HW2

Nicholas Alexeev

8/29/2021

1.How does the speed of a runner vary over the course of a marathon (a distance of 42.195 km)? Consider determining both the time to run the first 5 km and the time to run between the 35-km and 40-km points, and then subtracting the former time from the latter time. [. . .] The accompanying histogram is based on times of runners who participated in several different Japanese marathons ("Factors Affecting Runners' Marathon Performance," Chance, Fall, 1993: 24–30). What are some interesting features of this histogram? What is a typical difference value? Roughly what proportion of the runners ran the late distance more quickly than the early distance?

The histogram is skewed to the right. A typical difference is 150. Roughly 1 percent of runners ran late distances faster then earlier distances. # 2. Mercury is a persistent and dispersive environmental contaminant found in many ecosystems around the world. [. . .] The accompanying blood mercury concentration (μg/g) for adult females near contaminated rivers in Virginia was read from a graph in the article, "Mercury Exposure Affects the Reproductive Success of a Free-Living Terrestrial Songbird, the Carolina Wren" (The Auk, 2011: 759–769; this is a publication of the American Ornithologists Union). # .20, .22, .25, .30, .34, .41, .55, .56, 1.42, 1.70, 1.83, 2.20, 2.25, 3.07, 3.25

a. Determine the values of the sample mean and median and explain why they are different. (Hint:

$$\sum x_i = 18.55$$

).

```
data <- c(.20, .22, .25, .30, .34, .41, .55, .56, 1.42, 1.70, 1.83, 2.20, 2.25, 3.07, 3.25);
mean(data)
```

```
## [1] 1.236667
```

```
median(data)
```

```
## [1] 0.56
```

The Mean is 1.23, the median is 0.56. the median is lower then the mean because the data is skewed to the right. ## b. Determine the value of the 10% trimmed mean and compare to the mean and median.

```
trimmed_data = c( .22, .25, .30, .34, .41, .55, .56, 1.42, 1.70, 1.83, 2.20, 2.25, 3.07);
trimmed_mean <-mean(trimmed_data)
print(trimmed_mean)
```

```
## [1] 1.161538
```

```
trimmed_median <-median(trimmed_data)
print(trimmed_median)
```

```
## [1] 0.56
```

The trimmed mean is lower then the sample mean because the extreme data is removed. The median stays the same because the middle of the data has not moved. ## c. By how much could the observation .20 be increased without affecting the value of the sample median? Observation 0.2 could be increased to 0.56 without changing the sample median because when the observation is less than or equal to 0.56 the position of 0.56 does not change.

3. The article "Oxygen Consumption During Fire Suppression: Error of Heart Rate Estimation" (Ergonomics, 1991: 1469–1474) reported the following data on oxygen consumption (mL/kg/min) for a sample of ten firefighters performing a fire-suppression simulation:

29.5, 49.3, 30.6, 28.2, 28.0, 26.3, 33.9, 29.4, 23.5, 31.6

Compute the following:

a. The sample range

Sample Range:

```
data <- c(29.5, 49.3, 30.6, 28.2, 28.0, 26.3, 33.9, 29.4, 23.5, 31.6);
max(data)-min(data)
```

```
## [1] 25.8
```

b. The sample variance s2 from the definition (i.e., by first computing deviations, then squaring them, etc.)

```
data <- c(29.5, 49.3, 30.6, 28.2, 28.0, 26.3, 33.9, 29.4, 23.5, 31.6);
mean_data <- mean(data)
s_2 <- 0.0;
for (i in data){
  s_2<- s_2+(i-mean_data)**2;
}
s_2 = s_2/(length(data)-1)
print(s_2)
```

```
## [1] 49.31122
```

c. The sample standard deviation

```
stdev <- function (data){
  mean_data = mean(data)
  s_2 = 0.0;
  for (i in data){
    s_2<- s_2+(i-mean_data)**2;
  }
  s_2 <- s_2/(length(data)-1);
  s_2**(0.5)
}
print(stdev(data))
```

```
## [1] 7.022195
```

d. s2 using the shortcut method

```
short<-0;
sum_x_2 <-0
sum_x <-0;
for (x in data){
  sum_x_2 <- sum_x_2 + x**2;
  sum_x  <- sum_x + x
}
short_s_2 <- (sum_x_2-((sum_x)^2)/length(data))/(length(data)-1)
print(short_s_2)
```

```
## [1] 49.31122
```

Exercise 34 presented the following data on endotoxin concentration in settled dust both for a sample of urban homes and for a sample of farm homes:

U: 6.0, 5.0, 11.0, 33.0, 4.0, 5.0, 80.0, 18.0, 35.0, 17.0, 23.0

F: 4.0, 14.0, 11.0, 9.0, 9.0, 8.0, 4.0, 20.0, 5.0, 8.9, 21.0 9.2, 3.0, 2.0, 0.3

a. Determine the value of the sample standard deviation for each sample, interpret these values, and then contrast variability in the two samples. [Hint: $\sum$xi = 237.0 for the urban sample and = 128.4 for the farm sample, and $\sum$x2i = 10,079 for the urban sample and 1617.94 for the farm sample.]

```
U<-c(6.0, 5.0, 11.0, 33.0, 4.0, 5.0, 80.0, 18.0, 35.0, 17.0, 23.0)
F<-c(4.0, 14.0, 11.0, 9.0, 9.0, 8.0, 4.0, 20.0, 5.0, 8.9, 21.0, 9.2, 3.0, 2.0, 0.3)
print("U Standard deviation")
```

```
## [1] "U Standard deviation"
```

```
stdev(U)
```

```
## [1] 22.29961
```

```r
print("F Standard deviation")
```

```
## [1] "F Standard deviation"
```

```r
stdev(F)
```

```
## [1] 6.087669
```

