance is a measure closely related to the mean. It gives the average squared amount that the random variable X deviates from the mean:

$$\sigma^2 = E[(X - \mu)^2] \tag{13}$$

The variance is often denoted as σ^2 because it is equal to the square of the standard deviation, which is denoted as σ.

2.3 Basic continuous distributions

While the list of all possible outcomes for discrete probability distributions must be a countable set. Continuous distributions are defined on an uncountably infinite set, usually the set of real numbers.

The normalization condition now becomes an integral

$$\int_{\mathcal{S}} p(x)dx = 1 \tag{14}$$

where $\mathcal{S}$ is the set the distribution is defined on (usually some interval of the real number line).

Note that if x has units of L, then $p(x)$ must have units of $1/L$ for the units in the integral to cancel. Thus $p(x)$ is called a probability density.

By far the most common continuous distribution is called the normal distribution.

$$X \sim \mathcal{N}(\mu, \sigma) \tag{15}$$

where μ is the mean of the distribution, and σ is the standard deviation.

The probability density is a normalized Gaussian function

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{16}$$

2.4 Joined probability distributions and conditional probability

Suppose we have two random variables X and Y with probability distributions $\mathbb{P}(X)$ and $\mathbb{P}(Y)$. If X and Y are independent variables, the probability of seeing x and y is

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \tag{17}$$

This is called the joint probability distribution of X and Y. Note that when X and Y are not independent, it is not possible to decompose the joint distribution into separate distributions for each variable.

For example, suppose that we have a joint distribution

$$p(x,y) = \frac{1}{\sqrt{6\pi}} e^{-\frac{2}{3}(x^2+y^2-xy)} \tag{18}$$

Because of the xy term in the exponential, we cannot write this distribution as a product of $p(x)$ and $p(y)$.

In this case, x and y are said to be correlated with each other.

We can find the probability of just y from a joint distribution by marginializing:

$$p(y) = \int_{\mathcal{S}} p(x,y)dx \tag{19}$$

For a discrete distribution, this takes the form of summing over all possible outcomes

$$\mathbb{P}(Y=y) = \sum_x \mathbb{P}(X=x, Y=y) \tag{20}$$

If X and Y are correlated, we may not be able to decompose the joint into a product of separate distributions, but we can write it as a product of one separate distribution and a *conditional* distribution:

$$\mathbb{P}(X=x, Y=y) = \mathbb{P}(X=x|Y=y)\mathbb{P}(Y=y) \tag{21}$$

The term $\mathbb{P}(X=x|Y=y)$ is read as the probability of $X=x$ given that $Y=y$. It is a way to write the effect of one random variable on the probability of another.

In this way, we can treat the parameter of a distribution as a random variable. For example, the mean of the Poisson distribution λ can be treated as a random variable also. Then if

$$X \sim \text{Pois}(\lambda) \tag{22}$$

the probability distribution can be written as

$$\mathbb{P}(X=k|\lambda=L) = \frac{L^k e^{-L}}{k!} \tag{23}$$

The notation $\mathbb{P}(X=k|\lambda=L)$ is cumbersome, so we often write this just as

$$\mathbb{P}(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{24}$$

3 What is statistics?

3.1 Data, model, and likelihood

After running an experiment, we will have collected some data that we represent as d. For example, we might have measured 100 radioactive decays in an hour, so $d = 100$.

As physicists, we need to construct a model that predicts our observed data. This can be as complicated as S matrix elements in the standard model, or in our radioactive decay example, as simple as a half-life.

With our model and data in hand, we can now find the probability of seeing this data given our model. This probability distribution is called the likelihood. It is a conditional distribution, giving the probability of d given a parameter for our model θ.

$$\mathbb{P}(d|\theta) \tag{25}$$

This is often written using the symbol $\mathcal{L}$ as shorthand:

$$\mathcal{L}(\theta) := \mathbb{P}(d|\theta) \tag{26}$$

In our example, the likelihood is the Poisson distribution where $\lambda = \ln(2)/t_{1/2}$:

$$\mathcal{L}(t_{1/2}) = \mathbb{P}_{\text{Pois}}(d|\ln(2)/t_{1/2}) \tag{27}$$

3.2 Multiple measurements

Suppose our experiment produced two results, d_1 and d_2. Most often, these measurements are independent, so we can write the likelihood as a product of the individual likelihoods for each measurement.

$$\mathcal{L}(\theta) = \mathbb{P}(d_1|\theta)\mathbb{P}(d_2|\theta) \tag{28}$$

Often we generalize this to arbitrary measurements using vector notation. Let

$$\vec{d} = (d_1, d_2, \ldots) \tag{29}$$

then

$$\mathcal{L}(\theta) = \mathbb{P}(\vec{d}|\theta) \tag{30}$$

While the methods discussed in the remainder of this document generalizes to many measurements, we will avoid littering the equations with vector notation.

