PSY 611 Homework #2

YOUR NAME HERE

Contents

Instructions

Please complete this assignment using the RMarkdown file provided [Link to RMarkdown file here]. Once you download the RMarkdown file please (1) include your name in the preamble, (2) rename the file to include your last name (e.g., "weston-homework-1.Rmd"). When you turn in the assignment, include both the .Rmd and knitted .html files.

To receive full credit on this homework assignment, you must earn **30 points**. You may notice that the total number of points available to earn on this assignment is 65 – this means you do not have to answer all of the questions. You may choose which questions to answer. You cannot earn more than 30 points, but it may be worth attempting many questions. Here are a couple things to keep in mind:

1. Points are all-or-nothing, meaning you cannot receive partial credit if you correctly answer only some of the bullet points for a question. All must be answered correctly.
2. After the homework has been graded, you may retry questions to earn full credit, but you may only retry the questions you attempted on the first round.
3. The first time you complete this assignment, it must be turned in by 9am on the due date (October 25). Late assignments will receive 50% of the points earned. For example, if you correctly answer questions totaling 28 points, the assignment will receive 14 points. If you resubmit this assignment with corrected answers (a total of 30 points), the assignment will receive 15 points.
4. You may discuss homework assignments with your classmates; however, it is important that you complete each assignment on your own and do not simply copy someone else's code. If we believe one student has copied another's work, both students will receive a 0 on the homework assignment and will not be allowed to resubmit the assignment for points.

Data: Some of the questions in this assignment use the dataset referred to as `homework-avengers`. Link to data here. This dataset contains data on every movie in the Marvel Cinematic Universe from *Iron Man* (2008) through *Captain Marvel* (2019), including the budget, number of screens it was released on, year of release, and gross earnings in the United States (`DomesticGross`), over the first weekend (`WeekendGross`), internationally (`OverseasGross`) and total (`WorldwideGross`). It also contains a character variable for each character (pun not intended) in the MCU, which takes the value of `"yes"` if that character was in that movie and `"no"` if that character was not in that movie.

2-point questions

Question 1

You draw a sample of 43 cases and calculate the mean to be 102. Determine how probable it is that you would get a sample mean of 102 or more in a normally distributed population with the following parameters:

- $\mu = 105$, $\sigma = 11$
- $\mu = 94$, $\sigma = 5$
- $\mu = 83$, $\sigma = 25$
- $\mu = 110$, $\sigma = 50$
- $\mu = 121$, $\hat{\sigma} = 22$
- $\mu = 115$, $\hat{\sigma} = 75$

Question 2

A colleague designs a classroom test, administers it year after year to a very large number of students, claims the distribution is normal, and tells you that 50% of the scores fall between 45 and 67.

- What is the mean and standard deviation of this test?
- How likely would it be for a student to receive a grade of 83 or higher?

Question 3

You randomly sample the cell phone records of 5000 college students to determine the number of text messages sent during a typical school day (Monday through Friday). You find the mean to be 124.6 with a standard deviation of 21.2. What is the 99% confidence interval for the true mean based on these data?

Question 4

- Use the `pirates` dataset in the `{{yarrr}}` package to create a graphic that communicates the associations between the time it takes to draw a sword and number of treasure chests found, by sword type. Use the `facet_wrap()` function, and set the argument `scales` to `"free"`. Be sure your graph includes a title, properly labelled axes, and a caption.

Question 5

Use the `homework-avengers` to answer the following questions (Note: you may choose to use tidyverse functions or not, whichever make more sense to you):

- What's the average amount of money Marvel makes per movie?
- How many movies is Nick Fury in?
- Do Black Widow movies make more domestically or internationally?
- Does the average number of domestic screens per movie increase by year of release?

5-point questions

Question 1

The probability distribution functions beginning with r (e.g., `rbinom`, `rnorm`, `rchisq`, etc) can be used to simulate drawing from a known distribution many times, if you set the `n` parameter to a large integer, like 10000. Using this method of simulation, demonstrate each of the following:

- The distribution of squared Z-scores is equal to χ^2_1.
- The distribution of summed Z-scores is equal to χ^2_N. (Do this for N = 2, 5, and 10).

For each demonstration, do the following: 1. Plot a histogram of your simulated data and a density distribution based on a theoretical probability distribution on the same plot. 2. Show that the mean of your simulated data are approximately equal to the mean of your theoretical probability distribution.

