Basic Probability and Statistics

SHALA-2020

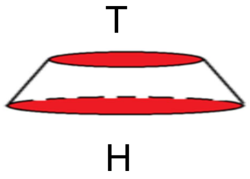
https://shala2020.github.io/

Learning Objectives

- Match PDF and PMF to outcomes of random experiments
- Write the relationship between PDF and CDF
- List common distributions and their parameters, means, and variances
- Write the condition for statistical independence of two variables
- Write the equation for the likelihood of a parametric distribution
- Write the steps for hypothesis testing

Random variables - discrete and continuous

- Say, we have an experiment with a random outcome (e.g. coin toss)
- We can map the set of outcomes to a variable that takes a unique value for each outcome (e.g. {0,1} for heads and tails)
- The value can change each time the experiment is conducted
- Such a variable is a random variable
- Random does not mean that all outcomes are equi-probable
- A biased coin can give more heads than tails, so somewhat predictable
- But, we still cannot predict each outcome with certainty



- Some random variables take continuous values
    - E.g. Height of the next person you will see on the road
    - Height measured in meters is a random variable
    - It still is random yet somewhat predictable; you won't see someone less than 1 foot, or more than 8 feet

Probability of a random variable taking a value

- Each outcome of an experiment is associated with a probability
- It is the expected proportion of the times you will see that outcome
- E.g. for a fair coin, prob(heads) = 0.5; or rather $p(X=0) = 0.5$
- For a biased coin, prob(heads) may be 0.53; prob(tails) = 0.47

- For a continuous variable, $\text{prob}(X=x) = 0$; prob(height = 1.6m *exactly*) is zero
- But, probability of an <u>interval</u> is finite; e.g. prob(1m < height ≤ 2.5m) = 0.66

Probability mass function (PMF) of discrete variables

- PMF maps discrete values of an RV to probabilities
- The probabilities sum up to 1
- E.g.
  - $p(X=1) = 0.2$,
  - $p(X=3) = 0.5$,
  - $p(X=7) = 0.3$,
  - $p(X \neq 1 \text{ and } X \neq 3 \text{ and } X \neq 7) = 0$

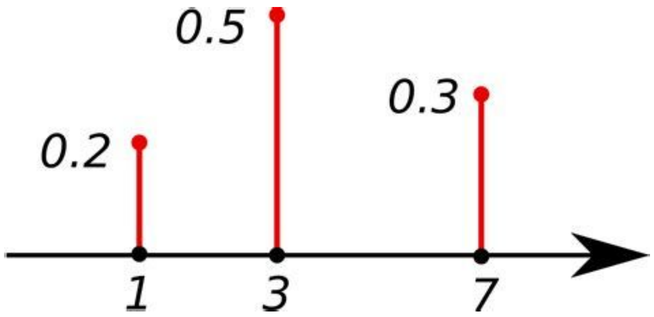


Image and example source: Wikipedia

Probability density function (PDF) of continuous variables

- Since a prob(X=x) = 0 for a continuous RV, we define a function for intervals
- p(x) is a function such that:

$$\int_{x_1}^{x_2} p(x)dx = \mathrm{prob}(x_1 < X \leq x_2) \qquad \int_{-\infty}^{+\infty} p(x)dx = 1$$

- Warning: Do <u>not</u> try to interpret the PDF at a single point!
- Always interpret it in relation to other points

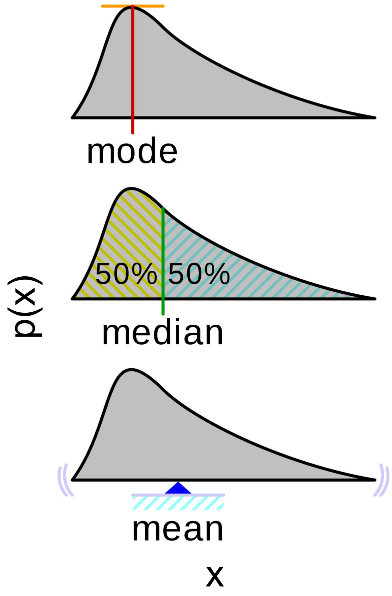
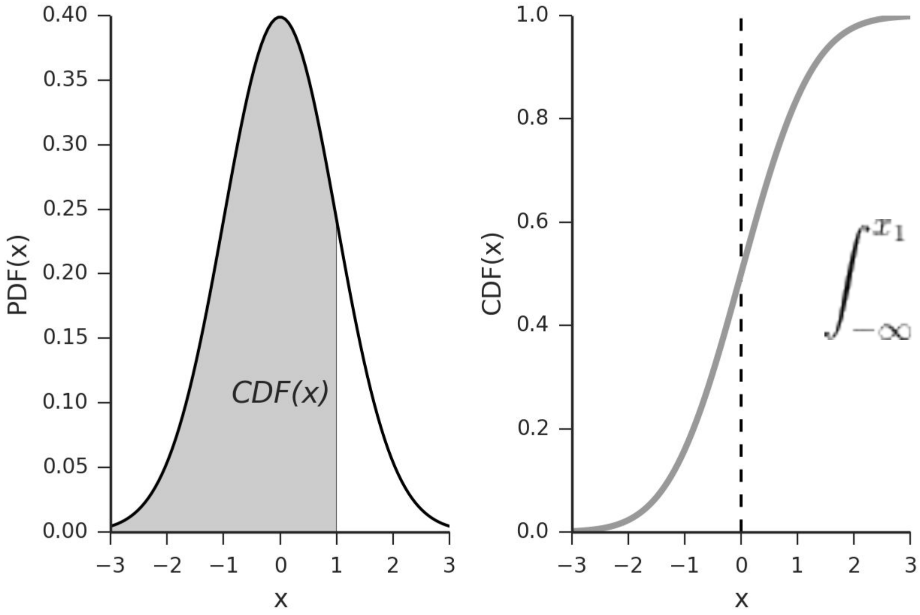