3.3 Binning

When an experiment produces large numbers of events, it is prudent to organize them into bins. Each bin spans a small range of an observable, and the value of each bin is equal to the number of events that land in this range.

For example, suppose we have a neutrino detector that measures neutrino events as a function of energy. We can define three bins:

$$\begin{aligned}\text{Bin}_1 &= [0\text{ TeV}, 100\text{ TeV}]\\ \text{Bin}_2 &= [100\text{ TeV}, 10\text{ PeV}]\\ \text{Bin}_3 &= [10\text{ PeV}, \infty]\end{aligned}$$

Note that the last bin extends to infinity. It is important that our bins the entire range of energies that the detector could measure, even if there are no events found there.

However, often an experiment's sensitivity becomes negligible below or above a certain energy, and the bins do not have to extent past these points as additional bins would no effect the final result.

We then define d_1, d_2, d_3 to be the number of events that land in each bin, and the likelihood can be written as a product of Poisson distributions for each d as the events are independent of each other.

4 Bayesian inference

So far we've introduced the likelihood: the probability of seeing some data given some model parameters. However, we are generally not interested in making statements about the measurements of an experiment. Instead, we want to talk about the model parameters and how our measurements affect our estimates of these parameters.

To do this, we need a conditional probability distribution of the model parameters given our data:

$$\mathbb{P}(\theta|d) \tag{31}$$

The process of finding this distribution, and how to interpret it is called Bayesian statistics.

4.1 Bayes' theorem

We start from the definition of conditional probability

$$\mathbb{P}(d, \theta) = \mathbb{P}(d|\theta)\mathbb{P}(\theta). \tag{32}$$

Notice that we can also write this as

$$\mathbb{P}(d, \theta) = \mathbb{P}(\theta|d)\mathbb{P}(d) \tag{33}$$

Equating the two definitions gives

$$\mathbb{P}(\theta|d)\mathbb{P}(d) = \mathbb{P}(d|\theta)\mathbb{P}(\theta) \tag{34}$$

This is known as Bayes theorem, and is usually written as

$$\mathbb{P}(\theta|d) = \frac{\mathbb{P}(d|\theta)\mathbb{P}(\theta)}{\mathbb{P}(d)} \tag{35}$$

The term $\mathbb{P}(d|\theta)$ is the likelihood that we recognize form earlier, so we can also write this as

$$\mathbb{P}(\theta|d) = \frac{\mathcal{L}(\theta)\mathbb{P}(\theta)}{\mathbb{P}(d)} \tag{36}$$

The unconditional distribution $P(\theta)$ is called the *prior*, while the conditional distribution $\mathbb{P}(\theta|d)$ is called the posterior. The names represent the probability of θ before and after taking into account our data d.

4.2 Priors

The prior encodes information about the parameter that we knew before measuring our data. For example, we might know what values a cross-section should take based on other experiments.

However, if we're testing an entirely new model (for example, some beyond the standard model) then we don't have any previous information about the parameter. In this case we need to choose an *uninformative prior*, one that has as little information as possible in it.

The uniform distribution has the least information of any probability distribution. And so the uniform distribution is often a good choice for the prior. However, the uniform distribution requires *boundaries* , otherwise it is not normalizable. This can pose a problem if the natural range of the parameter is infinite.

