1. What are the summary statistics used to describe data? Contrast between mean and average.

The three most popular summary statistics are mean, variance and median.

Mean

If we have a sample of n values, x_i , the mean μ, is the sum of values divided by hte number of values; ie:

$$\mu = \frac{1}{n} * \sum_i x_i$$

Variance and Std. Deviation

In the same way that the mean is intended to describe the central tendency, variance is intended to describe the spread. The variance of a set of values is

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$

The term x_i - μ is called the “deviation from the mean,” so variance is the mean squared deviation, which is why it is denoted σ^2 . The square root of variance, σ, is called the standard deviation.

Median

The median is the value separating the higher half of a data sample, a population, or a probability distribution, from the lower half.

Other kinds of Summary Statistics used include mode, interquartile mean, range, interquartile range, absolute deviation, mean absolute difference, gini coeeficient, percentiles, etc.

The words “mean” and “average” are sometimes used interchangeably, but the contrast is clear by the following:

- The “mean” of a sample is the summary statistic computed with the previous formula.
- An “average” is one of many summary statistics you might choose to describe the typical value or the central tendency of a sample.

Sometimes the mean is a good description of a set of values. For example, apples are all pretty much the same size (at least the ones sold in supermarkets). So if I buy 6 apples and the total weight is 3 pounds, it would be a reasonable summary to say they are about a half pound each. But pumpkins are more diverse. Suppose I grow several varieties in my garden, and one day I harvest three decorative pumpkins that are 1 pound each, two pie pumpkins that are 3 pounds each, and one Atlantic Giant® pumpkin that weighs 591 pounds. The mean of this sample is 100 pounds, but if I told you “The average pumpkin in my garden is 100 pounds,” that would be wrong, or at least misleading. In this example, there is no meaningful average because there is no typical pumpkin.

2. Bring out the limitations of summary statistics and explain their alternatives.

Summary statistics are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible. Thus they are concise, but dangerous, because they obscure the data. Such statistics are limited in ways so much so that they only allow you to make summations about the people or objects that you have actually measured. You cannot use the data you have collected to generalize to other people or objects (i.e., using data from a sample to infer the properties/parameters of a population). For example, if you tested a drug to beat cancer and it worked in your patients, you cannot claim that it would work in other cancer patients only relying on descriptive statistics (but inferential statistics would give you this opportunity).

An alternative is to look at the distribution of the data, which describes how often each value appears. Inferential statistics relies on this premise, where propositions about a population are made using data drawn from the larger population with some form of sampling. Some examples of such propositions are point estimates, interval estimates, credible intervals, clustering or classificaiton.

The data in inferential statistics are fit to mathematical models that can be parametric, non-parametric or semi-parametric in nature.

3. Illustrate the importance of PMF and write a pseudocode to derive them.

The probability mass function or PMF is a function that maps observations to probabilities. It is a function that gives the probability that a discrete random variable is exactly equal to some value. It is is often the primary means of defining a discrete probability distribution, and such functions exist for either scalar or multivariate random variables whose domain is discrete. They are also very useful for exploratory data analysis as they can help reveal patterns, differences and other features that can be otherwise non-obvious.

We can obtain a PMF for a discrite sample, by dividing the frequency of a value by the size of the sample - i.e the normalized frequencies of the sample.

The following psuedocode illustrates this:

In [6]:
```
#let sample be the list of observations/values
sample = [2,3, 4,4,4,4,4,5,2,10]
freq = {} # freq is a dictionary/map
for x in sample:
    freq[x] = freq.get(x,0)+1
n = float(len(sample))
pmf = {}
for key in freq:
    pmf[key] = freq[key]/n
print pmf
# PMF is the resulting probability mass function for sample.
```

```
{10: 0.1, 3: 0.1, 4: 0.5, 2: 0.2, 5: 0.1}
```

4. Describe two methods of exploratory data analysis with a suitable case study.

There are a number of techniques used in Exploratory Data Analysis (EDA) such as Histograms, Box plots, Scatter Plots, Pareto Charts, etc. Here we shall consider two examples - histograms and box plots.

Histograms

A histogram is a representation that is used to roughly assess the probability distribution of a given variable by depicting the frequencies of observations occurring in certain ranges of values. It is an accurate representation of the distribution of numerical data.To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent, and are often (but are not required to be) of equal size. If the bins are of equal size, a rectangle is erected over the bin with height proportional to the frequency—the number of cases in each bin. A histogram may also be normalized to display "relative" frequencies. It then shows the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1.

