Lesson 22 Sleep and Nightmares (6.1)

CDT I. M. Smrt

Background In a 2003 study in the Journal of Sleep and Hypnosis, researchers were interested in whether the side a person slept on impacted the dreams that they experienced and if they had nightmares. Researchers selected participants who fell asleep and awoke on one side and who could remember their dreams. They found 63 participants, of whom 41 were right- side sleepers and 22 were left-side sleepers. The researchers would like to determine if a larger proportion of left side sleepers had nightmares than right side sleepers. The data set `nightmares.csv` contains the reported side a participant woke up on and whether or not they experienced a nightmare.

```r
# Load packages
library(tidyverse)

### Data management
nightmare <- read.csv("https://raw.githubusercontent.com/jkstarling/MA376/main/nightmares.csv",
                      header=T,
                      stringsAsFactors = TRUE)

# setwd("C:/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons_AY23-1/Block III/LSN2
# nightmare <- read.csv("nightmares.csv", header = TRUE, stringsAsFactors = TRUE)
```

(1) What are our hypotheses (in words)? What is our explanatory/response variable? What type of data are we dealing with for the explanatory/response variables?

(2) How are the types of data different than what we have dealt with before? What is statistical model we will deal with now?

(3) A common way to depict data for this sort of study is using a 2 x 2 contingency table.

The order of the contingency table matters. Make sure you place the explanatory variable (groups) on the x-axis (columns) and the response variable (success/failure) on the y-axis.

What is success in this case?

```r
T1 <- table(nightmare$Nightmare, nightmare$Side)
T1
```

```
##
##        L  R
##   No  11 31
##   Yes 11 10
```

Drawing inferences beyond the data We have several tests that we can use to draw inferences beyond the data.

Choice 1: Two-sample z test (Difference in Proportions)

(4) What is the parameter of interest?

(5) Write the hypotheses in symbols.

(6) What is the statistic? Calculate it here using the table above.

```
phat_left <- T1[2,1] / sum(T1[,1])
phat_right <- T1[2,2] / sum(T1[,2])
phat_dif <- phat_left-phat_right
c(phat_left, phat_right, phat_dif)
```

```
## [1] 0.5000000 0.2439024 0.2560976
```

```
prop.table(T1,2)
```

```
##
##               L         R
##   No  0.5000000 0.7560976
##   Yes 0.5000000 0.2439024
```

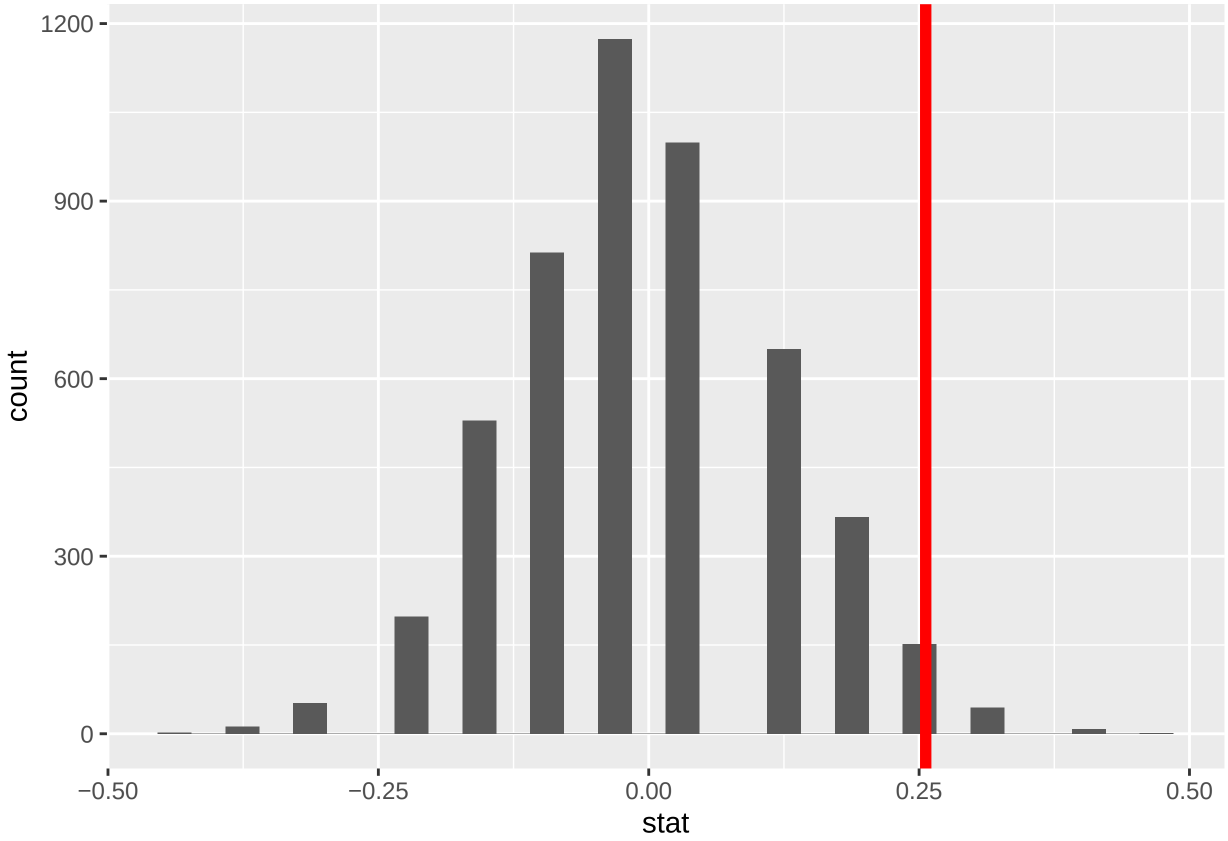
(7) How rare is this under H_0? We can use our shuffling simulation strategy to estimate:

```
set.seed(376)

M <- 5000
results <- data.frame(stat = rep(NA,M))

for(i in 1:M){
  nm.mod <- nightmare %>% mutate(shuff.cat = sample(Side))
  my.table <- table(nm.mod$Nightmare, nm.mod$shuff.cat)
  p.tabl <- prop.table(my.table,2)
  results$stat[i] <- p.tabl[2,1] - p.tabl[2,2]
}

results %>% ggplot(aes(x=stat)) + geom_histogram()+
  geom_vline(xintercept = phat_dif, lwd = 2, color = "red")
```



```
sum(results$stat >= phat_dif) / M
```

```
## [1] 0.041
```