Image source: Wikipedia

Cumulative distribution function (CDF) of continuous variables



- CDF represents cumulative density

$$\int_{-\infty}^{x_1} p(x)dx = prob(X \leq x_1) = CDF(X = x_1)$$

Image source: deepai.org

Cumulative distribution function (CDF) of discrete RVs

- Consider the following event of a football match with the set of outcomes :

  S = {Win,Draw,Lose} and P(Win) = 0.7 , P(Draw) = 0.2 , P(Lose) = 0.1

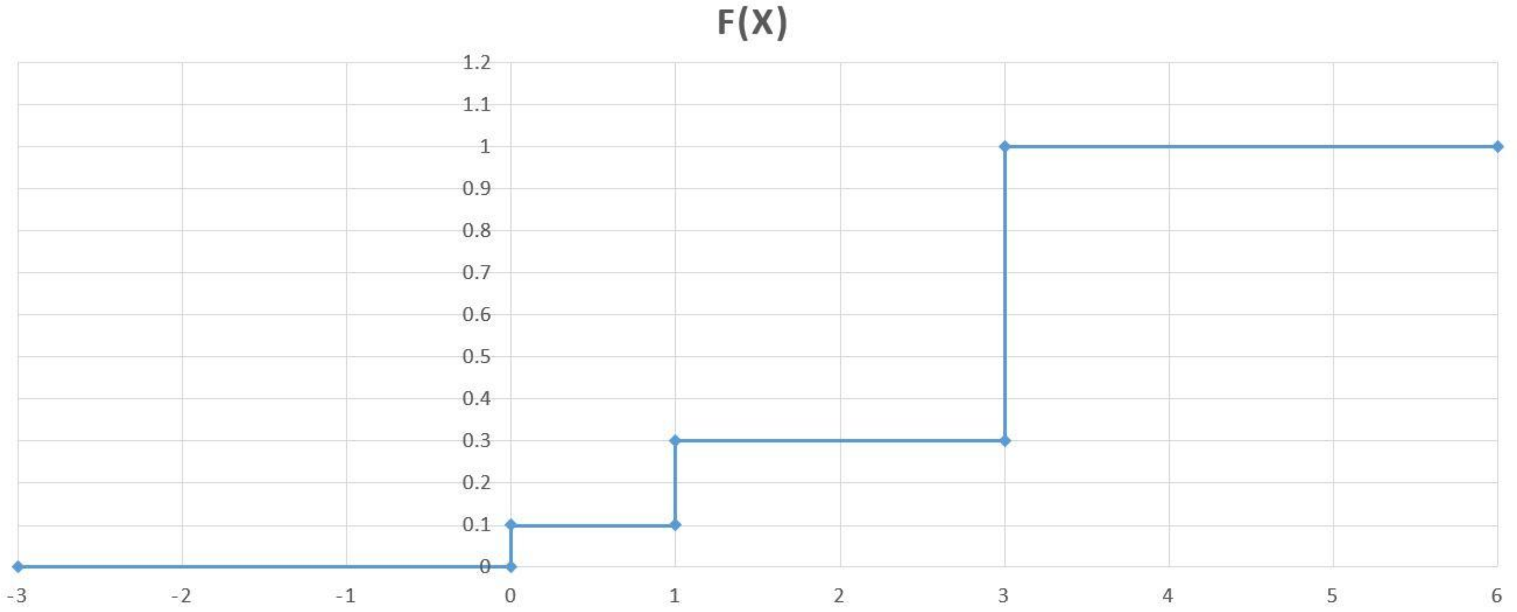
  Let the random variable X capture the points obtained in that match :

  X = { 0 : Lose, 1 : Draw, 3 : Win}

  Can you plot the distribution function for X?

  (CDF(X=x) as a function of x?)

Cumulative Distribution Function



CDF properties

- $0 \le CDF(x) \le 1$: CDF(X) is a probability value
- CDF(x) is non-decreasing : $prob(X \le x1) \le prob(X \le x1+x2)$ for $x2 \ge 0$
- $CDF(-\infty) = 0$, $CDF(+\infty) = 1$

Mean is expected value of a random variable

- Center of mass of the PDF

$$\int_{-\infty}^{\infty} x \; p(x) dx = \mu_x$$

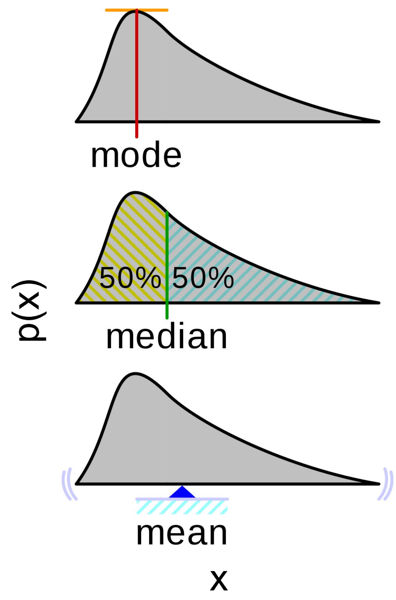


Image source: Wikipedia

Expected Value of the RV

Consider the example for the RV capturing points obtained after a football match. We had:

$$X = \begin{cases} 0 & , p_X(0) = 0.1 \\ 1 & , p_X(1) = 0.2 \\ 3 & , p_X(1) = 0.7 \end{cases}$$

