week3

kangzq

2021/8/26

Introduce the t-distribution and comparing means as well as a simulation based method for creating a CI: bootstrapping.

t-distribution and comparing 2 means

t-distribution

When sample size is small…

- When σ unknown, t-distribution could address the uncertainty of the standard error estimate. t-distribution has thicker tails, thus CI will be wider.
- t-distribution only have ONE parameter: degrees of freedom; as df (df=n−1) increases the distribution will approaches the normal distribution.
- df describes the thickness of tails.

Calculate t value just as calculate z value. T = (obs - null)/SE.

- How to obtain a p-value for a t-test? Using R, pt/pnorm(z, df = , lower.tail =)*2.

Inference for a mean

- How to obtain a critical t-score (t⋆df) for a confidence interval?

One sample mean: df = n-1. Then using a table to find the corresponding value or using R by `qt(half-area, df = )` function.

Inference for comparing 2 independent means

As for comparing 2 independent means.

If within groups, the sampled observations must be independent by using random sample/assignment and n < 10% (sampling without replacement); if between groups then the groups must be independent of each other (non-paired).

- df = min(n1 - 1, n2 - 1)
- H0: μ1 − μ2 = 0.

Inference for comparing 2 paired means

Define observations as paired if each observation in one dataset has a special correspondence or connection with exactly one observation in the other data set.

- Carry out inference for paired data by first subtracting the paired observations from each other, and then treating the set of differences as a new numerical variable on which to do inference (such as a confidence interval or hypothesis test for the average difference).
- H0:μ_diff = 0
- A good interpretation of a confidence interval for the difference between two parameters includes a comparative statement (mentioning which group has the larger parameter).

Power

Calculate the power of a test for a given effect size and significance level:

1. Find the cutoff for the sample statistic that will allow the null hypothesis to be rejected at the given significance level.
2. Calculate the probability of obtaining that sample statistic given the effect size.

How power changes for changes in effect size, sample size, significance level, and standard error?

Use bootstrap methods for confidence intervals for categorical variables with at most two levels.

ANOVA (analysis of variance) and Bootstrapping

Comparing more than 2 means

ANOVA could be used to determine many groups at once.

H0: μ1=μ2=…=μk HA: At least one mean is different

ANOVA

Test statistic as F statistic: mean square between groups (MSG, variability between groups) and mean square error (MSE, variability within errors).

F statistic has a right skewed distribution with two different measures of degrees of freedom:

- one for the numerator (df_G = k-1, where k is the number of groups);
- one for the denominator (df_E = n-k, where nn is the total sample size).

calculate p-value by `pf(F, dfG, dfE, lower.tail = FALSE)` . No need to multiple 2 because only 1 tail in F distribution.

Conditions for ANOVA

Requirements:

1. The observations should be independent within and across groups

2. The data within each group are nearly normal.
3. The variability across the groups is about equal and use graphical diagnostics to check if these conditions are met.

Multiple comparisons

Conducting many t-tests for differences between each pair of means leads to an increased Type 1 Error rate.

We use a corrected significance level (Bonferroni correction, ^= /K, and K = k(k-1)/2.

Bootstrapping

For constructing confidence intervals (CI).

1. A random sample taken with replacement from the original sample of the same size as the original sample.
2. Calculate the bootstrap statistic such as mean, median, proportion…
3. Repeat 1 and 2 and create a bootstrap distribution.

Limitations:

1. Not as rigid as CLT based methods.
2. If severe skewed, the interval might be unreliable.
3. A representative sample is still required.

Week3 Lab

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.
`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` .

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 x 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`.

