Biostatistics
Week VII – Review before MT

Ege Ülgen, MD, PhD

17 November 2022

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

Statistics

- The study of data
- A discipline concerned with
  - **Collecting data** for a certain purpose
  - **Analysis** of the collected data
  - **Reaching conclusions** based on the analysis

Descriptive/Inferential Statistics

- Descriptive Statistics
  - Organization of collected data, calculation of mean and dispersion, presentation as tables, graphics, etc.
- Inferential Statistics
  - Building hypothesis concerning the population based on sample findings, hypothesis testing, interpretation.

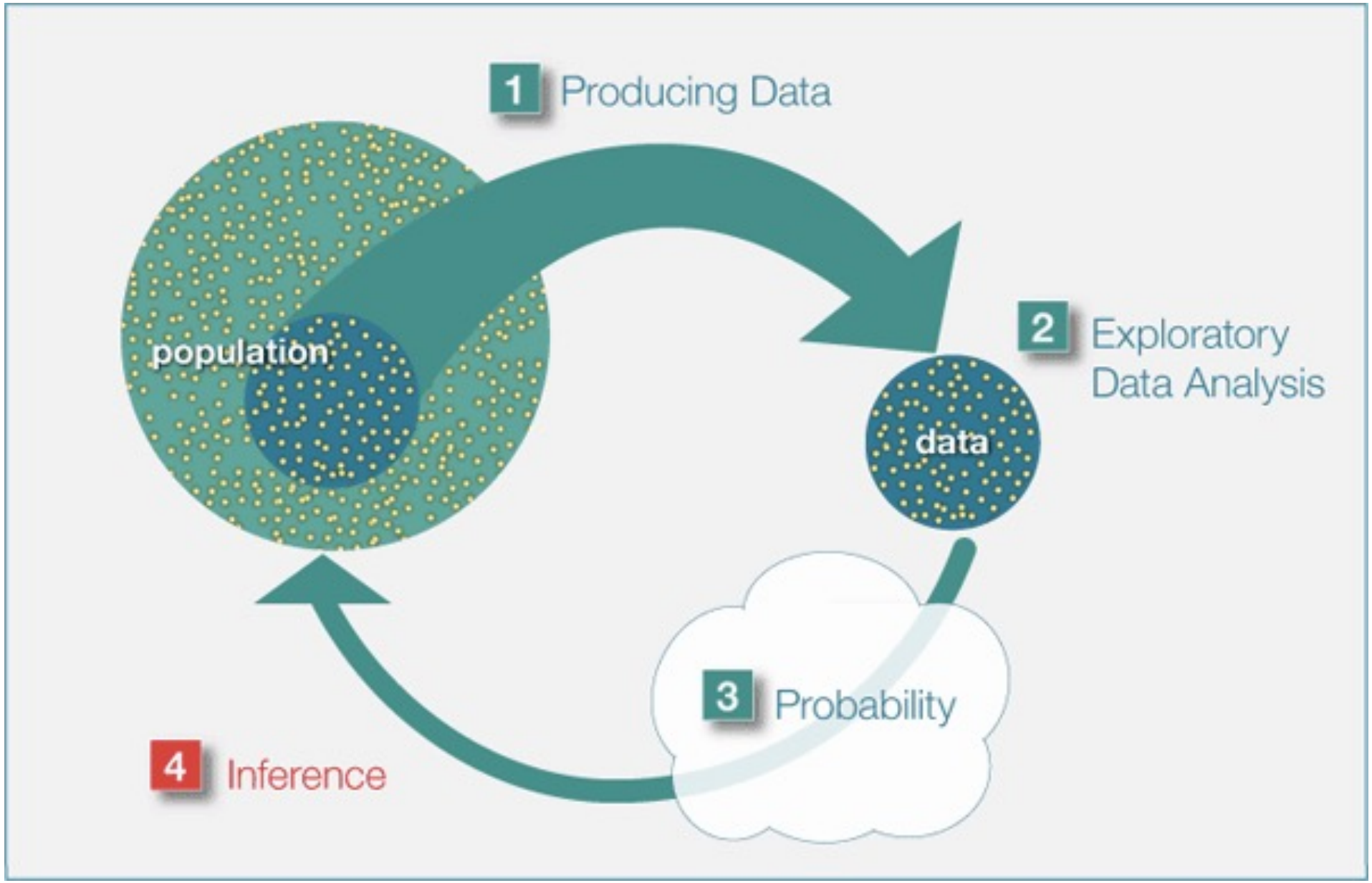
Population vs. Sample

- Population
  - All subjects under consideration that have the same properties
  - E.g., everyone living in Istanbul

  N = 15.52 million (as of 31 Dec 2019)

- Sample
  - A proportion of the population (ideally randomly selected)
  - E.g., **n** = 500, 1000, 5000, …

  (n might be decided based on sample size calculations)



Unit 4a: introduction to statistical inference [Internet]. [cited 2021 Sep 27]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-4/

Terminology/Notation

	Sample Statistic	Population Parameter
Size	n	N
Mean	$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum X}{N}$
Variance	$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$	$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$
Standard Deviation	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$
Proportion	$\hat{p} = \frac{n \text{ of successes}}{n \text{ of trials}}$	$p = \frac{N \text{ of successes}}{N \text{ of trials}}$

Variable Types

- **Discrete/Categorical/Qualitative**
  - Measured in a discrete manner

  - **Nominal**: no natural ordering. E.g., eye color, zip-code
  - **Dichotomous/binary**: only takes two values. E.g., dead/alive, female/male
  - **Ordinal**: natural ordering. E.g., agree/neutral/disagree, bad/fair/good
  - **Count**: counted values. E.g., number of tumor occurrences in one month

Variable Types

- **Continuous/Quantitative**
  - Measured in a continuous manner

  - **Interval:** real number (+/- including 0). E.g., temperature, location
  - **Ratio:** positive values (**0 indicates none**). E.g., height, age, daily calcium consumption (mg).

Example Study (cont.)

id	treatment	age	gender	week_1	cd4_1	week_2	cd4_2
1	trt2	36.43	male	0	22	7.57	20
2	trt4	47.85	male	0	21	8.00	48
4	trt3	36.60	male	0	61	7.14	60
5	trt1	35.95	male	0	35	8.00	30
6	trt2	38.40	male	0	10	7.29	10

Discrete - nominal

Contin.- ratio

Discrete – nominal /binary

Discrete - count
Contin. - ratio

Applied longitudinal analysis [Internet]. [cited 2021 Sep 27]. Available from: https://content.sph.harvard.edu/fitzmaur/ala2e/

Exploratory Data Analysis (EDA)

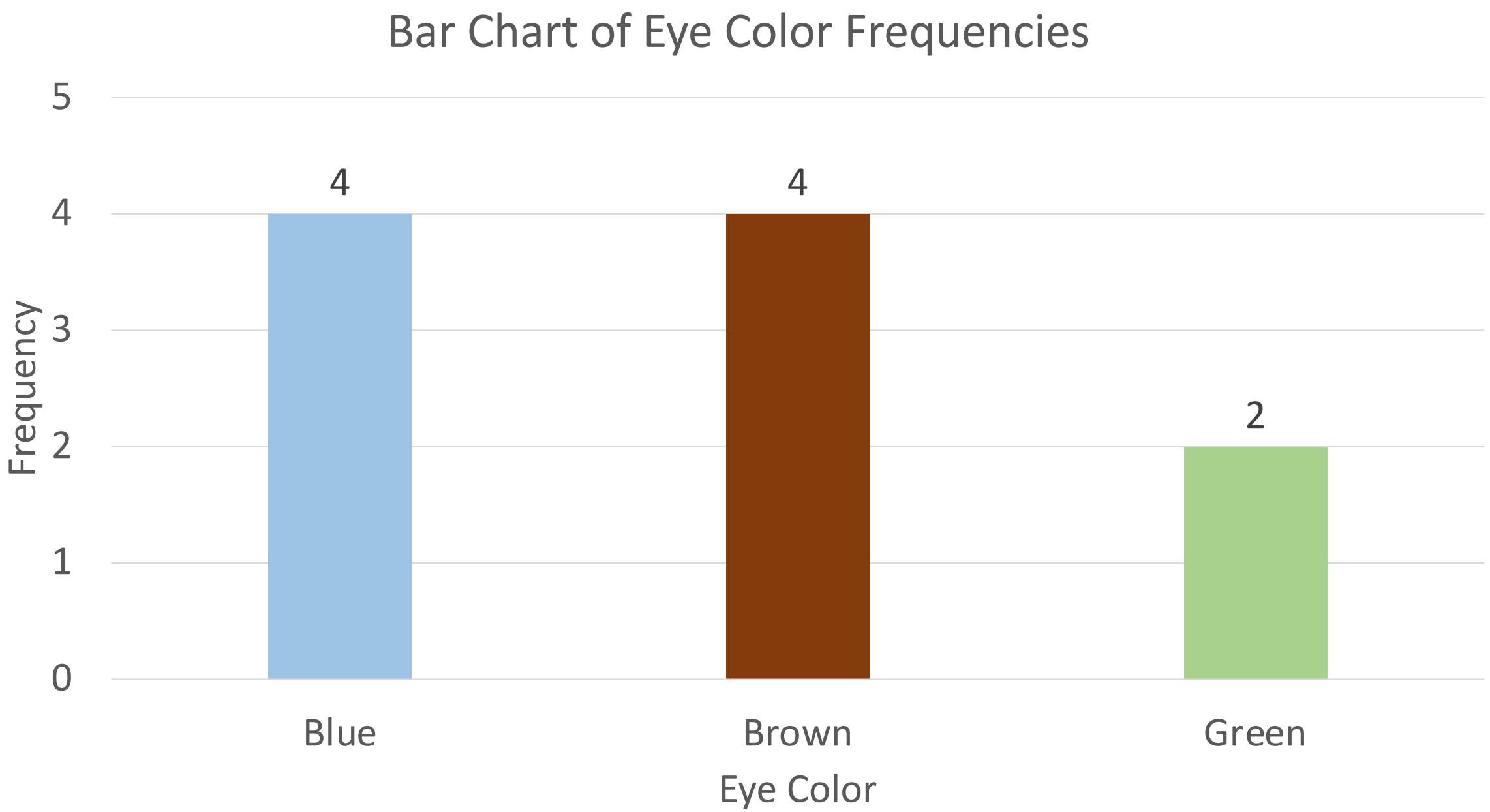
- **Examining Distributions — exploring data one variable at a time**
- Examining Relationships — exploring data two variables at a time

Frequency Tables – Categorical Variable

- Eye colors of 10 individuals:

  blue, green, brown, blue, brown, blue, blue, green, brown, brown

Eye Color	Frequency	Relative Freq.	%
Blue	4	4/10 = 0.4	40
Brown	4	4/10 = 0.4	40
Green	2	2/10 = 0.2	20


Bar Chart of Eye Color Frequencies
5
4
3
2
1
0
Frequency
4
4
2
Blue
Brown
Green
Eye Color

Frequency Tables – Continuous Variable

Cholesterol levels of 40 patients:

Original data

213, 174, 193, 196, 220, 183, 194, 200, 192, 200, 200, 199, 178, 183, 188, 193, 187, 181, 193, 205, 196, 211, 202, 213, 216, 206, 195, 191, 171, 194, 184, 191, 221, 212, 221, 204, 204, 191, 183, 227

Sorted data

171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, 227

Interval	Frequency	Relative Freq.	%
[170-180)	3	3/40 = 0.075	7.5
[180-190)	7	7/40 = 0.175	17.5
[190-200)	13	13/40 = 0.325	32.5
[200-210)	8	8/40 = 0.200	20.0
[210-220)	5	5/40 = 0.125	12.5
[220-230)	4	4/40 = 0.100	10.0

Ross, S. M. (2017). Introductory statistics. Amsterdam: Elsevier/AP.

Histogram

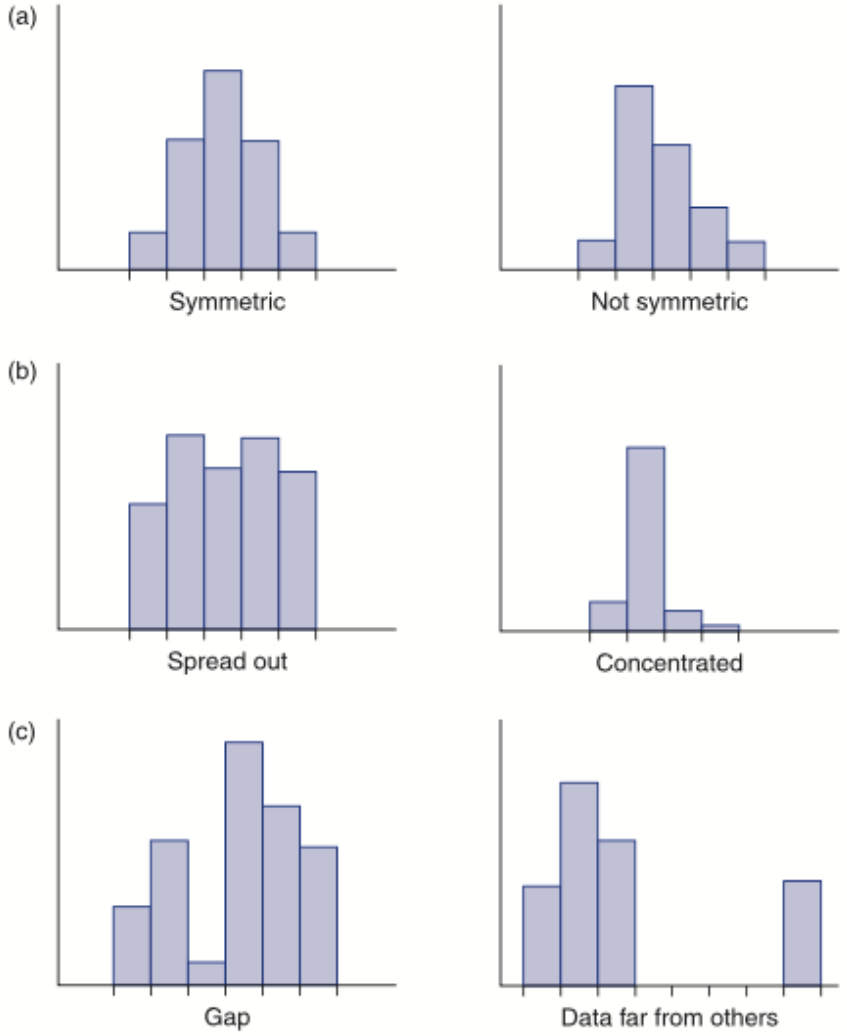


FIGURE 2.8

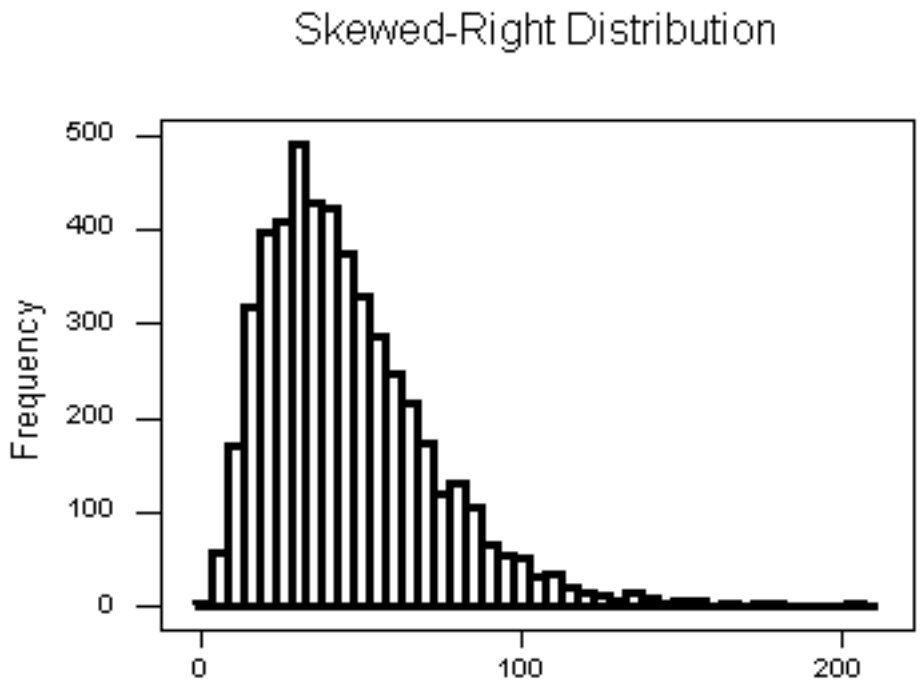
Characteristics of data detected by histograms. (a) symmetry, (b) degree of spread and where values are concentrated, and (c) gaps in data and data far from others.

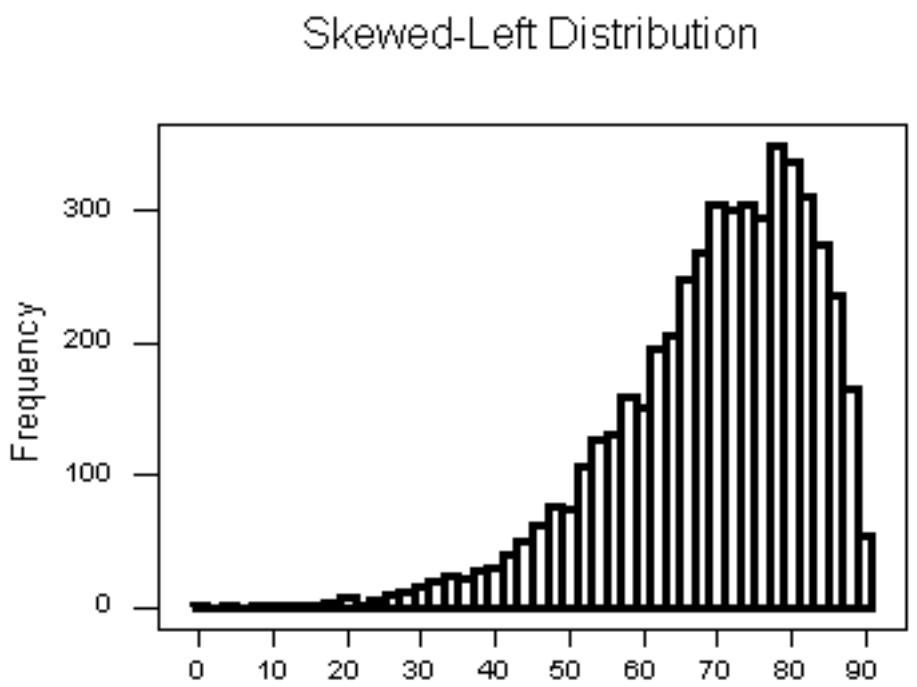
Describing a Dataset

- **Shape**
- Center
- Spread
- Outliers

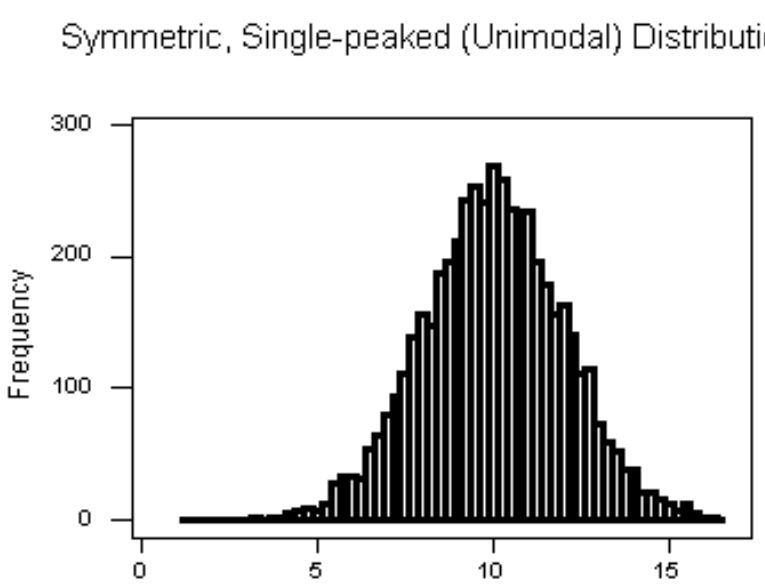
Shape

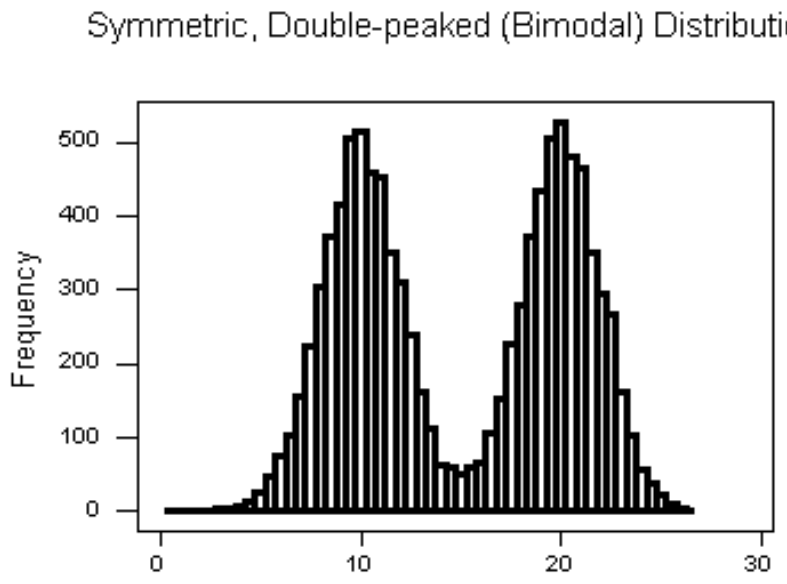
- **Symmetry/Skewness** of the distribution
- **Peakedness (modality)**
  - The number of peaks (modes) the distribution has

