Published by CERN in the Proceedings of the 2012 School of High-Energy Physics, La Pommeraye, Anjou, France, 6 – 19 June 2012, edited by C. Grojean and M. Mulders, CERN–2014–008 (CERN, Geneva, 2014)

Practical Statistics for Particle Physicists

H. B. Prosper
Florida State University, Tallahassee, USA

Abstract
We introduce a few of the key ideas of statistical analysis using two real-world examples to illustrate how these ideas are used in practice.

1 Introduction

These lectures introduce to two broad classes of theories of inference, the frequentist and Bayesian approaches. Two points should be made immediately. The first is that there is no such thing as “the" answer in statistics. Instead there are answers based on assumptions on which reasonable people may disagree. Second, none of the current theories of inference is perfect. It is worth appreciating these points in order to avoid fruitless arguments that cannot be resolved because they are ultimately about intellectual taste and not mathematical correctness.

For in-depth expositions of statistical analysis, we highly recommend the excellent books on statistics written for physicists, by physicists [1–4] and the very insightful book on the history of the ideas by Chatterjee [5].

2 Lecture 1: descriptive statistics, probability and likelihood

2.1 Descriptive statistics

Suppose we have a sample of N data $X = x_1, x_2, \cdots, x_N$. It is often useful to summarize these data with a few numbers called statistics. A **statistic** is any number that can be calculated from the data and known parameters. For example, $t = (x_1 + x_N)/2$ is a statistic, but if the value of θ is unknown $t = (x_1 - \theta)^2$ is not. However, we particle physicists tend to refer to *any* function of the data as a statistic including those that contain unknown parameters.

The two most important statistics are

the **sample mean** (or average)
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{1}$$

and the **sample variance**
$$\begin{aligned} s^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2, \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2, \\ &= \overline{x^2} - \bar{x}^2. \end{aligned} \tag{2}$$

The sample average is a measure of the center of the distribution of the data, while the sample variance is a measure of its spread. Statistics that merely characterize the data are called **descriptive statistics**, of which the sample average and variance are the most important.

Descriptive statistics can always be calculated because they depend only on a data sample X. We now consider numbers that cannot be calculated from the data alone. Imagine the repetition, infinitely many times, of the data generating system that yielded our data sample X, thereby creating an infinite set of data sets. We shall refer to the data generating system as an experiment and the infinite set of the results of the experiments as an infinite ensemble. This is clearly an abstraction.

http://dx.doi.org/10.5170/CERN-2014-008.195

The most common operation to perform on an ensemble is to compute the **ensemble average** of the statistics, which yield numbers such as the following.

$$\begin{array}{ll}
\text{Ensemble average} & \langle x\rangle \\
\text{Mean} & \mu \\
\text{Error} & \epsilon = x - \mu \\
\text{Bias} & b = \langle x\rangle - \mu \\
\text{Variance} & V = \langle (x - \langle x\rangle)^2\rangle \\
\text{Standard deviation} & \sigma = \sqrt{V} \\
\text{Mean square error} & \text{MSE} = \langle (x-\mu)^2\rangle \\
\text{Root MSE} & \text{RMS} = \sqrt{\text{MSE}}
\end{array} \tag{3}$$

None of these numbers can be calculated from data because the data needed do not objectively exist. Even in an experiment simulated on a computer, there are very few of these numbers we can calculate. If we know the mean μ, perhaps because we have chosen its value, we can certainly calculate the error ϵ for any simulated datum x. But, we can only *approximate* the ensemble average $\langle x\rangle$, bias b, variance V, and MSE, since the ensembles available either on our computers or in the real world are always finite. The point is that the numbers that characterize the infinite ensemble are also abstractions.

The MSE is the most widely used measure of the closeness of an ensemble of numbers to some parameter μ. The square root of the MSE is called the root mean square (RMS)[1]. The MSE can be written as

$$\text{MSE} = V + b^2, \tag{4}$$

Exercise 1: Show this

the sum of the variance and the square of the bias, a very important result with practical consequences. For example, suppose that μ represents the mass of the Higgs boson and x is a complicated function that is considered an **estimator** of the mass. An estimator is any function, which when data are entered into it, yields an **estimate** of the quantity of interest.

As noted, many of the numbers listed in Eq. (3) cannot be calculated because the information needed is unknown. This is true, in particular, of the bias. However, sometimes it is possible to relate the bias to another ensemble quantity. Consider the ensemble average of the sample variance, Eq. (2),

$$\begin{aligned}
\langle s^2\rangle &= \langle \overline{x^2}\rangle - \langle \bar{x}^2\rangle, \\
&= V - \frac{V}{N},
\end{aligned}$$

Exercise 2a: Show this

Exercise 2b: Use the method Rndm() of the `Root` class `TRandom3` to approximate the quantities in Eq. (3).

2.2 Probability

When the weather forecast specifies that there is a 80% chance of snow tomorrow at CERN, most people have an intuitive sense of what this means. Likewise, most people have an intuitive understanding of

[1] Sometimes, the RMS and standard deviation are using interchangeably. However, the RMS is computed with respect to μ, while the standard deviation is computed with respect to the ensemble average $\langle x\rangle$. The RMS and standard deviations are identical only if the bias is zero.

what it means to say that there is a 50-50 chance for a tossed coin to land heads up. Probabilistic ideas are thousands of years old, but, starting in the sixteenth century these ideas were formalized into increasingly rigorous mathematical theories of probability. In the theory formulated by Kolmogorov in 1933, Ω is some fixed mathematical space, $E_1, E_2, \cdots \subset \Omega$ are subsets (called events) defined in some reasonable way[2], and $P(E_j)$ is a number associated with subset E_j. These numbers satisfy the

Kolmogorov Axioms

$$\begin{array}{ll} 1. & P(E_j) \geq 0 \\ 2. & P(E_1 + E_2 + \cdots) = P(E_1) + P(E_2) + \cdots \quad \text{for disjoint subsets} \\ 3. & P(\Omega) = 1. \end{array}$$

Consider two subsets $A = E_1$ and $B = E_2$. The quantity AB means A *and* B, while $A + B$ means A *or* B, with associated probabilities $P(AB)$ and $P(A + B)$, respectively. Kolmogorov assumed, not unreasonably given the intuitive origins of probability, that probabilities sum to unity; hence the axiom $P(\Omega) = 1$. However, this assumption can be dropped so that probabilities remain meaningful even if $P(\Omega) = \infty$ [6].

Figure 1 suggests another probability, namely, the number $P(A|B) = P(AB)/P(B)$, called the **conditional probability** of A given B. This permits statements such as: “the probability that this track was created by an electron given the measured track parameters" or “the probability to observe 17 events given that the mean background is 3.8 events". Conditional probability is a very powerful idea, but the term itself is misleading. It implies that there are two kinds of probability: conditional and unconditional. In fact, *all* probabilities are conditional in that they always depend on a specific set of conditions, namely, those that define the space Ω. It is entirely possible to embed a family of subsets of Ω into another space Ω' which assigns to each family member a different probability P'. A probability is defined only relative to some space of possibilities Ω.

A and B are said to be mutually exclusive if $P(AB) = 0$, that is, if the truth of one denies the truth of the other. They are said to be exhaustive if $P(A) + P(B) = 1$. Figure 1 suggests the theorem

$$P(A + B) = P(A) + P(B) - P(AB), \qquad (5)$$

Exercise 3: Prove theorem

which can be deduced from the rules given above. Another useful theorem is an immediate consequence of the commutativity of “anding" $P(AB) = P(BA)$ and the definition of $P(A|B)$, namely,

Bayes Theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}, \qquad (6)$$

which provides a way to convert the probability $P(A|B)$ to the probability $P(B|A)$. Using Bayes theorem, we can, for example, deduce the probability $P(e|x)$ that a particle is an electron, e, given a set of measurements, x, from the probability $P(x|e)$ of a set of measurements given that the particle is an electron.

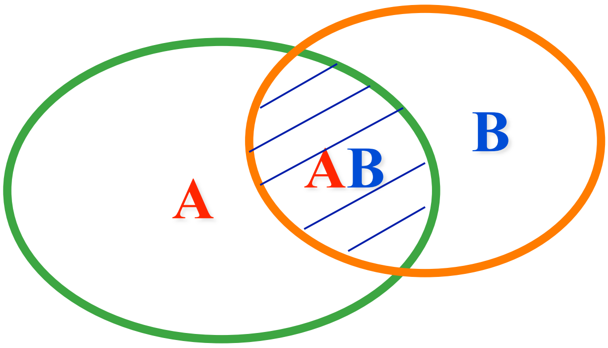

Fig. 1: Venn diagram of the sets A, B, and AB. $P(A)$ is the probability of A, while $P(A|B) = P(AB)/P(B)$ is the probability of AB relative to that of B, i.e., the probability of A given the condition B.

