Lab5 - Inference for numerical data

Michael Y.

March 24th, 2019

North Carolina births

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.

Exploratory analysis

Load the `nc` data set into our workspace.

```
load("more/nc.RData")
```

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. What are the cases in this data set? How many cases are there in our sample?

There are 1000 cases in this data set. Each case represents information associated with the birth of a child in North Carolina in 2004. The 13 features contain information about the baby, the mother and limited information about the father (just his age, if known.)

As a first step in the analysis, we should consider summaries of the data. This can be done using the `summary` command:

```
summary(nc)
```

```
##       fage            mage            mature        weeks
##  Min.   :14.000   Min.   :13   mature mom :133   Min.   :20.000
##  1st Qu.:25.000   1st Qu.:22   younger mom:867   1st Qu.:37.000
##  Median :30.000   Median :27                     Median :39.000
##  Mean   :30.256   Mean   :27                     Mean   :38.335
##  3rd Qu.:35.000   3rd Qu.:32                     3rd Qu.:40.000
##  Max.   :55.000   Max.   :50                     Max.   :45.000
##  NA's   :171                                     NA's   :2
##       premie        visits              marital        gained
##  full term:846   Min.   : 0.000   married    :386   Min.   : 0.000
##  premie   :152   1st Qu.:10.000   not married:613   1st Qu.:20.000
##  NA's     :  2   Median :12.000   NA's       :  1   Median :30.000
##                  Mean   :12.105                     Mean   :30.326
##                  3rd Qu.:15.000                     3rd Qu.:38.000
##                  Max.   :30.000                     Max.   :85.000
##                  NA's   :9                          NA's   :27
##      weight       lowbirthweight    gender          habit
##  Min.   : 1.000   low    :111    female:503   nonsmoker:873
##  1st Qu.: 6.380   not low:889    male  :497   smoker   :126
##  Median : 7.310                                NA's     :  1
##  Mean   : 7.101
##  3rd Qu.: 8.060
##  Max.   :11.750
##
##       whitemom
##  not white:284
```

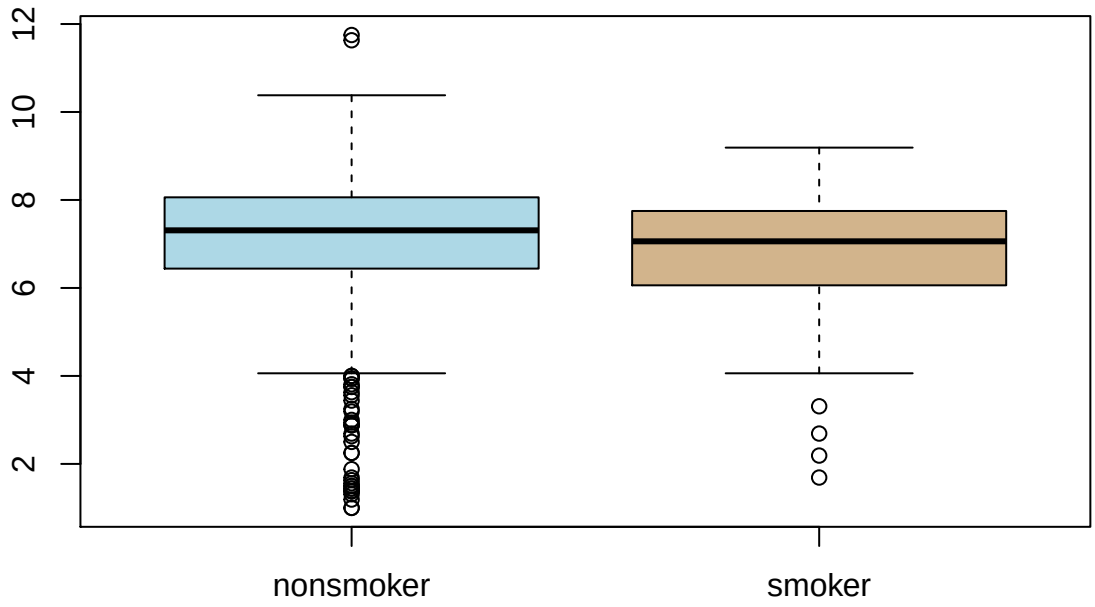
```
##  white    :714
##  NA's     :  2
##
##
##
##
```

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.

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.

2. Make a side-by-side boxplot of `habit` and `weight`. What does the plot highlight about the relationship between these two variables?

```
boxplot(nc$weight~nc$habit,col=(c("lightblue","tan")))
```



The plots show that the median birth weight for children born to nonsmoking mothers appears to be slightly greater than the median birth weight for children born to mothers who smoke.

The sizable number of outliers at the bottom indicate that the distributions are left-skewed, as the means for each group should be lower than the respective medians.

There are a much larger number of outliers associated with non-smoking mothers, but from the above box plots alone it is not apparent that the two subsets are of very different size, with 873 nonsmokers vs. 126 smokers, a ratio of nearly 7:1 .

```
nc[is.na(nc$habit),] %>%
  kable() %>%
  kable_styling(full_width = T) #%>%
```

	fage	mage	mature	weeks	premie	visits	marital	gained	weight	lowbirthweight	gender	habit	whitemom
988	NA	41	mature mom	NA	NA	NA	NA	NA	3.63	low	female	NA	white

```
  #scroll_box(width="1000px",height="150px")
```

NB: there is also one case where it is unknown whether the mother was a smoker or not; that observation is excluded from the above plots. In such case, the weight of the child was low (3.63 lbs) .

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 function to split the `weight` variable into the `habit` groups, then take the mean of each using the `mean` function.

```
by(nc$weight, nc$habit, mean)
```

```
## nc$habit: nonsmoker
## [1] 7.1442726
## ------------------------------------------------------------
## nc$habit: smoker
## [1] 6.8287302
```

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

Inference

3. Check if the conditions necessary for inference are satisfied. Note that you will need to obtain sample sizes to check the conditions. You can compute the group size using the same `by` command above but replacing `mean` with `length`.

```
samplesize <- by(nc$weight, nc$habit, length)
samplesize
```

```
## nc$habit: nonsmoker
## [1] 873
## ------------------------------------------------------------
## nc$habit: smoker
## [1] 126
```

There are two conditions necessary to apply the t-distribution to the difference in sample means.

(1) Because the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent. There were 119773 babies born in North Carolina in 2004.

(Source: https://schs.dph.ncdhhs.gov/schs/births/babybook/2004/northcarolina.pdf)

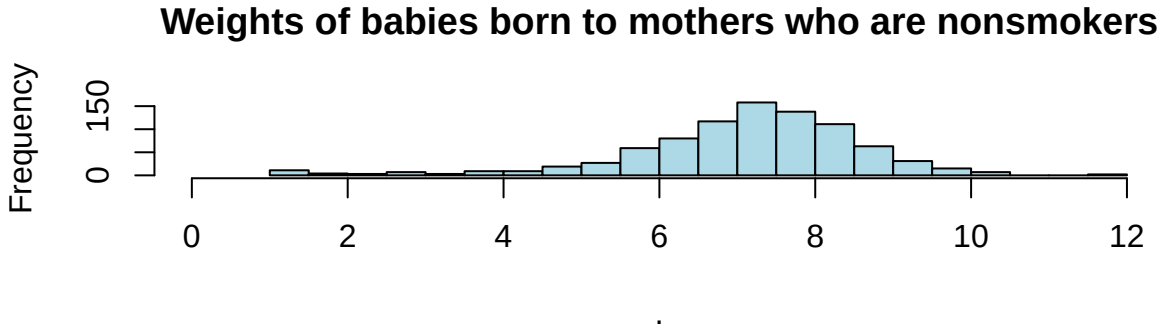
Summary of nonsmoker and smoker:

```
nc %>% drop_na(habit) %>% spread(key = habit, value=weight) %>%  select(nonsmoker,smoker) -> nc_smoke
summary(nc_smoke)
```

