



FACULTY OF SCIENCES

PHYSICS DEPARTMENT

THESIS:

Study and Application of Maximum Likelihood Limits in the Phenomenological Analysis of Particle Physics

AUTHOR:

JUAN DIEGO BERMEO ORTIZ

DIRECTOR:

CARLOS ANDRÉS FLÓREZ BUSTOS, PH.D.

Contents

List of Figures	2
List of Tables	3
1 Introduction	4
2 Objectives	6
3 Descriptive Statistics	7
3.1 Basic Properties of PDF's	7
3.1.1 Product and Sum Rule	10
3.2 Descriptive Measurements: Point values	11
3.2.1 Central Tendency	12
3.2.2 Spread: Variance	17
3.2.3 Spread: Excess Kurtosis	21
3.2.4 Skewness	22
3.3 Characteristic functions	24
3.4 Commonly used PDF's	27
3.4.1 Uniform Distribution	27
3.4.2 Binomial and Multinomial Distribution	27
3.4.3 Poisson Distribution	29
3.4.4 Gaussian Distribution	31
3.5 Central Limit Theorem (CLT)	32
3.6 Law of Large Numbers (LLN)	36
3.7 Law of Small Numbers	38
3.8 Histograms	39
4 Statistical Inference: Bayesians vs. Frequentists	41
4.1 Bayes' Theorem and its implications on both kinds of inference . .	41
4.1.1 Likelihood	42
4.1.2 Prior Probability	42
4.1.3 Marginal Likelihood	44
4.1.4 Posterior Probability	46

4.2	Bayesian Inference	48
4.2.1	Prediction intervals	49
4.2.2	Decision Theory and Maximum A Posteriori (MAP)	50
4.2.3	Bayes' Factor	55
4.3	Frequentist Inference	58
4.3.1	Estimation of parameters: Bias, Sample Variance, Consistency, Efficiency, and Maximum Likelihood	60
4.3.2	Confidence and Prediction Intervals	64
4.3.3	The logic behind Null Hypothesis Testing, p-values, and p-value controversy and caveats	66
4.3.4	Neyman and Pearson's Hypotheses Testing and Decision Theory	72
4.3.5	Likelihood Ratio and Wilk's Theorem	73
5	Statistical Inference in HEP	75
5.0.1	Hypotheses Tests using the Extended Maximum Likelihood	84
5.0.2	Extended Likelihood for Histograms or Binned Likelihood .	84
5.0.3	Nuisance Parameters and the Beeston Barlow method . . .	86
5.0.4	Profile Likelihood Ratio	88
6	Scripts using RooFit and RooStats	89
6.1	Script for the Phenomenology Group	93
7	Conclusions	96

List of Figures

1	Venn Diagram of the addition of two random variables X_1 and X_2	8
2	Gaussian distribution centered in $\mu = 2$ and with a standard deviation of $\sigma = 1$	12
3	Suitable measurements of central tendency in different types of distributions.	14
4	The tails of a Gaussian distribution. The tails are painted blue. . .	18

5	Example of a Fat Tailed distribution. The tails of a Gaussian distribution are compared to those of a fat or heavy tailed distribution	18
6	Probabilities of ranges of the population space in terms of σ . ¹	19
7	Comparison of two identical PDF's that only differ in the value of their variance.	20
8	Shapes of distributions according to the value of their kurtosis.	22
9	Visualization of skewness in a PDF.	23
10	Set of confidence intervals calculated for a parameter θ from different samples generated from the same distribution with a fixed value for the parameter. The red dotted line shows the value of the parameter, while the blue lines show the confidence intervals. ¹¹	65
11	How a p-vale looks for two of the three types of p-values for hypothesis tests. ¹²	69
12	Explanation for one of the problems of using p-values, especially to how the inference changes given a different prior probability [22].	71
13	How the different hypothesis tests for several signal points are represented in order to find regions of discovery and exclusion [4]	82
14	Flow Chart explaining roughly the process in a phenomenology study in HEP	83
15	Flow chart summarizing how to implement the shell script to carry out the hypotheses tests on a series of input histograms.	97

List of Tables

1	Common features of discrete and continuous probability distribution functions	8
2	Central tendency measurements	13

1 Introduction

In science, theories and models have to be validated experimentally, in order to be accepted as true until new evidence or data suggests otherwise. The process of experimentation or observation is one of the pillars and a key step in the scientific method [1]. The experimental observations must be quantified rigorously through probabilistic and statistical methods, in order take into account the dispersion of data and inherent uncertainty of every measurement. This permits the comparison and reproducibility of experiments, for even if they are not exactly the same, equivalence can now be tested within the calculated uncertainty. Perhaps more importantly, statistical methods allow us to contrast the values measured against what they should be, under a given hypothesis with a certain statistical significance. Statistical inference serves as an indicator of whether our hypothesis, if tested correctly, is a true indication of how nature behaves [2].

A clear example of just how important experimental validation is, is that Nobel Prizes are not awarded until there is experimental evidence that constitutes an official discovery or confirmation of a theory. It does not matter how sound or well established the theory is. An iconic example of this is the Higgs Mechanism proposed in 1964 [3], an important part of the standard model which has had a wide acceptance ever since. The Nobel Prize for the Higgs Mechanism was awarded only until 2012, when the data gathered in the LHC by the CMS [4] and the ATLAS [5] collaborations achieved enough statistical significance to establish the discovery of the Higgs Boson.

Currently in science, the way to assess and quantify inferences made from observations is through probability, specifically a branch called statistical inference². This is common to all areas of sciences, ranging from the natural sciences like physics, to social sciences like psychology. According to the two leading interpretations of probability, Frequentist and Bayesian, probability is a way to quantify how often certain specific types of processes occur or the degree of belief one has on them, respectively [6]. Of course, there are some underlying assumptions about the processes in order to facilitate the mathematical analysis [1] such as independence, dependence, randomness, and non-negativity, among others. The quantification of

²Source: <https://es.coursera.org/learn/statistical-inference>

a probability is made through the use and construction of mathematical functions called probability distribution functions (PDF). They can either be continuous or discrete and have a range between zero and one, obtained either by normalization, or deduced from the product and sum rule [1].

The present work seeks to introduce the reader to the conceptual background and rationale of the different methods applied to extract meaning from random processes, especially in science, through probability and statistics. It will be done with the following order: First, the usual methods of descriptive statistics will be explained to fully introduce the notion of a PDF both graphically and mathematically, in order to describe its key features such as shape, central tendency, and spread. Additionally, iconic and frequently used distributions will also be explained, as well as important theorems like the Central Limit Theorem, or the Law of Large Numbers. In the second part, the conceptual argument behind the methods of Bayesian and Frequentist inference will be approached, starting from how they use Bayes' Theorem and on how each understands differently the role of the data and the parameters of PDF's. This will shed light on how their statistical tests and decision theories are built.

The sections described so far are addressed to readers unfamiliar with probability and statistics, so they might seem trivial or unnecessary to the more experienced reader. If the reader is of the latter kind, please feel free to skip the sections described thus far. On the third part of this work, a closer look and emphasis will be made on the methods usually employed on High Energy Physics, (HEP), to test hypotheses relevant for this field. This hypotheses usually seek to establish the discovery of a particle, or exclusion limits for the particle's relevant variables such as its invariant mass or its cross section. Finally, we will implement these methods on a series of scripts with the software package ROOFit and RooStats, in order to leave an intuitive tool to test these two types of hypotheses for the HEP phenomenology group at Universidad de los Andes.

The most important reference in this work is [7]. The reader is strongly recommended to rely on it if at any point any explanation seems unsatisfactory, or if he or she wishes to look at a specific subject at a greater depth. Most of the information required either from probability and statistics, or about their appli-

cation on HEP, is properly summarized and explained in the reference. Although some previous knowledge on the subject is required.

2 Objectives

The main objectives of the present work are the following:

1. Write a short and easy to read introduction to probability and statistical inference.
2. The short introduction should result in a high yield in terms of understanding for the reader about the usual methods of statistical inference and their rationale.
3. Explain how statistical inference is usually conducted in High Energy Physics.
4. Write a script that has as input histograms for the background, signal, and measured data if available for a certain scattering process. The output of the script should be the p-value for discovery and for exclusion of the said scattering process.

3 Descriptive Statistics

Probability distributions functions, (PDF's), are used to describe the behavior of random variables. This mathematical functions assign a single probability to each value that the random variable might take, just like a normal function only assigns a single element of the range for each member of the domain. In other words, in the particular case of PDF's, their range are probabilities, and their domain is composed of the values the random variable can possibly take. This domain is referred to as the population space and is usually symbolized with a capital X . PDF's are characterized by their shapes, given that they provide a more evident and intuitive picture of how probability is distributed among the population space. Most descriptive measurements of a random variable are then aimed at describing the shape that its PDF takes, or the parameters it might depend on, in order to infer meaning about the random process, and the data gathered about it [2].

3.1 Basic Properties of PDF's

The basic properties of PDF's are summarized in Table 1. The PDF must be properly normalized, meaning that the sum of the probabilities of all the elements in the population space must add up to one, as seen in the second row of Table 1. On the other hand, in order to obtain the probability of a single value, we simply evaluate the PDF on that value. To find the probability of a range, we must find the probability of the sum of the elements in it, as is done in Eq. 1, which shows an example for a population space with only two elements. The term that is subtracted is due to the probability that they have of occurring at the same time. This can be understood much more intuitively by looking at the addition of the probabilities of the random variables in terms of venn diagrams, as is done in Fig. 1. For a range of multiple values, including subsets of real numbers, if the probability of every element is unrelated to the rest, then the probability of all the elements added together is the sum of the probability of each element, as is expressed in Eq. 2 and the third row of Table 1. The fact that the probability of the sum of these variables is the sum of the probability of each of them, when

they are not related, will be demonstrated later on.

$$P(x_1 + x_2) = P(x_1) + P(x_2) - P(x_1 \cap x_2) \quad (1)$$

$$P(x_1 + \dots + x_n) = \sum_i^n P(x_i) \text{ if } P(x_i \cap x_j) = 0 \text{ for all } x_i, x_j \in \{x_1, \dots, x_n\} \equiv X \quad (2)$$

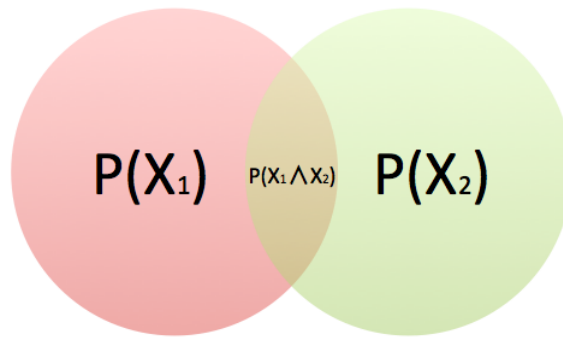


Figure 1: Venn Diagram of the addition of two random variables X_1 and X_2

	Discrete (PD)	Continuous (PdF)
PDF	$P(x)$	$p(x)dx$
Normalization	$\sum_i^n P(x_i) = 1$	$\int_{-\infty}^{\infty} p(x)dx = 1$
$P(x_0 \leq x \leq x_1)$	$\sum_{i=l}^k P(x_i)$ where $x_l \geq x_0$ and $x_k \leq x_1$	$\int_{x_0}^{x_1} p(x)dx$

Table 1: Common features of discrete and continuous probability distribution functions

As has been mentioned, PDF's can be discrete or continuous depending on the type of random process, taking the name of probability distribution (PD) and probability density function (PdF) respectively. An example of a probability

distribution is the probability for whether a coin will land on heads or tails, for there are only two possible values for the domain and range. Each possible value takes an specific probability, and hence the name probability distribution. On the other hand, an example of a Pdf is the distribution of the salaries in an industry, for they can take any value from the minimum wage to infinity. As it often happens, when going from discrete space to continuous space, our summatories become integrals as we can see in Table 1.

Furthermore, what we define as a probability and normalization changes for the continuous case, as the function $p(x)$ is multiplied by the differential dx to be considered an actual probability. The reason that the definition changes stems from the fact that in the continuous case a probability is given by the area under the curve of the Pdf. This means then, that the value $p(x)$ by itself has a measure of zero in terms of probability, because it is a point, while $p(x)dx$ is an area. That is why it takes on the name density, for a single value does not have a probability, just like a point in an object does not actually have a mass. Actual numerical values of probabilities can only be assigned to ranges of values.

The difference between $p(x)$ and $p(x)dx$ might seem pedantic or trivial, but it is not. Being rigorous with the definition allows us to appropriately apply a change of variables. Suppose we wish to find the probability of a continuous random variable, y , that is a function of a continuous random variable x as is seen in Eq. 3. If we had not taken into account the differentials, we would have omitted altogether the derivative that allows us to appropriately take into account how one random variable changes in terms of the other.

$$\begin{aligned}
 y &= f(x) \\
 p(y)dy &= p(f(x)) \frac{d(f(x))}{dx} dx = p(f(x))f'(x)dx = p(x)dx \\
 \Rightarrow p(y) &= p(x) \frac{1}{f'(x)} = p(x) \frac{dx}{dy}
 \end{aligned} \tag{3}$$

With these properties in mind, we can now take a look at what is called the cumulative distribution function (CDF). It is a function of the population space, and it tells us the probability of obtaining a value equal to or less than a value x' . This definition is clearer if we take a look at its mathematical definition in Eq. 4. We will often use this function when dealing with continuous variables, one-sided confidence intervals, and null-hypothesis tests.

$$F(x') = P(x \leq x') = \begin{cases} \int_{-\infty}^{x'} p(x) dx & \text{for PdF's} \\ \sum_{i=0}^k P(x_i) \text{ for } x_k \leq x' & \text{for PD's} \end{cases} \quad (4)$$

3.1.1 Product and Sum Rule

There are two general properties in probability theory coined by Laplace as the product and sum rules. If using them as axioms, many of the general properties of PDF's and CDF's, such as the requirement of a range between zero and one or between one and infinity, can be derived as is done in [1]. Through the rest of the document it is not required for you to know how the rules specifically imply normalization, although, it is strongly recommended as a separate reading for the sake of curiosity ³. However, we *will* use both rules in the section of statistical inference, and for this purpose we will explain what they are.

The sum rule relates the probability of an element x_i of the population space, that represents a clause, hypothesis, or value, and its complementary set x_i^C . The complementary set, or complement, is the rest of the elements of the population space, or set, aside from the element x_i . The sum rule relates the element x_i and its complement by requiring that the sum of their probabilities must add up to one, as seen in Eq. 5. Note that this implies the property of normalization stated in Table 1. If we instead use a notation, where A is a clause and \bar{A} is its negation, the sum of the probabilities can be defined as in Eq. 6. This type of notation is commonly referred to as Bool's notation.

$$P(x_i) + P(x_i^C) = 1 \text{ if } P(x_i \cap x_j) = 0 \text{ for all } x_i, x_j \in \{x_1, \dots, x_n\} \quad (5)$$

$$P(A) + P(\bar{A}) = 1 \text{ if } P(A \cap \bar{A}) = 0 \quad (6)$$

³I bet the normalization between one and infinity did the trick

The product rule on the other hand, tells us of the probability of two elements of the population space, x_i and x_j , of occurring at the same time. This is referred to as the joint probability, $P(x_i, x_j)$. The product rule states that we can separate the joint probability as the product of having one of the elements occur independently, for example $P(x_i)$, times the probability of the the other clause, x_j , given that the former one is true. The latter probability is written in the form $P(x_j|x_i)$, and is read as the probability of x_j given x_i . The converse separation also holds as shown in Eq. 7 [1]. This will lead to Bayes' Theorem as we will see in the section of statistical inference.

$$P(A, B) = P(AB) = P(A)P(B|A) = P(B)P(A|B) \quad (7)$$

3.2 Descriptive Measurements: Point values

The behavior of a random variable is usually characterized through the shape its PDF takes, and the point values that measure specific attributes of it. These attributes include how many peaks the shape has, if any, where they are located, how much is the probability spread in the population space, how symmetrical it is, and how much of the probability lies at the extremes. Among the point values that measure these characteristics of the shape are respectively: the mean, variance, skewness, and kurtosis. Skewness measures how symmetric a PDF is around its mean, while the kurtosis tells us if the tails of a PDF can be considered heavy or thin when compared to those of a Gaussian distribution. Among the shapes, perhaps the most famous is the bell-shape curve characteristic of the Gaussian distribution shown in Fig. 2. We can see that this PDF is symmetrical, that most of the probability is concentrated around its center, and values far from it have little chance of occurring. So, if a process distributes according to a Gaussian PDF, what is referred to as normally distributed, most of the values that the random variable takes will be around the center, whereas values at the extremes will be extremely unlikely and almost never seen.

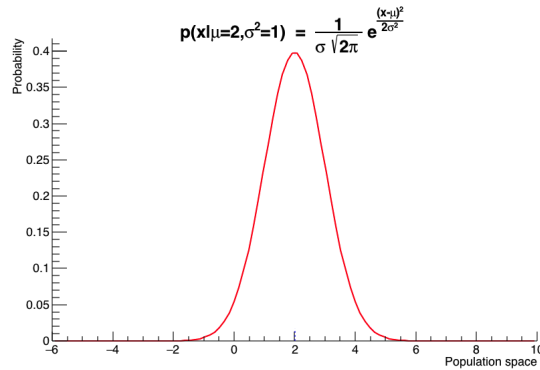


Figure 2: Gaussian distribution centered in $\mu = 2$ and with a standard deviation of $\sigma = 1$.

3.2.1 Central Tendency

The most common point values of a PDF's shape are the mean, the median, and the mode. The mode refers to the most probable value, while the median refers to the point where the probability of obtaining values greater or lesser than it, is 50%. The mean refers to the sum of each of the values weighed by the probability. In Table 2 we can take a look at how they are calculated for discrete and continuous variables. Together they provide an idea of the usual values that one might expect from a random element of the population, and it is referred to as the central tendency of the distribution.

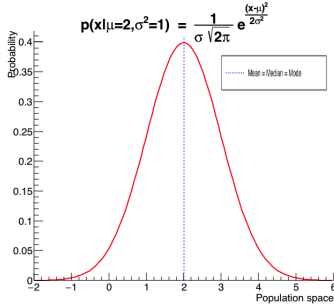
	Discrete	Continuous
Mean $\equiv \mu$	$\mu = \sum_i^n P(x_i)x_i$	$\mu = \int_{-\infty}^{\infty} p(x)x dx$
Median $\equiv \tilde{x}$	$P(x \leq x_{k-1}) \leq \frac{1}{2} \leq P(x \geq x_k = \tilde{x})$	$\int_{-\infty}^{\tilde{x}} p(x)dx = \int_{\tilde{x}}^{\infty} p(x)dx = \frac{1}{2}$
Mode $\equiv M_0$	$M_0 : P(M_0) = \sup_{x \in X} P(x)$	$M_0 : p(M_0)dx = \sup_{x \in X} p(x)dx$
$E[x^k] \equiv \langle x^k \rangle$	$\sum_i^n P(x_i)x_i^k$	$\int_{-\infty}^{\infty} p(x)x^k dx$

Table 2: Central tendency measurements

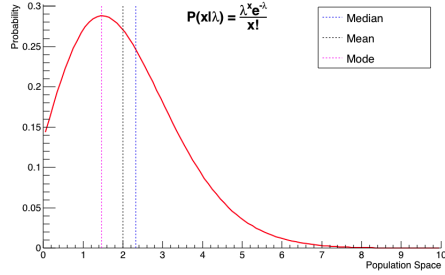
The fourth row refers to a measurement of the central tendency of the random variable elevated to the k th power. It is referred to as the expected value of x^k or $E[x^k]$, and it is an extension of the notion of a the mean, but instead of having $k = 1$, it is for any exponent k . It is not to be confused with an expansion of the PDF $p(x)dx$ or $P(X)$ in terms of several polynomials, given that the summatory in the discrete case is done over the index i and the integral over the variable x , so k can only take one value at a time. An expected value $E[x^k]$ with $k > 4$ do describe the distribution, but their qualitative meaning is more difficult to interpret, and thus are not usually named or used.

These expected values are referred to as the moments of a distribution. The collections of all moments uniquely characterize a PDF that is bounded. That a PDF is bounded means that none of the expected values of any order diverge. When a expected value is calculated only in terms of x^k , it is called a raw moment and symbolized by m_k . When it is defined as in Eq. 8, it is called a central moment. Central moments are used much more frequently than raw moments for they describe the shape of the PDF regardless of their absolute location. The absolute location is pinpointed by the PDF's mean. The second central moment is the variance, the third and fourth central moments, normalized over the square root of the second moment, are the skewness and excess kurtosis respectively.

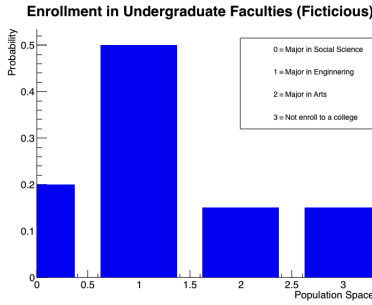
$$\mu_k \equiv E[(x - \mu)^k] \quad (8)$$



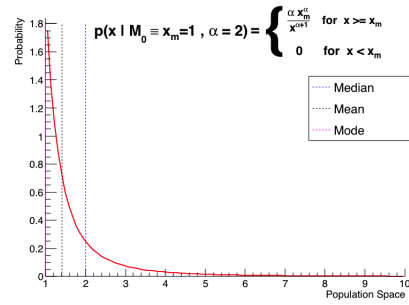
(a) Gaussian Distribution



(b) Poisson Distribution



(c) Nominal Variable Distribution



(d) Pareto Distribution

Figure 3: Suitable measurements of central tendency in different types of distributions.

There are three measurements of central tendency because, depending on the type of PDF, one of them will work better as the indicator of it. In Fig. 3 we can see specific examples about which of the measurements provide a better sense of the central tendency for specific PDF's, and where the notion of central tendency does not even make sense. In each case, both the mathematical expression and shape are shown. For the Gaussian distribution the appropriate central tendency is the mean, while for the Poisson distribution it is the median. On the other hand, for Nominal variables and the Pareto distribution shown, there is not a true central tendency.

All of the PDF's shown in the figure are unimodal, meaning that they have a single mode, or put simply, a single peak. In Fig. 3(a) we have a Gaussian distribution, that as we stated earlier is symmetrical. Being symmetrical and unimodal implies that the mean, mode, and median are equal [8]. This is an example of how these three measurements can tell us about a significant characteristic of the shape, given that by themselves they can discard the possibility that a PDF is symmetrical and unimodal.

