The Great Bean Experiment!

Joe Nese and EDLD 651

9/29/2021

Introduction

Introduction!

This is the great bean experiment! An under-powered (mediocre) scientific replication of the famous historical event (folktale?) in statistics in which each person in town was asked to guess the weight of the ox. Having no knowledge of oxen, no person correctly guessed the weight, but the average of all guesses was within one pound of the ox's weight! What an inspiring story of collectivism, the strength in numbers, and the power of data, data science, and statistics! We are better together.

```
truth <- 3992

dta <- read_csv(here("nopublish", "beans", "bean_data.csv")) %>%
  mutate(id = ifelse(uo_id %% 2 == 0, "even", "odd"),
         initials = paste0(str_sub(first, 1, 1), str_sub(last, 1, 2)),
         dist = guess - truth) %>%
  select(-uo_id)

answ <- tibble(
  name = c("Class Mean", "Truth"),
  value = c(round(mean(dta$guess, na.rm = TRUE)), truth)
)
```

Method

We used the following packages: equatiomatic (Anderson, ND), ggrepel Slowikowski (2021), ggthemes Arnold (2021), here Müller (2020), knitr Xie (2014), parameters Lüdecke et al. (2020), reactable Lin (2020), and tidyverse Wickham et al. (2019).

Research Question

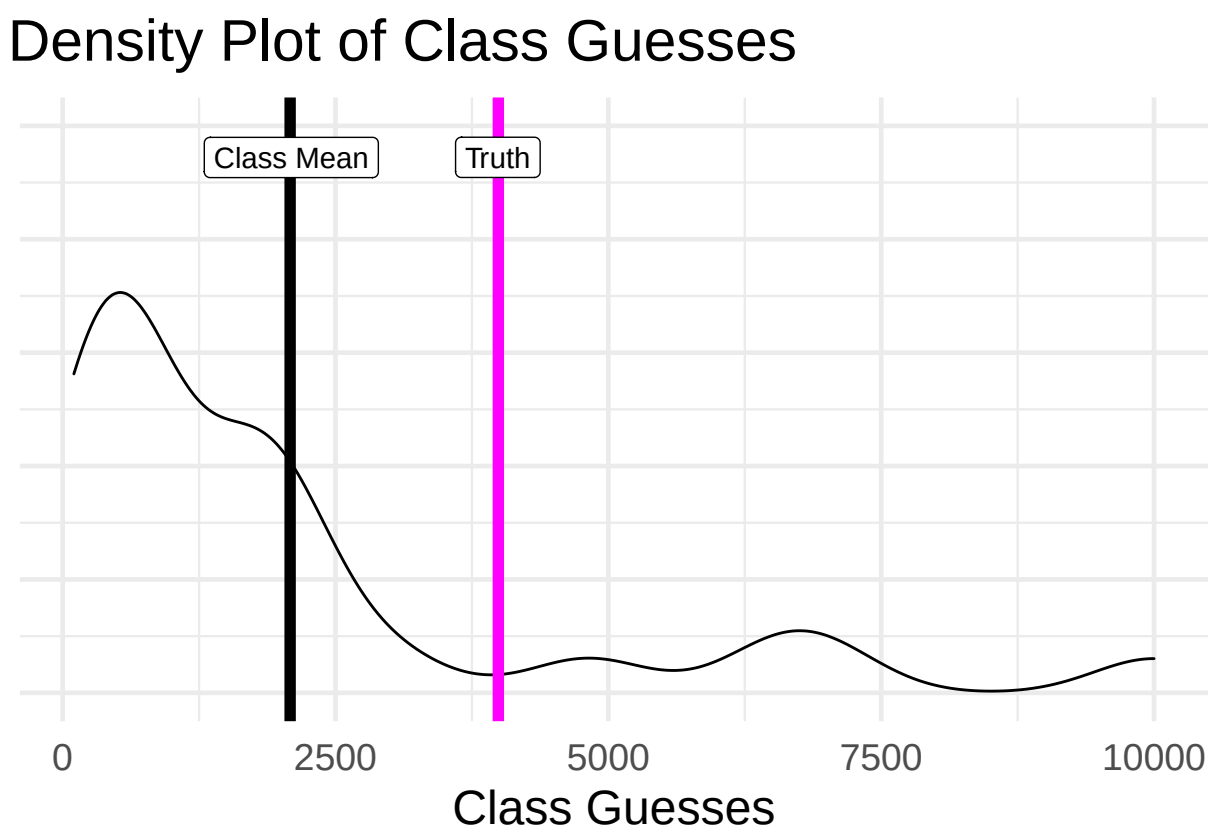
1. Is the average of the class's bean guesses closer to the actual number of beans in the jar than any one person?