Then E[X] = 0*0.1+1*0.2+3*0.7 = 2.3

Expected value of a function of an RV

- Expected value of any function of f(x) of x

$$\int_{-\infty}^{\infty} f(x)\ p(x)dx = \mathbb{E}_{x\sim p}[f(x)]$$

Expected value of a function of an RV

For the previous example of a football match points :

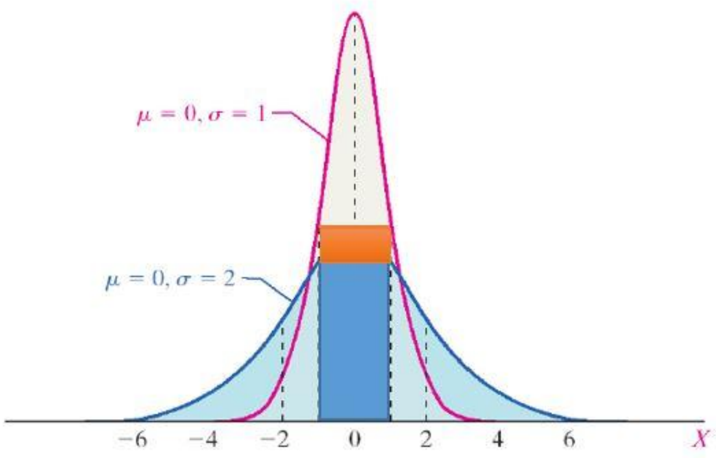
$$X = \begin{cases} 0 & , p_X(0) = 0.1 \\ 1 & , p_X(1) = 0.2 \\ 3 & , p_X(1) = 0.7 \end{cases}$$

If $Y = X^3$

E[Y] = 0*0.1 + 1*0.2 + 27*0.7 = 19.1

Variance as a measure of dispersion

Two RVs with the same mean can have very different distributions. Consider the pdfs of two RVs shown below :



Both are Gaussian distributions with the same mean. But the probability of them being in [-1,1] is very different with the red curve having larger area under the curve (the orange rectangle is the difference between areas). This corresponds to lesser variance of the pink curve

Image source : lumenlearning.com

Variance of a random variable

- Variance is the second central moment of an RV

$$\int_{-\infty}^{\infty} (x - \mu_x)^2 \ p(x)dx = \mathbb{E}_{x \sim p}[(x - \mu_x)^2]$$

- Similarly, we can define the n^{th} central moment (3 is skew, 4 is kurtosis)

$$\int_{-\infty}^{\infty} (x - \mu_x)^n \ p(x)dx = \mathbb{E}_{x \sim p}[(x - \mu_x)^n]$$

Variance of a random variable

- A well known form of variance is

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

So we can calculate the variance for the football match example as:

$E[X^2] = 0*0.1 + 1*0.2 + 9*0.7 = 6.5$; $(E[X])^2 = (2.3)^2 = 5.29$

$var(X) = 6.5 - 5.29 = 1.21$

Bernoulli

Bernoulli variables describe events with two outcomes. So the outcome of a coin toss can be modelled by a Bernoulli variable. A Bernoulli variable is described with a probability p of the outcome for X = 1.

$$X \in 0, 1$$

$$P(X) = \begin{cases} p & , X = 1 \\ 1 - p & , X = 0 \end{cases}$$

$$P(X = x) = p^x (1 - p)^{1-x}$$
$$E[X] = 1 \times p + 0 \times (1 - p) = p$$
$$var(X) = (1 - p)^2 \times p + p^2 \times (1 - p) = p(1 - p)$$

Binomial

A binomial variable can be seen as a sum of Bernoulli variables. A binomial RV is represented as $X \sim b(N, p)$ where N is the number of experiments performed and p is the probability of X = 1 for each outcome
eg. If I toss a coin 10 times, what is the probability of getting 6 heads?

$$\text{X} \in \{0, 1, ..., N\}$$

$$P(X) = \begin{cases} \binom{N}{x} p^x (1-p)^{N-x} & , x \in 0, 1, ..., N \\ 0 & \text{otherwise} \end{cases}$$

$$E[X] = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x} = np$$

$$var(X) = E[X^2] - (E[X])^2 = \sum_{x=0}^{N} x^2 \binom{N}{x} p^x (1-p)^{N-x} - (np)^2 = np(1-p)$$

