Ilmu Komputer
Universitas Pendidikan Ganesha

Statistics

Dr. Komang Setemen, S.Si., M.T.

Descriptive Statistics

Outline

- What are descriptive statistics
- Understanding descriptive statistics
- Types of descriptive statistics
- Main purpose of descriptive statistics

What are descriptive statistics

- Descriptive statistics summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
- Measures of central tendency describe the center of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).
- Measures of frequency distribution describe the occurrence of data within the data set (count).

Understanding descriptive statistics

- Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.
- The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics.
- The mean, or the average, is calculated by adding all the figures within the data set and then dividing by the number of figures within the set.

Understanding descriptive statistics

- For example, the sum of the following data set is 20: (2, 3, 4, 5, 6). The mean is 4 (20/5).
- The mode of a data set is the value appearing most often, and the median is the figure situated in the middle of the data set.
- It is the figure separating the higher figures from the lower figures within a data set.

Understanding descriptive statistics

Important!

- Descriptive statistics, especially in fields such as medicine, often visually depict data using scatter plots, histograms, line graphs, or stem and leaf displays

Types of descriptive statistics

- Central tendency
- Measures of variability
- Distribution

Central tendency

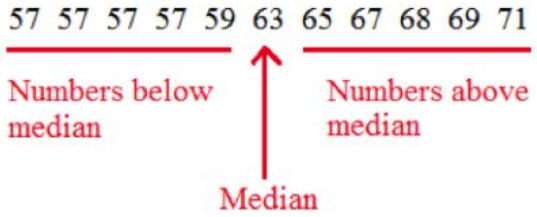
- Measures of central tendency focus on the average or middle values of data sets
- Measures of central tendency describe the center position of a distribution for a data set.
- A person analyzes the frequency of each data point in the distribution and describes it using the **mean**, **median**, or **mode**, which measures the most common patterns of the analyzed data set

Central tendency

- We have the data: 71 59 69 68 63 57 57 57 57 65 67
- Finding the mean, median, and mode!

$$Mean = \bar{x} = \frac{71 + 59 + 69 + 68 + 63 + 57 + 57 + 57 + 57 + 65 + 67}{11} = 62.7$$

Mode: This is the data value that occurs most often
First put the data in order from smallest to largest.
57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71
Now work from the outside in, until you get to the middle number.



Mode: From the ordered list it is easy to see that 57 occurs four times and no other data values occur that often. So the mode is 57

Measures of variability

- Measures of variability (or the measures of spread) aid in analyzing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set.
- Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.
- Consider the following data set: 5, 19, 24, 62, 91, 100. The range of that data set is 95, which is calculated by subtracting the lowest number (5) in the data set from the highest (100).

Measures of variability (example)

- **Finding the Range**

  Find the range for each data set.

  a. 10, 20, 30, 40, 50 -> Range = 50 – 10 = 40

  b. 10, 35, 36, 37, 50 -> Range = 50 – 10 = 40
- **Deviation from the Mean: data value – mean($\bar{x}$)**
- To see how this works, let's use the data set from Example below. The mean was about 62.7.

x	71	59	69	68	63	57	57	57	57	65	67
x-$\bar{x}$	8.3	-3.7	6.3	5.3	0.3	-5.7	-5.7	-5.7	-5.7	2.3	4.3

Measures of variability (example)

- **Squared Deviations from the Mean:** To find these values, square the deviations from the mean. Also, you can think of this as being the squared distance from the mean

x	71	59	69	68	63	57	57	57	57	65	67	
x-$\bar{x}$	8.3	-3.7	6.3	5.3	0.3	-5.7	-5.7	-5.7	-5.7	2.3	4.3	
$(\mathrm{x}-\bar{x})^2$	68.89	13.69	39.69	28.09	0.09	32.49	32.49	32.49	32.49	5.29	18.49	= 304.19

Measures of variability (example)

- **Sample Variance:** This is the sum of the squared deviations from the mean divided by **n-1** . The symbol for sample variance is S^2, and the formula for the sample variance is:

- $S2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{304.19}{11-1} = 30.419$

- The variance measures the average squared distance from the mean. Since we want to know the average distance from the mean, we will need to take the square root at this point.

Measures of variability (percentiles)

- There are other calculations that we can do to look at spread. One of those is called percentile. This looks at what data value has a certain percent of the data at or below it.
- **Percentiles:** A value with k-percent of the data at or below this value.
- For example, if a data value is in the 80th percentile, then 80% of the data values fall at or below this value.
- There are three percentiles that are commonly used. They are the first, second, and third quartiles, where the quartiles divide the data into 25% sections.
- **First Quartile (Q1):** 25th percentile (25% of the data falls at or below this value.)
- **Second Quartile (Q2 or M):** 50th percentile, also known as the median (50% of the data falls at or below this value.).
- **Third Quartile (Q3):** 75th percentile (75% of the data falls at or below this value)

Measures of variability (formula)

- $\boldsymbol{P_i = \frac{i(n+1)}{100}}$,

  where P_i = the value at which i percentage of data lie below that value

  n = total number of data

- $\boldsymbol{Q_i = \frac{i(n+1)}{4}}$,

  where Q_i = i Quartile (like Q1, Q2, or Q3)

  n = total number of data

- **Ex:** 6,5,8,7,9,4,5,8,4,7,8,5,8,4,5
- Find the P65, Q1, Q2, and Q1

Measures of variability (example)

- First, order the data from least to greatest -> **4,4,4,5,5,5,5,6,7,7,8,8,8,8,9**
- Then, find the position like the formula above

n=15, P65=65(15+1)/100=10.4 ->**10**. So, the P65 is the data which position to **10**, that is **7**

Q1=1(15+1)/4=4, So, the Q1 is the data which position to **4**, that is **5**
Q2=2(15+1)/4=8, So, the Q1 is the data which position to **8**, that is **6**
Q3=3(15+1)/4=12, So, the Q1 is the data which position to **12**, that is **8**

Measures of variability (distribution)

- Distribution (or frequency distribution) refers to the quantity of times a data point occurs.
- Alternatively, it is the measurement of a data point failing to occur.
- Consider a data set: **male, male, female, female, female, other**. The distribution of this data can be classified as:

The number of males in the data set is 2.
The number of females in the data set is 3.
The number of individuals identifying as other is 1.
The number of non-males is 4.

Main Purpose of Descriptive Statistics

- The main purpose of descriptive statistics is to provide information about a data set.
- In the example above, there are hundreds of baseballs players that engage in thousands of games.
- Descriptive statistics summarizes the large amount of data into several useful bits of information.

Thank you