Biostatistics
Week XIV

Ege Ülgen, M.D.

6 January 2022

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

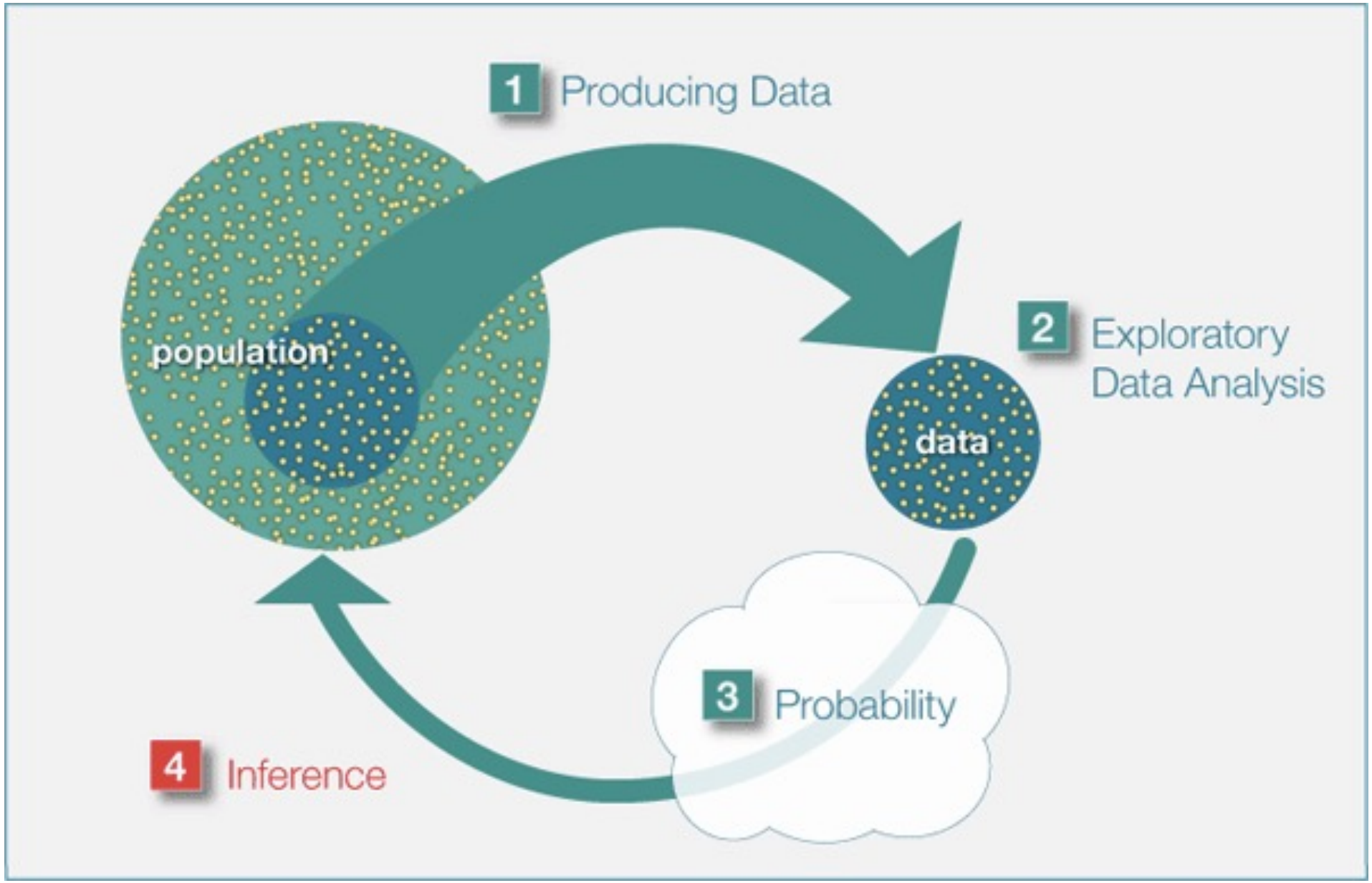
Population vs. Sample

- 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/

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

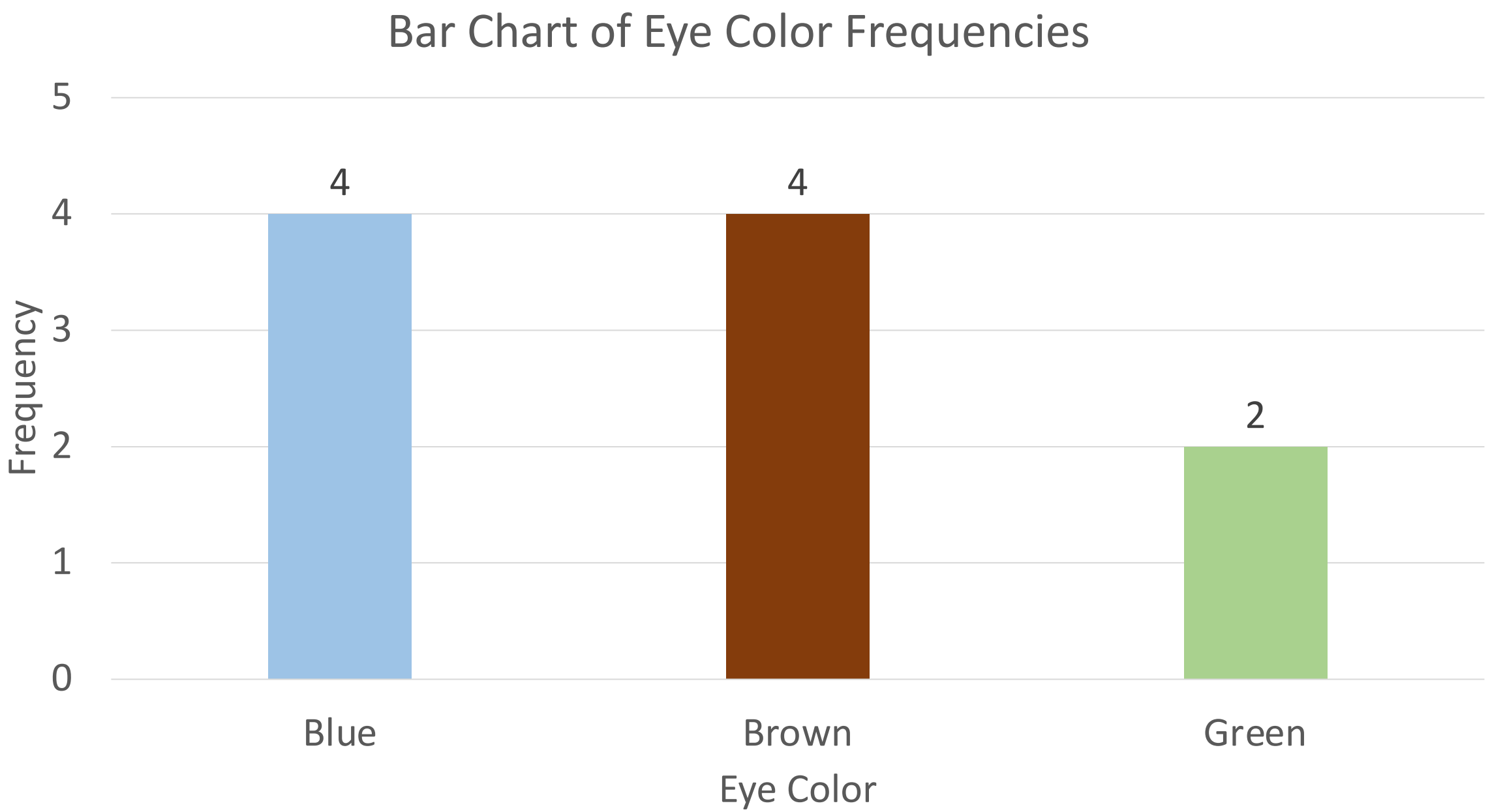
  - **Interval:** real number (+/- including 0). E.g., temperature, location
  - **Ratio:** positive values (**0 indicates none**). E.g., height, age, daily calcium consumption (mg).

Frequency Tables – Categorical Variable

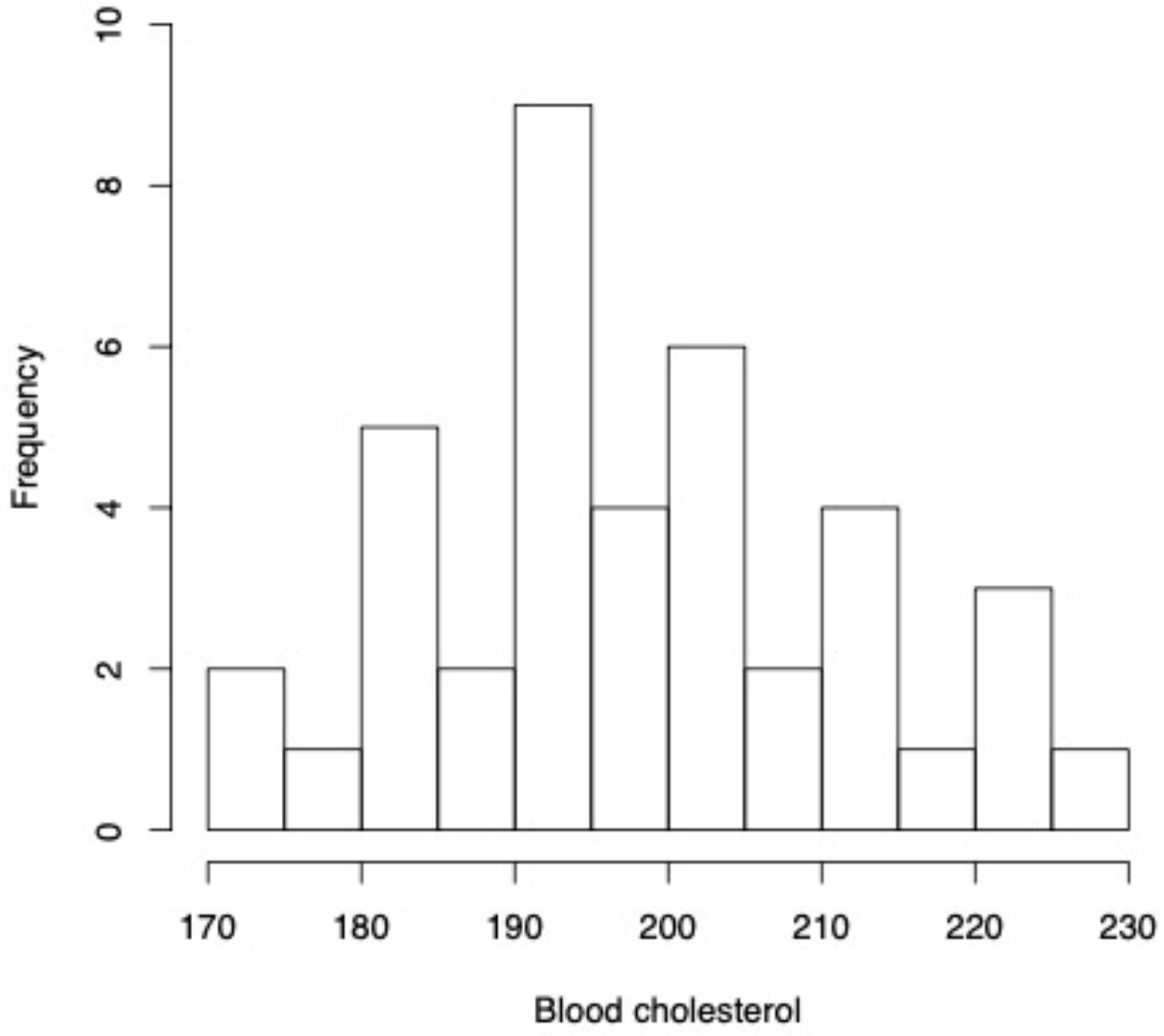
- Eye colors of 10 individuals:

  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

Histogram


Frequency
0
2
4
6
8
10
170
180
190
200
210
220
230
Blood cholesterol

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

Histogram

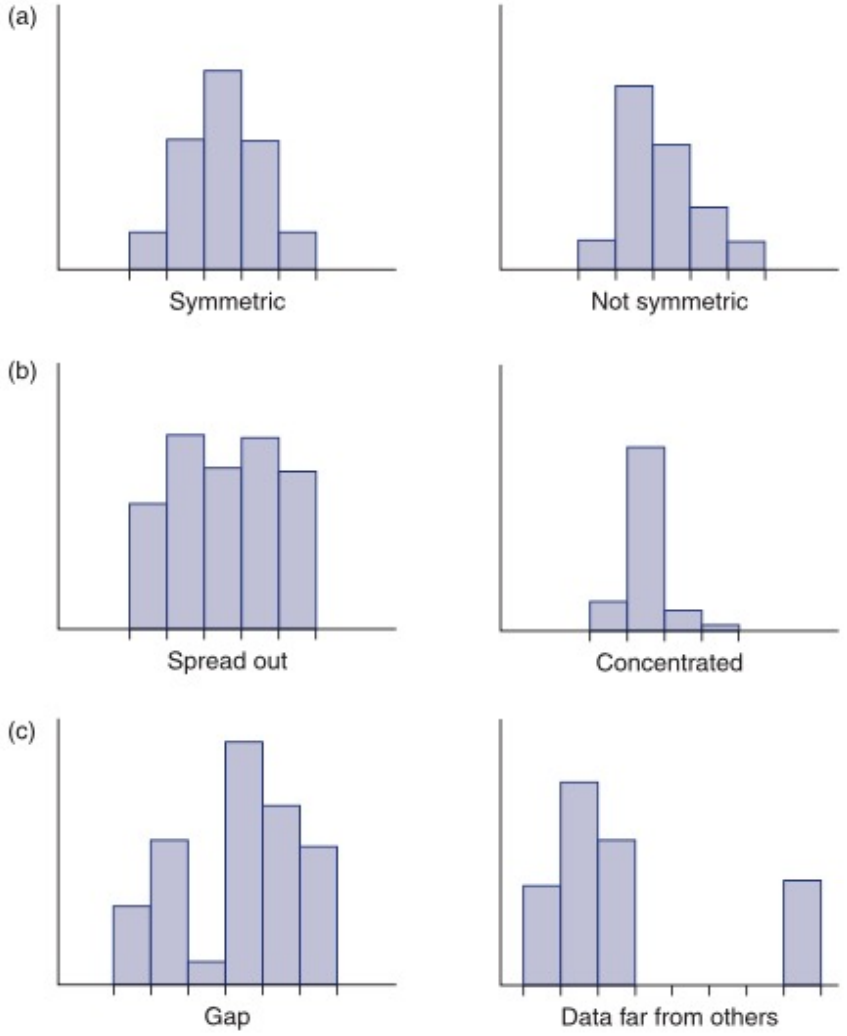


FIGURE 2.8

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.