Additionally, the uniform distribution is not much of a help if the distribution is continuous. Suppose we have a prior and likelihood defined as

$$\int \mathcal{L}(\theta)\mathbb{P}(\theta)d\theta \tag{37}$$

and we choose $\mathbb{P}(\theta) = 1/L$ as a uniform distribution in $[0, L]$. Thus, the integral is

$$\int_0^L \mathcal{L}(\theta)\frac{1}{L}d\theta \tag{38}$$

Now, we can make a transformation of variables $\phi = e^\theta$. So

$$d\phi = e^\theta d\theta$$
$$\frac{d\phi}{\phi} = d\theta$$

Now the integral can be written as

$$\int_1^{e^L} \mathcal{L}(\phi) \frac{1}{L\phi} d\phi \tag{39}$$

In these coordinates, the prior is no longer uniform

$$\mathbb{P}(\phi) = \frac{1}{L\phi}, \tag{40}$$

and is now a power law.

For these reasons, there is often no single best choice of prior. Instead, we should aim to select a *weak prior*: one that has a weak effect on our final results. If the parameter range is finite, then a uniform distribution can work. If it is infinite, then a very wide normal distribution is usually the best choice.

4.3 Posterior

After selection of our prior, all that remains is to calculate the posterior $\mathbb{P}(\theta|D)$. To do this, we need to find

$$\mathbb{P}(d) = \int \mathbb{P}(d|\theta) P(\theta) d\theta \tag{41}$$

This is known as the *evidence integral* (for reasons that are explained later) and is generally very hard to compute directly when the number of parameters is high.

Note that, for now, we don't actually care about the value of $\mathbb{P}(d)$; we only need it for normalizing the posterior. We could sidestep the evidence integral entirely by drawing random samples directly out of the posterior distribution. This is possible because the shape of the posterior only depends on the likelihood and prior, which are easy to calculate.

The algorithm for drawing these samples is called a Markov Chain Monte-Carlo (MCMC). There are many different kinds of MCMCs, each one designed to optimize for a specific use case.

Measures of the posterior (such as mean, standard deviation, or those discussed in the next section) can be estimated from these samples. The posterior itself can even be estimated using the density of the samples.

4.4 Parameter estimation and credibility regions

When reporting results, it is useful to be able to condense the posterior down into a few numbers of easy to read plots.

The easiest to understand is the location in parameter space that maximizes the posterior:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(\theta|d) \tag{42}$$

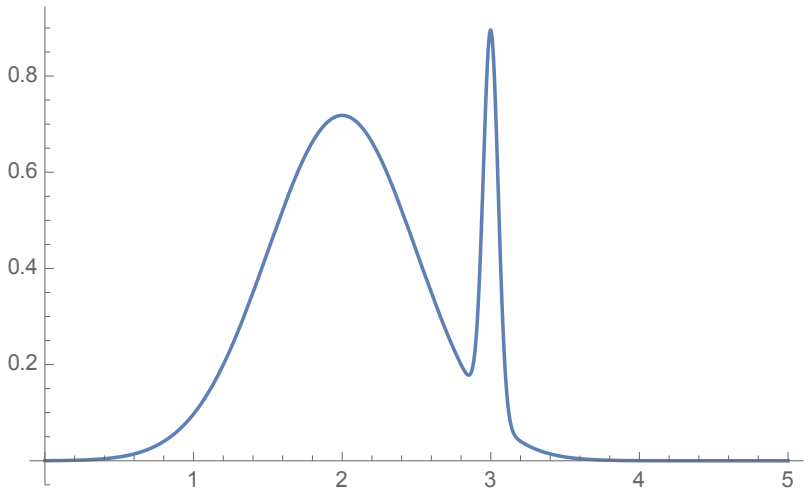

Figure 1: A pathological distribution.

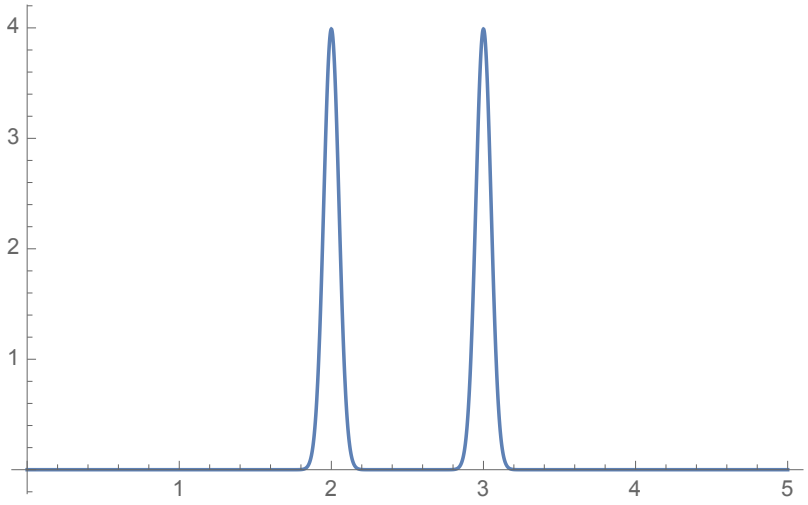

Figure 2: Another pathological distribution.

