Inference for Categorical Variables

Math 141

```
library(tidyverse)
library(infer)
```

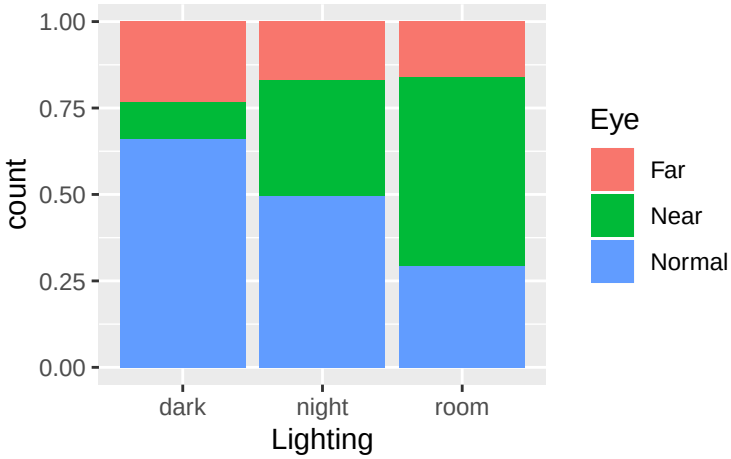
Example: Near-sightedness typically develops during the childhood years. Quinn, Shin, Maguire, and Stone (1999) examined the type of light children were exposed to and their eye health based on questionnaires at a university pediatric ophthalmology clinic. Below are the results.

```
# Import data
eye_data <- read_csv("/home/courses/math141f19/Data/eye_lighting.csv")
```

- **Cases**:
- **Variable of interests**:
- **Hypotheses**:

Does there appear to be a relationship?

```
ggplot(data = eye_data, mapping = aes(x = Lighting, fill = Eye)) +
  geom_bar(position = "fill")
```



Need to construct a test statistic which quantifies the likelihood of the sample results or more extreme under Ho.

```
count(eye_data, Eye)
count(eye_data, Lighting)
count(eye_data, Lighting) %>%
  summarise(sum(n))
```

Simulation-based method

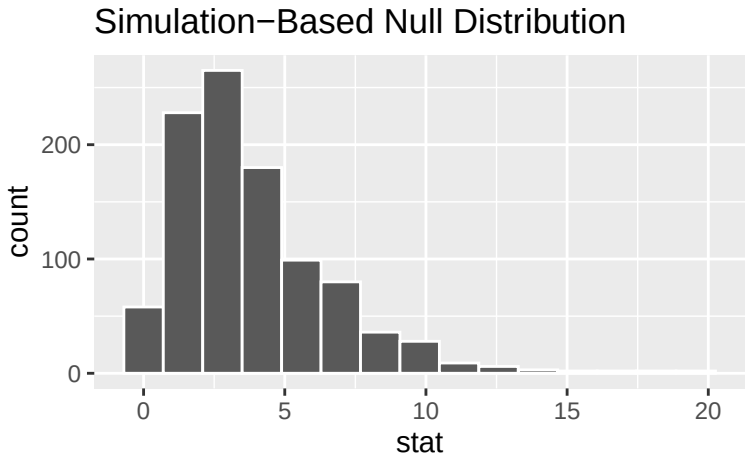
```
#Compute Chi-square test stat
test_stat <- eye_data %>%
  specify(Eye ~ Lighting) %>%
  calculate(stat = "Chisq")
test_stat
```

```
## # A tibble: 1 x 1
```

```
##    stat
##   <dbl>
## 1  56.5
```

```
# Construct null distribution
null_dist <- eye_data %>%
  specify(Eye ~ Lighting) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")

# Plot distribution
null_dist %>%
  visualize()
```

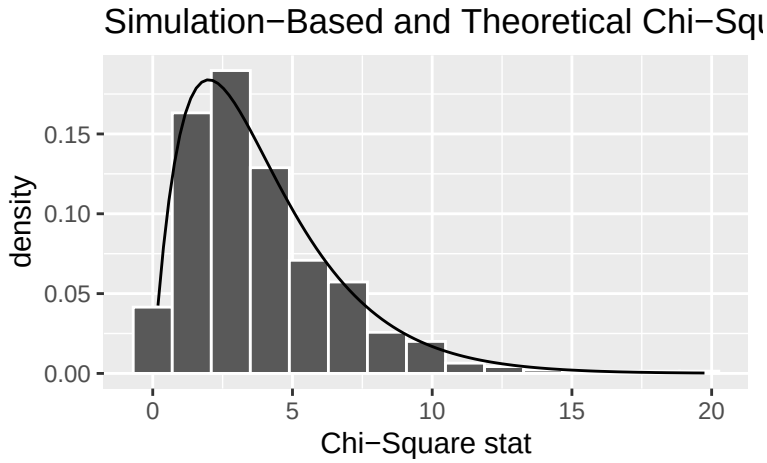



```
# Compute p-value
null_dist %>%
  get_pvalue(obs_stat = test_stat, direction = "greater")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

Theory-based method

```
# Add theoretical distribution
null_dist %>%
  visualize(method = "both")
```




```
chisq_test(eye_data, formula = Eye ~ Lighting)
```

```
## # A tibble: 1 x 3
##   statistic chisq_df  p_value
##       <dbl>    <int>    <dbl>
## 1      56.5        4 1.56e-11
```