Inference for numerical data

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.
`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?

```
dim(nc)[1]
```

```
## [1] 1000
```

1K observations

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

```r
summary(nc)
```

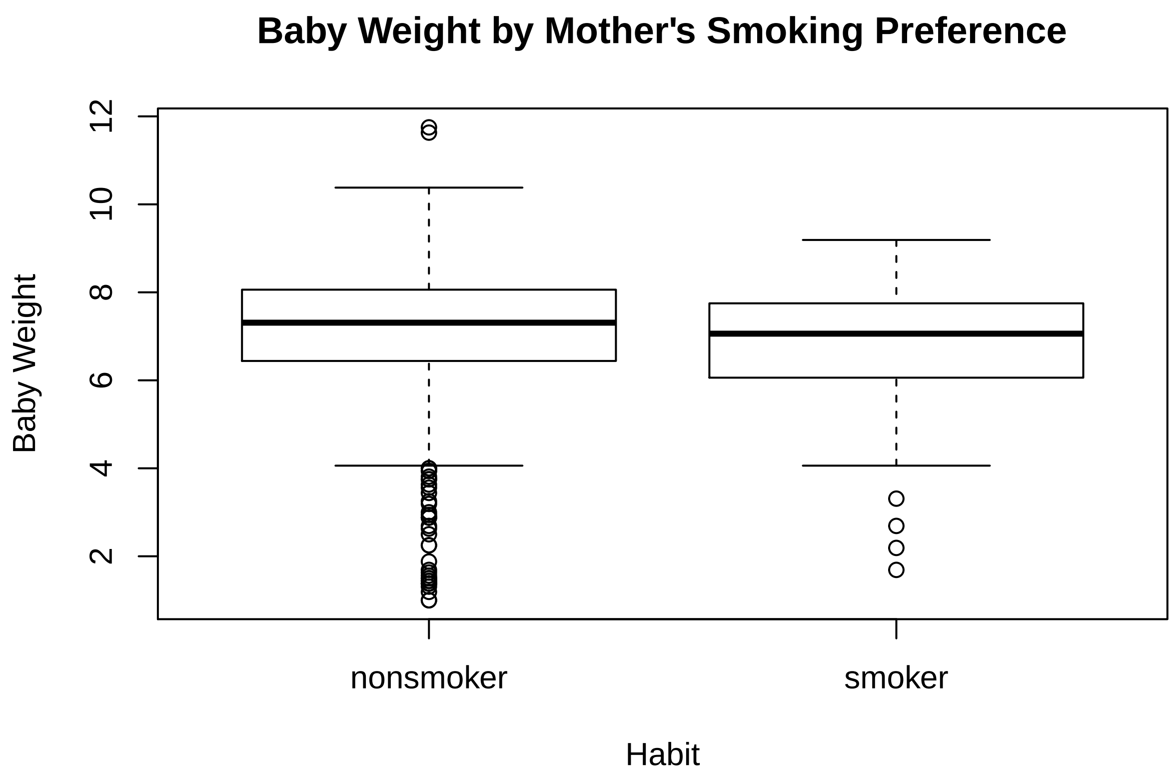
```
##       fage            mage            mature        weeks             premie    
##  Min.   :14.00   Min.   :13   mature mom :133   Min.   :20.00   full term:846  
##  1st Qu.:25.00   1st Qu.:22   younger mom:867   1st Qu.:37.00   premie   :152  
##  Median :30.00   Median :27                     Median :39.00   NA's     :  2  
##  Mean   :30.26   Mean   :27                     Mean   :38.33                  
##  3rd Qu.:35.00   3rd Qu.:32                     3rd Qu.:40.00                  
##  Max.   :55.00   Max.   :50                     Max.   :45.00                  
##  NA's   :171                                    NA's   :2                      
##      visits            marital        gained          weight      
##  Min.   : 0.0   married    :386   Min.   : 0.00   Min.   : 1.000  
##  1st Qu.:10.0   not married:613   1st Qu.:20.00   1st Qu.: 6.380  
##  Median :12.0   NA's       :  1   Median :30.00   Median : 7.310  
##  Mean   :12.1                     Mean   :30.33   Mean   : 7.101  
##  3rd Qu.:15.0                     3rd Qu.:38.00   3rd Qu.: 8.060  
##  Max.   :30.0                     Max.   :85.00   Max.   :11.750  
##  NA's   :9                        NA's   :27                      
##  lowbirthweight    gender          habit          whitemom  
##  low    :111    female:503   nonsmoker:873   not white:284  
##  not low:889    male  :497   smoker   :126   white    :714  
##                              NA's     :  1   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?

