Introduction to Scientific Computing for Biologists

ISCB20.09 – Interpreting Data Using Descriptive Statistics with R

Md. Jubayer Hossain
https://jhossain.me/
jubayer@hdrobd.org

Founder
Health Data Research Organization
Lead Organizer
Scientific Computing for Biologists

February 2, 2021

Statistics and Biostatistics: Definitions–1

Definition–1: What is Statistics?
Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied. (Source: https://en.wikipedia.org/wiki/Statistics)

Definition–2: What is Statistics?
Simply, Statistics is a branch of mathematics that deals with collecting, organizing, analyzing, and interpreting data. (Source: https://app.pluralsight.com/library/courses/interpreting-data-descriptive-statistics-python/)

Statistics and Biostatistics: Definitions–2

Definition–3: What is Biostatistics?
Biostatistics is the application of statistics to a variety of topics in biology. In this course, we tend to focus on biological topics in the health sciences as we learn about statistics. (Source: `https://bolt.mph.ufl.edu/6050-6052/`)

Definition–4: What is Biostatistics?
Biostatistics is the application of statistics to problems in the biological sciences, health, and medicine. (Source: `https://ocw.jhsph.edu/index.cfm/go/viewCourse/course/MethodsInBiostatisticsI/`)

What is Data?

- **Definition–1:** Data is a collection of facts, such as numbers, images, words, measurements, observations, audios, videos or just descriptions of things.
- **Definition–2:** Data is a tool to reach suitable conclusion.

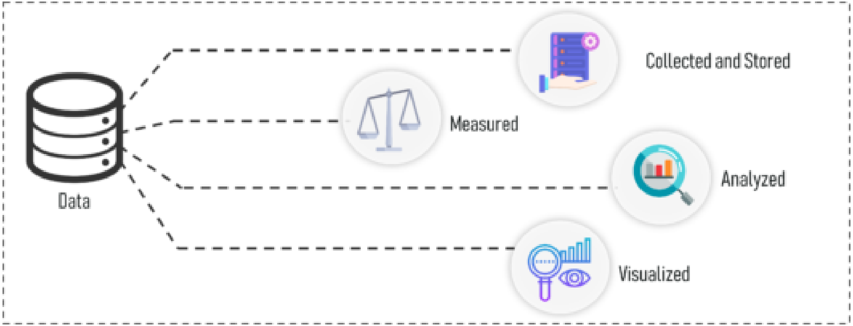
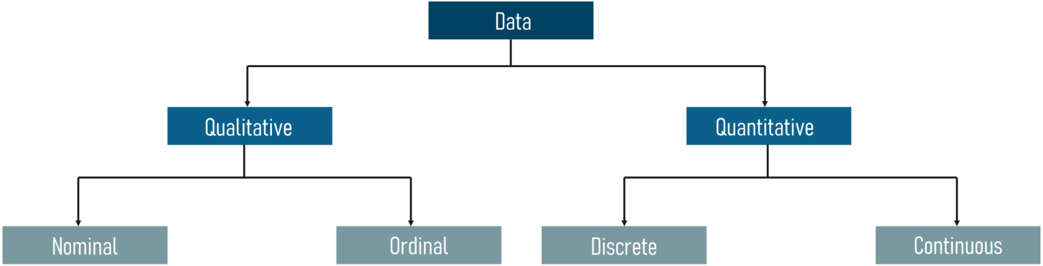


Figure: https://www.edureka.co/blog/statistics-and-probability/

Types of Data



Qualitative or Categorical Data

- Classifies individuals or items into different groups.
- Qualitative data is further divided into two types of data
  - **Ordinal:** groups have an order or ranking.
  - **Nominal:** groups are merely names, no ranking.

Customer ID	Rating
001	Good
002	Average
003	Average
004	Bad

(a) Ordinal Data

Gender
Male
Female
Male
Male

(b) Nominal Data

Figure: https://www.edureka.co/blog/statistics-and-probability/

Quantitative or Numeric Data

- Numerical, measurable quantities in which arithmetic operations often make sense.
- Quantitative data is also further divided into two types of data
  - **Continuous:** could take on any value within an interval,many possible values.
    - A person's height: could be any value (within the range of human heights), not just certain fixed heights.
    - Time in a race: you could even measure it to fractions of a second.
    - Blood pressure, mmHg.
    - Weight, pounds (kilograms, ounces, etc.)
  - **Discrete:** countable value, finite number of values.
    - The number of students in a class.
    - The results of rolling a die.

Binary Data

- Yes/No
- Polio: Yes/No
- Cure: Yes/No
- Sex: Female/Male(0 or 1)

Types of Variables

- **Independent Variable(IV)**
  A variable whose value does not change by the effect of other variables and is used to manipulate the dependent variables. It is often denoted as X.
- **Dependent Variable(DV)** A variable whose value change when there is any manipulation in the values of independent variables. Is is often denoted as Y

X **Causes** Y

X (effect) $\rightarrow$ Year of Experience $\rightarrow$ Independent

Y (cause) $\rightarrow$ Salary $\rightarrow$ Dependent

Other Names for IV and DV

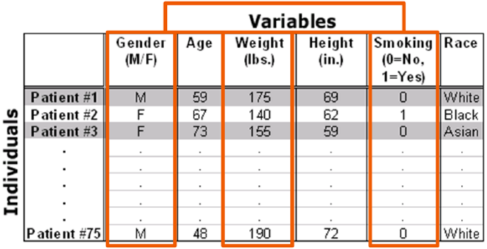
Other Names for Independent Variables

- Explanatory Variables (they explain an event or outcome)
- Predictor Variables (they can be used to predict the value of a dependent variable)

