Assignment 4

Sri Harsha Sudalagunta

2023-03-01

Problem 1. Hemophilia.

Hemophilia is a sex-linked bleeding disorder that slows the blood clotting process. In severe cases of hemophilia, continued bleeding occurs after minor trauma or even in the absence of injury. Hemophilia affects 1 in 5,000 male births. In the United States, about 400 males are born with hemophilia each year; there are approximately 4,000,000 births per year. Note: this problem is best done using statistical software. (2 pts)

(a) What is the probability that at most 380 newborns in a year are born with hemophilia? (0.5 pts)

```r
ppois(380,lambda = 400)
```

```
## [1] 0.164859
```

(b) What is the probability that 450 or more newborns in a year are born with hemophilia? (0.5 pts)

```r
1-ppois(449, lambda = 400)
```

```
## [1] 0.007454327
```

(c) Consider a hypothetical country in which there are approximately 1.5 million births per year. If the incidence rate of hemophilia is equal to that in the US, how many newborns are expected to have hemophilia in a year, with what standard deviation? (1 pt)

```r
#The Probability of hemophilia is 1/ 5000 = 0.0002

# No of success i.e., no of new borns = 1500000*0.0002 = 300

#variance = n*p*(1-p)
n = 1500000
p = 0.0002
Variance = n*p*(1-p)
Variance
```

```
## [1] 299.94
```

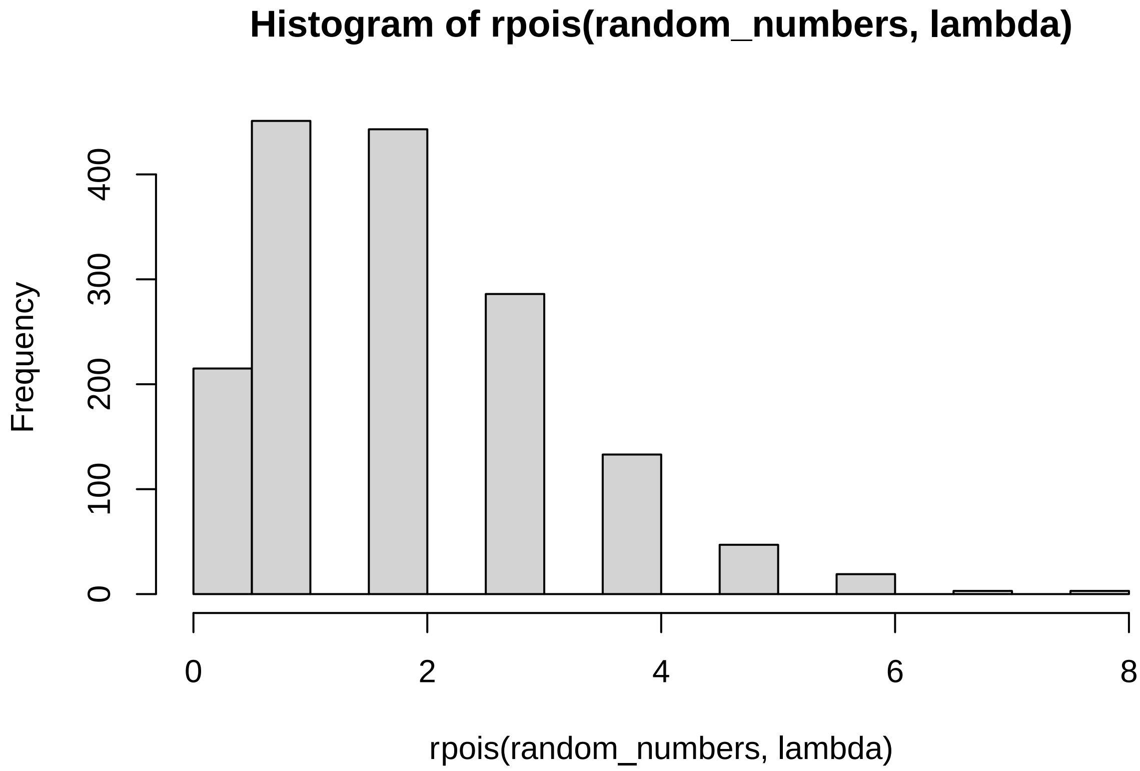
```
Standard_deviation = sqrt(Variance)
Standard_deviation
```

```
## [1] 17.31878
```

Problem 2. (2 pts)

