Inference for numerical data

Jimmy Ng

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? # JN: there are 1000 cases and 13 columns in this sample, i.e. dim(nc) [1] 1000 13

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.   :14.00   Min.   :13   mature mom :133   Min.   :20.00  
##  1st Qu.:25.00   1st Qu.:22   younger mom:867   1st Qu.:37.00  
##  Median :30.00   Median :27                     Median :39.00  
##  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      
##        premie        visits            marital        gained     
##  full term:846   Min.   : 0.0   married    :386   Min.   : 0.00  
##  premie   :152   1st Qu.:10.0   not married:613   1st Qu.:20.00  
##  NA's     :  2   Median :12.0   NA's       :  1   Median :30.00  
##                  Mean   :12.1                     Mean   :30.33  
##                  3rd Qu.:15.0                     3rd Qu.:38.00  
##                  Max.   :30.0                     Max.   :85.00  
##                  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?

```r
with(nc, boxplot(weight ~ habit, main = "weight by habit"))
```

JN: although the median of weight is similar between nonsmoker and smoker, nonsmoker has a larger variation within group, and tend to skew stronger to the left. It seems impossible to have weight close to 1 or even below 2 among the nonsmokers' babies. That raises the concern for data quality.

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.

```r
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
```

JN: the conditions are met: large enough sample size (n >=30), independent and random sampling.

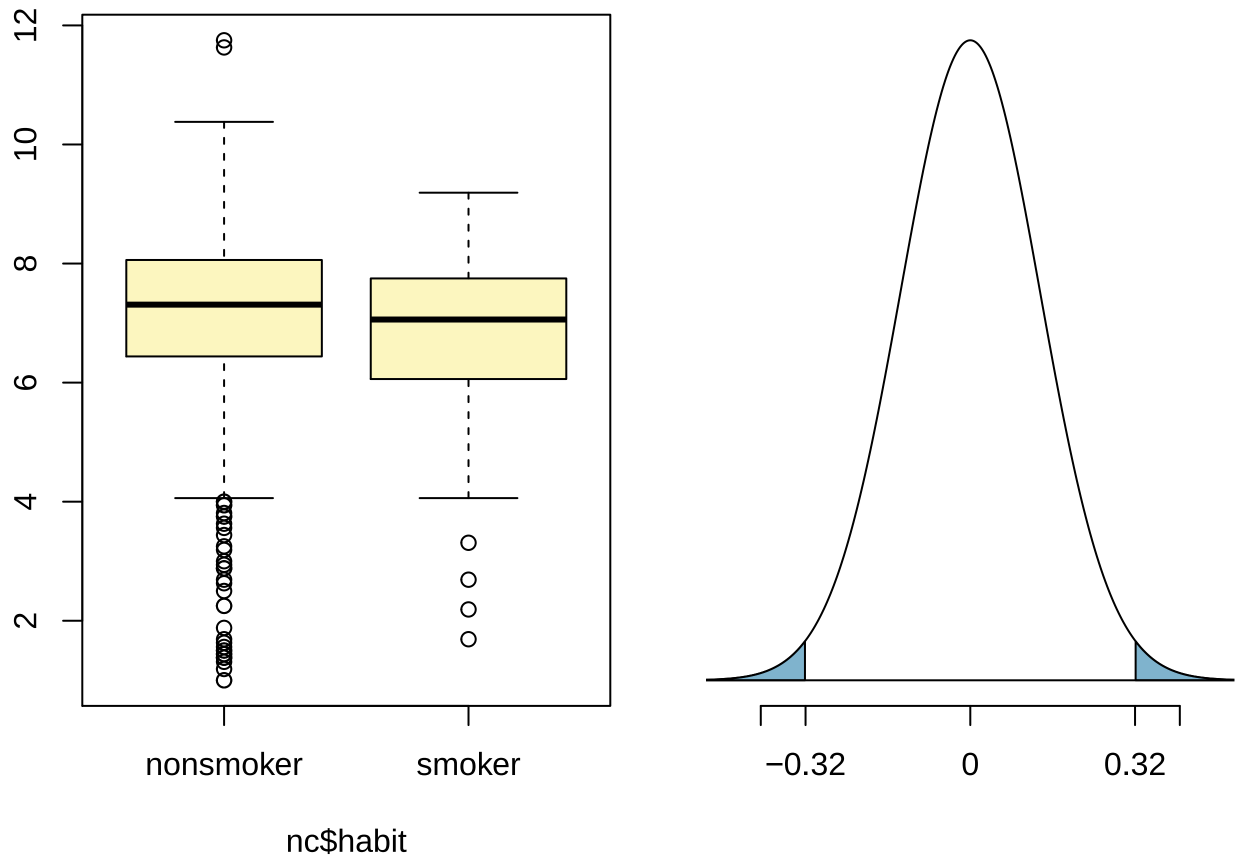
4. Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different. # JN: null hypothesis would be there is no difference between the average weight of babies born to smoking and non-smoking mothers; alternative hypothesis would be there is a difference.

