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.

```r
library(ggplot2)
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?

The cases are related to information on births. There are 1000 cases in the data set

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

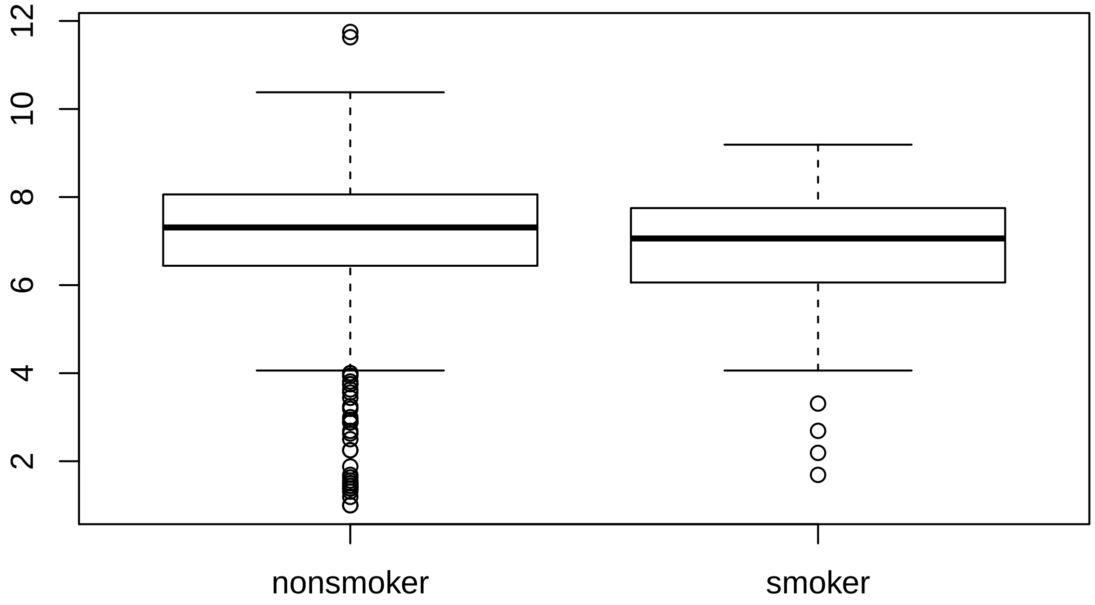
```r
summary(nc)
```

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)
```



