STAT 120 Midterm II

Bastola

May 25 2022

Name:

Point Distribution (Total: 100 points)

	Problem	Points	Score
1	a	5	
1	b	5	
1	c	5	
1	d	5	
1	e	5	
1	f	5	
1	g	5	
2	a	5	
2	b	5	
2	c	5	
3	a	5	
3	b	5	
3	c	5	
4	a	4	
4	b	5	
4	c	5	
4	d	5	
4	e	5	
5	a	5	
6	a	2	
6	b	2	
6	c	2	

- Calculators, writing utensils, and one-sided cheatsheet (max A4 size) allowed.
- Cheating is strictly prohibited.

Data: NHANES

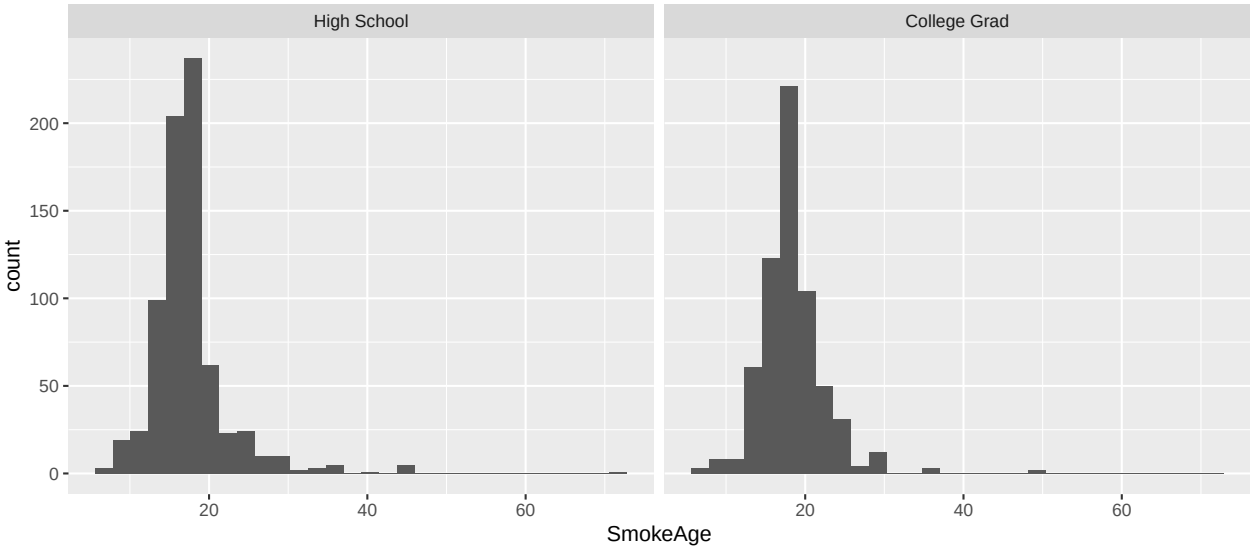
The US National Health and Nutrition Examination Survey (NHANES) periodically collects heath and nutrition data on a randomly selected sample of Americans. The data used on this exam is a compilation of data from surveys from 2009-2012. This data set will be used for Problems 1, 2, 4, and 5.

Problem 1: Smoking age by education

We subsetted the data to only include respondents 20 years or older who are current or former smokers and who have either completed only high school or completed college. We are interested in comparing the age that high school and college educated smokers first started to smoke (`NHANES.edu`).

```
table(NHANES.edu$Education)

 High School College Grad
         732          630
tapply(NHANES.edu$SmokeAge, NHANES.edu$Education, mean)
 High School College Grad
    17.52732     18.41746
tapply(NHANES.edu$SmokeAge, NHANES.edu$Education, sd)
 High School College Grad
    5.201080     4.193114
library(ggplot2)
ggplot(NHANES.edu, aes(x = SmokeAge)) + geom_histogram() + facet_wrap(~Education)
```




```
t.test(SmokeAge ~  Education, data=NHANES.edu)

    Welch Two Sample t-test

data:  SmokeAge by Education
t = -3.4951, df = 1354.3, p-value = 0.0004892
alternative hypothesis: true difference in means between group High School and group College Grad is no
95 percent confidence interval:
 -1.3897542 -0.3905216
sample estimates:
 mean in group High School mean in group College Grad
                  17.52732                   18.41746
```

(1a) State the hypotheses needed to address the following research question: In the US, is there a difference in the average age that people start smoking between those who only complete high school and those who complete college? Make sure to define the parameter(s) of interest.

