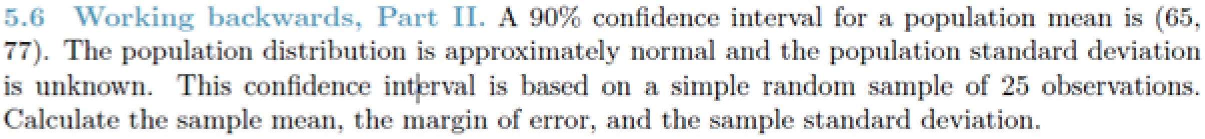
# Chapter 5 HW

*Kishore Prasad*

*March 31, 2016*

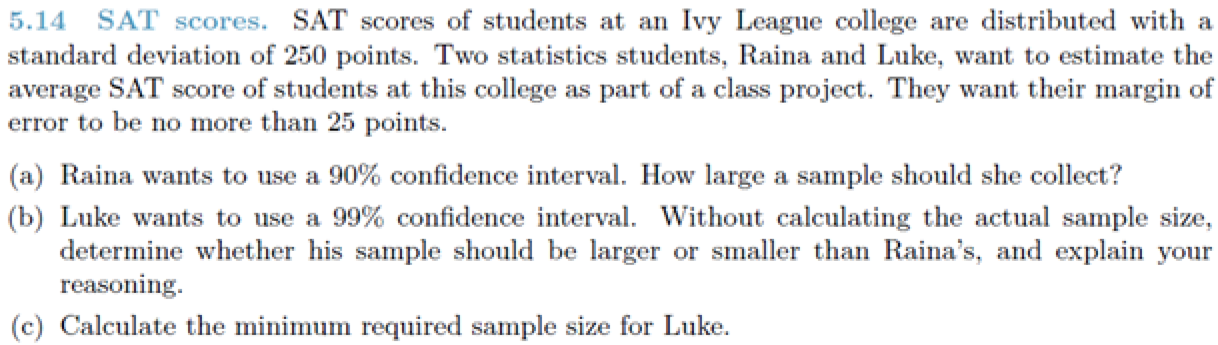


**5.6 Working backwards, Part II.** A 90% confidence interval for a population mean is (65, 77). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation.

```
margin_of_error <- (77 - 65) / 2

sample_mean <- 65 + margin_of_error

number_obs <- 25

degrees_freedom <- number_obs - 1


# t value lookup for 90% CI

confidence_interval <- 0.90
alpha <- 1 - confidence_interval
t_lookup <- (1 - (alpha / 2))


t <- qt(t_lookup, degrees_freedom)

standard_error <- margin_of_error / t

std_dev <- standard_error * sqrt(number_obs)
```

**Following are the calculated values:**

**sample mean** = 71  
**standard deviation** = 17.53  
**margin of error** = 6



**5.14 SAT scores.** SAT scores of students at an Ivy League college are distributed with a standard deviation of 250 points. Two statistics students, Raina and Luke, want to estimate the average SAT score of students at this college as part of a class project. They want their margin of error to be no more than 25 points.

(a) Raina wants to use a 90% confidence interval. How large a sample should she collect?

(b) Luke wants to use a 99% confidence interval. Without calculating the actual sample size, determine whether his sample should be larger or smaller than Raina's, and explain your reasoning.

(c) Calculate the minimum required sample size for Luke.

(a) Solution

```
margin_of_error <- 25
Std_dev <- 250

# z-score for 90% CI

z_Score <- qnorm(0.95)

z_Score
```

```
## [1] 1.644854
```

```
number_obs  <- ((Std_dev*z_Score) / margin_of_error)^2
number_obs
```

```
## [1] 270.5543
```

**Sample size needed =** 270.5543454

(b) Solution

**We can see from the formula that number_obs is directly proportional to the z_score.This means that as we increase the Z_score, the number_obs also increases. Thus Luke's sample should be larger than Raina's.**

