PROBABILITY & STATISTICS

Important Topics:

Probability

- Definitions: Frequentist, Bayesian
- Discussions of Probability
- Probability Density Functions (pdf)
  Cumulative Density Functions (cdf)
- Properties of Probability Distributions
- Popular Probability Distributions
  - Binomial, Poisson, Gaussian, χ^2, ..

STATISTICS

- Parameter Estimation
  - → Max. Likelihood
  - → Bayesian
  - → Method of least squares
  - → Other (Neural Networks, adaptive methods)
- Propagation of errors
- Statistical Tests
  - → Hypothesis Tests
  - → Goodness of fit Tests
- Confidence Intervals & Limits
  - → Frequentist
  - → Bayesian

MONTE CARLO TECHNIQUES

PROBABILITY AND STATISTICS

"Probability theory is nothing but common sense reduced to calculation." Laplace, 1819

"The goal of an experiment is to get a number."

— ?

But, the number isn't very useful, isn't telling you much if you do not know what the "uncertainty" in the number is.

There are uncertainties inherent in all measured data. So, the result of measurement must be reported with a corresponding uncertainty.

There are two approaches to defining and handling probability which lead to two different approaches of Statistics:

1. Conventional or Frequentist Approach
   (Also called "Classical" or "Orthodox")

   Based on the outcomes of a repeatable experiment

2. Modern, Probabilistic or Bayesian Approach
   Based on set theory, plausibility of propositions or hypotheses.

Frequentist Definition:

Consider an experiment in which a series of events are observed. Suppose that the total number of events is N, and the number of events of type A is n, Then,
the probability that any single event will be of type A, is,

$$P(A) = \frac{n}{N} \quad \text{as } N \to \infty.$$

(empirical limit of the frequency!)

Modern Probabilistic Definition: (Bayesian)

Consider a set S, which is the sample space of observed events, and possible subsets A, B, etc.
Then,

1. For every subset A in S, $P(A) \geq \emptyset$
2. For disjoint subsets, i.e., $A \cap B = \emptyset$,
$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$$
3. $P(S) = 1$.

Conditional Probability:

$$P(A \cap B) = P(A|B) \cdot P(B)$$
$$= P(B|A) \cdot P(A)$$

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

Probability of A given B

↖ Bayes' Theorem

In general, if there are $A_1, A_2, \ldots A_N$ exclusive sets, and X is a given event,

$$P(A_i|X) = \frac{P(X|A_i)\,P(A_i)}{P(X)}$$

From the law of total probability,

$$P(X) = \sum_i P(X|A_i)\,P(A_i)$$

$$\therefore P(A_i|X) = \frac{P(X|A_i)\,P(A_i)}{\sum_i P(X|A_i)\,P(A_i)}$$

↳ Bayes' Theorem.

A_i could be various hypotheses, X the observation of data

Probability Density Function (pdf):

Let x be the possible outcome of an observation. x is a random variable since it has many possible outcomes. If x forms a continuous set, then

$$P(x|\theta) = f(x|\theta)\,dx$$ is the probability

that the outcome of the measurement is between x and $x+dx$, and $f(x|\theta)$ is the "probability density function", which may depend on parameters θ.

If x takes discrete values then

$$P(x|\theta) = f(x|\theta)$$

x and θ can both have multiple components. And, given x, we would normally want to estimate θ.

Cumulative Distribution Function (cdf):

$$F(a) = \int_{-\infty}^{a} f(x)\,dx$$

probability that $x \leq a$

If x is discrete, $F(a) = \sum f(x)$

Properties of Probability Distributions:

The pdf's are used as weighting functions to obtain information about random variables.

If $g(x)$ is some function of a random variable x, with pdf $f(x)$, then the "expectation value" of $g(x)$

$$E[g(x)] = \int_{\Omega} g(x) f(x)\, dx$$

Ω represents entire space of x.

E is a linear operator:

So, $E[a\, g(x) + b\, h(x)] = a\, E[g(x)] + b\, E[h(x)]$

Various moments of the probability distributions are used to characterize it.

