Notes and Exercises Chapter 18 – Inference for Two-Way Tables

Name ____________________________
Date ____________________________

Math 217
Saidi

Computing expected counts in a two-way table

To calculate the expected count for the ith row and jth column, compute

$$Expected\ Count_{row_i,col_j} = \frac{(row\ i\ total) * (column\ j\ total)}{table\ total}$$

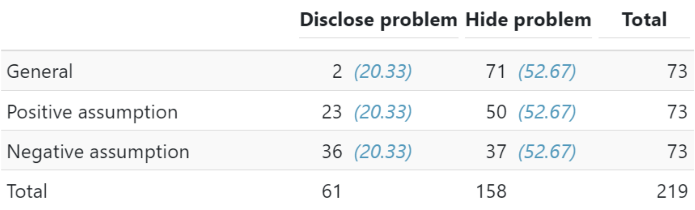
Example 1: Unbeknownst to the participants who were the sellers in the study, the buyers were collaborating with the researchers to evaluate the influence of different questions on the likelihood of getting the sellers to disclose the past issues with the iPad. The scripted buyers started with "Okay, I guess I'm supposed to go first. So you've had the iPad for 2 years …" and ended with one of three questions:

- General: What can you tell me about it?
- Positive Assumption: It does not have any problems, does it?
- Negative Assumption: What problems does it have?

The question is the treatment given to the sellers, and the response is whether the question prompted them to disclose the freezing issue with the iPod. The results are shown in the table below, and the data suggest that asking the, *What problems does it have?*, was the most effective at getting the seller to disclose the past freezing issues. However, you should also be asking yourself: could we see these results due to chance alone if there really is no difference in the question asked, or is this in fact evidence that some questions are more effective for getting at the truth?

Question	Disclose problem	Hide problem	Total
General	2	71	73
Positive assumption	23	50	73
Negative assumption	36	37	73
Total	61	158	219

Compute and include the expected counts for each cell:



	Disclose problem	Hide problem	Total
General	2 *(20.33)*	71 *(52.67)*	73
Positive assumption	23 *(20.33)*	50 *(52.67)*	73
Negative assumption	36 *(20.33)*	37 *(52.67)*	73
Total	61	158	219

$$\chi^2 = \sum_{i=1}^{k} \frac{(observed_i - expected_i)^2}{expected_i}$$

Test for Independence

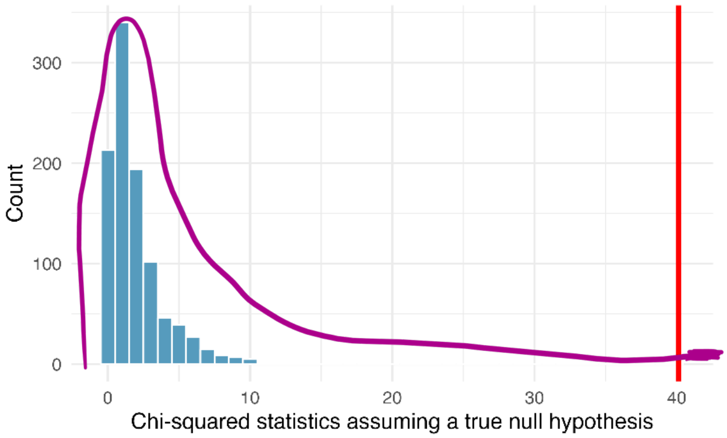
Chi-Square tests may be used to assess independence between two groups. Used in this sense, the null and alternative are:

Ho: The two different variables being measured are independent of each other
Ha: There is some dependence between the two groups being measured