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

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

Mean

If $y_i = x_i + c$ (c is a constant) $\bar{y} = \bar{x} + c$

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

$$\bar{y} = \frac{(213+5)+(174+5)+\ldots+(227+5)}{40} = 202.625$$

Mean

If $y_i = x_i \; x \; c$ (c is a constant) $\quad \bar{y} = \bar{x} \; x \; c$

x: 1, 2, 3, 4, 5

y: 3 (1 * 3), 6 (2 * 3), 9 (3 * 3), 12 (4 * 3), 15 (5 * 3)
$\Rightarrow c = 3$

$\bar{x} = 3, \; \bar{y} = 9 \Rightarrow \bar{y} = 3 * \bar{x}$

Mean

- Even a small change in a single value affects the mean

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

- If the maximal value was 700 (instead of 227), the mean would be 209.45 (instead of 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 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

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:

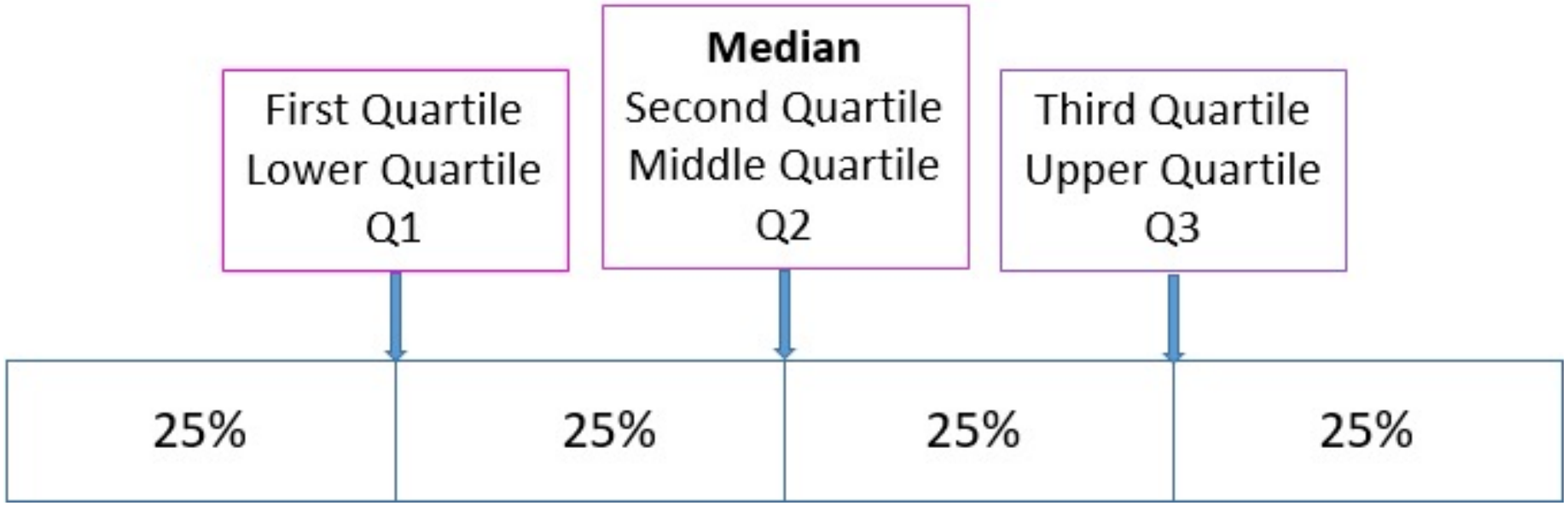
90, 80, **100**, 110, **100**, 120, **100**, 90, **100**, 110, 120, 110

Mode = 100

Mean = 102.5

Median = 100

Quartiles

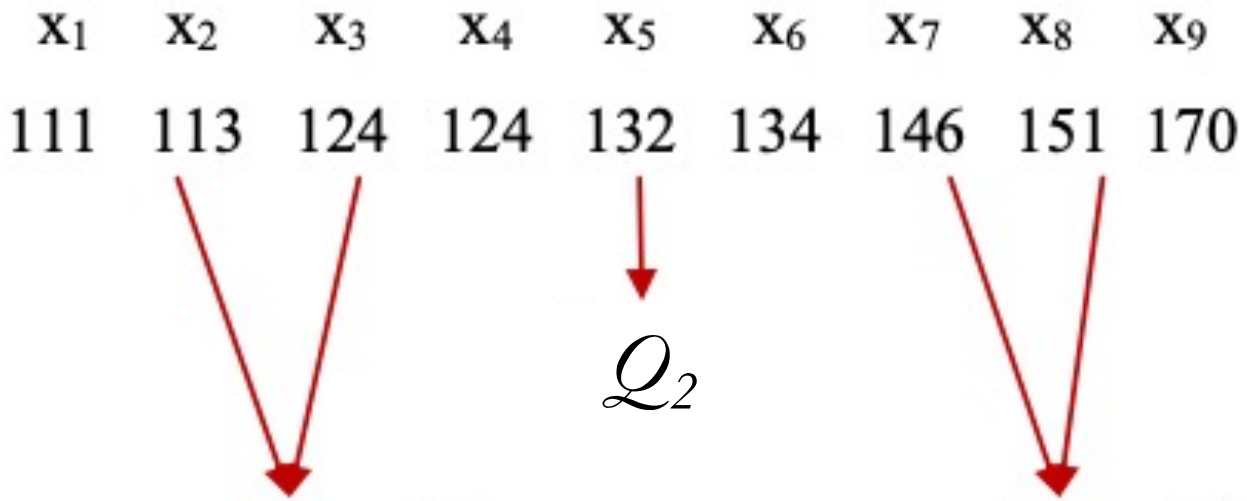


Median, quartiles, percentiles(Video lessons, examples, solutions) [Internet]. www.onlinemathlearning.com. [cited 2021 Oct 4]. Available from: https://www.onlinemathlearning.com/quartile.html

Quartiles

- Systolic blood pressure measurements of 9 patients:

151, 124, 132, 170, 146, 124, 113, 111, 134



$$Q_1 = \frac{113 + 124}{2} = 118.5 \qquad Q_3 = \frac{146 + 151}{2} = 148.5$$

Percentiles - Definition

$100 * p$ percentile ($0 \leq p \leq 1$) is the data value for which:

- at least $100 * p$ of the data values are less than or equal to it
- at least $100 * (1 - p)$ of the data values are greater than or equal to it

* If there are two values that satisfy the above conditions, the average of these values is taken as the $100 * p$ percentile

Percentiles - Algorithm

- Sort values in ascending order
- Calculate n * p
  - If n * p is not an integer, take the smallest integer greater than n * p
  - If n * p is an integer take the average of n * p^{th} and $(n * p + 1)^{th}$ values

Percentiles - Example

- 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

- 25th percentile (1st quartile, Q1): 189.5 (40 * 0.25 = 10)
- 50th percentile (median, Q2): 195.5 (40 * 0.5 = 20)
- 75th percentile (3rd quartile, Q3): 205.5 (40 * 0.75 = 30)
- 90th percentile : 218 (40 * 0.9 = 36)
- 95th percentile: 221 (40 * 0.95 = 38)
- 97.5th percentile: 224 (40 * 0.975 = 39)

Range

- The difference between the maximal and minimal value

$$R = maximum - minimum$$

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

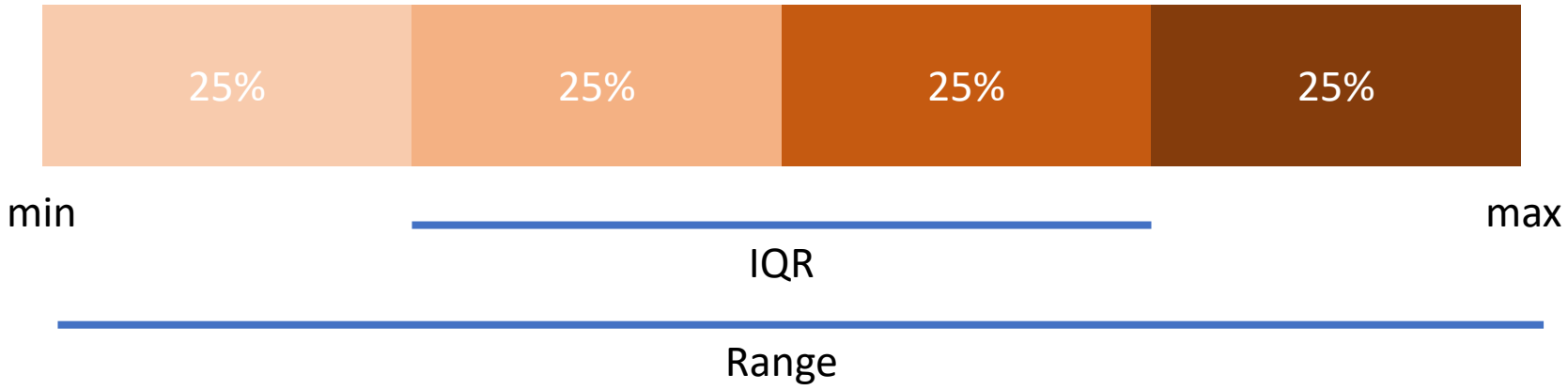
30, 12, 15, 22, 40, 55, 20, 58, 25, 60, 23, 72

$$R = 72 - 12 = 60$$

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 and Standard Deviation

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

Sample Variance

If $y_i = x_i + c$ (c is a constant), Var(y) = Var(x)

If $y_i = x_i * c$ (c is a constant), Var(y) = c^2 Var(x)

Units

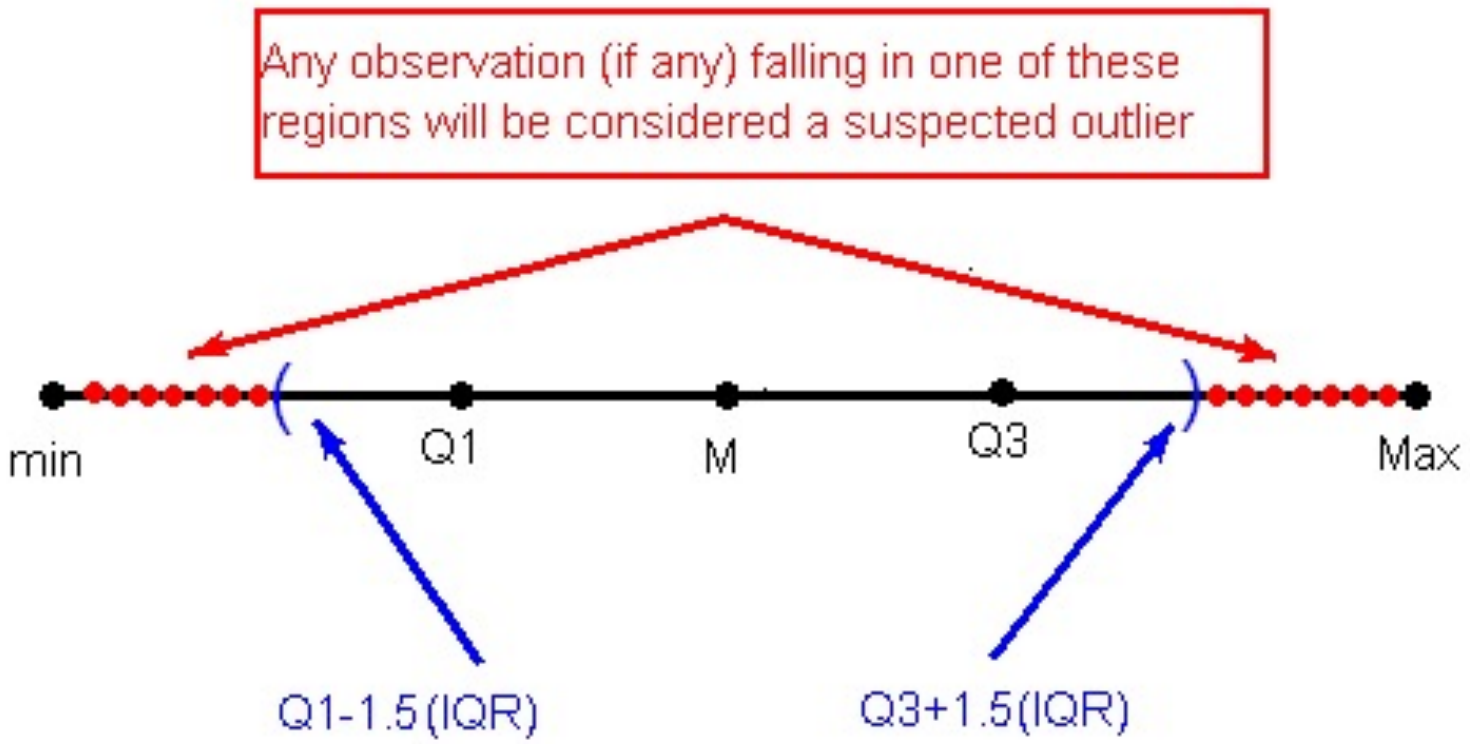
- Mean: same unit with the data
- Median: same unit with the data
- Mode: same unit with the data
- Quartiles: same unit with the data
- Percentiles: same unit with the data
- Variance: square of the unit of the data
- Standard deviation: same unit with the data

Outliers

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

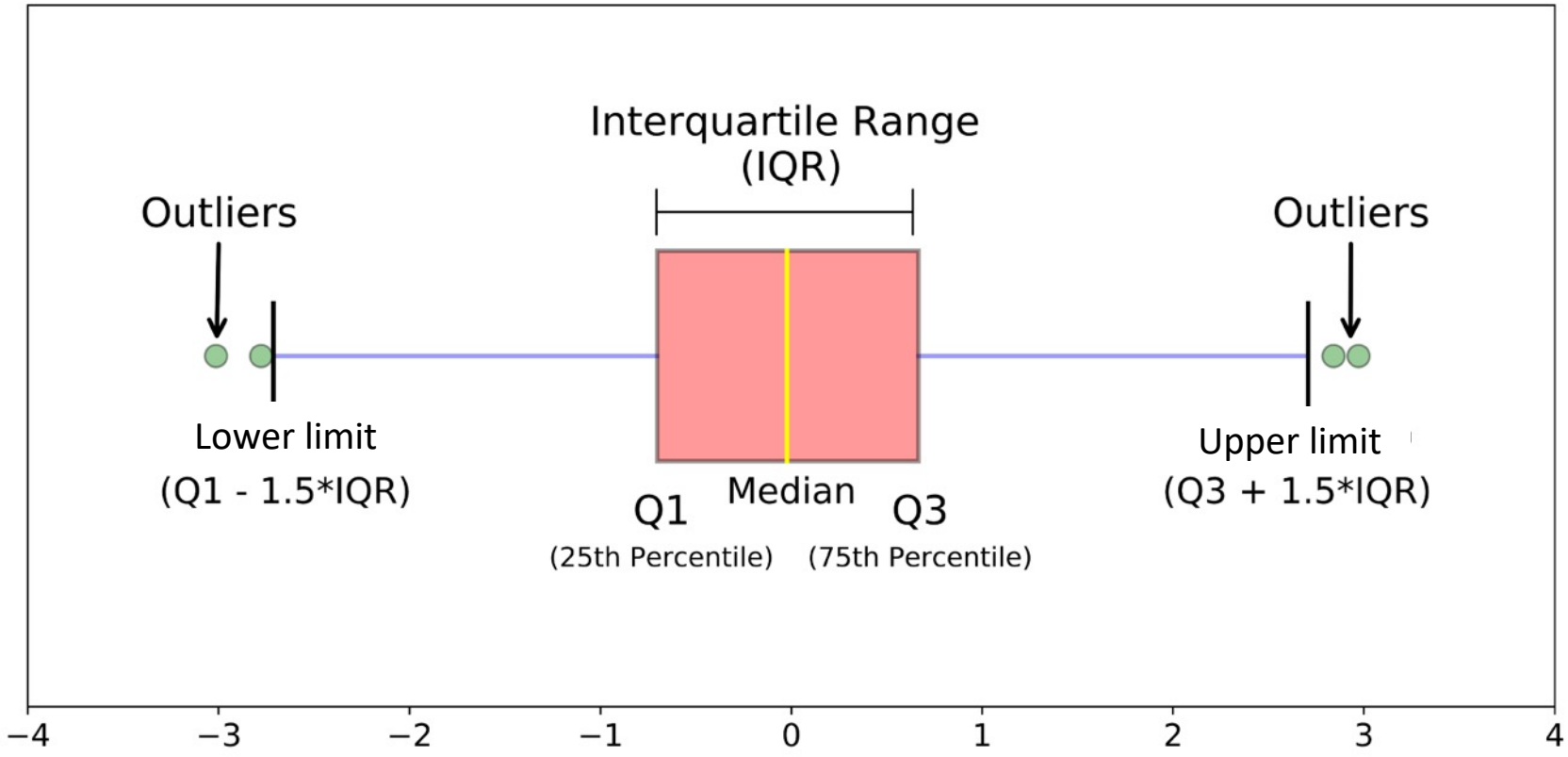
- For
  - Lower Limit = Q_1 - 1.5 * IQR
  - Upper Limit = Q_3 + 1.5 * IQR
- Outliers are defined as any value(s) larger than the upper limit or smaller than the lower limit