However, bins need not be of equal width; in that case, the erected rectangle is defined to have its area proportional to the frequency of cases in the bin.

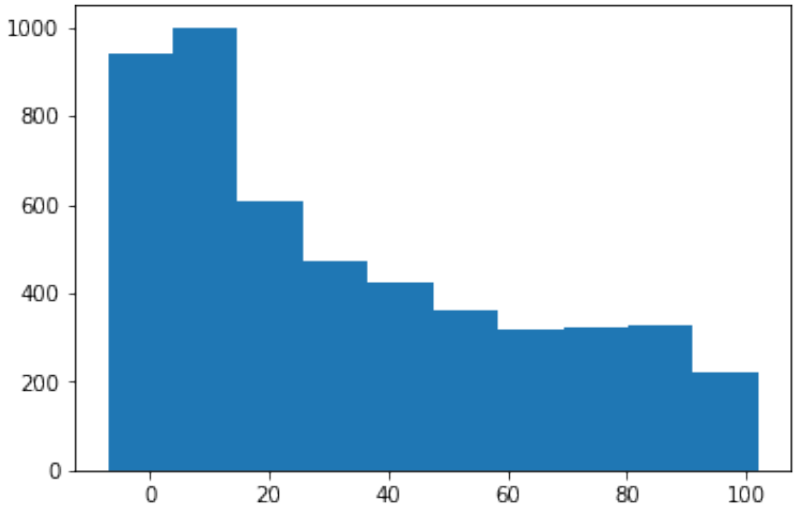
Case Study: Age of a person from Census data in an year. We can see that there are a lot more young people than old.

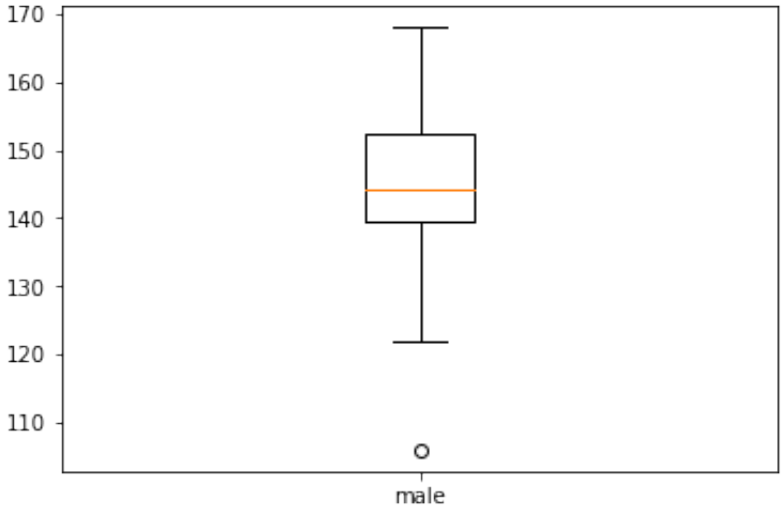
Box Plots

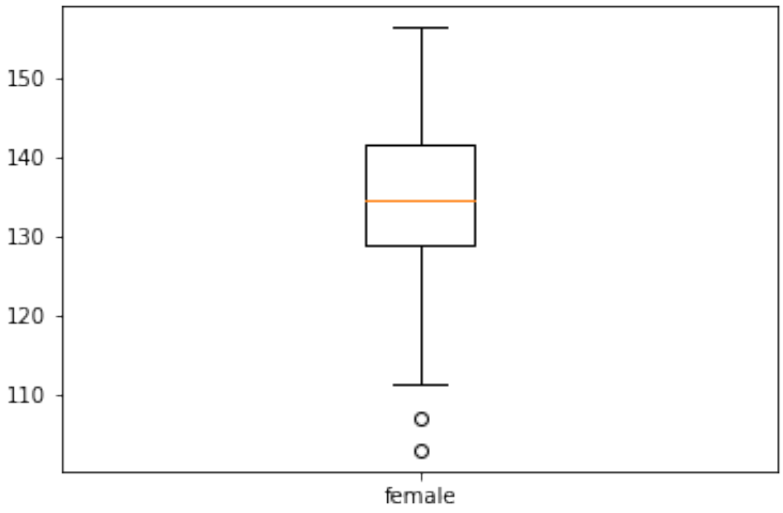
A box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles. After we check the distribution of the data by plotting the histogram, the second thing to do is to look for outliers. Identifying the outliers is important because it might happen that an association we find in our analysis can be explained by the presence of outliers A box plot is one of the best tools to identify such outliers. Through box plots we find minimum, 25th percentile, 50th percentile (median) and 75th percentile along with maximum value of a continuous variable.

