Midterm 1 Revisions

David Teng

Oct 15

1. (revised)

Misunderstanding : I used Google suggestions rather than my own words.

Revision: I intend to segment the U.S. by state and randomly sample working-age women (ages 18–65) from each state through carefully structured surveys or interviews to ensure minimal bias. This approach will help me explore the factors that affect women's involvement in the workforce. For this research, I will implement a stratified sampling method.

Justification: I included working and non-working women (ages 18–65) in the study and divided the U.S. by state to make it easier to obtain a representative sample while reducing bias and ensuring inclusivity.

2. (revised)

Misunderstanding : I used Google suggestions but didn't include the potential biases (insufficient sample size).

Revisions: A sample size of 793 might be insufficient to reflect the full diversity of women in the U.S. Since individuals are shaped by unique life experiences, no two people are exactly the same. If certain groups of women are underrepresented or missing from the sample, the findings may be too limited to apply broadly (generalize) to the entire population.

Justification: This revised version better addresses the potential biases in terms of sample size and representation. Ideally, in stratified sampling, if the sample is large enough and properly stratified, it can still lead to more generalizable (or trustful) results.

3. (revised)

Misunderstanding : I used Google suggestions instead of my own words, and I only believe that parents' educational level aligns with the existing variables 'wife_college' and 'husband_college' but didn't create a new useful factor.

Revisions: In my opinions parental education could be a key factor in influencing women's participation in paid employment. Well-educated parents are more likely to offer valuable guidance and childcare, helping their children to overcome obstacles and feel more self-assured when entering the workforce. This support can contribute to stronger performance in paid employment.

Justification: This is a strong answer because it introduces a relevant variable: parents' educational level that likely influences job-seeking skills and opportunities. It shows how early family influences can shape career choices and job-seeking skills, which is important when analyzing workforce participation.

4. (revised)

Misunderstanding : My conclusion is correct, but I used too many professional terms like 'categorical nature,' 'interpretability,' 'predictive modeling,' and etc. which didn't really fit the question.

Revisions: I regard the variable 'kids_under6' should be considered a categorical factor. It is a nominal variable that reflects how many children a woman has under the age of 6. The values (0, 1, and 2) simply indicate whether she has 0, 1, or 2 children, without any inherent ranking or order. Additionally, it is not a quantitative variable, as it cannot be used for mathematical calculations in a meaningful way.

Justification: This answer fits the question well because it clearly explains reasons for treating 'kids_under6' as a factor variable. It emphasizes that the variable is categorical and lacks a natural order and also justifies why it is inappropriate to treat it as an "integer". It explains the difference between categorical and quantitative , nominal and ordinal as well.

5. (revised)

```
library(ggplot2)
library(dplyr)
library(gtsummary)
library(openintro)
library(readr)
labor <- read_csv(here::here("data/labor.csv"))
glimpse(labor)
```

```
## Rows: 753
## Columns: 7
## $ labor_force     <chr> "No", "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", "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~
```

Misunderstanding : I didn't divide the variables into 'Does not participate' and 'Participates,' as indicated in the prompt.

My original answer:

```
# my original answer
summary_table <-
  labor %>%
  select(age,
         kids_under6,
         wife_college,
         husband_college,
       family_income) %>%
 tbl_summary(

 statistic = list(all_continuous() ~ "{mean} ({sd})",
 all_categorical() ~ "{n} ({p}%)"))

 summary_table
```

Revisions:

```
labor %>%
  select(!kids6_18) %>%
  mutate(labor_force = recode(labor_force, `No` = "Does not participate", `Yes` = "Participates")) %>%
  tbl_summary(
    by = labor_force,
    digits = list(all_continuous() ~ c(2, 2)),
    statistic = all_continuous() ~ "{mean} ({sd})"
  )
```

Justification: I created a table showing summary statistics that matches the prompt: it calculates the mean and standard deviation for quantitative data, as well as the counts and percentages for categorical data and

Characteristic	**N = 753**[1]
age	43 (8)
kids_under6	
0	606 (80%)
1	118 (16%)
2	29 (3.9%)
wife_college	212 (28%)
husband_college	458 (61%)
family_income	20 (12)