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)
```

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

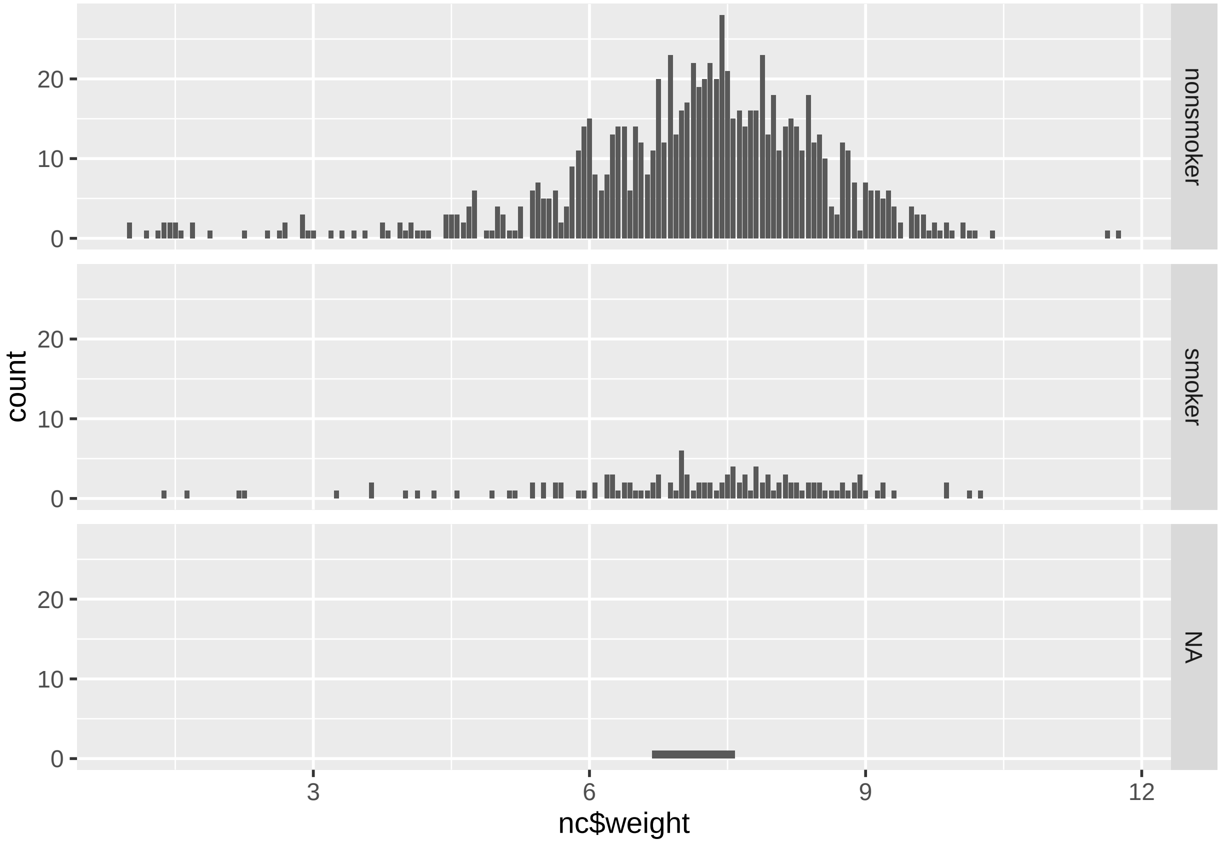
Inference

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`.

The number of cases in the sample is 873, 126

Below is the distribution for the weight by habit

```r
ggplot(data=nc) + geom_bar(aes(nc$weight)) + facet_grid(habit ~ .)
```



As we can see from the histogram above. The distribution seems to have a slight left skew. Given that we have a large enough sample size, we can assume normality for the distribution.

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

H_0: There is no difference between mean weights of smokers and non-smokers.

H_A: There is a difference between mean weights of smokers and non-smokers.

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

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

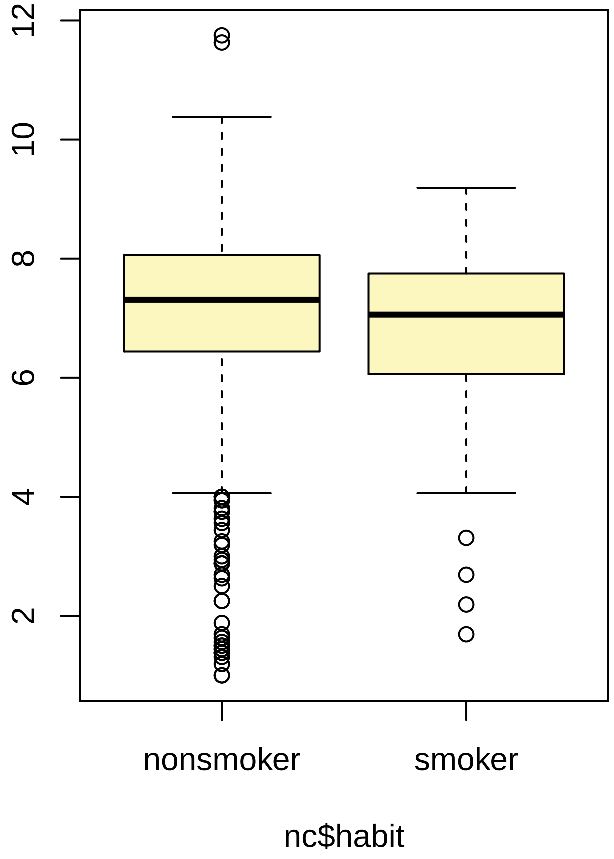
Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and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