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
```

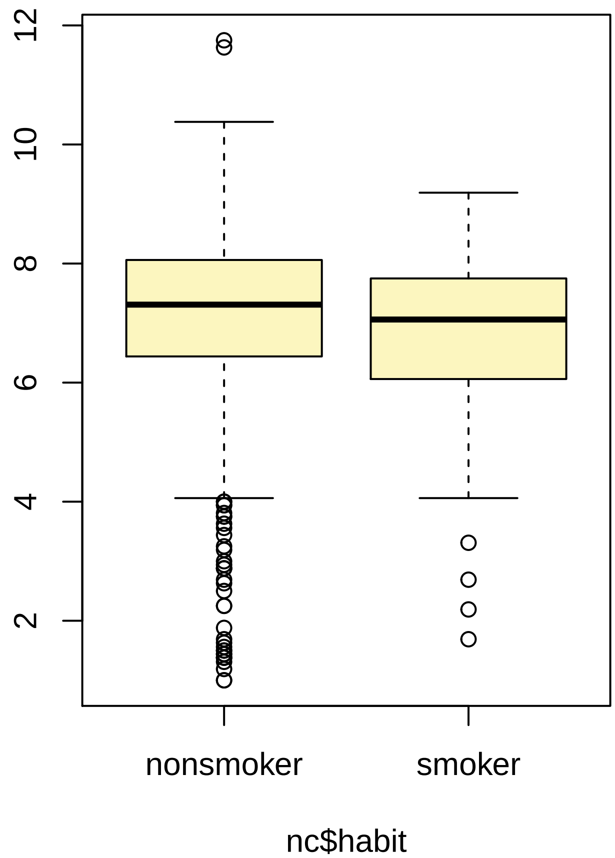


Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and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

```
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
```

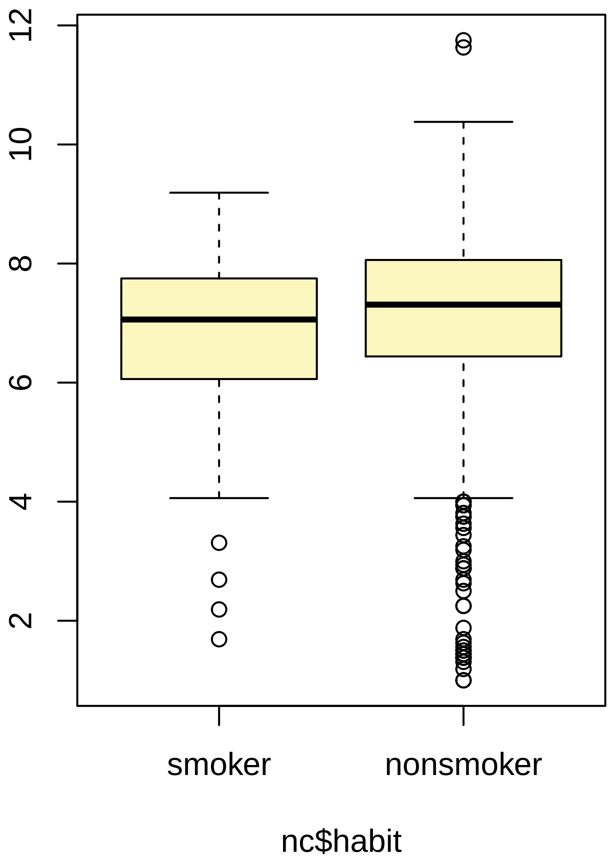



