1 Introduction

We typically conduct an investigation to learn about specific characteristics of a collection of objects, known as a **population** of interest. In one study, the population might consist of all the people in the United States or all products made by a factory in a given year. Constraints on time, money, and other scarce resources usually make the study of the entire population impractical or infeasible. Instead, a subset of the population—a **sample**—is selected. Thus we might obtain a sample of bearings from a particular production run as a basis for investigating whether bearings are conforming to manufacturing specifications. We use statistical analysis to make estimates about the characteristics of the entire population, from a measurement of the characteristic in a sample. If we know the characteristics of the population, however, we can use probability to predict the characteristics of a sample. If I have a bucket of marbles of different colours, and I know how many marbles of each colour are in the bucket, I can use probability to guess what would be in my hand if I grabbed a handful.If I knew nothing about the marbles in the bucket, but I have a handful of marbles, I could use statistics to guess what kinds of marbles are in the bucket.

2 Data

2.1 Types of Data

At the highest level, we classify data into

- **Quantitative**: dealing with numbers or values that can be objectively measured, like height, length, temperature etc.
- **Qualitative**: dealing with features that can only be subjectively classified, like smells, colours etc.

When dealing with Quantitative or numeric data, one may further distinguish between **discrete** and **continuous** data. Discrete data is used to measure countable objects, which cannot be subdivided, such as the number of passengers on a plane, or the number of neutrino events in a detector. Continuous data, on the other hand, is used to measure quantities that are divisible onto a continuum of increasing precision, such as the height of a person, which can be measured in metres, or centimetres, or the energy deposited by a neutrino in a detector.

Qualitative data, on the other hand, can be further classified into

- **Binomial**:classification into one of two mutually exclusive categories, a question where the answer is either a yes or a no. E.g. Did the neutrino deposit energy in the detector?
- **Nominal**:data linked to distinct labels, which cannot be ordered. E.g. Favourite flavour of ice-cream
- **Ordinal**:data which is classified into categories that can be ordered, such as clothes sizes.

2.2 Histograms

Consider data consisting of observations of a variable x. We may visualize the data using a histogram, a type of plot where the data is placed into discrete bins and for each bin a value proportional to the number of counts in the bin is assigned. The value we assign may be:

- Counts per bin: the number of times values lying in a particular bin occur in the data set. We use n_i to denote the number of values in bin i
- Relative Frequency:the fraction or proportion of times the value occurs
$$RF = \frac{n_i}{\sum_i n_i}$$
- Density: the Relative Frequency divided by the bin width w_i. This can be useful when the bins have different widths, and the number of counts is therefore proportional to the area of the bin
$$\rho_i = \frac{n_i}{\sum_i w_i n_i}$$

2.3 Measures of Location and Variability

We often summarize numerical datasets via certain quantities that describe the distribution of the data. One class of descriptors are measures of location, which provide information about the centre of a distribution of data. Some common measures of location are

- The **mean**, expectation value or arithmetic average is a useful measure of the centre of a distribution, but can be unduly affected by extreme outliers
$$\bar{x} = \frac{\sum_i x_i}{n}$$
- The **median** is obtained by ordering the n observations from smallest to largest, and selecting the midpoint there is equal probability of lying above or below.
- The **mode** is the value with the highest probability of appearing, the peak of the distribution

Apart from knowing where a distribution is centred, one often requires information about the spread of the distribution around the centre. This information is provided by the measure of variability. The most commonly used measure of variability is the sample standard deviation σ, given by

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N}}$$

Aside from these, one is often interested in measuring the degree to which two measurements depend on each other. This is quantified by the **linear correlation coefficient**(r)

$$r = \frac{\sum_i (x_i - \bar{x}) * (y_i - \bar{y})}{\sigma_x \sigma_y}$$

An r value of 1 means that x and y are strongly positively correlated, i.e. they rise and fall together. A value of -1, however, implies negative linear correlation. An r value of 0 implies that there is no *linear* correlation between x and y. This is **not** the same as x and y being independent of each other!

