Inference for numerical data

Complete all **Exercises**, and submit answers to **Questions** on the Coursera platform.

Getting Started

Load packages

In this lab we will explore the data using the `dplyr` package and visualize it using the `ggplot2` package for data visualization. The data can be found in the companion package for this course, `statsr`.

Let's load the packages.

```
library(statsr)
library(dplyr)
library(ggplot2)
```

The data

In 2004, the state of North Carolina released a large data set containing information on births recorded in this state. This data set is useful to researchers studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of observations from this data set.

Load the `nc` data set into our workspace.

```
data(nc)
```

We have observations on 13 different variables, some categorical and some numerical. The meaning of each variable is as follows.

variable	description
`fage`	father's age in years.
`mage`	mother's age in years.
`mature`	maturity status of mother.
`weeks`	length of pregnancy in weeks.
`premie`	whether the birth was classified as premature (premie) or full-term.
`visits`	number of hospital visits during pregnancy.
`marital`	whether mother is `married` or `not married` at birth.
`gained`	weight gained by mother during pregnancy in pounds.
`weight`	weight of the baby at birth in pounds.

variable	description
`lowbirthweight`	whether baby was classified as low birthweight (`low`) or not (`not low`).
`gender`	gender of the baby, `female` or `male` .
`habit`	status of the mother as a `nonsmoker` or a `smoker` .
`whitemom`	whether mom is `white` or `not white` .

1. There are 1,000 cases in this data set, what do the cases represent?
    1. The hospitals where the births took place
    2. The fathers of the children
    3. The days of the births
    4. The births

As a first step in the analysis, we should take a look at the variables in the dataset. This can be done using the `str` command:

```
str(nc)
```

```
## tibble [1,000 × 13] (S3: tbl_df/tbl/data.frame)
##  $ fage          : int [1:1000] NA NA 19 21 NA NA 18 17 NA 20 ...
##  $ mage          : int [1:1000] 13 14 15 15 15 15 15 15 16 16 ...
##  $ mature        : Factor w/ 2 levels "mature mom","younger mom": 2 2 2 2 2 2 2 2 2 2
...
##  $ weeks         : int [1:1000] 39 42 37 41 39 38 37 35 38 37 ...
##  $ premie        : Factor w/ 2 levels "full term","premie": 1 1 1 1 1 1 1 2 1 1 ...
##  $ visits        : int [1:1000] 10 15 11 6 9 19 12 5 9 13 ...
##  $ marital       : Factor w/ 2 levels "married","not married": 1 1 1 1 1 1 1 1 1 1
...
##  $ gained        : int [1:1000] 38 20 38 34 27 22 76 15 NA 52 ...
##  $ weight        : num [1:1000] 7.63 7.88 6.63 8 6.38 5.38 8.44 4.69 8.81 6.94 ...
##  $ lowbirthweight: Factor w/ 2 levels "low","not low": 2 2 2 2 2 1 2 1 2 2 ...
##  $ gender        : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 2 2 2 1 ...
##  $ habit         : Factor w/ 2 levels "nonsmoker","smoker": 1 1 1 1 1 1 1 1 1 1 ...
##  $ whitemom      : Factor w/ 2 levels "not white","white": 1 1 2 2 1 1 1 1 2 2 ...
```

As you review the variable summaries, consider which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren't sure or want to take a closer look at the data, make a graph.

Exploratory data analysis

We will first start with analyzing the weight gained by mothers throughout the pregnancy: `gained` .

Using visualization and summary statistics, describe the distribution of weight gained by mothers during pregnancy. The `summary` function can also be useful.

```
summary(nc$gained)
```

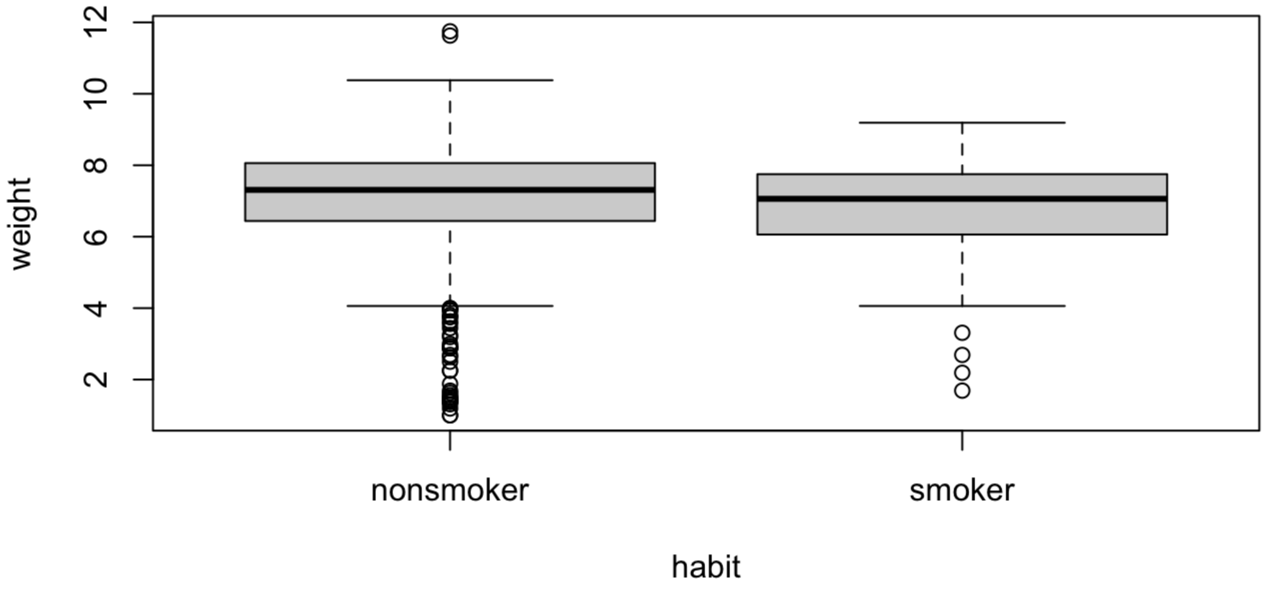
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.00   20.00   30.00   30.33   38.00   85.00      27
```

2. How many mothers are we missing weight gain data from?
   1. 0
   2. 13
   3. 27
   4. 31

Next, consider the possible relationship between a mother's smoking habit and the weight of her baby. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

3. Make side-by-side boxplots of `habit` and `weight` . Which of the following is false about the relationship between habit and weight?
   1. Median birth weight of babies born to non-smoker mothers is slightly higher than that of babies born to smoker mothers.
   2. Range of birth weights of babies born to non-smoker mothers is greater than that of babies born to smoker mothers.
      **
   3. Both distributions are extremely right skewed.
      **
   4. The IQRs of the distributions are roughly equal.