In Fig. 3(b) we can see that the PDF is highly asymmetrical, and thus the three measurements take different values. In this case the mean and mode are not good measurements of central tendency, for they are too skewed to the right of the figure. For PDF's that are highly skewed, the median provides the better sense of which is the tendency of the values of the PDF ⁴. Regarding Fig. 3(c), in it we are dealing with the distribution of nominal variables. It is a fictitious PDF of into which type of undergraduate majors will students from a hypothetical high school enroll. Given that we cannot assign an order or multiply probabilities by the names of our nominal variables, measurements like the mean or median actually have no sense or cannot be calculated in this case. Because of this, the mode is the only indication we have of the general behavior of the random variable.

Finally, for Fig. 3(d), we have a Pareto PDF with an exponent of 2. This is a very particular kind of PDF, for it has a rather wild behavior. With the exponent it has, the variance of the distribution diverges, and there is not a convincing central tendency as can be seen qualitatively in the figure. One might think that this type of PDF's are irregular or that they do not reflect many actual random processes. Quite to the contrary, Pareto distributions are very useful to model how wealth is distributed in societies, or to describe characteristics of social networks. For instance these PDF's fit very well the way the number of followers in Twitter is distributed per user ⁵. They actually have a particular point measurement called the Gini coefficient, that measures the variability or lack of uniformity of probability with respect to the values of the population space [9]. In other words, it provides a measurement of inequality, and is used for such purpose. The Gini co-

⁴Source: <https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php>

⁵Source: <http://radar.oreilly.com/2013/12/tweets-loud-and-quiet.html>

efficient is not exclusive of Pareto distributions, and can be used for other PDF's. Pareto distributions are characteristic of random variables in which the probability of an extremely high value is not that unlikely or negligible, compared to what one might expect of a Gaussian distribution for instance.

In the case of wealth, there are a few hundred billionaires, that amass wealth 6 to 7 orders of magnitude higher than the rest of the world. Actually, as of 2015, the top 1% of the human population amasses more wealth than the remaining 99%⁶. On Twitter, the number of followers resembles a Pareto distribution because there are celebrities that have a considerably higher number of followers than the normal Twitter user. In both cases it means that in average, everyone else has more Twitter followers or money than you. In the case of Twitter, the average number of followers is around 443, but only the top 10% of users have over 450 followers. The fact that the mean does not represent with a high probability an element selected a random, what is colloquially referred to as an average user or element, seems like a nonsensical statement that contradicts the very notion of a mean. What happens is that the celebrities and billionaires become outlier values with a non-negligible chance of occurring, that drag the mean to a value where it does not represent the tendency or behavior of most of the elements in the population space, and thus the idea of a central tendency tends to become meaningless for this types of distributions.

As for the median, it is indeed a better measurement for a central tendency than the mean, given that it is not as widely affected by outliers. For example the median number of followers in Twitter is 61, a much more reasonable value. However, the distribution is extremely resulting in that a very large portion of the probability is concentrated around the small values, and smooths towards higher values. This implies that its validity as a central tendency diminishes as the exponent of the distribution becomes smaller. This was the case for Fig. 3(d), where most of the probability is concentrated around the smallest values.

Pareto distributions provide a clear example of why point measurements should not be reported by themselves as is often taught, because without knowing the type of distribution, they might lack meaning and be misleading. The fact that

⁶Source:<http://www.theguardian.com/money/2015/oct/13/half-world-wealth-in-hands-population-inequality-report>

the mean yearly income of humans is \$10,721 USD⁷, even though 71% of the population makes less than that⁸, does not mean that the mean is incorrectly calculated, it means that it does not make sense for this type of distributions. Statements like "the average human has one testicle and one ovary" are in fact true, but meaningless nonetheless, for it does not represent a random human [2]. When using and reporting point measurements, one should always report the PDF that it represents, or its shape when the PDF is unknown, and why that PDF describes adequately the random variable or process, so as to not mislead and misrepresent it. Any subsequent inferences about the random variable depend on the initial assumptions of why the PDF describes it. These assumptions and justifications usually change on a case to case basis. If the assumptions are wrong, or fail to encompass relevant features of the process, then the inferences drawn will be wrong from the very beginning.

3.2.2 Spread: Variance

The second characteristic to describe a PDF is through measurements of how much is the probability spread in the population space. The two most common measurements for this are the variance, that tells us of the distance of all points to the mean of the distribution; and the excess kurtosis, that measures how probable tail values are in comparison to what they would be in a Gaussian distribution. Tail values refer to elements that lie far from the mean at the extremes of certain distributions, that visually seem like tails as shown in Fig. 4.

⁷Source: <http://knoema.es/sijweyg/gdp-per-capita-ranking-2015-data-and-charts>

⁸Source: <http://inequality.org/global-inequality/>

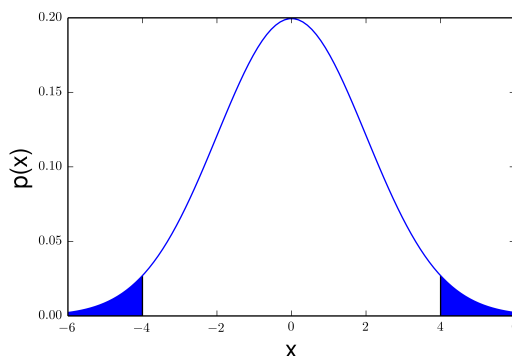


Figure 4: The tails of a Gaussian distribution. The tails are painted blue.

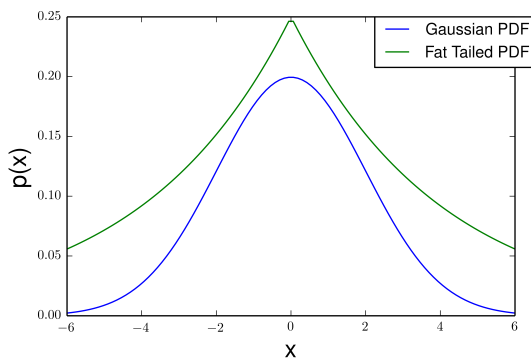


Figure 5: Example of a Fat Tailed distribution. The tails of a Gaussian distribution are compared to those of a fat or heavy tailed distribution

The variance is a fairly common measurement that is almost as ubiquitous as the mean, and it is usually represented with the symbol σ^2 . Almost all measurements in experiments from freshmen undergraduate lab courses include it. This is because the measurements usually taken vary similarly to the normal distribution. In this case, and many others, the square root of the variance provides an accurate measurement of the error or uncertainty, and is referred to as the standard deviation, $\pm\sigma$. The standard deviation establishes a bound on terms of probability

for the population space, as can be appreciated in Fig. 6. Note that the standard deviation σ has the same dimension or units as its random variable X , so it is usually preferred by physicists to present the measurement of uncertainty. The range of values $(\mu - \sigma, \mu + \sigma)$ is a much more intuitive bound for the error around the actual value μ , than simply using σ^2 .

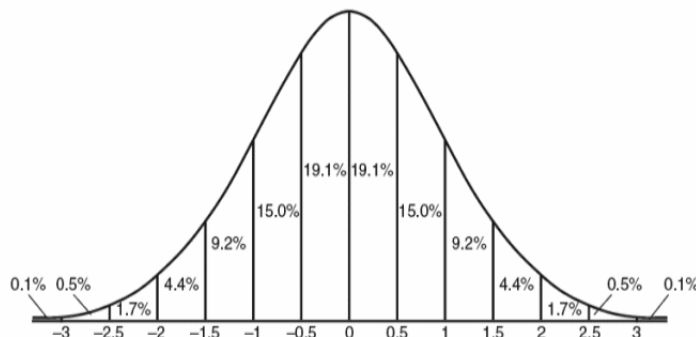


Figure 6: Probabilities of ranges of the population space in terms of σ .⁹

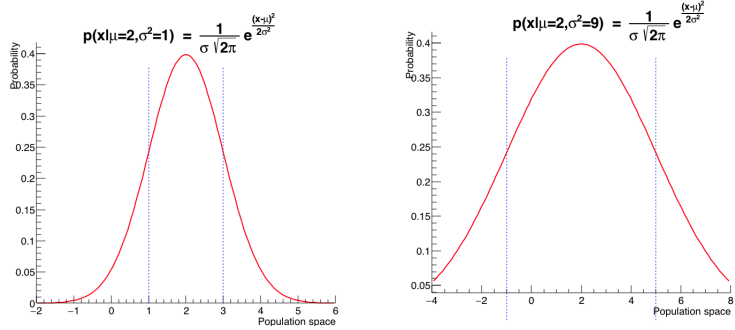
From Fig. 6 we can tell that values further than one standard deviation to either side of the mean have a probability of only 31.8%, while values 2σ away from it have only 4.6%¹⁰. Values further than 5σ have an staggering $3 \cdot 10^{-5}\%$ chance of occurring! The caveat should be made though, that the standard deviation is a good measurement of uncertainty only when values away from the mean decrease in terms of probability similarly to the Gaussian distribution. If this condition is not met, values further than one standard deviation will be seen more often as a result of chance or uncertainty, and thus σ would not be a proper bound in terms of probability.

In other words, the variance tells us how likely it is that we find measurements far from the mean and, in many cases, how wide is the shape of the PDF around it, as in Fig. 7. One can see that a larger variance, Fig. 7(b), visually results in a wider shape of a PDF than in Fig. 7(a). The variance is mathematically defined

⁹The image was taken from <http://goo.gl/gxa1Do>

¹⁰Source: <http://www.regentsprep.org/regents/math/algtrig/ats2/normallesson.htm>

as shown in Eq. 9. In terms of discrete and continuous PDF's, it is specifically calculated as shown in Eq. 10. A low variance means that our random variable is highly concentrated around the mean. A high variance means that either the range of values that are likely around the mean is large compared to the mean itself, or that there are some values extremely far from the mean, compared to its magnitude, with a non negligible probability.



(a) Normal distribution with $\mu = 2$
and $\sigma^2 = 1$

(b) Normal distribution with $\mu = 2$
and $\sigma^2 = 9$

Figure 7: Comparison of two identical PDF's that only differ in the value of their variance.

$$\sigma^2 \equiv \mathbb{E}[(x - \mu)^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \quad (9)$$

$$\mathbb{E}[(x - \mu)^2] = \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2$$

$$\mathbb{E}[(x - \mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

$$\sigma^2 = \begin{cases} \sum_{i=0}^n P(x_i)(x_i - \mu)^2 = \sum_{i=0}^n P(x_i)x_i^2 - \mu^2 \\ \int_{-\infty}^{\infty} p(x)x^2dx - \left(\int_{-\infty}^{\infty} p(x)xdx\right)^2 = \int_{-\infty}^{\infty} p(x)x^2dx - \mu^2 \end{cases} \quad (10)$$

3.2.3 Spread: Excess Kurtosis

The kurtosis, or excess kurtosis, tells us of the distribution's tail behavior, and thus of the probability of outliers in it. Recall that outliers are the values that are further from the mean and from where you would normally expect to see instances of the random variable. Put simply, they are the values that lie at the tail of the distribution. There is not a consensus for where does a tail actually start, or from what value is an element considered an outlier. The definition is based on heuristics and on a case to case basis. An example of a heuristic to define outliers, is that all values that lie a distance greater than $1.5 \cdot (x_{75\%} - x_{25\%})$, where $P(x \leq x_{25\%}) = 25\%$ and $P(x \leq x_{75\%}) = 75\%$, are outliers [2].

The excess kurtosis is particularly useful, for it allows us to identify the existence of heavy tails in a particular distribution, given that it is often tricky to identify this by inspection. It is important whether or not a distribution is heavy tailed, because it means that outlier values are not as unlikely, and that given enough attempts or measurements of the process, they can occur. Failing to identify heavy tails, as has happened in the estimation of the distribution of risk in investments, can have costly consequences. In the financial crisis of 2008, the risk of many assets were wrongfully considered to follow a Gaussian PDF when in reality they had heavy tails. As a result, several credits and packages of mortgage debt that were assumed to be safe, turned into toxic assets and contributed greatly to the Housing Bubble [10].

The expression to calculate the excess kurtosis, is that shown in Eq. 11 and Eq. 12. The adjective excess is used because the kurtosis of PDF's is compared to that of the Gaussian distribution, that has $\kappa = 3$. The tails of the Gaussian distribution serve then as a gauge of whether a tail is heavy or thin as shown in Fig. 8. A positive excess kurtosis implies the tail is thin, whilst a negative value implies the tail is heavy or fat.

$$\kappa \equiv \frac{\mathbb{E}[(x - \mu)^4]}{\sigma^4} - 3 = \frac{\mathbb{E}[(x - \mu)^4]}{(\mathbb{E}[(x - \mu)^2])^2} - 3 \quad (11)$$

$$\kappa = \begin{cases} \frac{\sum_{i=0}^n P(x_i)(x_i - \mu)^4}{\left(\sum_{i=0}^n P(x_i)x_i^2 - \mu^2\right)^2} - 3 \text{ for PD's} \\ \frac{\int_{-\infty}^{\infty} p(x)(x - \mu)^4}{\left(\int_{-\infty}^{\infty} p(x)x^2 dx - \mu^2\right)^2} - 3 \text{ for PdF's} \end{cases} \quad (12)$$

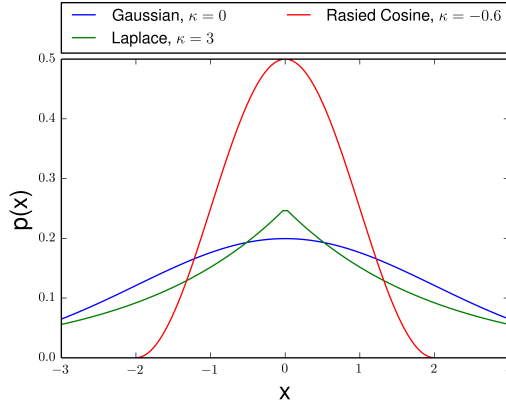


Figure 8: Shapes of distributions according to the value of their kurtosis.

3.2.4 Skewness

This characteristic refers in most cases to how asymmetrical a PDF is around the mean. It is a quantitative measurement that can take on both negative and positive values. To better grasp its meaning, let's take a look at Fig. 9, where we can see that a positive skewness means that most of the probability is concentrated towards values more positive than the mean. A negative skewness means the probability is more concentrated to the left of the mean. It is usually stated that skewness implies whether the mean is larger or not than the mean. However this only applies to certain PDF's, but does not generalize to all of them.

There is some ambiguity as to whether the statement of more skew implies if the tail will be fatter, or a long tail with non-negligible probability like those of Pareto distributions. The answer is that it does not imply anything or differentiate

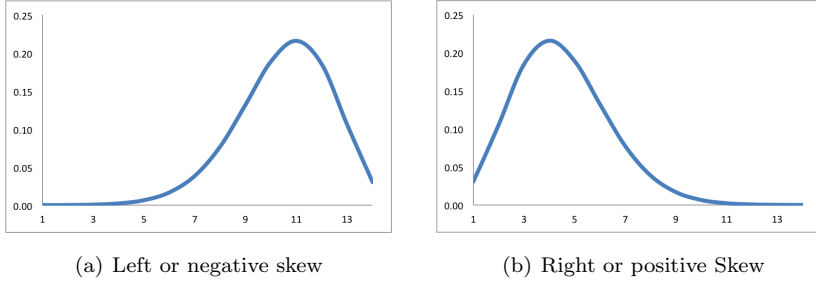


Figure 9: Visualization of skewness in a PDF.

between types of tails. In fact, it could happen that the fatness of one side evens out with how long the other tail is, and end up with a skewness of zero, and a highly asymmetrical figure. However, that is a rather unlikely case. The skewness is mathematically defined as shown in Eq. 13, and calculated as is done in Eq. 14

$$\gamma \equiv \frac{E[(x - \mu)^3]}{\sigma^3} \quad (13)$$

$$\gamma = \begin{cases} \frac{\sum_{i=0}^n P(x_i)(x_i - \mu)^3}{\left(\sum_{i=0}^n P(x_i)(x_i - \mu)^2\right)^{3/2}} & \text{for PD's} \\ \frac{\int_{-\infty}^{\infty} p(x)(x - \mu)^3}{\left(\int_{-\infty}^{\infty} p(x)x^2 dx - \mu^2\right)^{3/2}} & \text{for PdF's} \end{cases} \quad (14)$$

The skewness and excess kurtosis are what are called normalized central moments. They have been normalized by σ^k and relocated with its mean so that they are location-scale independent. This implies that these measurements will be comparable between distributions regardless of the location of their mean, or how wide they are in terms of the scale given by its standard deviation. Defining them this way allows for a standardized scale under which we can compare distributions, facilitating the analysis of a given PDF based on our previous experience.

3.3 Characteristic functions

There are additional ways to analyze analytically the behavior of a random variable aside from its PDF. These other approaches include the characteristic function, the cumulants, and the moment generating function [11]. The characteristic function is particularly important, for through it we can find the probability that the addition of random variables follow much more easily than through the convolution of their PDF's. Also, it becomes useful in one of the proofs of the Central Limit Theorem (CLT), as well as the Law of Large Numbers (LLN), and the Law of Small Numbers (LSN). However, how we find the characteristic function is usually through the PDF as shown in Eq. 15 and Eq. 16. The integral and summatory of Eq. 16 are called the Fourier Transform for the discrete and continuous case respectively. This transformation can be inverted, and thus, given the characteristic function, we can obtain the PDF through Eq. 17

$$\phi_x(t) = E[e^{itx}], \quad (15)$$

$$\phi_x(t) = \begin{cases} \sum_{i=0}^n P(x_i)e^{itx} & \text{for PD's} \\ \int_{-\infty}^{\infty} p(x)e^{itx}dx & \text{for PdF's} \end{cases} \quad (16)$$

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(t)e^{itx}dt \quad (17)$$

Usually the way to calculate the PDF of the addition of two random variables, $Z = X + Y$, is through an operation known as convolution, shown in Eq. 18. Usually this integral can be cumbersome to define and calculate. This is where the characteristic function becomes really useful by taking advantage of a property of the fourier transform, that states that the convolution of two variables is the multiplication of their fourier transforms. Applied to PDF's, this means that the characteristic function of the addition of variables is the multiplication of their characteristic functions as shown in Eq. 19. So if we wish to find what PDF does the addition of variables follow, we only have to find the multiplication of their characteristic functions and then invert it back using Eq. 17. For instance, using this property the random variable $Z = a + bX$ yields Eq. 20.

$$p(Z) = p(X) * p(Y) = \begin{cases} \sum_{k=-\infty}^{\infty} P(X = k)P(Y = Z - k) \\ \int_{-\infty}^{\infty} p(x = t)p(y = z - t)dt = \int_{-\infty}^{\infty} p(x = z - t)p(y = t)dt \end{cases} \quad (18)$$

$$\phi_{x_1+\dots+x_n}(t) = \prod_{i=1}^n \phi_{x_i}(t) \quad (19)$$

$$\phi_{a+bX}(t) = \int_{-\infty}^{\infty} p(x)e^{itbx}e^{iat}dx = e^{iat} \int_{-\infty}^{\infty} p(x)e^{it(bx)}dx = e^{iat}\phi_X(bt) \quad (20)$$

The characteristic function is related to two other analytical tools that provide different ways to calculate expected values, that in some cases might be more efficient and easier than the method shown in Table 2. These tools are the cumulant and the moment generating function. Cumulants are a quantity closely related to the moments of a function. The first three cumulants (the mean, variance and skewness) are equal to the first three central moments. From the fourth cumulant onward they are not the same, but they are related. For instance, the fourth cumulant is $\kappa_4 = \mu_4 - 3\mu_2^2$ and the fifth is $\kappa_5 = \mu_5 - 10\mu_3\mu_2$. The main advantage of using cumulants of independently distributed variables, instead of their moments, is that cumulants are additive. That cumulants are additive means that one does not have to recalculate the cumulant for the distribution of the sum of random variables, just add their cumulants directly.

A cumulant is obtained and defined from the Taylor Series expansion of the natural logarithm of the characteristic function as seen in Eq. 21 and Eq. 22. The coefficient of the n th term of the Taylor series is what is defined as the n th cumulant, κ_n , of the distribution [11]. Because of the definition of a cumulant, it can be calculated by taking the n th derivative of $\log(\phi_x(t))$, and then setting $t = 0$ as is done in Eq. 22. This is why the term $\log(\phi_x(t))$ is often referred to as the cumulant generating function, $K_X(t)$, of the random variable X , for through its derivatives we can calculate any cumulant we want.

$$K_X(t) \equiv \log(\phi_X(t)) = \sum_{n=1}^{\infty} \frac{d^n(\log \phi_X(0))}{dt^n} \frac{(it)^n}{n!} \quad (21)$$

$$\kappa_n \equiv \frac{d^n(\log \phi_X(0))}{dt^n} \quad (22)$$

As for the moment generating function, it is usually defined as the characteristic function of the random variable ix , where $i \equiv \sqrt{-1}$. However in the present work we will define it as equivalent to the characteristic function just like Eq. 23 shows. The end result will be the same except for a factor of i^{-n} . The steps from Eq. 24 to Eq. 26 show that a moment m_n can be calculated up to a factor of i^{-n} by using the n th derivative of the characteristic function and setting $t = 0$ after taking the derivative. This of course is equivalent to finding the coefficients of the Taylor Series expansion of the characteristic function. Taking either of the approaches, that technically end up in the same procedure, allows us to calculate the n th moment of the PDF that describes the random variable X as Eq. 27 shows.

$$M_X(t) = \exp(K_X(t)) = \mathbb{E}[e^{itX}] = \phi_X(t) = \int_{-\infty}^{\infty} p(x)e^{itx} dx \quad (23)$$

$$\frac{d^n(M_X(t))}{dt^n} = \frac{d^n}{dt^n} \int_{-\infty}^{\infty} p(x)e^{-tx} dx = \int_{-\infty}^{\infty} p(x) \frac{d^n(e^{itx})}{dt^n} dx \quad (24)$$

$$\frac{d^n(M_X(t))}{dt^n} = i^n \int_{-\infty}^{\infty} p(x)x^n e^{-tx} dx \quad (25)$$

$$\text{if } t \rightarrow 0 \quad \frac{d^n(M_X(0))}{dt^n} = i^n \int_{-\infty}^{\infty} p(x)x^n dx = i^n \mathbb{E}[x^n] \quad (26)$$

$$\Rightarrow m_n = \mathbb{E}[x^n] = i^{-n} \frac{d^n(M_X(0))}{dt^n} = i^{-n} \frac{d^n(\phi_X(0))}{dt^n} \quad (27)$$

3.4 Commonly used PDF's

3.4.1 Uniform Distribution

This one of the most frequently used distributions and the one that is usually taught first in probability courses, for most people are already familiar with it. The Uniform Distribution assigns the same probability to every element of the probability space, whether it is discrete or continuous. It is frequently used when nothing is known about the population space as a toy distribution or starting model. For instance, it is employed to model the probabilities of a coin of landing heads or tails, or of obtaining a certain face of a fair dice. The PD for the discrete cases is then the one shown in Eq. 28. As for the continuous case, its PdF is that shown in Eq. 29. Notice that this PdF is not scale invariant, for a change in scale would shrink or increase the probability for every element in the population space.

$$P(X_i) = \frac{1}{n} \quad \forall X_i \in (X_1, \dots, X_n) \quad (28)$$

$$P(x)dx = \frac{1}{b-a}dx \quad \forall x \in (a, b) \quad (29)$$

3.4.2 Binomial and Multinomial Distribution

The Binomial distribution is used for events that have only two possible outcomes, and it can answer the question of what is the probability of obtaining a number k of one of these two outcomes, given that n attempts were made. It is a discrete PDF, and has the mathematical form shown in Eq. 30. The term p refers to the probability that the outcome has of happening one time. The first term in the expression is known as the combination of n in k , and it accounts for all the combinations of the sequence of n outcomes in which the outcome with probability p occurs k times, while the other possible outcome occurs $n - k$ times. The probability for the other outcome is often expressed by a probability q , but given that we only have two possible outcomes, by the sum rule we have that $q = 1 - p$.

$$P(X \equiv k) = \binom{n}{k} p^k q^{n-k} = \binom{n}{k} p^k (1-p)^{n-k} \equiv B(n, p) \quad (30)$$

$$P(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \equiv B(n, p) \quad (31)$$

Note that this is the probability of obtaining the entire sequence, but for *each* of the attempts the chance for one of the outcomes remains p as well as $1-p$ for the other outcome. This is regardless of the sequence. Thinking otherwise is known as the gambler's fallacy, that in its more general form represents the belief that previous results will affect future outcomes. In the context of the binomial distribution, the fallacy takes form when one assumes that because a certain sequence has a given chance of occurring, then the previous results in a certain sequence should affect future ones, making a future outcome more or less likely. This belief is based on the probability of k number of outcomes in a given sequence. This is not the case. Each result that goes into the sequence is independent of the rest, and their probabilities are fixed to p and $1-p$. Systems that follow binomial distributions do not have memory, and their probabilities are not affected by previous or future outcomes.