```
## Observed difference between means (nonsmoker-smoker) = 0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( 0.0534 , 0.5777 )
```

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
```



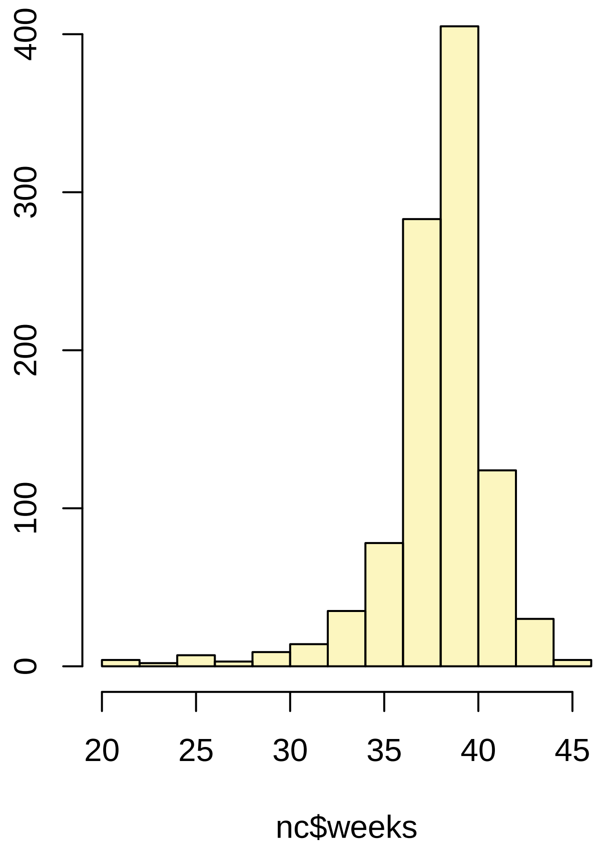

```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

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:
```



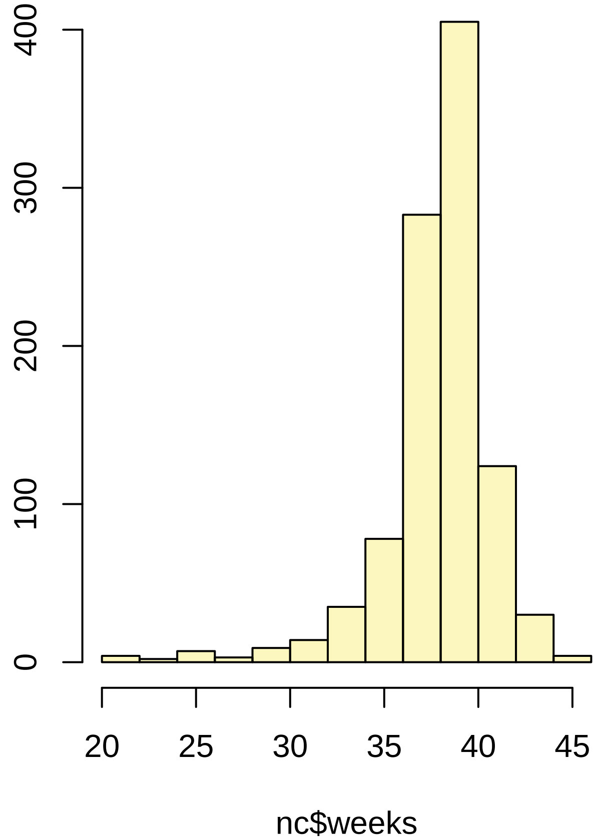

```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 95 % Confidence interval = ( 38.1528 , 38.5165 )
```

