HackBio
Statistics: Lecture 1

Winfred Gatua

03/02/2021

Objectives

HackBio

- Key definitions
- Sampling
- Measures of Data
- Data distribution
- Discrete and Continuous Probability

Key definitions

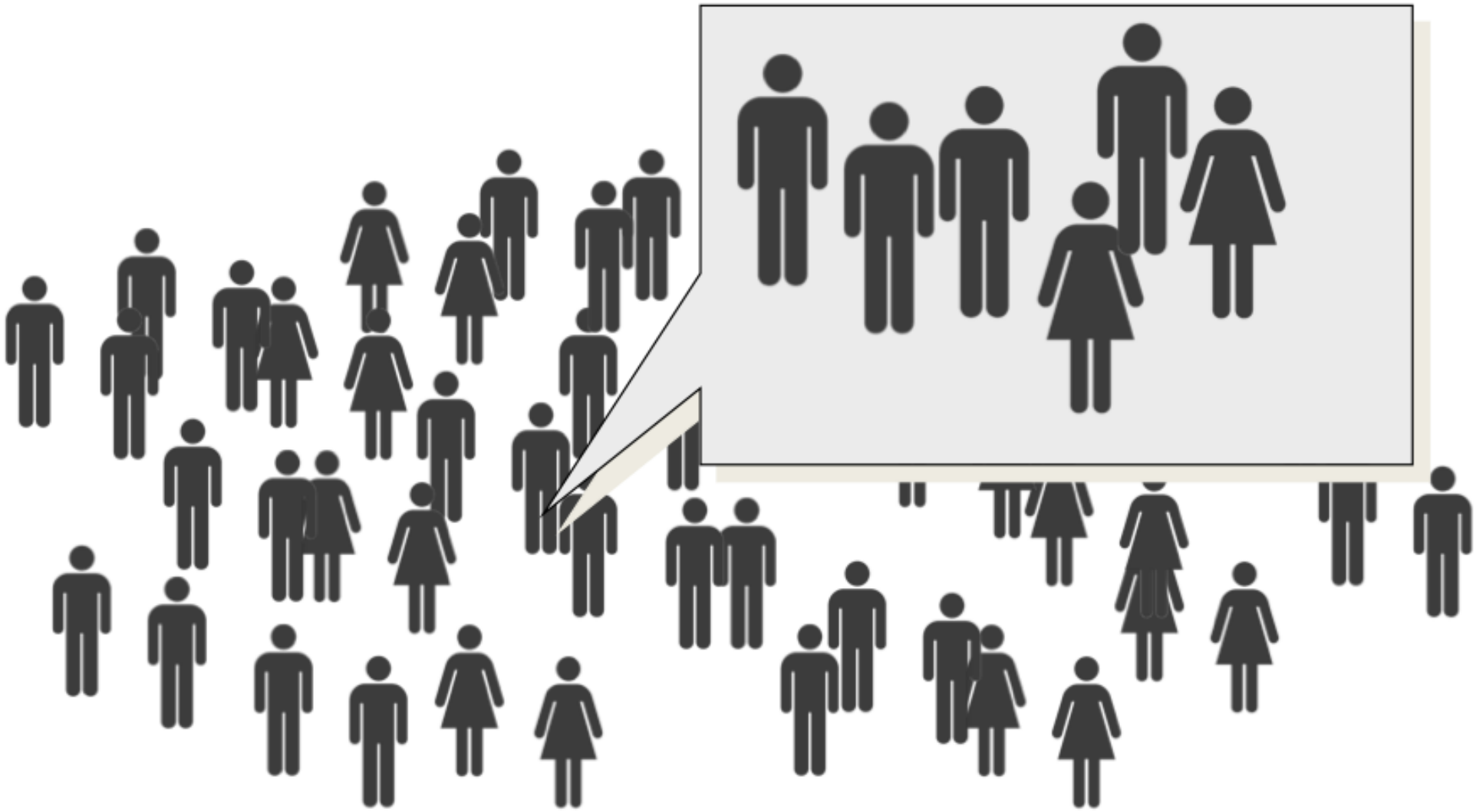
- **Statistics** is the science of Collecting, Summarizing, Presenting and Interpreting data
- **Descriptive statistics** interested in basic features, obtain simple summaries, frequencies
- **Inferential statistics** entails making inference about a population from the data
- **Population** a collection of elements about which we wish to make inference
- **Sample** a collection of sampling units drawn from the sampling frame

Continue

- **Parameter** numerical characteristic of a population
- **Statistic** numerical characteristic of a sample
- **Element** an object from which a measurement is taken

Sampling

HackBio



Why Sample?

