Midterm1_Keck_Brandon_STAT630

Brandon Keck

2024-10-13

```
library(ggplot2)
library(dplyr)
library(readr)
```

```
# You copy and paste this in from when you import the data. But it has to be in the same folder.
library(readr)
labor <- read_csv("labor.csv")
```

```
glimpse(labor)
```

```
## Rows: 753
## Columns: 7
## $ labor_force      <chr> "No", "No", "No", "No", "No", "No", "No", "No", …
## $ kids_under6      <dbl> 0, 0, 0, 2, 0, 0, 1, 2, 0, 0, 0, 0, 1, 0, 0, 1, …
## $ kids6_18         <dbl> 3, 0, 0, 3, 2, 2, 2, 0, 3, 1, 1, 1, 1, 0, 2, 0, 0, 1, …
## $ age              <dbl> 39, 60, 43, 31, 40, 36, 32, 39, 42, 53, 48, 44, 31, 48…
## $ wife_college     <chr> "No", "No", "No", "No", "Yes", "No", "No", "No", "No",…
## $ husband_college <chr> "Yes", "Yes", "Yes", "Yes", "No", "Yes", "Yes", "Yes",…
## $ family_income    <dbl> 28.363, 24.984, 9.952, 10.000, 28.200, 5.330, 6.800, 7…
```

```
#install.packages('gtsummary') # Make sure to comment out once ran
library(gtsummary) # This calls the gtsummary library
```

Revised

(1)

If you had all the resources and time in the world, how would you obtain a sample of women to be a part of your study? Describe your proposed sampling method in some detail.

Since we are focusing on a specific group (women) within a population, I believe that using a stratified sampling method would be the most appropriate. This method ensures that we account for variability within subgroups of the population. To ensure that all age groups of women are adequetly represented, I would divide the women into the following age strata: 18-25, 26-35, 36-45, 46-55, 56+. By radonly sampling from each of these age groups, we can capture the different stages of workforce participation, ensuring representation from younger women just entering the workforce, mid-career women, and older women nearing retirement age. This approach would provide a balanced view of labor force participation across different life stages, avoiding bias toward any specific age group.

Revised

(2)

Ideally, we want to generalize the results of our study to all women who are working age in the U.S. Based on your chosen sampling method in the previous question, what is the population of women that you can generalize your results to (i.e., is there any bias you may be concerned about)?

One possible bias could arise if certain groups of women, particularly younger women aged 18-25 are under-represented in the sample. Younger women might be harder to reach for various reasons-they may be less likely to participate in surveys or they may still be in school. They may not be fully engaged in the workforce at all. If we do not sample this group adequately the results of the study could be skewed toward older women who are more established in their careers.

(3) What is a variable (NOT listed in the dataset) that you think would be a useful factor in determining whether or not a woman participates in the paid labor force? Explain.

It might be useful to know what industries women are working in. That way we could see what areas of the workforce most women worked in 1975. This would be useful because then we might be able to predict from a women's college education what industry they were most likely to pursue after college. We could also witness any correlation between college education and that specific industry.

(4) The variable kids_under6 has the values 0, 1, and 2. Do you think this variable should be treated as an integer or a factor variable? Explain your reasoning.

I believe that this variable should be treated as a factor variable. This is because this has the values 0,1, and 2 which means that kids_under6 has three levels and would be easier to code if we were to change it to a factor variable.

Revised

(5) Using the R package of your choice (or manually creating in markdown), create the following table of summary statistics. Calculate the mean and standard deviation for quantitative data and the counts and percentages for categorical data.

```
labor %>%
  tbl_summary(by = labor_force,
              include = c(kids_under6, age, wife_college, husband_college, family_income),
              digits = list(
                all_continuous() ~ c(2,2)
              ),
              statistic = list(
                all_continuous() ~ "{mean} ({sd})"
                ))
```

Characteristic	No N = 325[1]	Yes N = 428[1]
kids_under6		
0	231 (71%)	375 (88%)
1	72 (22%)	46 (11%)
2	22 (6.8%)	7 (1.6%)
age	43.28 (8.47)	41.97 (7.72)
wife_college	68 (21%)	144 (34%)
husband_college	207 (64%)	251 (59%)
family_income	21.70 (12.73)	18.94 (10.59)

[1] n (%); Mean (SD)

