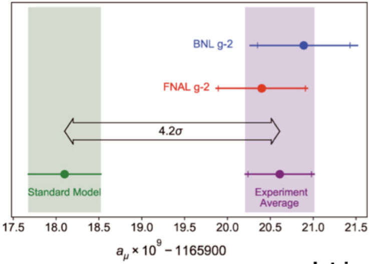
Statistical Methods (I)

Sadhana Dash

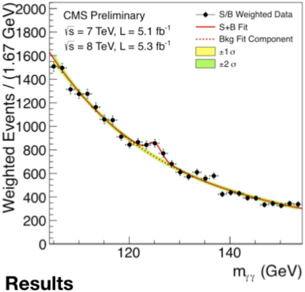
November 5, 2022

Our primary goal

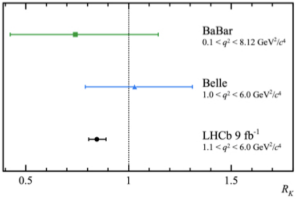
Experimental Measurements



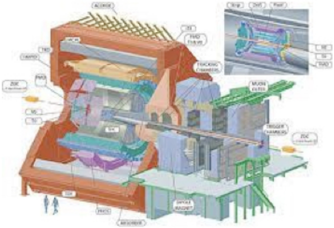
Discovery

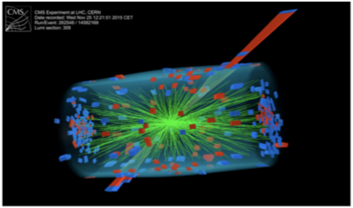


Intriguing Results

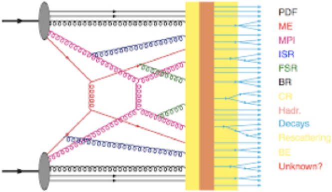


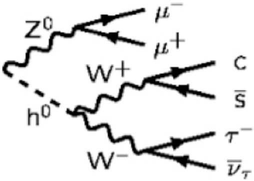
Initial ingredients





Guidance by existing theory





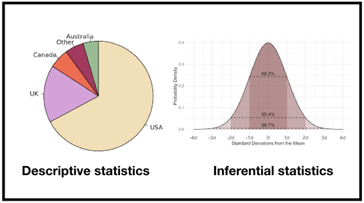
Outline

Statistics : Definition

The art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.

Two Types :

1. Descriptive Statistics : The part of statistics, concerned with the description and summarization of data, is called descriptive statistics.
2. Inferential Statistics : The part of statistics, concerned with the drawing of conclusions, is called inferential statistics.



Population and Sample

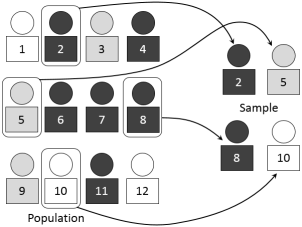


Figure: A schematic of population and sample

Population is the entire set of possible cases for which a study is conducted.
Sample refers to a subgroup of population from which we try to learn about the population.
A measure concerning a population is called **parameter** while that of a sample is called a **statistic**.

Sample Spaces and Events

Random Experiment : An experiment that can result in different outcomes, even though it is repeated in similar manner every time, is called a random experiment.
Sample Space : The set of all possible outcomes of a random experiment is called the sample space of the experiment. The sample space is denoted as S
Event : An event is a subset of the sample space of a random experiment .

Why probability ?

Probability is used to quantify the likelihood, or chance, that an outcome of a random experiment will occur. The likelihood of an outcome is quantified by assigning a number from 0 to 1 to the outcome (or a percentage from 0 to 100%). The higher numbers indicate that the outcome is more likely than lower numbers.

Suppose we perform a random experiment, the sample space consists of all possible outcomes of the random experiment. An event is a subset of the sample space.

How do we assign probability to the occurrence of an event ?

- Subjective Approach
- Frequency Approach
- Bayesian Approach

Axioms of Probability

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties: If S is the sample space and E is any event in a random experiment,

- P(S) = 1
- $0 \leq P(E) \leq 1$

- If two events E_1 and E_2, which have no outcomes in common , $P(E_1 U E_2) = P(E_1) + P(E_2)$

Random Variable : Definition

A **Random Variable** is a variable that associates a number with the outcome of a random experiment.
A Random variable is a function that assigns a real number to each outcome in the sample space of a random experiment.
They are generally denoted by upper case alphabets **X or Y** to distinguish from algebraic variables.
After an experiment is conducted, the measured value of the random variable is denoted by a lower case letter.
The **range** of a random variable X, shown by Range(X) or R_X, is the set of possible values of X.