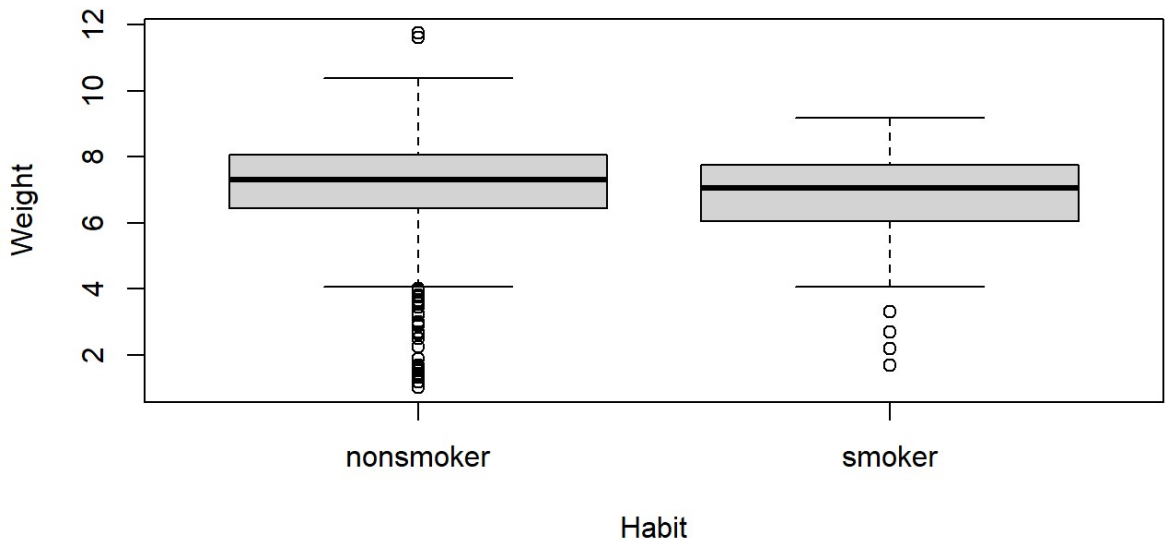
Using visualization and summary statistics, describe the distribution of weight gained by mothers during pregnancy. The `summary` function can also be useful.

```
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
```

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.

```
# type your code for the Question 3 here, and Knit
boxplot(weight ~ habit, data = nc, xlab = "Habit", ylab = "Weight")
```



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 x 2
##   habit     mean_weight
##   <fct>           <dbl>
## 1 nonsmoker        7.14
## 2 smoker           6.83
## 3 <NA>             3.63
```

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()`.

```
nc %>%
   group_by(habit) %>%
   summarise(group_size = n())
```

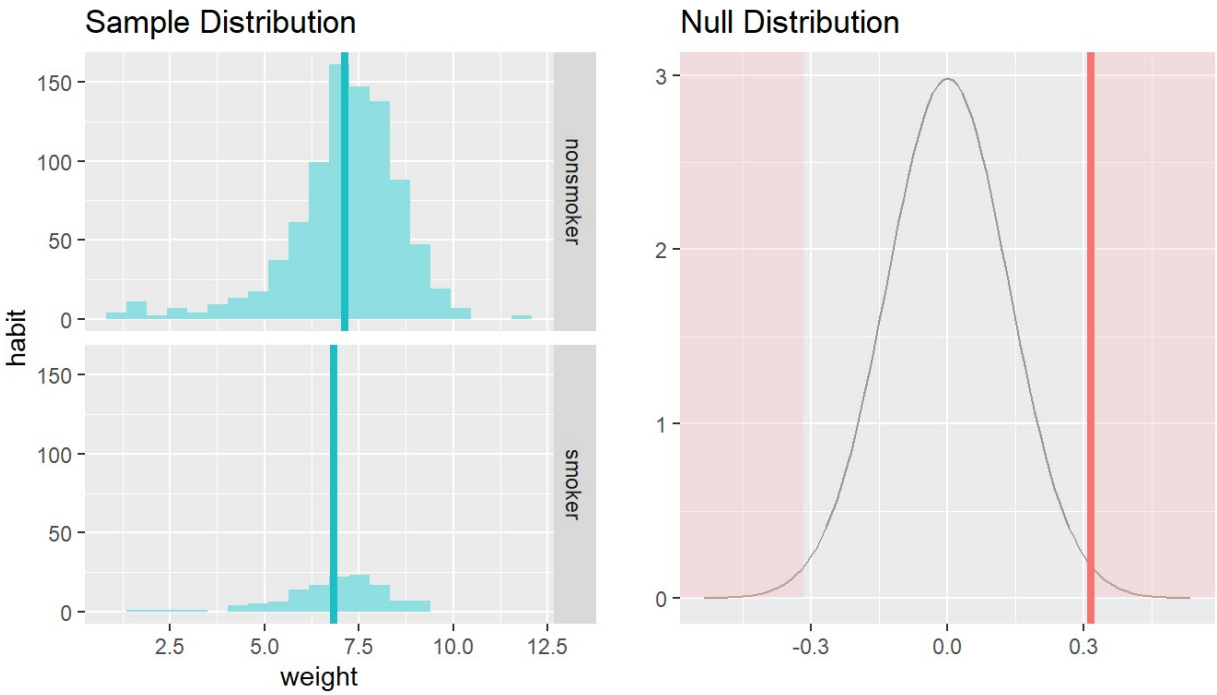
```
## # A tibble: 3 x 2
##   habit     group_size
##   <fct>          <int>
## 1 nonsmoker        873
## 2 smoker           126
## 3 <NA>               1
```

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
```



Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and non-smokers: `habit`. The third argument, `data`, is the data frame these variables are stored in. 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

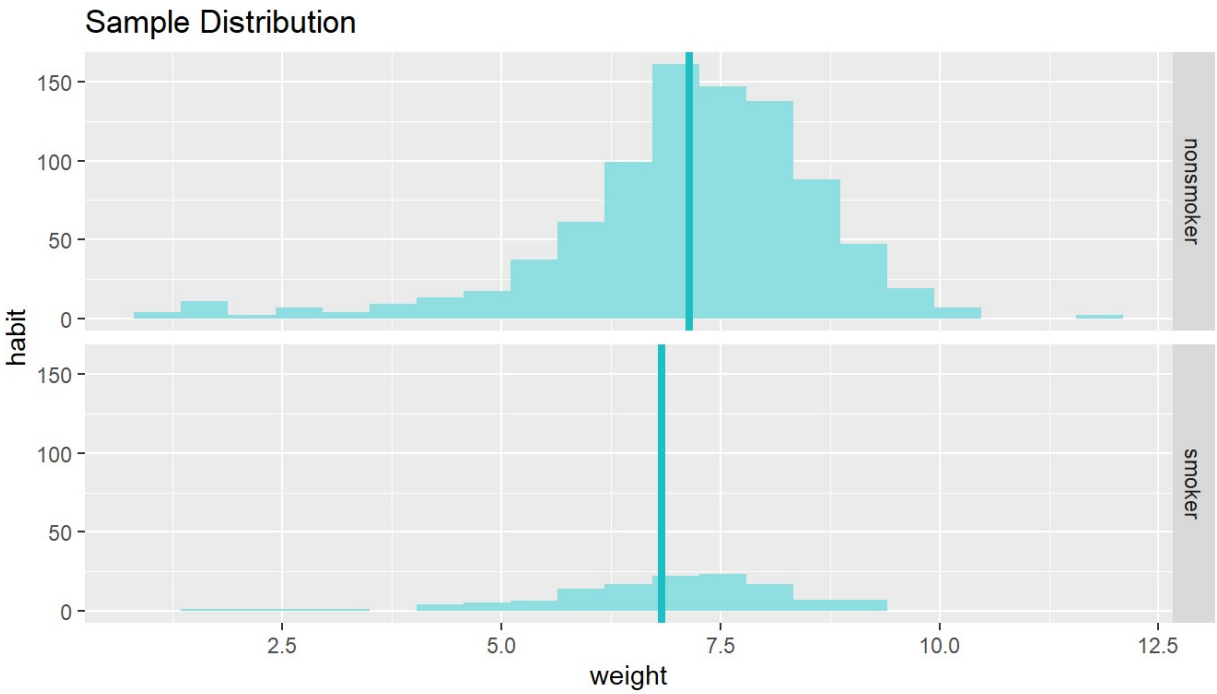
"median" and "proportion". Next we decide on the `type` of inference we want: a hypothesis test (`"ht"`) or a confidence interval (`"ci"`). When performing a hypothesis test, we also need to supply the `null` value, which in this case is `0` , since the null hypothesis sets the two population means equal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

```
# type your code for the Question 5 here, and Knit
inference(y = weight, x = habit, data = nc, statistic = "mean", type = "ci", null =
          0,
           alternative = "twosided", method = "theoretical")
```