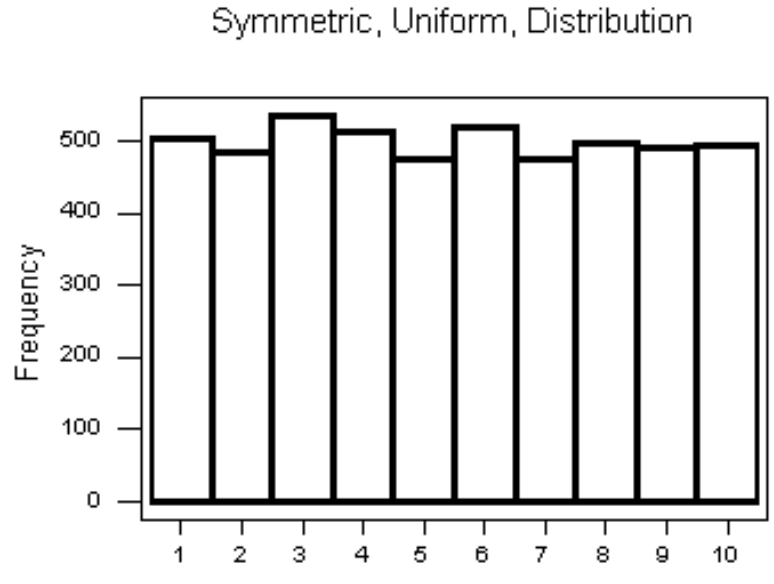




Describing distributions [Internet]. [cited 2021 Oct 1]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/







Describing distributions [Internet]. [cited 2021 Oct 1]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/

Center - Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Cholesterol levels of 40 patients:

213, 174, 193, 196, 220, 183, 194, 200, 192, 200, 200, 199, 178, 183, 188, 193, 187, 181, 193, 205, 196, 211, 202, 213, 216, 206, 195, 191, 171, 194, 184, 191, 221, 212, 221, 204, 204, 191, 183, 227

$$\bar{x} = \frac{213+174+\ldots+227}{40} = 197.625$$

Median

- It is calculated as the:
  - middle value of the sorted values (if n is odd)
  - average of two middle values of the sorted values (if n is even)

2, 5, 3, 10, 4

2, 3, 4, 5, 10 => median = 4

5, 3, 10, 4

3, 4, 5, 10 => median = 4.5

Median

Cholesterol levels of 40 patients:

Original data

213, 174, 193, 196, 220, 183, 194, 200, 192, 200, 200, 199, 178, 183, 188, 193, 187, 181, 193, 205, 196, 211, 202, 213, 216, 206, 195, 191, 171, 194, 184, 191, 221, 212, 221, 204, 204, 191, 183, 227

Sorted dataa

171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, 227

Mean = 197.625

Median = 195.5

Median

171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, **227**

Mean = 197.625

Median = 195.5

171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, **700**

Mean = 209.45

Median = 195.5

Mode

- The mode is the value that appears most often in a set of data values

- Systolic blood pressures of 12 patients:

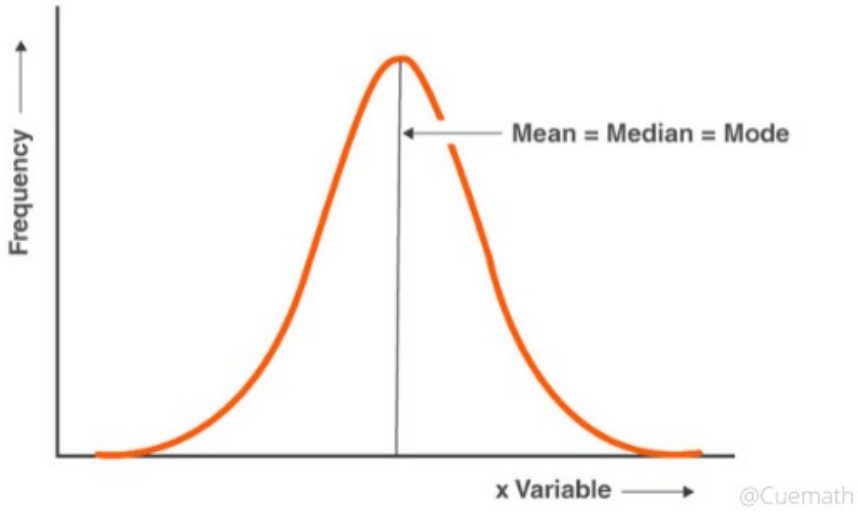
90, 80, **100**, 110, **100**, 120, **100**, 90, **100**, 110, 120, 110

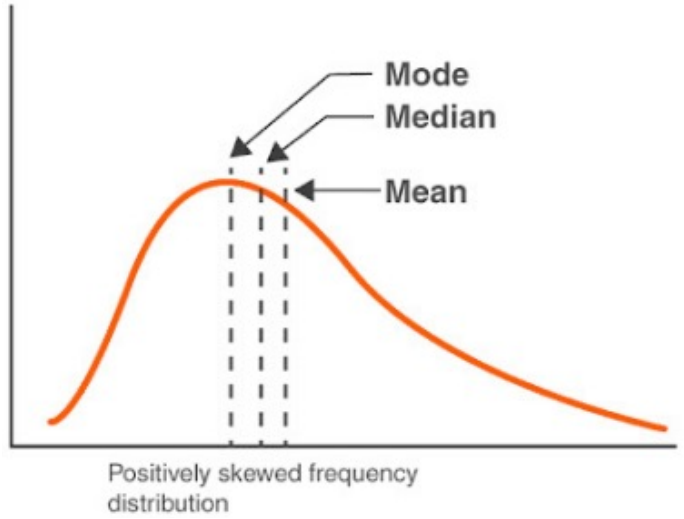
Mode = 100

Mean = 102.5

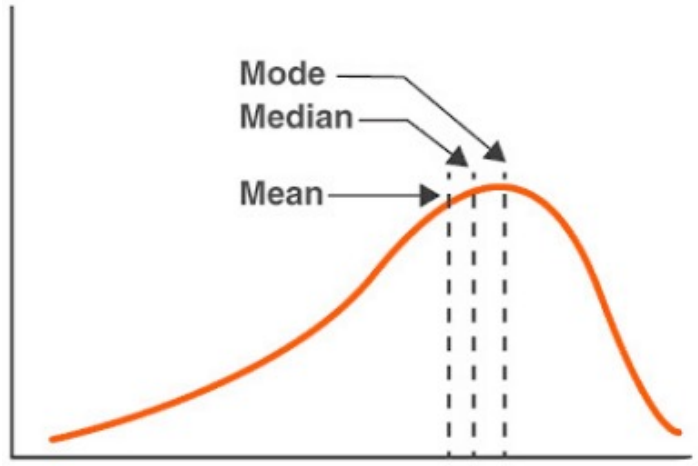
Median = 100

Mean – Median – Mode Relationship





Positively skewed frequency distribution

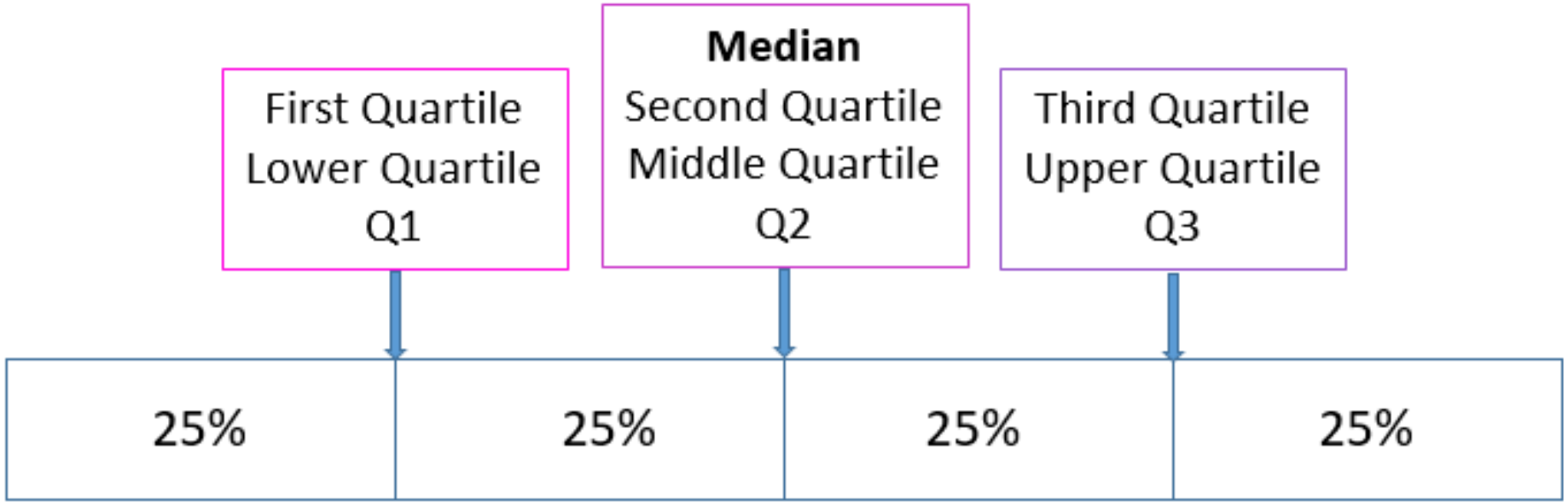


Negatively skewed frequency distribution

Describing Distributions

- Shape
- Center
- **(Measures of position)**
- Spread
- Outliers

Quartiles



Median, quartiles, percentiles(Video lessons, examples, solutions) [Internet]. www.onlinemathlearning.com. [cited 2021 Oct 4]. Available from: https://www.onlinemathlearning.com/quartile.html

Percentiles - Algorithm

1. Sort data X in ascending order
2. Calculate $n \times p$
3. If np is not an integer, return $X_{ceiling(np)}$
4. Else (if $n \times p$ is an integer), return $(X_{np} + X_{np+1})/2$

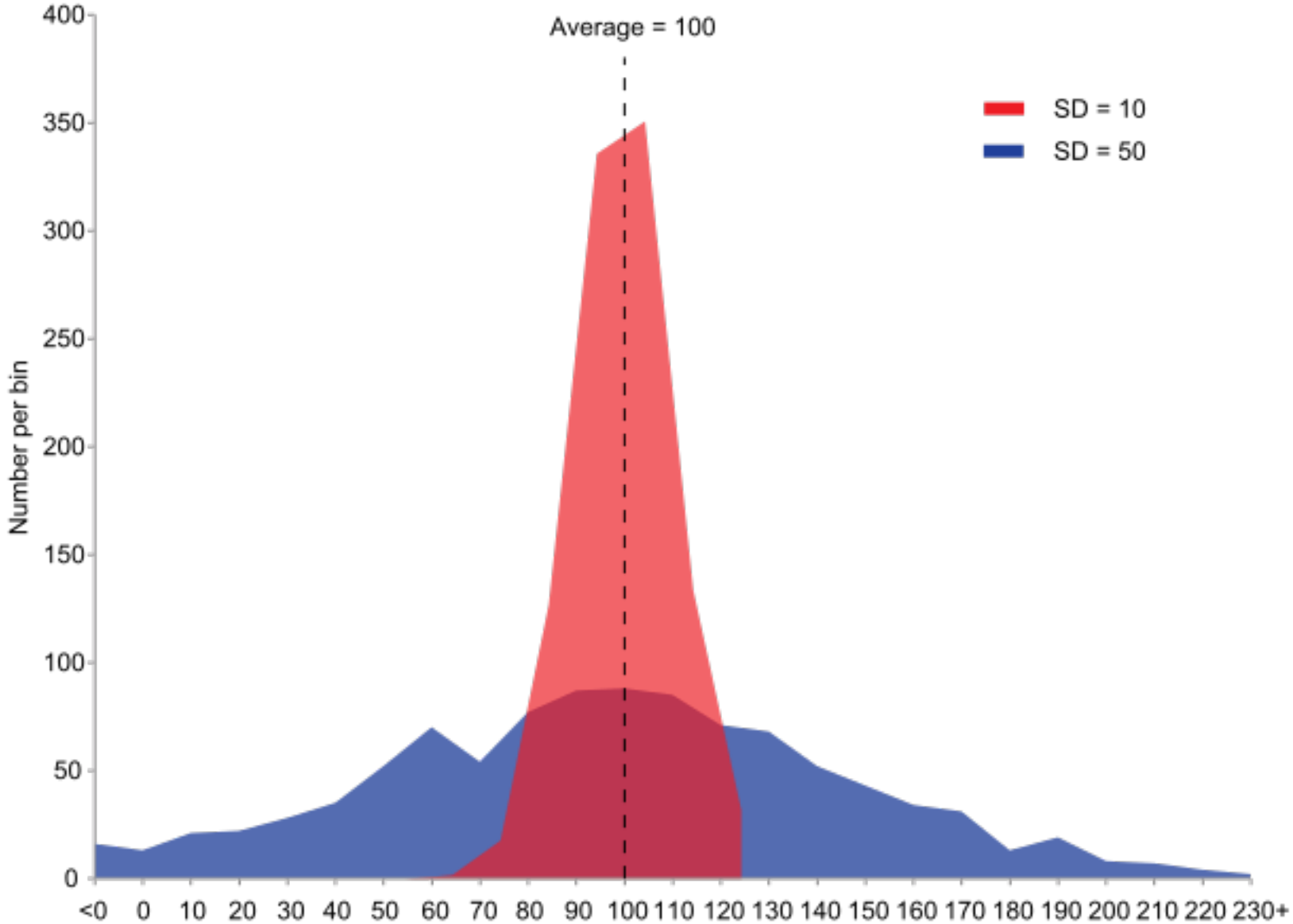
Percentiles – simple example

- Original data: 13, 14, 12, 11, 19, 15, 18, 16, 17, 20 (n = 10)
- Sorted data: 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

- 25th percentile (1st quartile): 13 (10 * 0.25 = 2.5 >> 3)
- 50th percentile (median): 15.5 (10 * 0.5 = 5 >> 5 & 6)
- 75th percentile (3rd quartile): 18 (10 * 0.75 = 7.5 >> 8)
- 90th percentile: 19.5 (10 * 0.9 = 9 >> 9 & 10

Describing Distributions

- Shape
- Center
- **Spread**
- Outliers


Average = 100
SD = 10
SD = 50
Number per bin
400
350
300
250
200
150
100
50
0
<0 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230+

Range

- The difference between the maximal and minimal value

$$R = maximum - minimum$$

e.g., The ages of 12 arthritis patients:

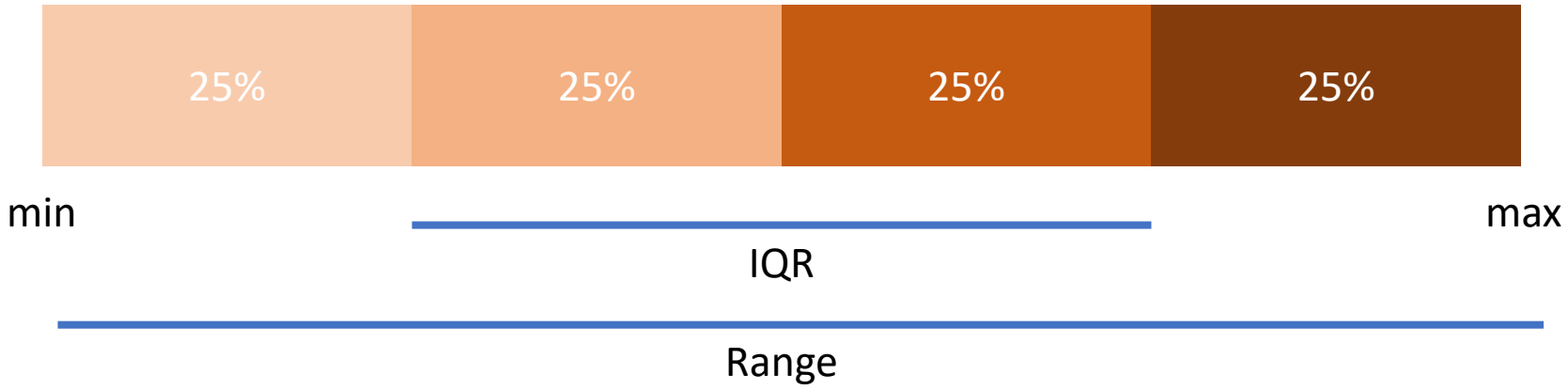
30, 12, 15, 22, 40, 55, 20, 58, 25, 60, 23, 72

$$R = 72 - 12 = 60$$

Inter-Quartile Range

- The range quantifies the variability by using the range covered by **all** the data
- the **Inter-Quartile Range (IQR)** measures the spread of a distribution by describing the range covered **by the middle 50%** of the data

$$\boldsymbol{IQR} = Q3 - Q1$$



Variance and Standard Deviation

- Variance
  - A measure of how distant observations are from the mean
  - Population variance: σ^2
  - Sample variance: s^2
- Because **the unit of variance is quadratic**, standard deviation is more widely used

- Standard deviation (sd)
  - Defined as the square-root of variance
  - Population sd: σ
  - Sample sd: s

Sample Variance

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

Describing Distributions

- Shape
- Center
- Spread
- **Outliers**

Outliers

- Extreme observations that are distant from the rest of the data

- For
  - Lower Limit = Q_1 - 1.5 * IQR
  - Upper Limit = Q_3 + 1.5 * IQR
- Outliers are defined as any value(s) larger than the upper limit or smaller than the lower limit

Outliers – Cholesterol Level Example (cont.)

