Hypothesis Test, Univariate, and Bivariate Analysis with R

Humayun Kabir, BScN, MPH, MSc (Student)
MSc in Health Research Methodology

Department of Health Research Methods, Evidence, and Impact, McMaster University, Canada

DASH: Data Analysis Support Hub Workshop Series
Date: February 13, 2024

Lewis & Ruth Sherman Centre for Digital Scholarship
scds.ca

McMaster University | Library

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Laslovarga, “Webster Falls in Winter, Waterdown, Hamilton, Ontario, Canada - Spencer Gorge / Webster's Falls Conservation Area,” 23 January 2011, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Waterdawn_Webster_Falls_in_Winter8.jpg

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information: scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Attendance that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

Verify your participation at a session: https://u.mcmaster.ca/verification

At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

DASH: Data Analysis Support Hub Workshops

Register for upcoming DASH events:

February 20: Machine Learning with R: Linear Regression– Humayun Kabir

February 24: Introduction to Python – Vivek Jadon

February 27: Multivariable Analysis with R – Humayun Kabir

March 22: Machine Learning with R: Logistic Regression– Humayun Kabir

March 28: Intermediate Python Programming – Seyed Amirreza Mousavi

April 30: Survival Analysis with R – Humayun Kabir

Book an Appointment with the DASH Team

Receive help from a member of the DASH team! DASH can assist with the following topics:

- Creating data visualizations, including charts, graphs, and scatter plots
- Figuring out which statistical tests to run (e.g., t-test, chi-square, etc.).
- Analyzing data with software including SPSS, Python, R, SAS, ArcGIS, MATLAB, and Excel
- Choosing which software package to use, including free and open-source software
- Troubleshooting problems related to file formats, data retrieval, and download
- Selecting methodology and type of data analysis to use in a thesis project

Book an appointment: https://library.mcmaster.ca/services/dash

Session Recording and Privacy

- This session is being recorded with the intention of being shared publicly via the web for future audiences. In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

- Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Hypothesis Test, Univariate, and Bivariate Analysis with R

- We have two parts of the session; theoretical and analysis part.
- Very basic statistics and analysis with R.

Objectives

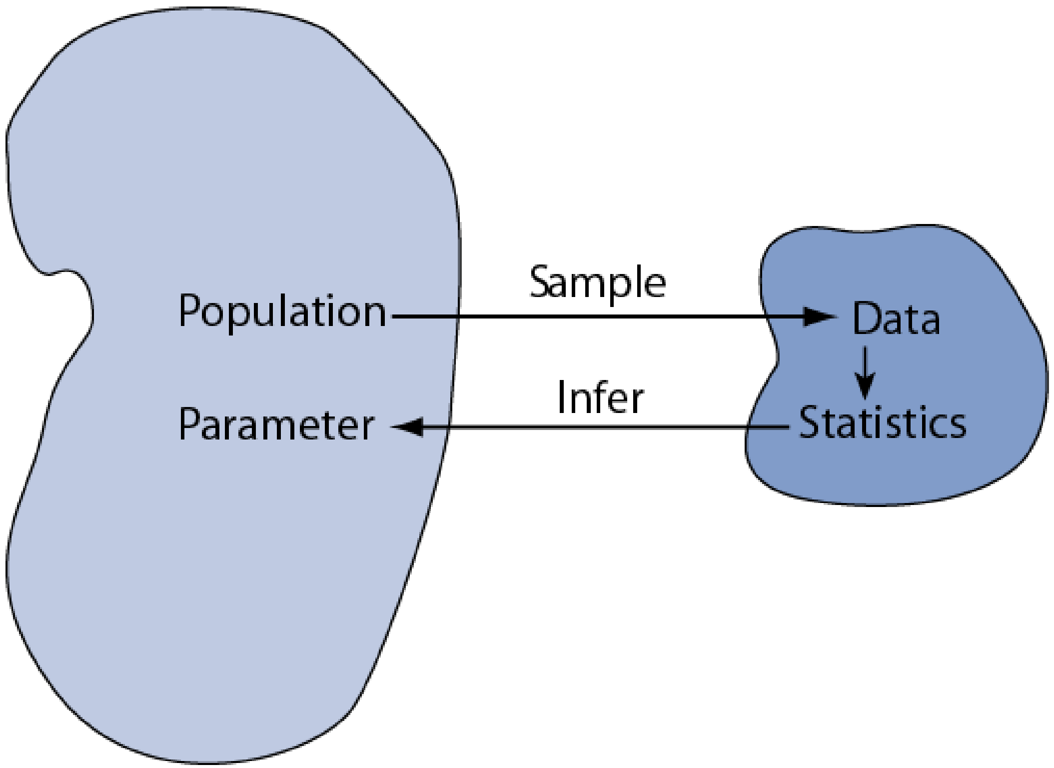
- Hypothesis test and how to do in R
- Univariate analysis and how to do in R
- Bivariate analysis and how to do in R

Parameters and Statistics

	Parameters	Statistics
Source	Population	Sample
Notation	Greek (e.g., μ)	Roman (e.g., *xbar*)
Vary	No	Yes
Calculated	No	Yes

Statistical inference

- **Statistical inference:** generalizing from a sample to a population with calculated degree of certainty
- Two forms of statistical inference
  - Hypothesis testing
  - Estimation
    - Point estimation
    - Interval estimation (95% CI)