The nth algebraic moment:

$$\alpha_n \equiv E[x^n] = \int x^n f(x)\, dx$$

The first moment or the expectation value of the random variable x itself is called the "mean" of pdf.

$$\alpha_1 \equiv \mu = \int x f(x)\, dx = \langle x \rangle$$

The nth central moments

$$m_n = E\left[(x-\mu)^n\right] = \int (x-\mu)^n f(x)\,dx$$

$\mu \equiv \alpha_1$

The expectation value of $(x-\mu)^2$ is called the variance

$$\sigma^2 \equiv E\left[(x-\mu)^2\right] = \int (x-\mu)^2 f(x)\,dx$$

$\sigma \equiv$ standard deviation

The mean is the "centre of mass" of the distribution and σ is a measure of the spread.

A distribution may be "skew"ed, i.e., not symmetrical about its mean. A measure of the "skewness" is given by

$$\beta_1 = m_3^2 / m_2^3$$

Kurtosis

$$\beta_2 = m_4 / m_2^2$$

Other useful quantities:

Median $\mu_{1/2}$:

$$P(X < \mu_{1/2}) = P(X > \mu_{1/2}) = \frac{1}{2}$$

Most Probable Value: μ_{max}

$$P(\mu_{max}) \geq P(x \neq \mu_{max})$$

Sample Mean:

$$\bar{x} = \frac{1}{N} \sum x_i$$

Sample Variance:

$$s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2$$

If x and y are two random variables.

Then joint pdf $= f(x, y)$

Marginal p.d.f $\quad f_1(x) = \int f(x,y)\, dy$

$$f_2(y) = \int f(x,y)\, dx$$

As in the case of one variable / dimension, in 2 dimension:

$$E\left[g(x,y)\right] = \iint g(x,y) f(x,y)\,dx\,dy$$

$$\mu_x = \iint x f(x,y)\,dx\,dy = \int x f_1(x)\,dx$$

$$\mu_y = \iint y f(x,y)\,dx\,dy = \int y f_2(y)\,dy$$

$$\sigma_x^2 = E\left[(x-\mu_x)^2\right] \;;\; \sigma_y^2 = E\left[(y-\mu_y)^2\right]$$

The <u>covariance</u> is defined by:

$$\mathrm{Cov}(x,y) = E\left[(x-\mu_x)(y-\mu_y)\right]$$

$$= E\left[xy\right] - \mu_x\mu_y$$

The <u>Correlation Coefficient</u> is given by:

$$\mathrm{Corr}\left[x,y\right] \equiv \rho(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

$$= \frac{\sum (x-\mu_x)(y-\mu_y)}{\sigma_x \cdot \sigma_y}$$

$-1 \le \rho(x,y) \le 1$

↳ also written ρ_{xy}

Some Important Probability Distributions

Binomial Distribution:

Used when there are two mutually exclusive outcomes of a trial/experiment.
The binomial distribution is used to obtain the probability of r "successes" in N trials, when the probability of success (or a certain outcome) in a single trial is denoted by p. (fixed for all trials)

$$P(r, N\,p) = \binom{N}{r} p^r (1-p)^{N-r}$$

$$\binom{N}{r} = \frac{N!}{r!(N-r)!}$$

Poisson Distribution represents an approximation to the binomial distribution for the case where $p << 1$ and $\mu << N$.

$$P(x,\mu) = \frac{\mu^x e^{-\mu}}{x!}$$

$$P(n,\mu) = \frac{e^{-\mu}\mu^n}{n!}$$

Very useful in counting experiments

$P(n,\mu) \equiv$ Probability of observing n events in a time interval with an expected mean number of events $= \mu$

Gaussian / Normal Distribution

- One of the most important distributions (almost ubiquitous); many processes follow Gaussian distribution, experimental errors most often.

  Gaussian assumption simplifies data analysis methods.