This is called the maximum a posteriori (MAP) probability estimator. It is kind of like the best fit, but also includes information from the prior.

The MAP can sometimes be misleading, consider the distribution in figure **??**. The MAP gives a value of 3, characterizing an important part of the distribution (the spike). However, it ignores a more important part of the distribution, that most of the probability lies in the area around 2.

The mean of this posterior is 2.1, and while this is a better indicator of where the probability lies, the mean is not a reliable measure either. Consider the distribution in Figure 2. The mean of this distribution is 2.5, completely missing all of the probability.

4.4.1 Credibility regions

Given this, it seems prudent to report an interval where most of the probability lies, instead of just a single number. For figure **??** this could be a single interval of $[0.8, 3.2]$, and for figure **??** this could be two intervals $[1.9, 2.1]$ and $[2.9, 3.1]$.

To be more specific, we define a *credibility interval* $\mathcal{C}(\alpha)$ by

$$1 - \alpha = \int_{\mathcal{C}(\alpha)} p(\theta|d)d\theta \tag{43}$$

which encloses $(1-\alpha)\%$ of the probability of the posterior. For example, a 90% credibility interval would be $\mathcal{C}(0.1)$.

There is no single solution for $\mathcal{C}(\alpha)$; i.e. there are many equally valid regions for any values of α.

The most common solution when multiple parameters are involved is the *highest posterior density* credibility interval. It is defined so that all points inside the interval have a higher probability density than all points outside the interval.

Mathematically, we find a value f_α that defines the interval by all points that have a probability density greater than f_α:

$$\mathcal{C}_{\text{HPD}}(\alpha) = \{\theta : p(\theta|d) \geq f_\alpha\} \tag{44}$$

The value of f_α has to be chosen such that equation **??** is satisfied.

Graphically, this can be thought of as drawing a horizontal line through the distribution such that the integral over all areas where the posterior is higher than the line equals $1-\alpha$.

4.5 Bayes factor and model selection

Suppose we have two models (M_0 and M_1) for our data (e.g. a power law and broken power law) and we wish to figure out which one is favoured by the data we measure.

To do this, we need to find the posterior probability of our model *given* the measured data: $\mathbb{P}(M_0|d)$. This can be done using Bayes theorem,

$$\mathbb{P}(M_0|d) = \frac{\mathbb{P}(d|M_0)\mathbb{P}(M_0)}{\mathbb{P}(d)} \tag{45}$$

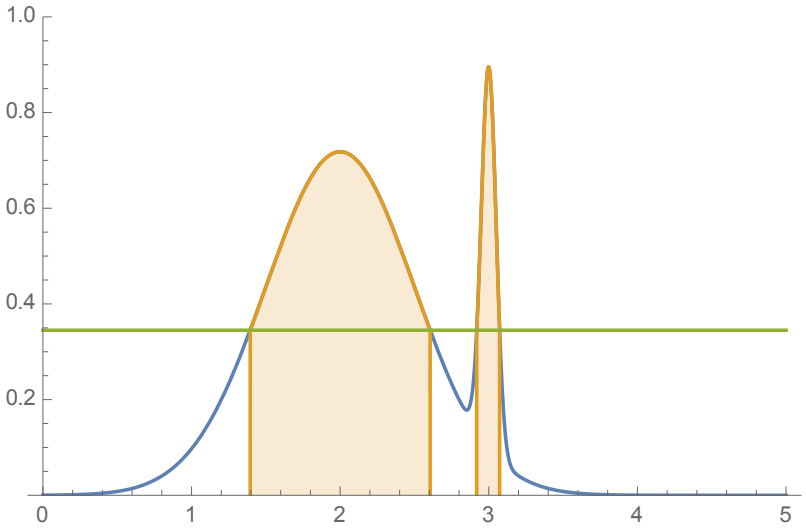

Figure 3: An example of the graphical method for determining the 80% HPD credibility interval.

