PHP 2510 Principles of Biostatistics & Data Analysis

Week 11: Categorical Data Comparisons

OUTCOMES

After this week's classes, along with the required readings (CHIHARA Chapter 10, 10.1-10.5.1), you should be able to:

- Construct a contingency table and calculate expected cell counts under the hypothesis of no association
- Perform a Chi-squared test in R and interpret all components of the output
- Determine when non-parametric tests are more appropriate and implement them appropriately

Note: SPEEGLE Ch 10 (up to and including 10.5), covers the same material if you prefer that resource

What are some common things to test?

1. ✅ Group mean (note: a proportion is a mean too)
2. ✅ Difference in group means
3. ✅ Number of successful trials
4. ✅ Variance (Chi-sq test)
5. Ratio of variances (F-test)
6. ⭐ **Differences in counts (Chi-sq test)**
7. …

- $Z^2 \sim \chi^2_1$.
- $\chi^2_\nu + \chi^2_\eta \sim \chi^2_{\nu+\eta}$ if independent.
- $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ when $X_1, \ldots, X_n$ iid normal.
- $\dfrac{Z}{\sqrt{\chi^2_\nu/\nu}} \sim t_\nu$ if independent.
- $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ when $X_1, \ldots, X_n$ iid normal.

SPEEGLE 5.5.4 Summary (pg. 143) - we are going to keep using these results!!

What is a contingency table?

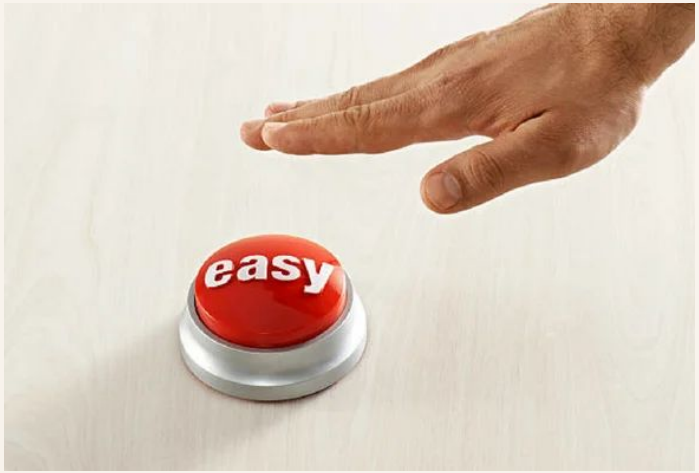
a. Displays the frequency (the count) of how often different combinations of categories occur
b. Helps us answer the question: "Are these two variables related?"
c. A 2×2 is already familiar (to summarize groups by binary outcomes), and we use prop.test(); Now we think about how to extend testing to work for any table

When do contingency tables get created?

1. “Tests of Independence”: Take a random sample, then tabulate two categorical variables
2. “Homogeneity”: Case/Control study, with categorical outcome
3. “Goodness of Fit”: Does my count data align with my predictions?

Come up with a hypothetical scenario for one of these; don’t do just a 2×2 table.

GOOD NEWS: all these designs actually result in the same test statistic and null distribution



$$\sum \frac{(Observed \ - \ Expected)^2}{Expected} \sim \chi^2_{df}$$

- For tests of independence and homogeneity:
  - "Expected" means: under the null of no association
  - df: (#Rows - 1) x (#Columns -1)
- For goodness of fit:
  - "Expected" means: under the null that the model is correct
  - Df: #categories - 1 - #parameters_estimated
- Sum is over all the cells in the table

Let's look at this closer:

$$\sum \frac{(Observed \ - \ Expected)^2}{Expected} \sim \chi^2_{df}$$

- $Z^2 \sim \chi^2_1$.
- $\chi^2_\nu + \chi^2_\eta \sim \chi^2_{\nu+\eta}$ if independent.
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ when $X_1, \ldots, X_n$ iid normal.
- $\frac{Z}{\sqrt{\chi^2_\nu/\nu}} \sim t_\nu$ if independent.
- $\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$ when $X_1, \ldots, X_n$ iid normal.