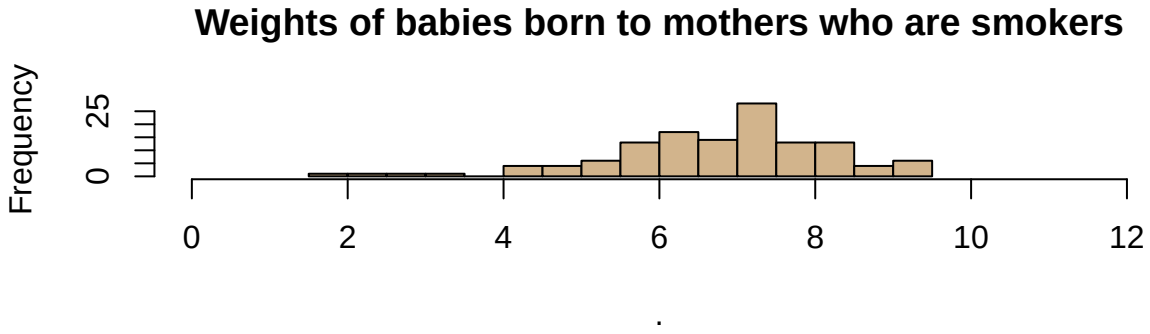
```
##    nonsmoker           smoker      
##  Min.   : 1.0000   Min.   :1.6900  
##  1st Qu.: 6.4400   1st Qu.:6.0775  
##  Median : 7.3100   Median :7.0600  
##  Mean   : 7.1443   Mean   :6.8287  
##  3rd Qu.: 8.0600   3rd Qu.:7.7350  
##  Max.   :11.7500   Max.   :9.1900  
##  NA's   :126       NA's   :873
```

Histograms, one above the other:

```
par(mfrow = c(2,1))
nc_smoke %>% pull(nonsmoker) %>% hist(main="Weights of babies born to mothers who are nonsmokers", breal
nc_smoke %>% pull(smoker) %>% hist(main="Weights of babies born to mothers who are smokers", breaks=24,
```



.

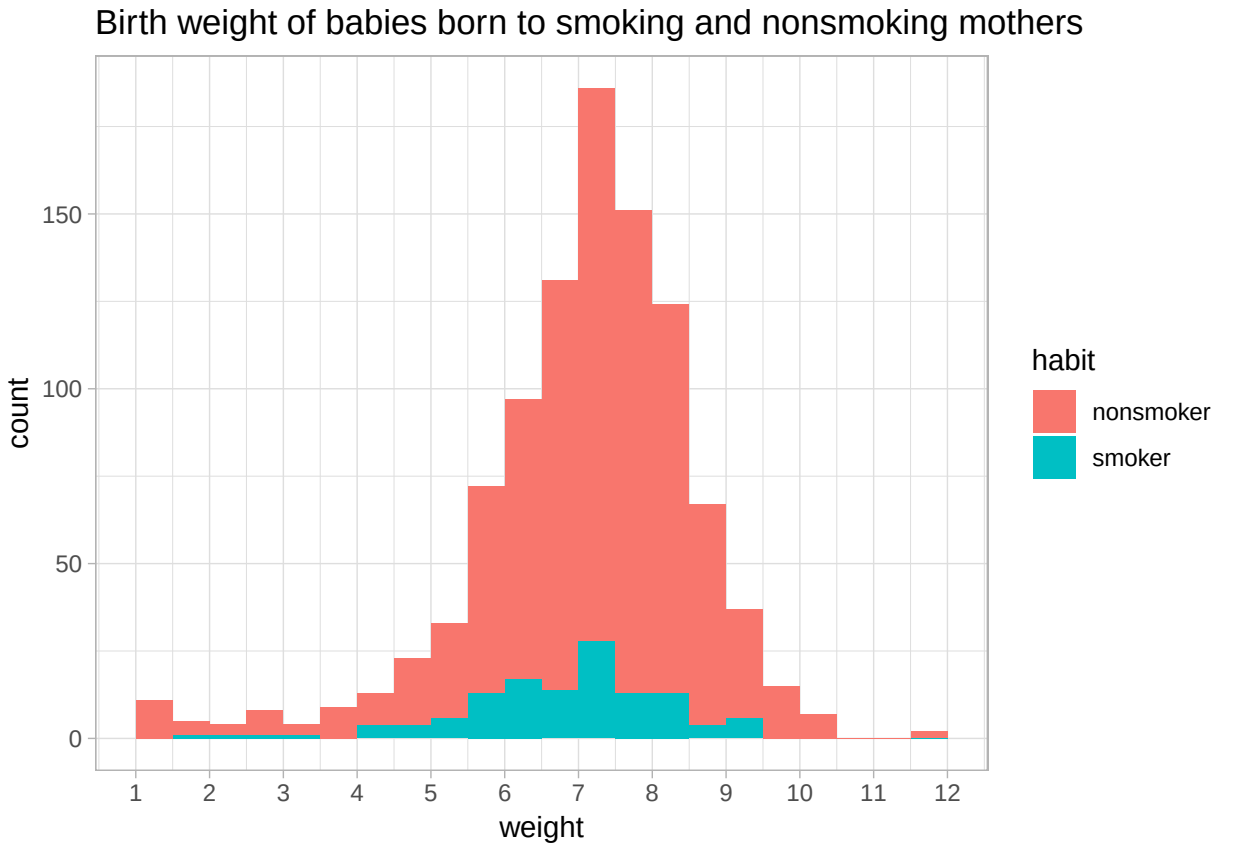


.

Histograms superimposed on same scale:

```
nc %>% drop_na(habit) %>%
ggplot(.,aes(x=weight,fill=habit)) +
    geom_histogram(binwidth=0.5, center=0.25) +
    theme_light() +
```

```
  scale_x_continuous(breaks=seq(0,12, by = 1))+
  ggtitle("Birth weight of babies born to smoking and nonsmoking mothers")
```



While each distribution is strongly left-skewed, the sample sizes of 126 and 873 make it reasonable to model each mean separately using a t-distribution. The skew is reasonable for these sample sizes of 126 and 873 .

(2) The independence reasoning applied above also ensures the observations in each sample are independent. Since both conditions are satisfied, the difference in sample means may be modeled using a t-distribution.

However, because the sample sizes are substantially larger than 30, and because the t-distribution converges to the Normal distribution for sufficiently large sample sizes, the results from using the Normal distribution will be substantially the same as those from the t-distribution.

4. Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.

Null Hypothesis: There is no difference in average birth weight for newborns from mothers who did and did not smoke.

In statistical notation:

$$H_0: \quad \mu_{nonsmoker} = \mu_{smoker} \quad \Rightarrow \quad \mu_{nonsmoker} - \mu_{smoker} = 0$$

where $\mu_{nonsmoker}$ represents the average weight of babies born to non-smoking mothers and μ_{smoker} represents the average weight of babies born to mothers who smoked.