(a)Generate 1600 random numbers from a Poisson distribution with lambda = 2. Plot the histogram of the generated numbers. (0.5 pts)

```
random_numbers = 1600
lambda = 2
hist(rpois(random_numbers, lambda))
```



(b) In a hospital, the number of patients admitted each day follows a Poisson distribution with a mean of 20 patients per day. What is the probability that the hospital will admit at most 25 patients in a day? (0.5 pts)

```
ppois(25,20)
```

```
## [1] 0.887815
```

(c) A hospital records an average of 5 patient arrivals per hour in the emergency room. What is the probability that at least 8 patients will arrive in the next 2 hours? (0.5 pts)

```
1-(ppois(7,lambda =10)) #considering lambda as 10 as we are calculating it for 2 hours
```

```
## [1] 0.7797794
```

(d) The number of cases of COVID-19 reported in a city follows a Poisson distribution with a mean of 10 cases per day. What is the probability that there will be no more than 70 cases reported in a week? (0.5 pts)

```
#lambda per day = 10
#lambda per week = 7*10
ppois(70,70)
```

```
## [1] 0.5317305
```

Problem 3.Hen eggs.

The distribution of the number of eggs laid by a certain species of hen during their breeding period is on average, 35 eggs, with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a distribution of sample means. (2 pts)

(a) What is this distribution called? (0.5 pts)

The distribution is called sampling distribution of sample mean.

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning. (0.5 pts)

According to central limit theory, I would anticipate the distribution to be symmetric, for the given sample size (n=45)

(c) Calculate the variability of this distribution and state the appropriate term used to refer to this value. (0.5 pts)

```
#it can be calculated using the standard error of the mean:

n =45
sd = 18.2

se= sd/sqrt(n)
se
```

```
## [1] 2.713096
```

(d) Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution? (0.5 pts)

Based on the formula of the standard error, the decrease in the number of sample will increase the variability of the distribution.

Problem 4.

The 2010 General Social Survey asked the question: “For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?” Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010. (2 pts)

(a) What does “95% confident” mean? Explain in the context of the application. (1 pt)

The “95% confident” mean that, we can be 95% sure that the 95% of the true population mean lies between 3.40 and 4.24 days.

(b) If a new survey were to be done with 500 Americans, would the standard error of the estimate be larger,smaller, or about the same? Assume the standard deviation has remained constant since 2010. (1 pt)

The standard error of the estimate would be smaller with a sample size of 500 compared to a sample size of 1,151, assuming the standard deviation remains constant.

Problem 5.

Write the null and alternative hypotheses in words and then symbols for each of the following situations. (2 pts)

(a) New York is known as “the city that never sleeps”. A random sample of 25 New Yorkers were asked how much sleep they get per night. Do these data provide convincing evidence that New Yorkers on average sleep less than 8 hours a night? (1 pt)

Null hypothesis: The average amount of sleep per night for New Yorkers is equal to 8 hours.

Alternative hypothesis: The average amount of sleep per night for New Yorkers is less than 8 hours.

H0: mu = 8

Ha: mu < 8 We need to conduct the hypothesis testing, and based on the p value we may conclude for the evidence. However, as the sample size looks smaller we may not able to generalize the results for the whole newyork.

(b) Employers at a firm are worried about the effect of March Madness, a basketball championship held each spring in the US, on employee productivity. They estimate that on a regular business day employees spend on average 15 minutes of company time checking personal email, making personal phone calls, etc. They also collect data on how much company time employees spend on such non- business activities during March Madness. They want to determine if these data provide convincing evidence that employee productivity decreases during March Madness. (1 pt)

Null hypothesis: The average amount of non-business activity time per employee during March Madness is the same as the average amount of non-business activity time per employee on a regular business day.

Alternative hypothesis: The average amount of non-business activity time per employee during March Madness is greater than the average amount of non-business activity time per employee on a regular business day.

H0: mu = 15min

Ha: mu > 15min

To conduct the test, the employers could collect a random sample of employees and record the amount of non-business activity time for each employee during both a regular business day and a day during March Madness. They could then calculate the mean and standard deviation of the differences between the two samples.

Next, they could use a one-sample t-test to compare the mean difference between the two samples to zero. If the p-value for the t-test is less than 0.05, they could reject the null hypothesis and conclude that there is convincing evidence that employee productivity decreases during March Madness.

Problem 6.