Hypothesis testing

General Concept hypothesis testing

- A hypothesis is a claim or statement about a property of a population
- A hypothesis test is a standard procedure for testing a claim about a property of a population using sample data
- We would like to compare the observed statistic with a preconceived parameter value and conclude whether data are consistent with preconceived idea
- Generally, we evaluate the role of chance in getting the observed sample statistic very different from the claim

Motivation for hypothesis testing

- Hypothesis testing provides an objective framework for making decisions using probabilistic methods, rather than relying on subjective impressions
- People can form different opinions by looking at data, but a hypothesis test provides a uniform decision-making criterion that is consistent for all people

Null Hypothesis: H_0

- The null hypothesis (denoted by H_0) is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is equal to sample value

Alternative Hypothesis: H_1

- The alternative hypothesis (denoted by H_1 or H_a or H_A) is the statement that the parameter (such as proportion, mean, or standard deviation) somehow differs from the sample value

Specification of hypothesis

- H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$ or $\mu < \mu_0$ or $\mu > \mu_0$
- We test the null hypothesis directly
- We compute the difference between the observed value and the expected value under null hypothesis and examine whether the difference is too large or too small
- Either reject H_0 or accept H_0 based on the size of the difference
- If the difference is too small, we tend to accept null hypothesis
- If the difference is too large, we tend to reject null hypothesis

Four possible outcomes in hypothesis testing

Decision based on test		**Truth**	
		H_0	H_1
	Accept H_0	H_0 is true H_0 is not rejected	H_1 is true H_0 is not rejected
	Reject H_0	H_0 is true H_0 is rejected	H_1 is true H_0 is rejected

- You can notice that incorrect “rejection” or “failure to rejection” of H_0 is possible in hypothesis testing
- These are the errors associated with hypothesis testing
- Two types of errors can occur- type 1 error and type 2 error

General example of hypothesis testing

- Consider mean daily expenditure of all university students μ_0 is claimed to be Tk. 350.00 and population standard deviation σ is Tk. 75.00
- We would like to test the claim using hypothesis testing

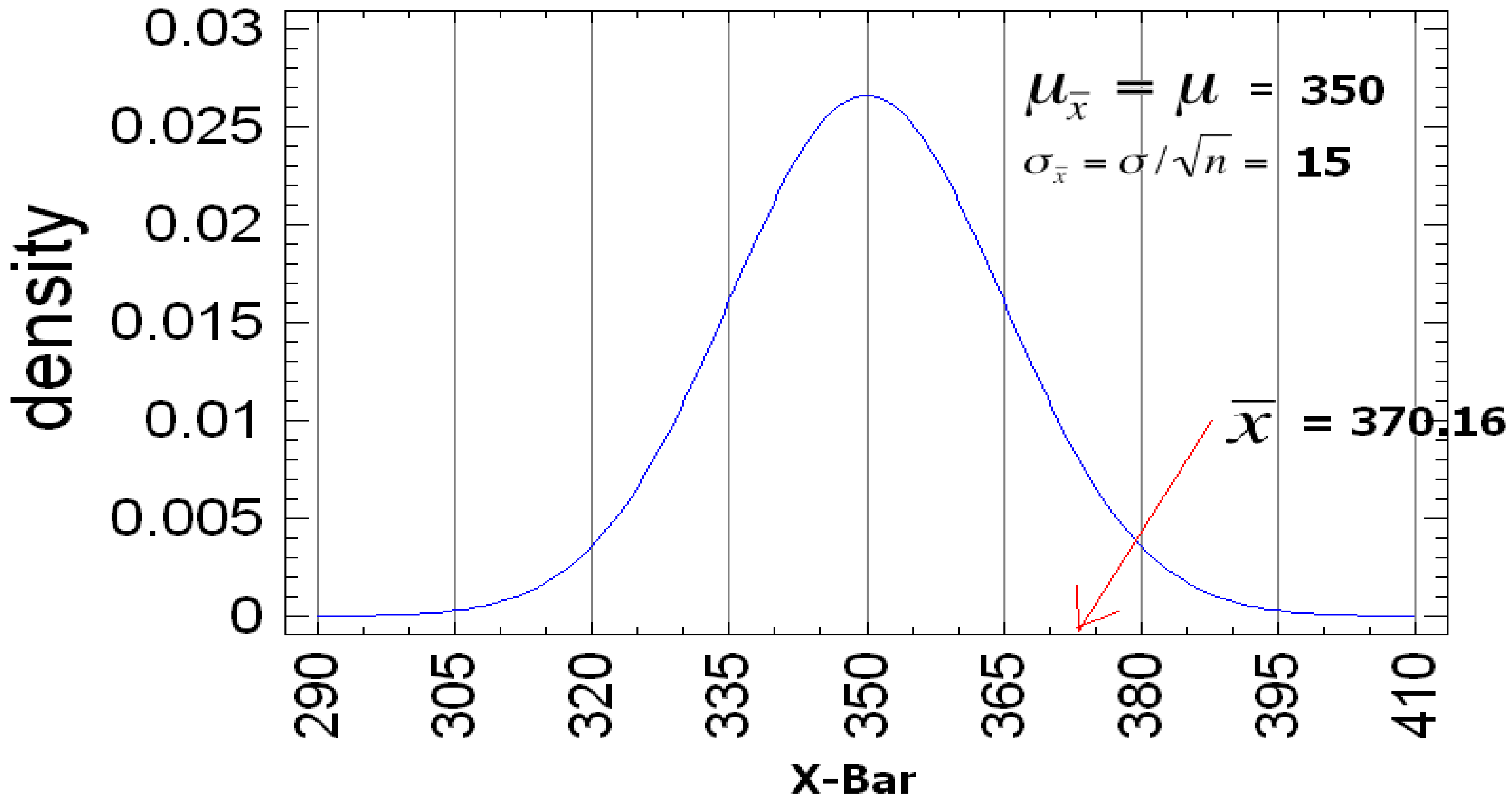
  H_0: $\mu = 350$
- Thus, our alternate hypothesis is:

