Birth Study

We will study the relation between habits and practices of expectant mothers and the birth of their children from the birth records collected in 2004 and released by the state of North Carolina.

Getting Started

Load packages

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, `statsr`.

Let's load the packages.

```r
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.

```r
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:

```r
str(nc)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1000 obs. of  13 variables:
##  $ fage          : int  NA NA 19 21 NA NA 18 17 NA 20 ...
##  $ mage          : int  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  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  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  38 20 38 34 27 22 76 15 NA 52 ...
##  $ weight        : num  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 ...
```

In this data set, there are 1,000 cases, which represent the births.

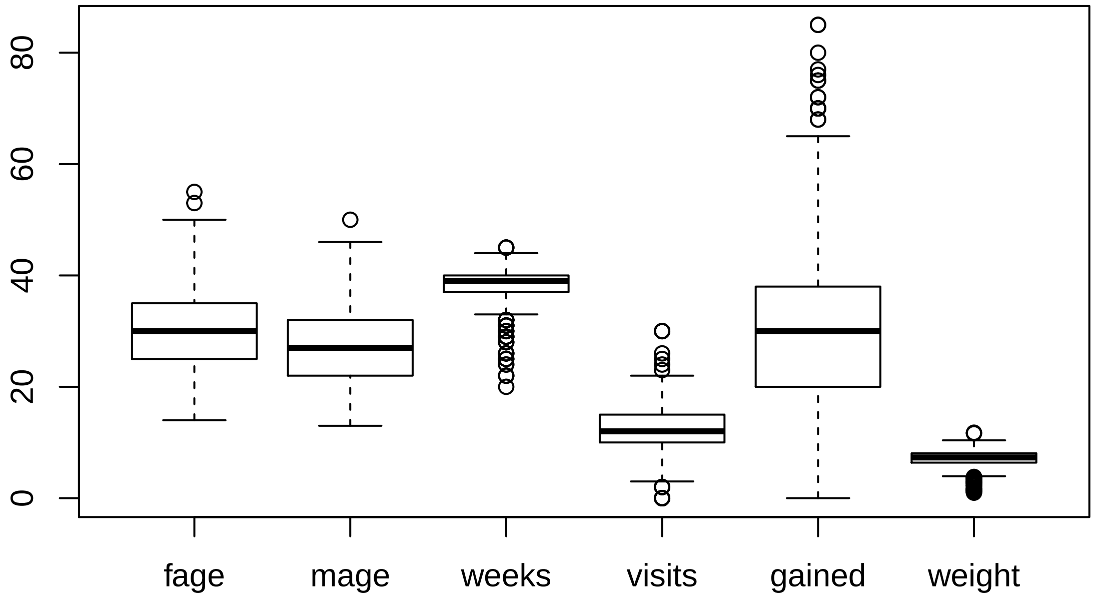
As we review the variable summaries, we know that `fage`, `mage`, `weeks`, `visits`, `gained`, `weight` are numerical variables, and others are categorical variables.

For numerical variables, are there outliers? We can take a closer look at the data by making a side-by-side boxplot.

```r
# Boxplot for numerical variables
names <- c("fage","mage", "weeks", "visits", "gained", "weight")

num_col <- nc[names]

# Draw side-by-side boxplot for numerical variables
boxplot(num_col, names=names)
```



From the above boxplot, we can see that there are some outliers in each numerical variables.

In what follows, let's study the relation between habits and practices of expectant mothers and the birth of their children via visualization and inferential statistics.

Q1: Does a mother's smoking habit have relationship with weight gained by mother during pregnancy?

Exploratory data analysis

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

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

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

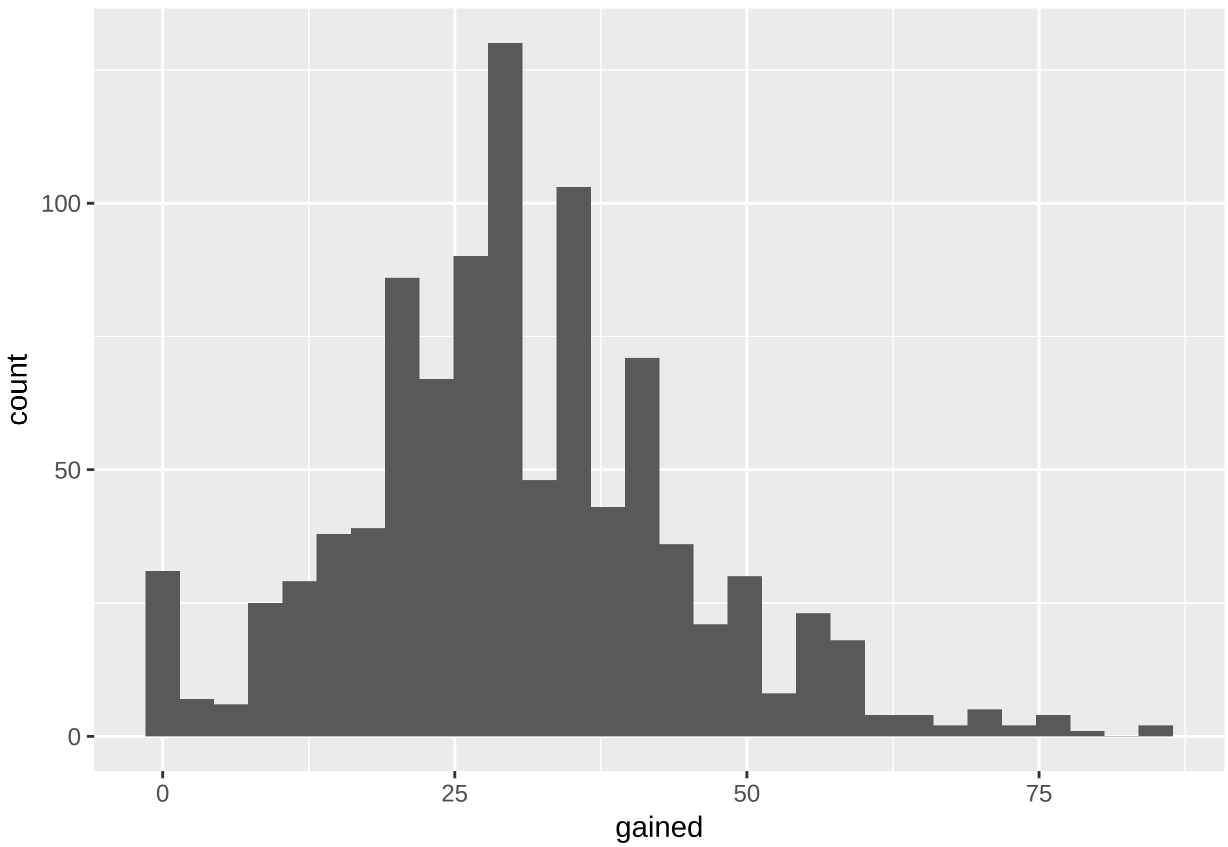
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   20.00   30.00   30.33   38.00   85.00      27
```

From above summaries, there are **27** missing weight gained data.

