Stats for Data Analysis

Jubayer Hossain

Health Data Research Organization

September 17, 2020

Describing Data

- Numerical Summary
  - Frequency Table
  - Measure of Center / Location
  - Measure of Spread / Variability
- Graphical Summary – Categorical Data
  - Categorical Data– Bar Chart, Pie Chart
  - Numerical Data – Histogram, Boxplot, Scatter Plot

Frequency Table

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 VALUE	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 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

https://openstax.org/books/introductory-statistics/

Measure of Center

- Mean – The "average" number; found by adding all data points and dividing by the number of data points.
- Mode – The most frequent number—that is, the number that occurs the highest number of times.
- Median – The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

Mean

- Add up data, then divide by sample size (n)
- The sample size n is the number of observations (pieces of data)

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

Median: Odd Sample Size

- The median is the middle number (also called the 50 th percentile)
- Other percentiles can be computed as well, but are not measures of center

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

Median: Odd Sample Size

- If the sample size is an even number

Example: Find the median systolic blood pressures (mmHg)
125, 110,90, 80, 120, 95

Why Median?

- The sample mean is sensitive to extreme values
- The sample median is not sensitive to extreme values

Before: 110,90, 80, 120, 95
After: 80, 90, 95, 110, 200
if 120 became 200, the median would remain the same, but the mean would change to 115

Mode

The most frequent number—that is, the number that occurs the highest number of times.
Example: 120, 80,80, 80 90, 110, 95

Measure of Spread

- Range
- Percentile
- Variance
- Standard Deviation

Range

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

The Range Can Be Misleading

Five systolic blood pressures (mmHg) (n = 5)
120, 80, 90, 110, 95, 1200

Percentile: Divides Data into 100 Equal Parts

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

Quartiles: Divides Data into 4 Equal Parts

Example: 5, 7, 4, 4, 6, 2, 8 (Odd Sample)

Quartiles: Divides Data into 4 Equal Parts

Example:1, 3, 3, 4, 5, 6, 6, 7, 8, 8 (Even Sample)

Interquartile Range

Example: 5, 7, 4, 4, 6, 2, 8 (Odd Sample)
Example:1, 3, 3, 4, 5, 6, 6, 7, 8, 8 (Even Sample)

The Significance of the Interquartile Range

The range gives us a measurement of how spread out the entirety of our data set is. The interquartile range, which tells us how far apart the first and third quartile are, indicates how spread out the middle 50% of our set of data is.

Interquartile Range Resistance to Outliers

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 = 5)
120, 80, 90, 110, 95, 125, 500

Variance / Variation

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

Standard Deviation