Our hypothesis is that the class average will be more accurate than the guess of any one person.

Results

First, let's look at a density plot, which shows the distribution of the class guesses.

```
dta %>%
  ggplot(aes(guess)) +
  geom_density() +
  geom_vline(xintercept = truth, color = "magenta", size = 2) +
  geom_vline(xintercept = mean(dta$guess, na.rm = TRUE), size = 2) +
  ggrepel::geom_label_repel(data = answ, aes(x = value, y = .0005, label = name)) +
  theme(plot.title.position = "plot",
        axis.text.y = element_blank()) +
  labs(
    x = "Class Guesses",
    y = NULL,
    title = "Density Plot of Class Guesses"
  )
```



And here is a table of the class guesses. You can filter and sort it!

```
dta %>%
  select(first, guess) %>%
  reactable(filterable = TRUE, searchable = TRUE)
```

Search

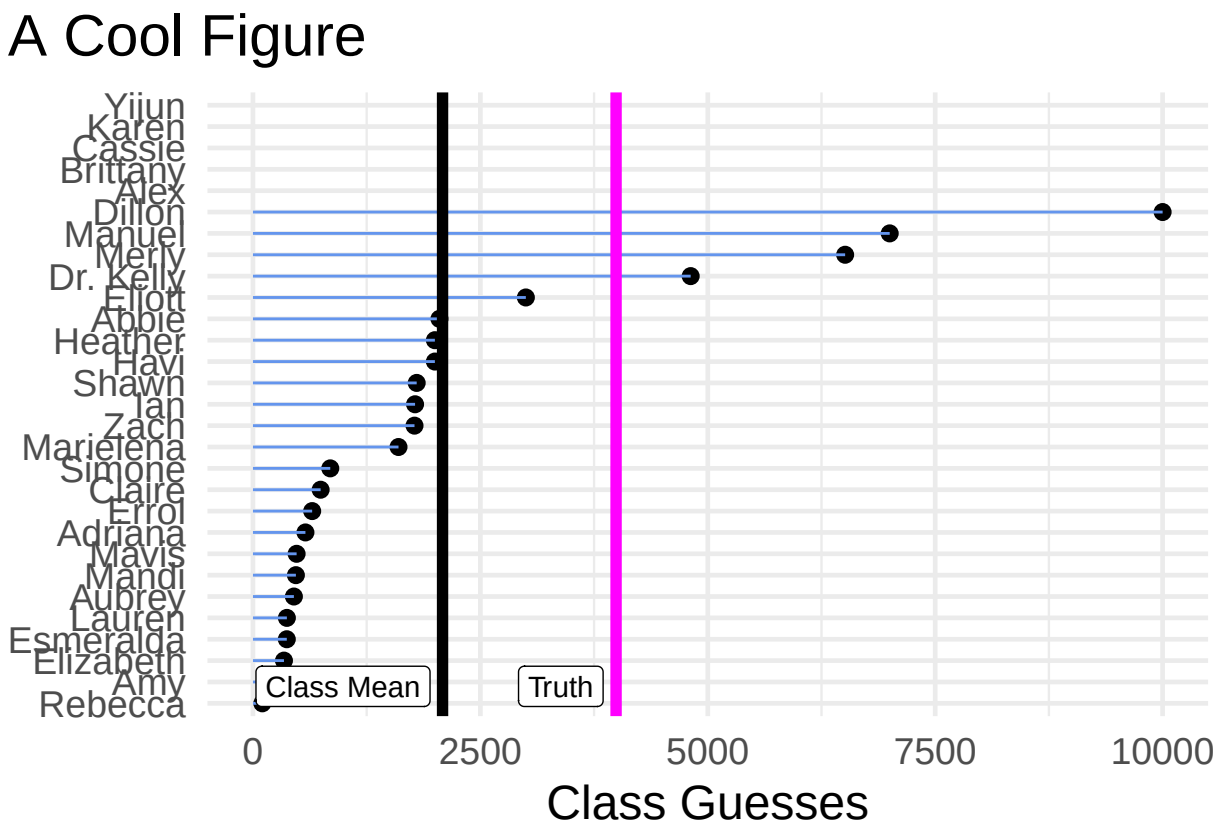
first	guess
Elizabeth	341
Lauren	373
Esmeralda	372
Yijun	
Adriana	577
Eliott	3000
Zach	1776
Mavis	480
Aubrey	450
Claire	744

