Fundamentals of Statistical Analysis in Physics
Lectures 1, 2: Overview of Statistical Analysis

African School of Physics 2022

Harrison B. Prosper

Kirby W. Kemper Endowed Professor of Physics
Department of Physics, Florida State University

28 November, 2022

Outline

Outline

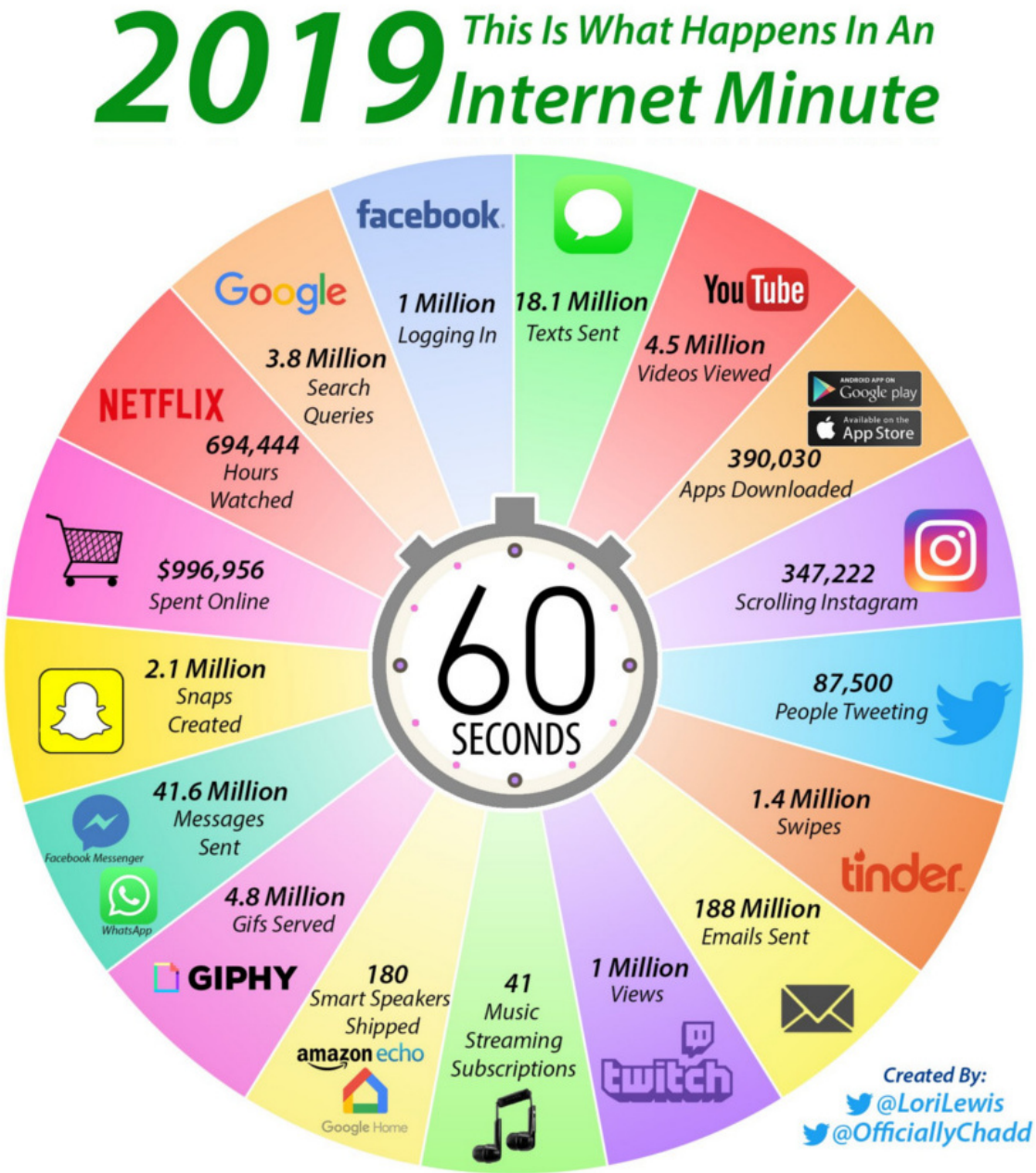
According to wikipedia:

> *Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data.*

Data, Data, Everywhere

Project	Data Size	Period
SDSS	100 TB	2000 – 2015
LSST	100,000 TB	2023 – 2033
LHC	15,000,000 TB	2010 – 2038

And even these enormous data sets are tiny compared with the global generation of data...