SPEEGLE 5.5.4 Summary (pg. 143)

Remember, we have options:

We can either:

1. Rely on large sample sizes and assume our test statistic does indeed follow a ChiSq distribution
    a. Rule of thumb: All cells have expected counts $\geq$ 5
2. Use non-parametrics
    a. Permutation Test
        i. Built into *chisq.test()* via *simulate.p.value* argument
    b. Fisher's Exact Test (remember the chocolate experiment!?)
        i. *fisher.test()*

THINK – PAIR – SHARE

1. Why is it uncommon to report a confidence interval for the chi-square test in this setting?

2. The “rule of thumb” was about the expected, not observed counts. Do you think we need the observed counts to be ≥ 5 as well. Why or why not?

Advanced Topic: Paired Binary Data

A public health department wants to know if a new, intensive 6–month educational campaign about the benefits of the HPV vaccine was effective in changing parents' intent to vaccinate their children. They survey the same 500 parents before and after and collect their intent to vaccinate (yes/no).

Are the two groups independent?
What does our data look like?
Can we derive a test statistic and it’s null distribution?

This is called McNemar’s Test.
What others tests are available in this situation?

NOW IT'S TIME FOR PRACTICE PROBLEMS

Remember, all our theory about hypothesis testing (null hypothesis, alpha-level, p-value, etc.) still holds; we just have a new data type and test statistic

Recall from week 2:

The world series is a best-of-7 competition. What's the likelihood of requiring exactly n = {4, 5, 6, 7} games to declare a champion?
We calculated the theoreticals under a model of: 50–50 chance & independence between games.

	Theoretical	~~Actuals (n=116)~~	Actuals (n=117)
4 games	12.5%	~~19 (16.4%)~~	19 (16.2%)
5 games	25%	~~30 (25.9%)~~	30 (25.6%)
6 games	31.25%	~~27 (23.3%)~~	27 (23.1%)
7 games	31.25%	~~40 (34.5%)~~	41 (35.0%)

DATA UPDATE: 2025 DODGERS

What is the null hypothesis? Test statistic? Test statistic's null distribution? … Let's solve "by hand" then use the built-in R function

CHIHARA 10.8*

Consider the following survey data. Conduct the appropriate test for independence in R and state your conclusions.

		Health Screening Frequency				
		Never	Once	< Annually	Annually	> Annually
Program Enrollment	Yes	180	260	137	96	52
	No	210	266	145	85	49

#2

*same data but new context to be public health focused

CHIHARA 10.14

For a given contingency table, we have a test statistic T given by the standard formula

$$\sum \frac{(Observed \ - \ Expected)^2}{Expected} \sim \chi^2_{df}$$

What happens when **every** entry in the contingency table is multiplied by k > 1?

- Do the marginal probabilities change?
- Do the degrees of freedom change?
- Does the test statistic change?
- Do the conclusions changes?

Let's confirm via applying some k > 1 to previous problem

CHIHARA 10.21*

The following data were collected to study the annual number of Emergency Department visits for patients with severe COPD

	Visits						
	0	1	2	3	4	5	6+
Number	57	20	11	4	6	2	2

Model this using a Poisson distribution and perform a goodness of fit test.

Let's do this both "by hand" & using chisq.test()

Let's roadmap the solution before we get started

*same data but new context to be public health focused

HANDOUT

OUTCOMES

Q&A

After this week's classes, along with the required readings (CHIHARA Chapter 10, 10.1-10.5.1), you should be able to:

- Construct a contingency table and calculate expected cell counts under the hypothesis of no association
- Perform a Chi-squared test in R and interpret all components of the output
- Determine when non-parametric tests are more appropriate and implement them appropriately

Note: SPEEGLE Ch 10 (up to and including 10.5), covers the same material if you want another resource to help

Exam #2

Topics Covered: Sampling Distributions & the Central Limit Theorem, Hypothesis Testing, Experimental Design, Confidence Intervals, Non-Parametric Tests, Categorical Comparisons

You may not use a computer, calculator, nor any reference materials. You have the full 80 minutes. The exam is graded out of 50 points (although there are 58 points available). There are 5 sections:

- CONCEPTS: assesses your understanding of statistical concepts via short-form answers
- DERIVATIONS & FORMULAE: assesses your ability to use and explain common formulae
- END-TO-END ANALYSIS: assesses your ability to solve multi-part word problems
- INTERPRETING R CODE: assesses your ability to interpret R code and output
- EXTRA CREDIT: additional challenge questions