Lab 7: Inference for Numerical Data

Julia Ferris

2023-10-22

```
knitr::opts_chunk$set(eval = TRUE, message = FALSE, warning = FALSE)
library(tidyverse)
library(openintro)
library(infer)
library(ggplot2)

data('yrbss', package='openintro')
```

Question 1

What are the cases in this data set? How many cases are there in our sample?

The cases are individual high schoolers who participated in the survey. 13,583 cases are in the sample.

```
nrow(yrbss)
```

```
## [1] 13583
```

Question 2

How many observations are we missing weights from?

1,004 observations are missing weights.

```
yrbss |>
  filter(is.na(height)) |>
  nrow()
```

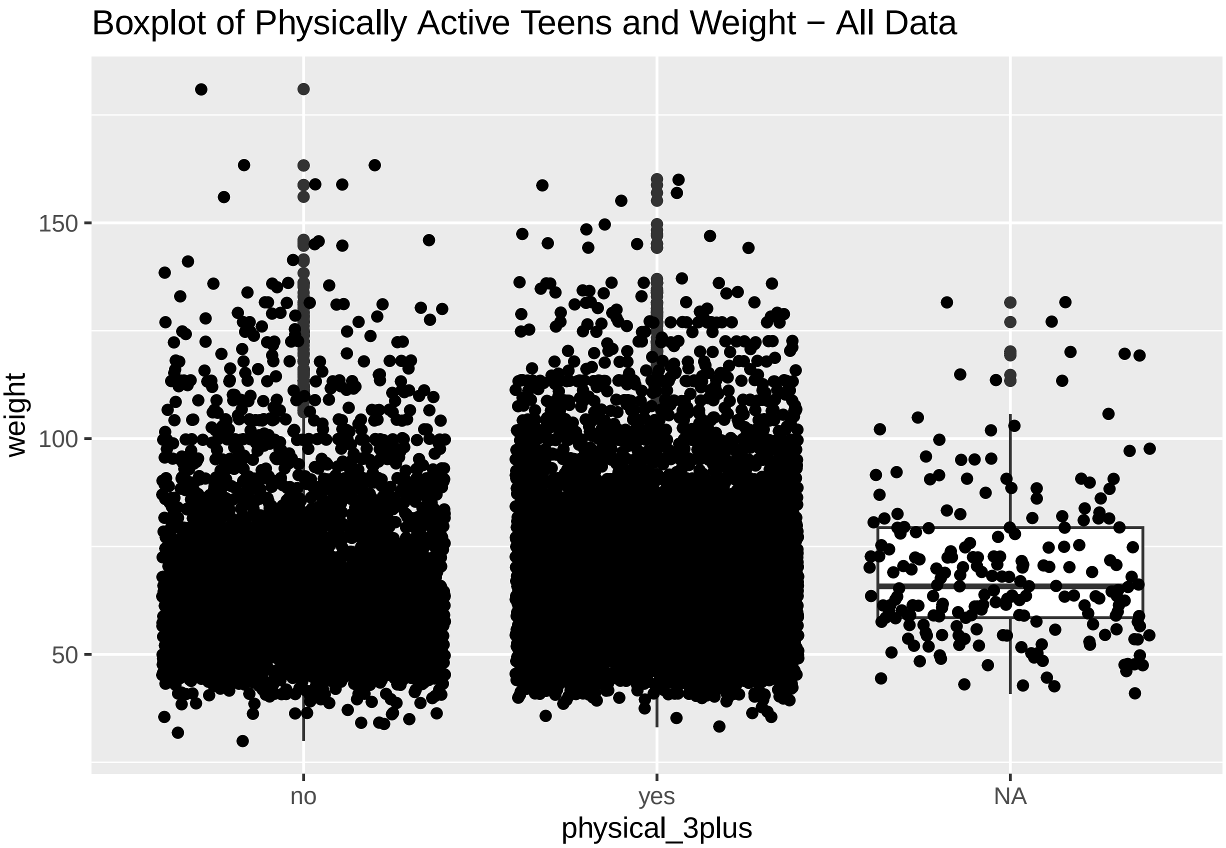
```
## [1] 1004
```

Question 3

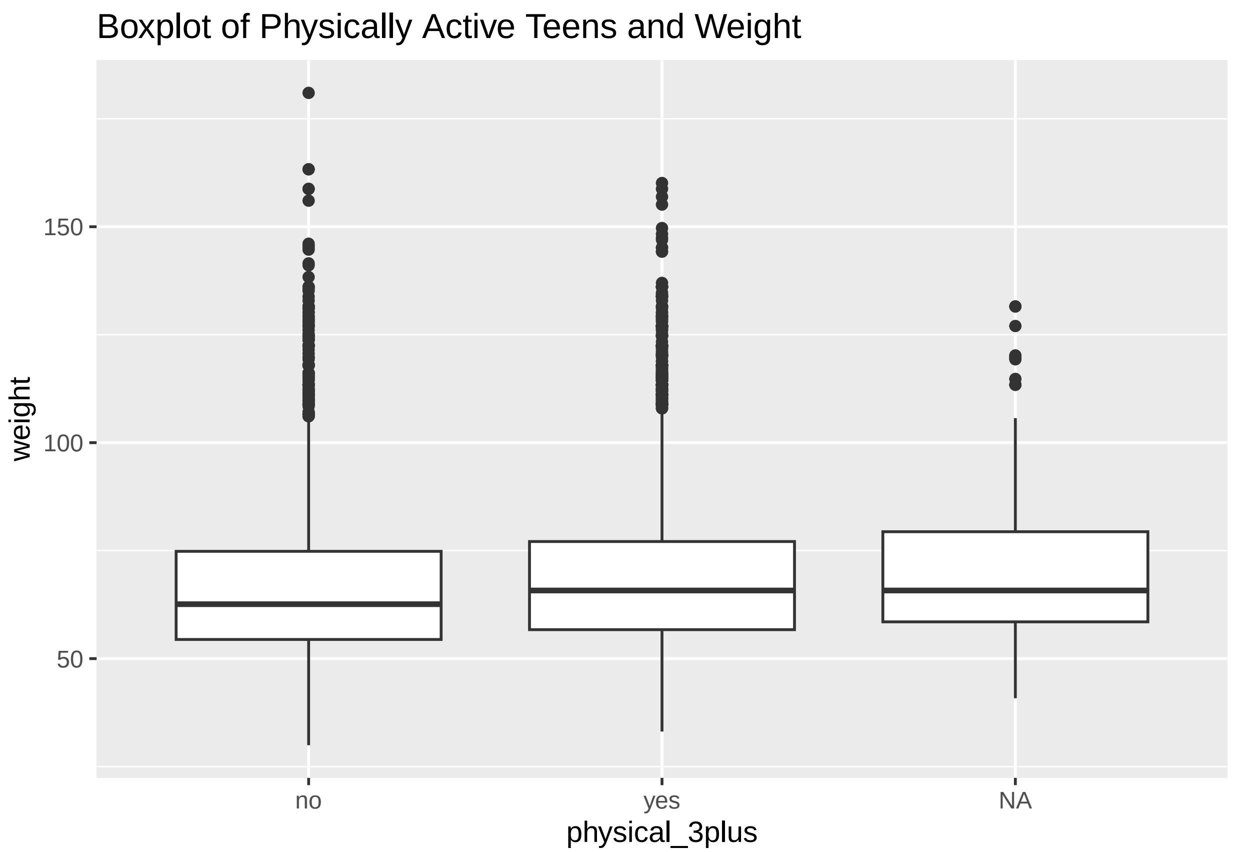
Make a side-by-side boxplot of physical_3plus and weight. Is there a relationship between these two variables? What did you expect and why?