1–10 of 29 rows

Previous 1 2 3 Next

Here is another cool data visualization.

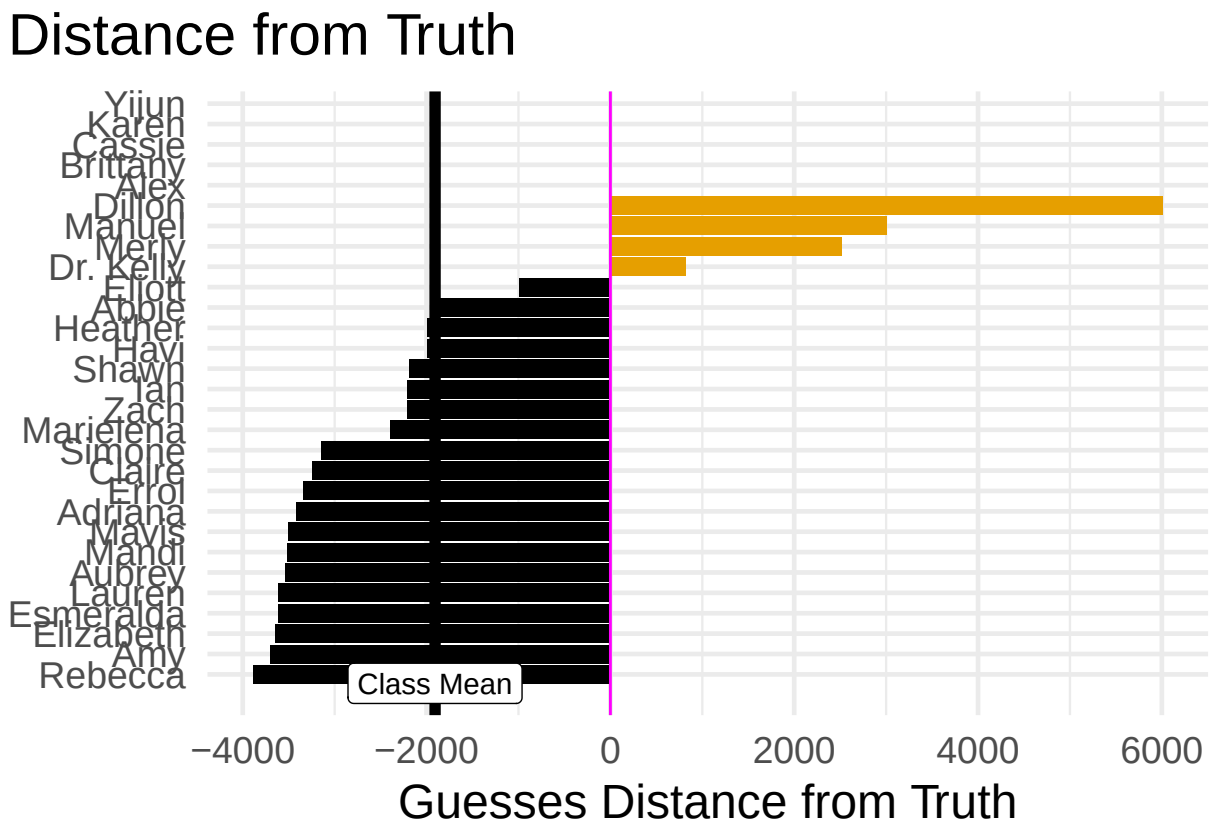
```
dta %>%
  ggplot(aes(guess, reorder(first, guess), group = 1)) +
  geom_point(size = 2.5) +
  geom_segment(aes(x = 0, xend = guess, y = first, yend = first), color = "cornflowerblue") +
  geom_vline(xintercept = truth, color = "magenta", size = 2) +
  geom_vline(xintercept = mean(dta$guess, na.rm = TRUE), size = 2) +
  geom_label_repel(data = answ, aes(x = value, y = 2, label = name)) +
  theme(plot.title.position = "plot") +
  labs(
    x = "Class Guesses",
    y = NULL,
    title = "A Cool Figure"
  )
```



