Homework 1

STT 461

02-24-2024

Contents

```r
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

1 Task 1 - Data science jobs and salaries 2020-2022

The dataset ds_salaries, includes information about salaries in the Data Science field during the years 2020 until 2022.

The variable employment_type has four levels:

- PT: Part-time
- FT: Full-time
- CT: Contract
- FL: Freelance

The variable experience_level has also four levels:

- EN: Entry-level / Junior
- MI: Mid-level / Intermediate
- SE: Senior-level / Expert
- EX: Executive-level / Director

Among the different variables recorded in this dataset, we can also find the type of job based on the specialization in Data Science, the salary_in_usd, the work_year, the country of the company and if the job was remote at a 0% to 100%.

1.1 Load the *ds_salaries* data:

To load the ds_salaries data:

1. Find your current working directory by using getwd()

```r
getwd()
```

2. Download and save the file ds_salaries.csv from D2L in your current working directory
3. Read the data with read.csv()

```
ds_salaries <- read.csv('ds_salaries.csv')[-1]
```

1.2 Assume that this dataset is representative of the population of employees in the data science field. Use simulations and the variables: *experience_level*, *salary_in_usd* and *company_location*, to estimate the probability of finding at least 8 employees with salary greater or equal to 100,000 USD in a group of 10 data entry level employees of companies in the US. Notice that companies located in the US have company_location = US.

```
set.seed(20)

# Filter dataset by company location in US and experience level of Entry
entries_us <- ds_salaries %>%
  filter(company_location == "US" & experience_level == "EN")

# Store salaries greater than or equal to 100,000
salaries_above_100K <- entries_us$salary_in_usd >= 100000

# Simulate 10,000 times to estimate the probability
sims <- replicate(10000, sum(sample(salaries_above_100K, 10, replace = TRUE)) >= 8)
mean(sims)
```

```
## [1] 0.0111
```

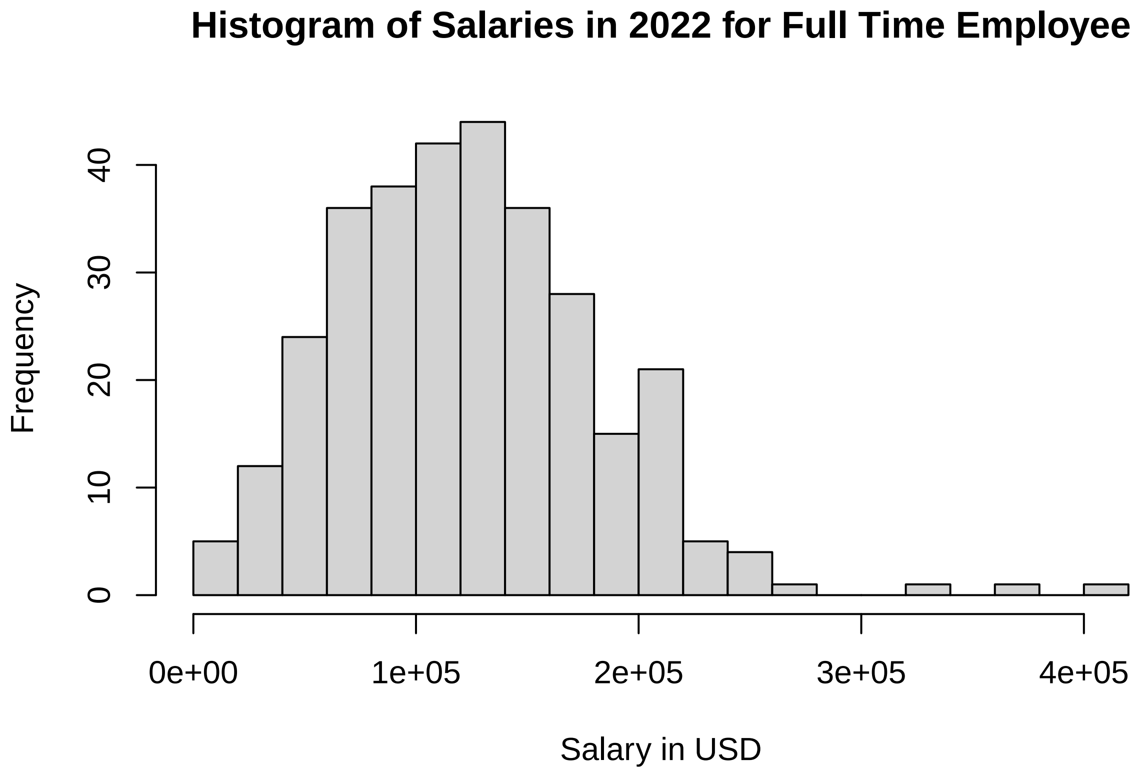
The estimated probability of finding at least 8 Entry Level employees with a salary greater than or equal to 100,000 USD in a group of 10 is 0.0111.

1.3 Plot the histrogram of *salary_in_usd* for the year 2022 for employees in the general data science field with full time employment. Observe that the salary data for this group are skewed. In those cases, the median is a better estimate of the center of the data than the mean. Provide a 90% confidence interval for the median *salary_in_usd* of the aforementioned group of employees, to estimate the center of their salary.