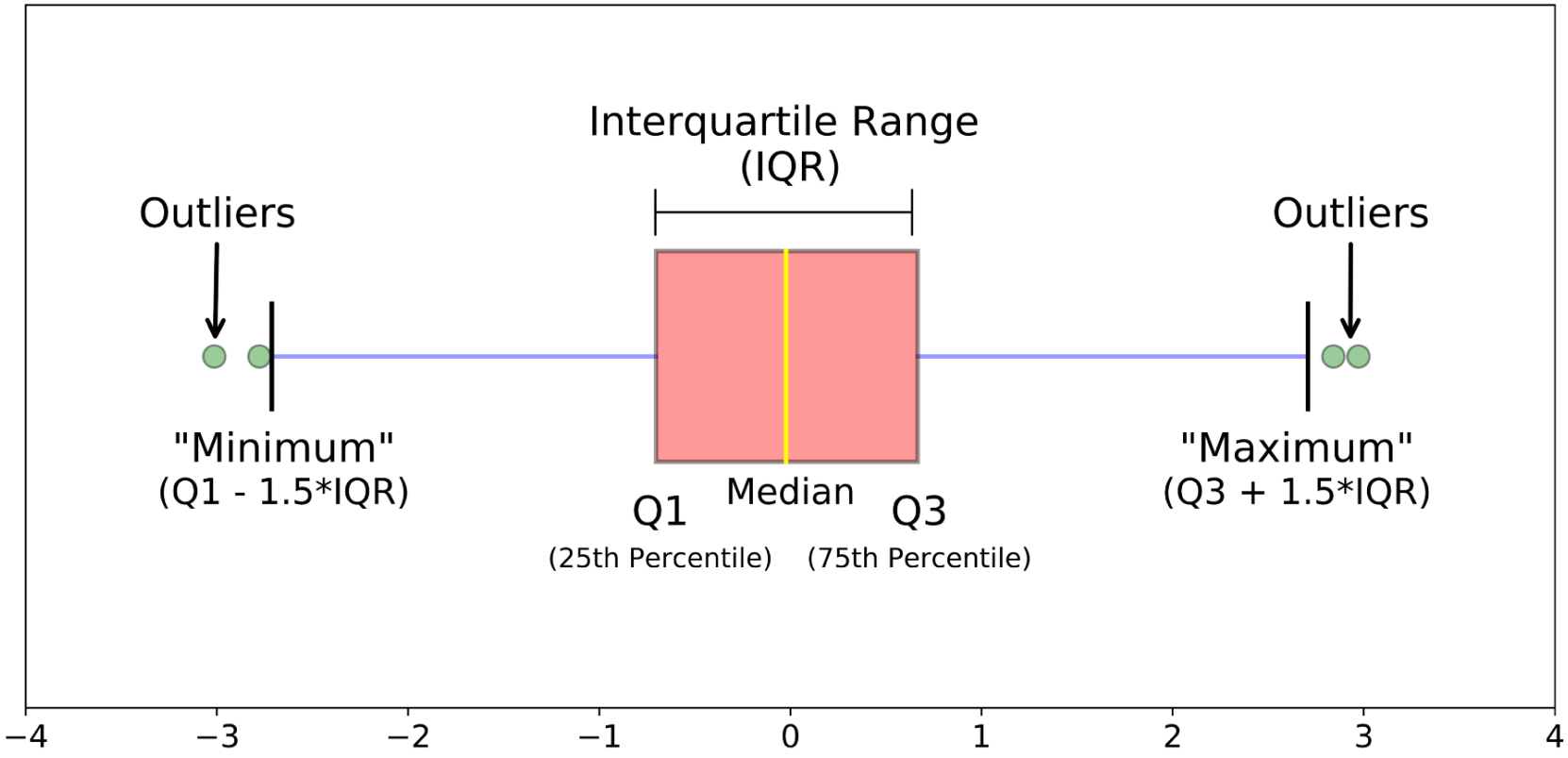
- Sorted data: 171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, **300**

- 25^{th} percentile (1st quartile, Q_1): 189.5 (40 * 0.25 = 10)
- 75th percentile (3rd quartile, Q_3): 205.5 (40 * 0.75 = 30)
- IQR = 205.5 - 189.5 = 16

- LL = Q_1 - 1.5 * IQR = 189.5 - 1.5 * 16 = 165.5
- UL = Q_3 + 1.5 * IQR = 205.5 + 1.5 * 16 = 229.5

- **300 > UL => outlier**

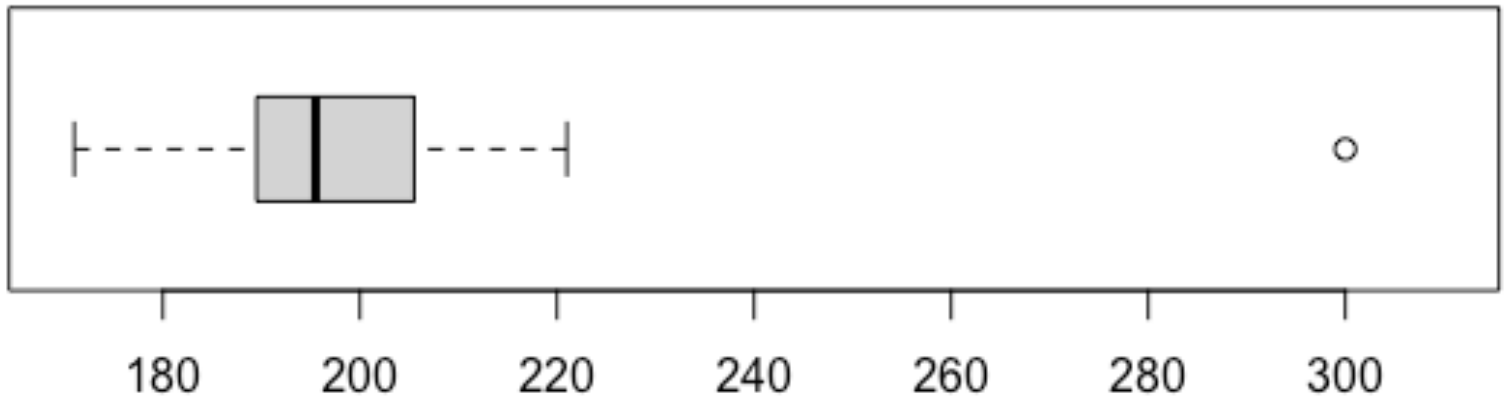
Box Plot



Galarnyk M. Understanding boxplots [Internet]. Medium. 2020 [cited 2021 Oct 4]. Available from: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Box Plot – Example

- 171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, **300**



Exploratory Data Analysis (EDA)

- Examining Distributions — exploring data one variable at a time
- **Examining Relationships — exploring data two variables at a time**

Contingency table/Cross tabulation/Crosstab

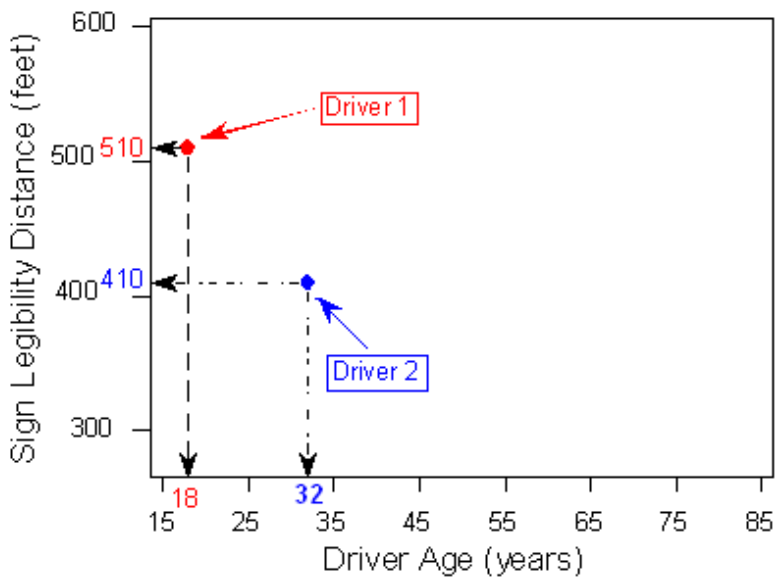
- Tables in which two categorical variables are investigated together

	Male	Female
No education	4	10
Primary school	3	5
High school	2	8
Bachelor's degree	7	9

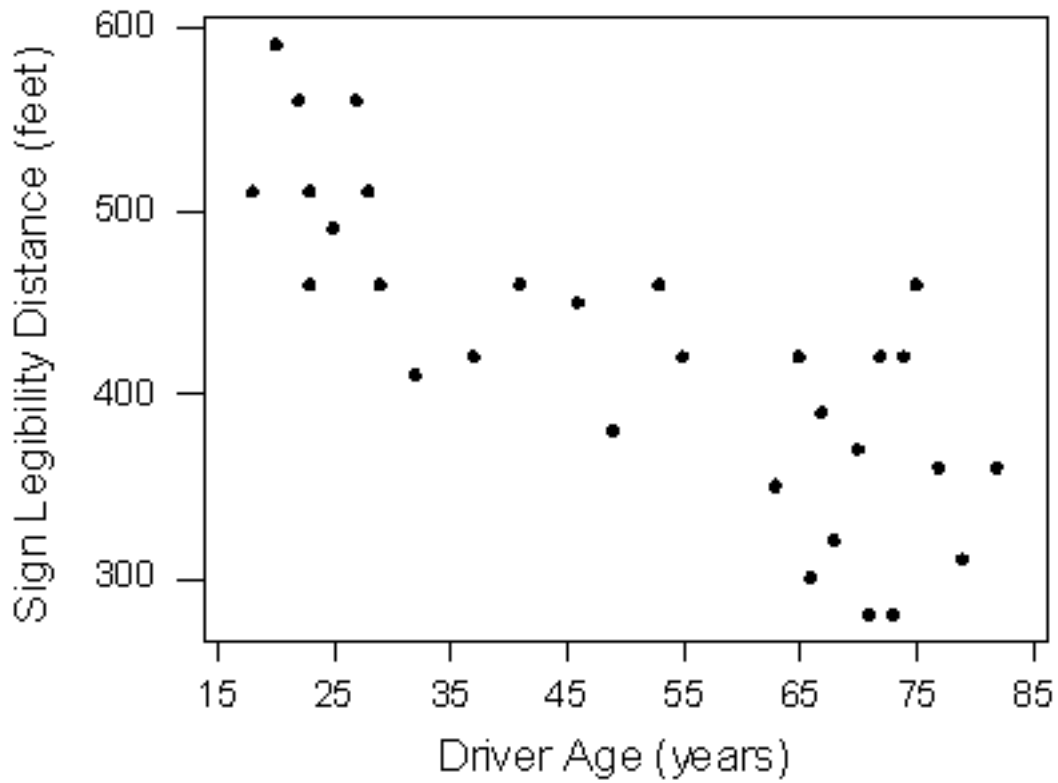
https://bolt.mph.ufl.edu/6050-6052/unit-1/case-q-q/scatterplots/

Scatter Plots

	Age (X)	Distance (Y)
Driver 1	18	510
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360



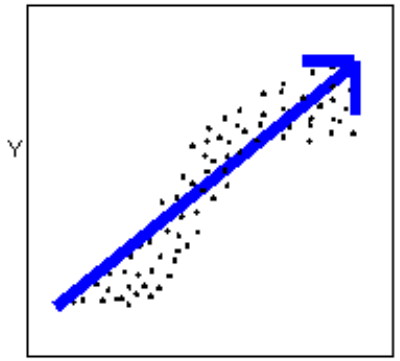
X – Explanatory
Y – Response



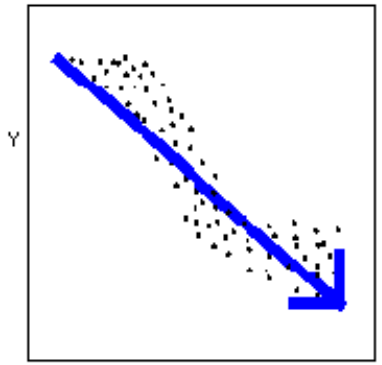
https://bolt.mph.ufl.edu/6050-6052/unit-1/case-q-q/scatterplots/

Interpreting Scatter Plots

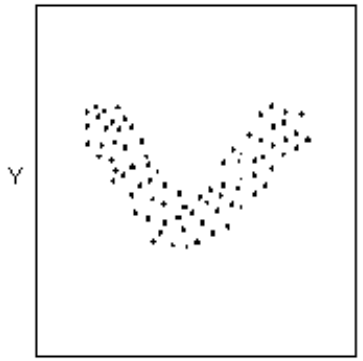
Direction



Positive relationship

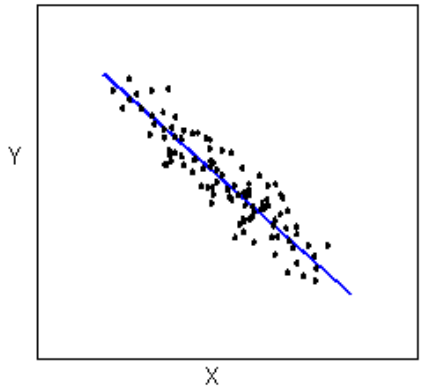


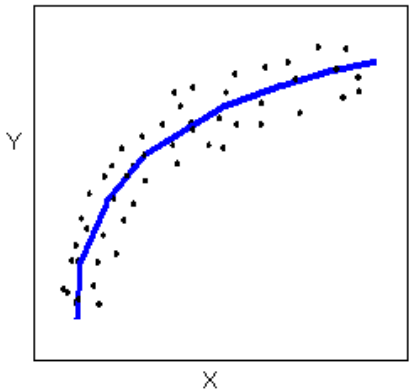
Negative relationship



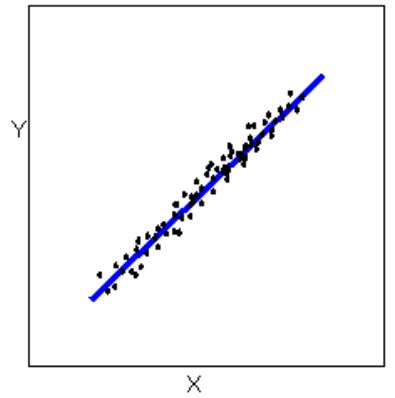
Neither positive nor negative

Form

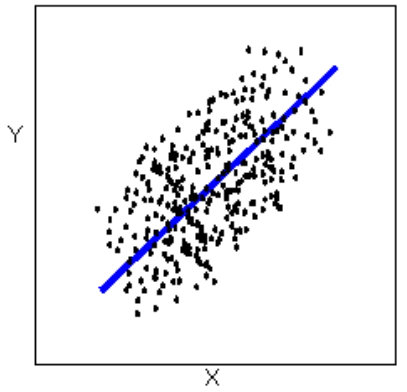




Strength



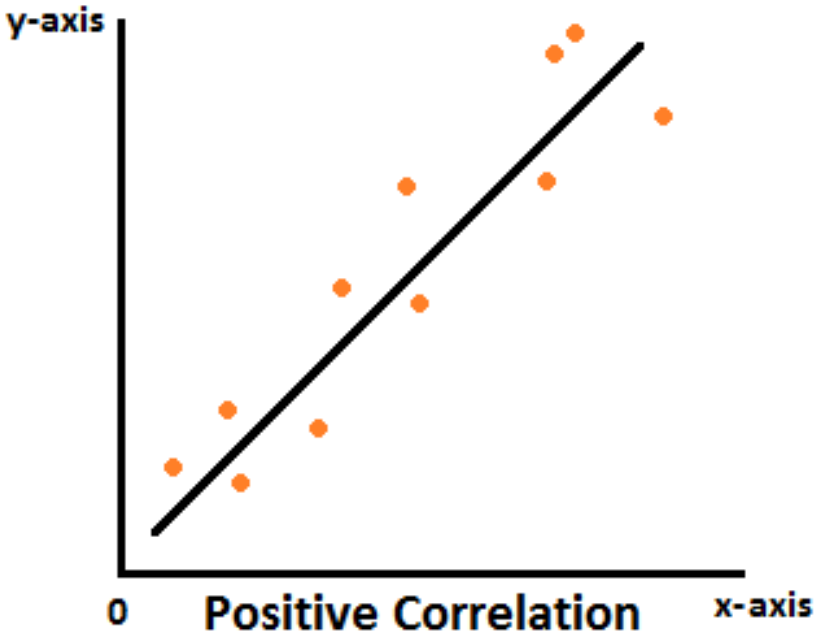
strong relationship

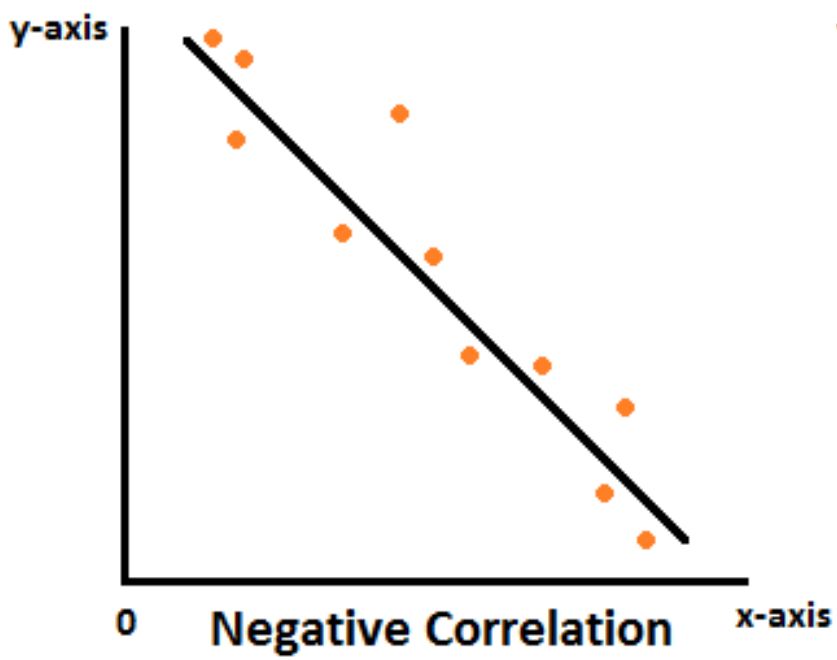


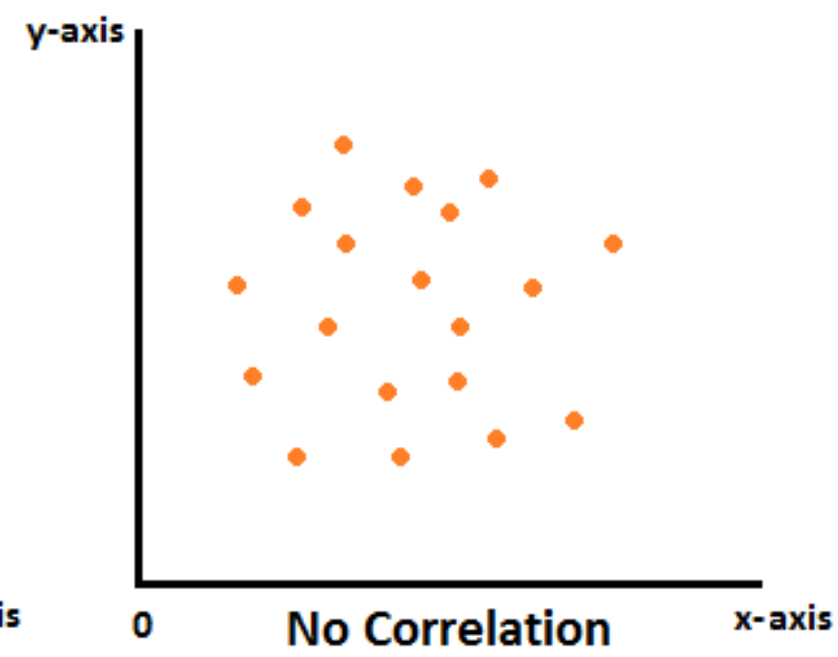
weaker relationship

Correlation

- Correlation is a bivariate analysis that measures **the strength of association** between two variables and **the direction** of the relationships
- In terms of the strength of relationship, the value of the correlation coefficient varies **between +1 and -1**
- **Correlation does not mean causation**