  H_1: $\mu \neq 350$

Concepts of Hypothesis Testing

- We randomly select 25 students and estimate sample mean $\bar{x}$=370.16
- If $\bar{x}$ is close to the claimed population mean of 350 that the difference is small, we are more convinced to believe the null hypothesis
- If $\bar{x}$ is much larger than 350 (say 600) or much less than 350 (say 150) that the difference is large, we are more convinced to reject the null hypothesis
- The question is how large is large and how less is less
- If that is allowed to be decided by individual perception, people will decide very differently on the same sample mean
- Hypothesis testing provides a standardized procedure without any subjective bias

Sampling distribution of $\bar{x}$ given null hypothesis is true



Set the significance level

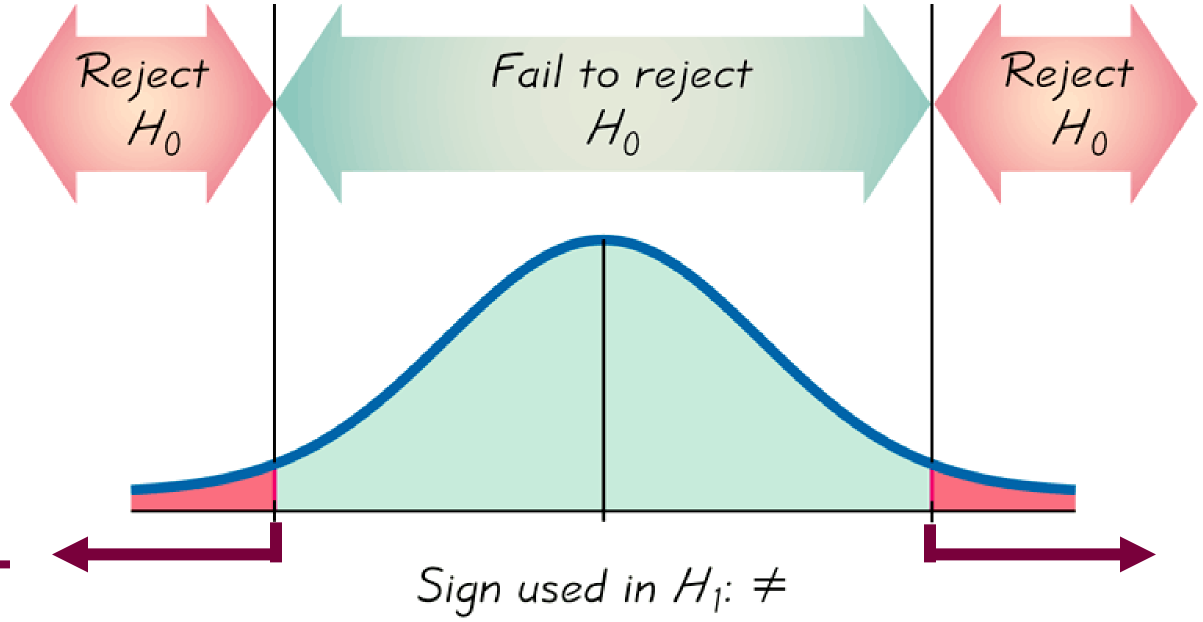
- Significance level is the α that refer to what proportion of sample means that are too extremes in the sampling distribution will be considered for rejection of null hypothesis
- Extreme values are located towards the tails of the distribution
- Usually, the significance level is 5%
- Two–tailed hypothesis: μ can be both greater or less than μ_0
- One-tailed hypothesis: μ can only be greater or less than μ_0

