Ch 4 Key

Jacob C. Cooper

2024-10-07

Overview

For the homework this week, we were looking at *descriptive statistics* over large datasets. These help us get an idea of what the data are like. We required two packages to perform these analyses:

```r
# allows for internet downloading
library(curl)
```

```
## Using libcurl 8.7.1 with LibreSSL/3.3.6
```

```r
# enables data management tools
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
## x readr::parse_date() masks curl::parse_date()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Answer Key: Descriptive Statistics

For the homework, you were to create an *RMarkdown* document rendering as an `html` file. For the homework, we need to find the following values. Here, I indicate:

- mean
  - `mean`
- median
  - `median` or `summary`
- range
  - distance between `min` and `max` values
- interquartile range
  - we can use `quantile` to find the 25th and 75th percentiles
- variance
  - `var`

- standard deviation
    - sd
- coefficient of variation
    - We can create a custom code to calculate this. Remember, $cv = \frac{\sigma}{\mu}$.

```
cv <- function(x){
  sigma <- sd(x)
  mu <- mean(x)
  val <- sigma/mu
  return(val)
}
```

- standard error
    - We can create a custom code to calculate this. Remember, $se = \frac{\sigma}{\sqrt{n}}$.

```
se <- function(x){
  n <- length(x) # calculate n
  s <- sd(x) # calculate standard deviation
  se_val <- s/sqrt(n)
  return(se_val)
}
```

- whether there are any “outliers”
    - Values that are 1.5x the IQR and beyond

Keymaker I have written a custom code to produce all of the above values, which I called `key_maker`.

```
key_maker <- function(x, # specify data
                      roundval){ # specify rounding!

  # calcualte and print mean value
  print(paste0("Mean: ", round(mean(x),roundval)))

  # calcualte and print median value
  print(paste0("Median: ", round(median(x),roundval)))

  # calculate and print range
  # distance between min and max
  print(paste0("Range: ", round(max(x) - min(x),roundval)))

  # get IQRs, use method of book!
  quantiles_x <- quantile(x, type = 6) %>% as.numeric()
  # get 25th and 75th quantiles
  print(paste0("IQR: ",round(quantiles_x[4] - quantiles_x[2],roundval)))

  # calculate variance
  print(paste0("Variance: ",round(var(x),roundval)))

  # calculate standard deviation
  print(paste0("SD: ", round(sd(x),roundval)))

  # calculate coefficient of variation
  # use previously defined function
```

```r
  print(paste0("CV: ",round(cv(x),roundval)))

  # calculate standard error
  # use previously defined function
  print(paste0("SE: ", round(se(x),roundval)))

  # calculate 25th quantile
  lowquant <- quantile(x,0.25,type = 6) %>%
    as.numeric() # convert to numeric

  # calculate 75th quantile
  hiquant <- quantile(x,0.75,type = 6) %>%
    as.numeric() # convert to numeric

  # get the IQR
  iqr <- hiquant - lowquant

  # calculate 1.5*IQR
  lowbound <- lowquant - (1.5*iqr)
  hibound <- hiquant + (1.5*iqr)

  # low outliers?
  # select elements that match
  # identify using logical "which"
  lows <- sum(x < lowbound)
  if(lows > 0){
    low.vals <- x[which(x < lowbound)]
    print("Low outlier(s): ")
    print(paste0(length(low.vals)," values found."))
    # only print if 75 or fewer vals
    if(lows <= 75){print(low.vals)}
  }else{print("No low outliers.")}

  # same procedure for high outliers
  highs <- sum(x > hibound)
  if(highs > 0){
    hi.vals <- x[which(x > hibound)]
    print("High outlier(s):")
    print(paste0(length(hi.vals)," values found."))
    # only print if 75 or fewer vals
    if(highs <= 75){print(hi.vals)}
  }else{print("No high outliers.")}
}
```

1: UNK Nebraskans Ever year, the university keeps track of where students are from. The following are data on the number of students admitted to UNK from the state of Nebraska:

```r
# create dataset in R
nebraskans <- c(5056,5061,5276,5244,5209,
                5262,5466,5606,5508,5540,5614)

years <- 2023:2013

nebraskans_years <- cbind(years,nebraskans) %>%
  as.data.frame()

nebraskans_years
```

```
##    years nebraskans
## 1   2023       5056
## 2   2022       5061
## 3   2021       5276
## 4   2020       5244
## 5   2019       5209
## 6   2018       5262
## 7   2017       5466
## 8   2016       5606
## 9   2015       5508
## 10  2014       5540
## 11  2013       5614
```

Using these data, please calculate the mean, median, range, interquartile range, variance, standard deviation, coefficient of variation, standard error, and whether there are any "outliers" for the number of UNK students from Nebraska.

