Introduction to Statistics: Probability

Sarah Mancina
IceCube Bootcamp 2020

June 15, 2020

1 Introduction

In order to study statistics, it helps to have a solid understanding of probability. When taking data we have our population and our sample. For example the population could be all people in the United States, and our sample are the collection of people we took data on. Another example could be our population is all neutrinos which are governed by some underlying physical properties and our sample is a selection of events seen by the IceCube detector. Probability and statistics are the tools we use to understand the relationship between our population and our sample. Probability maps from the population to the sample where as statistics maps from the sample to the population. Say I have a bucket of marbles of different colors, if I know the distribution of all of the colors in the bucket I can use probability to guess what would be in my hand if I grabbed a handful. If I don't know the distribution of all of the colors in the bucket, but I have a handful of marbles I can use statistics to guess what the distribution of colors is in the bucket.

2 Data

2.1 Types of Data

To start with, we'll talk about the different types of data. This is pretty simple, but it's helpful to think about the basics of data to understand probability and statitics.

First we can break it down into quantatative and qualitative.

- Quantatative: Numerical data. E.g. How tall are you? How many pairs of socks do you own?
- Qualitative: Categorical data. E.g. What is your favorite color? Do you like pizza?

Qualitative data can be broken down into a couple of sub-categories:

- Nominal: The data are linked to discrete labels which cannot be ordered. E.g. What's your favorite pizza topping? [pepperoni, spinach, olives, pinnapple]
- Binary: There are only two answers to the question. E.g. Did the neutrino leave light in the DOM? [no (0), yes(1)]
- Oridnal: The labels can be ordered, but the difference in values is not meaningful beyond order. E.g. What flavor of neutrino? [(anti) electron (-)12), muon ((-)14), tau ((-)16)]

Quantatitve data can be broken down into two sub-categories:

- Discrete: Data that has distinct values and can be counted. E.g. How many neutrino events did we see? [1, 2, 3, 4, ...]
- Continuous: Data that is measured on a continuum. E.g. What was the energy of the neutrino? [10.5 TeV, 47.9 TeV, 188.3 TeV, ...]

2.2 Displaying Data

How we display our data is a very important part in understanding and communicating our results. There are several ways to display data. Some tools are better suited for qualitative or quantatative data. Common ways to display qualitative data are tables, bar charts, pie charts, and more. We will focus more on quntatative data for the rest of this exercise.

2.2.1 Histograms

In physics we commonly use histograms to display our data. A histogram is a type of plot where the data is separated into discrete bins and for each bin a value proportional to the number of counts in the bin is assigned. The value assigned to the bin is usually one of the three following values:

- Counts per Bin: The total number of counts in the bin.
- Relative Frequency: The normalized number of counts in each bin calculated through the following formula:

  $$f_i = \frac{n_i}{\sum_i^{bins} n_i} \tag{2.1}$$

  where f_i is the frequency for bin i and n_i is the number of counts in bin i.
- Density: The frequency divided by the width of the bin as shown in the following formula:

  $$d_i = \frac{n_i}{w_i \sum_i^{bins} n_i} \tag{2.2}$$

  where d_i is the density for bin i, n_i is the number of counts in bin i, and w_i is the width of bin i.

The density can be quiet useful especially when using uneven histogram bins. When using the density the area of the bin, instead of the height, is proportional to the number of counts.

To create and plot histograms in python we can use a couple of different tools from numpy and matplotlib. To read more about these objects you can go to their documentation sites by google searching: numpy.histogram, numpy.histogram2d, matplotlib.pyplot.hist, and matplotlib.pyplot.hist2d.

2.3 Descriptive Statistics

When we have a population or sample, there are ways to combine the information we have together to describe the distribution of the data. The most common are below:

- Aritmetic Mean: $\bar{x} = \sum_{i=0}^{N} \frac{1}{n} x_i$
- Median: The value lying at the midpoint of a distribution of observed values where there is equal probability of lying above or below.

- Standard Deviation: A measure of the spread $\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(x_i - \bar{x})^2}$
- Correlation Coefficient: Measure the linear relationship between to variables

$$\rho = \frac{\sum_{i=0}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

$\rho = 0$ means there is no linear correlation, $\rho = 1$ means a positive linear correlation, and $\rho = -1$ means there is a negative linear correlation.

