MORINGA
Discover · Grow · Transform

# Probability distributions

# Session Expectations

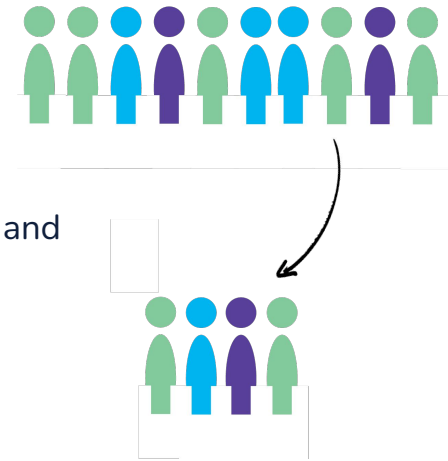
| Online edition | The usual |
| --- | --- |
| → Switch on your video when speaking (if possible)<br>→ Participate in the discussions<br>→ Mute yourself when you are not speaking | → Be present<br>→ Be honest<br>→ Be open<br>→ Be curious |

# What to expect

- ❑ Types of data
- ❑ Discrete Vs Continuous
- ❑ Statistical distribution
- ❑ Break
- ❑ PMF, PDF and CDF

# Sampling Keywords

1. **Population**: It is the complete set of all possible subjects of interest.
2. **Sample**: a subset of the population that is selected for study. should ideally be representative of the population
3. **Parameter**: a numerical characteristic of a population
   - Population mean = $\mu$, size = N
4. **Statistic**: a numerical characteristic of a sample. It is calculated from sample data and are used to estimate population parameters.
   - Sample mean = $\bar{x}$, size = n
5. **Sampling Distribution**: is the probability distribution of a statistic based on all possible samples of a fixed size drawn from a population.
6. **Bias**: the systematic error in a study that leads to incorrect estimates of population parameters.
7. **Variance** is a measure of the dispersion or spread of a set of data points around their mean.
8. **Standard Deviation**: *sqrt(variance)*. Also, a measure of dispersion because it is in the same units as the original data.