The term in the denominator $\mathbb{P}(d)$ is, in general, impossible to compute, as it involves marginalizing over all possible models. This term can be neatly cancelled out by taking the ratio of the posteriors for each of our two models:

$$\frac{\mathbb{P}(M_0|d)}{\mathbb{P}(M_1|d)} = \frac{\mathbb{P}(d|M_0)\mathbb{P}(M_0)}{\mathbb{P}(d|M_1)\mathbb{P}(M_1)} \tag{46}$$

This can be seen as the product of two factors: the ratio of the priors, and a term we call the Bayes factor.

$$\mathcal{B}_{01} = \frac{\mathbb{P}(d|M_0)}{\mathbb{P}(d|M_1)} \tag{47}$$

The $\mathbb{P}(d|M_0)$ can be recognized as the *evidence integral* mentioned previously. It is calculated by marginalizing the likelihood for M_0 over the model parameters

$$\mathbb{P}(d|M_0) = \int \mathbb{P}(d|\theta, M_0)\mathbb{P}(\theta|M_0)d\theta \tag{48}$$

A similar construction is used to find $\mathbb{P}(d|M_1)$. The likelihood functions for M_0 and M_1 could be different, and have different numbers of parameters, hence the need to condition the likelihood on the model as well.

The Bayes factor tells us how much the data favours one model over another in a prior independent way. Note that this is only independent of the model priors, not the priors on the parameters.

If the model priors are equal (their ratio is unity) then the Bayes factor can be used alone to weight the evidence. The following table provides a suggestion on how to interpret the computed value.

B	Strength of evidence for M_0
1 - 3	Not worth mentioning
3 - 20	Positive
20 to 150	Strong
> 150	Overwhelming

Factors less than 1 favour M_1 in a similar fashion.

If model priors differ, then they must be taken into account before deciding which model is more probable. When the difference in priors is substantial (for example, when considering the probability of a teapot orbiting the sun) then the Bayes factor will have to also be substantial to compensate. Thus, extraordinary claims require extraordinary evidence integrals.

5 Frequentist approach

In the previous section we discussed the Bayesian paradigm of statistical inference. In this section we will discuss the frequentist counterpart. As the name indicates the frequentist approach defines the probability of an event by means of its chance of repetition. Thus, *e.g.,* when we say that the probability that a fair coin will yield head is $1/2$ it means that if I throw the coin a large number of times approximately half of the time it will yield head. Thus remember that frequentist statements are not statements about the physical parameters, but of the distribution of random variables known as *test statistics*.

In what follows we will describe various techniques accomplish the following tasks: obtain *best* model parameters, confidence intervals, compare models, and estimate the goodness-of-fit (gof). These procedures will make use of a test statistic, which often is either the likelihood function or a chi-square statistic. Even though these objects are used across all of this tasks its not to confuse them [1].

5.1 Test statistics

A frequently used test statistics is the χ^2 statistics which is defines as follows

$$\chi^2_k(\theta) = \sum_{i=1}^{k} \frac{(x_i - \mu_i(\theta))^2}{\sigma_i^2}, \tag{49}$$

where x_i are the observation, μ_i the expected observation given the model, and σ_i the uncertainty in the measurement of x_i.

5.2 p-values and goodness-of-fit

For a χ^2 test-statistics we can define the goodness-of-fit p-value as

$$p_{value} = \mathbb{P}(\chi^2 > \chi^2_{obs}), \tag{50}$$

this definition is illustrated in Figure 4. This definition can be generalized for any $\mathcal{TS}$ in