```
# Plot sample distribution of gained
ggplot(data = nc, aes(x = gained)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

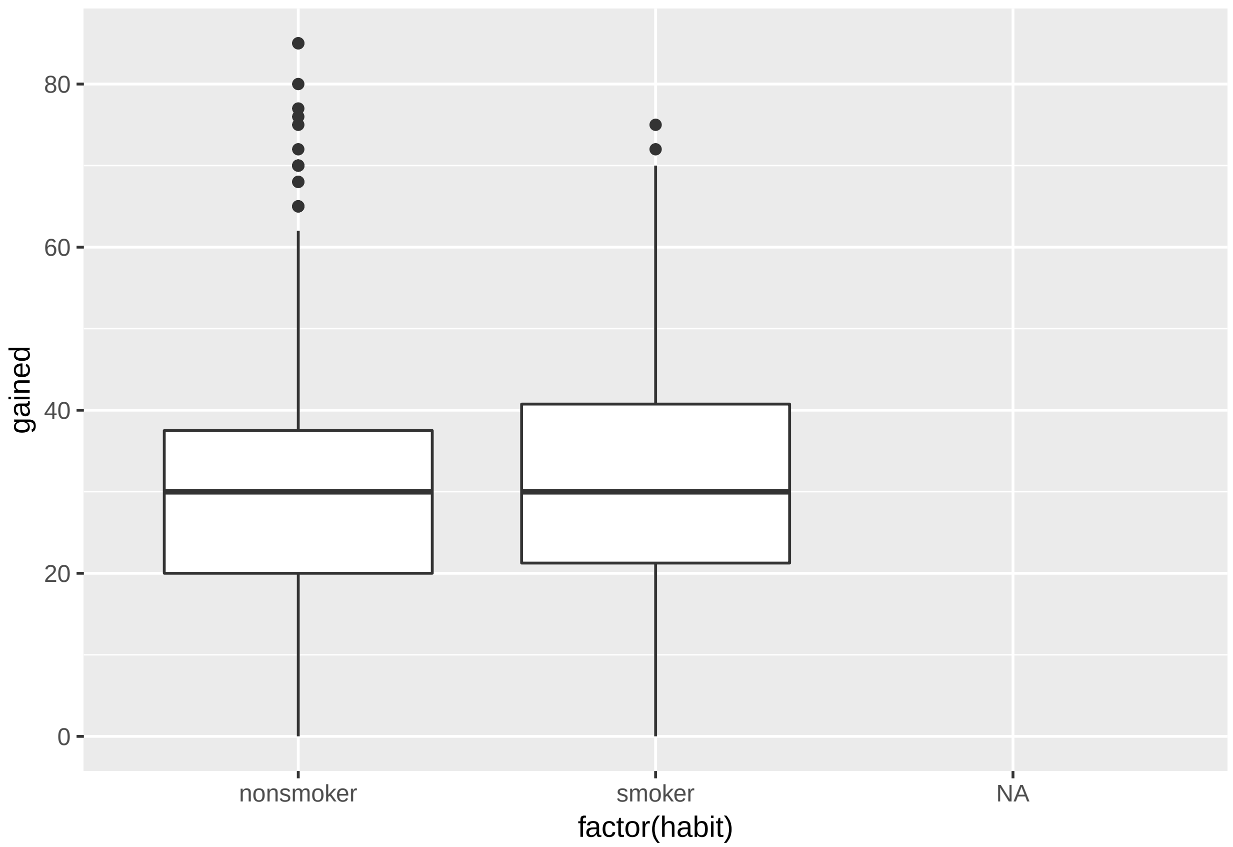
```
## Warning: Removed 27 rows containing non-finite values (stat_bin).
```



In order to study the relationship between `gained` and `habit`, we show the boxplot in the following.

```
# Gained vs habit
ggplot(nc, aes(x = factor(habit), y= gained)) +
  geom_boxplot()
```

```
## Warning: Removed 27 rows containing non-finite values (stat_boxplot).
```



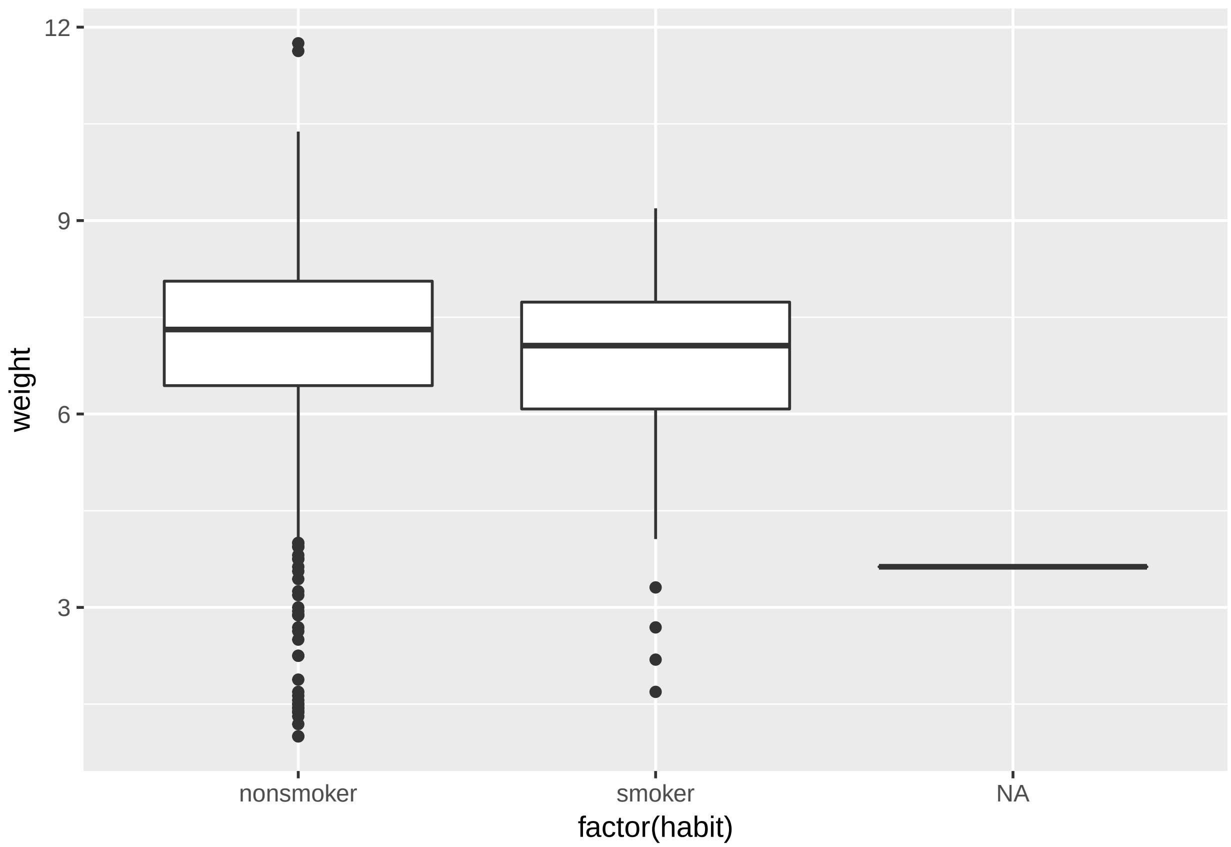
The medians for nonsmoker and smoker are very close so **there is no clear relationship between a mother's smoking habit and the weight gained by mother during pregnancy.**

Q2: Does a mother's smoking habit have relationship with weight of the baby at birth?