Suresh S. From correlation to causation [Internet]. Medium. 2020 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/from-correlation-to-causation-49f566eea954

Correlation Coefficient

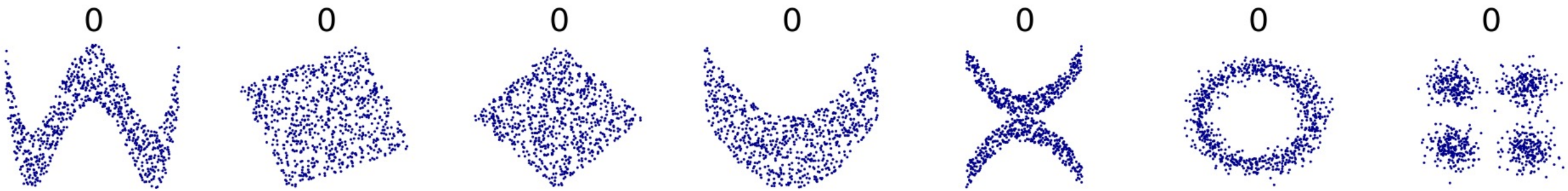
- A statistic that measures the relationship between two variables

- Pearson's r
  - Measures **linear** relationship
  - Both variables have to be normally distributed
- Spearman's ρ
  - Measures **monotonic** relationship
  - Based on rank – non-parametric

Pearson Correlation Coefficient

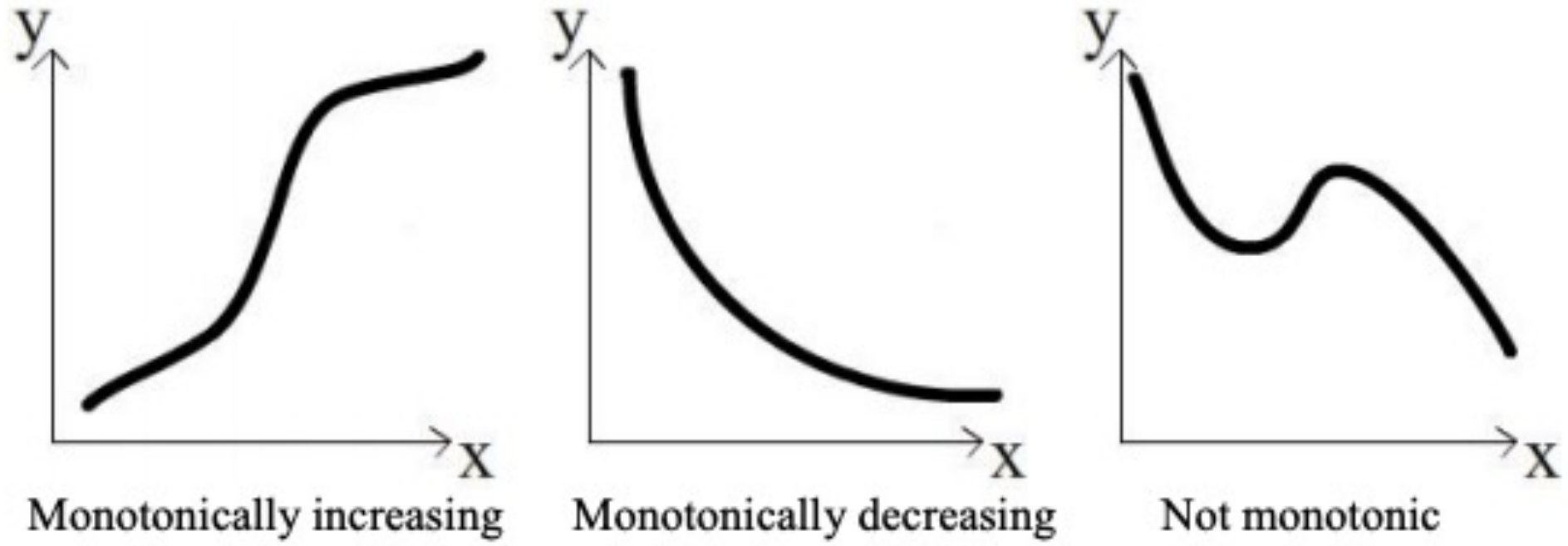
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- A measure of the **linear** correlation between two variables X and Y
- takes values between -1 and 1
- unitless
- $r_{X,Y} = r_{Y,X}$
- $r_{X,Y} = 0$ means **no linear relationship**



https://en.wikipedia.org/wiki/Correlation_and_dependence

Spearman Rank Correlation

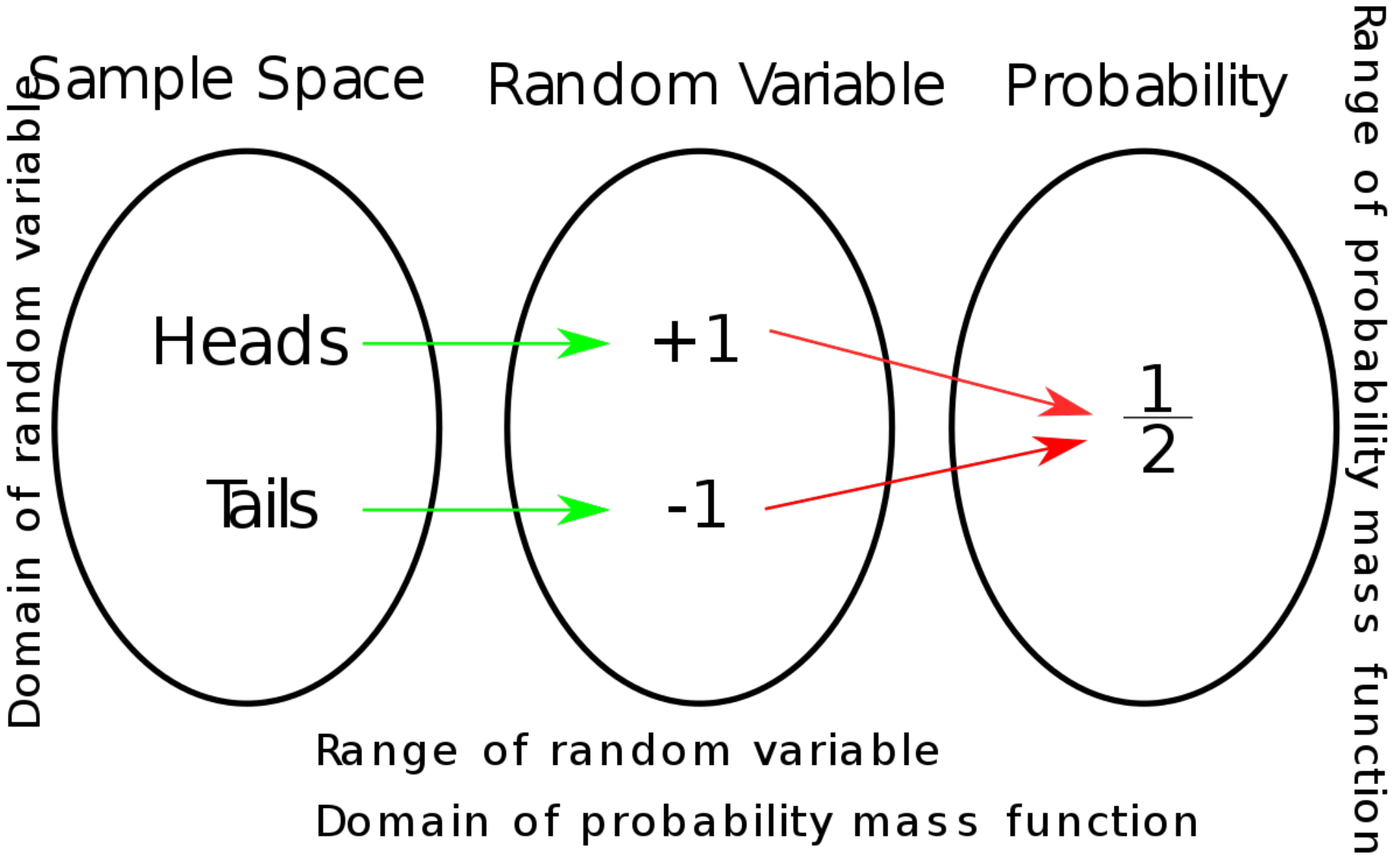


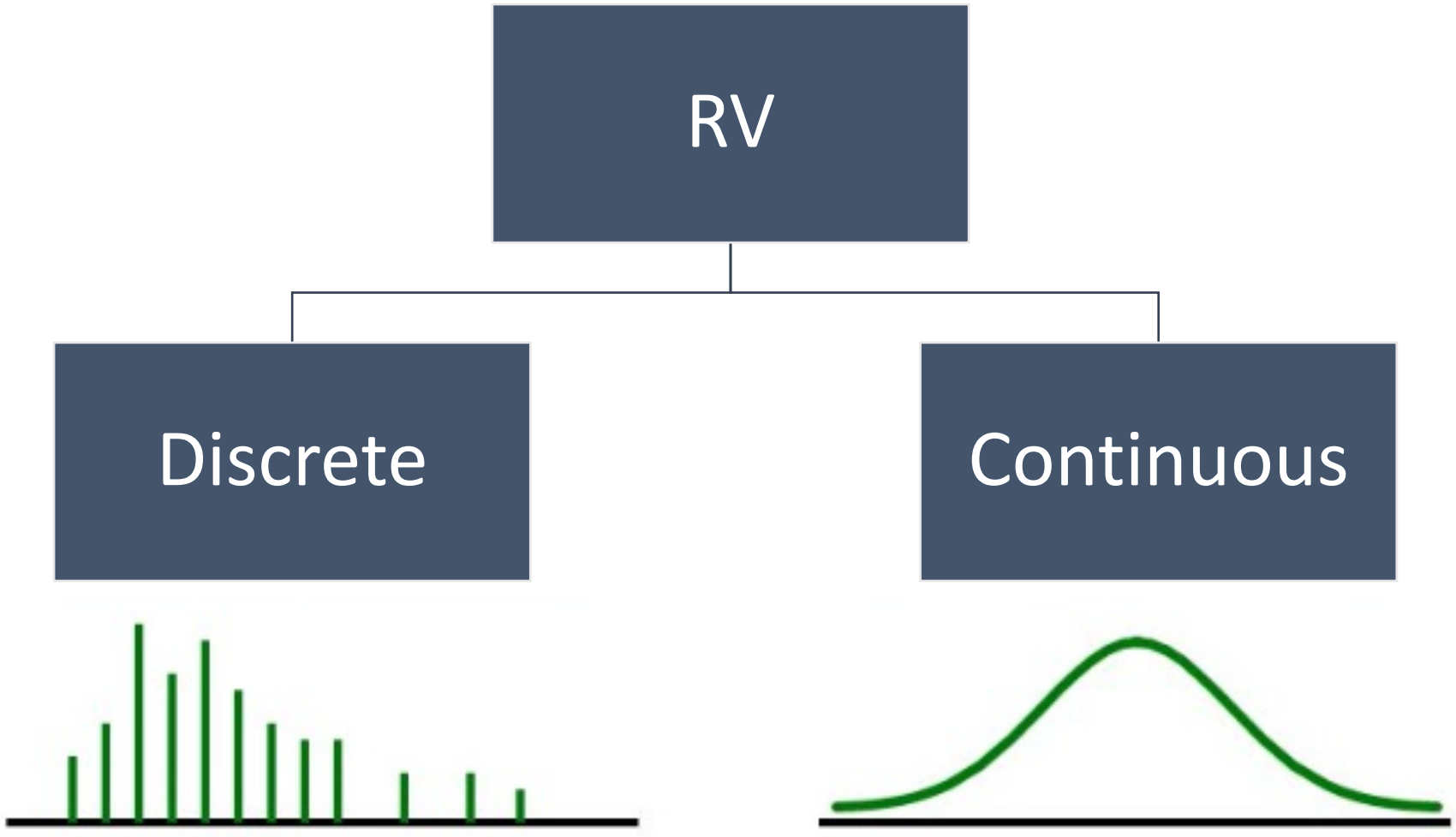
Random Variable

- A random variable (RV) is a variable whose possible values are **numerical outcomes of a random phenomenon**

- There are two types of random variables:
  - ***Discrete*** – flipping a coin, rolling a die, number of pancreatic cancer cases in a year ...
  - **Continuous** – systolic blood pressures of hypertensive patients, progression-free survival time of glioblastoma patients, expression level of a certain gene …

https://en.wikipedia.org/wiki/Random_variable

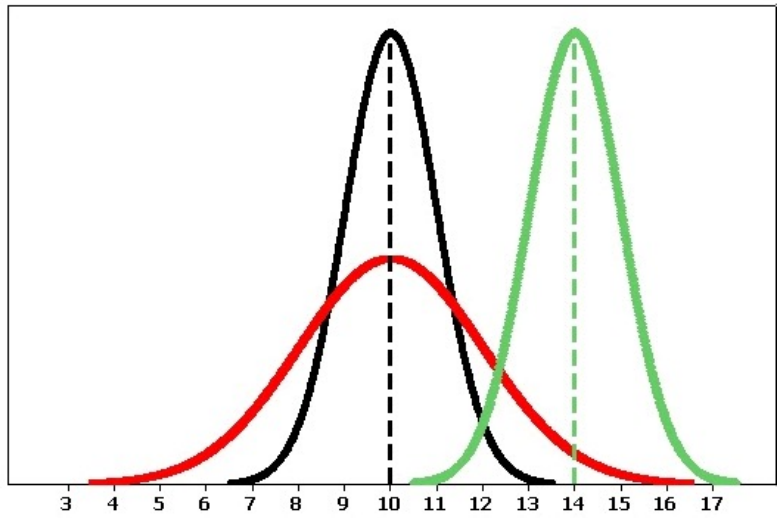



RV
Discrete
Continuous

Normal Distribution

- The distributions of many variables follow a “normal distribution”

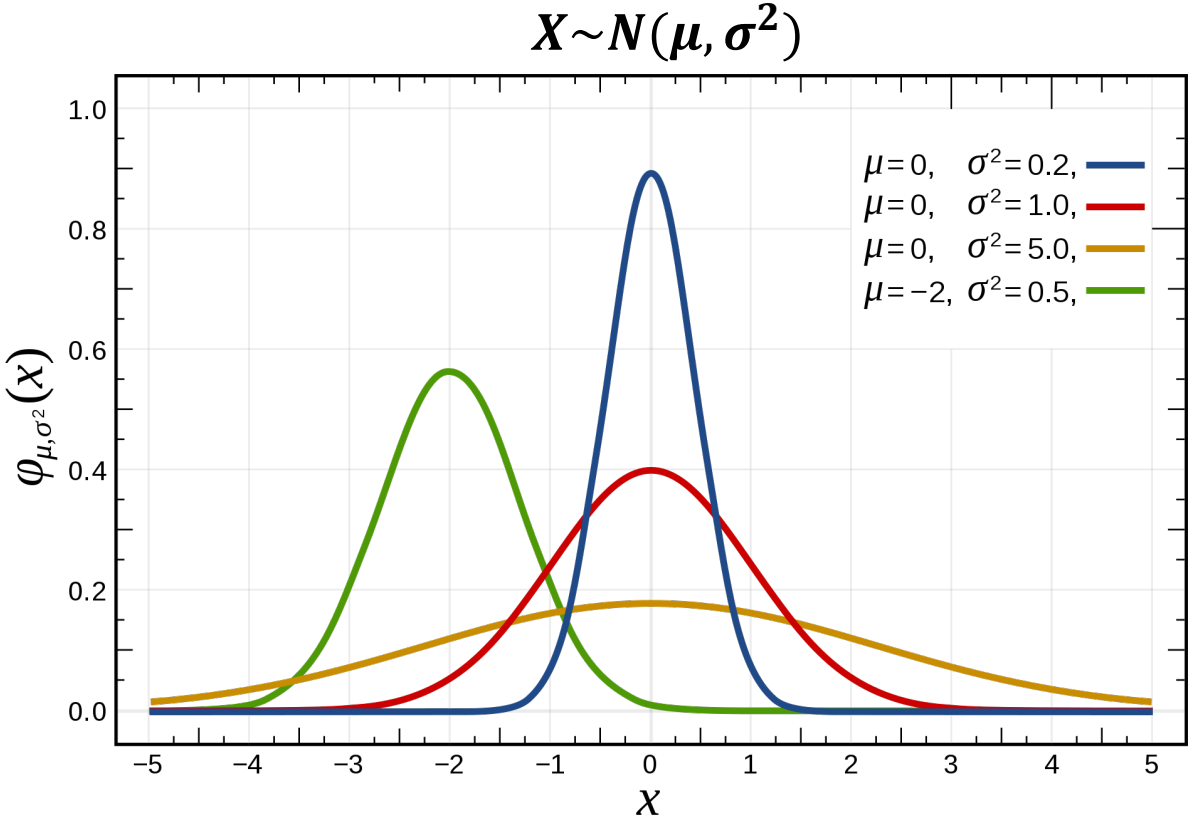
- The **bell-shape** indicates that values closer to the mean are more likely, and it becomes increasingly unlikely to take values far from the mean in either direction

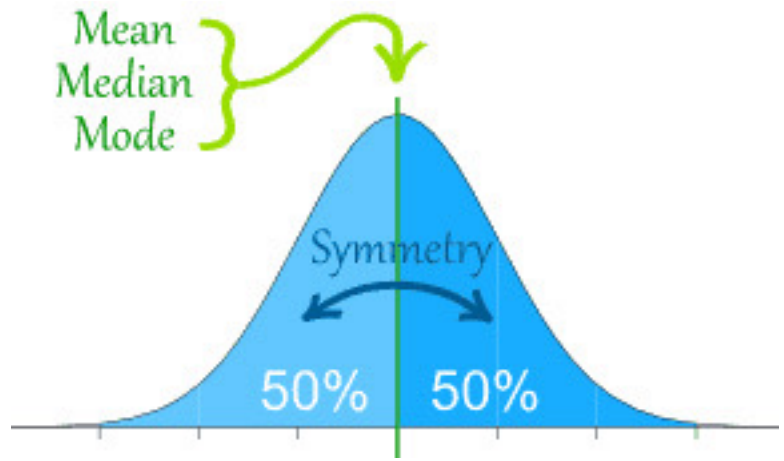


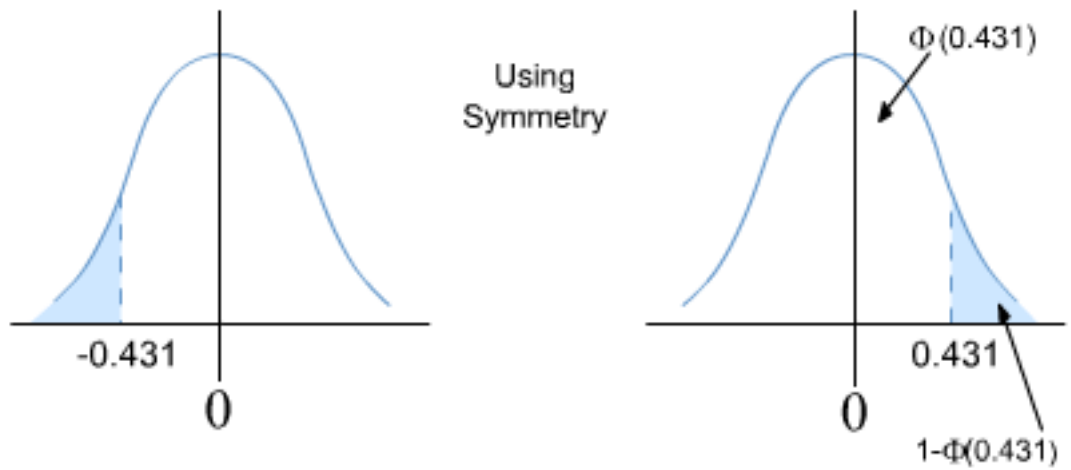
Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma^2 > 0$$