- **Pragmatic reasons**
  - Budget and time constraints
  - Limited access to the total population

- **Accurate and reliable results**
  - Strong similarity in population elements make it possible
  - Samples can yield reasonably accurate results

- **Destruction of test units**
  - Reduces the cost of research in finite populations

Sampling methods

- **Probabilistic**

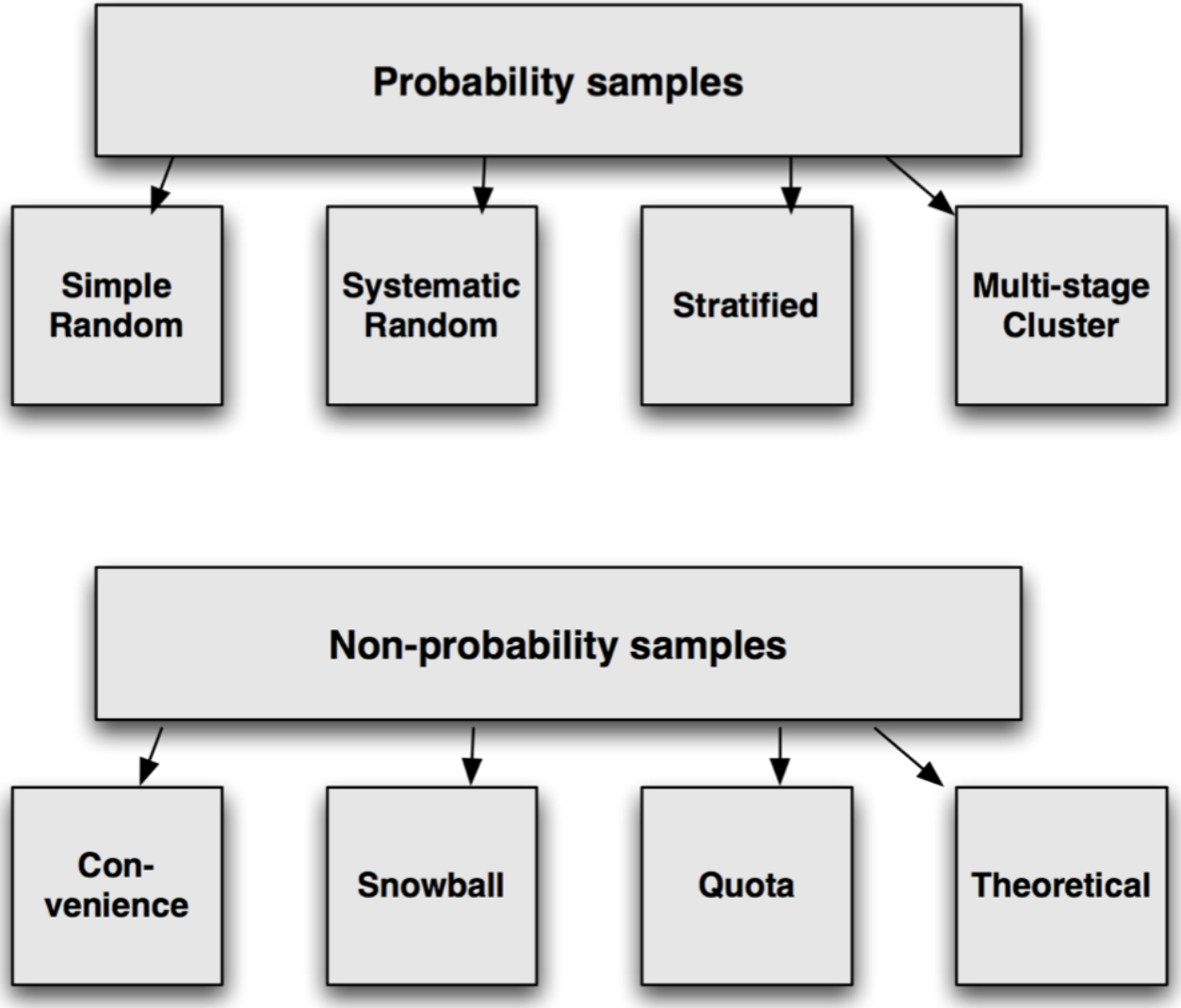
- Each member of the population has a non-zero probability of being selected

