Project Homework 3

STAT011-S23

Due: 10/27/23

Introduction and Purpose

You will analyze two data sets in this homework assignment. The first data set contains information about a clinical trial testing drug to treat arthritis. This data set consists of 84 patients who either received a drug or a placebo and who rated the level of improvement in their arthritis symptoms. The drug or placebo label is the first categorical variable of interest, and the level of improvement in symptoms (none, some, or marked) is the second categorical variable of interest. This data set is referenced in Problems 1 – 3.

The second data set you will analyze in Problem 4 is the data you saw in the second week of class on the survival of different class passengers aboard the Titanic. This data consists of 2201 passengers who either survived the wreck or died and their status as a passenger in first, second, or third class, or as a crew member of the ship.

Problem 5 requires no data analysis.

The purpose of this assignment is to practice exploring categorical data, designing an experiment, and the using the laws of probability.

Required Tech

Excel

The skills necessary to complete this assignment in Excel are covered in the following seven videos:

- Excel 2016 with Data Analysis Toolpak Introduction to Excel 2016 with Data Analysis Toolpak (1:52)
- Excel 2016 with Data Analysis Toolpak Introduction to Excel 2016 with Data Analysis Toolpak: Common Procedures (3:08)
- Excel 2016 with Data Analysis Toolpak Descriptive Statistics and Confidence Intervals for a Mean (2:57)
- Excel 2016 with Data Analysis Toolpak Histogram (3:08)
- Excel 2016 with XLSTAT Video - Frequency Tables, Contingency Tables, Chi-square Test of Independence and Homogeneity (2:42)
- Excel 2010 with XLSTAT Video - Finding the Area Under the Normal Curve and Inverse Normality (4:04)

RStudio

The skills necessary to complete this assignment in RStudio are covered in the following seven videos:

- R Studio Video Introduction to R and RStudio (1:52)
- R Studio Video Getting Started (3:51)

- R Studio Video Working with Data Objects 1 (3:29)
- R Studio Video Working with Data Objects 2 (5:15)
- R Studio Video Importing Data (4:36)
- R Studio Video Descriptive Statistics (3:33)
- R Studio Video Plotting - Histograms, Bar Charts, Boxplots, Scatterplots (4:08)
- RStudio Video Probability Distributions (4:39)

Instructions

If you are analyzing this data in Excel you first need to download the arthritis data set for HW 3 and download the Titanic data set for HW 3 from our Stat 11 Github Data page. Do this by right clicking on the link "View Raw". Save each link with the names `arthritis.csv` and `titanic.csv`, respectively.

If you are analyzing this data in RStudio, you will import the data with the following commands

```
arthritis <- read.delim(
        "https://raw.githubusercontent.com/dr-suz/Stat11/main/Data/arthritis.csv",
        sep=",")

titanic <- read.delim(
        "https://raw.githubusercontent.com/dr-suz/Stat11/main/Data/titanic.csv",
        sep=",")
```

The first data object is called `arthritis` and the second data object is called `titanic`.

Once you have access to the data, complete all parts of the first four problems in this assignment. There is not data set associated with problem five. You are encouraged to work with your classmates on this assignment but you must hand in your own, unique write up of the solutions. In a Word document, clearly label each problem's solution. Most solutions will include graphics which can be copied from Excel or RStudio and pasted into your solution document. All solutions require a written component. When you are ready to submit your assignment, save the Word document as a PDF and upload it to the Moodle link for Project Homework #3.

Problem 1

Create either a proportions table or a contingency table for the 84 patients in this study and use it to answer the next two questions.

(a) What proportion of the patients received a placebo and had marked improvement?

(b) Of all the patients who received a placebo, what proportion of them saw

  (b.1) no improvement?

  (b.2) some improvement?

  (b.3) marked improvement?

Problem 2

Name and describe the kind of bias that might be present if administrators decide that instead of subjecting people to random drug testing they'll just

(a) Hold hospital-wide meetings and drug test the employees that attend; or instead

(b) Offer additional employee discounts for those employees who agree to be drug tested.

Problem 3

Researchers believe that a new drug called AX319 will help bones heal after children have broken or fractured a bone. The researchers believe that AX319 will work differently on bone breaks than on bone fractures. AX319 will be used in conjunction with traditional casts. To test the impact of AX319 on bone healing, the researchers recruit 18 children with bone breaks and 30 children with bone fractures.

Describe the design of an appropriate experiment to determine if AX319 will help bones heal.

Problem 4

Create either a proportions table or a contingency table for the Titanic data. Then answer the following questions about this data.

(a) What is the probability that a crew member survived the wreck?

(b) What is the probability that someone aboard survived the wreck, given that they were a third class passenger?

(c) What is the probability that someone aboard was a third class passenger?

(d) Does it appear that class status and survival are independent variables? Why/why not?

Problem 5

Answer all parts to the following three questions. Show your work and/or justify your reasoning in addition to providing an answer.

(a) The American Red Cross says that about 45% of the U.S. population has Type O blood, 40% Type A, 11% Type B, and the rest Type AB. Someone volunteers to give blood. Calculate the probability that this donor (a.1) has Type AB blood, (a.2) has Type A or Type B blood, and (a.3) does not have Type O blood.

(b) Suppose IQ scores tend to follow a $N(100, 16)$ distribution. What is the percent of people that you would expect to have an IQ between 112 and 132? (Hint: Draw a picture of the Normal curve and shade the region of interest. Note, RStudio software can compute this percentage for you exactly with a quantile function. If you're using Excel, you may wish to use a Z-table instead of R.)

(c) Consider the made-up probability distribution for a (discrete) random variable X shown in the table below. (c.1) Find the mean and standard deviation of this distribution and (c.2) Suppose another random variable Y is found by $Y = 3X - 2$. Use the rules for means (or expectations) and variances to find the mean and standard deviation of Y.

X	0	1	2
P(X=x)	**0.2**	**0.4**	**0.4**