HW5 R

Likhit Raj Yerramsetty

North Carolina births

In 2014, 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("nc.RData")
```

We have observations on 13 different variables, some categorical and some numerical. The variable descriptions are given below.

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`.

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

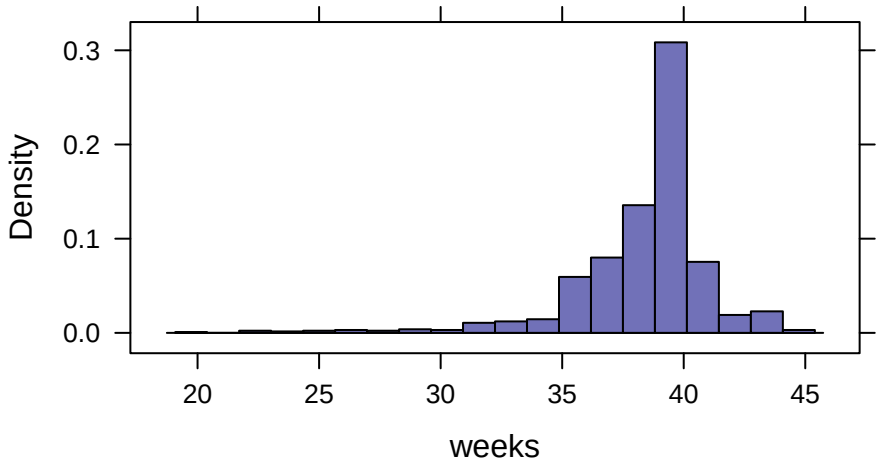
```
summary(nc)
```

```
##       fage            mage            mature        weeks            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
##
##
##
##
```

Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

Make a histogram of **weeks**, the length of each pregnancy in weeks.

```
histogram(~weeks, data=nc, nint=20)
```



Part 1 - Inference

The human pregnancies typically last 38 weeks. Use $\alpha = 0.05$ in each case.

Example 1:

Test the hypothesis that the mean duration of pregnancies is not equal to 38 weeks. What is your decision?

```
tstar <- qt(.975, df=997)
t.test(~ weeks, data=nc, alternative="two.sided", mu=38)
```

```
##
##  One Sample t-test
##
## data:  weeks
## t = 3.6065, df = 997, p-value = 0.0003258
## alternative hypothesis: true mean is not equal to 38
## 95 percent confidence interval:
##  38.15257 38.51677
## sample estimates:
## mean of x
##  38.33467
```

Looking at p-value what is your decision?

Now look at the Confidence Interval and state your decision base on the CI. A two-sided hypothesis test at significance level alpha is equivalent to using a confidence interval at (1-alpha)*100% confidence level and looking to see if the hypothesized value is in the confidence interval. If it is, you cannot reject the null hypothesis. If it isn't, you can reject the null.

Example 2:

Test the hypothesis that the mean duration of pregnancies is greater than 38 year. What is your decision?

```
m <- mean(~ weeks, data=nc, na.rm=TRUE)
s <- sd(~ weeks, data=nc, na.rm=TRUE)
tstar <- qt(.95, df=997)
t <- (m-38)/(s/sqrt(998))
t.test(~ weeks, data=nc, alternative="greater", mu=38)
```

```
##
##  One Sample t-test
##
## data:  weeks
## t = 3.6065, df = 997, p-value = 0.0001629
## alternative hypothesis: true mean is greater than 38
## 95 percent confidence interval:
##  38.18189      Inf
## sample estimates:
## mean of x
##  38.33467
```

Now look at the Confidence Interval or rather LCB. What is your decision?

Note: If you don't know how to calculate one sided bounds, you can use the regular CI (this is just for illustration, because we covered this material).For a one-sided test at significance level alpha you need

a confidence interval at (1 - 2alpha)*100% and again check for the hypothesized value in the confidence interval. If it is in the interval you cannot reject the null. In addition, the confidence interval must either be entirely below or above the hypothesized value (depending on situation) to reject the null. You can also obtain a 1-sided confidence interval (in fact called a 'confidence bound'- this is what we use) - R does this automatically, as seen. What is your decision?