Case Study: Heights of 11 year olds. We can remove outlier examples.

In [7]:
```
import matplotlib.pyplot as pyplot
import random
population_estimate = [int(random.gauss((random.random()**2)*100,2)
) for x in range(5000)]
data = population_estimate
histogram = pyplot.hist(data)
pyplot.figure()
heights = [random.gauss(140,10) for x in range(200)]
male = [x+random.uniform(0,10) for x in heights]
female = [x-random.uniform(0,10) for x in heights]
pyplot.boxplot(male,labels=["male"])
pyplot.figure()
pyplot.boxplot(female,labels=["female"])
pyplot.show()
```

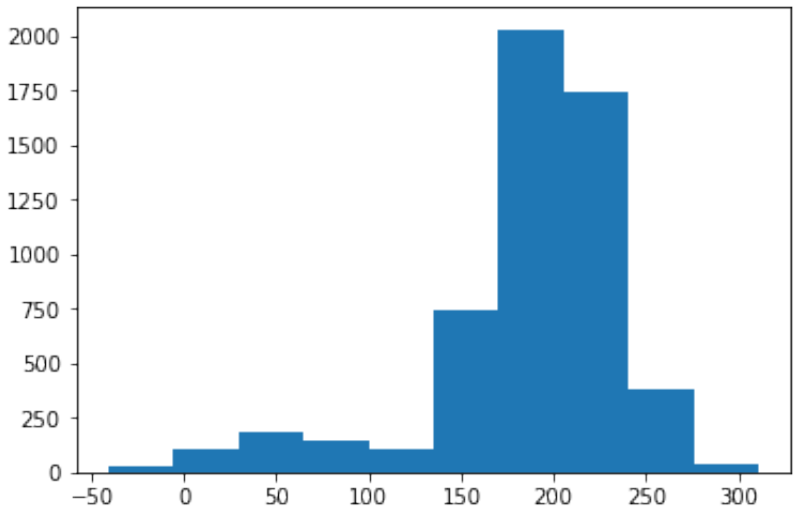






5. Draw a sample histogram and find Mode, Shape and Outliers

In [8]:
```
pyplot.figure()
data = [int(random.gauss(200,30)) for x in range(5000)]
data.extend([int(random.gauss(random.uniform(1,100),20)) for x in r
ange(500)])
histogram = pyplot.hist(data)
pyplot.show()
```



The mode is the value 200 (on x axis), shape is asymmetric about the mode, and outliers are values from -50 to 120.

6. Discuss how Relative Risk and Conditional Probability helps to understand the data better.

Relative Risk is the ratio between two probabilites -eg: probabilities that are representative of two groups or classes of events. This is a summary statistic that can help highlight the differences between the two groups or the significance of an effect.

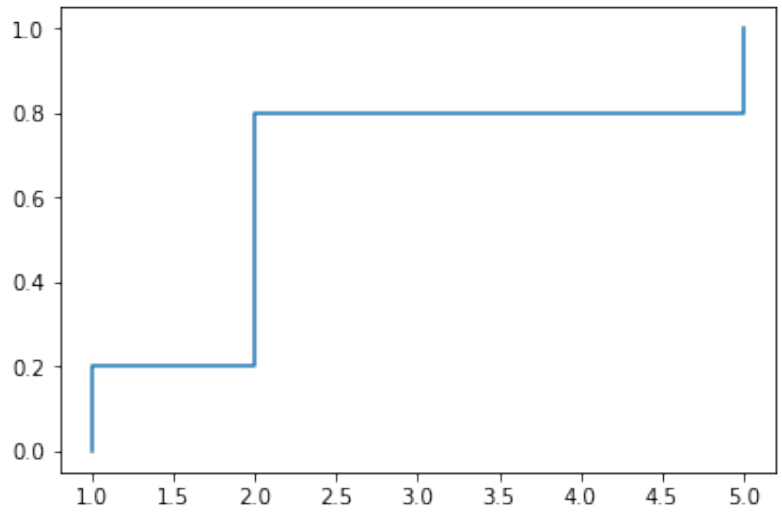
Conditional probability is the probability of an event occuring subject to certain constraints. Real world events are often subject to constraints, or bound by some condition. Conditional Probabilities help model such situations more accurately by taking into account the constraints of the system.