This type of distribution can be extended for a discrete and finite number of outcomes in what is known as the multinomial distribution. This PDF is used for when have k independent outcomes and we want to know the probability of the i th outcome of occurring an x_i number of times in a sequence. In this case, instead of having the probabilities p and q , we have the set of probabilities (p_1, \dots, p_k) for each outcome. Since it is assumed that there is no dependence between outcomes, by the product rule their joint probability should depend on the productory of $p_i^{x_i}$ in order to account for the probability p_i of each outcome and on the number of times x_i it appears in the sequence as shown in Eq. 32.

Notice that we still have to take into account all the possible combinations in which the sequence may occur, just like we did for the binomial distribution with the binomial coefficient. This is carried out with the remaining terms shown in Eq. 32 that make use of the gamma function. The gamma function is an analytic extension of the factorial function with the mathematical form seen in Eq. 33. The gamma function can be used to calculate the factorial for real numbers, and in this

case serves the same purpose of allowing us to count all possible combinations in which the i th term occurs x_i times in the sequence by taking the ratio expressed.

$$P(x_1, \dots, x_k | p_1, \dots, p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^k p_i^{x_i} \quad (32)$$

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx \quad (33)$$

3.4.3 Poisson Distribution

First of all, the mathematical representation of this distribution is the following:

$$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (34)$$

This is a continuous PDF that depends only on the number of attempts x , and a mean rate of occurrence λ over a certain unit, usually time or space, at which the types of events occur. This rate is assumed to be constant, and the units of both λ and x must match for the description of the PDF to be valid. The variance of this Pdf is also λ . The probability of any event is independent of any other that might have happened at other instances, meaning it behaves as a system without memory just like the Binomial Distribution.

The Poisson distribution is often referred to as a Binomial or Bernoulli Process in continuous space. This is true if the number of attempts is sufficiently large and the probability of each trial is small enough. This is referred to as the Law of Rare Events or Law of Small Numbers. There are two ways to show this, one is starting directly with a binomial distribution and taking n intervals in which the probability of having more than one event inside the interval is approximately zero, and thus each interval will have rate of occurrence equal to $\frac{\lambda}{n}$. This means we would have a binomial distribution, $B(n, p = \frac{\lambda}{n})$ that as n tends to infinity becomes a Poisson distribution. The other option to prove it is by using the characteristic function of a Binomial process and arriving at the characteristic function of the Poisson Distribution. This approach will be shown on a later subsection. For now, let us take a look at the former way step by step.

First off, the expected value for a Binomial distribution, with n events and probability p of occurrence for the outcome of interest, is equal to np . If we equate this to a mean rate of λ , we have then that $p = \frac{\lambda}{n}$. If n is sufficiently large, we can take n sub intervals such that for each interval it is extremely unlikely to obtain more than one event, and so the maximum number of possible events is n . So we effectively have the following binomial distribution

$$P(x) = \frac{n!}{(n-x)!x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \quad (35)$$

$$P(x) = \frac{(n)(n-1)\dots(n-x+1)}{x!} \frac{\lambda^x}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x} \quad (36)$$

Given that there is a number x of elements in the numerator of the first factor, where x is the number of attempts whose probability of occurring we wish to know, we can factorize a denominator of n for each element from n^x , effectively taking the following form:

$$P(x) = \frac{(n)}{n} \frac{(n-1)}{n} \dots \frac{(n-x+1)}{n} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \quad (37)$$

If we take the limit when n tends to infinity we have that all this factors will converge to one by L'Hôpital's rule. Also by doing the same to the other two factors and taking the Taylor Expansion of one of them as done in Eq. 39, we will arrive at the Poisson distribution.

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{-x} = (1 - 0)^{-x} = 1 \quad (38)$$

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = \lim_{n \rightarrow \infty} \left(\sum_{i=0}^{\infty} \frac{\lambda^n}{n!} \frac{d^n}{d\lambda^n} \left[\left(1 - \frac{\lambda}{n}\right)^n \right]_{\lambda=0} \right) \quad (39)$$

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = \lim_{n \rightarrow \infty} \left(1 + \frac{\lambda - n}{1!} \left(1 - \frac{0}{n}\right)^{n-1} + \frac{\lambda^2}{2!} \frac{(n-1)(n)}{n^2} \left(1 - \frac{0}{n}\right)^{n-2} + o(\lambda^3) \right) \quad (40)$$

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = 1 - \frac{\lambda}{1!} + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + o(\lambda^4) = e^{-\lambda} \quad (41)$$

By putting together all this, we arrive then at the result that we were after:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (42)$$

3.4.4 Gaussian Distribution

This is by far the most widely used distribution, and as we have seen, it serves as the frame of reference on which the interpretation of point value measurements is built and justified. The Pdf of the Gaussian distribution is shown in Eq. 43. It depends only on the mean μ and its variance σ^2 , and because of that, the usual notation for its Pdf, $p(x, |\mu, \sigma^2)$, is often abbreviated as $N(\mu, \sigma^2)$. As has already been said, it is unimodal and symmetric around its mean. Additionally, all moments of the function higher than four are equal to zero. This means that the distribution is completely determined by its first four moments (mean $\equiv \mu$, variance $\equiv \sigma^2$, skewness = 0, and kurtosis $\equiv \kappa = 3$).

The Gaussian distribution is frequently used to model errors, and as such it is sometimes referred to as the error function. For instance the least squares method that is used for linear regressions seeks to minimize the squared error, because it is assumed that the error distributes Gaussian. The most commonly taught methods for optimization are aimed at quadratic forms, which is the form that the Gaussian distribution takes in its multivariate form.

$$p(x, |\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (43)$$

Almost everything seems to be addressed in statistical inference to the Gaussian distribution, because as has already been said, it is the most frequently used. This of course is not entirely a coincidence. While sometimes things are assumed to distribute Gaussian as a starting model because it is simply easier given all the mathematical tools that have been developed so far to analyze it, it is also due to the Central Limit Theorem. This theorem will be explained in the next section, but it roughly states that when taking samples from a set of random variables that are identically and independently distributed, for many families of PDF's, their arithmetic mean will tend to distribute Gaussian as the number of samples grows to infinity.

3.5 Central Limit Theorem (CLT)

It is one of the most prominent theorems of probability theory and among the most widely used and applied. It states that the arithmetic mean \bar{X} of a set of n independent identically distributed (iid) random variables, such as $\mathbf{X} \equiv \{X_1, \dots, X_n\}$, where each X_i has a mean μ that does not diverge and a positive variance σ^2 that also does not diverge, will converge asymptotically in probability to a normal distribution with a mean μ and a variance $\frac{\sigma^2}{n}$, written as $N(\mu, \frac{\sigma^2}{n})$. This means that the arithmetic mean of the sample \mathbf{X} of size n , defined in Eq. 44, approximately bears a single point of the Gaussian distribution $N(\mu, \frac{\sigma^2}{n})$. In order to be able to plot the shape of this normal distribution, a k number of samples \mathbf{X} of size n must be taken to obtain k points. With these k points we can build a histogram by defining an appropriate number of bins, and with the relative frequency of each bin given by the number of points that fall within the range of the bin, we can plot the shape and verify that it indeed does resemble a Gaussian Distribution.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (44)$$

So far this is not the actual and formal form of the theorem, but just a practical description of it. Let us take a look at the previous description step by step. By identically distributed, we mean that all the $X_i \in \mathbf{X}$ have exactly the same PDF with the same parameters. By independent, it is meant that the probability of any sample X_i of occurring is independent of any value taken by another sample X_j . On the other hand, asymptotic convergence means that as the number n tends to infinity, the PDF of the arithmetic mean of sample \mathbf{X} converges in probability to $N(\mu, \frac{\sigma^2}{n})$. Convergence in probability means that the probability of obtaining a certain value tends to be the same as that obtained from $N(\mu, \frac{\sigma^2}{n})$ for that value. It is written as $P(\bar{X}) \xrightarrow{d} N(\mu, \frac{\sigma^2}{n})$.

The CLT holds regardless of the shape of the distribution, as long as it complies with the the restrictions on the mean and variance stated, and that the number of iid variables added tends to infinity or is large enough. This is a *very* important caveat, for it means that it will not hold for Pareto distributions with $\alpha < 2$, and thus the arithmetic mean of an iid sample of them will not converge to a Gaussian distribution, but to their respective attractor function. Some PDF's will converge

more rapidly to a Gaussian distribution than others. It is frequently quoted that an appropriate convergence to the Gaussian distribution occurs for samples with $n = 30$, but it actually depends on how rapidly the type of distribution converges. Depending on the type of PDF and application, the number of required samples for the arithmetic mean to be considered to distribute approximately Gaussian within a certain order of error varies.

Taking a single sample \mathbf{X} from a distribution is precisely what one does when taking a measurement in a laboratory in most scenarios, for it will have an error associated to it that can be modeled as a random variable. You may have noticed that most measurements registered are usually redundant, given that each of them is usually an observation of several repetitions of the event of interest, that are averaged to produce a single measurement. Meaning that each single measurement usually has the form of an arithmetic mean \bar{X} from a sample \mathbf{X} . Now think back, you must have surely at some point in your career measured the period of a pendulum. Do you recall that for each individual measurement you registered, you did not just take the time of a single period but of several periods. I insist, for *each* measurement you *registered*. Let us suppose that someone doing this experiment has to take the time for 10 periods for each measurement that he or she writes in the Table of their lab report. This means that their Table has k rows for a k number of samples, each with $n = 10$. According to the CLT, this meant that the measurements for the period had an approximate error of $\pm \frac{\sigma}{\sqrt{10}}$, even though we knew nothing of how specifically the mechanics of our setup produced errors around the measurement, but nevertheless an estimate is obtained for it.

This is where the power of the CLT lies, in that it allows for good approximations of errors for a wide variety of cases, with few exceptions, where we do not know much about the underlying PDF for the random variable, and obtaining it could be time consuming or impractical. Think of the distribution of political preferences, or of what people prefer on a first date, how should these features distribute? The CLT applies to many important distributions under which we can reasonably model these types of random variables. This includes PDF's like the uniform distribution, or the binomial distribution, that characterizes a random walk or the probability of obtaining a certain number of heads or tails in a se-

quence of coin tosses; or the Poisson distribution that gives the probability for an event that has a certain average rate of occurring in time.

Additionally, the reader might have noticed that most random processes are characterized in terms of features of the Gaussian distribution in probability. For instance, the use of the standard deviation as measure of uncertainty, or the kurtosis of the Gaussian distribution to define the excess kurtosis. The assumption that the meaning behind these point value measurements can be extrapolated from what they mean for a Gaussian Distribution to a sample \mathbf{X} , that obeys the assumptions of the CLT but that we do not know its PDF, stems and is valid precisely because of the CLT. Given that most of the distributions we work with will end up behaving like the Gaussian distribution, using it as our unit of measurement is appropriate because it provides a general framework and mold to describe other distributions. By understanding it, we can understand features that work with most types of random processes given enough samples. This might be taking it too far, but think of it as the ring to rule them all, because if you learn to work with it, you will know how to work with many random processes that tend to it.

Now that the theorem has been explained, hopefully thoroughly from the reader's point of view, let us take a look at it formally, and one of the paths to prove it through the use of the characteristic functions.

Theorem 1. *If \mathbf{X} is a set of $\{X_1, \dots, X_n\}$ iid random variables defined on a probability space, where $\forall X_i, E[X_i] = \mu < \infty$ and $E[|X_i|^2] < \infty$ and $\text{Var}(X_i) = \sigma^2 > 0$, then [12]:*

$$Z_n = \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0, 1) \text{ as } n \rightarrow \infty$$

Meaning that:

$$P(Z_n \leq x) \rightarrow \frac{1}{\sqrt{n}\sigma} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy \text{ as } n \rightarrow \infty$$

Proof. Let us first make the following change of variables, changing from X_i to a dimensionless quantity usually referred to as the score, and symbolized in this case by Y_i .

$$Y_i = \frac{X_i - \mu}{\sigma} \rightarrow Z_n = \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{Y_1 + \dots + Y_n}{\sqrt{n}} \quad (45)$$

Let us express now the characteristic function of Z_n in terms of the characteristic function ϕ_i of each Y_i , where:

$$\phi_i \equiv \phi_{\frac{Y_i}{\sqrt{n}}} \left(\frac{t}{\sqrt{n}} \right) \equiv \mathbb{E} \left[e^{it \frac{Y_i}{\sqrt{n}}} \right] \quad (46)$$

$$\phi_{Z_n}(t) = \prod_{i=1}^n \phi_i = \left(\phi_{\frac{Y_i}{\sqrt{n}}} \left(\frac{t}{\sqrt{n}} \right) \right)^n \quad (47)$$

Now we want to expand $\phi_{Z_n}(t)$ in terms of the taylor expansion of the productory. For this purpose, we need to state a few facts form the expected values of Y_i

$$\mathbb{E} [|X_i|^2] < \infty \rightarrow \mathbb{E} [|Y_i|^2] < \infty \quad (48)$$

$$\text{Var}(Y_i) = \text{Var} \left(\frac{X_i - \mu}{\sigma} \right) = \frac{\text{Var}(X_i) - \cancel{\text{Var}(\mu)}^0}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1 \quad (49)$$

$$\mathbb{E} [Y_i] = \mathbb{E} \left[\frac{X_i - \mu}{\sigma} \right] = \frac{1}{\sigma} \mathbb{E} [X_i] - \frac{\mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0 \quad (50)$$

$$(51)$$

Because of Eq. 48, we can assert that $\phi''_{Z_n}(t)$ exists given its relation to the second moment established by the moment generating function from Eq. 22. This also means that we can expand the characteristic function to second order in the from shown in Eq. 52.

$$\phi_i(t) = 1 + \phi'_i(0)t + \frac{1}{2}\phi''_i(0)t^2 + o(t^2) \quad (52)$$

Again, given the relationship between the moment generating function and the derivatives of the characteristic function shown in Eq. 52 we have the following:

$$\mathbb{E} [|Y_i|^2] = 0 = i \frac{d(\phi_i(0))}{dt} \Rightarrow \phi_i(0)' = 0 \quad (53)$$

$$\text{Var}(Y_i) = \mathbb{E} [Y_i^2] - \cancel{\mathbb{E} [|Y_i|]^2}^0 = \mathbb{E} [Y_i^2] = 1 = i^2 \frac{d^2(\phi_i(0))}{dt^2} \quad (54)$$

$$\Rightarrow \phi_i(0)'' = -1$$

This results in that the Taylor expansion becomes:

$$\phi_i(t) = 1 + -\frac{1}{2}t^2 + o(t^2) \Rightarrow \phi_{Z_n}(t) = (1 + -\frac{t^2}{2n} + o(\frac{t^2}{n}))^n \quad (55)$$

$$\lim_{n \rightarrow \infty} \phi_{Z_n}(t) = e^{-\frac{t^2}{2}} \quad \forall t \in R_1 \quad (56)$$

$$\Rightarrow Z_n \xrightarrow{d} Z_{N(\mu=0, \sigma^2=1)} \Rightarrow P(Z_n \leq x) \xrightarrow{d} N(\mu = 0, \sigma^2 = 1) \quad (57)$$

The result of expressed in Eq. 56 tells us that the characteristic function $\phi_{Z_n}(t)$ converges point wise for every t to the characteristic function of a Gaussian distribution. There is a theorem for characteristics functions called the continuity theorem that states that one distribution converges in probability to another if their respective characteristic functions converge to each other point wise. This means then, that any PDF that complies with the hypotheses of our theorems will converge to a Gaussian distribution as the number of samples tend to infinity.

□

In spite of the important properties and uses mentioned for the Gaussian distribution, it is important to recall that assuming indiscriminately that the theorem holds for any random variable will lead to mistakes, like erroneously estimating the risk of debt packages, a world-financial-crisis-contributing type of mistake. That being said, the theorem is extremely useful for it applies in many cases and experimental setups used. Even when it does not strictly apply, for example when the variables are dependent but not that strongly, it serves as a good first approximation to the problem because it simplifies the analysis and calculations. Additionally, some extensions of the CLT hold for less stringent conditions such as different variances, means, and even dependent random variables [11]. This extensions will not be approached on this text, but if working with random variables that obey such conditions, it is strongly recommend that the reader studies about these extensions of the CLT, for instance from the reference cited [11].

3.6 Law of Large Numbers (LLN)

This is one of the most important theorems of probability theory, for from it many important properties have been derived. For instance theorems like the

Source Coding Theorem or Noisy Channel Theorem rely upon it [13], as well as the Law of Large Deviations [1]. It is also of the utmost importance for Frequentist Inference, for it is what justifies that as the number of samples grows, so does the precision of the inference if the statistic is unbiased. Additionally, it serves as the philosophical justification for the frequentist interpretation of probability, for it is assumed that as the number of samples grows, the relative frequency of a certain outcome x of the experiment converges to its probability $p(x)$.

Without further delay, what the LLN states is that as the number of samples n tends to infinity, the arithmetic mean tends to the actual mean of the distribution. There are two forms of the LLN, the weak form and the strong form. They are called this way, for the weak form of proves that that the mean converges weakly in probability whilst the strong form converges almost surely in probability. The LLN is not only restricted to the mean, but it is actually a collection of theorems about convergence of quantities, among which the convergence of the mean is the most famous result. Let us take a look at the proof of the weak form of the LLN for the arithmetic mean of a sample using characteristic functions.

Theorem 2. *If $E[X_i] < \infty$ and $E[X_i] = a$ where $\forall X_i \in \{X_1, \dots, X_n\}$ are iid random variables defined on a probability space, then [12]:*

$$\overline{X_n} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{d} a \text{ as } n \rightarrow \infty$$

Proof. Let us define then the characteristic function for a single variable X_i and its arithmetic mean $\overline{X_n}$.

$$\phi_i(t) = e^{itX_i} \tag{58}$$

$$\phi_{X_n}(t) = e^{it\overline{X_n}} = \left(\phi_i\left(\frac{t}{n}\right) \right)^n \tag{59}$$

Given that $E[X_i] < \infty$, we can expand the characteristic function to first order because of the relationship between the derivatives of the characteristic function and the moments of a distribution. It is the same argument that was used to prove the CLT. We have then:

$$\phi_i(t) = 1 + \phi'(0)t + o(t) \tag{60}$$

given that $\phi'(0) = i E[X_i] = ia$ by hypothesis

$$\Rightarrow \phi_i(t) = 1 + iat + o(t) \quad (61)$$

$$\Rightarrow \phi_{X_n}(t) = \left(1 + iat + o\left(\frac{t}{n}\right)\right)^n \quad (62)$$

By arguments similar to those used in the previous proof, when $n \rightarrow \infty$, the function converges point wise to an exponential function:

$$\lim_{n \rightarrow \infty} \phi_{X_n}(t) = \lim_{n \rightarrow \infty} \left(1 + iat + o\left(\frac{t}{n}\right)\right)^n = e^{at} \quad \forall t \in R_1 \quad (63)$$

$$\Rightarrow \overline{X_n} \xrightarrow{d} a \quad \text{as } n \rightarrow \infty \quad \text{By the continuity theorem} \quad (64)$$

□

3.7 Law of Small Numbers

As was stated earlier, the convergence in probability of the binomial distribution $B(n, p)$ to a Poisson distribution $P(x|\lambda)$ as n tends to infinity can also be proved through the use of characteristic functions. First of all, let us state the hypotheses that will be used: the probability p is such that $pn > 0$, which implies that $\lambda < \infty$ [12]. We will prove that when we take the probabilities of all the range of outcomes, meaning we add the probabilities of all possible $k = 0, 1, \dots, n$, it will converge to a Poisson Distribution as shown next. First we implement the notation shown in Eq. 65 and Eq. 66

$$P(X_{n,k}) \equiv \binom{n}{k} p^k (1-p)^{n-k} \quad (65)$$

$$N_n = X_{n,1} + X_{n,2} + \dots + X_{n,k} + \dots + X_{n,n} \quad (66)$$

First of all, note that the random variable $X_{n,k}$ represents whether the k th attempt in n events was successful or not. Think of it in terms of coin tosses, where the value of $X_{n,k}$ tells you whether or not the k th coin toss in a series of n tosses landed in heads. Consequently, the random variable can only take two possible values. In this case, and in order to simplify the proof, these two possible values are chosen to be 0 and 1. The random variable N_n represents the total

number of successful outcomes when taking all possible attempts into account, and that is why it is constructed from the sum of all $X_{n,k}$ from 1 to n as shown in Eq. 66.

