Distribution & Hypothesis Testing

Statistics Tutorial

Day (3) + 4

Prabesh Dhakal
29 April 2019

REVIEW FROM LAST SESSION

1. **Data types for statistics:**
   - Qualitative Data
     - ➢ Nominal, Ordinal, Binary
   - Quantitative Data
     - ➢ Discrete: (interval)
     - ➢ Continuous: Ratio and Interval
2. **Data types in R**: character, numeric, integer, Boolean …

WHAT ARE WE DOING TODAY?

1. **Day 3 Summarized**
   - Data distribution
   - Central tendencies and dispersion
   - Box plots
2. **Probability and Probability Distribution**
3. **Hypothesis Testing & Chi-squared (χ^2) Test**

BEFORE WE START: **WORKING DIRECTORIES**

A **directory** is a fancy way of saying a folder

- In case of research projects, a **working directory** is the folder that you have created as your project folder.
- Simply put, a working directory contains your raw data and the outputs you save will be saved on the working directory.

- You set the working directory by using **`setwd()`** command
- You check what is set as your working directory by using **`getwd()`** command

(we will do this next week)

CLASS EXERCISE - 1

TASK:

Please identify the type of data each column from the data set on the left side contains:

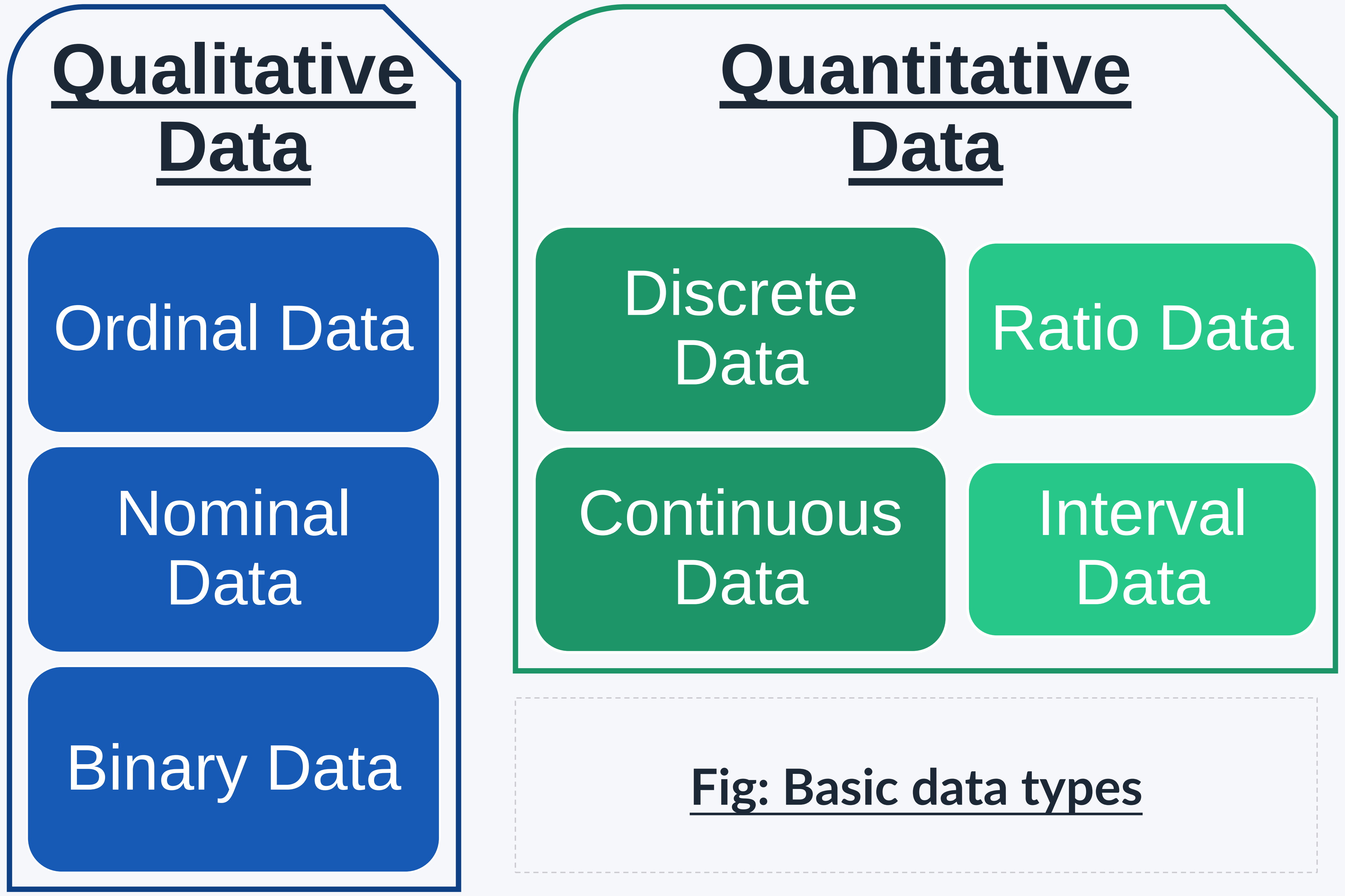
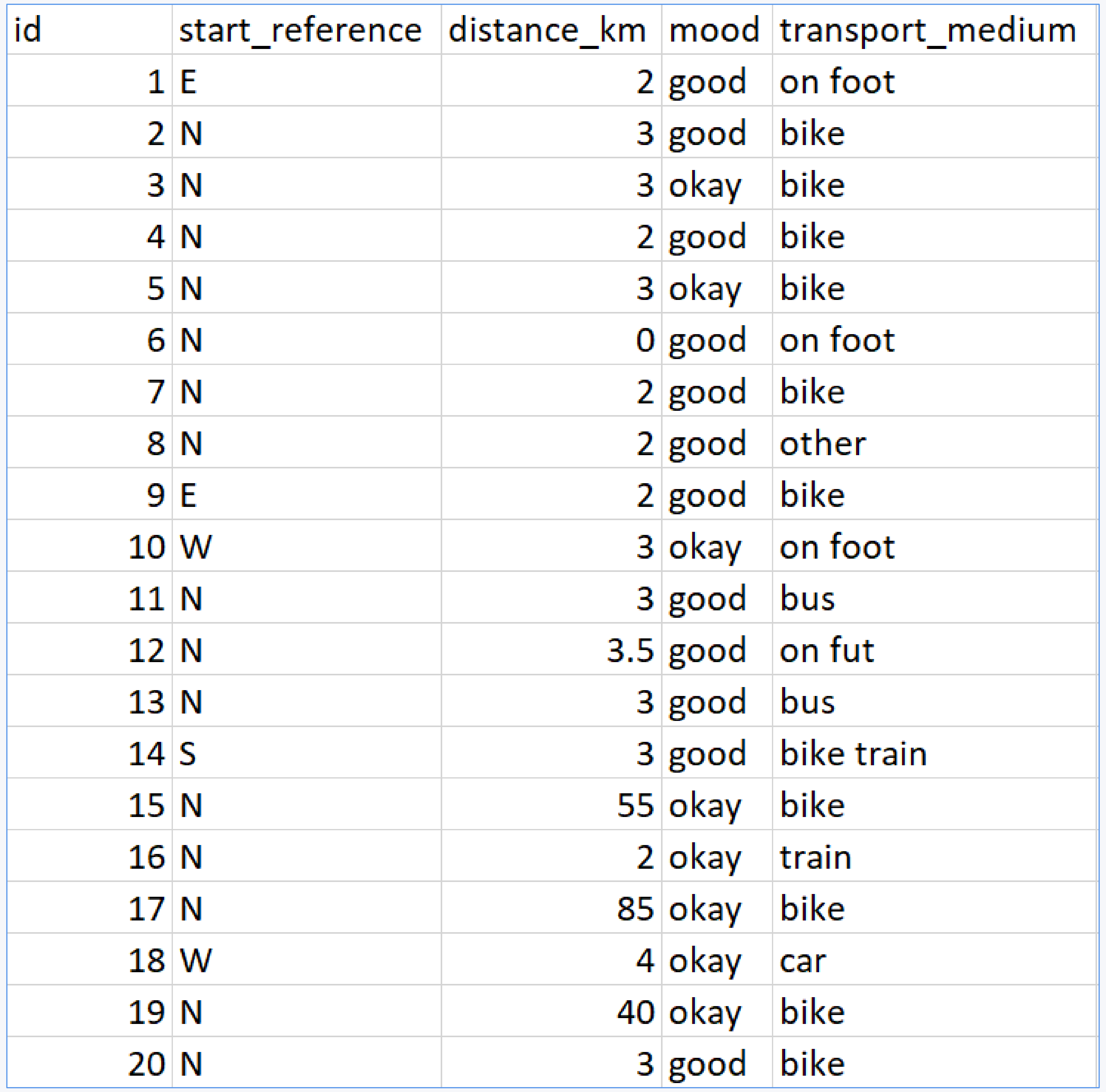