Gaussian

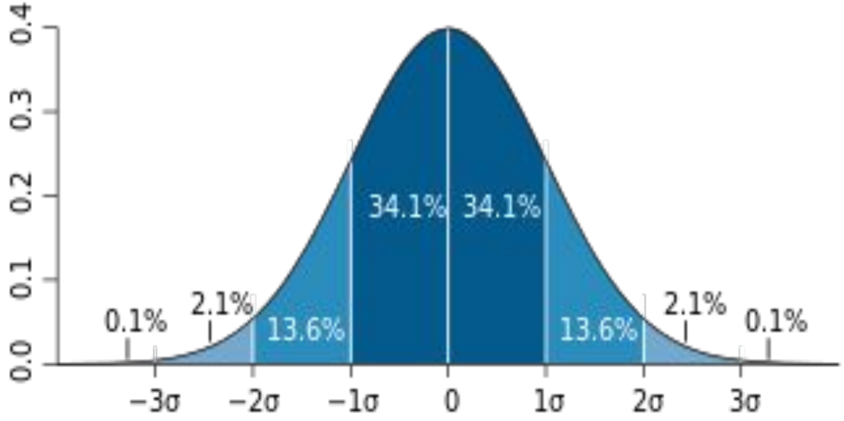
A Gaussian distribution is one of the most important and widely seen distributions. It explains the behaviour of many natural phenomena like the distribution of people's heights, shoe size etc. A Gaussian RV is described by the mean and variance, $X \sim N(\mu, \sigma^2)$.

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

$$E[X] = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

$$var(X) = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

Gaussian distribution



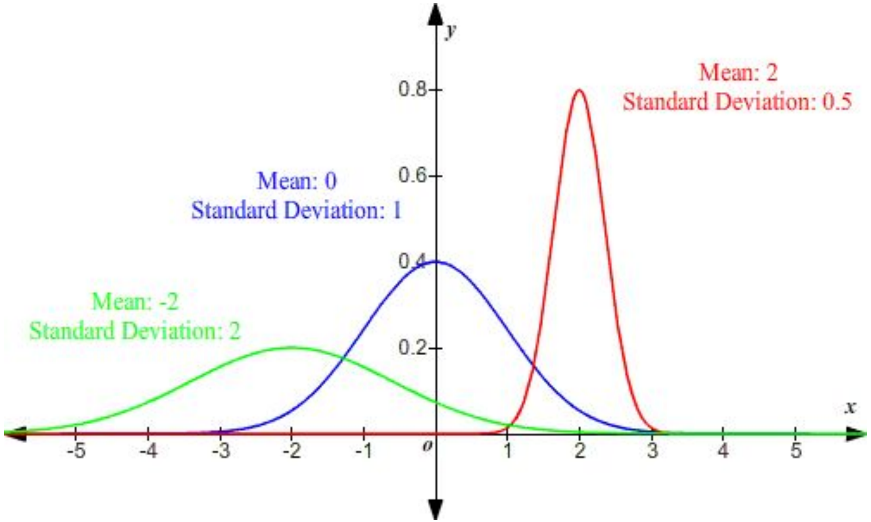


Image source : Wikipedia

Image source : varsitytutors.com

Exponential

An exponential distribution is generally used to model the time until an event occurs. It can model time until decay of a radioactive substance, time difference between the arrivals of two busses etc. An exponential RV is described as $X \sim Exp(\lambda)$.

$$f(X) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

$$E[X] = \int_0^\infty \lambda x e^{-\lambda x} dx = \frac{1}{\lambda}$$

$$var(X) = \int_0^\infty (x - \frac{1}{\lambda})^2 \lambda\, e^{-\lambda x} dx = \frac{1}{\lambda^2}$$

Exponential distribution

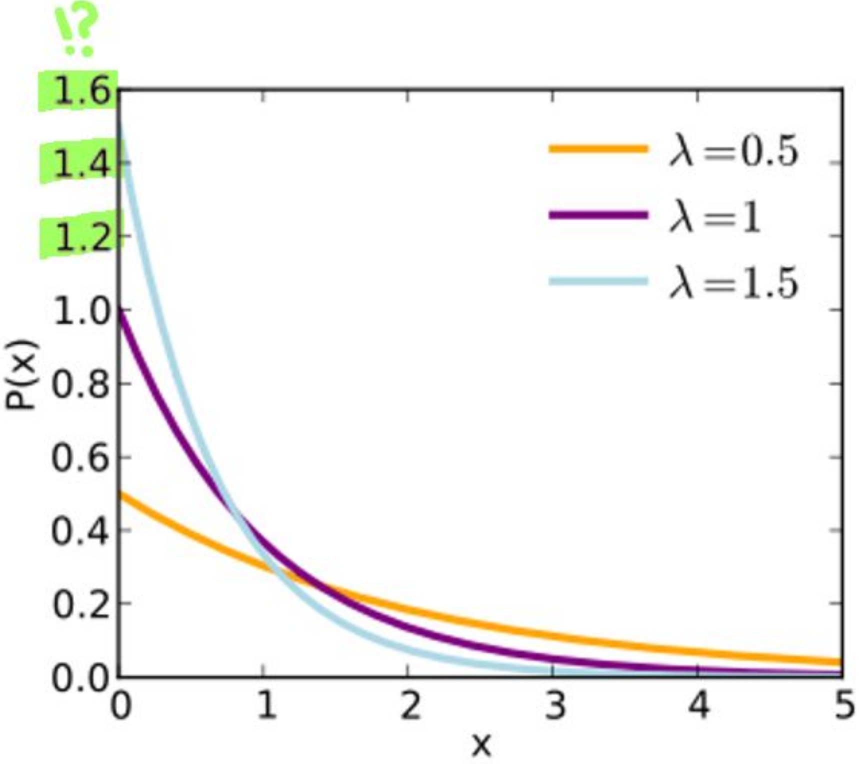


Image source : towardsdatascience.com

Gamma

A Gamma distribution is used to model aggregate insurance claims. A Gamma RV is described as $X \sim \Gamma(\alpha, \beta)$. Recall that the Gamma function is defined as