The characteristic functions for $X_{n,k}$ and N_n are obtained and shown in Eq. 67 and Eq. 70 respectively.

$$\phi_{X_{n,1}}(t) = E[e^{itX_{n,1}}] = pe^{it} + (1-p) \quad (67)$$

$$\Rightarrow \phi_{N_n}(t) = (pe^{it} + (1-p))^n = (p(e^{it} - 1) + 1)^n \quad (68)$$

Just like in Eq. 41, the term in Eq. 70 will converge to an exponential.

$$\lim_{n \rightarrow \infty} \phi_{N_n}(t) = \lim_{n \rightarrow \infty} (p(e^{it} - 1) + 1)^n = \left(e^{p(e^{it}-1)}\right)^n \quad (69)$$

$$\lim_{n \rightarrow \infty} \phi_{N_n}(t) = e^{np(e^{it}-1)} = e^{\lambda(e^{it}-1)} = \phi_{Poisson}(t) \quad (70)$$

By the continuity theorem, this means that the binomial distribution indeed converges to a Poisson distribution when n tends to infinity, thus proving the LSN.

3.8 Histograms

The usual method to check what type of PDF a random variable follows is by graphing the number of observed values that it takes within specific ranges, and inferring a PDF that matches the shape of the graph [2]. This graph is called a histogram and is usually represented with bars for each grouping, that are referred to as bins. From it, we can tell whether our distribution is symmetrical, is spread over a wide range of values, how many peaks it has, and so on. It is used mainly in statistical inference, when one wishes to represent the shape of data both visually and in terms of the relative frequency of the bin. The relative frequency refers to how many points of data fall within the range of the bin, divided by the total number of points of data used in the entire histogram. This way, the relative frequency can be associated to the actual probability of the random event given enough sampling.

This approach of interpreting histograms is usually identified as frequentist, for it counts the frequency with which the random variable takes values within the range of each bin, and assumes this frequency is closely related to the actual probability based on the LLN. Think of the usual experiment to check whether a coin is fair or not. Being fair means that each side has the same probability of occurring, in the case of coins it is 0.5 for each side. To test this we would throw the coin a certain number of times and expect the frequencies of both results to be approximately equal within its uncertainty. How to conclude whether the coin is fair or not will be approached when explaining frequentist inference, but suffice it to say, histograms are a very important tool that can help visualize the shape of the PDF and reduce computational cost when processing data by collecting the raw data in bins.

4 Statistical Inference: Bayesians vs. Frequentists

In order to estimate the probabilities of certain phenomena to decide if an inference or hypothesis about it is appropriate or probable, we measure the phenomena in terms of certain variables that describe it. From these variables there are certain functions that help us obtain parameters that describe the PDF of our phenomena. This description or estimation is dependent upon the conditions under which the variables were measured and the assumptions made such as the kind of distribution it follows. These functions are called statistics or estimates. The set of X_n observations are drawn from a group referred to as the population space or set [9]. It is often the case that it is not possible or practical to take measurements exhaustively of every single element of the population space, so they are taken from a subset. This subset is referred to as the sample space. From these measurements made on the sample space we can draw the statistics to make inferences that describe the population space.

In other words, statistical inference is the predilect choice of science to apply inductive reasoning and attempt to generalize the inferences drawn from the observations of the sample space to the population space. As Professor Brian Caffo of the Coursera class, *statistical inference*, said: "Without statistical inference, we are simply living within our data".

4.1 Bayes' Theorem and its implications on both kinds of inference

The two main methods of inference are Bayesian and Frequentist inference. The difference in their approaches stem from their respective interpretations of probability, which translate into how and which terms of Bayes' Theorem, Eq. 71, they use. Bayes' theorem relates the conditional probabilities of two clauses, A and B, through the use of the product rule, Eq. 7.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (71)$$

4.1.1 Likelihood

The second term in the numerator of Eq. 71, $P(B|A)$, is called the likelihood and is of the utmost importance, especially for frequentist inference. By likelihood, it is conceptually meant how well certain parameters and hypotheses fit or explain the set of data sampled. The name was chosen because it tells us then how probable or likely it is that we obtain the observed values of the sample, $\{X_1, \dots, X_n\}$, given a hypothesis H_i that we mean to assess. Each X_i represents a single measurement of a random variable with which we mean to test our hypotheses H_i . This becomes evident when evaluating the probability of H_i through $P(H_i|X_1, \dots, X_n)$. As is shown in Eq. 72, the term $P(B|A)$ becomes $P(X_1, \dots, X_n|H_i)$, and is read as the probability of observing the data given that H_i is true. The expression shown for the denominator in Eq. 73 is a marginalization over all possible hypotheses, which will be explained and demonstrated in a further section of the text.

$$P(H_i|X_1, X_2, \dots, X_n) = \frac{P(H_i)P(X_1, \dots, X_n|H_i)}{P(X_1, X_2, \dots, X_n)} \quad (72)$$

$$P(H_i|X_1, X_2, \dots, X_n) = \frac{P(H_i)P(X_1, X_2, \dots, X_n|H_i)}{\sum_j P(X_1, X_2, \dots, X_n|H_j)P(H_j)} \quad (73)$$

The likelihood is generally used to characterize how well certain parameters that define a pdf, such as the mean and variance for the normal distribution, help the pdf actually describe the set of data observed $\{X_1, \dots, X_n\}$ [14]. That our hypothesis matches the data with a high probability is a minimal requirement that should be fulfilled if it is in fact true. It is not a definitive answer in terms of inference, but it is a good signal that we are on the right path. The branch of frequentist inference referred nowadays as statistical hypothesis testing is based on the likelihood, whether it is through the use of Fisher's Null Hypothesis or Significance Testing [15], or through Neyman and Pearson's Hypothesis Testing [16], which will be explained later on.

4.1.2 Prior Probability

The first term of Bayes' Theorem in the numerator, $P(A)$, is referred to as the prior. It is the probability that we suppose or have estimated for clause A be-

forehand, hence the name prior. In the case of inference, this term is then $P(H_i)$, and refers to the probability of our hypothesis before having any knowledge from the current evidence or sample. This term is one of the main points of digression between Bayesians and Frequentists. Frequentists, especially in respect to the use of a uniform prior distribution, claim that assigning a pdf to the hypothesis before having any knowledge that came from the data is a form of bias, and that it is not a proper or transparent way of assessing whether one's views agree with nature [17]. For them, deciding the probability before having access to the data tampers with the required skepticism to evaluate a hypothesis, and that knowing the actual probability of a hypothesis is not possible. This is for instance why Ronald Fisher changed his take on probability, and went from being a Bayesian to one of the leading authors that developed key frequentist statistical methods to assess hypotheses [18].

Nevertheless, the assignment of priors is not done at whim and without proper ponder and thought by Bayesians. Priors may be constructed based on previous observations and estimations [1], or built so that they represent adequately a state of almost complete ignorance [19], or even defined in a way such that when multiplied by the likelihood it retains the form or family of the prior's PDF [20]. The first kind are called informative priors, for they incorporate previous information to provide a starting point that is updated after using Bayes' Theorem with a new set of observations. The second kind are called non-informative or Jeffrey's priors. The Jeffrey's Priors may be based on the principle of indifference, that states that when having no knowledge of how the probability distributes among the population space, one should assign equal probabilities to all elements through a Uniform Distribution. This is to avoid any form of bias by preferring any of them without having access to data about them [1]. Other kinds of non-informative priors are the ones that use general and broad restrictions that may apply to the entire population space such as non-negativity, normalization, or the existence of a mean and variance, among others. In this case it retains the name of non-informative, for it only uses information that applies to the entire population space in the form of restrictions. The third kind are called conjugate priors, and are used when one assumes that new data only updates the parameters of the PDF of our hypothesis H_i , but does not change the family that the PDF belongs to.

Given that Bayesians characterize probabilities as a degree of belief in a given hypothesis or clause, the probability need not be exactly and exhaustively descriptive of the population. Rather, the probability characterizes what we are logically allowed to believe of the sample or phenomena [1]. That is why, when we have very little or no information, the principle of maximum entropy is also used to estimate priors. Shannon's entropy is the measurement of the average uncertainty of a distribution [13]. It is different from the variance and is not dependent on it, for one can obtain the shannon entropy for distributions that have no variance. The PDF that maximizes entropy, is the one that maximizes the uncertainty or lack of information on a clause or random variable. Thus if our state is such that we do not have any prior knowledge of the subject, then our best approach is to start with the PDF that is characteristic of a state of maximum uncertainty possible of the variable [1].

4.1.3 Marginal Likelihood

The term in the denominator $P(B)$ of Eq. 71, or $P(X_1, X_2, \dots, X_n)$ of Eq. 72, is referred to as the evidence or marginal likelihood [9]. It represents the probability of obtaining the observed data by itself, without linking it to any hypothesis. Note that this term is common to all hypothesis H_i , for it does not depend on them. It is precisely because of this, that it does not play a key role in statistical inference when deciding between alternate hypotheses, for it is a common factor to all terms. So for instance, when comparing whether a hypothesis H_i has a higher probability than hypothesis H_j , the term is canceled as is shown in Eq. 74.

$$\begin{aligned} \frac{P(H_i|X_1, X_2, \dots, X_n)}{P(H_j|X_1, X_2, \dots, X_n)} &= \frac{P(H_i)P(X_1, X_2, \dots, X_n|H_i)}{P(X_1, X_2, \dots, X_n)} \frac{P(X_1, X_2, \dots, X_n)}{P(H_j)P(X_1, X_2, \dots, X_n|H_j)} \\ &= \frac{P(H_i)P(X_1, X_2, \dots, X_n|H_i)}{P(H_j)P(X_1, X_2, \dots, X_n|H_j)} \end{aligned} \quad (74)$$

Given that we not only want to choose between competing hypotheses, but also calculate the degree of belief that we have on a hypothesis H_i given the observations $\{X_1, \dots, X_n\}$, it is necessary to find a way to calculate this term. If

we have a complete set of mutually exclusive hypotheses, then we can marginalize over the likelihood and prior of each of them as was done in the denominator of Eq. 73. By a complete or exhaustive set, it is meant that the probability of all of its elements add up to one as shown in Eq. 75 [6]. Mutually exclusive means that the probability of any of them being true simultaneously is zero, Eq. 76 [9].

$$P(H_1 + H_2 + \dots, H_n) = 1 \quad (75)$$

$$P(H_i H_j) = 0 \text{ if } i \neq j \quad (76)$$

On the other hand, marginalizing is using the dependence of a clause A on the elements of the complete set, by adding the dependent probabilities of each element, weighed by the probability of the element occurring in the first place, to obtain then the probability of the clause itself. The procedure of how to marginalize is shown in Eq. 77. The proof that this approach is valid is outlined in Eq. 78.

$$P(A) = \sum_i^n P(A|H_i)P(H_i) \quad (77)$$

$$P(A) = P(A) \left(P(H_1 + \dots + H_n|A) + P(\overline{H_1 + \dots + H_n}|A) \right) \text{ by use of the sum rule} \quad (78)$$

$$P(A) = P(A, (H_1 + \dots + H_n)) + P(A, \overline{(H_1 + \dots + H_n)}) \text{ by use of the product rule} \quad (79)$$

$$P(A) = P(A, (H_1 + \dots + H_n)) + P(A, \overline{H_1 + \dots + H_n})P(\overline{H_1 + \dots + H_n}) \quad (80)$$

$$P(\overline{H_1 + \dots + H_n}) = 0 \text{ because of Eq. 75}$$

$$P(A) = P(A, (H_1 + \dots + H_n)) = P(AH_1 + \dots + AH_n) = \sum_i^n P(AH_i) \quad (81)$$

because of Eq.75 and 76

$$P(A) = \sum_i^n P(AH_i) = \sum_i^n P(A|H_i)P(H_i) \text{ by use of the product rule} \quad (82)$$

4.1.4 Posterior Probability

Through these terms, likelihood, prior, and evidence or marginal likelihood, we arrive at the final result of Bayes' Theorem, and the jewel in the crown of Bayesian Inference, the posterior probability. This probability, $P(H_i|X_1, \dots, X_n)$, is the degree of belief we should have for H_i after the evidence is observed and quantified. It is not limited to only signal whether we are on the right path or not, as occurs with the likelihood. It is actually quantifying how much should we believe that we are on it after gathering the new evidence X_1, \dots, X_n . The difference may seem subtle, and it is, but it is definitely substantial. It is the difference between the probability of our evidence matching our hypothesis, the likelihood, and the probability of our hypothesis being true after observing the evidence, the posterior. Calculating the latter is a step beyond the former, that helps assess directly the probability of the hypothesis, at the cost of certain assumptions in order to calculate the additional terms of Bayes' theorem.

Usually the hypotheses H_i are assumptions of the values that the parameters θ of a PDF might take. Throughout the rest of the text, θ will represent the set of parameters on which a Pdf depends. For instance, for a Gaussian distribution $\theta = \{\mu, \sigma^2\}$. With the information given so far, we can consider *why* Bayesians and Frequentists use the posterior probability and likelihood respectively. This key difference in form, perhaps the most important one aside from their interpretations of probability, is that Bayesians assume that the data is fixed and that the parameters remain undetermined and of a probabilistic nature and form determined by the data. On the other hand, Frequentists assume that the parameters are fixed and that they presuppose how the data is distributed, and thus it is the data that retains a probabilistic nature. For them, any uncertainty on the value of the parameters arises from the imperfection of our measurements and from the statistical tools used to estimate them.

To Bayesians it is not adequate to estimate parameters as just point estimates. They do not think that events actually follow specific PDF's with specific parameters. They are not as positivist as Frequentists are in this regard. For them, parameters are just how we construct the PDF's that describe our degree of belief on an event or parameter, based on the evidence and prior information about it.

This is why for them it is the sample that must be taken as fixed and as the actual facts that we obtain from nature. The parameters are just how we interpret this information. It is because of this that Bayesians go through the trouble of finding a posterior distribution $P(\boldsymbol{\theta}|X_1, \dots, X_n)$ for the parameters, for in it, the parameters are a function of the data

For Frequentists events occur according to a PDF, that dictate the frequency of the samples that we observe from a given population. For them it is the way that Nature's laws take form. From the samples, and in concordance with the LLN [13], certain form of estimates of the parameters will converge in probability or almost surely to the actual value for a number of samples that tends to infinity. According to them, with the appropriate statistical tools, for large enough samples, it is possible to infer the true value of the parameters within a margin of error from a sample. This is why frequentists rely solely on the likelihood to choose parameters and infer hypotheses, for it takes the form $P(X_1, \dots, X_n|\boldsymbol{\theta})$, where the sample is a function of the parameters.

4.2 Bayesian Inference

We have that using Bayes' theorem provides an actual quantification of how probable a hypothesis is after certain observations, and not just how compatible they are. One of the challenges of using the theorem, as has already been stated, is how to correctly quantify a prior in order to obtain a proper posterior probability, and what happens if the initial assumed prior is not correct. The answer to this is that the prior need not be completely exhaustive and informed, for the calculated posterior probability $P(H|\mathbf{X}_1)$, where $\mathbf{X}_1 \equiv \{X_1, \dots, X_n\}$, can be used as the prior for new set of observations $\mathbf{X}_2 \equiv \{X_{n+1}, \dots, X_{n+k}\}$. The previous observations now make part of the set of assumptions of the new problem. Both sets of data should belong to the same population, but not necessarily the same sample space, for otherwise they would not be compatible and consistent. Our new posterior probability would have the following form:

$$P(H|\mathbf{X}_1, \mathbf{X}_2) = \frac{P(H|\mathbf{X}_1)P(\mathbf{X}_2|H, \mathbf{X}_1)}{P(\mathbf{X}_2|\mathbf{X}_1)}$$

$$P(H|\mathbf{X}_1, \mathbf{X}_2) = \frac{P(H|\mathbf{X}_1)P(\mathbf{X}_2|H, \mathbf{X}_1)}{P(\mathbf{X}_2)} \text{ If } \mathbf{X}_2 \text{ is independent of } \mathbf{X}_1$$

Bayes' theorem can be used iteratively with different data sets, and incorporate every new set of observations, whilst pondering and using what was learned from previous ones. This makes it a great tool for meta-studies and meta-analysis. Even if starting with a prior probability that does not represent too adequately the belief that we should have on the hypothesis, through subsequent applications of Bayes' theorem on new groups of evidence, our initial estimates may improve. This is where the caveat for bayesian inference stands out, for if the chosen prior is grossly mistaken, updating through observations can be inefficient and unfruitful, to the point where the inferences are just wrong. So there is a margin of error that can be corrected by iterative implementation of the theorem, but if exceeded, it will lead to invalid inferences. To state it plainly, it does follow the maxim of computer science and logic of "Garbage in, Garbage out" [21].

4.2.1 Prediction intervals

With the posterior or prior probability it is possible to obtain predictions of future values of the population space. This can be carried out by marginalizing the probability of a future value, under the assumption of our hypothesis, either through the prior probability or the posterior. Our estimate of a future value will be reliable to the extent that both the prior and posterior probabilities are reliable themselves through enough sampling and updating. Once again, recall *Garbage in, Garbage out*.

To marginalize over a probability, we have to use the property shown in Eq. 78 and use it with the prior or posterior probability to obtain a PDF for a possible new element x_{n+1} as seen in Eq. 85.

$$\begin{aligned}
 X_{n+1} &\equiv \text{new element sampled from the population} \\
 \mathbf{X} &= \{X_1, \dots, X_n\} \text{ Previous information collected} \\
 H_i &\text{ such that for every } H_i, H_j \ P(H_i H_j) = 0
 \end{aligned} \tag{83}$$

In terms of the posterior probability for the discrete and continuous case, one marginalizes the probability $P(X_{n+1}|\mathbf{X})$ in terms of each of the posterior probabilities $P(H_i|\mathbf{X})$ for each hypothesis H_i :

$$\begin{aligned}
 P(X_{n+1}|\mathbf{X}) &= \sum_i^n P(X_{n+1}, H_i|\mathbf{X}) = \sum_i^n P(X_{n+1}|H_i, \mathbf{X})P(H_i|\mathbf{X}) \\
 p(x_{n+1}|\mathbf{X}) &= \int_H p(x_{n+1}|H, \mathbf{X})p(H|\mathbf{X})dH
 \end{aligned} \tag{84}$$

In terms of the prior probability, instead of marginalizing over the posterior probability $P(H_i|\mathbf{X})$, one marginalizes over the prior $P(H_i)$. This in order to obtain a prediction interval when one cannot calculate the posterior distribution. Prediction intervals constructed with the prior $P(H_i)$ are also carried out in order to compare whether the prediction improves with the updated probability for H_i

through the posterior probability or becomes worse.

$$\begin{aligned}
 P(X_{n+1}|\mathbf{X}) &= \sum_i^n P(X_{n+1}, H_i) = \sum_i^n P(X_{n+1}|H_i)P(H_i) \\
 p(x_{n+1}|\mathbf{X}) &= \int_H p(x_{n+1}|H)p(H)dH
 \end{aligned}
 \tag{85}$$

When our hypotheses belongs to a space where they distribute continuously, then we have to express our probability in terms of continuous variables, and our summatory should become an integral as in Eq. 85. An example of a continuous hypothesis space, is the set of hypotheses of the values that the mean of a distribution might take. Recall that marginalizing is taking into account all the possible values, weighed by their respective probabilities, of a continuous variable.

4.2.2 Decision Theory and Maximum A Posteriori (MAP)

In Bayesian inference, the problem of which Hypothesis to choose from a group of competing hypothesis is based on a criteria referred to as Bayes' Rule or Maximum a Posteriori, MAP. The criteria states that one should choose the hypothesis that has highest posterior probability. This is used in a wide variety of areas, for instance when decoding messages from noisy communication channels [13]. This is commonly used when there is not an strictly defined cost function that tells how is the error penalized. For different cost functions, there are different point value measurements that minimize it. This affects what hypothesis should be chosen in order to minimize the cost function. For instance, if the cost function is quadratic with the error, then the hypothesized value for the random variable that reduces this error is that the random variable equals the mean. If instead the cost function is a Dirac Delta function, the hypothesis that should be chosen is the mode [1].

Decision theory is a wide and deep subject, with a great variety of approaches and consequences on how inferences should be drawn. However, on the present section we will limit ourselves to one of the approaches taken by Bayesians through MAP estimation. If dealing with just two hypothesis, in the usual scenario of Null Hypothesis H_0 against an Alternate Hypothesis H_1 , the approach is to take the ratio between competing hypothesis as shown in Eq. 86. Note that the result

does not depend on the evidence as was mentioned earlier, for the evidence is independent of the hypotheses under consideration. When dealing with more than one alternate model, one can calculate the posterior directly for each of them, or use the ratio for several combinations to then order them using the transitive property. The last option is useful if one wants to avoid calculating the evidence directly. The last two methods mentioned rely on brute force to find the MAP hypothesis, and without a doubt there must be more intelligent ways to carry this out.

$$\begin{aligned} \frac{P(H_0|X_1, X_2, \dots, X_n)}{P(H_1|X_1, X_2, \dots, X_n)} &= \frac{P(H_0)P(X_1, X_2, \dots, X_n|H_0)}{P(X_1, X_2, \dots, X_n)} \frac{P(X_1, X_2, \dots, X_n)}{P(H_1)P(X_1, X_2, \dots, X_n|H_1)} \\ &= \frac{P(H_0)P(X_1, X_2, \dots, X_n|H_0)}{P(H_1)P(X_1, X_2, \dots, X_n|H_1)} \end{aligned} \quad (86)$$

Let us take a look at the following conceptual, and quite possibly superfluous example, that has the purpose of illustrating more clearly how MAP works. Suppose we have certain observations that match a hypothesis. Let us say that a group of irregular lights in the sky that move haphazardly could be a good match for a hypothesis, H_{LGM} , that they belong to alien spacecrafts. Assuming then that the likelihood is high, they could pass some of the statistical tests used to determine if we can accept H_{LGM} , or rather reject its near contrapositive, that the lights might be due to commercial airplanes for example. The inference would be then that we are in fact seeing alien spacecrafts.

