S21CSE471: Statistical Methods in AI **Spring 2021**

Some Basics Required in Statistics & Probability

Prepared by :

Aditi Aggarwal (2020201034), Satyam Singh (2020201035), Shweta Arya (2020201047)

1 Sample Vs. Population

A Population includes all of the elements from a set of data, whereas a Sample consists one or more observations drawn from the population using some sampling method. We are normally interested in knowing measurable characteristic of the population because our population contains all the values we are interested in. However, in statistics, we are usually presented with a sample from which we wish to estimate a population.

We use slightly different formula for calculating measurable characteristic in case of Sample or Population. A detailed analysis can be found ***here***. In this document we will be using formula for measurable characteristic for whole population unless specified otherwise.

2 Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

1. **Mean** : The mean μ is calculated as the sum of all the observations in the data set divided by the total number of observations n in the data set. A important properties of Mean is that it minimises the error in the prediction of any one value in the data set. That is, it is the value that produces the lowest amount of error from all other values in the data set. It can be formulated as

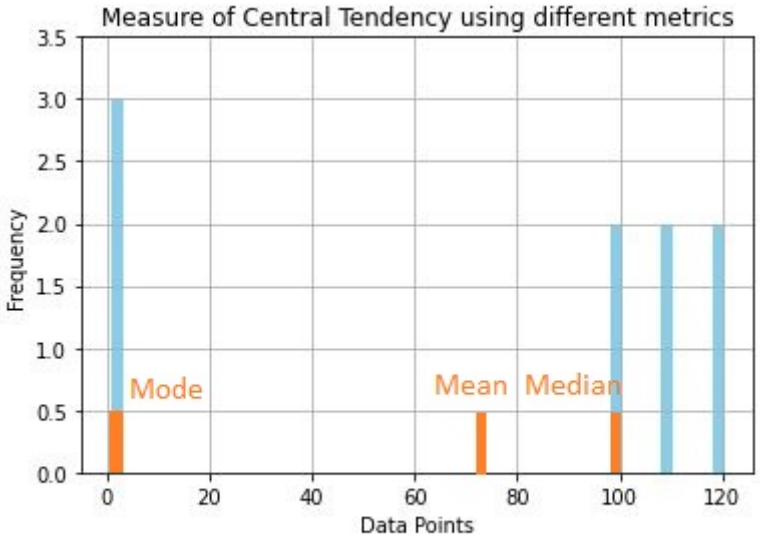
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

   One thing to be cautious about while considering Mean as a measurement of central tendency is that, it is susceptible to the influence of outliers. Also if our data is skewed then mean lies away from the typical value towards the skew.

   For example, consider the following data [10, 15, 20, 20, 30, 127]. Here we get $\mu = 37$, but most of our data points lie in the range of 10 to 30. So it is better to consider Median in this case than Mean.

2. **Median** : The median is the data point lying in the middle of data set that has been arranged in order of their magnitude. If the number of observations is even then we find Median by averaging out the $(\frac{N}{2})^{th}$ and $(\frac{N}{2}+1)^{th}$ elements. The good thing about median is that it is less affected by outliers and skewed data.

   For example, if we consider the example from above [10, 15, 20, 20, 30, 127] then we get the Median as $\frac{20+20}{2} = 20$ which turns out to be a good measure of central tendency in this case.




```
Mean = 73.67 , Median = 100 , Mode = 1
```

3. **Mode** : The mode is the most frequently occurring value in our data set. Normally, the mode is used for categorical data where we wish to know which is the most common category. Mode however, is not a very popular way of measuring central tendency because of following problems. One of the problems with the mode is that it is not unique, so we can have two or more values that share the highest frequency. Also when we have continuous data we are not so likely to have any one value that is more frequent than the other. Another problem with the mode is that when the most common mark is far away from the rest of the data in the data set we can not get a good measure of central tendency.

   For example, consider data set [1, 1, 1, 100, 100, 110, 110, 120, 120]. Here Mode is clearly 1, however the central measure is far from 1 and is rather shifted more towards 100.