It appears that these two variables are not highly related. If physical fitness had a higher criteria, then it might have shown a more obvious comparison. I expected students who worked out 3 days or more per week to weigh less, but they did not. One reason could be that people who are more active have more muscle, and muscle weighs more. Another explanation could be that the kids have high metabolisms, so a greater difference in weight might be seen in adults or the elderly than in teens.

```r
yrbss <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no"))
ggplot(yrbss, aes(physical_3plus, weight)) +
  geom_boxplot() +
  labs(title = "Boxplot of Physically Active Teens and Weight - All Data") +
  geom_jitter()
```



```r
ggplot(yrbss, aes(physical_3plus, weight)) +
  geom_boxplot() +
  labs(title = "Boxplot of Physically Active Teens and Weight")
```



Question 4

Are all conditions necessary for inference satisfied? Comment on each. You can compute the group sizes with the summarize command above by defining a new variable with the definition n().

The conditions necessary are independence, normality, and equal variance. Independence should be met because the Youth Behavior Risk Survey is from a simple random sample of high schoolers. Normality is met because the sample is much larger than 30, so it is assumed to be a normal distribution. Equal variance does not appear to be met. The variance of “no” is 311.1009, and the variance of “yes” is 271.5351. However, their spread is similar because their standard deviations are very close.

As for the group sizes, the group no contained 4022 people, the group yes contained 8342 people, and the group NA contained 273 people.

```
desc <- psych::describeBy(yrbss$weight, group = yrbss$physical_3plus, mat = TRUE, skew = FALSE)
names(desc)[2] <- 'Physically Active 3+ Days' # Rename the grouping column
desc$Var <- desc$sd^2 # We will need the variance latter, so calculate it here
desc
```

```
##     item Physically Active 3+ Days vars    n     mean       sd   min    max
## X11    1                        no    1 4022 66.67389 17.63805 29.94 180.99
## X12    2                       yes    1 8342 68.44847 16.47832 33.11 160.12
##      range        se      Var
## X11 151.05 0.2781183 311.1009
## X12 127.01 0.1804172 271.5351
```

```r
groups <- group_by(yrbss, physical_3plus)
summarise(groups, n())
```

```
## # A tibble: 3 x 2
##   physical_3plus ‘n()‘
##   <chr>          <int>
## 1 no              4404
## 2 yes             8906
## 3 <NA>             273
```

Question 5

Write the hypotheses for testing if the average weights are different for those who exercise at least times a week and those who don’t.