Other Names for Dependent Variables

- Response Variables (they respond to a change in another variable)
- Outcome Variables (they represent the outcome you want to measure)

A Typical Dataset



Variables

Individuals

	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (0=No, 1=Yes)	Race
Patient #1	M	59	175	69	0	White
Patient #2	F	67	140	62	1	Black
Patient #3	F	73	155	59	0	Asian
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Patient #75	M	48	190	72	0	White

- **Variables** contain the information about a particular characteristic for all individuals in a dataset.
- An **observation** in statistics is a value of something of interest you're measuring or counting during a study or experiment: a person's height, a bank account value at a certain point in time, or number of animals.

Terminologies In Statistics: Population and Sample

Population: The population is the entire group that you want to draw conclusions about.
Sample: The sample is the specific group of individuals that you will collect data from.

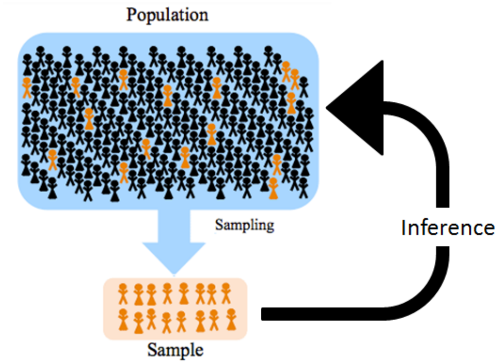


Figure: `https://online.stat.psu.edu/stat500/`

Terminologies In Statistics: Sampling Frame and Sample Size

Sampling Frame
The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

Sample Size
The number of individuals in your sample depends on the size of the population, and on how precisely you want the results to represent the population as a whole.

Sample Size Calculator
Surveymonkey–`https://www.surveymonkey.com/mp/sample-size-calculator/`

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.
- **Acurate representative of the population:** A sample should be an accurate representative of the population from which it is taken.

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.
- **Acurate representative of the population:** A sample should be an accurate representative of the population from which it is taken.
- **Proportional:** A sample should be proportional. It should be large enough to represent the population properly. In general, the larger the sample size, the more accurately and confidently you can make inferences about the whole population.

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.
- **Acurate representative of the population:** A sample should be an accurate representative of the population from which it is taken.
- **Proportional:** A sample should be proportional. It should be large enough to represent the population properly. In general, the larger the sample size, the more accurately and confidently you can make inferences about the whole population.
- **Random Selection:** A sample should be selected at random. This means that any item in the group has a full and equal chance of being selected and included in the sample.This makes the selected sample truly representative in character.

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.
- **Acurate representative of the population:** A sample should be an accurate representative of the population from which it is taken.
- **Proportional:** A sample should be proportional. It should be large enough to represent the population properly. In general, the larger the sample size, the more accurately and confidently you can make inferences about the whole population.
- **Random Selection:** A sample should be selected at random. This means that any item in the group has a full and equal chance of being selected and included in the sample.This makes the selected sample truly representative in character.
- **Economical:** A sample should be economical.The objective of the survey should be achieved with minimum cost and effort.

Characteristics of a Good Sample–2

- **Practical:** A sample should be practical. The sample design should be simple. It should be capable of being understood and followed in the fieldwork.

Characteristics of a Good Sample–2

- **Practical:** A sample should be practical. The sample design should be simple. It should be capable of being understood and followed in the fieldwork.
- **Actual information provider:** A sample should be designed so as to provide actual information required for the study and also provide an adequate basis for the measurement of its own reliability.

Characteristics of a Good Sample–2

- **Practical:** A sample should be practical. The sample design should be simple. It should be capable of being understood and followed in the fieldwork.
- **Actual information provider:** A sample should be designed so as to provide actual information required for the study and also provide an adequate basis for the measurement of its own reliability.

Types of Statistics

There are two kinds of statistics, the kind you look up and the kind you make up
–Rex Stout

- **Descriptive Statistics –** Identify important elements in a dataset.
- **Inferential Statistics –** Explain those elements via relationships with other elements.

Descriptive Statistics

Descriptive statistical methods provide an exploratory assessment of the data from a study.

- Descriptive statistical methods provide a exploratory data analysis.
  - Frequency Distribution Table
  - Graphs / Charts
  - Summary
- Descriptive statistical methods divide into 3 categories.
  - **Univariate analysis** summarize only one variable at a time.
  - **Bivariate analysis** compare two variables.
  - **Multivariate analysis** compare more than two variables.

Inferential Statistics

Assess the strength of evidence for/against a hypothesis; evaluate the data

- Inferential statistical methods provide a confirmatory data analysis
  - Generalize conclusions from data from part of a group (sample) to the whole group (population)
  - Assess the strength of the evidence
  - Make comparisons
  - Make predictions
- Inferential statistical methods divide into 2 categories.
  - **Hypothesis Testing:** Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.
  - **Model Fitting:** Model fitting is a measure of how well a statistical learning model generalizes to similar data to that on which it was trained. A model that is well-fitted produces more accurate outcomes.