Type of Random Variables :

Discrete Random Variables

Random variables whose set of possible values can be written either as a finite sequence $x_1, ..., x_n$ or as an infinite sequence $x_1, ...$ are said to be discrete.

A discrete random variable is a random variable with a finite (or countably infinite) range.

For instance, a random variable whose set of possible values is the set of non-negative integers is a discrete random variable.

Example: outcome of coin toss or random dice experiments, number of persons affected with covid-19 in a year etc.

Continuous Random Variables

Random variables that take on a continuum of possible values are known as continuous random variables.

A continuous random variable is a random variable with an interval (either finite or infinite) of real numbers for its range.

We can think of lifetime of a laboratory instrument, when the lifetime is assumed to take on any value in some interval (a, b).

Probability Distributions

The probability distribution of a random variable X is a description of the probabilities associated with the possible values of X.

Probability Mass Function (PMF) For a discrete random variable **X** with possible values $x_1, x_2, ..x_n$, a **probability mass function** is a function such that

- $f(x_i) \geq 0$
- $\sum_{i=0}^{\infty} f(x_i) = 1$
- $f(x_i) = P(X = x_i)$

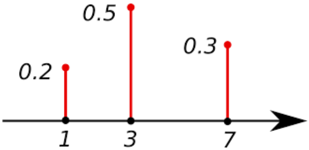


Figure: PMF of a discrete random variable

Cumulative Distribution Function (CDF)

The cumulative distribution function (or simply the distribution function) of a discrete random variable X, denoted as F(x), is the probability that the random variable X takes on a value that is less than or equal to x

$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$

For a discrete random variable X, F(x) satisfies the following properties.

- $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$
- $0 \leq F(x) \leq 1$
- If x $\leq$ y, then F(x) $\leq$ F(y)

Probability Density Function (PDF)

For a continuous random variable X, a probability density function is a function such that

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx =$ area under $f(x)$ from a and b

  It also follows

  $P(X = a) = \int_{a}^{a} f(x)dx = 0$

  The probability that a continuous random variable will assume any particular value is **zero**.

  If X is a continuous random variable, for any x_1 and x_2,

  $P(x_1 \leq X \leq x_2) = P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X < x_2)$

The cumulative distribution function (cdf) of a continuous random variable X is

$$F(a) = P(X < a) = \int_{-\infty}^{a} f(x)dx \text{ , for } -\infty < a < \infty$$

Differentiating both the sides:

$$\frac{d}{da}F(a) = f(a)$$

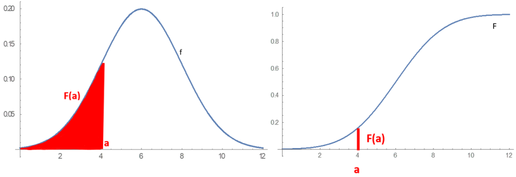


Figure: (Left panel) : PDF (Right panel) : CDF of a continuous random variable

$$P\left\{a - \epsilon/2 \leq X \leq a + \epsilon/2\right\} = \int_{a-\epsilon/2}^{a+\epsilon/2} f(x)dx \sim \epsilon f(a)$$ when ϵ is small. In other words, the probability that X will be contained in an interval of length ϵ around the point a is approximately $\epsilon f(a)$.

Expectation Value

If X is a discrete random variable taking on the possible values $x_1, x_2, ..,$ the mean or expected value of the discrete random variable X, denoted as μ or E[X], is

$$\mu = E[X] = \sum_i x_i f(x_i)$$

Thus, the expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability associated with X.

If X is a continuous random variable with the probability density function $f(x)$, the mean or expected value of X, denoted as E[X], is given by

$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Variance

Variance quantifies the variation, or spread, of the values associated with the random variable X. It measures the dispersion, or variability in the distribution.