```
boxplot(weight~habit,data=nc,xlab="habit",ylab="weight")
```



https://www.datamentor.io/r-programming/box-plot (https://www.datamentor.io/r-programming/box-plot)

The box plots show how the medians of the two distributions compare, but we can also compare the means of the distributions using the following to first group the data by the `habit` variable, and then calculate the mean `weight` in these groups using the `mean` function.

```
nc %>%
  group_by(habit) %>%
  summarise(mean_weight = mean(weight))
```

```
## # A tibble: 3 × 2
##   habit     mean_weight
##   <fct>           <dbl>
## 1 nonsmoker        7.14
## 2 smoker           6.83
## 3 <NA>             3.63
```

```
# we can compare two groups using by() function with the summary function
by(nc$weight,nc$habit,summary)
```

```
## nc$habit: nonsmoker
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   6.440   7.310   7.144   8.060  11.750 
## ------------------------------------------------------------ 
## nc$habit: smoker
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.690   6.077   7.060   6.829   7.735   9.190
```

```
# we can compare two groups using by() function with a specific metric too
by(nc$weight,nc$habit,sd)
```

```
## nc$habit: nonsmoker
## [1] 1.518681
## ------------------------------------------------------------ 
## nc$habit: smoker
## [1] 1.38618
```

There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test.

Inference

Exercise: Are all conditions necessary for inference satisfied? Comment on each. You can compute the group sizes using the same `by` command above but replacing `mean(weight)` with `n()`.

4. What are the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different?
    1. $H_0 : \mu_{\text{smoking}} = \mu_{\text{non-smoking}}; H_A : \mu_{\text{smoking}} > \mu_{\text{non-smoking}}$

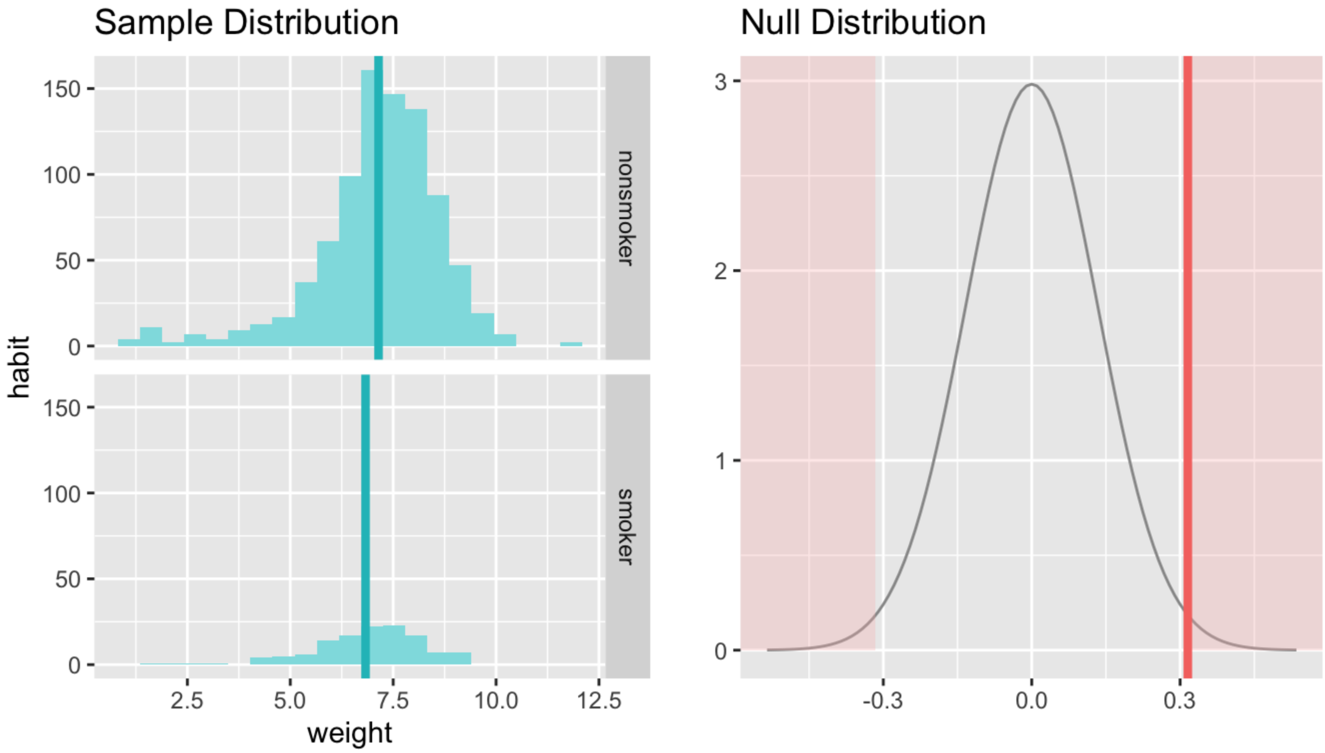
2. $H_0 : \mu_{\text{smoking}} = \mu_{\text{non-smoking}}; H_A : \mu_{\text{smoking}} \neq \mu_{\text{non-smoking}}$
3. $H_0 : \bar{x}_{\text{smoking}} = \bar{x}_{\text{non-smoking}}; H_A : \bar{x}_{\text{smoking}} > \bar{x}_{\text{non-smoking}}$
4. $H_0 : \bar{x}_{\text{smoking}} = \bar{x}_{\text{non-smoking}}; H_A : \bar{x}_{\text{smoking}} > \bar{x}_{\text{non-smoking}}$
5. $H_0 : \mu_{\text{smoking}} \neq \mu_{\text{non-smoking}}; H_A : \mu_{\text{smoking}} = \mu_{\text{non-smoking}}$

Next, we introduce a new function, `inference`, that we will use for conducting hypothesis tests and constructing confidence intervals.

Then, run the following:

```
inference(y = weight, x = habit, data = nc, statistic = "mean", type = "ht", null = 0,
          alternative = "twosided", method = "theoretical")
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_nonsmoker = 873, y_bar_nonsmoker = 7.1443, s_nonsmoker = 1.5187
## n_smoker = 126, y_bar_smoker = 6.8287, s_smoker = 1.3862
## H0: mu_nonsmoker =  mu_smoker
## HA: mu_nonsmoker != mu_smoker
## t = 2.359, df = 125
## p_value = 0.0199
```


Sample Distribution

Null Distribution

Let’s pause for a moment to go through the arguments of this custom function.

1. The first argument is `y`, which is the response variable that we are interested in: `weight`.
2. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and non-smokers: `habit`.
3. The third argument, `data`, is the data frame these variables are stored in.
4. Next is `statistic`, which is the sample statistic we’re using, or similarly, the population parameter we’re estimating. In future labs we can also work with “median” and “proportion”.

5. Next we decide on the `type` of inference we want: a hypothesis test (`"ht"`) or a confidence interval (`"ci"`).
6. When performing a hypothesis test, we also need to supply the `null` value, which in this case is `0` , since the null hypothesis sets the two population means equal to each other.
7. The `alternative` hypothesis can be `"less"` , `"greater"` , or `"twosided"` .
8. Lastly, the `method` of inference can be `"theoretical"` or `"simulation"` based.

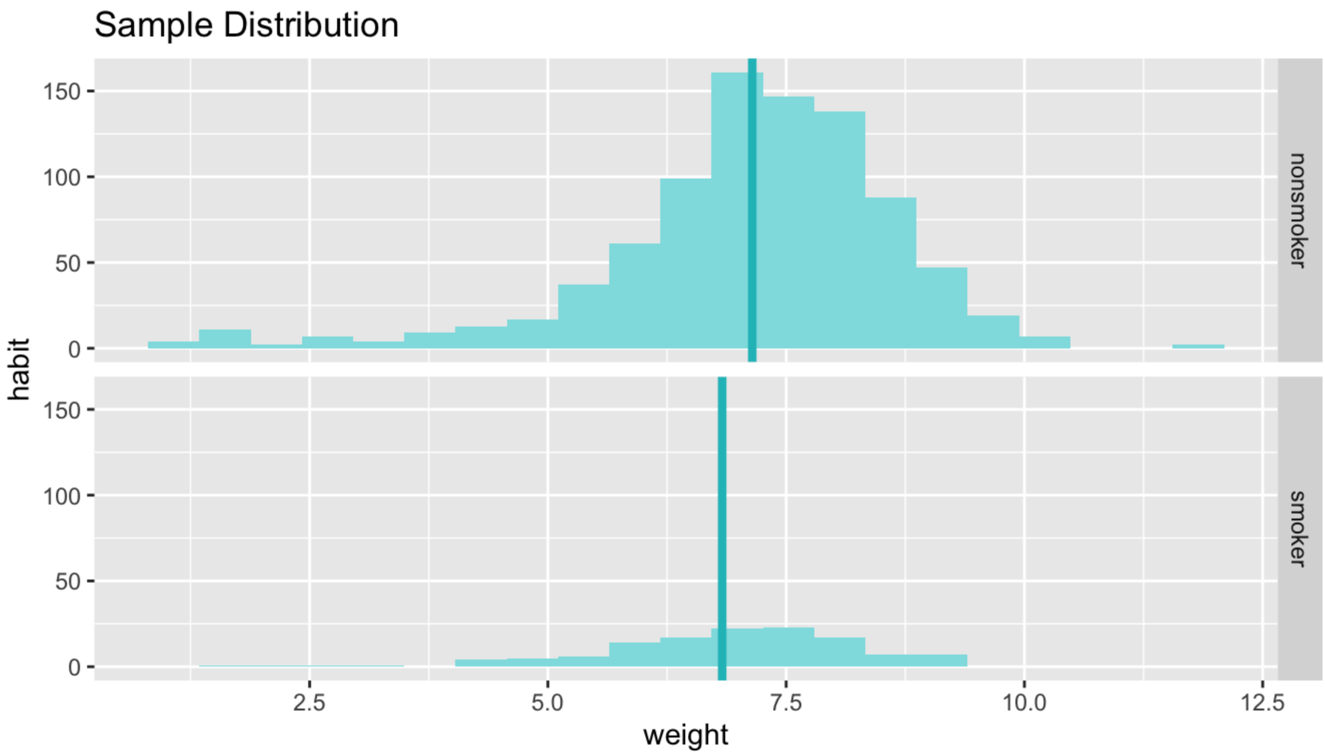
For more information on the inference function see the help file with `?inference` .

Exercise: What is the conclusion of the hypothesis test? Reject the null hypothesis because p_value is less than 0.05

5. Change the `type` argument to `"ci"` to construct and record a confidence interval for the difference between the weights of babies born to nonsmoking and smoking mothers, and interpret this interval in context of the data. Note that by default you'll get a 95% confidence interval. If you want to change the confidence level, add a new argument (`conf_level`) which takes on a value between 0 and 1. Also note that when doing a confidence interval arguments like `null` and `alternative` are not useful, so make sure to remove them.
    1. We are 95% confident that babies born to nonsmoker mothers are on average 0.05 to 0.58 pounds lighter at birth than babies born to smoker mothers.
    2. We are 95% confident that the difference in average weights of babies whose moms are smokers and nonsmokers is between 0.05 to 0.58 pounds.
    3. We are 95% confident that the difference in average weights of babies in this sample whose moms are smokers and nonsmokers is between 0.05 to 0.58 pounds.
    4. We are 95% confident that babies born to nonsmoker mothers are on average 0.05 to 0.58 pounds heavier at birth than babies born to smoker mothers.