Understand the variability

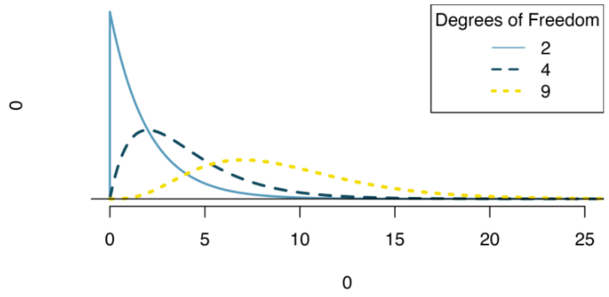
As before, one randomization will not be sufficient for understanding if the observed data are particularly different from the expected chi-squared statistics when Ho is true. To investigate whether 40.13 is large enough to indicate the observed and expected counts are substantially different, we need to understand the variability in the values of the chi-squared statistic we would expect to see if the null hypothesis was true.

The figure below plots 1,000 chi-squared statistics generated under the null hypothesis. We can see that the observed value is so far from the null statistics that the simulated p-value is zero. That is, the probability of seeing the observed statistic when the null hypothesis is true is virtually zero. In this case we can conclude that the decision of whether to disclose the iPod's problem is changed by the question asked. We use the causal language of "changed" because the study was an experiment. Note that with a chi-squared test, we only know that the two variables (question_class and response) are related (i.e., not independent). We are not able to claim which type of question causes which type of response.



This histogram of chi-squared statistics from 1,000 simulations produced under the null hypothesis, H0, where the question is independent of the response. The observed statistic of 40.13 is marked by the red line. None of the 1,000 simulations had a chi-squared value of at least 40.13. In fact, none of the simulated chi-squared statistics came anywhere close to the observed statistic!

The Chi-Square Distribution



Conditions for Chi-Square:

- Independent observations
- Sufficiently large samples: 5 expected counts in each cell
- df = (number of rows minus 1)×(number of columns minus 1)

R can be used to find the p-value with the function `chisq.test()`.

```
general <- c(2,71)
pos <- c(23, 50)
neg <- c(36, 37)
chi_df <- data.frame(general, pos, neg)
chi_df

chisq.test(chi_df)
```

Output:

```
> general <- c(2,71)
> pos <- c(23, 50)
> neg <- c(36, 37)
> chi_df <- data.frame(general, pos, neg)
> chi_df
  general pos neg
1       2  23  36
2      71  50  37
> chisq.test(chi_df)

        Pearson's Chi-squared test

data:  chi_df
X-squared = 40.128, df = 2, p-value = 1.933e-09
```

Conclusion: The p-value is very small. Reject the null. There is very strong evidence that the type of question does effect the rates of the customer disclosing the problem.

Or

There is very strong evidence that rates of disclosing problems for the iPod are dependent on the way the questions were phrased.

Example 2

Extracorporeal membrane oxygenation (ECMO) is a potentially life-saving procedure that is used to treat newborn babies who suffer from severe respiratory failure. An experiment in which 145 babies were treated with ECMO and 30 babies were treated with conventional therapy (CMT) show results:

	Treatment		
	CMT	ECMO	Total
Died	12	5	17
Lived	18	140	158
Total	30	145	175

If the null hypothesis is true that there is no difference in babies who survive, independent of the treatment, then we can think of the two column headings as arbitrary labels. **Therefore, we could conduct a randomization (non-parametric) test** to find the probability that either of these two labels would arise by chance, given the marginal totals are fixed.

Try this by simulating shuffling cards with these outcomes:

http://www.rossmanchance.com/applets/ChisqShuffle.htm?FET=1

Try it for the CMT versus ECMO treatment data.

1. State the null and alternative hypotheses in words.

2. Enter the data as a matrix on the calculator
3. Perform the Chi-Square test and get a p-value.
4. Write your conclusion in context.