```
# Method 1
t.test(~weeks,data=nc,
       alternative = "two.sided",
       mu = 38,
       conf.level = 0.90)
```

```
##
##  One Sample t-test
##
## data:  weeks
## t = 3.6065, df = 997, p-value = 0.0003258
## alternative hypothesis: true mean is not equal to 38
## 90 percent confidence interval:
##  38.18189 38.48745
## sample estimates:
## mean of x
##  38.33467
```

```
# Method 2
m <- mean(~ weeks, data=nc, na.rm=TRUE)
s <- sd(~ weeks, data=nc, na.rm=TRUE)
tstar1 <- qt(.95, df=997)
SE=s/sqrt(998)

c(m -tstar1*SE, m+tstar*SE)
```

```
## [1] 38.18189 38.48745
```

```
m +c(-1,1)*tstar1*SE
```

```
## [1] 38.18189 38.48745
```

Example 3:

Test the hypothesis that the mean duration of pregnancies is less than 38 weeks. What is your decision?

```
tstar <- qt(.05, df=997)
t.test(~ weeks, data=nc, alternative="less", mu=38)
```

```
##
##  One Sample t-test
##
## data:  weeks
## t = 3.6065, df = 997, p-value = 0.9998
## alternative hypothesis: true mean is less than 38
## 95 percent confidence interval:
```

```
##      -Inf 38.48745
## sample estimates:
## mean of x
##  38.33467
```

Question 1 - On your own

Use $\alpha = 0.1$ in each case.

(a) Test the hypothesis that the mean age of a mother is less than 27 years old. What is your decision?

SOLUTION:

```
alpha <- 0.1
mu <- 27

m <- mean(~ mage, data=nc, na.rm=TRUE)
s <- sd(~ mage, data=nc, na.rm=TRUE)

t <- (m - mu) / (s / sqrt(length(nc$mage)))
df = length(nc$mage) - 1
t_critical <- qt(1 - alpha, df)

t_test_result <- t.test(~ mage, data=nc, alternative = "less", mu = mu)

if (t < t_critical) {
  decision <- "Fail to reject the null hypothesis. So the mean is age is not less than 27."
} else {
  decision <- "Reject the null hypothesis."
}
decision
```

```
## [1] "Fail to reject the null hypothesis. So the mean is age is not less than 27."
```

```
t_test_result
```

```
##
##  One Sample t-test
##
## data:  mage
## t = 0, df = 999, p-value = 0.5
## alternative hypothesis: true mean is less than 27
## 95 percent confidence interval:
##     -Inf 27.3235
## sample estimates:
## mean of x
##        27
```

```
cat("Critical t-value (alpha = 0.1): ", t_critical)
```

```
## Critical t-value (alpha = 0.1):  1.2824
```

(b) Calculate a 95% confidence interval for the average age of a mother and interpret it in context.

SOLUTION:

```
t.test(~mage,data=nc,
       alternative = "two.sided",
       mu = 27,
       conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  mage
## t = 0, df = 999, p-value = 1
## alternative hypothesis: true mean is not equal to 27
## 95 percent confidence interval:
##  26.61442 27.38558
## sample estimates:
## mean of x
##        27
```

We are 95% confident that the range in which the average age of mothers lies is (26.61442, 27.38558)

Part 2. Test Population Proportions and Counts

Testing one sample proportion to population value - z test for one sample proportion

Example 4: Birth rate for boys in hospital We know that 51.7% of babies born are male in the population. We observed that 313 boys were born to 550 singleton deliveries in one hospital Is this different that would be expected by chance?

