HW #1: Review of Basic Statistical Concepts, Descriptive Statistics, & Normal Distribution

Eddie Chapman

September 10, 2019

1. A psychologist records how many words participants recalled from a list under three different conditions: large reward for each word recalled, small reward for each word recalled, and no reward.

   a. What is the independent variable?
   Reward condition

   b. What is the dependent variable?
   Word recall count

   c. What kind of scale is being used to measure the dependent variable?
   Ratio scale

2. Which of the following would be called a statistic and which would be called a parameter?

   a. The average income for 100 US citizens selected at random from various telephone books
   Statistic

   b. The average income of citizens in the United States
   Parameter

   c. The highest age among respondents to a sex survey in a popular magazine
   Statistic (generously)

Exercises #3-6 are based on the following values for two variables, X and Y:

$$X_1 = 2 \quad X_2 = 4 \quad X_3 = 6 \quad X_4 = 8 \quad X_5 = 10$$
$$Y_1 = 3 \quad Y_2 = 5 \quad Y_3 = 7 \quad Y_4 = 9 \quad Y_5 = 11$$

3. By hand (you may use a calculator), find the value of each of the following expressions:

$$a. \quad \sum_{i=2}^{5} X_i \qquad 4 + 6 + 8 + 10 = 28$$

$$b. \quad \sum 5X_i \qquad (5 * 2) + (5 * 4) + (5 * 6) + (5 * 8) + (5 * 10) = 150$$

$$c. \quad \sum 3Y_i \qquad (3 * 3) + (3 * 5) + (3 * 7) + (3 * 9) + (3 * 11) = 105$$

$$d. \quad \sum X_i^2 \qquad 2^2 + 4^2 + 6^2 + 8^2 + 10^2 = 220$$

$$e. \quad (\sum Y_i)^2 \qquad (3 + 5 + 7 + 9 + 11)^2 = 1225$$

4. Now, use R to evaluate all of the expressions from #3.

```
X = c(2, 4, 6, 8, 10)
Y = c(3, 5, 7, 9, 11)
```

   a. $\sum_{i=2}^{5} X_i$

```r
sum(X[2:5])
```

```
## [1] 28
```

b. $\sum 5X_i$

```r
sum(5 * X)
```

```
## [1] 150
```

c. $\sum 3Y_i$

```r
sum(3 * Y)
```

```
## [1] 105
```

d. $\sum X_i^2$

```r
sum(X^2)
```

```
## [1] 220
```

e. $(\sum Y_i)^2$

```r
sum(Y)^2
```

```
## [1] 1225
```

5. By hand (you may use a calculator), find the value of each of the following expressions:

$$\begin{array}{lll}
a. & \sum(X+Y) & (2+3)+(4+5)+(6+7)+(8+9)+(10+11)=65 \\
b. & \sum XY & (2*3)+(4*5)+(6*7)+(8*9)+(10*11)=250 \\
c. & (\sum X)(\sum Y) & (2+4+6+8+10)*(3+5+7+9+11)=1050 \\
d. & \sum(Y-2) & (3-2)+(5-2)+(7-2)+(9-2)+(11-2)=25
\end{array}$$

6. Now, use R to evaluate all of the expressions from #5.

a. $\sum(X+Y)$

```r
sum(X + Y)
```

```
## [1] 65
```

b. $\sum XY$

```r
sum(X * Y)
```

```
## [1] 250
```

c. $(\sum X)(\sum Y)$

```r
sum(X) * sum(Y)
```

```
## [1] 1050
```

d. $\sum(Y-2)$

```
sum(Y - 2)
```

```
## [1] 25
```

7. A veterinarian is interested in the life span of golden retrievers. She recorded the age at death (in years) of the retrievers treated in her clinic. The ages were $12, 9, 11, 10, 8, 14, 12, 1, 9, 12$.