(c) Solution

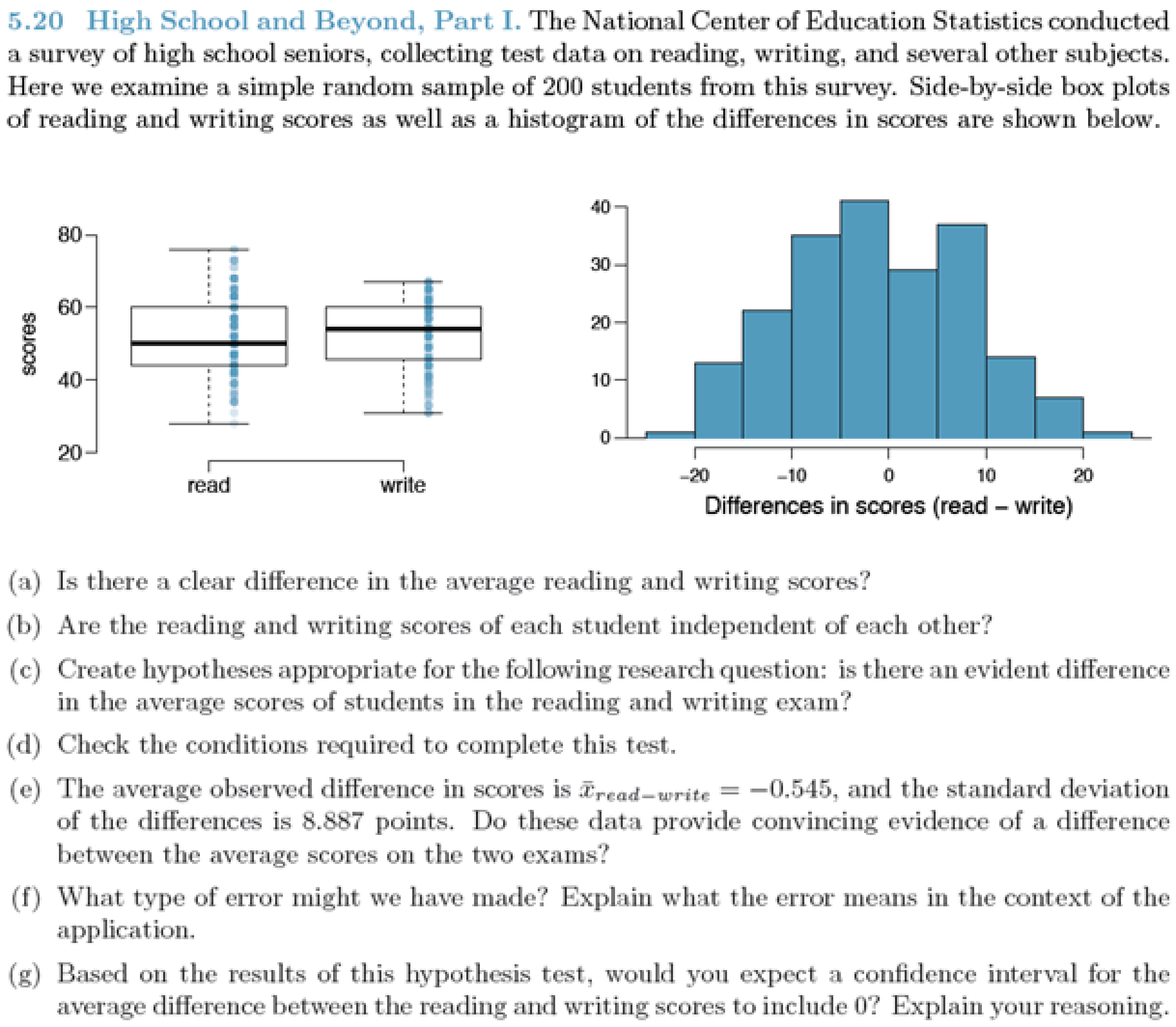
```
margin_of_error <- 25
Std_dev <- 250

# z-score for 99% CI

z_Score <- qnorm(0.995)

z_Score
```

```
## [1] 2.575829
```

```
number_obs  <- ((Std_dev*z_Score) / margin_of_error)^2
number_obs
```

```
## [1] 663.4897
```

**Sample size needed =** 663



**5.20 High School and Beyond, Part I.** The National Center of Education Statistics conducted a survey of high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of reading and writing scores as well as a histogram of the differences in scores are shown below.