Exploratory data analysis

Next, we 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.

```
# weight vs habit (boxplot)
ggplot(nc, aes(x = factor(habit), y= weight)) +
  geom_boxplot()
```



From the side-by-side boxplots of `habit` and `weight`, median birth weight of babies born to non-smoker mothers is slightly higher than that of babies born to smoker mothers. Also, the range of birth weights of babies born to non-smoker mothers is greater than that of babies born to smoker mothers. Finally, the IQRs of the distributions are roughly equal.

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 - Hypothesis Testing

First of all, we set up appropriate hypotheses to evaluate if the average weights of babies born to smoking and non-smoking mothers are different.

- H_0: The average weights of babies born to smoking and non-smoking mothers are same.

$$\mu_{\text{smoking}} = \mu_{\text{non}-\text{smoking}} \Rightarrow \mu_{\text{smoking}} - \mu_{\text{non}-\text{smoking}} = 0$$

- H_A: There is difference between average weights of babies born to smoking and non-smoking mothers.

$$\mu_{\text{smoking}} \neq \mu_{\text{non}-\text{smoking}} \Rightarrow \mu_{\text{smoking}} - \mu_{\text{non}-\text{smoking}} \neq 0$$

2.1 Check conditions necessary for inference

- Since the data come from random sample and consist of less than 10% of each class, smoking or non-smoking, **the observations are independent**.
- The sample size for each class is larger than 30, the distribution of sample mean is **nearly normal.**

```
# Check Normality
nc %>%
  group_by(habit) %>%
  summarise(sample_size = n())
```

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

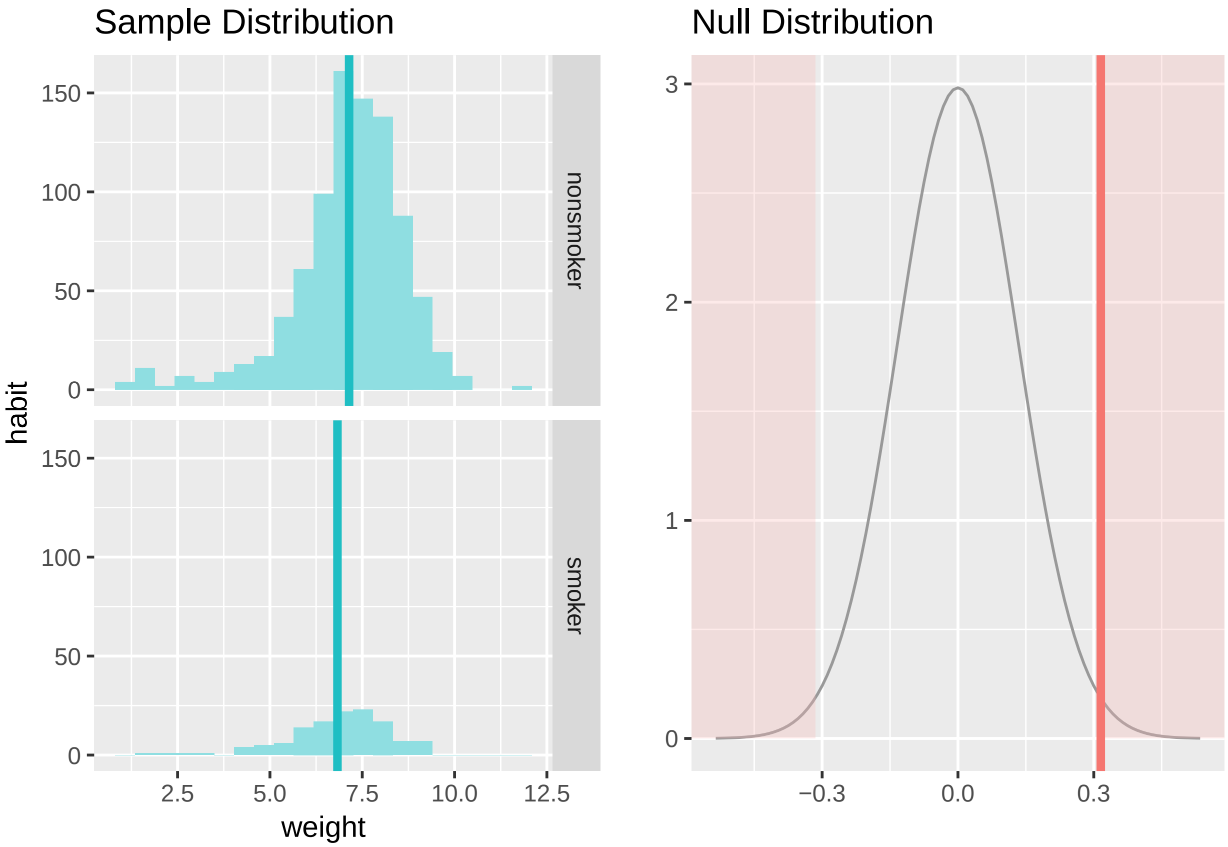
2.2 Calculate p-value

Next, we will use the function, `inference`, 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 stored in. Next one is `statistic`, which is the sample statistic we're using, or similarly, the population parameter we're estimating. Next we decide on the `type` of inference we want: a hypothesis test (`"ht"`) or a confidence interval (`"ci"`). Here, we are considering hypothesis test. 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"`. Here, we are considering **two-sided hypothesis test**. Lastly, the `method` of inference can be `"theoretical"` (CLT based) or `"simulation"` (randomization/bootstrap) based.