```r
boxplot(data= nc,weight~habit, main="Baby Weight by Mother's Smoking Preference",
   xlab="Habit", ylab="Baby 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 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$habit: nonsmoker
## [1] 7.144273
## --------------------------------------------------------------
## nc$habit: smoker
## [1] 6.82873
```

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`.

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

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

We have a big enough sample to assume normal data and samples were taken randomly. Green light for inference.

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

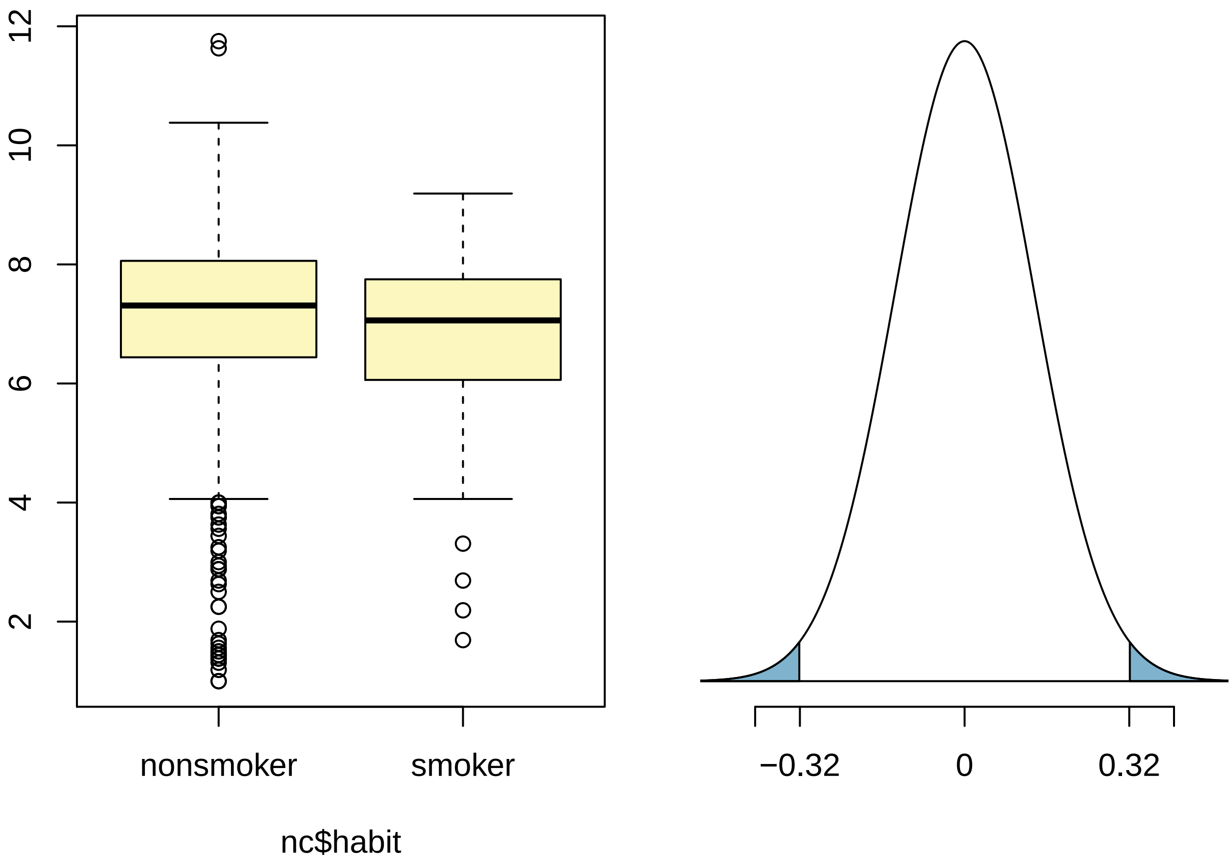
H0: There is no difference in weight between babies for smokers and that of non-smokers. HA: There is a difference in weight of babies for smokers and non-smokers.

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
```



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`.