JN: the average length of pregnancy is 38.3 weeks with SD equal to 2.9. The 95% confidence interval for the mean is between 38.2 and 38.5. In the other words, 95% of the time we expect that the true mean of the length of pregnancy is between such intervals; there's a 5% chance that we make such type 1 error.

- 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 = 0.9)
```

```
## 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
t.test(gained ~ mature, data = nc)
```

```
##
##  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 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
```

JN: not really, they are not statistically significantly different, i.e. p-value way above .5.

- 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.

```
ftable(nc$mature, nc$mage)
```

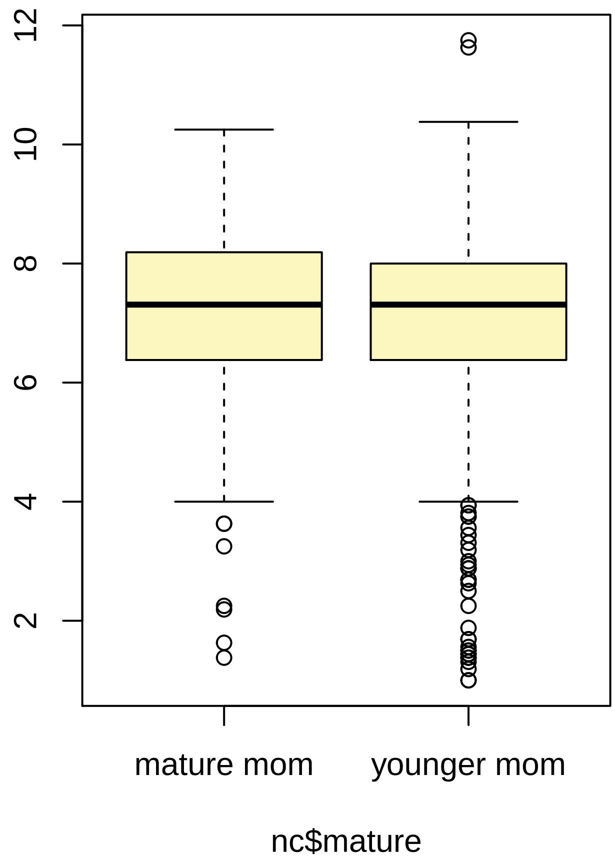
```
##              13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
##
## mature mom    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 35 31 26 12  7  9  8
## younger mom   1  1  6 10 19 38 35 66 51 60 51 53 54 51 47 53 52 39 52 38 45 45  0  0  0  0  0  0  0
```

JN: the cut-off is 35. Women aged 35 or above is considered to be mature. We can look at that from above frequency distribution using the ftable() function.

- 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.

```
inference(y = nc$weight, x = nc$mature, est = "mean", type = "ci", 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
##
## Standard error = 0.1525
## 95 % Confidence interval = ( -0.2705 , 0.3271 )
```

JN: the hypothesis is that women maturity would affect the weight of baby born. The null hypothesis is that there's no difference of baby weight given the age (mature vs younger mom), whereas the alternative hypothesis is that there's a significant difference. Using the inference() function, we can see that there's no statistical evidence in supporting the alternative hypothesis. The 95% confidence interval for the difference between mature mom-younger mom crossed 0, and we can compare the boxplots and see that the two distributions are not really different, except that there are more outliners for younger mom skewing more heavily to the left.