Or perhaps my favorite:

```
dta %>%
  ggplot(aes(dist, reorder(first, dist))) +
  geom_col(aes(fill = ifelse(dist > 0, "plus", "minus"))) +
  geom_vline(xintercept = 0, color = "magenta", size = .5) +
  geom_vline(xintercept = mean(dta$guess, na.rm = TRUE) - truth, size = 2) +
  ggrepel::geom_label_repel(data = filter(answ, name == "Class Mean"),
                            aes(x = value - truth, y = -1, label = name),
                            min.segment.length = 0) +
  ggthemes::scale_fill_colorblind() +
  theme(legend.position="none",
```

```
      plot.title.position = "plot") +
  labs(
    x = "Guesses Distance from Truth",
    y = NULL,
    title = "Distance from Truth"
  )
```



Now let's get to the results...

The actual number of beans in the jar was **3,992**.

The person closest to the truth was **Dr. Kelly** with a guess of **4,812** beans. This is a difference of 0.01%.

The class mean was **2,084.5** (SD = 2526). This is a difference of 0.01%.

So to answer our research question...

```
if(abs(mean(dta$guess, na.rm = TRUE) - truth) < abs(slice(dta, which.min(abs(dist))) %>% pull(guess) - t
   cat("Hypothosis confirmed! The class average was more accurate than any one person!")
} else if (abs(mean(dta$guess, na.rm = TRUE) - truth) > abs(slice(dta, which.min(abs(dist))) %>% pull(gu
   cat(paste0("**Hypthosis rejected! ", slice(dta, which.min(abs(dist))) %>% pull(first), "'s guess was
} else if (abs(mean(dta$guess, na.rm = TRUE) - truth) == abs(slice(dta, which.min(abs(dist))) %>% pull(g
  cat(paste0("What?! It was a tie! ", slice(dta, which.min(abs(dist))) %>% pull(first), "'s guess was th
}
```

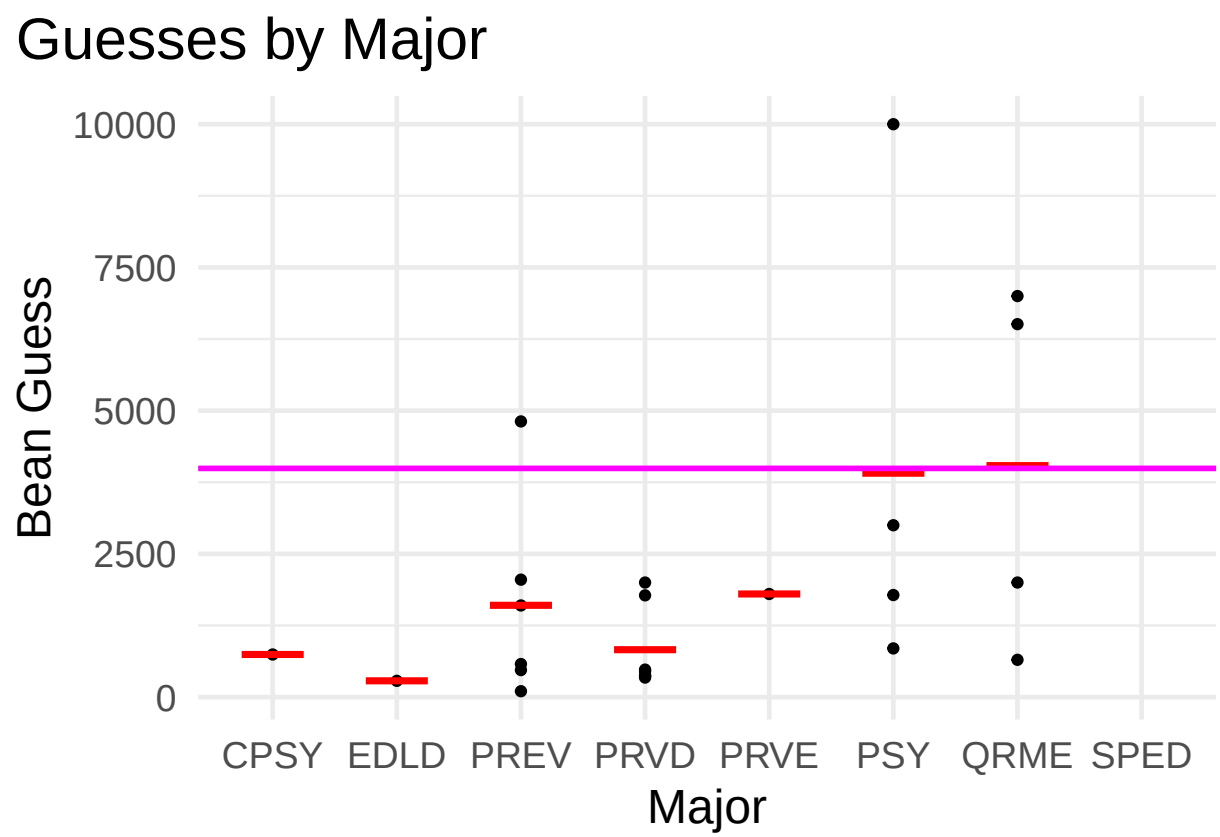
Hypthosis rejected! Dr. Kelly's guess was closer than the class average! Replication crisis?

