Chapter 10: Categorical Data

- **Categorical Variable**: Groups units into different categories
  - Oral contraceptive user vs. Non OC-user
  - has cancer vs. does not have cancer etc.
- Ex.) Test if cancer incidence is same between OC users and non-OC users?
  - Association b/t 2 categorical variables
- Ex.) Test if heavy OC users, light OC users, non-OC users have same cancer rates?
- Ex.)

Diastolic Blood Pressure	Frequency	Expected Prob
< 50	57	1%
$\geq 50, < 60$	330	7%
$\geq 60, < 70$	2132	30%
$\geq 70, < 80$	4584	62%

  - Are the observed frequencies consistent with the theoretical probabilities?

Two-Sample Test for Binomial Proportions

Ex.)

Age at first Birth vs. breast cancer incidence

		Control (No Cancer)	Case (Cancer)
Age at 1st Birth	≤ 29 years	8747	2537
	≥ 30 years	1498	683
	Total	10245	3220

Data Collection Scheme: Chose women with cancer and women without cancer. Then measured age.

- Goal: Test if Cancer is associated with age
- 2 populations: those w/ cancer and those without
- Let $p_1 = Pr(\text{age} \geq 30 \mid \text{cancer})$

  $p_2 = Pr(\text{age} \geq 30 \mid \text{No cancer})$
- $H_0: p_1 = p_2$ v.s. $H_1: p_1 \neq p_2$
- $\hat{p}_1 = \frac{683}{3220} = 0.212$, $\hat{p}_2 = \frac{1498}{10245} = 0.146$
- Idea: Base test on $\hat{p}_1 - \hat{p}_2$.

- Let $X_1 = \#\ \geq 30 \mid$ Cancer
  $X_2 = \#\ \geq 30 \mid$ No cancer

$X_1 \sim \text{Binom}(n_1, p_1)$ $\quad \hat{p}_1 = X_1/n_1, \quad n_1 = 10295$

$X_2 \sim \text{Binom}(n_2, p_2)$ $\quad \hat{p}_2 = X_2/n_2, \quad n_2 = 3220$

↳ Very large sample size so can use Normal theory

- $\hat{p}_1 \sim N\left(p_1, \frac{1}{n_1} p(1-p)\right), \quad \hat{p}_2 \sim N\left(p_1, \frac{1}{n_2} p(1-p)\right) \Big] p_1 = p_2 = p$

$\Rightarrow \hat{p}_1 - \hat{p}_2 \sim N\left(0, \left(\frac{1}{n_1} + \frac{1}{n_2}\right) p(1-p)\right)$

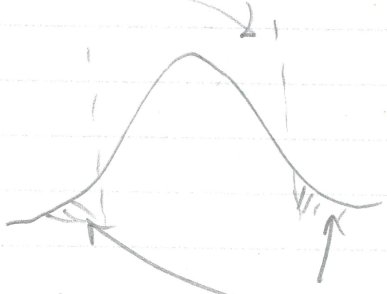
- $\Rightarrow \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) p(1-p)}} \sim N(0,1)$

- Estimate p by noting that, if H_0 is true, then

$X_1 + X_2 \sim \text{Binom}(n_1 + n_2, p)$

$\Rightarrow \hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$

- $\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}\, \hat{p}(1-\hat{p})}} \sim N(0,1)$



↳ Compare to $N(0,1)$ to get p-value

p-value

- Need to do continuity correction for better performance
- Ex.) $\hat{p}_1 = 0.212, \quad \hat{p}_2 = 0.146$

$x_1 = 683 \qquad x_2 = 1498$

$n_1 = 3220 \qquad n_2 = 10245$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{683 + 1498}{3220 + 10245} = 0.162$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} = \frac{0.212 - 0.146}{\sqrt{\left(\frac{1}{3220} + \frac{1}{10245}\right) 0.162(1-0.162)}}$$

$$= 8.866$$

$$p\text{-value} = 2 \cdot \text{pnorm}(-8.866) \approx 1 \times 10^{-15}$$

$\Rightarrow$ Strong evidence that women with curls are more likely to have first child after age 30.

2-sample binomial tests in R

- What I showed before was a 2×2 Contingency Table

	Age		
Status	≥30	≤29	Total
Case	683	2537	3220
Control	1498	8747	10245
Total	2181	11284	13465

Row Margins

Column Margins

Grand total

- An equivalent (exact same p-value) method is to test for "homogeneity" or "Independence" between the 2 variables.