7. How are percentiles used in CDF? Explain with an example.

The Cumulative Distribution Function (CDF) is the function that maps values to their percentile rank in a distribution.The CDF is a function of x, where x is any value that might appear in the distribution. To evaluate CDF(x) for a particular value of x, we compute the fraction of the values in the sample less than (or equal to) x. Here's what that looks like as a function that takes a sample, t, and a value, x:

In [9]:
```
def Render(x,p):
        """Generates a sequence of points suitable for plotting.

        An empirical CDF is a step function; linear interpolation
        can be misleading.

        Returns:
            tuple of (xs, ps)
        """
        xs = [x[0]]
        ps = [0.0]
        for i, v in enumerate(p):
            xs.append(x[i])
            ps.append(v)

            try:
                xs.append(x[i+1])
                ps.append(v)
            except IndexError:
                pass
        return xs, ps

def Cdf(t, x):
    count = 0.0
    for value in t:
        if value <= x:
            count += 1.0
    prob = count / len(t)
    return prob
d = [1,2,2,3,5]
x = [Cdf(d,i) for i in range(6)]
print x
pyplot.figure()
a,b = Render(d,x[1:])
pyplot.plot(a,b)
pyplot.show()
```

```
[0.0, 0.2, 0.6, 0.8, 0.8, 1.0]
```



8. Compare PMF and CDF and bring out their differences.

PMF stands for Probability Mass Function. The PMF is a function that maps observations to probabilities. It is a function that gives the probability that a discrete random variable is exactly equal to some value. The Sum of a PMF always adds upto 1. The PMF for a discrete random variable X, gives $P(X = x)$

Example, given a set of values {1,2,2,3,5}, the PMF is:

PMF(0) = 0

PMF(1) = 0.2

PMF(2) = 0.4

PMF(3) = 0.2

PMF(4) = 0

PMF(5) = 0.2

PMF can be susceptible to noise and have very low values when the number of observations is large. The histogram of PMF can sometimes have lots of spikes or valleys. CDF doesn't have such limitations.

CDF stands for Cumulative Distribution Function. The CDF is the function that maps values to their percentile rank in a distribution. The CDF is a function of x, where x is any value that might appear in the distribution. To evaluate CDF(x) for a particular value of x, we compute the fraction of the values in the sample less than or equal to x. The CDF function is always greater than or equal to 1. The CDF for a random variable X gives $P(X \leq x)$

Example, given a set of values {1,2,2,3,5}, the CDF is:

CDF(0) = 0

CDF(1) = 0.2

CDF(2) = 0.6

CDF(3) = 0.8

CDF(4) = 0.8

CDF(5) = 1

The CDF can be for discrite, continuous, or mixed distributions. PMF is only for discrete distributions.

9. How can conditional distributions help optimise the CDF?

Reference (https://www.colorado.edu/economics/morey/7818/jointdensity/NotesonConditionalCDFs/ConditionalCDF

A conditional distribution is the distribution of a subset of the data which is selected according to a condition.

The general form of conditional CDF is:

$$F_{XY}(x \mid y) = P_r[X \leq x \mid Y \leq y] = \frac{F_{XY}(x, y)}{F_Y(y)}$$

Dividing the joint CDF by the marginal CDF allows us to normalize our conditional distribution so that it maintains the necessary properties of the CDF only in terms of the variable we are not conditioning upon. For example, if we looked at the formula above, this particular conditional CDF only maintains the properties of the CDF in terms of X.

Conditional distributions are useful for comparing measurements from different tests, or tests applied to different groups. For example, people who compete in foot races are usually grouped by age and gender. To compare people in different groups, you can convert race times to percentile ranks. The reason for this is to reduce the variations in the data. It also helps us obtain some interesting insights.