```
## Warning: package 'BHH2' was built under R version 3.2.4
```

```
## 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:

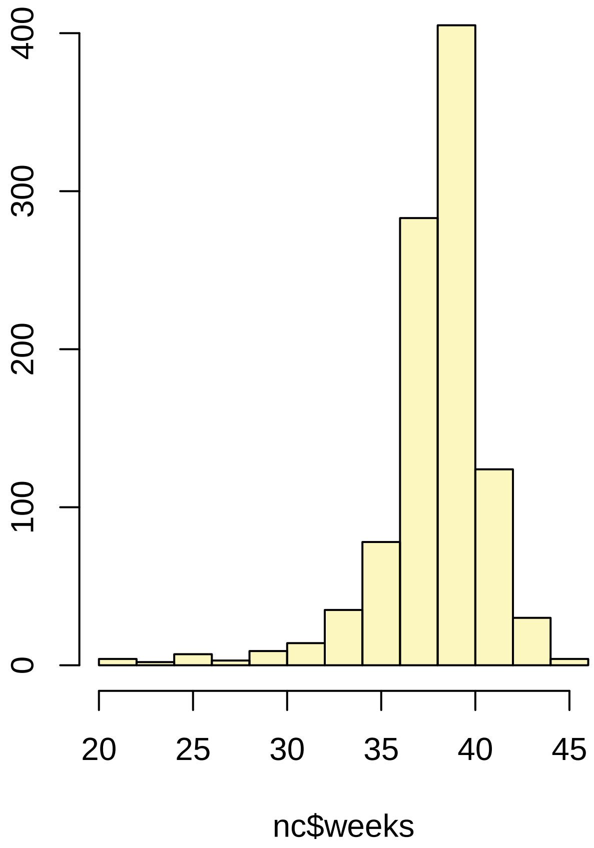
```r
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          order = c("smoker","nonsmoker"))
```

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.

```r
inference(y = nc$weeks, est = "mean", type = "ci", method = "theoretical")
```

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

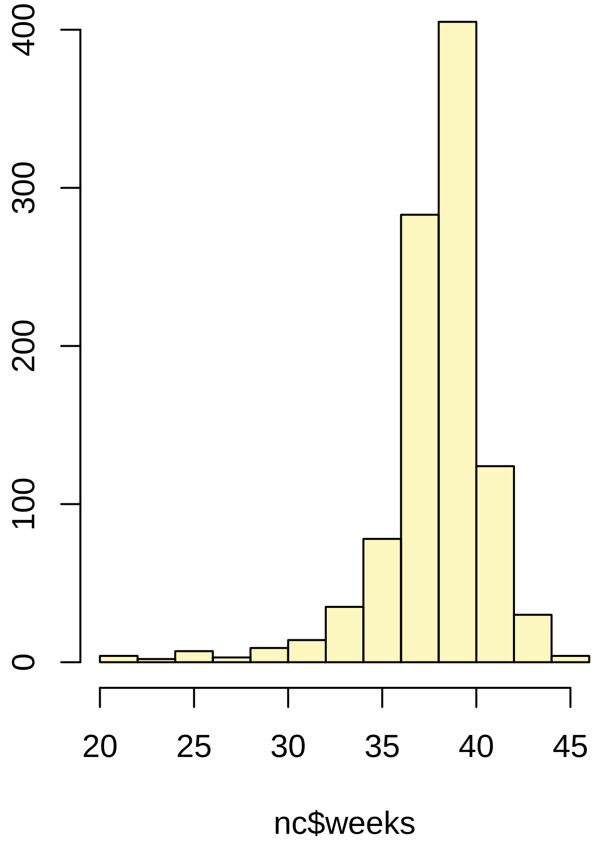


```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 95 % Confidence interval = ( 38.1528 , 38.5165 )
```