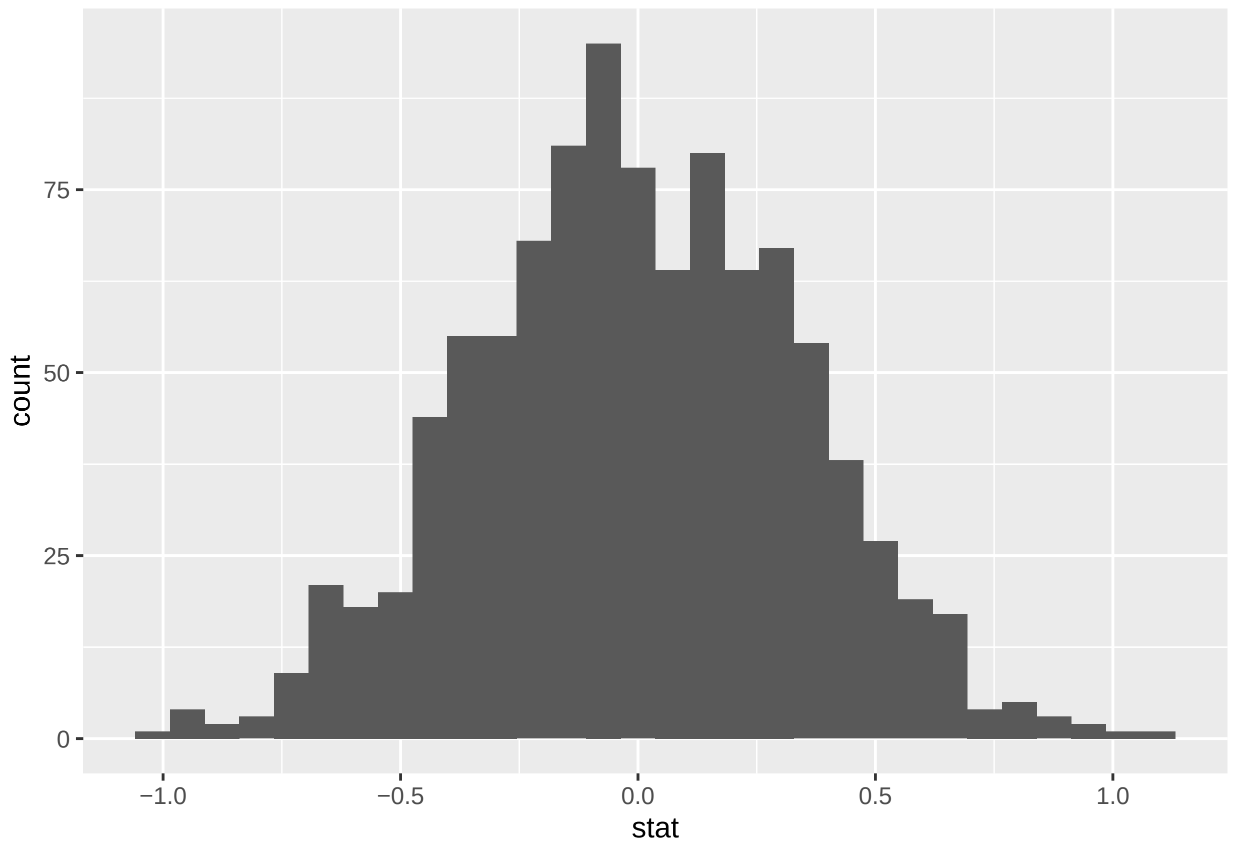
Null: Average weights are the same for those who exercise at least three times a week and those who don’t. Alternate: Average weights differ between those who exercise at least three times a week and those who don’t.

Question 6

How many of these null permutations have a difference of at least obs_stat?

obs_stat was approximately 1.77. None of the null permutations (obs_diff) were greater than 1.77, so none of them had a difference of at least obs_stat.

```r
set.seed(1234)
obs_diff <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
null_dist <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
ggplot(data = null_dist, aes(x = stat)) +
  geom_histogram()
```



```
obs_diff
```

```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  1.77
```

```
null_dist |>
  filter(stat >= obs_diff)
```

```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 0 x 2
## # i 2 variables: replicate <int>, stat <dbl>
```

Question 7

Construct and record a confidence interval for the difference between the weights of those who exercise at least three times a week and those who don't, and interpret this interval in context of the data.

(-0.02359907, 0.01866940)

Since the difference 0 falls within this confidence interval, there does not appear to be enough evidence to reject the null hypothesis. We are 95% confident that the actual difference falls within this confidence interval. Therefore, I would say that we cannot reject that the different groups of physical activity do not differ by weight.

```
(confidence_interval <- c(mean(null_dist$stat) - 1.96 * (sd(null_dist$stat) / sqrt(length(null_dist$sta
```

```
## [1] -0.02359907  0.01866940
```

