Name: _____ LA Initials: [ ]

Lab 9: Contingency Analysis

Learning Objectives

Getting started

Before you begin the lab activities, review the steps for setting up a project for a lab activity:

- Weekly Lab Setup

Then, complete the following steps to make sure you are fully set up.

1. **Get the Lab Worksheet.**

   Pick up a physical copy of the lab worksheet, or print one if you are working outside of class.
   Download Lab Worksheet (PDF, if needed)

2. **Open Posit Cloud and start Lab 9.**

3. Create a new R script and save it as **lab-9-script.R**.

4. Add the following code to your R script and save the script. Upon save, you should see a message in the source prompting you to install the package. Click "Install". Then run the line of code.

```
# load packages -----------------------------------------------------------

library(tidyverse)
library(haven)
```

> **i** Checkpoint
>
> At this point, you should have:
>
> - These instructions open in a web browser.
> - Your Lab 9 project open in Posit Cloud in another browser window.
> - The required packages installed and loaded without errors
> - The Lab 9 worksheet in front of you.
>
> Do not continue until all of the above steps are working correctly.

Overview

In this lab, you will analyze the relationship between two categorical variables using **contingency analysis**. A contingency table summarizes the joint distribution of two categorical variables and allows us to estimate measures of association such as **relative risk (RR)** and **odds ratio (OR)**. These measures describe the magnitude and direction of an association between two binary variables. We will also conduct a **chi-square test of independence** to determine whether the observed association is statistically significant.

You will use data from the National Health and Nutrition Examination Survey (NHANES), a large, nationally representative health survey conducted in the United States. For this lab, you will examine the relationship between sleep duration and depressive symptoms among adults.

Question 1

What are two ways to measure the magnitude of an association between two binary categorical variables?

Dataset and Variables

In this lab, you will use publicly available data from the **August 2021–August 2023 cycle** of the National Health and Nutrition Examination Survey (NHANES), conducted by the Centers for Disease Control and Prevention (CDC).

NHANES is a nationally representative survey of the civilian, non-institutionalized U.S. population. It combines interviews, physical examinations, and laboratory testing to assess health and nutritional status.

We will use two variables:

1. Depressive Symptoms (`DPQ020`)

File: Depression Questionnaire (`DPQ_L.xpt`)
Dataset:
https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/DPQ_L.xpt

Variable documentation:
https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/DPQ_L.htm#DPQ020

Question:
> "[Over the last 2 weeks, how often have you been bothered by the following problems:] feeling down, depressed, or hopeless?"

Response codes:

- 0 = Not at all
- 1 = Several days
- 2 = More than half the days
- 3 = Nearly every day
- 7 = Refused
- 9 = Don't know

For this lab, you will recode this variable into a binary outcome:

- **No depressive symptom** (0)
- **Depressive symptom present** (1–3)

Responses coded 7 or 9 will be treated as missing.

2. Sleep Duration (`SLD012`)

File: Sleep Disorders Questionnaire (`SLQ_L.xpt`)
Dataset:
https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/SLQ_L.xpt

Variable documentation:
https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/SLQ_L.htm#SLD012

Question:
> "How much sleep do you usually get at night on weekdays or workdays?"

Response format:

- 3 to 13.5 = Reported in half hours

- 2 = Less than 3 hours
- 14 = 14 hours or more
- 77 = Refused
- 99 = Don't know

For this lab, you will create a binary exposure variable:

- **Short sleep (< 7 hours)**
- **Adequate sleep (≥ 7 hours)**

Responses coded as 2 or 14 will be treated as-is, even though they may represent larger or smaller values.

Responses coded 77 or 99 will be treated as missing.

Research Question

In this lab, you will examine whether adults who report **short sleep duration** have a higher risk of reporting **depressive symptoms** compared to adults who report adequate sleep.

> **Question 1**
> What is the name of the sleep variable? The depression variable? The variable that serves as the unique identifier for individuals?

Lab Activity: Sleep and Depressive Symptoms

In this activity, you will investigate whether adults who report short sleep duration have a higher risk of reporting depressive symptoms compared to adults who report adequate sleep.

Work through the sections in order. Write all code in your `lab-9-script.R` file. Run each section before moving on.

Part 1 — Prepare the Data

Statistical goal: Create two clean binary variables:

- Sleep duration (risk factor)
- Depressive symptoms (outcome)

1. Read in the depression dataset.

- Use the `read_xpt()` function from the **haven** package to read the DPQ file.
- Select only the variables you need (`SEQN` and `DPQ020`).

Hint:

- Documentation: `read_xpt()`
- Don't download the data, just paste the URL to it directly into `read_xpt()`
- Use `select()` to keep only relevant columns: `select()`

2. Recode depressive symptoms into a binary variable.

Recode `DPQ020` as follows:

- 0 ⟶ "No"
- 1, 2, 3 ⟶ "Yes"
- 7, 9 ⟶ missing (`NA`)

Then remove rows with missing values.