- 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`.

```r
inference(y = nc$weeks, est = "mean", type = "ci", 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 )
```

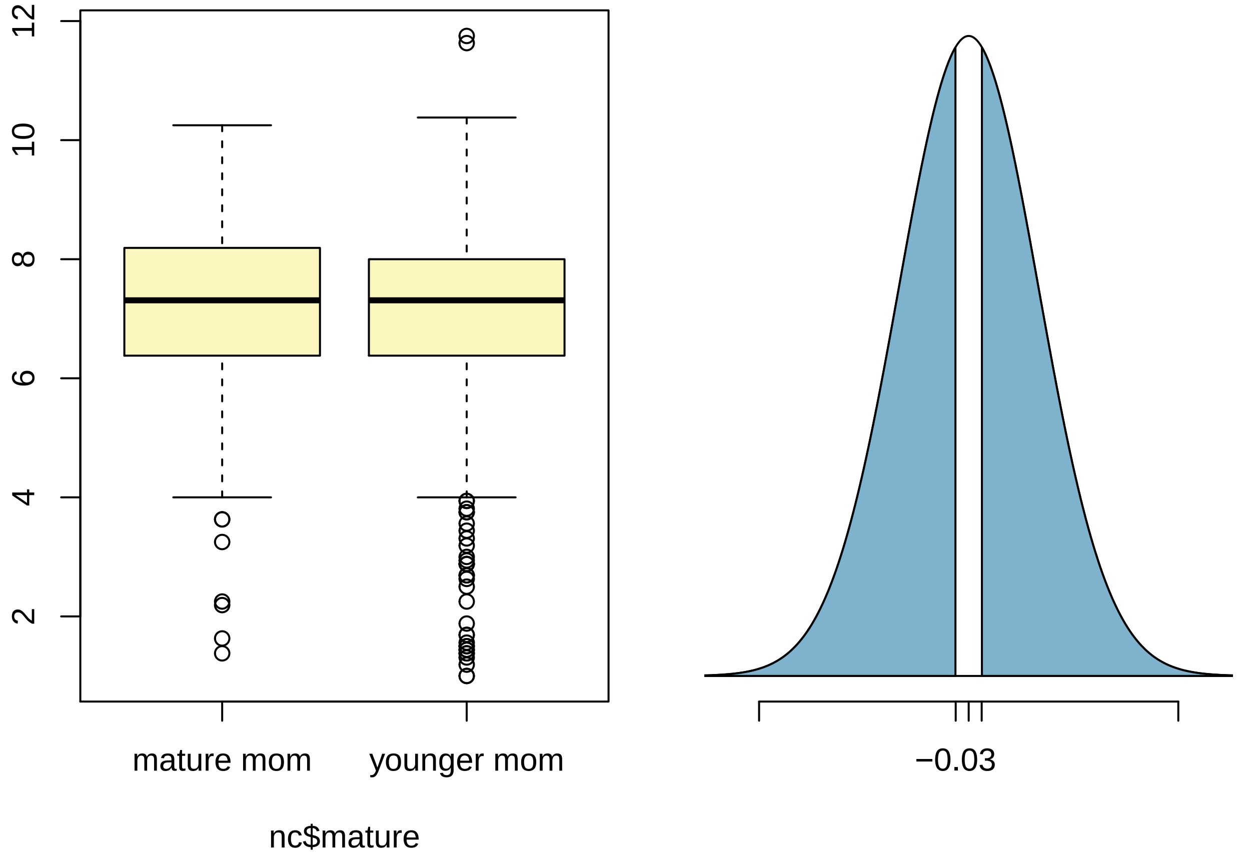
- Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

```r
inference(y = nc$weight, x = nc$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 = 133, mean_mature mom = 7.1256, sd_mature mom = 1.6591
## n_younger mom = 867, mean_younger mom = 7.0972, sd_younger mom = 1.4855

## Observed difference between means (mature mom-younger mom) = 0.0283
##
```

```
## H0: mu_mature mom - mu_younger mom = 0
## HA: mu_mature mom - mu_younger mom != 0
## Standard error = 0.152
## Test statistic: Z =  0.186
## p-value =  0.8526
```