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx, \quad \text{and an important property is}$$

$$\Gamma(x+1) = x\Gamma(x)$$

The Gamma distribution is defined as

$$f(X) = \begin{cases} \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} & , x > 0 \\ 0 & , x \le 0 \end{cases}$$

$$E[X] = \int_0^\infty \frac{\beta^\alpha x^\alpha e^{-\beta x}}{\Gamma(\alpha)} dx$$

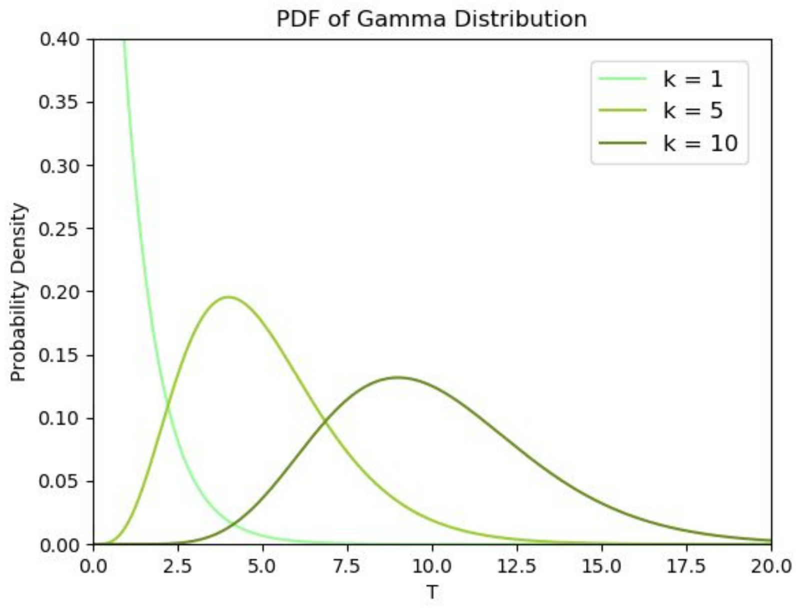
$$E[X] = \frac{\Gamma(\alpha+1)}{\beta\,\Gamma(\alpha)} \int_0^\infty \frac{\beta^{\alpha+1} x^\alpha e^{-\beta x}}{\Gamma(\alpha+1)} dx = \frac{\alpha}{\beta}$$

$$var(X) = E[X^2] - (E[X])^2$$

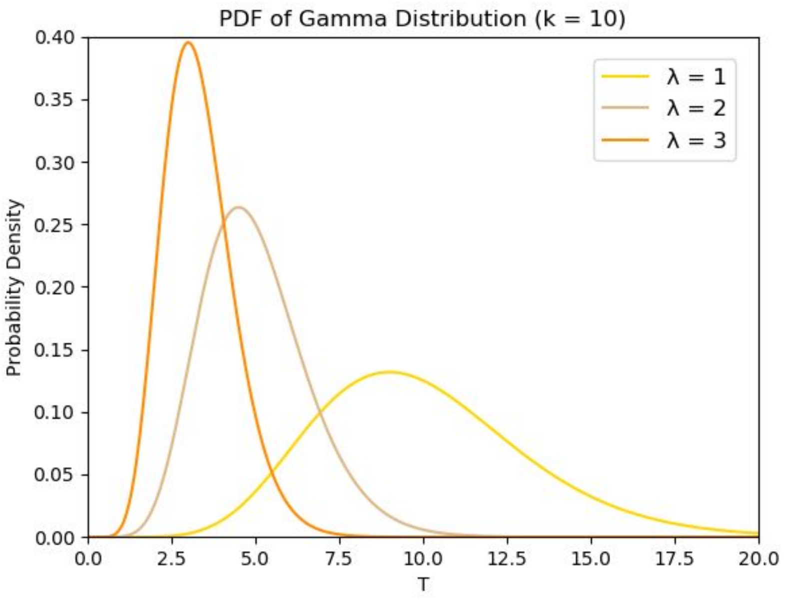
$$E[X^2] = \int_0^\infty \frac{\beta^\alpha x^{\alpha+1} e^{-\beta x}}{\Gamma(\alpha)} dx = \frac{\Gamma(\alpha+2)}{\beta^2\,\Gamma(\alpha)} \int_0^\infty \frac{\beta^{\alpha+2} x^{\alpha+1} e^{-\beta x}}{\Gamma(\alpha+2)} dx = \frac{(\alpha+1)\alpha}{\beta^2}$$

$$var(X) = E[X^2] - (E[X])^2 = \frac{\alpha}{\beta^2}$$

Gamma distributions



k is alpha



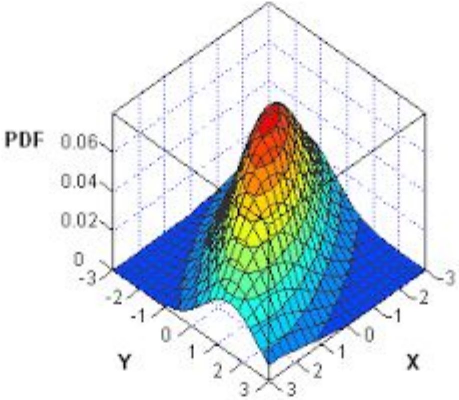
lambda is beta

image source : towardsdatascience.com

Multiple random variables and joint distribution

- Let there be two variables *x* and *y*
- The joint distribution is a surface parameterized by two variables *p(x,y)*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\, dx\ dy = 1$$



