Two-sample t-test

Jyotishka Datta

Updated: 2023-09-24

Birth-weight Data

- Source: https://www.openintro.org/data/index.php?data=births14

- Description: Every year, the US releases to the public a large data set containing information on births recorded in the country. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. This is a random sample of 1,000 cases from the data set released in 2014.

```
birth <- read.csv("https://www.openintro.org/data/csv/births14.csv")
dim(birth)
```

```
## [1] 1000   13
```

- Only consider the complete cases i.e. rows without any missing values.

```
birth <- birth[complete.cases(birth),]
```

Structure

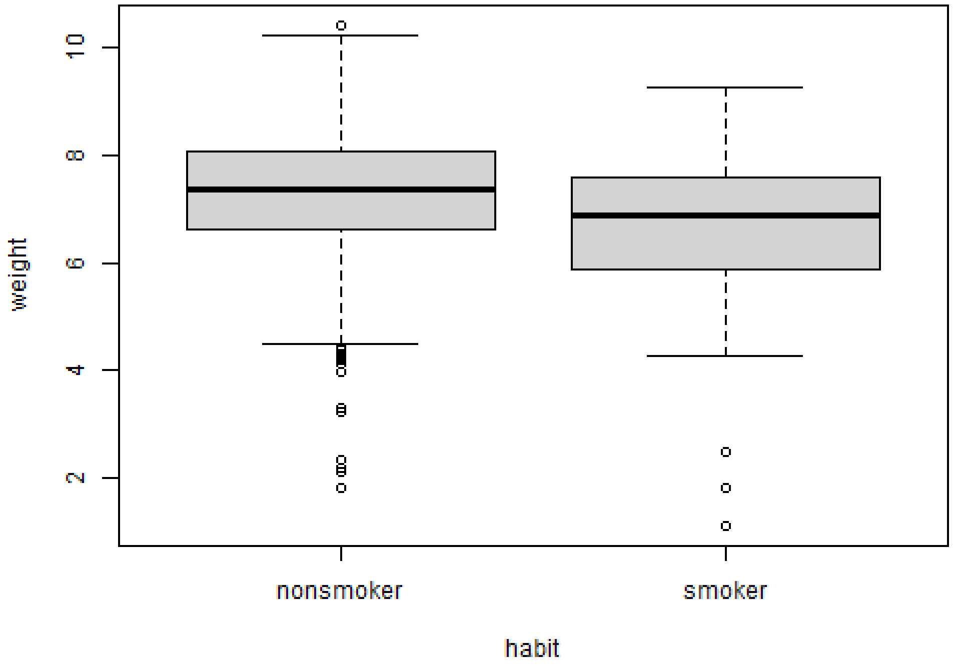
```
str(birth)
```

```
## 'data.frame':    794 obs. of  13 variables:
##  $ fage          : int  34 36 37 32 32 37 29 30 29 30 ...
##  $ mage          : int  34 31 36 31 26 36 24 32 26 34 ...
##  $ mature        : chr  "younger mom" "younger mom" "mature mom" "younger
##  $ weeks         : int  37 41 37 36 39 36 40 39 39 42 ...
##  $ premie        : chr  "full term" "full term" "full term" "premie" ...
##  $ visits        : int  14 12 10 12 14 10 13 15 11 14 ...
##  $ gained        : int  28 41 28 48 45 20 65 25 22 40 ...
##  $ weight        : num  6.96 8.86 7.51 6.75 6.69 6.13 6.74 8.94 9.12 8.91
##  $ lowbirthweight: chr  "not low" "not low" "not low" "not low" ...
##  $ sex           : chr  "male" "female" "female" "female" ...
##  $ habit         : chr  "nonsmoker" "nonsmoker" "nonsmoker" "nonsmoker" ..
##  $ marital       : chr  "married" "married" "married" "married" ...
##  $ whitemom      : chr  "white" "white" "not white" "white" ...
```