Fig: Basic data types



id	start_reference	distance_km	mood	transport_medium
1	E	2	good	on foot
2	N	3	good	bike
3	N	3	okay	bike
4	N	2	good	bike
5	N	3	okay	bike
6	N	0	good	on foot
7	N	2	good	bike
8	N	2	good	other
9	E	2	good	bike
10	W	3	okay	on foot
11	N	3	good	bus
12	N	3.5	good	on fut
13	N	3	good	bus
14	S	3	good	bike train
15	N	55	okay	bike
16	N	2	okay	train
17	N	85	okay	bike
18	W	4	okay	car
19	N	40	okay	bike
20	N	3	good	bike

DAY

3

Data Distribution

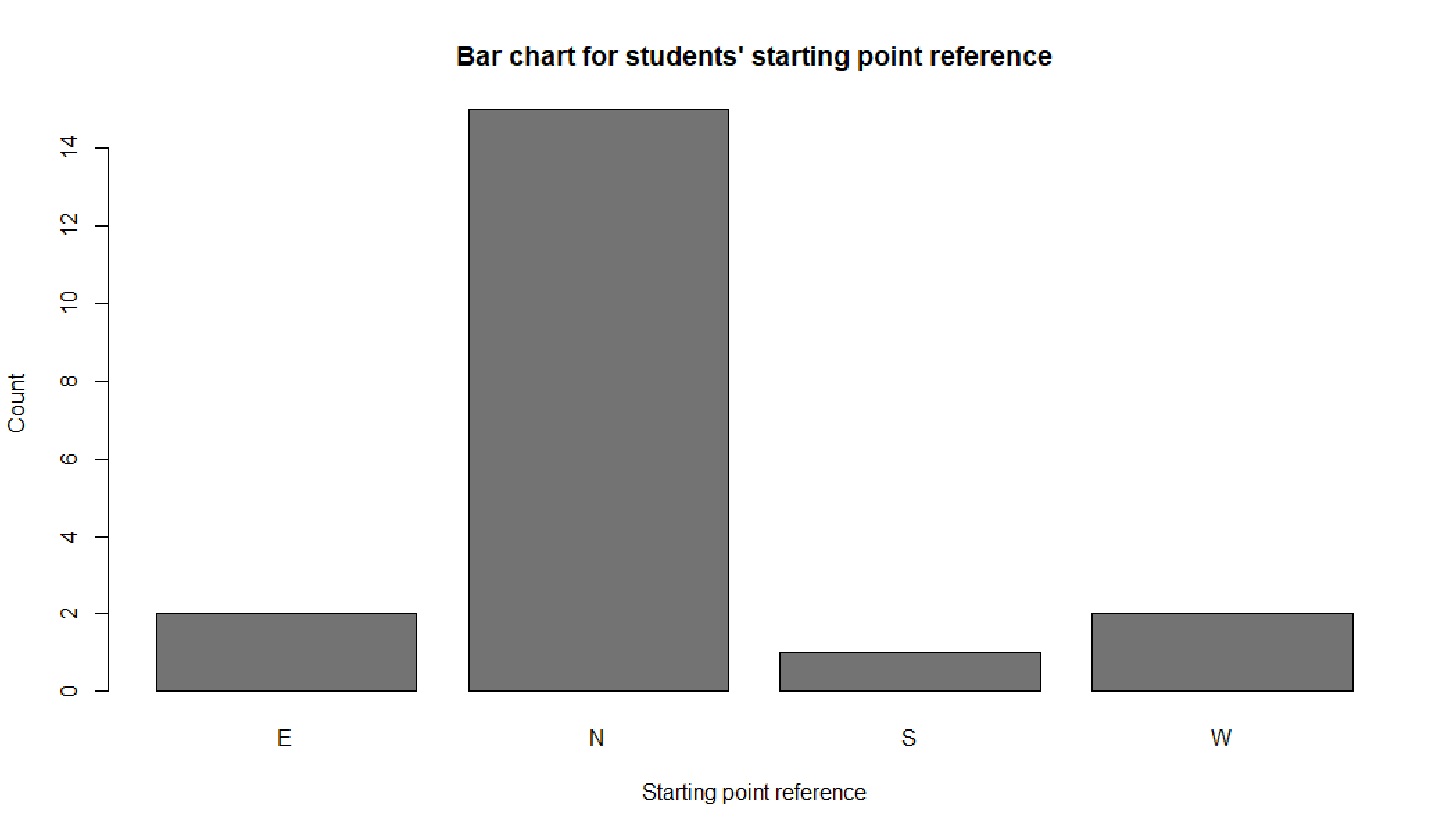
1. How to check the distribution of data
2. Measures of central tendencies and dispersion
3. Box plots

DISTRIBUTION OF THE DATA

1. **What?**
   - An arrangement of values of a variable showing their observed or theoretical frequency of occurrence
2. **Why?**
   - Shows how frequent each value is in a given data set
   - Enables us to get a better sense of the data than what just the numbers in the tables suggest
3. **How?**
   - *Discrete distribution*: bar chart
   - *Continuous distribution*: histogram

