Homework 10: Hypothesis Testing

XDASI Fall 2021

ANSWER KEY

11 December, 2021

COVID-19 cases in NYC

"NYC_covid_cases_by_age.csv" contains weekly data of cases in NYC over the course of the pandemic.

1) Import, tidy, and plot the data

a) Convert to long format Read in the data file and convert the data to long format (you will need this in order to make the plot below).

- You should preserve the `week_ending` column.
- Get rid of the `age_all_ages` column.
- Convert the rest of the columns to two new columns containing the `age_group` and `cases` information.
- Turn the `age_groups` column into a factor (watch out, R likes to put everything in alphabetical order! you can set the levels in the correct order of age group by making a vector of the column names from the original data table and then selecting just columns 3:12)
- Make sure the `week_ending` column is properly formatted as a Date type.

```
# import file
nyc_cases<- read.csv("NYC_covid_cases_by_age.csv")

# check structure
str(nyc_cases)
```

```
## 'data.frame':    85 obs. of  12 variables:
##  $ week_ending : chr  "3/14/20" "3/21/20" "3/28/20" "4/4/20" ...
##  $ age_all_ages: num  22.7 221.9 338.1 420.2 413.1 ...
##  $ age_0_4     : num  NA 14.1 23.1 26.2 27.3 ...
##  $ age_5_12    : num  1.07 17.66 19 25.55 31.84 ...
##  $ age_13_17   : num  6.7 57.3 47.4 59.4 68.2 ...
##  $ age_18_24   : num  16.5 148.9 170.6 191.4 200.7 ...
##  $ age_25_34   : num  32.3 259.2 324.6 361.1 327.1 ...
##  $ age_35_44   : num  33.5 321.9 432.1 488 437.5 ...
##  $ age_45_54   : num  28.8 315.2 481.4 607.6 575.4 ...
##  $ age_55_64   : num  26.1 304.4 525.2 667 681.3 ...
##  $ age_65_74   : num  26.4 268.6 490.7 664 650.4 ...
##  $ age_75up    : num  23.1 227.9 546.1 802.7 912.7 ...
```

```r
# make a data frame in long format
nyc_cases_long <- nyc_cases %>% gather("age_group", "cases", -week_ending)

# get rid of what used to be the age_all_ages column
nyc_cases_long <- nyc_cases_long %>% filter(age_group != "age_all_ages")

# format age_group as factor in the correct order of age
age_groups = colnames(nyc_cases)[3:12]
nyc_cases_long$age_group <- factor(nyc_cases_long$age_group, levels = age_groups)

# format week_ending as a Date type (hint: look up `as.Date()`)
nyc_cases_long$week_ending <- as.Date(nyc_cases_long$week_ending, "%m/%d/%y")
str(nyc_cases_long)
```

```
## 'data.frame':    850 obs. of  3 variables:
##  $ week_ending: Date, format: "2020-03-14" "2020-03-21" ...
##  $ age_group  : Factor w/ 10 levels "age_0_4","age_5_12",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ cases      : num  NA 14.1 23.1 26.2 27.3 ...
```

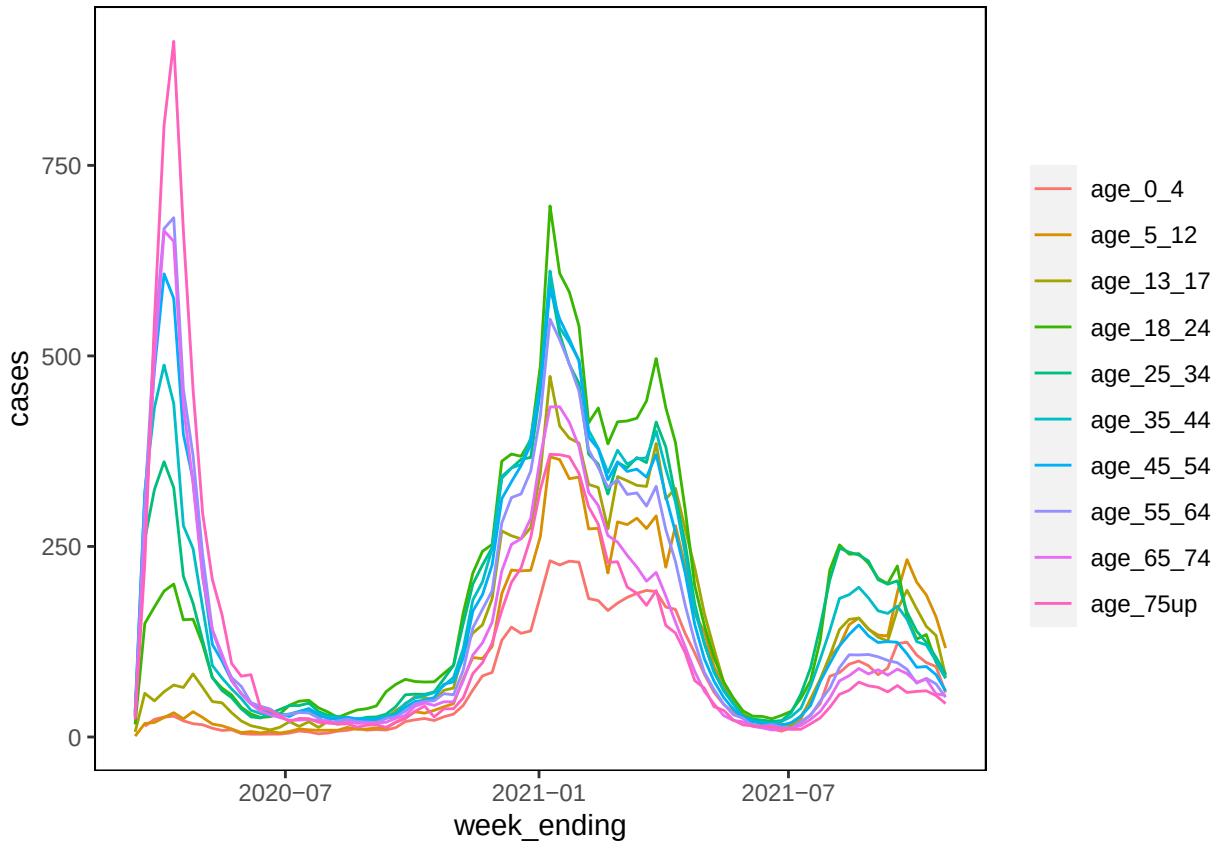
b) Percentage of cases in each age group As a convenience for questions to come, also add a column containing the percentage of cases observed in each age group out of the total cases per week