```
set.seed(20)

# Filter by full time and the year 2022
entries_2022 <- ds_salaries %>%
  filter(employment_type == "FT" & work_year == 2022)

# Plot the histogram
hist(entries_2022$salary_in_usd, main = "Histogram of Salaries in 2022 for Full Time Employee", xlab =
```



```r
# 90% confidence interval
# Simulate 10000
median_sims <- replicate(10000, median(sample(entries_2022$salary_in_usd, replace = TRUE)))

# Get the confidence interval
conf_interval <- quantile(median_sims, probs = c(0.05, 0.95))

# Print out the confidence interval
conf_interval
```

```
##       5%      95%
## 115821.2 129000.0
```

We can be 90% confident that the true median of salary lies between 115,821.20 USD and 129,000.00 USD.

1.4 In 2021, were employees with senior experience paid better than those with mid/intermediate level experience? Conduct a test with significance level of $\alpha = 0.01$ to answer the above question.

```r
set.seed(20)

alpha = 0.01
```

```r
# Filter for proper data sets
senior_exp <- ds_salaries %>%
  filter(experience_level == "SE" & work_year == 2021)

mid.int_exp <- ds_salaries %>%
  filter(experience_level == "MI" & work_year == 2021)

# Get the medians of salary of both experience levels
median_senior_sal <- median(senior_exp$salary_in_usd)
median_mid.int_sal <- median(mid.int_exp$salary_in_usd)

# Calculate the observed difference in median salary
obsv_diff <- median_senior_sal - median_mid.int_sal

# Bootstrap sampling
bootstrap <- replicate(10000, {
  seniors <- sample(senior_exp$salary_in_usd, length(senior_exp), replace=TRUE)
  mid.int <- sample(mid.int_exp$salary_in_usd, length(mid.int_exp), replace=TRUE)

  median(seniors) - median(mid.int)
})

# Calculate the p-value
p_value <- mean(abs(bootstrap) >= abs(obsv_diff))

p_value
```

```
## [1] 0.5271
```

The calculated p-value is larger than alpha. This leads us to reject the null hypothesis, telling us that there is not a statistically significant difference between the median salaries of senior and mid/intermediate level employees.

1.5 A friend of yours is interested in working for the first time in the data science field by finding a data entry position. Your friend can work either in Germany or in the US. Using the salary data for 2021, make a suggestion to your friend on which country has companies with better salary opportunities for data entry employees by conducting a test with significance level $\alpha = 0.01$. Notice that companies located in Germany have company_location = DE.