10. Demonstrate how CDF can be used to generate random numbers.

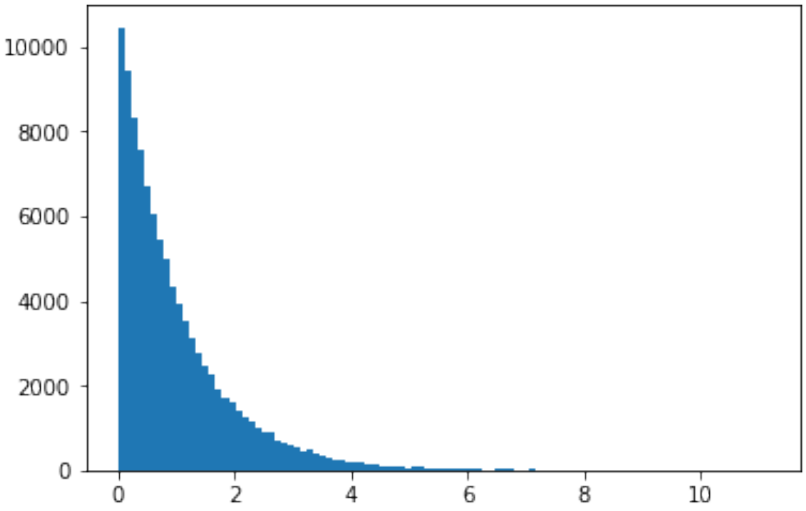
CDFs are useful for generating random numbers with a given distribution. Here's how:

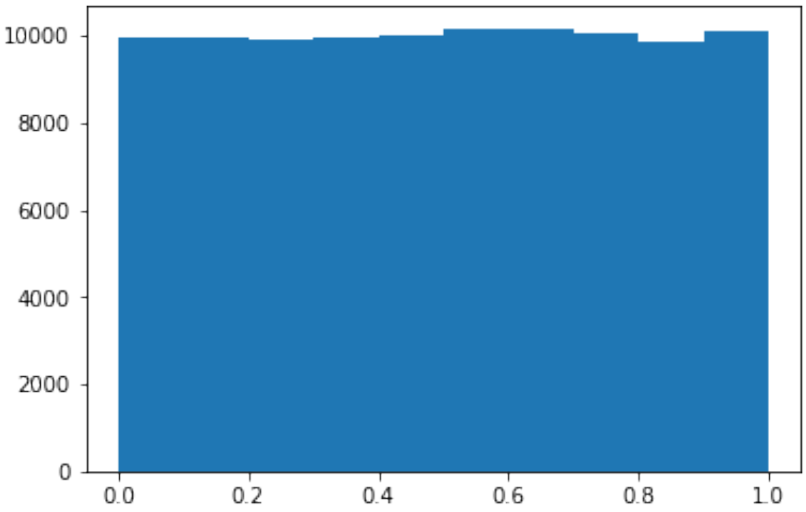
- Choose a random probability in the range 0–1.
- Use the inverse CDF to find the value in the distribution that corresponds to the probability you chose.

This is a by product of the properties of CDFs. If the CDF F is strictly increasing and continuous then $F^{-1}(p), p \in [0, 1]$, is the unique real number x such that $F(x) = p$. In such a case, this is called the inverse distribution function of F

Let's say, we have 10,000 random numbers drawn from exponential distribution. Note, the cdf of exponential is a simple analytic function, $F(p) = 1 - e^{-\lambda p}$. So, if you transform all those 10,000 randoms you just generated using this function $F(p)$, the resulting 10,000 numbers you get will be uniformly distributed. In other words, if you plot an histogram of these resulting numbers after transformation, you should see something like in figure below.

In [10]:
```
import numpy as np
import matplotlib.pyplot as plt

def cdf_of_exponential( x, scale=1.0 ): #only for x>0
    return 1.0 - np.exp(-scale*x)

_set = []
_transformed_set = []
for i in range(100000):
    expp = np.random.exponential() #exponential dist randoms
    texpp = cdf_of_exponential( expp )

    _set.append( expp )
    _transformed_set.append( texpp )

plt.hist( _set, 100 )
plt.figure()
plt.hist( _transformed_set )
plt.show()
```





11. Differentiate empirical and continuous distributions and explain the relevance

The distributions we have used so far are called empirical distributions because they are based on empirical observations, which are necessarily finite samples.

The alternative is a continuous distribution, which is characterized by a CDF that is a continuous function (as opposed to a step function). Many real world phenomena can be approximated by continuous distributions.