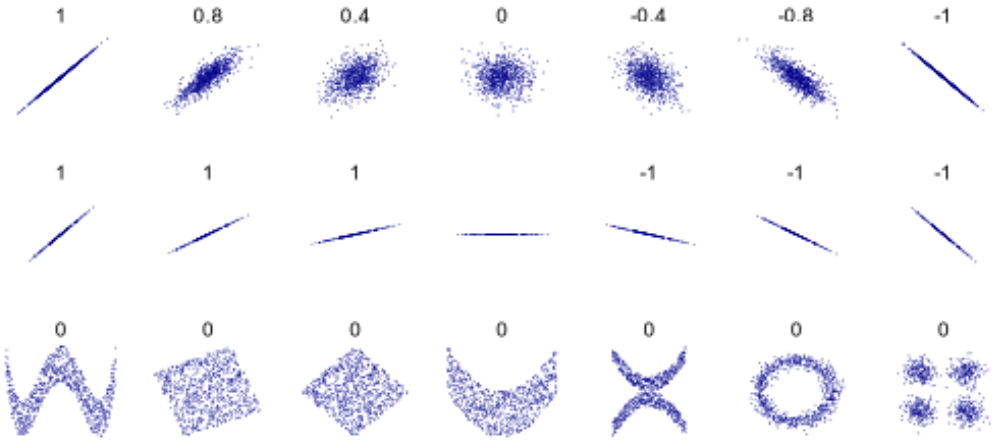

Figure 1: Several sets of (x, y) points, with the correlation coefficient of x and y for each set. (By DenisBoigelot, original uploader was Imagecreator - Own work, original uploader was Imagecreator, CC0, https://commons.wikimedia.org/w/index.php?curid=15165296)

3 Probability

In any situation with multiple outcomes, the study of probabilities provides a method to quantify the likelihood of each outcome. We define an **experiment** to be any activity with an uncertain outcome. The set of all outcomes of an experiment is known as the sample space for the experiment. For example, the sample space ($\mathcal{S}$) of a single flip is either Heads (H) or Tails (T). If we were to do two consecutive coin flips, the sample space would then be (HH;HT;TH;TT). Each outcome in the sample space has an associated probability, and we define an **event** to be any subset of the sample space, equivalent to any collection of outcomes. To advance further, we borrow some definitions from set theory.

- **Complement**: The complement of an event A, denoted by $\hat{A}$, is the set of all outcomes in the sample space, apart from those in A
- **Union**: The union of two events A and B, denoted by A $\bigcup$ B, is the event consisting of all outcomes that are *either* in A *or* in B *or* in both
- **Intersection**: The intersection of two events A and B, denoted by A $\bigcap$ B is the event consisting of all outcomes that are in *both* A and B. If there are no events in the intersection of A and B, they are **disjoint** or **mutually exclusive** events

The idea is to assign to each event A a number $\mathcal{P}(A)$, called the probability of the event A, which will measure the chance that A will occur. We further require P(A) to follow some logical axioms, known as the Kolmogorov Axioms

- For any event A in $\mathcal{S}$, $\mathcal{P}(A) \geq 0$.
- $\mathcal{P}(A) + \mathcal{P}(\hat{A}) = \mathcal{P}(\mathcal{S}) = 1$
- For an infinite collection of mutually exclusive events, $A_1, A_2, ...$ in $\mathcal{S}$,

$$\mathcal{P}(A_1 \bigcup A_2 \bigcup A_3 ...) = \sum_{i=1}^{\infty} \mathcal{P}(A_i)$$

If for example we had a bucket full of coloured marbles, with 30 red marbles, 20 yellow marbles, and 40 blue marbles, and we took out a single marble, the sample space would be (red, yellow, blue). Intuitively, we know we're more likely to pull out a blue marble than a red one, simply because there are more blue marbles, while we're least likely to pull out the sparse yellow marbles. The probability of each event in the sample space must therefore be proportional to the number of that kind of marble in the bucket, and inversely proportional to the total number of marbles. We therefore have

$$\begin{aligned} \mathcal{P}(red) &= \frac{30}{30+20+40} = \frac{30}{90} \\ \mathcal{P}(yellow) &= \frac{20}{30+20+40} = \frac{20}{90} \\ \mathcal{P}(blue) &= \frac{40}{30+20+40} = \frac{40}{90} \end{aligned}$$

We see that the axioms are satisfied.

