Chapter 2 Homework

Kaylynn Hiller

2026-01-26

Setup

Assume you are working in an RStudio Project. Use `here::here("data", "filename.csv")` to build file paths to datasets in the `data` folder.

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.0     v tibble    3.3.1
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
here() starts at /Users/kaylynnhiller/Desktop/Stats1


Attaching package: 'psych'


The following objects are masked from 'package:ggplot2':

    %+%, alpha


Rows: 110 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): Pulse1, Pulse2
dbl (9): Height, Weight, Age, Gender, Smokes, Alcohol, Exercise, Ran, Year
```

```
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.
```

> **Tip**
>
> If you ever get a "file not found" error, confirm (1) you opened the correct `.Rproj`, and (2) the dataset is inside your project's data folder (accessed via `here::here("data", ...)`).

Tutorial (Chapter 2, plus a bit of 3)

1) Estimates of variation around the mean

In class, we talked about quantifying variation. Variance and standard deviation are both based on the **sum of squared error (SSE)** around a model's predictions.

Start by computing the sample standard deviation and sample variance of height:

```
[1] TRUE
```

```
[1] 16.07687
```

```
[1] 258.4657
```

Convert variance to standard deviation

Recall: $\text{SD} = \sqrt{\text{Var}}$.

```
[1] TRUE
```

Compute SSE from the variance and sample size

For the *sample* variance,

$$s^2 = \frac{\text{SSE}}{n-1} \quad \Rightarrow \quad \text{SSE} = s^2(n-1).$$

```
     count
[1,]  TRUE
```

```
     count
1 28172.76
```

> **Tip**
>
> In R, `var(x)` uses the sample variance with denominator $n - 1$. That's why the formula above uses $n - 1$.

2) Estimates of central tendency

Compute mean and median:

```
  mean_height median_height
1     171.582         172.5
```

Mode (custom helper)

R's base `mode()` is **not** the statistical mode. Below is a simple helper that returns the most frequent value(s). If the data are multimodal, it returns the average of the modes.

3) Group and summarize a dataset

Compute mean and SD of height by gender using tidyverse verbs:

```
# A tibble: 2 x 4
  gender count mean_height sd_height
   <dbl> <int>       <dbl>     <dbl>
1      1    59        178.      16.3
2      2    51        164.      12.7
```

If you want a richer descriptive table, `psych::describeBy()` can do that:

```
    item group1 vars  n     mean       sd median  trimmed   mad min max range
X11    1      1    1 59 177.7119 16.27358    180 179.1837 7.413  68 195   127
X12    2      2    1 51 164.4902 12.67339    165 166.0000 7.413  93 180    87
         skew kurtosis       se
X11 -5.109844 32.12679 2.118639
X12 -3.540802 17.34468 1.774630
```

> **Tip**
>
> Even when you use a helper like `describeBy()`, prefer doing **wrangling** (filtering, selecting, mutating, grouping) with tidyverse verbs first, then pass the result to the summary function.

4) Estimates of variation revisited: SSE from the data

One way to compute SSE around the **mean model** (predict $\bar{Y}$ for everyone):

```
[1] 28172.76
```

Now modify that idea to compute SSE around the **median model** (predict $\tilde{Y}$ for everyone).

```
[1] 28265.5
```

Answer in words: is SSE around the median larger or smaller than SSE around the mean for these data?

The SSE around the median is larger that the SSE around the mean.

5) Missing data

Many functions accept `na.rm = TRUE` to ignore missing values.

You can also drop missing values explicitly:

```
# A tibble: 1 x 1
  mean_height
        <dbl>
1        172.
```

HW 02 Questions

1) Central tendency and variability (USNEWS)

The dataset `USNEWS.csv` contains data used by *U.S. News and World Report* to make its college rankings. Two variables of interest are:

- `gradRate`: graduation rate (percent from 0 to 100)
- `accptRate`: acceptance rate (proportion from 0 to 1)

For each variable, obtain estimates of central tendency (mean, median, mode) and variability (variance and standard deviation). Write a few sentences describing what you found.

The mean of the graduation rate is relatively close to the median. There is a large amount of standard deviation in the graduation rate as the deviation is 18.8. The mean of the acceptance rate is relatively farther from the median. There is a large amount of standard deviation in the acceptance rate as well as the deviation is 0.161.

