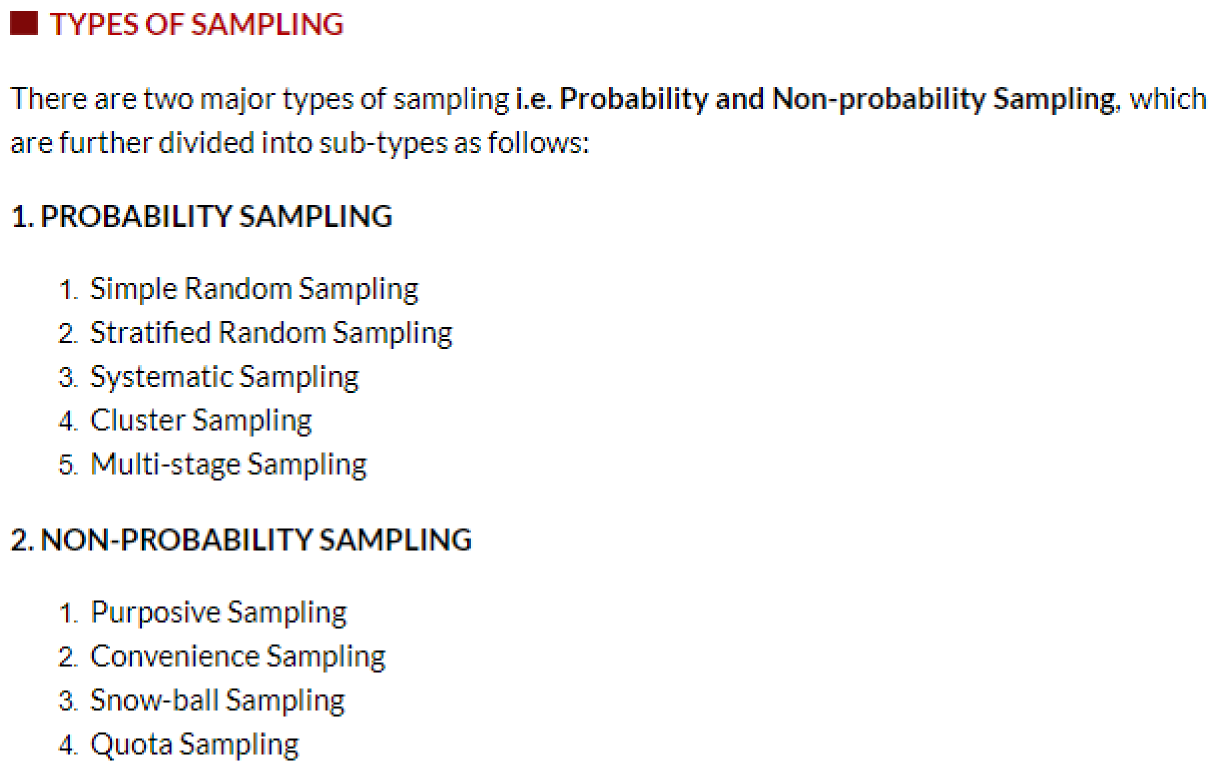
Range :

In **statistics**, the **range** of a set of data is the difference between the largest and smallest values. However, in descriptive **statistics**, this concept of**range** has a more complex meaning. The **range** is the size of the smallest interval which contains all the data and provides an indication of **statistical**dispersion.



■ TYPES OF SAMPLING

There are two major types of sampling **i.e. Probability and Non-probability Sampling**, which are further divided into sub-types as follows:

1. PROBABILITY SAMPLING

1. Simple Random Sampling
2. Stratified Random Sampling
3. Systematic Sampling
4. Cluster Sampling
5. Multi-stage Sampling

2. NON-PROBABILITY SAMPLING

1. Purposive Sampling
2. Convenience Sampling
3. Snow-ball Sampling
4. Quota Sampling

Descriptive statistics :are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics **analysis**, they form the basis of virtually every quantitative **analysis** of data. Descriptive statistics are typically distinguished from inferential statistics. With descriptive statistics you are simply describing what is or what the data shows. With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone

Univariate Analysis

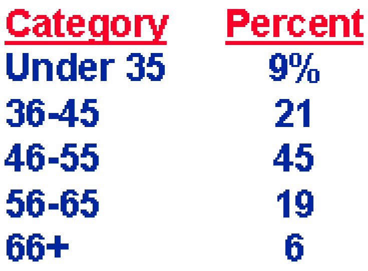
Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

- the distribution
- the central tendency

- the dispersion

In most situations, we would describe all three of these characteristics for each of the variables in our study.