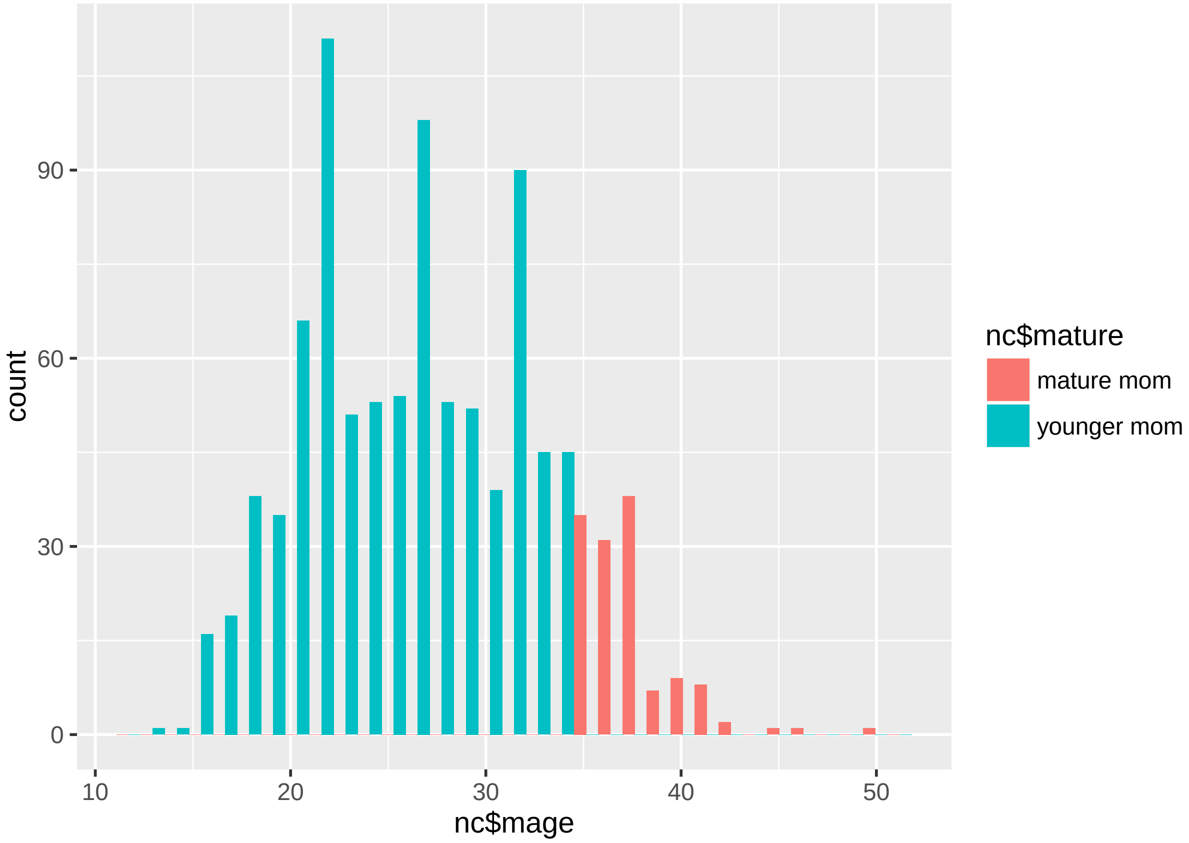


The p-value is 0.8526. This is high and hence we fail to reject the null hypothesis. Hence we can conclude that the mother's maturity has no effect on weight of the babies.

- 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.

```
ggplot(nc, aes(nc$mage, fill = nc$mature)) + geom_histogram(position = "dodge")
```

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



We can see from the above plot that there is a clear cutoff in ages between the mature and younger mom. Finding the max age for the younger mom (or the min age for the mature mom) will give us the cutoff value which is 34/35

```
max(nc[nc$mature=="younger mom", c("mage")])
```

```
## [1] 34
```

```
min(nc[nc$mature=="mature mom", c("mage")])
```

```
## [1] 35
```

- 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.