If X is a discrete random variable with mean μ, then the variance of X , denoted by Var(X) or σ^2, is defined by

$$Var(X) = \sigma^2 = E[(X-\mu)^2]$$

$$=\sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2$$

If X is a continuous random variable with probability density function f(x), V(X) or σ^2 is defined as

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

The standard deviation, σ is

$$\sigma = \sqrt{\sigma^2}$$

It can be shown

$$Var(X) = E[X^2] - \mu^2$$

If a and b are constants, $Var(aX+b) = a^2 Var(X)$ and. $Var(b) = 0$

Covariance

The covariance between the random variables X and Y, denoted as $cov(X, Y)$ or σ_{XY}, is

$\sigma_{XY} = Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

$=$ E[XY] - $\mu_x E[Y] - \mu_y E[X] + \mu_x \mu_y = E[XY] - E[X]E[Y]$

$\sigma_{XY} = \sigma_{YX}$

$\sigma_{XX} = Var(X)$

Properties :

- Cov(X,X)=Var(X)
- Cov(X,Y) = Cov(Y,X)
- Cov(aX,Y) = aCov(X,Y)
- Cov(X+c,Y) = Cov(X,Y)
- Cov(X+Y, Z) = Cov(X,Z) + Cov(Y,Z)

  Cov(X+Y, Z)= E[(X+Y)Z] - E(X+Y)E[Z]

  =E[XZ + YZ] - (E[X]+E[Y])E[Z]

  =E[XZ] - E[X]E[Z] + E[YZ] - E[Y]E[Z]

  = Cov(X,Z) + Cov(Y,Z)

We can show that

$$Cov\left(\sum_{i=1}^{n} X_i, Y\right) = \sum_{i=1}^{n} Cov(X_i, Y)$$

$$Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$$

Variance of sum of random variables

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} Cov(X_i, X_j)$$

if X and Y are independent, then Cov(X,Y)=0

As they are independent, Cov(X.Y) = E[XY] - E[X]E[Y]

=E[X]E[Y] - E[X]E[Y] = 0

Cov(X,Y) = 0

Therefore, for independent variables $X_1 ... X_n$, $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$

Correlation Coefficient

The correlation coefficient, ρ_{XY} of two random variables X and Y is obtained by dividing the covariance by the product of standard deviations of X and Y .

$\rho_{XY} = Corr(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Properties of the correlation coefficient:

- $-1 \leq \rho_{X,Y} \leq 1$
- If $\rho_{X,Y} = 1$, then $Y = aX + b$, where $a > 0$
- If $\rho_{X,Y} = -1$, then $Y = aX + b$, where $a < 0$
- $\rho(aX + b, cY + d) = \rho(X, Y)$

If $\rho(X, Y) = 0$, we say that X and Y are uncorrelated.
If $\rho((X, Y) > 0$, we say that X and Y are positively correlated.
If $\rho(X, Y) < 0$, we say that X and Y are negatively correlated.

Moment Generating function

Moments: Let X be any random variable. The moments are the expected values of various powers of X.
E[X] = first moment
E[X^2] = second moment
E[X^k] = k^{th} moment
The moment generating function (MGF), $M_X(t)$ of a random variable X is defined for all the values of t (provided the expectation exists for some t in a neighbourhood of zero) by

$$M_X(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx} f(x) \text{ , if X is a discrete rv} \\ \int_{-\infty}^{+\infty} e^{tx} f(x) dx \text{ , if X is continuous rv} \end{cases}$$

Finding Moments

One can write

$$e^{tX} = \sum_{k=0}^{\infty} \frac{(tX)^k}{k!} = \sum_{k=0}^{\infty} \frac{X^k t^k}{k!}$$

$$M_X(t) = E[e^{tX}] = \sum_{k=0}^{\infty} \frac{E[X^k] t^k}{k!}$$

$$E[X^k] = \frac{d^k}{dt^k} M_X(t)|_{t=0}$$

Properties

(1)If X is a random variable and a is a constant, then

(i) $M_{X+a}(t) = e^{at} M_X(t)$

(ii)$M_{aX}(t) = M_X(at)$

(2) The moment generating function of the sum of independent random variables is just the product of the individual moment generating functions.

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$$

(3) The moment generating function uniquely determines the distribution.

Summarizing Data Sets

Sample Mean :
The sample mean characterizes the central tendency in the data by the arithmetic average.
Let us consider a sample of size n with numerical values, $x_1, x_2, x_3 ... x_n$.
The arithmetic average of the values is called sample mean.

$$\mathbf{\bar{x}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i}}{\mathbf{n}} \tag{1}$$

If a data set is presented in a frequency table where the k distinct values, $v_1, ..., v_k$, having corresponding frequencies $f_1, ..., f_k$ such that $n = \sum_{i=1}^{k} f_i$, the sample mean of these n data values is

$$\mathbf{\bar{x}} = \frac{\sum\limits_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{v_i f_i}}{\mathbf{n}} \tag{2}$$

Therefore, the sample mean is a weighted average of the distinct values, where the weight given to the value v_i is equal to the proportion of the n data values that are equal to v_i , i = 1, . . . , k.