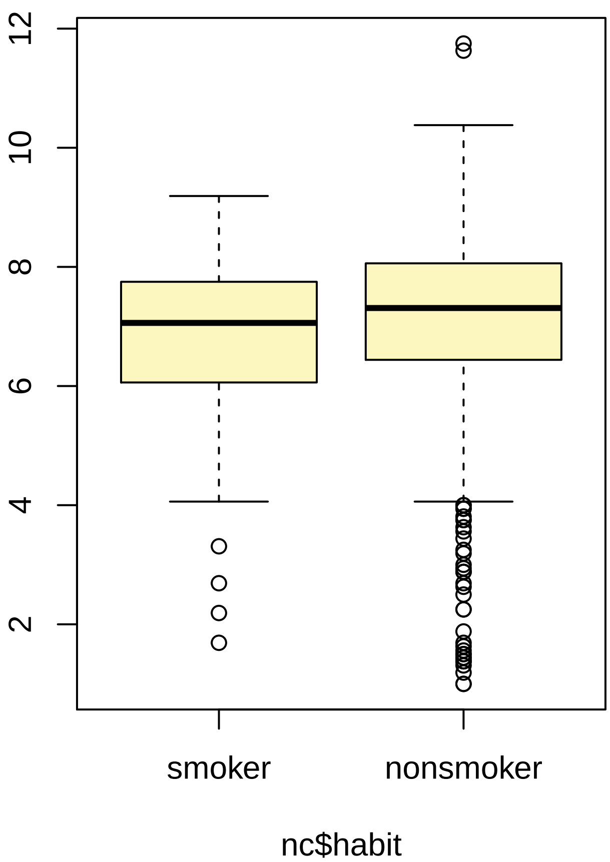
The third argument, `est`, is the parameter we're interested in: `"mean"` (other options are `"median"`, or `"proportion"`.) Next we decide on the `type` of inference we want: a hypothesis test (`"ht"`) or a confidence interval (`"ci"`). When performing a hypothesis test, we also need to supply the `null` 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.

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
```



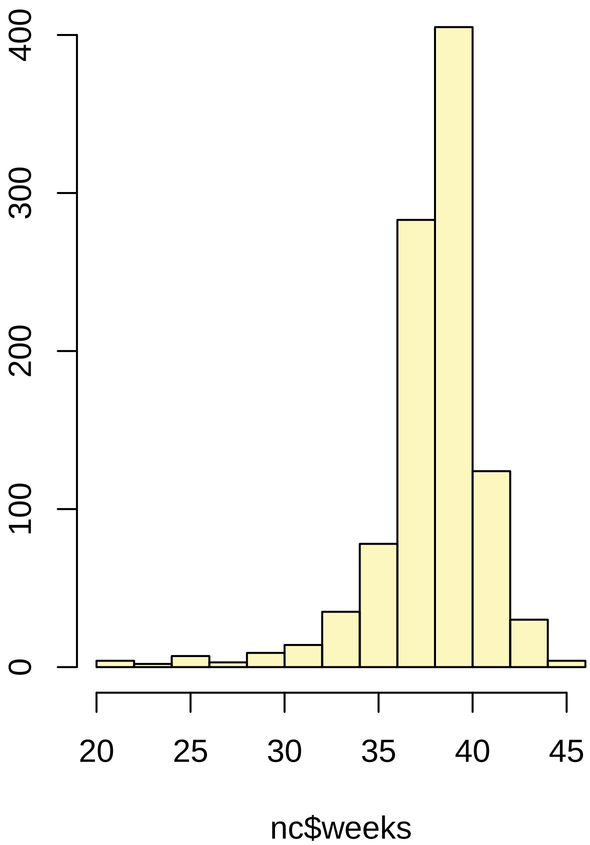

```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

On your own

- 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 = "theoretical")
```

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

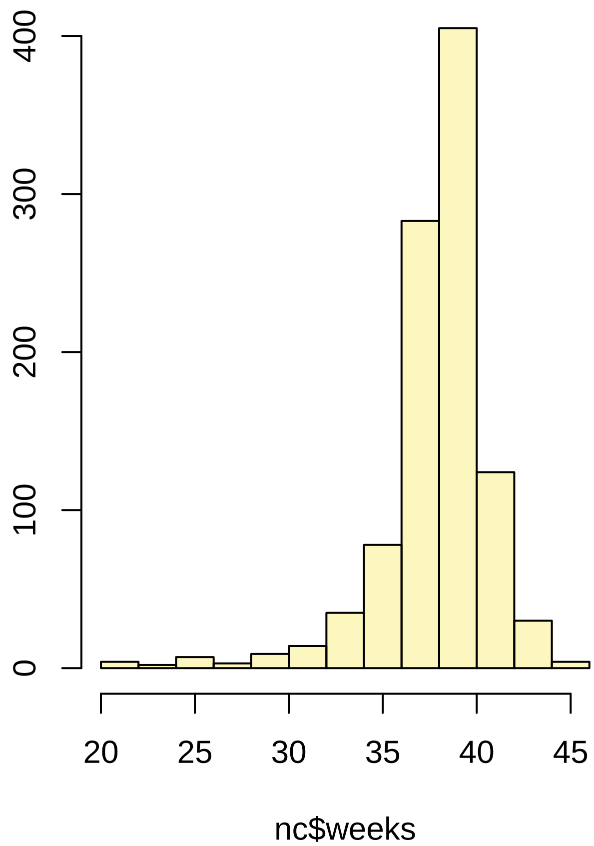