(8) We can use the normal distribution as our reference distribution in a theory based test provided we have at least 10 successes and 10 failures **in each group**. If that is the case we can say that the CLT has kicked in and we use the two-sample z-test.

Conduct the 2-sample z-test below and create a 95% CI for the difference in proportions:

```
### Theory-based approach
## 2-sample z - test for equality of proportions
n_left <- sum(T1[,1])
n_right <- sum(T1[,2])
phat <- sum(T1[2,])/ sum(T1)

z <- (phat_dif)/sqrt(phat*(1-phat)*(1/n_left + 1/n_right))
2*(1-pnorm(abs(z)))                 # Ha two-sided
```

```
## [1] 0.03981831
```

```
# Confidence interval
se <- sqrt( ( phat_left*(1-phat_left)/n_left) + ((phat_right*(1-phat_right))/n_right) )
multiplier <- qnorm(.975)

c((phat_dif)-multiplier*se , (phat_dif)+multiplier*se)
```

```
## [1] 0.009254591 0.502940531
```

Choice 2: Chi-square test **The statistic**:

$$\chi^2 = \sum_{Cells} \frac{(Obs - Exp)^2}{Exp}$$

where the Observed are the values in the contingency table and the Expected values are calculated from what we would have expected to get in each cell if $\pi_{left} = \pi_{right}$ (if the null hypothesis was true).

(9) Assuming the null hypothesis is true, use the overall proportion of success you found earlier to calculate the expected values for the 2 x 2 table below.

```
# Number of left side sleepers who have/do not have nightmares
c(n_left*phat, n_left*(1-phat))
```

```
## [1]  7.333333 14.666667
```

```
# Number of right side sleepers who have/do not have nightmares
c(n_right*phat, n_right*(1-phat))
```

```
## [1] 13.66667 27.33333
```

(10) How many left-side (right-side) sleepers do we expect to have/not have nightmares?

