PS303: Week 5

pp.358-363; 370-374

Categorical data

- Our previous discussion involved how to assess whether observed distributions of *binary* data (*successes* relative to *failures*) varied statistically from expected distributions using p-values generated from *binomial* tests.
  - Useful for data that can be split into 'successes' and 'fails'.
  - When observations vary from expectations with $p < .05$, the effect is considered statistically significant.
- Not all categorical data is binary however (e.g., seasons of the year, personality traits, political ideologies). Non-binary categorical data can be tested against null hypotheses using **Chi-Square** (χ^2) tests. Some versions of this are:
  - χ^2 test of *independence*, which measures whether measured variables are statistically independent of one other.
  - χ^2 test of *goodness-of-fit*, which measures 'how close' observed data is relative to expected data (*expected* patterns can include multiple probabilities, implied by the null H_O, which *must* add up to 1).
  - Fisher's exact test for estimating differences across samples with low sample size (e.g., if any cell in your data structure is less than 5, do not estimate χ^2).
  - McNemar test for *paired* categories (since χ^2 is a statistic for *independent* categories)

- *Remember*: *NHST*s inform you how likely the observed data would be *if* the fictional null hypothesis is true. *NHST* p-values cannot tell us whether the alternative hypothesis is 'true'. It's good practice to include *power* estimations and *effect sizes* to increase the interpretability of our findings

χ^2 test of *independence*

Are two categories statistically related to each other?

- You've been watching rugby matches between Namosi and Serua teams for the past year. You suspect that the weather may be be disproportionately influencing one team's performance over the other.
- Your *research hypothesis* is that the Namosi and Serua teams are variably influenced by alternating weather patterns. Your statistical H_O is that there is *no significant relationship* between the two categorical variables (Weather ~ Location).

You have collected the following data:

Weather	*Namosi*	*Serua*
Sunny	23	26
Raining	36	19
Windy	44	68

We can run a χ^2 test of independence to determine whether the difference in wins between Namosi and Serua is statistically significant.

- H_O : Namosi and Serua are equally likely to win independently of weather conditions
- H_A : Namosi and Serua wins are influenced by weather conditions

3 steps to running the test on *R*...

```
# Step 1: Bind data together and assign to a variable
V <- cbind(c(23,36,44), # Namosi
           c(26,19,68)) # Serua

# Step 2: Name dimensions and their individual
dimnames(V)<-list(weather=c("sunny","raining","windy"), # Row identifiers in 'V'
                  location= c("Namosi","Serua"))        # Column identifiers in 'V'

# Step 3: Run chi-square test
chisq.test(V)
```

```
##
##  Pearson's Chi-squared test
##
## data:  V
## X-squared = 10.14, df = 2, p-value = 0.006283
```

We can report this as follows:

A χ^2 test of independence indicated a significant association between location and weather, $\chi^2(2, n = 216) = 10.14$, $p = .006$. Inspection of findings revealed Serua was more likely to win when windy, whereas Namosi was more likely to win when raining.

Across a test of *independence*, we determine whether sample parameter distributions are associated with each other. A non-significant difference implies that the tested distributions vary independently of each other. χ^2 tests are *always* one-sided - the larger the χ^2 estimate, the greater our chances of rejecting the null (which describes our *expected* distribution - see p. 358 for details).

χ^2 test of *goodness-of-fit*

- Across a *goodness-of-fit* (GOF) test, we want to know whether sample distributions vary relative to an *expected* population distribution. In the first example below, we assume wins are equally likely for the null hypothesis. Later on, we describe how to alter our parameters if we interested in non-symmetrical null hypotheses (e.g., some regions may be expected to produce greater wins/losses - *more on this in a moment*).

How closely do observed sample distributions match ('fit with') expected population distributions?