Alternative Hypothesis: There is some difference in average newborn weights from mothers who did and did not smoke:

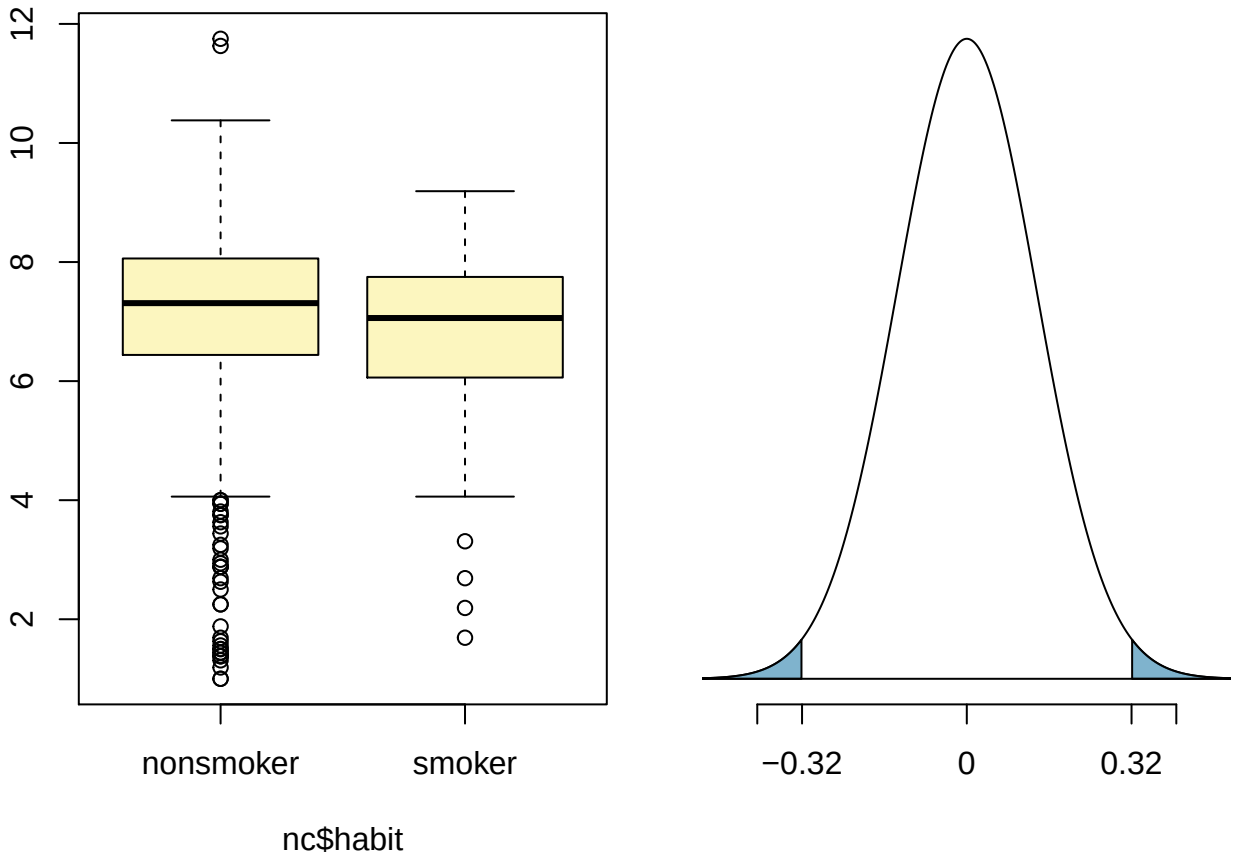
$$H_A :: \quad \mu_{nonsmoker} \neq \mu_{smoker} \quad \Rightarrow \quad \mu_{nonsmoker} - \mu_{smoker} \neq 0$$

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

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

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_nonsmoker = 873, mean_nonsmoker = 7.1443, sd_nonsmoker = 1.5187
## n_smoker = 126, mean_smoker = 6.8287, sd_smoker = 1.3862

## Observed difference between means (nonsmoker-smoker) = 0.3155
##
## H0: mu_nonsmoker - mu_smoker = 0
## HA: mu_nonsmoker - mu_smoker != 0
## Standard error = 0.134
## Test statistic: Z =  2.359
## p-value =  0.0184
```



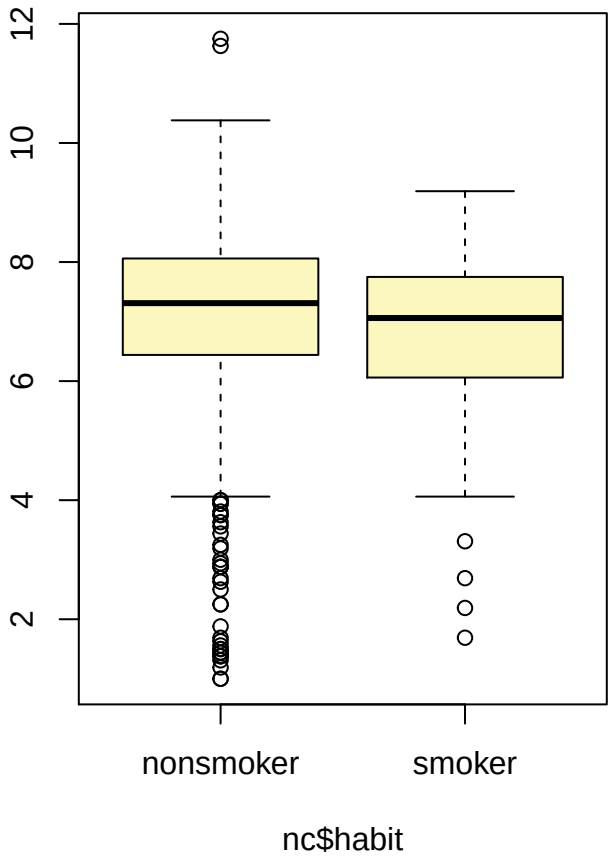
Since the p-value is so low, this gives evidence to reject the null hypothesis which conjectured that the mean weights are the same for babies born to smoking and non-smoking mothers. Thus, we accept the alternative hypothesis, which states whether a mother smokes does impact the average birth weight.

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: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and non-smokers: `nc$habit`. The third argument, `est`, is the parameter we're interested in: `"mean"` (other options are `"median"`, or `"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.

5. Change the `type` argument to `"ci"` to construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

```
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_nonsmoker = 873, mean_nonsmoker = 7.1443, sd_nonsmoker = 1.5187
## n_smoker = 126, mean_smoker = 6.8287, sd_smoker = 1.3862
```

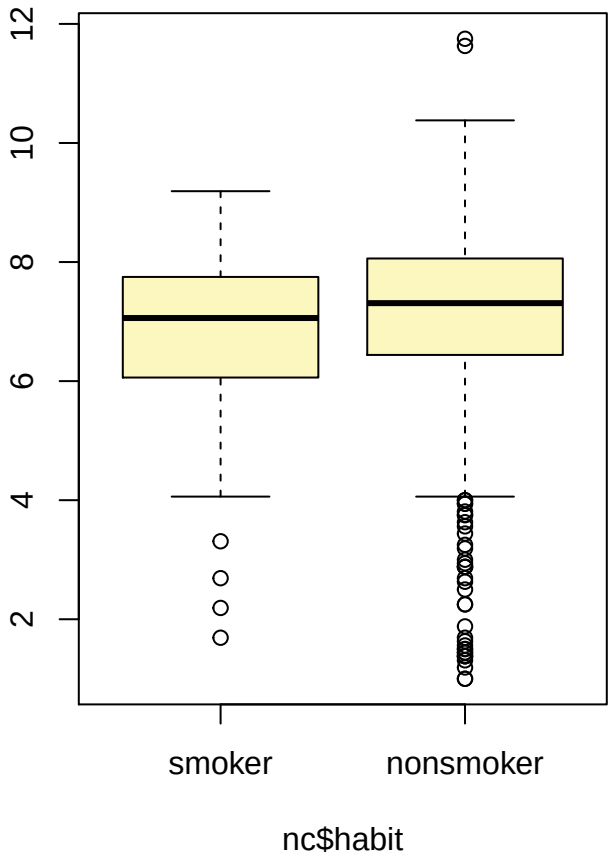