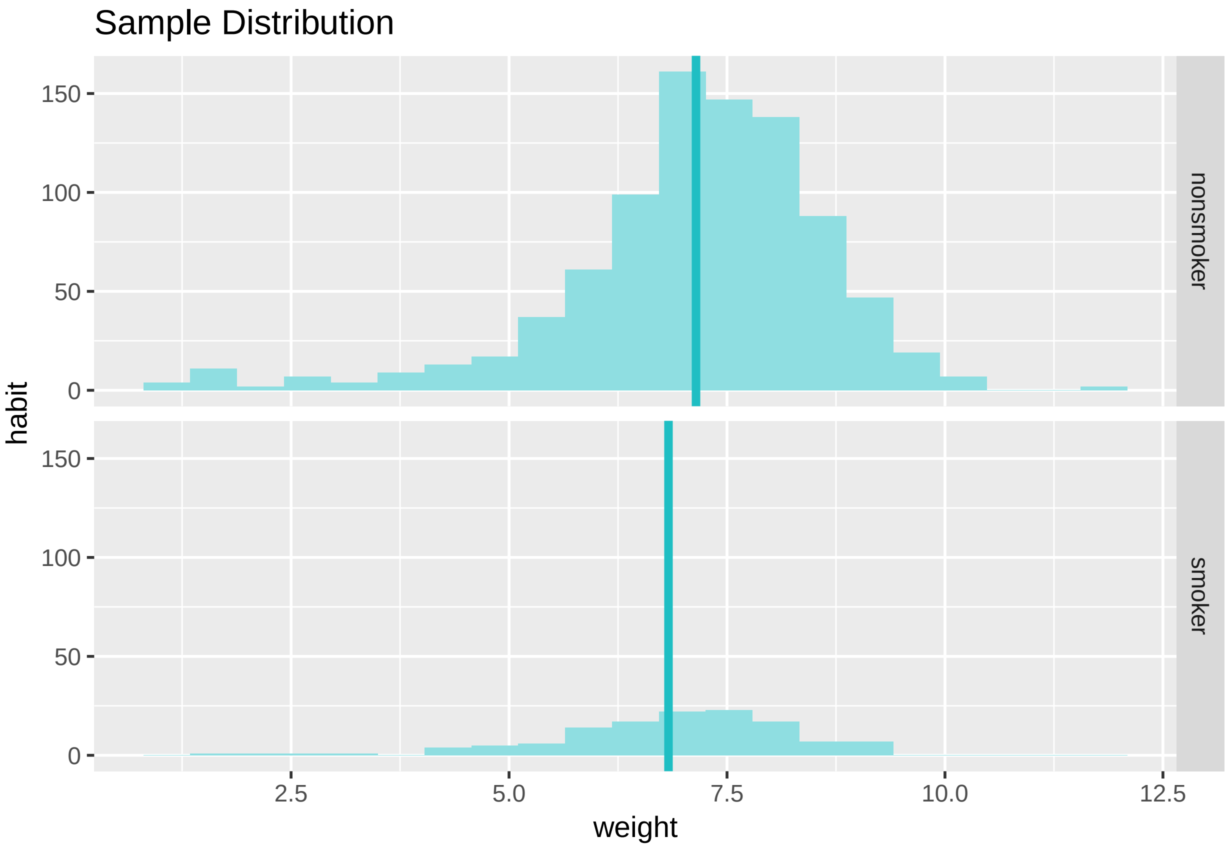
From the result of hypothesis testing, p-value is less than significance level ($\alpha = 0.05$) so we reject the null hypothesis H_0, and accept the alternative hypothesis H_A. **We conclude that the data provide convincing evidence that there is difference between average weights of babies born to smoking and non-smoking mothers.**

2.3 Confidence Interval (CI)

Now we use `inference` function to calculate 95% confidence interval by changing 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 we'll get a 95% confidence interval. If we 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.

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



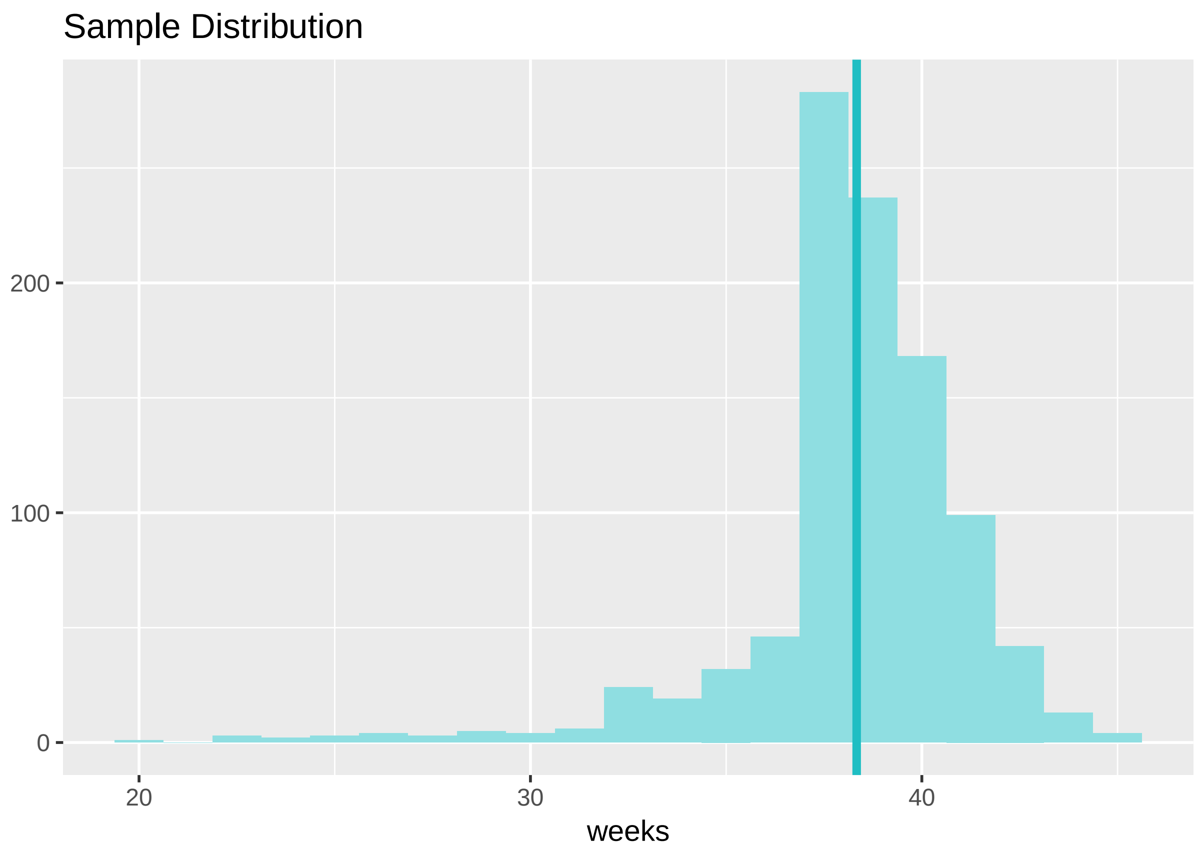
We are 95% confident that babies born to nonsmoker mothers are on average 0.05 to 0.58 pounds heavier at birth than babies born to smoker mothers. Also note that the null value 0 does not be included in confidence interval so it agrees with the conclusion from p-value calculation.

Q3 What is the average length of pregnancies in unit of weeks?