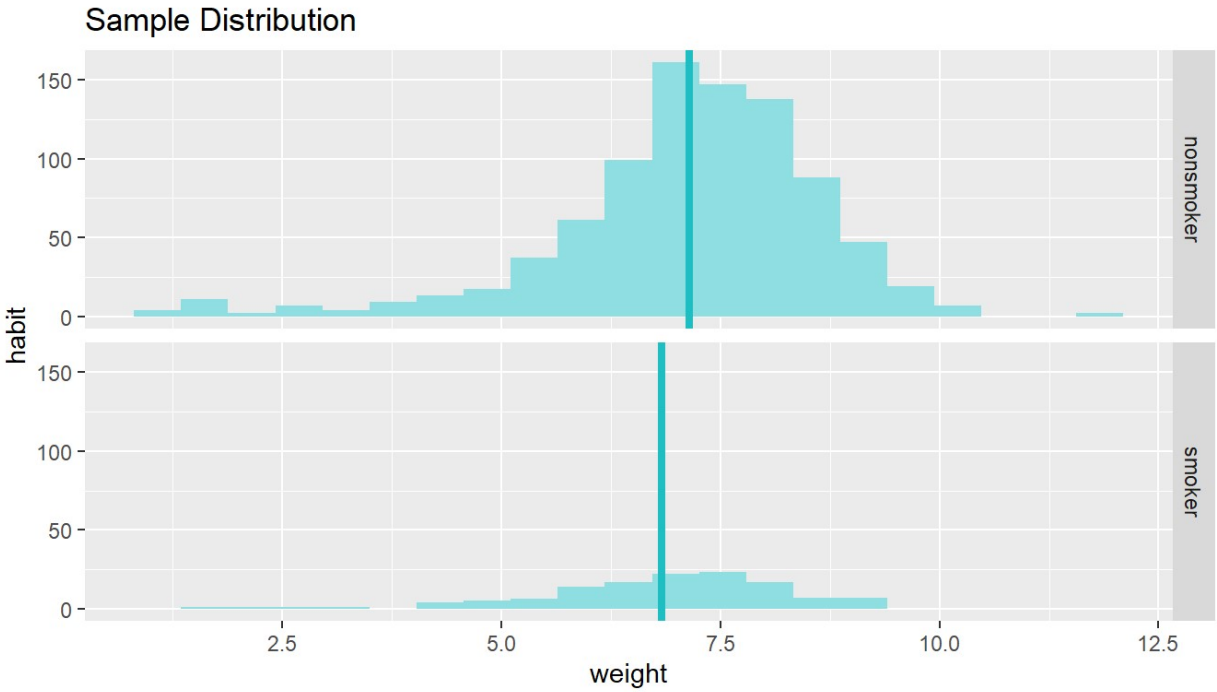
```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_nonsmoker = 873, y_bar_nonsmoker = 7.1443, s_nonsmoker = 1.5187
## n_smoker = 126, y_bar_smoker = 6.8287, s_smoker = 1.3862
## 95% CI (nonsmoker - smoker): (0.0508 , 0.5803)
```