- **Non-probabilistic**

- Members are selected form the population in a non-random manner

Sampling methods

HackBio

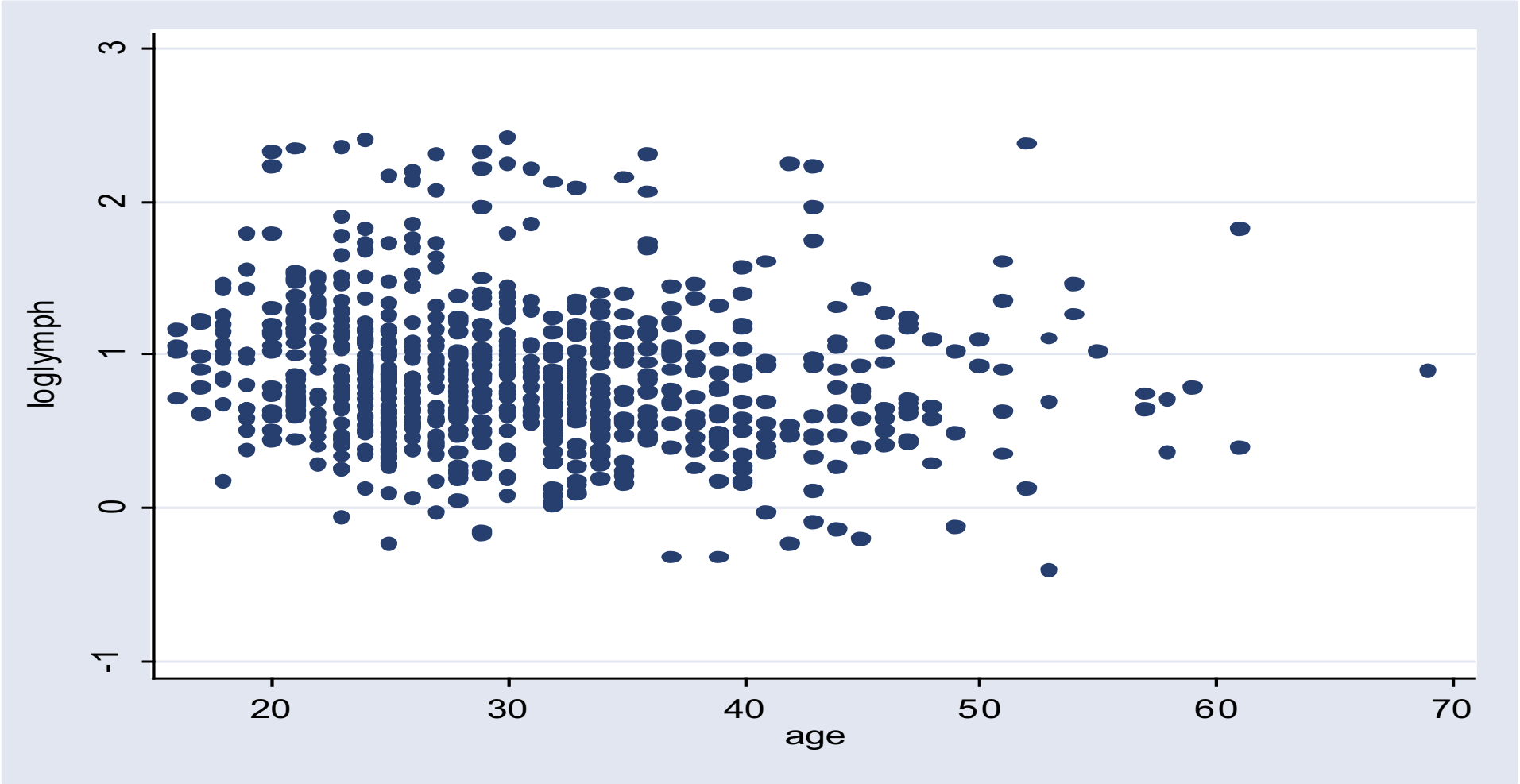


Classes of variables

- **Exposures** that which we wish to investigate in relationship with an outcome e.g. age, diet, treatment.
- A given exposure may be associated with increased or decreased outcome
- **Outcome** variable of interest e.g. weight gain, recovery from illness
- The two should be defined before the study starts
- Outcome determines the type of analysis to be done

