2×2 Contingency Tables and Test for 2 Independent Proportions

For a 2×2 contingency table, the χ^2 test of independence gives *the same* conclusion as the significance test for comparing two independent proportions. Why? How are these related?

	Y	N
Group 1		
Group 2		

Sig Test for $p_1 - p_2$:

H_0: $p_1 - p_2 = 0$

H_a: $p_1 - p_2 \neq 0$

Equivalent

χ^2 test :

H_0: Independence

H_a: Association

Example: Suppose we want to check if Gender is independent of Vegetarianism. We collect the following data on 679 randomly selected individuals.

	Vegetarian?		
Gender	Yes	No	**Total**
Male	5	154	159
Female	38	482	520
Total	43	636	679

Assumptions and Hypotheses for χ^2 test of independence:

- SRS – 679 ppl random + representative of poplation of interest
- Min 5 expected counts under H_0 in every cell

Assumptions and Hypotheses for 2 independent proportions test:

- SRS – same
- 10 succ + fail expected under H_0 for each group

Example: The Physician's Health Study I was a randomized, double-blind, placebo-controlled trial whose participants were all male doctors in the USA. It studied the relationship between heart attacks (Y/N) and taking aspirin (Y/N). One treatment group took Aspirin, while another took a placebo daily, for several years. https://phs.bwh.harvard.edu/phs1.htm

The results are summarized in the contingency table below:

	Heart Attack?		
	Yes	No	**Total**
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037
Total	293	21,778	22,071

- Have the assumptions been met to conduct a χ^2 test for independence?

•SRS of participants (Not representative of all American adults, at best to males, doctors (or maybe well-educated + rich, higher interest in health))

• min 5 expected per cell → Check obs.

- State the hypotheses for this problem.

Ho: No association between Aspirin and Heart attack

Ha: Association . . . .

- Conduct the statistical analysis and state the conclusion.

Pearson's TS: $X^2 = 25.014$

p-value ≈ 0 from X^2 distribution with 1 df.

Rej Ho at all usual α's → Very strong evidence of association between Aspirin and Heart attack.

- Compute Conditional Probabilities to describe the association.

% of Aspirin had heart attacks: 104/11037 = .0094 = 0.94%

% of Placebo had heart attacks: 189/11034 = .0171 = 1.71%

- Compute the Relative Risk of heart attack for the placebo group vs aspirin.

$$RR = \frac{\hat{p}_1}{\hat{p}_2} = \frac{1.71\%}{0.94\%} = 1.82$$

(1.82 times)

Interpretation: Ppl who take placebo are almost twice as likely as ppl who take Aspirin of having a heart attack

Not all ppl though !!

Identify the test statistics, p-values, and draw conclusions based on the Minitab Output below:

Chi-Square Test: Veg, Not

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

		Veg	Not	Total
M →	1	5	154	159
		10.07	148.93	
		2.552	0.173	
F →	2	38	482	520
		32.93	487.07	
		0.780	0.053	
	Total	43	636	679

Chi-Sq = 3.558, DF = 1, P-Value = 0.059

exp: $\frac{43\times159}{679}=10.07$ $\quad \frac{636\times159}{679}=148.93$

$\frac{43\times520}{679}=32.93$ $\quad \frac{636\times520}{679}=487.07$

contribution to TS:

$\frac{(5-10.07)^2}{10.07}=2.552$ $\quad \frac{(154-148.93)^2}{148.93}=0.173$

$\frac{(38-32.93)^2}{32.93}=0.780$ $\quad \frac{(482-487.07)^2}{487.07}=0.053$

$\chi^2 = 2.552+0.173+0.780+0.053=3.558$

$DF=(2-1)\times(2-1)=1$

$p\text{-value} = P(\chi^2_1 \geq 3.558) \xrightarrow{\text{use software}} 0.059$

a random variable that follows χ^2 distribution with degrees of freedom 1.

