Live Session 11

Week	Date	Plan
11	07/18	Categorical data analysis
12	07/25	Project Working Day
13	08/01	Project Presentations Final Review
14	08/08	Final Exam (Inclass portion)
15	08/15	**Final Exam (Take home part)**

Final Exam

- Week 14 : During the live session (In class portion)

- Part II (Take Home)
- Thursday August 10, 10.00a.m CT

- Submit on Monday August 14, midnight CT

Live Session 13

- Aug 01
- July 31 or Aug 04

Review from Statistical Methods: Inference for categorical data from iid sample

- $X_1, \ldots, X_n$ are a series of 0's and 1's, where

$$X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ unit has attribute} \\ 0 & \text{otherwise} \end{cases}$$

- Then $X = \sum_{i=1}^{n} X_i$ is the number of units in your sample that has the attribute

1. Estimate proportions
   – Confidence intervals

Confidence interval for proportion

$$\hat{p} = \frac{X}{n} \quad , \quad sd = \sqrt{\frac{p(1-p)}{n}}$$

Normal approximation sufficiently accurate if
np > 5
n(1 - p) > 5

$(1-\alpha)100\%$ Confidence Interval is

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Confidence interval for proportion

Simple random sample of 1,000 Voters
550 Prefer Candidate A
Estimate the True Voter Preference

$$\hat{p} = \frac{550}{1000} = 0.55$$

95% Confidence Interval is

$$0.55 \pm 1.96\sqrt{\frac{0.55(1-0.55)}{1000}}$$

$$0.55 \pm 0.0308 = (0.5192, 0.5808)$$

Review from Statistical Methods: Inference for categorical data

- $X_1, \ldots, X_n$ are a series of 0's and 1's, where

$$X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ unit has attribute} \\ 0 & \text{otherwise} \end{cases}$$

- Then $X = \sum_{i=1}^{n} X_i$ is the number of units in your sample that has the attribute

1. Estimate proportions
   - Confidence intervals
2. Examine relationships between two categorical variables
   - Are they independent?

Independence

Table of badhealth by GENDER			
badhealth	GENDER(gender)		
	1 (male)	2 (female)	Total
1 (yes)	65	134	199
2 (no)	574	754	1328
Total	639	888	1527

Is gender independent of self reported bad health?

Independence

The FREQ Procedure					
Frequency		Table of badhealth by GENDER			
Percent		badhealth	GENDER(gender)		
Row Pct			1 (male)	2 (female)	Total
Col Pct		1 (yes)	65	134	199
			4.26	8.78	
			32.66	67.34	
			10.17	**15.09**	**13.03**
		2 (no)	574	754	1328
			37.59	49.38	
			43.22	56.78	
			89.83	**84.91**	**86.97**
		Total	639	888	1527
			41.85	**58.15**	100

Is gender independent of self reported bad health?

Chi-Square Tests for Count Data

– Two categories are independent if the probability of having both attributes is equal to the product of the probabilities of having each attribute; i.e.,

$$H_0:\ p_{ij} = p_i * p_j$$

$$H_a:\ p_{ij} \neq p_i * p_j$$

$$\boxed{P(A \text{ and } B) = P(A) \times P(B)}$$

Notation for Observed Frequencies

Column Categories

		1	...	j	...	c	Total
	1						
Row	...						
Categories	i			O_{ij}			Row i Total (R_i)
	...						
	r						
	Total			Column j Total (C_j)			T

Notation for Expected Frequencies under H_0

Column Categories

		1	...	j	...	c	Total
Row Categories	1						
	...						
	i			$E_{ij} = T^*(R_i/T)(C_j/T)$			Row i Total (R_i)
	...						
	r						
	Total			Column j Total (C_j)			T

(Pearson's) Chi-square Test Statistic

Reject Ho if $X^2 > \chi^2_\alpha$

$$X^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

X^2 ~ Chi-Square with df = (r-1)*(c-1)

(Good approximation as long as $E_{ij} > 1$ and 80% >=5)

Health and Gender

Observed Frequencies

	male	female	TOTAL
BAD HEALTH	65	134	199
~ BAD HEALTH	574	754	1328
TOTAL	639	888	**1527**