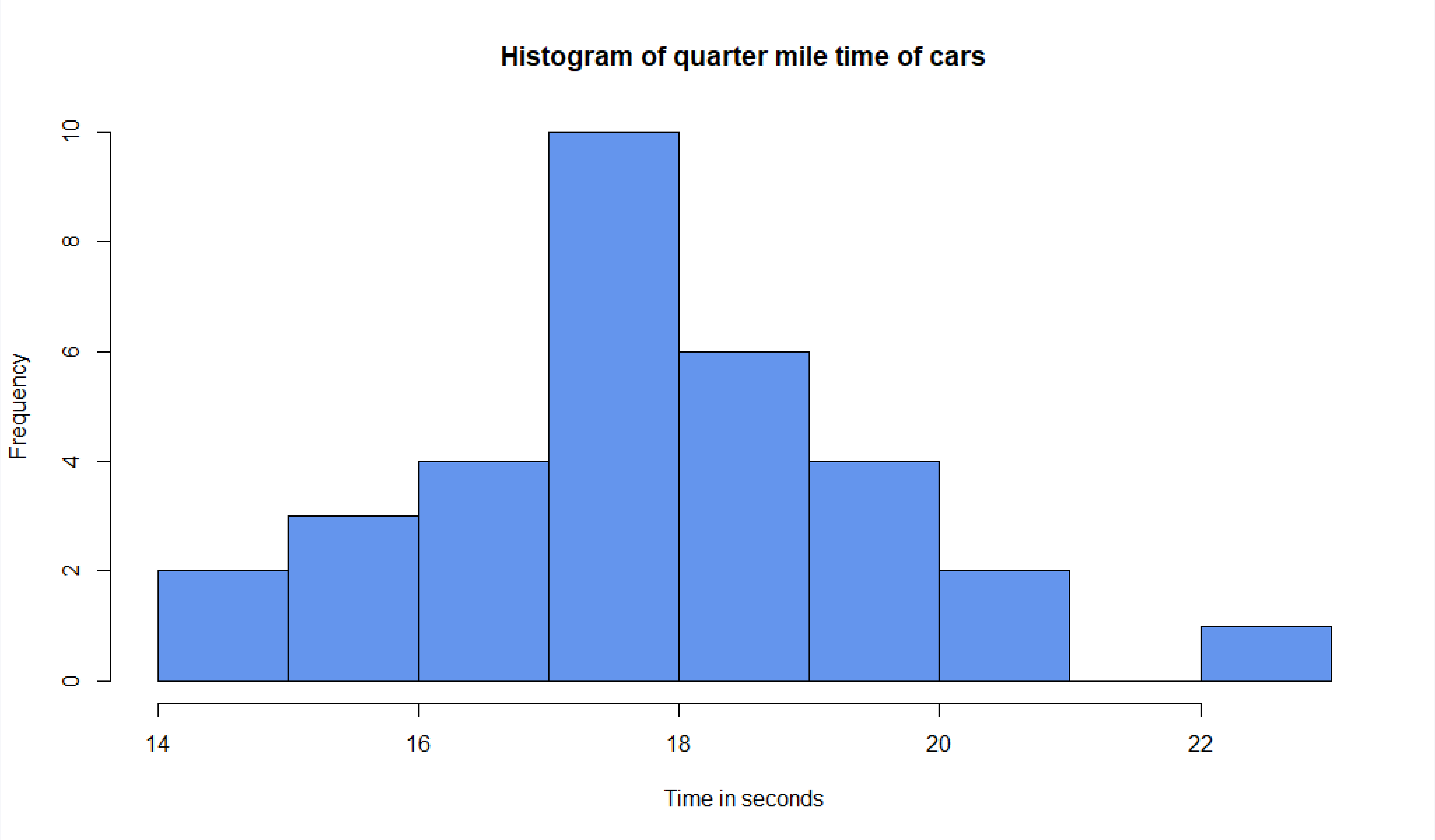
DISCRETE DISTRIBUTION

- Takes only certain values (discrete values)
- Are represented by bar charts
  - There are gaps between the bars



CONTINUOUS DISTRIBUTION

- Takes any value within some range
- Are represented by histograms
  - There are no gaps between the bars, and the distribution will look a little smoother.



BASIC PROPERTIES OF DISTRIBUTION

- All statistical distributions have inherent properties, the most basic of which are:
  - Mean
  - Median
  - Mode
  - Variance
  - Standard deviation

Good news: most of these concepts are intuitive to understand

MEASURES OF CENTRAL TENDENCIES

- Central tendencies signify the “average” of the data
  - Mode, mean, and median
- **Mode** = the most frequent number in the data
- **Mean** = arithmetic average of a set of numeric values

$$\boldsymbol{mean = \bar{x} = \frac{\sum(x)}{N}}$$

$$where, x = each\ data\ point\ and$$

$$N = total\ number\ of\ data\ points$$

MEDIAN (CENTRAL TENDENCY)

- The value whose occurrence lies in the middle of a set of observations (divides the data into two “equal” parts)
- Steps:

1. Arrange the data in an ascending order
2. If N is odd:

$$\boldsymbol{median = \left(\frac{N+1}{2}\right)^{th}\, item}$$

3. If N is even:
   - Identify the middle two numbers and take their average

$$\boldsymbol{median = \frac{\left(\frac{N}{2}\right)^{th}\, item \;+ \left(\frac{N}{2}+1\right)^{th}\, item}{2}}$$

QUARTILE

- Quartiles divide the data into 4 “equal” parts
- Median is the second quartile
- 1st Quartile = lower quartile: $\boldsymbol{Q_1 = \left(\frac{N+1}{4}\right)^{th}\ term}$
- 2nd Quartile = $\boldsymbol{Q_2 = \left(\frac{N+1}{2}\right)^{th}\ term = median}$
- 3rd Quartile = upper quartile: $\boldsymbol{Q_3 = \left(\frac{3(N+1)}{4}\right)^{th}\ term}$

MEASURES OF DISPERSION: RANGE & IQR

- Dispersion = measure of how much the data varies from the mean; e.g. range, variance, standard deviation, interquartile range

  - **Range** =

    $$\textbf{largest value} - \boldsymbol{smallest\ value} = \boldsymbol{L} - \boldsymbol{S}$$