(a) Is there a clear difference in the average reading and writing scores?

(b) Are the reading and writing scores of each student independent of each other?

(c) Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?

(d) Check the conditions required to complete this test.

(e) The average observed difference in scores is $\bar{x}_{read-write} = -0.545$, and the standard deviation of the differences is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams?

(f) What type of error might we have made? Explain what the error means in the context of the application.

(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

**(a) The distribution of the difference around the mean looks to be normal. There seems to be no difference in the average reading and writing scores.**

**(b) Given that this is a simple random sample, I would assume that each student's score is independent of another student's score. However, the reading skills when compared to the writing skills of the same student would not be independent.**

**(c) We could form the hypothesis as below assuming that we are talking about difference in average reading and writing scores (and not average difference between reading and writing scores):**

$\mathbf{H_0}$**:** There is no difference in the averages of reading and writing score i.e $(\mu_{\text{reading}} - \mu_{\text{writing}}) = 0$

$\mathbf{H_A}$**:** There is a difference in the averages of reading and writing score i.e $(\mu_{\text{reading}} - \mu_{\text{writing}}) \neq 0$

**(d) The below conditons need to be satisfied:**

Independence of observations: As we noted before, each student score is independent of another student. Hence this condition is satisfied.

Observations come from nearly normal distribution: Both the reading and the writing scores boxplots do not seem to show any outliers. They seem to be reasonably normally distributed.

**(e) We can form the below hypothesis for the average difference in scores:**
$\mathbf{H_0}$**:** $\mu_{\textbf{ diff}} = 0$

$\mathbf{H_A}$**:** $\mu_{\textbf{ diff}} \neq 0$

**Assuming that the samples are less than 10% of the students and are from a simple random sample and the difference is normally distributed, we can apply the t-distribution as below:**

```
std_dev_diff <- 8.887
mean_diff <- -0.545
number_obs_diff <- 200

# standard error
standard_error_diff <- std_dev_diff / sqrt(number_obs_diff)

# T statistic
t_diff <- (mean_diff - 0) / standard_error_diff

degrees_freedom <- number_obs_diff - 1

p_Value <- pt(t_diff, df=degrees_freedom)

p_Value
```

```
## [1] 0.1934182
```

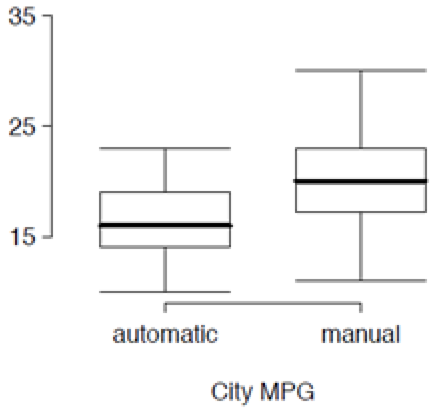
**Since, the p value is not less than 0.05 we fail to reject the null hypothesis. It can be concluded that there is no convincing evidence of a difference in the scores of reading and writing.**

**(f) We have rejected the alternative hypothesis. If we have incorrectly rejected the altenative hypothesis, we may have made a Type II error.**

**(g) Since our null hypothesis is that there is no difference between the scores, we would like the confidence interval for the average difference include 0. If the confidence interval includes 0 then it means that the difference is not positive or negative. Therefore this results in the failure to reject the null hypothesis.**

**5.32 Fuel efficiency of manual and automatic cars, Part I.** Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.[42]

|  | City MPG |  |
|---|---|---|
|  | Automatic | Manual |
| Mean | 16.12 | 19.85 |
| SD | 3.58 | 4.51 |
| n | 26 | 26 |



**Lets form the below hypothesis:**

**$H_0$:** There is no difference in the mean fuel economy of automatic and manual cars i.e $(\mu_{\text{automatic}} - \mu_{\text{manual}}) = 0$

