Prob Stats Cheat Sheet

• **Experiment:** any procedure that can be infinitely repeated is known as an experiment. There are two types of experiments:
- **Deterministic Experiment:** when the exact outcomes can be determined
- **Probabilistic Experiment:** when the outcomes are uncertain and cannot be determined

• **Sample Space:** the collection of all the possible outcomes of the experiment, known as the sample space (S).

• **Event (E):** a subset of the sample space of an experiment. i.e., E ⊆ S

• **Probability (P)**: the likelihood of an event occurring.

$$P(A) = \frac{\text{no. of favourable outcomes to A}}{\text{Total no. of possible outcomes}}$$

• **Mutually Exclusive Events** (Disjoint events)**:** If two events are mutually exclusive then the probability of both events occurring at the same time is equal to zero.

$$P(A \cap B) = 0$$

• **Mutually exhaustive** events when combined cover all the possible outcomes, known as mutually exhaustive events.

$$P(A \cup B) = 1$$

• **Non Mutually Exclusive Events** (Joint events)**:** If two events are not mutually exclusive then the probability of both events occurring at the same time is not equal to zero.

$$P(A \cap B) \neq 0$$

• **Independent Events**: The occurrence of one does not change the probability of the other occurring.

$$P(A \cap B) = P(A).P(B)$$

• **Basic rules of probability:**

Addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- In the case of disjoint events:
  $P(A \cup B) = P(A) + P(B)$
  as $P(A \cap B) = 0$ while events are disjoint.

Complement rule: $P(A^c) = 1 - P(A)$

• **Measures of Central Tendency:**

Mean: The average value of a set of numbers

$$\mu = \frac{\Sigma X}{n}$$

Median: The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

For odd no. of observations:

$$(\frac{n+1}{2})th \text{ observation's value}$$

For even no. of observations:

$$\frac{(\frac{n}{2})th\ observation + (\frac{n}{2}+1)th\ observation}{2}$$

Mode: It is the most occurring value in the dataset.

• **Measures of Variability:**

Range: Range is nothing but Maximum value - Minimum value

Inter Quartile Range (IQR)**:** a measure of the middle 50% of a data set.

IQR = Q3 - Q1

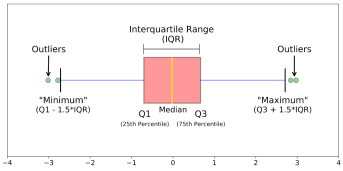
Q3 =third quantile value, Q1 = first quantile value

Used to detect outliers.

All the points having values greater than $(Q3 + 1.5IQR)$ or less than $(Q1 - 1.5IQR)$ are considered to be outliers.

Boxplot: a standardized way of displaying the distribution of data based on a five-number summary.

These are "minimum", first quartile [Q1], median, third quartile [Q3], and "maximum".



• **Random Variables:**

Discrete Random Variables: It represents distinct, separate outcomes, and it can only take on specific values within a countable set.

- E.g. The number of people in a household.
- The number of goals scored in a soccer match.

Continuous Random Variables: It can take any value within a range and has an uncountable set of possible outcomes.

- E.g. The height of an adult male or female.
- The weight of an object.

• **Distribution functions:**

Probability Mass Function (PMF): PMF is a function that gives the probability of a discrete random variable taking on a specific value.

Probability Density Function (PDF): PDF is a function that describes the likelihood of a continuous random variable falling within a particular range.

Cumulative Distribution Function (CDF): CDF gives the probability that a random variable takes a value less than or equal to a given point for both discrete and continuous variables.

• **Expectation:** For a discrete random variable(X) having a Probability mass function P(x).

$$E(X) = \sum_{i=1}^{n} x_i . P(x_i)$$

• **Cross tab:** Crosstab, or cross-tabulation, generates a contingency table showcasing the relationship between two or more categorical variables and provides a count or frequency of occurrences for pairs of categorical variables.

- There is a function called pd.crosstab() to generate the table

• **Conditional Probability:**

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

• **Marginal Probability:** the probability of a single event occurring without consideration for the occurrence of other events.

E.g. $P(A)$, $P(B)$, $P(C)$

• **Joint Probability:** Joint probability is the likelihood of two or more events occurring simultaneously.

- E.g. $P(A \cap B)$, $P(A \cap C)$, $P(B \cap C)$

• **Multiplication rule:**

$$P(A \cap B) = P(A|B) . P(B) = P(B|A) . P(A)$$

• **Bayes Theorem:**

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

• **Permutation and Combination:**

Permutation: A permutation is an arrangement of items or elements in a specific order, where the order of the arrangement matters.

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

Combination: A combination is a selection of items or elements where the order of the arrangement does not matter.

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

Variance: The spread of numbers in a data set w.r.t mean.