We calculate a 99% confidence interval for the average length of pregnancies (`weeks`) in the following. Note that since we're doing inference on a single population parameter, there is no explanatory variable, so we can omit the `x` variable from the function.

```r
inference(y = weeks, data = nc, statistic = "mean", type = "ci",
          method = "theoretical", conf_level = 0.99)
```

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

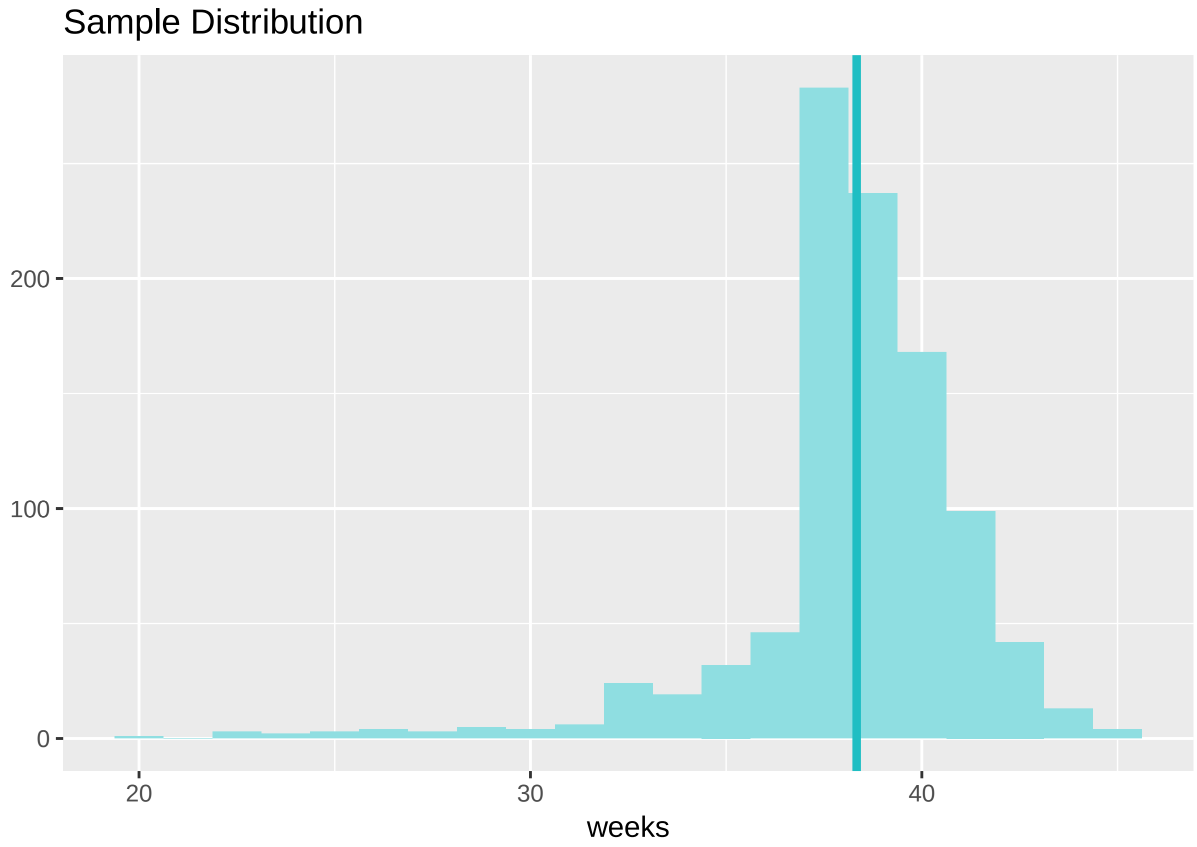


We are 99% confident that the population mean of length of pregnancies is between 38.0952 and 38.5742 weeks.

In order to narrow down the confidence interval, we calculate a new confidence interval for the same parameter at the 90% confidence level.

```r
inference(y = weeks, data = nc, statistic = "mean", type = "ci",
          method = "theoretical", conf_level = 0.90)
```

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



We are 90% confident that the population mean of length of pregnancies is between 38.1819 and 38.4874 weeks. Note that the confidence interval is narrower than the one obtained in the the previous calculation.

Q4 Is the average weight gained by younger mothers different from the average weight gained by mature mothers?

The Age Cutoff for Younger and Mature Mothers

First of all, we determine the age cutoff for younger and mature mothers.

```
# Determine cutoff age
nc %>%
  group_by(mature) %>%
  summarise(max_age = max(mage), min_age = min(mage))
```

```
## # A tibble: 2 x 3
##   mature      max_age min_age
##   <fct>         <dbl>   <dbl>
## 1 mature mom       50      35
## 2 younger mom      34      13
```

The maximum age of younger moms is 34 and minimum age of mature moms is 35. Hence, the age cutoff is 35 for younger and mature mothers.

Exploratory data analysis

We consider the possible relationship between a maturity status of mother (`mature`) and the weight gained by mother during pregnancy in pounds (`gained`).

Note that there are 27 missing values in `gained` field.