The Distribution. The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of persons who had each value. For instance, a typical way to describe the distribution of college students is by year in college, listing the number or percent of students at each of the four years. Or, we describe gender by listing the number or percent of males and females. In these cases, the variable has few enough values that we can list each one and summarize how many sample cases had the value. But what do we do for a variable like income or GPA? With these variables there can be a large number of possible values, with relatively few people having each one. In this case, we group the raw scores into categories according to ranges of values. For instance, we might look at GPA according to the letter grade ranges. Or, we might group income into four or five ranges of income values.



Category	Percent
Under 35	9%
36-45	21
46-55	45
56-65	19
66+	6

Central Tendency. The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

Dispersion. Dispersion refers to the spread of the values around the central tendency. There are two common measures of dispersion, the range and the standard deviation. The **range** is simply the highest value minus the lowest value. In our example distribution, the high value is 36 and the low is 15, so the range is 36 - 15 = 21.

Chi square test : A chi-square statistic is one way to show a relationship between two categorical variables. In statistics, there are two types of variables: numerical (countable) variables and non-numerical (categorical) variables. The chi-squared statistic is a single number that tells you how much difference exists between your observed counts and the counts you would expect if there were no relationship at all in the population.

What is a Chi Square Test?

There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:

- A chi-square goodness of fit test determines if a sample data matches a population. For more details on this type, see: Goodness of Fit Test.
- A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
  - A very small chi square test statistic means that your observed data fits your expected data extremely well. In other words, there is a relationship.
  - A very large chi square test statistic means that the data does not fit very well. In other words, there isn’t a relationship.

Back to Top

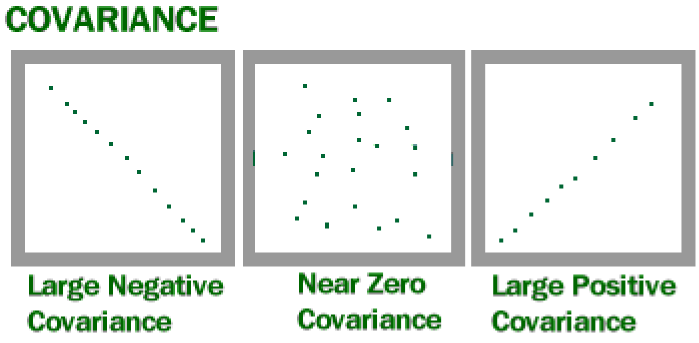
What is a Chi-Square Statistic?

The formula for the chi-square statistic used in the chi square test is:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalizes the t-test to more than two groups. ANOVA is useful for comparing (testing) three or more group means for statistical significance.

ovariance is a measure of how much two random variables vary together. It’s similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.



Advantages of the Correlation Coefficient

The Correlation Coefficient has several advantages over covariance for determining strengths of relationships:

- Covariance can take on practically any number while a correlation is limited: -1 to +1.
- Because of it’s numerical limitations, correlation is more useful for determining how strong the relationship is between the two variables.
- Correlation does not have units. Covariance always has units
- Correlation isn’t affected by changes in the center (i.e. mean) or scale of the vari