[2] If $E_1, E_2, \cdots$ are meaningful subsets of Ω, so to is the complement $\overline{E}_1, \overline{E}_2, \cdots$ of each, as are countable unions and intersections of these subsets.

2.2.1 *Probability distributions*

In this section, we illustrate the use of these rules to derive more complicated probabilities. First we start with a definition:

> A **Bernoulli trial**, named after the Swiss mathematician Jacob Bernoulli (1654 – 1705), is an experiment with only two possible outcomes: $S =$ success or $F =$ failure.
>
> *Example*
>
> Each collision between protons at the Large Hadron Collider (LHC) is a Bernoulli trial in which something interesting happens (S) or does not (F). Let p be the probability of a success, which is assumed to be the *same for each trial*. Since S and F are exhaustive, the probability of a failure is $1-p$. For a given order O of n proton-proton collisions and exactly k successes, and therefore exactly $n-k$ failures, the probability $P(k, O, n, p)$ is given by
>
> $$P(k, O, n, p) = p^k (1-p)^{n-k}. \tag{7}$$
>
> If the order O of successes and failures is judged to be irrelevant, we can eliminate the order from the problem by summing over all possible orders,
>
> $$P(k, n, p) = \sum_O P(k, O, n, p) = \sum_O p^k (1-p)^{n-k}. \tag{8}$$
>
> This procedure is called **marginalization**. It is one of the most important operations in probability calculations. Every term in the sum in Eq. (8) is identical and there are $\binom{n}{k}$ of them. This yields the **binomial distribution**,
>
> $$\text{Binomial(k, n, p)} \equiv \binom{n}{k} p^k (1-p)^{n-k}. \tag{9}$$
>
> By definition, the mean number of successes a is given by
>
> $$\begin{aligned} a &= \sum_{k=0}^{n} k \, \text{Binomial(k, n, p)}, \\ &= pn. \end{aligned} \tag{10}$$
>
> **Exercise 4:** Show this
>
> At the LHC n is a number in the trillions, while for successes of interest such as the creation of a Higgs boson the probability $p << 1$. In this case, it proves convenient to consider the limit $p \to 0, n \to \infty$ in such a way that a remains constant. In this limit
>
> $$\begin{aligned} \text{Binomial(k, n, p)} &\to e^{-a} a^k / k!, \\ &\equiv \text{Poisson}(k, a). \end{aligned} \tag{11}$$
>
> **Exercise 5:** Show this

Below we list the most common probability distributions.

Discrete distributions

Binomial(k, n, p) $\binom{n}{k} p^k (1-p^{n-k}$

$$\begin{array}{ll}
\text{Poisson}(k,a) & a^k \exp(-a)/k! \\
\text{Multinomial}(k,n,p) & \dfrac{n!}{k_1!\cdots k_K!}\displaystyle\prod_{i=1}^{K} p_i^{k_i}, \quad \sum_{i=1}^{K} p_i = 1, \sum_{i=1}^{K} k_i = n \\
\textbf{Continuous densities} & \\
\text{Uniform}(x,a) & 1/a \\
\text{Gaussian}(x,\mu,\sigma) & \exp[-(x-\mu)^2/(2\sigma^2)]/(\sigma\sqrt{2\pi}) \\
\text{(also known as the Normal density)} & \\
\text{LogNormal}(x,\mu,\sigma) & \exp[-(\ln x-\mu)^2/(2\sigma^2)]/(x\sigma\sqrt{2\pi}) \\
\text{Chisq}(x,n) & x^{n/2-1}\exp(-x/2)/[2^{n/2}\Gamma(n/2)] \\
\text{Gamma}(x,a,b) & x^{a-1}a^b\exp(-ax)/\Gamma(b) \\
\text{Exp}(x,a) & a\exp(-ax) \\
\text{Beta}(x,n,m) & \dfrac{\Gamma(n+m)}{\Gamma(m)\,\Gamma(n)} x^{n-1}\,(1-x)^{m-1}
\end{array} \qquad (12)$$

Particle physicists tend to use the term probability distribution for both discrete and continuous functions, such as the Poisson and Gaussian distributions, respectively. But, strictly speaking, the continuous functions are probability *densities*, not probability distributions. In order to compute a probability from a density we need to integrate the density over a finite set in x.

2.3 Likelihood

Let us assume that $p(x|\theta)$ is a **probability density function** (pdf) such that $P(A|\theta) = \int_A p(x|\theta)\,dx$ is the probability of the statement $A = x \in R_x$, where x denotes possible data, θ the parameters that characterize the **probability model** (that is the probability together with all the assumptions on which it is based), and R_x is a finite set. We shall use probability model as shorthand for probability density function (for continuous variables) or probability mass function (pmf) (basically, probabilities for discrete variables). If x is discrete, then both $p(x|\theta)$ and $P(A|\theta)$ are probabilities. The **likelihood function** is simply the probability model $p(x|\theta)$ evaluated at the data x_O actually obtained, i.e., the function $p(x_O|\theta)$. The following are examples of likelihoods.

Example 1

In 1995, CDF and DØ discovered the top quark [8, 9] at Fermilab. The DØ Collaboration found $x = N = 17$ events. For a counting experiment, the datum can be modeled using

$$\begin{aligned}
p(x|n) &= \text{Poisson}(x,n) \quad \text{probability to get } x \text{ events} \\
p(N|n) &= \text{Poisson}(N,n) \quad \text{likelihood of } N \text{ events} \\
&= n^N \exp(-n)/N!
\end{aligned}$$

We shall analyze this example in detail in Lectures 2 and 3.

Example 2

Figure 2 shows a plot of the distance modulus versus redshift for $N = 580$ Type 1a supernovae [7]. These heteroscedastic data[3] $D = \{z_i, x_i \pm \sigma_i\}$ are modeled using the likelihood

$$p(D|\Omega_M, \Omega_\Lambda, Q) = \prod_{i=1}^{N} \text{Gaussian}(x_i, \mu_i, \sigma_i),$$

[3] Data in which each item, x_i, or group of items has a different uncertainty.

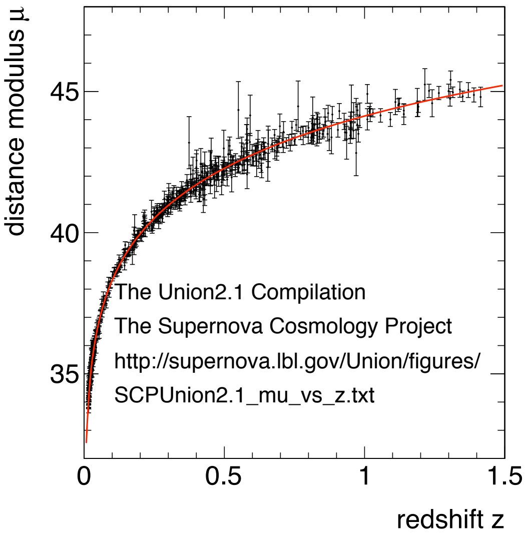

Fig. 2: Plot of the data points $(z_i, x_i{\pm}\sigma_i)$ for 580 Type 1a supernovae [7] showing a fit of the standard cosmological model (with a cosmological constant) to these data (curve).

which is an example of an *un-binned* likelihood. The cosmological model is encoded in the distance modulus function μ_i, which depends on the redshift z_i and the matter density and cosmological constant parameters Ω_M and Ω_Λ, respectively. (See Ref. [10] for an accessible introduction to the analysis of these data.)

Example 3

The discovery of a Higgs boson by ATLAS [11] and CMS [12] in the di-photon final state ($pp \to H \to \gamma\gamma$) made use of an un-binned likelihood of the form,

$$p(x|s, m, w, b) = \exp[-(s + b)] \prod_{i=1}^{N} [s f_s(x_i|m, w) + b f_b(x_i)]$$

$$\begin{aligned} \text{where } x &= \text{di-photon masses} \\ m &= \text{mass of boson} \\ w &= \text{width of resonance} \\ s &= \text{expected (i.e., mean) signal count} \\ b &= \text{expected background count} \\ f_s &= \text{signal probability density} \\ f_b &= \text{background probability density} \end{aligned}$$

Exercise 6: Show that a binned multi-Poisson likelihood yields an un-binned likelihood of this form as the bin widths go to zero

The likelihood function is arguably the most important quantity in a statistical analysis because it can be used to answer questions such as the following.