```
## Observed difference between means (nonsmoker-smoker) = 0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( 0.0534 , 0.5777 )
```

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 = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          order = c("smoker","nonsmoker"))
```

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_smoker = 126, mean_smoker = 6.8287, sd_smoker = 1.3862
## n_nonsmoker = 873, mean_nonsmoker = 7.1443, sd_nonsmoker = 1.5187
```




```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

Conclusion

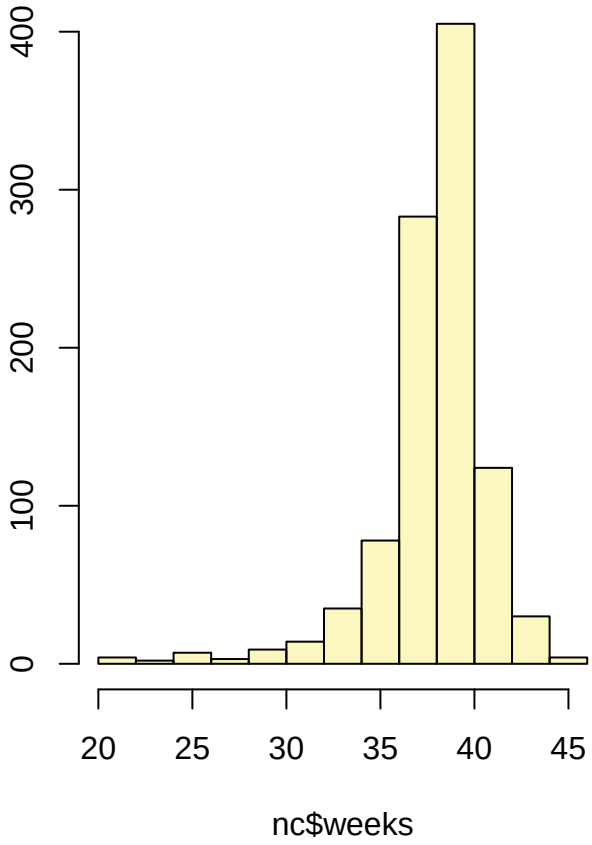
Since the confidence interval does not include the null (i.e., zero), this gives evidence to reject the null hypothesis which conjectured that the mean weights are the same for babies born to smoking and non-smoking mothers. Thus, we accept the alternative hypothesis, which states that whether a mother smokes does impact the average birth weight.

On your own

(1) Calculate a 95% confidence interval for the average length of pregnancies (`weeks`) and interpret it in context. 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.

```
inference(y = nc$weeks,  est = "mean", type = "ci", null = 0,alternative = "twosided", method = "theore
```

```
## Single mean
## Summary statistics:
```




```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 95 % Confidence interval = ( 38.1528 , 38.5165 )
```

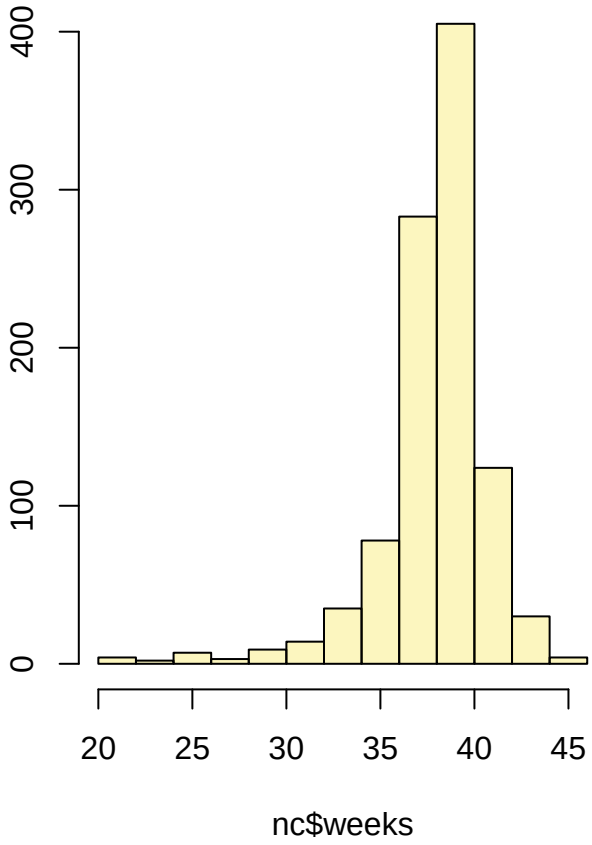
Because this is a random sample, the observations are independent. The sample size (998) is large enough to conduct inference using the normal distribution despite the fact that the distribution is left-skewed.

We are 95% confident that the average length of pregnancies in weeks falls in the interval (38.1528 , 38.5165) .

(2) Calculate a new confidence interval for the same parameter at the 90% confidence level. You can change the confidence level by adding a new argument to the function: `conflevel = 0.90`.

```
inference(y = nc$weeks,  est = "mean", type = "ci", null = 0, conflevel = 0.90,alternative = "twosided"
```

```
## Single mean
## Summary statistics:
```




```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 90 % Confidence interval = ( 38.182 , 38.4873 )
```

We are 90% confident that the average length of pregnancies, in weeks, falls in the interval (38.182 , 38.4873) – which is a narrower interval than that associated with 95% confidence above. .

(3) Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

Null Hypothesis: There is no difference in average weight gained during pregnancy between younger mothers vs. mature mothers.

In statistical notation:

$$H_0: \quad \mu_{younger} = \mu_{mature} \quad \Rightarrow \quad \mu_{younger} - \mu_{mature} = 0$$

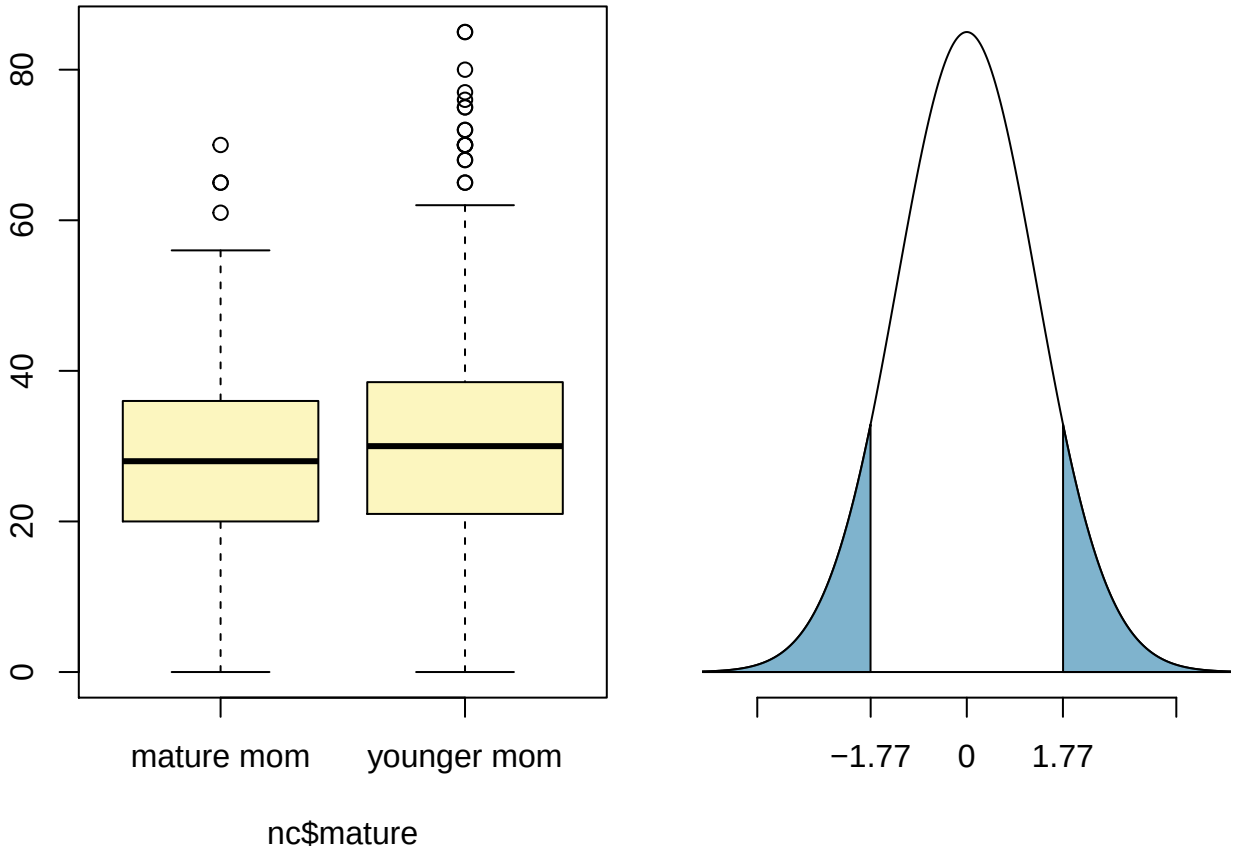
where $\mu_{younger}$ represents the average weight gained during pregnancy by younger mothers and μ_{mature} represents the average weight gained during pregnancy by mature mothers.