   a. Use R to calculate the mean, median, and mode for age at death.

```
# Return a vector's most frequently occurring value
#
# First, a new vector is created containing the input values
# with duplicates removed. The two vectors are compared to find
# the index positions where each input value occurs in the unique
# vector. The values (positions) in the index vector are counted
# for frequency and the position of the most frequent position
# value is returned. This is used to retrieve a value from the
# unique vector.
#
# https://stackoverflow.com/questions/2547402/is-there-a-built-
# in-function-for-finding-the-mode
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
```

```
dogs <- c(12, 9, 11, 10, 8, 14, 12, 1, 9, 12)
```

```
mean(dogs)
```

```
## [1] 9.8
```

```
median(dogs)
```

```
## [1] 10.5
```

```
Mode(dogs)
```

```
## [1] 12
```

   b. After examining her records, the veterinarian discovered that the dog that had died at 1 year was killed by a car. Using R, recalculate the mean, median, and mode without that dog's data.

```
dead_dogs <- c(1)
dogs <- dogs[-which(dogs == dead_dogs)]
dogs
```

```
## [1] 12  9 11 10  8 14 12  9 12
```

```
mean(dogs)
```

```
## [1] 10.77778
```

```
median(dogs)
```

```
## [1] 11
```

```r
Mode(dogs)
```

```
## [1] 12
```

c. Which measure of central tendency in part *b* changed the most, compared to the values in part *a*? Why?

   The one-year-old dog was an outlier in the dataset. None of the other dogs lived less than 8 years. Removing this value affected the mean most of all, because it is sensitive to outliers. The median was also impacted because the removed value happened to be the minimum value. This change was not as dramatic.

8. By hand, calculate the mean, sums of squares (SS), and variance (s^2) for the following sample of scores: $11, 17, 14, 10, 13, 8, 7, 14$.

   mean: $\frac{11+17+14+10+13+8+7+14}{8} = 11.75$

   sums of squares:

   $(11-11.75)^2+(17-11.75)^2+(14-11.75)^2+(10-11.75)^2+(13-11.75)^2+(8-11.75)^2+(7-11.75)^2+(14-11.75)^2 = 79.5$

   variance: $\frac{79.5}{7} = 11.357$

9. Now using R, calculate all of the quantities from #8.

```r
scores <- c(11, 17, 14, 10, 13, 8, 7, 14)
```

```r
mean(scores)
```

```
## [1] 11.75
```

```r
sum((scores - mean(scores))^2)
```

```
## [1] 79.5
```

```r
sum((scores - mean(scores))^2) / (length(scores) - 1)
```

```
## [1] 11.35714
```

10. If you convert each score in a set of scores to a z-score, which of the following will be true about the resulting set of z-scores?

    a. The mean will equal 1.
    b. The variance will equal 1.
    c. The distribution will be normal in shape.
    d. All of the above.
    e. None of the above.

11. The SAT has a mean of 500 and a standard deviation of 100 in the population. What SAT score corresponds to

    a. $z = -0.2$
    b. $z = +1.3$
    c. $z = -3.1$
    d. $z = +1.9$

12. Use the z-table to find the area under the normal distribution beyond z when z equals

    a. $+0.09$
    b. $+1.05$
    c. $+1.96$

13. On a normal distribution, find the area between

    a. $z = -0.5$ and $z = +1.0$
    b. $z = -1.5$ and $z = +0.75$
    c. $z = +0.75$ and $z = +1.5$

14. Assume that the resting heart rate in humans is normally distributed with $\mu = 72$bpm (beats perminute) and $\sigma = 8$bpm.

    a. What proportion of the population has resting heart rates above 82 bpm?
    b. What proportion of the population has resting heart rates below 75 bpm?
    c. What proportion of the population has resting heart rates between 80 and 85 bpm?

15. A set of reading scores for fourth grade children has a mean of 25 and a standard deviation of 5. A set of scores for ninth grade children has a mean of 30 and a standard deviation of 10. Assume that the distributions are normal.

    a. Draw a rough sketch of these data, putting both groups in the same figure.
    b. What percentage of the fourth graders score better than the average ninth grader?
    c. What percentage of ninth graders score worse than the average fourth grader?