Outliers



Outliers [Internet]. [cited 2021 Oct 4]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/understanding-outliers/

Box Plot



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

Hypothesis Testing

Hypothesis Testing

- **Hypothesis**: an assumption that can be tested based on the evidence available
  - A novel drug is efficient in treating a certain disease
  - Regular smoking leads to lung cancer
  - Overweight individuals who (1) consume greasy food and (2) consume a low amount vegetables (1) have high levels of cholesterol and (2) have a higher risk of cardiovascular diseases
- **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 (or smaller) than 0.5

One-tailed

H_0: $p \leq 0.5$ (or $p \geq 0.5$)

H_a: $p > 0.5$ (or $p < 0.5$)

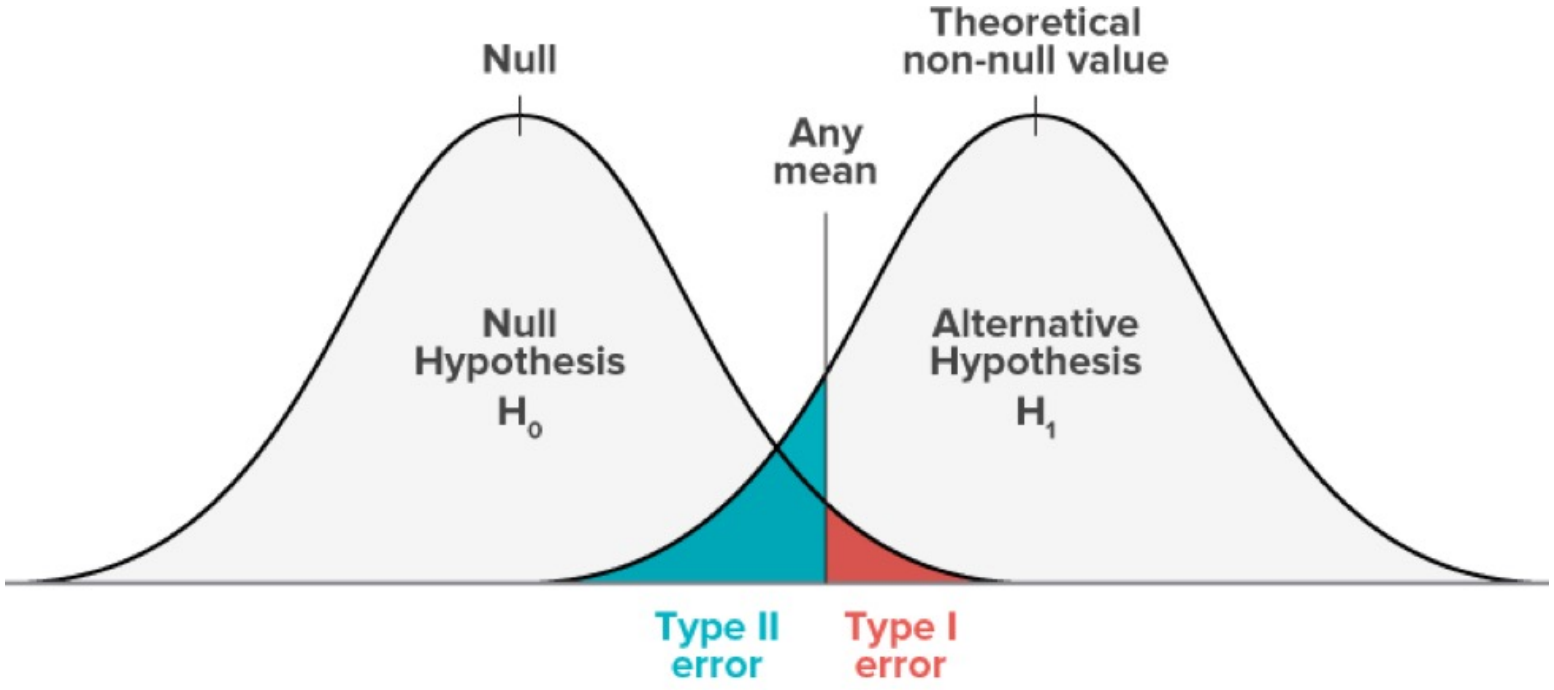
H_0	Decision	
	Fail to reject	Reject
True	Correct decision	**Type I Error** α
False	**Type II Error** **ß**	Correct decision

Barai A. Hypothesis Testing | All a beginner needs to know [Internet]. Medium. 2020 [cited 2021 Oct 4]. Available from: https://medium.com/analytics-vidhya/hypothesis-testing-all-a-beginner-needs-to-know-dd1bdeebdbf1

Hypothesis Testing

- P(Type 1 error) = α = P(reject H_0| H_0 is true)
- P(Type 2 error) = β = P(fail to reject H_0| H_0 is false)

- As α gets larger β gets smaller, vice versa
- As *n* gets large, both α and β get smaller



Hypothesis Testing

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

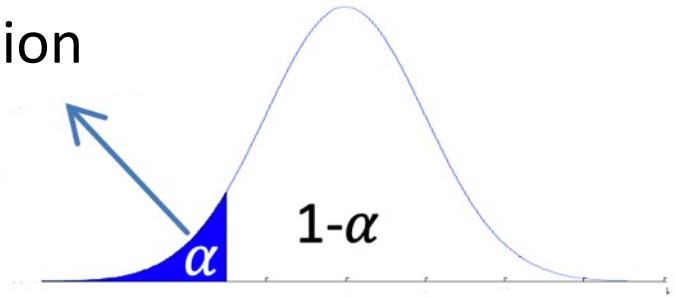
- **Confidence level** = $1 - \alpha$
  - P(fail to reject H_0| H_0 is true)
- **Statistical power** = $1 - \beta$
  - P(reject H_0| H_0 is false)

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/$p \leq \alpha$

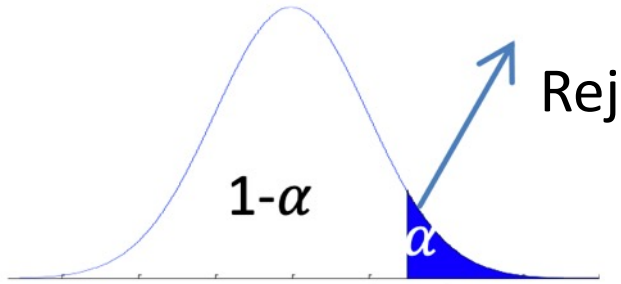
$$H_0: \mu \geq \mu_0$$
$$H_1: \mu < \mu_0$$

Rejection region



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

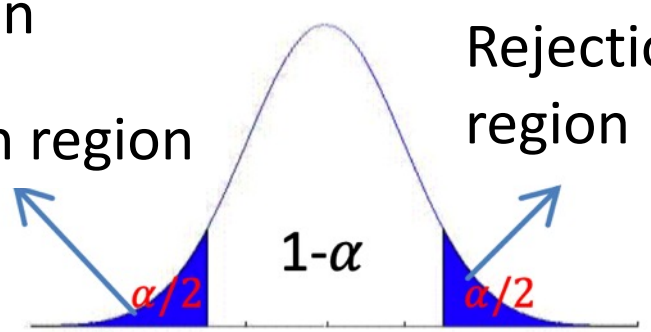
Rejection region



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

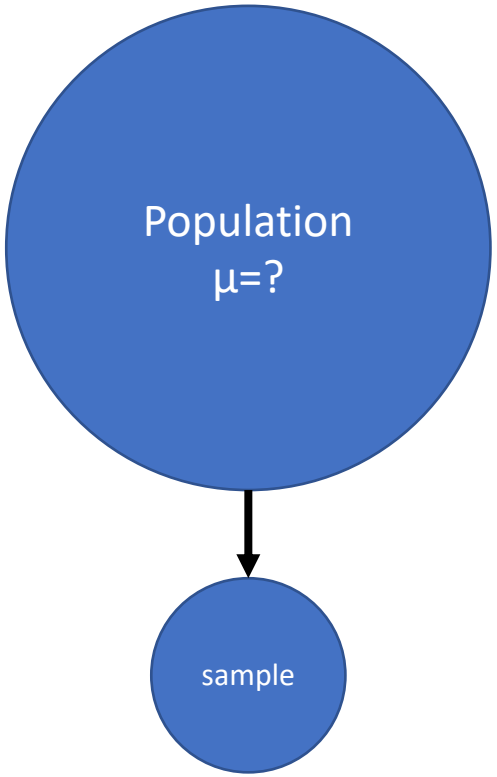
Rejection region

Rejection region



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

- Mean percentage benefit 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)
   - $H_0: \mu = 0$ $H_a: \mu \neq 0$
   - $\alpha = 0.05$

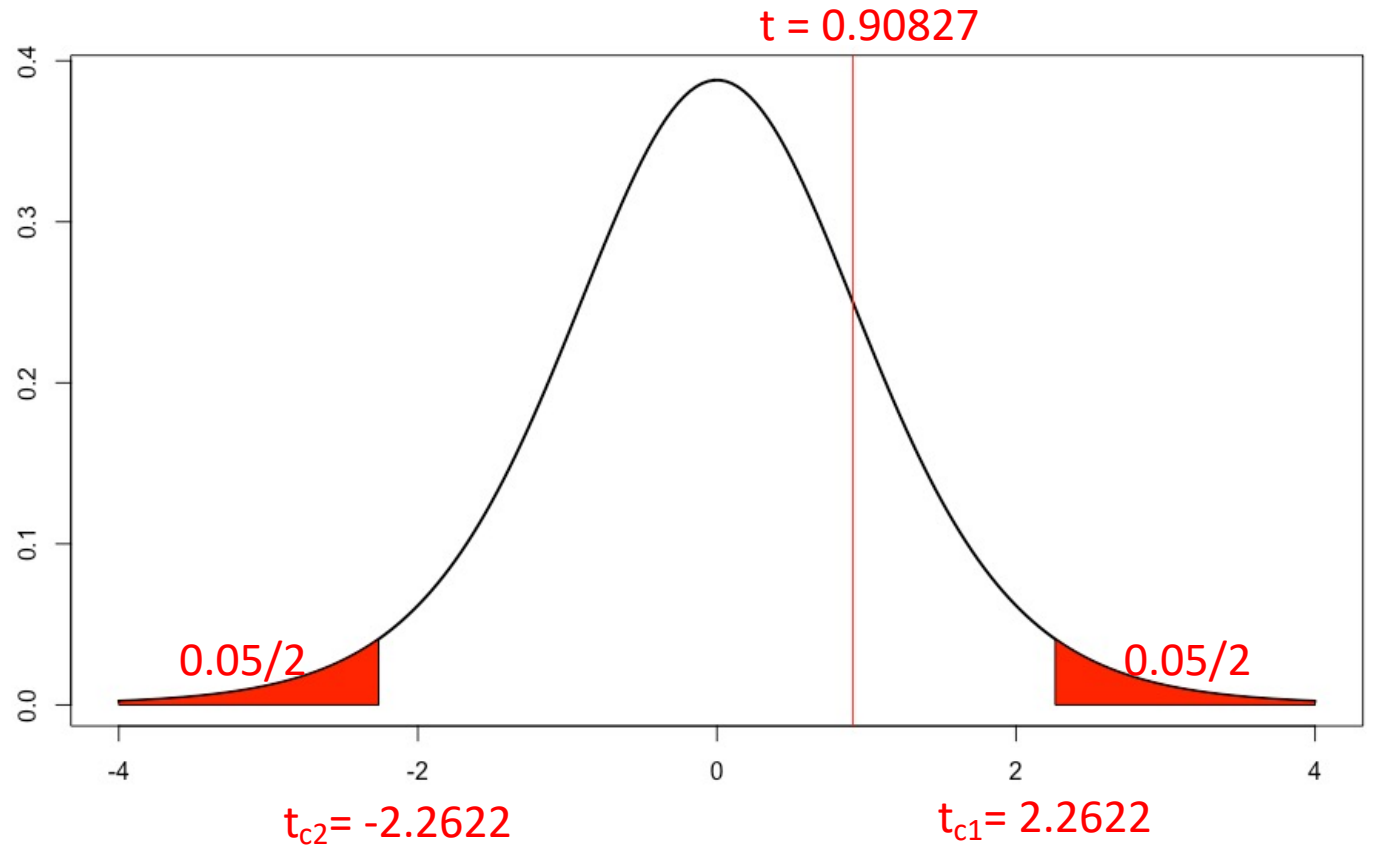
One-Sample t-Test – Example I (cont.)