Sampling

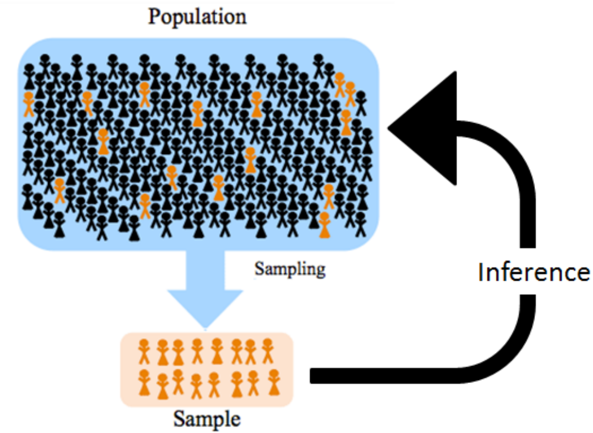


Figure: https://online.stat.psu.edu/stat500/

Types of Sampling Methods

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. There are two types of sampling methods

- Probability Sampling
- Non-Probability Sampling

Probability Sampling Methods

- Probability sampling involves random selection, allowing you to make statistical inferences about the whole group.
- Probability sampling means that every member of the population has a chance of being selected.
- It is mainly used in quantitative research.
- If you want to produce results that are representative of the whole population, you need to use a probability sampling technique.

Non-Probability Sampling Methods

- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect initial data.

Types of Probability Sampling Methods

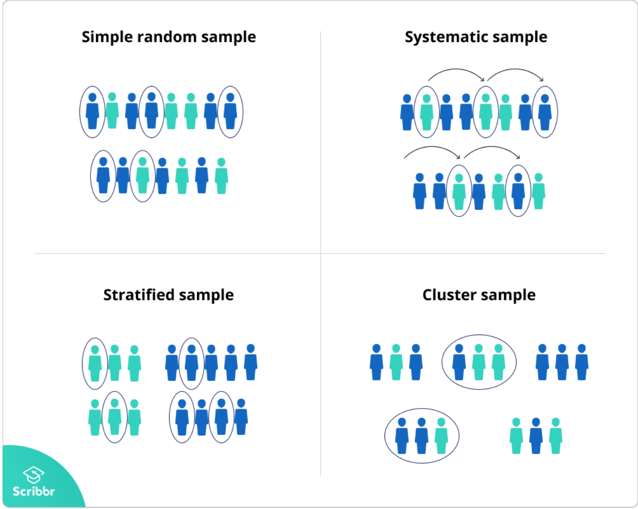


Figure: https://www.scribbr.com/methodology/sampling-methods/

Simple Random Sampling(SRS)

- In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.
- To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Systematic Sampling

- Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct.
- Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Stratified Sampling

- Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.
- To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).
- Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Cluster Sampling

- Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.
- If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above.
- This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Types of Statistical Methods

- **Descriptive Statistics:** Identify important elements in a dataset.
- **Inferential Statistics:** Explain those elements via relationships with other elements.

Descriptive Statistics

Descriptive Statistics divide into 3 categories.

- **Univariate Analysis:** summarize only one variable at a time.
- **Bivariate Analysis:** compare two variables.
- **Multivariate Analysis:** compare more than two variables.

Characteristics of Descriptive Statistical Methods

- Descriptive statistical methods provide an exploratory assessment of the data from a study
- Exploratory data analysis techniques
- Organization and summarization of data
  - Tables
  - Graphs
  - Summary measures

Inferential Statistics

Inferential Statistics divide into 2 categories.

- **Hypothesis Testing:** A hypothesis is a statement that can be tested by scientific research.
- **Model Fitting:** Model fitting is a measure of how well a statistical learning model generalizes to similar data to that on which it was trained.

Characteristics of Inferential Statistical Methods

- Assess the strength of evidence for/against a hypothesis.
- Inferential statistical methods provide a confirmatory data analysis.
- Generalize conclusions from data from part of a group (sample) to the whole group (population)
- Assess the strength of the evidence
- Make comparisons.
- Make predictions.
- Ask more questions; suggest future research.

A Quick Review of Data and Variables–1

- **Variable**
  - A characteristic taking on different values.
- **Random Variable**
  - A variable taking on different possible values as a result of chance factors.
- **Quantitative or Numerical Data**
  - Implies amount or quantity
- **Discrete**
  - Random variable with values that comprise a countable set
  - There can be gaps in its possible values

A Quick Review of Data and Variables–2

- **Continuous**
  - Random variable with values comprising an interval of real numbers
  - There are no gaps in its possible values
- **Qualitative or Categorical Data**
  - Implies attribute or quality
- **Nominal**
  - Classifications based on names
- **Ordinal**
  - Classifications based on an ordering or ranking

Descriptive Statistics

- Also known as Exploratory data analysis(EDA)
- Summarize data as it is
- Do not posit any hypothesis about data
- Do not try to fit models to data
- Very important initial step
- Often neglected
- Detect outliers
- Plan how to prepare data
- Precursor to feature engineering
- Descriptive visualization

Scale of Measurement–1

Counts

- Numbers represented by whole numbers.
  - For example, number of births, number of relapses