- You have been hired to scout new talent for the Fijian 7s from local teams in the Naitasiri, Namosi, Rewa, Serua and Tailevu regions. However, you only have time to visit a couple of regions before making your recommendations to the team manager. The only information provided to you is the following:
  - The **total games** won by all the teams together ($N_{Total} = 200$).
  - How many games **each region** has won (see table below).

Assuming each region is equally likely to win, we can expect each team should win $\frac{N_{Total}}{N_{Teams}} = \frac{200}{5} = 40$ games. Can you use this information to determine which region(s) may be worth visiting?

Regions	$Observed_{wins}$	$Expected_{wins}$
Naitasiri	36	40
Namosi	38	40
Rewa	41	40
Serua	40	40
Tailevu	45	40

We can use a *goodness-of-fit* test to determine whether the **observed distribution** of games won varies from your **expected distribution**.

As there are five regions, we should have five probabilities which should sum up to 1 ($\sum_{i=1}^{k=5}(P_i) = 1$). If we assume a symmetrical null hypothesis (all regions are equally likely to be associated with wins), the probability for any single region (P_i) winning a game would be 0.2 (because $P_{i=1} + P_{i=2}...P_{i=5} = 1$). The *expected* frequency of wins can be estimated by multiplying the probability with the total frequency ($.2 \times 200 = 40$).

Manually estimating *goodness-of-fit*

- We subtract our expected frequency of games won ($Expected_{wins}$, or E_{wins}) from the observed/actual number of games won ($Observed_{wins}$, or O_{wins}) to produce a difference score.
- We **square** this difference to remove all negative signs, then divide the difference by the expected wins to produce **error** terms (final column).
- The latter represents 'how much' our null H_O was in error with respect to predicting the outcome of each region's performance.

Regions	O_{wins}	E_{wins}	$O_{wins} - E_{wins}$	$(O_{wins} - E_{wins})^2$	$(O_{wins} - E_{wins})^2/E_{wins}$
$\text{Naitasiri}_{i=1}$	36	40	-4	16	.4
$\text{Namosi}_{i=2}$	38	40	-2	4	.1
$\text{Rewa}_{i=3}$	41	40	1	1	.025
$\text{Serua}_{i=4}$	40	40	0	0	0
$\text{Tailevu}_{i=5}$	45	40	5	25	.625

We can report the χ^2 GOF statistic by adding up each ith error across the total (k) number of categories, giving us $\chi^2 = .4 + .1 + .025 + 0 + .625 = 1.15$.

The formula can be summarized as follows:

$$\chi^2 = \sum \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

where i represents each value iteration.

Let's run the same procedure in R...

```
chisq.test(x = c(36,38,41,40,45),    # Vector with observed frequencies
           p = c(40,40,40,40,40),    # Vector of expected frequencies
           rescale.p = T)            # Tell R to convert expected frequencies into probabilities
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(36, 38, 41, 40, 45)
## X-squared = 1.15, df = 4, p-value = 0.8863
```

Is this large enough to *reject* our null hypothesis, which claimed each team had an equal probability of winning (so $H_O : P_{wins} = .2$ & $H_A : P_{wins} \neq .2$?). Since $p > .05$, we **cannot** reject H_O. We can report these results as follows:

Out of 200 games won by regional teams, Naitasiri won 36 games, Namosi won 38 games, Rewa won 41 games, Serua won 40 games and Tailevu won 45 games. A chi-square goodness of fit test indicated the probability of winning a rugby game did not statistically vary between regions, $\chi^2(4) = 1.15, p = .89$. The null hypothesis (that all regions were equally associated with wins) is retained.

For the sake of illustration, how would we alter the probabilities if we assumed (for whatever reason) that twice the number of games are likely to be won in Naitasiri relative to remaining regions?