Hint: dplyr will be really useful for this. It's a good idea to drop rows with `NA` at this point (look up `drop_na`).

```r
# add pct_cases column
nyc_cases_long = nyc_cases_long %>%
  drop_na() %>%
  group_by(week_ending) %>%
  mutate(pct_cases = (cases / sum(cases)) * 100)

# check weekly totals (should = 100)
#nyc_cases_long %>% group_by(week_ending) %>% summarize(sum(pct_cases))
```

c) Plot cases over time by age group Create a plot that shows the trajectory of cases over the course of the pandemic broken down by age group.

```r
cases_weekly_plot<- ggplot(nyc_cases_long,
                           aes(x=week_ending,
                               y = cases,
                               color= age_group)) +
  geom_line() +
  theme(panel.background=element_blank(),
        strip.background = element_blank(),
        strip.placement = "outside",
        strip.text=element_blank(),
        panel.border=element_rect(colour="black",fill=NA),
        legend.title = element_blank())

cases_weekly_plot
```



2) First and second waves

The plot above shows that the number of cases over time forms "waves". Did the first and the second wave have a similar number of weekly recorded cases in all age groups?

a) Total cases in each wave Subset the data that corresponds to the first wave. You have to decide for yourself where to draw the cutoffs on the time axis. Make a new data frame containing the total weekly number of cases in the first wave (sum them up across all age groups).

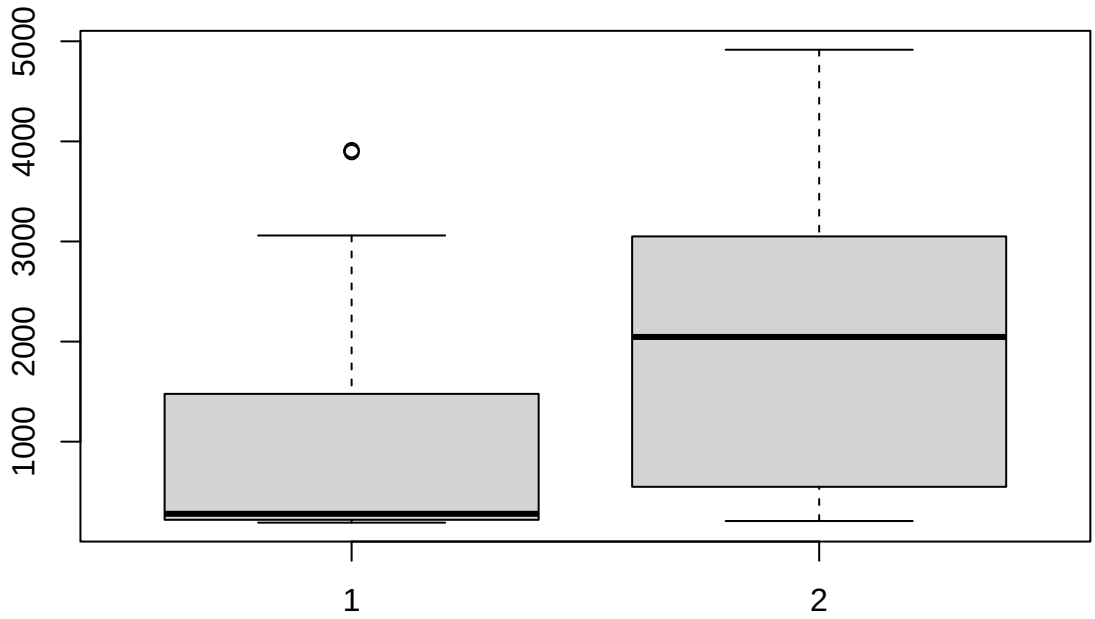
Then, do the same for the second wave.

```
# ============================================================================ #
# First wave data
# ============================================================================ #
first_wave_all_ages <- nyc_cases_long %>%
  drop_na() %>%
  filter(week_ending < as.Date("9/1/20", "%m/%d/%y")) %>%
  group_by(week_ending)

# summary table (total cases per week in 1st wave)
first_wave_all_cases = first_wave_all_ages %>%
  summarize(case_total = sum(cases))

# ============================================================================ #
# Second wave data
```

```r
# ============================================================================ #
second_wave_all_ages <- nyc_cases_long %>%
  drop_na() %>%
  filter(week_ending >= as.Date("9/1/20", "%m/%d/%y") &
           week_ending < as.Date("6/1/21", "%m/%d/%y"))

# summary table (total cases per week in 2nd week)
second_wave_all_cases = second_wave_all_ages %>%
  group_by(week_ending) %>%
  summarize(case_total = sum(cases))
```