Interval

- The same distances or intervals between values are equal.
  - For example, temperature, altitude

Ratio

- The same ratios of values are equal.
  - For example, weight, height, time, hospital length of stay
  - A true zero point indicates the absence of the quantity being measured

Scale of Measurement–2

Nominal

- Classifications based on names.
  - Binary or dichotomous
    - For example, gender, alive or dead
  - Polychotomous or polytomous
    - For example, marital status, ethnicity

Ordinal

- Classifications based on an ordering or ranking
  - For example, ratings, preferences

Methods for Organizing and Summarizing Data

- **Numerical Summary**
  - Frequency Distributions
  - Measure of Central Tendency
  - Measure of Spread or Dispersion
  - Correlation and Covariance
  - Confidence Intervals
  - Skewness and Kurtosis
- **Graphical Summary**
  - Tables
  - Histograms
  - Bar Charts
  - Box-and-whiskers plots
  - Scatter Plots
  - Pie Chart

Univariate Analysis

- Measures of Frequency, Relative Frequency
- Measures of Central Tendency
- Measures of Dispersion

Measures of Frequency

Frequency: Frequency is how often something occurs.

Example

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3

Data Values	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

Measures of Relative Frequency

Relative Frequency: How often something happens divided by all outcomes.

Example

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3

Data Values	Frequency	Relative Frequency
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or o.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Measures of Central Tendency

- Average (Mean)
- Median
- Mode
- Other infrequently used measures
  - Geometric Mean
  - Harmonic Mean

Mean

- Single best value to represent data
- Need not actually be data point itself
- Considers every point in data
- Discrete as well as continuous data
- Vulnerable to outliers

Arithmetic Mean of a Dataset

- The arithmetic mean is calculated as the sum of the values divided by the total number of values, referred to as n.

$$AM = \frac{(x_1 + x_2 + \ldots + x_n)}{n}$$

- A more convenient way to calculate the arithmetic mean is to calculate the sum of the values and to multiply it by the reciprocal of the number of values ($\frac{1}{n}$)

$$AM = (\frac{1}{n}) \times (x_1 + x_2 + \ldots + x_n)$$

- The arithmetic mean is appropriate when all values in the data sample have the same units of measure, e.g. all numbers are heights, or dollars, or miles, etc.
- When calculating the arithmetic mean, the values can be positive, negative, or zero.

Arithmetic Mean of a Dataset–1

Example: Five systolic blood pressures (mmHg) (n = 5)
120, 80, 90, 110, 95

$$Mean = \frac{120 + 80 + 90 + 110 + 95}{5} = \frac{495}{5} = 99 mmHg$$

$$Mean = \overline{x} = \frac{\sum x_i}{n}$$

- $\overline{x}$ = mean of a dataset
- x_i = data points
- n = number of sample

Arithmetic Mean of a Dataset–2

Example: Five systolic blood pressures (mmHg) (n = 5)
120, 80, 90, 110, 95

$$
\begin{aligned}
&AM = \tfrac{1}{5}(120) + \tfrac{1}{5}(80) + \tfrac{1}{5}(90) + \tfrac{1}{5}(110) + \tfrac{1}{5}(90) \\
&= \tfrac{1}{5}(120 + 80 + 90 + 110 + 95) \\
&= \tfrac{1}{5}(495) \\
&= 99 mmHg
\end{aligned}
$$

Population vs Sample Mean

Population	Sample
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
μ = number of items in the population	$\overline{x}$ = number of items in the sample

Impact of Outliers

Example: Six systolic blood pressures (mmHg) (n = 6)
120, 80, 90, 110, 95, 500

$$Mean = \frac{120 + 80 + 90 + 110 + 95 + 500}{6} = \frac{995}{6} = \boxed{165.83mmHg}$$

$$Mean = \overline{x} = \frac{\sum x_i}{n}$$

- $\overline{x}$ = mean of a dataset
- x_i = data points
- n = number of sample

Median

- Value such that 50either side
- Sort data, then use middle element
- For even number of data points, average two middle elements
- More robust to outliers than mean
- However does not consider every data point
- Makes sense for ordinal data (data that can be sorted)

Median of a Dataset: Odd Sample Size

Example: Find the median systolic blood pressures (mmHg) (n=5)
120, 80, 90, 110, 95

Median of a Dataset: Odd Sample Size

Example: Find the median systolic blood pressures (mmHg) (n=5)
120, 80, 90, 110, 95

1. **Sort Data:** 80, 90, 95, 110, 120

Median of a Dataset: Odd Sample Size

Example: Find the median systolic blood pressures (mmHg) (n=5)
120, 80, 90, 110, 95

1. **Sort Data:** 80, 90, 95, 110, 120
2. **Find the Middle Value:** $\boxed{95}$

Median of a Dataset: Even Sample Size

Example: Find the median systolic blood pressures (mmHg) (n=6)
120, 80, 90, 110, 95, 85

Median of a Dataset: Even Sample Size

Example: Find the median systolic blood pressures (mmHg) (n=6)
120, 80, 90, 110, 95, 85

1. **Sort Data:** 80, 85, 90, 95, 110, 120

Median of a Dataset: Even Sample Size

Example: Find the median systolic blood pressures (mmHg) (n=6)
120, 80, 90, 110, 95, 85