(11) For the theory-based approach we compare the statistic to our reference distribution which is the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns in the contingency table, provided we have at least 10 observations **in each cell**. Calculate the degrees of freedom for this example:

(12) Calculate the χ^2 statistic and p-value.

```
NM <- data.frame(L=c(11,11),
                 R=c(31,10),
                 row.names = c("no","yes"))
EXP <- data.frame(L=c(n_left*(1-phat),n_left*phat),
                  R=c(n_right*(1-phat),n_right*phat),
                  row.names = c("no","yes"))
NM
```

```
##      L  R
## no  11 31
## yes 11 10
```

```
EXP
```

```
##             L        R
## no  14.666667 27.33333
## yes  7.333333 13.66667
```

```
my.chisq <- sum( (NM-EXP)^2 / EXP)
my.chisq
```

```
## [1] 4.22561
```

```
pchisq(my.chisq, 1, lower.tail = FALSE)
```

```
## [1] 0.03981831
```

```
pchisq(z^2, 1, lower.tail = FALSE)
```

```
## [1] 0.03981831
```

```
## Chi-square test (same as z test for two sample proportions)
# my.table <- table(nightmare$Nightmare,nightmare$Side)   # Variable order matters; don't use this one!
T2 <- table(nightmare$Side,nightmare$Nightmare)  # use this one!
pptest <- prop.test(T2,correct = FALSE)
pptest
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  T2
## X-squared = 4.2256, df = 1, p-value = 0.03982
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.502940531 -0.009254591
## sample estimates:
##    prop 1    prop 2
## 0.5000000 0.7560976
```

(13) What do you notice about the value for z from the two-sample z-test and the value for χ^2? 1

Choice 3: Relative Risk Ratio (Ratio of two proportions) Used especially for rare diseases. This is because, as your book points out, if you are working with a rare disease and you find that the incidence rate increases from 1% to 2%, this is a difference of only one percentage point, but it represents a doubling of the rate (a 100% increase).

(14) **The statistic**: How do we calculate and interpret the RR? Ratio of conditional proportions of success $\frac{\hat{p}_{left}}{\hat{p}_{right}}$ with the larger proportion in the numerator for ease of interpretation,i.e., how many times the conditional proportion of success is larger in one group compared to the other. Under H_0 there is no difference in the proportion of successes between groups. What is the relative risk in this situation?

(15) Calculate the relative risk ratio below and interpret the result.

```
## Relative risk ratio
phat_left/phat_right
```

```
## [1] 2.05
```

(16) What are some issues with having a theory based approach with the relative risk ratio?

Choice 4: Odds Ratio

(17) **The statistic**: Yet another statistic (that we will see is quite useful in some cases) is the odds ratio.

- The odds ratio is formed by comparing the odds of success from group 1 to the odds of success from group 2. So from our data set, we compute the odds of success for subjects who sleep on their left and those that sleep on their right, and then use those odds to calculate the odds ratio. Odds are $p/1-p$ where p is a proportion or probability.
- In English it is common to use probability and odds interchangeably but they are NOT the same.
- Calculate the odds ratio for our nightmare dataset and interpret the results.

```
## Odds ratio
odds_left <- T1[2,1]/T1[1,1]       # Odds of having a nightmare sleeping on the left side
odds_right <- T1[2,2]/T1[1,2]       # Odds of having a nightmare sleeping on the right side.
OR <- odds_left/odds_right
c(odds_left, odds_right, OR)
```

```
## [1] 1.0000000 0.3225806 3.1000000
```