For measuring the central tendency in a data set, choosing a appropriate metric is crucial. We can observe the same in the below visualization of the data set we just discussed in above example.

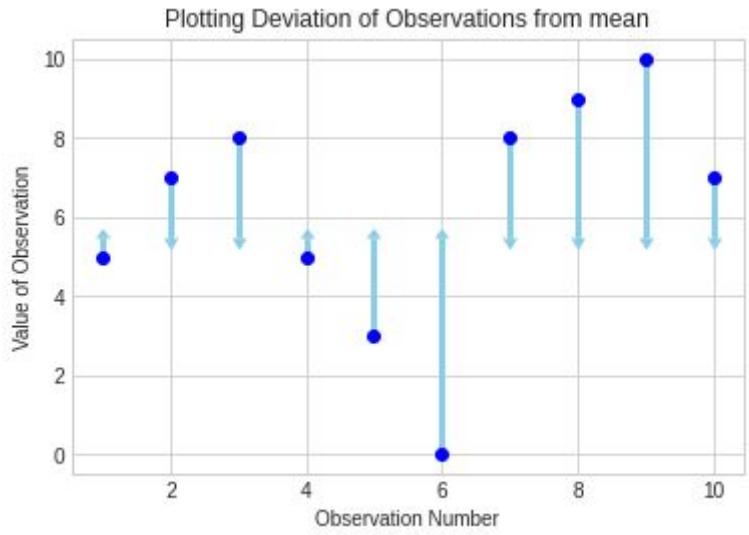
3 Measures of Spread

Along with the measure of central tendency we often use a Measure of spread of data to describe the variability in a sample or population. A measure of spread gives us an idea of how well the our choice of metric to measure of central tendency represents the data. We will discuss some of the most popular metrics here.

1. **Mean Absolute Deviation** : It is a measure of deviation of the data points from the mean value of the data set. The Deviation can be above or below the mean value, so we will consider only the absolute value of deviations. Adding up all of these absolute deviations and dividing them by the size of the data set will give us the mean absolute deviation. We can formulate it as:

$$M.A.D = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

2. **Variance** : A difference between Mean Absolute Deviation and Variance is that, Variance squares up the deviation term instead of performing the absolute operation on it. We can then add up all



squared deviation from mean and then divide by the size of the data set. If the points in our data set are spread out, the variance will be a very large number. Conversely, if the points stick closely around the mean, the variance will be a smaller number. The formula for calculating Variance σ^2 is given as:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

There are a few points to note about variance. First, as the deviations terms here are 'squared', this gives more weight to extreme values i.e. outliers. Also, the variance does not match with the values in our data set because of the 'squaring'. This means we cannot place it on our frequency distribution or relate its value directly to other values in our data set. Calculating the standard deviation rather than the variance rectifies this problem.

3. **Standard Deviation** : The average distance between the values of the data in the collection and the mean is measured by Standard Deviation σ. It is a "natural" indicator of statistical dispersion.

   - A low standard deviation means the data points are similar to the mean whereas, a high standard deviation means that the data points are spread out over a wide variety of values.
   - Standard deviation is widely used to assess confidence in mathematical conclusions in addition to expressing the heterogeneity of a population.

   It is calculated by finding out the square root of variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

4 Random Variable

A random variable X, is a variable whose possible values are numerical outcomes of a random phenomenon. It is different than Algebraic variables in the sense that a Random Variable has a whole set of values (called Sample Space) and it could take on any of those values, randomly.
For example, If we define a Random Variable X = "The score shown on the top face of a die", then X

could be 1, 2, 3, 4, 5 or 6. Here probability of getting 6 on top face is given by $P[X = 6] = 1/6$.
There are two types of random variables:

- Discrete : A Discrete Random Variable can take only a finite number of distinct values. For examples number of faces in a die, Number of face cards in a game of cards, etc.
- Continuous : A continuous random variable is one which takes an infinite number of possible values. They are usually measurements. For examples height or weight of a person, etc. it is not defined at specific values but over an interval of values, and is represented by the area under the curve.