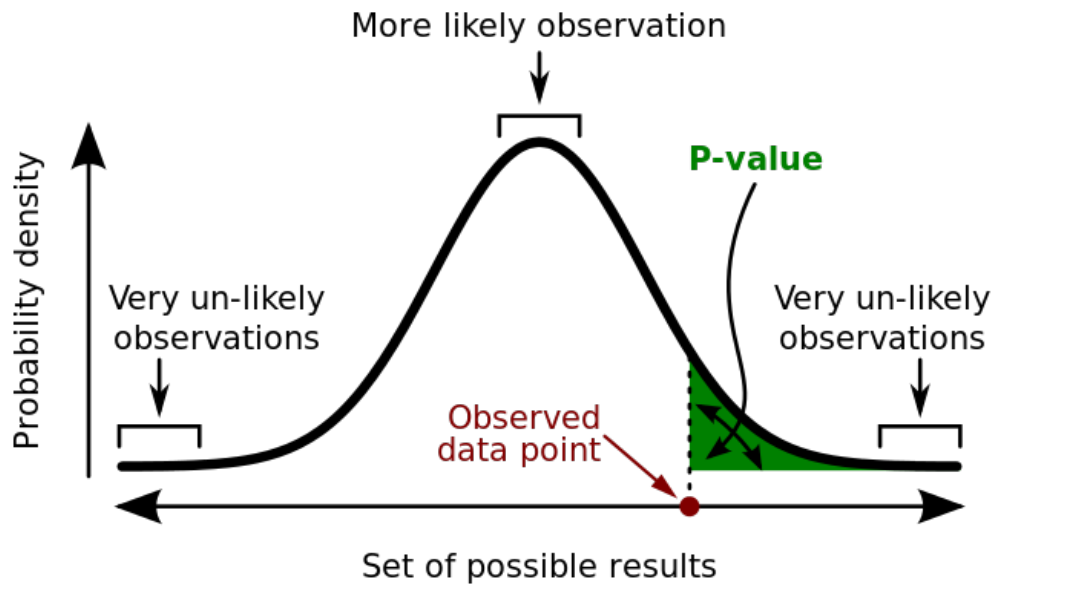

Figure 4: p-value sketch. Source wikipedia!

the following way

$$p_{value} = \mathbb{P}(\mathcal{TS} > \mathcal{TS}_{obs}) \tag{51}$$

Another useful test statistics to calculate goodness-of-fit is called the *deviance* which is defined as follows

$$\mathcal{D}(\theta) = -2(\log \mathcal{L}(\theta) - \log \mathcal{L}^s) \tag{52}$$

where $\log \mathcal{L}^s$ is known as the saturated likelihood. In the case of a poisson likelihood this is just given by

$$\log \mathcal{L}^s = \log \mathbb{P}(x|\lambda = x) \tag{53}$$

it can be shown that, with large statistics, $\mathcal{D} \sim \chi^2_k$, which makes the *deviance* a *natural* extension of the χ^2-goodness-of-fit test.

5.3 Parameter estimation and confidence regions

5.3.1 Parameter estimation

A parameter estimator is defined as

$$\hat{\theta} := \underset{\theta}{\operatorname{argmin}}\ \mathcal{TS}(\theta). \tag{54}$$

Common choices of $\mathcal{TS}$ that provide unbiassed estimators are χ^2 and $\mathcal{D}(\theta)$ or $\mathcal{L}$. In the case of the deviance or likelihood usage this estimator is known as the *maximum likelihood estimator* (MLE). Since the likelihood values are often very small to manage numerically we often maximize the $\log \mathcal{L}$, i.e.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \log \mathcal{L}(\theta), \tag{55}$$

or, equivalentely,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ \left(-\log \mathcal{L}(\theta)\right). \tag{56}$$

5.3.2 Confidence region construction

The Neyman confidence interval construction that we will briefly described here consist on two steps:

1. construction of the *confidence belt*,
2. inversion of the confidence belt to obtain the confidence region.

The confidence belt construction is related to two important elements. The first of them is to state the the confidence region satisfies the following condition

$$1 - \alpha = \mathbb{P}(\mathcal{TS}_0 < \mathcal{TS} < TS_1) = \int_{\mathcal{TS}_0}^{\mathcal{TS}_1} p_{\mathcal{TS}}(\mathcal{TS}; \theta) d\mathcal{TS}. \tag{57}$$

Its clear that this does not uniquely define $\mathcal{TS}_i$. This fact brings us to the next element known as the *ordering rule*. The ordering rule is the prescription that will be used to include elements into the confidence belt so as to satisfy 57. Three popular choices exist

- Two-sided interval: $\mathbb{P}(TS < TS_0) = \mathbb{P}(TS > TS_1) = \alpha/2$,
- One-sided upper interval: $\mathcal{TS}_0 = 0$ and set $\mathbb{P}(TS > TS_1) = \alpha$,
- One-sided lower interval: $\mathcal{TS}_1 = \infty$ and set $\mathbb{P}(TS < TS_0) = \alpha$.