3.1 Conditional Probability

The probabilities assigned to various events depend on what is known about the experimental situation prior to an experiment. After a single experiment, one may modify the probabilities on the basis of new information. For example, the probability that IceCube observes a neutrino from a blazar may vary depending on whether the blazar is flaring or not. We say the probability of A given B, denoted by $A|B$ is

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \bigcap B)}{\mathcal{P}(B)}$$

From this definition, we see that

$$\mathcal{P}(A|B)\mathcal{P}(B) = \mathcal{P}(B|A)\mathcal{P}(A)$$

known as Bayes' Theorem. For example, suppose that of all individuals buying a certain digital camera, 60% buy a memory card, 40% include an extra battery, and 30% buy both. Let A: memory card purchased and B: battery purchased. then $\mathcal{P}(A) = 0.6$, $\mathcal{P}(B) = 0.4$, and $\mathcal{P}(A \bigcap B) = 0.3$. Given that we know an individual has purchased an extra battery, the probability that an optional card was also purchased is

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \bigcap B)}{\mathcal{P}(B)} = \frac{0.3}{0.4} = 0.75$$

4 Probability Distribution Functions

Some physical problems have well-defined probabilities, given by a probability distribution function (PDF) for continuous variables, or a probability mass function (PMF) for discrete variables. From the Kolmogorov axioms

$$\sum_{i=1}^{max} PMF(x_i) = 1$$
$$\int_{-\infty}^{\infty} PDF(x)dx = 1$$

It is also useful to define cumulative probabilities, which represent the probability of getting a value less than or equal to a specific $\mathbf{x}$

$$CDF(\mathbf{x}) = \sum_{i=1}^{\mathbf{x}} PMF(x_i)$$
$$CDF(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} PDF(x)dx$$

We can evaluate expectation values from the PDFs using

$$g(\bar{x}) = \sum_{i=1}^{max} g(x_i)PMF(x_i)$$
$$g(\bar{x}) = \int_{-\infty}^{\infty} g(x_i)PDF(x_i)dx_i$$

Using $g(x) = x$, we find the mean of the distribution, while $g(x) = (x - \bar{x})^2$ gives us the variance σ^2. A plethora of important PDFs exist, and we explore here some of the most commonly encountered.

4.1 Binomial Distribution

A binomial distribution is a mass distribution which describes the probability of the number of successes in a series of independent experiments with a binary outcome, like tossing a coin or rolling a dice to obtain a specific number (Bernoulli trial).
If you a roll a dice 4 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:

$$\mathcal{P}(1;1;\hat{1};\hat{1}) = \frac{1}{6} \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6}$$

We observe that the two 1's could appear anywhere in our sequence, and we require a combinatoric factor to account for this

$$\mathcal{P} = \binom{4}{2}(\frac{1}{6})^2(\frac{5}{6})^2$$

The general form of the Binomial distribution can be obtained to be

$$\mathcal{P} = \binom{n}{k} p^k (1-p)^{n-k}$$

where k is the number of successes, p is the probability of a success in a single trial, and n is the total number of trials. We find the mean to be np and the standard deviation to be $\sqrt{np(1-p)}$
Exercise: You are doing a magic trick with a normal deck of cards with no jokers. What is the probability after repeating the trick 5 times that at least one ace was drawn by the participant? What is the probability that 2 aces were drawn?

4.2 Poisson Distribution

The Poisson Distribution is often used in counting experiments, and may be considered to be the limiting form of the Binomial Distribution in an experiment in which n is large and p is small, such that np is finite.
The form of the distribution is

$$\mathcal{P} = \frac{e^{-\lambda}\lambda^k}{k!}$$

where λ is the expected value, and k is the number of successes. For example, if we see 10 neutrinos a year on average, what is the probability we see 15 in a particular year?

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

Exercise: At a bus stop, the average rate of buses is one bus per 20 minutes. What is the probability that 5 buses come within an hour??

4.3 Normal Distribution

The normal distribution is the most important one in all of probability and statistics. Many numerical populations have distributions that can be fit very closely by an appropriate normal curve. In addition, even when individual variables themselves are not normally distributed, sums and averages of the variables will under suitable conditions have approximately a normal distribution; this is the content of the Central Limit Theorem

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

where the mean value is μ and the standard deviation is σ.