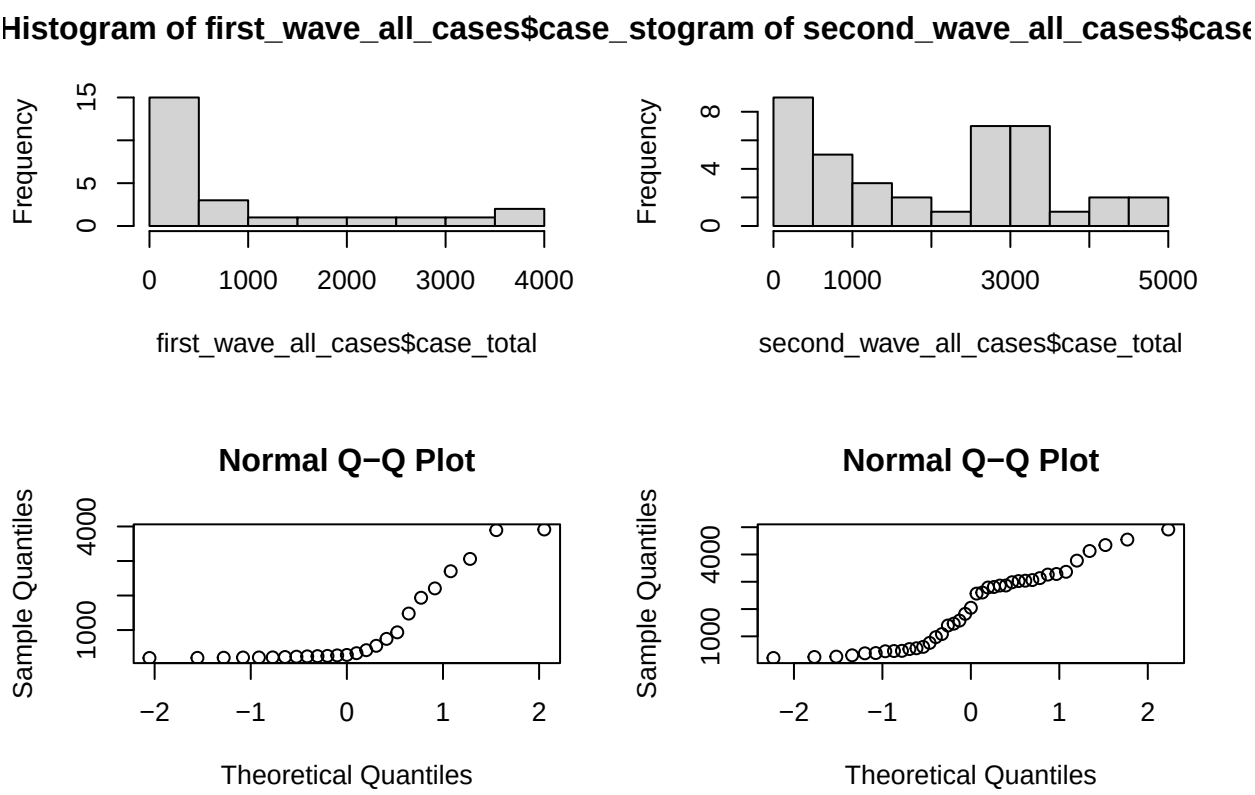
b) Exploratory analysis Explore the weekly number of cases for the two waves visually using a boxplot, histograms, and qq plots, and perform a test for normality.

```r
# boxplot
boxplot(first_wave_all_cases$case_total, second_wave_all_cases$case_total)
```




```r
# histograms and qqnorm plots
par(mfrow = c(2,2))
hist(first_wave_all_cases$case_total,
     breaks = 10)
hist(second_wave_all_cases$case_total,
     breaks = 10)
```

```
qqnorm(first_wave_all_cases$case_total)
qqnorm(second_wave_all_cases$case_total)
```




```
# normality test
shapiro.test(first_wave_all_cases$case_total)
```