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

Characteristic	**Does not participate** N = 325[1]	**Participates** 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)

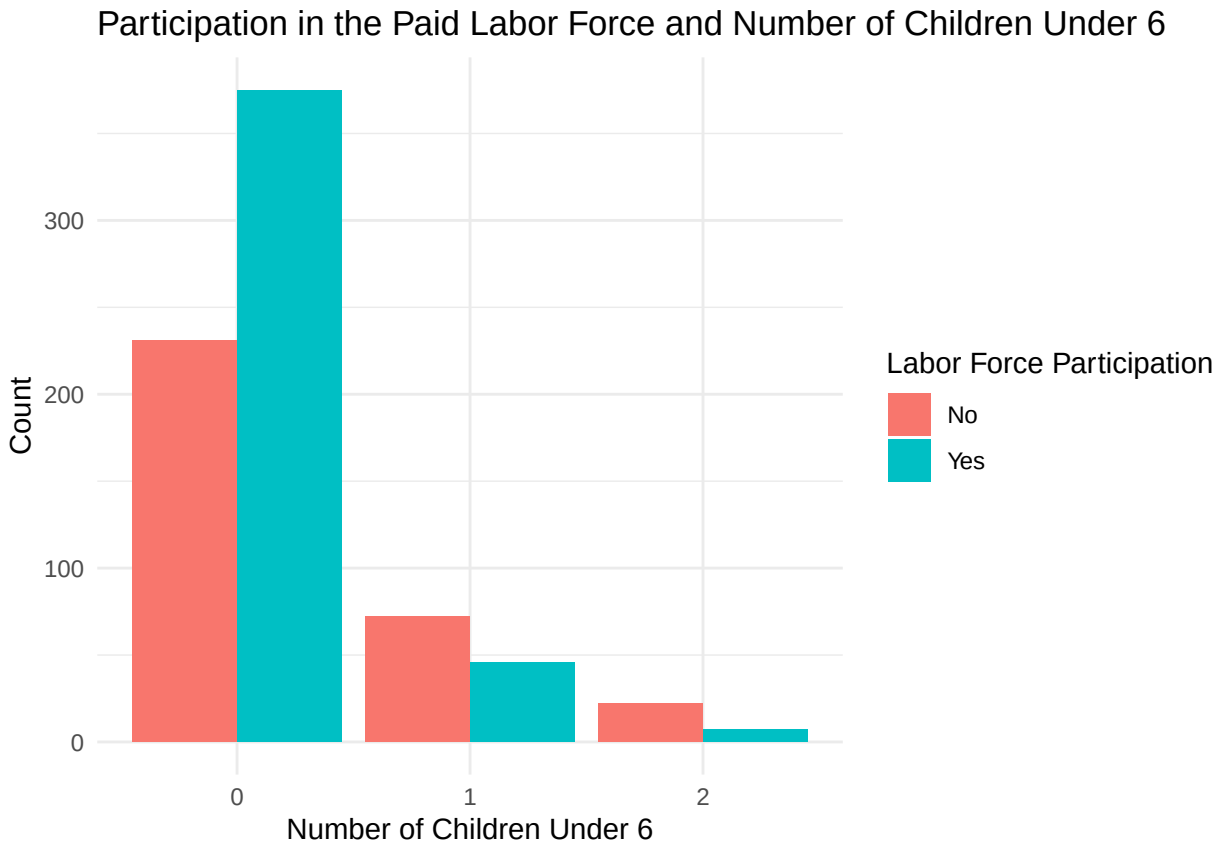
divided the variables into 'Does not participate' and 'Participates.'

6. (original)

Misunderstanding : No

(original answer)

```
#my original answer
labor <- labor %>%
mutate(
labor_force = factor(labor_force, levels = c("No", "Yes")),
kids_under6 = as.factor(kids_under6)
)
ggplot(labor, aes(x = kids_under6, fill = labor_force)) +
geom_bar(position = "dodge") +
labs(
title = "Participation in the Paid Labor Force and Number of Children Under 6",
x = "Number of Children Under 6",
y = "Count",
fill = "Labor Force Participation"
) +
theme_minimal()
```



Justification: I created a plot to visualize the relationship between participation in the paid labor force and number of children under 6.