- Suppose our table looks like this

		Z = A	Z = B	
W	C	X_{11}	X_{12}	M_1
	D	X_{21}	X_{22}	M_2
		n_1	n_2	N

$n_1 = X_{11} + X_{21}$

$n_2 = X_{12} + X_{22}$

$M_1 = X_{11} + X_{12}$

$M_2 = X_{21} + X_{22}$

$N = n_1 + n_2 = M_1 + M_2$

If $Z \perp\!\!\!\perp W$ then

$Pr(Z=A \cap W=C) = Pr(Z=A)\,Pr(W=C)$

$Pr(Z=A \cap W=D) = Pr(Z=A)\,Pr(W=D)$

etc...

- **Idea:** calculate expected counts (if independent) based on marginals

$$Pr(Z=A) = \frac{n_1}{N}, \quad Pr(Z=B) = \left(\frac{n_2}{N}\right)$$

$$Pr(W=C) = \frac{m_1}{N}, \quad Pr(W=D) = \left(\frac{m_2}{N}\right)$$

$$Pr(Z=A)\,Pr(W=C) = \frac{n_1}{N}\,\frac{m_1}{N}$$

$$E[X_{11} \mid \text{independent}] = N\,\frac{n_1}{N}\,\frac{m_1}{N} = \frac{n_1 m_1}{N}$$

Compare observed X_{11} to expected X_{11}

- Expected Counts from age at first birth example

	≥30	≤29
Case	$\frac{1281 \cdot 3220}{13465}$	$\frac{11784 \cdot 3220}{13465}$
Control	$\frac{1281 \cdot 10245}{13465}$	$\frac{11284 \cdot 10245}{13465}$

	≥30	≤29
= Case	521.6	2648.4
Control	1659.4	8585.6

↑ How do these compare to observed counts?

- Whenever you compare observed counts to expected counts, do a pearson χ^2 test

$$\chi^2 = \sum_{\text{categories}} \frac{(o-e)^2}{e}$$

o = observed count
e = expected count

If H_0 true, $\chi^2 \sim \chi^2_\nu$

ν = degrees of freedom

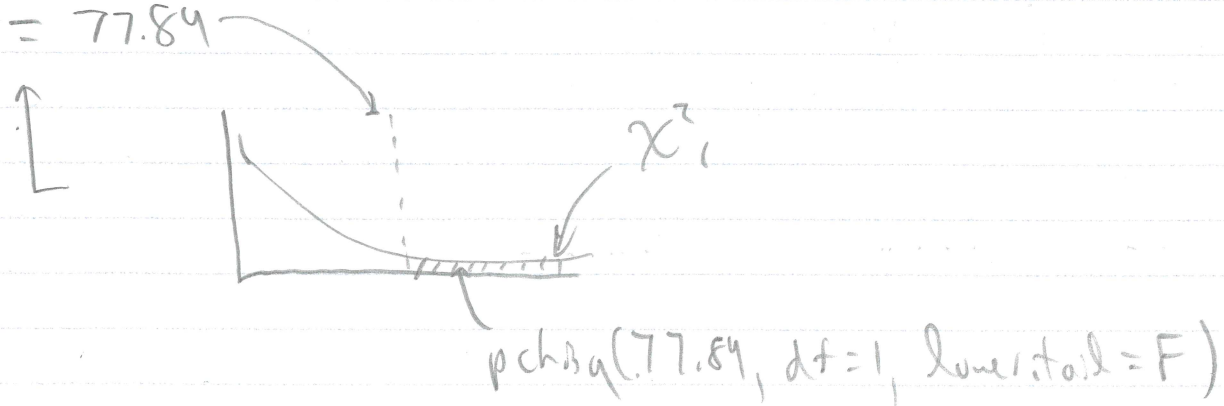
For test for homogeneity in 2×2 tables, $\nu = 1$

- Ex.) Birth Example

$$\chi^2 = \frac{(683-521.6)^2}{521.6} + \frac{(2537-2698.4)^2}{2698.4}$$

$$+ \frac{(1498-1659.4)^2}{1659.4} + \frac{(8747-8585.6)^2}{8585.6}$$

$$= 77.84$$



- Exercise: What are the expected counts of the following OC use vs. MI table

		MI: Yes	MI: No	
OC	Yes	13	4987	5000
	No	7	9993	10000
		20	14980	15000

Solution:

$$\frac{20 \cdot 5000}{15000} \quad \frac{14980 \cdot 5000}{15000}$$

$$\frac{20 \cdot 10000}{15000} \quad \frac{14980 \cdot 10000}{15000}$$

$$= \begin{matrix} 6.7 & 4993.3 \\ 13.3 & 9986.7 \end{matrix}$$

Exercise

- what is the χ^2-test statistic

$$\frac{(13-6.7)^2}{6.7} + \frac{(4987-4993.3)^2}{4993.3} + \frac{(7-13.3)^2}{13.3} + \frac{(9993-9986.7)^2}{9986.7}$$

Contingency test perspective in R

Fisher's Exact Test

- What if n is small?
- "Small": Any expected (Not observed) count ≤ 5

Ex) Salt diet vs. cardiovascular Disease (CVD) death

	Type of diet		
Cause of death	High salt	Low salt	Total
Non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

$$E_{11} = \frac{7}{60} \cdot \frac{25}{60} \cdot 60 = 2.92 < 5$$

Use 2.92, not 2, to determine normality approximation

- Exact Test: Controls Type I error for any n, not just large n.