1. How do I estimate a parameter?
2. How do I quantify its accuracy?
3. How do I test an hypothesis?
4. How do I quantify the significance of a result?

Writing down the likelihood function requires:

1. identifying all that is *known*, e.g., the observations,
2. identifying all that is *unknown*, e.g., the parameters,
3. constructing a probability model for *both*.

Many analyses in particle physics do not use likelihood functions explicitly. However, since the data we use are stochastic, the failure to reflect deeply on their probabilistic nature and to model it explicitly leads to analyses that may not as good as they could be. Deconstructing carefully what is being done in an analysis is a habit that should be encouraged so that an accurate probabilistic model of the analysis can be constructed.

3 Lecture 2: the frequentist approach

In this lecture, we consider statistical inference from the frequentist viewpoint. In lecture 3, we consider the Bayesian approach. In our opinion, both are needed to make sense of statistical inference, though this is not the dominant opinion in particle physics.

The most important principle in the frequentist approach is that enunciated by the Polish statistician Jerzy Neyman in the 1930s, namely,

> **The frequentist principle**
>
> The goal of a frequentist analysis is to construct statements so that a fraction $f \geq p$ of them are guaranteed to be true over an infinite ensemble of statements.

The fraction f is called the **coverage probability**, or coverage for short, and p is called the **confidence level** (C.L.). A procedure which satisfies the frequentist principle is said to *cover*. The confidence level as well as the coverage is a property of the ensemble of statements. Consequently, the confidence level may change if the ensemble changes. In a seminal paper published in 1937, Neyman [13] invented the concept of the confidence interval, a way to quantify uncertainty, that respects the frequentist principle. The confidence interval is such an important idea that it is worth working through the concept in detail.

3.1 Confidence intervals

Consider an experiment that observes D events with expected (that is, mean) signal s and no background. Neyman devised a way to make statements of the form

$$s \in [l(D),\, u(D)], \tag{13}$$

with the *a priori* guarantee that at least a fraction p of them will be true over an ensemble of statements of this kind. A procedure for constructing such intervals is called a **Neyman construction**. The frequentist principle must hold for *any* ensemble of experiments, not necessarily all making the same kind of observations and statements. For simplicity, however, we shall consider the experiments to be of the same kind and to be completely specified by a single unknown parameter s. The Neyman construction is illustrated in Fig. 3.

The construction proceeds as follows. Choose a value of s and use some rule to find an interval in the space of observations (or, more generally, a region), for example, the interval defined by the two

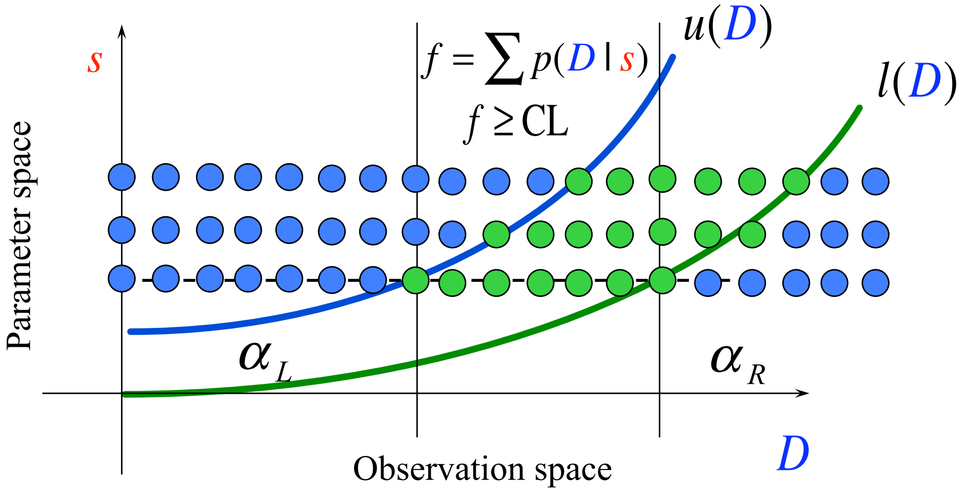

Fig. 3: The Neyman construction. Plotted is the Cartesian product of the parameter space, with parameter s, and the space of observations with potential observations D. For a given value of s, the observation space is partitioned into three disjoint intervals, such that the probability to observe a count D within the interval demarcated by the two vertical lines is $f \geq p$, where p = C.L. is the desired confidence level. The inequality is needed because, for discrete data, it may not be possible to find an interval with $f = p$ exactly.

vertical lines in the center of the figure, such that the probability to obtain a count in this interval is $f \geq p$, where p is the desired confidence level. Then move to another value of s and repeat the procedure. The procedure is repeated for a sufficiently dense set of points in the parameter space over a sufficiently large range in s. When this is done, as illustrated in Fig. 3, the intervals of probability content f will form a band in the Cartesian product of the parameter space and the observation space. The upper edge of this band defines the curve $u(D)$, while the lower edge defines the curve $l(D)$. These curves are the outcome of the Neyman construction.

For a given value of s, the interval with probability content f in the space of observations is not unique since different rules for choosing the interval will, in general, yield different intervals. Neyman suggested choosing the interval so that the probability to obtain an observation to the right or left of the interval are the same (for a given value of s), which yields the so-called **central intervals**. One virtue of these intervals is that their boundaries can be more efficiently calculated by solving the equations,

$$\begin{aligned} P(x \leq D|u) &= \alpha_L, \\ P(x \geq D|l) &= \alpha_R, \end{aligned} \tag{14}$$

a mathematical fact that becomes clear if we stare at Fig. 3 long enough.

Another rule was suggested by Feldman and Cousins [14]. For our example, the Feldman-Cousins rule requires that the potential observations $\{D\}$ be ordered in descending order, $D_{(1)}, D_{(2)}, \cdots$, of the likelihood ratio $p(D|s)/p(D|\hat{s})$, where $\hat{s}$ is the maximum likelihood estimator (see Sec. 3.2) of the parameter s. Once ordered, we compute the running sum $f = \sum_j p(D_{(j)}|s)$ until f equals or just exceeds the desired confidence level p. This rules does not guarantee that the potential observations D are contiguous, but this does not matter because we simply take the minimum element of the set $\{D_{(j)}\}$ to be the lower bound of the interval and its maximum element to be the upper bound.

Another simple rule is the mode-centered rule: order D in descending order of $p(D|s)$ and proceed as with the Feldman-Cousins rule. In principle, absent criteria for choosing a rule, there is nothing to prevent the use of *ordering rules* randomly chosen for different values of s! Figure 4 compares the widths of the intervals $[l(D), u(D)]$ for three different ordering rules, central, Feldman-Cousins, and mode-centered as a function of the count D. It is instructive to compare these widths with those provided by the well-known root(N) interval, $l(D) = D - \sqrt{D}$ and $u(D) = D + \sqrt{D}$. Of the three sets of intervals, the ones suggested by Neyman are the widest, the Feldman-Cousins and mode-centered ones are of similar width, while the root(N) intervals are the shortest. So why are we going through all the trouble of the Neyman construction? We shall return to this question shortly.

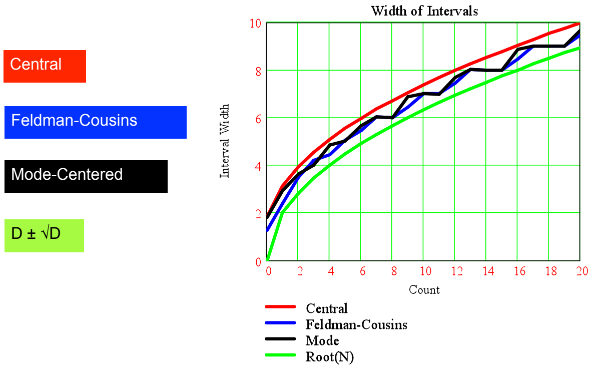

Fig. 4: Interval widths as a function of count D for four sets of intervals.

Having completed the Neyman construction and found the curves $u(D)$ and $l(D)$ we can use the latter to make statements of the form $s \in [l(D), u(D)]$: for a given observation D, we simply read off the interval $[l(D), u(D)]$ from the curves. For example, suppose in Fig. 3 that the true value of s is represented by the horizontal line that intersects the curves $u(D)$ and $l(D)$ and which therefore defines the interval demarcated by the two vertical lines. If the observation D happens to fall in the interval to the left of the left vertical line, or to the right of the right vertical line, then the interval $[l(D), u(D)]$ will not bracket s. However, if D falls between the two vertical lines, the interval $[l(D), u(D)]$ will bracket s. Moreover, by virtue of the Neyman construction, a fraction f of the intervals $[l(D), u(D)]$ will bracket the value of s whatever its value happens to be, which brings us back to the question about the root(N) intervals. Figure 5 shows the coverage probability over the parameter space of s. As expected, the three rules, Neyman's, that of Feldman-Cousins, and the mode-centered, satisfy the condition coverage probability $\geq$ confidence level over all values of s that are possible *a priori*; that is, the intervals cover. However, the root(N) intervals do not and indeed fail badly for $s < 2$.