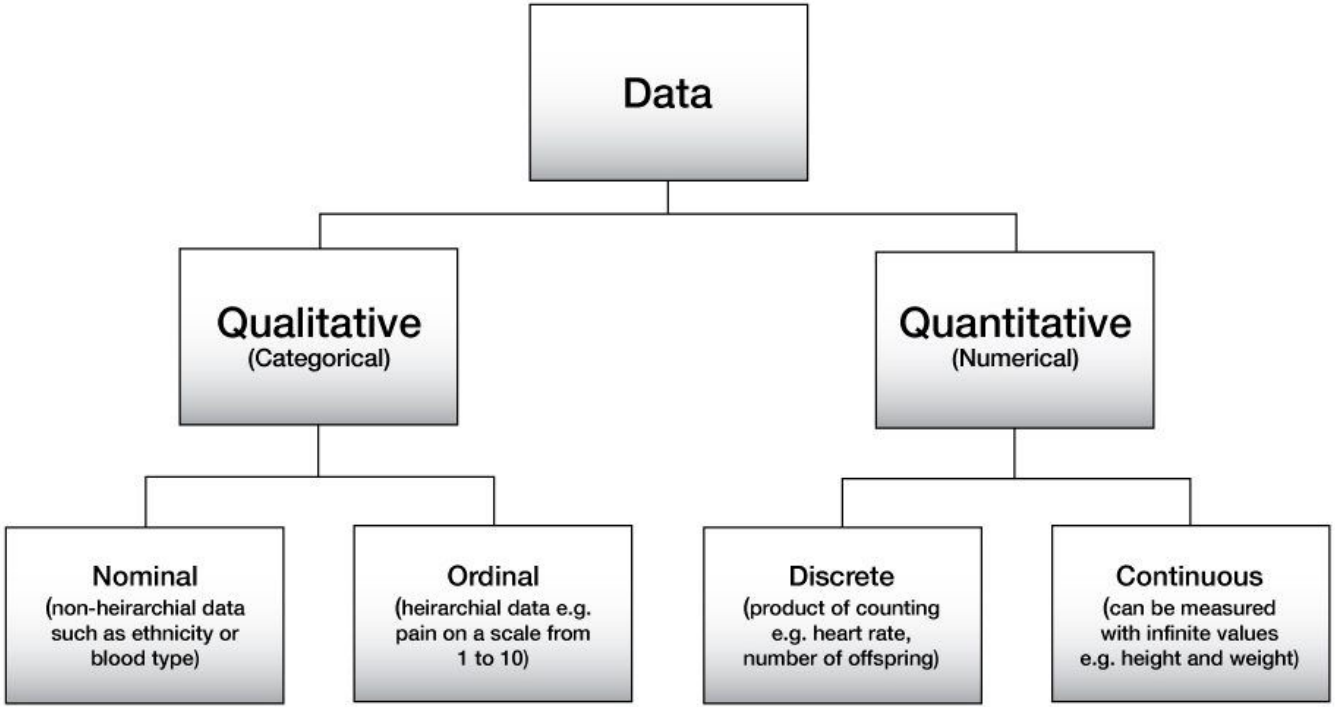
# Symbols in Sampling

| Name | Population Parameters | Sample Statistics |
|---|---|---|
| Mean | $\mu$ | $\overline{X}$ |
| Median | $\eta$ | $\tilde{X}$ |
| Mode | No symbol | No symbol |
| Range | R | R |
| Variance | $\sigma^2$ | $s^2$ |
| Standard Deviation | $\sigma$ | s |
| Sample Size | N | n |
| Estimates | $\hat{\sigma}$ | n/a |