However, if we were to calculate the posterior probability, we should have a rather low probable prior $P(H_{LGM})$, given that we have never seen aliens before on earth or neighboring planets. That would downplay the effect of the likelihood when calculating the posterior probability of H_{LGM} , especially in respect to competing hypotheses that also have a good likelihood. If these competing hypotheses have a much higher prior probability, for example the hypotheses that they are aircraft formations or satellites, then their posterior probability will be more likely and thus a more appropriate and logical inference. Using Bayes' theorem allows us then to choose or distinguish between competing hypotheses that yield the

same results in terms of likelihood, and discard the ones that should have a lower probability of actually being true.

The statistical tests based entirely on the likelihood are not actual estimates of the posterior probability of H_i , but conjectures or rules of thumb to accept an inference in terms of the likelihood. They are useful and have worked in both science and in the industry with great results. However, the caveat is that the likelihood is not an absolute replacement for the posterior probability. Misuse of these rules of thumb, especially in cases of hypothesis with a low prior probability, has led to problems of irreproducibility of inferences drawn in areas like psychology [22].

On the other hand, the caveat for the calculation of the posterior, is that the assumptions of the evidence and prior do not bias and blind the analysis to a point where it tampers greatly with the inference, such that statistical tests that resort to just the likelihood result in better estimates. There are instances where there are no viable models or reasonable assumptions to calculate the prior or the marginal likelihood. Special care must be had then in order to decide if the assumptions are conservative and valid enough to make a quantitative leap from the likelihood to the posterior or not. If not, it is better to stick to the quantitative leap of Frequentist statistical tests.

Now let us take a look at a more serious numerical example to further illustrate how MAP is used to choose between hypotheses, and to take a look at the nuances between likelihood and posterior probability. Let us work out a version of a numerical example that was used by professor Alonso Botero in his information theory class at Universidad de Los Andes. Suppose a new medical test came out for an extremely rare and lethal disease E with the following conditions:

$B = \{X_1, \dots, X_n\} = + \equiv$ Positive result of the medical test

$\overline{+} = - \equiv$ Negative result of the medical test

$A = H = E \equiv$ Actually having the disease

$P(+|E) = 1$ if the person is sick, the test will always mark a positive result

$P(-|E) = 1 - P(+|E) = 0$ By the sum rule the false negative rate is zero

$P(E) = 10^{-6}$ Frequency of occurrence of the disease

$P(\overline{E}) = 1 - P(E) = 0.9999999$ According to the sum rule

$P(+|\overline{E}) = 10^{-2}$ false positive rate

$P(-|\overline{E}) = 1 - P(+|\overline{E}) = 0.99$ true negative rate

$P(E|+) \equiv$ Probability of actually being sick after testing positive

$P(\overline{E}|+) \equiv$ Probability of not being sick after testing positive

Let us describe carefully each of the quantities given and how they relate and characterize the problem. As is shown above, this particular test has the advantage, that if the person has the disease E , then it will always detect it and mark a positive result. The quantity $P(+|E)$ is known in medicine as the sensitivity of the test [9]. It represents the likelihood of obtaining a positive result, given that it should be positive. In other words, that the effect is correctly identified as true given that it *is* true. This quantity is also referred to as statistical power [9]. The statistical power in the case of this test is absolute. By the sum rule, this also means that the false negative rate $P(-|E)$, usually written as β , is zero. False negative is the probability that a test does not detect an effect, given that it is true.

Note that so far we have not said anything about the test given that the effect is not happening. That is, we have not said anything of how it functions, in terms of positive and negative results, when the person does not have the disease. These probabilities also have to be taken into account. The probability $P(+|\overline{E})$ is known as the false positive rate. Its complement $P(-|\overline{E})$ is called specificity in medicine [9]. In statistical terms, especially frequentist ones, the false positive rate is referred to as the significance or α . This probability is extensively used in the most ubiquitous kind of statistical tests, null hypothesis testing [22].

The general idea behind null hypothesis testing is that if the likelihood of the contrapositive hypothesis is low, usually convened to be below 5% and 1%, we can reject it in favor of our original hypothesis. Think of it as an attempt of a proof by contradiction through a loose interpretation of the sum rule. A lower value of the significance of the null hypothesis implies that less compatibility between the hypothesis and the data is demanded, resulting in that the test becomes more stringent.

The null hypothesis in our example, \overline{E} , is that the person does not have the disease. The significance in our example is right on the 1% threshold. Our hypothetical medical test complies outstandingly with the usual requirements of frequentist inference to accept our hypothesis, low significance and high statistical power. Under these conditions based only on likelihoods, we should accept the inference that the person that gets a positive result is in fact sick. However let us use Bayes' theorem to calculate the actual value of this probability as is done in Eq. 87.

$$\begin{aligned}
 P(E|+) &= \frac{P(E)P(+|E)}{P(+)} = \frac{P(E, +)}{P(+, E) + P(-, E)} \\
 P(E|+) &= \frac{P(E)P(+|E)}{P(+|E)P(E) + P(+|\overline{E})P(\overline{E})} \\
 P(E|+) &= \frac{10^{-6} * 1}{10^{-6} * 1 + 10^{-2} * 0.999999} = 9.999 * 10^{-5} \\
 P(\overline{E}|+) &= 0.99990001
 \end{aligned} \tag{87}$$

With a probability in the order of 10^{-5} one should not believe that a positive result implies that the person has the disease. Additionally if one compares the posterior of both hypothesis, one obtains that the hypothesis that the person does not have the disease is far larger than the hypothesis of having it. This implies that one should favor the former hypothesis.

The posterior probability is extremely low due to the fact that our hypothetical disease has a really low frequency in the population and hence a low prior $P(E)$ in comparison with the significance of the test. This clearly illustrates how the rules of thumb based on just likelihoods can lead us to incorrect inferences in cases where the prior probabilities are really low. In order to correct for this, one must

keep in mind the effect of the prior, and thus require much more stringent values for α for the rules of thumb and inferences to apply [22].

The example shows that in order to have a good posterior, given a good likelihood but an extremely unlikely prior, the evidence must be nearly as rare as the prior. Please check again Eq. 72 to make sure you follow the train of thought. If the evidence is not as unlikely, it means that the alternatives to our hypothesis, that make up the set of complementary hypotheses, also have a high likelihood of explaining the evidence. The larger the difference in probability in favor of the evidence, or in other words the more likely the evidence in comparison to the prior, the more likely it is that the evidence is due to the complementary hypotheses. This of course translates into a lower posterior probability. This means that the posterior is actually quantifying how much of the probability of the evidence is explained by the probability that our hypothesis occurs along with the observed data. This can be clearly seen in Eq. 87, where the posterior is the proportion of obtaining a positive result along with having the disease, to this very same probability plus the probability of not having the disease and still obtaining a positive result. If our hypothesis accounts for much of the observed data, then it will revamp our prior belief in the hypothesis. If not, it will downplay our belief.

4.2.3 Bayes' Factor

Bayes Factor is the equivalent of Bayesian inference to hypothesis testing. It is the same approach as that of likelihood ratio between competing hypothesis or models, except that instead of taking the ratio over the parameter that bears maximum likelihood, it is taken over all parameters [23]. Now, it might seem unfair to try to explain it by drawing equivalents to something that has not been explained so far, so let us get into what a likelihood ratio is, and what it is used for. A likelihood ratio is similar to the approach of MAP, but instead of taking the ratio of posterior probabilities, it is the ratio of the likelihoods as shown in the right hand side, (RHS), of Eq. 88. The terms θ_0 and θ_1 represent the set of parameters that compose the model proposed by hypothesis H_0 and H_1 respectively.

$$\lambda = \frac{L(X_1 + \dots + X_n | H_0)}{L(X_1 + \dots + X_n | H_1)} = \frac{L(X_1 + \dots + X_n | \mathbf{H}_0, \boldsymbol{\theta}_0)}{L(X_1 + \dots + X_n | \mathbf{H}_1, \boldsymbol{\theta}_1)} \quad (88)$$

Broadly speaking, what is done in likelihood ratio tests, is that the parameters that are not fixed by hypothesis, are chosen to be the ones that bare the maximum likelihood. As a result, the likelihood function for hypothesis is optimized to find the parameters $\hat{\boldsymbol{\theta}}$ that maximize $L(\mathbf{X}|H)$, and are then fixed to that value. The value is fixed at $\hat{\boldsymbol{\theta}}$, the optimal value of $\boldsymbol{\theta}$, for they are the set of parameters assumed to be the ones closest to the real value because they result in the greatest coherence with the data for all possible values for the parameters. In Frequentist inference it is used to reject one of the hypothesis in favor of the other. In order to do this, the statistic λ must be quite small, meaning that hypothesis H_0 is rejected in favor of H_1 .

Now the bayesian approach to this kind of test differs in that they do not take just a single value for each parameter to evaluate the likelihood. Instead they marginlize it over all the parameters that compose the model as seen in Eq. 89. By doing so, it includes information about all possible ranges of the parameter and the uncertainty around the optimal choice $\hat{\boldsymbol{\theta}}$. The advantage this poses is that by using Bayes Factor, any uncertainty around the value or values of $\boldsymbol{\theta}$ is taken into account. A large uncertainty around $\hat{\boldsymbol{\theta}}$ can occur if the maximum of the likelihood function does not behave like a peak, but rather as a plateau. This means that there is a wide range of values of $\boldsymbol{\theta}$ that have a likelihood similar to that of $\hat{\boldsymbol{\theta}}$. Additionally, if there is a model that has many free parameters, then it is very likely to fit the data but in turn, it looses its predictive capabilities. Think about it, if a model fits perfectly to any set of data, what happens when you feed the model garbage, then it will also fit it as well. This phenomenon is known as overfitting. Calculating Bayes factor avoids this, becasue if the model has a great deal of parameters, then it will have to spread its likelihood among this parameters, to the point that when comparing them will lead to an overall reduction of likelihood when taking all parameters into account. So in a sense Bayes Factor takes care of overfitting in a quite natural way, and in accordance to Occam's Razor, a maxim that expresses that simpler model should be favored over complex ones [1].

$$K = \frac{P(\mathbf{X}|H_0)}{P(\mathbf{X}|H_1)} = \frac{\int P(\boldsymbol{\theta}_0|H_0)P(\mathbf{X}|\boldsymbol{\theta}_0, H_1) d\boldsymbol{\theta}_0}{\int P(\boldsymbol{\theta}_1|H_1)P(\mathbf{X}|\boldsymbol{\theta}_1, H_1) d\boldsymbol{\theta}_1} \quad (89)$$

This will be the topics that we will cover in the present work about Bayesian Inference. Of course there are many more methods of Bayesian Inference such as Bayesian Networks, Monte Carlo Markov Chains, and so on. Let us now take a look at Frequentist Inference, and the methods that we will apply from it to a phenomenological study in HEP.

4.3 Frequentist Inference

Frequentist inference, as its name suggests, is a way to draw meaning from a sample of data by applying the frequentist interpretation of probability in its methods. This means that the assumptions about a random process will contrast heavily to those of Bayesian Inference. Perhaps among the key assumptions, is that it supposes that there *exists* a PDF with fixed parameters that determines the behavior of the random variable, and thus the frequency with which we will observe the samples. Again, recall that Bayesians suppose that we usually cannot do this, and that the data or samples are the ones fixed. In Frequentist Inference it is key then to find the parameters of a PDF, given that they fully characterize it, and hence the behavior of our random variable.

According to the frequentist interpretation of probability, the frequency of a value is directly related to its probability and it is a consequence of it. So, if we have a fair coin with a 50/50 chance of landing on heads or tails, we will see heads in nearly 50% of the coin tosses for a large number of tosses. This means that the relative frequency, or the observed occurrences of a value divided by the total number of samples, will tend to be the same as the probability of that value, given enough observations or sampling. This statement is a consequence of the Law of Large Numbers (LLN) and the Law of Large Deviations (LLD) [13].

The Law of Large numbers states that for a large number of observations, the arithmetic mean of the sample will converge either strongly or weakly to the actual mean of the PDF, as shown in Eq. 90. The Law of Large deviations states on the other hand that for independently identical distributed variables (iid), the probability of obtaining a value larger than the mean has an approximately exponential dependence on the shannon entropy of the PDF, as seen on Eq. 91. This implies that for a large number of samples N , the probability of obtaining a value larger than the mean decreases exponentially. This along with the LLN implies that for very large samples of iid variables, the PDF that maximizes the Shannon entropy of the relative frequencies, subject to the restrictions of the problem such as normalization, will be the PDF that describes best or corresponds to the random variable [13]. In principle this is one approach to find the PDF our

sample follows, given enough data or samples to build a histogram and the relative frequencies.

$$\bar{X} = \frac{1}{n} \sum_{i=0}^n x_i$$

$$\text{Relative Frequency} \equiv \bar{f}_i = \frac{1}{N} \sum_{i=0}^N n_i \text{ where } \begin{cases} N \equiv \text{Number of samples taken} \\ n \equiv \text{Number of times } x_i \text{ is sampled} \end{cases}$$

$$\text{Law of Large Numbers} \equiv \begin{cases} \lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mu| > \varepsilon) = 0 \text{ Weak convergence.} \\ \Pr\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu\right) = 1. \text{ Strong convergence} \end{cases} \quad (90)$$

$$P(\bar{f}_i > f_i) \approx \exp(-NI(f_i)) \quad (91)$$

Between these two laws, the one that is used the most is the LLN, for it tells us that we can find the moments of the PDF of our random variable X , with weak or strong convergence, by using a sample $\mathbf{X} = \{X_1, \dots, X_n\}$ taken from the population space. Given that the parameters that many distributions depend on are actually moments of the PDF (think of the dependence of the Gaussian Distribution on μ and σ^2) through enough sampling we would find approximately the PDF itself.

Sometimes there is not enough access to large samples of data, so in order to enhance the estimation of parameters, the point estimate is calculated within a range. This range will have a certain probability associated to it, for it characterizes the number of times that the true value of the parameter falls within a newly calculated range from a new sample. As the reader might have noticed, the range bears no meaning unless a probability is associated to it. It is this probability that will give us a sense of whether it is likely that the parameter lies within that range. This probability is referred to as a confidence interval. Usually confidence intervals are required to have a probability of 95%, read as a 95% confidence. The other way to estimate the parameters of the PDF is through the likelihood, where the assumed hypothesis is the value that the parameters take. The values

that yield the maximum likelihood are assumed to be the best estimates for the parameters, given that they yield the highest compatibility between the data and the distribution.

4.3.1 Estimation of parameters: Bias, Sample Variance, Consistency, Efficiency, and Maximum Likelihood

The different ways and functions to estimate the value of a given parameter is called an estimator or an statistic, that is a function of the sample space as shown in Eq. 92. There can be different estimators for a same quantity, and they have different advantages and shortcomings. Several definitions have been constructed in order to quantify the performance of an estimator. For instance, the Bias characterizes how close the expected value of the estimator actually converges to the true value of the parameter, or the variance of the estimator around the true value [7]. With these definitions, concepts like the efficiency and robustness of an estimator are built. Given that there is no definition that allows to determine how close are we to the actual value for a given sample size, tests have been developed to reject a null hypotheses in favor of an alternate one in terms of statistics or estimators calculated from the sample. This tests rely on calculating probabilities with the likelihood term, either assuming the null hypothesis or the alternate one as true.

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n) = \hat{\theta}(\mathbf{X}) \quad (92)$$

Error

It is a function of the elements of the sample space, and it measures how far off is the estimator $\hat{\theta}(X_i)$, when evaluated on a given element X_i from the sample space, from the actual value that the parameter θ of the PDF has. The definition is given in Eq. 93.

$$e(X_i) = \hat{\theta}(X_i) - \theta \quad (93)$$

Variance

It is a measure of how dispersed is the estimator when evaluated for each datum X_i from the average, by calculating its variance as shown in Eq. 94. In other words, it is related to the precision in an experiment. It differs from the mean square error, in that it does not measure how far off it is from the actual value of the parameter, it just measures how dispersed are the outcomes of the estimator.

$$\text{Var}(\hat{\theta}) = \text{E} \left[(\hat{\theta} - \text{E} [\hat{\theta}])^2 \right] \quad (94)$$

Bias

The definition for the Bias is that shown in Eq. 95, and it measures the difference between the expected value of the estimator of a parameter and the actual value it has. In other words, it measures the discrepancy between the actual value of the parameter of the population space with the expected value yielded by our estimator that depends on how it calculates the parameter with the data gathered \mathbf{X} .

$$\text{B}_\theta[\hat{\theta}] = \text{E}_\theta[\hat{\theta}] - \theta = \text{E}_\theta[\hat{\theta} - \theta] \quad (95)$$

That an estimator is considered unbiased, meaning that the Bias equals zero, is a desirable property in most cases. An example of an unbiased estimator is the arithmetic mean of a sample of iid random variables. The standard deviation and the population variance on the other hand, usually calculated as shown in Eq. 96, are not unbiased. It is not to say that this estimators are wrong, is that they only apply when considering the entire population space.

When dealing only with samples, the unbiased estimator for the variance takes the name *sample variance*. The sample variance differs from the population variance in a factor known as Bessel's correction, as seen by comparing Eq.96 and Eq. 97. That the estimator for the variance is unbiased does not mean that the standard deviation is unbiased as well. As matter of fact, an unbiased estimator for the variance yields a biased estimator for the standard deviation and the mean.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{and} \quad \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (96)$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (97)$$

One usually seeks out unbiased estimators, and that is why it is sometimes frequent to take the expected value of a measurement. Allowing some bias is not necessarily bad, for it can allow in some cases a lower Mean Squared Error (MSE), than the unbiased estimator, thus having a higher efficiency and estimating better the parameter. So, it is a sort of a dilemma similar but not analogous to that of dispersion between time and frequency in a Fourier transform. But let us take a close look at this dilemma also known as the Variance-Bias trade off in the section of MSE.

Mean Squared Error

The Mean Squared Error measures how far off and how dispersed is the estimator when applied on the sampled data from the actual value it is estimating. It measures these properties by calculating the expected value of the squared difference between the estimator $\hat{\theta}(X_i)$ and the actual value of the parameter θ as shown in Eq. 98. The MSE is also related to both the sample variance and bias of the estimator. Usually decreasing one increases the other, so it becomes sort of a dilemma and a delicate balance when deciding which estimator to choose. This dilemma takes special importance in the context of Machine Learning, where the estimator represents a model $\hat{f}(x)$ that predicts the value of a variable y . Additionally, it is customary to add a term that represents an irreducible error in the problem represented by σ , as seen in Eq. 98.

$$MSE \equiv \mathbb{E} [(\theta - \hat{\theta}(X)(x))^2] = \mathbb{E} [(y - \hat{f}(x))^2] = \mathbb{B} [\hat{f}(x)]^2 + \text{Var} [\hat{f}(x)] + \sigma^2 \quad (98)$$

In this case the Bias is associated to underfitting and simpler models, given that it measures the expected value of the error between the predictions and the actual value. Simpler models tend not to overfit, but an excess bias because of oversimplification can ignore and not learn key relationships and characteristics between the sample (features) and the outcomes (eg: Classifications). This occurs because it cuts assumptions and parameters in order to have a simpler model that is easier to establish. While a simpler model can ignore noise, an excess of bias can result in underfitting and a limitation to the learning algorithm.

Variance on the other hand is linked to overfitting. Fitting a highly dispersed sample, with a dispersion that corresponds to noise and not to characteristics of the problem, can often lead to overfitting and rules that have nothing to do with the problem itself, but the noise associated with it. This is known as the bias variance tradeoff, and it frequently arises in supervised machine learning, given that in order to construct the functions that will predict classification or outcomes, having less of one usually implies increasing the other.

Consistency

The characteristic of consistency of an estimator quantifies whether an estimator converges or not, as the number of samples grows to infinity to the actual value of the parameter. The mathematical definition of this is that shown in Eq. ??.

$$\lim_{n \rightarrow \infty} \hat{\theta} \rightarrow \theta \quad (99)$$

Efficiency

A more efficient estimator requires less statistic to reach certain level of performance such as a given confidence interval. Efficiency is dependent on a low variance. The lower the variance, the more efficient the estimator. The lowest bound is imposed by something known as the Cramer Rao lower bound. This bound sets the theoretical minimum variance for an unbiased estimator, and it is equal to the inverse of the Fisher Information. The Fisher Information is a measure of how much can we know, from a random variables X_i , about a parameter θ that is part of the PDF that the random variable X supposedly obeys. It follows in the sense of the Shannon information, in that it measures uncertainty. The definition for

the Fisher information is given in Eq. 100. An efficient estimator is one whose variance equals the inverse of the fisher information. The mathematical definition of efficiency is that given in Eq. 101

$$\mathcal{I}(\theta) = \text{E} \left[\left(\frac{\partial}{\partial \theta} \log f(X, \theta) \right)^2 \right] \quad (100)$$

$$\text{Ef}(\hat{\theta}) = \frac{\frac{1}{\mathcal{I}(\theta)}}{\text{Var}(\hat{\theta})} \quad (101)$$

Maximum Likelihood Estimate (MLE)

Maximum likelihood consists in choosing the parameter that maximizes the likelihood as seen in Eq. 102. The rationale behind this is that the parameter that shows the greatest coherence between model and data should be the correct one. This is frequently used in what is known as maximum likelihood fits, where a shape of a distribution is assumed for a set of data, and then the negative of the likelihood is minimized in order to obtain the parameters that maximize it. MLE has the advantage that this estimators are both efficient and consistent, which is why they are frequently used.

$$\hat{\theta} = \theta' : P(X_1, \dots, X_n | \theta') = \sup_{\theta \in \boldsymbol{\theta}} P(X_1, \dots, X_n | \theta) \quad (102)$$

4.3.2 Confidence and Prediction Intervals

Confidence intervals, (CI), are a form of estimation used to infer a parameter of a PDF devised by Jerzy Neyman. Instead of obtaining a point estimate, an interval defined by the probability it accumulates is obtained for it. This implies that ones has to assume how the random variable distributes. Under these assumptions, the limits of the confidence interval are set around the point estimate, such that the interval contains a certain probability, usually 68% or 95 % [7]. The limits of the interval will vary from sample to sample, and so the probability refers to how frequently the actual value of the parameter will lie inside the intervals when repeating the sampling and its respective CI.

However, one must be careful, for a confidence level, say 95%, does not mean that the parameter lies within the interval with a probability of 95%, as is usually assumed and taught. What it means, is that if one were to repeat this experiment several times, the parameter would fall inside the 95% confidence interval, 95 out of every 100 times. The meaning behind a confidence interval is clearly illustrated in Fig. 10. This interpretation follows from the frequentist interpretation of probability, where parameters are fixed and cannot have a probabilistic interpretation. As stated by Neyman, once the confidence interval is defined, the parameter either lies within it or it does not. It is a deterministic question.