1. **Sort Data:** 80, 85, 90, 95, 110, 120
2. **Compute the Average of Middle 2 Values:** $\frac{90+95}{2} = 137.5$

Median of a Dataset: Even Sample Size

Example: Find the median systolic blood pressures (mmHg) (n=6)
120, 80, 90, 110, 95, 85

1. **Sort Data:** 80, 85, 90, 95, 110, 120
2. **Compute the Average of Middle 2 Values:** $\frac{90+95}{2} = 137.5$
3. **Computed Mean is the Median:** $\boxed{137.5}$

Impact of Outliers

Example: Five systolic blood pressures (mmHg) ($n = 5$)
120, 80, 90, 110, 500

Impact of Outliers

Example: Five systolic blood pressures (mmHg) $(n = 5)$
120, 80, 90, 110, 500

1. **Sort Data:** 80, 90,110, 120, 500

Impact of Outliers

Example: Five systolic blood pressures (mmHg) (n = 5)
120, 80, 90, 110, 500

1. **Sort Data:** 80, 90,110, 120, 500
2. **Find the Middle Value:** $\boxed{110}$

Mode

- Most frequent value in dataset
- Highest bar in histogram
- Winner in elections
- Typically used with categorical data
- Unlike mean or median, mode need not be unique
- Not great for continuous data
- Continuous data needs to be discretized and binned first

Mode of a Dataset

- **Candidate:** Abul, Akhi, Babul, Bithi, Dabul, Doli
- **Votes:** 60, 20, 10, 40, 50, 30

Mode represents the most frequent value in the data, so the winner is $\boxed{60}$

Other Measures of Central Tendency

- Geometric mean
  - Great for summarizing ratios
  - Compound Annual Growth Rate (CAGR)
- Harmonic mean
  - Great for summarizing rates
  - Resistors in parallel
  - P/E ratios in finance

Geometric Mean of a Dataset

- The geometric mean is calculated as the *nth* root of the product of all values, where n is the number of values.

$$GM = \sqrt{(x_1 \times x_2 \times \ldots \times x_n)}$$

- For example, if the data contains only two values, the square root of the product of the two values is the geometric mean. For three values, the cube-root is used, and so on.
- When calculating the arithmetic mean, the values can be positive, negative, or zero.
- The geometric mean is appropriate when the data contains values with different units of measure, e.g. some measure are height, some are dollars, some are miles, etc.
- The geometric mean does not accept negative or zero values, e.g. all values must be positive.

Harmonic Mean of a Dataset

- The harmonic mean is calculated as the number of values n divided by the sum of the reciprocal of the values (1 over each value).

$$HM = \frac{n}{(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n})}$$

- The harmonic mean is the appropriate mean if the data is comprised of rates.
- Recall that a rate is the ratio between two quantities with different measures, e.g. speed, acceleration, frequency, etc.
- The harmonic mean does not take rates with a negative or zero value, e.g. all rates must be positive.

Measures of Spread

- Range (max - min)
- Inter-quartile range (IQR)
- Standard deviation and variance

Minimum

Example: Five systolic blood pressures (mmHg) ($n = 5$)
120, 80, 90, 110, 95

- Minimum Value $= \boxed{80}$

Maximum

Example: Five systolic blood pressures (mmHg) ($n = 5$)
120, 80, 90, 110, 95

- Maximum Value $= \boxed{120}$

Range

Example: Five systolic blood pressures (mmHg) (n = 5)
120, 80, 90, 110, 95

$$\boxed{\text{Range} = \text{Maximum - Minimum}}$$

- Maximum = 120
- Minimum = 80
- $Range = 120 - 80 = \boxed{40}$

Impact of Outliers

Example: Six systolic blood pressures (mmHg) (n = 6)
120, 80, 90, 110, 95, 500

$$\boxed{\text{Range} = \text{Maximum - Minimum}}$$

- Maximum = 500
- Minimum = 80
- $Range = 500 - 80 = \boxed{420}$

Percentiles

- Divides data into 100 equal parts
- The pth percentile P is the value that is greater than or equal to p percent of the observations.
- Common percentiles are
  - 25th
  - 50th
  - 75th

Method for Calculating Percentiles

- $P_{50} = Q_2 =$ middle observation
- $P_{25} = Q_1 =$ middle observation of the lower half of observations
- $P_{75} = Q_3 =$ middle observation of the upper half of observations

Method for Calculating Percentiles

Odd Observations

- $P_{50} = Q_2 =$ middle observation
- $P_{25} = Q_1 =$ middle observation of the lower half of observations
- $P_{75} = Q_3 =$ middle observation of the upper half of observations

Even Observations

- $P_{50} = Q_2 =$ average of the middle two observations
- $P_{25} = Q_1 =$ middle observation of the lower half of n/2 observations
- $P_{75} = Q_3 =$ middle observation of the upper half of n/2 observations

Percentiles: Examples–1

Problem-1: Sample height(cm) of 9 graduate students 168, 170, 150, 160, 182, 140, 175, 180, 170(odd observations)

Percentiles: Examples–2

Problem-2: Sample height(cm) of 10 graduate students 168, 170, 150, 160, 182, 140, 175, 180, 170, 190(even observations)

Inter Quartile Range(IQR)

$$IQR = Q_3 - Q_1$$

Why IQR?