We have already discussed about Expectation, Variance and Standard Deviation, Now in this section we will see how we can introduce Random Variables in the context of Expectation, Variance and Standard Deviation.

1. **Expected Value of Random Variable :** Mathematical expectation $E(X)$, also known as the expected value, is the summation (Discrete Data) or integration (Continuous Data) of a possible values from a random variable. In other words it is the product of the probability of an event occurring, and the value corresponding to occurrence of the event. From the definition we can infer that Mean μ is actually a special case of Expectation when each event is equally likely to occur.

$$\text{For Discrete Random Variable : } E[X] = \sum_i x_i P(x_i)$$

$$\text{For Continuous Random Variable : } E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

2. **Variance of Random Variable :** We have already talked about what Variance means. So let us quickly see how we express it if we are given a Random Variable X.

$$\text{For Discrete Random Variable : } V[X] = E[X^2] - E[X]^2$$

$$\text{For Continuous Random Variable : } V[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

3. **Standard Deviation of Random Variable :** It is expressed as the square root of variance.

$$\sigma = \sqrt{V[X]}$$

5 Covariance

Variables can be connected by a linear relationship that is continuously proportional between the two variables. This interaction is referred to as covariance. It is determined as the sum of the product between the values of each sample where the values have been centered.

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

In terms of expected values,

$$cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y]$$

The sign of the covariance indicates whether the two variables are changing in the same direction (positive) or in opposite directions (negative). If two variables are independent, their covariance is zero. However, a covariance of zero does not guarantee independence.

Since the mean is used in the estimation, each data set should have a Gaussian or Gaussian-like distribution.

Drawback of covariance is that it scales with the random variables X and Y. So if we are to change the units of X and Y, we will scale the covariance which makes it diffult to know how strongly are two random variables connected.

6 Correlation

Correlation is a statistic that measures the degree to which two variables move in relation to each other. It is given by the formula:

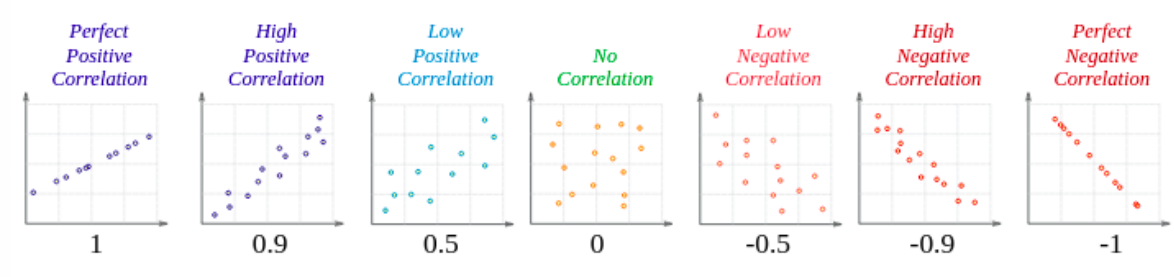
$$corr(x, y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2 \sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

A correlation may be positive, indicating that the two variables shift in the same direction, or negative, meaning that as the value of one variable increases, the values of the other variables decrease.

Correlation is broadly of three types:

- Positive Correlation: Both variables change in the same direction.
- Negative Correlation: Variables change in opposite directions.
- Neutral Correlation: No relationship in the change of the variables.

If the variables are unrelated, the correlation will be zero.



Correlation is obtained by normalizing the covariance. In particular, we define the correlation coefficient of two random variables X and Y as the covariance of the standardized versions of X and Y. Define the standardized versions of X and Y as:

$$U = \frac{X - \mu_X}{\sigma_X} \qquad V = \frac{Y - \mu_Y}{\sigma_Y}$$

$$corr(X, Y) = cov(U, V) = cov\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right)$$
$$= \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Hence, correlation is invariant to the units of X and Y unlike covariance.

A common saying is "Correlation Is Not Causation". What it really means is that a correlation does not prove one thing causes the other:

- One thing might cause the other
- The other might cause the first to happen
- They may be linked by a different thing
- Or it could be random chance!