```
inference(y = weight, x = habit, data = nc, statistic = "mean", type = "ci", method = "t
        heoretical")
```

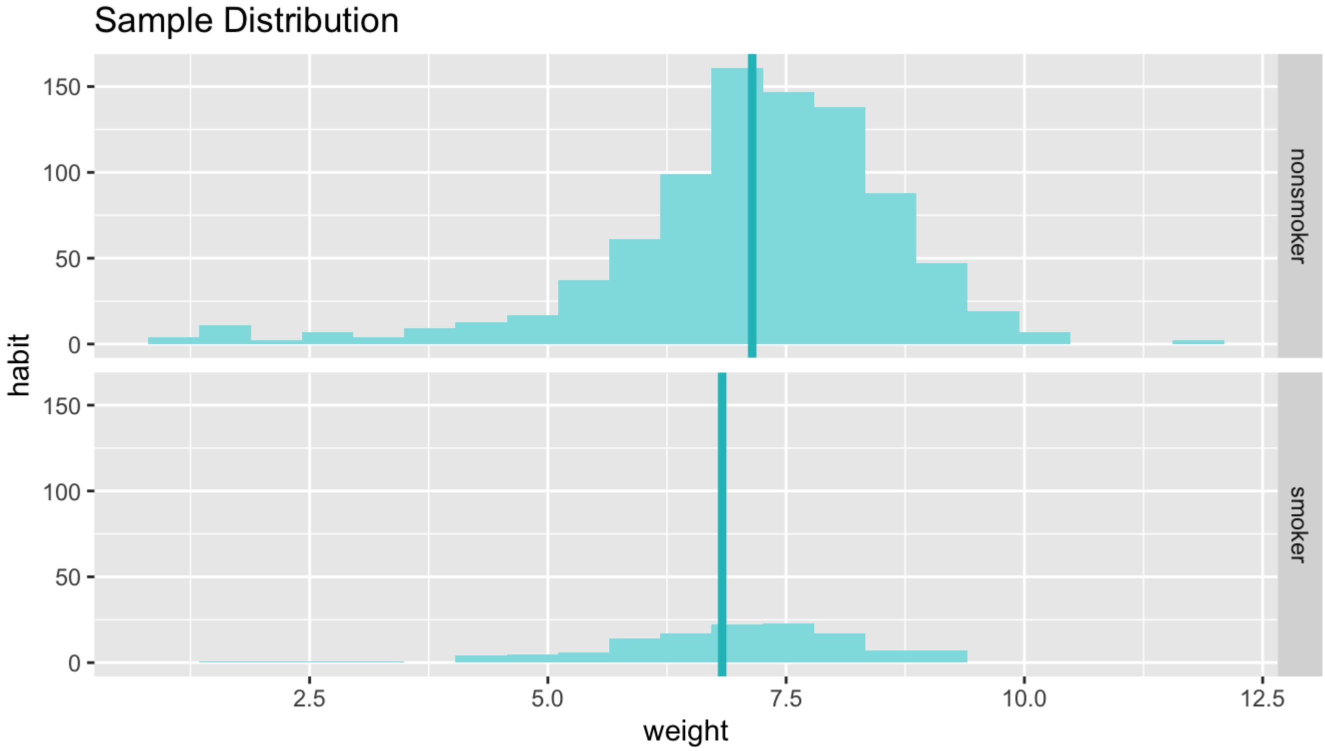
```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_nonsmoker = 873, y_bar_nonsmoker = 7.1443, s_nonsmoker = 1.5187
## n_smoker = 126, y_bar_smoker = 6.8287, s_smoker = 1.3862
## 95% CI (nonsmoker - smoker): (0.0508 , 0.5803)
```



By default the function reports an interval for ($\mu_{\text{nonsmoker}} - \mu_{\text{smoker}}$) . We can easily change this order by using the `order` argument:

```
inference(y = weight, x = habit, data = nc, statistic = "mean", type = "ci",
          method = "theoretical", order = c("smoker","nonsmoker"))
```

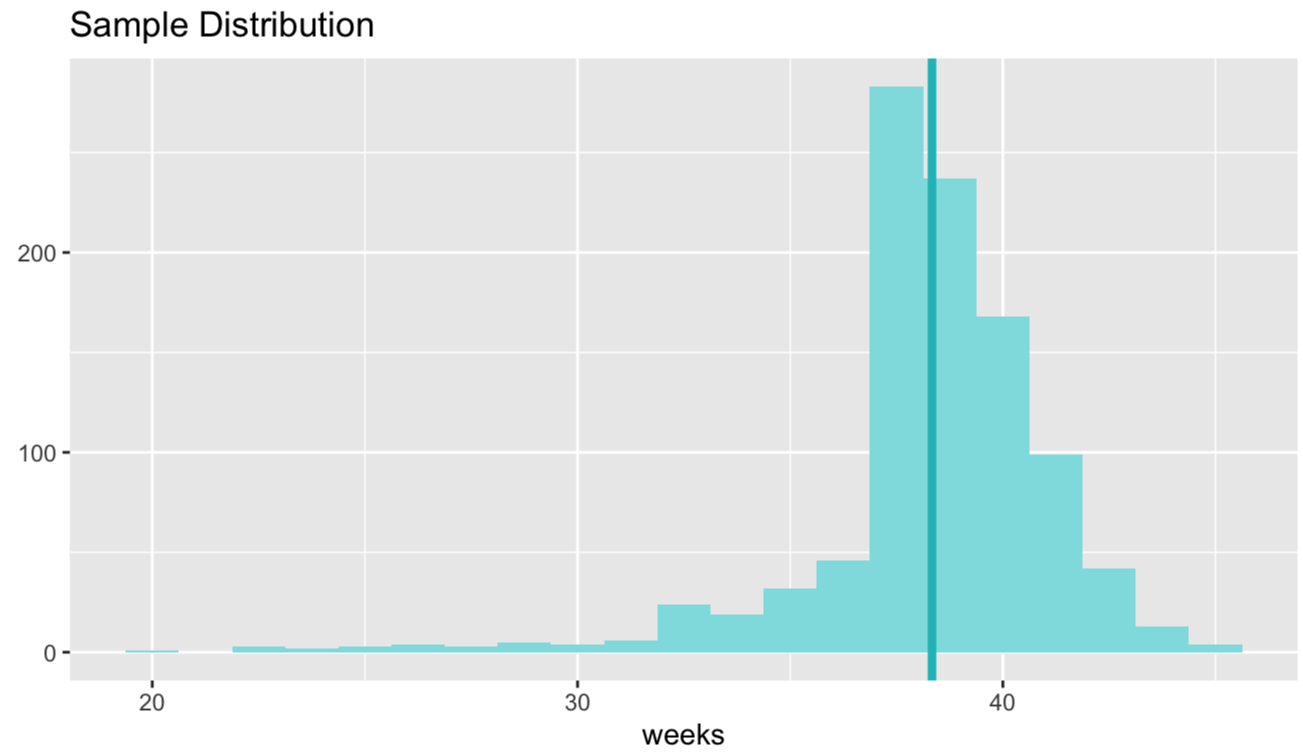
```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_smoker = 126, y_bar_smoker = 6.8287, s_smoker = 1.3862
## n_nonsmoker = 873, y_bar_nonsmoker = 7.1443, s_nonsmoker = 1.5187
## 95% CI (smoker - nonsmoker): (-0.5803 , -0.0508)
```



6. Calculate a 99% confidence interval for the average length of pregnancies (`weeks`). Note that since you're doing inference on a single population parameter, there is no explanatory variable, so you can omit the `x` variable from the function. Which of the following is the correct interpretation of this interval?
    1. (38.1526 , 38.5168)
    2. (38.0892 , 38.5661)
    3. (6.9779 , 7.2241)
       **
    4. (38.0952 , 38.5742)
       **