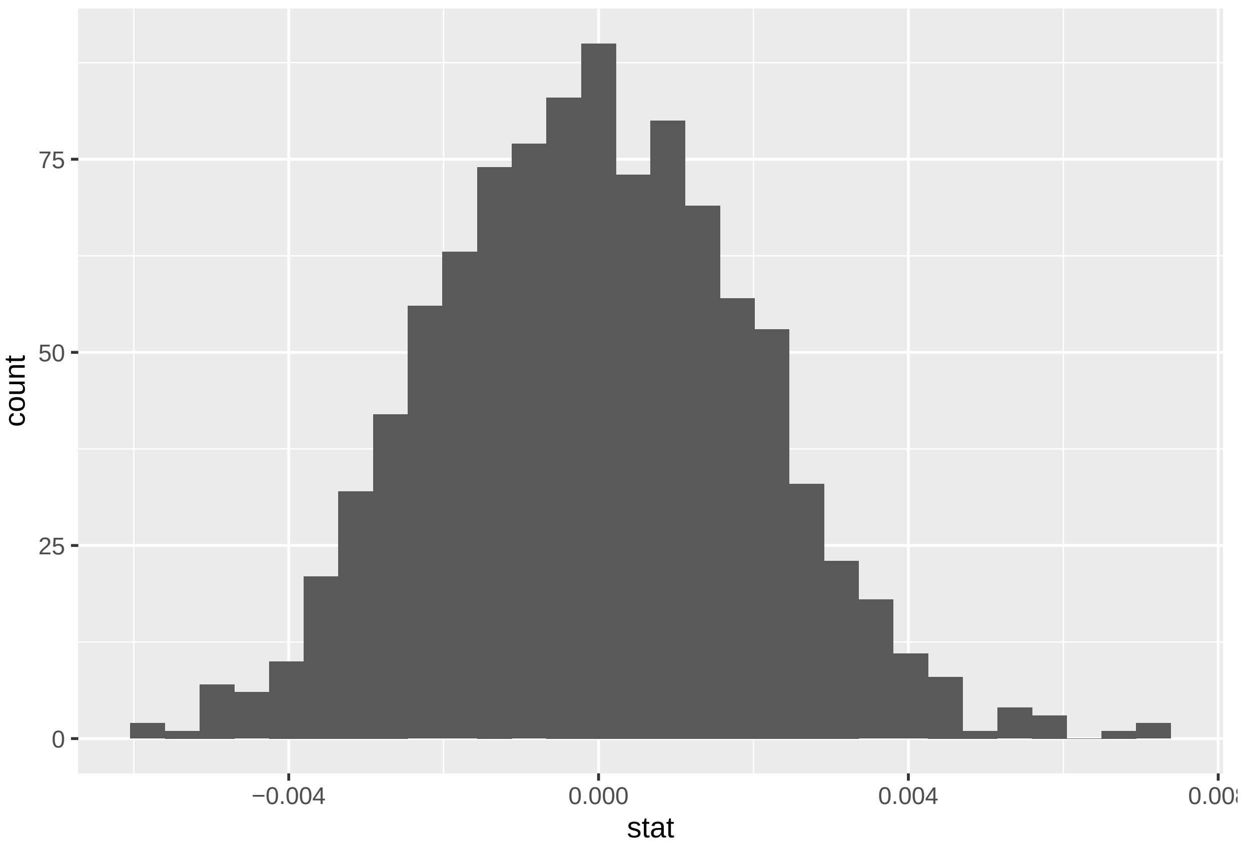
Question 8

Calculate a 95% confidence interval for the average height in meters (height) and interpret it in context.

(-0.00019285, 0.00006128537)

Since the difference 0 falls within this confidence interval, there does not appear to be enough evidence to reject the null hypothesis. We are 95% confident that the difference is within this interval. Therefore, I would say that we cannot reject that the different groups of physical activity do not differ by weight. The interval is smaller than the interval for weight, so it appears that the difference is smaller for heights than weights.

```
set.seed(1234)
obs_diff2 <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(height ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
null_dist2 <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
ggplot(data = null_dist2, aes(x = stat)) +
  geom_histogram()
```



```
(confidence_interval2 <- c(mean(null_dist2$stat) - 1.96 * (sd(null_dist2$stat) / sqrt(length(null_dist2
```

```
## [1] -1.928500e-04  6.128537e-05
```

Question 9

Calculate a new confidence interval for the same parameter at the 90% confidence level. Comment on the width of this interval versus the one obtained in the previous exercise.

Confidence interval for weights: (-0.02020249, 0.01527282) This interval is slightly smaller, which makes sense because we are less confident that the actual difference is within this interval.