7. (revised)

Misunderstanding : I didn't have enough time so I just copied my plot and the summary statistics.

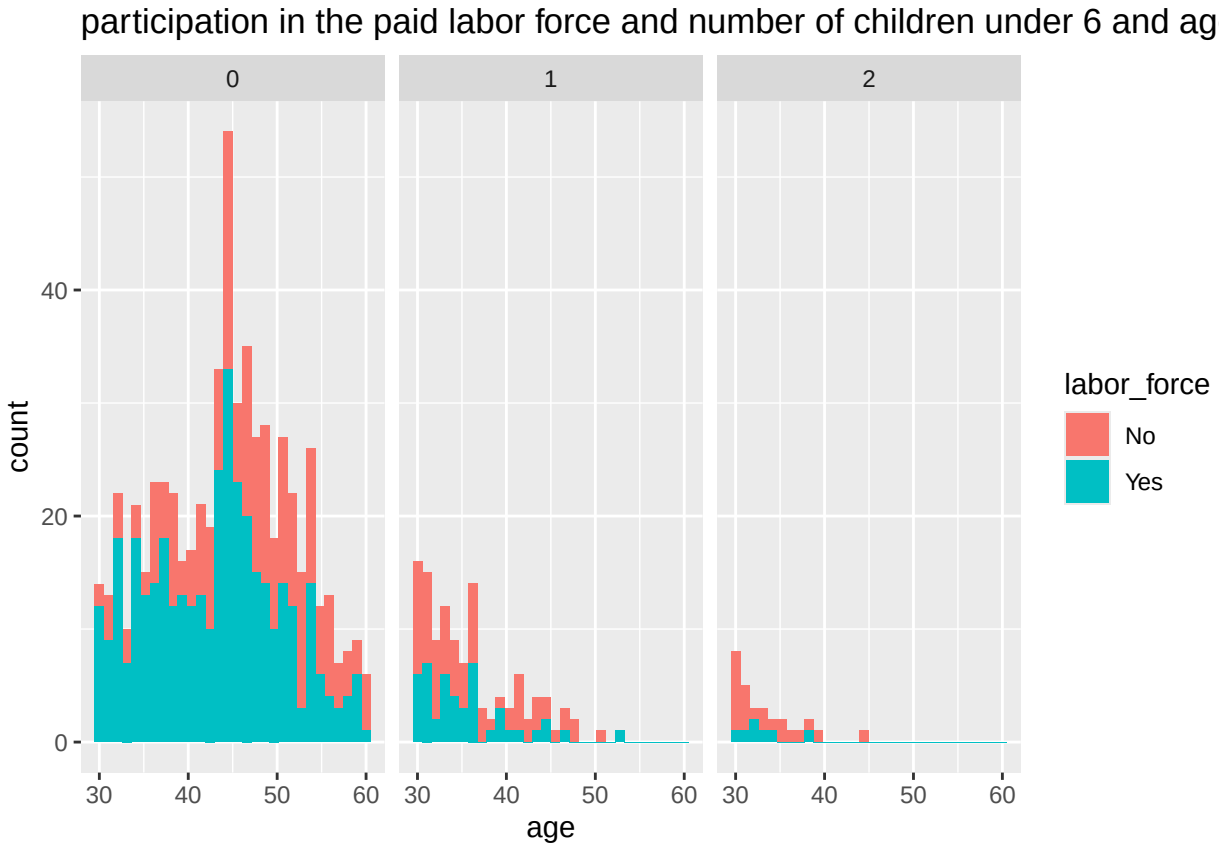
Revisions: Based on my analysis, there is a relationship between a woman's employment status and the number of children she has under the age of 6. Among those not employed, 71% have no children under 6, 22% have one child, and 6.8% have two. In comparison, 88% of employed women do not have children under 6, 11% have one, and 1.6% have two. These findings suggest that having fewer children under the age of 6 may increase a woman's likelihood of being employed.