Lymphocyte and Age

HackBio



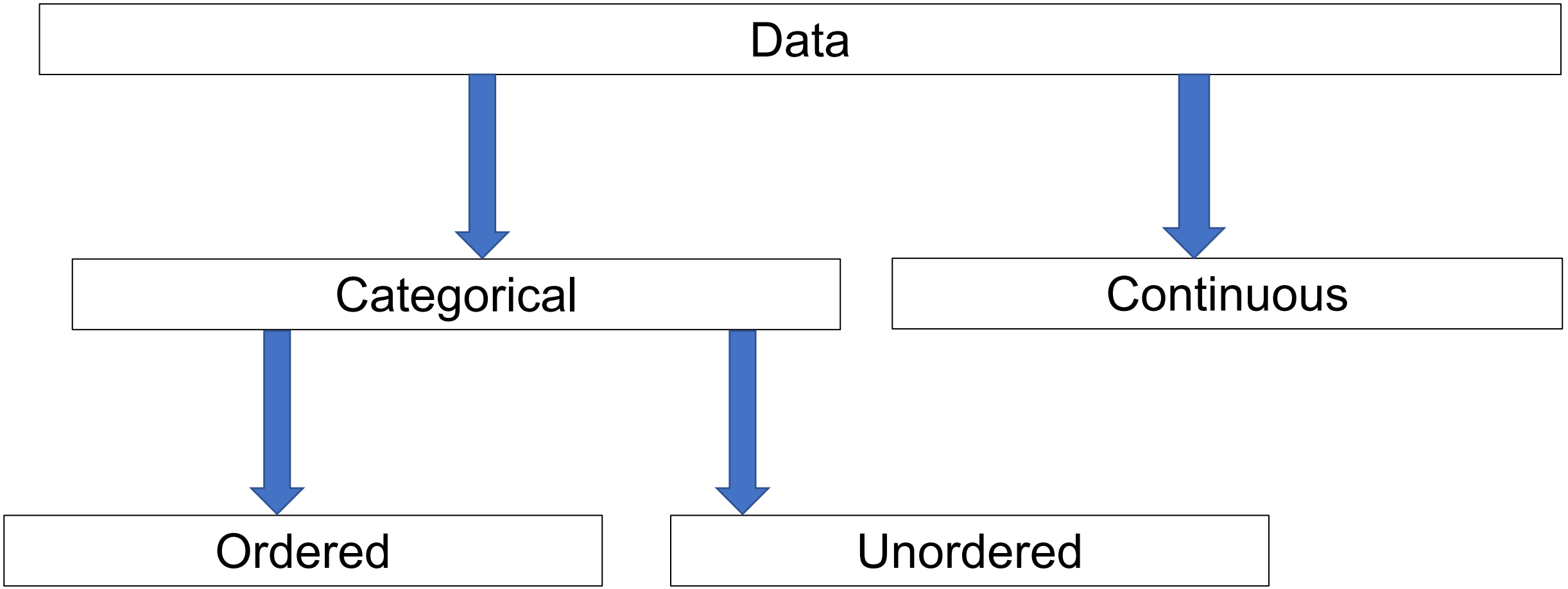
What is the exposure and outcome variable?

Variable types

- **Exposures** that which we wish to investigate in relationship with an outcome e.g. age, diet, treatment.
- A given exposure may be associated with increased or decreased outcome
- **Outcome** variable of interest e.g. weight gain, recovery from illness
- The two should be defined before the study starts
- Outcome determines the type of analysis to be done

Variable types

HackBio



Categorical Variables

- **Nominal categorical variables**
  - Marital status
  - Ethnic group
  - None is higher than the other
- **Ordinal Categorical variable**
  - Wealth status: High, Medium, Low
  - Education level: Primary, Secondary, College,

Quantitative Variable

- **Discrete:**
  - No. of toys: 0, 1, 2, 3, 4, 5, …
- **Continuous:**
  - Weight, Age
- **Interval:** data with an arbitrary zero
  - Measured along a scale in which each position is equidistant from another
- **Ratio:** data with an absolute zero. In a ratio scale numbers can be compared as multiples of one another
  - The number zero has meaning