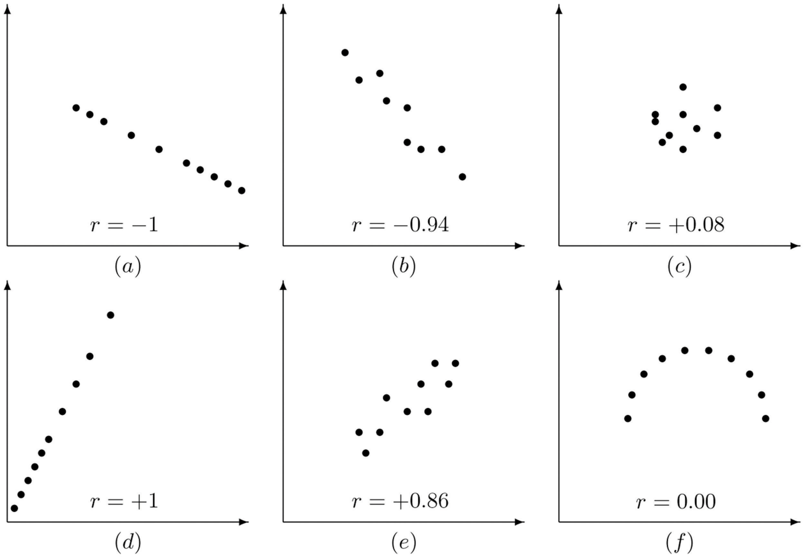

Fig. 2.1. Examples of the pearson correlation cooefficient.

3 Basic Probability

Now that we've covered some of the basics of data, we can think about the basics of probability. Going back to our bucket idea, we have a bucket full of 30 red marbles, 60 blue marbles, and 20 yellow marbles. If we grab two marbles from the bucket what is the probability that we grab 2 red marbles, P(2 red)? What are the properties of P(2 red) and how does it relate to other possible outcomes?

An experiment is any activity or process where the outcome is subject to uncertainty. The sample space is the set of all possible outcomes of an experiment. An event is a set of outcomes in the sample space. So in the marble example, our experiement is grabbing two marbles from the bucket. The sample space would be drawing {2 red, 1 red & 1 blue, 1 red & 1 yellow, 2 blue, 1 blue & 1 yellow, 2 yellow}. And possible events could be drawing 2 red: {2 red}, or drawing at least one red: {2 red, 1 red & 1 blue, 1 red & 1 yellow}. These events have some probability that follow the Kolmogorv Axioms.

First, here are some notations for sets which we use to describe events and our sample space:

- The complement of event A, $\bar{A}$, is the set of all a in the sample set, S, where a is not in A. E.g. The complement of {2 Red} is {1 red & 1 blue, 1 red & 1 yellow, 2 blue, 1 blue & 1 yellow, 2 yellow}.

- The union of events A and B, $A \cup B$, is all a in the sample set, S, if a is in A or B. E.g. the union of drawing at least one blue, {2 blue, 1 blue & 1 red, 1 blue & 1 yellow}, or at least one red, {2 red, 1 red & 1 blue, 1 red & 1 yellow}, is {2 red, 1 red & 1 blue, 1 red & 1 yellow, 2 blue, 1 blue & 1 yellow}.
- The intersection of events A and B, $A \cap B$, is all a in the sample set, S, if a is in A and B. E.g. the intersecton of drawing at least one blue, {2 blue, 1 blue & 1 red, 1 blue & 1 yellow}, or at least one red, {2 red, 1 red & 1 blue, 1 red & 1 yellow}, is {1 red & 1 blue}.

The foundation of probability are the Kolmogorov Axioms, which are as follows.

- For any event A in the sample space S, $P(A) \geq 0$.
- $P(A) + P(\bar{A}) = 1$, where $\bar{A}$ is the complement of A in S.
- For all possible mutually exclusive events, A_1, A_2, ... in S,

$$P(A_1 \cup A_2 \cup ...) = \sum_{i=1}^{\mathrm{inf}} P(A_i)$$

.

Lets think about another experiement where we measure the energy of an incoming neutrino. We could think of the probability the event that we measure a neutrino in the 10-20 TeV range. This probability, $P(E \in [10, 20]\ \mathrm{TeV})$, depends on some underlying physical distribution (our population).

3.1 Conditional Probability

Conditional probabilty is the probability that an event may occur given another event occuring. E.g. probability that IceCube sees a neutrino from the direction of a blazar given the blazar is flaring. The probability of event A given event B is the probability of the union of events A and B divided by the probability of B

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3.1}$$

From this comes another axiom $P(A \cap B) = P(A|B)P(B)$. Conditional probabilities can be reversed using Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3.2}$$

Conditional probabilities are used in statistical methods. Often we are looking for the probability that the underlying physical parameters are some value the given the data we observed. E.g. we want to find the probability that the neutrino powerlaw spectra has a $\gamma = 2$ ($\phi(E) \propto \left(\frac{E}{100\ \mathrm{TeV}}\right)^{-\gamma}$) given the reconstructed neutrino energies we observed in our data, $P(\gamma|E_i)$. This value is unknown to us, so we can use statistical methods, such as likelihood estimation or baeysian statistics to estimate it from $P(E_i|\gamma)$.