Justification: My revisions clearly show the differences in employment status and the number of young children women have. By providing specific percentages, it highlights a pattern that having fewer kids may make it easier for women to find jobs.

8. (revised) Misunderstanding : I used geom_jitter, but histogram is better (please see further reasons in the justification).

Revisions:

```
ggplot(labor,aes(x=age,fill = labor_force))+
geom_histogram(bins=30)+
labs(title="participation in the paid labor force and number of children under 6 and age",
x="age",
fill="labor_force")+
facet_wrap(~kids_under6)
```



Justification: "Geom_jitter" focuses on individual observations and can reveal specific relationships, but a histogram offers a broader view of the data distribution, making it easier to interpret overall trends. More, histogram can be particularly useful for identifying how age or the number of children relates to labor force participation across larger groups rather than individual data points.

9. (revised)

Misunderstanding :I didn't have enough time so I copied my plot and the summary statistics.

Revisions: After including age in the plot, it reveals that younger women, whether employed or not, are more likely to have more children. As women grow older, fewer have children under the age of 6, and their involvement in the workforce also declines. This trend reflects a gradual reduction in workforce participation as age increases.

Justification: My revisions show that younger women tend to have more kids, while older women generally have fewer young children and lower workforce participation. By connecting these dots, they show the relationship between age, the number of children, and employment status clearly.

10.(a) (original)

Misunderstanding : No

(original answer)

```
#my original answer
n <- nrow(labor)
p_hat <- mean(labor$labor_force== "Yes")
q_hat <- 1- p_hat
np <- n * p_hat
nq <- n * q_hat
```

```r
#I used 'cat' to clearly show that it meets the conditions for the Central Limit Theorem (CLT).

cat("(1) np=",np,'\n')
```

```
## (1) np= 428
```

```r
cat("check if np>=10 ?",np >= 10,'\n')
```

```
## check if np>=10 ? TRUE
```

```r
cat("(2) n(1-p)=",nq,'\n')
```

```
## (2) n(1-p)= 325
```

```r
cat("check if n(1-p)>=10 ?",nq >= 10,'\n')
```

```
## check if n(1-p)>=10 ? TRUE
```

```r
cat("Therefore, both the necessary conditions for the Central Limit Theorem are
met.")
```

```
## Therefore, both the necessary conditions for the Central Limit Theorem are
## met.
```

Justification:

I clearly demonstrate that both conditions for the Central Limit Theorem are satisfied by calculating np and n (1-p) and checking if they are greater than or equal to 10. This approach provides the justification for dealing with the confidence interval calculation.

10. (b) (original)

Misunderstanding : No

(original answer)

```r
se <- sqrt(p_hat * q_hat / n)
z_value <- qnorm(0.975)
ci <- p_hat + c(-1, 1) * z_value * se
print(paste("95% Confidence Interval for the proportion of women in the workforce:",
ci[1], "to", ci[2]))
```

```
## [1] "95% Confidence Interval for the proportion of women in the workforce: 0.533016241718983 to 0.603
```

Justification: This means that if your results fall between 0.5330 and 0.6038, I am 95% confident that they are accurate.

11. (original)

```r
# my original answer
true_proportion <- 0.40
tf <- true_proportion >= confidence_interval[1] && true_proportion <= confidence_
```

```
## Error: object 'confidence_interval' not found
```

```r
cat("Is the 1970 proportion within the confidence interval? ",tf, "\n")
```

```
## Error: object 'tf' not found
```

```r
cat("In conclusion, this data is not trustworthy or reasonable.","\n")
```

```
## In conclusion, this data is not trustworthy or reasonable.
```

Justification: My answer is correct because the 40% proportion does not fall within the calculated confidence interval, which suggests that the claim is not supported by the data.

Reflection Questions:

1. I realize that I didn't fully address the comment questions (short essays) as effectively as I could have, mainly because I struggled with some of the English. But I handled the coding and graph questions pretty well. I really did my best, and I understand how essential it is to express my thoughts in my own words. I'm determined to keep working on my language skills so I can do better in future exams. This experience has really motivated me to keep growing and learning.
2. Rating E - Excellent