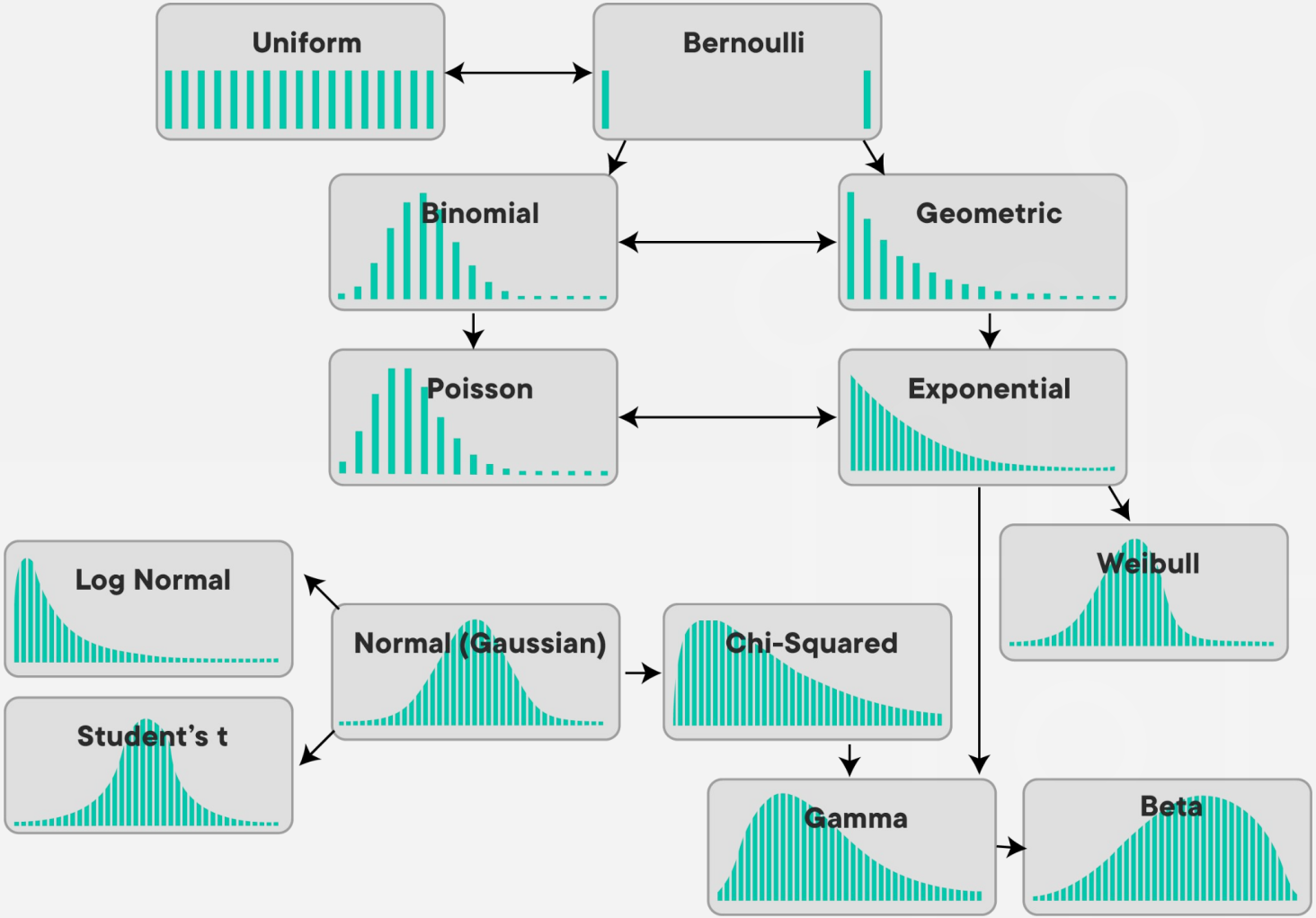
# Symbols in statistics

| Symbol | Name | Symbol | Name |
|---|---|---|---|
| ∑ | Sum | n | Size of a subsample |
| ^ | Hat, used above a parameter to denote an estimate | N | Total sample size |
| ANOVA | Analysis of variance | OR | Odds ratio |
| α | Alpha, probability of Type I error | *P* | Statistical probability |
| β | Beta, probability of Type II error; or population regression coefficient | $X^2$ | $X^2$ test or statistic |
| CI | Confidence interval | *r* | Bivariate correlation coefficient |
| CV | Coefficient of variation | *R* | Multivariate correlation coefficient |
| Δ | Delta, change | RR | Relative risk |
| δ | Delta, true sampling error | ρ | Rho, population coefficient |
| ε | Epsilon, true experimental error | SD | Standard deviation of a sample |
| $H_0$ | Null hypothesis | SE | Standard error |
| $H_1$ | Alternate hypothesis; specify whether 1 or 2 sided | SEM | Standard error of the mean |
| κ | Kappa statistic | *t* | Student t; specify α level |
| μ | Population mean | *U* | Mann-Whitney *U* (Wilcoxon) statistic |
| | | *z* | *z* score |

# Types of data



| Discrete | Continuous |
| --- | --- |
| It takes on distinct, countable values | It takes on any value within a range, and the number of possible values within that range is infinite. |
| - Countable | - Measureable |
| - Individually whole | - Ungrouped whole |
| - Can have hierarchy | - Cannot have hierarchy unless grouped |
| - Can be sorted | - Cannot be sorted |
| - Integers | - Any real numbers |
| - Numbers of cars, products, items, books, customers etc | - Weight, Height, Distance travelled, water volume etc |


Uniform
Bernoulli
Binomial
Geometric
Poisson
Exponential
Weibull
Log Normal
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta

# Statistical distributions

These are mathematical functions that describe the likelihood of different outcomes occurring in a population or sample

1. **Normal Distribution (Gaussian)**
   - Symmetric around their mean, bell-shaped distribution characterized by its mean and standard deviation.
   - μ is the mean and σ is the standard deviation.(mean, median & mode are equal)
   - Widely used in statistics due to the central limit theorem.
     - Naturally occurring phenomena such as heights, weights, test scores, temperature.
2. **Binomial Distribution**
   - Models the number of successes in a fixed number of independent Bernoulli trials, where each trial has the same probability of success.
   - the number of trials, n, and the probability of success, p.
     - success/failure experiments.
3. **Poisson Distribution**
   - Models the number of events occurring in a fixed interval of time or space, given a known average rate of occurrence.
   - the average rate of occurrence λ
     - waiting times between phone calls, durations of time until failure in reliability engineering, and inter-arrival times in queuing systems.
4. **Exponential Distribution**
   - Describes the time between events in a Poisson process, where events occur continuously and independently at a constant average rate.
   - λ.

•••

5. **Uniform Distribution**
   - Assigns equal probability to all outcomes within a specified range
   - Minimum , maximum.
     - selecting a random number between two values i.e. coin toss, die throw

6. **Bernoulli Distribution**
   - a single trial with two possible outcomes:
   - success (p) or failure (1–p).
     - success/failure experiments.

7. **Gamma Distribution**
   - Models the waiting time until a specified number of events occur in a Poisson process, where events occur continuously and independently at a constant average rate.
   - the shape parameter k and the rate parameter θ
     - queuing theory, finance, and insurance.