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

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

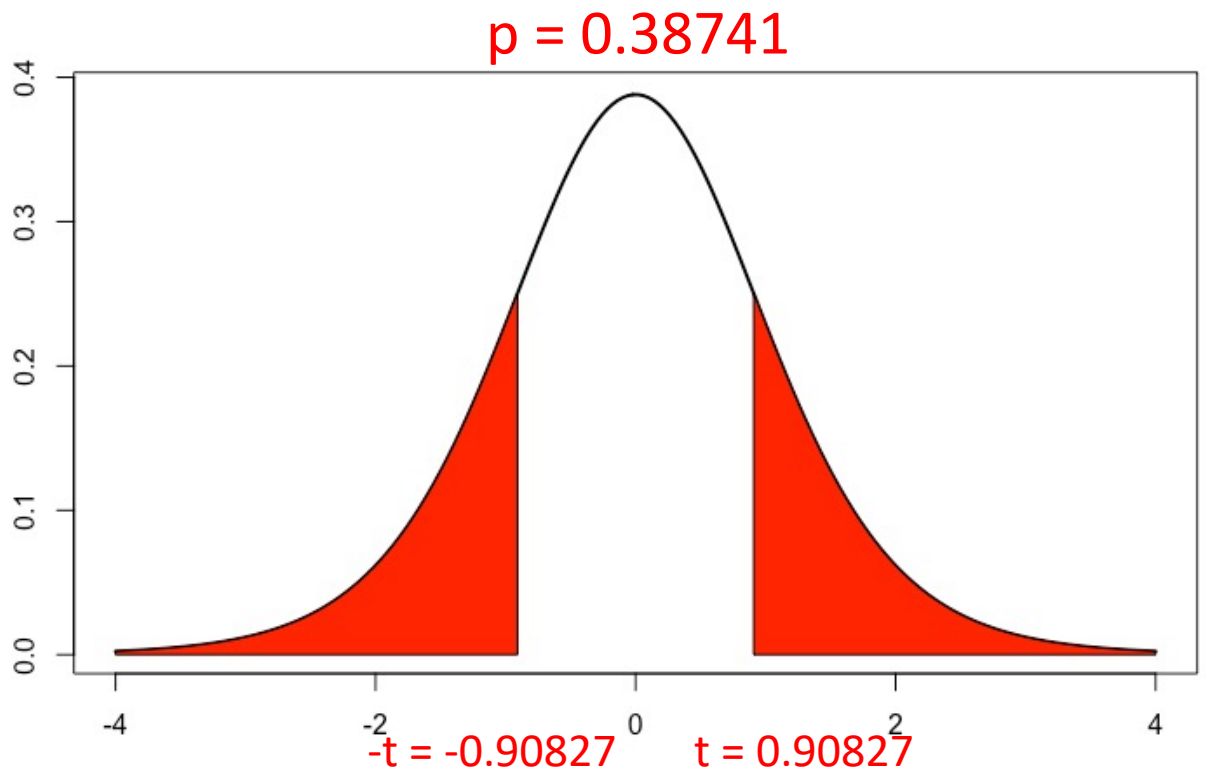
One-Sample t-Test – Example I (cont.)

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



One-Sample t-Test – Example II

- It is claimed that:
- A novel drug reduces the recovery time of patients to less than 10 days
- Recovery time for 7 randomly-selected patients:

  2, 4, 11, 3, 4, 6, 8 ($\bar{X}$ = 5.43, s = 3.15)

- Test the hypothesis using α = 0.01

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

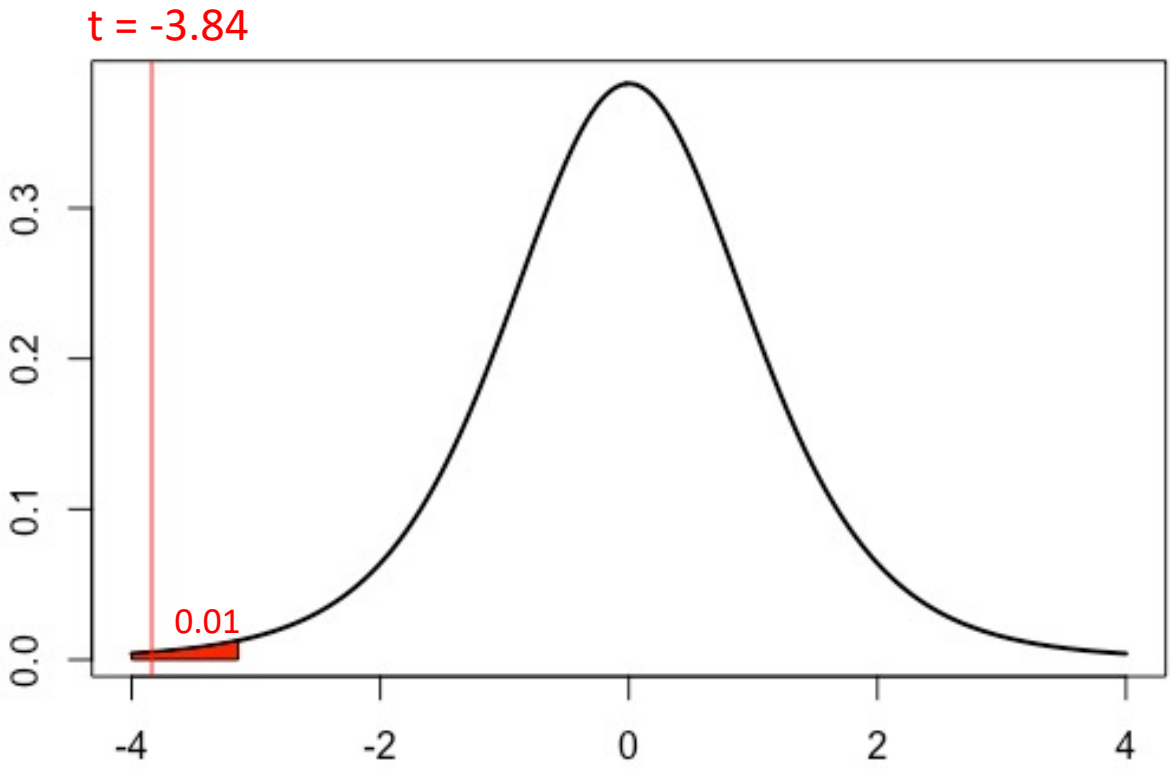
1. Check assumptions, determine H_0 and H_a, choose α
   - Normality of the variable is checked
   - H_0: $\mu \geq 10$ H_a: $\mu < 10$
   - $\alpha = 0.01$

2. Calculate the appropriate test statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{5.43 - 10}{3.15/\sqrt{7}} = -3.84 \quad (\sim t_{n-1} = t_6)$$

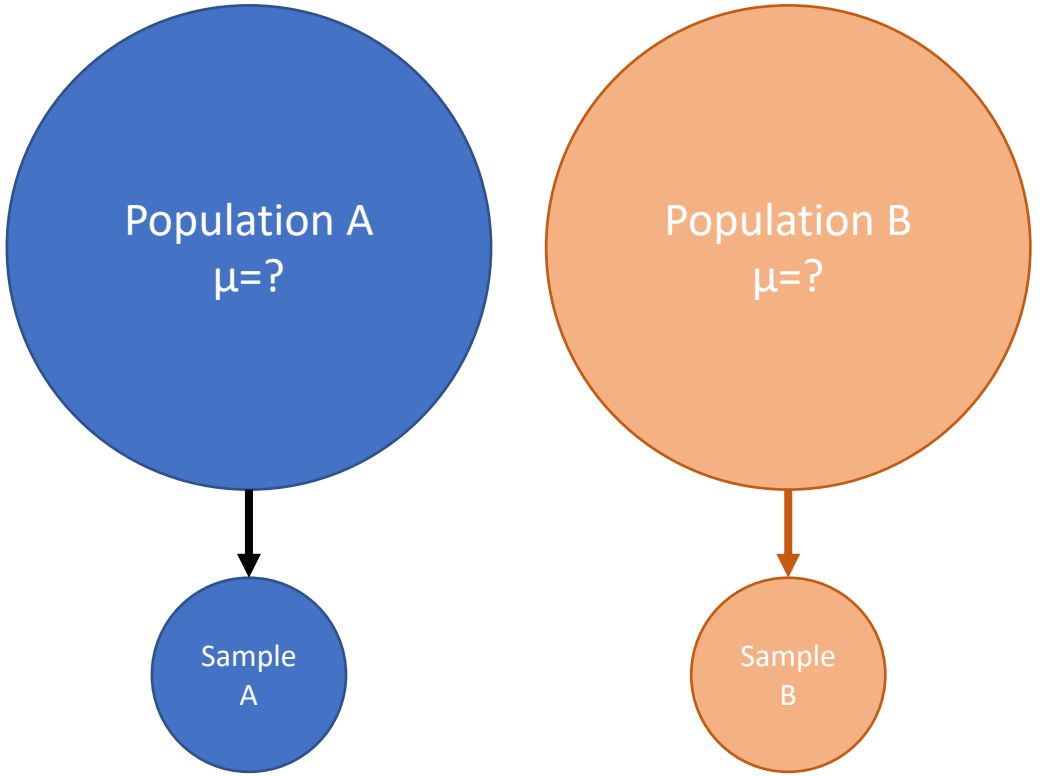
One-Sample t-Test – Example II (cont.)

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.)

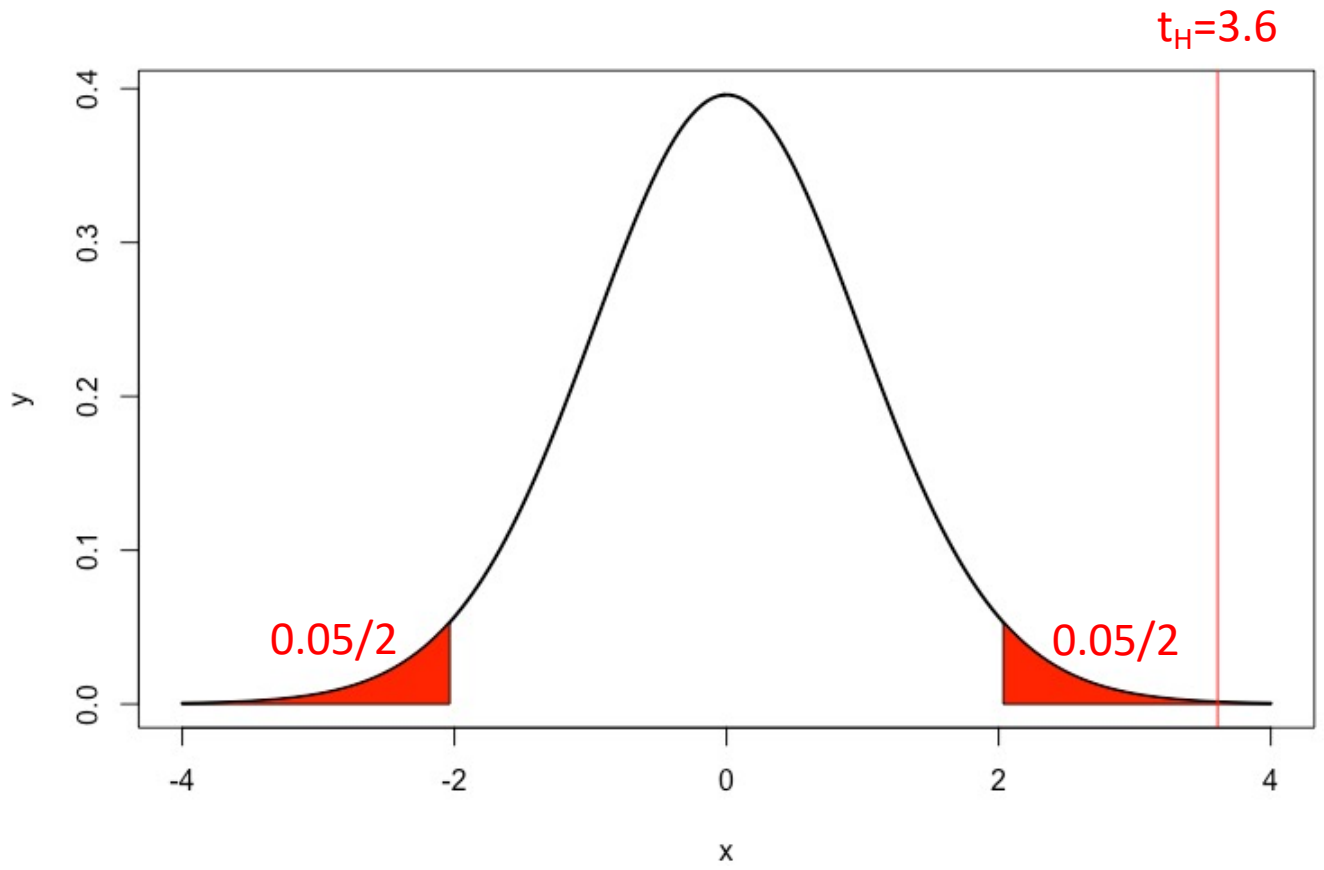
1. Check assumptions, determine H_0 and Ha, choose α
   - We check that the variables are normally distributed
   - $H_0: \mu_c = \mu_T$ $H_a: \mu_c \neq \mu_T$
   - $\alpha = 0.05$

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

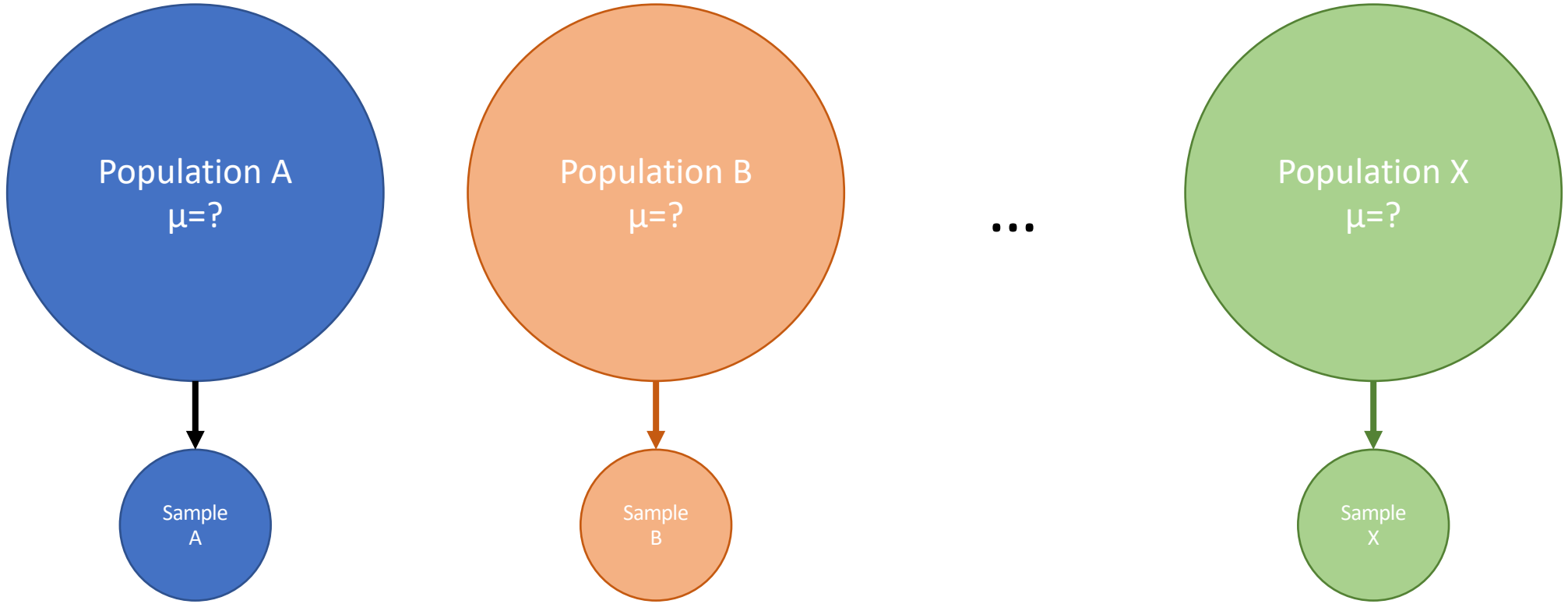
	Control Group (n = 20)	High-Flow Nasal Oxygenation Group (n = 20)	Mean Difference (95% CI)	*P* Value
Safe apnea time (s)	185.5 ± 53.0	261.4 ± 77.7	75.9 (33.3–118.5)	.001
Minimum Spo_2 (%)	87.9 ± 4.7	90.9 ± 3.5	3.1 (0.4–5.7)	.026
Plateau $ETco_2$ (mm Hg)	38.8 ± 2.5	37.9 ± 3.0	−0.8 (−2.6 to 0.9)	.33
Time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

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.