There can be many reasons the data has a good correlation.

7 Standard Deviation

The average distance between the values of the data in the collection and the mean is measured by standard deviation.

- A low standard deviation means the data points are similar to the mean; a high standard deviation means the data points are spread out over a wide variety of values.
- Standard deviation is widely used to assess confidence in mathematical conclusions in addition to expressing the heterogeneity of a population.
- To find the population standard deviation, divide each data point's variance from the mean by two and square the result. So, take the square root of the average of these values.
- To know if the center of the data is measured around the mean, the standard deviation is a "natural" indicator of statistical dispersion since the standard deviation from the mean is lower than from every other point.

The variance of a data sample drawn from a Gaussian distribution is calculated as the average squared difference of each observation in the sample from the sample mean.

$$var(x) = 1/(n-1) \times \sum_{i=1}^{n} (x_i - mean(x))^2$$

Sometimes, it is hard to interpret the variance because the units are the squared units of the observations. In that case when the spread of a Gaussian distribution is summarized, it is described using the square root of the variance. And this is called standard deviation.

8 Conditional Probability

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

If A and B are two events in a sample space S, then the conditional probability of A given B is defined as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ when } P(B) > 0 \tag{1}$$

Some Basic Axioms: Axiom 1: For any event A,

$$P(A \mid B) > 0 \tag{2}$$

Axiom 2: Conditional probability of B given B is 1, i.e.,

$$P(B \mid B) = 1 \tag{3}$$

Axiom 3: If A1,A2,A3,... are disjoint events, then

$$P(A1 \cup A2 \cup A3 \cdots \mid B) = P(A1 \mid B) + P(A2 \mid B) + P(AB \mid B) + \cdots \tag{4}$$

9 Bayes's Theorem for Conditional Probability

Bayes Theorem provides a principled way for calculating a conditional probability. Although it is a powerful tool in the field of probability, Bayes Theorem is also widely used in the field of machine learning. Including its use in a probability framework for fitting a model to a training dataset, referred to as maximum a posteriori or MAP for short, and in developing models for classification predictive modeling problems such as the Bayes Optimal Classifier and Naive Bayes.

Total probability Theorem: Given n mutually exclusive events $A_1, ..., A_2$ whose probabilities sum to unity, then $P(B) = P(B \mid A_1)P(A_1) + ... + P(B \mid A_n)P(A_n)$, where B is an arbitrary event, and $P(B \mid A_1)$ is the conditional probability of B assuming A_i.

If the random variable is independent, then it is the probability of the event directly, otherwise, if the variable is dependent upon other variables, then the marginal probability is the probability of the event summed over all outcomes for the dependent variables, called the sum rule.

The Bayes' theorem is expressed in the following formula:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)} \tag{5}$$

Where:

$P(A \mid B)$ – the probability of event A occurring, given event B has occurred.

$P(B \mid A)$ – the probability of event B occurring, given event A has occurred.

$P(A)$ – the probability of event A.

$P(B)$ – the probability of event B.

Note that events A and B are independent events (i.e., the probability of the outcome of event A does not depend on the probability of the outcome of event B).

A special case of the Bayes' theorem is when event A is a binary variable. In such a case, the theorem is expressed in the following way:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P\left(B \mid A^{-}\right) P\left(A^{-}\right) + P\left(B \mid A^{+}\right) P\left(A^{+}\right)} \tag{6}$$

.

$P(B \mid A^{-})$: the probability of event B occurring given that event A^{-} has occurred.

$P(B \mid A^+)$:the probability of event B occurring given that event A^+ has occurred.

In the special case above, events A^- and A^+ are mutually exclusive outcomes of event A.

Example of Bayes' Theorem

Imagine you are a financial analyst at an investment bank. According to your research of publicly-traded companies, 60% of the companies that increased their share price by more than 5% in the last three years replaced their CEOs during the period.