**$H_A$:** There is a difference in the mean fuel economy of automatic and manual cars i.e $(\mu_{\text{automatic}} - \mu_{\text{manual}}) \neq 0$
Next, lets calculate the point estimate of population difference.

```
number_obs <- 26

mean_automatic <- 16.12
mean_manual <- 19.85

std_dev_automatic <- 3.58
std_dev_manual <- 4.51


# Difference in means
means_diff <- mean_automatic - mean_manual
means_diff
```

```
## [1] -3.73
```

```
# Standard error
standard_error_diff <- sqrt( (std_dev_automatic^2 / number_obs) + (std_dev_manual^2 / number_obs) )
standard_error_diff
```

```
## [1] 1.12927
```

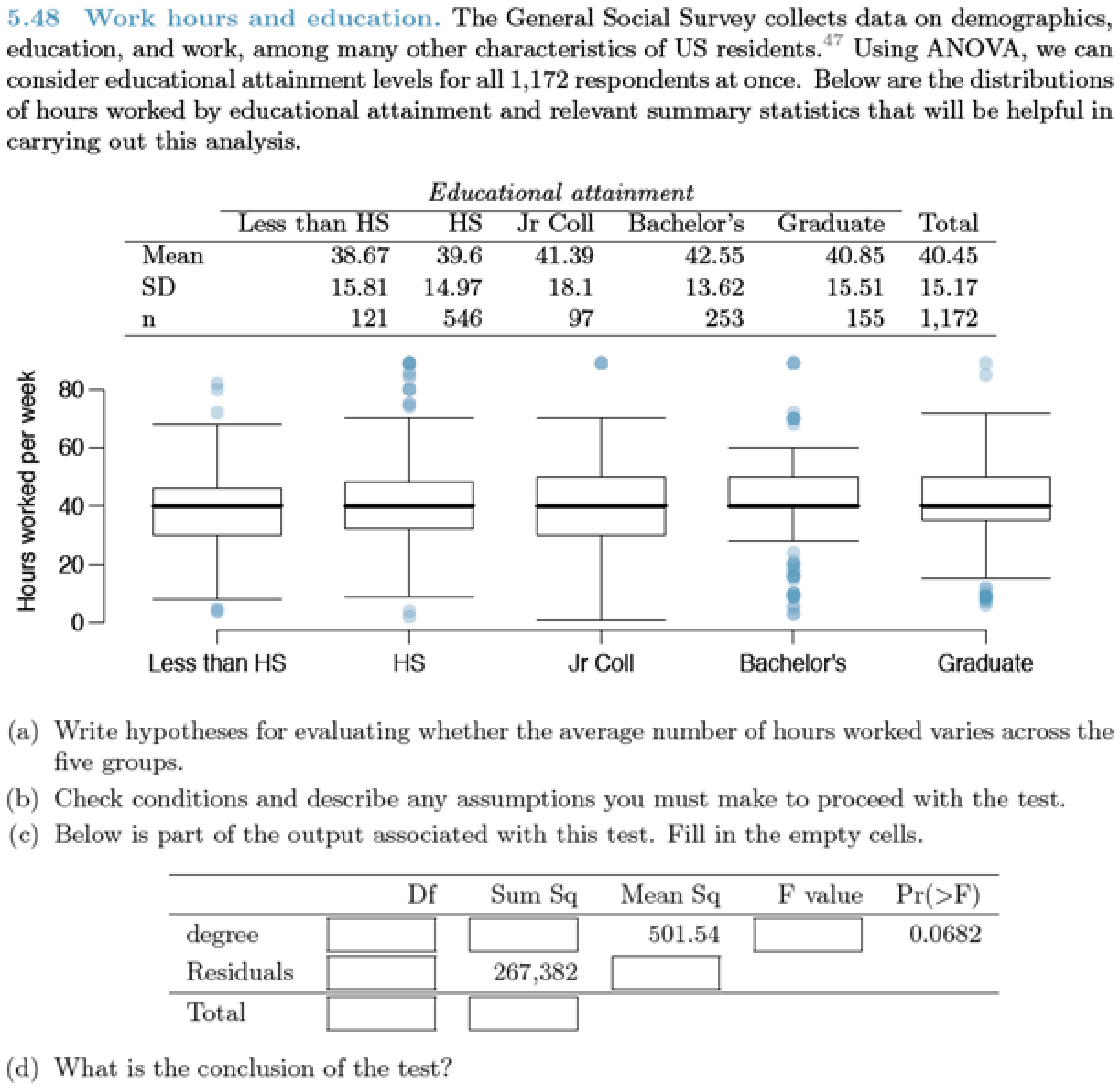
```
# t-statistic
t_diff <- (means_diff - 0) / standard_error_diff
t_diff
```

```
## [1] -3.30302
```

```
#p-value
p_Value <- pt(t_diff, df=number_obs-1)
p_Value
```

```
## [1] 0.001441807
```