Sample median

Sample median is the middle value when the data set is arranged in increasing order.
Determination of median :
Order the values of a data set of size n from smallest to largest.
If n is odd, the sample median is the value in position $(n + 1)/2$
if n is even, it is the average of the values in positions $n/2$ and $n/2 + 1$.
Sample mode
The value that occurs with the greatest frequency.

Sample Variance

This statistic describes the variability or scatter in the data.
The sample variance , s^2 (of a data set $x_1, x_2, x_3 ... x_n$) is given by
,

$$\mathbf{s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}} \tag{3}$$

The sample standard deviation is given by

$$\mathbf{s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}}} \tag{4}$$

Show that
$\sum_{i=1}^{n}(x_i - \overline{x})^2 = (\sum_{i=1}^{n} x_i^2) - n\overline{x}^2$ Note: While calculating variance for population, one divides by N, size of population .

Sample Correlation Coefficient

If s_x and s_y denote, respectively, the sample standard deviations of the x values and the y values in a paired data set, the sample correlation coefficient, r , of the data pairs (x_i, y_i) , i = 1, . . . , n is defined by

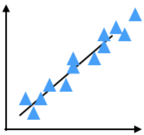
$$\mathbf{r} = \frac{\sum_{i=1}^{n}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{y_i} - \bar{\mathbf{y}})}{(\mathbf{n-1})\mathbf{s_x s_y}} \tag{5}$$

When $r > 0$, the sample data pairs are positively correlated, and when $r < 0$, they are negatively correlated.

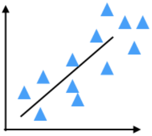
Properties of r

1. $-1 \leq r \leq 1$
2. if $y_i = a + bx_i$, for i = 1,2,.. n then $r = 1$ (for $b > 0$) and $r = -1$ (for $b < 0$) where a and b are constants.
3. If r is the sample correlation coefficient for the data pairs (x_i, y_i) , i= 1,..., n then it is also the sample correlation coefficient for the data pairs $a + bx_i, c + dy_i$, i=1,...,n provided that b and d are both positive or both negative constants.

Sample Correlation Coefficient

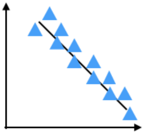


Strong positive correlation

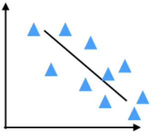


Moderate positive correlation

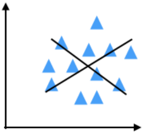
Guess the "r" values ?



Strong negative correlation



Moderate negative correlation



No correlation

Central moments

Central moments are moments about the mean, μ
μ_n = E[(X - μ)^n]
Raw moments
$\mu_n^{'}$ = E[(X)^n]
First central moment $= \mu_1 = 0$
Second central moment $= \mu_2 = \sigma^2 = \mu_2^{'} - \mu^2$
Third central moment $= \mu_3 = \mu_3^{'} - 3\mu_1^{'}\mu_2^{'} + 2\mu^3$
Fourth central moment $= \mu_4 = \mu_4^{'} - 4\mu\mu_3^{'} + 6\mu^2\mu_2^{'} - 3\mu^4$
The standardized moment of degree n is

$$\text{standardized moment} = \frac{\mu_n}{\sigma^n} \tag{6}$$

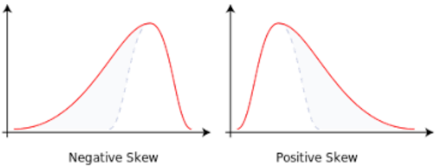
where σ is the standard deviation.

Skewness

The term skewness is a measure of lack of symmetry or departure from symmetry of the probability distribution function. If the distribution is not symmetrical (or is asymmetrical) about the mean, it is called a skewed distribution.

If the bulk of the data is at the left and the right tail is longer, the distribution is skewed right or positively skewed.

If the bulk of the data is towards the right and the left tail is longer, we say that the distribution is skewed left or negatively skewed.



Measure of Skewness

The skewness is quantified as the third standardized moment.

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

Pearson's coefficient of skewness

$\gamma_1 = \sqrt{\beta_1}$

Estimate the skewness of normal and exponential distribution.

Kurtosis

Kurtosis refers to the height and sharpness of the peak relative to a normal distribution. Sometimes, it is also interpreted as the tailedness of the pdf. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak.

The reference standard is a normal distribution, which has a kurtosis of 3. Therefore, excess kurtosis is defined as kurtosis - 3.