Discussion

Wait! I'm having fun with this, so let's look at major and ID number!

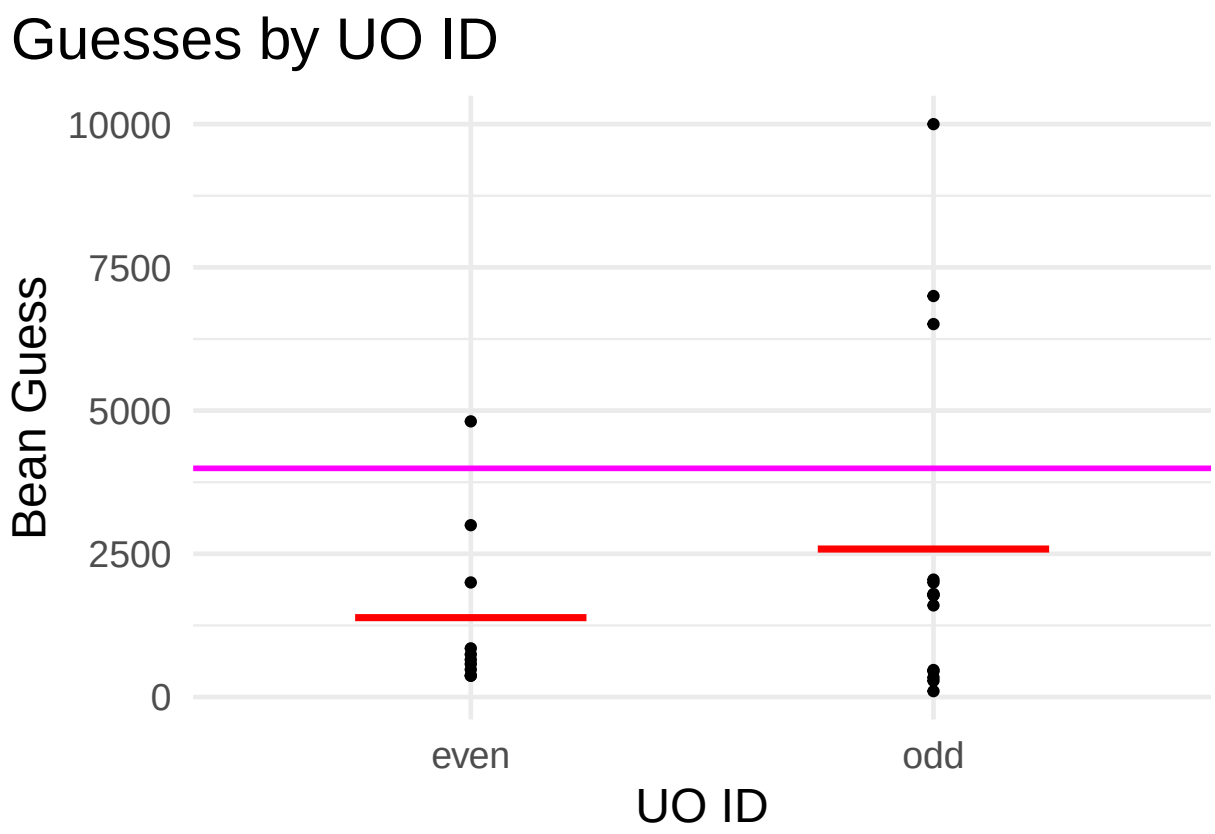
```
dta_smry <- dta %>%
  group_by(major) %>%
  summarize(mean_guess = mean(guess, na.rm = TRUE))

ggplot(dta, aes(major, guess, group = major)) +
  geom_point() +
  stat_summary(fun = "mean",
               geom = "crossbar",
               width = .5,
               color = "red") +
  geom_hline(yintercept = truth, color = "magenta", size = 1) +
  labs(
    x = "Major",
    y = "Bean Guess",
    title = "Guesses by Major"
  ) +
  theme(plot.title.position = "plot")
```




```
ggplot(dta, aes(id, guess, group = id)) +
  geom_point() +
  stat_summary(fun = "mean",
               geom = "crossbar",
               width = .5,
```

```
              color = "red") +
  geom_hline(yintercept = truth, color = "magenta", size = 1) +
  labs(
    x = "UO ID",
    y = "Bean Guess",
    title = "Guesses by UO ID"
  ) +
  theme(plot.title.position = "plot")
```