Expected Frequencies

	male	female	TOTAL
BAD HEALTH	83.28	115.72	199
~ BAD HEALTH	555.7	772.28	1328
TOTAL	639	888	1527

$$1527 * \frac{199}{1527} * \frac{639}{1527} = 83.28$$

Health and Gender

Observed Frequencies

	male	female	TOTAL
BAD HEALTH	65	134	199
~ BAD HEALTH	574	754	1328
TOTAL	639	888	1527

Expected Frequencies

	male	female	TOTAL
BAD HEALTH	83.28	115.72	199
~ BAD HEALTH	555.7	772.28	1328
TOTAL	639	888	1527

Chi-square Calculation

$$X^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

	male	female	TOTAL
BAD HEALTH	4.01	2.89	
~ BAD HEALTH	0.601	0.432	
TOTAL		TOTAL =	7.93

Health and Gender

Observed Frequencies

	male	female	TOTAL
BAD HEALTH	65	134	199
~ BAD HEALTH	574	754	1328
TOTAL	639	888	1527

Expected Frequencies

	male	female	TOTAL
BAD HEALTH	83.28	115.72	199
~ BAD HEALTH	555.7	772.28	1328
TOTAL	639	888	1527

Chi-square Calculation

	male	female	TOTAL
BAD HEALTH	4.01	2.89	
~ BAD HEALTH	0.601	0.432	
TOTAL		TOTAL =	7.93

$$\frac{(65-83.28)^2}{83.28} = 4.01$$

Gender and Self-reported Health Status

H_o: Health and gender are independent
H_a: Health and gender are not independent

Reject Ho if X^2 (7.93) > 6.635 ($\alpha = 0.01$, $df = 1$)

Conclusion: **There is sufficient evidence ($p < 0.01$), to conclude that gender and self-reported health are not statistically independent.**

Reason: **A lower proportion of males reported bad health than expected under the hypothesis of independence; also a greater proportion of females reported bad health than expected.**

Independence- (example)

	Under 40	Over 40	Total
Retained	1634	1091	2725
Terminated	391	627	1018
Total	2025	1718	3743

Is Employment Status Independent of Age?

Employment Discrimination

Observed Frequencies

	Under 40	Over 40	Total
Retained	1634	1091	2725
Terminated	391	627	1018
Total	2025	1718	3743

Expected Frequencies

	Under 40	Over 40	Total
Retained	1474.25	1250.75	2725
Terminated	550.75	467.25	1018
Total	2025	1718	3743

Chi-square Calculation

	Under 40	Over 40	Total
Retained	17.31	20.40	
Terminated	46.34	54.62	
Total			138.67

Employment Discrimination

H_o: Employment Status and Age are Independent
H_a: Employment Status and Age are Not Independent

Reject Ho if X^2 (138.67) > 6.635 ($\alpha = 0.01$, df = 1)

Conclusion: There is sufficient evidence ($p < 0.001$), using a significance level of 0.01, to conclude that employment status and age are not statistically independent.

Reason: A greater number of older employees were terminated than expected under the hypothesis of independence.

Drug Usage

Campus Group

Frequency of Drug Use

	Athlete	Campus Organization	Performing Arts	Total
Annually	104	76	101	281
Monthly	39	12	25	76
Total	143	88	126	357

Drug Usage

Observed Frequencies

	Athlete	Campus Organization	Performing Arts	Total
Annually	104	76	101	281
Monthly	39	12	25	76
Total	143	88	126	357

Expected Frequencies

	Athlete	Campus Organization	Performing Arts	Total
Annually	112.56	69.27	99.18	281
Monthly	30.44	18.73	26.82	76
Total	143	88	126	357

Chi-Square Calculation

	Athlete	Campus Organization	Performing Arts	Total
Annually	0.65	0.65	0.03	
Monthly	3.00	2.42	0.12	
Total				6.87

Drug Usage

H_o: Drug Usage and Campus Group are Independent
H_a: Drug Usage and Campus Group are Not Independent

Reject Ho if X^2 (6.87) > 5.991 (α = 0.05, df = 2)

Conclusion: Using a significance level of 0.05, there is sufficient evidence (p =0.032) to conclude that drug usage and campus group are not statistically independent.