Significance level in two-tailed Test

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

α is divided equally between the two tails of the critical region

Means less than or greater than

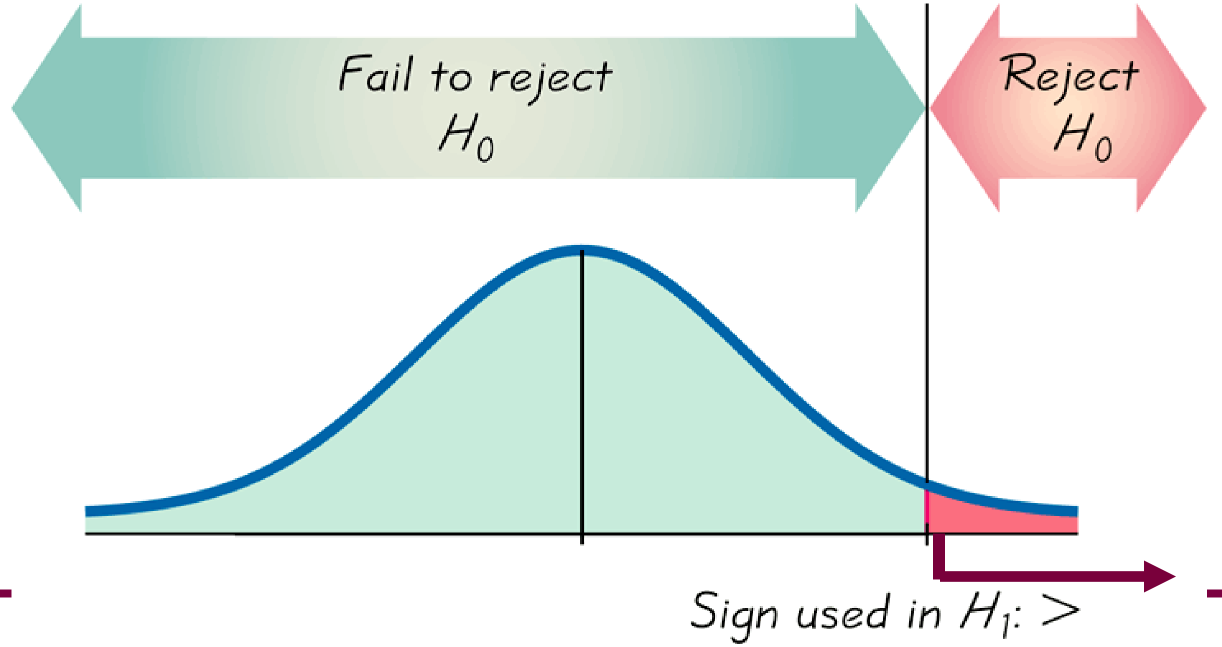


Significance level in right sided one-tailed test

$$H_0\colon\ \mu \leq \mu_0$$

$$H_1\colon\ \mu > \mu_0$$

Points Right

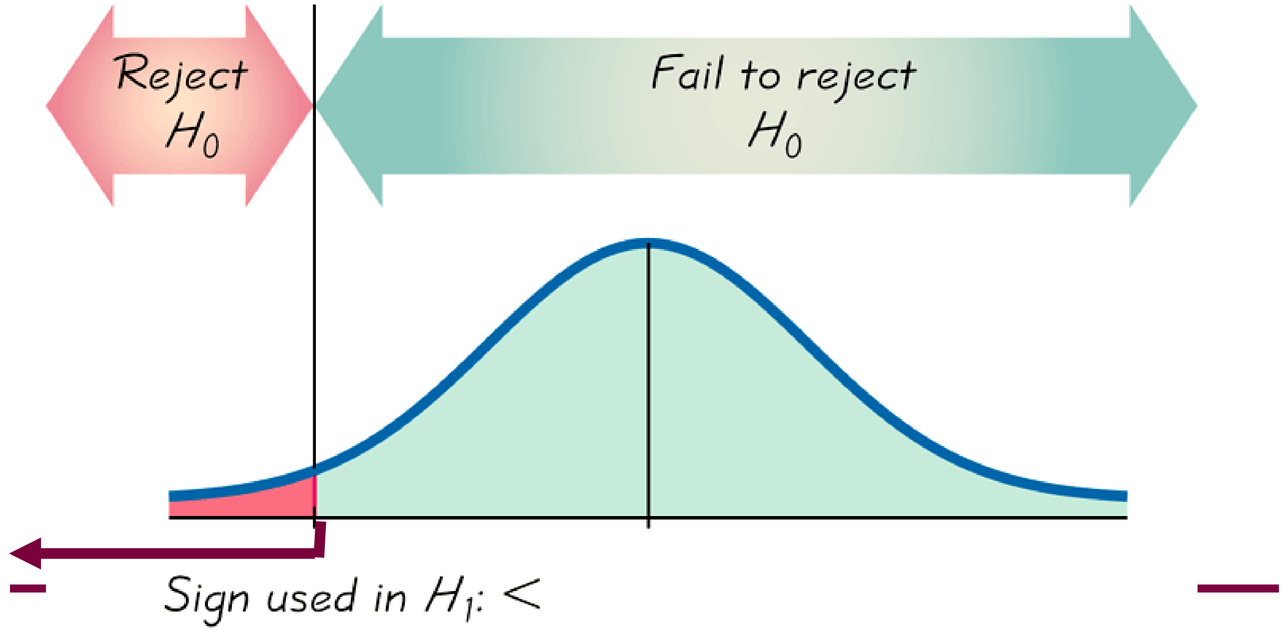


Significance level in left sided one-tailed test

$$H_0\colon\ \mu \geq \mu_0$$

$$H_1\colon\ \mu < \mu_0$$

Points Left



Three approaches for hypothesis testing

- Unstandardized critical value method
- Standardized critical value method (Also known as test statistic method)
- P-value method
- All are basically same