Let's run (bad) a regression! With some some help from the `{equatiomatic}` package by our own Daniel Anderson!

```
m1 <- lm(dist ~ major + id, data = dta)
```

```
extract_eq(m1, wrap = TRUE)
```

$$\begin{aligned} \text{dist} = \alpha &+ \beta_1(\text{major}_{\text{EDLD}}) + \beta_2(\text{major}_{\text{PREV}}) + \beta_3(\text{major}_{\text{PRVD}}) + \\ &\beta_4(\text{major}_{\text{PRVE}}) + \beta_5(\text{major}_{\text{PSY}}) + \beta_6(\text{major}_{\text{QRME}}) + \beta_7(\text{id}_{\text{odd}}) + \\ &\epsilon \end{aligned}$$

```
model_parameters(m1) %>%
  as_tibble() %>%
  select(-c(CI, df_error, t)) %>%
```

```
  mutate(across(c(2:5), ~round(., 1)),
         p = round(p, 3)) %>%
  kable(booktabs = TRUE,
        align = c("l", "r", "r", "r", "r", "r"),
        caption = "These are the Regression Results")
```

Table 1: These are the Regression Results

Parameter	Coefficient	SE	CI_low	CI_high	p
(Intercept)	-3248.0	2434.5	-8408.9	1912.9	0.201
majorEDLD	-1555.5	3617.9	-9225.0	6114.0	0.673
majorPREV	128.0	2732.0	-5663.5	5919.5	0.963
majorPRVD	-386.1	2645.8	-5994.9	5222.8	0.886
majorPRVE	-39.5	3617.9	-7709.0	7630.0	0.991
majorPSY	2616.3	2778.0	-3272.8	8505.4	0.360
majorQRME	2474.7	2846.6	-3559.9	8509.2	0.398
idodd	1095.5	1111.5	-1260.7	3451.7	0.339

References

Arnold, Jeffrey B. 2021. *Ggthemes: Extra Themes, Scales and Geoms for 'Ggplot2'*. https://CRAN.R-project.org/package=ggthemes.

Lin, Greg. 2020. *Reactable: Interactive Data Tables Based on 'React Table'*. https://CRAN.R-project.org/package=reactable.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, and Dominique Makowski. 2020. "Extracting, Computing and Exploring the Parameters of Statistical Models Using R." *Journal of Open Source Software* 5 (53): 2445. https://doi.org/10.21105/joss.02445.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Slowikowski, Kamil. 2021. *Ggrepel: Automatically Position Non-Overlapping Text Labels with 'Ggplot2'*. https://CRAN.R-project.org/package=ggrepel.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.