The primary advantage of using the interquartile range rather than the range for the measurement of the spread of a data set is that the interquartile range is not sensitive to outliers.

Example: Five systolic blood pressures (mmHg) $(n = 6)$

120, 80, 90, 110, 95, 85, 500

Outlier Detection

Example: Six systolic blood pressures (mmHg) $(n = 6)$
120, 80, 90, 110, 95,85,500

$$[Q_1 - 1.5IQR, Q3 + 1.5IQR]$$

Five Number Summary

Dataset: Sample height(cm) of 10 graduate students 168, 170, 150, 160, 182, 140, 175, 180, 170, 190

- Min
- Q_1
- Q_2 or Median or 50th Percentile
- Q_3
- Max

Variance

Dataset: Sample height(cm) of 10 graduate students 168, 170, 150, 160, 182, 140, 175, 180, 170, 190

1. Calculate the center value/mean
2. Subtract each value from the mean and square all of them
3. Calculate the sum of squared values
4. Divide the sum by the number of values

Population vs Sample Variance

Population	Sample
$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n-1}$
σ^2 = population variance	s^2 = sample variance

Standard Deviation

Dataset: Sample height(cm) of 10 graduate students 168, 170, 150, 160, 182, 140, 175, 180, 170, 190

$$SD = \sqrt{Variance}$$

Summary Statistics

Dataset: Sample height(cm) of 10 graduate students 168, 170, 150, 160, 182, 140, 175, 180, 170, 190

- Min
- Q_1 or 25th Percentile
- Q_2 or Median or 50th Percentile
- Q_3 or 75th Percentile
- Max
- Mean
- Standard Deviation

Point Estimation

The value of any statistic of any that estimates the value of a parameter is called a point estimation.
$\overline{x} = 2.9 \rightarrow \mu = 3.00$
We rarely know if our point estimate is correct because it is merely an estimation of the actual value.

Confidence Interval

A Confidence Interval is a range of values we are fairly sure our true value lies in.

Confidence Interval	Z-Value
90%	1.65
95%	1.69
99%	2.58
99.9%	3.291

Calculating Confidence Intervals

We measure the heights of 40 randomly chosen men, and get a mean height of 175cm,We also know the standard deviation of men's heights is 20cm.

- **Step-1**
  - the number of observations(n)
  - the mean $\overline{x}$
  - the standard deviation s
- **Step-2:**
  - number of observations n = 40
  - mean X = 175
  - standard deviation s = 20
- **Step-3:** decide what Confidence Interval we want: 95% or 99% are common choices. Then find the "Z" value for that Confidence
- **Step-4:** use that Z value in this formula for the Confidence Interval.

$$X \pm Z \frac{s}{\sqrt{n}}$$

Calculating Confidence Intervals

$$X \pm Z \frac{s}{\sqrt{n}}$$

$$175 \pm 1.960 \times \frac{20}{40} = 175cm \pm 6.20$$

Bivariate Analysis

- **Covariance:** Measures relationship between two variables specially whether greater values of one variable correspond to greater values in the other.
- **Correlation:** Similar to covariance; measures whether greater values of one variable correspond to greater values in the other. Scaled to always lie between $+1$ and -1

Covariance

- Covariance is a measure of how much two random variables vary together.
- It's similar to variance, but where variance tells you how a single variable varies, covariance tells you how two variables vary together.

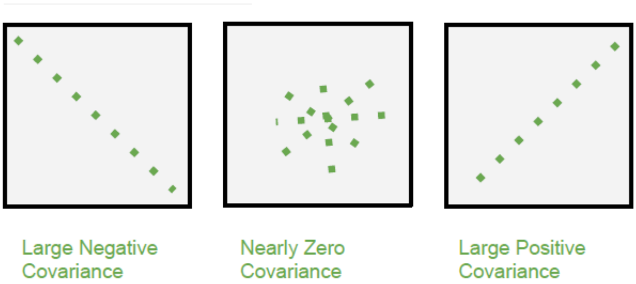


Figure: https://www.statisticshowto.com/covariance/

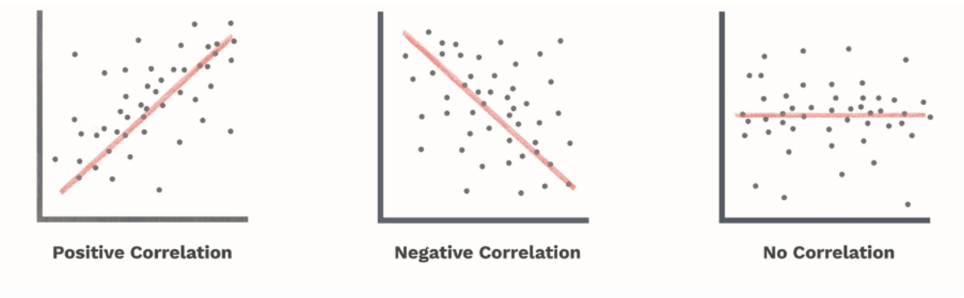
Covariance

$$cov(x,y) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

- $cov(x,y) \rightarrow$ covariance between x and y
- $x_i \rightarrow$ data value of x
- $y_i \rightarrow$ data value of y
- $\overline{x} \rightarrow$ mean of x
- $\overline{y} \rightarrow$ mean of y
- $n \rightarrow$ number of data values.