- Mean = Median = Mode= μ
- Variance = σ^2

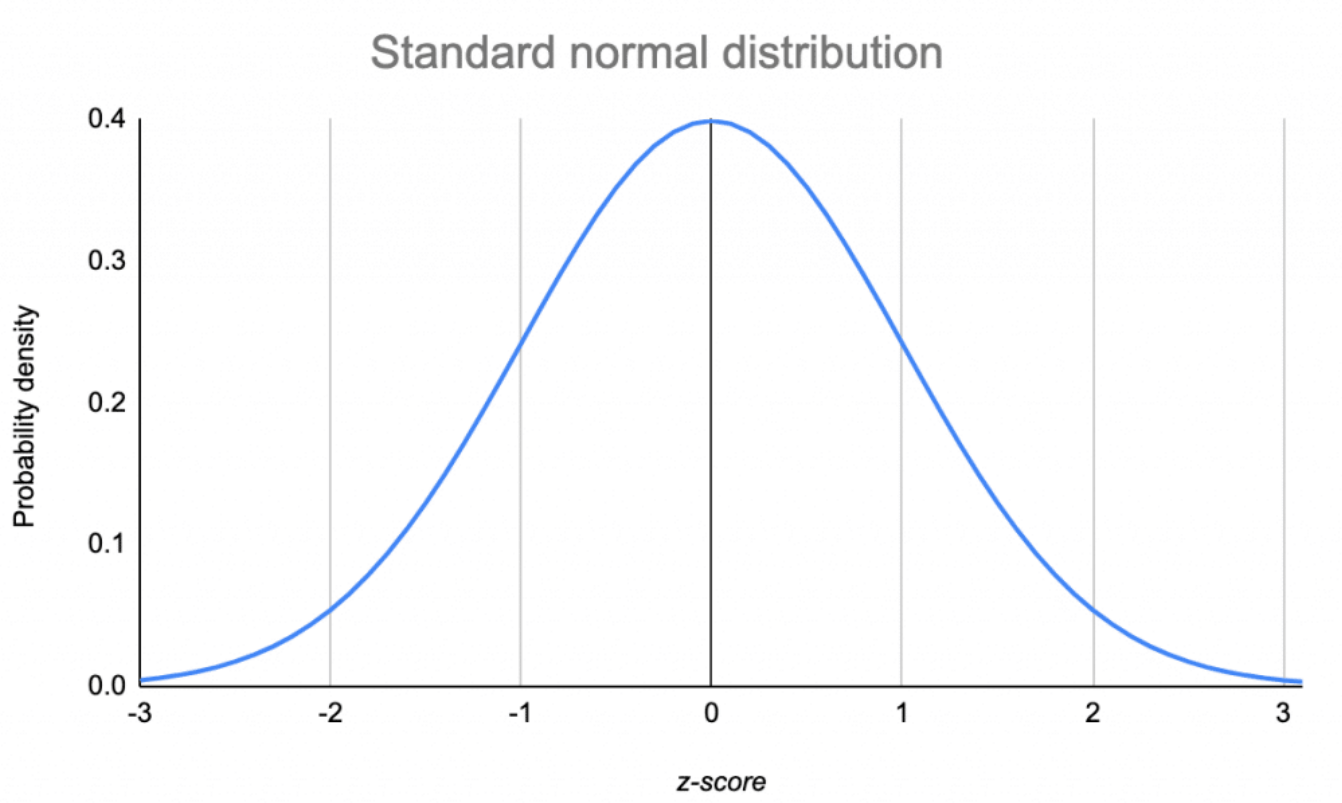



Mean
Median
Mode
Symmetry
50%
50%


-0.431
0
Using
Symmetry
Φ(0.431)
0.431
0
1-Φ(0.431)

Standard Normal Distribution

- Normal distribution for which $\mu = 0$ and $\sigma^2 = 1$
- Usually denoted with Z



https://byjus.com/maths/z-score-table/

STANDARD NORMAL PROBABILITIES

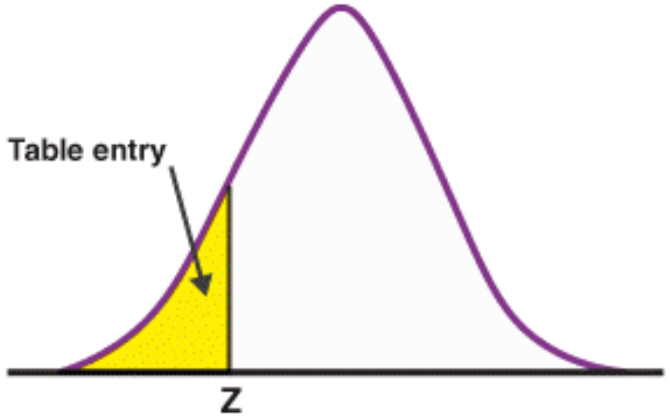


Table entry for z is the area under the standard normal curve to theleft of z.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143

Standard Normal Probabilities

https://byjus.com/maths/z-score-table/

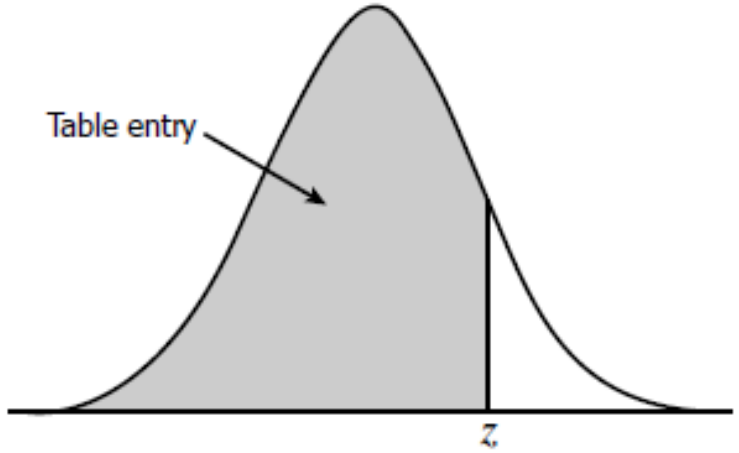


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Standardization

$$X \sim N(\mu, \sigma^2) \implies \boldsymbol{Z = \frac{X - \mu}{\sigma}} \sim N(0,1)$$

Normal distribution [Internet]. [cited 2021 Oct 4]. Available from: https://www.mathsisfun.com/data/standard-normal-distribution.html

Normal Distribution - Example

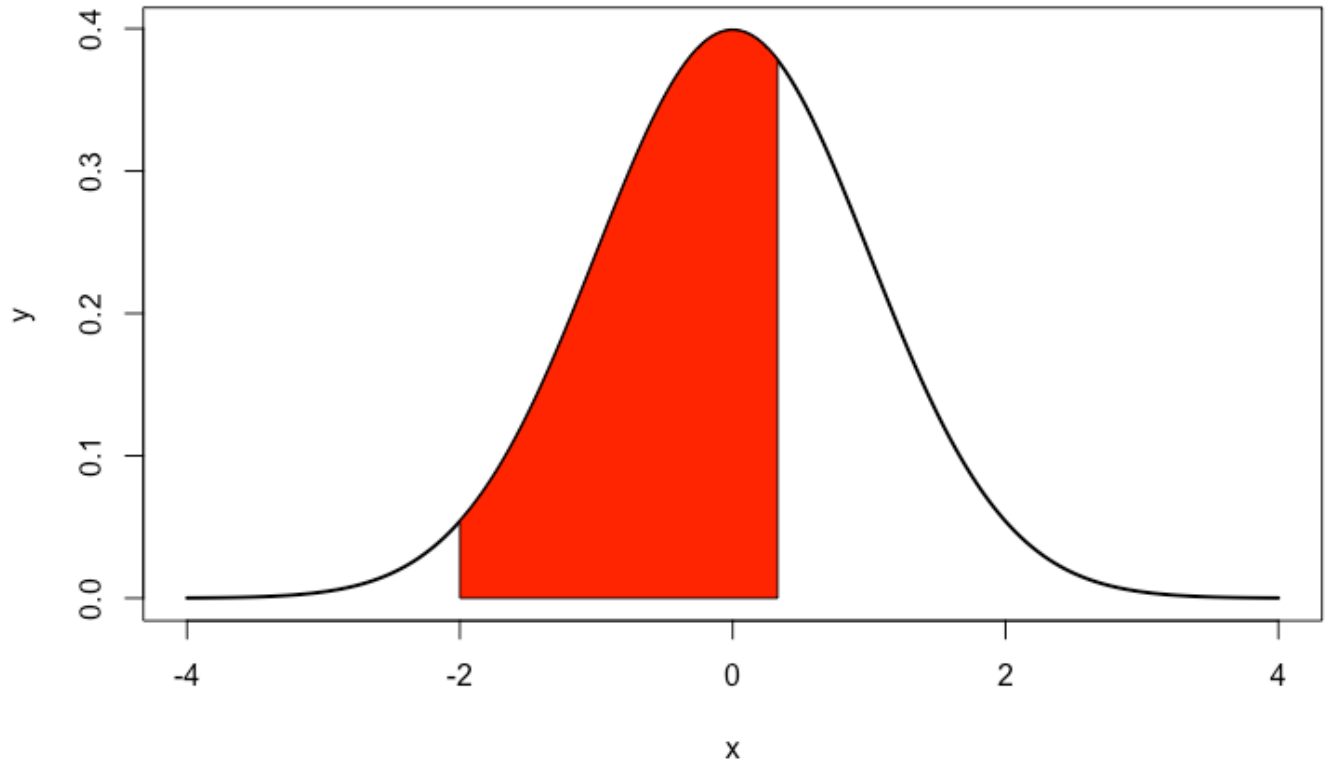
$$X \sim N(15,9)$$

- In a hospital, the systolic blood pressures of patients follow a normal distribution with mean = 15, variance = 9
- For a randomly selected patient, what is the probability that their SBP is **between 9 and 16**?

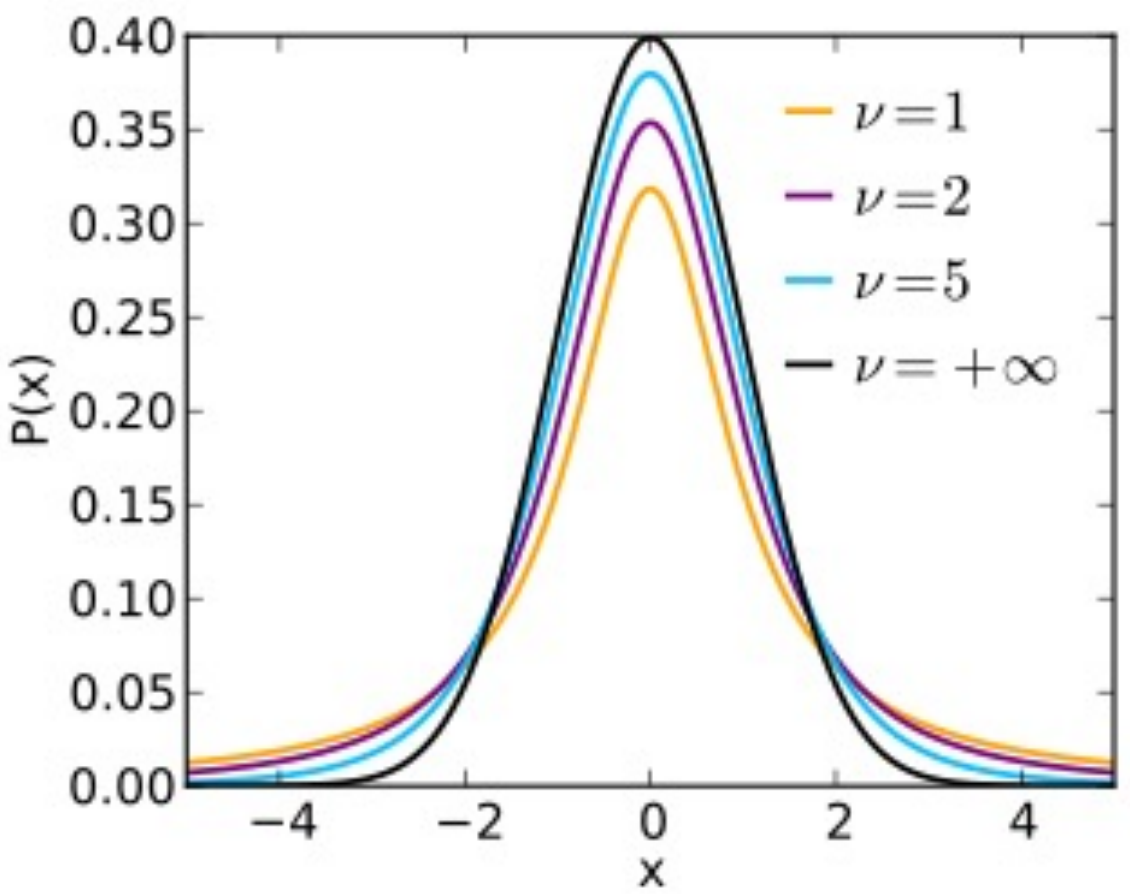
$X \sim N(15,9)$

SBP Between 9 and 16

$$P(9 < X < 16) = P\left(\frac{9-15}{3} < Z < \frac{16-15}{3}\right) = P(-2 < Z < 0.33) = P(Z < 0.33) - P(Z \leq -2) = 0.6065$$



(Student’s) t Distribution



https://en.wikipedia.org/wiki/Student%27s_t-distribution

Sampling Distribution

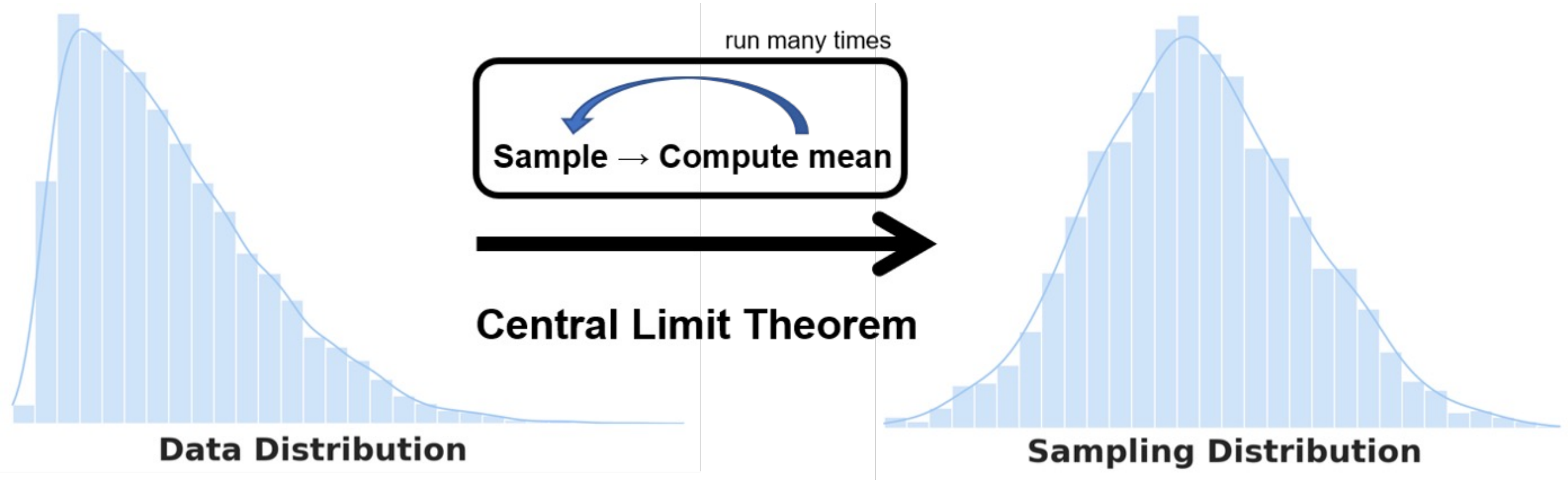
- Population Distribution
- Sample Distribution
- **Sampling Distribution**
  - theoretical probability distribution of a statistic obtained through a specific number of samples drawn from a specific population
  - if samples are randomly selected, the sample means will be somewhat different from the population mean (sampling error)

Sampling Distribution of the Sample Mean

- If sample size is large enough, the sampling distribution of the sample mean will be approximately normal

- the mean of the sample means will be the same as the population mean
- the standard deviation of the sample means = $\frac{\sigma}{\sqrt{n}}$

https://towardsdatascience.com/data-distribution-vs-sampling-distribution-what-you-need-to-know-294819109796