  - **Interquartile range** = the middle 50% of the data

    $$\boldsymbol{IQR} = \boldsymbol{Q_3} - \boldsymbol{Q_1}$$

MEASURES OF DISPERSION: VARIANCE

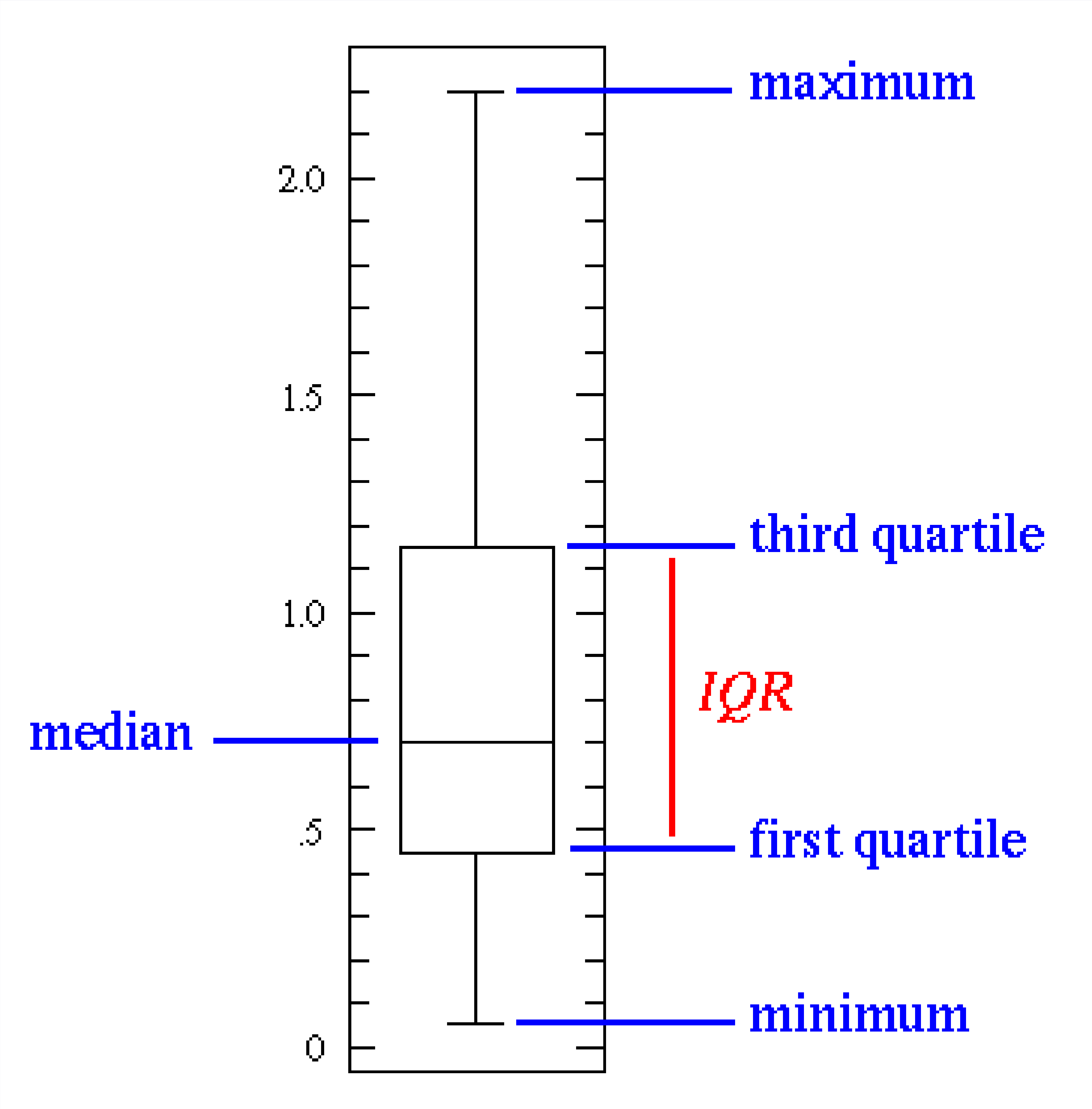
- **Variance** = a more robust, and widely accepted, measure of dispersion, and is defined as:

$$\boldsymbol{sample\ variance = s^2 = \frac{\sum(x_i - \ \bar{x})^2}{N-1}}$$

$$\boldsymbol{population\ variance = \sigma^2 = \frac{\sum(x_i - \ \bar{x})^2}{N}}$$

- **Standard** deviation (SD) = $\sqrt{variance} = \ \boldsymbol{\sigma}$ or $\boldsymbol{s}$
  - Measures the variability in the observations
  - Is easier to interpret because the values’ unit is in the scale of the data points

BOX PLOTS



- Summarize many measures of central tendencies and dispersion
- Learn more: http://www.physics.csbsju.edu/stats/box2.html

CLASS EXERCISE - 2

1. Form groups
2. Draw a box plot from this data:

17, 12, 14, 7, 8,
19, 23, 19, 10, 7,
12, 7, 12

SECTION II

Probability and Probability Distribution

PROBABILITY (NAÏVE DEFINITION)

- The likelihood of an <u>event</u> occurring
- The value lies between 0 and 1 (inclusive)

Naïve definition of probability:

- If X is an event for an experiment with a finite sample space S, probability of the event X occurring is:

$$P(X) = \frac{number\ of\ outcomes\ favorable\ to\ X}{total\ number\ of\ outcomes\ in\ S}$$

Why naïve, though?

→ This definition requires equally likely outcomes and cannot handle infinite sample space

PROBABILITY (GENERAL DEFINITION)