Hypothesis

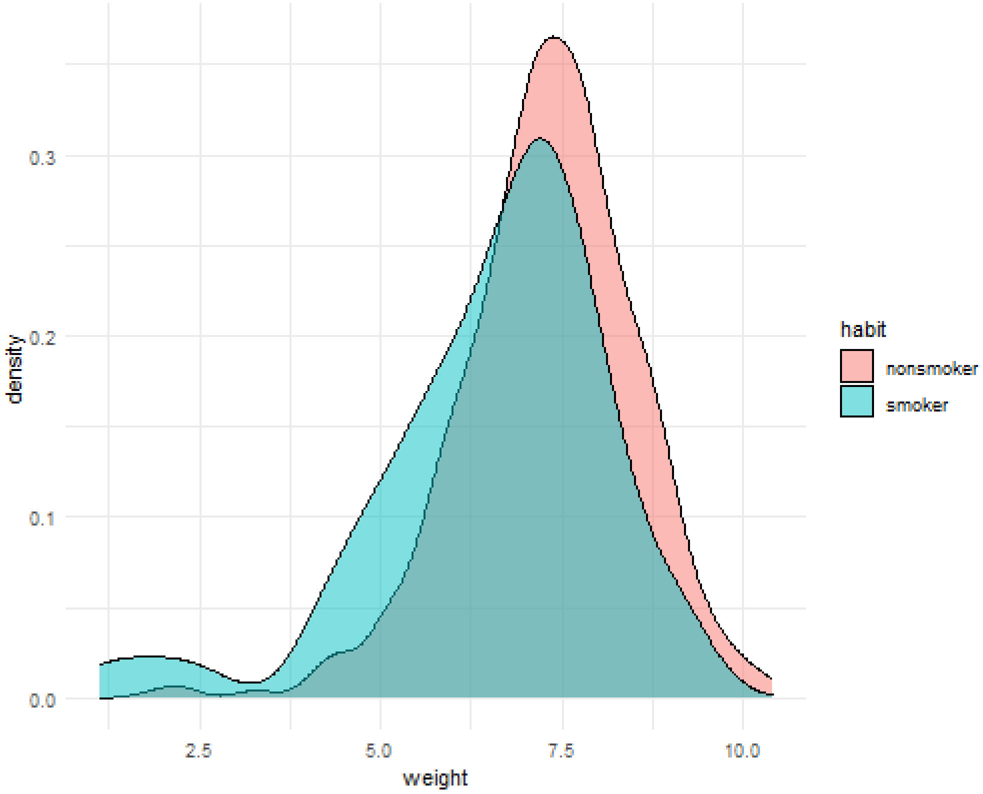
- Q: Is the mean birth weigth same across smoker and nonsmokers?

```
boxplot(weight ~ habit, data = birth)
```



Plot the densities

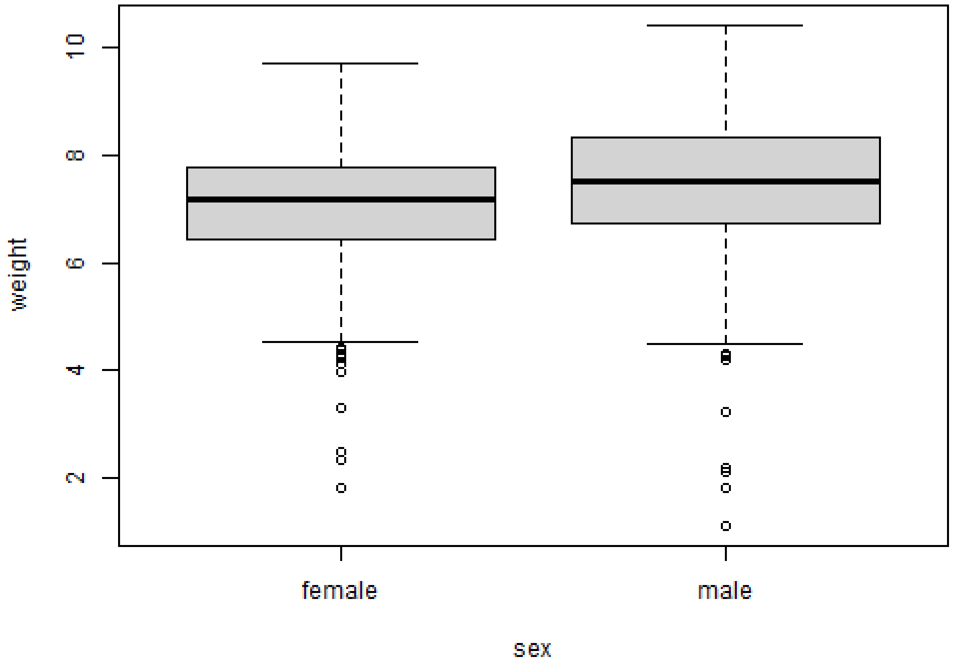
```
library(ggplot2)
ggplot(birth, aes(x = weight, group = habit, fill = habit)) + geom_de
```