5 Hypothesis Testing

5.1 Notation

Probability of A	P(A)
Probability of A given B	P(A|B)
Model Parameter Set	$\vec{\theta}$
Data	x_i
Probability of Data given Model	P(x_i|$\vec{\theta}$)

5.2 Frequentist vs Bayesian Statistics

A frequentist approach would be to use the data to evaluate whether a given model describes the data. We do this by calculating what the probability of the observed data is, assuming a certain underlying model is the truth. The Bayesian approach would be to find the probability of our model, given our data, and prior beliefs. i.e

$$\mathrm{P}(\vec{\theta}|\vec{x})=\frac{\mathrm{P}(\vec{x}|\vec{\theta})\mathrm{P}(\vec{\theta})}{\mathrm{P}(\vec{x})}$$

5.3 Log-Likelihood

For a binned analysis, the probability of seeing a certain data distribution is equal to the product of the probabilities of the individual bin counts

$$\mathcal{L}(\vec{\theta})=\prod_i \mathrm{P}(x_i|\vec{\theta})$$
$$\mathrm{Ln}(\mathcal{L}(\vec{\theta}))=\sum_i ln(\mathrm{P}(x_i|\vec{\theta}))$$

Assuming each bin has a gaussian error

$$\ln(\mathcal{L}(\vec{\theta}))=\sum_i \ln(e^{-\frac{(y_i-\mu_i(\theta))^2}{2\sigma_i^2}}))$$
$$\ln(\mathcal{L}(\vec{\theta}))=\sum_i \ln(e^{-\frac{(y_i-\mu_i(\theta))^2}{2\sigma_i^2}}))$$
$$\ln(\mathcal{L}(\vec{\theta}))=-\frac{\chi^2(\theta)}{2})$$

We then optimize these parameters to find the model most likely to produce our data.

5.4 Goodness of Fit

The goodness of fit is a metric used to test how well a model describes data. In the example above, suppose we measure 10 different observables, and our model predicts each of the 10 observables to be Gaussian distributed with a known mean and variance (so no free parameters in the model)

$$X^2=\sum_i(\frac{x_i-\mu}{\sigma})^2)$$

where X^2 follows a χ^2 with 10 degrees of freedom (χ^2_{10})

5.5 Test Statistics and p-values

Define a null-hypothesis, which is usually that the effect being studied does not exist. *Assuming the null hypothesis to be true*, how likely is it that your data looks the way it does? This is answered by the p-value. The first step is to define a test statistic, such as the delta chi-squared (among others).

Suppose you roll a 6-sided die, and you get a 3. This may happen by random 1 in 6 times, so it's not super unlikely. If we rolled the die 8 times, however, and we kept getting 3s, this would be progressively less likely, and we'd start getting suspicious! A p-value is a good metric to determine exactly how suspicious we should be Suppose in the above example, we determine, from data, a $\chi^2_{data} = 25$. We evaluate the p-value to be

$$p = \int_{25}^{\infty} \chi^2_{10}(x)dx \approx 5 \times 10^{-3}$$

and how we interpret this is that, if the model was correct, we would see an observation like the one we made 0.5% of the time.
In a slightly more advanced example, let's say our model predicts that our n observables x_i are Gaussian distributed depending on y parameters θ. Then our distribution follows a χ^2 with n-p degrees of freedom

$$X^2 = \sum_i (\frac{x_i - \mu_i(\theta)}{\sigma})^2)$$

A good fit to data should have a p-value of about 50%, while a small p-value would indicate bad agreement between the data and the model. A very high p-value would indicate that the model errors are overestimated, and should not be as large. We often convert the p-value to an equivalent σ assuming a gaussian distributed null hypothesis. Important significance thresholds are

- 3σ: 1/300 chance of observation occurring at random, evidence
- 5σ: Discovery, 3×10^{-7} chance of observation occurring at random

When you go much higher than 5σ, you want to be careful about claiming significance, because the chances are you're at the tail of your null hypothesis distribution, and that is likely to deviate from a Gaussian

5.6 Systematics

Observational uncertainties can be broadly split into statistical uncertainties, and systematics ones. Statistical uncertainties occur due to finite number of data measurements, and are uncorrelated. Observing for longer/adding more data reduces the statistical uncertainty by $\frac{1}{\sqrt{N}}$. There are various aspects of the detector operations which we do not understand, which we must model as systematic effects. In addition to these, there are uncertainties associated with the various models used, which must also be accounted for. These generally correlate between different datasets, and are not rectified by adding more data. These are pull terms in the Log-Likelihood, penalizing the likelihood when it moves systematic parameters away from where they should physically be

$$\ln(\mathcal{L}(\vec{\theta})) \to \ln(\mathcal{L}(\vec{\theta})) - \frac{(\theta_2 - \mu_{\theta_2})^2}{2\sigma^2_{\theta_2}}$$

Misidentified systematics may affect data results, and you need to carefully evaluate these. A good way to characterize your systematics is to run trials, and plot the correlation between your various systematic parameters as a matrix.