(1b) Give the test statistic value given in the R output for this test and interpret this value in context.

(1c) Give the p-value given in the R output for this test and interpret this value in context.

(1d) State your conclusion for the test in context without using statistical jargon like "null" or "alternative". Use a 5% significance level for your test.

(1e) Interpret the 95% confidence interval given in the R output for this problem. Interpret the CI in context without using the word "difference" (or symbols) to explain the relationship between education level and initial smoking age.

(1f) Briefly explain if the assumptions needed to use t-test and CI inference methods are satisfied for this problem.

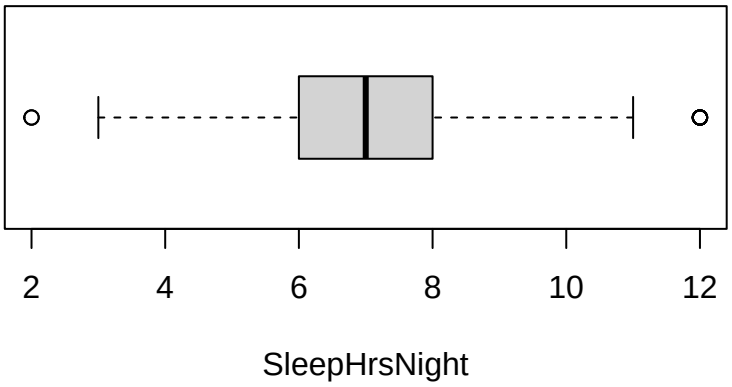
(1g) Results from a t-test were used in this problem to compare mean initial smoking ages for people in the two education levels. Alternatively, we could have used a randomization test to compute our p-value. Carefully explain how a randomization distribution could be formed to test your hypotheses in part (a).

Problem 2: Sleep hours at night

The `SleepHrsNight` variable records the number of hours slept at night. There are 2093 college grads in the data (`NHANES.am`) who have `SleepHrsNight` recorded.

```
summary(NHANES.am$SleepHrsNight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   6.000   7.000   7.038   8.000  12.000
sd(NHANES.am$SleepHrsNight)
[1] 1.128236
```

```
boxplot(NHANES.am$SleepHrsNight, horizontal = TRUE, xlab="SleepHrsNight")
```



(2a) A study claims that college grads in US sleep more than 7 hrs per night. Is the average sleep hours per night for US college grads more than 7 hours? State the hypotheses for this test. Define the parameter of interest.

```
t.test(NHANES.am$SleepHrsNight, mu=7, alt = "greater")
```

```
            One Sample t-test

data:  NHANES.am$SleepHrsNight
t = _____, df = 2092, p-value = 0.06066
alternative hypothesis: true mean is greater than 7
95 percent confidence interval:
 6.99764     Inf
sample estimates:
mean of x
 7.038223
```

(2b) Compute the missing standardized test statistic for the test in (2a) and interpret it in context.

(2c) Does your test statistic from (2b) provide evidence for the alternative or is it consistent with the null hypothesis? Briefly explain.

Problem 3: Sample sizes and point estimates.

(3a) Suppose you want to estimate the proportion of all Carleton students who spent money in downtown Northfield in the last month. How many students would you need to randomly sample to obtain a margin of error of 3% with a confidence level of 95%? You may use $z^* = 1.96$.

(3b) Suppose the margin of error in estimating the average number of hours spent in studying for Stat 120 class per week at Carleton is 0.3 hrs. The standard deviation in the number of hours spent in studying for Stat 120 per week based on a random sample is 2.2 hrs. How many students are in the sample? Use 95% confidence level and $z^* = 1.96$ instead of t^*.

