

Midterm1_Keck_Brandon_STAT630

Brandon Keck

2024-10-15

```
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, 1, 0, 0, 1, ~
## #> #> $ kids6_18          <dbl> 3, 0, 0, 3, 2, 2, 2, 6, 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", "Yes", "No", "No", "No", "No", "No", ~
## #> #> $ husband_college    <chr> "Yes", "Yes", "Yes", "Yes", "No", "Yes", "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 gysummary library
```

(1) Revised

My misunderstanding stems from the inability to accurately articulate my response. I knew that I wanted to stratified sampling method and was wanting to sample particular age groups. However, I did not go into enough detail into how one would perform this experiment. The reason why I believe this response is correct is because I have no detailed how I would do a stratified sample of women and their age. In order for us to adequately capture women in the working class we have stratified all age groups of women.

(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 adequately represented, I would divide the

women into the following age strata: 18-25, 26-35, 36-45, 46-55, 56+. By randomly 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.

(2) Revised

Again my misunderstanding was not going into enough detail about how there might be sampling bias. I was worried about only sampling women of a particular age. Perhaps women who were only going to sample women aged 36-45, 46-55, 56+ and not 18-25, 26-35 for some reason. I don't think that I had thought the whole experiment through in my head. Why I believe this to be correct is because now I have had enough time to really think about how I would conduct a stratified sample for this experiment. Because I have thought about it I now have a good example of how I would stratify a certain group. In this case that group is working age of women.

(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 woman'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.

(5) Revised

I was confused about what the question was asking me. I had all of the code at the ready for the exam and once I saw to create a summary statistics table my mind immediately went to HW2 where we created a summary statistics table. However I did not realize that there were certain variables that we were not interested in generating a table for. Why I believe this is now correct is because I have looked at the quantitative data and the counts of percentages for categorical data and have generated a summary statistics table to reflect

the question more appropriately. In general I have created a summary table of the variables that we are interested in.

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

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

¹Mean (SD); n (%)

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

(6) Revised

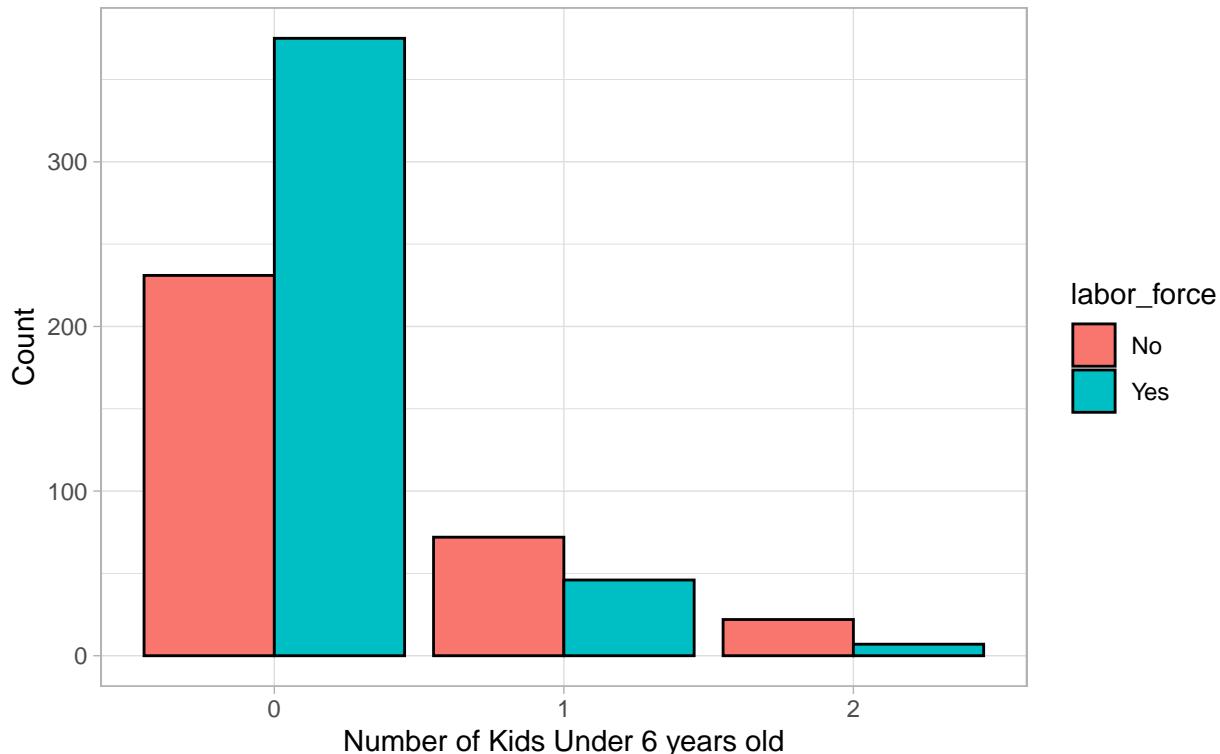
This is where I started to have difficulty as I was not as confident about the subject as I had thought. I had spent a great deal of time on this problem and wasn't getting the results I was hoping for. Why I believe this is now correct is from the help of Dr. Moore suggesting to coerce the kids_under6 variable and looking back at homework and notes in order to create a plot that visualizes the relationship between labor_force and kids_under6.