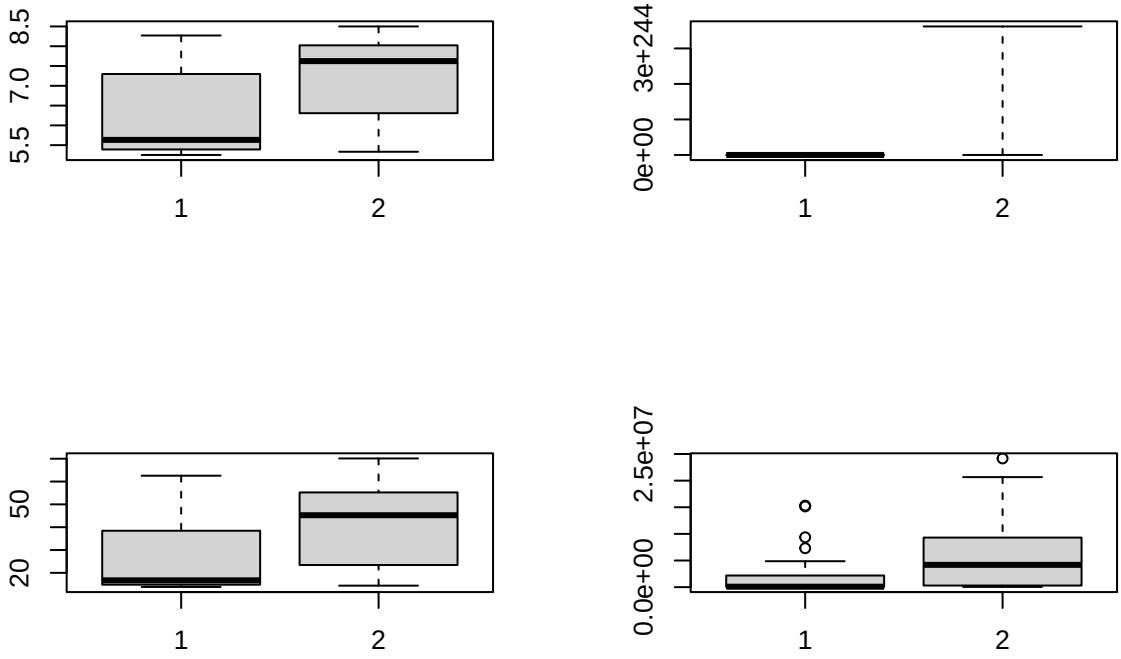
```
##
##  Shapiro-Wilk normality test
##
## data:  first_wave_all_cases$case_total
## W = 0.7041, p-value = 8.279e-06
```

```
shapiro.test(second_wave_all_cases$case_total)
```

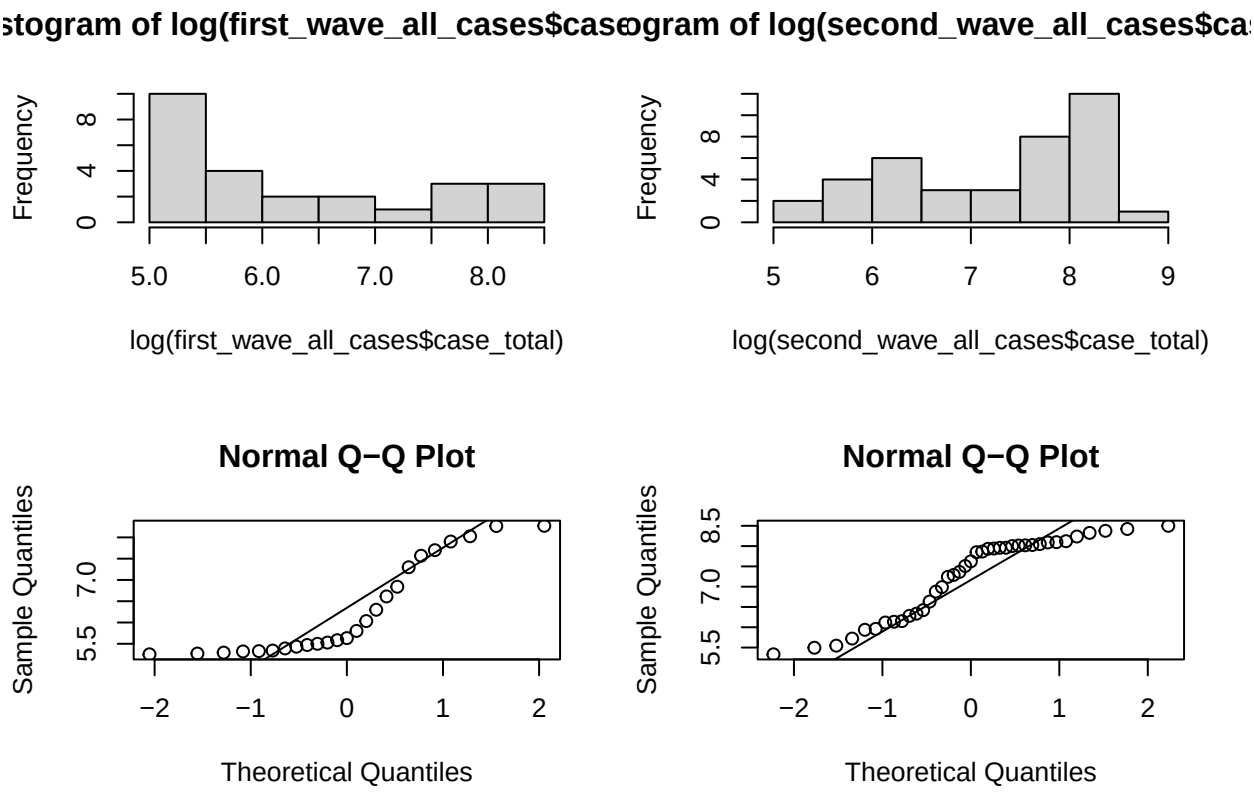
```
##
##  Shapiro-Wilk normality test
##
## data:  second_wave_all_cases$case_total
## W = 0.91062, p-value = 0.004513
```