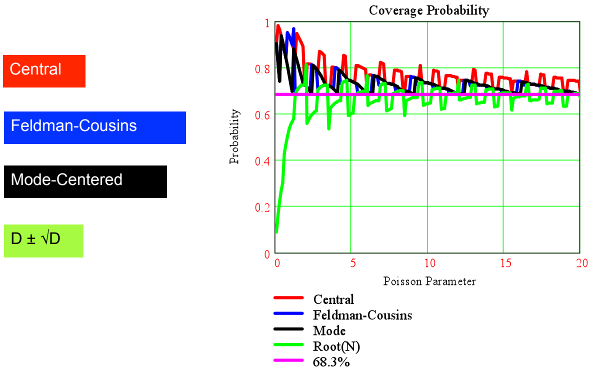

Fig. 5: Interval widths as a function of count D for four sets of intervals.

However, the coverage probability of the root(N) intervals bounces around the (68%) confidence level for vaues of $s > 2$. Therefore, if we knew for sure that $s > 2$, it would seem that using the root(N) intervals may not be that bad after all.

So what, after all this, does the statement $s \in [l(D), u(D)]$ at $100p\%$ C.L. mean in this approach, given that p is a property of the ensemble to which this statement belongs? In means this: $s \in [l(D), u(D)]$ is a member of an ensemble of statements a fraction $f \geq p$ of which are true. In principle, in order to verify this we need just count how many statements of the form $s \in [l(D), u(D)]$ are true and divide by the total number of statements. Unfortunately, this requires that we know which statements are true. But if we knew that we would not need a theory of statistical inference!

Neyman required a procedure to cover whatever the value of *all* the parameters, be they known or unknown, of the probability models that describe the data generation mechanisms. This is a very tall order, which cannot be met in general. In practice, we resort to approximations, the most widely used of which is the profile likelihood to which we now turn.

3.2 The profile likelihood

As noted in Section 2.3, likelihood functions can be used to estimate the parameters on which they depend. The method of choice to do so, in a frequentist analysis, is called **maximum likelihood**, a method first used by Karl Frederick Gauss and developed into a formidable statistical tool in the 1930s by Sir Ronald A. Fisher [15], perhaps the most influential statistician of the twentieth century. The DØ top quark discovery example illustrates the method.

Example: Top Quark Discovery Revisited

We start by listing

the knowns

$D = N, B$ where

$N = 17$ observed events

$B = 3.8$ estimated background events with uncertainty $\delta B = 0.6$

and the unknowns

b mean background count

s mean signal count.

Next, we construct a probability model for the data $D = N, B$. Since this is a counting experiment, we shall assume that $p(x|s, b)$ includes a Poisson distribution with mean count $s + b$. In the absence of details about how the background B was arrived at, the standard assumption is that data of the form $y \pm \delta y$ can be modeled with a Gaussian (or normal) density. However, we shall do something slightly better.

Suppose that the observed count in the control region is Q and the mean count is bk, where k (ideally) is the known scale factor between the control and signal regions. But, since we are given B and δB rather than Q and k, we need to relate the two pairs of numbers. The simplest model is $B = Q/k$ and $\delta B = \sqrt{Q}/k$ from which we can infer an effective count Q using $Q = (B/\delta B)^2$. Since the scale factor k is not given, we shall use the obvious estimate $k \sim Q/B = B/\delta B^2$. With these assumptions, our likelihood function is

$$\begin{aligned} p(D|s, b) &= \text{Poisson}(N, s + b)\,\text{Poisson}(Q, bk), \qquad (15)\\ \text{where} & \\ Q &= (B/\delta B)^2 = 41.11, \\ k &= B/\delta B^2 = 10.56. \end{aligned}$$

The first term in Eq. (15) is the likelihood for the count $N = 17$, while the second term is the likelihood for $B = 3.8$, or equivalently the count Q. The fact that Q is not an integer causes no difficulty; we merely continue the Poisson distribution to non-integer Q using $(bk)^Q \exp(-bk)/\Gamma(Q + 1)$.

The maximum likelihood estimators for s and b are found by maximizing Eq. (15), that is, by solving the equations

$$\begin{aligned} \frac{\partial \ln p(D|s, b)}{\partial s} &= 0 \quad \text{leading to } \hat{s} = N - B, \\ \text{and } \frac{\partial \ln p(D|s, b)}{\partial b} &= 0 \quad \text{leading to } \hat{b} = B, \end{aligned}$$

as expected.

A more complete model would account for the uncertainty in k.

The maximum likelihood method is the most widely used method for estimating parameters because it generally leads to reasonable estimates. But the method has features, or encourages practices, which, somewhat uncharitably, we label the good, the bad, and the ugly!

- *The Good*
  - Maximum likelihood estimators are consistent: the RMS goes to zero as more and more data are included in the likelihood. This is an extremely important property, which basically says it makes sense to take more data because we shall get more accurate results. One would not knowingly use an inconsistent estimator!
  - If an unbiased estimator for a parameter exists the maximum likelihood method will find it.
  - Given the MLE for s, the MLE for any function $y = g(s)$ of s is, very conveniently, just $\hat{y} = g(\hat{s})$. This is a very nice practical feature which makes it possible to maximize the likelihood using the most convenient parameterization of it and then transform back to the parameter of interest at the end.
- *The Bad (according to some!)*
  - In general, MLEs are biased.

Exercise 7: Show this
Hint: Taylor expand $y = g(\hat{s} + h)$ about the MLE $\hat{s}$, then consider its ensemble average.

- *The Ugly (according to some!)*
  - The fact that most MLEs are biased encourages the routine application of bias correction, which can waste data and, sometimes, yield absurdities.

Here is an example of the seriously ugly.

Example

For a discrete probability distribution $p(k)$, the **moment generating function** is the ensemble average

$$\begin{aligned} G(x) &= \langle e^{xk} \rangle \\ &= \sum_k e^{xk}\, p(k). \end{aligned}$$

For the binomial, with parameters p and n, this is

$$G(x) = (e^x p + 1 - p)^n, \qquad \boxed{\textbf{Exercise 8a:}\text{ Show this}}$$

which is useful for calculating **moments**

$$\mu_r = \left.\frac{d^r G}{dx^r}\right|_{x=0} = \sum_k k^r\, p(k),$$

e.g., $\mu_2 = (np)^2 + np - np^2$ for the binomial distribution. Given that k events out n pass a set of cuts, the MLE of the event selection efficiency is the obvious estimate $\hat{p} = k/n$. The equally obvious estimate of p^2 is $(k/n)^2$. But,

$$\langle (k/n)^2 \rangle = p^2 + V/n, \qquad \boxed{\textbf{Exercise 8b:}\text{ Show this}}$$

so $(k/n)^2$ is a biased estimate of p^2 with positive bias V/n. The unbiased estimate of p^2 is

$$k(k-1)/[n(n-1)], \quad \boxed{\textbf{Exercise 8c:} \text{ Show this}}$$

which, for a single success, i.e., $k = 1$, yields the sensible estimate $\hat{p} = 1/n$, but the less than useful $\widehat{p^2} = 0$!

In order to infer a value for the parameter of interest, for example, the signal s in our 2-parameter likelihood function in Eq. (15), the likelihood must be reduced to one involving the parameter of interest only, here s, by getting rid of all the **nuisance** parameters, here the background parameter b. A nuisance parameter is any parameter that is not of current interest. In a strict frequentist calculation, this reduction to the parameter of interest must be done in such a way as to respect the frequentist principle: *coverage probability* $\geq$ *confidence level*. In general, this is very difficult to do exactly.

In practice, we replace all nuisance parameters by their **conditional maximum likelihood estimates** (CMLE). The CMLE is the maximum likelihood estimate conditional on a *given* value of the current parameter (or parameters) of interest. In the top discovery example, we construct an estimator of b as a function of s, $\hat{b}(s)$, and replace b in the likelihood $p(D|s, b)$ by $\hat{b}(s)$ to yield a function $p_{PL}(D|s)$ called the **profile likelihood**.