The standard deviation represents the spread of the sample data. Urban homes have larger variability in concentration of toxins. ## b. Compute the fourth spread for each sample and compare. Do the fourth spreads convey the same message about variability that the standard deviations do? Explain.

```r
spread <-function(data){
  sorted <-sort(data);
  lower <- median(sorted[1:(length(sorted)/2)]);
  upper <- median(sorted[(length(sorted)/2+1):length(sorted)])
  upper-lower

}
U_spread <- spread(U)
print("U fourth spread")
```

```
## [1] "U fourth spread"
```

```r
print(U_spread)
```

```
## [1] 18
```

```r
print("F fourth spread")
```

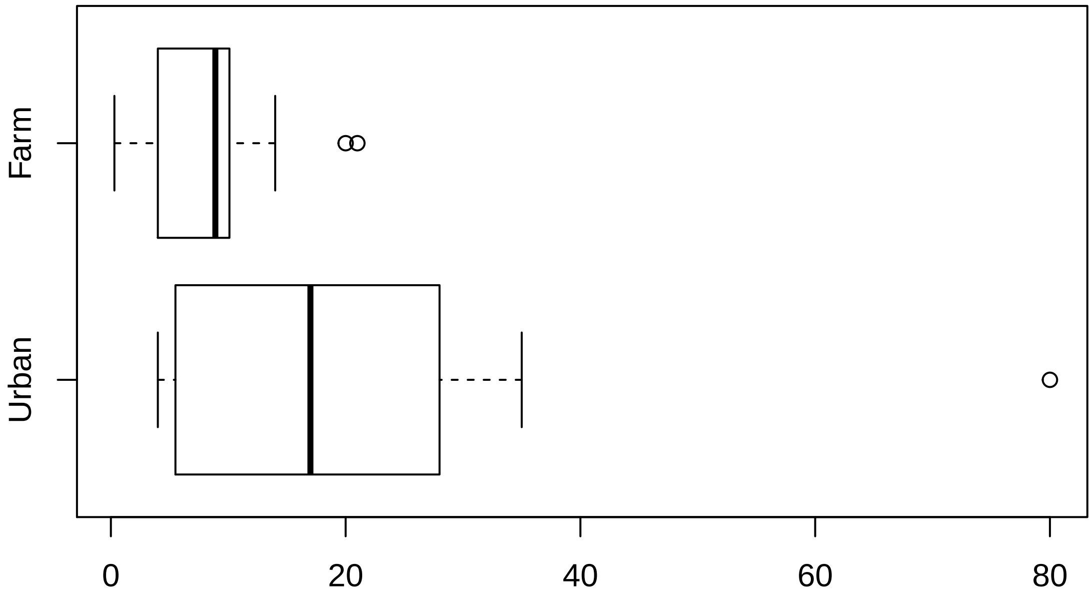
```
## [1] "F fourth spread"
```

```r
print(spread(F))
```

```
## [1] 5.2
```

The spread conveys the same information about spread as the standard deviation because the standard deviation for U is still larger then the standard deviation for F. ## c. Create side-by-side boxplots of these two data sets by hand or using R; compare and contrast. (This replaces part (c) in the text.)

```r
boxplot(U,F,names = c("Urban","Farm"),horizontal = TRUE)
```



The farm data has a lower mean and spread then the urban toxin concentration.

5. Devore §1.4 # 50. In 1997 a woman sued a computer keyboard manufacturer, charging that her repetitive stress injuries were caused by the keyboard (Genessy v. Digital Equipment Corp.). The injury awarded about \$3.5 million for pain and suffering, but the court then set aside that award as being unreasonable compensation. In making this determination, the court identified a "normative" group of 27 similar cases and specified a reasonable award as one within two standard deviations of the mean of the awards in the 27 cases. The 27 awards were (in \$1000s)

37, 60, 75, 115, 135, 140, 149, 150, 238, 290, 340, 410, 600, 750, 750, 750, 1050, 1100, 1139, 1150, 1200, 1200, 1250, 1576, 1700, 1825, 2000

from which $\sum xi = 20{,}179$ and $\sum x2i = 24{,}657{,}511$. What is the maximum possible amount that could be awarded under the two standard deviation rule

```
cases = c(37, 60, 75, 115, 135, 140, 149, 150, 238, 290, 340, 410, 600, 750, 750, 750, 1050, 1100, 1139
(2.0*stdev(cases)+mean(cases))*1000
```

```
## [1] 1961158
```

The maximum amount that can be awarded is \$1,961,158

6. Without doing any computations, which of the following data sets has the smallest standard deviation (sd), and which has the largest sd? Explain; your answer should include terms such as "center" and "spread", and how far are many / most observations from the mean of the data. Note that each data set has 10 observations. Draw a dot plot (by hand is fine) for each and use it to help explain how you know your answer is correct (without ever finding the value of the sd's)

A: {2,2,2,2,2,20,20,20,20,20}

B: {2,11,11,11,11,11,11,11,11,20}

C: {2,4,6,8,10,12,14,16,18,20}

A has the largest standard deviation because its values are spread far away from the mean. B has the lowest standard deviation because most of the values are close to the mean.