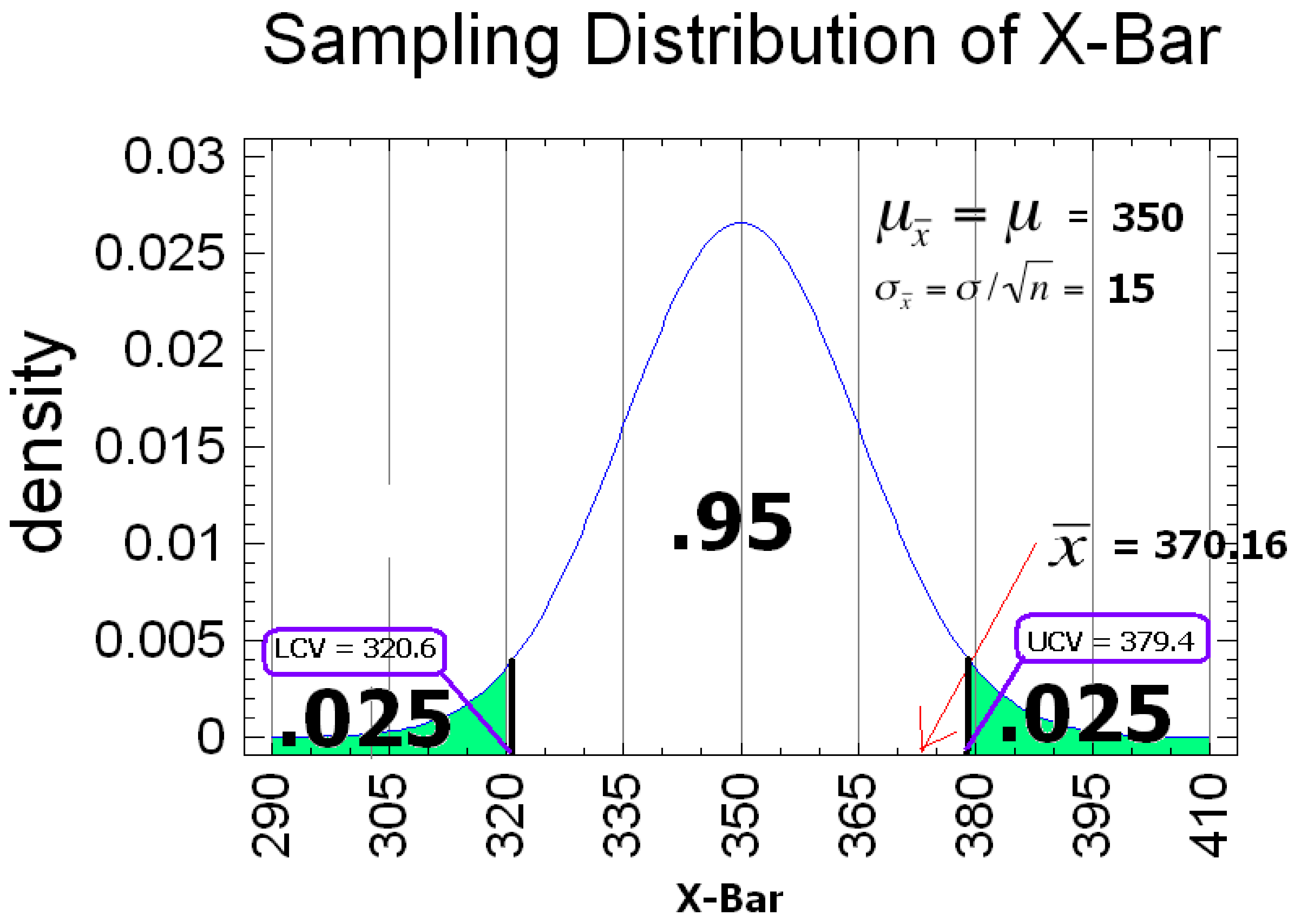
Unstandardized critical value method

- Identify the critical (cut point) value(s) that demarcates the extreme values from the non-extreme values
- If we define the non-extreme values as the middle 95% of the distribution [this means $\alpha = 0.05$], then the critical values that demarcate the non-extreme values will be 1.96 standard deviations of X-Bar on either side of the mean of the sampling distribution [350], or

  UCV = 350 + 1.96*15 = 350 + 29.4 = 379.4

  LCV = 350 – 1.96*15 = 350 – 29.4 = 320.6

Unstandardized sampling distribution approach

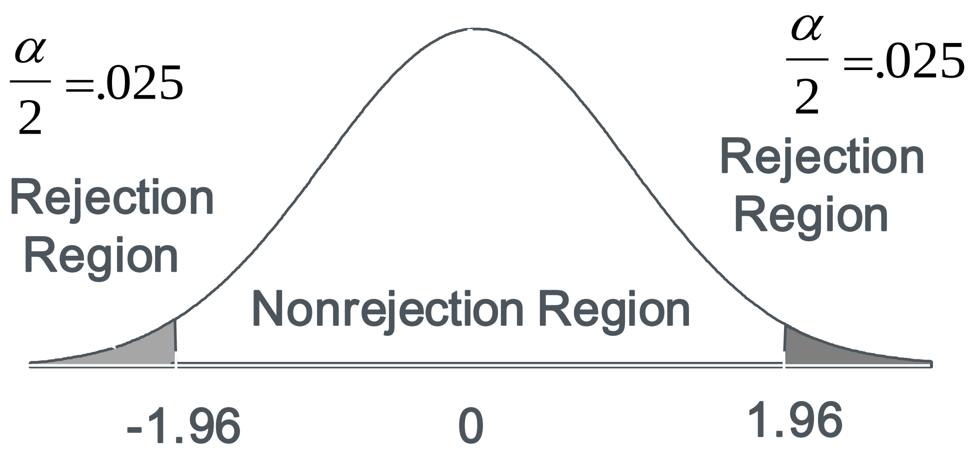


Standardized critical value (test statistic) method

- Compute the Z-statistic for the observed sample mean.
- If it is greater than 1.96 or less than -1.96, we know that will be in the rejection region on the either side respectively.

Z score=$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$=1.344

Rejection region in a two-tailed test



P value method

- The *p-value* approach (which is generally computed with a computer and statistical software or distribution table) and compared with the preset significance level
- P value is the probability of obtaining a test statistic as extreme or more extreme than the actual test statistic obtained, given that the null hypothesis is true
- For this example, since the sample mean is to the right side of the mean, calculate

  $P(\bar{x} \geq 370.16) = P(Z \geq 1.344) = 0.0901$ (computed by excel)

  Since this is a two tailed test, we must double this area for the p-value

  p-value = 2*(0.0901) = 0.1802
- Since we defined the significance level at 0.05 and p value is 0.1802, we cannot reject the null hypothesis

Statistical conclusions in our example

Unstandardized sampling distribution:

Since LCV (320.6) < $\bar{x}$ (370.16) < UCV (379.4), we fail to reject the null hypothesis at a 5% level of significance.