```
Rows: 1302 Columns: 37
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (35): College, State, Type, satMath, satVerbal, satTotal, act, math1q, m...
dbl  (2): FICE, typeCode

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.

Warning: There were 2 warnings in `mutate()`.
The first warning was:
i In argument: `gradRate = as.numeric(gradRate)`.
Caused by warning:
! NAs introduced by coercion
i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.

# A tibble: 1 x 6
  mean_gradRate median_gradRate mode_gradRate var_gradRate sd_gradRate count
          <dbl>           <dbl>         <dbl>        <dbl>       <dbl> <int>
1            NA              NA            57           NA          NA  1302

# A tibble: 1 x 6
  mean_accptRate median_accptRate mode_accptRate var_accptRate sd_accptRate
           <dbl>            <dbl>          <dbl>         <dbl>        <dbl>
1             NA               NA              1            NA           NA
# i 1 more variable: count <int>
```

2) Conditional vs unconditional predictions (USNEWS)

Another interesting variable is `Type` (public vs private).

1. Write **Model C** that makes a constant prediction for every school.

```
# A tibble: 1 x 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)     60.4     0.544      111.       0
```

2. Write **Model A** that makes predictions of `gradRate` conditional on `Type`.

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     50.2     0.829      60.6 0
2 Private         16.0     1.04       15.4 4.06e-49
```

3. As a first look at whether Model A might be useful, compute the mean `gradRate` for private and public schools.
4. Write a sentence or two describing these results and whether it *appears* useful to move from Model C to Model A.

Model A appears more useful than Model C. The graduation rate for private schools is 16% higher compared to public schools. This large difference means control for the type of school would likely reduce error in the model.

Use both a verbal description and a model statement in LaTeX.

Model statements (fill in)

- Model C (compact):

$$gradRate_i = b_0 + \varepsilon_i$$

- Model A (augmented; conditional means by type):

$$gradRate_i = b_0 + b_1 X_i + \varepsilon_i$$

  where X_i is an indicator you define (e.g., $X_i = 1$ if Private, 0 if Public).

```
# A tibble: 2 x 3
  Type    count mean_gradRate
  <chr>   <int>         <dbl>
1 Private   770          66.2
2 Public    434          50.2
```

3) Coupon campaign and store sales (stores)

At 10 grocery stores, a market researcher records the number of cases of a product sold both before and after coupons were mailed to households in the area. The data below are the **changes** in the number of cases sold (positive = more after coupons, negative = fewer, zero = no change).

Store	Change
A	5
B	4
C	-2
D	6
E	1
F	0
G	-4
H	3
I	2
J	7

3a) By hand

Calculate the **mean** and **standard deviation** for the change scores by hand (or in Excel).

mean = 2.2

standard deviation = 3.521363

(Optional check in R after you finish your by-hand work:)

```
# A tibble: 10 x 2
   store change
   <chr>  <dbl>
 1 a          5
 2 b          4
 3 c         -2
 4 d          6
 5 e          1
 6 f          0
 7 g         -4
 8 h          3
 9 i          2
10 j          7
```

```
# A tibble: 10 x 4
   store change mean_change sd_change
   <chr>  <dbl>       <dbl>     <dbl>
 1 a          5         2.2      3.52
 2 b          4         2.2      3.52
 3 c         -2         2.2      3.52
 4 d          6         2.2      3.52
 5 e          1         2.2      3.52
 6 f          0         2.2      3.52
 7 g         -4         2.2      3.52
 8 h          3         2.2      3.52
 9 i          2         2.2      3.52
10 j          7         2.2      3.52
```

3b) Model C (no change)

Specify a Model C that predicts **no change** for every store.

This model predicts that the change in cases sold is 0 for stores.

