PSTAT 5LS Lab 8

TA NAME HERE

Week of November 25, 2024

Section 1

Announcements

Upcoming Deadlines

- HW 7 due by 11:59pm on **Wednesday, November 27**
- HW 8 due by 11:59pm on Friday, December 6
- ECoach Journal Entry 9 due by 11:59 on Monday, November 25
- no ECoach Journal Entry 9 due on Monday, December 2
- ECoach Journal Entry 10 due by 11:59 on Monday, December 9

Section 2

Learning Objectives

R Learning Objectives

1. Create a variable consisting of the differences in corresponding observations.
2. Interpret R output providing confidence intervals and hypothesis tests for inference on the mean of a population of differences.

Statistical Learning Objectives

1. Continue discussing quantitative data, this week in regards to a mean of differences (paired data).
2. Understand confidence intervals and hypothesis test for the mean of differences when we have paired data.

Functions Covered in this Lab

1. `pt()`
2. `qt()`
3. `t.test()`

Section 3

Lab Tutorial

Mean of the Differences for Paired Data

When we are working with paired data, the parameter is the *mean of the differences*

$$\mu_d$$

The point estimate (sample statistic) is

$$\bar{x}_d$$

where the d's represent the difference in observations.

Mean of the Differences for Paired Data

The conditions are

1. **Independence**: The observed differences should be independent of one another. This condition is typically satisfied with a random sample of the differences, but you might need to think about whether the differences are independent from observation to observation.
2. **Normality**: The observed differences in observations should come from a (nearly-)normal population of differences. We will check by making a histogram of the differences in the sample. If the sample looks like it came from a population that is normally distributed, then this condition is met. Note that we can relax the condition more and more for larger and larger sample sizes.

Births from North Carolina in 2004

The `penguins` data set we have used doesn't contain anything that is paired data. So we turn our attention to a data set called `births.csv`.

This data set was collected by taking a random sample of 800 entries in a database containing information about babies born to cisgender heterosexual couples in North Carolina in 2004. Variables include the father's age, the mother's age, the weight of the baby, etc.

Let's say we want to estimate the average difference in the age of the father and the mother of babies born in North Carolina in 2004. Do you think that the father's age and the mother's age in this data set are independent of each other, or, do you think that the father's age and the mother's age in this data set are dependent on each other?

Paired Data

We should **pair** the data for father's age and mother's age together, as both ages were collected from the same observation (here, same baby). So we have one sample from one population of babies born in North Carolina in 2004.

Thus our parameter of interest is μ_{d}, which represents the population mean of the difference in the age of the father and the mother of a baby born in North Carolina in 2004.

Normality Condition

Recall that our normality condition requires that the observations of the difference in age need to come from a (nearly-)normal population.

So we will need to create a histogram of the **differences** in the parents' ages from the sample, and see if the distribution is approximately bell-shaped.

Creating a Variable of the Differences

In order to make this histogram, we need to create a new variable that is defined as the difference in the father's age and mother's age for the babies born in North Carolina in 2004.

First, let's read in the data. Run the `tryIt1` code chunk in your notes file.

```
babies <- read.csv("births.csv", stringsAsFactors = TRUE)
```

Creating a Variable of the Differences

Let's look at structure of the `babies` data set so we can figure out which variables indicate the father's age and the mother's age.

Write the code you need to peek at the structure of the data set in the `tryIt2` code chunk and then run the code chunk.

```
str(babies)
```

What are the variables that indicate the father's age and the mother's age?

Creating a Variable of the Differences

Next, let's add a variable called `ageDiff` to this `babies` data set. We will define the difference as father's age (`fage`) minus mother's age (`mage`).

Create this new variable in the `tryIt3` code chunk in your notes file.

```
babies$ageDiff <- babies$fage - babies$mage
```