Confidence interval for heights: (-0.0001724284, 0.00004086378) This interval is slightly smaller, which makes sense because we are less confident that the actual difference is within this interval.

```
(confidence_interval3 <- c(mean(null_dist$stat) - 1.645 * (sd(null_dist$stat) / sqrt(length(null_dist$s
```

```
## [1] -0.02020249  0.01527282
```

```
(confidence_interval4 <- c(mean(null_dist2$stat) - 1.645 * (sd(null_dist2$stat) / sqrt(length(null_dist
```

```
## [1] -1.724284e-04  4.086378e-05
```

Question 10

Conduct a hypothesis test evaluating whether the average height is different for those who exercise at least three times a week and those who don't.

Null: The average height is the same for those who exercise at least three times a week and those who don't.
Alternate: The average height is not the same for those who exercise at least three times a week and those who don't.

The t-test shows a significant p-value and a confidence interval that does not include 0. Therefore, we can reject the null hypothesis and accept that alternative hypothesis that the average height is different for those who exercise at least three times a week and those who don't. The true difference does not appear to equal 0.

```
new <- yrbss |> filter(!is.na(physical_3plus))

t.test(height ~ physical_3plus, data = new)
```

```
##
##  Welch Two Sample t-test
##
## data:  height by physical_3plus
## t = -19.029, df = 7973.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -0.04150183 -0.03374994
## sample estimates:
##  mean in group no mean in group yes
##          1.665587          1.703213
```

Question 11

Now, a non-inference task: Determine the number of different options there are in the dataset for the hours_tv_per_school_day there are.

There are 7 options for hours_tv_per_school_day (not including NA). There are 8 options if you include NA as an option.

```
unique(yrbss$hours_tv_per_school_day)
```

```
## [1] "5+"           "2"            "3"            "do not watch" "<1"
## [6] "4"            "1"            NA
```

```
yrbss$hours_tv_per_school_day |>
  unique() |>
  length()
```

```
## [1] 8
```

Question 12

Come up with a research question evaluating the relationship between height or weight and sleep. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Report the statistical results, and also provide an explanation in plain language. Be sure to check all assumptions, state your α level, and conclude in context.

Do students who sleep 8 or more hours weigh less than students who sleep less than 8 hours at an alpha level of 0.05?

The conditions necessary are independence, normality, and equal variance. Independence should be met because the Youth Behavior Risk Survey is from a simple random sample of high schoolers. Normality is met because the sample is much larger than 30, so it is assumed to be a normal distribution. The variance is not equal, but it is very close. The variance differs by only about 13. The variance of "no" is 291.5521, and the variance of "yes" is 278.6469.

The results of the t-test are shown below. The p-value is 0.003999, which is less than 0.05. Also, the 95% confidence interval does not contain 0. Therefore, I reject the null hypothesis and accept the alternate hypothesis. I would say that the difference between the groups is significant. The mean for "yes" was lower, so it does appear that students who sleep 8 or more hours on school nights weigh less than students who sleep less than 8 hours.

```
yrbss <- yrbss |>
  mutate(hours_sleep_8plus = ifelse((yrbss$school_night_hours_sleep == "8") | (yrbss$school_night_hours_

desc2 <- psych::describeBy(yrbss$weight, group = yrbss$hours_sleep_8plus, mat = TRUE, skew = FALSE)
names(desc2)[2] <- 'Sleep 8 Hours or More School Nights' # Rename the grouping column
desc2$Var <- desc2$sd^2 # We will need the variance latter, so calculate it here
desc2
```

```
##     item Sleep 8 Hours or More School Nights vars    n     mean       sd   min
## X11    1                                  no    1 8016 68.19018 17.07490 29.94
## X12    2                                 yes    1 3465 67.20623 16.69272 31.75
##        max  range        se      Var
## X11 180.99 151.05 0.1907125 291.5521
## X12 156.95 125.20 0.2835799 278.6469
```

```
t.test(weight ~ hours_sleep_8plus, data = yrbss)
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by hours_sleep_8plus
## t = 2.8792, df = 6712.5, p-value = 0.003999
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  0.3140248 1.6538777
## sample estimates:
##  mean in group no mean in group yes
##          68.19018          67.20623
```

Sources:

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). R for Data Science (2nd ed.). O'Reilly.