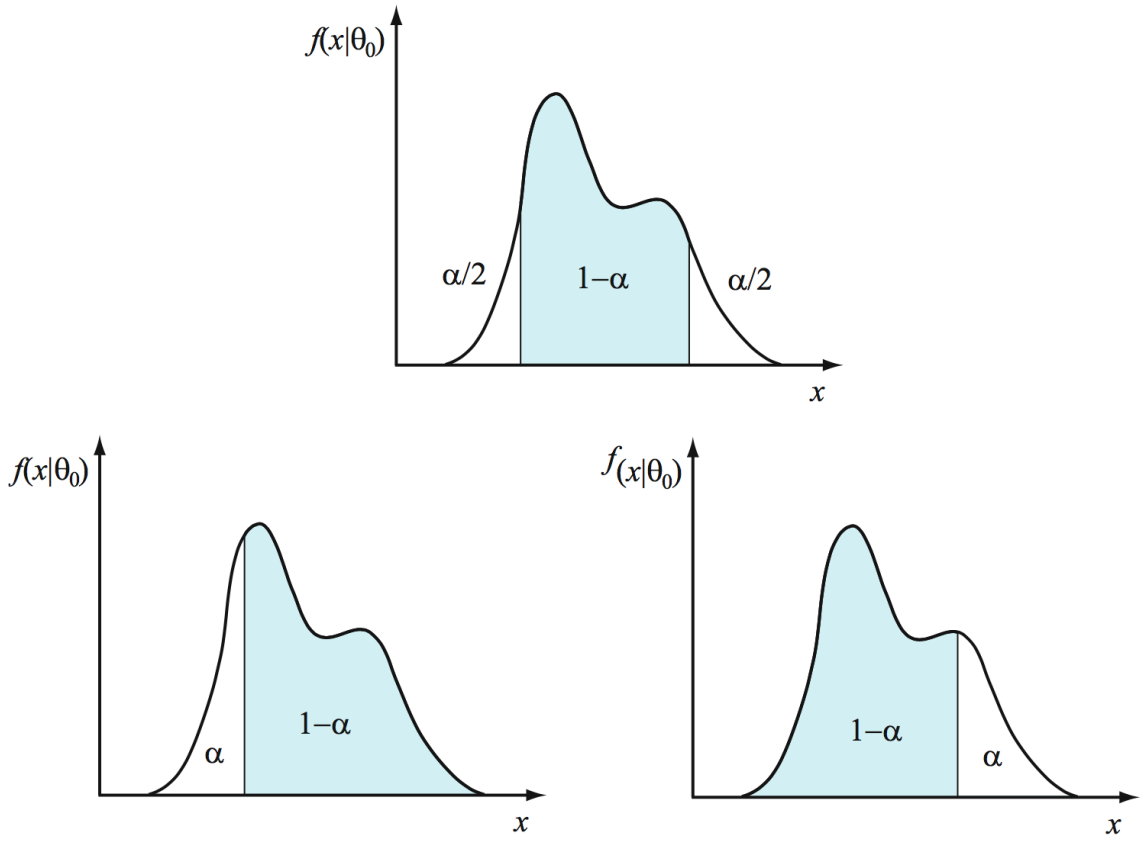

Figure 5: Three possible choices of ordering rule described in the text. Note that in this note notation $x = \mathcal{TS}$. Graphic obtained from [2].

These choices are illustrated in Figure 5.

We often use Wilks' theorem to construct confidence region. In this case its practical to define the following $\mathcal{TS}$

$$\mathcal{TS}(\theta) = \chi^2(\theta) - \chi^2_{min}. \tag{58}$$

Then $\mathcal{TS} \sim \chi^2_{dof}$ where the numebr of dof is the number of observations minus the number of parameters. in this case the confidence region is given by

$$\mathcal{C} = \{\theta : \mathcal{TS} < TS_1\} \tag{59}$$

where the threshold values for a $\Delta\chi^2$ and $2\Delta \log \mathcal{L}$ $\mathcal{TS}$ can be found in Table 5.3.2.

$(1-\alpha)(\%)$	dof = 1	dof = 2	dof = 3
68.27	1.00	2.30	3.53
90.00	2.71	4.61	6.25
95.00	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.00	6.63	9.21	11.34
99.73	9.00	11.83	14.16

Table 1: From [3] values of $\Delta\chi^2$ and $2\Delta \log \mathcal{L}$ corresponding to a coverage probability of $1-\alpha$.