> *Since the profile likelihood entails an approximation, namely, replacing unknown parameters by their conditional estimates, it is no longer the likelihood but rather an approximation to it. Consequently, the frequentist principle is not guaranteed to be satisfied exactly.*

But, if certain conditions are met (Wilks' theorem, 1938), roughly that the MLEs do not occur on the boundary of the parameter space and the likelihood becomes ever more Gaussian as the data become more numerous — that is, in the so-called **asymptotic limit**, then if the true density of x is $p(x|s, b)$ the random number

$$t(x, s) = -2 \ln \lambda(x, s), \tag{16}$$

$$\text{where } \lambda(x, s) = \frac{p_{PL}(x|s)}{p_{PL}(x|\hat{s})}, \tag{17}$$

has a probability density that converges to a χ^2 density with one degree of freedom. More generally, if the numerator of λ contains m free parameters the asymptotic density of t is a χ^2 density with m degrees of freedom. Therefore, we may take $t(D, s)$ to be a χ^2 variate, at least approximately, and solve $t(D, s) = n^2$ for s to get approximate n-standard deviation confidence intervals. In particular, if we solve $t(D, s) = 1$, we obtain approximate 68% intervals. This calculation is what `Minuit`, and now `TMinuit`, has done countless times since the 1970s! Wilks' theorem provides the main justification for using the profile likelihood. We again use the top discovery example to illustrate the procedure.

Example: Top Quark Discovery Revisited Again

The conditional MLE of b is found to be

$$\hat{b}(s) = \frac{g + \sqrt{g^2 + 4(1+k)Qs}}{2(1+k)}, \tag{18}$$

where

$$g = N + Q - (1+k)s.$$

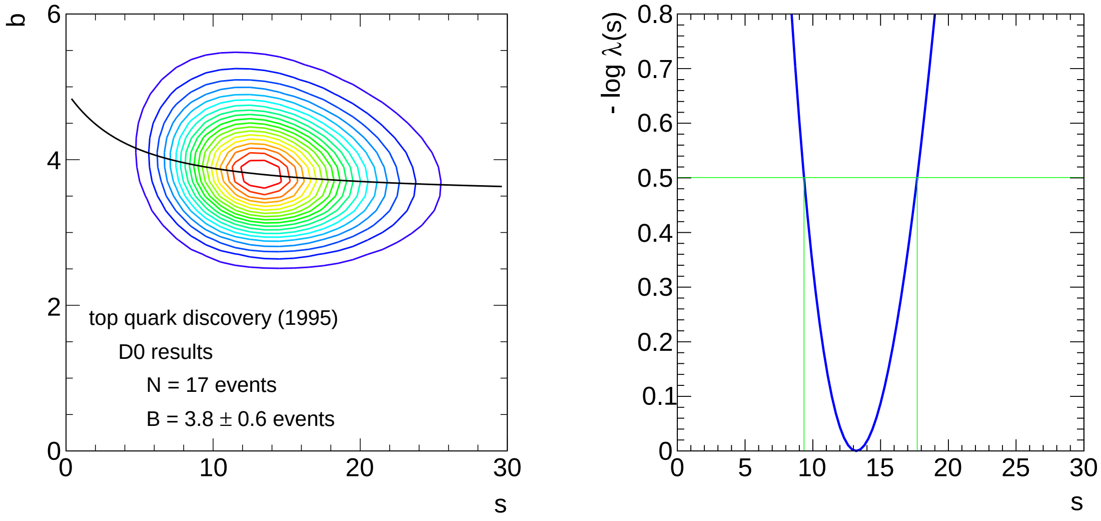

Fig. 6: (a) Contours of the DØ top discovery likelihood and the graph of $\hat{b}(s)$. (b) Plot of $-\ln\lambda(17, s)$ versus the expected signal s. The vertical lines show the boundaries of the approximate 68% interval.

The likelihood $p(D|s, b)$ is shown in Fig. 6(a) together with the graph of $\hat{b}(s)$. The mode (i.e. the peak) occurs at $s = \hat{s} = N - B$. By solving

$$-2\ln\frac{p_{PL}(17|s)}{p_{PL}(17|17-3.8)} = 1$$

for s we get two solutions $s = 9.4$ and $s = 17.7$. Therefore, we can make the statement $s \in [9.4, 17.7]$ at approximately 68% C.L. Figure 6(b) shows a plot of $-\ln\lambda(17, s)$ created using the RooFit [16] and RooStats [17] packages.

Exercise 9: Verify this interval using the RooFit/RooStats package

Intervals constructed this way are not guaranteed to satisfy the frequentist principle. In practice, however, their coverage is very good for the typical probability models used in particle physics, even for modest amounts of data.

3.3 Hypothesis tests

It is hardly possible in experimental particle physics to avoid testing hypotheses, testing that invariably leads to decisions. For example, electron identification entails hypothesis testing; given data D we ask: is this particle an isolated electron or is it not an isolated electron? In the discovery of the Higgs boson, we had to test whether, given the data available in early summer 2012, the Standard Model without a Higgs boson, a somewhat ill-founded background-only model, or the Standard Model the new boson in July 2012, the background + signal model, was the preferred hypothesis. We decided that the latter model was preferred and announced the discovery of a new boson. Given the ubiquity of hypothesis testing, it is important to have a grasp of the methods that have been invented to implement it.

One method was due to Fisher [15], another was invented by Neyman, and a third (Bayesian) method was proposed by Sir Harold Jeffreys [18], all around the same time. We first describe the method of Fisher, then follow with a description of the method of Neyman. For concreteness, we consider the problem of deciding between a background-only model and a background + signal model.

3.3.1 *Fisher's approach*

In Fisher's approach, we construct a **null hypothesis**, often denoted by H_0, and *reject* it should some measure be judged small enough to cast doubt on the validity of this hypothesis. In our example, the null hypothesis is the background-only model, for example, the SM without a Higgs boson. The measure is called a **p-value** and is defined by

$$\text{p-value}(x_0) = P(x > x_0|H_0), \tag{19}$$

where x is a statistic designed so that large values indicate departure from the null hypothesis. This is illustrated in Fig. 7, which shows the location of the observed value x_0 of x. The p-value is the probability that x could have been higher than the x actually observed. It is argued that a small p-value implies that either the null hypothesis is false or something rare has occurred. If the p-value is extremely small, say $\sim 3 \times 10^{-7}$, then of the two possibilities the most common response is to presume the null to be false. If we apply this method to the DØ top quark discovery data, and neglect the uncertainty in the null hypothesis, we find

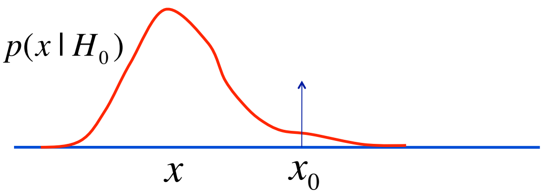

Fig. 7: The p-value is the tail-probability, $P(x > x_0|H_0)$, calculated from the probability density under the null hypothesis, H_0. Consequently, the probability density of the p-value under the null hypothesis is $\text{Uniform}(x, 1)$.

$$\text{p-value} = \sum_{N=17}^{\infty} \text{Poisson}(N, 3.8) = 5.7 \times 10^{-7}.$$

We usually report a more intuitive number by converting the p-value to the scale defined by

$$Z = \sqrt{2}\,\text{erf}^{-1}(1 - 2\text{p-value}). \tag{20}$$

This is the number of Gaussian standard deviations away from the mean[4]. A p-value of 5.7×10^{-7} corresponds to a Z of 4.9σ. The Z-value can be calculated using the `Root` function

$$Z = \texttt{-TMath::NormQuantile(p-value)}.$$

3.3.2 *Neyman's approach*

In Neyman's approach *two* hypotheses are considered, the null hypothesis H_0 and an alternative hypothesis H_1. This is illustrated in Fig. 8. In our example, the null is the same as before but the alternative hypothesis is the SM with a Higgs boson. Again, one generally chooses x so that large values would cast doubt on the validity of H_0. However, the Neyman test is specifically designed to respect the frequentist principle, which is done as follows. A *fixed* probability α is chosen called the significance (or size) of the test, which for a specific class of experiments corresponds to some threshold x_α defined by

$$\alpha = P(x > x_\alpha|H_0). \tag{21}$$

Should the observed value $x_0 > x_\alpha$, or equivalently, p-value$(x_0) < \alpha$, the hypothesis H_0 is rejected in favor of the alternative. In particle physics, in addition to applying the Neyman hypothesis test, we also report the p-value. This is sensible because there is a more information in the p-value than merely reporting the fact that a null hypothesis was rejected at a significance level of α.

[4] $\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} \exp(-t^2)\, dt$ is the error funtion.