```
y <- 313; n <- 550; phat <- y/n; phat
```

```
## [1] 0.5690909
```

```
nullp <- 0.517
sdp <- sqrt(nullp*(1-nullp)/n); sdp
```

```
## [1] 0.02130775
```

```
onesidep <- 1-pnorm(phat, mean=nullp, sd=sdp); onesidep
```

```
## [1] 0.007248761
```

```
twosidep <- 2*onesidep; twosidep
```

```
## [1] 0.01449752
```

or we can carry out the exact test (not described by the book):

```
binom.test(y, n, p=nullp)
```

```
##
##
##
## data:  y out of 550
## number of successes = 313, number of trials = 550, p-value = 0.01499
## alternative hypothesis: true probability of success is not equal to 0.517
## 95 percent confidence interval:
##  0.5265178 0.6109179
## sample estimates:
## probability of success
##              0.5690909
```

What can we conclude for the above example ?

Testing for a difference in proportions - Two sample z-test for a difference in proportions

Example 5: Use data from the NYC Maternal Infant HIV Transmission Study We are given two qualitative variables (AZT use & Transmission) for a prospective study of HIV transmission to infants among 321 mothers. Of the 47 women on AZT, 6 transmitted and of the 274 mothers who did not take AZT, 64 transmitted.

```
n1 <- 47; y1 <- 6
n2 <- 274; y2 <- 64
ppooled <- (y1+y2)/(n1+n2); ppooled
```

```
## [1] 0.2180685
```

```
sepooled <- sqrt(ppooled*(1-ppooled)/n1 + ppooled*(1-ppooled)/n2); sepooled
```

```
## [1] 0.06519423
```

```
z <- (y1/n1 - y2/n2)/sepooled; z
```

```
## [1] -1.624639
```

```
pval <- 2*(1-pnorm(z, lower.tail = FALSE)); pval
```

```
## [1] 0.1042396
```

Question 2 - On your own

a) Was AZT effective? Based on what information? Make sure that you read the correct p-value.

SOLUTION: Since the p-value is greater alpha (0.104 > 0.1). We are using the data from example 5 to derive the p-value for this question. We cannot reject the null hypothesis and say that there is a difference between using AZT and not using AZT.

In the previous example of MTC HIV transmission, we had counts. If instead, we had counts and proportions we could use this code to calculate the standard error of the difference and create a confidence interval.

```
n1 <- 47; p1 <- 0.13
n2 <- 274; p2 <- 0.23
sediff <- sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2); sediff
```

```
## [1] 0.05525155
```

```
(p2 - p1) + c(-1.96, 1.96)*sediff
```

```
## [1] -0.00829303  0.20829303
```

b) Was AZT effective?

SOLUTION: Since the intervals (-0.00829, 0.20829) include a 0 in the range. We cannot reject the null hypothesis. We do not have enough evidence to say that AZT is effective.

Part 3. Test Population Menas

Example 6: Acquire the WNBA & NBA datafile

```
bball <- read.csv("Basket.csv")
glimpse(bball)
```

```
## Rows: 24
## Columns: 5
## $ X        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ PLAYER   <chr> "Cynthia Cooper         ", "Ruthie Bolton-Holifield", "Lisa L~
## $ GENDER   <chr> "Female", "Female", "Female", "Female", "Female", "Female", "~
## $ HEIGHTIN <int> 70, 69, 77, 74, 75, 70, 70, 75, 74, 68, 72, 75, 78, 81, 85, 7~
## $ WEIGHTLB <int> 150, 150, 170, 165, 180, 158, 145, 178, 178, 132, 172, 185, 2~
```

Take a peek at the data