```r
set.seed(20)

us_salaries <- ds_salaries %>%
  filter(experience_level == "EN" & company_location == "US" & work_year == 2021)

de_salaries <- ds_salaries %>%
  filter(experience_level == "EN" & company_location == "DE" & work_year == 2021)

median_us_salary <- median(us_salaries$salary_in_usd)
```

```r
median_de_salary <- median(de_salaries$salary_in_usd)

# Observed difference
obsv_diff <- median_us_salary - median_de_salary

# Bootstrap sampling
bootstrap <- replicate(10000, {
  us_sal <- sample(us_salaries$salary_in_usd, length(us_salaries), replace=TRUE)
  de_sal <- sample(de_salaries$salary_in_usd, length(de_salaries), replace=TRUE)

  median(us_sal) - median(de_sal)
})

# Calculate the p-value
p_value <- mean(abs(bootstrap) >= abs(obsv_diff))

p_value
```

```
## [1] 0.5033
```

Due to the high p-value calculated after performing bootstrap sampling and comparing the medians of the salaries from the US and Germany, I would suggest to my friend that there is not a large statistical significance between salaries of the two countries. So, they could choose where they would want to go.

2 Task 2 - Robustness of the parametric t confidence interval for a mean and the t-test for a mean.

In class we explored the robustness of the parametric t-test when the assumption of independent values is not met.

In this task you will use simulations to demonstrate how Type I error and confidence level are affected when the assumption of normal population data is not met.

2.1 Estimate the type I error rate in a t-test of Ho : $\mu = 1$ versus Ha : $\mu \neq 1$ when the underlying population is exponential with rate 1. Use a sample of size n = 20 and test at the $\alpha = 0.05$ significance level. Did the skewness of the sample data affected the type I error rate (hint: compare the estimated type I error rate with α.)

```r
set.seed(20)

n <- 20
alpha <- 0.05

# Perform simulations
errors_sim <- replicate(10000, {
  sample <- rexp(n, rate = 1)
  test <- t.test(sample, mu = 1)
  test$p.value < alpha     # Test if p-value from t.test is less than alpha
})

mean(errors_sim)
```

```
## [1] 0.0797
```

The mean p-value from the simulations is greater than alpha. This suggests that the skewness affects the test, leading to more Type I errors where the null hypothesis is incorrectly rejected.

2.2 Choose 10 random values of X having an exponential distribution with rate 1/3. Use t.test to compute the 95% confidence interval for the mean. Is the true mean, 3, in your confidence interval?

```r
set.seed(20)

n <- 10

sample <- rexp(n, rate = 1/3)

conf_int <- t.test(sample, conf.level = 0.95)$conf.int
conf_int
```

```
## [1] -0.1723597  7.9744486
## attr(,"conf.level")
## [1] 0.95
```

The 95% confidence interval is (-0.17, 7.97). The true mean of 3 is within the confidence interval.

2.3 Check if the assumptions for the appropriate t.test are met. Replicate the experiment in part (2.2) 10,000 times and compute the percentage of times that the population mean 3 was included in the confidence interval. Explain why this number is not 95%.

```r
set.seed(20)

n <- 10

ci_sims <- replicate(10000, {
  sample <- rexp(n, rate = 1/3)
  conf_int <- t.test(sample, conf.level = 0.95)$conf.int

  # Test if 3 is within the confidence interval
  conf_int[1] <= 3 && 3 <= conf_int[2]
})

# Calculate the percentage from the simulations
mean(ci_sims) * 100
```

```
## [1] 89.81
```

The result is 89.81%, meaning that the true mean of 3 is only within the bounds of the 95% confidence interval 89.81% of the time instead of 95%. This is because due to the small value of n, 10, and the population being exponential and heavily skewed, the population cannot follow the Central Limit Theorem. The t.test expects a normal distribution which the sample data is not.

2.4 Repeat part (2.3) using 100 values of X to make each confidence interval. What percentage of times did the confidence interval contain the population mean? Why is it closer to 95%?

```r
set.seed(20)

n <- 100

ci_sims <- replicate(10000, {
  sample <- rexp(n, rate = 1/3)
  conf_int <- t.test(sample, conf.level = 0.95)$conf.int

  # Test if 3 is within the confidence interval
  conf_int[1] <= 3 && 3 <= conf_int[2]
})

# Calculate the percentage from the simulations
mean(ci_sims) * 100
```

```
## [1] 94.28
```

The percentage of times that the confidence interval contained the true population mean in this test is 94.28%, which is much closer to 95%. This is because, with a larger sample size, the distribution becomes slightly closer to an approximately normal distribution due to the Central Limit Theorem. This slightly improves the t.test results since t.test assumes an approximately normal distribution.