**The p-value of 0 is less than 0.05, we reject the null hypothesis in favour of alternative hypothesis. We can conclude that there is strong evidence of a difference in fuel efficiency between manual and automatic transmissions.**



5.48 Work hours and education. The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents.[47] Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

| | *Educational attainment* | | | | | |
|---|---|---|---|---|---|---|
| | Less than HS | HS | Jr Coll | Bachelor's | Graduate | Total |
| Mean | 38.67 | 39.6 | 41.39 | 42.55 | 40.85 | 40.45 |
| SD | 15.81 | 14.97 | 18.1 | 13.62 | 15.51 | 15.17 |
| n | 121 | 546 | 97 | 253 | 155 | 1,172 |

(a) Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

(b) Check conditions and describe any assumptions you must make to proceed with the test.

(c) Below is part of the output associated with this test. Fill in the empty cells.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| degree | □ | □ | 501.54 | □ | 0.0682 |
| Residuals | □ | 267,382 | □ | | |
| Total | □ | □ | | | |

(d) What is the conclusion of the test?

**(a) We will form the below hypothesis:**

$H_0$: There is no difference in the mean / average hours worked across the 5 groups i.e $\mu_{\text{LeHS}} = \mu_{\text{HS}} = \mu_{\text{JrCol}} = \mu_{\text{Bch}} = \mu_{\text{Grad}} = 0$

$H_A$: At least one group has a difference in the mean / average hours worked with the others

### (b) The following conditions must be met:

The observations are independent within and across groups: Assuming that the survey has been carried out in an unbiased manner with random sampling, we will consider that this condition has been met.

The data within each group are nearly normal: The Bachelor's distribution seems to be skewed with significant outliers on lower end of the scale. Similarly, the HS boxplot indicates a skew at the higher end of the scale. Most of the other groups also have some outliers. We cannot consider that this condition has been met.

The variability across the groups is about equal: The standard deviation indicates that both Bachelor's and Jr Coll have a different pattern of variation than the other groups.All other groups seem to have the standard deviation around 15.

### (c) Calculating for the missing values:

```
n <- 1172
k <- 5

degrees_freedom_G <- k - 1
degrees_freedom_E <- n - k

degrees_freedom_G
```

```
## [1] 4
```

```
degrees_freedom_E
```

```
## [1] 1167
```

```
mean_Total <- 40.45

groups_data <- data.frame(n=c(121,546,97,253,155),
                          sd=c(15.81,14.97,18.1,13.62,15.51),
                          mean=c(38.67,39.6,41.39,42.55,40.85))
# Compute the SSG
SSG <- sum( groups_data$n * (groups_data$mean - mean_Total)^2 )
SSG
```

```
## [1] 2004.101
```

```
# Compute the MSG
MSG <- (1 / degrees_freedom_G) * SSG

# Compute the MSE
SSE <- 267382
MSE <- SSE / degrees_freedom_E
MSE
```

```
## [1] 229.1191
```

```
# Compute the F statistic
F <- MSG / MSE
F
```

```
## [1] 2.186745
```

**Below is the completed table:**

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| degree | **4** | **2004.1** | 501.54 | **2.186745** | 0.0682 |
| Residuals | **1167** | 267382 | **229.12** | | |
| Total | **1171** | **269386.1** | | | |

**(d) The p-value = 0.0682 is greater than 0.05. Hence we fail to reject the null hypothesis. We conclude that there is no significant difference between the groups.**