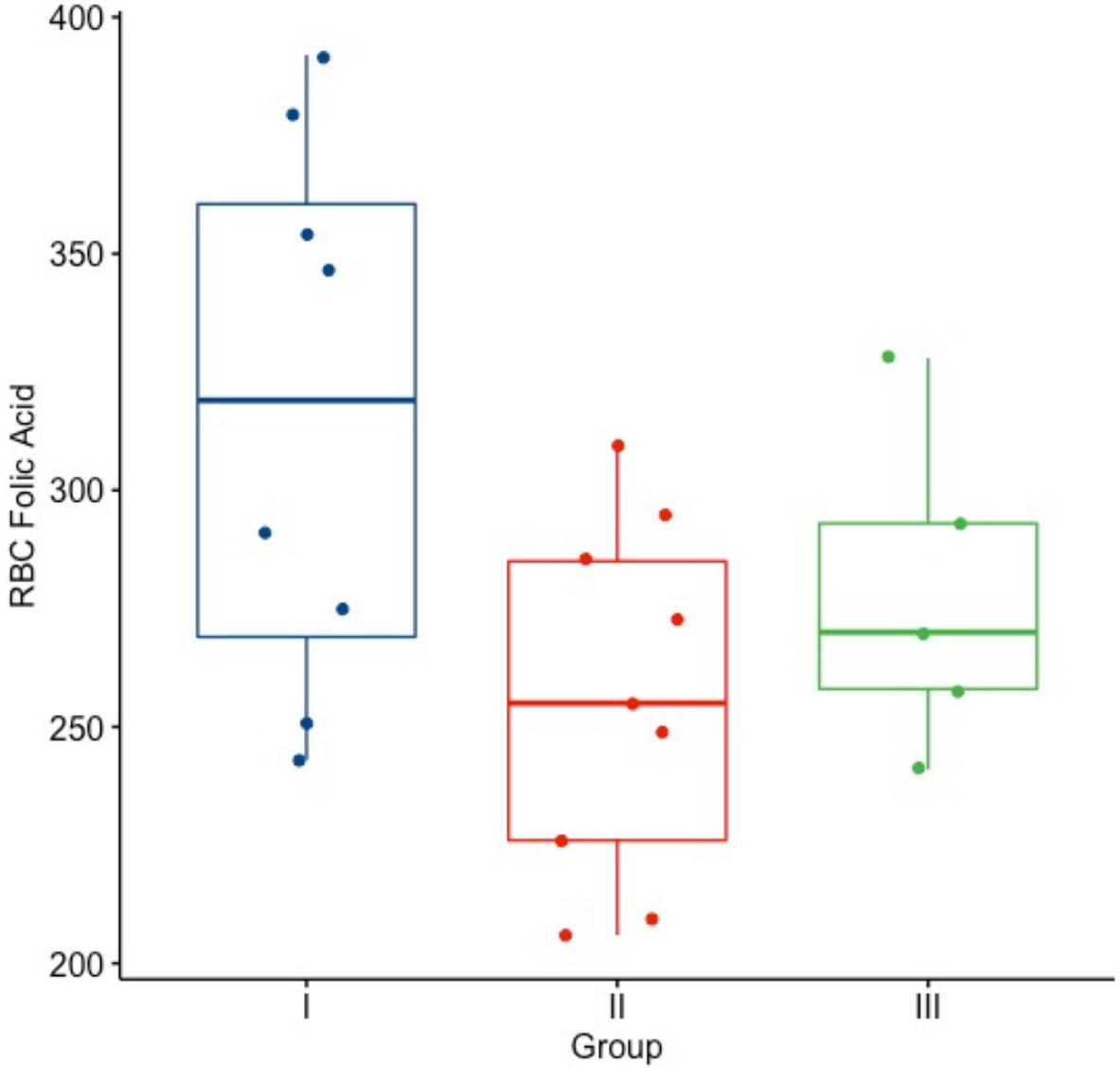
Group II.—9 patients received approximately 50% nitrous oxide and 50% oxygen mixture only during the operation (5–12 h) and thereafter 35–50% oxygen for the remainder of the 24 h period.

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.)



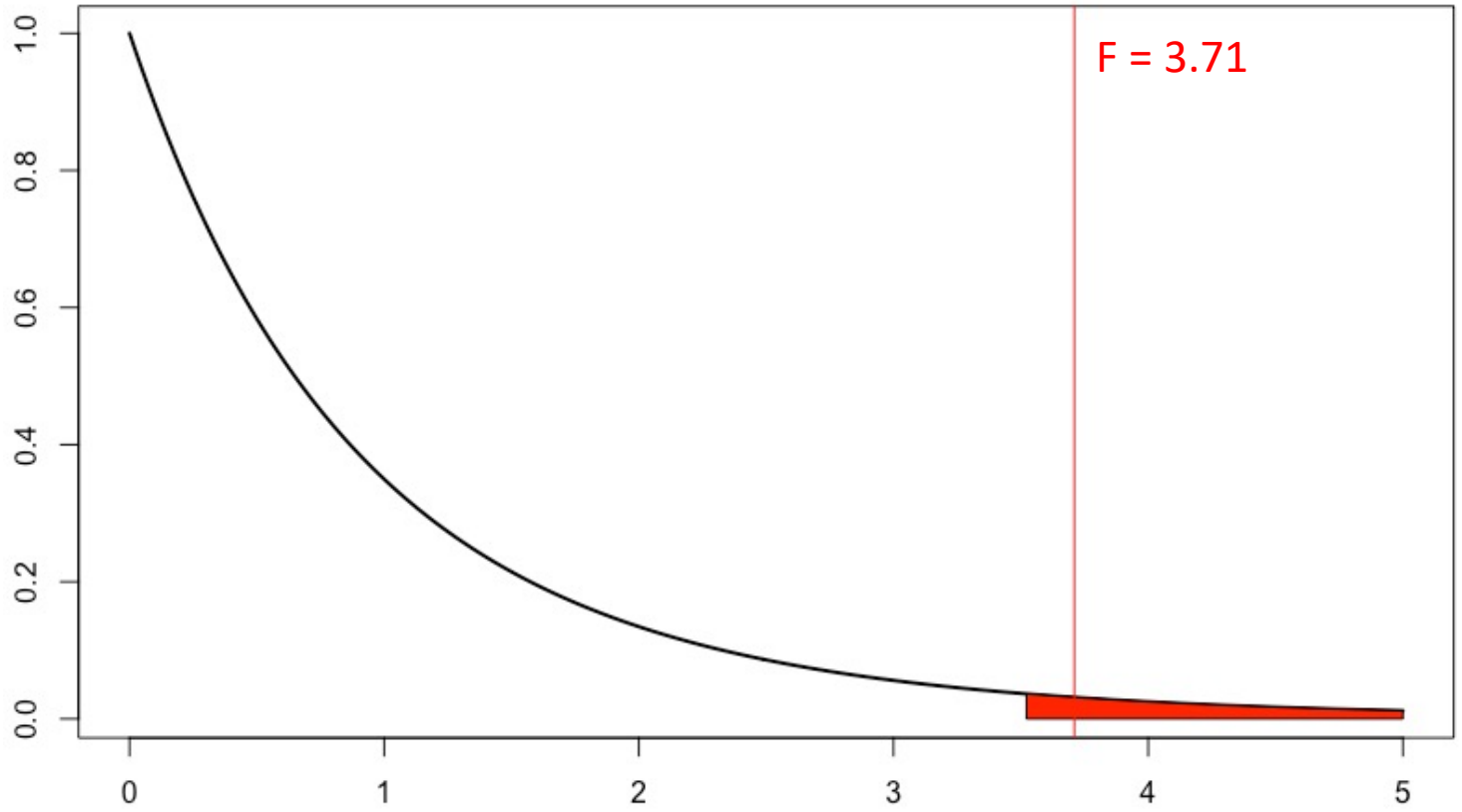
One-way ANOVA – Example II (cont.)

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 $\sim F_{2,19}$

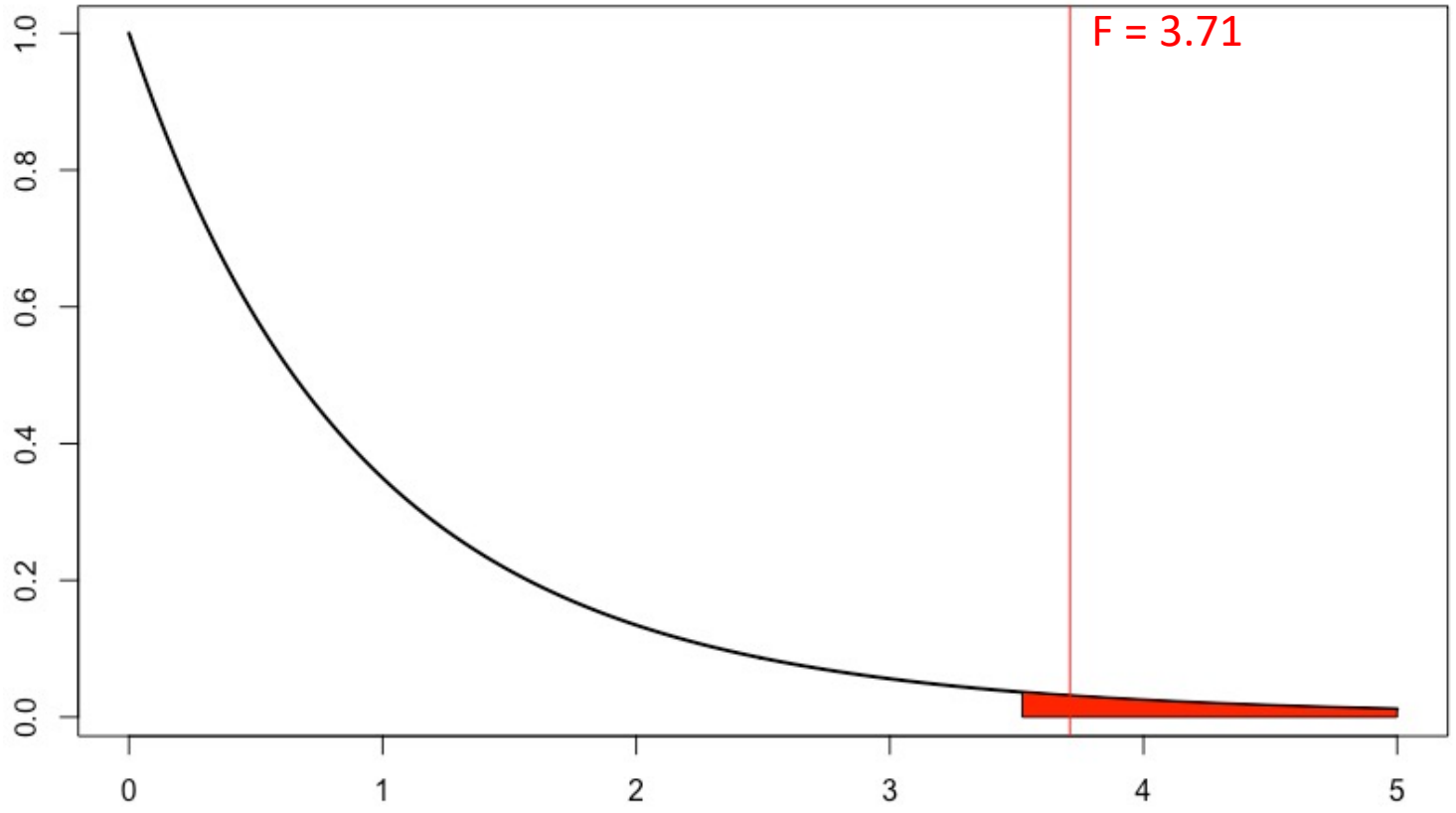
One-way ANOVA – Example II (cont.)

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



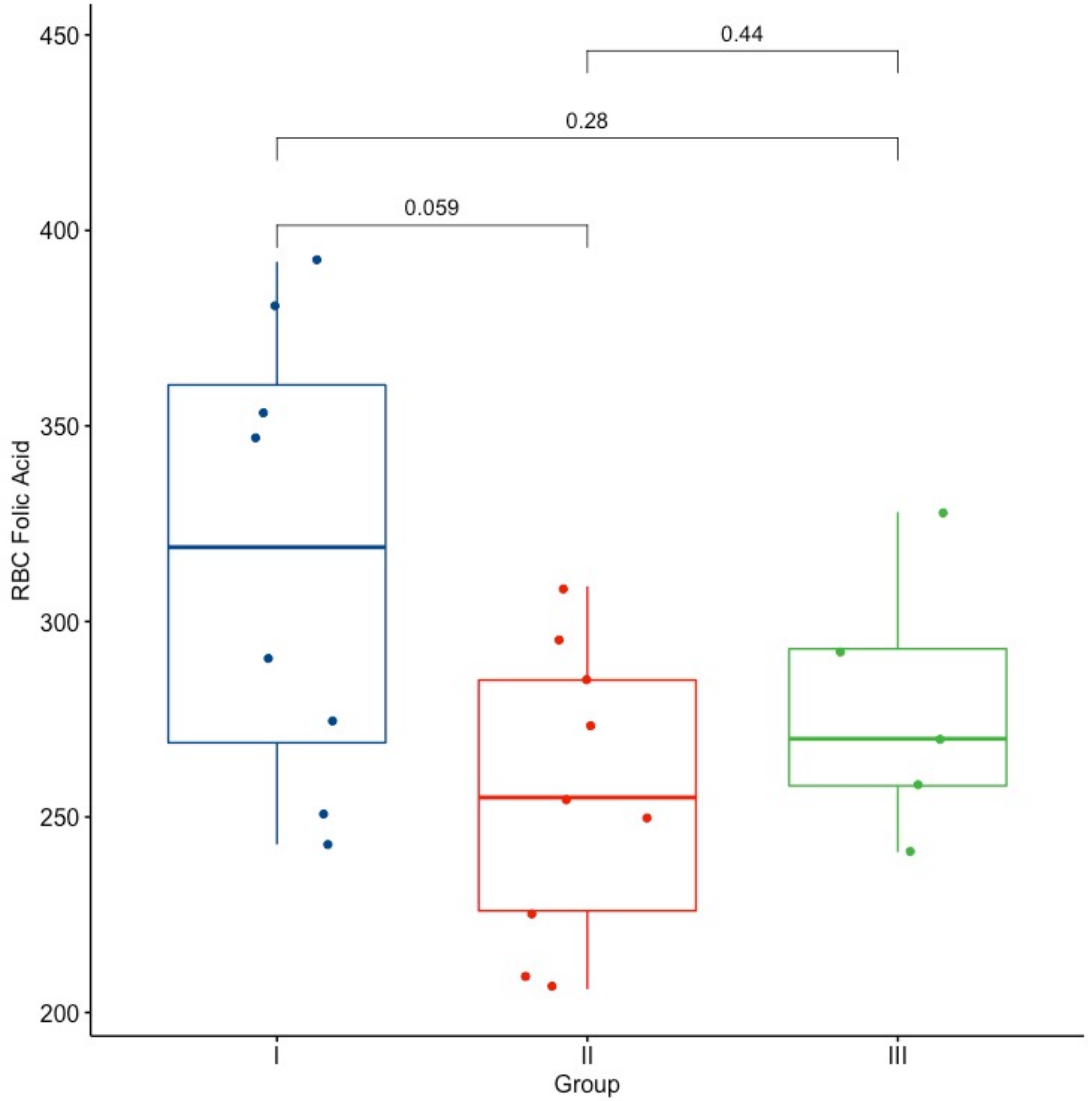
p = 0.043631

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

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

Cupples ME, McKnight A. Randomised controlled trial of health promotion in general practice for patients at high cardiovascular risk. BMJ. 1994 Oct 15;309(6960):993–6.