A normal distribution has kurtosis exactly 3 (excess kurtosis $=0$).

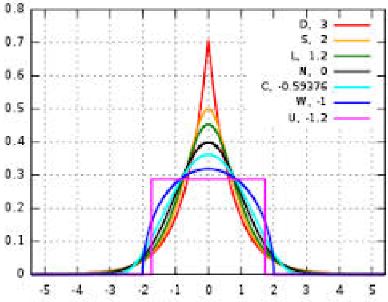
Any distribution with kurtosis $= 3$ (excess kurtosis $= 0$) is called mesokurtic.

A distribution with $kurtosis < 3$ (excess kurtosis < 0) is called platykurtic.

A distribution with $kurtosis > 3$ (excess kurtosis > 0) is called leptokurtic.

Measure of Kurtosis

$\beta_2 = \frac{\mu_4}{\mu_2^2}$
Excess Kurtosis
$\gamma_2 = \beta_2 - 3$



Estimate the kurtosis of a normal distribution

Bernoulli Random Variable : Definition

A Bernoulli trial (named after the 17th century Swiss mathematician Jacob Bernoulli), is an experiment that can result in two outcomes, which can be denoted as a success and as a failure. If we let $X = 1$ when the outcome is a success and $X = 0$ when it is a failure, then the probability mass function of X is given by

$P(X = 0) = 1 - p$

$P(X = 1) = p$

where p, $0 < p < 1$, is the probability that the trial is a success.

A random variable X is said to be a Bernoulli random variable if its probability mass function is given by above equation for some $p \in (0, 1)$.

The expectation of a Bernoulli random variable is the probability that the random variable equals 1.

$E[X] = 1P(X = 1) + 0P(X = 0) = p$

The probability of a success is denoted p, and the probability of failure is therefore 1 - p.

Binomial Random Variable : Definition

A random experiment consists of n Bernoulli trials such that
(1) The trials are independent
(2) Each trial results in only two possible outcomes, labeled as success and failure
(3) The probability of a success in each trial, denoted as p, remains constant
The random variable X that equals the number of trials that result in a success has a binomial random variable with parameters $0 < p < 1$ and $n = 1, 2, ...$ The probability mass function of X is
$f(x) = \binom{n}{x} \cdot p^x(1-p)^{n-x}$, x = 0,1,2, ...n

Properties of Binomial Random Variable

(1) $f(x) > 0$

(2) $\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = [p + (1-p)]^n = 1$

If X is a binomial random variable with parameters p and n,

$E(X) = np$

We know that X is the sum of n identical Bernoulli random variables, each with expected value p.

$X = X_1 + \cdots + X_n$

From the linearity of the expected values

$\mathrm{E}[\ \mathrm{X}\ ] = \mathrm{E}\ [\mathrm{X}_1 + X_2 \cdots + X_n]$

$= \mathrm{E}\ [\mathrm{X}_1] + \cdots + E[X_n]$

$= \mathrm{p} + \cdots + p = np$

Similarly we can show that

Var(X) = np(1-p)

(the variance of a sum of independent random variables is the sum of the variances.)

Poisson Distribution

Given an interval of real numbers, assume counts occur at random throughout the interval. If the interval can be partitioned into subintervals of small enough length such that
(1) the probability of more than one count in a subinterval is zero,
(2) the probability of one count in a subinterval is the same for all subintervals and proportional to the length of the subinterval, and
(3) the count in each subinterval is independent of other subintervals, the random experiment is called a Poisson process.
The random variable X that equals the number of counts in the interval is a Poisson random variable with parameter $\lambda > 0$, and the probability mass function of X is
$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$, $x = 0, 1, 2, ..$

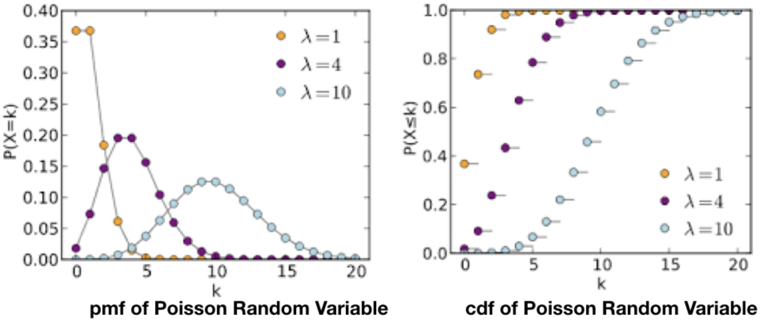


Figure: A schematic of pdf and cdf of Poisson distribution

Normal Random Variable

A random variable X is normal random variable if the pdf is given by
$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, $-\infty < x < \infty$
with parameters μ ($-\infty < \mu < \infty$) and $\sigma > 0$.

