Homework 1: Descriptive Statistics and Intro to R

Caitlin Jagla

Monday Feb 28, 2022

Exercise 1 – Mean v. Median

Suppose we have the following measurements for the weights (in grams) of 10wk old male mice:

`27.3  18.6  23.4  22.8  19.5  28.3`

A. Create a vector called `mice` that contains these values. Call `mice` to print the vector

```
mice <- c(27.3, 18.6, 23.4, 22.8, 19.5, 28.3)

mice
```

```
## [1] 27.3 18.6 23.4 22.8 19.5 28.3
```

B. Use the `summary()` function to see the 6 number summary of the `mice` vector

```
summary(mice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.60   20.32   23.10   23.32   26.32   28.30
```

C. Suppose we had an additional male mouse who weighs 39.3 g. Add this observation to the `mice` vector and re-save it as `mice2`

```
mice2 <- c(mice, 39.3)
```

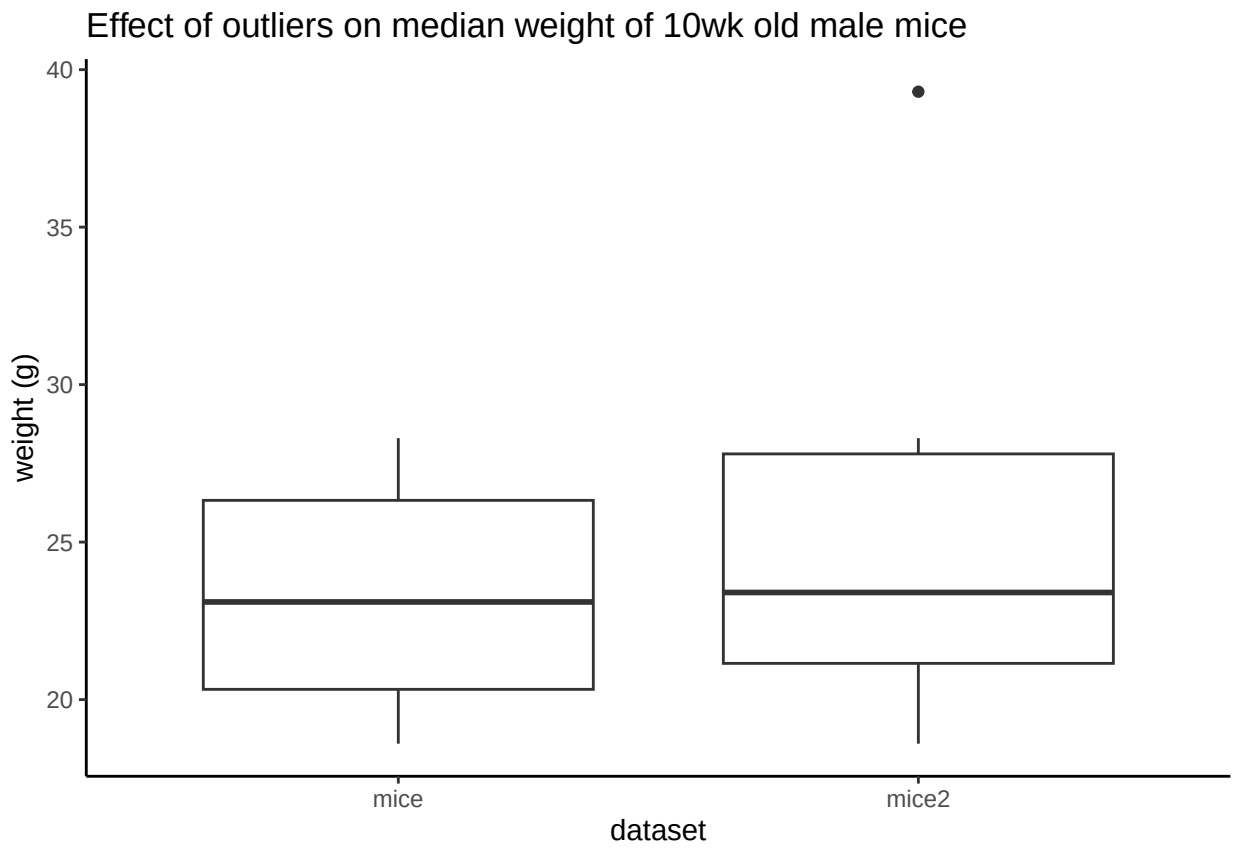
D. Use the `summary()` function to see the 6 number summary of the `mice2` vector

```
summary(mice2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.60   21.15   23.40   25.60   27.80   39.30
```