Alternative Hypothesis: There is some difference in average weight gained during pregnancy between younger mothers vs. mature mothers:

$$H_0: \quad \mu_{younger} \neq \mu_{mature} \quad \Rightarrow \quad \mu_{younger} - \mu_{mature} \neq 0$$

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

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_mature mom = 129, mean_mature mom = 28.7907, sd_mature mom = 13.4824
## n_younger mom = 844, mean_younger mom = 30.5604, sd_younger mom = 14.3469
## Observed difference between means (mature mom-younger mom) = -1.7697
##
## H0: mu_mature mom - mu_younger mom = 0
## HA: mu_mature mom - mu_younger mom != 0
## Standard error = 1.286
## Test statistic: Z =  -1.376
## p-value =  0.1686
```



Because the p-value is larger than 0.05, we do not have sufficient evidence to reject the null hypothesis, and conclude that younger and mature mothers, on average, experience similar weight gain during pregnancy.

(4) 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.

Examine the variables in the nc data set, subsetting between "younger mom" vs. "mature mom" to see what numeric value determines the age cutoff:

```
### Summary of "younger mom" subset:
nc %>% subset(mature=="younger mom") %>% summary()
```

```
##       fage            mage              mature        weeks
##  Min.   :14.000   Min.   :13.000   mature mom :  0   Min.   :22.000
##  1st Qu.:24.000   1st Qu.:21.000   younger mom:867   1st Qu.:37.000
##  Median :29.000   Median :25.000                     Median :39.000
##  Mean   :28.857   Mean   :25.438                     Mean   :38.382
##  3rd Qu.:33.000   3rd Qu.:30.000                     3rd Qu.:40.000
##  Max.   :48.000   Max.   :34.000                     Max.   :45.000
##  NA's   :160                                         NA's   :1
##        premie        visits             marital        gained
##  full term:737   Min.   : 0.000   married    :361   Min.   : 0.00
##  premie   :129   1st Qu.:10.000   not married:506   1st Qu.:21.00
##  NA's     :  1   Median :12.000                     Median :30.00
##                  Mean   :12.028                     Mean   :30.56
##                  3rd Qu.:15.000                     3rd Qu.:38.25
##                  Max.   :30.000                     Max.   :85.00
##                  NA's   :7                          NA's   :23
##      weight        lowbirthweight    gender          habit
##  Min.   : 1.0000   low    : 93    female:435   nonsmoker:752
##  1st Qu.: 6.3800   not low:774    male  :432   smoker   :115
##  Median : 7.3100
##  Mean   : 7.0972
##  3rd Qu.: 8.0000
##  Max.   :11.7500
##
##       whitemom
##  not white:255
##  white    :611
##  NA's     :  1
##
##
##
##
```

```
### Summary of "mature mom" subset:
nc %>% subset(mature=="mature mom") %>% summary()
```

```
##       fage            mage              mature        weeks
##  Min.   :26.000   Min.   :35.00   mature mom :133   Min.   :20.000
##  1st Qu.:35.000   1st Qu.:35.00   younger mom:  0   1st Qu.:38.000
##  Median :38.000   Median :37.00                     Median :39.000
##  Mean   :38.361   Mean   :37.18                     Mean   :38.023
##  3rd Qu.:41.000   3rd Qu.:38.00                     3rd Qu.:40.000
##  Max.   :55.000   Max.   :50.00                     Max.   :44.000
##  NA's   :11                                         NA's   :1
##        premie        visits             marital        gained
##  full term:109   Min.   : 3.000   married    : 25   Min.   : 0.000
```

```
##  premie   : 23   1st Qu.:10.000   not married:107   1st Qu.:20.000
##  NA's     :  1   Median :12.000   NA's       :  1   Median :28.000
##                  Mean   :12.611                     Mean   :28.791
##                  3rd Qu.:15.000                     3rd Qu.:36.000
##                  Max.   :30.000                     Max.   :70.000
##                  NA's   :2                          NA's   :4
##      weight        lowbirthweight    gender         habit
##  Min.   : 1.3800   low    : 18    female:68   nonsmoker:121
##  1st Qu.: 6.3800   not low:115    male  :65   smoker   : 11
##  Median : 7.3100                             NA's     :  1
##  Mean   : 7.1256
##  3rd Qu.: 8.1900
##  Max.   :10.2500
##
##       whitemom
##  not white: 29
##  white    :103
##  NA's     :  1
##
##
##
##
```

```
### summary of "mage" (mother's age) for "younger mom" subset
summary(nc$mage[nc$mature=="younger mom"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  13.000  21.000  25.000  25.438  30.000  34.000
```

```
maxyounger <- summary(nc$mage[nc$mature=="younger mom"])["Max."]
cat("**Maximum** age for **younger** moms: ", maxyounger, "\n")
```

```
## **Maximum** age for **younger** moms:  34
```

```
### summary of "mage" (mother's age) for "mature mom" subset
summary(nc$mage[nc$mature=="mature mom"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35.00   35.00   37.00   37.18   38.00   50.00
```

```
minmature <- summary(nc$mage[nc$mature=="mature mom"])["Min."]
cat("**Minimum** age for **mature** moms: ", minmature, "\n")
```