Standardized Test Statistic:

Since $-Z_{\alpha/2}(-1.96) < Z(1.344) < Z_{\alpha/2}$ (1.96), we fail to reject the null hypothesis at a 5% level of significance.

P-value:

Since p-value (0.1802) > 0.05 [α], we fail to reject the null hypothesis at a 5% level of significance.

Univariate Analysis

Univariate Analysis

"Univariate is a term commonly used in statistics to describe a type of data which consists of observations on only a single characteristic or attribute." Wikipedia

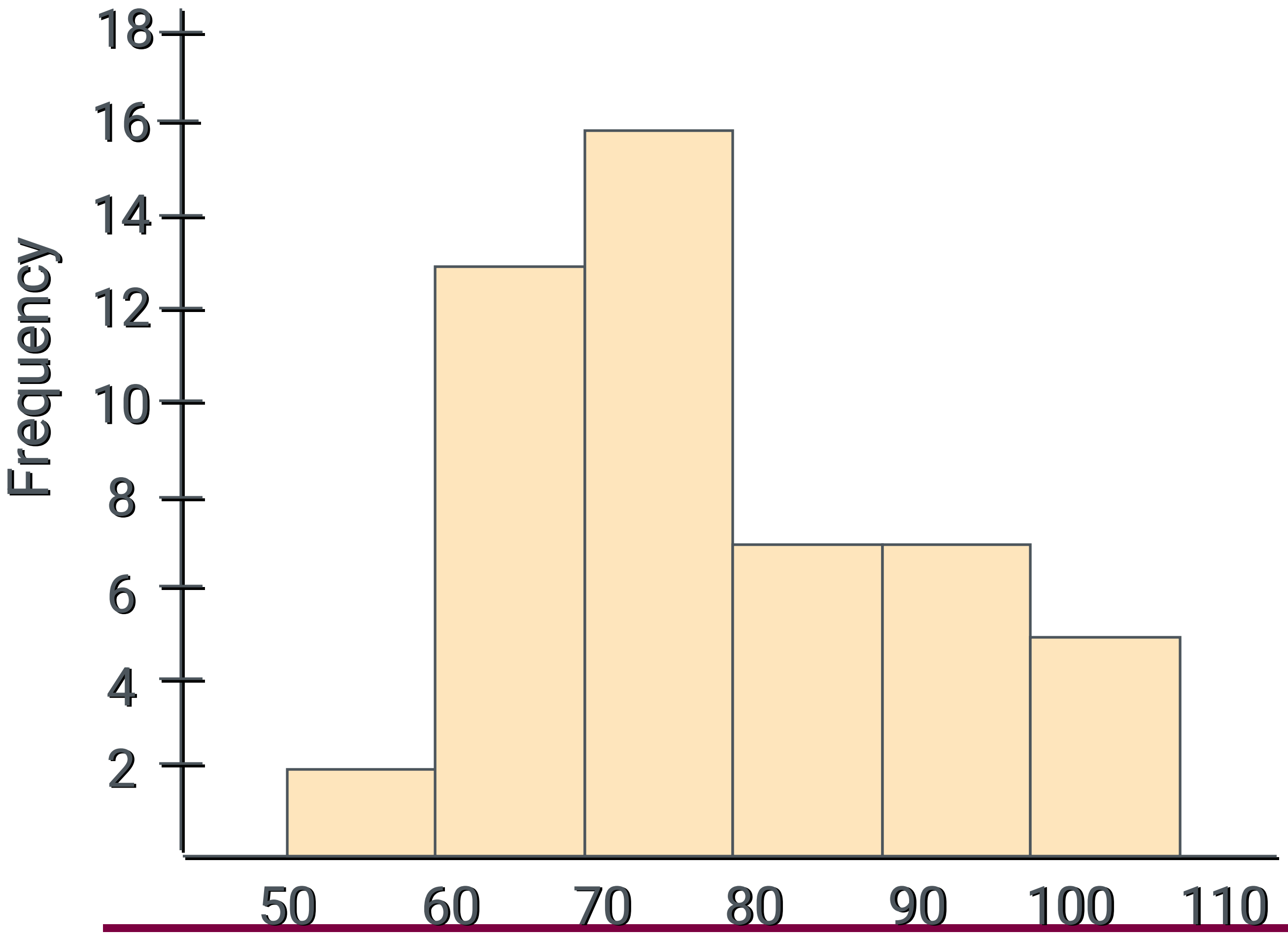
Descriptive Statistics:

- Measures of Central Tendency: Mean, median, and mode.
- Measures of Dispersion: Range, variance, and standard deviation.
- Position: Quartile, and interquartile range.
- Frequency Distribution: How often each value or range of values occurs.