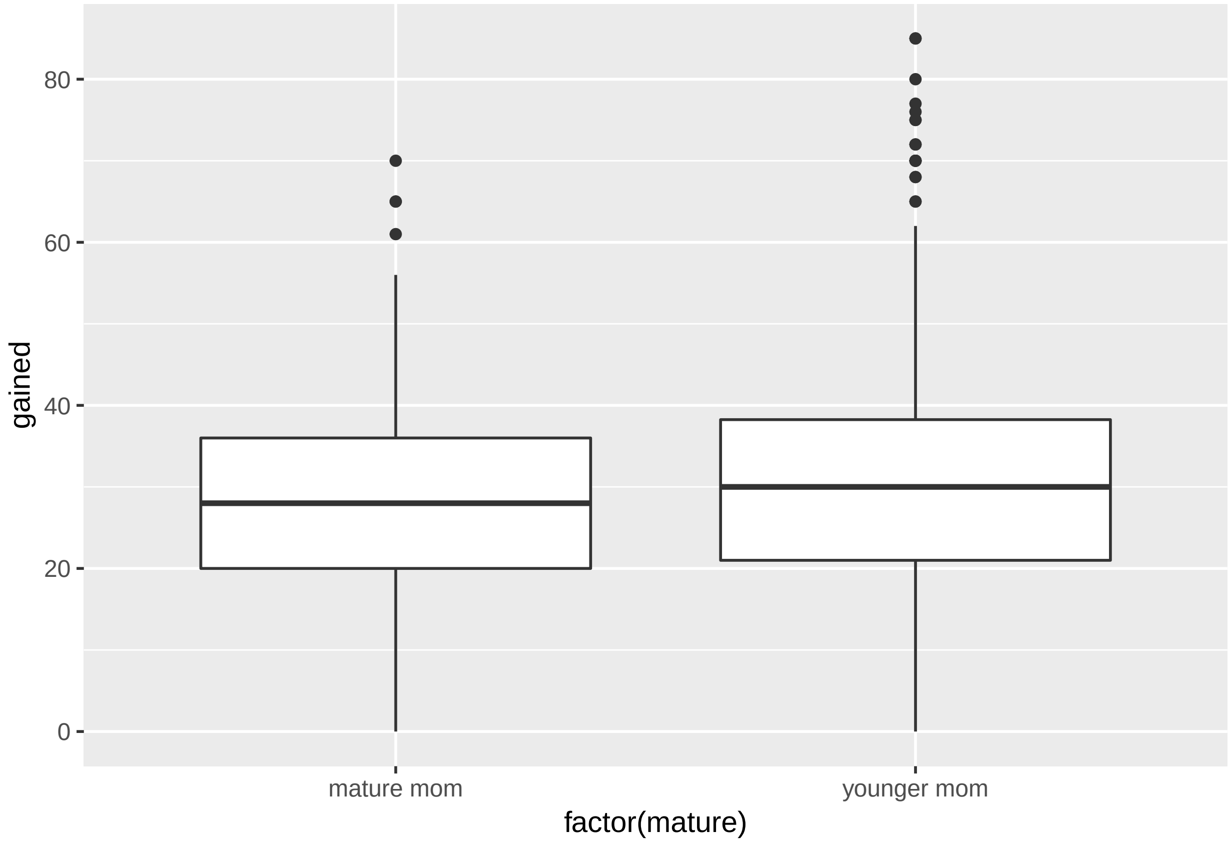
```
summary(nc$gained)
```

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

Let's remove the missing values in `gained` field before doing further analysis.

```
df_gained <- nc[!is.na(nc$gained),]
```

```
# Mature vs Gained (boxplot)
ggplot(df_gained, aes(x = factor(mature), y= gained)) +
  geom_boxplot()
```



The side-by-side boxplots of `mature` and `gained` show that **median weight gained by younger mothers is slightly higher than weight gained by mature mothers**. Also, the IQRs of the distributions are roughly equal.

The box plots show how the medians of the two distributions compare, but we can compare the means of the distributions as well.

```
df_gained %>%
  group_by(mature) %>%
  summarise(mean_weight = mean(gained))
```

```
## # A tibble: 2 x 2
##   mature      mean_weight
##   <fct>             <dbl>
## 1 mature mom         28.8
## 2 younger mom        30.6
```

There is an observed difference. Let's do hypothesis test to examine whether this difference is statistically significant.

Inference - Hypothesis Testing

First of all, we set up appropriate hypotheses to evaluate whether the average weight gained by younger mothers is different from the average weight gained by mature mothers.

- H_0: The average weights gained by younger and mature mothers are same.

  $$\mu_{\text{younger}} = \mu_{\text{mature}} \Rightarrow \mu_{\text{younger}} - \mu_{\text{mature}} = 0$$

  so the null value is 0.

- H_A: The average weights gained by younger and mature mothers are different.

  $$\mu_{\text{younger}} \neq \mu_{\text{mature}} \Rightarrow \mu_{\text{younger}} - \mu_{\text{mature}} \neq 0$$

  Here, we consider two-sided test.

4.1 Check conditions necessary for inference

- Since the data come from random sample and consist of less than 10% of each class, smoking or non-smoking, **the observations are independent**.
- The sample size for each class is larger than 30, the distribution of sample mean is **nearly normal.**

```
# Check Normality
df_gained %>%
  group_by(gender) %>%
  summarise(sample_size = n())
```

```
## # A tibble: 2 x 2
##   gender sample_size
##   <fct>        <int>
## 1 female         488
## 2 male           485
```

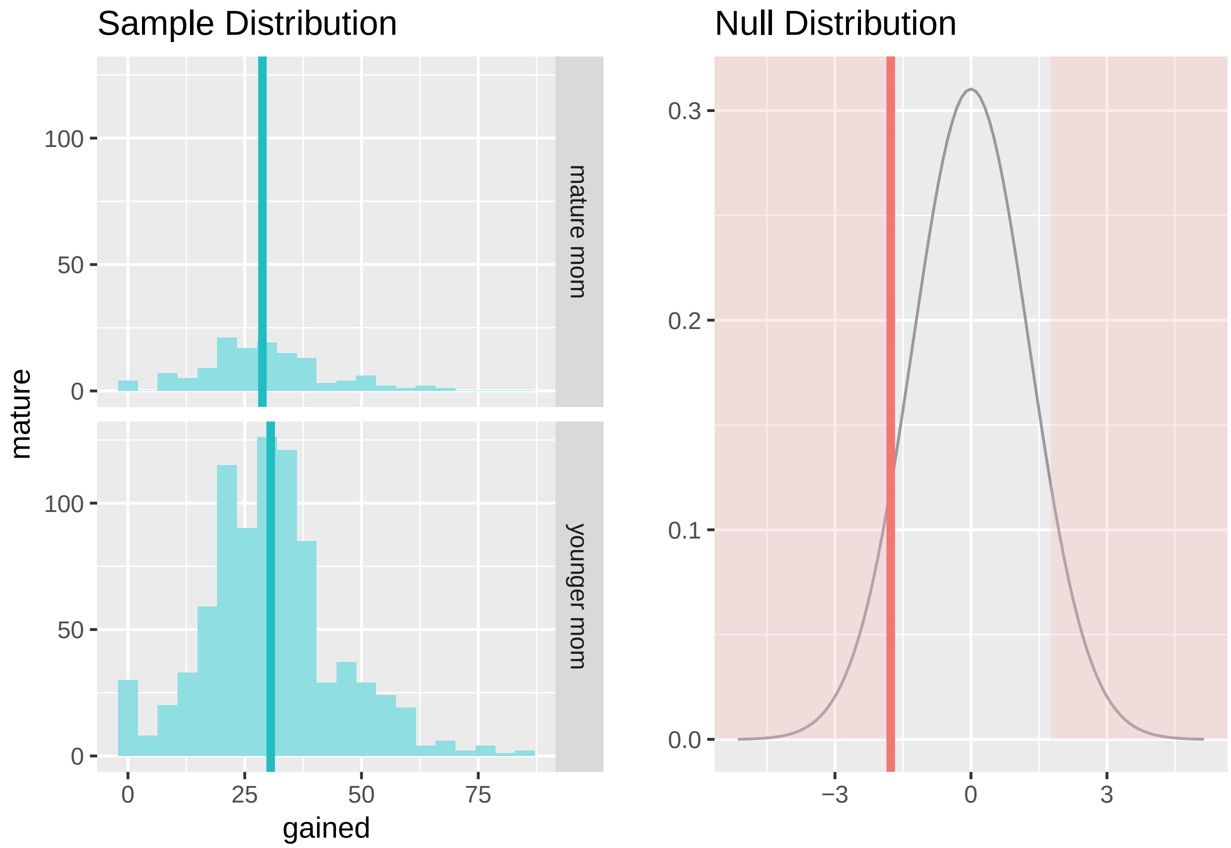
4.2 Calculate p-value

We will use the function, `inference`, for conducting hypothesis tests and constructing confidence intervals.