At the same time, only 35%of the companies that did not increase their share price by more than 5% in the same period replaced their CEOs. Knowing that the probability that the stock prices grow by more than 5% is 4% , find the probability that the shares of a company that fires its CEO will increase by more than 5% .

Before finding the probabilities, you must first define the notation of the probabilities.

P(A) – the probability that the stock price increases by 5% .

P(B) – the probability that the CEO is replaced.

$P(A \mid B)$ – the probability of the stock price increases by 5% given that the CEO has been replaced.

$P(B \mid A)$ – the probability of the CEO replacement given the stock price has increased by 5% .

Using the Bayes' theorem, we can find the required probability:

$$P(A \mid B) = \frac{0.60 \times 0.04}{0.60 \times 0.04 + 0.35 \times (1 - 0.04)} = 0.067 \text{ or } 6.67\% \tag{7}$$

10 Distributions

Data types:

- Discrete Data, as the name suggests, can take only specified values. For example, when you roll a die, the possible outcomes are 1, 2, 3, 4, 5 or 6 and not 1.5 or 2.45.
- Continuous Data can take any value within a given range. The range may be finite or infinite. For example, A girl's weight or height, the length of the road. The weight of a girl can be any value from 54 kgs, or 54.5 kgs, or 54.5436kgs.

Statisticians divide probability distributions into the following types:

- Discrete Probability Distributions
- Continuous Probability Distributions

Different distributions:

- Bernoulli Distribution
- Uniform Distribution
- Binomial Distribution

- Normal Distribution
- Poisson Distribution

10.1 Bernoulli Distribution

Let's start with the easiest distribution that is Bernoulli Distribution. Let's say if you toss a coin, it results in a head, you win. Else, you lose if it is a tail. There's no midway. A Bernoulli distribution has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

Here, the occurrence of a head denotes success, and the occurrence of a tail denotes failure. Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes. The probability mass function is given by: px(1-p)1-x where x € (0, 1). It can also be written as:

$$P(x) = \begin{cases} 1-p, x=0 \\ p, x=1 \end{cases} \tag{8}$$

The probabilities of success and failure need not be equally likely, like the result of a fight between me and Undertaker. He is pretty much certain to win. So in this case probability of my success is 0.15 while my failure is 0.85. Here, the probability of success(p) is not same as the probability of failure. So, the chart below shows the Bernoulli Distribution of our fight.

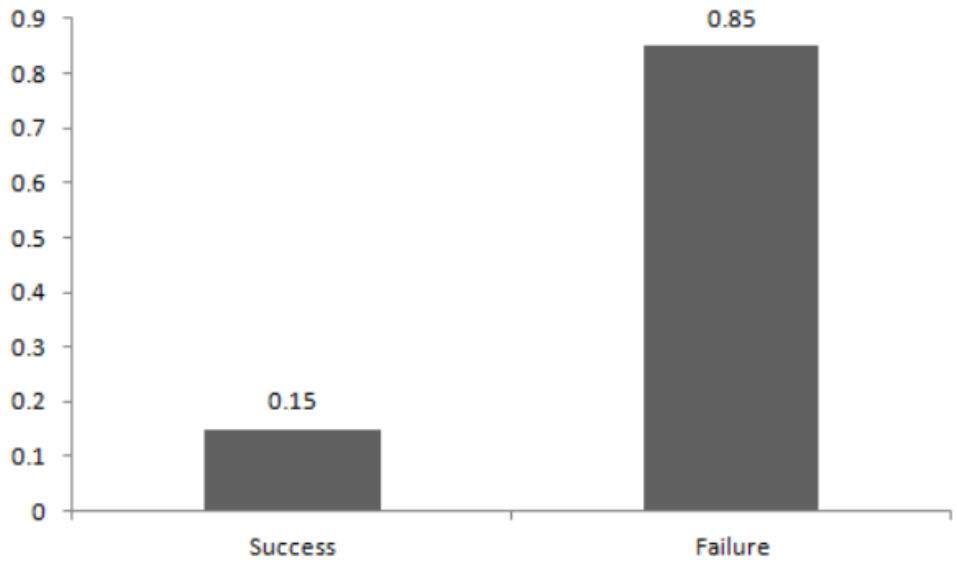


Figure 1: Bernoulli Distribution

Here, the probability of success = 0.15 and probability of failure = 0.85. The expected value is exactly what it sounds. If I punch you, I may expect you to punch me back. Basically expected value of any distribution is the mean of the distribution. The expected value of a random variable X from a Bernoulli distribution is found as follows