χ^2 Test – Example

TABLE III—*Changes in frequency of physical exercise in patients with angina between baseline and review at two years*

	No (%) of patients	
	Intervention group	Control group
Increased	108 (34)	63 (21)
No change	120 (38)	74 (25)
Decreased	89 (28)	163 (54)

χ^2 Test – Example

	Intervention Group	Control Group	Total
Increased	108	63	**171**
No change	120	74	**194**
Decreased	89	163	**252**
Total	**317**	**300**	**617**

$$expected_{1,1} = 317 \times \frac{171}{617} \qquad expected_{1,2} = 300 \times \frac{171}{617}$$

$$expected_{2,1} = 317 \times \frac{194}{617} \qquad expected_{2,2} = 300 \times \frac{194}{617}$$

$$expected_{3,1} = 317 \times \frac{252}{617} \qquad expected_{3,2} = 300 \times \frac{252}{617}$$

χ^2 Test – Example

OBSERVED	Intervention Group	Control Group
Increased	108	63
No change	120	74
Decreased	89	163

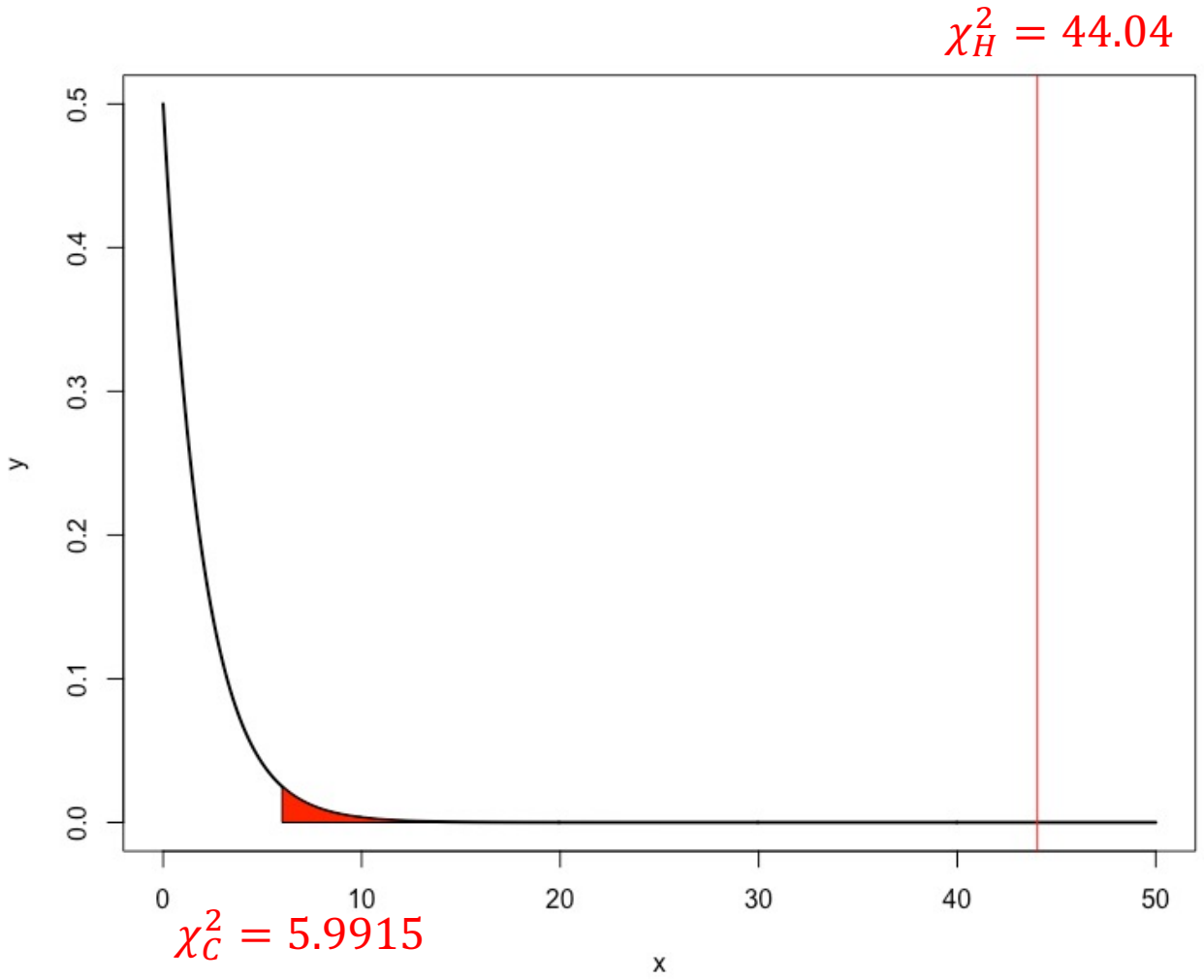
EXPECTED	Intervention Group	Control Group
Increased	87.86	83.14
No change	99.67	94.33
Decreased	139.47	122.53

χ^2 Test – Test Statistic

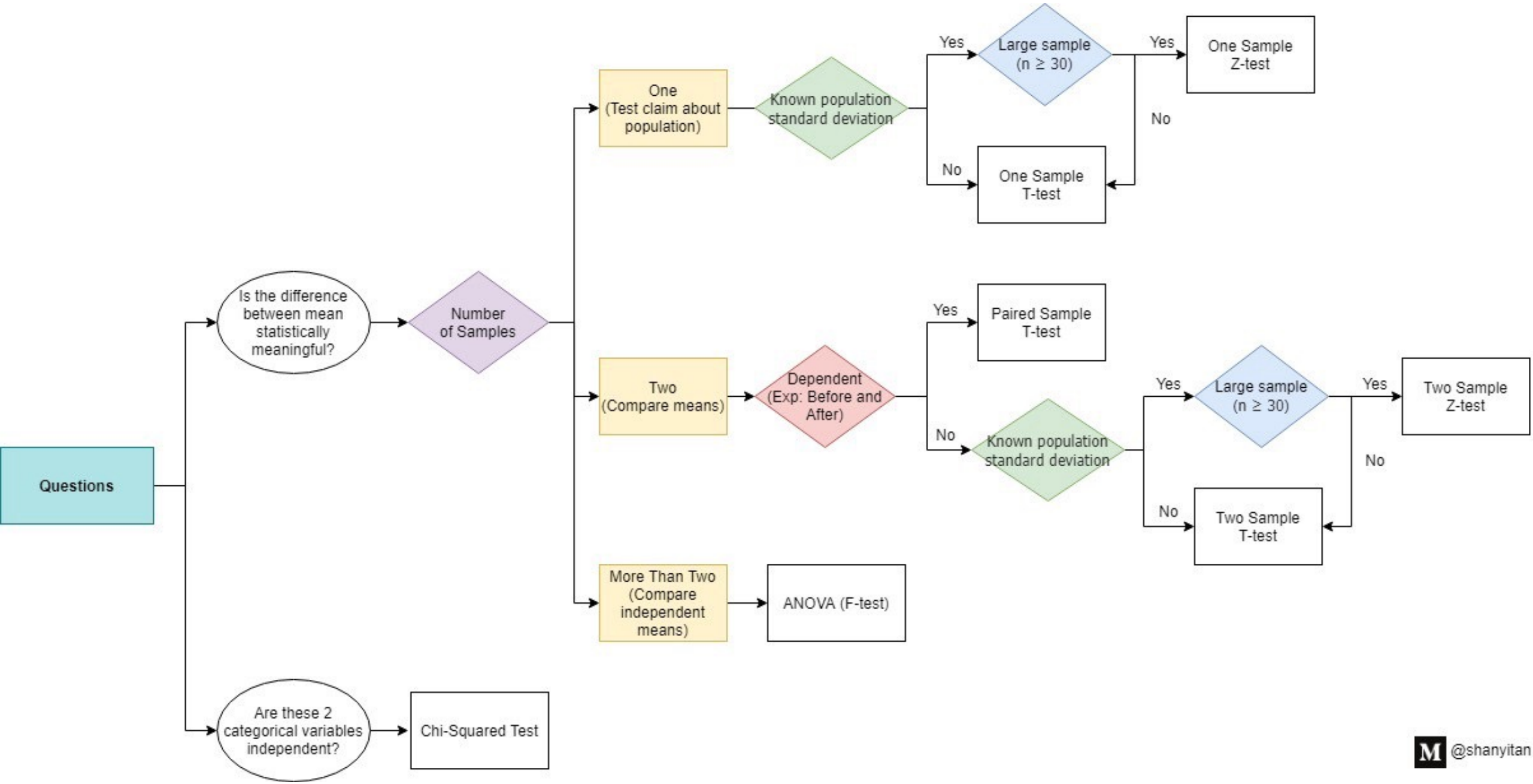
$$\chi^2 = \sum \frac{(observed\ - expected)^2}{expected}$$

$$\chi^2_H = 44.04 \sim \chi^2_{(3-1)(2-1)=2}$$

χ^2 Test – Test Statistic



p < 0.001


Questions
Is the difference between mean statistically meaningful?
Number of Samples
One (Test claim about population)
Known population standard deviation
Yes
Large sample (n ≥ 30)
Yes
One Sample Z-test
No
No
One Sample T-test
Two (Compare means)
Dependent (Exp: Before and After)
Yes
Paired Sample T-test
No
Known population standard deviation
Yes
Large sample (n ≥ 30)
Yes
Two Sample Z-test
No
No
Two Sample T-test
More Than Two (Compare independent means)
ANOVA (F-test)
Are these 2 categorical variables independent?
Chi-Squared Test
@shanyitan

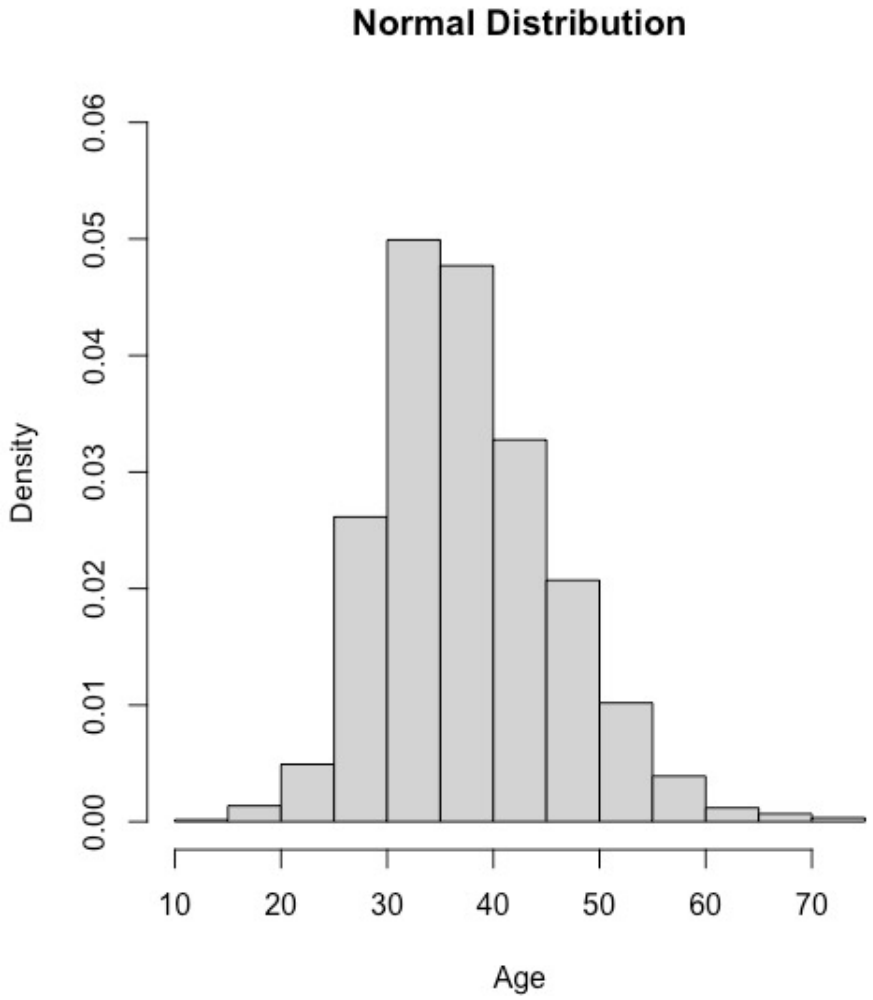
General Assumptions of Parametric Tests

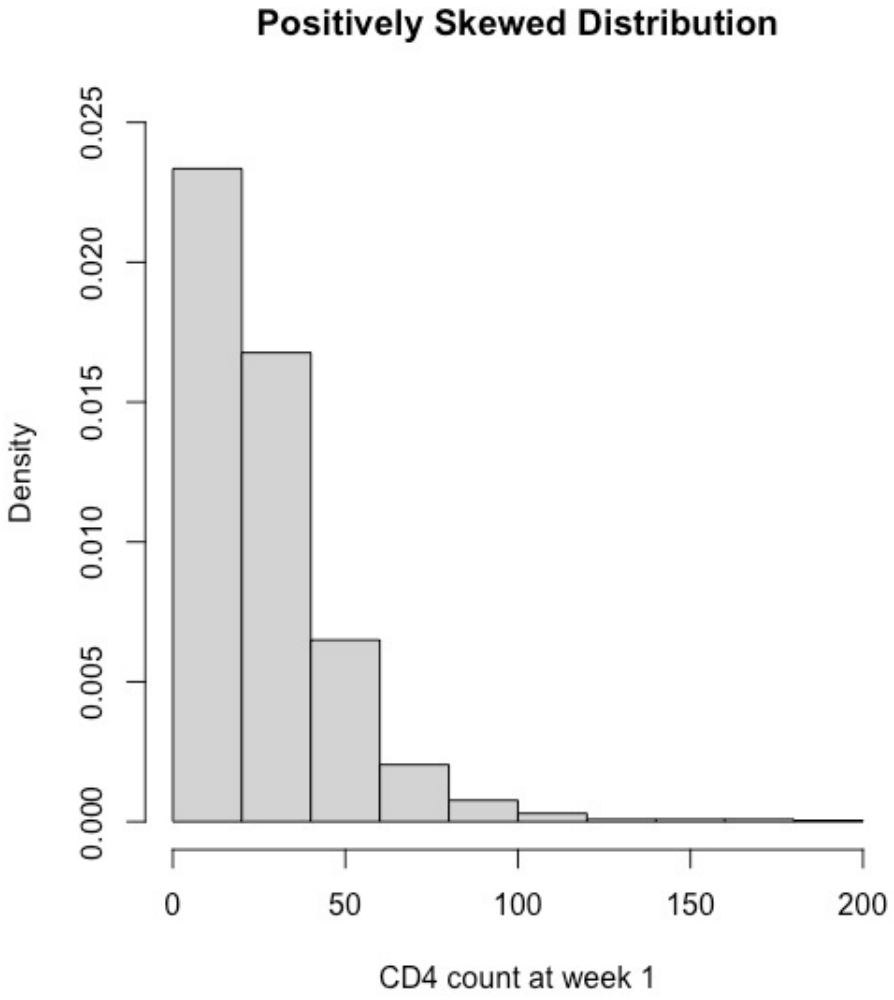
- The population(s) are **normally distributed**
- The selected sample is **representative of general population**
- The data is **continuous**