(3c) Suppose the 95% confidence interval for the parameter of interest in (3b) is (14.7, 15.3). What is the point estimate for the parameter of interest?

Problem 4: Marriage and Sleep

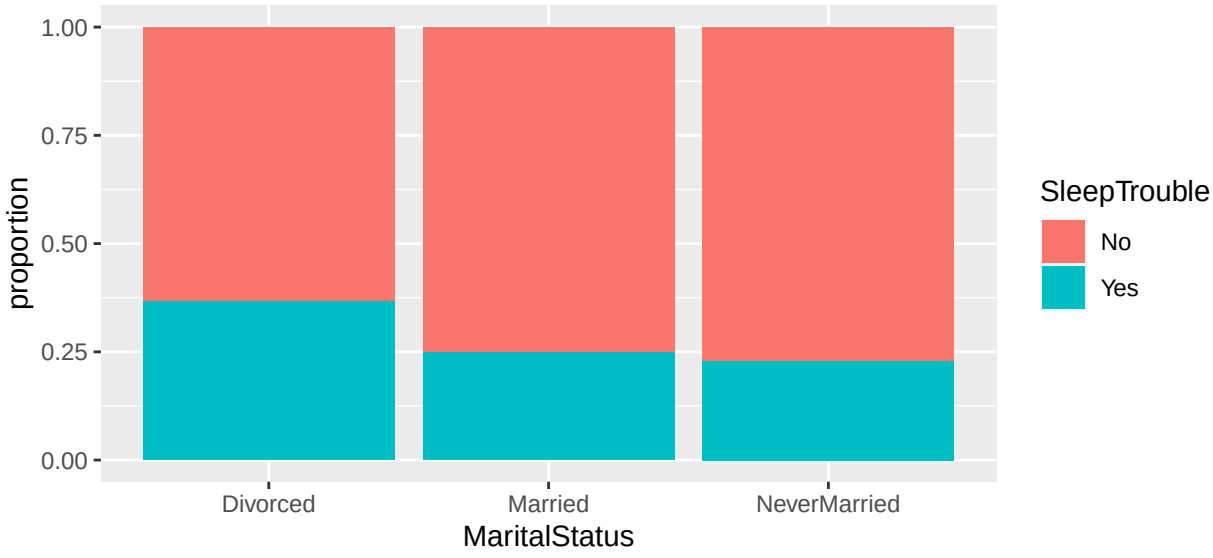
The variable `SleepTrouble` records whether a participant has told a doctor or other health professional that they had trouble sleeping. The variable `MaritalStatus` records whether a participant is `Married`, `Divorced` or `NeverMarried`. There are 6,032 participants who are married, divorced, or never married. Note that we are omitting any other type of participant for this problem (e.g. widows).

```
table(nhanes$MaritalStatus, nhanes$SleepTrouble)

                 No  Yes
  Divorced      448  259
  Married      2966  979
  NeverMarried 1064  316
table(nhanes$MaritalStatus)

    Divorced      Married NeverMarried
         707         3945         1380
table(nhanes$SleepTrouble)

  No  Yes
4478 1554
ggplot(nhanes, aes(x=MaritalStatus, fill = SleepTrouble)) +
  geom_bar(position = "fill") +
  labs(y = "proportion")
```



(4a) If there is no association between marital status and sleep troubles, how many responses in this sample would you expect to be "Never Married" with no sleep troubles?

(4b) Complete the following incomplete R-output of the Chi-square test of association between marital status and sleep trouble status of the respondents.

```
chisq.test(nhanes$MaritalStatus, nhanes$SleepTrouble)
```

```
      Pearson's Chi-squared test

data:  nhanes$MaritalStatus and nhanes$SleepTrouble
X-squared = ______, df = ______, p-value = 6.716e-12
```

(4c) Is there an association between marital status and sleep troubles? State your hypotheses and use a p-value to support your conclusion. Use a 5% significance level for your test.