Hints:

- Use `mutate()` to create a new variable.
  - Use `case_when()` to recode. Briefly, this is coded as:

```
case_when(
  var_name == "old value 1" ~ "new value 1",
  var_name == "old value 2" ~ "new value 1",
  var_name == "old value 3" ~ "new value 1",
  var_name == "old value 4" ~ "new value 2",
  var_name == "old value 5" ~ NA_character_
)
```

  See the function help page for details and examples. Use `NA_character_` instead of the simpler `NA` (this specifies a missing text value, rather than a generic missing value).
- Use `drop_na()` to remove missing rows.

Question 2
How many rows are there in the depression dataset (after removing rows with missing values)?

3. Read in the sleep dataset.

- Use `read_xpt()` again to read the SLQ file.
- Select only `SEQN` and `SLD012`.

Hint:

- Use `select()`.

4. Create a binary sleep variable.

Create a new variable that classifies:

- `<` 7 hours → "Short"
- `>=` 7 hours → "Adequate"
- 77 and 99 → missing (`NA`)

Remove rows with missing values.

Hints:

- Use `mutate()` to create a new variable.
  - Inside mutate, use `case_when()`. to recode the values.
- Remove missing values with `drop_na()`.

Question 3
How many rows are there in the sleep dataset (after removing rows with missing values)?

5. Join the datasets.

Join the cleaned depression and sleep datasets using `SEQN`. After joining, keep only your two new binary variables.

Hint:

- Use `inner_join()`.
- Use `select()` to keep only the two new binary variables. Here the order matters, put the explanatory variable (risk factor) first, e.g. sleep, followed by the response variable (the health outcome), e.g. depression.
- Assign the joined data a new name.

Question 4
How many rows are there in the joined dataset?

Part 2 — Construct the Contingency Table

Statistical goal: Summarize the joint distribution of sleep and depressive symptoms.

6. Create a contingency table.

- Put sleep (risk factor) in the rows.
- Put depressive symptoms (outcome) in the columns.

Hint:

- Use `table()`.
- If you already have a tibble with the outcome and risk factors in that order, you can just give that whole tibble to `table()` . The first variable will be in the rows of the contingency table, the second variable will be in the columns.

Inspect the table carefully. Each cell represents a count of individuals in that category combination.

Question 5
Draw the contingency table here. (two columns, labeled; two rows, labeled)

7. Convert counts to conditional proportions.

We are interested in:

$$\Pr(\text{Depressed} \mid \text{Sleep Group})$$

Compute row-wise proportions.

Hint:

- Use `prop.table()`.
- Set `margin = 1` to compute proportions within rows. This assumes your group variable (risk factor, i.e. sleep) is on the rows of your contingency table.

Part 3 — Estimate Relative Risk

Statistical goal: Quantify the magnitude of association.

8. Extract the risk of depression for each sleep group.

Identify:

- Risk among short sleepers
- Risk among adequate sleepers

Then compute the relative risk:

$$RR = \frac{\text{Risk in Short Sleepers}}{\text{Risk in Adequate Sleepers}}$$

Hint:

- Extract specific cells from your proportion table using indexing, e.g.

  ```
  short_risk <- risk_table["Short", "Yes"]
  ```

  In this example, the square brackets are used to extract a single cell. "Short" identifies the row, and "Yes" identifies the column. Use the row and column names found in your risk table, which may vary depending on how you recoded the variables when preparing the data.

- Division (`/`) computes the relative risk.

Question 6
What is the risk of depressive symptoms in short sleepers? Long sleepers?

Question 7
What is the relative risk? Interpret your result in a complete sentence.

Part 4 — Test for Independence

Statistical goal: Determine whether the association is statistically significant.

9. State the hypotheses.

Write the null and alternative hypotheses in words.

Question 8
Write the null and alternative hypotheses in words

10. Perform the chi-square test of independence.

Use your contingency table to run the test.

Hint:

- Use `chisq.test()`.
- Give it the frequency table you calculated previously.

11. Interpret the results.

Report:

- The chi-square statistic
- The p-value
- Whether you reject or fail to reject the null hypothesis

Question 9
Write the χ^2 test statistic, P-value, and whether you reject or fail to reject the null hypothesis.

Then answer:

Question 10

Can we conclude that short sleep causes depression? Why or why not?

Part 5 (optional) — Mosaic Plot

If you finish the previous sections with time to spare, try the following optional challenge:

Create a mosaic plot of the contingency table.

Hints:

- Use the base `mosaicplot()` function
- Give it the contingency table you created with `table()`
- You can spruce it up by adding
  - a title `main = "Title"`
  - x and y labels e.g. `xlab = "X Axis Label"`
  - rotating the y axis label with `las = 1`
  - colors, e.g. `color = c("#0072b2", "#e69f00")`

If you get a beautiful, working graph, be sure the show the instructor, LAs, and your peers on every side. Great job!

Wrap-up and Submission

1. Make sure your script is saved in your project on Posit Cloud.
2. Ask the instructor to explain anything you aren't sure about.
3. **Show your handout to a Learning Assistant for a completion grade before you leave lab**. You may do this as soon as you finish. Keep the handout for yourself.