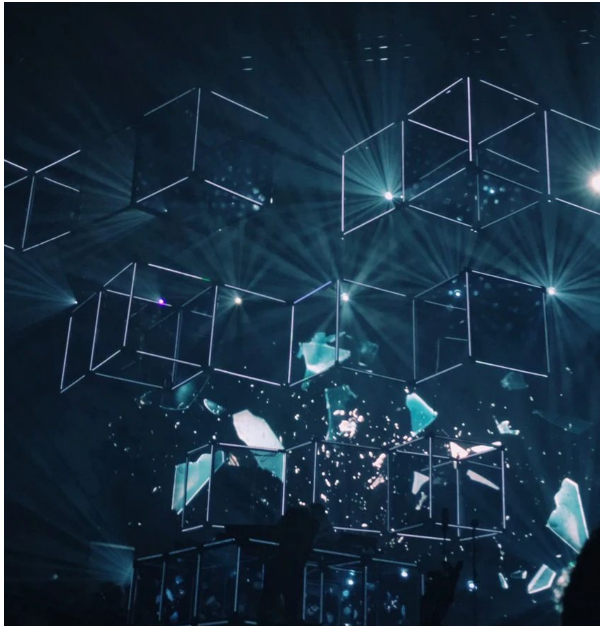

# Basic statistical concepts.

DATA ANALYTICS | IRONHACK



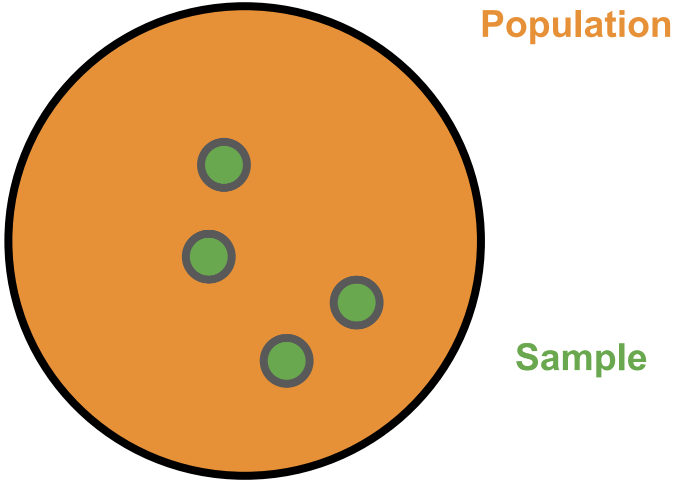
Credit: Unsplash

# Agenda

- Population.
- Samples.
- Measures of Central Tendency and Dispersion
- Random variables.
- Probabilities and probability distributions

# POPULATIONS AND SAMPLES

- A population is a COMPLETE collection of items/individuals/products




- A sample is SUBSET of a population

# POPULATION

- A population is an aggregate/collection of creatures, things, cases, etc…
- A population commonly contains too many individuals to study conveniently, so an investigation is often restricted to a reduced subset of the population.

# SAMPLES

- A sample of a population is a subset of elements of the population.

- Based on information from the sample, we can make **assumptions** about the population.

- A **random sample** means that the observations from the population are picked randomly and without any bias.

# STATISTICAL DESCRIPTOR OF A SAMPLE VS. THE POPULATION

- The difference between both is that the population statistical descriptor is **FIXED**, while the the statistical descriptor of the **SAMPLE** **will change from sample to sample.**

    - The population mean and standard deviation are fixed (assumed to be fixed) and are called population parameters.
    - The sample mean, sample std. deviation varies every time we calculate them as a random sample will have different values every time. They are called sample statistics.