E. Use a graphing function to create a boxplot showing `mice`. Do the same for `mice2`.

```r
rbind(enframe(mice, value = "weight", name = "dataset") |> mutate(dataset = "mice"),
      enframe(mice2, value = "weight", name = "dataset") |> mutate(dataset = "mice2")) |>
  ggplot(aes(y=weight, x=dataset)) +
  geom_boxplot() +
  labs(title = "Effect of outliers on median weight of 10wk old male mice",
       y = "weight (g)") +
  theme_classic()
```



F. Which statistic (mean or median) do you think better represents this new sample of 7 mice? Why?

```
Median better represents the new sample of 7 mice, because the additional
datapoint is an outlier which skews the mean more than it affects the median.
```

G. If we add 25 to each of the 7 observations, what will happen to the mean and what will happen to the standard deviation? Optionally, you may do it to see what happens. Why is this is the case?

```r
mice3 <- mice2 + 25
```

```r
mean(mice2)
```

```
## [1] 25.6
```

```r
sd(mice2)
```

```
## [1] 7.03278
```

```r
mean(mice3)
```

```
## [1] 50.6
```

```r
sd(mice3)
```

```
## [1] 7.03278
```

The mean will increase by 25, because you have shifted the entire dataset higher by 25.
The standard deviation will not change when you add 25 to every observation,
because it describes variation between the observations and the mean,
rather than the magnitude of the observations themselves.

Exercise 2 – Measures of spread

Nine men were measured for testosterone levels with the following values (in ng/dL):

634 521 616 784 542 705 810 597 623

```
tlvl <- c(634, 521, 616, 784, 542, 705, 810, 597, 623)
```

A. Calculate the Sum of Squares (SS). *Note: there is no built in function for SS.* Look at the formula in the notes

```
ss <- sum((tlvl - mean(tlvl))^2)
ss
```

```
## [1] 79800
```

B. Calculate the variance (s^2).

```
var(tlvl)
```

```
## [1] 9975
```

C. Calculate the standard deviation (s).

```
sd(tlvl)
```

```
## [1] 99.87492
```

D. Add an additional testosterone value of 950 to the original vector and save the result to a new object.

```
tlvl2 <- c(tlvl, 950)
```

E. Calculate the standard deviation (s) of the new vector.

```
sd(tlvl2)
```

```
## [1] 134.1159
```

F. What effect did adding the observation 950 ng/dL have on the standard deviation? Why?

```
# manually calculate original dataset upper fence
summary(tlvl)[5][[1]] + # extract 3rd quartile from summary() output for original dataset
  IQR(tlvl)*1.5
```

```
## [1] 867
```

```r
range(tlvl) # get original range
```

```
## [1] 521 810
```

Standard deviation increased by a substantial amount (34.24095) because
950 ng/dL is well outside the previous range of observations (521 - 850)
and even above the upper fence of the original dataset.

Exercise 3 – Descriptive Statistics in R

For this problem we will load a dataset about GDP per capita and life expectancy in various countries.

A. Using `read_csv()`, load the `gapminder.csv` dataset. Try adding the `message=FALSE` option to this R chunk too. Then use **two functions** of your choice to investigate the dataset.

```
gm <- read_csv("gapminder.csv")

glimpse(gm)
```

```
## Rows: 76
## Columns: 5
## $ country     <chr> "Albania", "Argentina", "Australia", "Austria", "Bahrain",~
## $ continent   <chr> "Europe", "Americas", "Oceania", "Europe", "Asia", "Europe~
## $ lifeexp     <dbl> 76.423, 75.320, 81.235, 79.829, 75.635, 79.441, 65.554, 74~
## $ population  <dbl> 3600523, 40301927, 20434176, 8199783, 708573, 10392226, 91~
## $ gdp_per_cap <dbl> 5937, 12779, 34435, 36126, 29796, 33693, 3822, 7446, 9066,~
```

```
summary(gm)
```

```
##    country           continent           lifeexp        population
##  Length:76          Length:76          Min.   :43.49   Min.   :   199579
##  Class :character   Class :character   1st Qu.:72.35   1st Qu.:  4403455
##  Mode  :character   Mode  :character   Median :75.55   Median :  9637865
##                                        Mean   :74.49   Mean   : 23288718
##                                        3rd Qu.:79.34   3rd Qu.: 27869468
##                                        Max.   :82.60   Max.   :190010647
##   gdp_per_cap
##  Min.   :  470
##  1st Qu.: 6974
##  Median :11468
##  Mean   :17425
##  3rd Qu.:29065
##  Max.   :49357
```