Examples of continuous distributions are exponential distribution, normal distribution, etc

12. Model a real world scenario with exponential distribution and bring out the inferences.

Example:We will look at the interarrival time of babies recorded on a day in a hospital in Brisbane, Australia. On December 18, 1997, 44 babies were born. The times of birth for all 44 babies were reported in the local paper; you can download the data from here (http://thinkstats.com/babyboom.dat)

The CDF of the interarrival times in minutes can be found, which seems to have the general shapte of an exponential distribution.

The CDF of an exponential distribution is given by:

$$F(p) = 1 - e^{-\lambda p}$$

The parameter, λ, determines the shape of the distribution. In general, the mean of an exponential distribution is $1/\lambda$ and the median is given by $ln(2)/\lambda$.

In the real world, exponential distributions often come up when we look at a series of events and measure the times between events, which are called interarrival times. If the events are equally likely to occur at any time, the distribution of interarrival times tends to look like an exponential distribution.

13. Why CCDF should be considered along with CDF? Illustrate with an example.

When checking if a CDF is exponential, it is often convenient to use the Complimentary CDF (CCDF),which is 1-CDF(x), instead of directly plotting the CDF value. If the plot of CCDF on the log-y scale is a straight line, then we can easily conclude that the CDF is an exponential distribution. The example below illustrates this.

Example: Let the CDF of a distribution be given by the function $F(p) = 1 - e^{-\lambda p}$ then the CCDF of p is $e^{-\lambda p}$.

If we were to plot this, we would get

$$y \approx e^{\lambda p}$$

Taking log on both sides gives

$$logy \approx -\lambda x$$

which is a straight line with slope $-\lambda$

14. How to determine whether a data set exhibits Pereto distribution?

The CDF of the Pareto distribution is given by

$$CDF(x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}$$

The parameters x_m and α determine the location and shape of the distribution. x_m is the minimum possible value. To determine whether a data set exhibits Pereto distribution we can plot the values of the CCDF on log-log scale. This should turn out to be a straight line if the CDF of the distribution follows Pereto Distribution.

To see this, consider the CCDF of Pereto distribution being plotted:

$$y \approx \left(\frac{x}{x_m}\right)^{-\alpha}$$

Taking log on both sides, we get

$$log(y) \approx -\alpha(log(x) - log(x_m))$$

On a logarithmic scale, this is just a straight line with slope $-\alpha$ and intercept $\alpha * log(x_m)$

Thus to check if the data set exhibits Pereto distribution, plot the CCDF of this function (1 - CDF(x)) on a log-log scale. If it is a straight line, then its following Pereto distribution.

15. Explain the popularity of Normal distribution in real world.

This is mainly because of the Central Limit Theorem. In probability theory, the central limit theorem (CLT) establishes that, in most situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions. It turns out that if we add up a large number of values from almost any distribution, the distribution of the sum converges to normal. More specifically, if the distribution of the values has mean and standard deviation μ and σ, the distribution of the sum is approximately $N(n\mu, n\sigma^2)$ provided that:

- The values have been drawn independently.
- The values have to come from the same distribution (although thisrequirement can be relaxed)
- The values have to be drawn from a distribution with finite mean and variance, so most Pareto distributions are out
- The number of values you need before you see convergence depends on the skewness of the distribution. Sums from an exponential distribution converge for small sample sizes. Sums from a lognormal distribution do not.

This explains the prevalence of normal distributions in the natural world. Most characteristics of animals and other life forms are affected by a large number of genetic and environmental factors whose effect is additive. The characteristics we measure are the sum of a large number of small effects, so their distribution tends to be normal.

16. How to model a real world phenomenon as a particular continuous distribution?

Like all models, continuous distributions are abstractions, which means they leave out details that are considered irrelevant. For example, an observed distribution might have measurement errors or quirks that are specific to the sample; continuous models smooth out these idiosyncrasies.

Continuous models are also a form of data compression. When a model fits a dataset well, a small set of parameters can summarize a large amount of data.

It is sometimes surprising when data from a natural phenomenon fit a continuous distribution, but these observations can lead to insight into physical systems.

When modelling real world phenomenon as a continuous distribution we can often examine the plot of the CCDF of the observations to see if it fits one of the known continuous distributions such as Exponential, Pereto, etc. We can also plot the normal probability plot of the dataset by using rankits. If we generate n values from a normal distribution and sort them, the kth rankit is teh mean of the distribution for the kth value.

An approximate way to generate this is to generate a sample with the same size as the dataset from N(0,1). Sorting the values of the dataset, we plot the sampled values with the sorted dataset values. The resulting curve is a good approximation of the normal probability plot.