Hypothesis

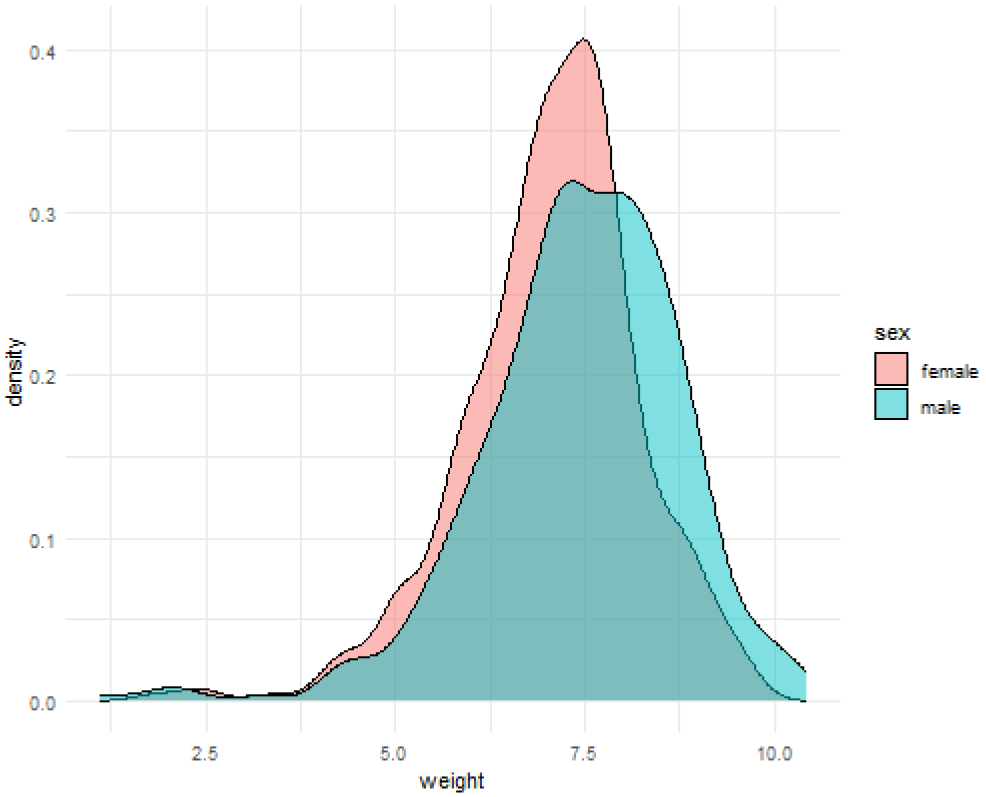
- Q: Is the mean birth weigth same across male and female children?

```
boxplot(weight ~ sex, data = birth)
```



Plot the densities

```
library(ggplot2)
ggplot(birth, aes(x = weight, group = sex, fill = sex)) + geom_densit
```



T-test (Equal Variance)

```r
x<- birth$weight[birth$habit =="nonsmoker"]
y <- birth$weight[birth$habit =="smoker"]
t.test(x,y, alternative = "greater", var.equal = T)
```

```
##
##      Two Sample t-test
##
## data:  x and y
## t = 4.5821, df = 792, p-value = 2.672e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4817562       Inf
## sample estimates:
## mean of x mean of y
##  7.306544  6.554516
```

T-test (Equal Variance, Using formula)

```
t.test(weight ~ habit, data = birth, alternative = "greater", var.equ
```

```
##
##      Two Sample t-test
##
## data:  weight by habit
## t = 4.5821, df = 792, p-value = 2.672e-06
## alternative hypothesis: true difference in means between group nonsmoker a
## 95 percent confidence interval:
##  0.4817562       Inf
## sample estimates:
## mean in group nonsmoker    mean in group smoker
##                7.306544                6.554516
```

Manually?

```r
s1=sd(x)
s2=sd(y)
m=length(x)
n=length(y)
ntotal=m+n
spooled=sqrt(((m-1)*(s1^2))+((n-1)*(s2^2)))/sqrt(ntotal-2)
spooled
```

```
## [1] 1.240811
```

```r
tm=mean(x)-mean(y)
tval=tm/(spooled*sqrt((1/m)+(1/n)))
tval
```

```
## [1] 4.58215
```

```r
(pval = 1 - pt(tval, df = ntotal-2))
```

```
## [1] 2.672421e-06
```

T-test (Unequal Variance)

- Pay attention to the value of `t`, `df` and `p-value` in the output.

```
x<- birth$weight[birth$habit =="nonsmoker"]
y <- birth$weight[birth$habit =="smoker"]
t.test(x,y, alternative = "greater", var.equal = F)
```

```
## 
##      Welch Two Sample t-test
## 
## data:  x and y
## t = 3.5993, df = 66.935, p-value = 0.0003033
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4035308       Inf
## sample estimates:
## mean of x mean of y 
##  7.306544  6.554516
```

Manual calculation for unequal variance

- The `df` calculation is a little tricky! See notes.

```
s1=sd(x)
s2=sd(y)
m=length(x)
n=length(y)
tm=mean(x)-mean(y)
tval=tm/(sqrt((s1^2/m)+(s2^2/n)))
tval
```

```
## [1] 3.599276
```

```
(df = ((s1^2/m)+(s2^2/n))^2/(1/(m-1)*(s1^2/m)^2+1/(n-1)*(s2^2/n)^2))
```

```
## [1] 66.93501
```

```
(pval = 1 - pt(tval, df = df))
```

```
## [1] 0.0003032835
```

Exercise

- Test if there is any difference in birth-weights for male and female born babies?
- Test if the weights are normally distributed?