# COMMON STATISTICS NOTATION

- **POPULATION PARAMETERS:**
  - N: Number of elements in the population
  - $\mu$: mean of the population
  - $\sigma$: standard deviation of the population
- **SAMPLE STATISTICS: ( they are random by nature )**
  - n: Number of elements in the sample
  - $\overline{x}$: sample mean
  - S: sample standard deviation

# POPULATION, SAMPLE AND COLLECTION OF SAMPLES

- A population can be characterized by a distribution which will have:

$$(\mu, \sigma)$$

- A sample drawn from a population will be characterized by a sample mean (changes from sample to sample; **is a random variable**), and a sample standard deviation:

$$(\bar{x}, s)$$

- We can analyze what mean we will get if we compute the **mean of the means of many collected samples** from the population:

$$(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

# RANDOM VARIABLES

- A random variable, usually written as X, is a variable whose possible values are numerical outcomes of a random phenomenon.(for eg. height people in the US, marks scored in a test, etc.)
- This random variable can be either continuous or discrete in nature.
    - ***Discrete random variable*:** The set of values that his random variable can take are discrete (usually but not necessarily counts)
    - ***Continuous random variable***: The set of values that his random variable can take are continuous

# HOW TO CHARACTERIZE A SAMPLE/POPULATION

- There are two types of measures to characterize a sample / population
  - Central tendency
  - Dispersion

# MEASURES OF CENTRAL TENDENCY

- [1,2,2,4,6,8,50]

- Mean: (1+2+2+4+6+8+50)/7 = 10.43 ( **heavily affected by outliers**)
- Median: 4 ( **unaffected by outliers** )
- Mode: 2

# MEASURES OF DISPERSION

- **Residuals:** ( real - predicted )
- **Variance:** sum of squared residuals divided by n - 1
- **Standard deviation:** root square of variance
- **Range:** difference between the maximum and minimum
- **Percentile:** the value such P percent takes this value or less
- **Interquartile range:** difference between percentiles 75% and 25%

# VARIANCE AND STANDARD DEVIATION

- Variance: $$var(x) = \sum_i \frac{(x - \bar{x})^2}{n - 1}$$

- Standard deviation: (we divide by n-1 to avoid underestimating the Standard deviation) $$std(x) = \sqrt{\sum_i \frac{(x - \bar{x})^2}{n - 1}}$$
  demo

# QUANTILES AND IQR

- The quantiles of a sample are defined in the following way:
    - Q1: after sorting the values from the smallest to highest, which value is bigger than 25% of the values:
    - Q2: after sorting the values from the smallest to highest, which value is bigger than 50% of the values:
    - Q3: after sorting the values from the smallest to highest, which value is bigger than 75% of the values:

- The IQR = Q3 - Q1 is called the “interquartile range”

IRON HACK

# Probability and probability distributions

# EVENTS AND PROBABILITY

- The probability of an outcome is the proportion of times that the outcome will occur if we observe the random process an infinity amount of times

# COMBINING PROBABILITIES

- Two different events can be broadly classified as:
  - Joint -> They can happen together
  - Disjoint -> They can't happen together

- Combination rules for probabilities are different for join and disjoint events:

A coin and you have dice

# COMBINING PROBABILITIES: DISJOINT EVENTS

- To compute the probability of two disjoint events, we add the probabilities of each of them:

$$P(A or B) = P(A) + P(B)$$

$$P_T = \sum_i P_i$$

# COMBINING PROBABILITIES: JOINT EVENTS

- To compute the probability of two joint events, we add the probabilities of each of them but subtracting the intersections:

$$P(A or B) = P(A) + P(B) - P(A and B)$$

# EXAMPLES

| | Barcelona | Berlin | Total |
|---|---|---|---|
| Data | 26 | 21 | 47 |
| UX | 19 | 10 | 29 |
| Total | 45 | 31 | 76 |

**P(Barcelona)** = 45/76 = 0,59

**P(Data)** = 47/76 = 0,62

**P(Barcelona and Data)** = 26/76 = 0,34

**P(Bcn or Data) = P(Bcn) + P(Data) – P(Bcn and Data)**

**P(Bcn or Data) =** 0,59 + 0,62 – 0,34 = **0,87**

# RANDOM VARIABLES:

- Random variable → Discrete
- Random variable → Continuous

<u>Discrete</u>

X = { heads, 1
{ tail , 0

123.8976 ????

<u>Continuous:</u>

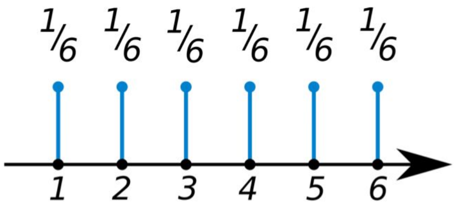
Y = {What is the **EXACT** mass of a random animal selected at the zoo?