```
## **Minimum** age for **mature** moms:  35
```

The distinction is observed when subsetting the dataset into the 133 cases where `mature=="mature mom"` vs. subsetting into the 867 cases where `mature=="younger mom"` . By visual inspection of the summary results, "**mature mom**" is associated with those cases where mother's age (mage) is greater than or equal to **35**, while "**younger mom**" is associated with those cases where mother's age is less than or equal to **34**.

(5) Pick a pair of numerical and categorical variables and 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.

Numerical value: `visits` : number of hospital visits during pregnancy.

Categorical variable: `lowbirthweight` : whether baby was classified as low birthweight (`low`) or not (`not low`).

Question: Is there a relationship between the number of hospital visits made by the mother during pregnancy and whether the baby is born with "low birth weight" or not?

Null Hypothesis: There is no difference between the average number of hospital visits during pregnancy by mothers whose babies were born with low birth weight vs. the average number of hospital visits during pregnancy by mothers whose babies were not born with low birth weight.

In statistical notation:

$$H_0: \quad \mu_{low} = \mu_{notlow} \quad \Rightarrow \quad \mu_{low} - \mu_{notlow} = 0$$

where μ_{low} represents the average number of hospital visits during pregnancy by mothers whose babies **were** born with low birth weight and μ_{notlow} represents the average number of hospital visits during pregnancy by mothers whose babies were **not** born with low birth weight.

Alternative Hypothesis: There is some difference between the average number of hospital visits during pregnancy by mothers whose babies were born with low birth weight vs. the average number of hospital visits during pregnancy by mothers whose babies were not born with low birth weight. :

$$H_0: \quad \mu_{low} \neq \mu_{notlow} \quad \Rightarrow \quad \mu_{low} - \mu_{notlow} \neq 0$$

Summary of the numeric variable:

```
summary(nc$visits)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000  10.000  12.000  12.105  15.000  30.000       9
```

```
sd(nc$visits, na.rm = T)
```

```
## [1] 3.9549337
```

The average number of hospital visits by the mother while pregnant was 12.1, with a median value of 12.

Summary of the categorical variable:

```
summary(nc$lowbirthweight)
```

There were 111 babies for whom their birth weight was categorized as "low", while 889 were "not low."

What is the cutoff weight?

```
### Summary of "low" subset:
nc %>% subset(lowbirthweight=="low") %>% summary()
```

```
##       fage            mage               mature       weeks
##  Min.   :16.000   Min.   :15.000   mature mom :18   Min.   :20.000
##  1st Qu.:24.000   1st Qu.:21.000   younger mom:93   1st Qu.:31.000
##  Median :30.000   Median :27.000                    Median :34.000
##  Mean   :30.309   Mean   :26.964                    Mean   :33.427
##  3rd Qu.:35.000   3rd Qu.:32.000                    3rd Qu.:37.000
##  Max.   :55.000   Max.   :46.000                    Max.   :43.000
##  NA's   :30                                         NA's   :1
##        premie       visits             marital       gained
##  full term:30   Min.   : 0.000   married    :61   Min.   : 0.000
##  premie   :80   1st Qu.: 8.000   not married:49   1st Qu.:15.000
##  NA's     : 1   Median :10.000   NA's       : 1   Median :25.000
##                 Mean   :10.796                    Mean   :26.077
##                 3rd Qu.:14.000                    3rd Qu.:35.000
##                 Max.   :30.000                    Max.   :65.000
##                 NA's   :3                         NA's   :7
##      weight       lowbirthweight    gender         habit         whitemom
##  Min.   :1.0000   low    :111    female:59   nonsmoker:92   not white:43
##  1st Qu.:3.0950   not low:  0    male  :52   smoker   :18   white    :68
##  Median :4.5600                              NA's     : 1
##  Mean   :4.0348
##  3rd Qu.:5.1600
##  Max.   :5.5000
##
```

```
### Summary of "not low" subset:
nc %>% subset(lowbirthweight=="not low") %>% summary()
```

```
##       fage            mage               mature        weeks
##  Min.   :14.00   Min.   :13.000   mature mom :115   Min.   :32.000
##  1st Qu.:25.00   1st Qu.:22.000   younger mom:774   1st Qu.:38.000
##  Median :30.00   Median :27.000                     Median :39.000
##  Mean   :30.25   Mean   :27.004                     Mean   :38.943
##  3rd Qu.:35.00   3rd Qu.:32.000                     3rd Qu.:40.000
##  Max.   :50.00   Max.   :50.000                     Max.   :45.000
##  NA's   :141                                        NA's   :1
##        premie       visits             marital        gained
##  full term:816   Min.   : 0.000   married    :325   Min.   : 0.000
##  premie   : 72   1st Qu.:10.000   not married:564   1st Qu.:22.000
##  NA's     :  1   Median :12.000                     Median :30.000
##                 Mean   :12.265                     Mean   :30.834
##                 3rd Qu.:15.000                     3rd Qu.:39.000
##                 Max.   :30.000                     Max.   :85.000
##                 NA's   :6                          NA's   :20
##      weight        lowbirthweight    gender          habit
##  Min.   : 5.5600   low    :  0    female:444   nonsmoker:781
##  1st Qu.: 6.7500   not low:889    male  :445   smoker   :108
##  Median : 7.4400
##  Mean   : 7.4838
##  3rd Qu.: 8.1300
##  Max.   :11.7500
##
```

```
##       whitemom
##  not white:241
##  white    :646
##  NA's     :  2
##
##
##
##
```

```
### summary of "weight"  for "low" subset
summary(nc$weight[nc$lowbirthweight=="low"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  1.0000  3.0950  4.5600  4.0348  5.1600  5.5000
```

```
maxlow <- summary(nc$weight[nc$lowbirthweight=="low"])["Max."]
cat("**Maximum** weight for **low** babies: ", maxlow, "\n")
```

```
## **Maximum** weight for **low** babies:  5.5
```

```
### summary of "weight"  for "not low" subset
summary(nc$weight[nc$lowbirthweight=="not low"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  5.5600  6.7500  7.4400  7.4838  8.1300 11.7500
```

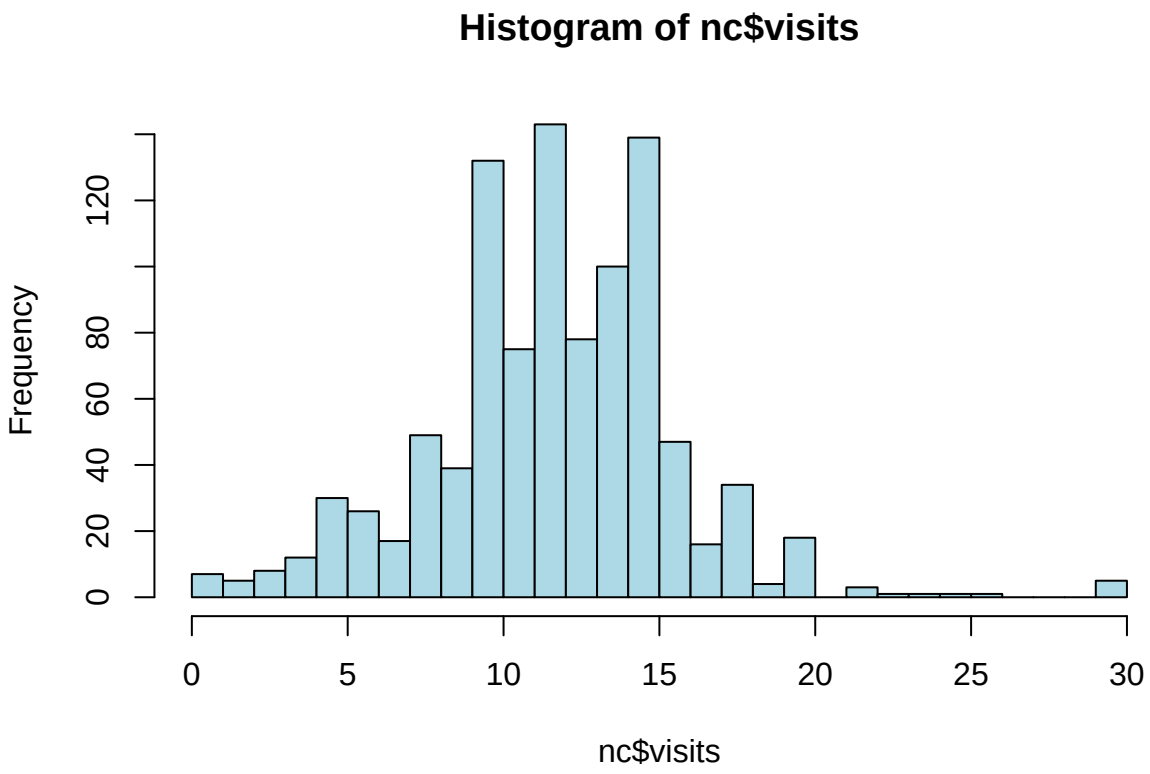
```
minnotlow <- summary(nc$weight[nc$lowbirthweight=="not low"])["Min."]
cat("**Minimum** weight for **not low** babies: ", minnotlow, "\n")
```