Does the gender of the baby have an effect on the weight gained by the mother

Lets formulate the below hypothesis to evaluate this relationship:

H_0: There is no difference between mean weights gained by mothers of male and female babies.

H_A: There is a difference between mean weights gained by mothers of male and female babies.

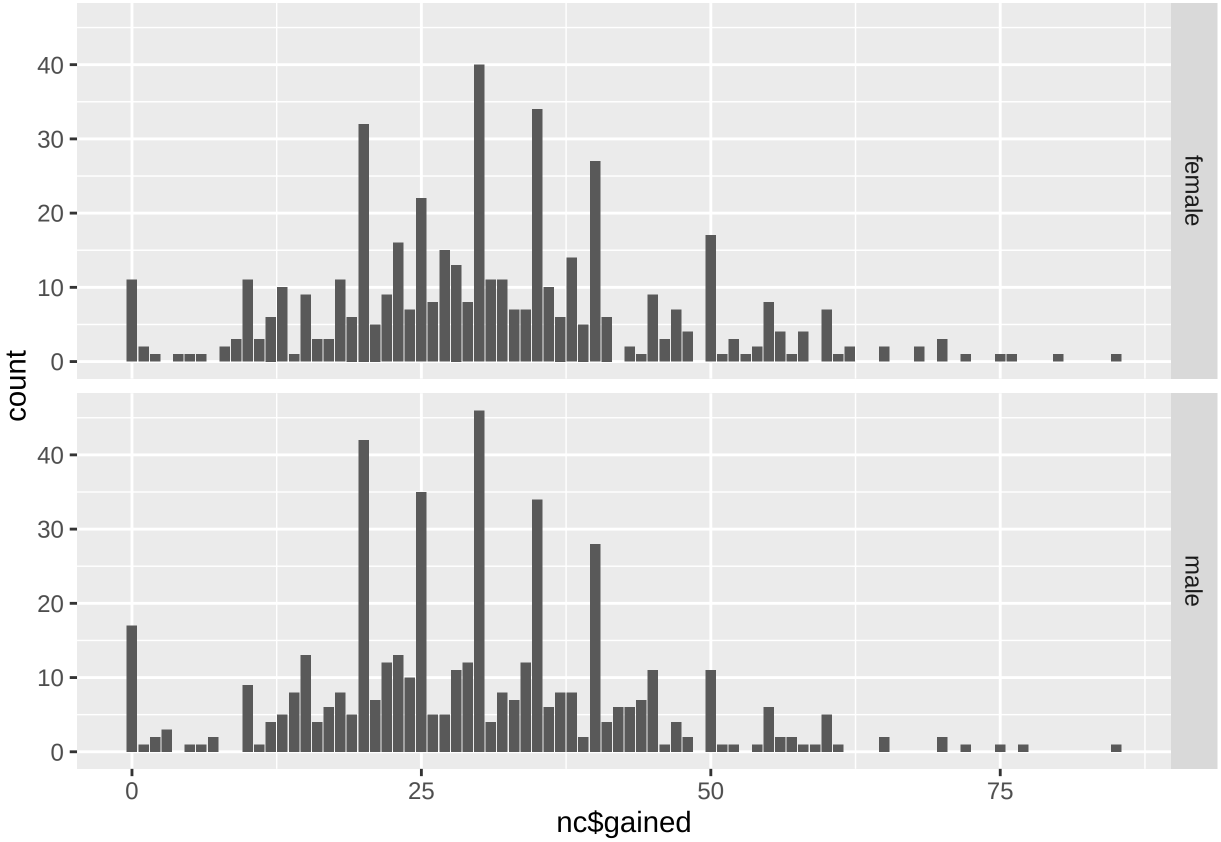
Lets next check for conditions of normality