Assessing Normality

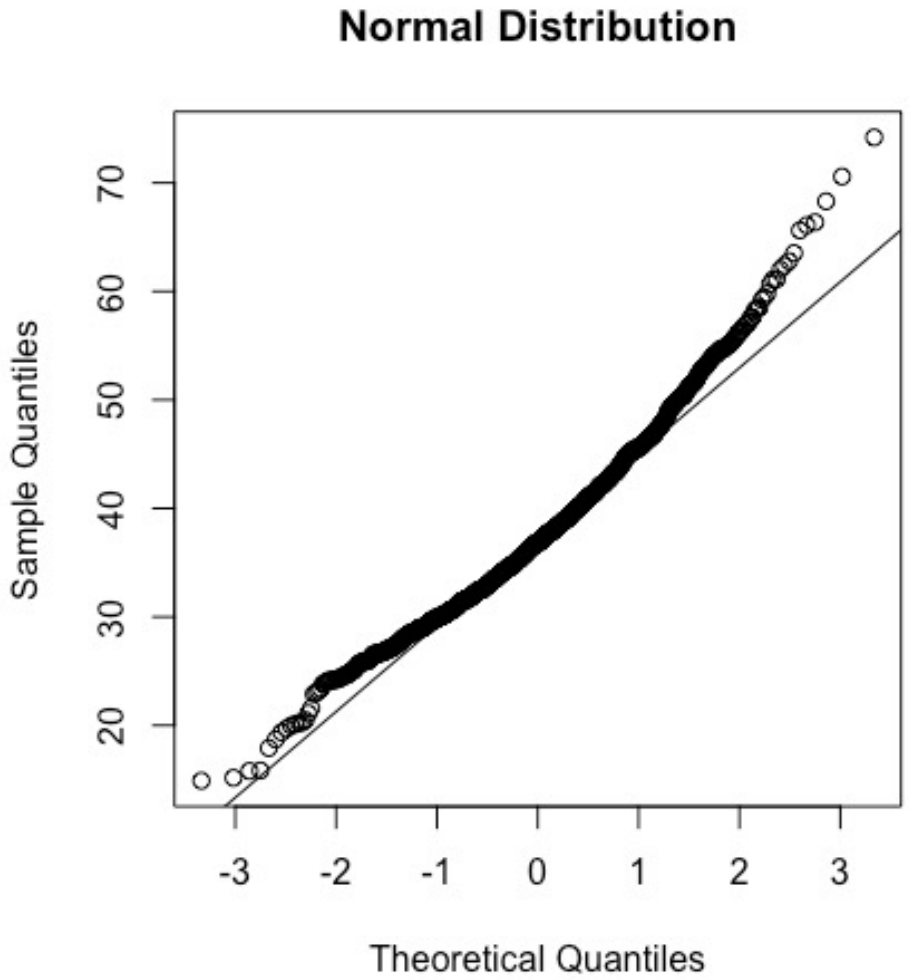
- Inspecting the **histogram** of the variable
- **Quantile-quantile plots**
- **Shapiro-Wilk test**
  - $p > 0.05$ indicates normal distribution
- …

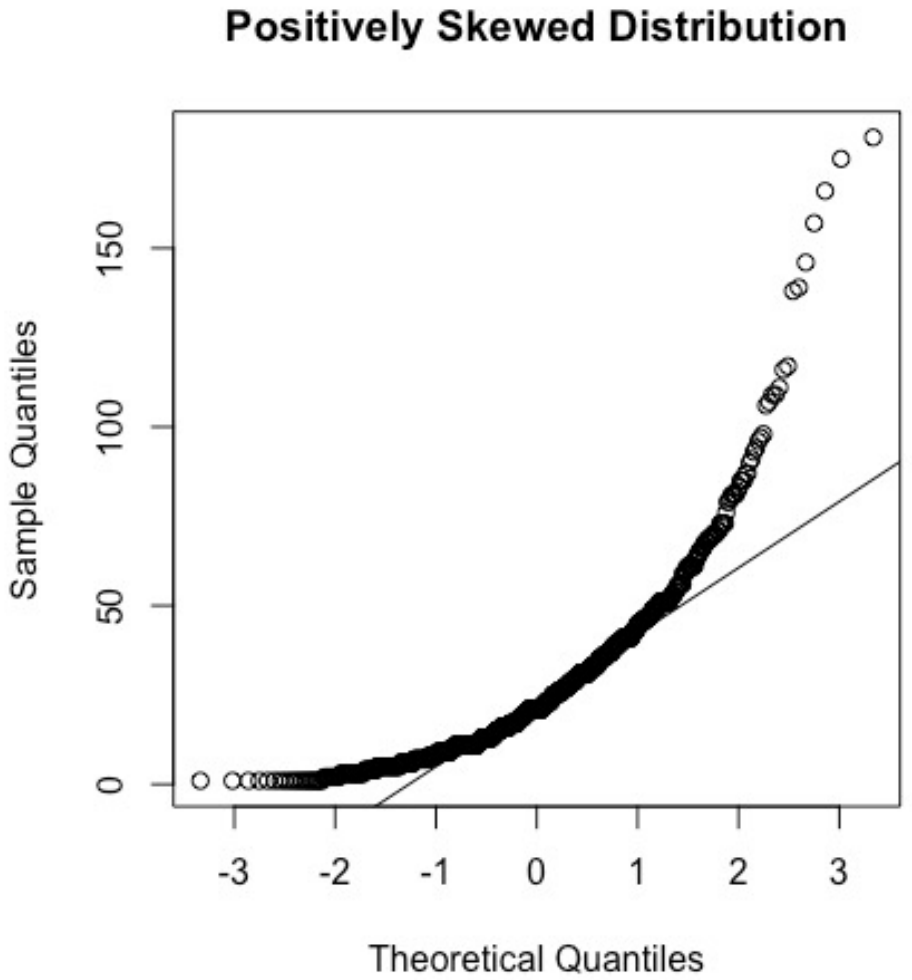
Inspecting Histogram





Quantile-Quantile Plots





Non-parametric Tests

- Used when assumptions of parametric tests are not met
- **Not dependent on the distribution**
- **Less assumptions**
  - e.g., they do not depend on the assumption of normality

- **Less statistical power** compared to parametric tests
  - Higher risk of type II errors (e.g., high probability of accepting there is no difference between the groups where there is a difference)

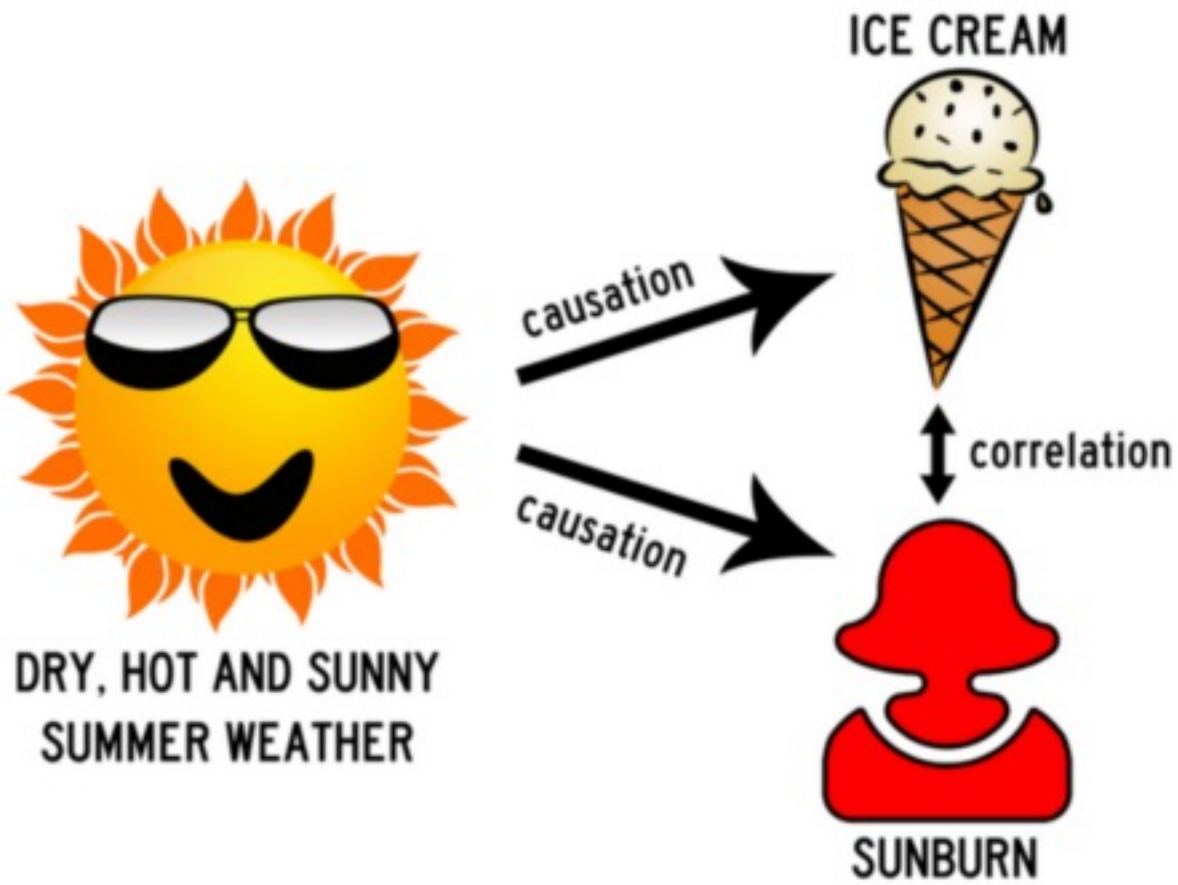
Non-parametric Tests

- **χ^2 test**
- **Wilcoxon rank-sum test (Mann–Whitney U test)** ~ t-test
- **Kruskal-Wallis test** ~ANOVA
- **Spearman's rank correlation test** ~ Pearson correlation test
- …

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**

Correlation does not mean causation



Figueroa A. Correlation is not causation [Internet]. Medium. 2019 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53

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

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

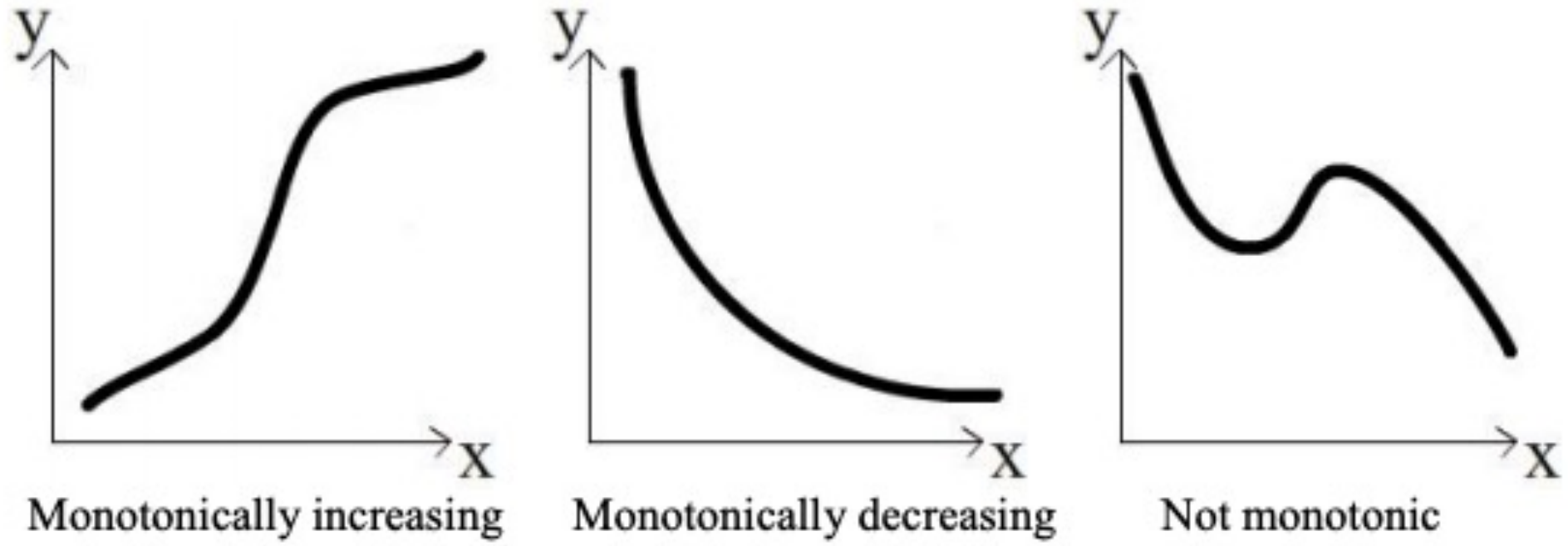
- A measure of the **linear** correlation between two variables X and Y
- takes values between -1 and 1
- unitless
- $r_{X,Y} = r_{Y,X}$
- $r_{X,Y} = 0$ means **no linear relationship**

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

Spearman Rank Correlation

- It assesses how well the relationship between two variables can be described **using a monotonic function**
- It **does not carry any assumptions about the distribution** of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal

Spearman Rank Correlation



Regression Analysis

- Regression can be used to
  - Understand the relationship between variables
  - Predict the value of one variable based on other variables
- Examples:
  - Quantifying the relative impacts of age, gender, and diet on BMI
  - Predicting whether the treatment will be successful or not based on age, tumor stage, tumor volume, …

Linear Regression

E.g., quantifying the relative impacts of age, gender, and diet on BMI

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

where *Y* is the dependent variable, X_1 to X_p are *p* independent variables, β_0 to β_p are the coefficients, and ε is the error term

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffffacc68d53/slides-04.pdfc

Example - Prognostic factors for body fat - Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-60.045	5.365	-11.192	0.000
bmi	0.123	0.236	0.519	0.605
abdomen	0.438	0.105	4.183	0.000
waist_hip_ratio	38.468	10.262	3.749	0.000

$R^2 = 0.681, \boxed{R^2_{\text{adj}} = 0.677}$

the proportion of the variation in the dependent variable that is predictable from the independent variable

$$Estimated\ Body\ Fat = -60.045 + 0.123 * bmi + 0.438 * abdomen + 38.468 * waist\_hip\_ratio$$