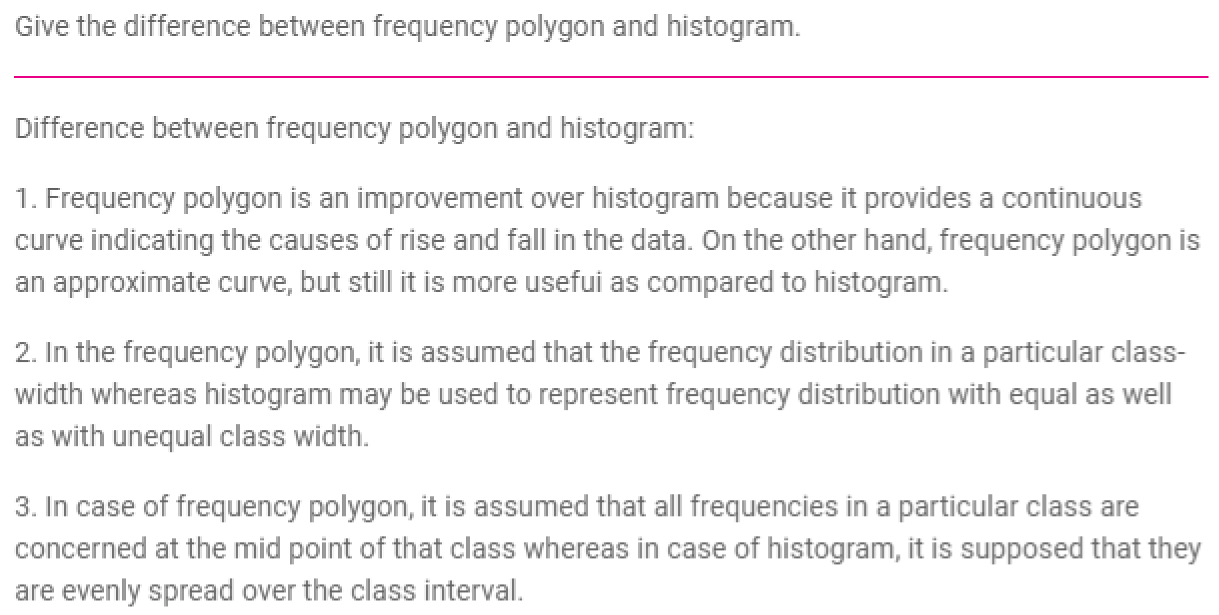
Standard deviation is a measure of dispersement in statistics. “Dispersement” tells you how much your data is spread out. Specifically, it shows you how much your data is spread out around the mean or average. For example, are all your scores close to the average? Or are lots of scores way above (or way below) the average score?

What is the Standard Deviation Symbol?
Which symbol you use depends on if you have a sample or a population:

- The symbol for a sample is s.
- The symbol for a population is σ

Formula of SD:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$



Give the difference between frequency polygon and histogram.

Difference between frequency polygon and histogram:

1. Frequency polygon is an improvement over histogram because it provides a continuous curve indicating the causes of rise and fall in the data. On the other hand, frequency polygon is an approximate curve, but still it is more usefui as compared to histogram.

2. In the frequency polygon, it is assumed that the frequency distribution in a particular class-width whereas histogram may be used to represent frequency distribution with equal as well as with unequal class width.

3. In case of frequency polygon, it is assumed that all frequencies in a particular class are concerned at the mid point of that class whereas in case of histogram, it is supposed that they are evenly spread over the class interval.

Application of chi squared test:

In cryptanalysis, chi-squared test is used to compare the distribution of plaintext and (possibly) decrypted ciphertext. The lowest value of the test means that the decryption was successful with high probability.[10][11] This method can be generalized for solving modern cryptographic problems.[12]

In bioinformatics, chi-squared test is used to compare the distribution of certain property of genes (e.g., genomic content, mutation rate, interaction network clustering, etc.) belonging different categories (e.g., disease genes, essential genes, genes on a certain chromosome etc.).

Advantages and Disadvantages
Each probability sampling method has its own unique advantages and disadvantages.

Advantages

- Cluster sampling: convenience and ease of use.
- Simple random sampling: creates samples that are highly representative of the population.
- Stratified random sampling: creates strata or layers that are highly representative of strata or layers in the population.
- Systematic sampling: creates samples that are highly representative of the population, without the need for a random number generator.

Disadvantages

- Cluster sampling: might not work well if unit members are not homogeneous (i.e. if they are different from each other).

- Simple random sampling: tedious and time consuming, especially when creating larger samples.
- Stratified random sampling: tedious and time consuming, especially when creating larger samples.
- Systematic sampling: not as random as simple random sampling,

What is a Population in Statistics?