Write it in the same form as class:

$$\Delta_i = 0 + \varepsilon_i$$

3c) Model A (best constant prediction from the data)

This model predicts that the change in cases sold is equal to the average number of cases sold in this sample

Specify a Model A with one parameter that makes the best possible constant prediction of `Change` based on the data.

$$\Delta_i = b_0 + \varepsilon_i$$

3d) Null hypothesis and interpretation

State the null hypothesis tested by comparing Model C and Model A, and give a non-technical interpretation of what the comparison asks.

The null hypothesis assumes there is no change in the number of cases sold when there are coupons. We are testing to see if the change in the number of cases sold is different from 0 (or there is change compared to no change).

3e) Compute error for both models (SSE)

Using SSE as your aggregate measure of error, compute:

- SSE(C) for the compact model
- SSE(A) for the augmented model

```
[1] 160
```

```
[1] 111.6
```

3f) Proportional reduction in error (PRE)

Compute:

$$\text{PRE} = \frac{\text{SSE}(C) - \text{SSE}(A)}{\text{SSE}(C)}.$$

Based on this PRE, do you think Model C should be rejected in favor of Model A? (No formal test yet—explain your reasoning.)

We should reject Model C in favor of Model A. While the sample size here is small, the PRE is .3025. This is a relatively large reduction in error for the model.

```
[1] 0.3025
```

3g) Do it again using `stores.csv`

These data are also available in `stores.csv`. Read in the data and use R to obtain all the numbers you calculated above (including PRE if you can).

```
Rows: 10 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): Store
dbl (1): Change

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] 1062
```

```
[1] 572
```

```
[1] 0.4613936
```

> **Tip**
>
> A convenient SSE pattern is:
>
> - If predictions are stored in `y_hat`, then `sum((y - y_hat)^2)`.
> - For Model C here, `y_hat` is a constant 0.
> - For Model A here, `y_hat` is a constant equal to `mean(y)`.

4) Concept questions

In your own words (**not** using the example from class):

1. Why is $\text{ERROR}(\text{Model A}) \leq \text{ERROR}(\text{Model C})$?

The error for model A is always less than or equal to the error of model C because conditional models always have less error than unconditional models. Any information, such as the mean, allows us to make a better prediction than no information where we have to randomly guess.

1. What is a **degree of freedom**?

A degree of freedom is the number of independent data points in a sample that can vary when estimating parameters. When you have an estimation (such as standard deviation or mean) and know all the data points besides one, the unknown data point can only be one number which produces the known estimation.

5) Sampling distributions and small samples

Visit the app:

- https://correll.shinyapps.io/centralTendency/

5a) Generate four or five iterations using a normally distributed population and samples of $n = 5$.

- Where do the sampling distributions of the mean peak?
  - The sampling distributions of the mean peak around 10 with slight variance.
- How much do the means and medians vary? What minimum and maximum values do you see for each?

  - The means and medians vary slightly. The medians vary more compared to the means. The minimum value for means is 9.94 and the maximum value is 10.22. The minimum value for the medians is 9.6 and the maximum value is 10.4.
  - Means: 10.14, 10.18, 9.94, 10.22, 10.03
  - Medians: 10.4, 9.7, 9.9, 10, 9.6

5b) Answer the same questions for the mean using $n = 500$. (Rescale the histograms so you can see variability.)

- The sampling distributions of the mean peak around 10.1 with slight variance.
- The means vary slightly while the medians had no variation. The minimum value for means is 10.07 and the maximum value is 10.12. I got 10.1 for all median values.
- Means: 10.09, 10.11, 10.12, 10.07, 10.09
- Medians: 10.1, 10.1, 10.1, 10.1, 10.1

6) Central Limit Theorem intuition

Visit the app:

- https://correll.shinyapps.io/centralLimit/

6a) Set a **skewed** population and draw samples of different sizes. What is the shape of the sampling distribution of the mean for samples of $n = 2$?

The sampling distribution is very right skewed.

6b) What is the shape for $n = 10$?

This sampling is slightly right skewed compared to the previous sampling.

6c) What is the shape for $n = 100$?

This sampling is even less right skewed, but still slightly skewed.