The Neyman method satisfies the frequentist principle by construction. Since the significance of the test is fixed, α is the relative frequency with which *true* null hypotheses would be rejected and is called the **Type I** error rate.

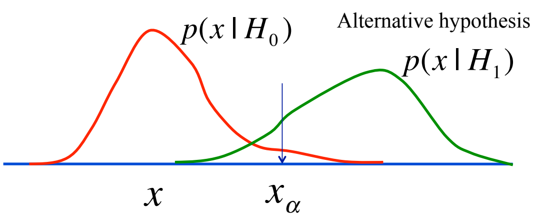

Fig. 8: Distribution of a test statistic x for two hypotheses, the null H_0 and the alternative H_1. In Neyman's approach to testing, $\alpha = P(x > x_\alpha|H_0)$ is a *fixed* probability called the significance of the test, which for a given class of experiments corresponds the threshold x_α. The hypothesis H_0 is rejected if $x > x_\alpha$.

However, since we have specified an alternative hypothesis there is more that can be said. Figure 8 shows that we can also calculate

$$\beta = P(x \leq x_\alpha|H_1), \qquad (22)$$

which is the relative frequency with which we would reject the hypotheses of the form of H_1 if they are true. These mistakes are called **Type II** errors. The quantity $1 - \beta$ is called the **power** of the test and is the relative frequency with which we would accept H_1 if true. Obviously, for a given α we want to maximize the power. Indeed, this is the basis of the Neyman-Pearson lemma (see for example Ref. [2]), which asserts that given two simple hypotheses — that is, hypotheses in which all parameters have well-defined values — the optimal statistic t to use in the hypothesis test is the likelihood ratio $t = p(x|H_1)/p(x|H_0)$.

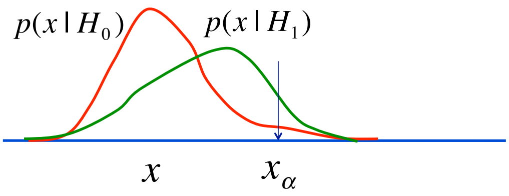

Fig. 9: See Fig. 8 for details. Unlike the case in Fig. 8, the two hypotheses H_0 and H_1 are not that different. It is then not clear whether it makes practical sense to reject H_0 when $x > x_\alpha$ only to replace it with an hypothesis H_1 that is not much better.

Maximizing the power seems sensible. Consider Fig. 9. The significance of the test in this figure is the same as that in Fig. 8, so the Type I error rate is identical. However, the Type II error rate is much greater in Fig. 9 than in Fig. 8, that is, the power of the test is considerably weaker in the former. In that case, there may be no compelling reason to reject the null since the alternative is not that much better. This insight was one source of Neyman's disagreement with Fisher. Neyman objected to the possibility that one might reject a null hypothesis regardless of whether it made sense to do so. He insisted that the task is always one of deciding between competing hypotheses. Fisher's counter argument was that an alternative hypothesis may not be available, but we may nonetheless wish to know whether the only hypothesis that is available is worth keeping. As we shall see, the Bayesian approach also requires an alternative, in agreement with Neyman, but in a way that neither he nor Fisher agreed with!

We have assumed that the hypotheses H_0 and H_1 are simple, that is, fully specified. Unfortunately, most of the hypotheses that arise in realistic particle physics analyses are not of this kind. In the Higgs boson discovery analyses by ATLAS and CMS the probability models depend on many nuisance parameters for which only estimates are available. Consequently, neither the background-only nor the background + signal hypotheses are fully specified. Such hypotheses are called **compound hypotheses**. In order to illustrate how hypothesis testing proceeds in this case, we again turn again to the top discovery example.

Example

As we saw in Sec. 3.2, the standard way to handle nuisance parameters in the frequentist approach is to replace them by their conditional MLEs and thereby reduce the likelihood function to the profile likelihood. In the top discovery example, we obtain a function $p_{PL}(D|s)$

that depends on the single parameter, s. We now treat this function as if it were a likelihood and appeal to both the Neyman-Pearson lemma, which suggests the use of likelihood ratios, and Wilks' theorem to motivate the use of the function $t(x, s)$ given in Eq. (17) to distinguish between two hypotheses: the hypothesis H_1 in which $s = \hat{s} = N - B$ and the hypothesis H_0 in which $s \neq \hat{s}$, for example, the background-only hypothesis $s = 0$. In the context of testing, $t(x, s)$ is called a **test statistic**, which, unlike a statistic as we have defined it (see Sec. 2.1), usually depends on at least one unknown parameter.

In principle, the next step is the computationally arduous task of simulating the distribution of the statistic $t(x, s)$. The task is arduous because *a priori* the probability density $p(t|s, b)$ can depend on *all* the parameters that exist in the original likelihood. If this is the case, then after all this effort we seem to have achieved a pyrrhic victory! But, this is where Wilks' theorem saves the day, at least approximately. We can avoid the burden of simulating $t(x, s)$ because the latter is approximately a χ^2 variate.

Using $N = 17$ and $s = 0$, we find $\sqrt{t_0} = \sqrt{t(17, 0)} = 4.6$. According to the results shown in Fig. (6)(a), $N = 17$ may can be considered "a lot of data"; therefore, we may use t_0 to implement a hypothesis test by comparing t_0 with a fixed value t_α corresponding to the significance level α of the test.

4 Lecture 3: the Bayesian approach

In this lecture, we introduce the Bayesian approach to inference, again using the top quark discovery data from DØ to illustrate the ideas.

The Bayesian approach is merely applied probability theory (see Section 2.2). A method is Bayesian if

- it is based on the degree of belief interpretation of probability and
- it uses Bayes theorem

$$p(\theta, \omega|D) = \frac{p(D|\theta, \omega)\, \pi(\theta, \omega)}{p(D)}, \tag{23}$$

$$\begin{aligned} &\text{where} \\ D &= \text{ observed data}, \\ \theta &= \text{ parameters of interest}, \\ \omega &= \text{ nuisance parameters}, \\ p(\theta, \omega|D) &= \text{posterior density}, \\ \pi(\theta, \omega) &= \text{prior density (or prior for short)}. \end{aligned}$$

for *all* inferences. The result of a Bayesian inference is the posterior density $p(\theta, \omega|D$ from which, if desired, various summaries can be extracted. The parameters can be discrete or continuous and nuisance parameters are eliminated by marginalization,

$$\begin{aligned} p(\theta|D) &= \int p(\theta, \omega|D)\, d\omega, \qquad (24) \\ &\propto \int p(D|\theta, \omega)\, \pi(\theta, \omega)\, d\omega. \end{aligned}$$

The function $\pi(\theta, \omega)$, called the prior, encodes whatever information we have about the parameters θ and ω independently of the data D. A key feature of the Bayesian approach is recursion: the use of the posterior density $p(\theta, \omega|D)$ or one, or more, of its marginals such as $p(\theta|D)$ as the prior in a subsequent analysis.

These simple rules yield an extremely powerful and general inference model, a model that was used, for example, in the discovery of single top quark production at the Tevatron [19,20].

4.1 Model selection

Conceptually, hypothesis testing in the Bayesian approach (also called model selection) proceeds exactly the same way as any other Bayesian calculation: we compute the posterior density,

$$p(\theta, \omega, H|D) = \frac{p(D|\theta, \omega, H)\, \pi(\theta, \omega, H)}{p(D)}, \tag{25}$$

and marginalize it with respect to all parameters except the ones that label the hypotheses or models, H,

$$p(H|D) = \int p(\theta, \omega, H|D)\, d\theta\, d\omega. \tag{26}$$

Equation (26) is the probability of hypothesis H given the observed data D. In principle, the parameters ω could also depend on H. For example, suppose that H labels different parton distribution function (PDF) models, say CT10, MSTW, and NNPDF, then ω would indeed depend on the PDF model and should be written as ω_H.

It is usually more convenient to arrive at the probability $p(H|D)$ in stages.

1. Factorize the prior in the most convenient form,

$$\begin{aligned} \pi(\theta, \omega_H, H) &= \pi(\theta, \omega_H|H)\, \pi(H), \\ &= \pi(\theta|\omega_H, H)\, \pi(\omega_H|H)\, \pi(H), \end{aligned} \tag{27}$$

or

$$= \pi(\omega_H|\theta, H)\, \pi(\theta|H)\, \pi(H). \tag{28}$$

Often, we can assume that the parameters of interest θ are independent, *a priori*, of both the nuisance parameters ω_H and the model label H, in which case we can write, $\pi(\theta, \omega_H, H) = \pi(\theta)\, \pi(\omega_H|H)\, \pi(H)$.