```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 95 % Confidence interval = ( 38.1528 , 38.5165 )
```

- 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,
          alternative = "twosided", method = "theoretical", conflevel=0.90)
```

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




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

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

```
by(nc$gained,nc$mature, length)
```

```
## nc$mature: mature mom
## [1] 133
## ------------------------------------------------------------
## nc$mature: younger mom
## [1] 867
```

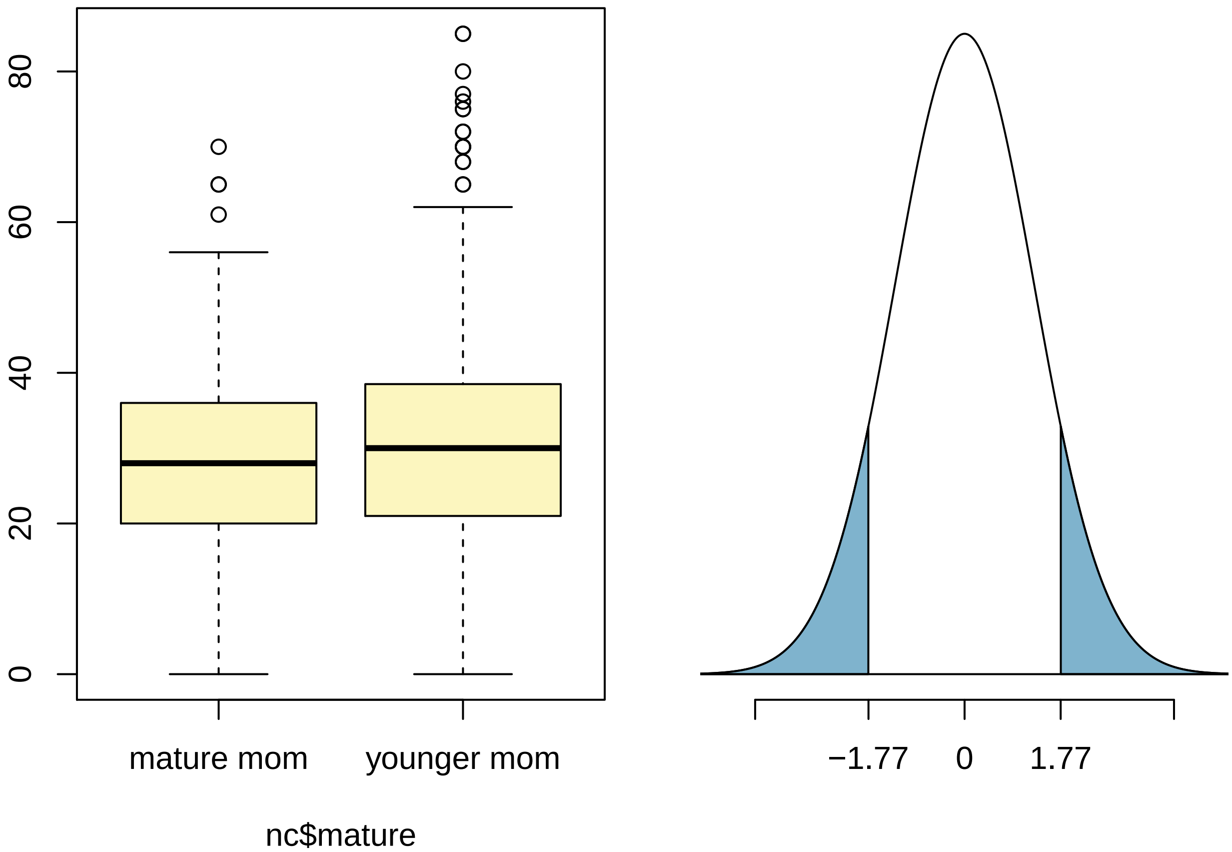
Got a good sample size and the otther independence assumptions are still true for this dataset. H0: There is no difference in gained weight for younger moms compared to older moms. $\mu_{diff} = 0$. HA: There is a difference in gained weight for younder moms compared to older moms. $\mu_{diff} \neq 0$.

```
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
```



For a critical value of 0.05 we can not reject the null hypothesis, there is no difference between mean mother weight gain and younger mother weight gain.