(4d) What type of testing error (I or II) may you have made in your conclusion in part (4b). Briefly explain your answer.

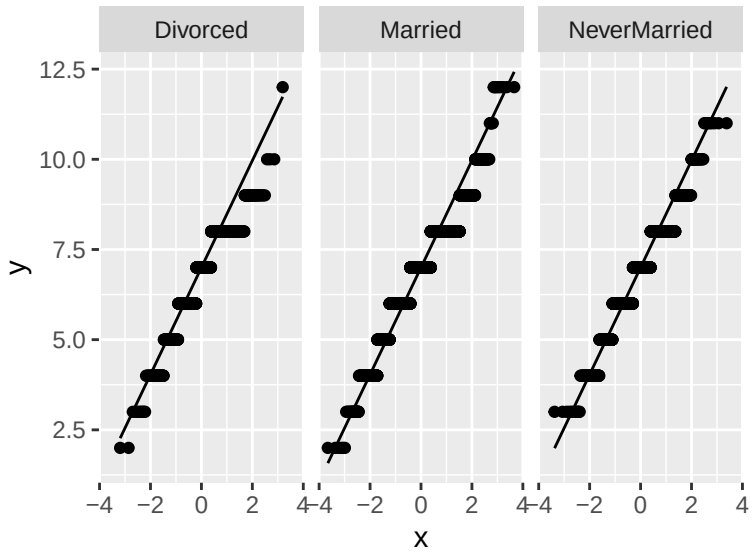
(4e) Set up the formula that shows how to compute a 95% confidence interval for the difference in the proportion divorced people with sleep troubles and the proportion of married people with sleep trouble. *Do not compute this interval, just show how it is computed using appropriate numbers from this data.*

Problem 5: The variable `SleepHrsNight` records the night time sleep hours. The variable `MaritalStatus` records whether a participant is `Married`, `Divorced` or `NeverMarried`.

(5a) Conduct a hypothesis test to see if the average night time sleep hours depends on the marital status. Please refer to the following R-outputs and remember to check if the assumptions for the test are satisfied.

```
marital_aov <- aov(SleepHrsNight~MaritalStatus, data = nhanes)
summary(marital_aov)
                Df Sum Sq Mean Sq F value   Pr(>F)
MaritalStatus    2     29  14.527    8.52 0.000202 ***
Residuals     6013  10253   1.705
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(nhanes, aes(sample=SleepHrsNight)) + geom_qq() + geom_qq_line() +
  facet_grid(~MaritalStatus)
```




```
nhanes %>% group_by(MaritalStatus) %>% summarize(mean = mean(SleepHrsNight),
                                                  sd = sd(SleepHrsNight),
                                                  n = length(SleepHrsNight))

# A tibble: 3 x 4
  MaritalStatus  mean    sd     n
  <fct>         <dbl> <dbl> <int>
1 Divorced       6.72  1.38   706
2 Married        6.94  1.27  3938
3 NeverMarried   6.88  1.37  1372
```

Problem 6: Multiple choice

(a) Suppose we want to compare average initial smoking age across four different education levels: less than high school, high school only, some college and college grad. Which hypothesis testing method is appropriate?

(i) Chi-square test for a two-way table

(ii) One-way ANOVA test

(iii) Randomization test for regression slope

(iv) t-test for two independent samples

(b) Suppose we want to compare sleep trouble status across four different education levels: less than high school, high school only, some college and college grad. Which hypothesis testing method is appropriate?

(i) Chi-square test for a two-way table

(ii) One-way ANOVA test

(iii) Randomization test for regression slope

(iv) t-test for two independent samples

(c) In the context of inference of the response for simple linear regression, which of the following statement is **NOT** correct?

(i) The prediction interval for a particular response is narrower than a confidence interval for a mean response

(ii) The prediction interval for a particular response is wider than a confidence interval for a mean response

(iii) The confidence interval for the mean response is narrower for the extreme predictor values

(iv) Both (i) and (iii)