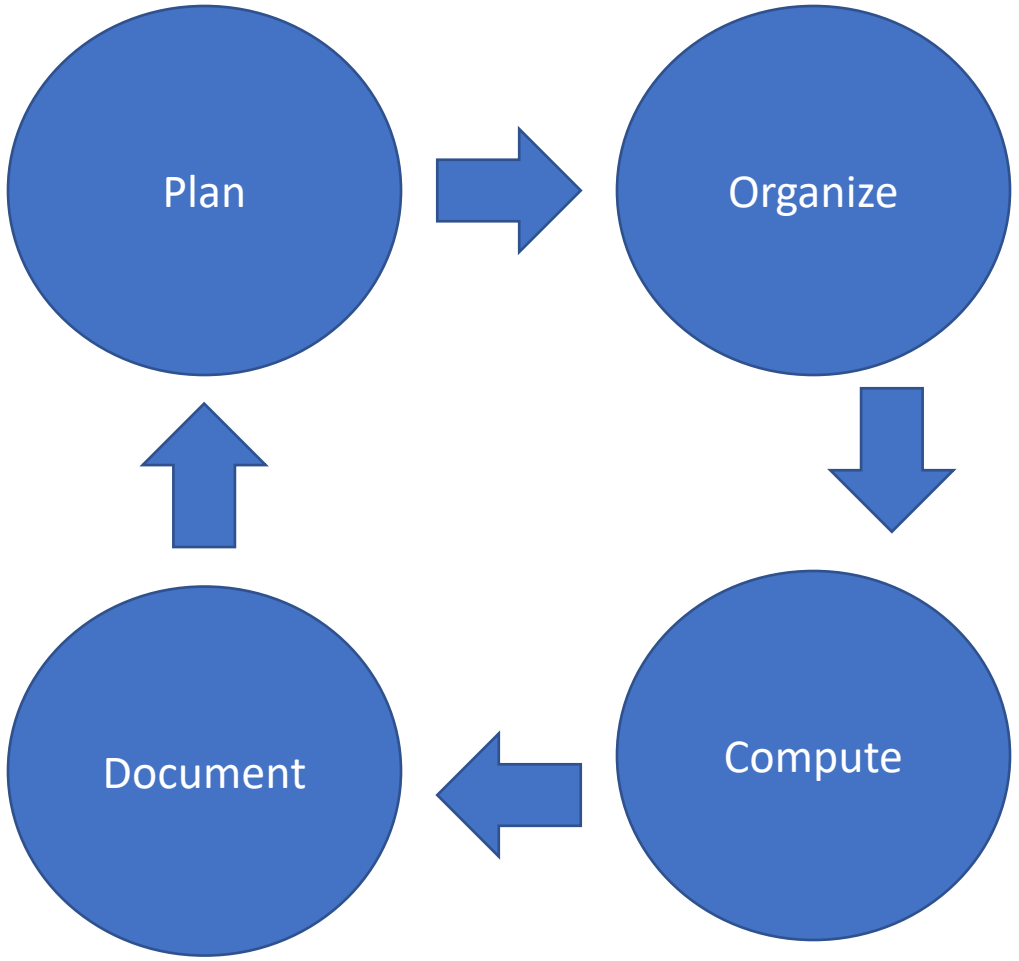
Workflow of Data

HackBio

Why plan?
So that one remembers what is to be done and not end up confused

Why document?
For reproducibility.

The data and code used to make findings must be available for an independent researcher to recreate the findings



Measures of Data

- **Measures of central tendency**
  - Median
  - Mode
  - Mean (Arithmetic Average)

- **Measures of dispersion**
  - Range
  - Interquartile range (IQR)
  - Variance
  - Standard deviation

Measures of central tendency

- A summary measure that attempts to describe whole set of data with a single value that represents the middle or the centre of its distribution

- Sometimes referred to as **Measures of central location**; classified as summary statistics

- They include:
- **Median:** Middle value in an ordered set of data, not influenced strongly by outliers
  - For **odd n**, median is in position: **n+1/ 2**
  - 12,0,5,13,0,0,5,10,5,1,5,6,7,5,7,8,10,5,11,14
    - **Arrange in order**
  - 0,0,0,1,5,5,5,5,5,**5,6**,7,7,8,10,10,11,12,13,14
  - **Median?**

Continue

- **Mode**
- Value with the highest frequency, most commonly occurring value in the dataset

- Not a very good measure when the most common mark is very far away from the rest of the data

- What is the mode?

- Mode is **13**

Value	Count
10	3
11	1
12	2
13	6
15	4
18	2
14	4

Continue

- **Mean** (also Arithmetic average)
- Sum of all values divided by the total number of values (n)
  - $x_1 + x_2 + x_3 ... + x_n / n = \sum x_i / n$
- It includes every value in the dataset

Continue

- *E.g* - 0,0,0,1,5,5,5,5,5,**5,6**,7,7,8,10,10,11,12,13,14
- Mean = 129/20 = 6.45

- **Limitation:**
  - Mean includes all values and thus is **influenced by outliers**

  - For instance:
  - 0,0,0,1,5,5,5,5,5,5,6,7,7,8,10,10,11,12,13,14, **55,60**
  - Mean 244/29 = 11.09

Measures of Dispersion

- Measure of spread (variability) of data

- The variability in a sample displays how the observations spread out from the average