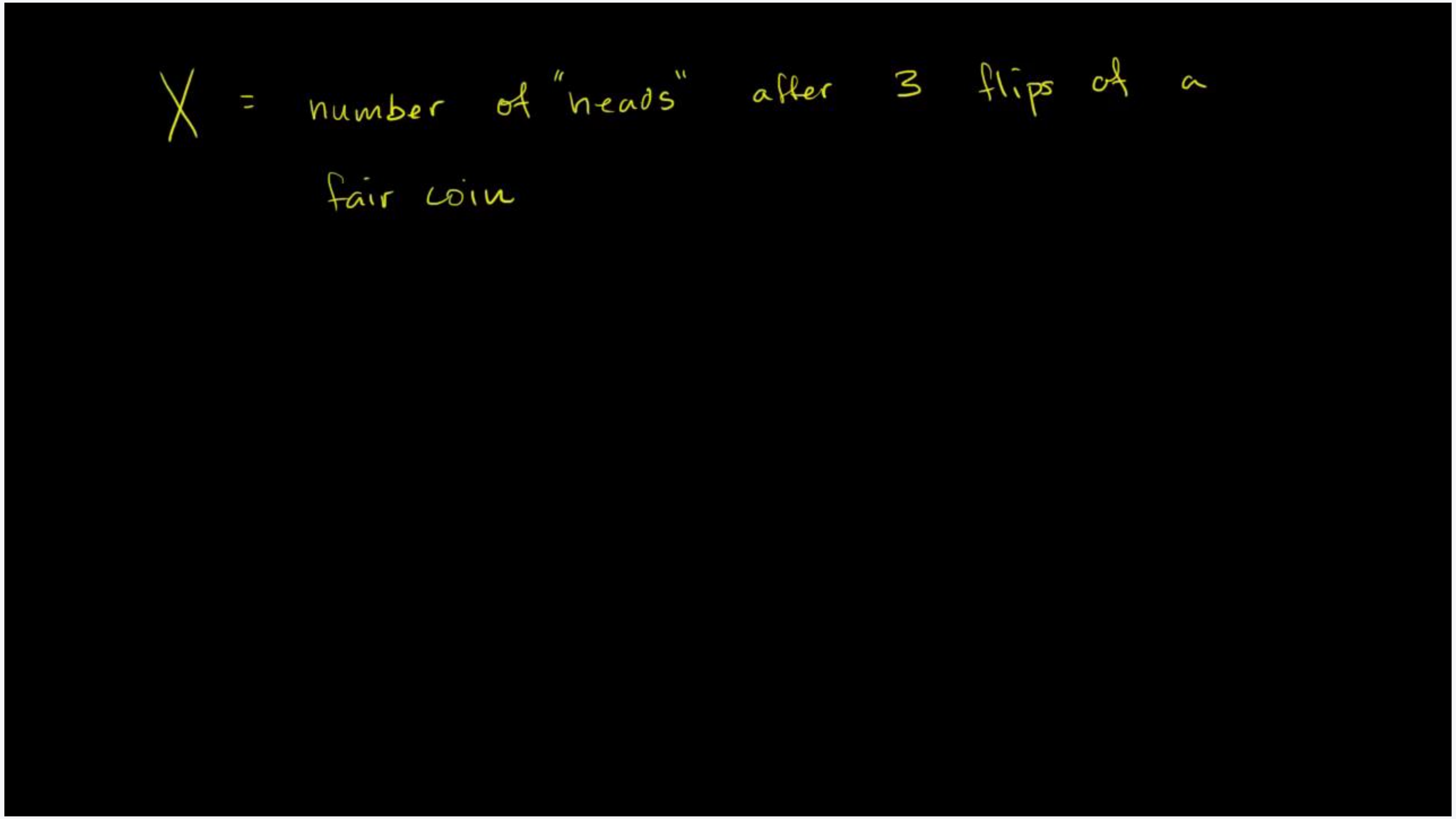
General definition of probability:

Given a *probability space* consists of a sample space S, a probability function P takes an event $X \subseteq S$ as input and returns $P(X)$, a real number between 0 and 1, as output. The function P must satisfy the following:

1. $P(\emptyset) = 0\,; P(S) = 1$
2. If $X_1, X_2, \ldots$ are mutually exclusive events, then:

$$P(all\ events\ X_i's\ occuring) = \sum_{j=1}^{\infty} P(X_j)$$

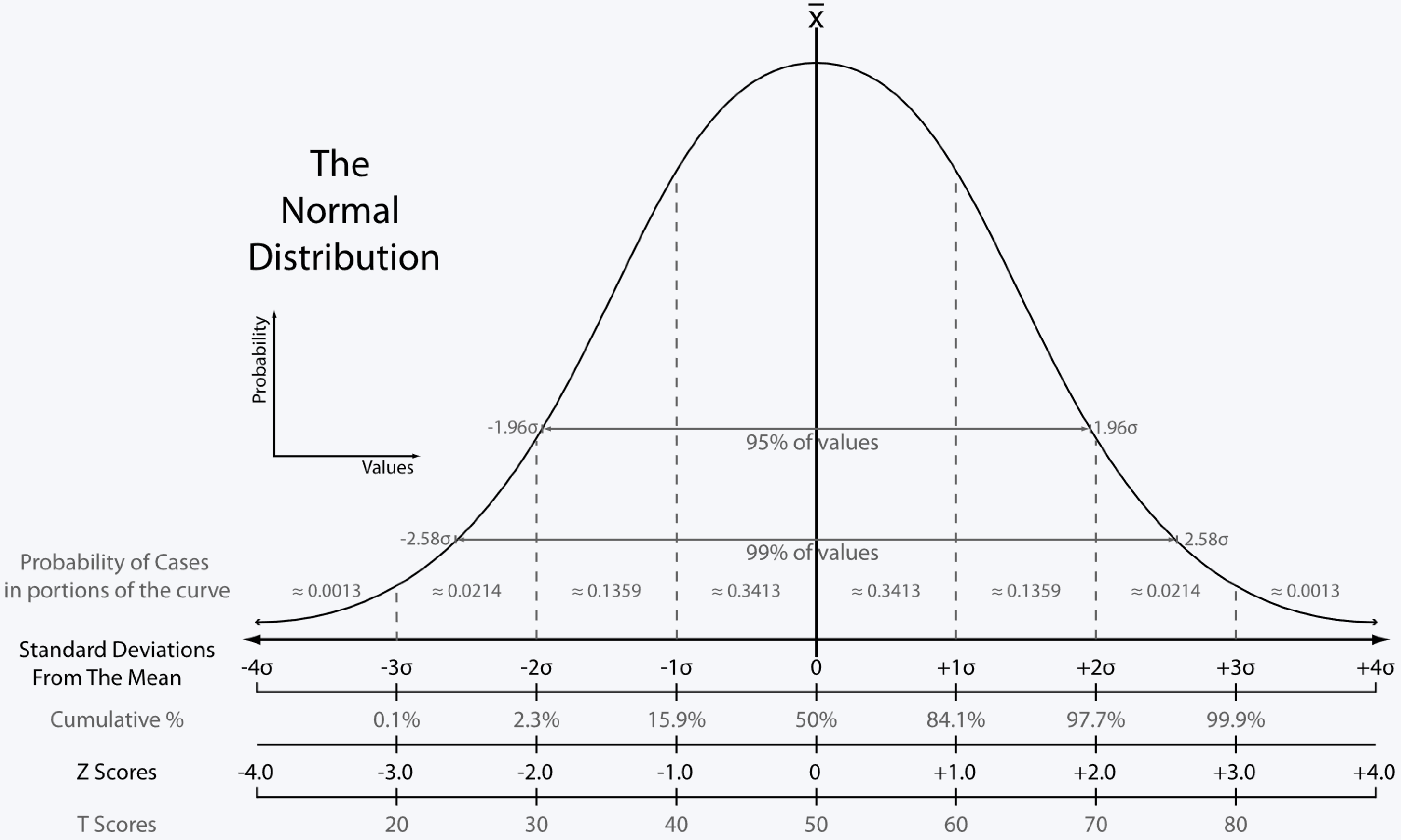
MAKING A PROBABILITY DISTRIBUTION



X = number of "heads" after 3 flips of a fair coin

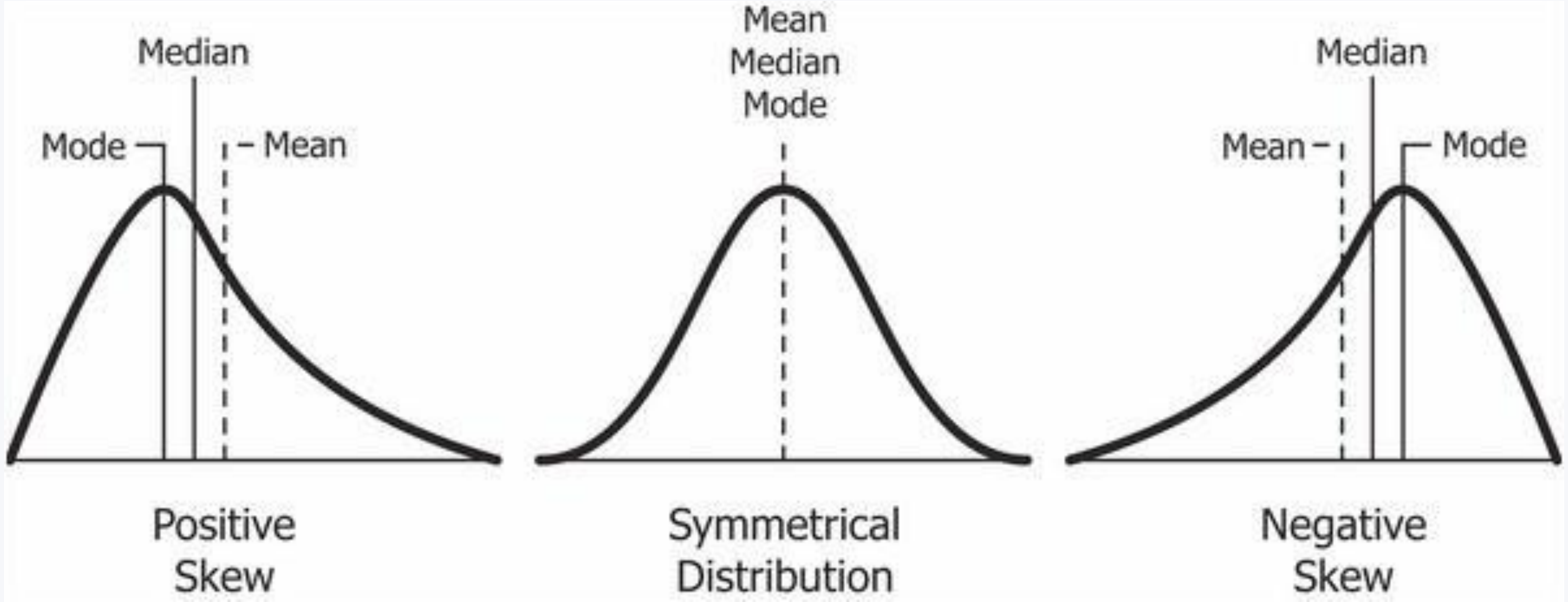
Source: https://www.youtube.com/watch?v=cqK3uRoPtk0&t=327s

The Normal Distribution



Standard Deviations From The Mean	-4σ	-3σ	-2σ	-1σ	0	+1σ	+2σ	+3σ	+4σ
Cumulative %		0.1%	2.3%	15.9%	50%	84.1%	97.7%	99.9%	
Z Scores	-4.0	-3.0	-2.0	-1.0	0	+1.0	+2.0	+3.0	+4.0
T Scores		20	30	40	50	60	70	80	

SKEWED DISTRIBUTION



SECTION III

Hypothesis Testing

HYPOTHESIS TESTING: WHAT & WHY

- A hypothesis is a proposed explanation for a phenomenon
  - Null hypothesis $\boldsymbol{H_0} = a\ statement\ about\ a\ population\ parameter$
    - Test the likelihood of this statement being true in order to decide whether to accept or reject the alternative hypothesis
    - Can include $=, \leq, or\ \ \geq$ sign.
  - Alternative hypothesis $\boldsymbol{H_1} = a\ statement\ that\ contradicts\ the\ null\ hypothesis$
    - Only true when null hypothesis is rejected
    - Can include a $\neq\ ,\ >\ ,\ or < $ sign.
- Why perform hypothesis tests?