B. What type of variable is `country`? What type of variable is `gdp_per_cap`? (Multiple functions can be used to tell you this)

```
gm |>
  pull(country) |> # select column (variable) to be queried
  type_sum()        # check variable type
```

```
## [1] "chr"
```

```
gm |>
  pull(gdp_per_cap) |>
  type_sum()
```

```
## [1] "dbl"
```

```r
str(gm$gdp_per_cap) # confirm that `gdp_per_cap` is a numeric vector by checking structure
```

```
##  num [1:76] 5937 12779 34435 36126 29796 ...
```

‘country‘ is a character variable.
‘gdp_per_cap‘ is a double (numeric) variable.

C. Calculate the median life expectancy (`lifeexp`) across the whole dataset

```r
gm |> pull(lifeexp) |> median()
```

```
## [1] 75.55
```

D. What is the range of gdp per capita (`gdp_per_cap`) across the whole dataset?

```r
gm |> pull(gdp_per_cap) |> range()
```

```
## [1]   470 49357
```

E. What is the mean and standard deviation for population?

```r
gm |>
  summarize(mean = mean(population),
            stdev = sd(population))
```

```
## # A tibble: 1 x 2
##        mean     stdev
##       <dbl>     <dbl>
## 1 23288718. 33209649.
```

Exercise 4 – Practice with dplyr and ggplot

A. How many distinct countries (`country`) are there in each continent (`continent`)?

```
gm |>
  group_by(continent) |>
  summarize(n_unique = length(unique(country)))
```

```
## # A tibble: 5 x 2
##   continent n_unique
##   <chr>        <int>
## 1 Africa           8
## 2 Americas        23
## 3 Asia            14
## 4 Europe          29
## 5 Oceania          2
```

B. Show the names of the 8 distinct countries in Africa.

```
gm |>
  filter(continent == "Africa") |>
  distinct(country)
```

```
## # A tibble: 8 x 1
##   country
##   <chr>
## 1 Egypt
## 2 Mauritius
## 3 Morocco
## 4 Reunion
## 5 Sao Tome and Principe
## 6 South Africa
## 7 Tunisia
## 8 Zimbabwe
```

C. For each `continent` calculate the mean of `lifeexp`, the median `lifeexp`, and the standard deviation of `lifeexp`. Put the output in order by median `lifeexp`.

```
gm |>
  group_by(continent) |>
  summarize(mean = mean(lifeexp),
            median = median(lifeexp),
            stdev = sd(lifeexp)) |>
  arrange(median)
```

```
## # A tibble: 5 x 4
##   continent  mean median  stdev
##   <chr>     <dbl>  <dbl>  <dbl>
## 1 Africa     65.5   71.3 12.3
## 2 Americas   73.4   72.9  4.53
## 3 Asia       73.8   73.5  5.93
## 4 Europe     77.8   78.9  2.98
## 5 Oceania    80.7   80.7  0.729
```