```
key_maker(nebraskans_years$nebraskans, # data
          0) # rounding
```

```
## [1] "Mean: 5349"
## [1] "Median: 5276"
## [1] "Range: 558"
## [1] "IQR: 331"
## [1] "Variance: 42404"
## [1] "SD: 206"
## [1] "CV: 0"
## [1] "SE: 62"
## [1] "No low outliers."
## [1] "No high outliers."
```

Synthesis question: Do you think that there are any really anomalous years? Do you feel data are similar between years? *Note* we are not looking at trends through time but whether any years are outliers.

> Data appear to be similar between years. If looking directly at dataset, values are decreasing through time, but there are no major anomalies. **NOTE** values may seem weird because of rounding!

2: Piracy in the Gulf of Guinea The following is a dataset looking at oceanic conditions and other variables associated with pirate attacks within the region between 2010 and 2021 [@Moura2023]. Using these data, please calculate the mean, median, range, interquartile range, variance, standard deviation, coefficient of variation, standard error, and whether there are any "outliers" for distance from shore for each pirate attack (columns Distance_from_Coast).

```
url_file <- curl("https://figshare.com/ndownloader/files/42314604")
pirates <- url_file %>% read_csv()
```

```
## New names:
## Rows: 595 Columns: 40
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (18): Period, Season, African_Season, Coastal_State, Coastal_Zone, Navi... dbl
## (20): ...1, Unnamed: 0, Year, Month, Lat_D, Lon_D, Distance_from_Coast,... dttm
## (2): Date_Time, Date
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
key_maker(pirates$Distance_from_Coast,
          2) # round to two decimal places!
```

```
## [1] "Mean: 33.58"
## [1] "Median: 13"
## [1] "Range: 680"
## [1] "IQR: 43.5"
## [1] "Variance: 2502.82"
## [1] "SD: 50.03"
## [1] "CV: 1.49"
## [1] "SE: 2.05"
## [1] "No low outliers."
## [1] "High outlier(s):"
## [1] "33 values found."
##  [1] 136 120 174 205 140 680 220 257 180 130 142 195 116 200 120 135 200 208 132
## [20] 208 155 120 149 136 118 127 150 196 212 157 133 116 124
```

We have no low outliers but 50 high outliers. We have a very high variance and range.

Synthesis question: Do you notice any patterns in distance from shore? What may be responsible for these patterns? *Hint*: Think about what piracy entails and also what other columns are available as other variables in the above dataset.

> Most attacks are close to shore. We have no low bounds because we are very close to shore, and high bounds are the very distant attacks.

3: Patient ages at presentation The following is a dataset on skin sores in different communities in Australia and Oceania, specifically looking at the amount of time that passes between skin infections [@lydeamore_estimation_2020]. This file includes multiple different datasets, and focuses on data from children in the first five years of their life, on househould visits, and on data collected during targeted studies [@Lydeamore2020].

```
url_file <- curl("https://doi.org/10.1371/journal.pcbi.1007838.s006")

ages <- url_file %>% read_csv()
```

```
## Rows: 17150 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): dataset
## dbl (1): time_difference
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Let's see what this file is like real fast. We can use the command `dim` to see the `rows` and `columns`.

```
dim(ages)
```

```
## [1] 17150     2
```

As you can see, this file has only two columns but 17,150 rows! For the column `time_difference`, please calculate the mean, median, range, interquartile range, variance, standard deviation, coefficient of variation, standard error, and whether there are any “outliers”.

```r
key_maker(ages$time_difference,7)
```

```
## [1] "Mean: 32.824105"
## [1] "Median: 16.988195"
## [1] "Range: 718.000995"
## [1] "IQR: 32.0026956"
## [1] "Variance: 2526.193286"
## [1] "SD: 50.2612503"
## [1] "CV: 1.5312299"
## [1] "SE: 0.3837967"
## [1] "No low outliers."
## [1] "High outlier(s):"
## [1] "1651 values found."
```

Synthesis question: Folks will often think about probabilities of events being “low but never zero”. What does that mean in the context of these data? What about these data make you feel like probabilities may decrease through time but never become zero?

> There is always a change of reoccurrence, as shown by the massive variance and range. However, reoccurrence almost always happens on a shorter time scale. We can never say with complete certainty that the chance of something is 0%.