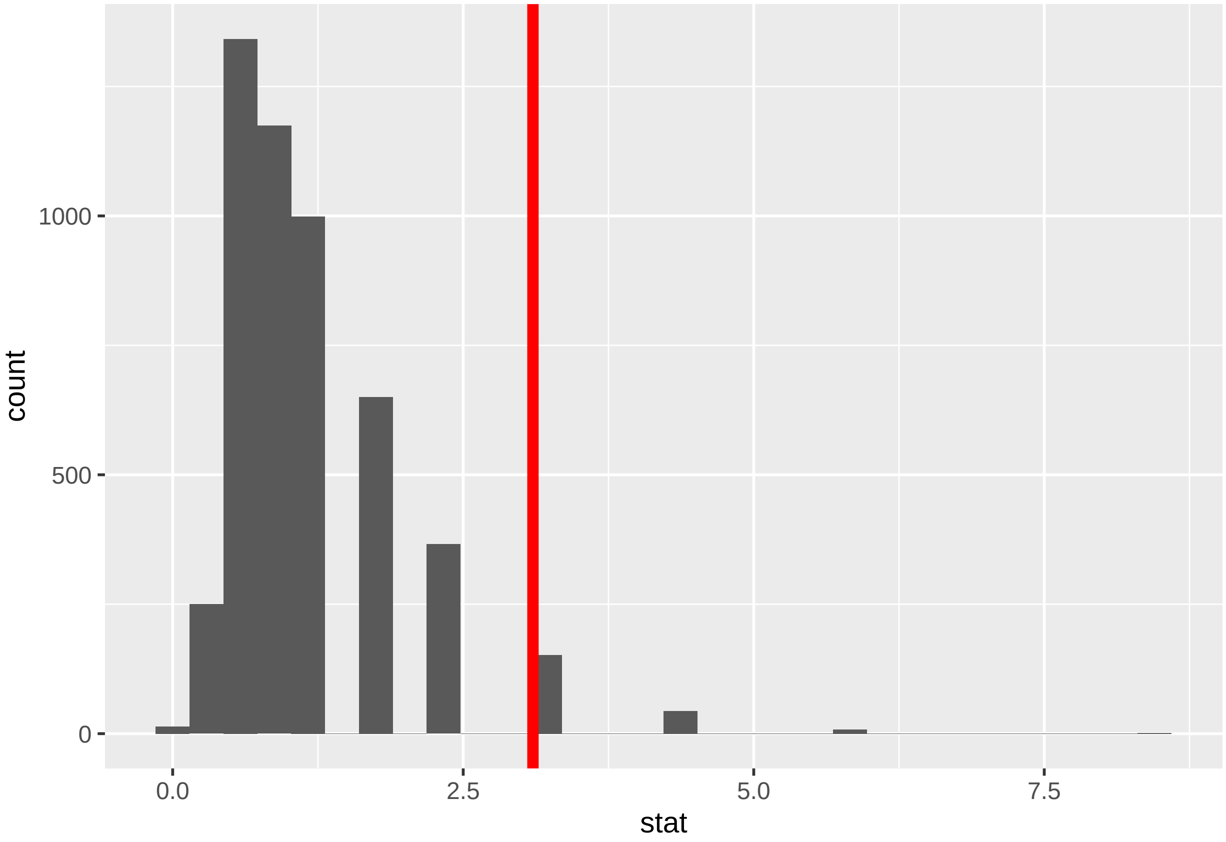
(18) How rare is this under H_0? We can use our shuffling simulation strategy to estimate:

```
set.seed(376)

M <- 5000
results <- data.frame(stat = rep(NA,M))

for(i in 1:M){
  nm.mod <- nightmare %>% mutate(shuff.cat = sample(Side))
  my.table <- table(nm.mod$Nightmare, nm.mod$shuff.cat)
  p.tabl <- prop.table(my.table,2)
  odds.sim.left <- p.tabl[2,1] / p.tabl[1,1]
  odds.sim.right <- p.tabl[2,2] / p.tabl[1,2]
  results$stat[i] <- odds.sim.left / odds.sim.right
}

results %>% ggplot(aes(x=stat)) + geom_histogram()+
  geom_vline(xintercept = OR, lwd = 2, color = "red")
```



```
sum(results$stat >= OR) / M
```

```
## [1] 0.041
```

(19) Under the null hypothesis H_0 (there is no difference in the proportion of successes between groups), what should the the Ratio between `Odds_left/odds_right` be equal to?

(20) In order to use a theory based test involving Odds, it turns out that it often times is better to use log-Odds. Though harder to interpret. The OR is generally symmetric, it turns out that log-Odds converges super quickly under the CLT to a normal distribution, which makes life really nice.

Next lesson we will continue to use log-odds as a statistic. Our statistical model will called a logistic regression model. This can be fit in R using an extremely flexible class of models known as a Generalized Linear Models.

```
nightmare <- nightmare %>% mutate(nite.bin = ifelse(Nightmare == "Yes", 1, 0))
my.glm <- glm(nite.bin ~ Side, data = nightmare, family = "binomial")
summary(my.glm)
```

```
##
## Call:
## glm(formula = nite.bin ~ Side, family = "binomial", data = nightmare)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1774  -0.7478  -0.7478   1.1774   1.6799
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.597e-15  4.264e-01   0.000   1.0000
## SideR       -1.131e+00  5.604e-01  -2.019   0.0435 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80.201  on 62  degrees of freedom
## Residual deviance: 76.052  on 61  degrees of freedom
## AIC: 80.052
##
## Number of Fisher Scoring iterations: 4
```

```r
my.glm.coefs <- coef(my.glm)
my.glm.coefs
```

```
##   (Intercept)         SideR
## -1.597444e-15 -1.131402e+00
```

```r
exp(abs(my.glm.coefs[1]))
```

```
## (Intercept)
##           1
```

```r
exp(sum(my.glm.coefs))
```

```
## [1] 0.3225806
```

(21) What is the statistical model we are considering now?