- 1.) Fix margins
  2.) Find all possible tables with these margins
  3.) Each table has a known probability (if H_0 true)
  4.) Find how likely our table is (if H_0 true)

observed table

0 25 / 7 28	1 24 / 6 29	2 23 / 5 30	3 22 / 4 31	4 21 / 3 32	5 20 / 2 33	6 19 / 1 34	7 18 / 0 35
0.017	0.105	0.252	0.312	0.214	0.082	0.016	0.001

Sum to get p-value

- These probabilities come from the hypergeometric distribution (don't need to know)
- Exercise: what are the possible tables with fixed margins from the following

2	1
2	2

solution: 3 0 / 1 3 ; 2 1 / 2 2 ; 1 2 / 3 1 ; 0 3 / 4 0

Fisher Exact Test in R

McNemar's Test

- Study design: have matched samples, each has 2 binary variables.

- Eg. Treatment A vs. Treatment B
  Survive vs. Not

  - Choose woman to go into A
  - Match another woman with similar characteristics to go into B
  - See survival of both

- "similar characteristics": Age, weight, clinical condition, etc..

- Naive (wrong) Way: Each individual is an observational unit, put these in 2×2 table

Treatment	Survive	Die	Total
A	526	95	621
B	515	106	621
total	1041	201	1242

[Run a χ^2-test for homogeneity, get large p-value

- But observations are matched (not independent) so this is the wrong contingency table

	Treat	survive
pairs	A	Y
	B	N
	A	Y
	B	Y
	⋮	⋮

→

Pair	A survive	B survive
1	Y	N
2	Y	Y
⋮	⋮	⋮

use this to make contingency table

Correct Way: Each match pair is a unit

A outcome	B outcome: Survive	B outcome: Die	Total
Survive	510	16	526
Die	5	90	95
total	515	106	621

- Concordant Pair: Same outcome (both survive or both die)
- Discordant Pair: Different Outcome (one dies one survives)
- Concordant Pair tells you nothing about which treatment is better
- Treatments are if two discordant cells have about same counts
- Let X = # A survive and B die
  n = # discordant pairs

$X \sim \text{Binom}(n, p)$

$H_0: p = \frac{1}{2}$ ← treatments A and B equally effective

$H_1: p \neq \frac{1}{2}$

Don't use χ^2 on this table b/c we know they are associated (by matching). Just want to know which one does better.

- So just use binomial methods on discordant pairs

↑ Normal approximation for large counts
↑ Exact test for small counts.

McNemar's Test in R

- Exercise: Hypertensive diagnosis. Each person is assessed ① by trained observer and ② by a mall machine. Here, each person is a unit and the assessments are matched. Test if same result (on average)

Mall \ Trained Observer	+	−
+	3	7
−	1	9

$X \sim Binom(8, p)$ $H_0: p = \frac{1}{2}$, $H_1: p \neq 1/2$

```
binom.test(x=7, size=8, p=0.5)
```

↑ Cannot use mcnemar.test() because that uses ^ large approach

- Summary: If you can group folks into concordant vs. discordant pairs, then use McNemar's test.

- Skip Power calculations §10.5

Larger Contingency Tables

- Instead of 2 binary variables, we have 2 categorical variables.

Variable 1 has R levels
Variable 2 has C variables

- Ex) Cancer v.s. age at first birth

	Age					
	<20	20-24	25-29	30-34	≥35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10245
Total	1742	5638	3904	1555	626	13465

- H_0: Cancer status $\perp\!\!\!\perp$ age at first birth
  H_1: 2 variables are related

- Use same contingency table approach as from 2x2 tables

① Calculate expected counts assuming independence.