By default the function reports an interval for ($\mu_{nonsmoker} - \mu_{smoker}$). 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"))
```

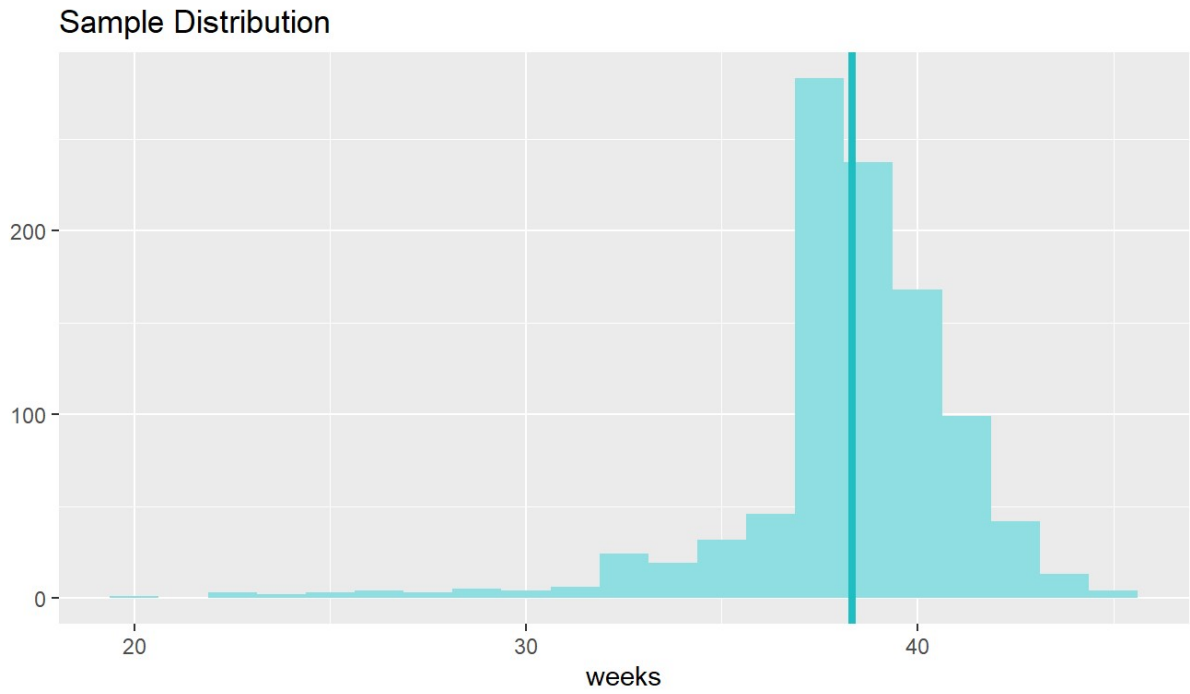
```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_smoker = 126, y_bar_smoker = 6.8287, s_smoker = 1.3862
## n_nonsmoker = 873, y_bar_nonsmoker = 7.1443, s_nonsmoker = 1.5187
## 95% CI (smoker - nonsmoker): (-0.5803 , -0.0508)
```



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)

```
# type your code for Question 6 here, and Knit
inference(y = weeks, data = nc, statistic = "mean", type = "ci", conf_level = 0.99,
          method = "theoretical")
```

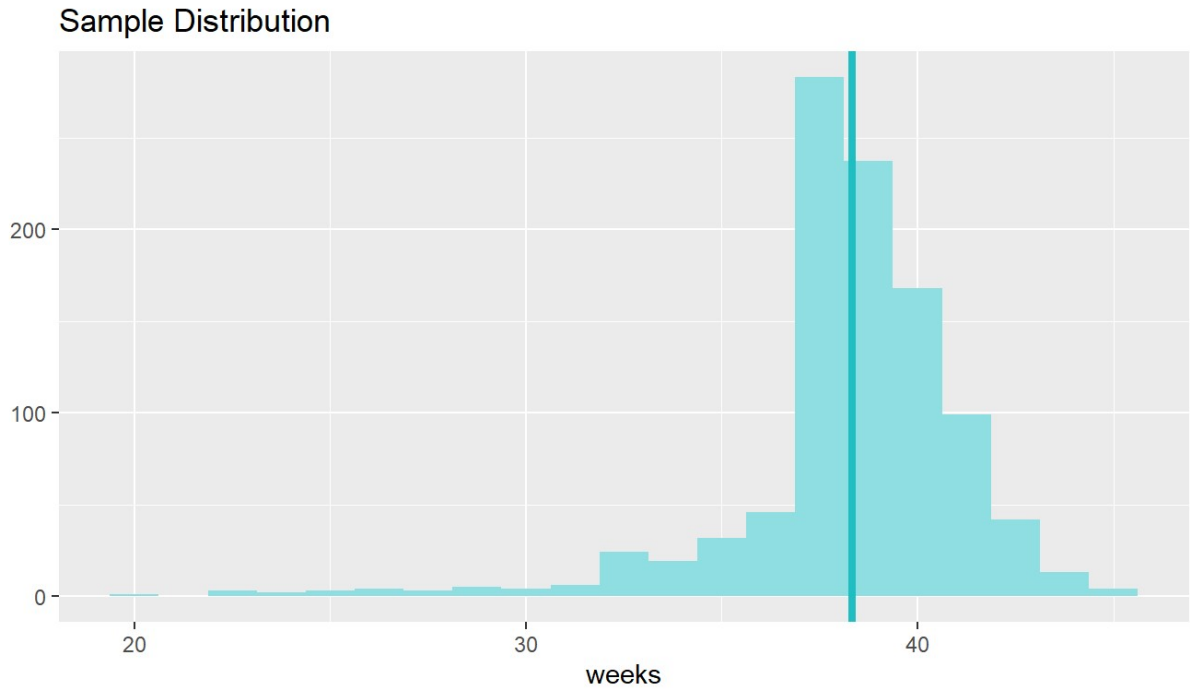
```
## Single numerical variable
## n = 998, y-bar = 38.3347, s = 2.9316
## 99% CI: (38.0952 , 38.5742)
```



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.