Using R

```
died <- c(12,5)
lived <- c(18,140)
test <- data.frame(died, lived)

test

chisq.test(test)
```

OUTPUT:

```
> died <- c(12,5)
> lived <- c(18,140)
> test <- data.frame(died, lived)
> test
  died lived
1   12    18
2    5   140
> chisq.test(test)

        Pearson's Chi-squared test with Yates' continuity correction

data:  test
X-squared = 33.812, df = 1, p-value = 6.07e-09

Warning message:
In chisq.test(test) : Chi-squared approximation may be incorrect
```

Conclusion:____

Example 3: Table 18.4 summarizes the results of an experiment evaluating three treatments for Type 2 Diabetes in patients aged 10-17 who were being treated with metformin. The three treatments considered were continued treatment with metformin (met), treatment with metformin combined with rosiglitazone (rosi), or a lifestyle intervention program. Each patient had a primary outcome, which was either lacked glycemic control (failure) or did not lack that control (success). What are appropriate hypotheses for this test? Complete steps to conclude whether the treatment type has an effect on glycemic control. The diabetes2 data can be found in the **openintro** R package.

Treatment	Failure	Success	Total
Lifestyle	109	125	234
Met	120	112	232
Rosi	90	143	233
Total	319	380	699

Ho:

Ha:

Relative Risk and the Odds Ratio

There are 4 ways to test whether 2 population proportions $p_1 - p_2$ are equal

1. 2 x 2 Chi Square Table
2. Hypothesis Test/Confidence interval $(p_1 - p_2)$
3. Relative Risk
4. Odds Ratio

Relative Risk

When the ratio of the probabilities has a negative outcome, this ratio is called the **relative risk,** $\frac{p_1}{p_2}$**.**

Example 4: (from Significance Magazine 2/21 pg 35)

"One aspect of the disparity in Covid impact that remains mysterious is what has happened within healthcare. For example, a study of more than 100 deaths of health services staff as of 22 April 2020 found that 94% of doctors who died were non-white (bit.ly/35eTxcO). But only 44% of doctors are non-white. The same study found that non-white nurses and midwives are only 20% of their profession, but they made up 71% of those who died with Covid."

$\Pr(Non - white|\, Doctor\ who\ died) = .94$
$\Pr(Non - white|Doctor) = .44$

The following table provides simulated data from the values provided in the article with a sample size of 1000.

Simulated Data	Non-white Doctors	White Doctors	Totals
Died of Covid	53	3	57
Survived Covid	387	557	943
Totals	440	560	1000

a. Find $\hat{p}_1$: Pr(died | non-white)

b. Find $\hat{p}_2$: Pr(died | white)

c. Calculate the ***estimated relative risk* for a non-white doctor to die of Covid.** $\hat{p}_1 / \hat{p}_2$

d. Interpret the relative risk value.

Example 5: An observational study middle-age male smokers and former smokers tracked over many years to see how many developed lung cancer.

a. Find $\hat{p}_1$: Pr(lung cancer | smoker)

b. Find $\hat{p}_2$: Pr(lung cancer | former smoker)

	Smoking history	
	Smoker	Former Smoker
Lung Cancer	89	37
No Lung Cancer	6063	5711
Total	6152	5748

c. What is the ***estimated relative risk:*** $\hat{p}_1 / \hat{p}_2$

d. Interpret the relative risk value.

Odds Ratio

Odds of an event *E* is defined as: $odds\ of\ E = \dfrac{\Pr\{E\}}{1-\Pr\{E\}}$

Example: if the probability of an event is 2/3, then the odds of the event are: $\dfrac{2/3}{1/3} = 2:1$

The **odds ratio, $\hat{\theta}$,** is the ratio of odds under 2 conditions.

$$\hat{\theta} = \frac{\dfrac{\hat{p}_1}{1-\hat{p}_1}}{\dfrac{\hat{p}_2}{1-\hat{p}_2}}$$

Odds Ratio Shortcut:

$$\hat{\theta} = \frac{a * d}{b * c}$$

	B	B’
A	a	b
A’	c	d

From the same two-way table above, the odds of developing lung cancer.