For the case of a single variable,

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

with $\mu = E[x] = \int_{-\infty}^{+\infty} x\, p(x)\, dx$; $\sigma^2 = E[(x-\mu)^2]$

$$= \int (x-\mu)^2\, p(x)\, dx$$

In "d" dimensions,

$$p(\underline{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu}) \right\}$$

$\underline{x}$ and $\underline{d}$ are d-dimensional vectors

$$\mu = E[\underline{x}] \; ; \; \Sigma = E\left[(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})^T \right]$$

Σ is a $d \times d$ covariance matrix

If Σ is diagonal, then

$$p(\underline{x}) = \prod_{i=1}^{d} p(x_i)$$

χ^2 Distribution

If $(x_1, \ldots x_n)$ are independent Gaussian random variables, then,

$$\chi^2(n) = \sum_{i=1}^{n} (x_i - \mu_i)^2 / \sigma_i^2$$ follows the χ^2 p.d.f. with n degrees of freedom

For large n, the χ^2 p.d.f. approaches a Gaussian with $\mu = n$ and $\sigma^2 = 2n$

χ^2 p.d.f is often used in evaluating the level of compatibility between observed data and a hypothesis or a fit.

Statistical Inference

We, Experimentalists want to <u>infer</u> parameters of the theory or test a model/theory using our observations/measurements.

P.D.F $f(x|\theta)$ i.e., Prob. of x given θ, the parameters.

But, we need to solve the inverse problem

i.e., obtain $P(\theta|x)$ where θ = mass of the top quark, for example

So, the aim of Statistical inference is to extract useful information from data, as efficiently as possible, or, in other words, establish "facts" about the world.

We, generally, start with one or more hypotheses or theoretical models $M(\theta)$; θ = parameters of the model.

- We want to find out the degree to which the model(s) agree(s) with the observation
  — hypothesis testing
- We want to find/estimate values of the parameters θ
  — Parameter Estimation

The Method of Maximum Likelihood

We have a data set $\{x_i\}$ $i = 1, \ldots N$

Let the p.d.f. be $p(x_i|\theta)$ where θ are the unknown parameters of the theory/model.

The likelihood of making these observations is, simply,

$$\mathcal{L}(\theta) = p(x_1, x_2, \ldots x_n|\theta)$$

[x_i are vectors]

joint probability distribution of all measurements

If we assume the observations x_i to be independent, then

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

To replace "products" by "sums", we can use log of likelihood i.e.,

$$\ln \mathcal{L}(\theta) = \sum_{i=1}^{N} \ln \left[p(x_i|\theta) \right]$$

The logarithms are monotonic functions of the arguments θ and maximized for the same values of $\theta \Rightarrow$ ML estimates $\hat{\theta}$

In practice, we minimize negative log likelihood, rather than maximizing likelihood.

$$E = -\ln \mathcal{L}(\theta) = -\sum_{i=1}^{N} \ln\left[p(x_i|\theta)\right]$$

The maximum likelihood (ML) estimates $\hat{\theta}$ can be found by analytically solving

$$\frac{\partial}{\partial \theta_i} \ln \mathcal{L} = \emptyset$$

For most density functions, the $\hat{\theta}$ are found by iterative numerical procedure.

For multivariate Gaussian density

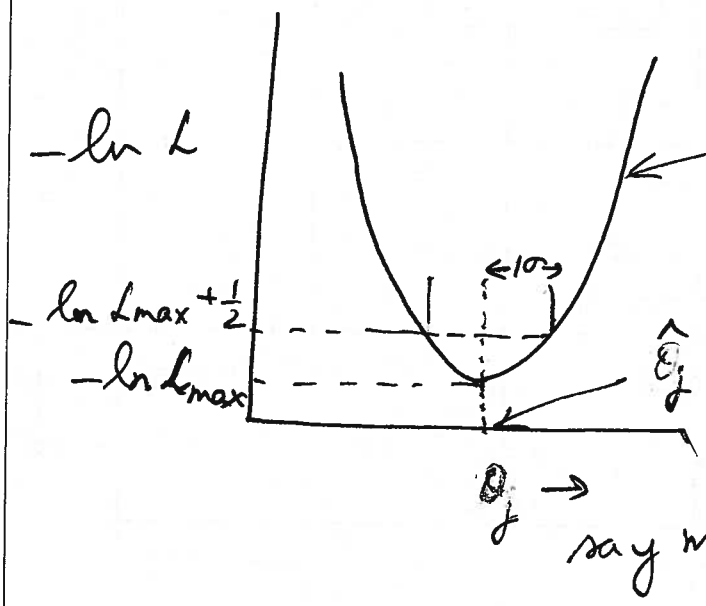
$$p(x_i \mid \sigma_i, \mu_i) = \frac{1}{\left(\sqrt{2\pi\sigma_i^2}\right)} \exp\left[-\frac{1}{2\sigma_i^2}(x_i-\mu_i)^2\right]$$