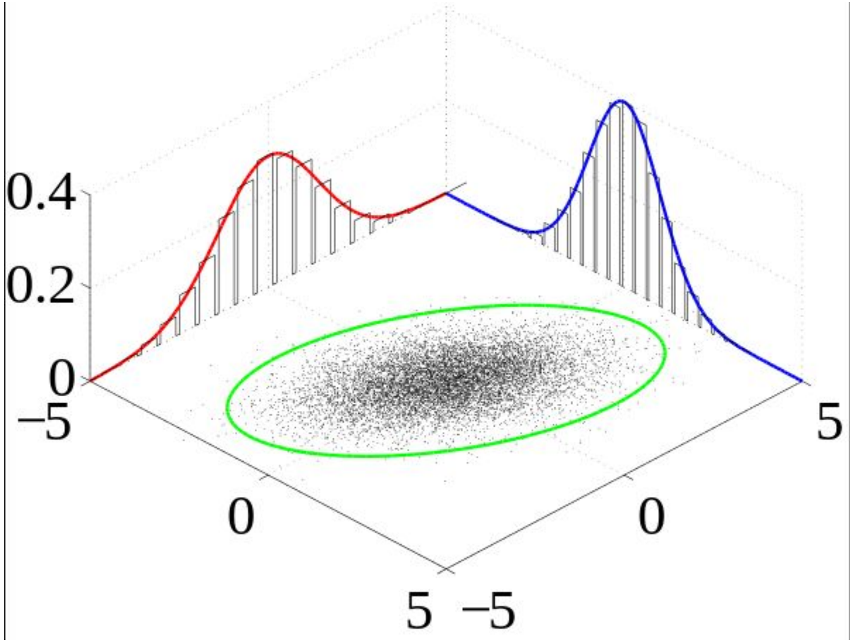
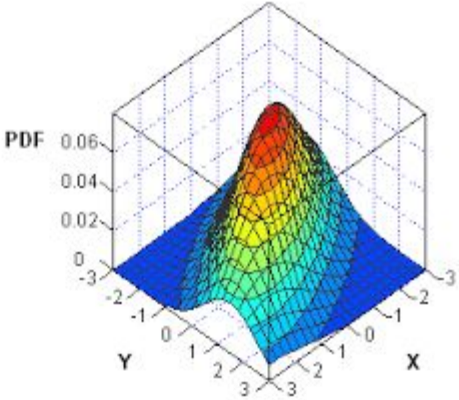


Image sources: sas.com, Wikipedia

Marginal probability eliminates one variable

- If we eliminate *y*, then we get a one-dim function *p(x)*,
- That is the marginal density function; a projection of the 2-d function onto 1-d

$$\int_{-\infty}^{\infty} p(x,y)\; dy = p(x)$$



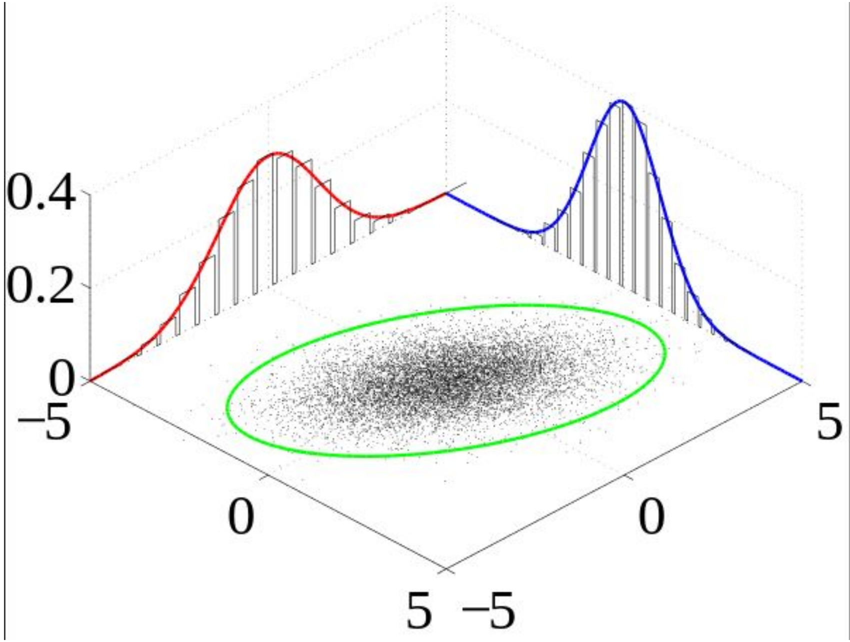
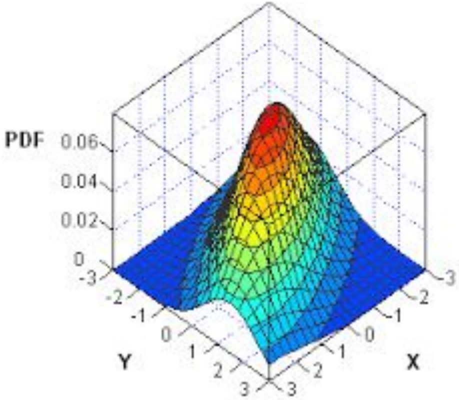


