Brief Article

The Author

February 9, 2016

Chapter 1

Why should we use data science?

—

Solution 1.7-1
All of the choices is the correct answer.

—

Solution 1.7-2
The level of measurement of marital status is nominal.

—

Solution 1.7-3
The level of measurement of marital status is category - nominal.

—

Solution 1.7-4
The number of reported robberies in June 2014 is Kalamazoo County is a continuous variable.

—

Solution 1.7-5

Numeric - interval

—

Solution 1.7-6
Numeric - ratio

—

Solution 1.7-7
all adult residents of the U.S.

—

Solution 1.7-8
birth weight.

—

Solution 1.7-9

Variable: Homicide rate
Level of Measurement: Interval-ratio
Type: Continuous
Application: Descriptive (two variables)

—

Solution 1.7-10

Variable: party, gender, opinion
Level of Measurement: nominal, nominal, ordinal
Type: discrete, discrete, discrete
Application: inferential, NA, NA

Chapter 2

Dataset Descriptive Information

—

Solution 2.6-1

$$(2.)p = \frac{250 - 195}{250} = 0.22 \tag{2.1}$$

—

Solution 2.6-2

$$(3.)ratio = \frac{195}{250 - 195} = 3.55 \tag{2.2}$$

—

Solution 2.6-3

$$(2.)rate = \frac{13}{25000} \times 100000 = 52 \tag{2.3}$$

—

Solution 2.6-4

$$(1.)PC = \frac{(83 - 89)}{89} \times 100 = -6.7\% \tag{2.4}$$

—

Solution 2.6-5
(1.) The percentage of nurses who are female is 36.8%. **(2.)** The proportion of orderlies who are males is 0.367. **(3.)** Ratio is 18 females docs to 83 males docs or approximately 1 female doc for every 5 males docs. **(4.)** Percentage of females on the staff is 43.3%.

—

Solution 2.6-6
The measure of the lost hours due to traffic is interval-ratio.

—

Solution 2.6-7

$$pc = \frac{169.53 - 159.90}{159.90} \times 100 = 6.02\% \tag{2.5}$$

—

Solution 2.6-8

Min. 1st Qu. Median Mean 3rd Qu. Max. 15.70 17.60 18.30 18.24 19.28 20.70

—

Solution 2.6-9
$p = \frac{(19+15)}{50} \times 100 = 68$

—

Solution 2.6-10
violent crime rate is [1] 5.625563

property crime rate is [1] 46.97676

Chapter 3

Measures of Location

—

Solution 3.10-1

Compute the mean, median, and mode for the weekly grocery budget mean = 32.167, median = 32.5, mode = 35

—

Solution 3.10-2

The third value must be 22 so that the average is 20 for all three quizzes.

—

Solution 3.10-3

The mean is 472.6 The median is 555 There is no mode since there are no duplicates.

1. The greater value is the median (555). **2.** There is a negative skewness in the middle half the of the dataset. (Refer to Figure 3.1.)

3. When we compare the median to the mean, we find the median (555) is greater the mean (472.6). Therefore, we can say the the dataset is left skewed.

—

Solution 3.10-4

(1.) Mode **(2.)** Median **(3.)** Mean **(4.)** Mean **(5.)** Median **(6.)** Mode

—

Solution 3.10-5
Birth Mode="North"; **Legal** Median=2.5; **Expense** Mean=48.5; **Movies** Mean=5.8; **Food** Median=6; **Religion** Mode="Protestant"

—

Solution 3.10-6

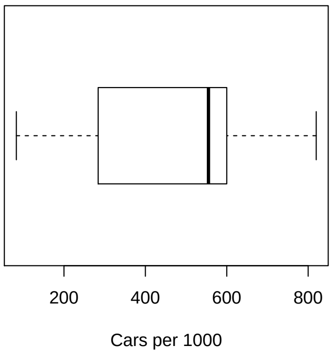

Figure 3.1: Boxplot of Cars per 1000

The mean for 2005 is 55.4. The median for 2005 is 54.5. The mean for 2015 is 57.1. The median for 2015 is 57.

—

Solution 3.10-7

The mean is [1] 31.8
The median is [1] 35

—

Solution 3.10-8

The mean is 28.72, and the median is 30

—

Solution 3.10-9

The pretest mean is 9.3333333
The pretest median is 10
The posttest mean is 12.9333333
The pretest median is 12

—

Solution 3.10-10

The sex ed pretest mean is 9.3333333
The sex ed posttest mean is 12.9333333.

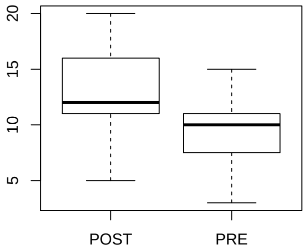

Figure 3.2: Box plot of Post - Pre Test Scores

After reviewing the box plot of the differences between posttest and pretest scores, it appears that the students learned the subject material.

Chapter 4

Measures of Spread

—

Solution 4.10-1

The range is $(X_{max} - X_{min}) = 40 - 25 = 15$

—

Solution 4.10-2

LB = Q_1 - 1.5(IQR) = 28 - 1.5 (35 - 28) = 17.5
UB = Q_3 + 1.5(IQR) = 35 + 1.5 (35 - 28) = 45.5

An outlier is any value that is less than 17.5 or greater than 45.5; therefore, no outliers in this set of data.

—

Solution 4.10-3

The standard deviation is 236.7524373

—

Solution 4.10-4

The range is 14.
The IQR is 4.775
The standard deviation is 3.2241029
The coefficient of variation is 38.8583832
The variance is 10.3948396

—

Solution 4.10-5

The LB outlier is No LB outlier and the UP outlier is No UB outlier.

—

Solution 4.10-6

First we must determine the data types for each variable:

- height is numeric, ratio continuous
- GPA is numeric, interval
- gender is categorical, nominal
- major is categorical, nominal

Note: numerical variables can use the mean, standard deviation, median, histograms, and line charts; while categorical variables can use frequency tables, mode, bar and pie charts. For example, the graphics for gender and major should be a bar or pie chart because they are categorical variables, and the graphics for GPA and height should be box plots or histograms.

—

Solution 4.10-7

The range is (r47), interquartile range is 15, standard deviation is 12.6918609, and variance is 161.0833333.

—

Solution 4.10-8

The range is 41, interquartile range is 25, standard deviation is 13.6763787, and variance is 187.0433333.

—

Solution 4.10-9

The pretest range is 12 and standard deviation is 3.6187343
The posttest range is 15 and standard deviation is 4.558613

—

Solution 4.10-10

The sex ed pretest mean is 9.3333333
The sex ed posttest mean is 12.9333333.
After reviewing the box plot of the differences between posttest and pretest scores, it appears that the students learned the subject material.

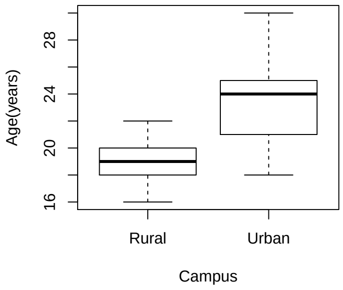

Figure 4.1: Box plot of Post - Pre Test Scores