8. **Beta Distribution**
   - Describes the probability distribution of a random variable bounded between 0 and 1.
   - two shape parameters, α and β
     - Bayesian statistics, modeling proportions in epidemiology, and A/B testing in marketing

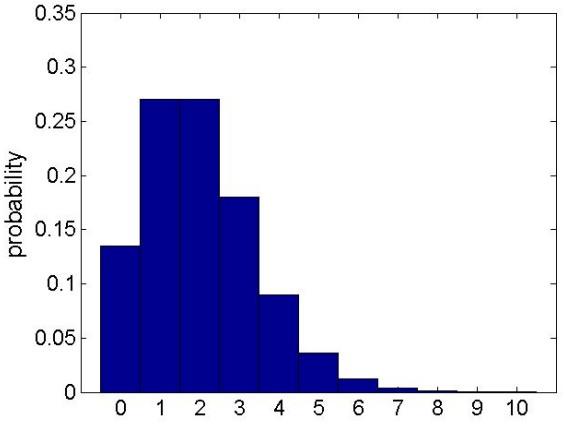
# Probability mass function (PMF)

- A function that associates probabilities with discrete random variables.
- Its a function over the sample space of a discrete random variable X which gives the probability that X is equal to a certain value.
- PMF is strictly positive.
- Based on your experience of rolling a dice, you can develop a PMF showing the probabilities of each possible value between 1 and 6 occurring.

$$f(x) = \mathrm{P}[X = x]$$

$$f(x) \geq 0 \text{ for all } x \in S,$$

$$\sum_{x \in S} f(x) = 1.$$




## Applications

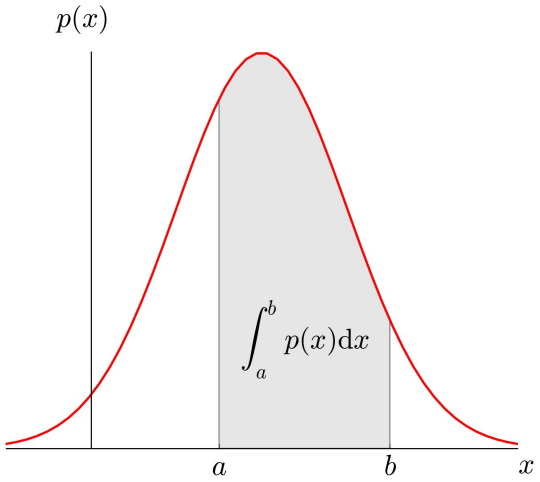
1. It is used to calculate the mean and variance of the discrete distribution.
2. It is used in binomial and Poisson distribution to find the probability value where it uses discrete values.

# Probability density function (PDF)

$$\int_a^b f(x)\mathrm{d}x = \mathrm{P}[a < X \leq b]$$

(i) $f(x) \geq 0$ for all $x \in \mathbb{R}$,

(ii) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1.$

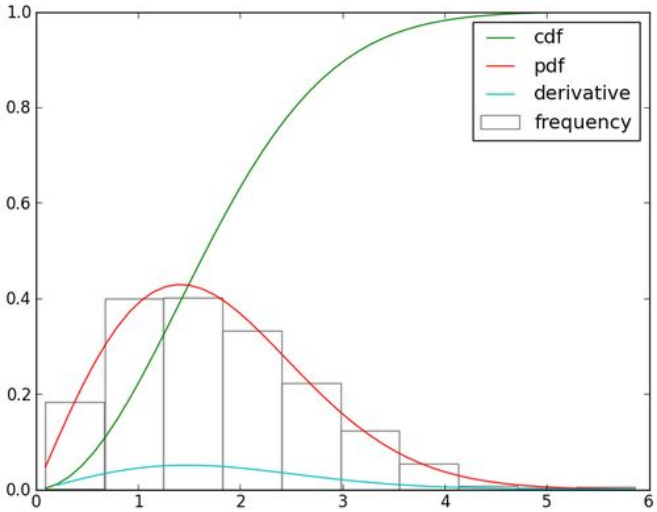



- Its a function that specifies the likelihood of a continuous random variable falling within a particular range of values.
- PDF is also positive.
- Unlike the PMF, the PDF does not directly give probabilities but rather density values.

## Applications

1. used to calculate the probabilities associated with the random variables

# Cumulative density functions (CDF)




- Its a function that gives the probability that a random variable is less than or equal to a specified value.
- It is cumulative, adding up probabilities.
- It is denoted as F(x), where x is the value.

**PMF/CDF**

$$F(x) = \mathrm{P}[X \leq x]$$

$$P[a < X \leq b] = F(b) - F(a)$$

**PDF/CDF**

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t \qquad \frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x)$$


The End