Image sources: sas.com, Wikipedia

Conditional probability assumes a fixed value for some variables

- If we fix *y* = *Y*, then we get a one-dim function *p(x | y=Y)*
- It is a normalized slice of the 2-dim function for *y=Y*

$$\int_{-\infty}^{\infty} p(x|y = Y)\ dx = 1$$



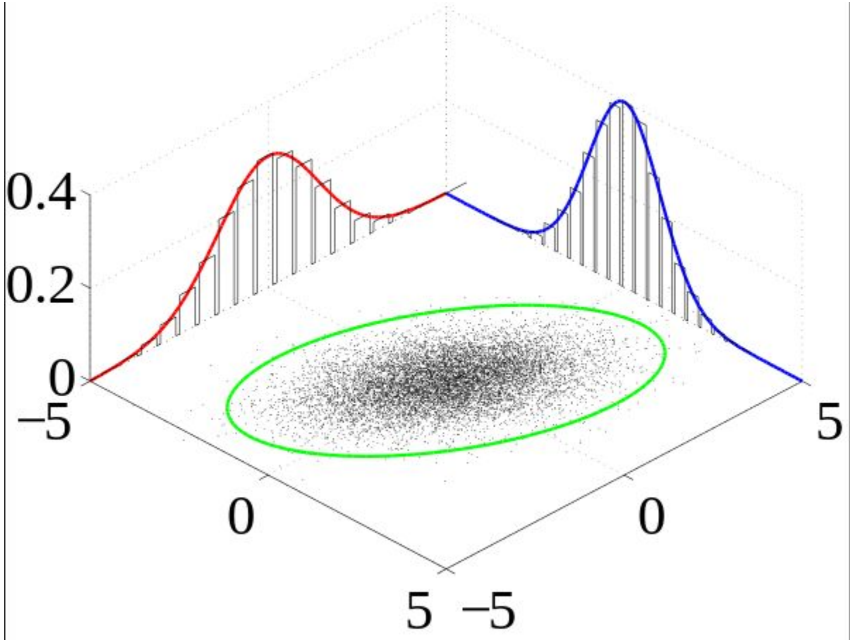


Image sources: sas.com, Wikipedia

Statistical independence of two random variables

- Two variables x and y are independent if and only if:

$$p(x,y) = p(x)\, p(y), \qquad \text{for all values of } x \text{ and } y$$

- Else, for dependent variables:

$$p(x,y) \neq p(x)\, p(y) \qquad \text{at least not for all values of } x \text{ and } y$$

Independently identically distributed (I.I.D.) sample

- Treat each sample drawn as a random variable
- Each such random variable is identically distributed
- Each random variable is also independent of each other
- E.g., heights of two people sampled at random at a metro station is independent of each other
- And, they are likely to have the same distribution

Likelihood of single instance of an RV

- If we assume a distribution $p_\theta(x)$ parameterized by θ
- Then, if we observe a single instance labeled x_1, then its distribution is: $p_\theta(x_1)$
- Then, the likelihood of x_1 taking value X_1 is $p_\theta(X_1)$

Likelihood of an I.I.D. sample

- For an IID sample of $x_1, x_2, \ldots, x_n$
- The likelihood is going to be a product of all these (because of independence)
- That is the likelihood of parameter θ given observations $X_1, X_2, \ldots, X_n$ is
- $p_\theta(X_1)\, p_\theta(X_2) \ldots p_\theta(X_n) = \prod_{i=1 \text{ to } n} p_\theta(X_i)$
- Log likelihood of θ is $\sum_{i=1 \text{ to } n} \log p_\theta(X_i)$

Comparing likelihood of two distributions

- Between two parameters θ_1 and θ_2 the one with higher log likelihood better explains the data
- This is the basis of statistical tests