Then, run the following:

```
inference(y = gained, x = mature, data = df_gained, 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
```



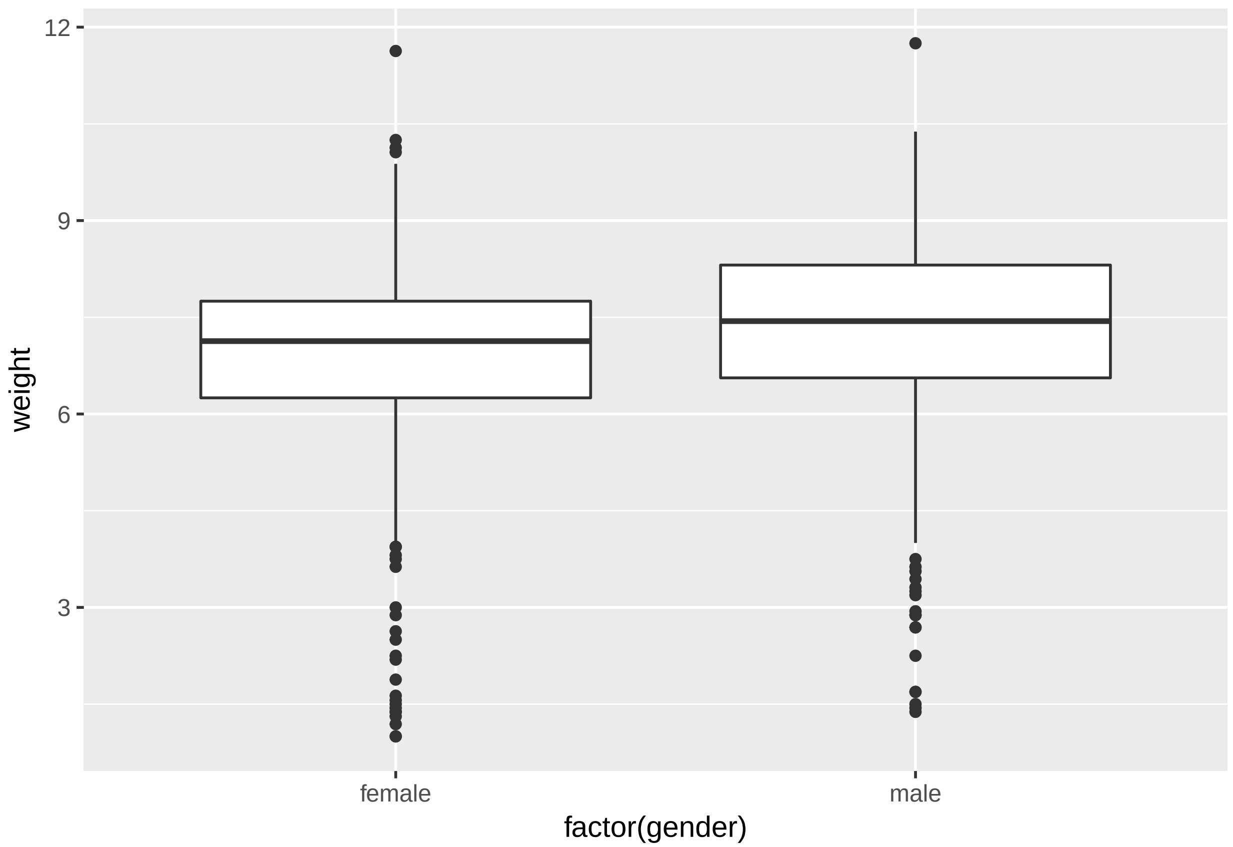
The p-value is much lager than significance level ($\alpha = 0.05$) so we fail to reject the null hypothesis H_0. **We conclude that the data do not provide convincing evidence that there is difference between average weights gained by younger and mature mothers.**

Q5 Is the average weight of male babies identical to female babies?

Exploratory data analysis

We consider the possible relationship between a gender of baby (`gender`) and the weight of baby (`weight`).

```
# Gender vs Weight (boxplot)
ggplot(nc, aes(x = factor(gender), y= weight)) +
  geom_boxplot()
```



From the side-by-side boxplots of `gender` and `weight`, **median birth weight of male babies is slightly higher than that of female babies**.

The box plots show how the medians of the two distributions compare, but we can compare the means of the distributions as well.

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

```
## # A tibble: 2 x 2
##   gender mean_weight
##   <fct>        <dbl>
## 1 female        6.90
## 2 male          7.30
```

There is an observed difference. Let's do hypothesis test to examine whether this difference is statistically significant.

Inference - Hypothesis Testing

First of all, we set up appropriate hypotheses to evaluate if the average weights of male babies and female babies are different.

- H_0: The average weights of male and female babies are same.

  $$\mu_{\text{male}} = \mu_{\text{female}} \Rightarrow \mu_{\text{male}} - \mu_{\text{female}} = 0$$

  so the null value is 0.

- H_A: The average weights of male and female babies are different.

  $$\mu_{\text{male}} \neq \mu_{\text{female}} \Rightarrow \mu_{\text{male}} - \mu_{\text{female}} \neq 0$$

  Here, we consider two-sided test.

5.1 Check conditions necessary for inference

- Since the data come from random sample and consist of less than 10% of each class, smoking or non-smoking, **the observations are independent**.
- The sample size for each class is larger than 30, the distribution of sample mean is **nearly normal.**

```r
# Check Normality
nc %>%
  group_by(gender) %>%
  summarise(sample_size = n())
```

```
## # A tibble: 2 x 2
##   gender sample_size
##   <fct>        <int>
## 1 female         503
## 2 male           497
```

