Worksheet 4: Chi-Squared Tests

Useful Fact: The pdf of a chi-squared random variable with d degrees of freedom is

$$f(x) = \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d}{2}-1} e^{-x^2/2}, \quad x > 0.$$

Goodness-of-Fit Test

Suppose we have a multinomial random variable with k outcomes, labelled $1, \ldots, k$, constructed from n trials such that in each trial outcome i occurs with probability p_i. The *chi-squared statistic* is the quantity

$$\chi^2 := \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i}$$

where E_i is the expected number of times outcome i occurs, and N_i is the number of times outcome i is observed. For a goodness of fit test using this statistic, χ^2 has a chi-squared distribution with $k-1$ degrees of freedom.

Problem 1 (Goodness-of-fit test)**.** According to the Mars Wrigley Confectionery Company, the color distribution of a bag of skittles is skittles, the null hypothesis

$$H_0 : p_{\text{green}} = p_{\text{yellow}} = p_{\text{red}} = p_{\text{orange}} = p_{\text{purple}} = 0.2$$

the alternative hypothesis is

$$H_1 : \text{at least one of the probabilities is different from } 0.2$$

(a) Assuming that H_0 is true, what is E_i for all $i \in \{\text{green}, \text{yellow}, \text{red}, \text{orange}, \text{purple}\}$?

(b) Fill in the following table with expected and observed counts:

	green	**yellow**	**red**	**orange**	**purple**
expected					
observed					

(c) Compute the chi-squared statistic χ^2 for your bag of skittles

(d) Circle the correct phrase to complete the sentence: *A larger chi-square value means the observed data is (very different from / very similar to) the expected data.*

(e) From theorem stated in class, χ^2 has a chi-squared distribution with $k-1$ degrees of freedom. Set up and evaluate an integral to determine the p-value $\mathbb{P}\left[\chi^2 \geq u\right]$, where u is the value you got in the previous part of the problem.

(f) Write a sentence or two interpreting your result

Test of Independence

Problem 2 (Test of Independence)**.** A 2007 study of $n = 186$ Japanese childen examined a possible relationship between the onset of autism-related developmental regression and exposure to the measles mumps rubella (MMR) vaccine.[1] One of their analyeses involved the following dataset:

Table of observed values

	Vaccinated	**Not Vaccinated**
Autism	15	39
No Autism	45	87

(a) Find the following values
- total number of children exhibiting autistic regression
- total number of children not exhibiting autistic regression
- total number of vaccinated children
- total number of unvaccinated children

(b) Let p be the probability that a child exhibits autistic regression, and let q be the probability that a child is vaccinated. Use your answers from the previous part to estimate p and q.

(c) Under the assumption that autism and vaccination are independent, we can compute expected values for each of the four categories of outcomes, which we'll call e_{11}, e_{12}, e_{21}, and e_{22}, and arrang them in the following table:

Table of expected counts

	Vaccinated	**Not Vaccinated**
Autism	$e_{11} =$	$e_{12} =$
No Autism	$e_{21} =$	$e_{22} =$

For example, under this assumption, the expected number of children who are both vaccinated and exhibit autistic regression is

$$e_{11} = npq$$

Fill in the values of the table.

(d) Compute the chi-squared statistic

$$\chi^2 = \sum_{\text{all } i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{11}, o_{12}, o_{21}, and o_{22} are the observed values.

(e) Set up and evaluate an integral to compute a p-value for your χ^2 statistic.

(f) Write a sentence or two interpreting your result.

[1] Uchiyama et al, *J Autism Dev Disord*, 2007. https://doi.org/10.1007/s10803-006-0157-3

Problem 3 (Simpson's Paradox: The Perils of Aggregation)**.** In this problem we compare the performance of two medical centers, Hospital A and Hospital B. We have the following survival data for surgery patients from these two hospitals from the last 6 weeks:

Outcome	**Hospital A**	**Hospital B**
Died	63	16
Survived	2037	784

(a) Compare the survival rates of the two hospitals: what percentage of patients at Hospital A survive? What about Hospital B?

(b) Is there a relationship between patient survival and which hospital they go to? Do a chi-squared test of independence. What p-value do you get? In one or two sentences, state your null hypothesis and draw a conclusion about it.