Figure 10: Set of confidence intervals calculated for a parameter θ from different samples generated from the same distribution with a fixed value for the parameter. The red dotted line shows the value of the parameter, while the blue lines show the confidence intervals.¹¹

It is very important then, to keep in mind what the confidence interval does not mean. It does not mean that there is a probability of the confidence level that the parameter lies within the confidence interval. It also does not mean that a future value will lie within it. To calculate the probability of future values some modifications must be made, that result in more conservative measurements that introduces more variance in order to account for the error of a future sample. If this procedure is done then it is called a prediction interval, (PI). The easiest way to tell them apart is that CI are around a parameter, and PI are around a future and new element of the sample space.

The most common use for confidence intervals is defining the uncertainty in a measurement. This are usually represented as error bars which often correspond

to 1σ , which would be a 68% confidence interval for a Gaussian Distribution. This is useful to compare point estimates and establish if there are notable differences, or if they can be considered equal within the uncertainty established by the CI. Usually in lab reports for undergraduate courses, the CI is obtained from the experimental data to check if it includes the theoretical value that was expected.

The desired properties of a CI are that it is valid, optimal, and invariant. Valid means that the CI meets the required confidence level. Optimality means that the entire information is used as much as possible, meaning that no information is discarded without proper reasons. Invariance means that the same intervals are obtained should a transformation be applied to the limiting values of the interval. Confidence intervals are also used hypothesis testing. A certain null hypothesis is made, about the value of the parameter that is being estimated, and then a confidence interval is defined and a test statistic is drawn. If the test statistic falls inside the interval, then the hypothesis is not rejected. However let us take a closer look at this at the next section.

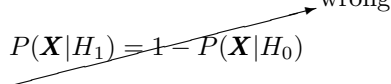
4.3.3 The logic behind Null Hypothesis Testing, p-values, and p-value controversy and caveats

The traditional way of hypothesis testing is frequently compared to how a jury reaches a guilty verdict [2]. Commonly in penal systems a person is presumed to be innocent until proven guilty, and guilty is established when the hypothesis of being innocent is very unlikely or illogical in light of the evidence or data. The rationale of Null hypothesis testing is the same, where the Null Hypothesis H_0 corresponds to the presumption of innocence, and the Alternate Hypothesis H_1 corresponds to being guilty. So just like a person is found guilty when the evidence makes the idea of innocence extremely unlikely, the alternate hypothesis is accepted once the null hypothesis is rejected for being unlikely.

When using Null Hypothesis testing, the first step is to decide an appropriate null hypothesis opposite to what one wishes to prove based on the data. Ideally, H_0 should be the contrapositive of H_1 . Think of it as though trying to prove something by contradiction. This is easier said than done, and sometimes a complete set of

¹¹This image was taken from: <https://goo.gl/sLe0cV>

hypothesis completely divisible between H_1 and H_0 is not available. However, even if both hypothesis compose a complete set, proving that one of the hypothesis is not compatible with the data only proves strictly that this hypothesis is not compatible with the data. In other words, It does not imply that the probability of the alternate hypothesis automatically takes the remaining probability, given that the sum rule does not apply to the likelihood conditioned on a hypothesis, it applies for its posterior as Eq. 103 shows. Recall that the sum rule over the hypothesis can be applied when dealing with the probability of the hypothesis itself, not on the complement of what it is conditioned. The right hand side of the equation indeed is wrong. However this is not to say that it is not useful, but rejecting the null hypothesis, as has been stated throughout the text so far, is in a strict sense only an indicator that ones inference about H_1 being true is on the right path and that H_0 is not. Knowing that one is on the right path might seem as a trivial or incomplete step, but it has proven extremely useful in both science and industry.

$$P(H_1|\mathbf{X}) = 1 - P(H_0|\mathbf{X}) \neq P(\mathbf{X}|H_1) = 1 - P(\mathbf{X}|H_0) \quad (103)$$


The matter remaining is how to decide if a hypothesis is rejected or not, and the answer to this is that it is done by calculating what is known as p-values. If the p-value is less than some pre-convened number α , then H_0 is rejected in favor of H_1 . This number α is called the significance level, and it refers to the number of times the null hypothesis could be correct given the data observed. It is closely related to a confidence interval, but let us take things slower and focus on what a p-value is. A p-value is the probability of obtaining a value at least as extreme as a given statistic or estimator $\hat{\theta}$. From here on the term statistic and estimator will be used interchangeably.

The statistic or estimator used will determine how the likelihood probability distributes according to the estimator chosen, that in turn depends on how the random variable distributes and what is known about it. All these characteristics are included on the assumptions made about the problem. For instance, if using a student's t-statistic, then this presupposes that the sample space distributes Gaus-

sian and iid, and results in that the likelihood distributes according to student's t distribution. In order to calculate the p-value correctly, then the data better comply with the assumptions of the estimator so that its use in its respective distribution is appropriate. Having chosen the statistic, the p-value can then be calculated as Eq. 104 shows. It can be calculated using both extremes of a distribution, in which case our test will be referred to as a one-sided or two-sided test. For instance, if our hypothesis H_1 is that our parameter is greater than the hypothesized value by H_0 , then we should use a one-sided sided test as shown in the first case of Eq. 104. If H_1 implies that the the value is less than that postulated by H_0 , then we should use the second case. Lastly, if H_1 states that it is simply different, then we must check the two options it has of being different, which are that it is greater or lesser than the two possible values given by the estimator.

$$p \equiv \begin{cases} P(X \geq \hat{\theta}(\mathbf{X}) | H_0) & \text{where } H_0 \Rightarrow \theta = a \text{ and } H_1 \Rightarrow \theta = b \text{ and } b > a \\ P(X \leq \hat{\theta}(\mathbf{X}) | H_0) & \text{where } H_0 \Rightarrow \theta = a \text{ and } H_1 \Rightarrow \theta = b \text{ and } b < a \\ P(X \geq \hat{\theta}(\mathbf{X}) | H_0) + P(X \leq -\hat{\theta}(\mathbf{X}) | H_0) & H_0 \Rightarrow \theta = a \text{ and } H_1 \Rightarrow \theta \neq a \end{cases} \quad (104)$$

In Fig. 11(a) we can see how a p-value looks like for a on sided test. The darkest shaded area is the p-value while the lighter shade of gray is the critical region, that groups the set of values that have of probability of at least α of lying in the right tail of distribution. If α is greater than the p-value, then H_0 is rejected. On Fig. 11(b) we can see the region that has a probability equal to the p-value coloured in blue.

The value for α is usually convened to be 1% or 5%. The statistic or estimator used on the data should appropriately summarize its information, for it is based on this number that we will test the hypothesis. If the statistic that summarizes all of the data falls in the rejection region, something that is assumed to happen only α percentage of times out of a hundred, then it is considered too unlikely to be explained by Hypothesis H_0 . For instance for $\alpha = 5\%$, H_0 would only bare result in this region 5% of the time. This why α is also known as the false positive rate,

¹²The one sided Null Hypothesis Test image was taken from: <http://goo.gl/9M0DPt>

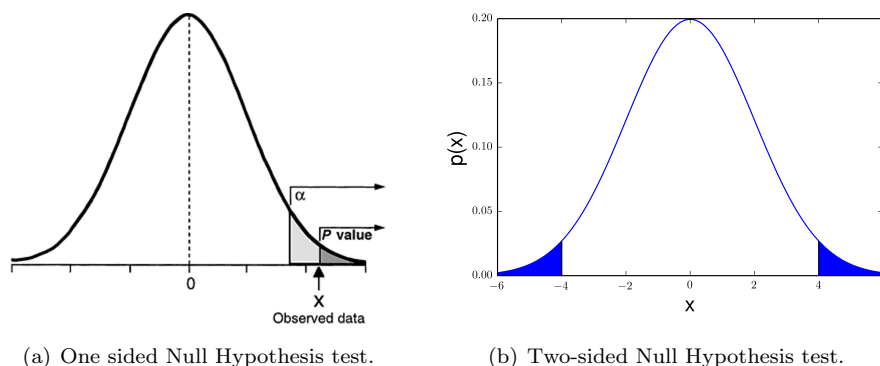


Figure 11: How a p-value looks for two of the three types of p-values for hypothesis tests.¹²

for 5 times out of a hundred the null hypothesis will produce data that results in a statistic that falls in this region. So if a test that should not result in the rejection of the hypothesis were to be carried out a 100 times, we would expect that it is rejected 5 times.

The reader might have noticed that confidence intervals are constructed in a very similar fashion, and indeed the CI's are closely related to hypothesis tests. When one finds the limits of a confidence interval, it is done by taking the estimate $\hat{\theta}((X))$ for the parameter θ and sets it as the true parameter of the distribution. The limits for the interval are then found by finding the values a and b for which $P(X \geq a) = \alpha$ and $P(X \leq b) = \alpha$. However you might think that it does not seem to follow the logic of proof by contradiction stated earlier. The procedure just outlined for CI is equivalent to setting θ to any of the values excluded and carrying out the null hypothesis test with $\hat{\theta}((X))$ as our statistic.

The quality of the inference depends greatly on the appropriate choice of the estimator, that as has been mentioned, in the context of hypothesis testing is usually referred to as an statistic. If the statistic is not chosen appropriately, then it might not use the data correctly by assuming information that is not present in it. An example of a misuse of a statistic is using a Z-statistic, that assumes that one knows the variance and mean of the iid samples of the random variables

X , when one does not know its variance. There are different statistics designed for different conditions of information on the data. Some statistic are to be used when one does not know the variance and the mean of the distribution, or when the sample is not composed of an iid set of random variables, and so on. Failing to choose the correct statistic for the specific conditions of the elements of the sample space will lead to an erroneous inference, for instance by failing to reject a null hypothesis that should be rejected, or vice versa.

The quality of the inference depends also on the quality of the data, and if it indeed was selected at random from the population space. If data is not selected at random, then it will introduce biases to the inferences, deviating them to the point where they can be plainly mistaken. We are all familiar with the fact that sometimes statistics may be used to lie or draw false conclusions, and this usually occurs when data is corrupted by biases when gathering them. The topic of biases is also a wide one, and there are many examples about them and how to avoid them, such as blind studies in medicine [2].

There have been some critiques to p-value since its conception by Ronald Fisher, and it has currently become a heated discussion. Many of the critiques are centered around the fact that the test only discards the null hypothesis and so a hypothesis test should be carried out under the assumption that it is true. This critique was formulated by Neyman and Pearson, and they postulated other measurements such as statistical power that are based on the likelihood conditioned on H_1 or any number of alternate hypotheses H_i being true. This is known as Neyman and Pearson's Decision Theory and will be explained on the next section.

Another of the critiques is that given its false positive rate, a study may be repeated until by chance alone the statistic lands on the critical region and so the Hypothesis is rejected even though it should not be. This is known as p-hacking and it presupposes dishonesty on part of the person carrying out the test. However even if the person is honest and the results from the likelihood are real, if the hypothesis has a low prior probability, then the probability of a real effect is downplayed just like in the numerical example of bayesian inference. The infographic shown in Fig. 12 appeared in an article in Nature, and it summarizes this point perfectly [22].

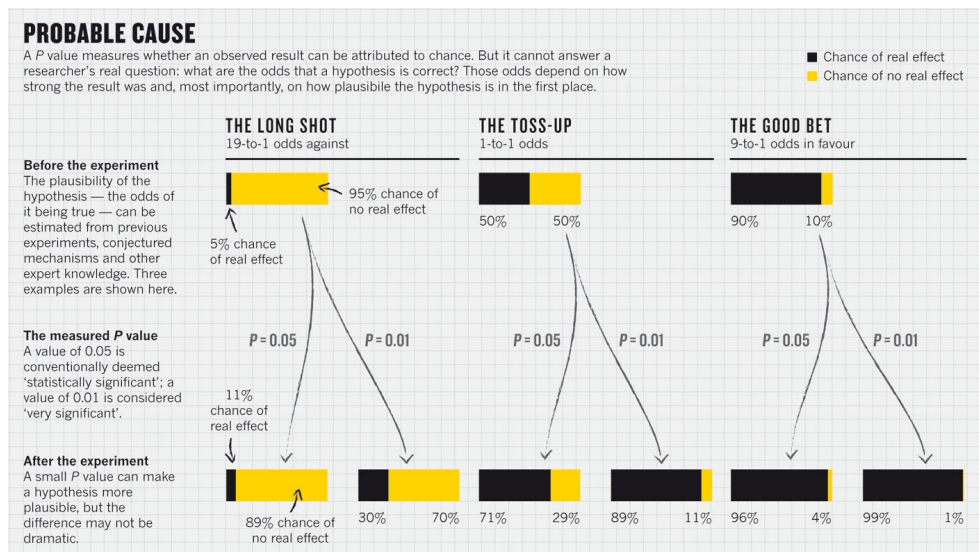


Figure 12: Explanation for one of the problems of using p-values, especially to how the inference changes given a different prior probability [22].

Most of the current critiques are addressed at the fact that only Null Hypothesis tests are used, when they were devised in the first place by Fisher as a first step verification. Some argue that if Neyman and Pearson's decision theory were taught more frequently, many of these problems might be avoided. However some articles also take statistical power into account and have arrived at some intriguing and even troubling results [24]. This article investigates the meta-distribution of p-values. By meta-distribution it is meant the distribution found for p-values should a Null Hypothesis test be carried out several times with different sets of simulated data generated from the same distribution. What was found is that the distribution of p-values does not necessarily follow a consistent behavior with that of the false positive rate and that of confidence intervals, meaning effects might be accepted by chance alone much more frequently than α given the statistics and tests used. Given that the p-value distribution does not follow a consistent behavior, the reproducibility of inferences made from them is seriously under question.

In spite of this, Null Hypothesis testing is a widely used technique, and not always a misleading one. Even though a pretty bleak evaluation for p-values was given, when the significance α takes on much more stringent values, as it does in High Energy Physics, then the possibility of a false rejection diminishes greatly. In order to declare the discovery of a particle on HEP, the critical region must amount a probability in the order of 10^{-7} that corresponds to being 5σ away from the mean in a Gaussian Distribution. On the other hand, in order to exclude a possible mass or any other physical parameter of a particle, the significance level is set at $\alpha = 5\%$. The pitfalls just mentioned for p-values in the context of HEP, especially for exclusion, might be avoided because of the extraordinary high number of data gathered on its experiments, perhaps much higher than those of any other field. It is nothing more than a conjecture, but this is probably the reason that inferences made on HEP based on p-values are much more consistent, because they are constantly being put the test.

4.3.4 Neyman and Pearson's Hypotheses Testing and Decision Theory

Neyman and Pearson proposed that one should have tests for both the null and alternate hypothesis. The idea remains that the two hypothesis should be their logical opposite in a best case scenario, but instead of quantifying only how much the H_0 does not agree with the data, one should also quantify how much H_1 does. To emphasize this, when using their decision theory one employs the notation H_1 and H_2 . This is the reason why their way of hypothesis testing is also called a decision theory, for it decides and tests both hypotheses with the data. Doing so allows one to calculate both the false positive rate α and the false negative rate β . These are also known as Type I error and Type II error respectively. To test how much the hypothesis that one defines as a positive result, one calculates the likelihood of having values higher or lower than the statistic $\hat{\theta}(\mathbf{X})$. Whether the test is for higher values depends on the kind of test carried out for what is defined as a positive result. The logic is that shown in Eq. 105.

$$\text{Statistical Power} = 1 - \beta \quad (105)$$

$$\text{Where } \beta = \begin{cases} 1 - P(X \geq \hat{\theta}(\mathbf{X}) | H_2) & \text{if } \alpha = P(X \geq \hat{\theta}(\mathbf{X}) | H_1) \\ 1 - P(X \leq \hat{\theta}(\mathbf{X}) | H_2) & \text{if } \alpha = P(X \leq \hat{\theta}(\mathbf{X}) | H_1) \end{cases} \quad (106)$$

Statistical power tells us then of the likelihood of hypothesis H_2 to explain our data given that it is true, not just how unlikely hypothesis H_1 is. This becomes particularly useful when the H_2 is not the logical opposite of H_1 , and thus the quantitative leap from rejection of H_1 implying H_2 becomes much more blurred. With statistical power and certain assumptions about how the statistical power distributes, it is possible to find the required sample size n for a fixed α and β . This is particularly useful for samples that follow the CLT, for their usually their mean μ is set by hypothesis, and the standard deviation σ depends on \sqrt{n} .

Effect Size

The Neyman-Pearson approach gives a sense and a measure of the magnitude of the effects being tested, while null hypothesis testing may not. For instance, the p-value significance test can establish a difference between populations, but does not say anything of the likelihood of the measured difference given by the statistic. This difference is called the effect size. It is not that a null hypothesis test does not give a quantitative figure for the effect size, because it does. What occurs is that the numerical value of the statistic is only used to reject the hypothesis H_1 or H_0 in null hypothesis testing, and nothing is said of how likely this value actually is when hypothesis H_2 is true.

4.3.5 Likelihood Ratio and Wilk's Theorem

This is a type of statistic used carry out tests based only on the likelihood of a model and null and alternate hypothesis with certain degrees of freedom as shown in Eq. 107. The idea is to compare directly how much more does one of the hypotheses agree with data than the other by taking their ratio. The comparison is done by squaring the ratio and taking its logarithm, because doing so leads to

obtaining a statistic that distributes according to a χ^2 distribution as the sample size tends to infinity. This is known as Wilk's Theorem, and it allows us to calculate the p-value of the likelihood ratio by using the distribution it tends to. This test is not only because it can be approximated to a χ^2 distribution, but also it is the test that bears the maximum statistical power for a given significance α .

$$t = -2 \log(\Lambda(x)) = -2 \log \frac{L(\theta_0|x)}{L(\theta_1|x)} = -2 \log \frac{f(x|\theta_0)}{f(x|\theta_1)} \quad (107)$$

5 Statistical Inference in HEP

On HEP the data gathered experimentally is usually represented in terms of histograms, and the convention is that inferences are made and reported according to standards of frequentist inference. To be specific, it is usually done in terms of null hypothesis testing, where the p-value is reported in terms of how many standard deviations σ it is from a mean μ in a standardized normal distribution $N(0, 1) = P\left(\frac{x-\mu}{\sigma}|\mu, \sigma^2\right)$. The significance levels are also reported in terms of what they would be for a Gaussian distribution, and the convened values for α to accept the discovery or exclusion of physical characteristics of a particle are 5σ and 2σ respectively. These two values correspond to $\alpha = 3*10^{-7}$ for discovery and $\alpha = 0.05$ for exclusion [25]. Some methods make use of marginalization over priors in order to have a better estimate of the likelihood and then carry out a hypothesis test with this newly constructed likelihood. This is known as a hybrid approach [26]. However, the fact remains that Null Hypothesis testing is the norm regarding the statistical tests carried out in this field, and that p-values and confidence levels based on them is what is usually reported.

The main focus of statistical inference on HEP is to prove experimentally with the data gathered the existence of physics beyond what has been demonstrated experimentally so far, or to reject with the same data the existence of such physics for the physical parameters under which the hypothesis test was carried out. At present time this means gathering data that proves the existence of physics beyond the standard model, abbreviated as BSM. Perhaps one of the most iconic examples of BSM is Super Symmetry (SUSY), which is the next target of many physicists working at the LHC, after having gathered enough data to declare the discovery of the Higgs Boson in 2012.

The usual way of carrying out a hypothesis test to declare a discovery in HEP is that the measurements that are expected from the Standard Model (SM) are the null hypothesis H_0 , usually referred to as the background hypothesis. The alternate hypothesis on the other hand is a variation of the background hypothesis, where H_1 includes both the detection of a signal generated by BSM physics and what is expected from SM physics. In order to test whether the data gathered is not consistent with BSM physics, one inverts the hypotheses. In other words, to test

if the physical parameters of hypothesis H_1 should be excluded, one takes H_1 as the null hypothesis and draws an appropriate statistic from the data and checks if it falls in the critical region. If it does, then the parameters are excluded, and one can assume that these parameters of BSM physics are not consistent with what is measured from nature, and thus are discarded and attributed to random behavior and statistical noise.

The null hypothesis test is carried out this way, because the detectors in particles accelerators and other experimental setups of HEP, measure kinematic variables of the final states of scattering processes in order to identify them. Doing this process is a gigantic task and a great challenge in both experimental design and data processing. Even if one were able to classify them perfectly, one still has that different scattering processes can have the same final states. This implies that an additional task remains, and that is to differentiate which of the detected final states came from an already demonstrated SM scattering, and which from the BSM physics that one is trying to measure. This is the reason why hypothesis testing is carried out as a background plus signal model, because we have a certain amount of final states detected that correspond to the background or SM physics, and the rest that corresponds to the signal or BSM Physics.

Additional to this, a scattering process can have different final states, each of them with a certain probability also called cross section or branching ratio. Each type of final state is referred to as a channel. The name is quite appropriate, since different channels carry different information, because each of them has different types of background from SM scatterings. Just like in a communications channel, it is not only about raw amount of information than can be transferred, it is also about the noise of the channel and how much it allows to appropriately interpret this information. Following our analogy, the raw capacity of the channel is the branching ratio of the BSM signal, and the noise is the background from SM Physics that produce the same kind of final state.

In order to obtain the most information, one has to juggle between signal intensity and the intensity of its background. For instance, in the case of the search for the Higgs Boson, some channels had a higher probability of producing the Higgs Boson itself, but the end state that was detected by the LHC detectors were also produced by several other processes. While these channels produced a

higher number of Higgs Bosons, it was much harder to identify them statistically given that the signal of their final state was dwarfed by that of the background.

Given that calculations of Quantum Field Theory of these types of processes are quite cumbersome and difficult, they are generated and quantified by using software such as MadGraph [27], Pythia [28] and Delphes [29]. The combination of this software can simulate the amount of final states, their respective kinematic variables, and how they are detected at the LHC either at CMS or ATLAS for each scattering process separately. The output of this software is usually presented on the form of histograms. The simulations have as input parameters the energy of the collision, the number of particles involved in it, as well as many other conditions such as the precision and error of the instruments used at these experiments. With these softwares, it is possible to have a quantitative theoretical prediction of what one hopes to detect on the LHC from BSM and SM physics independently. While the experimental measurements are of all processes added together, with the software we can quantify what we expect from each scattering process.