2019 This Is What Happens In An Internet Minute
facebook
1 Million Logging In
18.1 Million Texts Sent
YouTube
4.5 Million Videos Viewed
Google play
App Store
390,030 Apps Downloaded
347,222 Scrolling Instagram
87,500 People Tweeting
1.4 Million Swipes
tinder
188 Million Emails Sent
1 Million Views
twitch
41 Music Streaming Subscriptions
180 Smart Speakers Shipped
amazon echo
Google Home
4.8 Million Gifs Served
GIPHY
41.6 Million Messages Sent
Facebook Messenger
WhatsApp
2.1 Million Snaps Created
$996,956 Spent Online
694,444 Hours Watched
NETFLIX
3.8 Million Search Queries
Google
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

Topics to be covered

- Descriptive statistics
- Probability models and likelihoods (Binomial, Poisson, Gaussian, χ^2)
- Frequentist inference (profile likelihood, confidence intervals, statistical significance,...)
- Bayesian inference

To illustrate the application of the concepts to be discussed, we shall work through statistical analyses of the following data sets:

- DØ 1995 top quark discovery data
- Type 1a supernova data
- ATLAS Open Data ($H \to 4\ell$)

Fundamental Assumption

Data can be regarded as the outcome of a stochastic data generation mechanism.

Example

At Fermilab (USA) during the 1990s, the data generation mechanism was the Tevatron accelerator together with the data processing systems of the DØ and CDF Collaborations. The Tevatron produced proton-antiproton collisions at 1.8 TeV.

The DØ top quark discovery data consists of $N = 17$ observed proton-antiproton collision events of which $B = 3.8 \pm 0.6$ events were "estimated" to be non-$t\bar{t}$ events, that is, background events, thereby indicating "decisive evidence" of a top quark signal.

The Big Picture

- Collect raw data from a data generation mechanism.
- Process the raw data into meaningful observed data, D.
- Construct a statistical or probability model of all potentially observable data, X.
- Enter the observed data into the probability model thereby converting the latter into a likelihood function.
- Use the likelihood function to make inferences about the model.

A probability model usually depends on two kinds of parameter: parameters of interest and nuisance parameters.

Example

The cross section for $pp \to H \to ZZ \to 4\ell$ is a parameter of interest, while those for the background processes are considered nuisance parameters.

Descriptive Statistics

Population

Suppose that the data generation mechanisms (e.g., Nature) can supply, in principle, an infinite number of data instances, i.e., outcomes.

The infinite set of outcomes defines a population. Clearly, the latter is an abstraction.

Sample

An actual set of outcomes is called a sample.

Example: The set of $n = 50,785$ di-photon masses $m_{\gamma\gamma}$ that led to the CMS discovery of $H \to \gamma\gamma$ in 2012 concretely exists (e.g., on my laptop), while the associated population exists in the same sense as the set of real numbers.

Statistic

Any function of a sample $x_1, x_2, \cdots, x_n$ is a statistic.

Here are a few.

$$x_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r, \qquad \text{sample moments}$$

$$\bar{x} = x_1, \qquad \text{sample average}$$

$$\text{Var} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \text{sample variance}$$

Ensemble Average

The ensemble average of a statistic x, which we shall denote by $\mathbb{E}[x]$, is an average over infinitely many instances of the statistic.

Error, Bias, Variance

Here are a few standard quantities that characterize populations.