If the logarithms of a set of values have a normal distribution, the valueshave a lognormal distribution. The CDF of a lognormal distribution is the same as a normal distribution and thus it can be used to model a real world scenario.

17. Compare Frequentism with Bayesianism and bring out their relative merits.

There is general agreement that a probability is a real value between 0 and 1 that is intended to be a quantitative measure corresponding to the qualitative notion that some things are more likely than others. If we have a finite sample of n trials and we observe s successes, the probability of success is s/n. If the set of trials is infinite, defining probabilities is a little trickier, but most people are willing to accept probabilistic claims about a hypothetical series of identical trials, like tossing a coin or rolling a die.

Consider the probability of a candidate winning an election. This cannot be determined according to some people as they argue that as there is no series of identical trials to consider in such as situation. This position is called frequentism because it defines probability in terms of frequencies. If there is no set of identical trials, there is no probability.

Frequentism is philosophically safe, but it limits the scope of probability to physical systems that are either random (like atomic decay) or so unpredictable that we model them as random (like a tumbling die). Anything involving people is pretty much not considered.

However Bayesianism, which defines probability as a degree of belief that an event will occur. By this definition, the notion of probability can be applied in almost any circumstance. One difficulty with Bayesian probability is that it depends on a person’s state of knowledge; people with different information might have different degrees of belief about the same event. For this reason, many people think that Bayesian probabilities are more subjective than frequency probabilities.

18. Argue why “Switch” is the optimal strategy in the Monty Hall problem.

The problem is defined as follows:

- Monty shows you three closed doors and tells you that there is a prize behind each door: one prize is a car, the other two are less valuable prizes like peanut butter and fake finger nails. The prizes are arranged at random.
- The object of the game is to guess which door has the car. If you guess right, you get to keep the car.
- So you pick a door, which we will call Door A. We’ll call the other doors B and C.
- Before opening the door you chose, Monty likes to increase the suspense by opening either Door B or C, whichever does not have the car. (If the car is actually behind Door A, Monty can safely open B or C, so he chooses one at random).
- Then Monty offers you the option to stick with your original choice or switch to the one remaining unopened door.

The question is, should you “stick” or “switch” or does it make no difference?

Probability theory favors "Switch" over "Stick". This can be explained by considering the probability of find the car behind the doors. There are three possible scenarios: the car is behind Door A, B or C. Since the prizes are arranged at random, the probability of each scenario is 1/3.

If your strategy is to stick with Door A, then you will win only in Scenario A, which has probability 1/3.

If your strategy is to switch, you will win in either Scenario B or Scenario C, so the total probability of winning is 2/3.

This is because If two events are mutually exclusive, that means that only one of them can happen, the probability of either event occuring is:

$$P(A \ or \ B) = P(A) + P(B)$$

19. State and analyse the implications of PMF of Binomial distribution and its coefficient.

The PMF of Binomial distribution is given by

$$PMF(k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}$$

The binomial coefficient is pronounced “n choose k”, and it can be computed directly like this:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

or recursively:

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

The implications of the PMF of a binomial distribution is that if the probability of an event occuring in the distribution is given by p, and the probability of the event not occuring is given by q, then the probability of the event occuring k times for all values of $k \in [0, n]$ is given by the PMF of the binomial distribution. Where n is the number of trials.

The value of the coefficient is also the same as the number of ways to select k items from a set of n items. Which is an important measure in combinatorics and computer science. It also gives the expansion of $(x + y)^n$, the pascal triangle, finding the number of paths in a grid, etc.

20. Analyse the diachronic interpretation of Bayes's theorem.

Bayes’s theorem is often interpreted as a statement about how a body of evidence, E, affects the probability of a hypothesis, H:

$$P(H \mid E) = P(H)\frac{P(E \mid H)}{P(E)}$$

In words, this equation says that the probability of H after you have seen E is the product of P(H), which is the probability of H before you saw the evidence, and the ratio of P(E|H), the probability of seeing the evidence assuming that H is true, and P(E), the probability of seeing the evidence under any circumstances (H true or not).

This way of reading Bayes’s theorem is called the “diachronic” interpretation because it describes how the probability of a hypothesis gets updated over time, usually in light of new evidence. In this context, P(H) is called the prior probability and P(H|E) is called the posterior. P(E|H) is the likelihood of the evidence, and P(E) is the normalizing constant