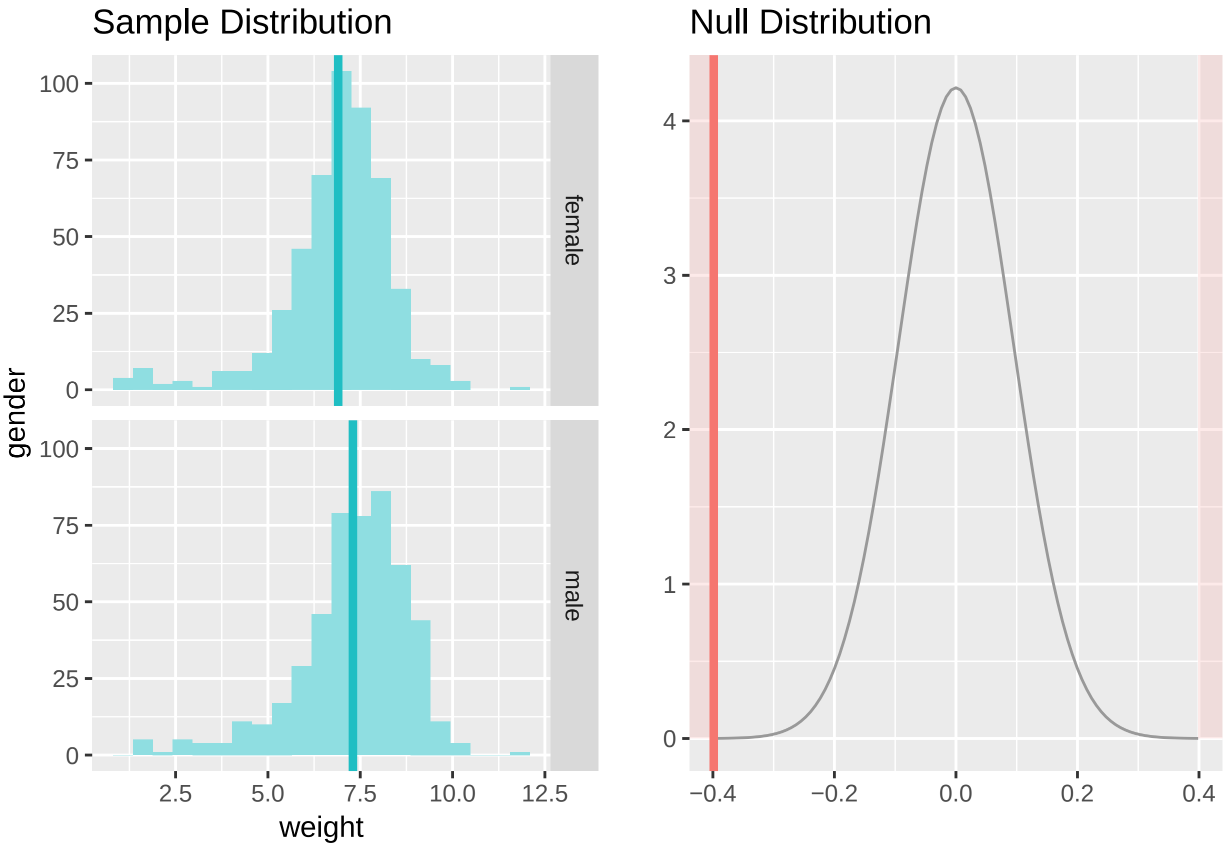
5.2 Calculate p-value

We will use the function, `inference`, for conducting hypothesis tests and constructing confidence intervals.

Then, run the following:

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

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_female = 503, y_bar_female = 6.9029, s_female = 1.4759
## n_male = 497, y_bar_male = 7.3015, s_male = 1.5168
## H0: mu_female =  mu_male
## HA: mu_female != mu_male
## t = -4.2113, df = 496
## p_value = < 0.0001
```



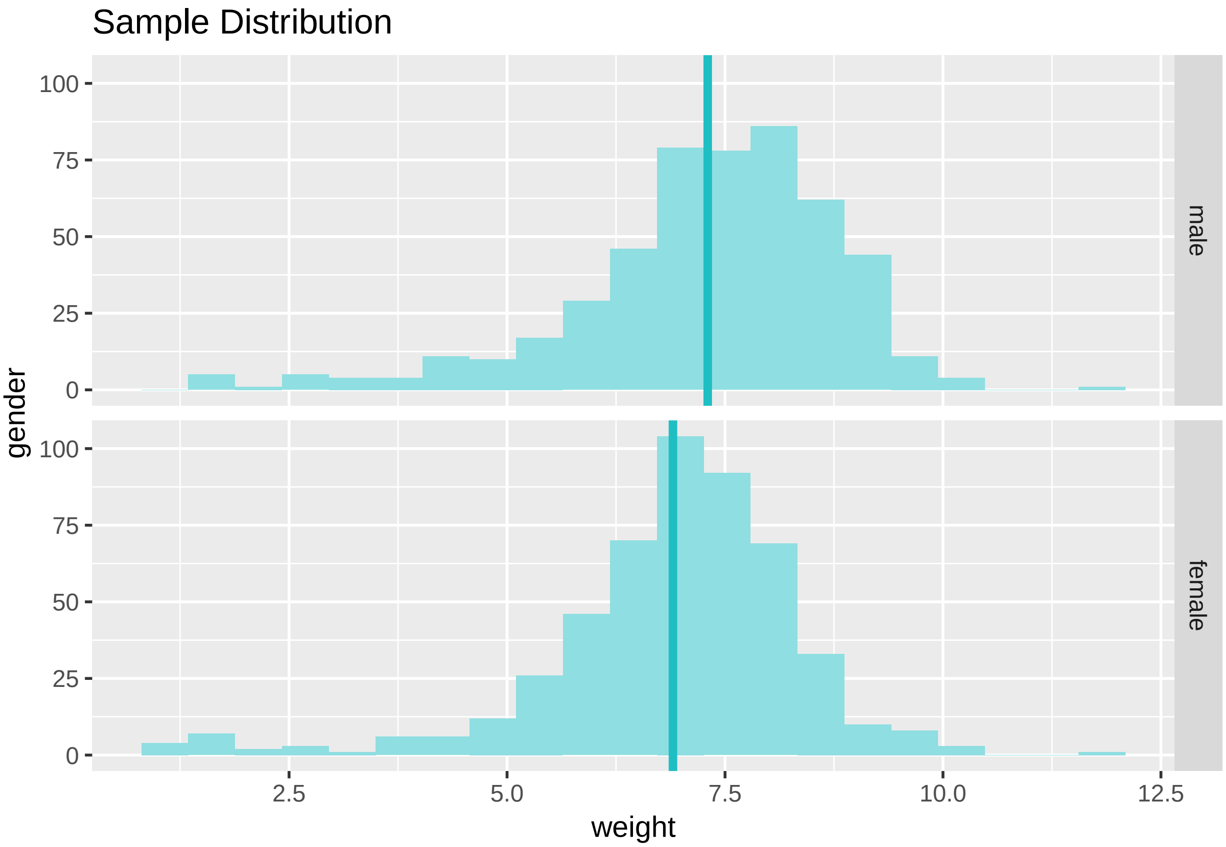
The p-value is much less than significance level ($\alpha = 0.05$) so we reject the null hypothesis H_0, and accept the alternative hypothesis H_A. **We conclude that the data provide convincing evidence that there is difference between average weights of male and female babies.**

5.3 Confidence Interval (CI)

Now we use `inference` function to calculate 95% confidence interval.

```
inference(y = weight, x = gender, data = nc, statistic = "mean", type = "ci", method = "theoretical", o
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_male = 497, y_bar_male = 7.3015, s_male = 1.5168
## n_female = 503, y_bar_female = 6.9029, s_female = 1.4759
## 95% CI (male - female): (0.2126 , 0.5846)
```



We are 95% confident that male babies are on average 0.2126 to 0.5846 pounds heavier at birth than female babies. Also note that the null value 0 does not be included in confidence interval so it agrees with the conclusion from p-value calculation.

This study was carried out by Hsuan-Hao Fan.