c) Data transformations If the data look highly skewed, it might be useful to transform the data. Try some different transformations and see if they help make the data look more normal.

```r
# try some different transformations
par(mfrow = c(2,2))
boxplot(log(first_wave_all_cases$case_total), log(second_wave_all_cases$case_total))
boxplot(exp(first_wave_all_cases$case_total), exp(second_wave_all_cases$case_total))
boxplot(sqrt(first_wave_all_cases$case_total), sqrt(second_wave_all_cases$case_total))
boxplot((first_wave_all_cases$case_total)^2, (second_wave_all_cases$case_total)^2)
```




```r
# histograms and qqnorm plots
par(mfrow = c(2,2))
hist(log(first_wave_all_cases$case_total),
     breaks = 10)
hist(log(second_wave_all_cases$case_total),
     breaks = 10)
qqnorm(log(first_wave_all_cases$case_total))
qqline(log(first_wave_all_cases$case_total))
qqnorm(log(second_wave_all_cases$case_total))
qqline(log(second_wave_all_cases$case_total))
```




```
# normality test
shapiro.test(log(first_wave_all_cases$case_total))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(first_wave_all_cases$case_total)
## W = 0.81158, p-value = 0.0003569
```

```
shapiro.test(log(second_wave_all_cases$case_total))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(second_wave_all_cases$case_total)
## W = 0.88387, p-value = 0.0007852
```

```
shapiro.test((first_wave_all_cases$case_total)^2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  (first_wave_all_cases$case_total)^2
## W = 0.59319, p-value = 3.569e-07
```

```
shapiro.test((second_wave_all_cases$case_total)^2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  (second_wave_all_cases$case_total)^2
## W = 0.85284, p-value = 0.0001267
```

Did any of the transformations you tried help at all?

```
# your answer here
No. The first-wave data are very right-skewed, so transformations like the log will not help since this

Moreover, the data in the two groups have a different skew, so applying the same transformation to both
```

d) Statistical test Run an appropriate statistical test to compare the number of cases in the first and second wave. Which wave was bigger?

```
wilcox.test(first_wave_all_cases$case_total, second_wave_all_cases$case_total)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  first_wave_all_cases$case_total and second_wave_all_cases$case_total
## W = 236, p-value = 0.0003975
## alternative hypothesis: true location shift is not equal to 0
```

```
# the second wave was bigger
```

3) Changes in case rates over time

We cannot directly compare the number of cases in each age category because we do not know the total number of population in each group (e.g. are there more cases among 75+ year olds than among 0-4 year olds because they get sick more or because there are more of them to begin with?)

But we can look whether the **proportion** of cases that fall within each age group (out of the total number of weekly cases summed up across all age groups) differs with time.

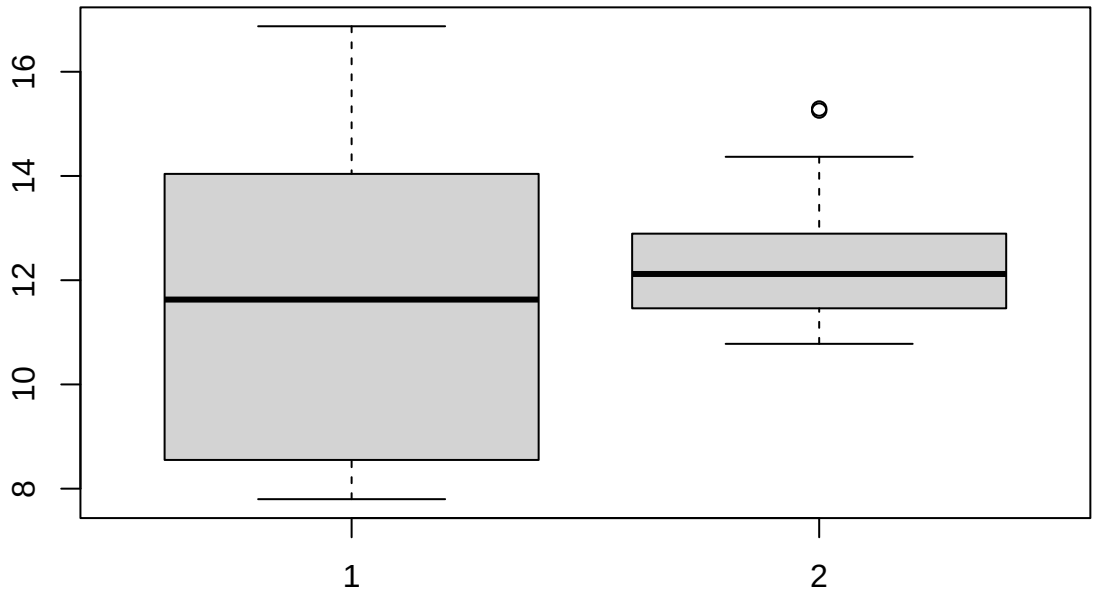
For example, is there a significant difference between the proportion of infections among younger adults vs. older adults in the first and second waves? What about before and after the appearance of the Delta variant?

a) Younger vs. older adults in waves 1 and 2 Did the age distribution of cases differ between the first two big waves of cases?