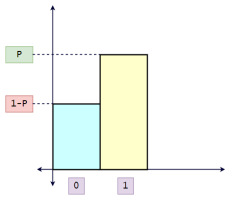
$$Var(X) = \sigma^2 = \frac{1}{n}\sum_{1}^{n}(x_i - \mu)^2$$

Standard Deviation: It represents how far the data point from the mean (μ)

SD = √variance = σ

Discrete probability distributions:

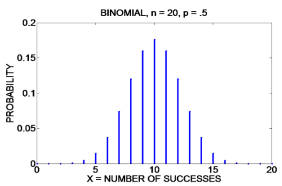
1. Bernoulli distribution



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

$$E[X] = p$$

2. Binomial distribution

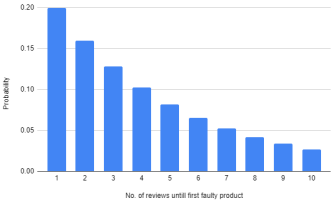


$$P(X = x) = {}^nC_x . p^x . (1-p)^{n-x}$$

$$E[X] = n . p$$

$$Var(X) = n . p . (1-p)$$

3. Geometric distribution:



$$P(k) = (1-p)^{k-1} . p$$

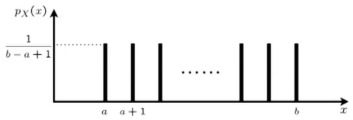
$$E(X) = \frac{1}{p}$$

4. Poisson distribution:

$$P(X = x) = \frac{\lambda e^{-x}}{x!}$$

$$E(X) = Var(X) = \lambda$$

5. Discrete Uniform distribution:



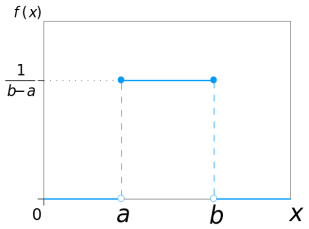
$$f(x) = \frac{1}{b-a+1}$$

$$\mu = E(X) = \frac{b+a}{2}$$

$$\sigma^2 = \frac{(b-a+1)^2 - 1}{12}$$

• Continuous Probability distributions:

1. Uniform distribution:

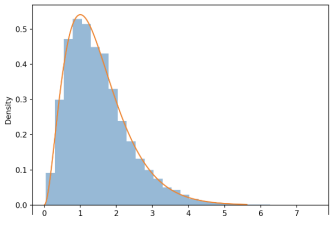


$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

$$E(X) = \frac{a+b}{2}$$

$$\text{Variance } (\sigma^2) = \frac{(b-a)^2}{12}$$

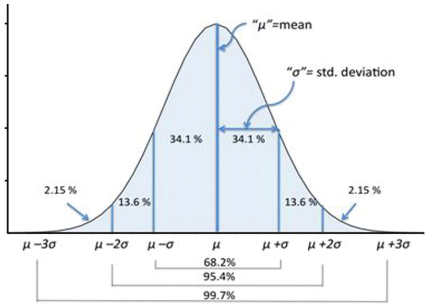
5. Exponential distribution:



pdf $f(x|\beta) = \frac{1}{\beta} e^{-x/\beta},$

Where β= 1/λ

2. Normal distribution:



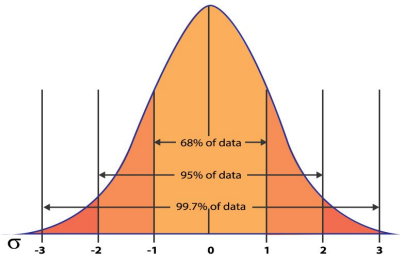
PDF = $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

Properties:

Symmetric about mean and has a bell-shaped distribution.
Mean (μ) = mode = median

Empirical rule:
68% of values lie within 1 standard deviation from the mean.
95% of values lie within 2 standard deviations from the mean.
99.7% of values lie within 3 standard deviations.

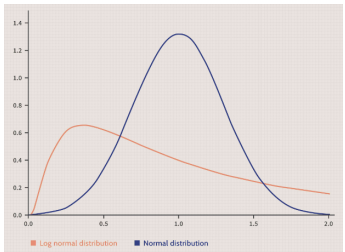
3. Standard Normal distribution:



Special case of Normal distribution when the mean=0 and standard deviation=1.

$$z - score = \frac{x-\mu}{\sigma}$$

4. Log Normal distribution:



$$\mu_X = \exp\left(\mu + \sigma^2/2\right), \qquad \sigma_X^2 = Var(X) = \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2}$$

• **Sample vs population:**

Population mean = $\mu = \frac{1}{N}\sum_{1}^{N} x_i$, *Population Variance* = $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Sample mean = $\overline{x} = \frac{1}{n}\sum_{1}^{n} x_i$, *Sample Variance* = $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Sample Standard Deviation = $\sqrt{s^2}$

• **Sampling Techniques:**

Simple random sampling: In this sampling method, each member of the population has an exactly equal chance of being selected which tends to produce representative, unbiased samples.

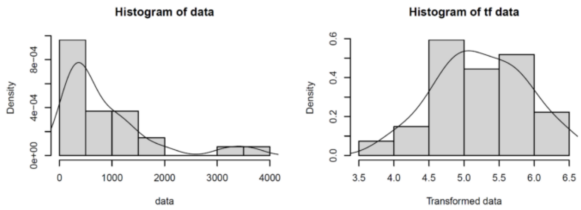
A simple random sample is a randomly selected subset of a population.

• **Standard error:** the spread of sample means around the population mean.

$$Standard\ error = \frac{\sigma}{\sqrt{n}}$$

• **Central Limit Theorem (CLT):**
the distribution of sample means is Gaussian, no matter what the shape of the original distribution is.
Assumptions: population mean and standard deviation should be finite and sample size >=30.

• **Box-Cox Transformation:** The basic idea behind this method is to find best value for λ such that the transformed data is as close to normally distributed as possible, using the following formula:



$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$