ISCB20.05–Introduction to Biostatistics

Md. Jubayer Hossain
https://jhossain.me/
December 16, 2020

Lead Organizer, Introduction to Scientific Computing for Biologists
Founder, Health Data Research Organization

HEALTH DATA
RESEARCH ORGANIZATION

Section–2.1: Interpreting Data Using Descriptive Statistics

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

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

Impact of Outliers

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

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

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

- $\overline{x}$ = 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: Five 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, 500

Outlier Detection

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

References

- https://www.formpl.us/blog/research/home
- https://www.scribbr.com/methodology/
- https://www.questionpro.com/blog/
- https://www.scribbr.com/category/research-process/
- https://ocw.jhsph.edu/index.cfm/go/viewCourse/course/FundEpi/coursePage/index/

Thank You

☺