```
# type your code for the Exercise here, and Knit
inference(y = weeks, data = nc, statistic = "mean", type = "ci", conf_level = 0.90,
          method = "theoretical")
```

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

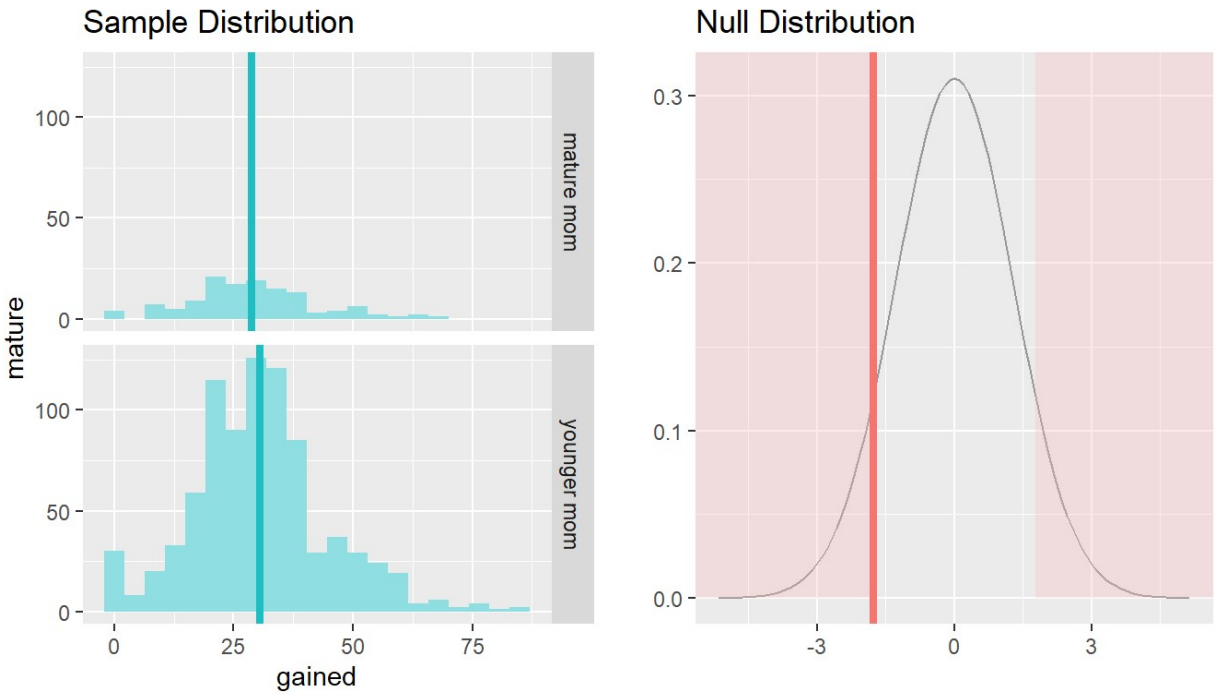


