

Problem Set 3

Statistics 100

Due July 08, 2022 at 11:59 pm

Problem set policies. Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., " $SD = 3.3$ ") without explanation is not sufficient. For problems involving R, be sure to include the code in your solution.

Please submit your problem set via Canvas as a PDF, along with the R Markdown source file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TAs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Discussion Prompt.

For this week, choose between one of the following options:

- Find an example of a named distribution not discussed in the class. Briefly summarize the features of the distribution and how it is used; if you have found your favorite distribution¹, feel free to share why you consider it your favorite. Some interesting distributions are the [Weibull distribution](#) and [Gamma distribution](#). The second page of this [article](#) provides the names of many probability distributions.
- Share an example of how random variables are used in a real-world context. For example, you might enjoy reading about [modern portfolio theory](#)² if you are interested in finance. [This article](#) discusses the distribution of the number of births in a single day in England and Wales, [this article](#) discusses the methodology behind an election forecasting model, and [this article](#) discusses how differential privacy works. For something a bit more light-hearted, [this article](#) discusses examining the distribution of times popcorn kernels pop. Briefly summarize how random variables are used in your chosen example and comment on what you find interesting.
- [Benford's Law](#) refers to the observation that the leading digit in sets of numerical data follows a particular distribution (that is non-uniform). A number of reports have applied Benford's Law to COVID-19 data, including this article [assessing whether there may be misreporting of COVID-19 deaths in the US](#), this article [examining COVID-19 statistics in the US and worldwide](#), and this article [looking at the association between a country's COVID-19 reporting accuracy and development](#). Benford's Law has also been used to [detect fraud](#). Share an idea related to the application of Benford's Law that you find interesting; feel free to use these linked articles or investigate how Benford's Law has been used in other contexts.

To receive full participation credit, 1) write a post in the #discussion channel on Slack and 2) respond to someone else's post. The deadline for posting is the same as the problem set deadline.

¹A must for any statistician!

²For a more technical discussion, see [this resource](#) from the University of Washington.

Problem 1.

According to data from the CDC, about 37.1% of adults (individuals 18 years of age or older) in the United States and 57.9% of children (individuals between 6 months and 17 years of age) in the United States received a flu vaccine during the 2017-2018 flu season.

For any calculations, be sure to demonstrate the reasoning behind your calculations, such as by defining the relevant random variable(s), using probability notation, and/or briefly writing out your thought process.

- a) Consider a random sample of 50 adults from the Boston area.
 - i. Calculate the probability that exactly 20 adults received a flu vaccine.
 - ii. Calculate the probability that exactly 30 adults did not receive a flu vaccine.
- b) Consider a random sample of 20 children from the Boston area.
 - i. What is the probability that at most 10 children received a flu vaccine?
 - ii. What is the probability that at least 11 children received a flu vaccine?
- c) State two assumptions you needed to make in order to answer parts a) and b). Briefly comment on the extent to which those assumptions were reasonable.
- d) Consider a random sample of $n = 70$ individuals, which consists of 50 adults and 20 children. Let Z represent the total number of individuals who received the flu vaccine in the sample of 70 individuals. Does Z follow a binomial distribution? Explain your answer.

Problem 2.

Assume the annual returns on a stock portfolio are normally distributed with a mean of 14.7% and a standard deviation of 33%. A return of 0% indicates the value of the portfolio does not change.

For any calculations, be sure to demonstrate the reasoning behind your calculations, such as by defining the relevant random variable(s), using probability notation, and/or briefly writing out your thought process.

- a) What is the probability that in any given year the portfolio will lose money?
- b) What is the probability that in any given year the portfolio will have at least a 50% return?
- c) What is the probability that in any given year the portfolio will have a return between 25% and 75%?
- d) Calculate the return value that marks off the lowest 10% of annual returns for this portfolio.
- e) What is the probability that four of the next ten years will have a return greater than 50%? Comment on the validity of any assumptions required to make this calculation.

Problem 3.

The World Series is a postseason play-off series between the two respective champions of the two major professional baseball leagues in North America: the American League (AL) and the National League (NL), which together constitute Major League Baseball. The World Series is a best-of-seven playoff; i.e., it consists of at most 7 games and the first team to win 4 games wins the championship. It is not necessary for the games to be won consecutively and draws are not permitted. Once a winner is decided, the Series ends.

- a) Suppose the two teams in the World Series have equal probability of winning each of the 7 games and that the games are independent.
 - i. Calculate the probability that the World Series ends in 4 games.
 - ii. Calculate the probability that the World Series ends in 5 games.
 - iii. Calculate the probability that the World Series ends in 6 games.
 - iv. Calculate the probability that the World Series ends in 7 games.
- b) Let X represent the number of games that the World Series will consist of this year. Compute the mean and variance of X .

Problem 4.

[TidyTuesday](#) is a weekly podcast and community activity organized by the R4DS Online Learning Community, with the goal of helping R learners learn in real-world contexts. One of the datasets shared by TidyTuesday comes from Spotify and consists of 32,833 songs.³

After filtering down to unique tracks, there are 28,356 songs remaining. The distribution of song duration for these songs is moderately right skewed, with mean 226.6 seconds and standard deviation of 61.1 seconds.

- a) For a random sample of 30 tracks from the list of unique songs, what is the expected mean song duration?
- b) For a random sample of 30 tracks from the list of unique songs, what is the standard deviation of the mean song duration?
- c) How many tracks would you need to (randomly) sample for the standard deviation of the sample mean song duration to equal 7 seconds?
- d) Suppose that you create a randomly generated playlist of 50 songs to listen to during an upcoming flight. Compute the probability that the playlist duration is longer than 3 hours.

³If you're interested, check out the dataset and data dictionary [here](#).

Problem 5.

The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 US residents. The survey is conducted face-to-face with an in-person interview of a randomly selected sample of adults. One of the questions on the survey is “After an average workday, about how many hours do you have to relax or pursue activities that you enjoy?” A 95% confidence interval from the 2010 GSS survey for the collected answers is 3.53 to 3.83 hours.

Identify each of the following statements as true or false. Explain your reasoning.

- a) If the researchers wanted to report a confidence interval with a smaller margin of error based on the same sample of 1,154 Americans, the confidence interval would be larger.
- b) We can be 95% confident that the interval (3.53, 3.83) hours contains the mean hours that the sampled adults have for leisure time after an average workday.
- c) The confidence interval of (3.53, 3.83) hours contains the mean hours that U.S. adults have for leisure time after an average workday.
- d) The survey provides statistically significant evidence at the $\alpha = 0.05$ significance level that the mean hours U.S. adults have for leisure time after the average workday is 3.6 hours.
- e) There is a 5% chance that the interval (3.53, 3.83) hours does not contain the mean hours that U.S. adults have for leisure time after an average workday.
- f) The interval (3.53, 3.83) hours provides evidence at the $\alpha = 0.05$ significance level that U.S. adults, on average, have fewer than 3.9 hours of leisure time after a typical workday.