First, subset the data to get the percentage of cases in adults under 35 in each wave. Make a box plot to compare them and perform an appropriate statistical test for significance between the rates in the two waves.

Then, do the same thing for adults 65 and up.

```r
# ============================================================================ #
# young adults (18-34) - 1st vs. 2nd wave
# ============================================================================ #
first_wave_under34 <- first_wave_all_ages %>%
  filter(age_group == "age_18-24" | age_group == "age_25_34" ) %>%
  select(week_ending, pct_cases)

second_wave_under34 <- second_wave_all_ages %>%
  filter(age_group == "age_18-24" | age_group == "age_25_34" ) %>%
  select(week_ending, pct_cases)

boxplot(first_wave_under34$pct_cases, second_wave_under34$pct_cases)
```

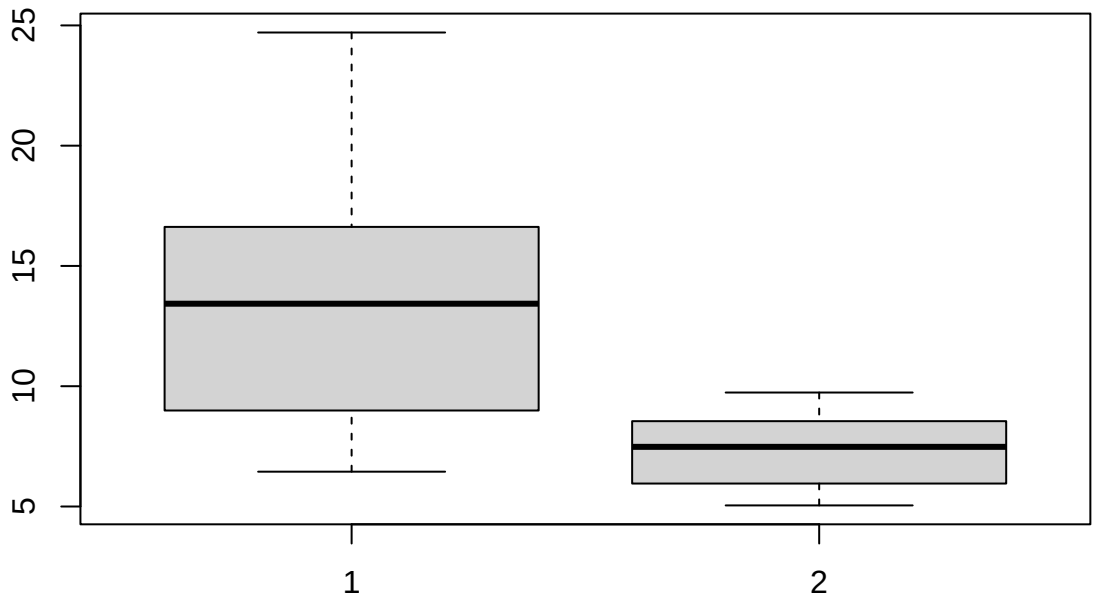



```r
wilcox.test(first_wave_under34$pct_cases, second_wave_under34$pct_cases)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  first_wave_under34$pct_cases and second_wave_under34$pct_cases
## W = 408, p-value = 0.2792
## alternative hypothesis: true location shift is not equal to 0
```

```r
# ============================================================================ #
# older adults (65 and up) - 1st vs. 2nd wave
```

```r
# ============================================================================ #
first_wave_65up <- first_wave_all_ages %>%
  filter(age_group == "age_65_74" | age_group == "age_75up" ) %>%
  select(week_ending, pct_cases)

second_wave_65up <- second_wave_all_ages %>%
  filter(age_group == "age_65_74" | age_group == "age_75up" ) %>%
  select(week_ending, pct_cases)

boxplot(first_wave_65up$pct_cases, second_wave_65up$pct_cases)
```




```r
wilcox.test(first_wave_65up$pct_cases, second_wave_65up$pct_cases)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  first_wave_65up$pct_cases and second_wave_65up$pct_cases
## W = 3597, p-value = 8.89e-16
## alternative hypothesis: true location shift is not equal to 0
```