Chapter 18 - Chi Square

R Saidi

Load the libraries and data

```r
library(tidyverse)
library(openintro)
library(tidymodels)
data("gss")
```

A question in two variables

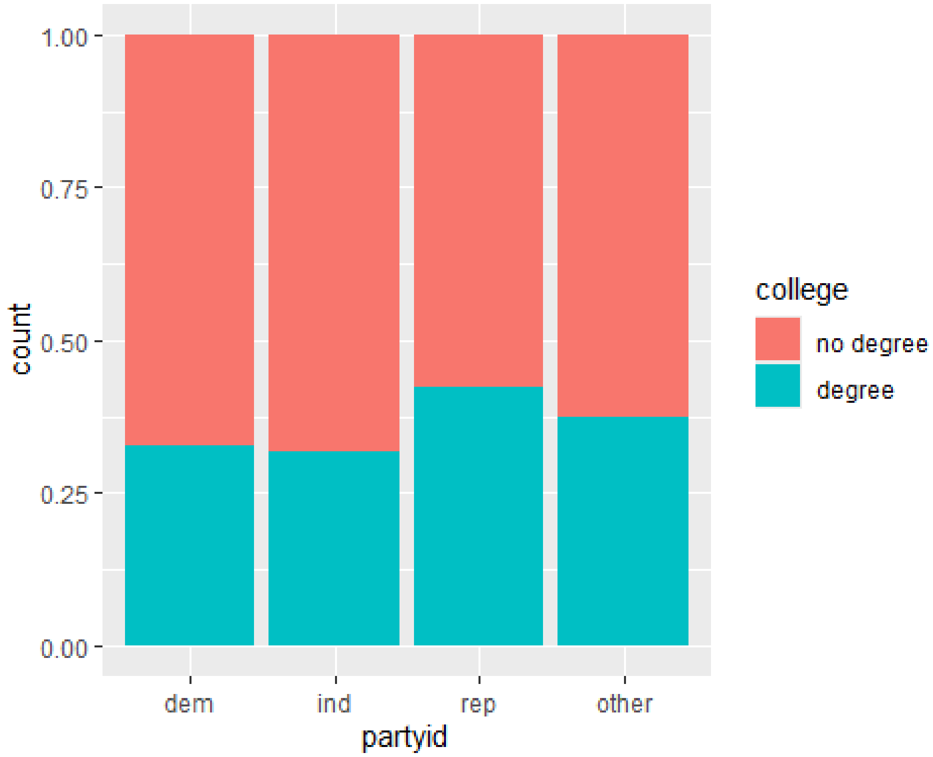
Does level of education have an association with political party affiliation?