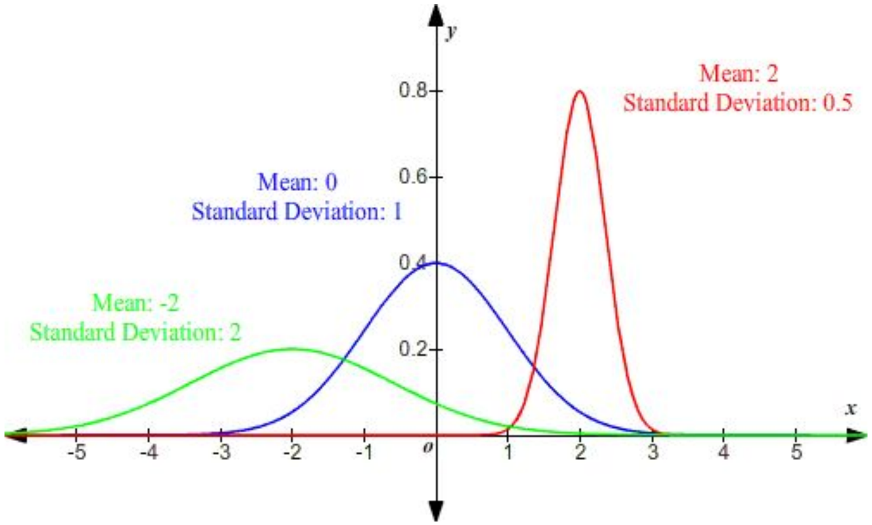


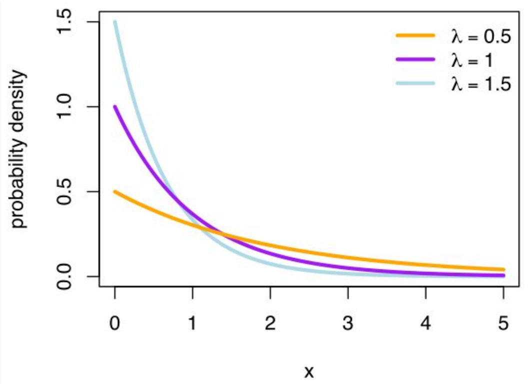
Image source : varsitytutors.com

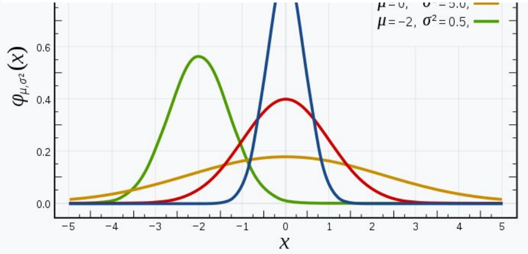
Maximum likelihood as an optimization problem

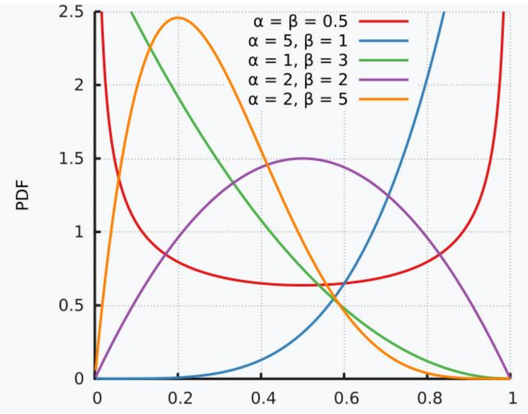
- One can form an optimization problem to maximize the likelihood by defining a likelihood function in terms of a continuous variable θ and setting its derivative with respect to θ to zero
- This is the basis of some ML techniques

Sufficient statistics

- For some distributions, estimating certain finite number of sample statistics is sufficient for maximizing the likelihood
- E.g. mean for an exponential distribution
- E.g. mean and variance for a Gaussian
- Generalized to “location” and “dispersion” parameter for several distributions of “fixed shape” that can be translated and stretched
- Some function families do not have a fixed shape, e.g. *beta*, and need more number of statistics to reach sufficiency; others may never reach sufficiency



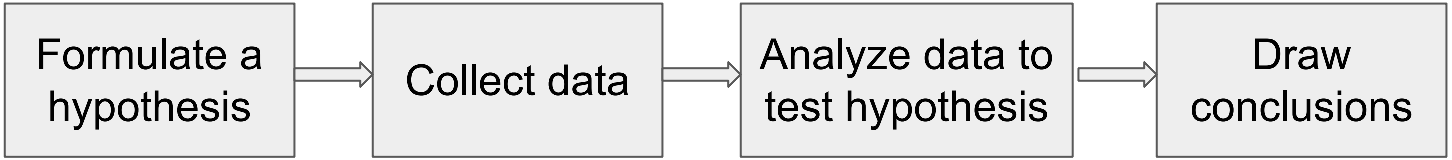




Images source: Wikipedia

Hypothesis testing

- Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. It is basically an assumption that we make about the population parameter.
- Hypothesis testing is used to establish whether a research hypothesis extends beyond the individuals examined in a single study.



Source: https://statistics.laerd.com/statistical-guides/hypothesis-testing.php

Key terms while using hypothesis testing

- **Null hypothesis** - two sample means are equal
- **Alternate hypothesis** - two sample means are NOT equal
- **Level of significance/ critical value/ p-value** - the degree of significance in which we accept or reject the null-hypothesis (usually 5% or 1%)
- **One-tailed predictions/values** - given statistical hypothesis is one value
- **Two-tailed predictions/values** - given statistical hypothesis assumes a less than or greater than value
- An analogy to describe hypothesis testing is a defendant on trial, since he/she is presumed innocent until proven guilty. This is equivalent to the null hypothesis being presumed true until proven false.

Student t-test

- It is a method of testing hypotheses about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown.
- t-test determines a probability that two populations are the same with respect to the variable tested.

Source: https://www.statisticshowto.com/probability-and-statistics/t-test/

Types of t-test

- An independent samples t-test compares the means for two groups.
- A paired sample t-test compares means from the same group at different times (say, one year apart).
- A one sample t-test tests the mean of a single group against a known mean.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$$

where

$$s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Sources: https://www.statisticshowto.com/probability-and-statistics/t-test/ and Wikipedia

Wilcoxon rank test

- The Wilcoxon test is a nonparametric statistical test that compares two paired groups, and comes in two versions, as given below:
    - the Rank Sum test or
    - the Signed Rank test.
- The goal of the test is to determine if two or more sets of pairs are different from one another in a statistically significant manner
- Basically you add the ranks of samples from one distribution, and check how likely they are to be high (or low) if they were from the combined distribution

Sources: https://www.investopedia.com/terms/w/wilcoxon-test.asp , https://www.stat.auckland.ac.nz/~wild/ChanceEnc/Ch10.wilcoxon.pdf