```
inference(y = weeks, data = nc, statistic = "mean", conf_level=0.99,type = "ci",
          method = "theoretical", order = c("smoker","nonsmoker"))
```

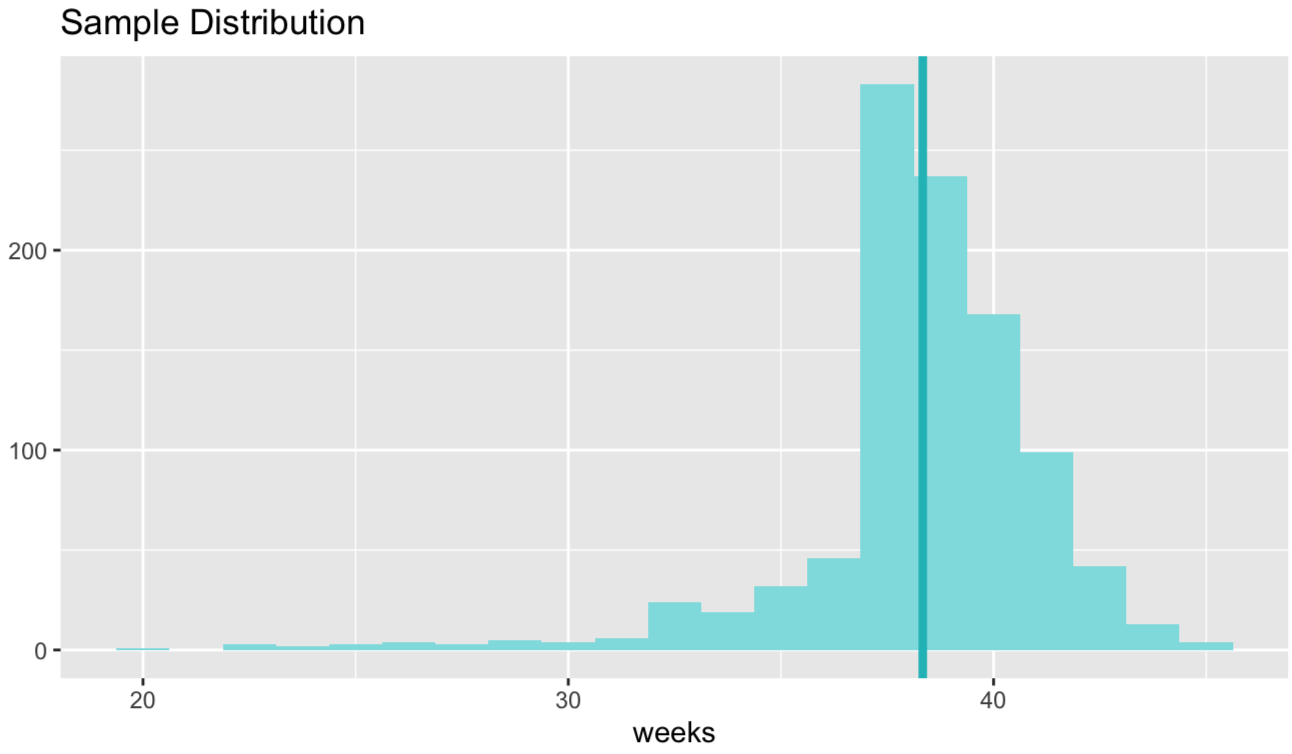
```
## Single numerical variable
## n = 998, y-bar = 38.3347, s = 2.9316
## 99% CI: (38.0952 , 38.5742)
```



Exercise: Calculate a new confidence interval for the same parameter at the 90% confidence level. Comment on the width of this interval versus the one obtained in the the previous exercise.

```
inference(y = weeks, data = nc, statistic = "mean", conf_level=0.90,type = "ci",
          method = "theoretical", order = c("smoker","nonsmoker"))
```

```
## Single numerical variable
## n = 998, y-bar = 38.3347, s = 2.9316
## 90% CI: (38.1819 , 38.4874)
```

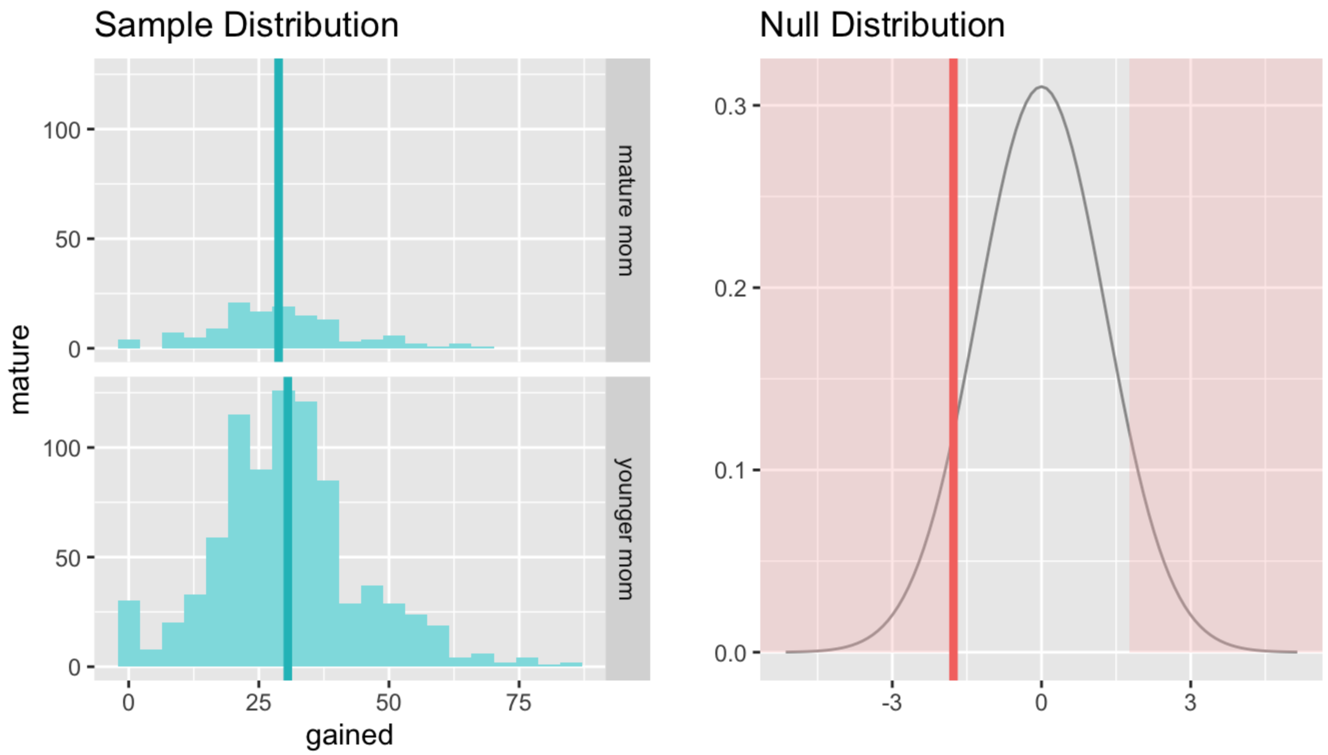