  → To determine whether there is enough statistical evidence in favor of a hypothesis

Hypothesis can only be rejected, they cannot be verified based on data.

UNCERTAINTY AND ERRORS IN HYPOTHESIS TESTING

		Actual Situation	
		H_0 **True**	H_0 **False**
Experimenter's Decision	Reject H_0	Type I Error (reject incorrectly)	Correct
	Retain H_0	Correct	Type II Error (accept incorrectly)

χ^2 TEST VIDEO

Source: https://www.youtube.com/watch?v=WXPBoFDqNVk

χ^2 TEST (Chi-squared Test)

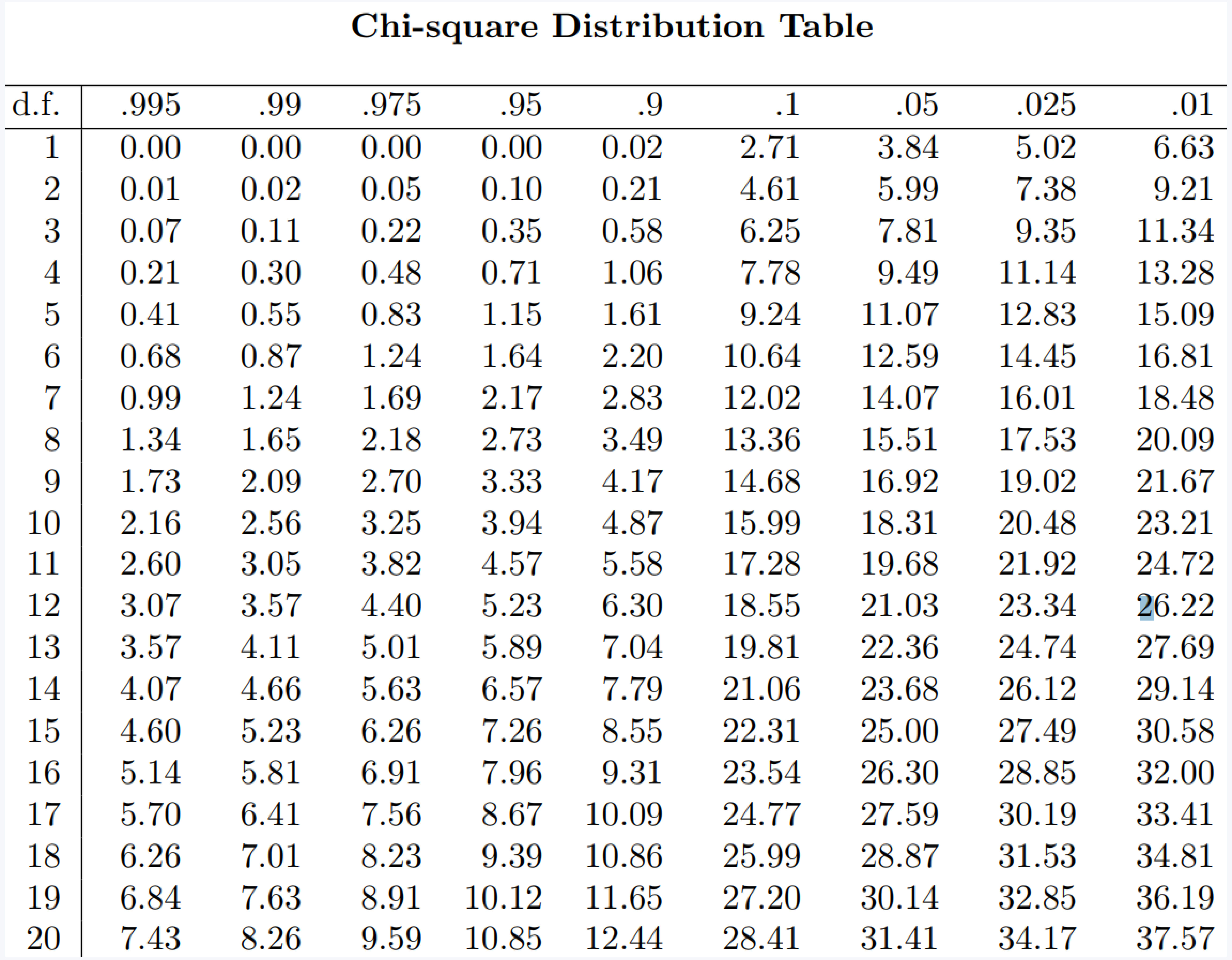
χ2 Test is used to test how likely is it that an observed distribution is due to chance/randomness.

Hypotheses:

- H_0 = *features are stochastically independent* *(patterns are random)*
- H_1 = *there is a statistically significant relationship*

Test:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - Ei)^2}{E_i}$$



Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57

Full table source: (*Click here*)

PLAN FOR NEXT WEEK

That’s it for today! :-)

Next week, we are going to discuss:

1. More on hypothesis tests
2. Correlation

If you want to reach me, mail me at:

`prabesh.dhakal@stud.leuphana.de`