Correlation

- When two sets of data are strongly linked together we say they have a High Correlation.
- Correlation is **Positive** when the values increase together.
- Correlation is **Negative** when one value decreases as the other increases
- A correlation is assumed to be linear.



Interpretation

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

Pearson's r Correlation

- Pearson's r measures the strength of the linear relationship between two variables.
- Pearson's r is always between -1 and 1

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- $r \rightarrow$ correlation between x and y
- $x_i \rightarrow$ data value of x
- $y_i \rightarrow$ data value of y
- $\overline{x} \rightarrow$ mean of x
- $\overline{y} \rightarrow$ mean of y

Correlation Is Not Causation

- A common saying is "Correlation Is Not Causation".
- What it really means is that a correlation does not prove one thing causes the other.
- Causation means that one variable causes something to happen in another variable.
- To say that two things are correlated is to say that they are not some kind of relationship.
- In order to imply causation, a true experiment must be performed where subjects are randomly assigned to different conditions.

Skewness

Skewness: A measure of asymmetry around the mean.

- Normally distributed data: skewness $= 0$
- Extreme values are equally likely on both sides of the mean
- Symmetry about the mean

Positive Skewness

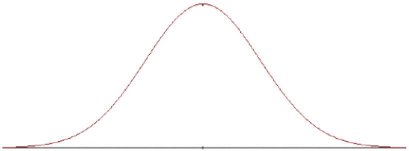
- Consider incomes of individuals
- Billionaires: positive skew
- Outliers greater than mean more likely than outliers less than mean
- Right-skewed distribution
- Often seen when lower bound but no upper bound

Kurtosis

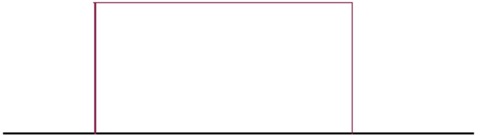
Kurtosis: Measure of how often extreme values (on either side of the mean) occur.
Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.

- Normally distributed data: kurtosis = 3
- Excess kurtosis = kurtosis - 3
- Kurtosis Tail risk
- High kurtosis extreme events more likely than in normal distribution.

Distribution



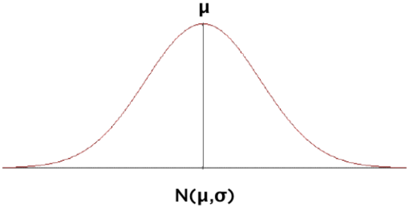
(a) Values close to the mean are more likely



(b) All values are equally likely

The Normal Curve

- The distributions of most continuous random variables will follow the shape of the normal curve.
- Mean, Median and Mode all exist at the center.



A formula which tells how likely a particular value is to occur in your data.

The Empirical Rule–1

The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean.

- 68% of all values fall within 1 standard deviation of the mean.
- 95% of the all values fall within 2 standard deviation of the mean.
- 99.7% of the all values fall within 3 standard deviation of the mean.

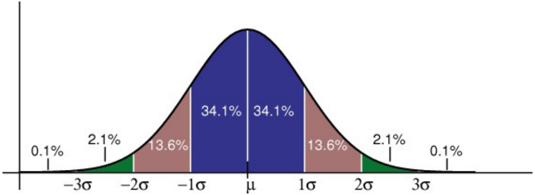
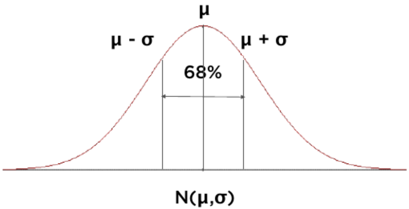


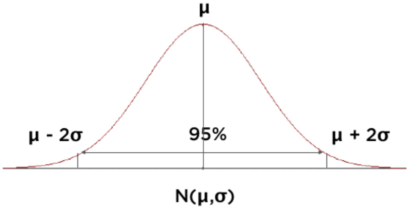
Figure:
https://www.statisticshowto.com/probability-and-statistics/normal-distributions/

The Empirical Rule–2



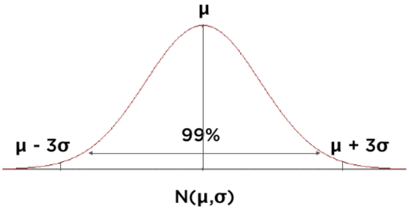
68% within 1 standard deviation of mean

The Empirical Rule–3



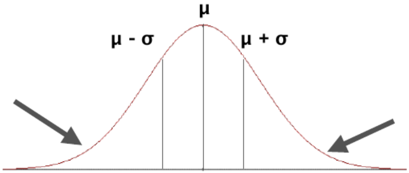
95% within 2 standard deviations of mean