```
#narrower interval
```

Exercise: Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

```
inference(x = mature, y=gained, data = nc, statistic = "mean", type = "ht", null = 0,
          alternative = "twosided", method = "theoretical")
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_mature mom = 129, y_bar_mature mom = 28.7907, s_mature mom = 13.4824
## n_younger mom = 844, y_bar_younger mom = 30.5604, s_younger mom = 14.3469
## H0: mu_mature mom =  mu_younger mom
## HA: mu_mature mom != mu_younger mom
## t = -1.3765, df = 128
## p_value = 0.1711
```

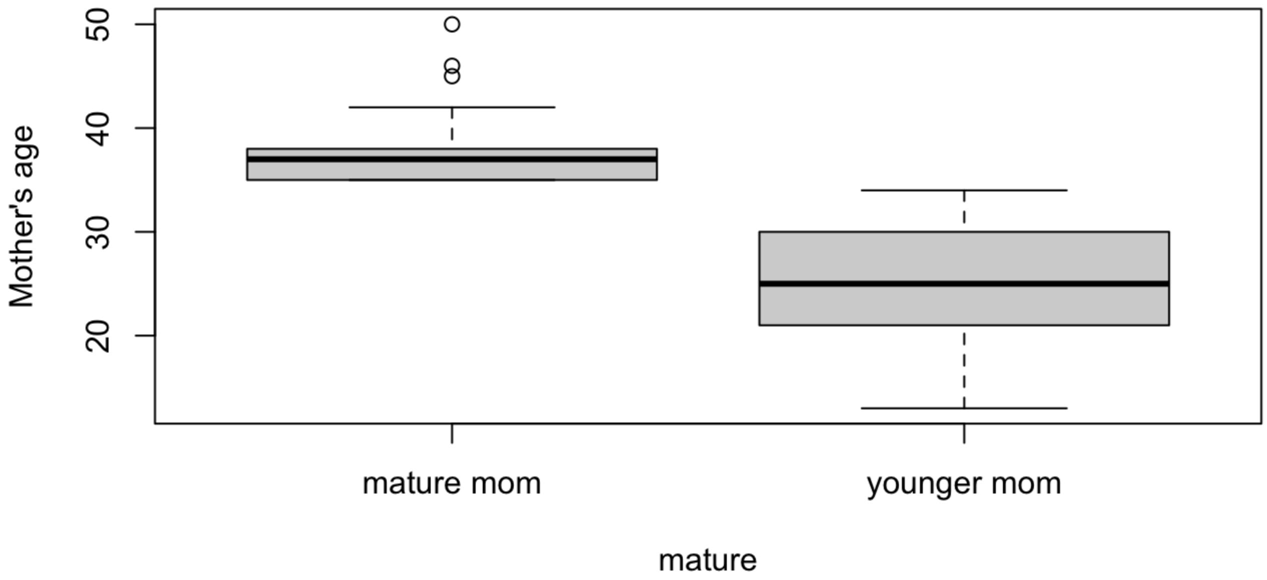


```
t.test(gained ~ mature, data = nc, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  gained by mature
## t = -1.3765, df = 175.34, p-value = 0.1704
## alternative hypothesis: true difference in means between group mature mom and group y
ounger mom is not equal to 0
## 95 percent confidence interval:
##  -4.3071463  0.7676886
## sample estimates:
##  mean in group mature mom mean in group younger mom
##                 28.79070                  30.56043
```

7. Now, a non-inference task: Determine the age cutoff for younger and mature mothers. Use a method of your choice, and explain how your method works.

```
boxplot(mage~mature, data=nc, ylab="Mother's age")
```



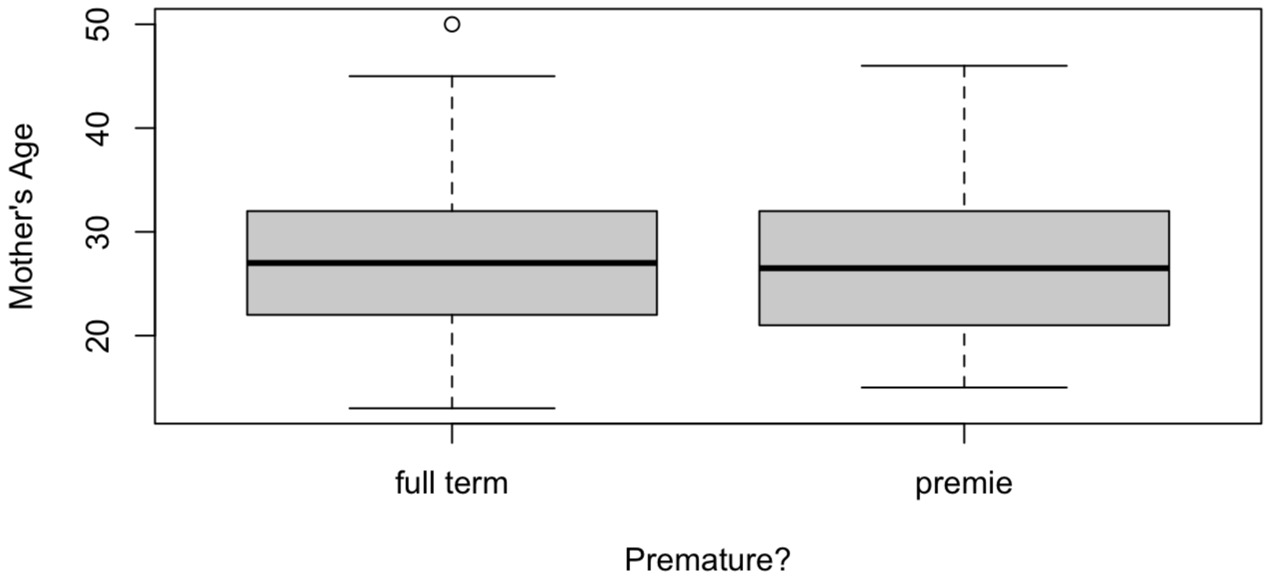
```
by(nc$mage, nc$mature, summary)
```