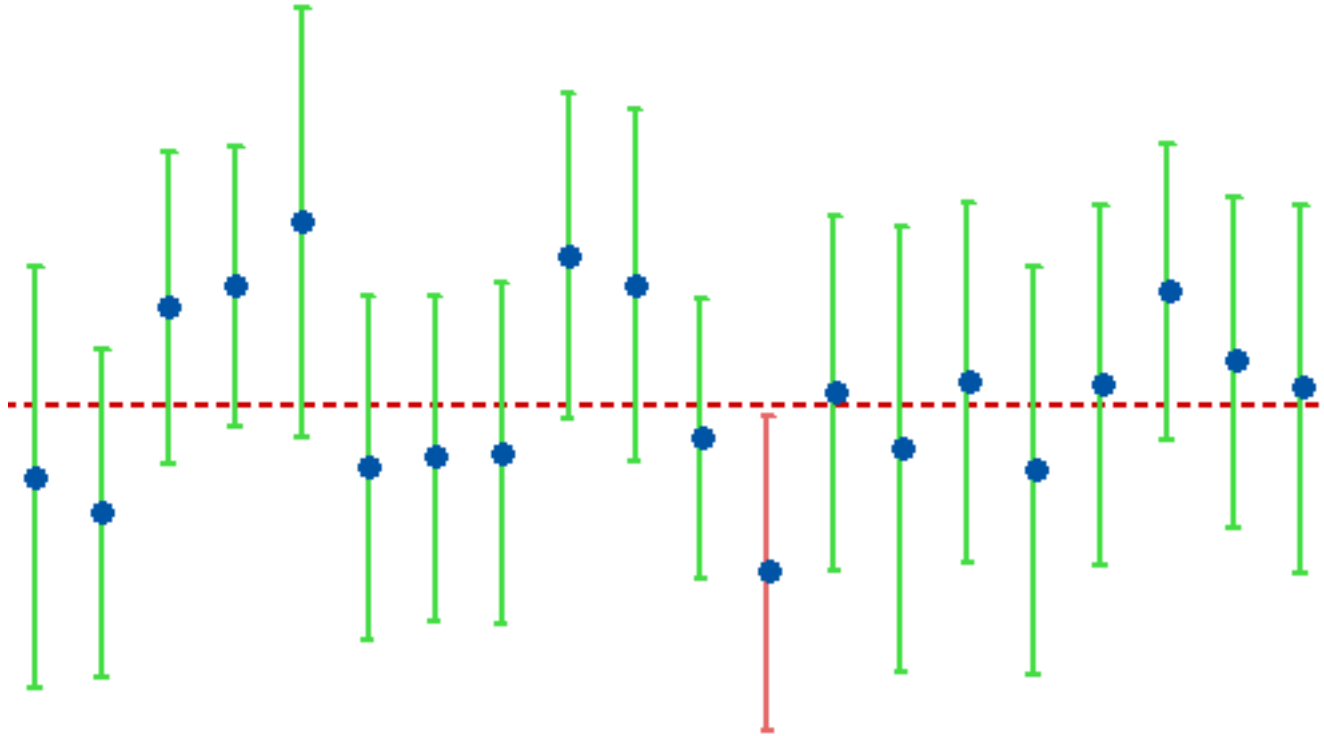


https://www.scribbr.com/statistics/confidence-interval/

Confidence Interval

- When you make an estimate in statistics, there is always uncertainty around that estimate because the number is based on a single sample
- The confidence interval is the **range of values that you expect your estimate to fall between a certain percentage of the time** if you run your experiment again (re-sample the population in the same way)

https://statisticsbyjim.com/hypothesis-testing/confidence-interval/



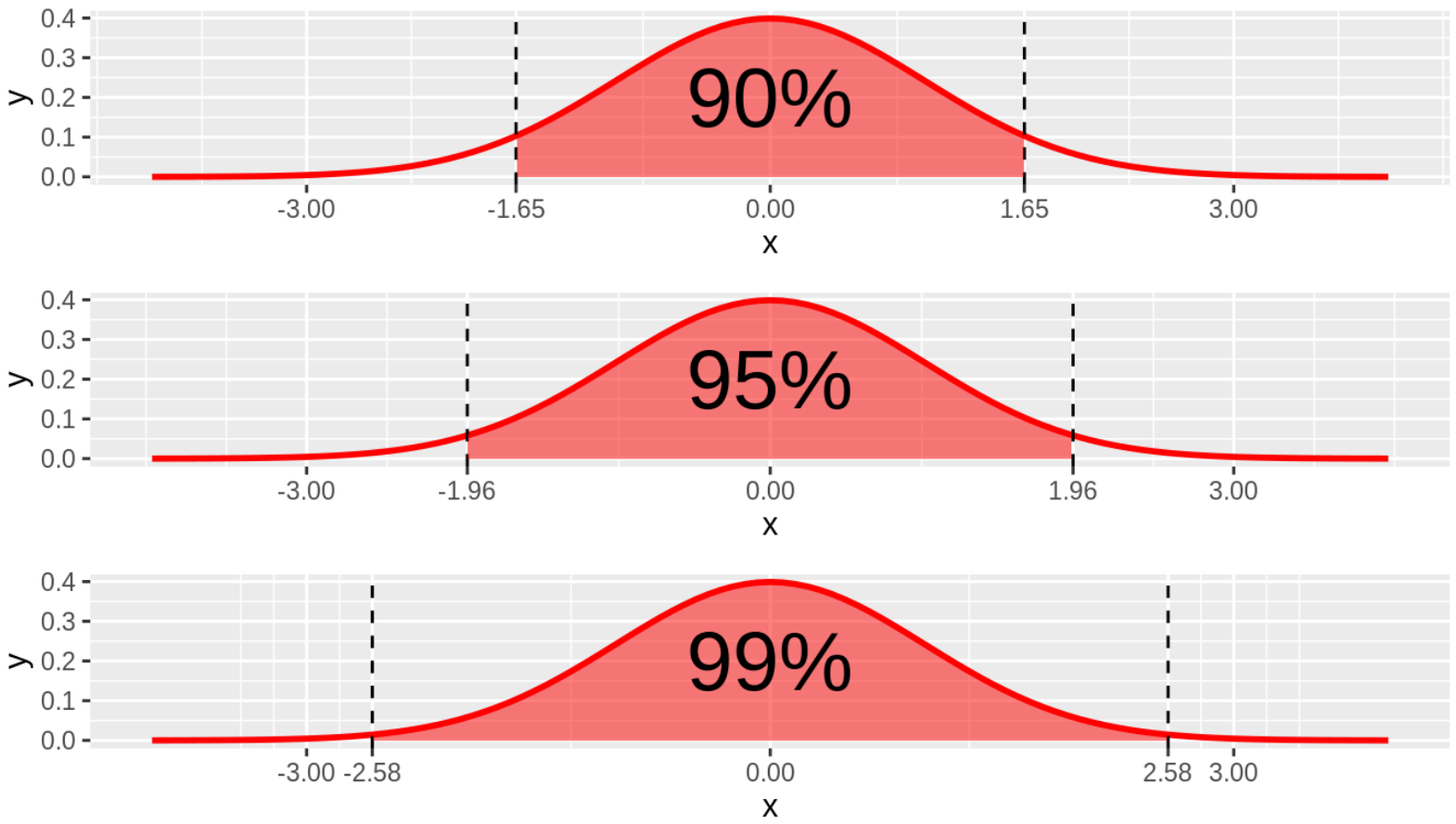
a 95% confidence interval [10 15] indicates that we can be 95% confident that the parameter is within that range

However, it does NOT indicate that 95% of the sample values occur in that range

Confidence Interval

$$CI = \bar{x} \pm Z * \frac{s}{\sqrt{n}}$$

$$CI = \bar{x} \pm t * \frac{s}{\sqrt{n}}$$



Hypothesis Testing

- **Hypothesis**: an assumption that can be tested based on the evidence available
  - A novel drug is efficient in treating a certain disease

- **Hypothesis test**: investigation of the hypothesis using the sample
  - Assessing evidence provided by the data against the null claim (the claim which is to be assumed true unless enough evidence exists to reject it)

Null and Alternative Hypotheses

- H_0 – Null hypothesis
  - The mean of a variable is not different than *c*
  - There is no difference between the two groups' means
  - There is no difference compared to baseline
  - …
- H_a or H_1 – Alternative hypothesis
  - There is a difference between the two groups' means
  - The mean in group A is higher than group B
  - …

One- vs. Two-tailed Tests

- The coin is biased

Two-tailed

H_0: $p = 0.5$

H_a: $p \neq 0.5$

- The probability of heads is larger than 0.5

One-tailed

H_0: $p \leq 0.5$

H_a: $p > 0.5$

Hypothesis Testing

H_0	Decision	
	Fail to reject	Reject
True	Correct decision	**Type I Error** α
False	**Type II Error** ß	Correct decision

- **Confidence level** = $1 - \alpha$
  - P(fail to reject H_0| H_0 is true)
- **Statistical power** = $1 - \beta$
  - P(reject H_0| H_0 is false)

Hypothesis Testing - Steps

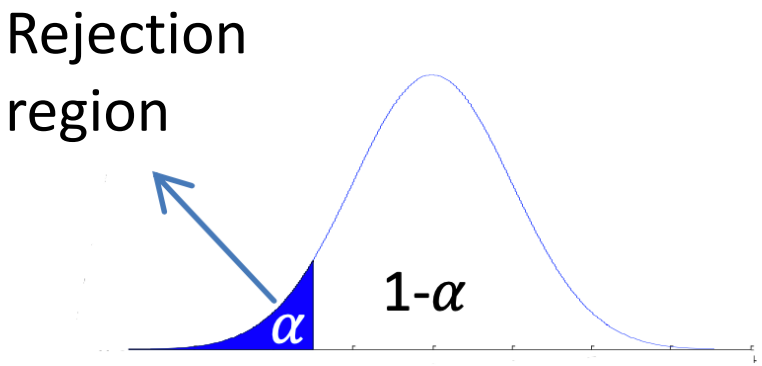
1. **Check assumptions, determine H_0 and H_a, choose α**
   - Assumptions differ based on the test
   - The null hypothesis always contains equality (=)
2. **Calculate the appropriate test statistic**
   - z, t, χ^2, …
3. **Calculate critical values/p value**
   - With the aid of precalculated tables/software
4. **Decide whether to reject/fail to reject H_0**
   - Reject if the statistic is within the critical region/$p \leq \alpha$

Test Statistic

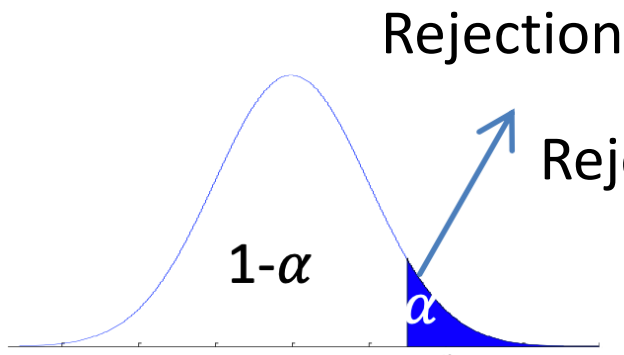
$$\text{test statistic} = \frac{\text{estimator} - \text{null value}}{\text{standard error of estimator}}$$

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

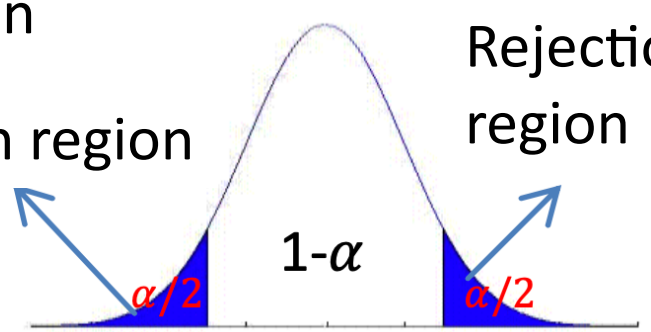
$$H_0: \mu = \mu_0$$
$$H_1: \mu < \mu_0$$



$$H_0: \mu = \mu_0$$
$$H_1: \mu > \mu_0$$

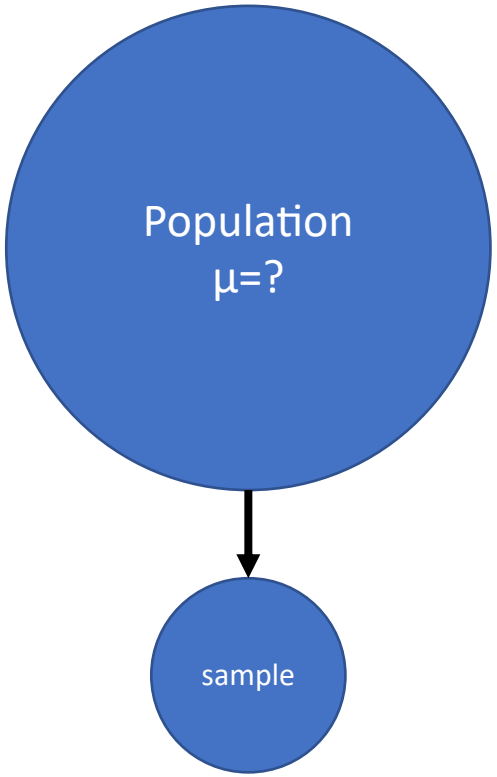


$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$



One-Sample t-Test

- a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value



One-Sample t-Test – Example I

id	week_1	cd4_1	week_2	cd4_2	perc_benefit
361	0	26	7.43	3	-11.905994
1017	0	13	7.00	10	-3.296703
519	0	3	8.14	5	8.190008
1147	0	65	33.00	97	1.491841
1216	0	36	8.00	31	-1.736111
52	0	16	9.43	31	9.941676
660	0	34	8.43	32	-0.697788
1145	0	41	8.00	71	9.146341
697	0	33	8.00	45	4.545455
560	0	21	8.00	27	3.571429

$\bar{X} = 1.925015$

$\mu = 0$ or not?

- Mean percentage benefit $\bar{X}$ is 1.925015
- Is it due to chance? Or does it indicate positive impact of the novel treatment?
  - What would be the value of mean percentage benefit what if you selected another set of 10 patients?

One-Sample t-Test – Example I (cont.)

1. Check assumptions, determine H_0 and H_a, choose α
   - Normality of the variable is checked (Quantile-quantile plot)
   - $H_0: \mu = 0$ $H_a: \mu \neq 0$
   - $\alpha = 0.05$

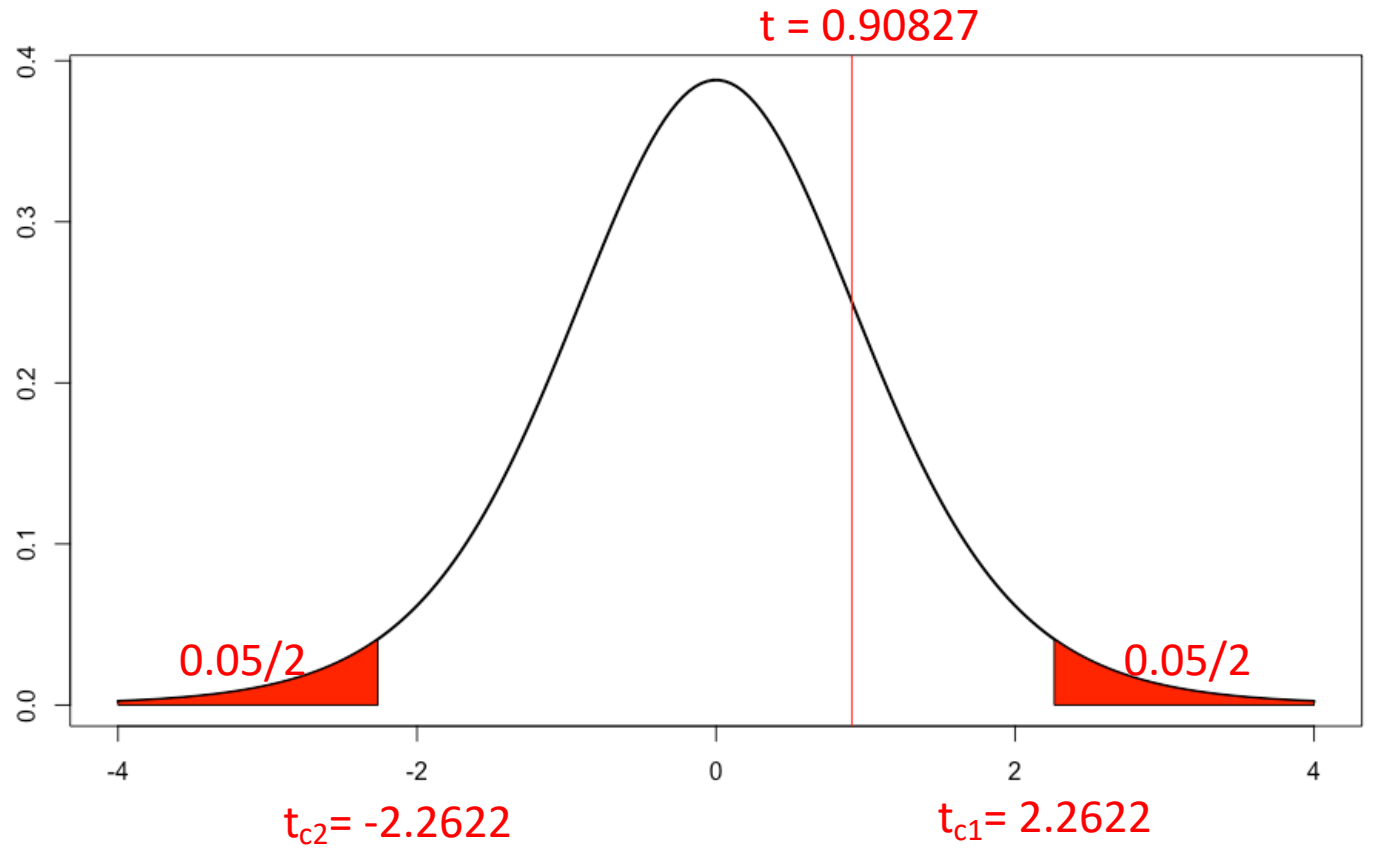
One-Sample t-Test – Example I (cont.)

2. Calculate the appropriate test statistic
   - Mean percentage benefit is 1.925015
   - Standard deviation is 6.702202
   - Sample size is 10

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{1.925015 - 0}{6.702202/\sqrt{10}} = 0.9082736 \quad (\sim t_{n-1} = t_9)$$

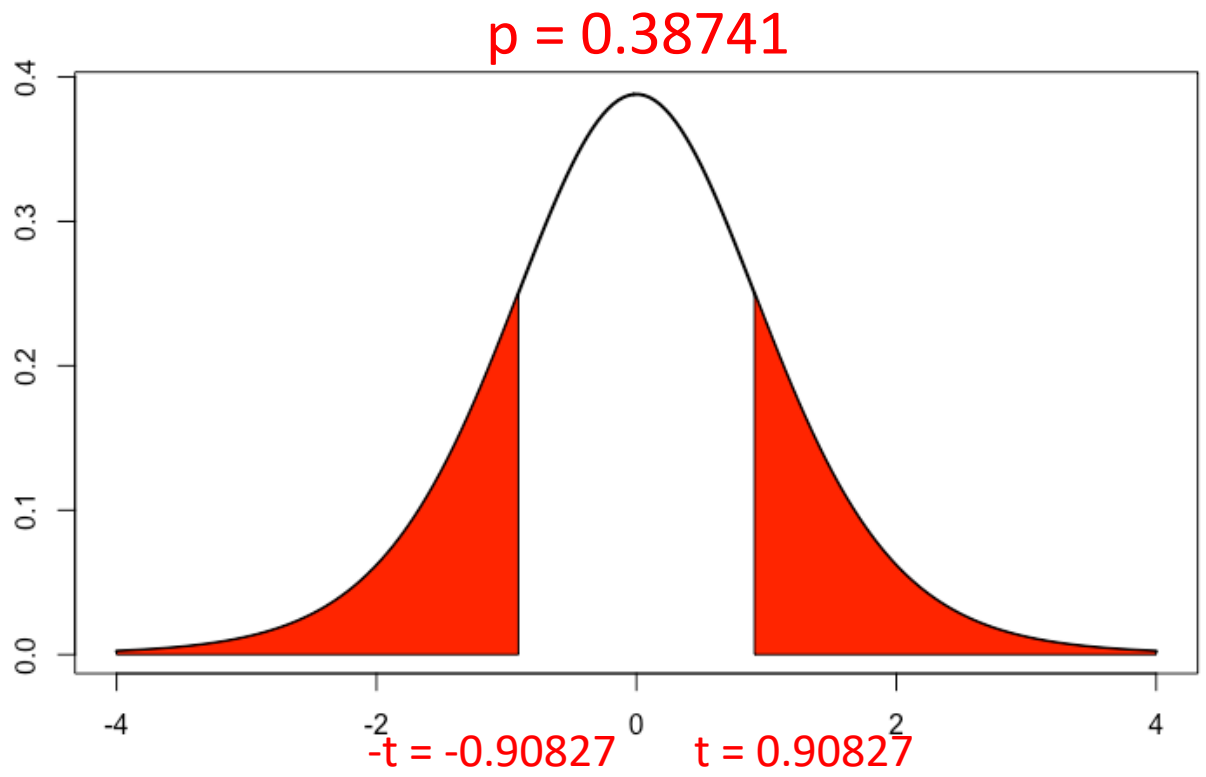
One-Sample t-Test – Example I (cont.)

