AP Statistics
Mr. Murphy
12.2 Test for Homogeneity and Independence in a Two-Way Table
HW 12.14, 12.15, 12.18, 12.19

GOALS:
1. Run a chi-square test for homogeneity.
2. Run a chi-square test for independence.
3. Understand what the next four weeks of class look like!

There are two more types of tests we can use with the χ^2 (Chi-square) test statistic. We can test:

- two or more populations for homogeneity (same proportions), or we can investigate
- whether two variables from the same population are independent

Both tests have identical mechanics and only differ in conclusion.

Comparing Two or More Populations for Homogeneity

1. State your target population, null and alternative hypothesis (in symbols and words).

   H_0 : The true category proportions are the same for all populations (homogeneity of the populations)

   $H_a : H_0$ is not true. The true category proportions are not the same for all of the populations.

2. Assumptions:
   1. The data consist of *independently chosen random samples*.
   2. The *sample size is large*. The sample size is large enough for the chi-square test to be appropriate as long as every expected cell count is at least 5.

3. Test Statistic: $\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count - expected cell count})^2}{\text{expected cell count}}$

   The expected cell counts are estimated from the sample data (assuming H_0 is true) using the formula:

   $$\text{expected cell count} = \frac{(\text{row marginal total})(\text{column marginal total})}{\text{grand total}}$$

   df = (number of rows – 1) (number of columns – 1)

   Calculation of P – value

4. Conclusion, in theory and context.

<u>Notes About Using Matrices</u>

Ex1 The table below shows the rank attained by male and female officers in the New York City Police Department (NYPD).

	Male	Female	Total
Officer	21,900	4,281	**26,181**
Detective	4,058	806	**4,864**
Sergeant	3,898	415	**4,313**
Lieutenant	1,333	89	**1,422**
Captain	359	12	**371**
Higher Ranks	218	10	**228**
Total	**31,766**	**5,613**	**37,379**

Is there evidence that the proportion of promotions differs for males and females. Test the relevant hypotheses using $\alpha = 0.01$.

Ex2 For the following table, compute the value of χ^2.

	C	D
A	15	25
B	10	30

(a) 2.63
(b) 1.22
(c) 1.89
(d) 2.04
(e) 1.45

Comparing Two or More Populations for Independence

1. State your target population, null and alternative hypothesis (in symbols and words).

   H_0 : The two variables are independent.

   $H_a : H_0$ is not true. The two variables are not independent.

2. Assumptions:

   1. The observed counts are from a *random sample*.
   2. The *sample size is large*. The sample size is large enough for the chi-square test to be appropriate as long as every expected cell count is at least 5.

3. Test Statistic: $\chi^2 = \sum_{\text{all cells}} \dfrac{(\text{observed cell count - expected cell count})^2}{\text{expected cell count}}$

   The expected cell counts are estimated from the sample data (assuming H_0 is true) using the formula:

   $$\text{expected cell count} = \frac{(\text{row marginal total})(\text{column marginal total})}{\text{grand total}}$$

   Degrees of Freedom, *df* = (number of rows – 1) (number of columns – 1)

   Calculation of P – value

4. Conclusion, in and out of context.

Ex3 Here is a table showing who survived the sinking of the Titanic based on whether they were crew members, or passengers booked in first-, second-, or third-class staterooms:

	Crew	**First**	**Second**	**Third**	**Total**
Alive	212	202	118	178	710
Dead	673	123	167	528	1491
Total	885	325	285	706	2201

(a) If an individual is drawn at random, what is the probability that we will draw a member of the crew?

(b) What's the probability of randomly selecting a third-class passenger who survived?

(c) If someone's chances of surviving were the same regardless of their status on the ship, how many members of the crew would you have expected to survive?

(d) Is there a relationship between a person's status on the Titanic and whether or not they survived the sinking? Perform the appropriate test.

Checkpoint
Multiple Choice

1. Find the expected value of the cell marked with the "***" in the following 3×2 table (the bold face values are the marginal totals):

observation	observation	**19**
observation	***	**31**
observation	observation	**27**
45	**32**	**77**

(a) 74.60 (b) 18.12 (c) 12.88 (d) 19.65 (e) 18.70

2. A study is to be conducted to help determine if race is related to blood type. Race groups are identified as White, African-American, Asian, Latino, or Other. Blood types are A, B, O, and AB. How many degrees of freedom are there for a chi-square test of independence between Race and Blood Type?

(a) $5 \times 4 = 20$
(b) $5 \times 3 = 15$
(c) $4 \times 4 = 16$
(d) $5 + 4 - 2 = 7$
(e) $4 \times 3 = 12$

3. A group separated into men and women are asked their preference toward certain types of television shows. The following table gives the results.

	Program Type A	Program Type B
Men	5	20
Women	3	12

Which of the following statements is/are true?

I. The variables gender and program preference are independent.
II. For these data, $\chi^2 = 0$.
III. The variables gender and program preference are related.

(a) I only
(b) I and II only
(c) II only
(d) III only
(e) II and III only

4. Is there a relationship between education level and sports interest? A study cross-classified 1500 randomly selected adults in three categories of education level (not a high school graduate, high school graduate, and college graduate) and five categories of major sports interest (baseball, basketball, football, hockey, and tennis). The χ^2 value is 13.95. Is there evidence of a relationship between education level and sports interest?

(a) The data prove there is a relationship between education level and sports interest.
(b) The evidence points to a cause-and-effect relationship between education level and sports interest.
(c) There is evidence at the 5% significance level of a relationship between education level and sports interest.
(d) There is evidence at the 10% significance level, but not at the 5% significance level, of a relationship between education level and sports interest.
(e) The *P*-value is greater than 0.10, so there is no evidence of a relationship between education level and sports interest.

5. A disc jockey wants to determine whether middle school students and high school students have similar music tastes. Independent random samples are taken from each group, and each person is asked whether he/she prefers hip-hop, pop, or alternative. A chi-square test of homogeneity of proportions is performed, and the resulting *P*-value is below 0.05. Which of the following is a proper conclusion?

(a) There is evidence that for all three music choices the proportion of middle school students who prefer each choice is equal to the corresponding proportion of high school students.
(b) There is evidence that the proportion of middle school students who prefer hip-hop is different from the proportion of high school students who prefer hip-hop.
(c) There is is evidence that for all three music choices the proportion of middle school students who prefer each choice is different from the corresponding proportion of high school students.
(d) There is evidence that for at least one of the three music choices the proportion of middle school students who prefer that choice is equal to the corresponding proportion of high school students.
(e) There is evidence that for at least one of the three music choices the proportion of middle school students who prefer that choice is different from the corresponding proportion of high school students.