$$\mu \qquad \text{mean}$$

$$\epsilon = x - \mu \qquad \text{error}$$

$$b = \mathbb{E}[x] - \mu, \qquad \text{bias}$$

$$\sigma^2 = \mathbb{E}[(x - \mathbb{E}[x])^2] \qquad \text{variance}$$

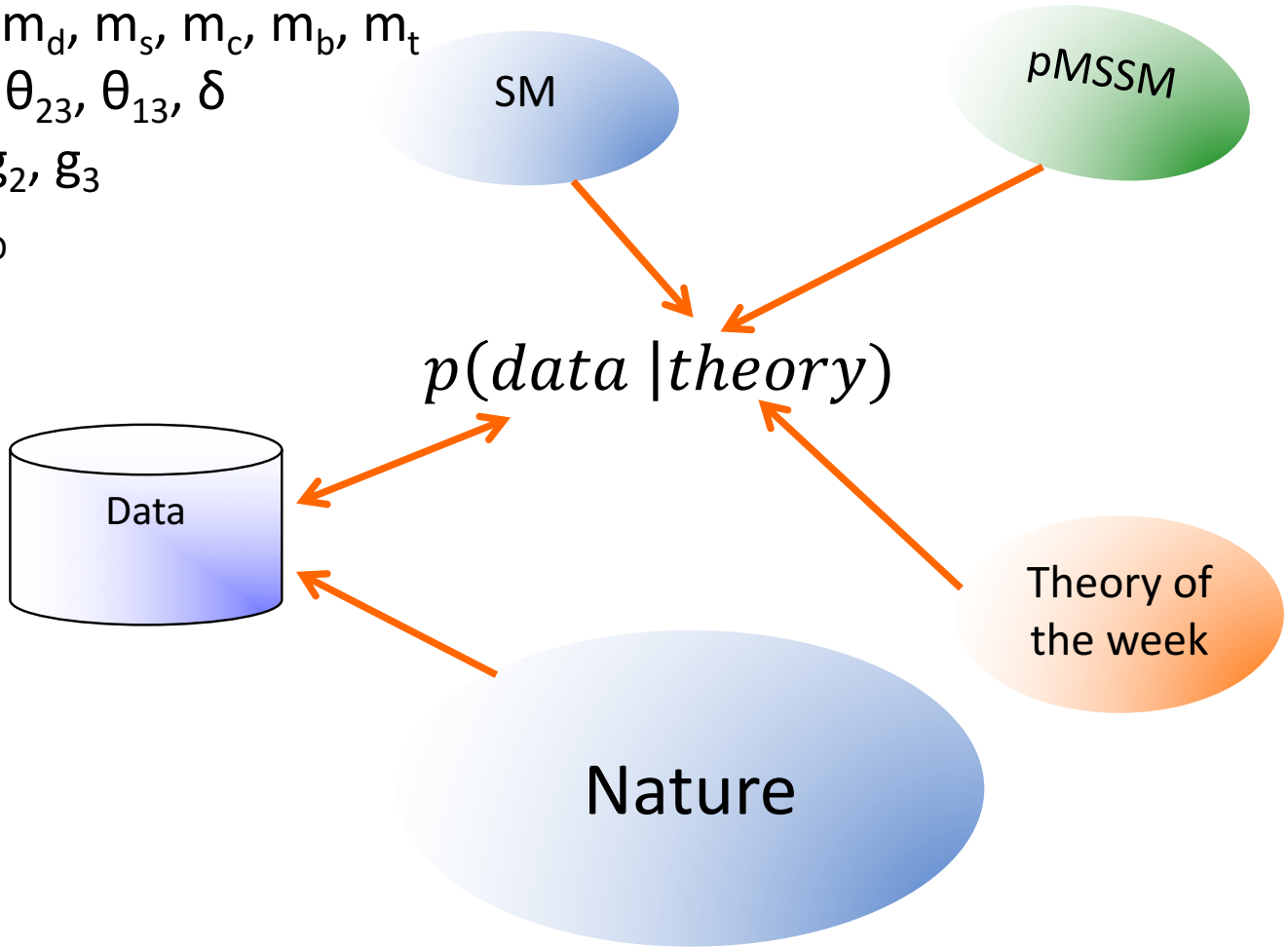
$$\text{MSE} = \mathbb{E}[(x - \mu)^2] = \sigma^2 + b^2 \qquad \text{mean square error}$$

While the above descriptors are useful, most physicists are interested in extracting information about the phenomenon that yielded the data; that is, we are interested in making inferences.

Outline

Inference

m_e, m_μ, m_τ
m_u, m_d, m_s, m_c, m_b, m_t
θ_{12}, θ_{23}, θ_{13}, δ
g_1, g_2, g_3
θ_{QCD}
μ, λ



Inference

The likelihood function is the fundamental mathematical quantity in most statistical analyses in particle physics, astrophysics, and cosmology.

Given the likelihood function we can answer several questions:

- How does one estimate (i.e., measure) a parameter?
- How does one quantify the uncertainty in the estimate?
- How does one test an hypothesis?
- How does one quantify the significance of a result?

How these questions are answered, however, depends on how one chooses to interpret probability. The two most common interpretations are:

$$\begin{aligned} \text{relative frequency} &\rightarrow \text{frequentist approach} \\ \text{degree of belief} &\rightarrow \text{Bayesian approach.} \end{aligned}$$

Frequentist Approach

This approach uses the probability model $p(x|\mu,\nu)$ and, therefore, the likelihood $p(D|\mu,\nu)$ for making inferences. D are the observations and μ and ν denote the parameters of interest and nuisance parameters, respectively.

The guiding principle of this approach is:

The Frequentist Principle

In an ensemble of statements, not necessarily about the same quantity, the fraction of true statements is guaranteed to be greater than or equal to a given value.

Frequentist Approach

Consider the statements below from the Particle Data Group (PDG),

$$m_t \in (172.46, 173.06]\,\text{GeV}$$

$$\text{BR}(t \to e\nu_e b) \in (10.80, 11.4)\%,$$

$$\tau_\tau \in (289.8, 290.8) \times 10^{-15}\,\text{s},$$

and the tens of thousands of other statements compiled by the PDG. Each of these statements is either True or False.

Although we do not know which are true and which are false, we can assert that about 68% of the statements are true if the frequentist principle is satisfied for the ensemble of statements compiled by the PDG.