$$E(X) = 1^{\star}p + 0^{\star}(1-p) = p \tag{9}$$

The variance of a random variable from a bernoulli distribution is

$$V(X) = E\left(X^2\right) - [E(X)]^2 = p - p^2 = p(1-p) \tag{10}$$

10.2 Uniform Distribution

When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

A variable X is said to be uniformly distributed if the density function is

$$\boldsymbol{f}(\boldsymbol{x}) = \frac{1}{b-\alpha} \quad \text{for } -\infty < \text{a} \leq \text{x} \leq \text{b} < \infty \tag{11}$$

The graph of a uniform distribution curve looks like is given on next page.

Figure 2: Uniform Distribution

For a Uniform Distribution, a and b are the parameters.The number of bouquets sold daily at a flower shop is uniformly distributed with a maximum of 40 and a minimum of 10. Let's try calculating the probability that the daily sales will fall between 15 and 30.

The probability that daily sales will fall between 15 and 30 is (30-15)*(1/(40-10)) = 0.5

Similarly, the probability that daily sales are greater than 20 is = 0.667

The mean and variance of X following a uniform distribution is:

Mean: E(X) = (a+b)/2

Variance: V(X) = (b-a)^2/12

The standard uniform density has parameters a = 0 and b = 1, so the PDF for standard uniform density is given by

$$f(x) = \begin{cases} 1, 0 \leq x \leq 1 \\ 0, \text{ otherwise} \end{cases} \tag{12}$$

10.3 Binomial Distribution

Suppose that you won the toss today and this indicates a successful event. You toss again but you lost this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow. Let's assign a random variable, say X, to the number of times you won the toss. What can be the possible value of X? It can be any number depending on the number of times you tossed a coin. There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, probability of getting

a head = 0.5 and the probability of failure can be easily computed as: q = 1- p = 0.5. A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution. The outcomes need not be equally likely. Remember the example of a fight between me and Undertaker? So, if the probability of success in an experiment is 0.2 then the probability of failure can be easily computed as q = 1 – 0.2 = 0.8. Each trial is independent since the outcome of the previous toss doesn't determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

Properties of a Binomial Distribution are

- Each trial is independent.
- There are only two possible outcomes in a trial- either a success or a failure.
- A total number of n identical trials are conducted.
- The probability of success and failure is same for all trials. (Trials are identical.)

The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!}p^x q^{n-x} \tag{13}$$

A binomial distribution graph where the probability of success does not equal the probability of failure looks like below.

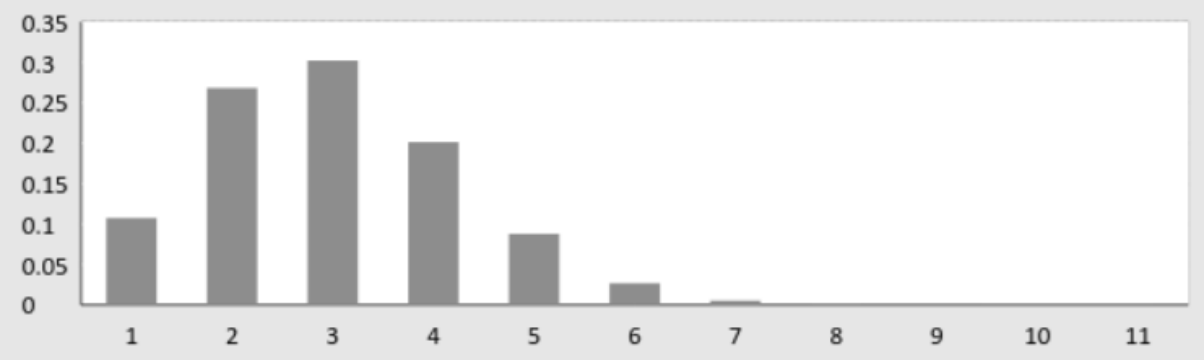


Figure 3: Binomial Distribution when P(success)!=P(failure)

Now, when probability of success = probability of failure, in such a situation the graph of binomial distribution looks like as in figure below.