(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 = kids_under6, fill = labor_force)) +
  geom_bar(position = "dodge", color = "black") +
  labs(title = "Relationship Between Paid Labor Force \n and Number of children under 6",
       x = "Number of Kids Under 6 years old",
       y = "Count") +
  theme_light()
```

**Relationship Between Paid Labor Force
and Number of children under 6**



(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 woman 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.

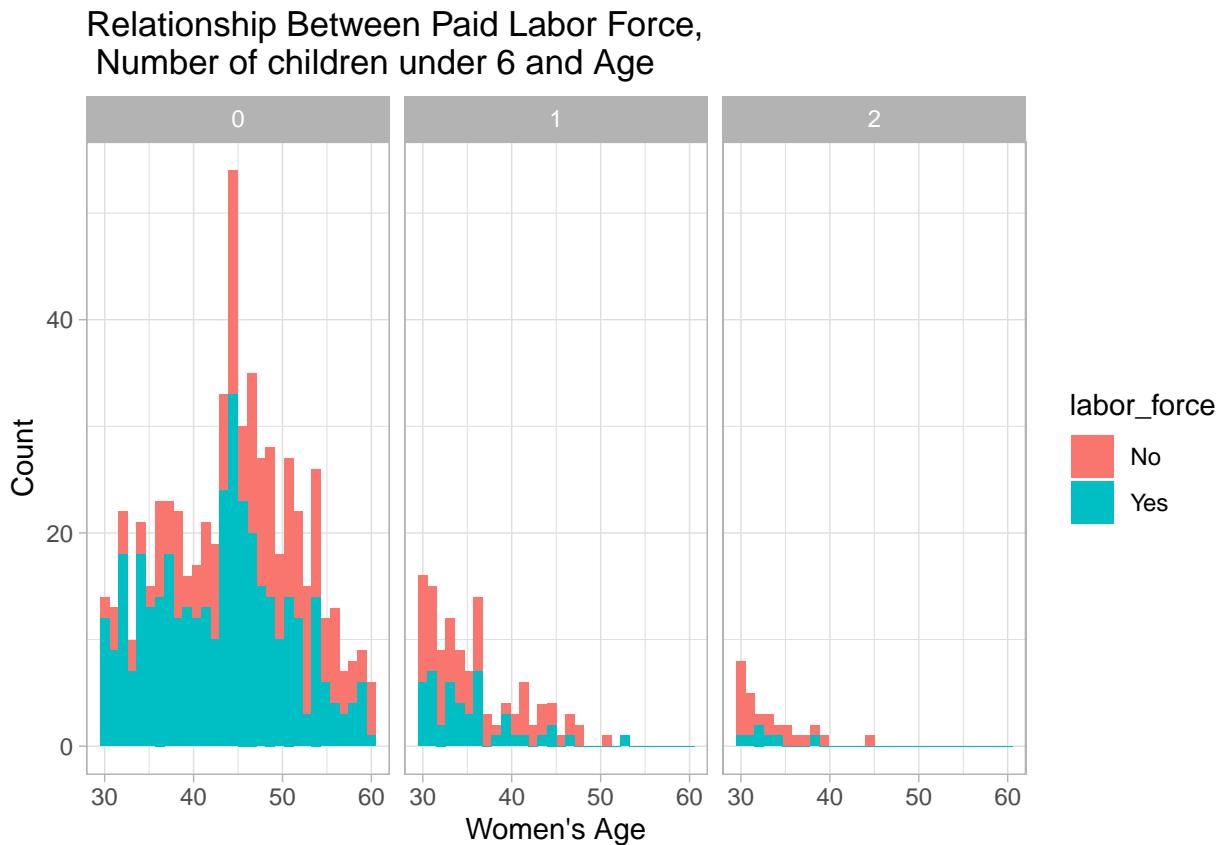
(8) Revised

I misunderstood what type of graph would be best used here. I believe initially I was attempting to create a boxplot however that wasn't working for me. I believe this graph to be correct because I am now calling the correct variables. I believe initially had my x variable as kids_under6 and my y variable as labor_force

and my fill as age. Now I have a nice histogram that represents the relationship between participation in the paid labor force, number of children under 6, and age.

(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 = age, fill = labor_force)) +  
  geom_histogram() +  
  facet_wrap(~kids_under6) +  
  labs(title = "Relationship Between Paid Labor Force, \n Number of children under 6 and Age",  
       x = "Women's Age",  
       y = "Count") +  
  theme_light()
```