3. Calculate **critical values**/p value
4. Decide whether to reject/fail to reject H_0



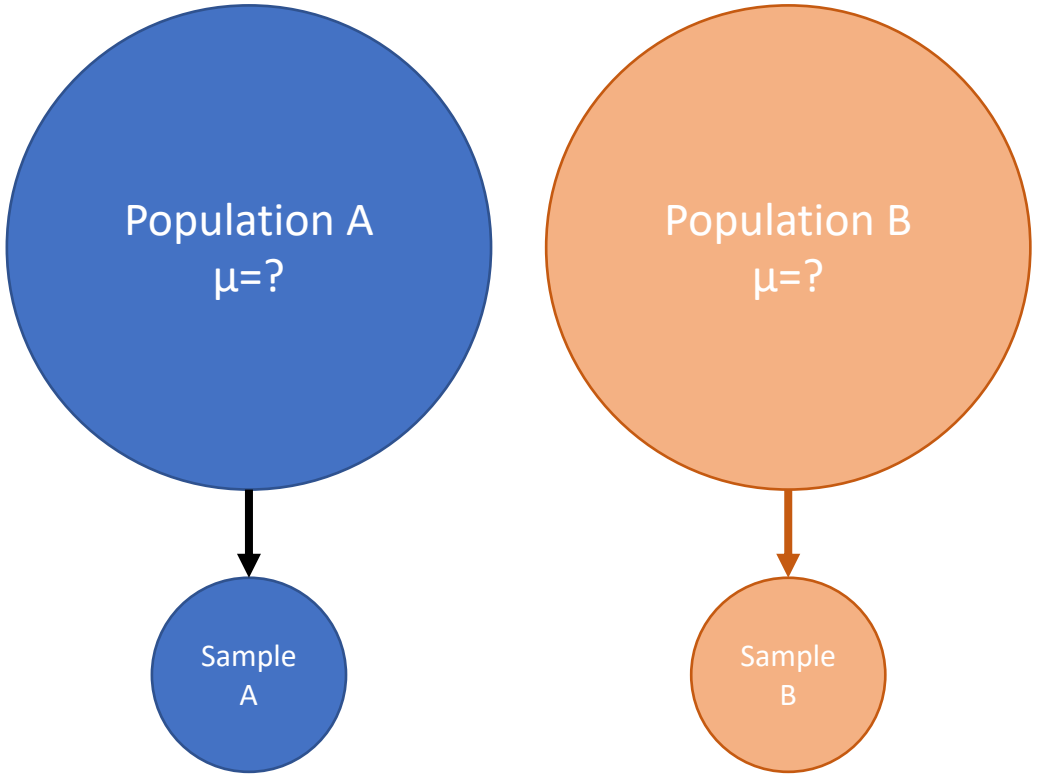
One-Sample t-Test – Example I (cont.)

3. Calculate critical values/**p value**
4. Decide whether to reject/fail to reject H_0



Two-Sample t-Test

- The **two-sample t-test** (also known as the **independent samples t-test**) is a method used to test whether the unknown population means of two groups are equal or not



Wong DT, Dallaire A, Singh KP, Madhusudan P, Jackson T, Singh M, et al. High-flow nasal oxygen improves safe apnea time in morbidly obese patients undergoing general anesthesia: a randomized controlled trial. Anesth Analg. 2019 Oct;129(4):1130–6.

Two-sample t-Test – Example III

- "Morbidly obese patients undergoing general anesthesia are at risk of hypoxemia during anesthesia induction"
- A randomized controlled trial investigating:
- Does high-flow nasal oxygenation provide longer safe apnea time compared to conventional facemask oxygenation during anesthesia induction in morbidly obese surgical patients?

Wong DT, Dallaire A, Singh KP, Madhusudan P, Jackson T, Singh M, et al. High-flow nasal oxygen improves safe apnea time in morbidly obese patients undergoing general anesthesia: a randomized controlled trial. Anesth Analg. 2019 Oct;129(4):1130–6.

Two-sample t-Test – Example III (cont.)

- Safe Apnea time in Control Group (n = 20)
  - $\overline{X_c} = 185.5$
  - $s_c = 53$
- Safe Apnea time in High-Flow Nasal Oxygenation Group (n = 20)
  - $\overline{X_T} = 261.4$
  - $s_T = 77.7$

Two-sample t-Test – Example III (cont.)

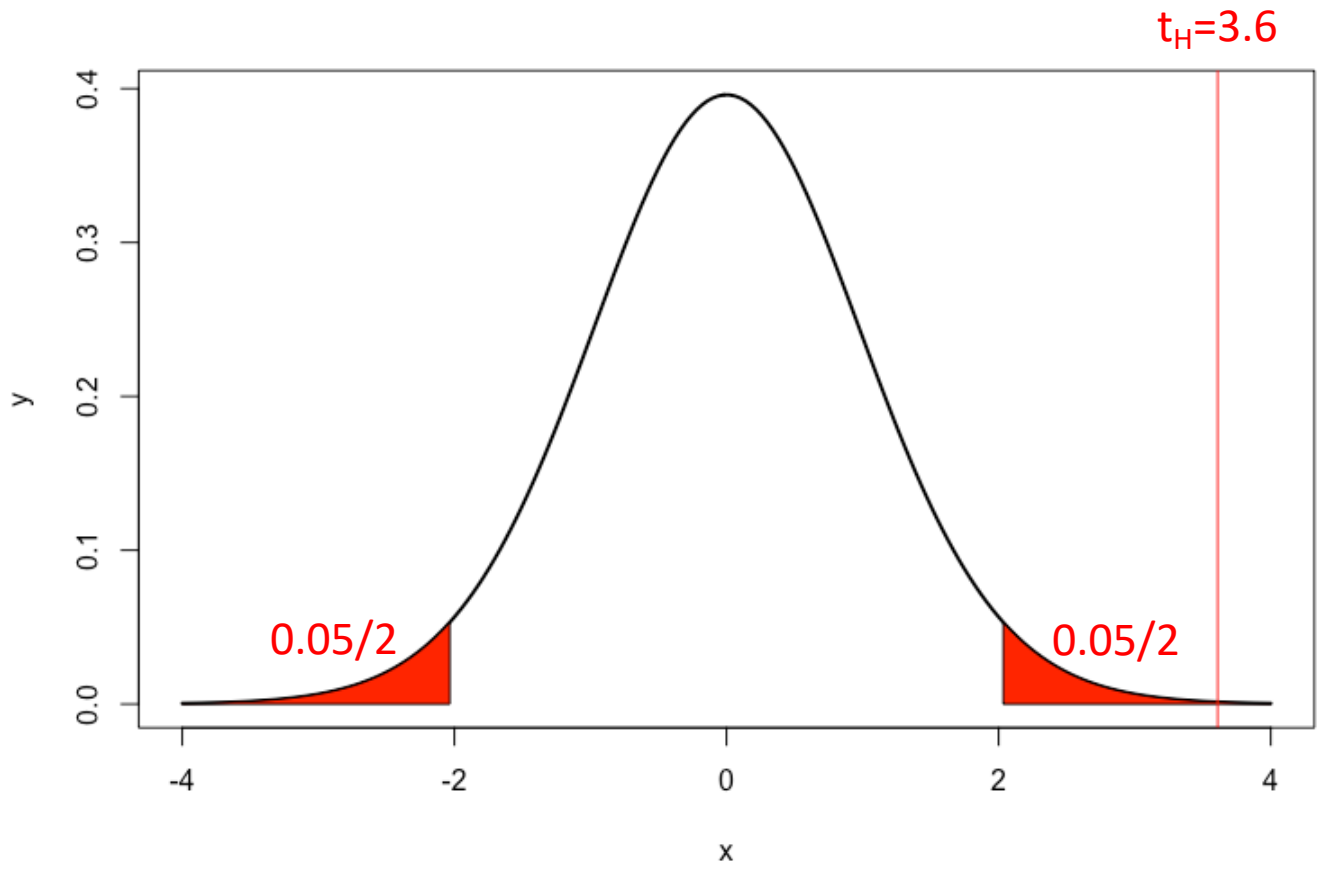
1. Check assumptions, determine H_0 and Ha, choose α
   - We check that the variables are normally distributed
   - $H_0: \mu_c = \mu_T$ $H_a: \mu_c \neq \mu_T$
   - $\alpha = 0.05$

2. Calculate the appropriate test statistic

   $t = 3.6 \qquad (\sim t_{33.53})$

Two-sample t-Test – Example III (cont.)

3. Calculate **critical values**/p value
4. Decide whether to reject/fail to reject H_0



Wong DT, Dallaire A, Singh KP, Madhusudan P, Jackson T, Singh M, et al. High-flow nasal oxygen improves safe apnea time in morbidly obese patients undergoing general anesthesia: a randomized controlled trial. Anesth Analg. 2019 Oct;129(4):1130–6.

Two-sample t-Test – Example III (cont.)

Table 2. Study Outcomes: Safe Apnea Time, Minimum Spo_2, Plateau $ETco_2$, and Time to Regain Baseline Spo_2

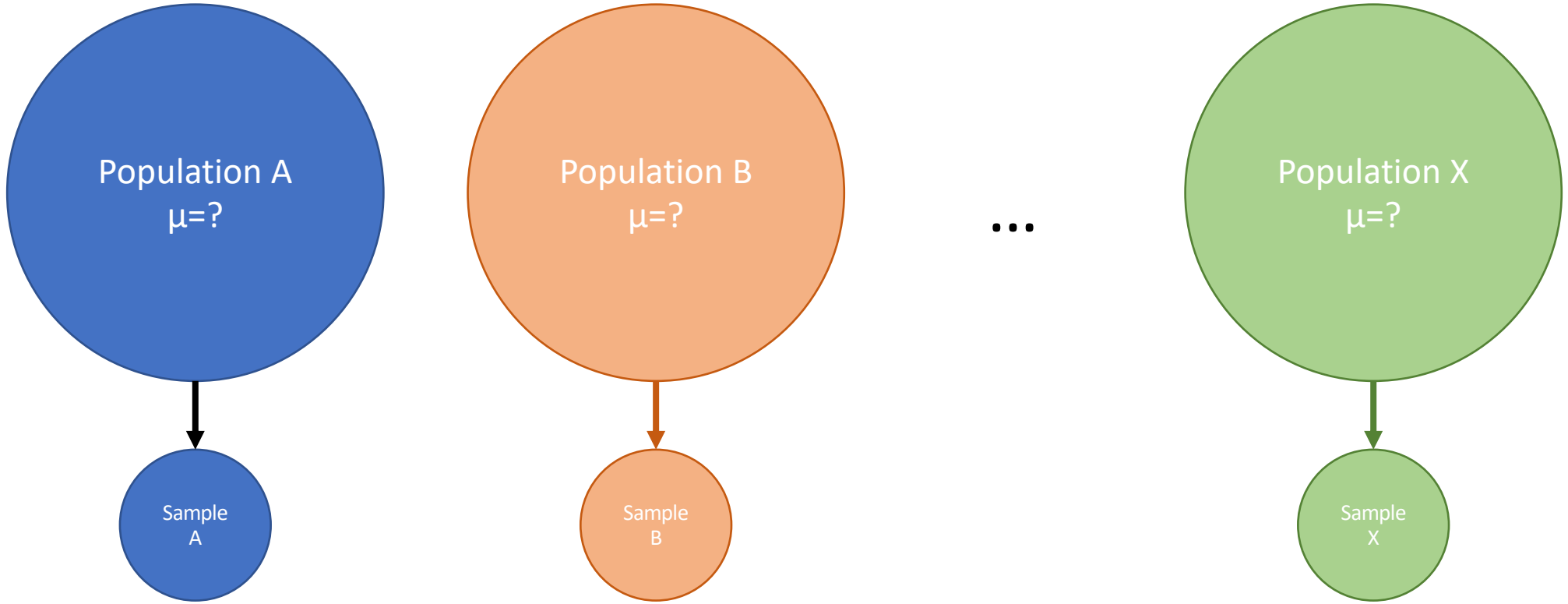
	Control Group (n = 20)	High-Flow Nasal Oxygenation Group (n = 20)	Mean Difference (95% CI)	*P* Value
Safe apnea time (s)	185.5 ± 53.0	261.4 ± 77.7	75.9 (33.3–118.5)	.001
Minimum Spo_2 (%)	87.9 ± 4.7	90.9 ± 3.5	3.1 (0.4–5.7)	.026
Plateau $ETco_2$ (mm Hg)	38.8 ± 2.5	37.9 ± 3.0	−0.8 (−2.6 to 0.9)	.33
Time to regain baseline Spo_2 (s)	49.6 ± 20.8	37.3 ± 6.8	−12.3 (−22.2 to −2.4)	.016

Values represent mean ± SD.
Control group: facemask oxygenation.
Abbreviations: CI, confidence interval; $ETco_2$, end-tidal carbon dioxide; Spo_2, oxygen saturation measured by pulse oximetry.

"Safe apnea time was significantly longer (261.4 ± 77.7 vs 185.5 ± 52.9 seconds; mean difference [95% CI], 75.9 [33.3–118.5]; *P* = .001)…"

Analysis of Variance (ANOVA)

- Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of **two or more groups** are significantly different from each other



One-way ANOVA

k: number of groups
n: total number of samples
n_i: number of samples in group i

Analysis of Variance(ANOVA)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Between	$\sum n_i(\bar{X}_i - \bar{X})^2$	$k - 1$	SS_b/df_b	$F = \frac{MS_b}{MS_w}$
Within	$SS_T - SS_b$	$n - k$	SS_w/df_w	
Total	$\sum (X_j - \bar{X})^2$	$n - 1$		

How to perform analysis of variance (Anova) - step by step procedure [Internet]. The Genius Blog. 2018 [cited 2021 Oct 6]. Available from: https://kindsonthegenius.com/blog/how-to-perform-analysis-of-variance-anova-step-by-step-procedure/

One-way ANOVA – Example II

THE LANCET, AUGUST 12, 1978

MEGALOBLASTIC HÆMOPOIESIS IN PATIENTS RECEIVING NITROUS OXIDE

J. A. L. AMESS J. F. BURMAN
G. M. REES D. G. NANCEKIEVILL
D. L. MOLLIN

Departments of Hæmatology, Cardiothoracic Surgery, and Anæsthetics, St. Bartholomew's Hospital, West Smithfield, London EC1A 7BE

- 22 patients who underwent coronary artery bypass graft surgery (CABG) are separated into 3 different treatment groups (different ventilation strategies)
- Is there a difference in red blood cell folic acid measurements at 24 hours between the 3 treatment groups?

One-way ANOVA – Example II (cont.)

Group I.—8 patients received approximately 50% nitrous oxide and 50% oxygen mixture continuously for 24 h. 1 patient received 2000 μg of hydroxocobalamin intramuscularly immediately before and after the operation.

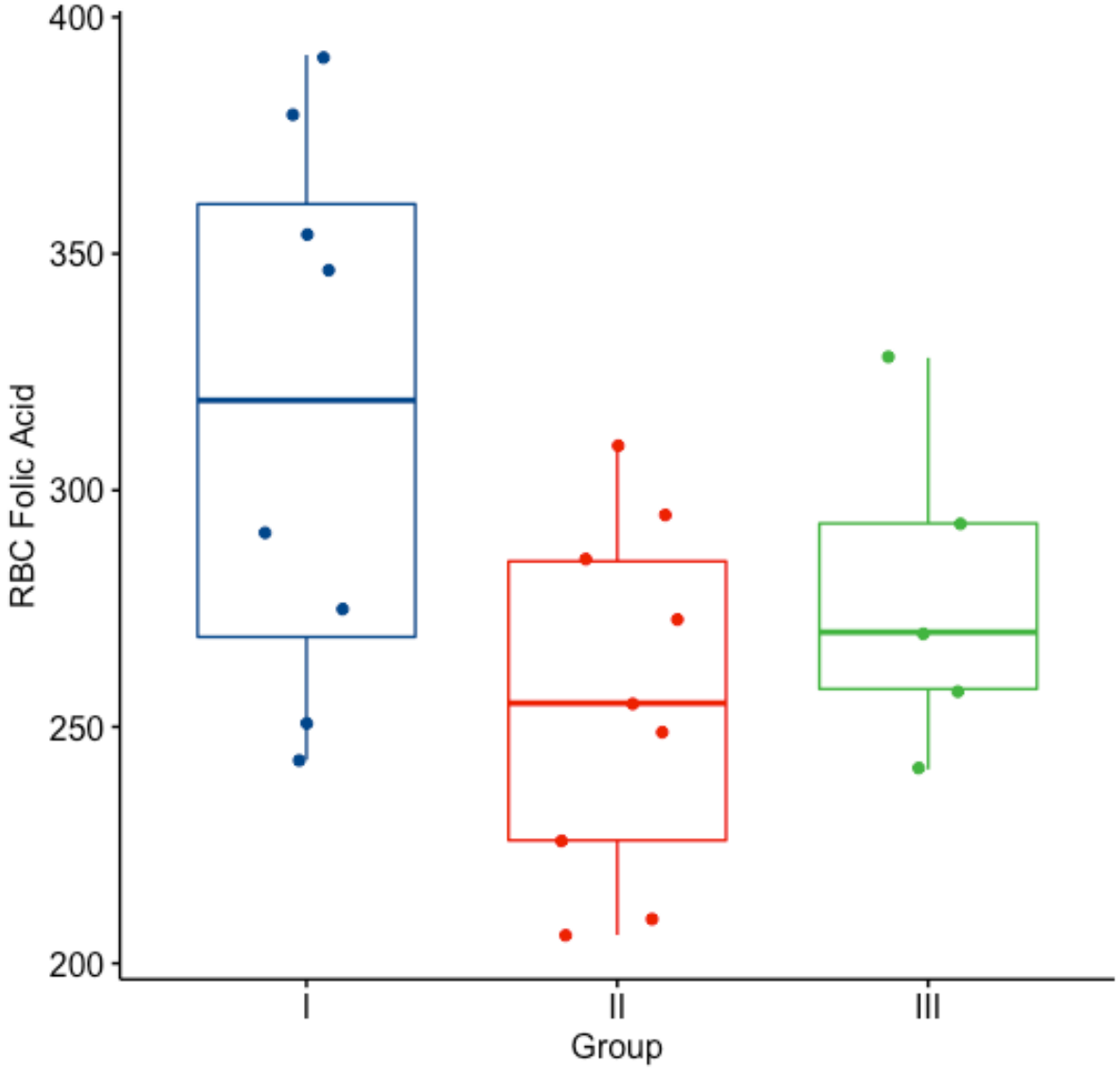
Group II.—9 patients received approximately 50% nitrous oxide and 50% oxygen mixture only during the operation (5–12 h) and thereafter 35–50% oxygen for the remainder of the 24 h period.

Group III.—5 patients received no nitrous oxide but were ventilated with 35–50% oxygen for 24 h.

Group I	Group II	Group III
243	206	241
251	210	258
275	226	270
291	249	293
347	255	328
354	273	
380	285	
392	295	
	309	

Amess JA, Burman JF, Rees GM, Nancekievill DG, Mollin DL. Megaloblastic haemopoiesis in patients receiving nitrous oxide. Lancet. 1978 Aug 12;2(8085):339–42.

One-way ANOVA – Example II (cont.)