In stats, a sample is a part of a population. A population is a whole, it's every member of a group. A population is the opposite to a sample, which is a fraction or percentage of a group. Sometimes it's possible to survey every member of a group. A classic example is the U.S. Census, where it's the law that you have to respond. Note: if you do manage to survey everyone, it actually is called a census: The U.S. Census is just one example of a census.
In most cases, it's impractical to survey everyone.
Imagine how long it would take you to call every dog owner in the U.S. to find out what their preferred brand of dog food was. In addition, sometimes people either don't want to respond or forget to respond, leading to incomplete censuses. Incomplete censuses become samples by definition.

Sample vs. Population Example

Sample
Population

If you go into a candy store, the owner might have samples of their products on display. It wouldn't be possible for you to sample everything in the store; Financially the owner wouldn't want you to taste everything for free. And you probably wouldn't want to eat a sample of candy from a couple hundred jars or you might get sick to your stomach. So, you might base your opinion about the entire store's candy line based on the samples they have to offer. The same logic holds true for most surveys in stats; You're only going to want to take a sample of the whole population ("population" in this example would be the entire candy line). The result is a statistic about that population.

MATLAB (matrix laboratory) is a multi-paradigm numerical computingenvironment and proprietary programming language developed by MathWorks. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, C#, Java, Fortran and Python.

Although MATLAB is intended primarily for numerical computing, an optional toolbox uses the MuPAD symbolic engine, allowing access to symbolic computingabilities. An additional package, Simulink, adds graphical multi-domain simulation and model-based design for dynamic and embedded systems.

As of 2018, MATLAB has more than 3 million users worldwide.[7] MATLAB users come from various backgrounds of engineering, science, and economics.

A Poisson distribution is a tool that helps to predict the probability of certain events from happening when you know how often the event has occurred. It gives us the probability of a given number of events happening in a fixed interval of time.

A textbook store rents an average of 200 books every Saturday night. Using this data, you can predict the probability that more books will sell (perhaps 300 or 400) on the following Saturday nights. Another example is the number of diners in a certain restaurant every day. If the average number of diners for seven days is 500, you can predict the probability of a certain day having more customers.

Because of this application, Poisson distributions are used by businessmen to make forecasts about the number of customers or sales on certain days or seasons of the year. In business, overstocking will sometimes mean losses if the goods are not sold.

What is Bayes’ Theorem?

Bayes’ theorem is a way to figure out conditional probability. Conditional probability is the probability of an event happening, given that it has some relationship to one or more other events. For example, your probability of getting a parking space is connected to the time of day you park, where you park, and what conventions are going on at any time. Bayes’ theorem is slightly more nuanced. In a nutshell, it gives you the actual probability of an event given information about tests.

- “Events” Are different from “tests.” For example, there is a test for liver disease, but that’s separate from the event of actually having liver disease.
- Tests are flawed: just because you have a positive test does not mean you actually have the disease. Many tests have a high false positive rate. Rare events tend to have higher false positive rates than more common events. We’re not just talking about medical tests here. For example, spam filtering can have high false positive rates. Bayes’ theorem takes the test results and calculates your real probability that the test has identified the event.

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

In probability, and statistics, a multivariate random variable or random vector is a list of mathematical variables each of whose value is unknown, either because the value has not yet occurred or because there is imperfect knowledge of its value. The individual variables in a random vector are grouped together because they are all part of a single mathematical system — often they represent different properties of an individual statistical unit. For example, while a given person has a specific age, height and weight, the representation of these features of an unspecified person from within a group would be a random vector. Normally each element of a random vector is a real number.

Random vectors are often used as the underlying implementation of various types of aggregate random variables, e.g. a random matrix, random tree, random sequence, stochastic process, etc.

Interface ghh::

A central challenge in machine learning is to quantify uncertainty via probabilistic models that capture statistical dependencies between many uncertainty quantities. We have recently initiated the study of learning and inference in Probabilistic Submodular Models, a rich class of probabilistic models defined through submodular functions. This class contains and generalizes a number of extensively studied subclasses, such as determinantal point processes, and Ising models. A major benefit of such models is that they allow to capture complex, long-range interactions between many variables, which is useful, for example in computer vision and information retrieval. We develop novel algorithms for efficient approximate inference, using, e.g., variational or sampling techniques, as well as learning such models from data.