- 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.

```r
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
```

Age appears to be a good cutoff value since there in no overlap between mature and younger after an age of 34.

- 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.

I'll look at the number of hospital visits in terms of baby birth weight. H0: There is no difference between the number of hospital visits and the birth weight of the baby (low or not low). HA: There is a difference between the number of visits and the birth weight of the baby.

```
by(nc$visits,nc$lowbirthweight, length)
```

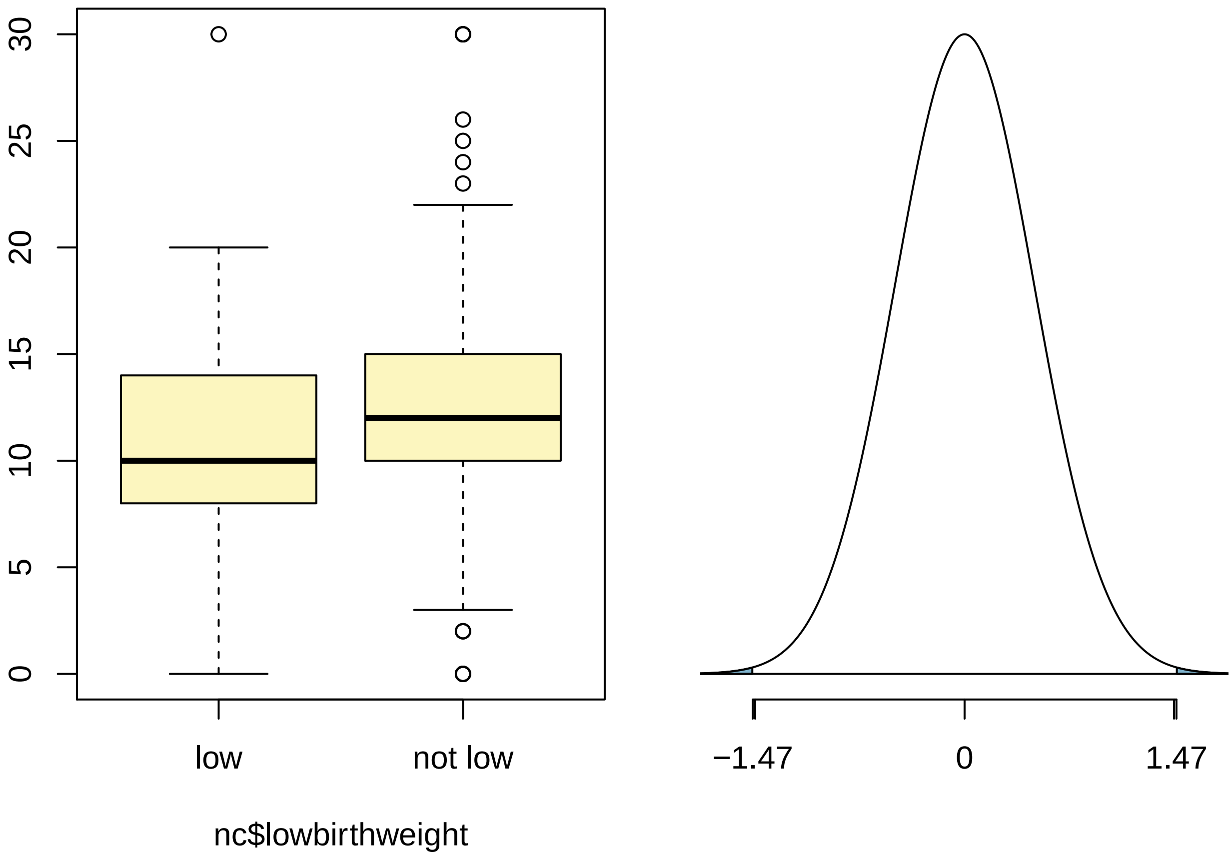
```
## nc$lowbirthweight: low
## [1] 111
## ------------------------------------------------------------
## nc$lowbirthweight: not low
## [1] 889
```

Good sample size for each to infer. Other assumptions still hold.

```
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
```



With a critical value of 0.05, and a p-value of 0.0024 we can reject the null hypothesis. There is a difference in normal birth weight and the number of hospital visits.