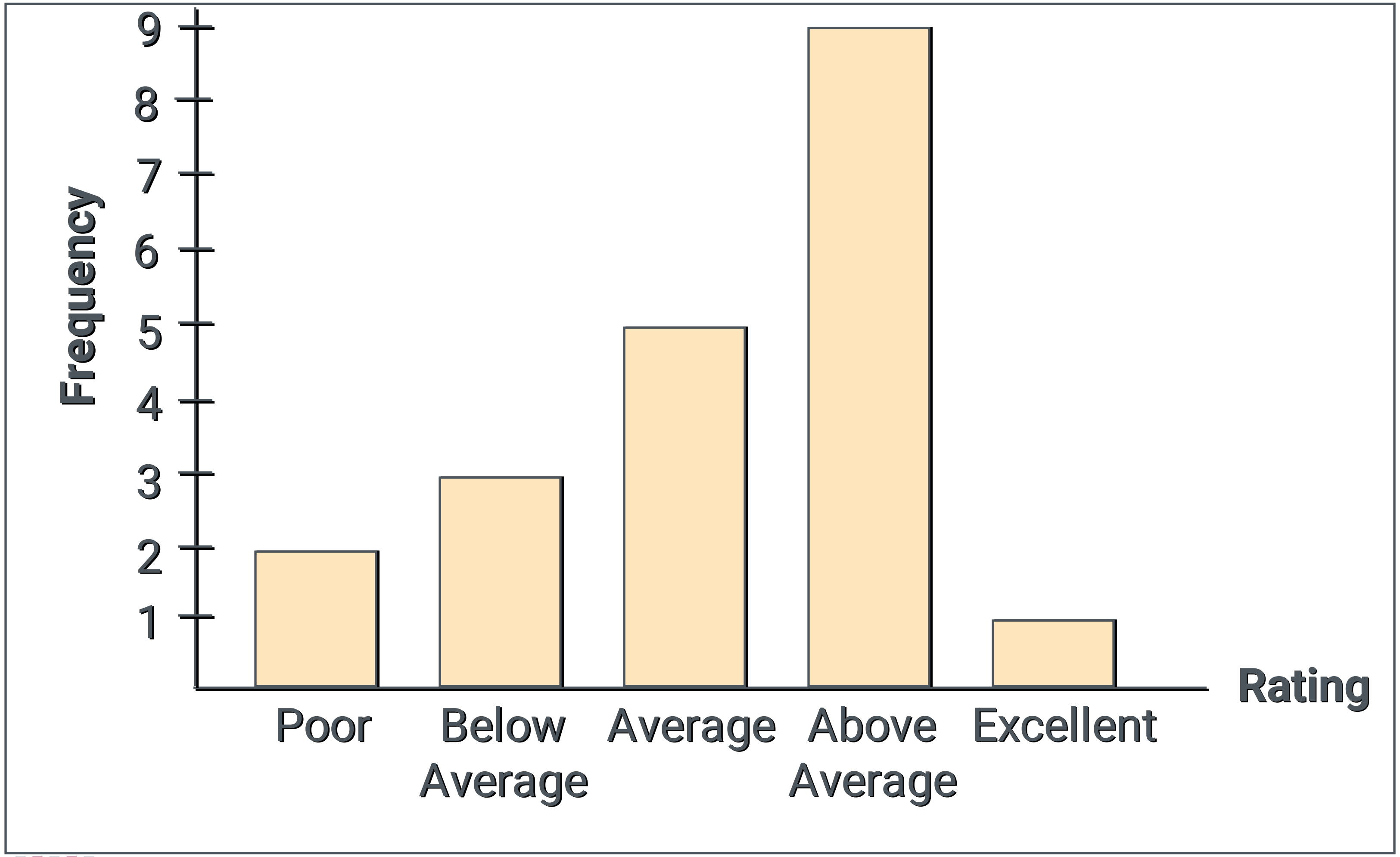
Graphical Representations:

- Histograms, box plots, bar chart, pie charts, etc.