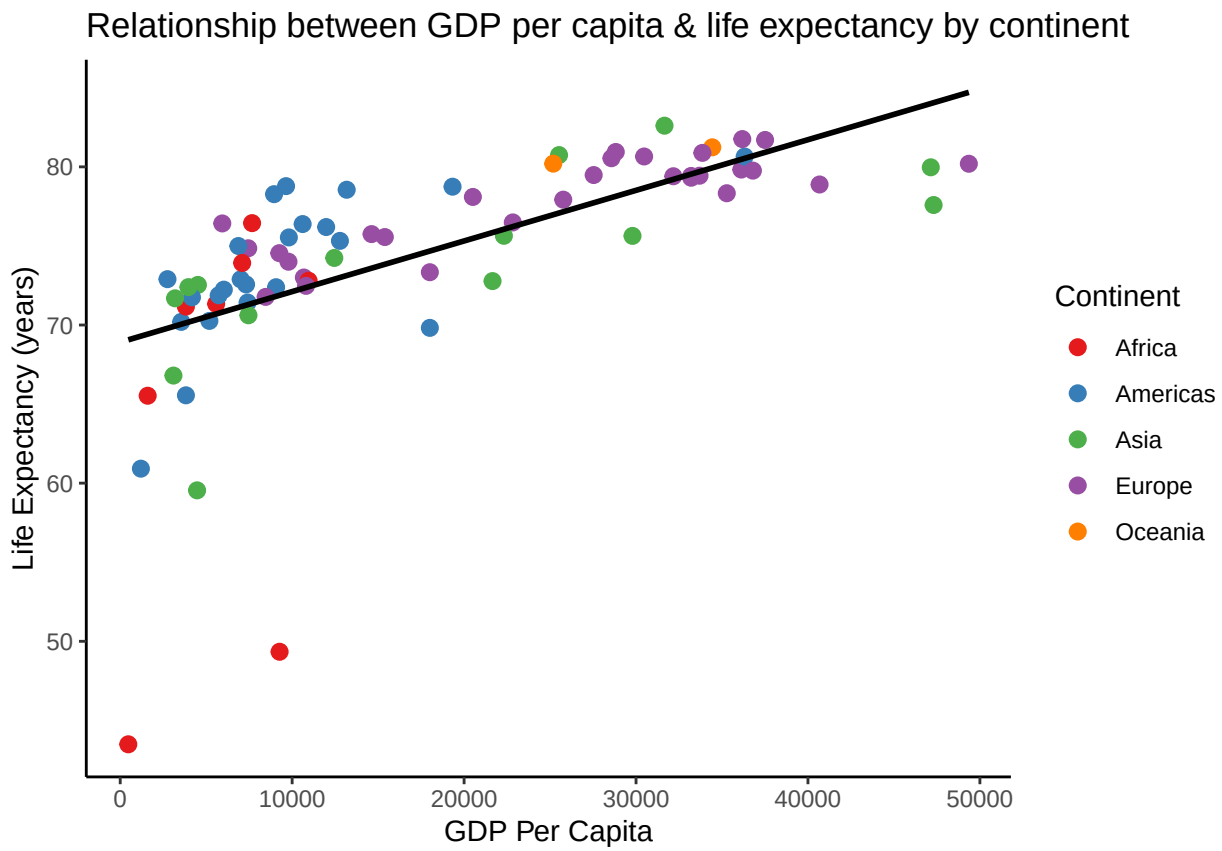
D. Which European nation had the lowest `gdp_per_capita`? Show only the `country` and `gdp_per_capita` columns.

```
gm |>
  filter(continent == "Europe") |>
  filter(gdp_per_cap == min(gdp_per_cap)) |>
  select(country, gdp_per_cap)
```

```
## # A tibble: 1 x 2
##   country gdp_per_cap
##   <chr>         <dbl>
## 1 Albania        5937
```

E. Create a scatter plot using `ggplot()` to look at the relationship between life expectancy and gdp per capita. Color each of the points by continent.

```
gm |>
  ggplot(aes(x = gdp_per_cap, y = lifeexp)) +
  geom_point(aes(color = continent), size = 2.5) +
  geom_smooth(method = "lm", color = "black", se = FALSE) +
  labs(title = "Relationship between GDP per capita & life expectancy by continent",
       x = "GDP Per Capita",
       y = "Life Expectancy (years)",
       color = "Continent") +
  scale_color_brewer(type = "qual", palette = "Set1") +
  theme_classic()
```



F. Create a boxplot using `ggplot()` that looks at the relationship between continent and population. Use `filter()` to remove Oceania from the graph.

```
gm |>
  filter(continent != "Oceania") |>
  ggplot(aes(x = continent, y = population, fill = continent)) +
  geom_boxplot(show.legend = FALSE) +
  labs(title = "Country Populations By Continent",
       x = "Continent",
       y = "Country Population") +
  scale_fill_brewer(type = "qual", palette = "Set1") +
  theme_classic()
```