The number of cases in the sample is 503, 497

Below is the distribution for the weight by habit

```
ggplot(data=nc) + geom_bar(aes(nc$gained)) + facet_grid(gender ~ .)
```

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

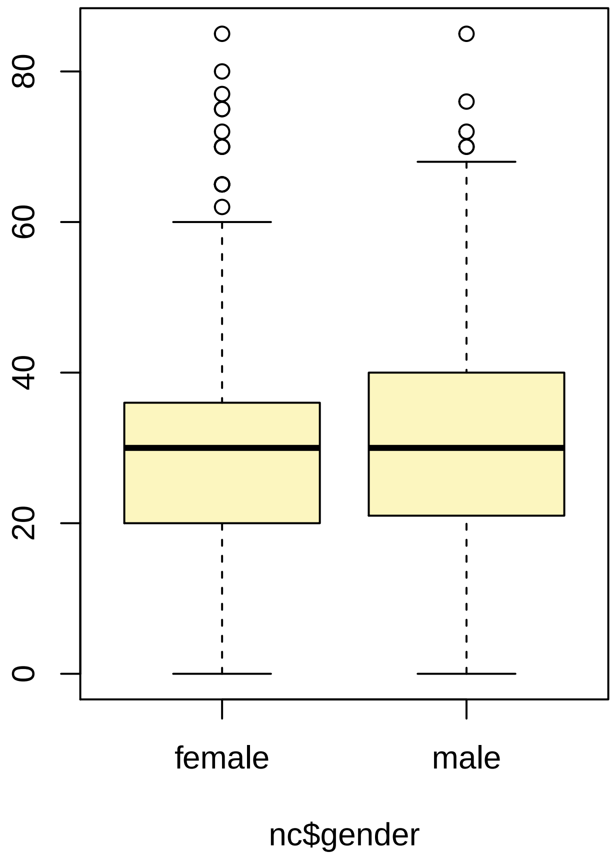


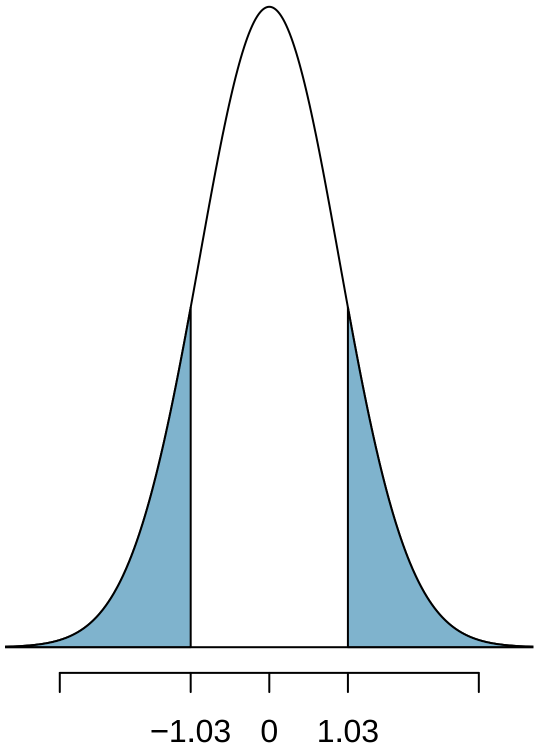
As we can see from the histograms above. The distribution seems to have a slight right skew. Given that we have a large enough sample size, we can assume normality for the distribution.

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

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_female = 488, mean_female = 29.8135, sd_female = 14.2506
## n_male = 485, mean_male = 30.8412, sd_male = 14.228

## Observed difference between means (female-male) = -1.0277
##
## H0: mu_female - mu_male = 0
## HA: mu_female - mu_male != 0
## Standard error = 0.913
## Test statistic: Z =  -1.126
## p-value =  0.2604
```





The p-value is 0.2604. This is high and hence we fail to reject the null hypothesis. Hence we can conclude that the weight gained by the mother has no effect on gender of the babies.