one gets

$$\hat{\mu}_i = \frac{1}{N}\sum x_i \; ; \; \hat{\sigma}_i^2 = \frac{1}{N}\sum (x_i-\mu_i)^2$$

$$\text{or } \hat{\Sigma} = \frac{1}{N}\sum (x_i-\mu_i)(x_i-\mu_i)^T$$

When $\mathcal{L}$ is Gaussian, $\ln \mathcal{L}$ is parabolic (or hyperparabolic in more than 1D)



$$-\ln \mathcal{L}(\theta_j) = -\ln \mathcal{L}_{max} + \frac{s^2}{2}$$

where $s \equiv$ s-standard deviation

So, for 1σ error

$$-\ln \mathcal{L}(\theta') = -\ln \mathcal{L}_{max} + \frac{1}{2}$$

In $t\bar{t} \rightarrow l l'$ + jets (dileptón + jets channel), one possible likelihood function would be for measured lepton energies (say in the rest-frame of the top quark)

$$\mathcal{L}(D | m_t, j) \propto f(x) f(\bar{x}) \; P(E_l^{cm} | m_t) \; P(E_{l'}^{cm} | m_t)$$

↑ top mass; ↑ particular configuration

$f(x)$, $f(\bar{x})$ are parton distribution functions.

So, Monte Carlo + numerical methods are used.

17a

$$E = -\ln L$$

$$= \sum \left(\frac{1}{2} \ln(2\pi) + \ln \sigma_i + \underbrace{\frac{1}{2\sigma_i^2} (x_i - \hat{x}_i)^2}_{\frac{S^2}{2}} \right)$$

The Method of Least Squares

- This coincides with Maximum Likelihood when the measurements are Gaussian distributed.

∴ It is a special case of ML method.

For simplicity, consider

$$\mathcal{L}(\theta) = p(x|\sigma,\mu) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\left(\frac{x-\mu}{2\sigma}\right)^2}$$

$$-\ln \mathcal{L}(\theta) = \text{Constant} + \frac{\chi^2(\theta)}{2}$$

(Assume σ to be constant, $\theta \equiv \mu$)

$$\therefore \chi^2(\theta) = -2 \ln \mathcal{L}(\theta) + \text{Constant}.$$

So, minimizing χ^2 in this case is same as minimizing -ve log Likelihood.

$$\chi^2(\hat{\theta}) = \chi^2_{min}$$

And, $\chi^2(\theta) = \chi^2(\hat{\theta}) + 1$ has tangent planes at $\pm 1\sigma$ from $\hat{\theta}$.

Bayesian Inference

The frequentist likelihood,

$$\mathcal{L}(\theta) = p(x|\theta) \equiv p(\text{data}|\text{model})$$

But, this is not we are after!
We are interested in the inverse,

$$p(\text{model}|\text{data})\ !$$

From Bayes theorem,

$$p(\text{model}|\text{data}) \propto p(\text{data}|\text{model})\, p(\text{model})$$

or in other words,

$$p(\theta|x) \propto p(x|\theta)\, p(\theta)$$

where $p(\theta)$ is the 'a priori' or 'prior' probability of the model parameters. (which does not involve any knowledge of the data).

Bayesian inference then, implies, updating our probabilistic knowledge of the model parameters, after making observations.