```
names (bball)
```

```
## [1] "X"        "PLAYER"   "GENDER"   "HEIGHTIN" "WEIGHTLB"
```

```
head (bball)
```

```
##   X                  PLAYER GENDER HEIGHTIN WEIGHTLB
## 1 1 Cynthia Cooper          Female       70      150
## 2 2 Ruthie Bolton-Holifield Female       69      150
## 3 3 Lisa Leslie             Female       77      170
## 4 4 Wendy Palmer            Female       74      165
## 5 5 Jennifer Gillom         Female       75      180
## 6 6 Andrea Stinson          Female       70      158
```

Look more closely at the variables

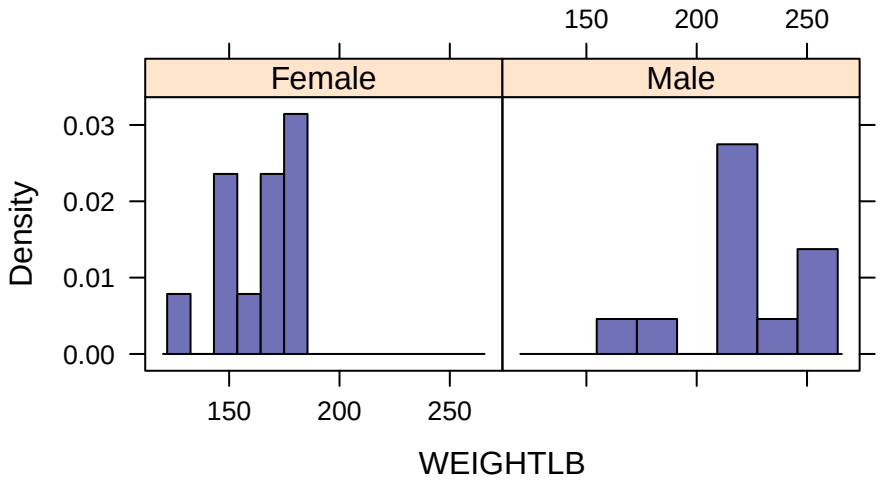
```
favstats (~HEIGHTIN |GENDER, data=bball)
```

```
##   GENDER min    Q1 median Q3 max     mean       sd  n missing
## 1 Female  68 70.00     73 75  77 72.41667 2.937480 12       0
## 2   Male  72 78.75     80 82  85 79.91667 3.423404 12       0
```

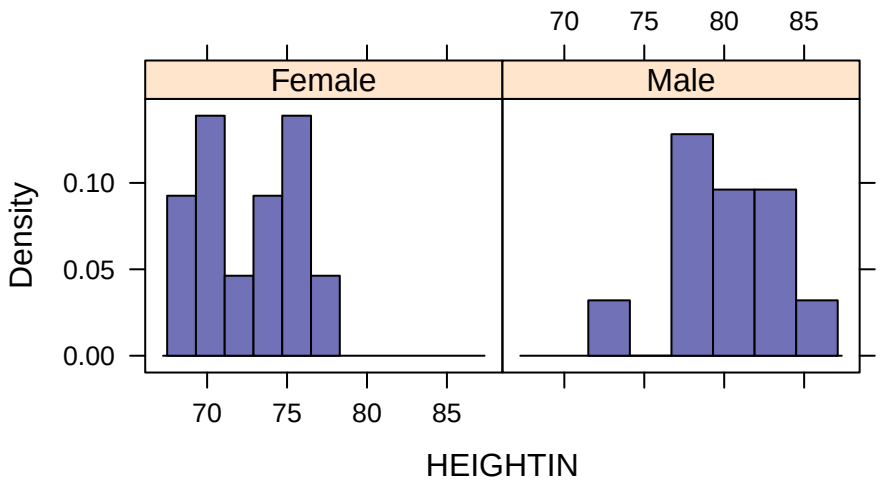
```
favstats (~WEIGHTLB |GENDER, data=bball)
```

```
##   GENDER min     Q1 median    Q3 max     mean       sd  n missing
## 1 Female 132 150.00  167.5 178.0 185 163.5833 16.51698 12       0
## 2   Male 165 215.75  220.0 242.5 256 221.5833 26.58933 12       0
```

```
histogram (~WEIGHTLB |GENDER, data=bball)
```




```
histogram (~HEIGHTIN |GENDER, data=bball)
```



Comparing one group mean to population - a one sample t-test

Example 7: Are Women in WNBA taller than US women in general? To answer this question we need to compare the sample of WNBA heights to NHANES mean height which has population values of: (mu=63.75, sd=3.423)