The mean and variance of a binomial distribution are given by:

$$\begin{gathered}\text{Mean} -> \mu = n^\star p \\ \text{Variance} -> \text{Var}(X) = n^\star p^\star q\end{gathered} \tag{14}$$

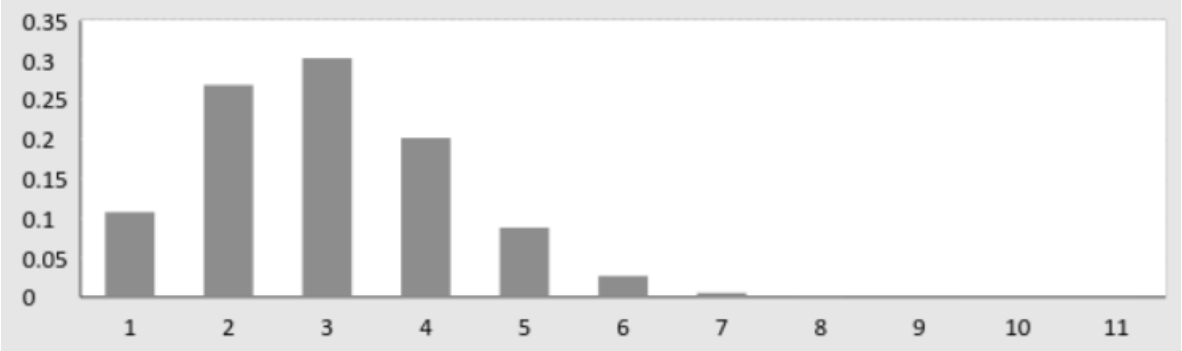


Figure 4: Binomial Distribution when P(success)=P(failure)

10.4 Normal Distribution

Normal distribution represents the behavior of most of the situations in the universe (That is why it's called a "normal" distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:

- The mean, median and mode of the distribution coincide.
- The curve of the distribution is bell-shaped and symmetrical about the line x=.
- The total area under the curve is 1.
- Exactly half of the values are to the left of the center and the other half to the right.
- A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity then the shapes will be quite similar.

The PDF of a random variable X following a normal distribution is given by:

$$\boldsymbol{f}(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}\sigma} \boldsymbol{e}^{\left\{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2\right\}} \quad \text{for } -\infty < x < \infty \tag{15}$$

The graph of a random variable X N (μ, σ) is shown below.

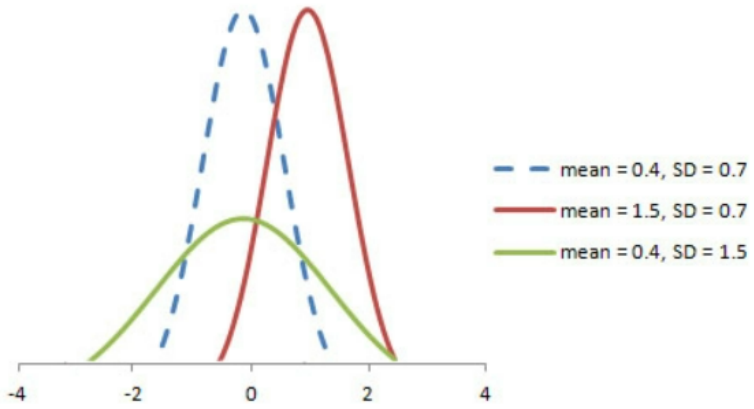


Figure 5: Normal Distribution

The mean and variance of a random variable X which is said to be normally distributed is given by: Mean - E(X) = μ Variance - Var(X) = σ^2 Here, μ (mean) and σ (standard deviation) are the parameters.