```
# it is not wider
```

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

```
# type your code for the Exercise here, and Knit
inference(x = mature , y = gained ,data =nc ,statistic = "mean",type = "ht", null =
         0,alternative = "twoside",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)
```

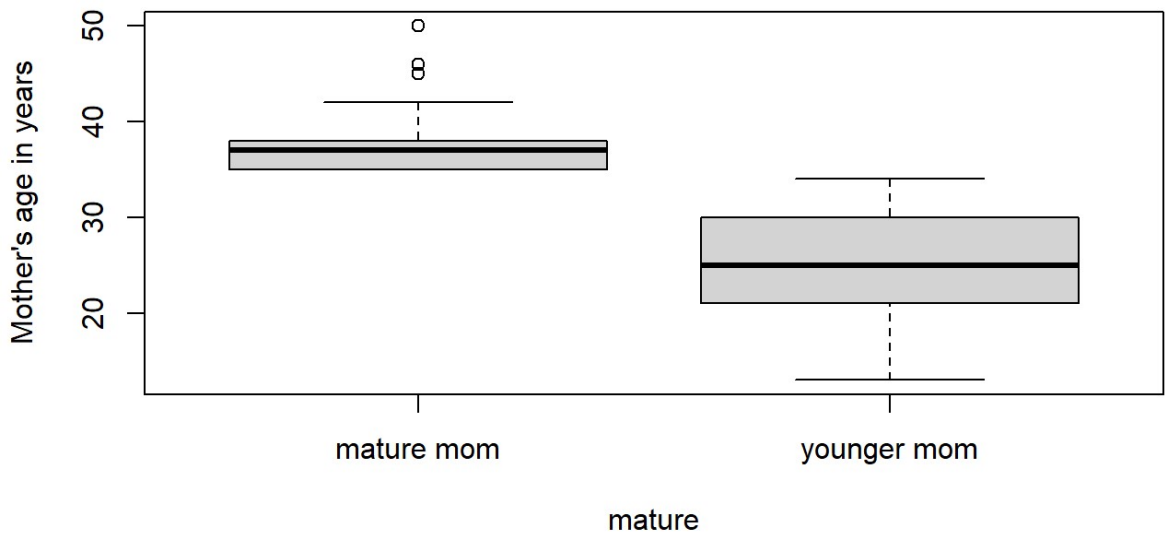
```
##
##  Welch Two Sample t-test
##
## data:  gained by mature
## t = -1.3765, df = 175.34, p-value = 0.1704
## alternative hypothesis: true difference in means between group mature mom and gr
oup younger mom is not equal to 0
## 95 percent confidence interval:
##  -4.3071463  0.7676886
## sample estimates: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

```
# type your code for Question 7 here, and Knit
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
```

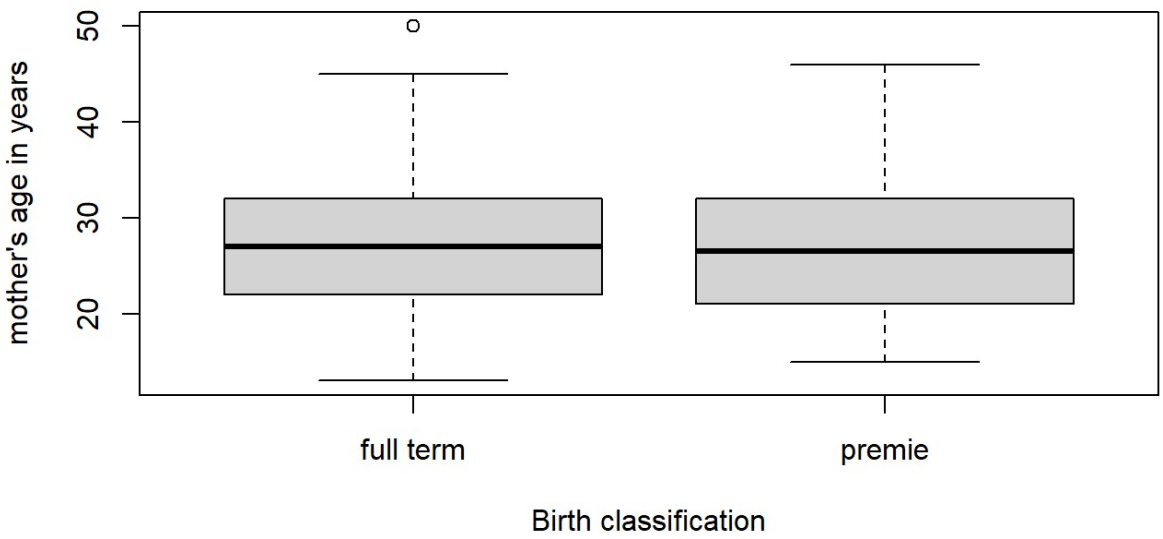
```
boxplot(mage ~ mature, data = nc, ylab = "Mother's age in years")
```