```
mu=63.75
sd=3.423
womenonlyds <- filter(bball, GENDER=="Female")
xpnorm(c(mu-3*sd, mu-2*sd, mu-sd, mu+sd, mu+2*sd, mu+3*sd), mean=mu, sd=sd)
```

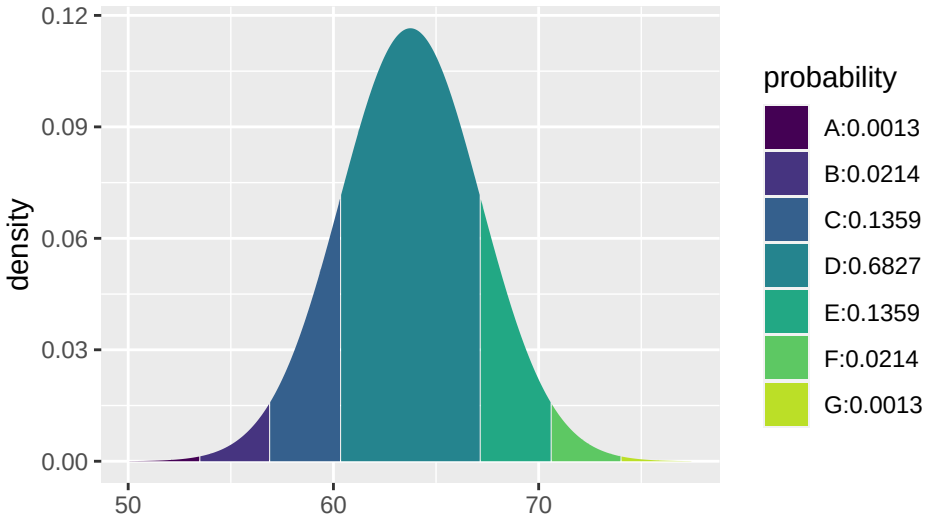
```
##

## If X ~ N(63.75, 3.423), then

##  P(X <= 53.48) = P(Z <= -3) = 0.00135    P(X <= 56.90) = P(Z <= -2) = 0.02275    P(X <= 60.33) = P(Z

##  P(X >  53.48) = P(Z >  -3) = 0.99865    P(X >  56.90) = P(Z >  -2) = 0.97725    P(X >  60.33) = P(Z

##
```




```
## [1] 0.001349898 0.022750132 0.158655254 0.841344746 0.977249868 0.998650102
```

```
t.test(womenonlyds$HEIGHTIN, alternative="greater", mu=63.75, data=womenonlyds)
```

```
##
##  One Sample t-test
##
## data:  womenonlyds$HEIGHTIN
## t = 10.22, df = 11, p-value = 2.972e-07
## alternative hypothesis: true mean is greater than 63.75
## 95 percent confidence interval:
##  70.8938     Inf
## sample estimates:
## mean of x
##  72.41667
```

Question 3 - On your own

What do we conclude? Are women playing professional basketball taller than American women?

SOLUTION: From the data above we can see that the LCB is 70.89 inches for WNBA players with a 95% confidence level. Which is much higher than the average for american women which is 63.75.The WNBA players fall in the F area in the graph where the probability is 0.0214 + 0.0013. So it is safe to conclude that women playing professional basketball are taller than 97.73% of the american women.

Comparing two groups to each other - a two Sample independent t-test

If people who play basketball tend to be exceptionally tall would we expect women playing professional basketball to be as tall as the men?

Using the bball data we can ask if women in WNBA different in terms of men in NBA?

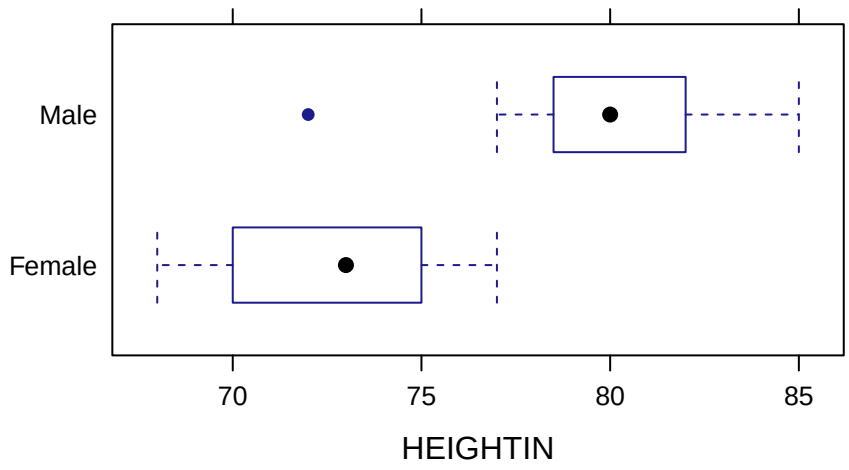
```
t.test(HEIGHTIN ~ GENDER, data=bball) # Unpooled
```