So, the Bayesian posterior probability,

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{\int p(x|\theta)\, p(\theta)}$$

It is also possible to split the model parameters into two sets

those that are of interest to us → θ
those that are not of interest → λ
nuisance parameters.

then

$$p(\theta, \lambda | x) = \frac{p(x|\theta,\lambda)\, p(\theta,\lambda)}{\int_{\theta,\lambda} p(x|\theta,\lambda)\, p(\theta,\lambda)}$$

We can remove the nuisance parameters, by integrating over them

$$P(\theta|x) = \int_{\lambda} P(\theta,\lambda|x)$$

This allows great flexibility and renders the Bayesian method very powerful in dealing with systematic uncertainties and such.

No such general method exists in frequentist approach.

Having obtained the posterior probability for the unknown parameters Θ, given the data, the best Bayesian estimator of a parameter, is the mean of the posterior probability. (This can be proven for a quadratic loss function which we shall discuss in a little bit).

i.e.,

$$\hat{\theta} = \int_{\theta} \theta \, p(\theta|x) \equiv d(x)$$

called the decision function

The accuracy of the estimator is the variance,

$$\sigma^2_{\hat{\theta}} = \mathrm{Var}[d(x)] = \int_{\theta} \theta^2 p(\theta|x) - \hat{\theta}^2$$

i.e., the variance of the posterior probability.

Loss and Risk Functions

Suppose someone asks you on a bright sunny day, what the time is. You don't have a watch, but you immediately notice that "the sun is high in the sky". This is your datum

- The sun is high in the sky
  → So, you could say it is not midnight
  → Or you could say it is noon.

There are two factors to consider in making the inference

— how much useful information you could get
  i.e., how can you minimize loss?

— how much safety margin you would rather have or how much risk can you take?

So, we need a way to quantify the losses and risks incurred in such decisions!

The loss function $L(\theta, d)$

— depends both on the data and the decision function!

The decision function provides mapping from data → model parameters

The most commonly used loss function is,

$$L(\theta, d) = (\theta - d)^2$$

The risk function is constructed by some sort of averaging of the loss function

Frequentist: Inferences can be drawn based on data observed and on data that could have been observed.

$$R_d(\theta) = E\left[L(\theta, d)\right]_x$$

Bayesian: Inferences must be based only on data observed. Averaging is done over all possible hypothesis but only observed data set is used.

$$R_d(x) = E\left[L(\theta, d)\right]_\theta$$

$$= \int_\theta L(\theta, d)\, P(\theta \mid x)$$

Is the best estimator the one that minimizes risk?

Or should it be unbiased or have small variance? (or smallest)

Some useful properties of the Estimator

- Consistency
  - the most desirable property
  - the estimate $\hat{\theta}$ converges to the true value θ as the amount of data increases.

- Bias
  - $b = E[\hat{\theta}] - \theta$

    $E[\hat{\theta}]$ is the expectation value taken over a hypothetical set of experiments.
  - if $b = 0$, the estimator is said to be unbiased
  - if $\hat{b}$ is estimated, then we can correct for the bias $\hat{\theta}' = \hat{\theta} - \hat{b}$

- Efficiency
  - defined as $\sigma_{\hat{\theta}}^2 / \sigma_{min}^2$

    where σ_{min} is the minimum possible variance

$$\sigma_{min}^2 = \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[\left(\frac{\partial}{\partial \theta} \sum \ln f(x|\theta)\right)^2\right]$$

A useful quantity to minimize is the mean-squared error $MSE = \sigma^2 + b^2 \equiv E\left[(\hat{\theta} - \theta)^2\right]$

- Robustness
  - the estimate has to be insensitive to assumptions regarding noise etc..

Statistical Tests

Hypothesis Testing:

Let us consider Signal hypothesis $\equiv$ 'S'

Background hypothesis $\equiv$ 'B'

One tries to minimize background efficiency for a given signal efficiency.

Neyman-Pearson Lemma

$$\Rightarrow \quad \lambda(x) = \frac{f(x/S)}{f(x/B)} > \text{constant}.$$

This could also be simply the Likelihood ratios for the two hypotheses.