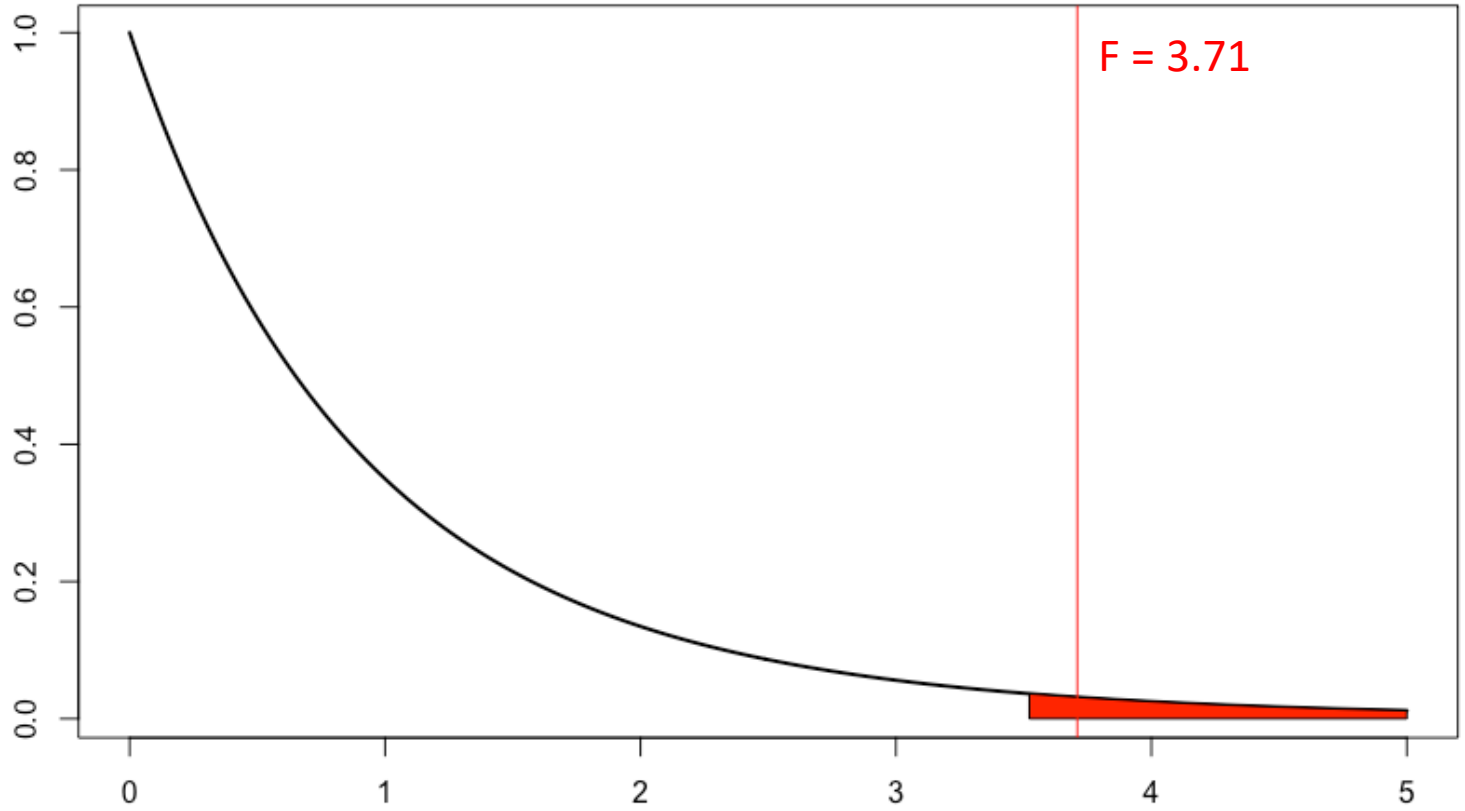
One-way ANOVA – Example II (cont.)

1. Check assumptions, determine H_0 and H_a, choose α
   - Check that data is normally distributed
   - H_0: $\mu_1 = \mu_2 = \mu_3$ H_a: at least one mean is different
   - $\alpha = 0.05$

2. Calculate the appropriate test statistic
   - F = 3.71 $\sim F_{2,19}$

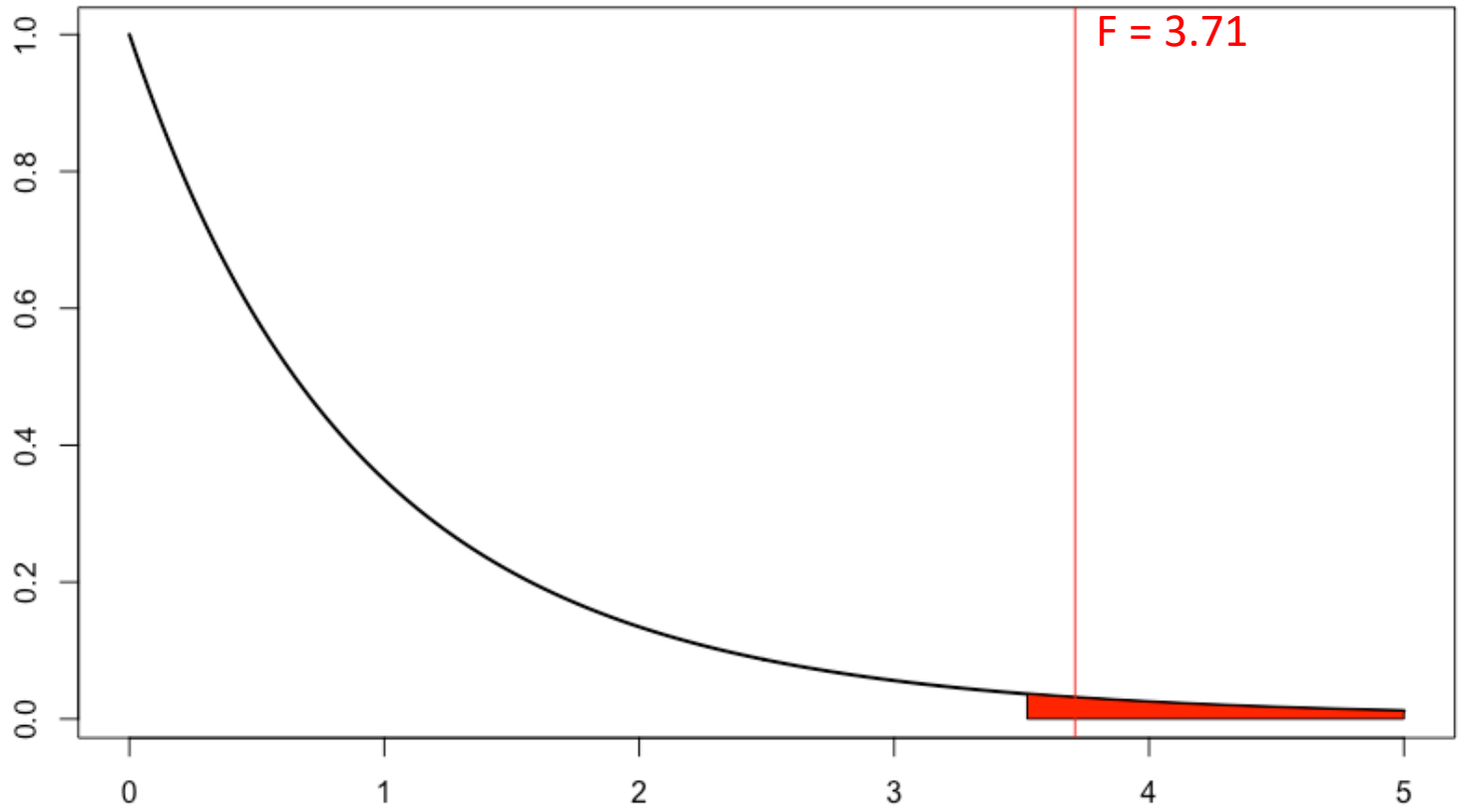
One-way ANOVA – Example II (cont.)

3. Calculate **critical values**/p value
4. Decide whether to reject/fail to reject H_0



One-way ANOVA – Example II (cont.)

3. Calculate critical values/**p value**
4. Decide whether to reject/fail to reject H_0



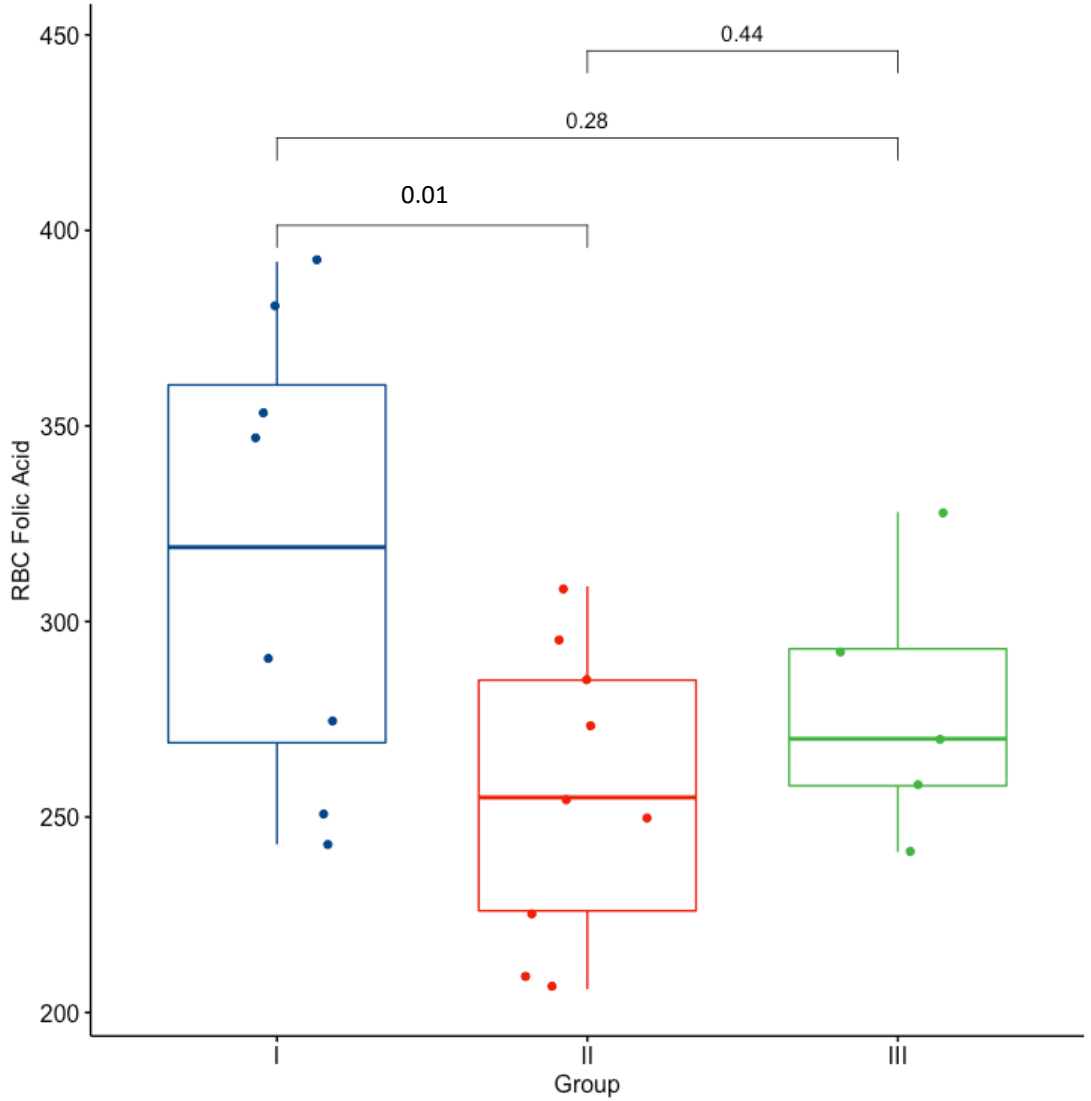
p = 0.043631

One-way ANOVA – Example II (cont.)

- With 95% confidence, we can conclude that the mean RBC folic acid level of at least one group is significantly different than the others

- Next, we perform 2-sample t-tests between all pairs of groups

One-way ANOVA – Example II (cont.)



χ^2 Test of Association

- Used to assess the association between two categorical variables
- More generally, used to investigate the significance of the difference between expected and observed values

- Are the 2 categorical variables **independent**?

χ^2 Test – Test Statistic

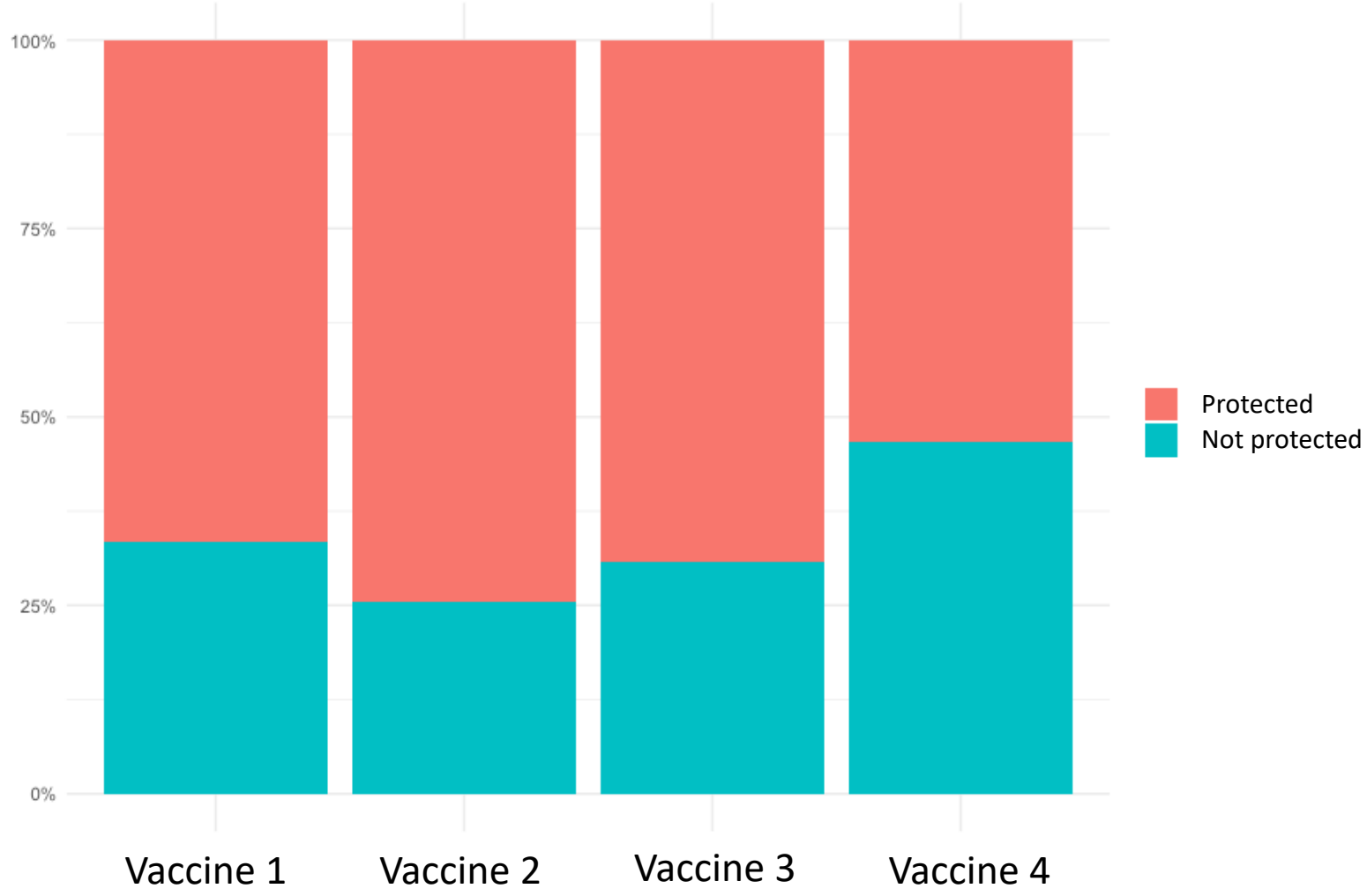
$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

χ^2 Test – Example

- Is the protection status dependent on different COVID vaccines?

	Protected	Not protected
Vaccine 1	82	41
Vaccine 2	70	24
Vaccine 3	45	20
Vaccine 4	48	42

χ^2 Test – Example



χ^2 Test – Example

1. Check assumptions, determine H_0 and H_a, choose α
   - H_0: there is **no association** between protection status and vaccine type

     H_a: there **is association** between protection status and vaccine type
   - $\alpha = 0.05$

2. Calculate the appropriate test statistic

   $\chi_H^2 = 9.297 \sim \chi_3^2$

χ^2 Test – Example

	Proctected	Not protected	Total
Vaccine 1	82	41	**123**
Vaccine 2	70	24	**94**
Vaccine 3	45	20	**65**
Vaccine 4	48	42	**90**
Total	**245**	**127**	**372**

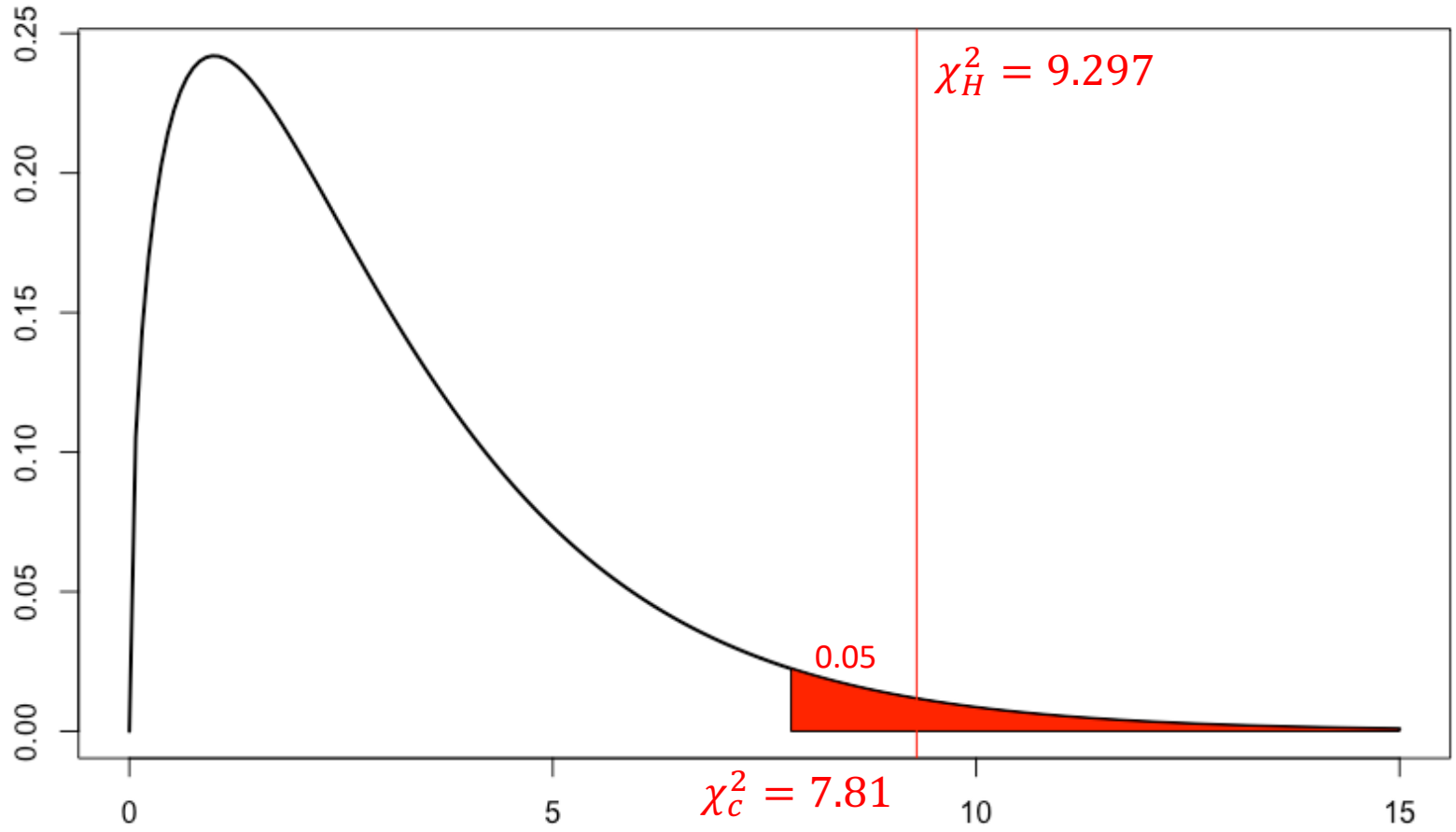
$$expected_{4,1} = 245 \times \frac{90}{372} = 59$$

$$expected_{4,2} = 127 \times \frac{90}{372} = 31$$

$$\chi^2_H = \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}} \sim \chi^2_{(m-1)(n-1)}$$

$$\boldsymbol{\chi^2_H = 9.297 \sim \chi^2_3}$$

χ^2 Test – Example



p = 0.025592