```
## nc$mature: mature mom
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   35.00   35.00   37.00   37.18   38.00   50.00 
## ------------------------------------------------------------ 
## nc$mature: younger mom
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   13.00   21.00   25.00   25.44   30.00   34.00
```

Exercise: Pick a pair of variables: one numerical (response) and one categorical (explanatory). Come up with a research question evaluating the relationship between these variables. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Answer your question using the `inference` function, report the statistical results, and also provide an explanation in plain language. Be sure to check all assumptions,state your α level, and conclude in context. (Note: Picking your own variables, coming up with a research question, and analyzing the data to answer this question is basically what you'll need to do for your project as well.)

Compare the age and and the length of term carried. Higher the age, higher chance of premature births

```
boxplot(mage~premie, data=nc, xlab="Premature?", ylab="Mother's Age")
```



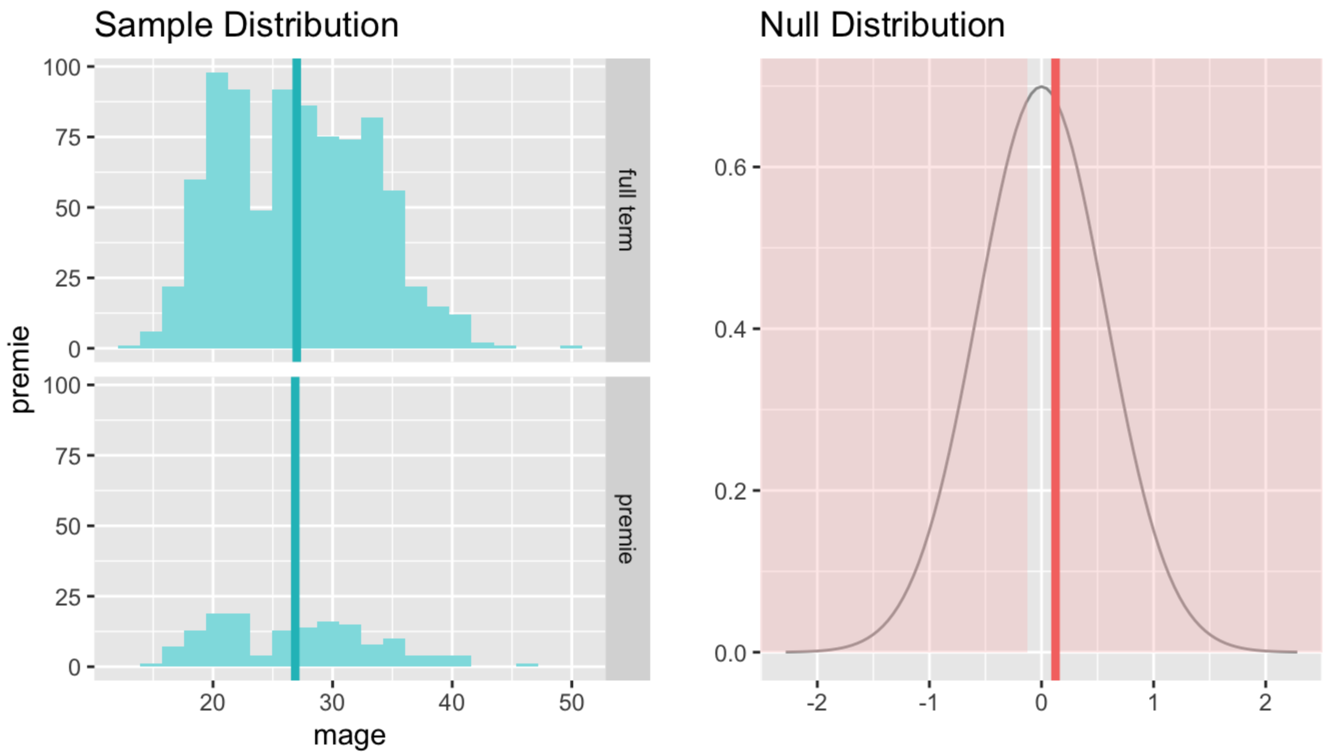
Conduct hypothesis test

$H_0 : \mu_{fullTerm} = \mu_{premie}$
$H_A : \mu_{fullterm} \neq \mu_{premie}$

Perform a test

```
inference(x = premie, y=mage, data = nc, statistic = "mean", type = "ht", null = 0,
          alternative = "twosided", method = "theoretical")
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_full term = 846, y_bar_full term = 27, s_full term = 6.1444
## n_premie = 152, y_bar_premie = 26.875, s_premie = 6.533
## H0: mu_full term =  mu_premie
## HA: mu_full term != mu_premie
## t = 0.2191, df = 151
## p_value = 0.8268
```



results show that we failed to reject the null hypothesis. So mother’s age does not impact the length carried to term

This is a product of OpenIntro that is released under a Creative Commons Attribution-ShareAlike 3.0 Unported (http://creativecommons.org/licenses/by-sa/3.0). This lab was written for OpenIntro by Andrew Bray and Mine Çetinkaya-Rundel.