The Bayes discriminant is a similar proposition but uses posterior probabilities

$$R(x) = \frac{p(S/x)}{p(B/x)} > \text{Constant}.$$

The most modern test statistics are based on Neural Networks and other multivariate methods.

Goodness of fit Tests

Construct a statistic whose value reflects the level of agreement between the data and the fit hypothesis.

If you define some

goodness-of-fit statistic, t and let

$g(t|H)$ be the p.d.f given a hypothesis H

then p-value is defined as

$$p = \int_{t_{obs}}^{\infty} g(t|H)\, dt$$

[probability to find t in the region of equal or lesser compatibility with H, than observed.

where t_{obs} is the observed value of the statistic.

"t" is defined such that larger values correspond to poor agreement with the hypothesis

$p \to 0 \Rightarrow$ poor fit ; $p \to 1$ too good fit ?

An example of such a statistic is

$P(\chi^2)$.

Confidence Intervals & Limits

The goal of an experiment may be to measure a parameter θ, and quote the statistical precision of the measurement.
In the simplest case, $\hat{\theta} \pm \sigma_{\hat{\theta}}$

If the p.d.f. is not a Gaussian, or due to other constraints, one usually quotes an interval for statistical uncertainty.
– the interval is constructed to cover the true value of the parameter with a certain probability

Bayesian Approach:

Determine an interval that contains a given fraction $(1-\alpha)$ of the probability:

$$1-\alpha = \int_{\theta_{lo}}^{\theta_{up}} p(\theta|x)\, d\theta$$

$\Rightarrow$ there is $(1-\alpha)\times 100\%$ probability that the true value is contained in the interval θ_{lo} to θ_{up}.

$\Leftarrow$ Very different from frequentist definition.

The Frequentist's Confidence Interval would say that

if the experiment is repeated a large number of times and the same analysis is carried out, then, $(100)(1-\alpha)\%$ of the experiments will give the answer within the interval

i.e., $100(1-\alpha)\%$ of the experimental results after a large number of similar experiment are conducted, will lie within the interval.

Frequentist methods therefore use ensembles to determine the intervals.

In Bayesian approach, ensembles are used to study the validity of the method rather than to construct intervals.

Confidence levels:

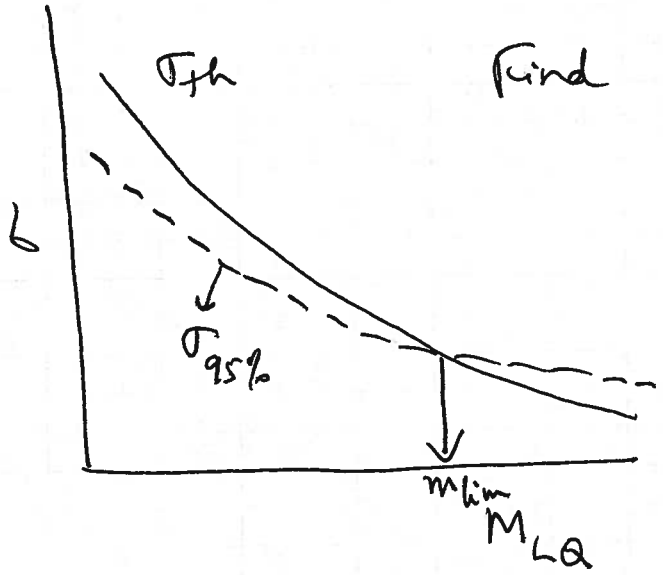
Upper limits, say, on cross section at 95% C.L.

$$\int_0^{\sigma_{95}} p(\sigma / N_s^{obs}) = 0.95$$

One can find 95% C.L. upper limit on the number of signal events and then estimate σ_{95}

$$1-\alpha = \int_{-\infty}^{n_{sup}} p(n_s|n)\,ds = \frac{\int_{-\infty}^{n_{sup}} L(n|n_s)\,p(s)\,ds}{\int_{-\infty}^{+\infty} L(n|n_s)\,p(s)\,ds}$$

Mass lower limits:



Find $\sigma_{95\%}$

$M_{LQ} > m_{lim}$ @ 95% C.L.