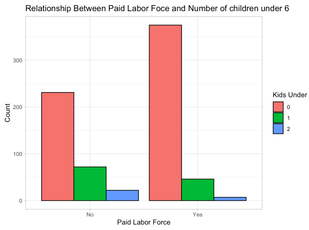
```
# Used from lectures notes with the help of Dr. Moore
```

Revised

(6) Create a well-labeled plot to visualize the relationship between participation in the paid labor force and number of children under 6.

```
labor$kids_under6 <- as.factor(labor$kids_under6) # Coerce kids_under6 to a factor variable
```

```
ggplot(data = labor, aes(x = labor_force, fill = kids_under6)) +
  geom_bar(position = "dodge", color = "black") +
  labs(title = "Relationship Between Paid Labor Foce and Number of children under 6",
       x = "Paid Labor Force",
       y = "Count",
       fill = "Kids Under 6") +
  theme_light()
```



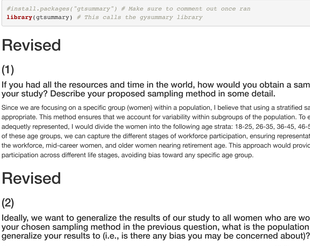
(7) Using your plot and the summary statistics you calculated in question 5, comment on any similarities or differences between whether or not a woman works and the number of kids she has under the age of 6.

The similarities between whether or not a women works and the number of kids she has under the age of 6 are for both groups that have 0 kids they are pretty similar in that 71% of women who do not work have 0 kids while 88% of women who do not have kids do work. The differences that we see between whether or not a woman works and the number of kids she has under the age of 6 is when women have 1 kid we see that 22% do not work while women who do have 1 kid under 6 only 11% work. Even more of a difference we see that women who have 2 kids under 6 6.8% do not work while only 1.6% do work.

Revised

(8) Create a well-labeled plot to visualize the relationship between participation in the paid labor force, number of children under 6, and age.

```
ggplot(data = labor, aes(x = labor_force, y = kids_under6, color = age)) +
  geom_boxplot() + # Used age
  # facet_wrap(~age) +
  labs(title = "Relationship Between Paid Labor Force, Number of children under 6 and Age",
       x = "Paid Labor Force",
       y = "Children under 6",
       fill = "age") +
  theme_light()
```



(9)

Revised

(10) In this problem, you will create a 95% confidence interval for the true proportion of women who participated in the paid labor force in the 1970's.

(a) First, check that both the necessary conditions for the Central Limit Theorem are met. Show all work/code and explain why the condition is met or not.

```
n <- 752 # Sample size of Women in the work force
p <- 428/752 # Proportion of Women working

# Calculate success and failures
success_condition <- n * p
failures_condition <- n * (1-p)

success_condition
```

```
## [1] 428
```

```
failures_condition
```

```
## [1] 324
```

In order for the Central Limit Theorem to be applied two conditions must be met

1. Independence this is met if the observation in our sample are independent of one another. We can assume the sample that was obtained is not influenced by the participation of other women.
2. Success-Failure Condition: We calculated the success and failure conditions: n * p = 752 * (326/752) = 428 > 10 and n * (1-p) = 752 * (324/752) = 324 > 10 Both conditions are met so now can apply the Central Limit Theorem.

```
n <- 752 # Sample size
phat <- 428/752 # Proportion of Women in Work force


ci_low <- phat - qnorm(0.975) * sqrt((phat*(1-phat))/n) # upper bound
ci_high <- phat + qnorm(0.975) * sqrt((phat*(1-phat))/n) # lower bound


print(c(ci_low, ci_high))
```

```
## [1] 0.5337561 0.6045418
```

We are 95% confident that the true proportion of women who participated in the paid labor force in the 1970's lies between 53.4% and 60.5%

Revised

(11) According to the internet (so this may or may not be true), 40% of married women were employed by 1970. Based on the confidence interval you calculated, does this percentage seem reasonable? Comment on why or why not.

Based on the 95% confidence interval that I calculated earlier, which ranges from 53-60% the 40% figure reported from the internet does NOT seem reasonable. This value falls outside of our range from our confidence interval suggesting that 40% is likely an underestimate of women who worked in the 1970s.

12) Give yourself a rating for this assignment using the EMRN rubric.

E - Excellent

M - Meeting expectations

R - Revision needed

N - Not assessible (mostly blank or did not complete)

R- I wasn't able to finish as much as I would have liked. I got stuck on a couple of things.