0 5000 kg

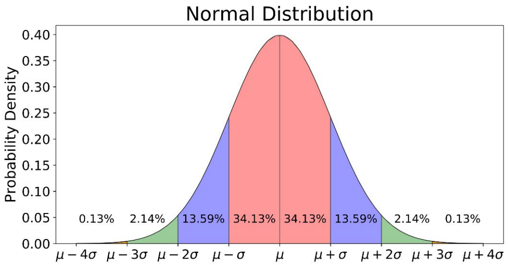
# PROBABILITY DISTRIBUTIONS:

- Mathematical function that gives us the probabilities of occurrence of different outcomes.

Discrete →

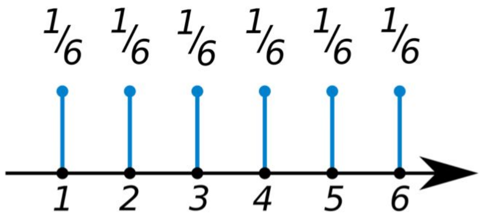



Continuous →



# PMF: PROBABILITY MASS FUNCTION

- Is a function that gives us the probability that a **DISCRETE** random variable takes a specific value



# PDF: PROBABILITY DENSITY FUNCTION

- The PDF is the equivalent of PMF but for **continuous variables**.
- As you can have an infinite amount of possible values between any two numbers, **it's not possible to define the probability for every value.**

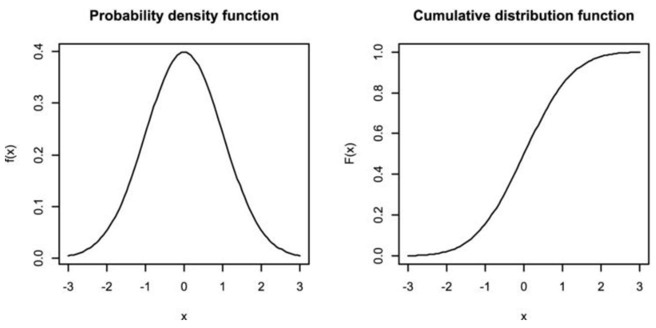
- Therefore, we define a **density of probability** rather than the probability mass. The concept is very similar to mass density in physics: its unit is **probability per unit length.** To get a feeling for PDF, consider a continuous random variable X

$$g(x) = lim_{\Delta \to 0} \frac{P(X < x < X + \Delta)}{\Delta}$$

# CDF: CUMULATIVE DISTRIBUTION FUNCTION

- The CDF is a function that gives the probability of getting any value smaller or equal to a given value:

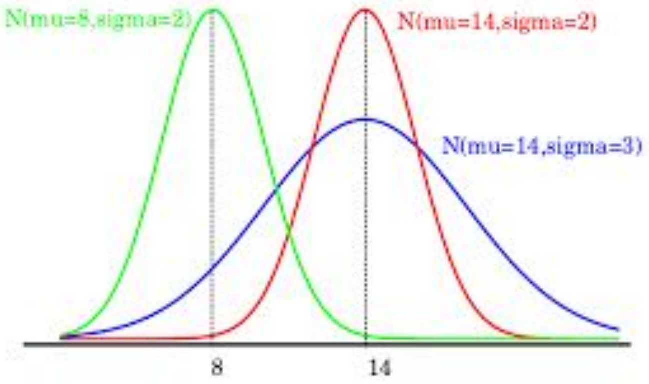
$$F(X) \equiv P(x <= X) = \int_a^x g(z)dz$$



# NORMAL DISTRIBUTION

- The normal distribution is characterized by the mean and the standard deviation.

$$\Phi_{\mu,\sigma} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad P(x <= z) = \int_{-\infty}^{z} \Phi_{\mu,\sigma}\, dx$$




- X can take values from (-inf,inf)

# PROBABILITIES IN PYTHON: SCIPY

- Scipy library contains the module ‘stats’ which allows you to compute probabilities, PDF, CDF,....for several distributions.
- Official documentation

    - scipy.stats.norm.pmf()
    - scipy.stats.expon.pmf()
    - …

Every distribution has its own pmf() function in scipy.

THANKS !