Frequentist Approach

Example (1.1 The Ebola Outbreak)

During the 2014 – 2016 Ebola outbreak in West Africa, $N = 4$ cases of Ebola were reported in the United States.

Since Ebola is extremely rare in the United States, we can model the probability to observe x Ebola cases in the US by a Poisson distribution,

$$p(x|\mu) = e^{-\mu}\mu^x \,/\, x!,$$

with mean μ. The likelihood function is therefore,

$$p(4|\mu) = e^{-\mu}\mu^4 \,/\, 4!$$

We wish to say something about the mean count μ. An obvious estimate is $\hat{\mu} = 4$; but how reliable is this estimate?

Frequentist Approach

Example (1.1 The Ebola Outbreak)

For a Poisson distribution with mean μ the standard deviation is $\sigma = \sqrt{\mu}$.

Therefore, one way to quantify the reliability of the estimate is to assert that

$$\mu \in [N - \sqrt{N}, N + \sqrt{N}],$$

with $N = 4$. Again, this statement is either True or False but we do not know which it is.

But on its own, the above statement it is not enough; we need to quantify our confidence in it.

Jerzy Neyman (1894 – 1981)

In 1937, the Polish mathematician and statistician introduced the concept of a confidence interval and the associated relative frequency, called a confidence level (CL).

For this example, the confidence level is the minimum fraction of true statements of the form $\mu \in [N - \sqrt{N}, N + \sqrt{N}]$.

The defining characteristic of (exact) confidence intervals, and more generally, confidence sets, is that over an infinite ensemble of intervals or sets, the fraction of true statements is guaranteed not to fall below the desired confidence level.

Moreover, for any given parameter, Neyman required this to hold whatever the true values of the other parameters of the probability model.

Example (1.1 The Ebola Outbreak)

In his 1937 paper, Neyman showed how to construct intervals $[\mu_L, \mu_U]$ with a given confidence level. For the Poisson problem, the method entails solving the equations

$$D_R(N, \mu) \equiv \sum_{k=N}^{\infty} \text{Poisson}(k, \mu) = (1 - CL)/2,$$

$$D_L(N, \mu) \equiv \sum_{k=0}^{N} \text{Poisson}(k, \mu) = (1 - CL)/2,$$

for the lower and upper bounds, μ_L and μ_U, respectively[a] By convention, we choose $CL = 0.683$.

[a] Define $P(\alpha, x) = \int_0^x t^{\alpha-1} e^{-t}\, dt/\Gamma(\alpha)$, the normalized lower incomplete gamma function. Then $D_R(N, \mu) = P(N, \mu)$ and $D_L(N, \mu) = 1 - P(N + 1, \mu)$. $P(\alpha, x)$ can be computed with the ROOT function `TMath::Gamma(`α`, x)`.

Bayesian Approach

In the Bayesian approach the interpretation of the statements,

$$m_t \in (172.46, 173.06]\,\text{GeV}$$

$$\text{BR}(t \to e\nu_e b) \in (10.80, 11.4)\%,$$

$$\tau_\tau \in (289.8, 290.8) \times 10^{-15}\,\text{s},$$

is that the degree of belief in the truth of these specific statements is about 0.68.

Bayesian Approach

The Bayesian approach also uses the likelihood function $p(D|\mu, \nu)$.

But, in addition, a probability $\pi(\mu, \nu) d\mu\, d\nu$ is assigned to the parameters. This is meaningful because probability is interpreted not as a relative frequency but as a degree of belief.

Contrast the two statements:

1. The chance of creating a Higgs boson when two protons collide at 13 TeV is $\approx 10^{-10}$.
2. The chance that in 2023 reason will prevail among all *Homo sapiens* is $\approx 10^{-10}$.

The first probability can reasonably be interpreted as a relative frequency, while the only operationally meaningful interpretation of the second is a degree of belief.

Outline

- The most important task in a statistical analysis is constructing an accurate probability model. (This is also known as a statistical model.)
- When data are entered into the model we arrive at the likelihood function.
- The likelihood function (together with a prior density in the Bayesian approach) is the usual basis for inferences in particle physics and cosmology.

Outline

All Jupyter notebooks will be posted to:

http://www.hep.fsu.edu/~harry/ASP2022

Today, if time permits, we shall work through topdiscovery.ipynb

Exercise 1

Generate $T = 10,000$ experiments each with a different mean μ sampled from an exponential density with mean $b = 5$ and a count N sampled from a Poisson distribution. Since this is a simulation, we can determine which statements of the form $\mu \in [N - \sqrt{N}, N + \sqrt{N}]$ are True or False. The fraction of true statements is called the coverage probability. The frequentist principle requires that this probability, interpreted as a relative frequency, be no smaller than the desired confidence level CL, typically, CL=0.683.

Can you find a modification of the statements, which depends only on N, such that the coverage probability is closer to 0.68s?