(c) We now consider a third factor: the condition of the patient when they are admitted for surgery. Patients are either classified as "poor condition" or "good condition". Here's the more detailed data:

Good Condition

Outcome	**Hospital A**	**Hospital B**
Died	6	8
Survived	594	592

Poor Condition

Outcome	**Hospital A**	**Hospital B**
Died	57	8
Survived	1443	192

(d) What fraction of Hospital A's patients were classified as good condition? What about Hospital B?

(e) Out of the patients in **good condition** who went to hospital A, what fraction survived? What about hostpial B? Compare these two survival rates.

(f) Out of the patients in **bad condition** who went to hospital A, what fraction survived? What about hostpial B? Compare these two survival rates.

(g) Compare your answers to parts (a) (e) (f). What you are observing is called *Simpson's Paradox*. How can this discrepancy be explained? If you are facing surgery, should you go to Hospital A or Hospital B?

(h) Fill in the following contingency table:

Outcome	**Good Condition**	**Bad Condition**
Died		
Survived		

(i) Is there a reslationship between patient survival and the condition that they are admitted for surgery. Perform a chi-squared test of independence and report a p-value.

Problem 4 (Another example of Simpson's Paradox)**.** The influence of race on imposition of the death penalty for murder has been much studied and contested in the courts. The following three-way table classifies 326 cases in which the defendant was conficted of murder.[2] The three varaibles are the defendant's race, the victim's race, and whether or not hte defendant was sentenced to death.

White Defendants

Victim Race	Yes	No
White	19	132
Black	0	9

Black Defendants

Victim Race	Yes	No
White	11	52
Black	6	97

(a) From these data make a 2-way table of defendant's race by death penalty. Does it look like there is a relationship between these two things?

(b) Suppose you performed a chi-squared test of independence and for the table you just made. You get $p = .64$. How do you interpret this result?

(c) Show that Simpson's paradox holds: specifically, that
- a higher percent of white defendants are sentenced to death overall
- but for both black and white victims, a higher percent of black defendants are sentenced to death.

(d) Basing your reasoning on the data, explain why the paradox holds in language that a judge could understand.

[2]Data from M. Radelet, *American Sociological Review*, 1981. https://doi.org/10.2307/2095088

Problem 5 (One True Love)**.** A poll is taken of $n = 2625$ individuals; people were survedy on some demographic information and also asked whether or not they believed in "the one true love".

In this problem, we test whether or not there is a relationshp between the following two factors

- Factor 1: education level
- Factor 2: sentiment about true love

We can construct a two-way contingency table for these two factors, as follows:

Response	**HS**	**Some College**	**College**	**Total**
Agree	363	176	196	735
Disagree	557	466	789	1812
Don't Know	20	26	32	78
Total	940	668	1017	2625

This contingency table has $r = 3$ rows and $c = 3$ columns (plus a row and column for totals). The data counts themselves take the form of a 3×3 matrix

$$\begin{matrix} Y_{11} & Y_{12} & Y_{13} & \cdots & Y_{1c} \\ Y_{21} & Y_{22} & Y_{23} & \cdots & Y_{2c} \\ \vdots & & & & \vdots \\ Y_{r1} & Y_{r2} & Y_{r3} & \cdots & Y_{rc} \end{matrix}$$

where Y_{ij} is the number of samples for which factor 1 is i and factor 2 is j. We've also added row and column totals in the margins. We can construct a coresponding table of "estimated expected values" of the form

$$\begin{matrix} E_{11} & E_{12} & E_{13} & \cdots & E_{1c} \\ E_{21} & E_{22} & E_{23} & \cdots & E_{2c} \\ \vdots & & & & \vdots \\ E_{r1} & E_{r2} & E_{r3} & \cdots & E_{rc} \end{matrix}$$

by the formula

$$E_{ij} = n \times \frac{(\text{sum of row } i)}{n} \times \frac{(\text{sum of column } j)}{n} \tag{1}$$

which gives us a new table of estimated expected values.

Fact: For contingency tables with r rows and c columns, the chi-squared statistic

$$\chi^2 := \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}}$$

has chi-squared distribution with $(r-1) \times (c-1)$ degrees of freedom.

Questions:

(a) Explain why the formula in Equation (1) holds.

(b) What is the degrees of freedom for the chi-squared statistic in this problem? Write down the formula for the pdf of the chi-squared statistic.

(c) Do the χ^2 test of independence (i.e., compute a table of expected values, find χ^2, and use the known distribution of the statistic to compute a p-value).

(d) Intepret your result.