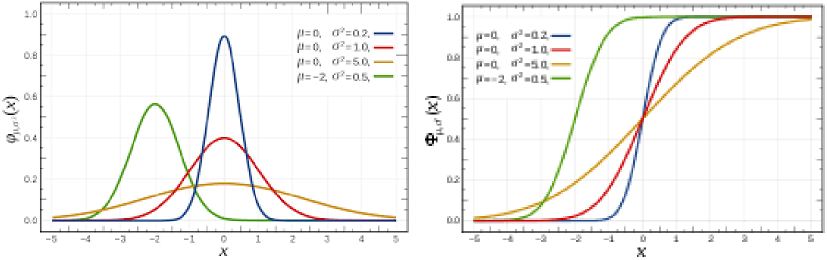


Figure: A schematic of pdf and cdf of normal continuous distribution

Mean and Variance

$$M_X(t) = exp(\mu t + \frac{t^2\sigma^2}{2}) \tag{7}$$

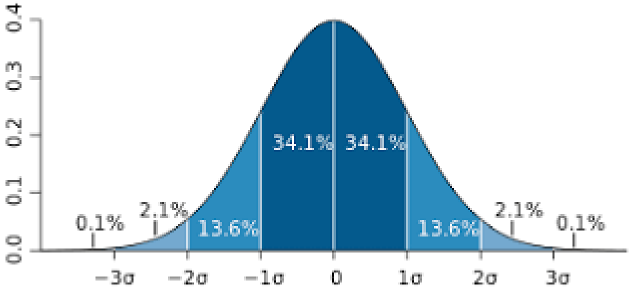
Differentiating w.r.t t

$$M_X^{'}(t) = (\mu + t\sigma^2)exp(\mu t + \frac{t^2\sigma^2}{2}) \tag{8}$$

$$M_X^{''}(t) = \sigma^2 exp(\mu t + \frac{t^2\sigma^2}{2}) + (\mu + t\sigma^2)^2 exp(\mu t + \frac{t^2\sigma^2}{2}) \tag{9}$$

$E[X] = M_X^{'}(0) = \mu$
$E[X^2] = M_X^{''}(0) = \sigma^2 + \mu^2$
$Var[X] = E[X^2] - (E[X])^2 = \sigma^2$

For a normal distribution one can show that
approximately 68% of the data fall within one standard deviation of the mean
approximately 95% of the data fall within two standard deviations of the mean
approximately 99.7% of the data fall within three standard deviations of the mean



Standard Normal Random Variable

A normal random variable with $\mu = 0$ and $\sigma^2 = 1$ is called a **standard normal random variable** and is denoted by $Z \sim N(0, 1)$

$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$, $-\infty < z < \infty$

The cumulative distribution function of a standard normal random variable is denoted as $\Phi(x)$ is

$$\Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt$$

Properties

- $\lim_{x \to \infty} \Phi(x) = 1$
- $\lim_{x \to -\infty} \Phi(x) = 0$
- $\Phi(0) = 1/2$
- $\Phi(-x) = 1 - \Phi(x)$

The Chi-Square Distribution:

If Z_1, Z_2, ... , Z_n are independent standard normal random variables, the random variable Y defined as

Y$= Z_1^2 Z_2^2 + \cdots +$Z$_n^2$

is said to have a chi-square (χ_n^2) distribution with n degrees of freedom shown by $Y \sim \chi_n^2$

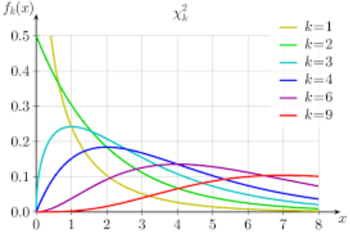


Figure: Schematic of χ_n^2 distribution for different degrees of freedom.

Properties:

If Y_1 and Y_2 are independent chi-square random variables with n_1 and n_2 degrees of freedom, respectively, then $Y_1 + Y_2$ is a chi-square random variable with $n_1 + n_2$ degrees of freedom.

If X is a chi-square random variable with n degrees of freedom, then for any $\alpha \in (0, 1)$ the quantity $\chi^2_{\alpha,n}$ is defined to be such that

$P(X \geq \chi^2_{\alpha,n}) = \alpha$