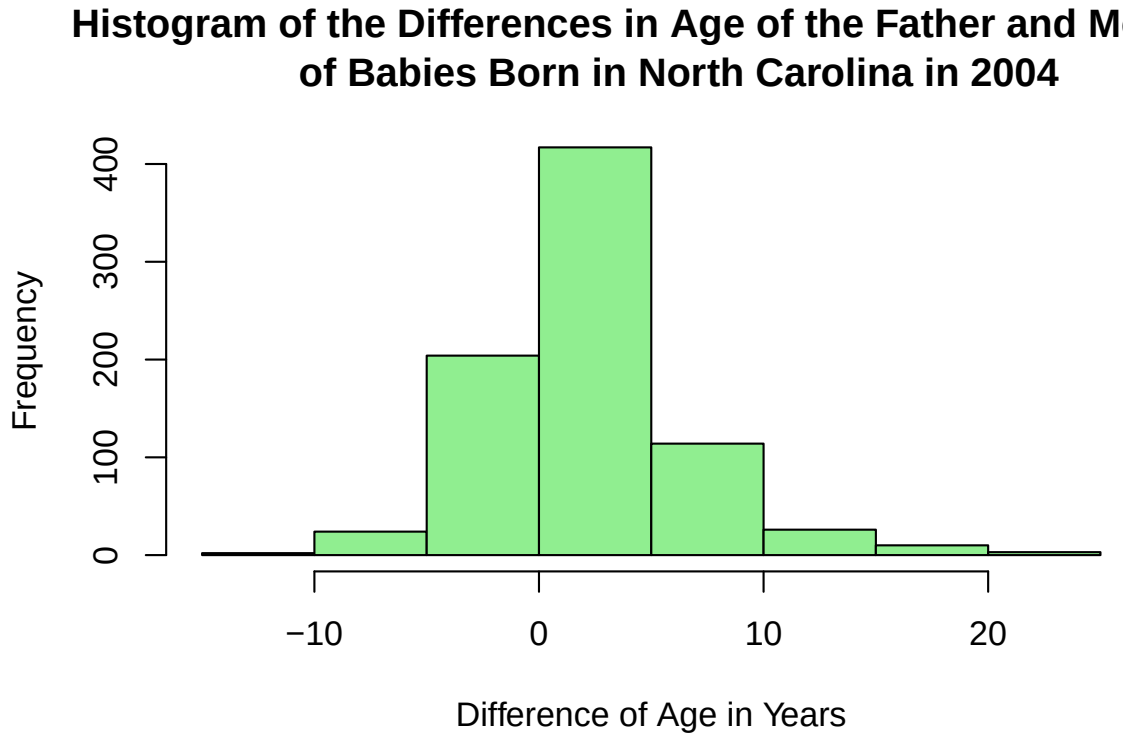
Creating a Histogram of the Differences

We can now create the histogram of the differences in age between the father and the mother for the sampled babies born in North Carolina in 2004.

The title is really long, so we will go to the next line where we want a line break. Specify the variable name in the `tryIt4` code chunk in your notes and then run it to create a histogram.

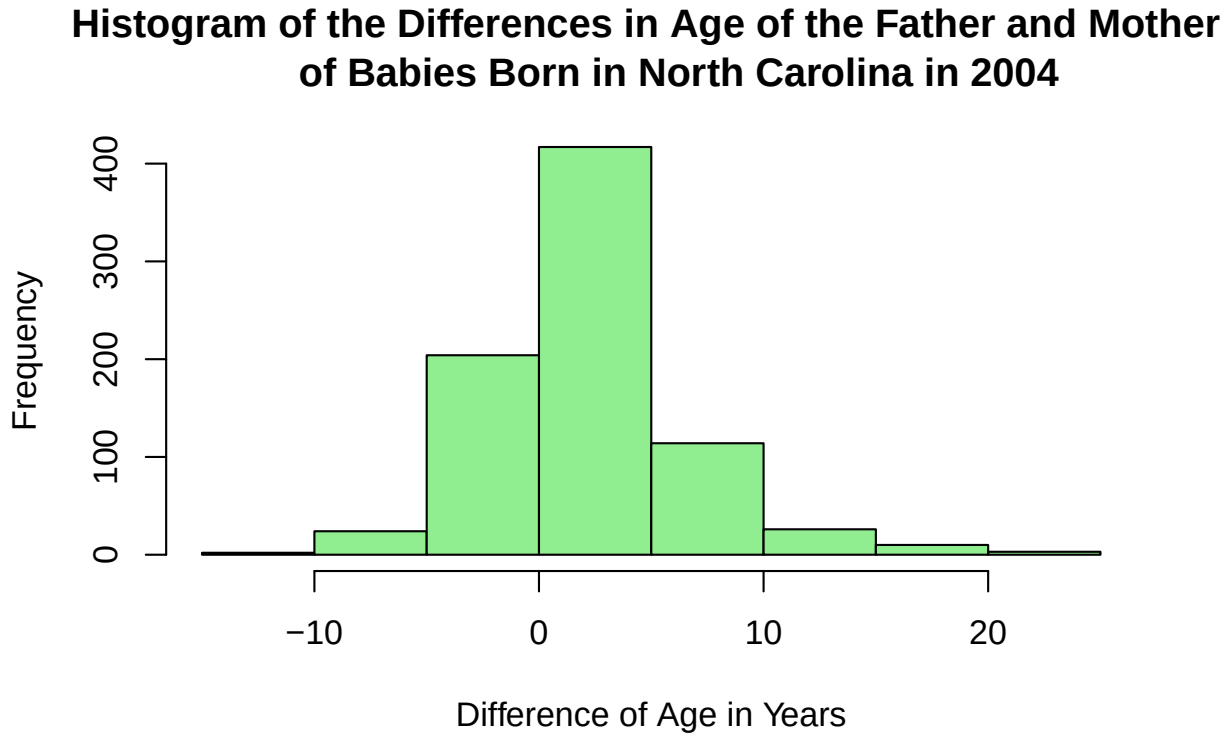
```
hist(babies$ageDiff,
     main = "Histogram of the Differences in Age of the Father and Mother
     of Babies Born in North Carolina in 2004",
     xlab = "Difference of Age in Years",
     col = "lightgreen")
```

Histogram of the Differences



Does the normality condition appear to be satisfied?

Histogram of the Differences



This histogram is unimodal and slightly right skewed. However, the sample size (800) is definitely large enough to relax the normality condition.

(Remember that the reason we can relax the normality condition is that the sample mean of the differences $\bar{x}_d$ will be approximately normal.)

Estimating the Mean of the Differences

We want to estimate the mean of the differences in age between fathers and mothers for all babies born in North Carolina in 2004. This requires creating a confidence interval. Let's create a 90% confidence interval. Try this out in the `tryIt5` code chunk in your notes file.

```
t.test(babies$ageDiff, conf.level = 0.90)
```

Confidence Interval R Output

```
t.test(babies$ageDiff, conf.level = 0.90)
```

```
##
##  One Sample t-test
##
## data:  babies$ageDiff
## t = 17.265, df = 799, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  2.380281 2.882219
## sample estimates:
## mean of x
##   2.63125
```

Interpreting the Confidence Interval

Interpret the confidence interval.

Interpreting the Confidence Interval

The 90% confidence interval suggests that, on average, fathers are between 2.38 and 2.88 years older than mothers for all babies born in North Carolina in 2004.

Section 4

Questions

What Questions Do You Have?