```
##
##  Welch Two Sample t-test
##
## data:  HEIGHTIN by GENDER
## t = -5.7595, df = 21.504, p-value = 9.339e-06
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -10.204201  -4.795799
## sample estimates:
## mean in group Female   mean in group Male
##             72.41667             79.91667
```

```
t.test(HEIGHTIN ~ GENDER, var.equal=TRUE, data=bball)   # Pooled
```

```
##
##  Two Sample t-test
##
## data:  HEIGHTIN by GENDER
## t = -5.7595, df = 22, p-value = 8.564e-06
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -10.200583  -4.799417
## sample estimates:
## mean in group Female   mean in group Male
##             72.41667             79.91667
```

```
bwplot(GENDER ~ HEIGHTIN, data=bball)
```



(We went over 2 independent t-tests , one assumes equal variance like the test above. If you don't include var.equal=TRUE you will have Welch test)

Question 4 - On your own

Are there difference in the heights of men and women professional basketball players?

SOLUTION: Yes there is a difference in the heights of men and women professional basketball players.We can see from the confidence interval (-10.2, -4.8) that there is no 0 in the range which means we can reject the null hypothesis. Also, the p value is 0.000008564 which is much smaller than alpha. Based on this evidence, we can say that there is a difference in the heights of men and women basketball players.

Comparing Groups when data is paired - Dependent t-test using difference scores

Example 8:

This is a special case of data.

We are not comparing the mean of one group vs. another. We have the same group of people over time with matched data. Paired data can also come from different people (twins, siblings, etc.), the unique aspect is that two numbers belong together.

For this analysis we will be using data from National Education Longitudinal Study (NELS). A nationally representative sample of eighth-graders were first surveyed in the spring of 1988. A sample of these respondents were then resurveyed through four follow-ups in 1990, 1992, 1994, and 2000. On the questionnaire, students reported on a range of topics including: school, work, and home experiences; educational resources and support; the role in education of their parents and peers; neighborhood characteristics; educational and occupational aspirations; and other student perceptions. Additional topics included self-reports on smoking, alcohol and drug use and extracurricular activities. For the three in-school waves of data collection (when most were eighth-graders, sophomores, or seniors), achievement tests in reading, social studies, mathematics and science were administered in addition to the student questionnaire.

Question: Many middle schoolers say their grades don't matter and they'll work harder in high school. We can answer this question using data from NELS. For this example we'll operationally define performance as reading achievement.

Create difference scores between reading achievement scores in 8th and 10th grade for each participant.

```
educ <- read.csv("NELS.csv")
educ <- mutate(educ, diff = ACHRDG10 - ACHRDG08);

favstats (~ACHRDG08, data=educ)
```

```
##    min      Q1 median      Q3   max     mean       sd   n missing
##  35.74 49.9875 56.445 63.1325 70.55 56.04906 8.829726 500       0
```

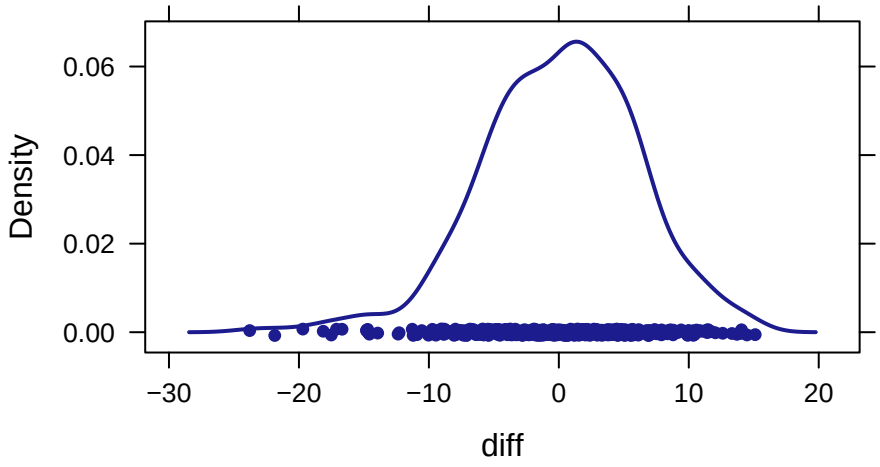
```
favstats (~ACHRDG10, data=educ)
```