In order to deal with the fact that sometimes the background signal has a much higher count or intensity than the signal, it is customary to look for sections where the signal surpasses the background. Once this sections are located, filters applied on their variables, usually known as cuts, are applied in order to select this area of interest. This is done with a software built upon C++ developed at CERN called ROOT [30]. It is then over this reduced set, that one constructs the model for the likelihood and carries out the null hypothesis test.

With a theoretical prediction for each process, one can construct a model for the PDF of the likelihood of each of them. Of course, the PDF's do not need to belong to the same family of distributions, and it is often not the case. In order to construct a model for the likelihood, one can assume the family of the distribution, and try to fit the parameters of it to the simulated data, and check if it bares a good fit. the goodness of fit is usually established through an statistic known as the χ^2 , that measures the squared error between the data and the proposed model for the PDF. The fit is usually made through a maximum likelihood estimation of the parameters given the data simulated. The family that bears the least value for χ^2 can be chosen, if its value is low enough. However, this is not the only option. Instead of building a model for each scattering process, which is a time

consuming task, one can build what is known as a binned likelihood directly from the histograms produced by the simulations. The option of a binned likelihood may result in a lower significance, but it is really useful to obtain a likelihood for a large group of simulated signals.

However, if one chooses to model the likelihood, all the parameters that are not set by the null hypothesis are usually left undetermined, given that they ought to be determined by the data collected experimentally. Once the model for the likelihood is defined, the parameters are left with their values unspecified, and they are determined by a maximum likelihood fit to the data. The reason behind this is that one wishes to have a model that represents the data as best as possible. Once the distribution of the likelihood has been fixed, one chooses a statistic and calculates it based on the data, and carries out the null hypothesis test. The usual statistic chosen is the likelihood ratio [25].

The simulated data then serves the purpose of being pseudo experiments that helps us formulate a model for each type of scattering process individually, and then fitting their sum to the actual data to later execute a hypothesis test. The background is simulated in order to be able to reject the background only hypothesis and attempt to establish discovery. If it fails, then one needs a simulated signal to add it to that of the background, and from its model construct the likelihood for the new null hypothesis, that if rejected will lead to the exclusion of the assumptions made when simulating the signal.

The procedure can be summarized in the following sequence of steps:

1. Define a channel for the process that one wishes to discover or establish exclusion regions for.
2. Identify the background from SM physics that have the same final states.
3. Choose the physical parameters for the BSM physics process or particle such as its invariant mass or cross section.
4. Simulate how each of this processes will be detected on CMS or ATLAS.

The procedure in the phenomenology group at Los Andes is the following.

- (a) Simulate the partons produced by the collision using MadGraph.
 - (b) Simulate the hadronization of this partons using Pythia.
 - (c) Simulate how this particles are detected by CMS or ATLAS using Delphes.
5. Apply a sequence of cuts in order to separate as much as possible the signal from the background.
6. Choose a model or family of distributions for the likelihood of each of the scattering processes simulated. Define the model without fixing its parameters yet. There are two options for this:
 - (a) Choose a pool of families of distributions and carry out Maximum Likelihood Fits to the simulated data. Choose the distribution that shows the best fit among the ones that have an acceptable value for χ^2 . Be careful of over-fitting.
 - (b) Construct a binned likelihood model.
7. Calculate a maximum likelihood estimate for the free parameters of the model by doing a maximum likelihood fit to the actual experimental data.
8. Calculate an appropriate statistic from the experimental data, usually a likelihood ratio statistic, and carry out the Null Hypothesis test for the PDF defined for the likelihood of the background hypothesis.
9. If discovery cannot be established, then do a new hypothesis test with the same statistic, but this time for the alternate hypothesis of signal plus background in order to establish exclusion.
10. Repeat the whole process from step 3 by changing one of the physical parameters and start constructing a rejection region, or nearing the true value that will bear a higher significance for discovery.

Phenomenology

The previous procedure outlined is for when one has access to experimental data. However it is often required to know beforehand where to look, given that even the processing of information in the LHC is a limited resource because of the sheer amount of data produced. This is where phenomenology comes in. The procedure outlined before to simulate both the background and the signal as how they would be measured at the LHC is a part of phenomenology. The process is the same as that outlined before up to step seven.

Given that there is no experimental data, pseudo-data must be generated in order to calculate a statistic and do the required hypotheses tests. The way to generate pseudo-data is to first add the histograms of all the backgrounds and the signal. In an experiment it is highly unlikely that we obtain exactly what we expect from theory, this means some randomness around the value of each bin should be introduced to account for it. In the present work, this randomness was taken into account by assuming it distributed Gaussian. A Gaussian distribution was defined by setting mean μ equal to the value of the bin n_i , and σ as the square root of that value. A number N of samples is generated from the previously defined Gaussian distribution, and the sample mean and sample variance from this sample is calculated. The sample variance is the new point of pseudo-data, while the sample mean is the estimated error or uncertainty for this new point.

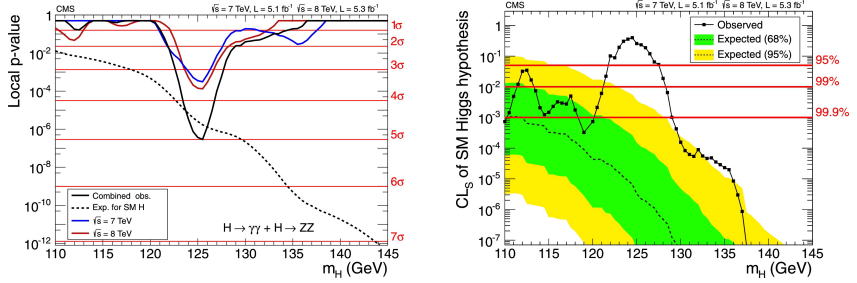
The pseudo-data is useful to find out what would happen if the nature behaves exactly as assumed when generating the simulated data for the signal. In a sense it is an analysis of a best case scenario where the agreement of our hypothesized signal and what one should measure from nature is exact, and only differs because of the inherent randomness and error of measurements. Once the pseudo-data is generated, the null hypotheses tests are carried out just the same way but by replacing the data by the pseudodata. Let us specify how the procedure changes from step seven.

7. Add all histograms of all the background processes and the histogram of the signal.

8. Define the PDF with which the pseudo-data for each bin will be generated by using the value or frequency of such bin.
9. Generate data from the defined PDF and set the value of the bin to the sample mean, and the error to the sample variance.
10. Calculate a maximum likelihood estimate for the free parameters of the model chosen by doing a maximum likelihood fit to the pseudo-data.
11. Calculate an appropriate statistic from the experimental data, usually a likelihood ratio statistic, and carry out the Null Hypothesis test for the PDF defined for the likelihood of the background hypothesis.
12. If discovery cannot be established, then do a new hypothesis test with the same statistic, but this time for the alternate hypothesis of signal plus background in order to establish exclusion.
13. Repeat the whole process from step 3 by changing one of the physical parameters and start constructing a rejection region, or nearing the true value that will bear a higher significance for discovery.

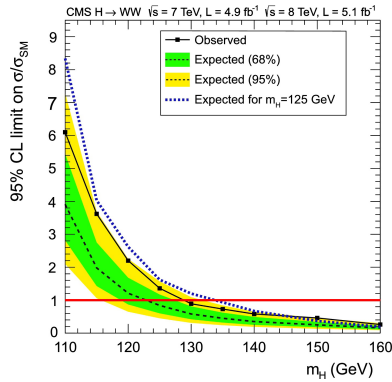
After repeating these steps several times for different simulated signals, one should obtain graphs similar to the ones shown in Fig. 13 and taken from reference [7]. The first graph in Fig. 13(a) shows how to present a hypothesis test for discovery that is carried out for several signal point from different channels. They almost reach the mark of 5σ for a certain range, which is quite significant. If one wishes to perhaps achieve discovery, one should focus on this segment of energy and gather more data in hopes of achieving a higher significance.

The second Fig. 13(b) shows the hypothesis test for also a group of signal points or histograms but by taking the signal plus background hypothesis as the null hypothesis. The values shown are a quantity called the confidence level of the signal and it is very similar to the p-value and serves the same function, if smaller than α then that point of signal is rejected. A collection of rejected points constitute then an exclusion region, as that shown in the graph for a mass higher than 129 GeV and lower than 123 GeV.



(a) P-values for Hypothesis tests aimed at discovery for several signal points.

(b) Special kind of p-value called CL_S used to test for exclude or reject a point when $CL_S \leq \alpha$.



(c) Finds a exclusion point by solving for the value of cross section where p-value $= \alpha$.

Figure 13: How the different hypothesis tests for several signal points are represented in order to find regions of discovery and exclusion [4]

Fig. 13(c) shows what is also called an inversion of a hypothesis test. Its purpose is to find for what value of a parameter of interest such as the signals cross section, or its invariant mass, is the signal rejected. From this point value onward, the signal is rejected, so it establishes regions of exclusions for a fixed value of other of the parameters of interest. In the case of the graph shown, the mass is fixed, and the

regions of rejection for signal cross section is found. The whole process is summed up in the flow chart shown in Fig. 14.

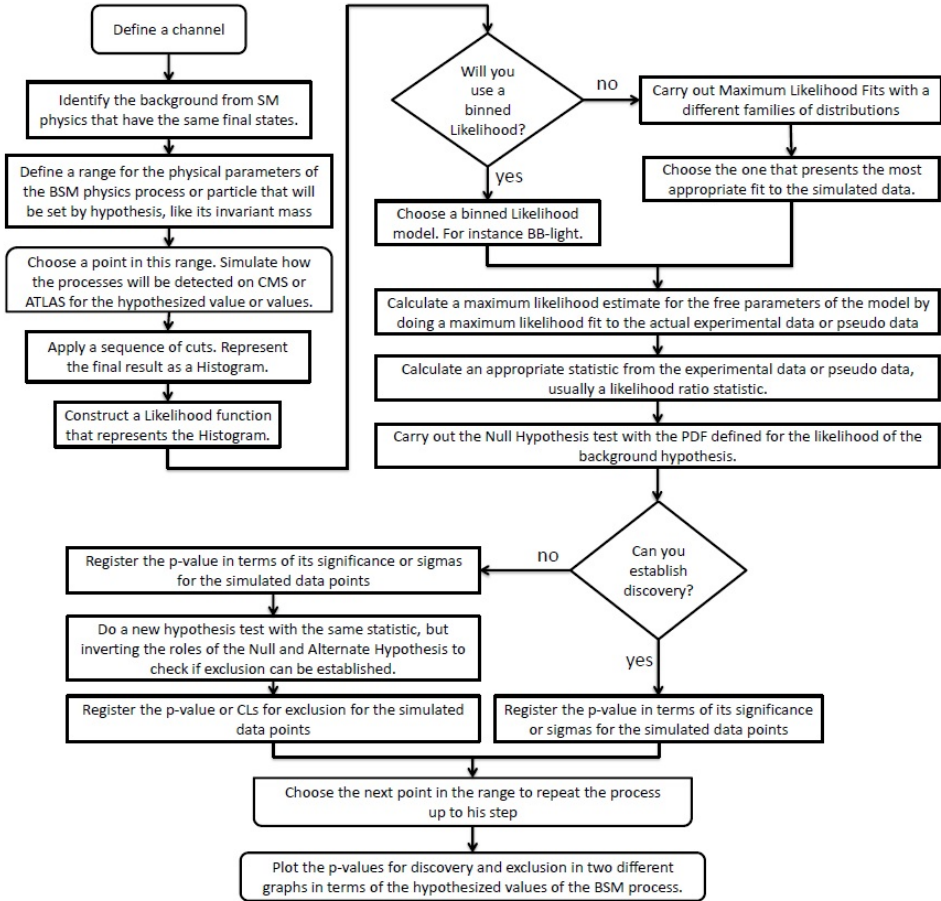


Figure 14: Flow Chart explaining roughly the process in a phenomenology study in HEP

CL_s Intervals

The quantity CL_s is calculated by using both the p-values p_0 and p_1 of both the background only hypothesis H_0 and the background plus signal hypothesis H_1 respectively [7]. With this two quantities the confidence level of the signal is calculated according to Eq. 108. One re-scales the quantity p_1 to take into account the fact that wishes to reject only the signal and not the signal plus background. The background has already been established as true, so by dividing by $1 - p_0$, new quantity CL_s becomes the actual p-value of just the signal given the data, not of the entire set of signal plus background.

$$CL_s = \frac{p_1}{1 - p_0} \quad (108)$$

5.0.1 Hypotheses Tests using the Extended Maximum Likelihood

There is a quantity related to the likelihood referred to as the extended likelihood. The difference lies in that aside from the distribution found to be the best fit and represent a given hypothesis, a rate of occurrence λ is introduced. In other words, one assumes that this data has a rate of occurrence, that as we saw very early in the text, is described by a Poisson distribution. This means that the likelihood changes as shown in Eq. 109. Adding this parameter and estimating it by maximizing the likelihood results in a better model altogether to represent the data and better estimates [7].

$$L(\mathbf{X}|\boldsymbol{\theta}) \rightarrow \frac{\lambda^N e^{-\lambda}}{N!} \prod_i^N L_i(x_i|\boldsymbol{\theta}) \quad (109)$$

5.0.2 Extended Likelihood for Histograms or Binned Likelihood

The method mentioned before to construct a model for the likelihood directly from the histogram consists of modeling the probability of obtaining the frequency for each of the bins independently. The motivation for obtaining a likelihood directly from the histogram without attempting to fit different families of distributions is that the second option is more labor intensive. The binned likelihood is particularly favored in preliminary phenomenological studies to identify regions of interest of

possible discovery and exclusion. Even though constructing a likelihood directly from the histogram may result in a lower significance for the hypothesis tests, this approach can be executed and automated for a large group of simulated signals. If more significance is required for certain signals, then a smaller group of signals may be fitted by proposing families of distribution that describe it.

The binned likelihood without extending is a multinomial distribution as shown in Eq. 110. This equation applies for a number of data N , distributed in a number of bins B . The parameter n_i corresponds to the number of elements from the data \mathbf{X} that lie in the range of bin i . The factor $P_i(\boldsymbol{\theta})$ is the probability that data falls within the bin i . If we extend the likelihood of Eq. 110 we will arrive at a PDF that represents better the scattering processes.

$$L(\mathbf{X}|\boldsymbol{\theta}) = N! \prod_{i=1}^B \frac{P_i(\boldsymbol{\theta})^{n_i}}{n_i!} \quad \text{where } P_i(\boldsymbol{\theta}) = \int_{x_i^{low}}^{x_i^{up}} f(x|\boldsymbol{\theta}) dx \quad (110)$$

To extend the likelihood we must multiply the previous term by a rate that distributes according to a Poisson distribution as seen in Eq. 111

$$L(\mathbf{X}|\boldsymbol{\theta}) = \frac{\lambda^N e^{-\lambda}}{N!} \cdot N! \prod_{i=1}^B \frac{P_i(\boldsymbol{\theta})^{n_i}}{n_i!} \quad (111)$$

Defining $\nu_i \equiv \lambda P_i(\boldsymbol{\theta}) \rightarrow \sum_i^B \nu_i = \sum_i^B \lambda P_i(\boldsymbol{\theta}) = \lambda \sum_i^B \cancel{P_i(\boldsymbol{\theta})}^{\rightarrow 1} = \lambda$ by the sum rule. Also, using the fact that $N = \sum_i^B n_i$ we can carry out the following changes:

$$L(\mathbf{X}|\boldsymbol{\theta}) = \lambda^{\sum_i^B n_i} e^{\sum_i^B \nu_i} \prod_{i=1}^B \frac{P_i(\boldsymbol{\theta})^{n_i}}{n_i!} = \prod_{i=1}^B \frac{(\lambda P_i(\boldsymbol{\theta}))^{n_i} e^{\nu_i}}{n_i!} = \prod_{i=1}^B \frac{\nu_i^{n_i} e^{\nu_i}}{n_i!} \quad (112)$$

With the result of Eq. 112 we can construct an appropriate likelihood directly from the histogram without going through the process of choosing from a pool of families which fits best the simulated data. What is needed is an appropriate choice for the rate of occurrence ν_i . The appropriate choice is that shown in Eq. 113, where s_i represents what the signal contributes to the total rate. The other term b_i has the same meaning but for the background. The parameter μ will be the parameter of interest, given that changing its value changes contribution of

the signal. Notice that if $\mu = 0$, then the likelihood will become the likelihood of the background only hypothesis. On the other hand if $\mu = 1$, then it will equal the signal plus background hypothesis. When doing a hypothesis inversion, the parameter μ is sampled until it results in a $CL_s \leq \alpha$.

$$\nu_i = s_i\mu + b_i \quad (113)$$

In either of the three tests, one fixes a value for μ and evaluates a likelihood ratio statistic, with the parameters b_i and s_i being set by the respective value they take in the histograms that result from simulating the channel in the procedure outlined before. One could attempt a Maximum Likelihood fit, but one has no guarantee that the parameter estimated corresponds or is consistent with what was observed in the simulated data. Let us take a closer look in the next section at how to allow the parameters to vary appropriately to look for a maximum fit by introducing constraints that use the simulated data as auxiliary measurements.

5.0.3 Nuisance Parameters and the Beeston Barlow method

In order to be able to treat the parameters b_i , s_i , and $s_i + b_i$ as variables to be maximized one must resort to treat the results from the simulated data as auxiliary measurements. This has the consequence that they can now vary according to the PDF of the auxiliary measurement. Introducing the auxiliary measurements allows one to treat a parameter as a constraint and doing a maximum likelihood estimate (MLE), effectively treating it as a nuisance parameters (NP). At first hand this seems like introducing more variance and unnecessary parameters to the problem, however the fact is that they *already are in* the problem in the form of $s_i + b_i$ and one is not taking properly into account their random behavior by not modeling the auxiliary measurements used.

A NP is a parameter that is of no physical interest or that is not being tested for rejection by the hypothesis test. An example of a NP is the variance or noise in the data. The NP makes more difficult the analysis and usually lowers the possible significance attained by the experiment. Given that one cannot remove them completely, the remaining option is to downplay their negative effect as much as possible by finding the MLE for this parameters and set them equal

to this estimate. This reduces the impact of the variance of our NP resulting in a likelihood model or PDF that better interprets the data. In our analysis, all parameters that are not the one that is fixed to test the hypotheses is a nuisance parameter. This means that in the case of the bin likelihood, the parameter of interest is μ , and all the other parameters are nuisance parameters.

Let us take a look at how to rewrite the likelihood to include the auxiliary measurements from the simulation. The procedure is to include the measurements made along with the experimental data as shows Eq. 114. We separate the likelihood for each of the different measurements by using the product rule to obtain a multiplication of three likelihoods, one for each type of measurement as in Eq. 115

$$L\left(\mathbf{X}, \tilde{b}_i, \tilde{s}_i | \mu, b_i, s_i\right) = L\left(\mathbf{X}, \tilde{s}_i | \mu, b_i, s_i, \tilde{b}_i\right) L\left(\tilde{b}_i | \mu, b_i, s_i\right) \text{ by the product rule} \quad (114)$$

$$L\left(\mathbf{X}, \tilde{b}_i, \tilde{s}_i | \mu, b_i, s_i\right) = L\left(\mathbf{X} | \mu, b_i, s_i, \tilde{b}_i, \tilde{s}_i\right) L\left(\tilde{b}_i | \mu, b_i, s_i\right) L\left(\tilde{s}_i | \mu, b_i, s_i, \tilde{b}_i\right) \quad (115)$$

Given that the measurements \tilde{b}_i do not depend on μ, b_i, s_i one can discard them from the conditionals. The same line of reasoning applies for or \tilde{s}_i with respect to $\mu, b_i, s_i, \tilde{b}_i$. These two facts lead to the from shown on Eq. 116, that when replaced by their respective Poisson PDF results in Eq. 117. This form for the likelihood is known as the Beeston-Barlow method, and while it results in a greater significance, for large amounts of data can become computationally expensive [31]. When computational cost becomes an issue, there is a variation of this method that only adds one additional parameter known as the Beeston-Barlow light method [32]. The idea is to constraint the values of both b_i and s_i only by their relation to the total expected measurements according to the simulations. This means that

the NP that will constraint them will be $\tilde{s}_i + \tilde{b}_i$ as Eq. 118 shows.

$$L(\mathbf{X}, \tilde{b}_i, \tilde{s}_i | \mu, b_i, s_i) = L(\mathbf{X} | \mu, b_i, s_i, \tilde{b}_i, \tilde{s}_i) L(\tilde{b}_i | b_i) L(\tilde{s}_i | s_i) \quad (116)$$

$$L(\mathbf{X}, \tilde{b}_i, \tilde{s}_i | \mu, b_i, s_i) = \prod_{i=1}^B \frac{(s_i \mu_i + b_i)^{n_i} e^{s_i \mu_i + b_i}}{n_i!} \prod_{i=1}^B \frac{(b_i)^{\tilde{b}_i} e^{b_i}}{\tilde{b}_i} \prod_{i=1}^B \frac{(s_i)^{\tilde{s}_i} e^{s_i}}{\tilde{s}_i} \quad (117)$$

$$L(\mathbf{X}, \tilde{b}_i, \tilde{s}_i | \mu, b_i, s_i) = \prod_{i=1}^B \frac{(s_i \mu_i + b_i)^{n_i} e^{s_i \mu_i + b_i}}{n_i!} \prod_{i=1}^B \frac{(s_i + b_i)^{\tilde{s}_i + \tilde{b}_i} e^{s_i + b_i}}{\tilde{s}_i + \tilde{b}_i} \quad (118)$$

5.0.4 Profile Likelihood Ratio

It is an additional type of modification to the likelihood ratio, in which the ratio is not taken to the alternate hypothesis by setting $\mu = 1$, but it is taken to the MLE for μ represented as $\hat{\mu}$. The likelihood ratio obtained then is not that shown in the left hand side of Eq. 119, but that of the right hand side. In the notation used $\boldsymbol{\theta}$ represents the nuisance parameters of the model. The estimate $\hat{\boldsymbol{\theta}}$ is the MLE for a fixed value of μ , while $\hat{\boldsymbol{\theta}}$ is the MLE when considering all the possible values of μ . The motivation for changing the likelihood in the denominator is to compare the null hypothesis to the model that best fits the data, given that the alternate hypothesis may not represent the data well.

$$\frac{L(\mathbf{X} | H_0 \Rightarrow \mu = 0)}{L(\mathbf{X} | H_1 \Rightarrow \mu = 1)} \xrightarrow{\text{Profile LL}} \frac{L(\mathbf{X} | \mu = 0, \hat{\boldsymbol{\theta}})}{L(\mathbf{X} | \hat{\mu}, \hat{\boldsymbol{\theta}})} \quad (119)$$

The term profile likelihood is used because the likelihood is calculated for a fixed value of μ . This means that it is profiled in terms of the possible values of μ . This approach is specially popular when looking for the μ that results in the highest likelihood, but it is too computationally expensive to calculate directly.