```
## **Minimum** weight for **not low** babies:  5.56
```

By visual inspection of the summary results, "low" is associated with those cases where the baby's birth weight is less than or equal to 5.5, while "not low" is associated with those cases where the baby's birth weight is greater than or equal to 5.56 .

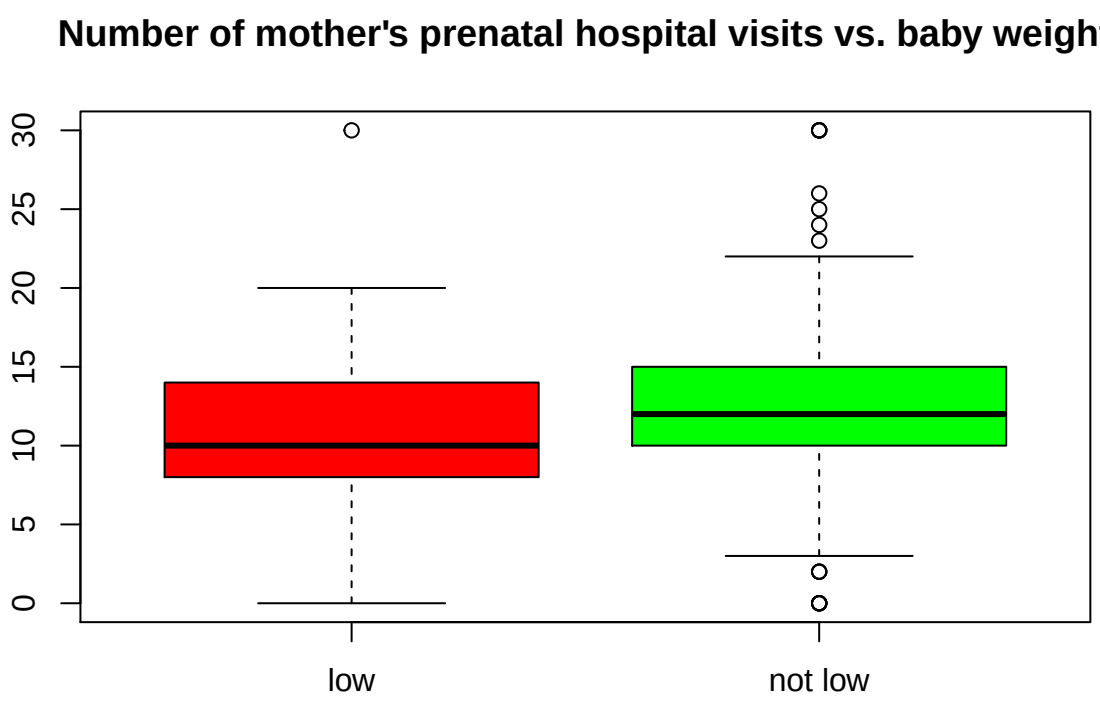
Histogram of number of hospital visits by mother during pregnancy:

```
hist(nc$visits, col="lightblue", breaks=31)
```



Boxplot of number of prenatal hospital visits vs. low or normal birthweight

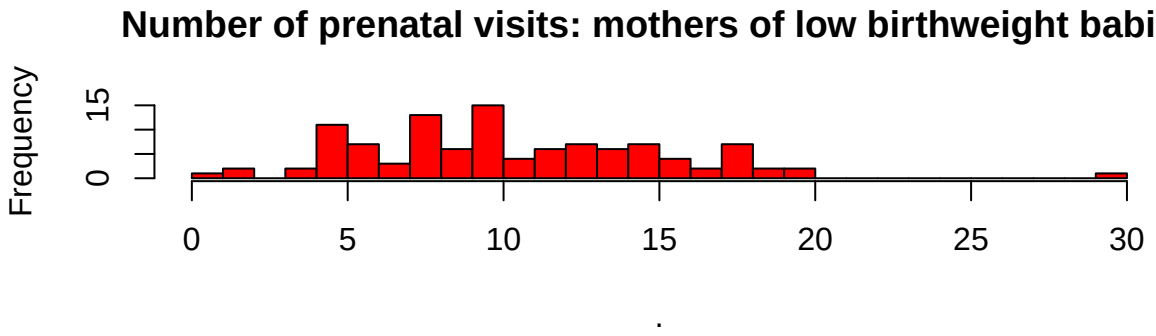
```
boxplot(nc$visits ~ nc$lowbirthweight, main="Number of mother's prenatal hospital visits vs. baby weigh
```



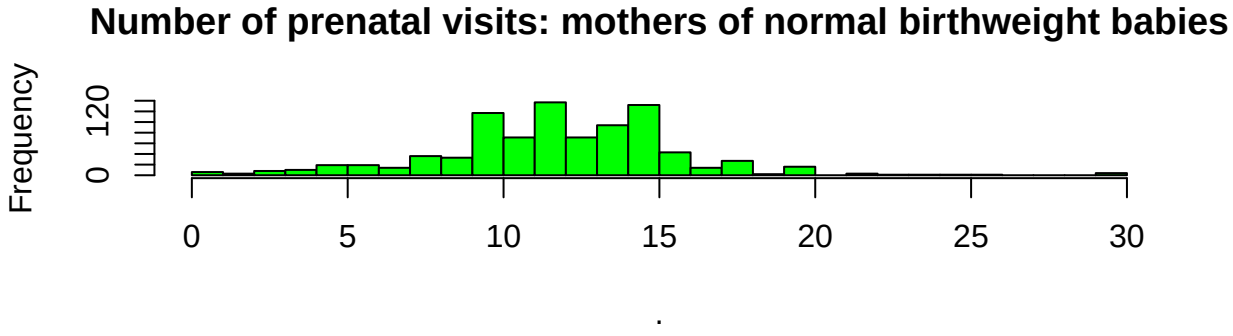
Histograms one-above-the-other:

```
nc %>% drop_na(visits) %>% spread(key = lowbirthweight, value=visits) %>%  select(low,`not low`) -> nc_v

par(mfrow = c(2,1))
nc_visits %>% pull(low) %>% hist(main="Number of prenatal visits: mothers of low birthweight babies", br
nc_visits %>% pull(`not low`) %>% hist(main="Number of prenatal visits: mothers of normal birthweight ba
```



.

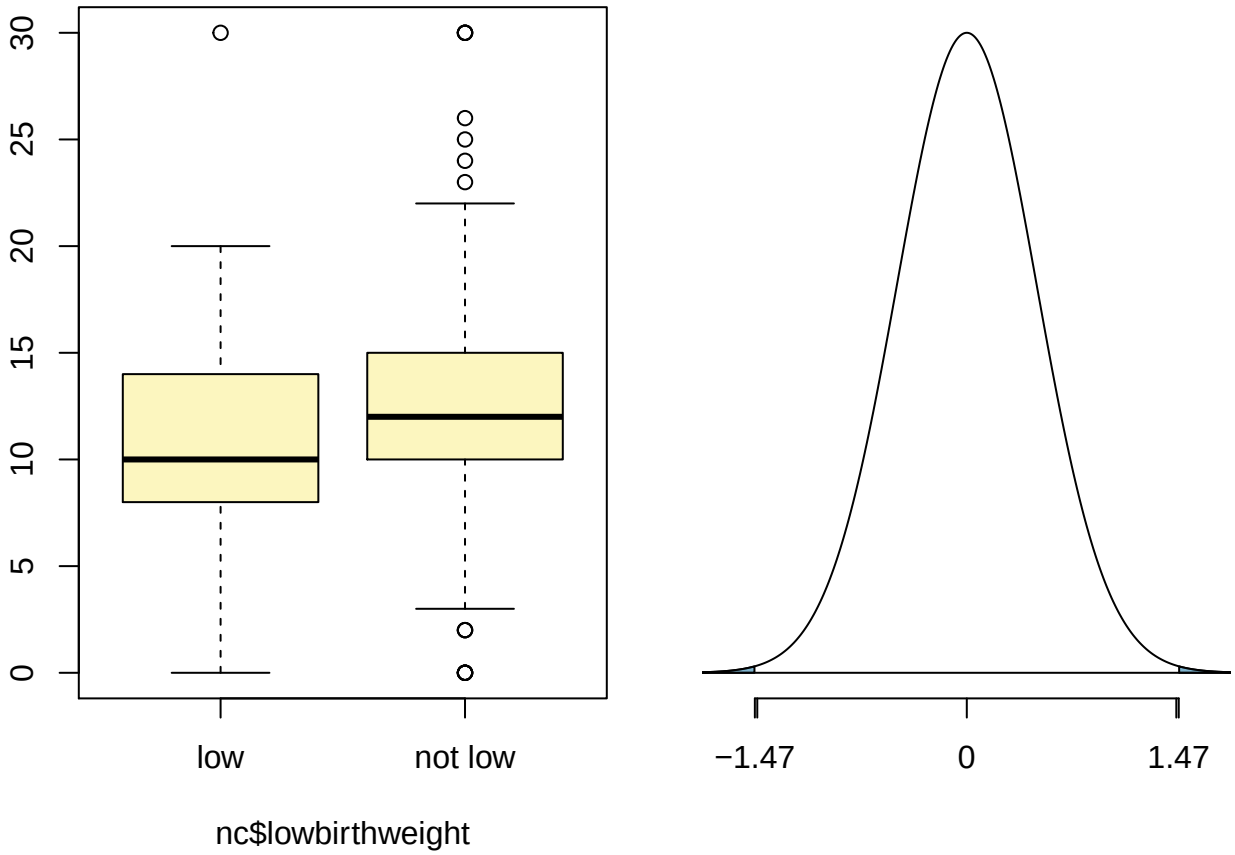


.

The number of prenatal visits by each mother are independent, as each case represents a different pregnancy. The above plots are somewhat normal (they may more closely resemble a normal distribution if the plots showed fewer, wider bands.) The sample sizes are large enough that we can perform inference assuming a normal distribution, and disregarding the question of skew.

```
inference(y = nc$visits, x = nc$lowbirthweight, est = "mean", type = "ht", null = 0,
          alternative = "twosided", method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_low = 108, mean_low = 10.7963, sd_low = 4.8506
## n_not low = 883, mean_not low = 12.265, sd_not low = 3.8036
## Observed difference between means (low-not low) = -1.4687
##
## H0: mu_low - mu_not low = 0
## HA: mu_low - mu_not low != 0
## Standard error = 0.484
## Test statistic: Z =  -3.035
## p-value =  0.0024
```



The mean number of prenatal visits made by mothers of low-weight babies was 10.7963, while the mean number of such visits made by mothers of babies for which the weight was `not low` was 12.265 .

The observed difference in the average number of prenatal visits, -1.4687 , reflects a Z-score of -3.035 , and yields an extremely small p-value (.0024) under the hypothesis test.

This leads us to reject the null hypothesis, which stated that there is no difference between the number of prenatal hospital visits made by mothers of babies born with low birth weight vs. the number of visits made by mothers of babies who have a greater birth weight.

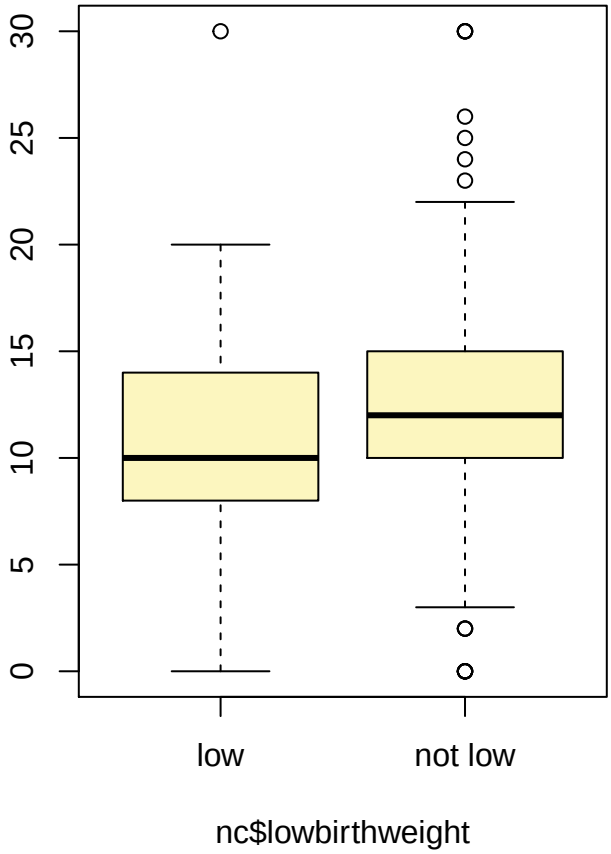
Instead we accept the alternative, which states that there is a significant difference between the number of prenatal hospital visits made by mothers of babies born with low birth weight vs. the number of visits made by mothers of babies who have a greater birth weight.

Check for 95% confidence interval of difference of means:

```
inference(y = nc$visits, x = nc$lowbirthweight, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
```

```
## n_low = 108, mean_low = 10.7963, sd_low = 4.8506
## n_not low = 883, mean_not low = 12.265, sd_not low = 3.8036
```




```
## Observed difference between means (low-not low) = -1.4687
##
## Standard error = 0.484
## 95 % Confidence interval = ( -2.4173 , -0.5201 )
```

With 95% confidence we infer that difference between the average number of prenatal hospital visits made by mothers of low-weight babies vs. the average number of prenatal hospital visits made by mothers of babies with greater birth weight falls in the range (-2.4173 , -0.5201) .

Conclusion

We observe that mothers who give birth to low-weight babies (5.5lbs or less) make, on average, fewer pre-natal hospital visits than do mothers whose babies are born with greater weight.

Of course, this does not imply causation, i.e., we cannot infer that children are born with lower birth weight because their mothers did not make sufficient prenatal hospital visits.

Indeed, there are a number of confounding effects which can be observed from the above summary tables.

In particular, for low-weight babies, average length of the pregnancy was only 33.4 weeks, while the average length was 38.9 weeks for babies of normal weight. Additionally, the low-weight baby set includes a majority of the number of "premies", i.e., prematurely-delivered infants. Although these variables were not rigorously examined here, the smaller number of prenatal hospital visits by the mother in the case of low-weight babies might be explained by the fact that such babies tend to arrive prematurely, which means there is a shorter time period over which such hospital visits could occur.