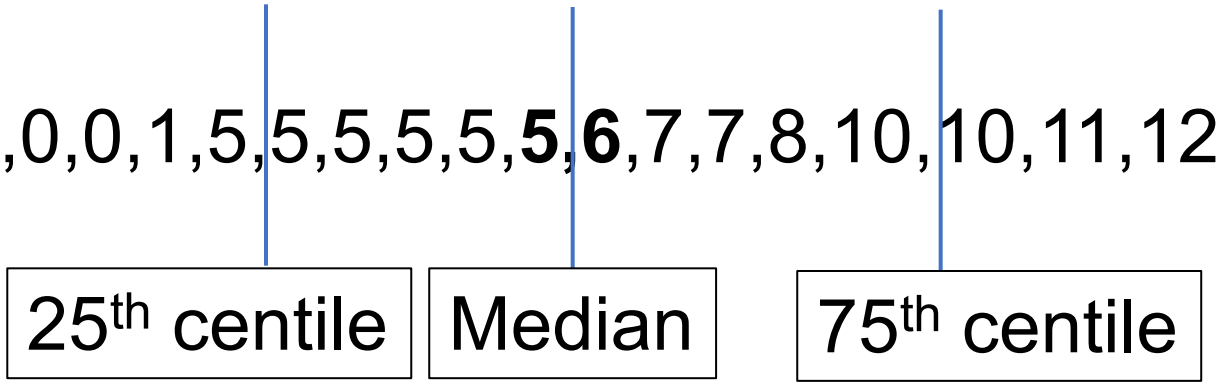
- **Range:**
- Entails the difference between the lowest and the highest values in an ordered dataset.

- 0,0,0,1,5,5,5,5,5,**5,6**,7,7,8,10,10,11,12,13,14
  - Range=14-0=14

Interquartile Range

HackBio

- Shows the spread of the middle 50%
- 0,0,0,1,5,5,5,5,5,**5**,**6**,7,7,8,10,10,11,12,13,14

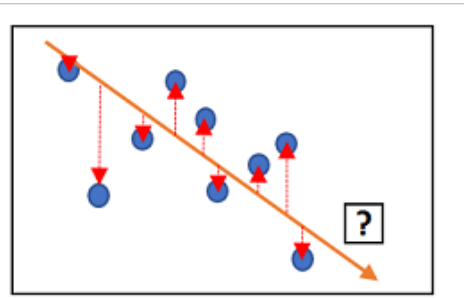


- IQR: 5 - 10

Standard deviation

- Measure that summarizes the amount by which every value within a dataset varies from the mean.

- Variance is calculated by summing the squares of the deviation's from the mean and dividing by n- (second moment)



Standard deviation

Mean squared deviations from the mean.

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

Standard deviation (sd).

$sd = \sqrt{s^2}$.

Standard deviation

HackBio

$$sqrt(\sum (x_i - x)^2/(n-1))$$

staff_id	x	x-mean	$(x\text{-mean})^2$
A	0	-6.45	41.6025
B	0	-6.45	41.6025
C	0	-6.45	41.6025
D	1	-5.45	29.7025
E	5	-1.45	2.1025
F	5	-1.45	2.1025
G	5	-1.45	2.1025
H	5	-1.45	2.1025
I	5	-1.45	2.1025
J	5	-1.45	2.1025
K	6	-0.45	0.2025
L	7	0.55	0.3025
M	7	0.55	0.3025
N	8	1.55	2.4025
O	10	3.55	12.6025
P	10	3.55	12.6025
Q	11	4.55	20.7025
R	12	5.55	30.8025
S	13	6.55	42.9025
T	14	7.55	57.0025
			346.95

Statistical distribution

- **Distribution** number of times a given quantity occurs in a dataset.

