Chapter 7 - Inference for Numerical Data

Md Forhad Akbar

Working backwards, Part II. (5.24, p. 203) 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.

Answer:

Sample mean:

```
a<- c(65,77)
sample_mean <- mean(a)
sample_mean
```

```
## [1] 71
```

Margin of error:

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

```
## [1] 6
```

Standard Deviation (sd):

```
n <- 25
df <- n - 1 #degree of freedom
p <- 0.9
p_2tails <- p + (1 - p)/2
t_val <- qt(p_2tails, df)

SE <- ME / t_val # ME = t * SE
sd <-  round((SE * sqrt(n)),2)# SE = sd/sqrt(n)
sd
```

```
## [1] 17.53
```

SAT scores. (7.14, p. 261) 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?

Answer:

z score:

```
sd <- 250
ME <- 25
z <- qnorm(0.95)
z
```

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

Sample size:

```
sample_size <- (((z * sd) / (ME))^2)
sample_size
```

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

(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.
Answer:
Luke's sample size should be larger as he wants to have higher confidence interval. Also, higher confidence interval require higher z score.

(c) Calculate the minimum required sample size for Luke.
Answer:
z score:

```
sd <- 250
ME <- 25
zscore_Luke <- qnorm(0.995)
zscore_Luke
```

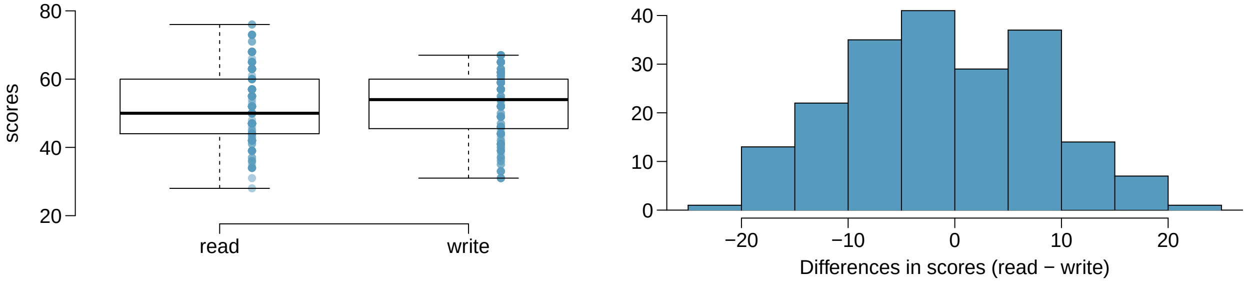
```
## [1] 2.575829
```

Sample size:

```
sample_size_Luke <- (((zscore_Luke * sd) / (ME))^2)
sample_size_Luke
```

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

High School and Beyond, Part I. (7.20, p. 266) 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?
Answer: There is no clear differences in average reading and writing, it is almost a normal distribution wtih slightly right skewed wuth mean at 0

(b) Are the reading and writing scores of each student independent of each other?
Answer:
The reading and writing scores of each student are independent, reading and writing test data will not affect 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?
Answer:

H0: Difference between average reading and writing scores is zero.
HA: Difference between average reading and writing scores is not zero.

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

Answer: The histrogram is almost Normal. So, the conditions of Central Limit Theorem were met, which are:
* Sample observations are independent.
* We have a large enough sample size

(e) The average observed difference in scores is $\widehat{x}_{read-write} = -0.545$, 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?

Answer:

Before determining the difference, let's compute some essential parameters:

```
n <- 200
mean.diff <- -.545
df <- n-1
SD <- 8.887
SE <- SD/sqrt(n)
T <- (mean.diff-0)/SE
pvalue <- pt(T, df)
pvalue
```

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

The null hypothesis is: H0: Difference between average reading and writing scores is zero.

We see that the p_value $>$ 0.5. So, we can not reject the null hypothesis. We can can coclude that data doesn't offer strong evidence of a difference between the average scores of two exams.

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

Type I error: Incorrectly rejecting the null hypothesis.
Type II error: Incorrectly rejecting the alternative hypothesis.

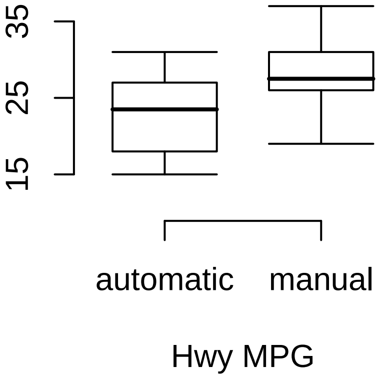
Since we did not reject the null hypothesis, we are at a risk of making a Type II error. In this instance, we should have noticed that we had convincing data that there is a difference in the reading and writing average scores but we did not find any evidence.