Chi Square

When we are looking at a two-way table, we can explore the χ^2 distribution

Start with a bar plot

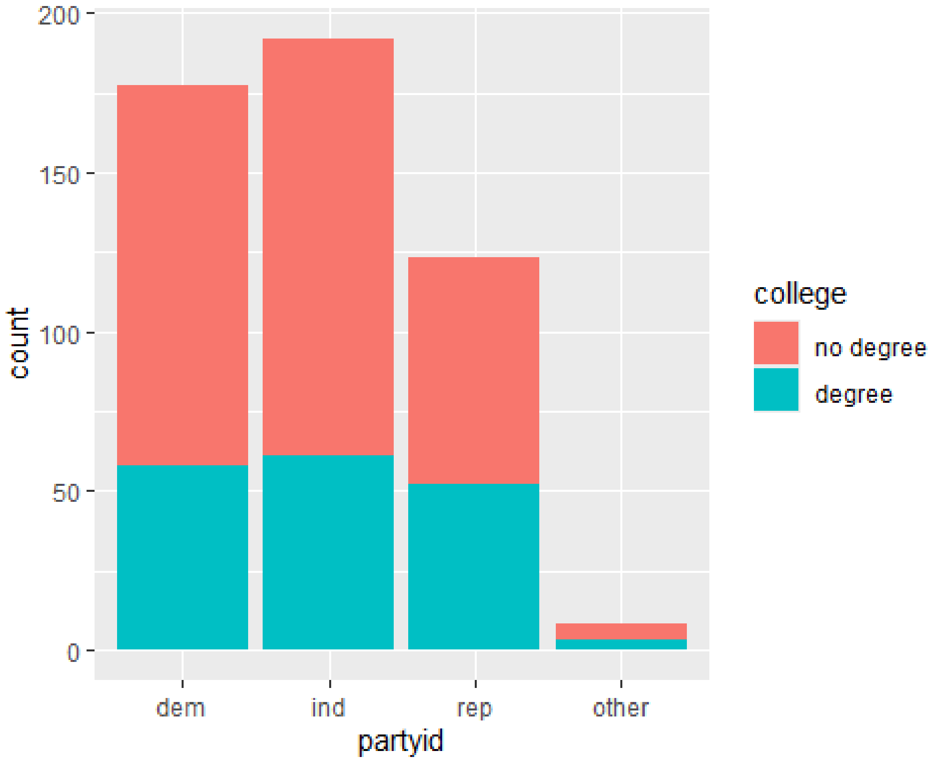
```r
# Visualize distribution
gss |>
  ggplot(aes(x = partyid, fill = college)) +
  # Add bar layer of proportions
  geom_bar(position = "fill")
```



The education proportions for each party look relatively similar.

Remove position = “fill”

```
gss |>
  ggplot(aes(x = partyid, fill = college)) +
  # Add bar layer of proportions
  geom_bar()
```



base R table

```
obs_table <- table(gss$college, gss$partyid)
obs_table


             dem ind rep other  DK
  no degree 119 131  71     5   0
  degree     58  61  52     3   0

# From previous step
obs <- gss |>
  select(college, partyid) |>
  tibble::as_tibble() |>
  table()
obs

           partyid
college      dem ind rep other  DK
  no degree 119 131  71     5   0
  degree     58  61  52     3   0
```

What is the DK?

DK seems to be some anomaly that needs to be removed.

```
gss$partyid <- as.character(gss$partyid) |>
  trimws() |>
  as.factor()

# Now check that DK is removed
unique(gss$partyid)

[1] ind   rep   dem   other
Levels: dem ind other rep
```

Convert table back to tidy df

```
obs |>
  # Tidy the table
  tidy() |>
  # Expand out the counts
  uncount(n)

Warning in tidy.table(obs): 'tidy.table' is deprecated.
Use 'tibble::as_tibble()' instead.
See help("Deprecated")

# A tibble: 500 × 2
   college   partyid
   <chr>     <chr>
 1 no degree dem
 2 no degree dem
 3 no degree dem
 4 no degree dem
 5 no degree dem
 6 no degree dem
 7 no degree dem
 8 no degree dem
 9 no degree dem
10 no degree dem
# ℹ 490 more rows
```