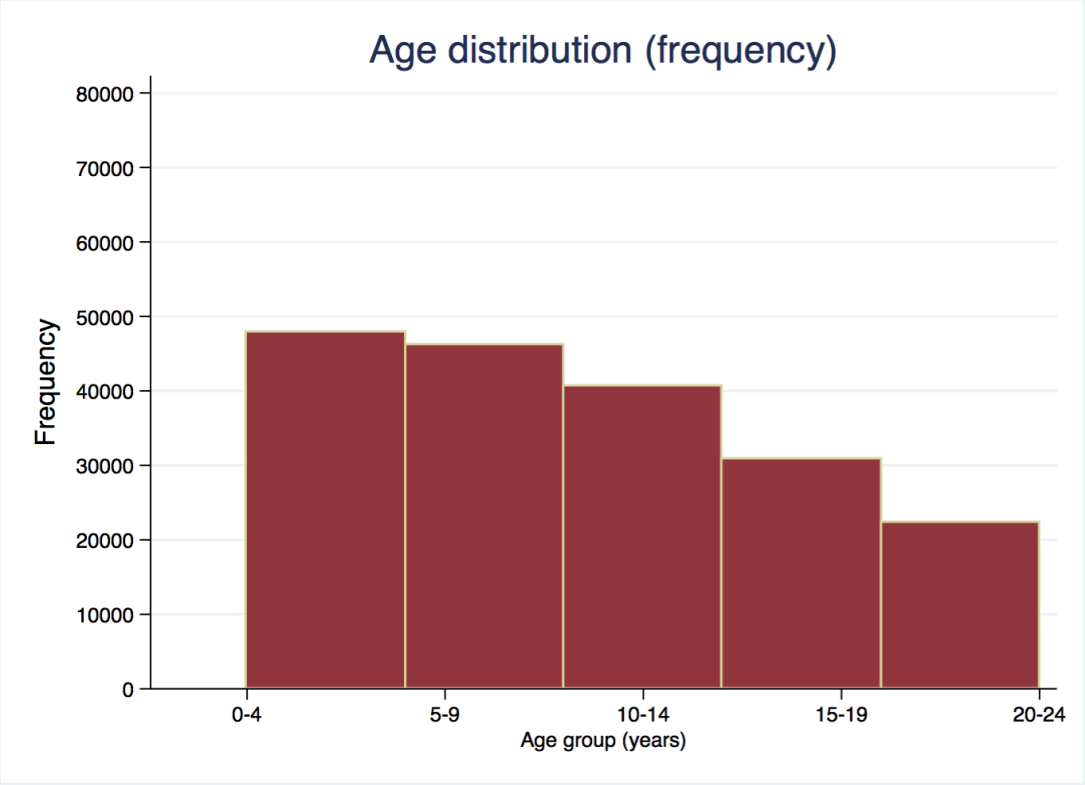
- Can be represented in two ways:
- **Frequency distribution**

- **Probability distribution**

Frequency distribution

HackBio

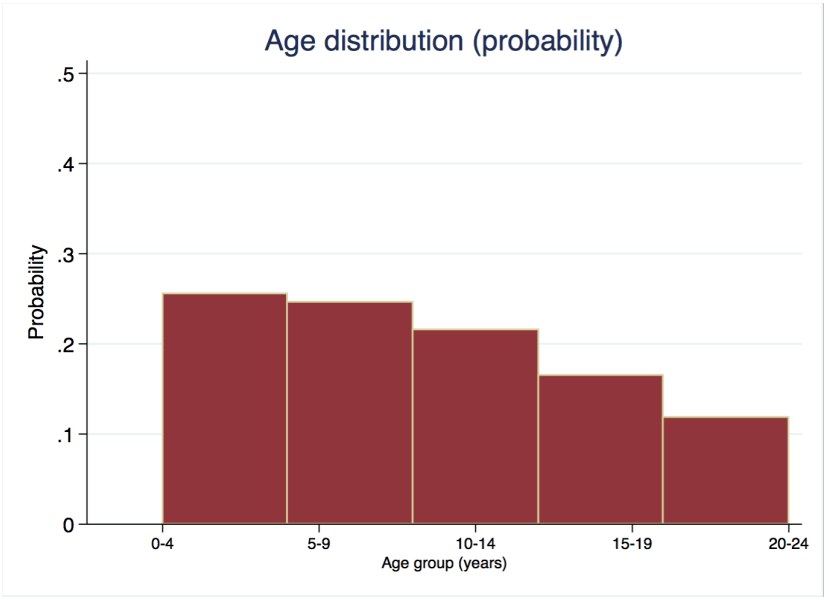
Age group (yrs)	Frequency
0-4	47,964
5-9	46,267
10-14	40,573
15-19	30,926
20-24	22,274



Probability distribution

HackBio

Age group (yrs)	Frequency	Proportion (%)
0-4	47,964	25.51
5-9	46,267	24.61
10-14	40,573	21.58
15-19	30,926	16.45
20-24	22,274	11.85