It was pointed out by Feldman-Cousin [4] that an a posteriori choice of ordering produces confidence regions with the wrong coverage, this is illustrated in Figure 6. They established a new ordering known as the Feldman-Cousin ordering principle which does not require to decide in ordering before making the measurement. This ordering is establishes that you should add $\mathcal{TS}$ values to your confidence belt according to

$$\mathcal{R}(\mathcal{TS}, \theta) := \frac{p_{\mathcal{TS}}(\mathcal{TS}; \theta)}{p_{\mathcal{TS}}(\mathcal{TS}; \hat{\theta})}. \tag{60}$$

5.4 Likelihood ratio and model selection

Given two models (hypothesis) H_0 and $H_1(\theta)$ where $H_0 = H_1(\tilde{\theta})$, *i.e.* they are *nested models*. Then the Neyman-Pearson lemma states that the likelihood ratio test defined as

$$\Lambda = \frac{\mathcal{L}(\tilde{\theta})}{\mathcal{L}(\hat{\theta})}. \tag{61}$$

has the strongest statistical power (i.e. correctly rejects H_0 when H_1 is true) while at the same time given the smallest probability false positive (i.e. rejecting H_0 in favor of H_1 when H_0 is true). Its customary to then define the $\mathcal{TS} = -2\log\Lambda$ if the p-value according for this $\mathcal{TS}$, drawn from null like realization, is small then the null is rejected.

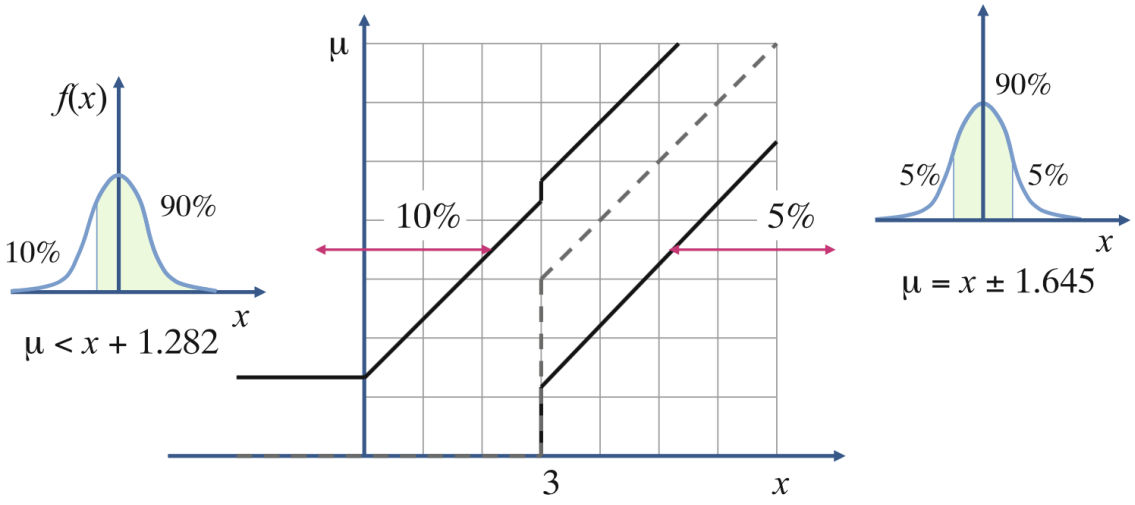

Figure 6: Illustration of the *flip-flop* problem from [2].

References

[1] Steve Baker and Robert D. Cousins. Clarification of the Use of Chi Square and Likelihood Functions in Fits to Histograms. *Nucl. Instrum. Meth.*, 221:437–442, 1984.

[2] Luca Lista. Statistical Methods for Data Analysis in Particle Physics. *Lect. Notes Phys.*, 909:pp.1–172, 2016.

[3] C. Patrignani et al. Review of Particle Physics. *Chin. Phys.*, C40(10):100001, 2016.

[4] Gary J. Feldman and Robert D. Cousins. A Unified approach to the classical statistical analysis of small signals. *Phys. Rev.*, D57:3873–3889, 1998.