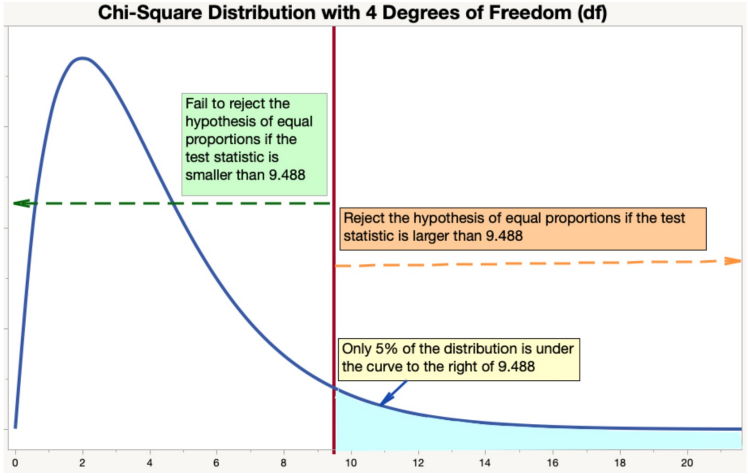
```
chisq.test(x = c(36,38,41,40,45),
           p = c(80,30,30,30,30), # Note the altered frequencies relative to our updated null
           rescale.p = T)         # Think about what a rejection/retention of the null here may imply
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(36, 38, 41, 40, 45)
## X-squared = 41.2, df = 4, p-value = 2.443e-08
```

How is χ^2 interpreted?

- Remember that every test statistic follows a distribution. Knowing the specific distribution tells us where the observed statistic stands relative to some critical value.
- A large χ^2 statistic implies that the null hypothesis did not do a good job predicting the data, meaning we can *reject* the null.

- p-values indicate whether the χ^2 estimate is less than ($p > .05$ - retain H_O) or more than ($p < .05$ - reject H_O) the critical value. Larger test statistics imply greater distance from the center of the distribution, meaning the null is less likely (corresponding with a smaller p-value).



The degrees of freedom (df) for estimating a χ^2 distribution is determined by $k-1$ where k is the total number of categories being measured.

Effect sizes

- Recall that we report effects to illustrate the **magnitude** of observed effects (also described as *practical* significance)
- For chi-square tests, Cramer's V is a typical measure of effect size. We can estimate V with known parameters:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

```
# Computing Cramer's V
chi   <- 1.15  # Value for Chi-square
n_tot <- 200   # Total sample
k <-5          # Total Number of groups
cram_v <- (chi^2/n_tot*(k-1))^.5  # Re-writing the formula in R
cram_v         # Print
```

- Inputting our previous values ($\chi^2 = 1.15; N = 200; k = 5$) generates $V = 0.16$, which seems small (interpret the effect size as you would a correlation coefficient). This is not unexpected since our test was non-significant!

Assumptions of a χ^2 test

- Expected frequencies (for each cell) should be greater than 5.
- Observations should be independent of each other.

If 20% of data cells contain 5 or fewer iterations, we can use Fisher's exact test. Alternatively, if we want to test across paired categories, we may use McNemar's test.

Fisher's exact test

You suspect that facebook-users are more depressed relative non-users. You distribute a survey to 14 friends to measure whether they have depression and use facebook. You create a table with your collected data:

Depression	Facebook user	Not a Facebook user	*Total*
Present	7	1	$\sum Row_1 = 8$
Absent	3	3	$\sum Row_2 = 6$
Total	$\sum Col_1 = 10$	$\sum Col_2 = 4$	$N = 14$

Because some of the cell counts are less than 5, a conventional χ^2 test may be unreliable. Instead, Fisher's exact test can determine if there is an association between the variables present (similar to a test of independence but for a small sample) by 'directly' estimating a p-value (no test statistics involved).

```
## Running Fisher's exact test

# Step 1: Combine cells into columns and bind them
V1 = cbind(c(7,3),c(1,4))

# Step 2: Name dimensions + levels
dimnames(V1) = list(Depression = c("Yes","No"),Facebook = c("User","Non-user"))

# Step 3: Run the test
fisher.test(V1,alternative="two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  V1
## p-value = 0.1189
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.5027782 521.2731861
## sample estimates:
## odds ratio
##   7.876343
```

From the outcome viewed, we can claim that the null hypothesis may *not* be rejected.