6 Scripts using RooFit and RooStats

The scripts written in order to carry out the hypothesis tests consist of using toolkits of Root designed for this purpose. The main toolkits are a couple of complementary toolkits developed by the same authors called RooFit and RooStats. The main purpose of RooFit is to easily create PDF's in a ROOT environment, and fit them to a set of data with a very small set of code. For instance one can create a PDF with a single line, and with a single line fit the PDF to the data. RooFit has implemented all the methods to generate data from this PDF's, to graph them, carry out fits through MLE or χ^2 minimization among several methods, or the addition or multiplication of PDF's. Basically any operation that one requires to carry out on a PDF except for hypothesis tests are implemented on RooFit. RooStats is a toolkit that is built upon RooStats, and uses the classes defined by RooFit to carry out the hypothesis tests also with a small number of lines.

There is a third toolkit that allows us to construct the binned likelihoods, and do the hypothesis tests directly from the histograms, and it is called HistFactory. This method will be the bridge between the histograms saved in .ROOT files, and the methods of RooStats to carry out hypotheses tests. Let us first list the most useful references used in from the author's perspective, an the ones that he resorted to the most when writing the script that will be presented further on.

ROOT The two references resorted to the most where [33] and [34]. The first one is directed at beginners as a tutorial, while the second one is mainly used when having to use something very specific. Given that the hypothesis tests are made on binned likelihoods constructed from histograms, I focused on the section of creating, manipulating, and importing and exporting histograms. The main classes used were:

- TH1F: This is the class for histograms.
- TCanvas: This is the class on which plots are built.
- TFile: This is the class to import and export files.
- TTree: This is the class to in which data is stored.

- `gROOT->ProcessLine("<ROOT Command">)`. This is not a class, but it is very useful command to use other macros from inside a macro.

RooFit The main references I relied upon for this toolkit where [35] and the tutorials that are in the tutorials directory of ROOT. This reference is from a course given by one of the lead developers of RooFit and RooStats, and it has a very good compression of the most useful commands of RooFit. The second reference I relied when I needed to look for something specific was RooFit's guide [36]. The most important classes used from this toolkit where:

- `RooRealVar`: It is the class on which the variables and their range is defined.
- `RooAbsPdf`: It is the class to create PDF's, and it uses objects from `RooRealVar` to define the ranges or values of the PDF.
- `RooDataSet`: It is the class on which unbinned data is stored. When fitting a `RooAbsPdf` object, it is fitted to an `RooAbsData` object.
- `RooDataHist`: It is the class that used to handle histograms of RooFit.
- `RooAddPdf` and `RooProdPdf`: they calculate the sum and products of two PDF's, respectively. If one wants to do the operation over more than two PDF's, one should group them using `RooArgset`.
- `RooWorkspace`: It is perhaps the most important class of RooFit, because it has built in it all the other classes mentioned above. With a single line one can create a PDF, a function, etc. If you learn how to use it, most of RooFits functions are at your hand.
- `RooParamHistFunc`: It is used to characterize treat histogram as a PDF. Using it along with `RooHistConstraint` allow one to construct a Binned Likelihood from a histogram, though it will not be the method we use.
- `ModelConfig`: It is used to construct a small set that contains a model with its parameters of interest and nuisance parameters identified. This class along with `RooWorkspace` are the most important ones for our scripts.

- TFrame: It is similar to TCanvas, and RooFit objects are graphed on a TFrame.
- The most useful tutorials were:
 - Tutorial rf101basics: It shows how to graph RooFit Objects.
 - Tutorial rf102dataimport: It shows how to import a Histogram from a .ROOT file into RooFit's objects.
 - Do the exercises from a CERN's Twiki. There are three, and they are all very useful [37]

RooStats This toolkit is designed to carry out hypothesis tests through Monte Carlo methods, Hybrid approaches that use priors, and approximations of the likelihood ratio to the χ^2 distribution. It uses especially objects from RooFit classes such as RooWorkspace and ModelConfig. The most important class to implement any of the methods of hypothesis testing is ModelConfig, given that in it the Parameters of Interest and Nuisance Parameters are identified, in order to know with respect to which variable the hypothesis test is to be carried out. The most important classes from my experience are the ones shown in the following list. They all take as input a ModelConfig for the Signal plus Background Model, and one for the background model, if there is not one, it creates from the signal plus background model by setting the parameter of interest to zero. They also take the RooDataSet or RooDataHist where the experimental data or pseudodata is stored in order to calculate the test statistic. One can also specify the type of statistic, for instance profiled or not.

- FrequentistCalculator: Calculates the p-values by methods based on Monte Carlo sample created. One specifies the number of toy samples for the background model and the signal plus background model. Calculates p-values as well as CL_s .
- HybridCalculator: Takes also as input a prior distribution for the nuisance parameters, and marginalizes them. It then functions like a FrequentistCalculator.

- **AsymptoticCalculator**: It makes use of Wilk's Theorem, and calculates the p-value assuming that the ratio of log-likelihood ratio distributes according to a χ^2 distribution. The most important references that I relied on were [26] to read a description of each of the methods.
- **HypoTestInverterResult**: It is used to solve for the parameter of interest for a fixed confidence level. It is implemented after having carried out the null hypothesis test with any of the previous methods.
- The most important scripts or tutorials are:
 - **StandardHypoTestDemo.C**: It is a script created by one of the developers of RooFit. It takes mainly as input a .ROOT file that contains a workspace with the data, ModelConfig for signal and background, ModelConfig for background, type of calculator, type of statistic, and number of toys if using. It gives as output the p-value for the background only hypothesis and CL_s to check if the signal point may be rejected or not.
 - **StandardHypoTestInvDemo.C**: It is practically the same as the previous macro, except it has more possibilities for test-statistics, and it carries out the test inversion. For the test inversion one can specify how many points are desired, and the macro will take as many different values for the parameter of interest and check their CL_s to see from which value it can be rejected.
 - The tutorial of the statistical school dictated on 2015 [38]. The exercises I found more useful were exercises 1,2,6, and 7. I strongly recommend exercise 1, without you probably won't be able to complete the rest.

HistFactory It is a separate toolkit of ROOT, but it was created by one of the developers of RooStats and RooFit [39]. This toolkit allows one to create objects named measurements, channels, and so on. These objects are closely related to their physical interpretation, and that was the purpose when the classes were created. The toolkit allows one to create a ModelConfig with an extended likelihood model for the histogram by only importing the signal and total background histograms independently. The toolkit incorporates

the error contained in the histograms, and one can define constraints on the error of their normalization, their luminosity, among other parameters. The greatest strength is that it allows group different channels and models in a seemingly invisible manner to the coder. This way giant models that incorporate measurements from a plethora of channels can be made, greatly increasing the significance. For instance, for the model created in ATLAS to test the discovery of the Higgs Boson contains over 1600 parameters [26]. The most useful references I found for HistFactory and how to implement it were:

- How to implement it: The reference I found for this is CERN Twiki that mainly contains information in xml format, which was the format in which the toolkit was originally used [40]. However, if you search for the following sentence "HistFactory configuration in C++" it will take you directly to the section where they implement in C++. In this section it also explains how to export model to a RooFit ModelConfig to then use with RooStats.
- In order to better understand the rationale and different errors that can be modeled with HistFactory, I recommend the following guide by the developer behind HistFactory [39].

6.1 Script for the Phenomenology Group

The script written for the phenomenology groups has the task of doing a hypothesis test by taking as input a set of histograms for different signal points, a set of histograms of the background common to all signal points, and a histogram of data if there is one. If there is not a histogram of experimental data, then pseudodata must be generated from each histogram of signal and the sum of all the histograms from the background. The output of the script so far, is to calculate the p-value and CL_s for each signal point and store them in a tree in a .ROOT file in order to be plotted later. The sequence of how the script functions and how to use it is the following.

- Clone from my Github the repository named HTHEP.¹¹
- Copy the .root files containing the histograms of the background to the directory named background. Make sure all histograms have the same binning and range.
- Copy the .root files containing the histograms for each simulated signal to the directory named signal. Make sure all histograms have the same binning and range as the background ones and between them.
- If an experimental data histogram is available, copy it to the data directory, if not leave it empty.
- Open the terminal and go to the directory HTHEP. Run the following command "vi hypothesisTest".
- Modify the parameters used in the script. The parameters are:
 - n events pd: If using pseudodata, declare the number of points generated from a gaussian pdf constructed around the content of the bin. If it is too high, then there will be no difference with the histogram of signal plus background.
 - If using data, declare the name of the .root file containing the data.
 - Specify the name that the histograms have in the .root file. If you do not know it, run the command ls() in root to find out the name of the histograms. The shell script assumes all background histograms have the same name as objects. The same applies to the entire set of signal files. Note that the name of the file is not the same as the histograms name within the .root file. They are often different.
 - calculator: specify a number according to the type of calculator you wish to use. The Hybrid calculator is not yet available.
 - ntoys: Declare the number of toys you wish to use if using the frequentist calculator

¹¹<https://github.com/Jdbermeo>

- Exit the shell script
- Run the command "chmod 755 hypothesisTest". Now run "./hypothesisTest"
- First of all the shell script will use the command hadd to sum all the background histograms into a histogram.
- If there is no data file, then it will create a new set of histograms by summing each of the signals with the total background.
- It will run the script generatePseudoData.C for each of the newly created histograms and generate a set of pseudodata, one file for each signal point.
- The script now will run the script histHypoTest.C. This script does the following:
 - The script takes as input the names of the files that contain the total background, the signal point being evaluated, and the pseudodata. It takes as parameter the name that each histogram takes within its respective file.
 - It creates the ModelConfig using HistFactory and taking into account the errors of each histogram.
 - It runs a modified version of Lorenzo Moretta's script standardHypoTestDemo.C. The shell script has also passed as parameters to histHypoTest.C the required parameters to run standardHypoTestDemo.C.
 - The modified version of standardHypoTestDemo.C. saves the results of the hypothesis test to the file "hypoTestDisc.root". It stores the calculated values for the the p-values p_0 , p_1 , and CL_S .
- The script deletes all unnecessary files that were created
- Access the hypoTestDisc.root file and graph the data for all the signal points from the file, one graph for the p-values of discovery, and one graph for CL_S .
- The basic functioning of the script is summarized in Fig. 15.

7 Conclusions

Hopefully the first five sections of the present work will serve as a high yield introduction to statistical inference for someone new to the subject, without demanding too much time or effort on behalf of the reader. When studying this subject, I frequently stumbled upon sources that were either too diluted or too oriented to mathematical rigor, except for source [7]. There does not seem to be many sources oriented to an audience that only wants sufficient understanding of the methods to apply them properly as a tool. That being said, there is nothing wrong with mathematical rigor, but the fact remains that it is not for everyone. Sometimes scientists and engineers only need to know how to appropriately employ the results and methods of statistical inference, and why they should choose this method or the other without going full depth into the subject. Whether this objective is in fact fulfilled in this work will be up to each reader to decide, but if at any point I have failed in this, please refer to [7], it is much less likely to fail you.

As for the remaining objective of writing a script that helps automatize the process of hypothesis testing directly from histograms, and by applying maximum likelihood fits, it was attained. The script written constructs likelihood functions directly from the histograms obtained from phenomenology, whose nuisance parameters are constrained by using auxiliary measurements and applying a maximum likelihood fit to them. The hypothesis test is then carried out using either actual experimental data, or generating pseudodata from the histograms taken as input. The end result is that the process need not be as labor intensive, and regions of interest can now be identified more rapidly, on which more careful models can be implemented if more significance is desired in the said region of interest.

I sincerely hope this work is of use to any reader who wants to carry out a null hypothesis test for exclusion or discovery on high energy physics, and that it helps digest much more easily other documents that dig more deeply and specifically into the subject.

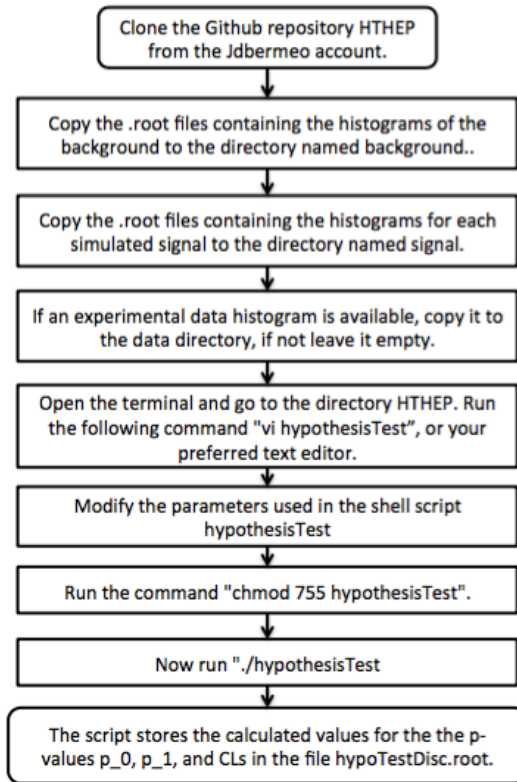


Figure 15: Flow chart summarizing how to implement the shell script to carry out the hypotheses tests on a series of input histograms.

References

- [1] E. T. Jaynes and G. L. Bretthorst, Eds., *Probability theory : The Logic of Science*. Cambridge, UK, New York: Cambridge University Press, 2003, ISBN: 0-521-59271-2.
- [2] M. Starbird, “Meaning from data: Statistics made clear”, The Teaching Company, 2006.
- [3] P. W. Higgs, “Broken symmetries and the masses of gauge bosons”, *Phys. Rev. Lett.*, vol. 13, pp. 508–509, 16 Oct. 1964. DOI: 10.1103/PhysRevLett.13.508. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevLett.13.508>.
- [4] T. C. Collaboration, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”, *ArXiv:1207.7235 [hep-ex]*, Jul. 2012, arXiv: 1207.7235. DOI: 10.1016/j.physletb.2012.08.021. [Online]. Available: <http://arxiv.org/abs/1207.7235> (visited on 02/15/2016).
- [5] T. A. Collaboration, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”, *Physics Letters B*, vol. 716, no. 1, pp. 1–29, Sep. 2012, arXiv: 1207.7214, ISSN: 03702693. DOI: 10.1016/j.physletb.2012.08.020. [Online]. Available: <http://arxiv.org/abs/1207.7214> (visited on 02/15/2016).
- [6] A. Hájek, “Interpretations of probability”, in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Winter 2012, 2012. [Online]. Available: <http://plato.stanford.edu/archives/win2012/entries/probability-interpret/> (visited on 12/27/2015).
- [7] J. H. Friedman *et al.*, *Data analysis techniques for high energy particle physics*. Stanford Linear Accelerator Center, Stanford University, 1974. (visited on 03/27/2016).
- [8] B. C. David S. Moore George P. McCabe, *Introduction to the Practice of Statistics (6th Edition)*, 6th Edition. W. H. Freeman, 2007.

- [9] B. Everitt and A. Skrondal, *The Cambridge dictionary of statistics*, English. Cambridge; New York: Cambridge University Press, 2010, ISBN: 978-0-511-78974-8 978-0-511-78713-3 978-0-511-78827-7. (visited on 01/22/2016).
- [10] C. Donnelly and P. Embrechts, “The devil is in the tails: Actuarial mathematics and the subprime mortgage crisis”, *Astin Bulletin*, vol. 40, no. 01, pp. 1–33, 2010. [Online]. Available: http://journals.cambridge.org/abstract_S0515036100000350 (visited on 03/29/2016).
- [11] P. Billingsley, *Probability and measure*, 3rd ed. J. Wiley and Sons, 1995, ISBN: 0471007102, 9780471007104.
- [12] D. Silvestrov, “Probability Theory IV: Lecture 10 Characteristic Functions”, Stockholm University, [Online]. Available: <http://kurser.math.su.se/course/view.php?id=94> (visited on 04/21/2016).
- [13] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003. (visited on 02/21/2016).
- [14] M. M. Fernández, *Fundamentals on Estimation Theory*. May, 2004. [Online]. Available: <http://lmi.bwh.harvard.edu/papers/pdfs/2004/martin-fernandezCOURSE04b.pdf> (visited on 02/18/2016).
- [15] R. A. Fisher, *Statistical Methods for Research Workers*, ser. Cosmo study guides. Cosmo Publications, 1925.
- [16] J. Neyman and E. S. Pearson, “On the problem of the most efficient tests of statistical hypotheses”, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 231, no. 694-706, pp. 289–337, 1933, ISSN: 0264-3952. DOI: 10.1098/rsta.1933.0009. [Online]. Available: <http://rsta.royalsocietypublishing.org/content/231/694-706/289>.
- [17] J. Aldrich *et al.*, “RA Fisher on Bayes and Bayes’ theorem”, *Bayesian Analysis*, vol. 3, no. 1, pp. 161–170, 2008. [Online]. Available: <http://projecteuclid.org/euclid.ba/1340370565> (visited on 02/19/2016).
- [18] S. Zabell, “R. a. fisher on the history of inverse probability”, *Statistical Science*, vol. 4, no. 3, pp. 247–256, 1989, ISSN: 08834237. [Online]. Available: <http://www.jstor.org/stable/2245634>.

- [19] H. Jeffreys, “An invariant form for the prior probability in estimation problems”, *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 186, no. 1007, pp. 453–461, 1946, ISSN: 0080-4630. DOI: 10.1098/rspa.1946.0056. [Online]. Available: <http://rspa.royalsocietypublishing.org/content/186/1007/453>.
- [20] H. Raiffa and R. Schlaifer, *Applied statistical decision theory*, ser. Studies in managerial economics. Division of Research, Graduate School of Business Administration, Harvard University, 1961, ISBN: 9780875840178.
- [21] W. Bolstad, *Understanding Computational Bayesian Statistics*, ser. Wiley Series in Computational Statistics. Wiley, 2010, ISBN: 9780470046098.
- [22] R. Nuzzo, “Scientific method: Statistical errors”, *Nature*, vol. 506, no. 7487, pp. 150–152, Feb. 2014, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/506150a. [Online]. Available: <http://www.nature.com/doifinder/10.1038/506150a> (visited on 02/22/2016).
- [23] R. E. Kass and A. E. Raftery, “Bayes factors”, *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, Jun. 1995, ISSN: 01621459. DOI: 10.2307/2291091. [Online]. Available: <http://dx.doi.org/10.2307/2291091>.
- [24] N. N. Taleb, “The Meta-Distribution of Standard P-Values”, *ArXiv preprint arXiv:1603.07532*, 2016. [Online]. Available: <http://arxiv.org/abs/1603.07532> (visited on 04/07/2016).
- [25] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, “Asymptotic formulae for likelihood-based tests of new physics”, *The European Physical Journal C*, vol. 71, no. 2, pp. 1–19, 2011. [Online]. Available: <http://link.springer.com/article/10.1140/epjc/s10052-011-1554-0> (visited on 05/09/2016).
- [26] L. Moneta, *Statistical Software Tools RooFit/RooStats. Part 2*, Terascale Statistics School, 2015. [Online]. Available: <https://indico.desy.de/getFile.py/access?contribId=10&resId=0&materialId=slides&confId=11244>.

- [27] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, “MadGraph 5: Going beyond”, en, *Journal of High Energy Physics*, vol. 2011, no. 6, Jun. 2011, ISSN: 1029-8479. DOI: 10.1007/JHEP06(2011)128. [Online]. Available: [http://link.springer.com/10.1007/JHEP06\(2011\)128](http://link.springer.com/10.1007/JHEP06(2011)128).
- [28] B. Andersson, “Pythia 6.4”, 2006.
- [29] S. Ovin, X. Rouby, and V. Lemaitre, “DELPHES, a framework for fast simulation of a generic collider experiment”, *ArXiv preprint arXiv:0903.2225*, 2009. [Online]. Available: <http://arxiv.org/abs/0903.2225>.
- [30] CERN, “Root tutorials: fitting tutorials”, 2015, <https://goo.gl/j0Kf5w>.
- [31] R. Barlow and C. Beeston, “Fitting using finite Monte Carlo samples”, *Computer Physics Communications*, vol. 77, no. 2, pp. 219–228, 1993. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/001046559390005W> (visited on 05/09/2016).
- [32] W. Verkeke, *Advanced classes with Roostats and HistFactory*. [Online]. Available: <http://agenda.nikhef.nl/getFile.py/access?contribId=2&resId=0&materialId=slides&confId=2544> (visited on 05/09/2016).
- [33] CERN, *A ROOT Guide For Beginners*, 2015. [Online]. Available: <https://root.cern.ch/root/htmldoc/guides/primer/ROOTPrimerLetter.pdf> (visited on 05/12/2016).
- [34] *User’s Guide — ROOT a Data analysis Framework*. [Online]. Available: <https://root.cern.ch/guides/users-guide> (visited on 05/12/2016).
- [35] L. Moneta, *Statistical Software Tools RooFit/RooStats. Part 1*, Terascale Statistics School, 2015. [Online]. Available: <https://indico.desy.de/getFile.py/access?contribId=6&resId=0&materialId=slides&confId=11244>.
- [36] W. Verkeke and D. Kirby, *RooFit_users_manual_2.91-33*, CERN. [Online]. Available: https://root.cern.ch/download/doc/RooFit_Users_Manual_2.91-33.pdf (visited on 05/12/2016).
- [37] *RooFit Tutorial*. [Online]. Available: <https://twiki.cern.ch/twiki/bin/view/RooStats/RooFitTutorialMarch2015> (visited on 05/12/2016).

- [38] *Roostats Exercises*. [Online]. Available: <https://twiki.cern.ch/twiki/bin/view/RooStats/RooStatsExercisesMarch2015> (visited on 05/12/2016).
- [39] K. Cranmer, A. Shibata, W. Verkerke, L. Moneta, and G. Lewis, *Histfactory: A tool for creating statistical models for use with RooFit and RooStats*, 2012. [Online]. Available: <http://cds.cern.ch/record/1456844/files/CERN-OPEN-2012-016.pdf?subformat=pdfa> (visited on 05/09/2016).
- [40] W. Breaden, *HistFactory and RooStats*, University of Galsgow: Experimental Particle Physics, Jan. 2016. [Online]. Available: <https://twiki.ppe.gla.ac.uk/bin/view/ATLAS/HiggsAnalysisAtATLASUsingRooStats> (visited on 04/24/2016).