(9) Revised

I was unable to create the graph above in order to respond to this question. If I was able to generate the graph I could have answered. The reason why I believe this response is correct is because I now have a graph where I can see the affect the relationship between workforce participation and the number of kids under the age of 6.

(9) Now that you have added age to the plot, how does that affect the relationship between workforce participation and the number of kids under the age of 6? Comment

From the histogram we can see that women who do not have children or rather children at least under 6 years old have a higher amount of participation in the workforce. What we can also see is that as women tend to get older the participation in the workforce also goes down.

(10) Revised

I was confused as to how to calculate the 95% confidence interval for the true proportion. I had my notes open and I had copied some of that down but by this point I was running out of time and was trying to make sure I had answered everything else correctly. The reason why I believe this is now correct is because I have a sample size of the women who participated and the proportion of women who were a part of the workforce. This is from the summary statistics that was generated earlier. So now I have the correct calculations for this problem.

(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.

```
N <- 753 # Sample size of Women in the work force
n <- 428 # women in labor force
p <- 428/753 # Proportion of Women working

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

success_condition
```

(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.

```
## [1] 243.2722
```

```
failures_condition
```

```
## [1] 184.7278
```

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: $n * p = 428 * (428/753) = 243.2722 > 10$ and $n * (1-p) = 428 * (1 - 428/753) = 184.7278 > 10$ Both conditions are met so now we can apply the Central Limit Theorem.

(10b) Revised

I did not have enough time to complete this question. In order to answer this we needed to calculate the 95% confidence interval and I kept struggling to answer that. Now that I believe I have correctly calculated the confidence interval I believe this is now correct. The reason why I believe it is correct is because I now have the correct proportion of women in the work force calculated.

```
phat <- p # 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))
```

(b)

```
## [1] 0.5214691 0.6153171
```

We are 95% confident that the true proportion of women who participated in the paid labor force in the 1970's lies between 52.15% and 61.53%

(11) Revised

I did not have enough time to answer this question. I was still trying to figure out earlier problems and so I never was able to complete this section. The reason why I believe this response to be correct is because I am now able to correctly calculate the 95% confidence interval. Using that information I am able to respond to the question.

(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 52-61% 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)

With the new revisions I believe that I earned an E.