The formula for Fisher's exact test is

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!N!}$$

where (a, b, c, d) are the cell values (so $a = 7, b = 3, c = 1, d = 4$) and N is the total frequency ($N = 14$). We can re-write the above as follows

$$p = \frac{(7+3)!(1+4)!(7+1)!(3+4)!}{7!3!1!4!14!} = .1189$$

McNemar Test for symmetry

> All our tests so far have relied on the assumption of independence between samples. What if you are collecting categorical data from the same sample over repeated iterations?

You want to set up an information campaign across USP students to reduce Styrofoam littering. You decide to show a video by a conservationist philosopher on the importance of beauty and *oikophilia* (love of one's home) to see if that can reduce littering behavior. Namely, you want to test whether realizing the value of natural beauty can help reduce littering behavior.

To test this idea, you recruit 58 students and monitor their littering behavior for one week. You then show them the video, then monitor their littering behavior during the following week. You want to know whether there is a difference (asymmetry) in littering instances before and after the video. As you are looking at behaviors of the *same* group across multiple time points, your samples are not independent of each other so you cannot run a conventional χ^2 test.

Do you litter?	Before video	After video
Yes	26	16
No	32	42

For *paired* nominal/categorical data, we can run Mcnemar's test for symmetry. We will examine the null hypothesis that the littering instances are not different (symmetrical) before and after watching the video.

```
## Running McNemar's test for symmetry

# Step 1: Prepare the data
V2 = cbind(c(26,32),c(16,42))

# Step 2: Naming dimensions and their levels
dimnames(V2) = list(Litter = c("Yes","No"), Time = c("Before video","After video"))

# Step 3: Run the test
mcnemar.test(V2)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  V2
## McNemar's chi-squared = 4.6875, df = 1, p-value = 0.03038
```

The test statistic is significant (because $p = .03$). We are now justified in estimating Cramer's V to assess whether the statistically significant difference is *practically* important.

```
# Forumula: chi^2/n_tot*(k-1))^.5
cram_v2<-(4.6875/58*(2-1))^.5

# Print (rounded to 2 decimal places)
round(cram_v2,2)
```

```
## [1] 0.28
```

A McNemar test of symmetry indicated the number of participants who reported not littering after watching the video ($N_{After/Yes} = 42$) was statistically different, $\chi^2(1) = 4.69$, $p = .030$; $V_{Cramer} = 0.28$, relative to participants who did not litter before watching the video.

Lab Activity

Copy the first column of the table in slide 4, and answer the following:

1. Enter a new series of observed wins (O_{wins}) for the five regions (Naitasiri = 28 wins, Namosi = 31 wins, Rewa = 37 wins, Serua = 59 wins, and Tailevu = 45 wins). Assume that each region still has an equal probability of winning ($p = .2$ for each team). Run a chi-square goodness of fit test and report whether the *newly* observed distribution of wins retains *or* rejects the null hypothesis. Ensure to construct a table similar to the one shown on slide 4.

2. Retaining the number of (O_{wins}) from Question 1, suppose you receive new information that alters your expectations of which teams are likely to win. Specifically, you learn that the last cyclone washed away the rugby gear of the teams in Rewa and Namosi. You consequently expect teams in Rewa and Namosi would have won fewer matches relative to remaining regions as they had less equipment. You speculate that the *expected* number of wins (E_{wins}) for Rewa and Namosi will be 10 each. This means that the E_{wins} for the three remaining regions will be 60 each (since the *total* win count remains 200 games). Run a chi-square goodness of fit test and report whether the *new* expected distribution of wins influences your outcomes. Make sure to update your probability vector (`p = c(num1, num2, num3...)`) before running the chi square.

Remember to report the effect size for any chi-square tests that are statistically significant ($p < .05$). Submit your completed activity in the dropbox on Moodle.