

HOMEWORK 14: CATEGORICAL

Instructions

This assignment is due in Canvas by **Wednesday, December 15 at 11:59pm**.

Your submission needs to include **your R code, the corresponding R output, and your narrative** interpretation/responses to the questions (in complete sentences). The easiest way to do this is to work in an **R Notebook**, as described in the first homework guide and demonstrated in the corresponding video. If you use an R Notebook, you will submit the notebook's **HTML file** to Canvas.

Exercise

Refer to the diabetes dataset on the homework assignment page on Canvas.

1. The diabetes dataset is an excerpt from a health survey conducted in the United States in 2013–14. The demographic breakdown in the U.S. at that time was 12.3 percent Black, 17.1 percent Hispanic, 8.2 percent Other, and 62.4 percent White. But, the population that the survey researchers were using may not have been the entire U.S. population. Conduct an investigation to determine whether it is plausible that the researchers were using the U.S. population for their research.
 - a. What are the null and alternative models for your statistical test?
 - b. What are the expected and observed counts for each demographic category?
 - c. Calculate the test statistic and find its p-value.
 - d. Explain how the statistical results answer the research question: Is it plausible that the researchers were using the entire U.S. population for their research?
 - e. Explain why you should be cautious about generalizing results from this sample to the entire U.S. population. When would such generalizations be appropriate?
2. Use the diabetes dataset to investigate whether diabetes and dietary quality (**diet**) are associated in the survey researchers' population.
 - a. What are the null and alternative models for this test?
 - b. Show how the number of degrees of freedom is found for this test.
 - c. Show how to calculate the expected counts for the "Excellent diet" and "Diabetes" cell of the contingency table.
 - d. How much does the "Excellent diet" and "Diabetes" cell contribute to the test statistic for this statistical test?
 - e. Find the test statistic and p-value for this test (see R code section, below).
 - f. Explain what conclusion you should draw from the statistical analysis.

3. Consider diet and poverty as possible predictors of diabetes.
 - a. What is the relative risk of diabetes for those with Fair dietary quality versus those with Good dietary quality?
 - b. What is the relative risk of diabetes for those who are Very Poor versus those who are Not Poor?
 - c. How would you explain what the relative risk means to someone who hasn't studied statistics?
 - d. Explain which variable (diet or poverty) is a better predictor of diabetes.

R Code

The `chisq.test()` function can be used to conduct the Test of Categorical association. For example, to test for an association between poverty and health status, you could run:

```
chisq.test(diabetes$health, diabetes$poverty)
```

Note that while the `chisq.test()` can, technically, be used for the Goodness of Fit test, it requires R syntax that we have not learned.