Perform a chi-squre hypothesis test

1. Calculate the observed Chi-Square statistic

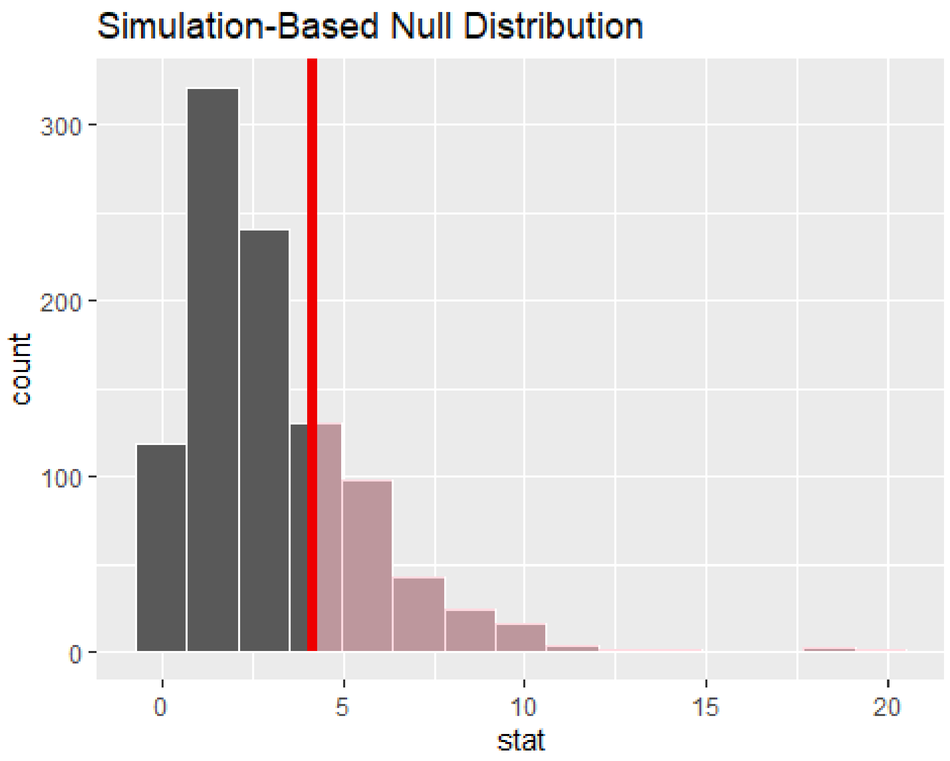
```
# calculate the observed statistic
observed_indep_statistic <- gss |>
  specify(partyid ~ college) |>
  hypothesize(null = "independence") |>
  calculate(stat = "Chisq")
observed_indep_statistic

Response: partyid (factor)
Explanatory: college (factor)
Null Hypothesis: ind...
# A tibble: 1 × 1
   stat
  <dbl>
1  4.15
```

```
# Create one permuted data set
perm_null <- gss |>
  specify(partyid ~ college) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "Chisq")
perm_null

Response: partyid (factor)
Explanatory: college (factor)
Null Hypothesis: ind...
# A tibble: 1,000 × 2
   replicate  stat
       <int> <dbl>
 1         1 2.62
 2         2 0.152
 3         3 8.85
 4         4 5.24
 5         5 6.08
 6         6 9.55
 7         7 7.78
 8         8 0.875
 9         9 7.87
10        10 1.42
# ℹ 990 more rows

perm_null |>
  visualize() +
  shade_p_value(observed_indep_statistic,
    direction = "greater"
  )
```



Compare to base R

```
# Compute chi-squared stat
chisq.test(gss$partyid, gss$college)

Warning in chisq.test(gss$partyid, gss$college): Chi-squared approximation may
be incorrect


    Pearson's Chi-squared test

data:  gss$partyid and gss$college
X-squared = 4.1543, df = 3, p-value = 0.2453
```

With a p-value of 0.3608, there is no compelling evidence that there is an association between having a college degree or not and the political party affiliation.

Fisher’s Exact Test

When basic assumptions for Chi-Square test (expected cell counts >= 5) are violated, we can try using the comparable non-parametric Fisher’s Exact Test.

(rowtot C success)*(rowtot C failure) / (samplesize C truetotal)

Example Below

yes 4 3 7 no 13 8 21 total 17 11 28

FE = ((7 C 4) * (21 C 13)) / (28 C 17)

```
# fisher.test(v1, v2)
fisher.test(gss$college, gss$partyid)
```

```
    Fisher's Exact Test for Count Data

data:  gss$college and gss$partyid
p-value = 0.2388
alternative hypothesis: two.sided
```

Homework Chapter 18

1. Review section 18.3 (the chapter review)
2. Suggested problems from textbook section 18.4 exercises: 1, 5, 10, 13, 14
3. Suggested tutorials:

3 - Chi-squared test of independence