3.2 Frequentist Probability

For this lecture, we will only cover the Frequentist definitions of probability and statistics. However, there is also Bayesian statistics which is a useful tool.

The frequentist definition of probaility is:

$$P(X) = \lim_{N \to \inf} \frac{n}{N}, \tag{3.3}$$

where n is the number of times outcome X occured over N trials. The law of large numbers says that as N approaches infinity in an experiement, the measurement of $P(X)$ approaches it's true value.

4 Probability Distribution Functions

Sometimes variables have a well defined probability distribution. If the variable is descrete it's distribution is called a Probability Mass Functions (PMF), otherwise if the variable is continuous it's distribution is called a Probability Density Function (PDF). For a PMF,

$$\sum_{i=0}^{max} PMF(x_i) = 1 \tag{4.1}$$

and for a PDF,

$$\int_{min}^{max} PDF(x)dx = 1. \tag{4.2}$$

The cummulative distribution is for a given value, r, the probability of all the outcomes $\leq r$.For a PMF,

$$CDF(r) = \sum_{i=0}^{r} PMF(x_i) \tag{4.3}$$

and for a PDF,

$$CDF(r) = \int_{min}^{r} PDF(x)dx. \tag{4.4}$$

If we wanted to find the expected value of and function of x, $g(x)$, we can find the expected value using the following formulations. For a PMF,

$$g(\bar{x}) = \sum_{i=0}^{max} g(x)PMF(x_i) \tag{4.5}$$

and for a PDF,

$$g(\bar{x}) = \int_{min}^{max} g(x)PDF(x)dx. \tag{4.6}$$

Often for distributions we describe them using their moments. The first moment is their mean, $\bar{x}$, which can be found using the equations above and setting $g(x) = x$. Another moment is the variance, $v = g(x) = (x - \bar{x})^2$, which can be used to find the standard deviation, $\bar{v} = \sigma^2$

There are many different named PMFs and PDFs. Here we'll cover the binomial, Poisson, and normal distributions. Wikipedia has good coverage of many other distributions.

4.1 Binomial Distribution

A binomial distribution is a PMF which describes the probability of the number of successes in a squence of independent experiments with a binary outcome (Bernoulli trial). An example would be rolling a dice and calling rolling 1 a success and rolling anything else a failure. Let's try to use this example to build the binomal distribution.

If you a roll a dice 5 times, what would be the probability that you would roll a 1 twice? One way to roll a 1 twice is to roll the 1 in the first two rolls:

$$P(1 \text{ \& } 1 \text{ \& not } 1 \text{ \& not } 1 \text{ \& not } 1) = 1/6 \times 1/6 \times 5/6 \times 5/6 \times 5/6.$$

But this isn't the only way to roll 1 twice. We can see that no matter what order we roll 1 twice in five rolls, the probability for each combination is the same. We must include the combinatorics to calculate the total probability

$$P(\text{Roll 1 twice}) = \binom{5}{2}(1/6)^2(5/6)^3,$$

where $\binom{5}{2} = \frac{5!}{2!3!}$.

The general form of the Binomial distribution is as follows:

$$P(k) = \binom{n}{k}p^k(1-p)^{(n-k)}, \tag{4.7}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Where k is the number of successes, n is the number of trials, and p is the probability of a single success.

We can find the moments of a distribution by solving for the expectation values. For the binomial distribution, the mean is np and the standard deviation is $\sqrt{np(1-p)}$.

4.2 Poisson Distribution

The poisson distribution is a PMF which gives the probability of a discrete number of events occuring. The form is as follows:

$$P(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \tag{4.8}$$

where λ is the average number of events that occur. The mean of the distribution is therefore λ and the standard deviation is $\sqrt{\lambda}$.

E.g.

Let's say on average we see 10 high energy neutrinos per year, what is the probability that in a give year we saw 15 nuetrinos? what is the probability that we see 1 neutrino in a month?

For this example, $\lambda = 10$ nuetrinos per year. So for the first question

$$P(15) = \frac{e^{(-10)} \times 10^{15}}{15!} = 0.03472.$$

For the second question, we can adjust our λ to be the expected value per month, $\lambda = 10/12 = 0.8333$ (assuming all months are approximately equal). So

$$P(1) = \frac{e^{(-0.8333)} \times (0.8333)^1}{1!} = 0.3622.$$

4.3 Normal Distribution

The normal distribution is one of the most common PDFs. This is because the central limit theorem explains that when independed random variables are added together, their normalized sum tends towards a normal distribution. It is often a good assumption to assume measurements are normally distributed. The normal distribution is also known as the Gaussian distribution and a bell curve and has the following form:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}. \tag{4.9}$$

where μ is the mean of the distribution and σ is the standard deviation. The normal distribution has a characteristic bell curve shape which is peaked at μ and it's width is affected by σ.