Example II

- We'll analyze the prostate cancer dataset
- The main aim of collecting this data set was to inspect the associations between **prostate-specific antigen (PSA)** and **prognostic clinical measurements** in men advanced prostate cancer
- Data were collected on 97 men who were about to undergo radical prostectomies

**PSA was transformed to logPSA for "normalization"*

Example II – Model 1

$$logPSA = 1.8 + 0.07 * \boldsymbol{vol} + 0.77 * I(\boldsymbol{invasion} = \mathbf{1})$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.8035	0.1141	15.81	<0.001
vol	0.0725	0.0133	5.43	<0.001
invasion1	0.7755	0.2541	3.05	0.003

Adjusted R-squared: 0.472

Example II – Model 2

$$logPSA = 1.67 + 0.1021 * \boldsymbol{vol} + 1.326 * I(\boldsymbol{invasion} = \mathbf{1}) - 0.056 * I(\boldsymbol{invasion} = \mathbf{1}) * \boldsymbol{vol}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6673	0.1289	12.94	<0.001
vol	0.1021	0.0191	5.35	<0.001
invasion1	1.326	0.3588	3.7	<0.001
vol:invasion1	-0.056	0.0262	-2.13	0.0354

Adjusted R-squared: 0.491

For a patient with invasion, there is an additional -0.056 change in PSA when vol changes one unit
= For a patient with invasion, one unit change in volume results in (0.1021 – 0.056) change in PSA

Example II – Model 3

$$logPSA = 1.55 + 0.076 * \boldsymbol{vol} + 0.45 * I(\boldsymbol{Gleason = 7}) + 0.9 * I(\boldsymbol{Gleason = 8})$$

(compared to **Gleason = 6**)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5523	0.1548	10.02	< 2e-16
vol	0.0758	0.0131	5.79	9.30E-08
Gleason7	0.4521	0.1928	2.34	0.0212
Gleason8	0.9043	0.2747	3.29	0.0014

Adjusted R-squared: 0.48

Example II – Model 4

$$logPSA = 1.57 + 0.076 * (\boldsymbol{vol - 0.26}) + 0.45 * I(\boldsymbol{Gleason = 7}) + 0.9 * I(\boldsymbol{Gleason = 8})$$

(compared to **Gleason = 6**)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1. 5719	0.1537	10.23	< 2e-16
vol	0.0758	0.0131	5.79	9.30e-08
Gleason7	0.4521	0.1928	2.34	0.0212
Gleason8	0.9043	0.2747	3.29	0.0014

Adjusted R-squared: 0.48

Logistic Regression

- Logistic regression is a specialized form of regression used when the dependent variable is **binary outcome**
  - Having a binary outcome (dependent variable) violates the assumption of linearity in linear regression

- The goal of logistic regression is to find the best fitting model to describe the relationship between the binary outcome and a set of independent variables
  - e.g., predicting whether the treatment will be successful or not, the presence/absence of a disease, etc.

Logistic Regression

- Logistic regression generates the coefficients of the following formula to predict a **logit transformation** of the probability of presence of the outcome:

$$logit(P(Y = 1)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where *P(Y = 1)* indicates the probability that the outcome is 1 (where the binary outcome variable is encoded as 0 and 1)

- *logit* is in fact the log of odds:

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

Logistic Regression – Example

- Identification of risk factors for lymph node metastases with prostate cancer
- n = 52 patients
- y = nodal metastases (0 = none, 1 = metastases)
- x = phosphatase, age , X-ray result, tumor size, tumor grade
  - The first two variables are continuous, the rest are binary

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffff9de0876f/slides-logistic.pdf

Lymph node metastases – **Univariate** Models

	Estimate	Std. Error	z value	Pr(>|z|)	OR
$\log_2$(phosph)	2.4198	0.8778	2.76	0.0058	11.2
Age	-0.0448	0.0468	-0.96	0.3379	1.0
X-ray	2.1466	0.6984	3.07	0.0021	8.6
Size	1.6094	0.6325	2.54	0.0109	5.0
Grade	1.1389	0.5972	1.91	0.0565	3.1

Lymph node metastases – Final Model

	Estimate	Std. Error	z value	Pr(>|z|)	OR
(Intercept)	-0.5418	0.8298	-0.65	0.5138	
$\log_2$(phosph)	2.3645	1.0267	2.30	0.0213	10.6
X-ray	1.9704	0.8207	2.40	0.0163	7.2
Size	1.6175	0.7534	2.15	0.0318	5.0

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffff9de0876f/slides-logistic.pdf

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)	OR
(Intercept)	-0.5418	0.8298	-0.65	0.5138	
$\log_2$(phosph)	2.3645	1.0267	2.30	0.0213	10.6
X-ray	1.9704	0.8207	2.40	0.0163	7.2
Size	1.6175	0.7534	2.15	0.0318	5.0

- With 95% confidence, it could be said that a patient with $\log_2$(phosphatase) = 0, negative X-ray result, size = 0 was equally-likely in terms of having nodal metastases (p = 0.5138)

- With 95% confidence, it could be said that $\log_2$(phosphatase) and having nodal metastases are associated (p = 0.0213)
  - A one unit increase in $\log_2$(phosphatase) was associated with approximately 963.87% increase in the odds of having nodal metastases
  - (exp(2.3645) – 1) * 100 = 963.87
- …

Poisson Regression

- Linear regression was for continuous outcome, whereas logistic regression for binary outcome
- For **count** outcome, Poisson regression can be used

Poisson Regression - Example

- For 59 epilepsy patients the following data were collected:
  - **treatment:** the **treatment group**, a factor with levels placebo and Progabide
  - **base**: the **number of seizures** collected during 8-week period **before** the trial started
  - **age**: the **age of the patient**
  - **seizure rate:** the **number of seizures** occurred during the 2-week period **after** the trial was started

Poisson Regression – Example (cont.)

- A Poisson regression with treatment group, previous seizures and age are related to the mean number of of seizure for patient i, λ_i, is given by:

$$log(\lambda_i) = \beta_0 + \beta_1 * I(treatment = Progabide) + \beta_2 * (base - 6) + \beta_3(age - 18)$$

Poisson Regression – Example (cont.)

$$log(\lambda_i) = \beta_0 + \beta_1 * I(treatment = Progabide) + \beta_2 * (base - 6) + \beta_3(age - 18)$$

	Estimate	Std. Error	z value	p
(Intercept)	0.75	0.14	5.33	<0.001
treament = Progabide	-0.12	0.09	-1.28	0.20
base	0.03	0.00	26.37	<0.001
age	0.05	0.01	5.95	<0.001

Poisson Regression – Example (cont.)

	Estimate	Std. Error	z value	p
(Intercept)	0.75	0.14	5.33	<0.001
treament = Progabide	-0.12	0.09	-1.28	0.20
base	0.03	0.00	26.37	<0.001
age	0.05	0.01	5.95	<0.001

- A patient in placebo group, with 6 previous seizures, and aged 18 had approximately 2 seizures on average in the first two weeks after the trial was started
  - exp(0.75)
- With 95% confidence, it could be said that there was no difference between placebo and progabide (p-value = 0.199)
  - Negative estimate for β_1 indicates lowered mean number of seizures for progabide, but the difference from placebo was not significant

Poisson Regression – Example (cont.)

	Estimate	Std. Error	z value	p
(Intercept)	0.75	0.14	5.33	<0.001
treament = Progabide	-0.12	0.09	-1.28	0.20
base	0.03	0.00	26.37	<0.001
age	0.05	0.01	5.95	<0.001

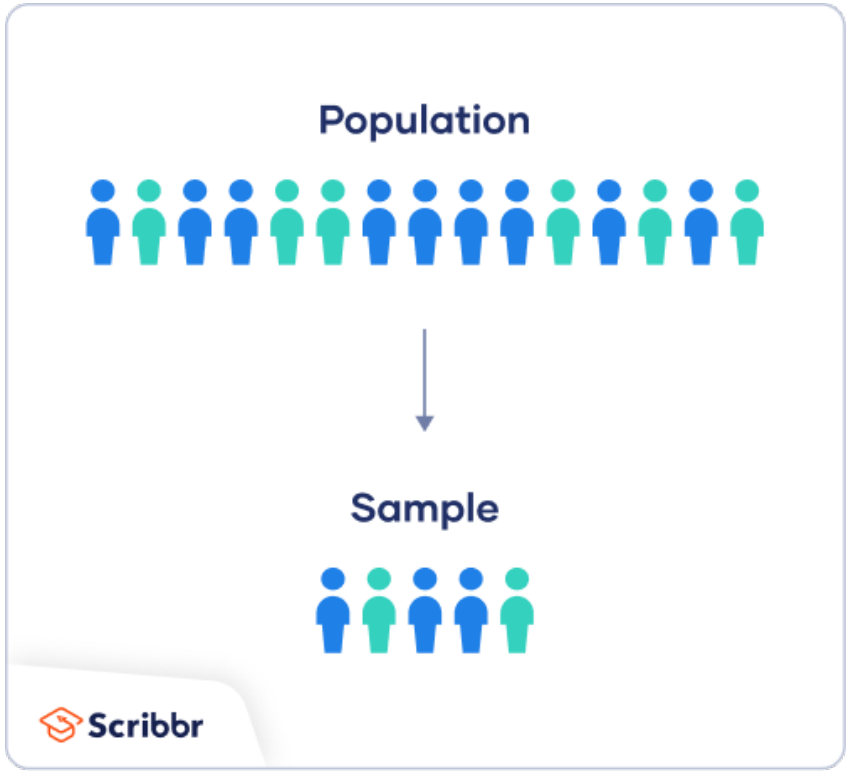
- With 95% confidence, it could be said that previous number of seizures occurred in the 8-week interval prior to the study start and mean seizure rate was significantly associated (p-value < 0.001)
- One unit increase in previous seizure is associated with approximately 2.6% increase in the mean number of seizures in the first two weeks of the trial
  - (exp(0.03) - 1) * 100

Poisson Regression – Example (cont.)

	Estimate	Std. Error	z value	p
(Intercept)	0.75	0.14	5.33	<0.001
treament = Progabide	-0.12	0.09	-1.28	0.20
base	0.03	0.00	26.37	<0.001
age	0.05	0.01	5.95	<0.001

- With 95% confidence, it could be said that age sand mean seizure rate was significantly associated (p-value < 0.001)
- One unit increase in age is associated with approximately 4.8% increase in the mean number of seizures in the first two weeks of the trial
  - (exp(0.05) - 1) * 100

Sampling Methods



- There are 2 main types of sampling methods:
  - Probability sampling
  - Non-probability sampling

Probability Sampling

- Probability sampling involves random selection of elements in which each element has a chance of being selected.

- Four main techniques used for a probability sample:
  - Simple random
  - Systematic
  - Stratified random
  - Cluster

Non-Probability Sampling

- **Non-probability sampling** involves non-random methods in the selection of elements in which not all have equal chances of being selected

- Four main techniques used for a non-probability sample:
  - Convenience
  - Purposive
  - Snowball
  - Quota

Missing Data

- Missing data, or missing values, occur when no data value is stored for the variable in an observation

- **complications** in handling and analyzing the data
- **bias** resulting from differences between missing and complete data

Brief Summary

- There are 2 main sampling methods:
  - Probability sampling
  - Non-probability sampling
- There are 3 kinds of missing data:
  - MCAR: nothing systematic
  - MAR: missingness associated with a variable
  - MNAR: missingness related with the outcome

Statistical Power

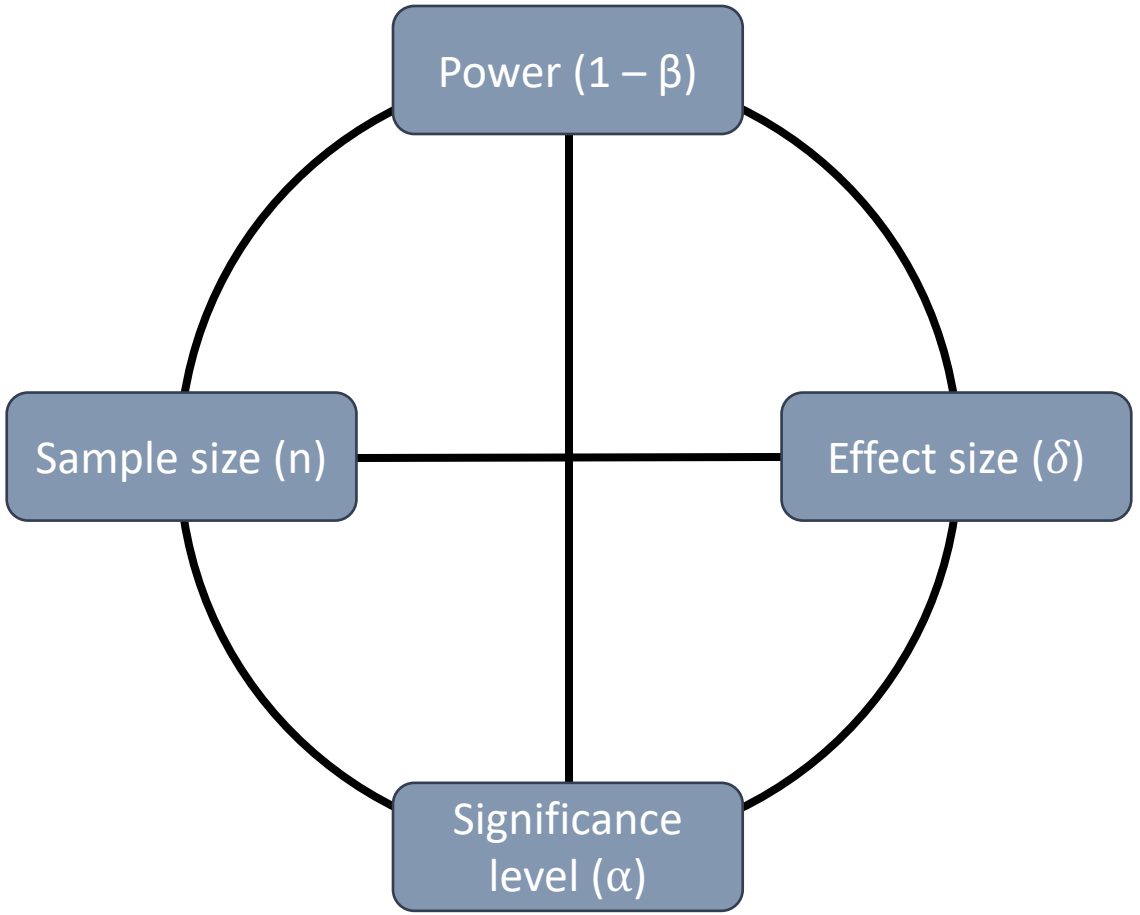
H_0	Decision	
	Fail to reject	Reject
True	Correct decision	**Type I Error** α
False	**Type II Error** **ß**	Correct decision

- **Statistical power** = $1 - \beta$
  - P(reject H_0| H_0 is false)

Statistical Power

- Power is affected by:
  - Significance level (α)
  - Effect size (δ)
  - Sample size (n)

Power Analysis/Sample Size Calculation



- Given any three, the fourth can be determined

Default Values

- Power = usually **0.80**, 0.90
- Significance level = usually **0.05**, 0.01, 0.001
- Effect size
  - Literature review
  - Pilot study
  - Cohen’s recommendations