```
##    min      Q1 median      Q3  max     mean       sd   n missing
##  31.53 49.6175 57.545 62.9475 68.8 56.11404 8.304459 500       0
```

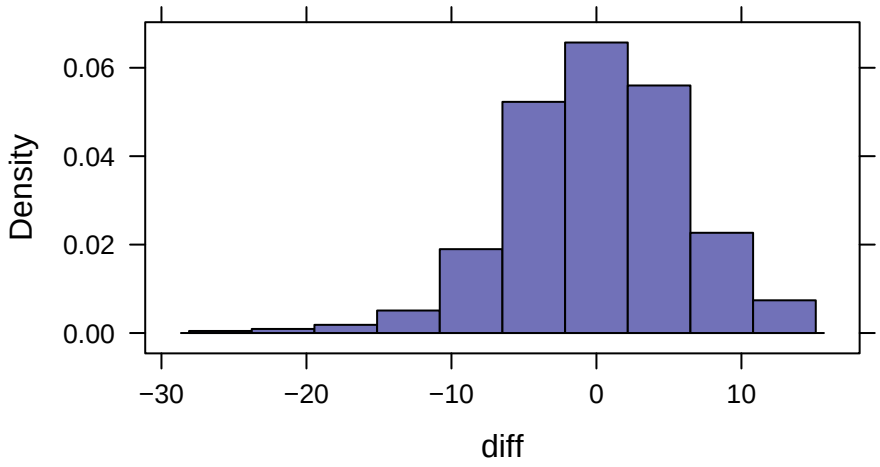
```
favstats (~diff, data=educ)
```

```
##     min      Q1 median     Q3   max    mean       sd   n missing
##  -23.79 -3.7575  0.525 4.2625 15.11 0.06498 6.031412 500       0
```

```
densityplot (~diff, data=educ)
```




```
histogram (~diff, data=educ)
```



Question 5 - On your own

Create a confidence interval for the difference and perform the correct hypothesis test - you are back to one - mean t test , with variable diff. Did reading achievement scores differ over time (between 8th and 10th grade)? If so, why? If not, why not?

SOLUTION:

```
mean_diff <- mean(educ$diff, na.rm = TRUE)
sd_diff <- sd(educ$diff, na.rm = TRUE)

se_diff <- sd_diff / sqrt(length(educ$diff))

df <- length(educ$diff) - 1
```

```r
t_critical <- qt(0.975, df)

margin_of_error <- t_critical * se_diff

confidence_interval <- c(mean_diff - margin_of_error, mean_diff + margin_of_error)

t_test_result <- t.test(educ$diff, mu = 0)

confidence_interval
```

```
## [1] -0.4649723  0.5949323
```

```r
t_test_result
```

```
##
##  One Sample t-test
##
## data:  educ$diff
## t = 0.2409, df = 499, p-value = 0.8097
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.4649723  0.5949323
## sample estimates:
## mean of x
##   0.06498
```

The confidence intervals at 95% confidence level is (-0.465,0.595). This includes 0 in the range so we cannot reject the null hypothesis. Furthermore, we can see that the p-value is 0.81 which is much higher than alpha whether it is at 0.1 or 0.05. So, based on this evidence, we cannot conclude that there is a difference in reading achievement scores between 8th grade and 10th grade.