② Compare to Observed counts using χ^2-statistic

$$\chi^2 = \sum_{\text{cells}} \frac{(e-o)^2}{e}$$

③ $\chi^2 \sim \chi^2_\nu$ if H_0 true

$$\nu = (R-1)((-1)$$

- $E_{11} = \underbrace{\frac{1742}{13465}}_{Pr(<20)} \cdot \underbrace{\frac{3220}{13465}}_{Pr(case)} \cdot \widehat{13465} = \frac{1742 \cdot 3220}{13465} = 416.6$

$E_{12} = \underbrace{\frac{1742}{13465}}_{Pr(<20)} \cdot \underbrace{\frac{10245}{13465}}_{Pr(control)} \cdot \widehat{13465} \approx 1348.3$

- Exercise: E_{25}

$E_{25} = \frac{626 \cdot 10245}{13465} = 476.3$

- $X^2 = \sum_{cell} \frac{(E-O)^2}{E} = \frac{(416.6-320)^2}{416.6} + \cdots + \frac{(476.3-406)^2}{476.3} = 130.3$

$\nu = (2-1)\cdot(5-1) = 4$

pchisq(130.3, df = 4, lower.tail = FALSE)

Larger Contingency tables in R

- Skip χ^2-test for Trend

χ^2 Goodness-of-Fit Test

- The χ^2 tests of homogeneity are special cases of χ^2 Goodness-of-fit tests.

- Let e be the expected counts under the null hypothesis.

$$\chi^2 = \sum_{\text{categories}} \frac{(e-o)^2}{e} \qquad o = \text{observed counts}$$

- Under H_0, $\chi^2 \sim \chi^2_\nu$

- ν = (# parameters under H_1) − (# parameters under H_0)

↳ H_1 has more parameters (more complicated)

- Eg. 2×2 table

	C	D
A	x_{11}	x_{12}
B	x_{21}	x_{22}

H_0: Independence

↳ 2 parameters: $\Pr(A)$ and $\Pr(C)$ ($\Pr(D) = 1 - \Pr(C)$)

↳ $\Pr(B) = 1 - \Pr(A)$

H_1: Association: 3 parameters, $\Pr(A \cap C)$, $\Pr(A \cap D)$, $\Pr(B \cap C)$

↳ $\Pr(B \cap D) = 1 - [\Pr(A \cap C) + \Pr(A \cap D) + \Pr(B \cap C)]$

$\nu = 3 - 2 = 1$

Cohen's Kappa

- Want a measure of how reliable a test is
- Or measure how similar 2 judges rate something

Ex) 2 surveys, ask amount of beef consumed

	Survey 2: ≤1 serving/week	Survey 2: >1 serving/week	
Survey 1: ≤1 serving/week	136	92	228
Survey 1: >1 serving/week	64	240	304
	205	332	537

- Idea: "Amount" of concordance is how reliable the survey is

- p_o = proportion concordant = $\frac{136 + 240}{537} = 0.7$

- Compare it to expected amount of concordance under independence

$$p_e = \frac{205}{537} \cdot \frac{228}{537} + \frac{332}{537} \cdot \frac{304}{537} = 0.518$$

- $K = \frac{p_o - p_e}{1 - p_e}$

- Properties:

1.) $\frac{-p_e}{1-p_e} \leq K \leq 1$ (set $p_0 = 0$ or 1 for bounds)

2.) $K = 1 \Rightarrow$ perfect concordance

3.) Rules of thumb:

$K > 0.75 \Rightarrow$ excellent reproducibility
$0.4 \leq K \leq 0.75 \Rightarrow$ good reproducibility
$0 \leq K \leq 0.4 \Rightarrow$ marginal reproducibility

- Confidence intervals and tests for K are possible
- Use K for repeat measures of the same variable
- For two different variables, use sensitivity and specificity.

Cohen's Kappa in R

Exercises 10.8 - 10.13

- 2 Drugs (A, B)
  Patients matched based on Age, gender, condition
  $n = 200$ matched pairs

	A	
B	Effective	Ineffective
Effective	89	16
Ineffective	5	90

10.8.) What test? McNemar

10.9) Do test. prop.test($x = 16$, $n = 21$)

p-value = 0.0291

Evidence that tests differ in efficacy

- Focus on 160 males of same study
  - 52 pairs: both effective
  - 35 pairs: both ineffective
  - 4 pairs: A effective, B ineffective
  - 9 pairs: B effective, A ineffective

10.10.) How many concordant pairs? $52 + 35 = 87$

10.11.) How many discordant pairs? $4 + 9 = 13$

10.12.) Test if difference in effectiveness

binom.test($x = 4$, $n = 13$), $p = 0.2668$ ⇒ fail to reject H_0

Σ $n < 20$ so cannot use asymptotic approach

10.13.) - 160 of 200 patients diagnosed with gonorrhea have had previous episodes of urethritis.
- 50 of 105 patients diagnosed with nongonococcal urethritus have had previous episodes of urethritis
- Are present diagnosis and prior episodes of urethritis associated?

$$tab = \begin{pmatrix} 160 & 40 \\ 50 & 55 \end{pmatrix}$$

chisq.test(tab)

p-value = 1.4e-8

↳ very small, so strong evidence of association