2. Then, for each hypothesis, H, compute the function

$$p(D|H) = \int p(D|\theta, \omega_H, H)\, \pi(\theta, \omega|H)\, d\theta\, d\omega. \tag{29}$$

3. Then, compute the probability of each hypothesis,

$$p(H|D) = \frac{p(D|H)\, \pi(H)}{\sum_H p(D|H)\, \pi(H)}. \tag{30}$$

Clearly, in order to compute $p(H|D)$ it is necessary to specify the priors $\pi(\theta, \omega|H)$ and $\pi(H)$. With some effort, it is possible to arrive at an acceptable form for $\pi(\theta, \omega|H)$, however, it is highly unlikely that consensus could ever be reached on the discrete prior $\pi(H)$. At best, one may be able to adopt a convention. For example, if by convention two hypotheses H_0 and H_1 are to be regarded as equally likely, *a priori*, then it would make sense to assign $\pi(H_0) = \pi(H_1) = 0.5$.

One way to circumvent the specification of the prior $\pi(H)$ is to compare the probabilities,

$$\frac{p(H_1|D)}{p(H_0|D)} = \left[\frac{p(D|H_1)}{p(D|H_0}\right] \frac{\pi(H_1)}{\pi(H_0)}. \tag{31}$$

and use only the term in brackets, called the global **Bayes factor**, B_{10}, as a way to compare hypotheses. The Bayes factor specifies by how much the relative probabilities of two hypotheses changes as a result of incorporating new data, D. The word global indicates that we have marginalized over all the parameters of the two models. The *local* Bayes factor, $B_{10}(\theta)$ is defined by

$$B_{10}(\theta) = \frac{p(D|\theta, H_1)}{p(D|H_0)}, \tag{32}$$

where,

$$p(D|\theta, H_1) \equiv \int p(D|\theta, \omega_{H_1}, H_1)\, \pi(\omega_{H_1}|H_1)\, d\omega_{H_1}, \tag{33}$$

are the **marginal** or integrated likelihoods in which we have assumed the *a priori* independence of θ and ω_{H_1}. We have further assumed that the marginal likelihood H_0 is independent of θ, which is a very common situation. For example, θ could be the expected signal count s, while $\omega_{H_1} = \omega$ could be the expected background b. In this case, the hypothesis H_0 is a special case of H_1, namely, it is the same as H_1 with $s = 0$. An hypothesis that is a special case of another is said to be **nested** in the more general hypothesis. The example, discussed below, will make this clearer.

There is a notational subtlety that may be missed: because of the way we have defined $p(D|\theta, H)$, we need to multiply $p(D|\theta, H)$ by the prior $\pi(\theta)$ and then integrate with respect to θ in order to calculate $p(D|H)$.

4.1.1 Priors

Constructing a prior for nuisance parameters is generally neither controversial (for most parameters) nor problematic. The Achilles heal of the Bayesian approach is the need to specify the prior $\pi(\theta)$, for the parameters of interest, at the start of the inference chain when we know almost nothing about them. Careless specification of this prior can yield results that are unreliable or even nonsensical. The mandatory requirement is that the posterior density be **proper**, that is integrate to unity. Ideally, the same should hold for priors. A very extensive literature exists on the topic of prior specification when the available information is extremely limited. However, a discussion of this topic is beyond the scope of these lectures.

For model selection, we need to proceed with caution because Bayes factor are sensitive to the choice of priors and therefore less robust than posterior densities. Suppose that the prior $\pi(\theta) = Cf(\theta)$, where C is a normalization constant. The global Bayes factor for the two hypotheses H_1 and H_0 can be written as

$$B_{10} = C\frac{\int p(D|\theta, H_1)\, f(\theta)\, d\theta}{p(D|H_0)}. \tag{34}$$

Therefore, if the constant C is ill defined, typically because $\int f(\theta)\, d\theta = \infty$, the Bayes factor will likewise be ill defined. For this reason, it is generally recommended that an improper prior not be used for parameters θ that occur only in one hypothesis, here H_1. However, for parameters that are common to all hypotheses, it is permissible to use improper priors because the constants cancel in the Bayes factor.

The discussion so far has been somewhat abstract. The next section therefore works through an example of a possible Bayesian analysis of the DØ top discovery data.

4.2 The top quark discovery: a Bayesian analysis

In this section, we shall perform the following calculations as a way to illustrate a typical Bayesian analysis,

1. compute the posterior density $p(s|D)$,
2. compute a 68% credible interval $[l(D), u(D)]$, and
3. compute the global Bayes factor $B_{10} = p(D|H_1)/p(D|H_0)$.

Probability model

The first step in any serious statistical analysis is to think deeply about what has been done in the physics analysis and construct a probability model. The full probability model is the joint probability,

$$p(x, s, b|I),$$

which, as is true of *all* probability models, is conditional on the information and assumptions, I, that define the abstract space Ω (see Sec. 2.2). In these lectures, we have omitted the conditioning data I, but it should not be forgotten that it is always present and may differ from one probability model to another.

The full probability model $p(x, s, b)$ can be factorized is several mathematically valid ways. However, we find it convenient to factorize the model in the following way,

$$p(x, s, b) = p(x|s, b)\, \pi(s, b), \tag{35}$$

where we have introduced the symbol π in order to highlight the distinction we choose to make between this part of the model and the remainder. We shall compute the likelihood from $p(x|s, b)$ and view $\pi(s, b)$ as the prior for s and b. We assume $p(x|s, b)$ to be

$$p(x|s, b) = \text{Poisson}(x, s + b). \tag{36}$$

The interpretation of $p(x|s, b)$ is clear: it is the probability to observe x events *given* that the mean event count is $s+b$. What does $\pi(s, b)$ represent? This prior encodes what we *know*, or *assume*, about the mean background and signal independently of the potential observations x. The prior $\pi(s, b)$ can be factored in two ways,

$$\begin{aligned} \pi(s, b) &= \pi(s|b)\, \pi(b), \\ &= \pi(b|s)\, \pi(s). \end{aligned} \tag{37}$$

The factorizations remind us that the parameters s and b may not be independent. However, we shall assume that they are, at least at this stage of the analysis, in which case it is permissible to write,

$$\pi(s, b) = \pi(s)\, \pi(b). \tag{38}$$

What do we know about the background? We know the count Q in the control region and we have an estimate of the control region to signal region scale factor k. Since Q is a count, a reasonable model for the likelihood is

$$p(Q|k, b) = \text{Poisson}(Q, kb), \tag{39}$$

from which, together with a prior $\pi(k, b)$, we can compute the posterior density

$$p(b|Q, k) = p(Q|k, b)\, \pi(k, b)/p(Q, k). \tag{40}$$

As usual, we factorize the prior, $\pi(k, b) = \pi(k|b)\pi_0(b)$, where we have introduced the subscript 0 to distinguish $\pi_0(b)$ from the background prior associated with Eq. (36). But, now we need to construct $\pi(k|b)$ and $\pi_0(b)$ using whatever information we have at hand.

Clearly, $b \geq 0$. But, that miserable tidbit is all we know apart from the background likelihood, Eq. (39)! Today, after a century of argument and discussion, the consensus amongst statisticians is that there is no unique way to represent such vague information. However, well founded ways to construct such priors are available, see for example Ref. [21] and references therein, but for simplicity we take the prior $\pi_0(b) = 1$, that is, the **flat prior**. If the uncertainty in k can be neglected, the (proper!) prior for k is $\pi(k|b) = \delta(k - B/\delta B^2)$, which amounts to replacing k in Eq. (40) by $B/\delta B^2$. This yields,

$$p(b|Q, k) = \text{Gamma}(kb, 1, Q + 1) = \frac{e^{-kb}(kb)^Q}{\Gamma(Q + 1)}, \tag{41}$$

for the posterior density of b, which can serve as the prior $\pi(b)$ associated with Eq. (36).

By construction, $p(x, s, b)$ is identical in form to the likelihood in Eq. (15); we have simply availed ourselves of the freedom to factorize $p(x, s, b)$ as we wish and therefore to reinterpret the factors. This freedom is useful because it makes it possible to keep the likelihood simple while relegating the complexity to the prior. This may not seem, at first, to be terribly helpful; after all, we arrived at the same mathematical form as Eq. (15). However, the complexity can be substantially mitigated by sampling from the prior so that the model is represented by the relatively simple likelihood and an ensemble of points that collectively represent the prior. The likelihood, as we have conceptualized the problem, is given by

$$p(D|s, b) = \frac{e^{-(s+b)}(s + b)^D}{D!}, \tag{42}$$

where $D = 17$ events.