The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips? We have verified the independence, sample size, and skew conditions are satisfied. (2 pts)

```
mean <- 130
hypothesized_mean <- 134
sd <- 17
length <-(35)

t_stat <- (hypothesized_mean - mean) / (sd / sqrt(length))
t_stat
```

```
## [1] 1.392019
```

```
p_value <- 2 * pt(abs(t_stat), df = length - 1, lower.tail = FALSE)
p_value
```

```
## [1] 0.1729561
```

```
#based on the P-value and test statistics we cannot reject the null hypothesis.
```

Problem 7. Waiting at an ER,

A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning. (3.5 pts)

(a) This confidence interval is not valid since we do not know if the population distribution of the ER wait times is nearly Normal. (0.5 pts)

False, As the given sample size n =64 and the distribution of the sample mean, the distribution of the population will be approximately normal.

(b) We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.(0.5 pts)

False, we can only say that the 95% of times the mean of the population lies between the 128 minutes and 147 minutes.

(c) We are 95% confident that the average waiting time of all patients at this hospital's emergency room is between 128 and 147 minutes. (0.5 pts)

True, as the 95% confident interval means the same.

(d) 95% of random samples have a sample mean between 128 and 147 minutes. (0.5 pts)

False, the confidence interval is about the mean of the population lies in between 128 and 147 minutes in 95% of times.

(e) A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.(0.5 pts)

False, The 99% confidence interval would be wider than the 95% confidence interval as more data lies within the 99% confidence interval and to be more confident about the interval.

(f) The margin of error is 9.5 and the sample mean is 137.5. (0.5 pts)

True, the margin of error is half the width of the confidence interval, which is (147-128)/2 = 9.5. The sample mean is given as 137.5, which is the midpoint of the confidence interval.

(g) Halving the margin of error of a 95% confidence interval requires doubling the sample size. (0.5 pts)

False, Halving the margin of error of a 95% confidence interval requires increasing the sample size by a factor of four.

Problem 8.

Suppose an investigator takes a random sample of n = 50 birth weights from several teaching hospitals located in an inner-city neighborhood. In her random sample, the sample mean x is 3,150 grams and the standard deviation is 250 grams. (2 pts)

(a) Calculate a 95% confidence interval for the population mean birth weight in these hospitals.(1 pt)

```
n = 50
population_sample_mean = 3150
sd = 250

#standard error = standard deviation / sqrt(sample size)

#Assuming a normal population distribution and a 95% confidence level, the critical value for the z-sco

z = 1.96
se = sd/sqrt(n)

confidence_interval = c(population_sample_mean- z*se, population_sample_mean + z *se)

confidence_interval
```

```
## [1] 3080.704 3219.296
```

(b) The typical weight of a baby at birth for the US population is 3,250 grams. The investigator suspects that the birth weights of babies in these teaching hospitals is different than 3,250 grams, but she is not sure if it is smaller (from malnutrition) or larger (because of obesity prevalence in mothers giving birth at these hospitals). Carry out the hypothesis test that she would conduct. (1 pt)

```
#considering null hypothesis as Birth weight is equal to 3250.
#Alternative hypothesis as birth weight is less than or greater than 3250.

#two tailed t-test would be appropriate for the test

n = 50
population_sample_mean = 3150
sd = 250
Avg_mean = 3250

t_stats = (population_sample_mean- Avg_mean)/(sd/sqrt(n))

t_stats
```

```
## [1] -2.828427
```

```
#lets calculate the p value

#As we calculating the to tailed we can consider alpha = 0.025

p_value <- 2*pt(t_stats, df = n - 1, lower.tail = TRUE)
p_value
```

```
## [1] 0.006758541
```

Since the p_value is less than 0.025 she can reject the null hypothesis.

“‘

Problem 9.Testing for fibromyalgia.

A patient named Diana was diagnosed with fibromyalgia, a long-term syndrome of body pain, and was prescribed anti-depressants. Being the skeptic that she is, Diana didn’t initially believe that anti-depressants would help her symptoms. However after a couple months of being on the medication she decides that the anti-depressants are working, because she feels like her symptoms are in fact getting better. (1.5 pts)

(a) Write the hypotheses in words for Diana’s skeptical position when she started taking the anti-depressants. (0.5 pts)

Null Hypothesis: Anti-depressants do not help to reduce or treat symptoms of fibromyalgia.

Alternate Hypothesis: Anti-depresants do reduce or treat symptoms of fibromyalgia

(b) What is a Type 1 Error in this context? (0.5 pts)

Believing that anti-depressants do work for the fibromyalgia symptoms when they actually not.

(c) What is a Type 2 Error in this context? (0.5 pts)

Believing that anti-depressants do not work for the fibromyalgia symptoms when they actually do.