The Empirical Rule–4



99% within 3 standard deviations of mean

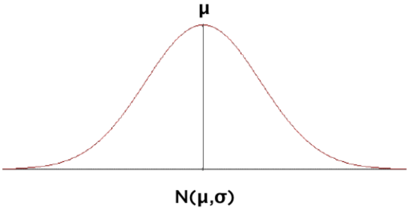
Impact of Outliers



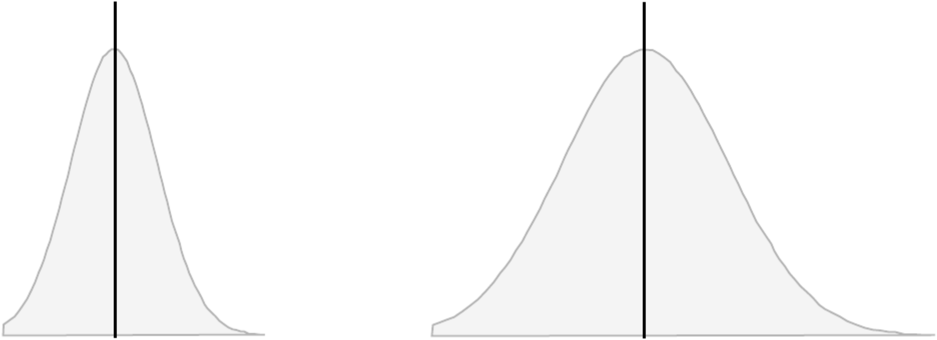
There will be few extreme values - the number of extreme values at either side of the mean will be the same.

Properties of a Normal Distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.



Role of Sigma(σ)



(a) Small Standard Deviation(σ)

Few points far from the mean

(b) Large Standard Deviation(σ)

Many points far from the mean

Z-Scores

- Z-Scores are standardized values that can be used to compare scores in different distributions.
- Simply put, a z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is.
- A z-score can be placed on a normal distribution curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve).
- In order to use a z-score, you need to know the population mean μ and also the population standard deviation σ.

Calculating Z-Score

$$Z = \frac{\overline{x} - \mu}{\sigma}$$

- $\overline{x}$ → mean
- μ → population mean
- σ → population standard deviation

Calculating Z-Score

For example, let's say you have a test score of 190. The test has a mean (μ) of 150 and a standard deviation (σ) of 25. Assuming a normal distribution, your z score would be

Skewness–1

- A measure of asymmetry around the mean.
- If one tail is longer than another, the distribution is skewed.
- These distributions are sometimes called asymmetric or asymmetrical distributions as they don't show any kind of symmetry.
- Symmetry means that one half of the distribution is a mirror image of the other half.

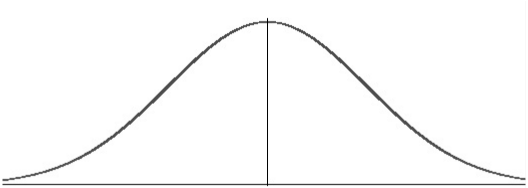
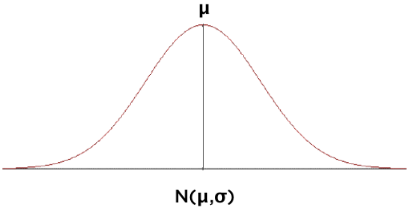


Figure: https://www.statisticshowto.com/probability-and-statistics/

Skewness–2

- Normally distributed data: skewness $= 0$
- Extreme values are equally likely on both sides of the mean.
- Symmetry about the mean



Negative Skewness

- A left-skewed distribution has a long left tail.
- Left-skewed distributions are also called negatively-skewed distributions.
- That's because there is a long tail in the negative direction on the number line. The mean is also to the left of the peak.

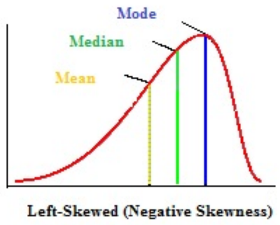


Figure: https://www.statisticshowto.com/probability-and-statistics/

Positive Skewness

- A right-skewed distribution has a long right tail.
- Right-skewed distributions are also called positive-skew distributions.
- That's because there is a long tail in the positive direction on the number line. The mean is also to the right of the peak.

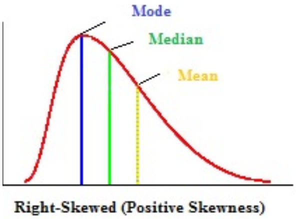


Figure: https://www.statisticshowto.com/probability-and-statistics/

Mean and Median in Skewed Distributions

In a normal distribution, the mean and the median are the same number while the mean and median in a skewed distribution become different numbers.

- A left-skewed, negative distribution will have the mean to the left of the median.

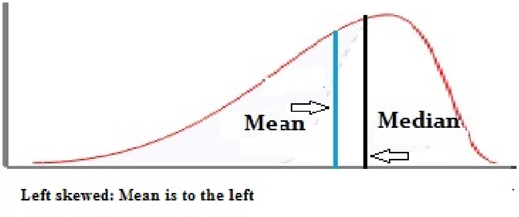


Figure: `https://www.statisticshowto.com/probability-and-statistics/`

Mean and Median in Skewed Distributions

- A right-skewed, negative distribution will have the mean to the right of the median.

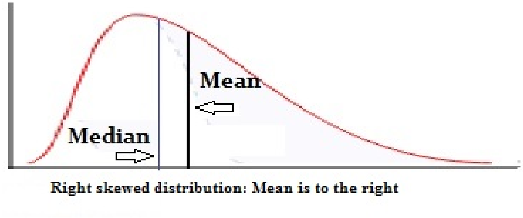


Figure: https://www.statisticshowto.com/probability-and-statistics/