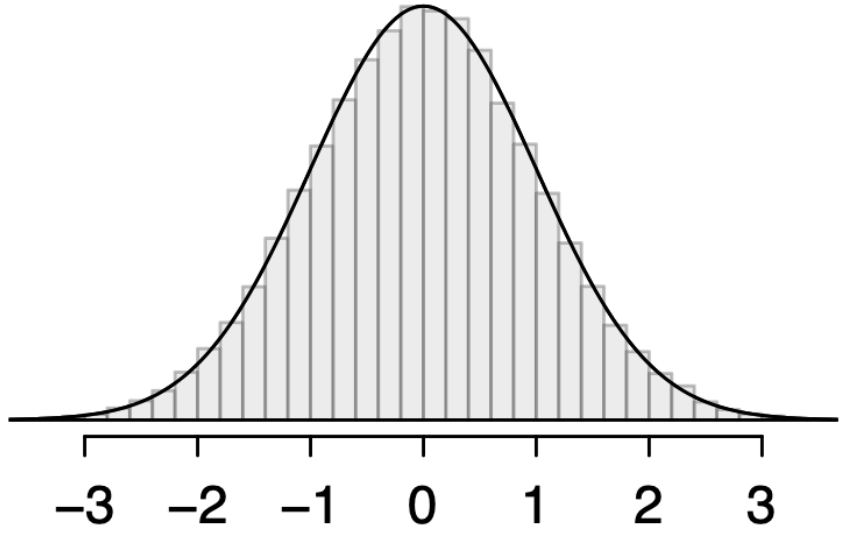


Normal distribution

HackBio

- For any given normal distribution characterized by by mean and standard deviation, can be represented as N(μ,σ): N(0,1)

- **Characteristics**
  - Bell shaped
  - Symmetrical about the mean
  - Mean, median and mode are ~ equal



Binomial distribution

- Entails multiple Bernoulli trials
- Each observation represents one of the two outcomes(success or failure)
- The same action is repeated n times
- Probability of success is denoted as p, same for every trial
- Trials are independent
- Examples:
  - Outcome of a coin toss
  - Outcome of a new drug

Poisson distribution

- Useful for estimating the number of events in a large population over a unit time

- **Characteristics**

- Used to describe **discrete** quantitative data e.g counts

- E.g number of accidents along the highway per week

Other distribution

- Uniform
- Exponential
- Geometric
- Hypergeometric
- Gamma
- Chi - square

Probability

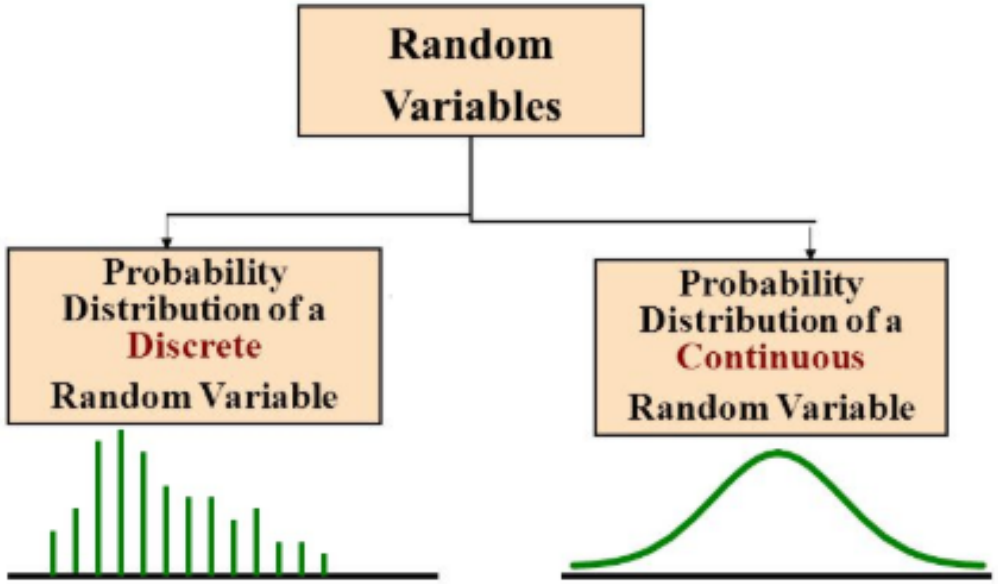
- Probability is a measure of the expectation that an event will occur
- Probability ranging between (0 and 1)
- The higher the probability the more certain we are that the event will take place or that the statement is true

Probability distribution

- Statistical function which links or lists all the possible outcomes a random variable can take, in any random process with its corresponding probability of occurrence

- Values of random variables changes based on the underlying probability of distribution

- Classified into two:
  - **Discrete probability distribution**
  - **Continuous probability distribution**



Continue

- Those that can take up any value between two values are referred to as **continuous variables** otherwise is **discrete variables**.

- E.g. weight between 150 and 200 pounds; **continuous variable**

- Toss a coin repeatedly and count the number of times we see the head; **discrete variable**

Discrete Probability distribution

- If a random variable is a **discrete variable**, its probability distribution is called **discrete probability distribution**

- Suppose you flip a coin twice, there are four possible outcomes: HH,HT,TT and TH

Continuous Probability distribution

- If a random variable is a **continuous variable** its probability is called a **continuous probability distribution**

- Involves an equation to compute for the continuous probability distribution referred to as **probability density function**

HackBio