The cutoff is 35 years old. If >= 35 yrs, the woman is mature mom.

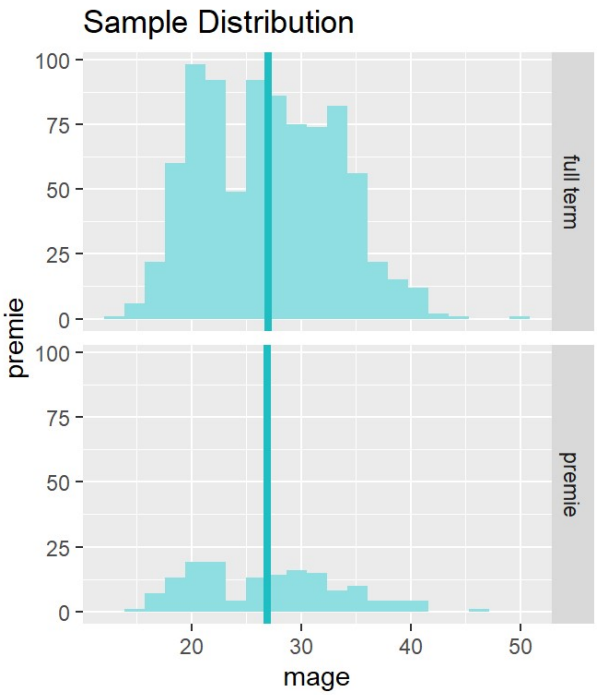
Exercise: Pick a pair of variables: one numerical (response) and one categorical (explanatory). Come up with a research question evaluating the relationship between these variables. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Answer your 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.)

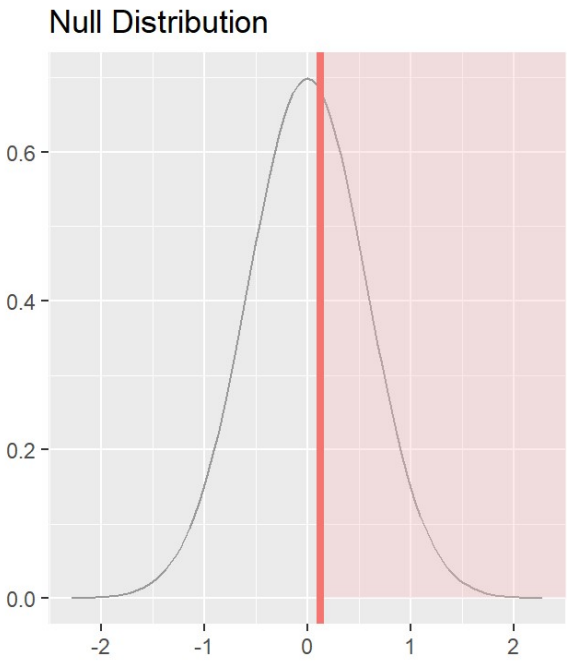
```
# type your code for the Exercise here, and Knit
boxplot(mage ~ premie, data = nc, xlab = "Birth classification", ylab = "mother's a
         ge in years")
```




```
inference(y = mage, x = premie, data = nc, statistic = "mean", type = "ht", null =
         0,
          alternative = "greater", 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 = 0.4134
```





Thus, we failed to reject the null hypothesis and the average of mother's age shows no difference in full term birth and premature birth.