The final ingredient is the prior $\pi(s)$. At this stage, all we know is that $s \geq 0$ and, again, there is no unique way to specify $\pi(s)$, though as noted there are well founded methods to construct it. We shall assume either the improper prior $\pi(s) = 1$ or the proper prior $\pi(s) = \delta(s - 14)$.

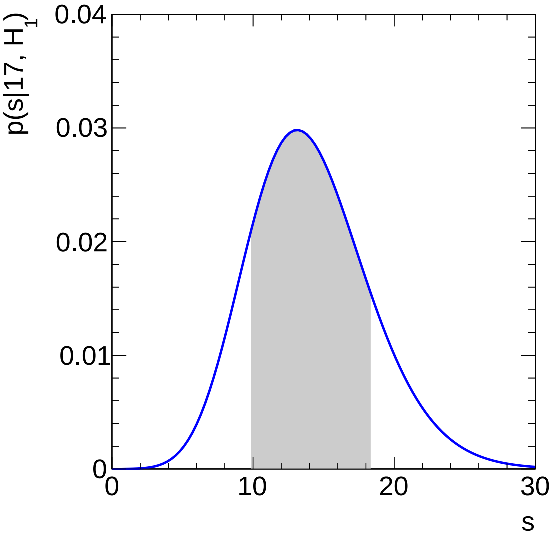

Fig. 10: Posterior density computed for DØ top quark discovery data. The shaded area is the 68% central credible interval.

Marginal likelihood

We have done the hard part: building the full probability model. Hereafter, the rest of the Bayesian analysis is mere computation.

It is convenient to eliminate the nuisance parameter b,

$$\begin{aligned} p(D|s, H_1) &= \int_0^\infty p(D|s, b)\, \pi(b) d(kb), \\ &= \frac{1}{Q}(1 - x)^2 \sum_{r=0}^{N} \text{Beta}(x, r + 1, Q)\, \text{Poisson}(N - r|s), \end{aligned} \tag{43}$$

where $x = 1/(1 + k)$,

$\boxed{\textbf{Exercise 10:}\text{ Show this}}$

and thereby arrive at the marginal likelihood $p(D|s, H_1)$.

Posterior density

Given the marginal likelihood $p(D|s, H_1)$ and a prior $\pi(s)$ we can compute the posterior density,

$$p(s|D, H_1) = p(D|s, H_1)\, \pi(s)/p(D|H_1), \tag{44}$$

where,

$$p(D|H_1) = \int_0^\infty p(D|s, H_1)\, \pi(s)\, ds.$$

Assuming a flat prior for the signal, $\pi(s) = 1$,we find

$$p(s|D, H_1) = \frac{\sum_{r=0}^{N} \text{Beta}(x, r+1, Q)\, \text{Poisson}(N - r|s)}{\sum_{r=0}^{N} \text{Beta}(x, r+1, Q)}, \tag{45}$$

Exercise 11: Derive an expression for $p(s|D, H_1)$ assuming $\pi(s) = \text{Gamma}(qs, 1, M + 1)$ where q and M are constants

from which we can compute the central **credible interval** $[9.9, 18.4]$ for s at 68% C.L., which is shown in Fig. 10. The statement $s \in [9.9, 18.4]$ at 68% C.L. means there is a 68% probability that s lies in $[9.9, 18.4]$. Unlike the frequentist statement, this statement is about this particular interval and the 68% is a degree of belief, not a relative frequency. That being said, the best Bayesian methods tend to produce credible intervals that also approximate confidence intervals.

4.2.1 Bayes factor

As noted, the number $p(D|H_1)$ can be used to perform a hypothesis test. But, as argued above, we need to use a proper prior for the signal, that is, a prior that integrates to one. The simplest such prior is a δ-function, e.g., $\pi(s) = \delta(s - 14)$. Using this prior, we find

$$p(D|H_1) = p(D|14, H_1) = 9.28 \times 10^{-2}.$$

Since the background-only hypothesis H_0 is nested in H_1, and defined by $s = 0$, the number $p(D|H_0)$ is given by $p(D|0, H_1)$, which yields

$$p(D|H_0) = p(D|0, H_1) = 3.86 \times 10^{-6}.$$

We conclude that the hypothesis $s = 14$ is favored over $s = 0$ by a Bayes factor of 24,000. In order to avoid large numbers, the Bayes factor can be mapped into a (signed) measure akin to the frequentist "n-sigma" [22],

$$Z = \text{sign}(\ln B_{10})\sqrt{2|\ln B_{10}|}, \tag{46}$$

which gives $Z = 4.5$. Negative values of Z correspond to hypotheses that are excluded.

Summary

We have given an overview of the main ideas of statistical inference in a form directly applicable to statistical analysis in particle physics. Two widely used approaches were covered, frequentist and Bayesian. Statistics is not physics. While Nature is the ultimate arbiter of which physics ideas are "correct", the ultimate arbiter of statistical ideas is intellectual taste. Therefore, we hope you take to heart the following advice.

> "Have the courage to you use your own understanding"
>
> Immanuel Kant

Acknowledgement

I thank Nick Ellis, Martijn Mulders, Helene Haller, and their counterparts from JINR, for organizing and hosting an excellent school, and the students for their energetic participation. These lectures were supported in part by US Department of Energy grant DE-FG02-97ER41022.

References

[1] L. Lyons, *Statistics for Nuclear and Particle Physicists* (Cambridge University Press, Cambridge, 1989).

[2] F. James, *Statistical Methods in Experimental Physics*, 2nd Ed. (World Scientific, Singapore, 2006).

[3] G. Cowan, *Statistical Data Analysis* (Oxford University Press, Oxford, 1998).

[4] R. J. Barlow, *Statistics: A Guide To The Use Of Statistical Methods In The Physical Sciences*, The Manchester Physics Series (John Wiley and Sons, New York, 1989).

[5] S. K. Chatterjee, *Statistical Thought: A Perspective and History* (Oxford University Press, Oxford, 2003).

[6] G. Taraldsen and B.H. Lindqvist, *The American Statistician*, **64**(2) (2010) 154.

[7] N. Suzuki, D. Rubin, C. Lidman *et al.*, *Astrophys. J.* **746** (2012) 85. [arXiv:1105.3470 [astro-ph.CO]].

[8] F. Abe *et al.* [CDF Collaboration], *Phys. Rev. Lett.* **74** (1995) 2626. [hep-ex/9503002].

[9] S. Abachi *et al.* [D0 Collaboration], *Phys. Rev. Lett* **74** (1995) 2632. [hep-ex/9503003].

[10] R. Dungan and H. B. Prosper, Varying-G Cosmology with Type Ia Supernovae, arXiv:0909.5416 [astro-ph.CO].

[11] G. Aad *et al.* [ATLAS Collaboration], *Phys. Lett. B* **716** (2012) 1. [arXiv:1207.7214 [hep-ex]].

[12] S. Chatrchyan *et al.* [CMS Collaboration], *Phys. Lett. B* **716** (2012) 30. [arXiv:1207.7235 [hep-ex]].

[13] J. Neyman, *Phil. Trans. R. Soc. London* **A236** (1937) 333.

[14] G. J. Feldman and R. D. Cousins, *Phys. Rev.* **D57** (1998) 3873.

[15] S. E. Fienberg and D. V. Hinkley, eds., *R.A. Fisher: An Appreciation*, Lecture Notes on Statistics, Vol. 1, (Germany, Springer Verlag, 1990).

[16] W. Verkerke and D. Kirkby, `RooFit`, `http://roofit.sourceforge.net`.

[17] K. Cranmer, G. Schott, L. Moneta and W. Verkerke, `RooStats`, `https://twiki.cern.ch/twiki/bin/view/RooStats`

[18] H. Jeffreys, *Theory of Probability*, 3rd Ed. (Clarendon Press, Oxford, 1961).

[19] V. M. Abazov *et al.* [D0 Collaboration], *Phys. Rev. Lett.* **103** (2009) 092001. [arXiv:0903.0850 [hep-ex]].

[20] T. Aaltonen *et al.* [CDF Collaboration], *Phys. Rev. Lett.* **103** (2009) 092002. [arXiv:0903.0885 [hep-ex]].

[21] L. Demortier, S. Jain and H. B. Prosper, *Phys. Rev. D* **82** (2010) 034002. [arXiv:1002.1111 [stat.AP]].

[22] S. Sekmen *et al.*, Phenomenological MSSM interpretation of the CMS 2011 5fb-1 results, CMS Physics Analysis Summary, CMS-PAS-SUS-12-030, CERN (2012).