Question 2

Which avengers character do you believe brings in the most money? Using the `homework-avengers` dataset, create a figure that demonstrates the relationship between budget and worldwide gross earnings. Color the geoms differently if the movie includes the your predicted character. Be sure your graph includes a title and appropriate labels.

- Based on this graph, what do you conclude regarding whether movies including this character make more than projected?

Question 3

Recall from lab 4 the data released by the Graduate Coffee Drinkers Association (GCDA) that the average graduate student drinks 5 cups of coffee a day and the distribution in the population is 1. You suspect that graduate students at the University of Oregon is unusual, in part because of the number of independent coffee roasters in the state and in part because they have to complete such arduous statistics assignments. You poll 10 of your fellow graduate students and find the average number of cups of coffee consumed is 5.7.

- What must be true of the way you conducted your poll in order for you to use NHST?
- What are your null and alternative hypotheses (in words or in math)?
- What kind of probability distribution will you use to test your hypotheses? **Justify your choice.**
- What alpha level will you set for your test. **Justify your choice.**
- What is the standard deviation of your sampling distribution?
- Conduct your statistical test (use a two-tailed test). What is the p-value? What do you conclude?
- Calculate a confidence interval around the population mean. The width of this confidence interval should be 1-[your alpha level]. Does it include your sample mean?
- Calculate a confidence interval around your sample mean. Does it include the population mean?

10-point questions

Question 1

Robert Downey Jr. has been on tour recently bragging that his presence alone brought millions of dollars to Marvel. (Not really, just for the sake of the homework example.) Use the `homework-avengers` dataset to answer the following questions.

- Do the movies featuring Iron Man (the character played by RDJ) make more money on average than movies that do not feature Iron Man?
- Maybe this is an artifact, driven by the fact that later Marvel movies made more money than earlier ones. If Iron Man appears in more movies per year over time, this would suggest there's a third variable (time) that accounts for this relationship. Create a figure that answers the question: does Iron Man appear in more movies over time? Interpret this figure: Could this explain why RDJ's movies made more?
- RDJ charges a lot of money for his acting services. Maybe we're less interested in the gross and more interested in the profit. Do the movies with Iron Man profit more than movies that don't? Create a figure to represent this, including the mean and 95% confidence intervals around the outcome(s). (Note that budget is reported in millions of dollars and gross is reported in dollars.) Interpret this graph: do movies with Iron Man have greater profit?

Question 2

A commonly used distribution is the Poisson distribution. This is useful when you want to estimate the number of times an event E occurs, but there is no specific set of trials, only a specific space or time interval. This distribution is especially useful for estimating the likelihood of experiencing a number of rare events, for example, the number of traffic accidents at a particular intersection. There is no set "number of trials" in this scenario – would it be the number of times a car approaches the interaction? the number of minutes a car is at an intersection? the number of seconds? You can see where this goes. We can't use the binomial, so instead, we turn to the Poisson.

The Poisson is a single-parameter distribution defined by λ ("lambda") which is the rate of event occurrence, or the typical number of events during the space or time interval of interest.

Use the distribution functions for the Poisson (`dpois`, `ppois`, `qpois` and `rpois`) to answer the following questions:

- Graph the probability distributions for $\lambda = 1$, $\lambda = 4$ and $\lambda = 10$. What do you notice about the shape of the Poisson as λ increases?
- Using `rpois` generate a large sample from the Poisson distribution defined by $\lambda = 1$. What is the expected value of the Poisson distribution defined by $\lambda = 1$? What about for the distributions defined by $\lambda = 4$ and $\lambda = 10$?
- Using `rpois` generate a large sample from the Poisson distribution defined by $\lambda = 1$. What is the variance of this distribution? Do the same thing for the distributions defined by $\lambda = 4$ and $\lambda = 10$.
- How would you express the mean and variance of the Poisson distribution in terms of its parameter, λ?
- Are the mean and variance of the Poisson distribution independent from each other?

20-point questions

Question 1

You would like to be 99% confident that the mean from a sample falls within K standard deviations of a population mean.

- Construct a graph showing how the magnitude of K changes with sample sizes ranging between 1 and 100.

- How does the relationship of K by change sample size if you're willing to be only 95% confident? Include both relationships in the same graph.
- How big does your sample size need to be if you want to be 99% confident that your sample mean is .5 standard deviations away from the population mean?