Final Project

David, Saixiao, Alex

2024-11-12

```r
library(dplyr)
library(ggplot2)
library(forcats)
library(tidyr)
```

1.

```r
# install.packages("tidytuesdayR")
# library(tidytuesdayR)
# tt_available()
```

2. links: https://github.com/WSJ/measles-data https://www.wsj.com/graphics/school-measles-rate-map/

```r
measles <- read.csv("/cloud/project/data/all-measles-rates.csv")
#summary(measles)
```

```r
measles$type <- as.factor(measles$type)
```

```r
measles<-measles %>%
 mutate(mmr_category= ifelse(mmr >=95,">=95%","<95%"))
```

Part 1 Data Cleaning

The minimum MMR rate is -1.00, which is unrealistic. Given that there are no missing values in the original "mmr" data, it is likely that missing values were mistakenly assigned the value of -1.00.

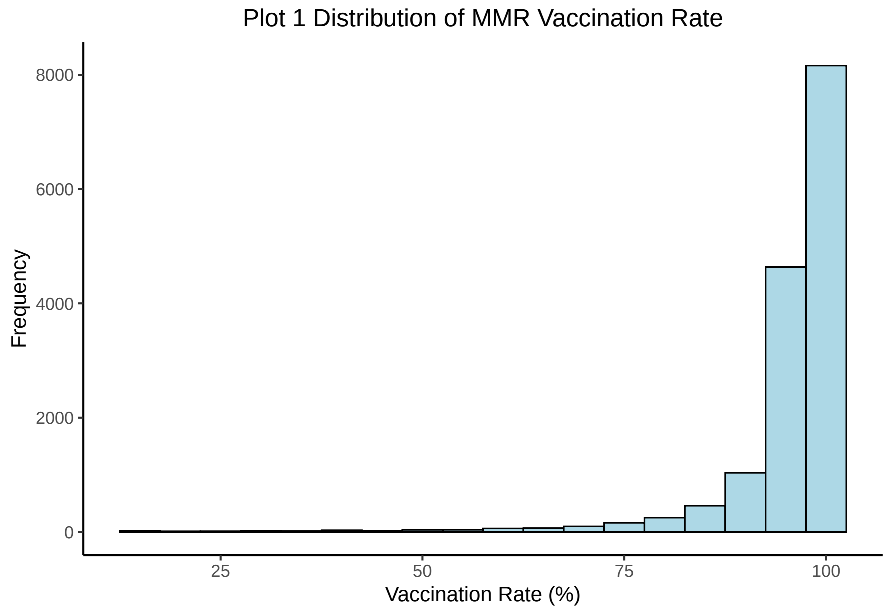
```r
measles_type <- measles%>%
filter(type != "Kindergarten" & type != "",mmr>=10) %>%
  mutate(school_type = fct_collapse(type,
  Public = c("BOCES", "Public"),
  Private = c("Nonpublic", "Private"),
  Charter = "Charter")) %>%
mutate(school_type = fct_drop(school_type))
```

Part 2: Exploratory Data Analysis

The data is left-skewed as most schools with mmr rate higher than 90%.

```r
ggplot(measles_type, aes(x = mmr)) +
  geom_histogram(binwidth = 5, fill = "lightblue",
                 col = "black", lwd = 0.4) +
  labs(title = "Plot 1 Distribution of MMR Vaccination Rate",
       x = "Vaccination Rate (%)",
       y = "Frequency") +
```

```r
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
```



Part 3: Data Analysis

3.1 One proportion hypothesis test

1. Write the hypotheses.

$H_0 : p_0 = 0.8$ (The proportion of elementary schools in the U.S. with an MMR vaccination rate of at least 95% is 80%.)

$H_A : p_0 < 0.8$ (The proportion of elementary schools in the U.S. with an MMR vaccination rate of at least 95% is less than 80%.)

2. Check conditions.

3. Independence condition is satisfied since we assume all the elementary schools are independent.

4. Success-failure conditions is satisfied since:

$np_0 = 15130 * 0.80 = 12104 \geq 10$ $n(1 - p_0) = 15130 * 0.2 = 3026 \geq 10$

```r
addmargins(table(measles_type$mmr_category))
```

```
##
##   <95% >=95%   Sum
##   3361 11769 15130
```

```r
p0 <- 0.8
n1 <- 11769
n <- 15130
n*p0
```

```
## [1] 12104
```

```r
n*(1-p0)
```

```
## [1] 3026
```

3. Test statistic

```r
# Hypothesis test for one-proportion
test1 <- prop.test(n1, n, p = 0.8, alternative = "less")
cat("Test statistic:", -sqrt(test1$statistic))
```

```
## Test statistic: -6.798556
```

4.

```r
p_val <- test1$p.value
cat("P-value =",p_val)
```

```
## P-value = 5.283654e-12
```

5. Decision and conclusion

Since p-value less than 0.05, we reject H_0, we have enough evidence that the true proportion of elementary schools in the US with MMR vaccination rate higher than 95% is less than 80%.

```r
cat("Confidence interval:", quantile(test1$conf.int, c(0.025, 0.975)), "\n")
```

```
## Confidence interval: 0.019585 0.763815
```

3.2 Independence test

1. Hypotheses

H_0 : school type and MMR vaccination rate are independent.

H_1: school type and MMR vaccination rate are not independent.

2. Test conditions:

a. Independence: we are reasonable to assume all schools are independent with each other.

b. Expected counts: all greater than 5

```r
# Hypothesis test for independence
test2 <- chisq.test(measles_type$school_type, measles_type$mmr_category)
# Compute the expected count
cat("Expected counts:", test2$expected)
```

```
## Expected counts: 2615.271 47.53827 698.1905 9157.729 166.4617 2444.809
```

3. Test statistics

```r
cat("Test statistic:", test2$statistic)
```

```
## Test statistic: 520.1772
```

4. P-value < 0.05

```r
cat("P-value:", test2$p.value)
```

```
## P-value: 1.109063e-113
```

5. Decision and conclusion

Decision: we reject H_0

Conclusion: we have enough evidence that school type and MMR vaccination rate are not independent with each other.

3.3 Exemption Analysis

```r
exemption_summary <- measles_type %>%
filter(xmed != -1 ,xper != -1 , xrel != -1)%>%
  group_by(school_type) %>%
  summarise(
    mean_med_exempt = mean(xmed, na.rm = TRUE),
    mean_pers_exempt = mean(xper, na.rm = TRUE),
    mean_rel_exempt = mean(xrel, na.rm = TRUE)
  )
exemption_summary
```

```
## # A tibble: 2 x 4
##   school_type mean_med_exempt mean_pers_exempt mean_rel_exempt
##   <fct>                 <dbl>            <dbl>           <dbl>
## 1 Public                0.248             7.22           0.252
## 2 Private               1.23             15.2            3.31
```

Limitations: Exemption data does not include charter schools.

Future Research: Investigating the reasons behind the lower vaccination rates in “charter schools” can provide actionable insights to policymakers.