Test and CI for Two Proportions

Sample	X	N	Sample p	
M 1	5	159 (n_1)	0.031447	$= 5 \div 159$
F 2	38	520 (n_2)	0.073077	$= 38 \div 520$

Difference = p (1) - p (2)
Estimate for difference: -0.0416304
95% CI for difference: (-0.0767909, -0.00646989)
Test for difference = 0 (vs not = 0): Z = -1.89 P-Value = 0.059

~~Fisher's exact test: P-Value = 0.063~~

$\hat{p}_1-\hat{p}_2 = \frac{5}{159}-\frac{38}{520} = -0.0416304$

95% CI for $p_1 - p_2$:

$\hat{p}_1-\hat{p}_2 \pm Z_{0.025}\cdot\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$=\frac{5}{159}-\frac{38}{520}\pm 1.96\times\sqrt{\frac{\frac{5}{159}\times(1-\frac{5}{159})}{159}+\frac{\frac{38}{520}\times(1-\frac{38}{520})}{520}}$

$=(-0.0767909, -0.00646989)$

Test Statistic:

$TS (= Z) = \frac{\hat{p}_1-\hat{p}_2-0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{\frac{5}{159}-\frac{38}{520}}{\sqrt{\frac{43}{679}\times(1-\frac{43}{679})\times(\frac{1}{159}+\frac{1}{520})}}$

$= -1.89$

$p\text{-value} = 2P(Z \leq -1.89) \xrightarrow{\text{Z-table}} 0.059$

$(-1.89)^2 = 3.558$

☆ $Z^2 = \chi^2_1$

☆ $t_n^2 = F_{1,n}$

$\chi^2_n = Z_1^2 + Z_2^2 + \cdots + Z_n^2$

$Z_i \perp\!\!\!\perp Z_j, (i \neq j)$

$Z_i \sim N(0,1)$

$t_n = \frac{X}{\sqrt{Y/n}}$ $\quad X \to N(0,1)$, $Y \to \chi^2_n$ $\quad X \perp\!\!\!\perp Y$

$F_{m,n} = \frac{X/m}{Y/n}$ $\quad X \to \chi^2_m$, $Y \to \chi^2_n$ $\quad X \perp\!\!\!\perp Y$

Extra Credit (0.5pts)

Example: Collect data in class, or use the General Social Survey to answer a question of interest involving two categorical variables.

Variables: Sexuality & Dating apps

Dating apps

Sexuality	No	Yes	
Hetero	80	40	120
Not hetero	7	17	24
	87	57	144

- Assumptions:
- Ho:
  Ha:
- TS:
- DF:
- p-value:
- conclusion:

- Conditional Prob: % who use Dating Apps for heterosexuals: / not:
- Marginal distributions for sexuality and Dating apps usage:
- RR of using dating apps for the heterosexuals v.s. not heterosexuals:
  Interpretation of the RR:
- Joint distribution:

Deadline: March 9th ~~10:25am~~ 11:59pm (No email; No paper)

Upload a <u>pdf</u> file on Canvas "Extra Credit 2"

Nonparametric Statistics (Ch. 15)

Overview of Nonparametric Methods

- **Normal-based procedures** (Z and t tests) assume that the statistic has an approximately Normal distribution, either because the sample sizes are large enough or because the original distribution is Normal.

- **When utilizing Nonparametric methods, we still have a parameter but we have no distribution assumptions** for the response variable. In order to use these methods, our response variable *Y* must must be **quantitative**.

- **Nonparametric methods make inferences about:** Populations

  But not about means, instead inferencens about median or entire distribution.

- **Nonparametric methods are useful when:**

  - small sample size (with outliers)
  - several groups with very different variances
  - data is quantitative but NOT continuous - particularly subjective ratings

- **Nonparametric methods generally only need to assume:**

  - SRS from population of interest

Remarks:

- Almost All the nonparametric methods we see in this class are based on RANKS.
- If assumptions of Normality and Equal Variances are met, then both parametric (normal-based) and nonparametric (distribution-free) methods can be used.
- However, parametric methods are **statistically more powerful** and are able to find true significant differences (assuming all assumptions are met), while nonparametric methods sometimes cannot.

power = $1-\beta$. $\beta = P(\text{Not rej } H_0 \mid H_a \text{ is true})$

$\beta \downarrow$ Type II error