R Assignment 6

Test for two means

optional name

2024-12-27

Intro/Goal

In this assignment you will carry out a hypothesis for the difference between two means. Specifically you will see if the average birthweight of a baby born from a smoking mother is different than the birthweight of a baby born from a non-smoking mother.

As with R Assignment 4 and 5 you will do an EDA, check the conditions and finally come to a conclusion. You will also have to consider which testing error is possible. Along the way I will introduce you to a new function called t.test() which will make computing the hypothesis test easier.

The data is from ncbirths. There are many variables in this data, but you will be studying two: weight and habit. In the code chunk below I've removed all other variables as well as the mothers who did not respond to the question of whether they smoked or not.

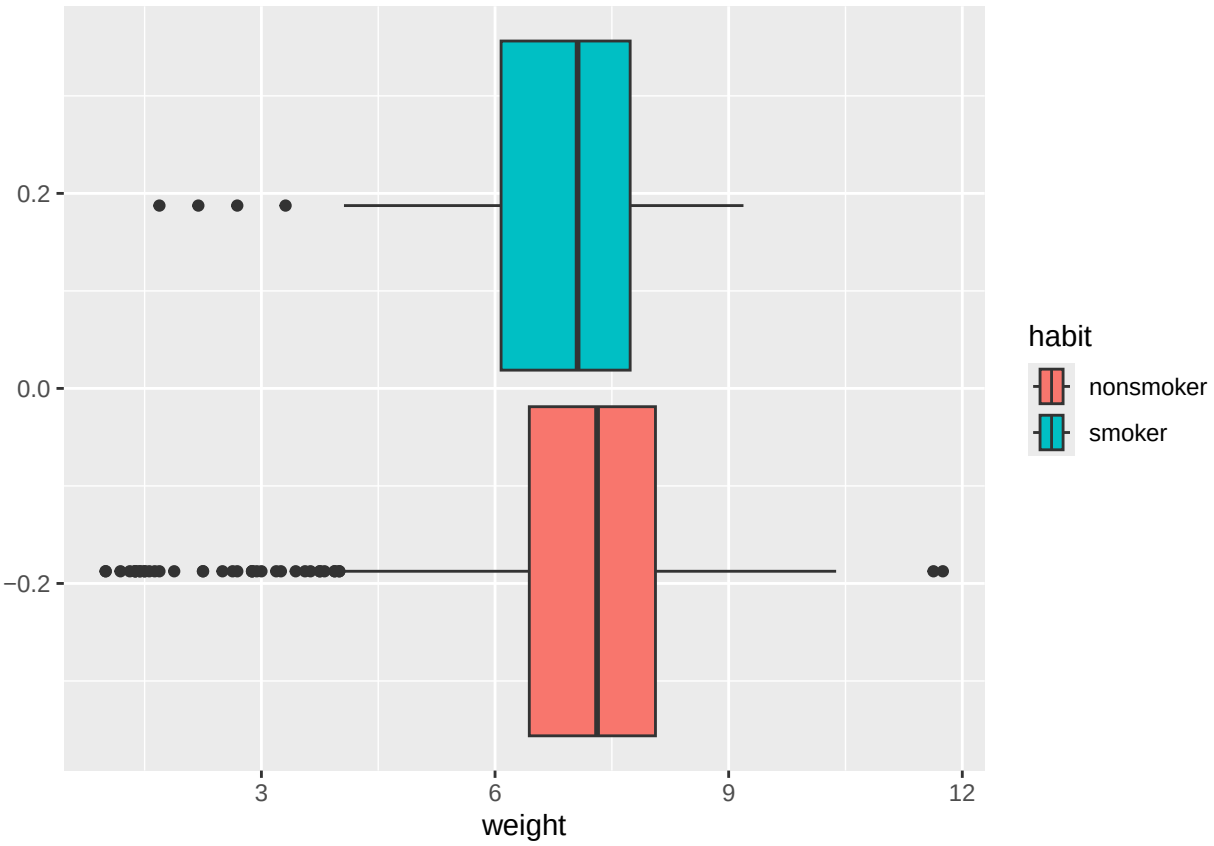
```
ncbirths = ncbirths |>
  filter(habit %in% c("nonsmoker","smoker"))|>
  select(weight,habit)
ncbirths
```

```
## # A tibble: 999 x 2
##    weight habit
##     <dbl> <fct>
##  1   7.63 nonsmoker
##  2   7.88 nonsmoker
##  3   6.63 nonsmoker
##  4   8    nonsmoker
##  5   6.38 nonsmoker
##  6   5.38 nonsmoker
##  7   8.44 nonsmoker
##  8   4.69 nonsmoker
##  9   8.81 nonsmoker
## 10   6.94 nonsmoker
## # i 989 more rows
```

Do an exploritory analysis.

At minimum you should make a boxplot to check for extreme outliers. You should plot weight, but fill by habit. (see R Assignment 1 or my example video on how to do this) You might want to check for other information that might be necessary for conditions.

```
ggplot(data = ncbirths)+
  geom_boxplot(aes(x=weight, fill= habit))
```



Check conditions

Check the conditions of the test. You can find these conditions in chapter 20. Do the test even if the conditions are not met.

Answer: There are several outliers, but our sample size is large so we can relax that condition. There is independence between groups because mothers cannot be smoking an non-smoking. There is likely independence within the group because the data is a random subset of all births from 2004. Independence would be broken for the case of multiple births from the same mother, which is unlikely due to randomness.

Write out the hypothesis notation. Include a significance level.

Edit the values below to match this test. I've done the null hypothesis for you. You need to edit the alternative hypothesis.

$$H_o : \mu_{smoker} = \mu_{non\_smoker} H_a : \mu_{smoker} \neq \mu_{non\_smoker} \alpha = 0.05$$

Do the test and state your p-value.

If you'd like to do a difference of means test like you did in the homework that is fine, you can disregard the t.test() function below. Otherwise use the t.test() function.

Use t.test() to have R calculate the p-value for you. State your p-value. note: t.test() needs four arguments for this test.

- formula of the form = numeric variable ~ categorical variable
- data = ncbirths
- alternative = the alternative hypothesis (use "less", "greater","two.sided").

- correct = FALSE (I've done this for you)

```
# This line of code turns off scientific notation, you don't need to do anything with it.
options(scipen = 100)

t.test(formula = weight~habit , data = ncbirths , alternative = "two.sided", correct = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by habit
## t = 2.359, df = 171.32, p-value = 0.01945
## alternative hypothesis: true difference in means between group nonsmoker and group smoker is not equa
## 95 percent confidence interval:
##  0.05151165 0.57957328
## sample estimates:
## mean in group nonsmoker    mean in group smoker
##                7.144273                6.828730
```

Conclusion

Write a conclusion be sure to state your pvalue and state what it means in terms of the weights of babies from smoking and non smoking mothers. Make sure to discuss whether the conditions were met and any issues that might cause your conclusion to be incorrect.

Answer: There is a statistically significant difference and we reject the null hypothesis. The average birth weights are different. However From this dataset is seems there is not much practical difference between the mean birth weight of babies born from smoking or non smoking mothers, just 0.3 pounds.

Discuss which testing error (I or II) may be possible.

Answer: We rejected our null hypothesis. If there were really no difference in weights we would have made a type 1 error.

Note: this does not mean your test is invalid. We are dealing with probabilities so it is always possible for for our conclusion to be incorrect.