How did the rates of infection in these two groups differ between the first and second waves? Can you speculate on the reason for any observed differences?

```
# your answer here
Overall rates of infection in young adults did not change much as a fraction of total cases, however th
```

b) Bootstrap confidence intervals Perform a bootstrap analysis to get estimates for the median and 95% confidence intervals for young adults and older adults in each of the two waves.

```
# ========================================================================== #
# young adults (18-34) - 1st vs. 2nd wave
# ========================================================================== #
median_first_under34 = do(1000) * median(sample(first_wave_under34$pct_cases,
                                                nrow(first_wave_under34),
                                                replace=T))
quantile(median_first_under34$median,c(0.025,0.5,0.975))
```

```
##      2.5%       50%     97.5%
##  9.276137 11.628024 13.888617
```

```
median_second_under34 = do(1000) * median(sample(second_wave_under34$pct_cases,
                                                 nrow(second_wave_under34),
                                                 replace=T))
quantile(median_second_under34$median,c(0.025,0.5,0.975))
```

```
##     2.5%      50%    97.5%
## 11.90674 12.11942 12.58080
```

```
# ========================================================================== #
# older adults (65 and up) - 1st vs. 2nd wave
# ========================================================================== #
median_first_65up = do(1000) * median(sample(first_wave_65up$pct_cases,
                                             length(first_wave_65up),
                                             replace=T))
quantile(median_first_65up$median,c(0.025,0.5,0.975))
```

```
##      2.5%       50%     97.5%
##  8.275076 13.246014 21.136289
```

```
median_second_65up = do(1000) * median(sample(second_wave_65up$pct_cases,
                                              length(second_wave_65up),
                                              replace=T))
quantile(median_second_65up$median,c(0.025,0.5,0.975))
```

```
##     2.5%      50%    97.5%
## 5.559296 7.302396 9.157573
```

```
# ========================================================================== #
# take a look at the bootstrap distributions for the sample medians
# par(mfrow=c(2,2))
# hist(median_first_under34$median, breaks=10)
# hist(median_second_under34$median, breaks=10)
# hist(median_first_65up$median, breaks=10)
# hist(median_second_65up$median, breaks=10)
```

4) Delta strain

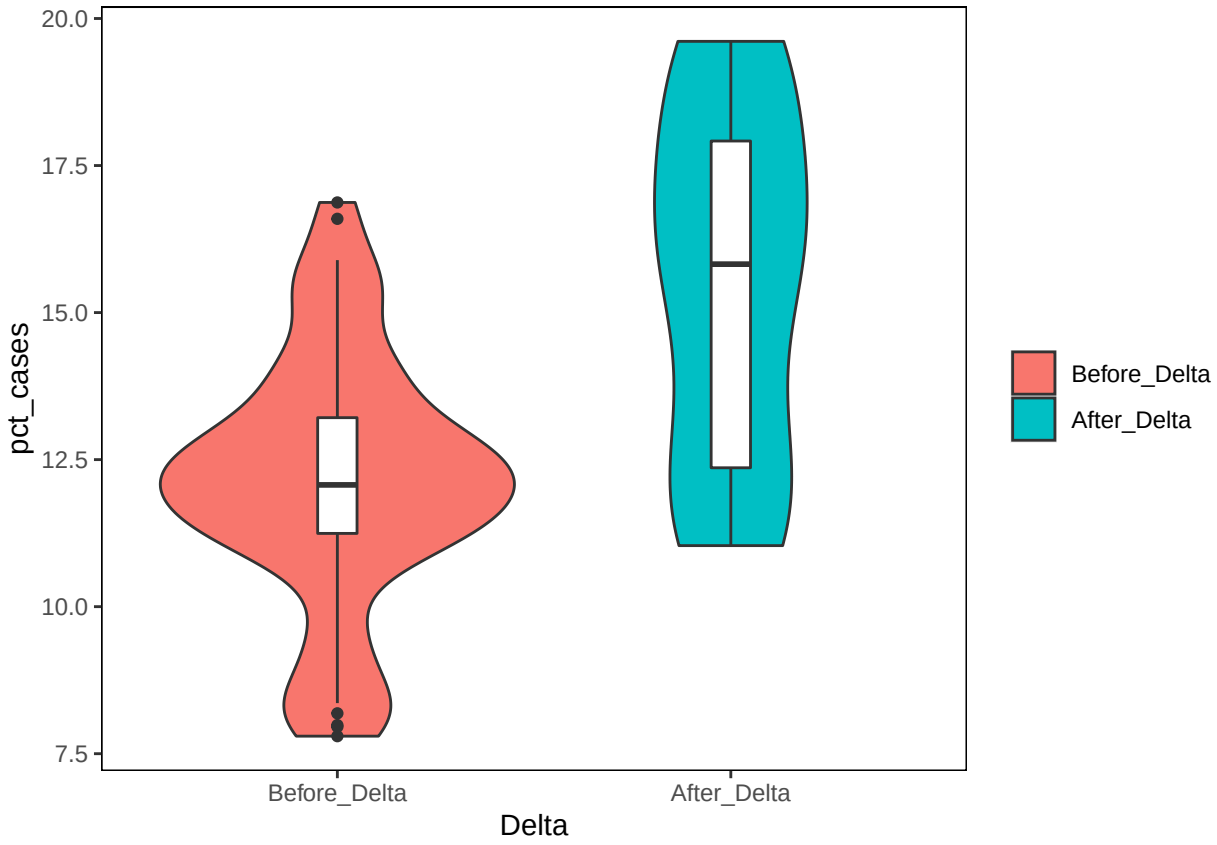
The Delta strain became prevalent - i.e. started accounting for >50% all cases - on 6/23/21. Is there a significant difference between the proportion of infections among 25-34 year-olds before and after the rise of the Delta strain?