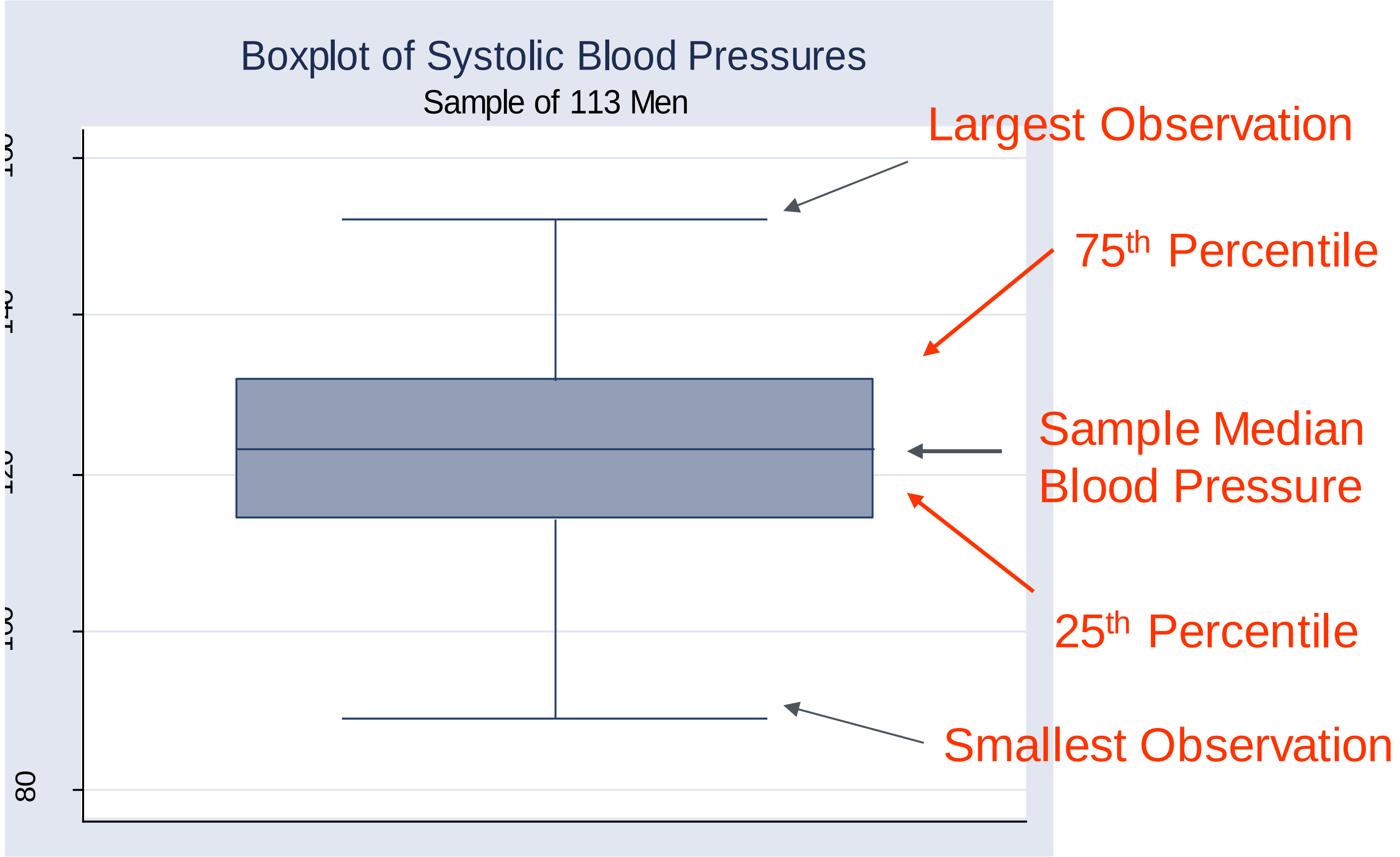
Histogram



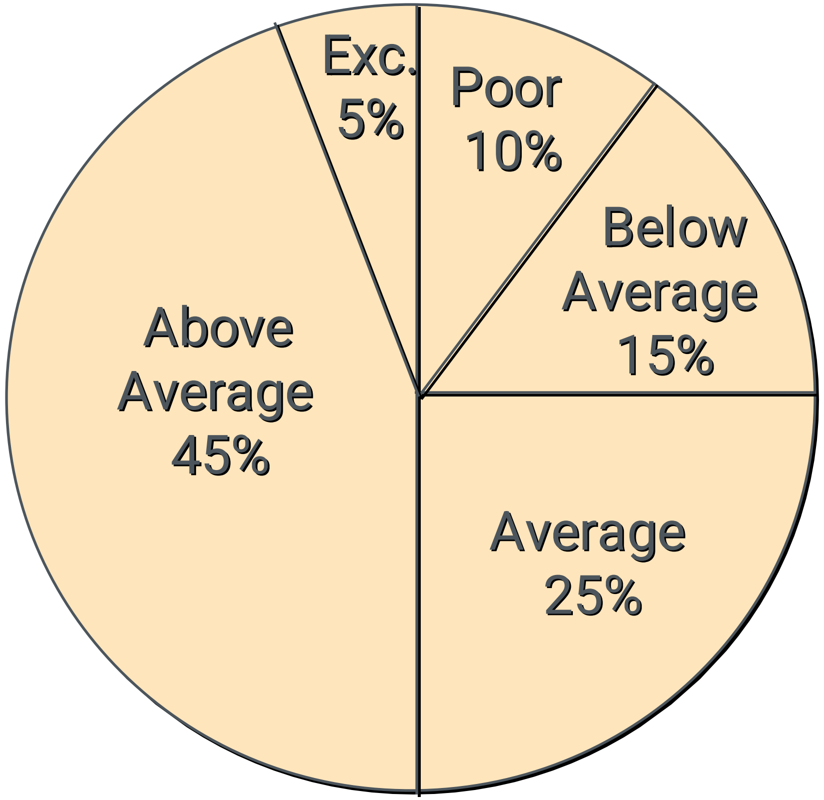
Bar Graph: Hotel rating



Boxplot: BP for 113 Males



Pie Chart: Hotel rating



Quality Ratings

Bivariate Analysis

Bivariate Analysis

Bivariate analysis is the analysis of two variables (often denoted as X, Y), for the purpose of determining the relationship.- Wikipedia

Purpose of bivariate analysis

Testing of association.

Predicting one variable (possibly a dependent variable) by another variable (possibly the independent variable).

Contrast with Univariate Analysis

Bivariate analysis involves the analysis of two variables, while univariate analysis focuses on only one variable.

Bivariate Analysis

Measures of Central Tendency, Dispersion, Position, Frequency Distribution of one variable (predictor) by another variable (outcome).

Histograms, box plots, and bar charts of one variable (predictor) by another variable (outcome).

Two sample t-test, Anova test, Simple regression analysis, etc.

Acknowledgement

I included some slides from Dr. Dipak Kumar Mitra, Professor, North South University, Bangladesh.

Contact

Book an appointment with DASH: https://library.mcmaster.ca/services/dash

Contact DASH: Data Analysis Support Hub: libdash@mcmaster.ca

Let move to the coding part

https://colab.research.google.com/drive/1R22VaodM2ehWrMMekXiObg4pxMJKcS3k#scrollTo=2poHgvqImUn3