(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.

Answer:
Since our results indicated that there is no difference in the reading and writing scores, I would expect that the confidence interval would include 0.

Fuel efficiency of manual and automatic cars, Part II. (7.28, p. 276) The table provides summary statistics on highway fuel economy of cars manufactured in 2012. Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, and interpret this interval in the context of the data.

	Hwy MPG	
	Automatic	Manual
Mean	22.92	27.88
SD	5.29	5.01
n	26	26



Answer:

```
n <- 26

mean_auto <- 22.92
sd_auto <- 5.29

mean_manual <- 27.88
sd_manual <- 5.01

point_estimate <- mean_manual - mean_auto
SE_Diff <- sqrt( (sd_auto^2 / n) + ( sd_manual^2 / n) )

df <- n - 1

critical_value_t<- qt(.99, df)

point_estimate-critical_value_t*SE_Diff
```

```
## [1] 1.409078
```

```
point_estimate+critical_value_t*SE_Diff
```

```
## [1] 8.510922
```

We are 98% confident that difference between average highway mileage of manual and automatic cars is 1.41% to 8.5%.

Email outreach efforts. (7.34, p. 284) A medical research group is recruiting people to complete short surveys about their medical history. For example, one survey asks for information on a person's family history in regards to cancer. Another survey asks about what topics were discussed during the person's last visit to a hospital. So far, as people sign up, they complete an average of just 4 surveys, and the standard deviation of the number of surveys is about 2.2. The research group wants to try a new interface that they think will encourage new enrollees to complete more surveys, where they will randomize each enrollee to either get the new interface or the current interface. How many new enrollees do they need for each interface to detect an effect size of 0.5 surveys per enrollee, if the desired power level is 80%?

Answer:

We start by identifying the Z-score that would give us a lower tail of 80%. For a moderately large sample size per group, the Z-score for a lower tail of 80% would be about Z = 0.84.

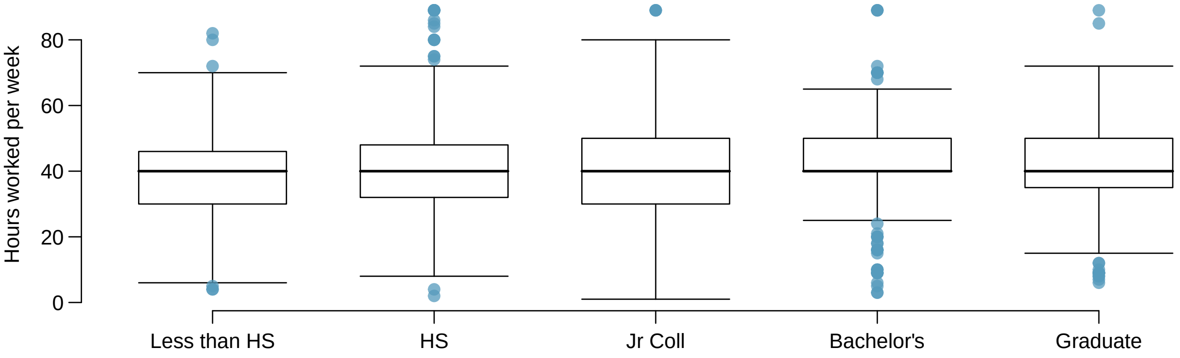
Additionally, the rejection region extends 1.96 SE from the center of the null distribution for sigma= 0.5. This allows us to calculate the target distance between the center of the null and alternative distributions in terms of the standard error:
0.84*SE + 1.96*SE = 2.8*SE

```
#SE<- sqrt( (2.2^2 / n) + ( 2.2^2 / n) )
#0.5 = 2.8 * SE
n = (2.8^2/(0.5)^2)*(2.2^2+2.2^2)
n
```

```
## [1] 303.5648
```

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.

Answer:
H0: The average number of hours worked do not varry across groups.
HA: The average number of hours worked do varry across groups

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

Answer:

ANOVA assumptions: 1) Based on the data provided (the number of respondents were 1172, which is quite large, but $<$ 10% of the total pupulation), I presume the observations are independent within and across groups. 2) The box plots don't suggest strong normality in every group, but it could be nearly normal. 3) However variability across the groups is about equal.

(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					

Answer: Couldn't open the latex to fillout the blank

(d) What is the conclusion of the test?
Answer:
The conclusion of this test is the average hours worked is different for at least one group