a) Before and after Delta Extract the data for the 25-34 age group from your long table and add a column that labels the data for each week as "Before_Delta" or "After_Delta" based on the date of appearance of the Delta variant.

```
cases_25_34 <- nyc_cases_long %>%
  drop_na() %>%
  group_by(week_ending) %>%
  filter(age_group == "age_25_34") %>%
  mutate(Delta = ifelse(week_ending < as.Date("6/23/21", "%m/%d/%y"),
                        "Before_Delta", "After_Delta"))

cases_25_34$Delta = factor(cases_25_34$Delta,
                           levels=c("Before_Delta","After_Delta"))
```

b) Violin plot Make a violin plot of the percentage of cases observed per week in the 25-34-year old group before and after the appearance of the Delta variant.

```
ggplot(cases_25_34, aes(x= Delta, y=pct_cases, fill= Delta)) +
  geom_violin() +
  geom_boxplot(width=0.1, fill="white") +
  theme(panel.background=element_blank(),
        strip.background = element_blank(),
        strip.placement = "outside",
        strip.text=element_blank(),
        panel.border=element_rect(colour="black",fill=NA),
        legend.title = element_blank())
```



c) Permutation test Perform a permutation ("shuffle") test for the observed median difference in the percentage of total cases in the 25-34 year-old age group before and after Delta.

```
# observed difference in rates
obs = diff(mosaic::median(pct_cases ~ Delta, data = cases_25_34))
obs
```

```
## After_Delta
##     3.75334
```

```
# shuffle test
permute_25_34 = mosaic::do(1000) * diff(mosaic::median(pct_cases ~ shuffle(Delta), data = cases_25_34))

# # For some reason, the shuffle renames the `pct_cases` column to `After_Delta`.
# # Change the name of the percent cases column back to `pct_cases`
# permute_25_34 = permute_25_34 %>% rename(pct_cases = After_Delta)
#
# # draw a histogram
# gf_histogram(gformula = ~ pct_cases, fill = ~ (pct_cases >= obs), data = permute_25_34,
#   binwidth = 0.4,
#   xlab = "Distribution of difference in medians under the null hypothesis")
```

What's the p-value you get from the shuffle test? Compare this with with the result of another appropriate statistical test.

```r
# p-value (permutation)
p_val = sum(permute_25_34 > obs) / 1000
p_val
```

```
## [1] 0
```

```r
# p-value (statistical test)
before = cases_25_34 %>% filter(Delta == "Before_Delta")
after = cases_25_34 %>% filter(Delta == "After_Delta")
wilcox.test(before$pct_cases,after$pct_cases)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  before$pct_cases and after$pct_cases
## W = 259, p-value = 0.00022
## alternative hypothesis: true location shift is not equal to 0
```

5) Breakthrough cases

Consider the following data of the number of NYC documented Covid cases, hospitalizations, and deaths among vaccinated and unvaccinated individuals:

- Cases:
  - Vaccinated: 820
  - Unvaccinated: 4140
- Hospitalizations:
  - Vaccinated: 32
  - Unvaccinated: 349
- Deaths:
  - Vaccinated: 4
  - Unvaccinated: 39

What is an appropriate test you could use to analyze these data? Explain.

```r
# your answer here
Either a Fishers Exact Test or a Chi-square test would work here.

Since the numbers are not so big, it is feasible to perform Fishers
and not too computationally intensive. (People tend to forget that Fishers can
be performed on tables that are larger than two-by-two, but it works fine when
the numbers are not too big).
```

Use this data to create a contingency table, and then perform the appropriate test to determine whether there is a significant difference between the vaccinated and unvaccinated groups.

```r
breakthrough_cases = data.frame(Cases = c(820,4140),
                                Hospitalizations = c(32, 349),
                                Deaths = c(4, 39))
row.names(breakthrough_cases) = c("Vaccinated","Unvaccinated")
breakthrough_cases
```

```
##              Cases Hospitalizations Deaths
## Vaccinated     820               32      4
## Unvaccinated  4140              349     39
```

```r
knitr::kable(breakthrough_cases)
```

	Cases	Hospitalizations	Deaths
Vaccinated	820	32	4
Unvaccinated	4140	349	39

```r
fisher.test(breakthrough_cases)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  breakthrough_cases
## p-value = 2.324e-05
## alternative hypothesis: two.sided
```

```r
chisq.test(breakthrough_cases)
```

```
##
##  Pearson's Chi-squared test
##
## data:  breakthrough_cases
## X-squared = 18.915, df = 2, p-value = 7.81e-05
```

The tests work the same either way the table is constructed. By convention, we put the groups we want to compare in the rows with the "focal" group first, and the conditions in the columns.

```r
vaccinated<- c(820,32,4)
unvaccinated<- c(4140,349,39)

breakthrough<- data.frame(vaccinated, unvaccinated)
rownames(breakthrough)<- c("cases", "hospitalizations", "deaths")

fisher.test(breakthrough)
chisq.test(breakthrough)
```