Reason: A greater number of athletes and fewer members of campus organizations reported monthly usage of drugs than expected under the hypothesis of independence.

Inference for categorical data from complex samples

- We will reproduce these analyses properly, accounting for complex designs.
- As when making any estimate from a complex sample, weights must be used to compensate for unequal probability of selection
- We will not learn the formulas for standard errors, as they are not straightforward. Instead, a statistical software package that correctly handles sampling data must be used.

Confidence interval for proportion from a complex sample

- The estimator of proportion p is:

$$\hat{p} = \frac{\sum w_i x_i}{\sum w_i} = \frac{\sum \text{sum of weights for units in category}}{\sum \text{sum of all weights}}$$

- Its standard error is estimated using TS for ratios or other method; denoted se($\hat{p}$)
- Confidence interval is $\widehat{p}$ +/- $z_{1-\alpha/2}$ * se($\hat{p}$)

SAS code using PROC SURVEYMEANS

```
proc surveymeans data=hispmales;
weight KWGTR;
strata stratum;
cluster  SECU;
class badhealth;
var badhealth;
run;
```

Data Summary	
Number of Strata	47
Number of Clusters	81
Number of Observations	686
Number of Observations Used	639
Number of Obs with Nonpositive Weights	47
Sum of Weights	2385958

Class Level Information

Statistics						
Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
badhealth	1	65	0.100569	0.013550	0.07303	0.12810
	2	574	0.899431	0.013550	0.87189	0.92696

Compare with 0.1017 and SE = 0.012 ignoring the design

Testing Independence or homogeneity with complex survey data

The FREQ Procedure				
Frequency	Table of badhealth by GENDER			
Percent	badhealth	GENDER(gender)		
Row Pct		1 (male)	2 (female)	Total
Col Pct	1 (yes)	65	134	199
		4.26	8.78	
		32.66	67.[illegible]4	
		10.17	**15.09**	**13.03**
	2 (no)	574	754	1328
		37.59	49.38	
		43.22	56.78	
		89.83	**84.91**	**86.97**
	Total	639	888	1527
		41.85	**58.15**	100

We need estimated percentages to be weighted

Adjusted Chi-square Test Statistic for testing for independence

Step 1: Replace unweighted with weighted estimates in Chi-square statistic

$$X^2 = \sum_{\text{all cells}} \frac{(n\hat{p}_{ij} - n\hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{n\hat{p}_{i\bullet}\hat{p}_{\bullet j}} = n \sum_{\text{all cells}} \frac{(\hat{p}_{ij} - \hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{\hat{p}_{i\bullet}\hat{p}_{\bullet j}}$$

Step 2: Deflate the statistic by a generalized degree of freedom gdeff

$$X^2_{\text{Rao-Scott}} = X^2/\text{gdeff}$$

Step 3: Compare to the usual reference distribution, $X^2_{(r-1)(c-1)}$

SAS code for Chi-square test

```
proc surveyfreq data=hisp;
weight KWGTR;
strata stratum;
cluster  SECU;
table badhealth*gender / chisq;
run;
```

SAS output

Rao-Scott Chi-Square Test	
Pearson Chi-Square	8.2962
Design Correction	0.8716
Rao-Scott Chi-Square	9.5184 = Pearson Chisq/gdeff
DF	1
Pr > ChiSq	0.0020
F Value	9.5184
Num DF	1
Den DF	44
Pr > F	0.0035
Sample Size = 1527	

Comparison

- In this case, the results of the test of independence are the same; we conclude gender and bad health are not independent, with women more likely to report bad health than men.
- This will not always be true; design features can change conclusions about relationships, just as it can change the inference about single means and proportions
- Thus, if a complex design is used, take account of it in analysis of relationships

Bottom line…

- Always use design information for estimation of parameters, whether they describe means, proportions, or relationships
- Ignoring weights can lead to bias, and (typically) underestimation of standard errors