10.5 Poisson Distribution

A distribution is called Poisson distribution when the following assumptions are valid:

- Any successful event should not influence the outcome of another successful event.
- The probability of success over a short interval must equal the probability of success over a longer interval.
- The probability of success in an interval approaches zero as the interval becomes smaller.

Now, if any distribution validates the above assumptions then it is a Poisson distribution. Some notations used in Poisson distribution are:

λ is the rate at which an event occurs, t is the length of a time interval, And X is the number of events in that time interval. Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.

Let μ denote the mean number of events in an interval of length t. Then, μ = λ *t.

The PMF of X following a Poisson distribution is given by:

$$\boldsymbol{P}(\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{e}^{-\mu}\frac{\mu^x}{x!} \quad \text{for } x = 0, 1, 2, \ldots\ldots \tag{16}$$

The mean and variance of X following a Poisson distribution:

Mean: E(X) = μ

Variance: Var(X) = μ

The mean μ is the parameter of this distribution. μ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown below:

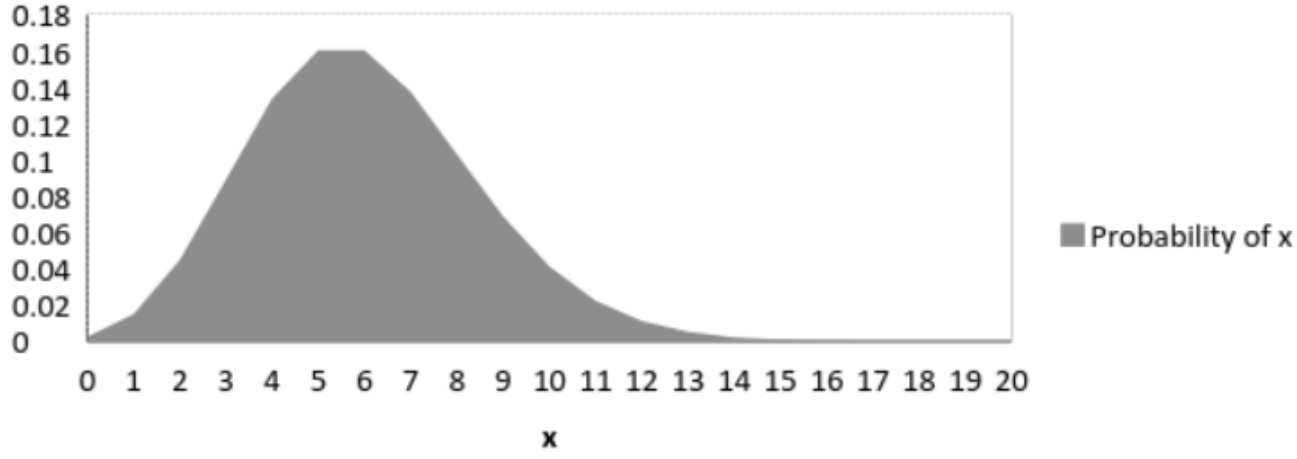


Figure 6: Poisson Distribution

References

[1] Class Lecture 12 : https://courses.iiit.ac.in/mod/forum/discuss.php?d=22997

[2] https://calcworkshop.com/continuous-probability-distribution/expected-value-variance-continuous-random-variable/

[3] https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

[4] https://stattrek.com/sampling/populations-and-samples.aspx

[5] http://vortex.ihrc.fiu.edu/MET4570/members/Lectures/Lect05/m10divideby_nminus1.pdf

[6] https://en.wikipedia.org/wiki/Covariance

[7] https://www.mathsisfun.com/search/search.html?query=covariancesubmit=search=1ff

[8] https://en.wikipedia.org/wiki/Covariance

[9] https://en.wikipedia.org/wiki/Correlation_and_dependence

[10] https://en.wikipedia.org/wiki/Standard_deviation

[11] https://en.wikipedia.org/wiki/Normal_distribution

[12] https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/