Chapter 7 - Inference for Numerical Data

Gehad Gad - 10/17/2020

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

```
HI90CI <- 77
LO90CI <- 65

#Get the sample mean
Sample_mean <- (HI90CI + LO90CI)/2
Sample_mean
```

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

The sample mean is 71.

```
#Get te margin of error.
MarginError <- (HI90CI - LO90CI) / 2
MarginError
```

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

The margin of error is 6.

```
n = 25
df <- n - 1
p <- 0.9
p_2tails <- p + (1 - p)/2

t_val <- qt(p_2tails, df)

# Since ME = t * SE
SE <- MarginError / t_val

# Since SE = sd/sqrt(n)
sd <- SE * sqrt(n)
sd
```

```
## [1] 17.53481
```

The Sample standard deviation is 17.53481.

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

```
SD <- 250
ME <- 25
Z <- 1.65

Sample_n <- ((Z * SD) / ME ) ^ 2
Sample_n
```

```
## [1] 272.25
```

The sample size should be 272 students.

(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 should be larger than Raina's since the CI is 99%, the sample should increase.

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

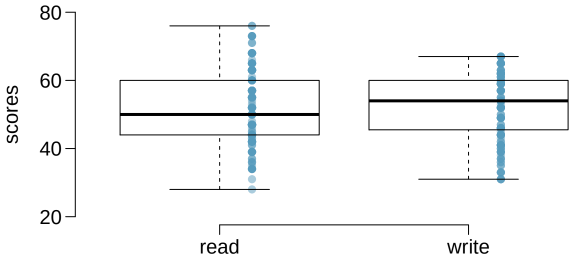
Answer

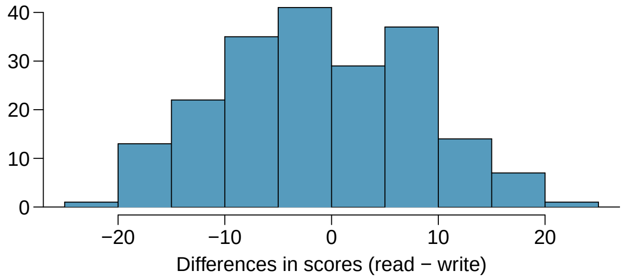
```
SD <- 250
ME <- 25
Z <- 2.58

Sample_n <- ((Z * SD) / ME ) ^ 2
Sample_n
```

```
## [1] 665.64
```

The sample size should be 665 students.

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

I do not see a clear different in the average reading and writing scores.

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

Answer

No, the reading and writing ability are some how related.

(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 Null hypothesis (H0): There is not a difference in the average scores of students in the reading and writing exam.

Alternate hypothesis (H1): There is a difference in the average scores of students in the reading and writing exam.

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

Answer

The difference histogram suggested the data are paired. If paired, then they wouldn't be independent.

(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

Calculate the p-value.

```
sd_diff <- 8.887
mean_diff <- -0.545
n <- 200

SE <- sd_diff / sqrt(n)

t_value <- (mean_diff)/ SE

p_value <- pt(t_value,n-1)
p_value
```

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

Since the p__value is higher than 0.05, we can not reject the NULL hypothesis.

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

Answer

We may have made error in rejecting the alternative hypothesis H1. We might have wrongly concluded that there is not a difference in the average student reading and writing exam scores.

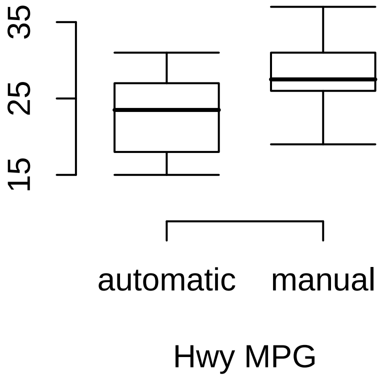
(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

Answer

Yes, I would expect a confidence interval for the average difference between reading and writing scores to include 0 because the hypothesis test indicates that the difference is not in one side or another.

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

We can calculate the P-value or the CI.

```
#Automatic
auto_mean <- 22.92
auto_sd <- 5.29
auto_n <- 26

#Manual
man_mean <- 27.88
man_sd <- 5.01
man_n <- 26

#Difference in sampe means
mean_diff <- auto_mean - man_mean

#Standard Error
SE_diff <- sqrt((auto_sd^2 / auto_n) + (man_sd^2 / man_n))
```

```
z_98 <- 2.326
LowCI <- mean_diff - SE_diff * z_98
LowCI
```

```
## [1] -8.283576
```

```
HighCI <- mean_diff + SE_diff * z_98
HighCI
```

```
## [1] -1.636424
```

The confidence inerval are less than zero, which indicate that there is a differnece of fuel efficiency between automatic and manual vehicles.

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

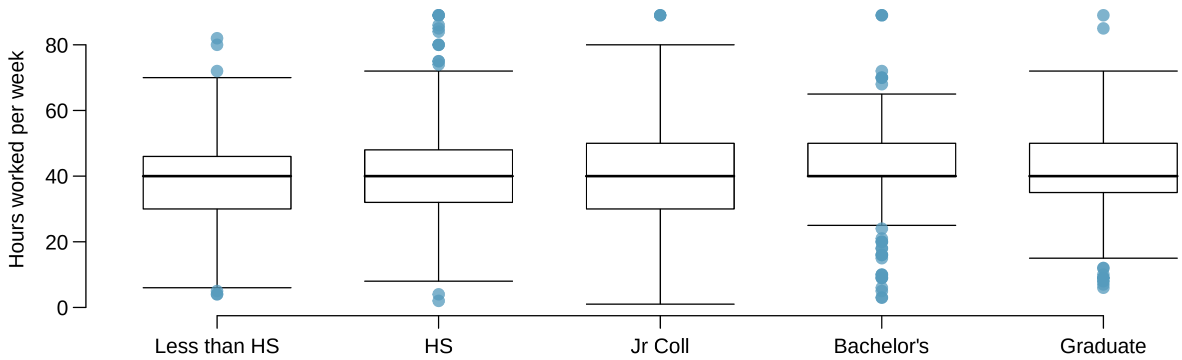
```
Mean <- 4
SD <- 2.2
ME <- 0.5
Z_score <- 1.28
N_Enrollees <- ((Z_score * SD) / ME)
N_Enrollees
```

```
## [1] 5.632
```

They need at least 6 enrollees.

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

Null hypothesis (H0): The averages across all five groups are the same. Alternate hypothesis (H1): The averages across some fo the five groups are different.

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

Answer

The sample taken is random.

The observations are independent.

The sample size is large enough to assume a normal distribution.

(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

```
mu <- c(38.67, 39.6, 41.39, 42.55, 40.85)
sd <- c(15.81, 14.97, 18.1, 13.62, 15.51)
n <- c(121, 546, 97, 253, 155)
data_table <- data.frame (mu, sd, n)
n <- sum(data_table$n)
k <- length(data_table$mu)

# Finding degrees of freedom
df <- k - 1
dfResidual <- n - k

# Using the qf function on the Pr(>F) to get the F-statistic:

Prf <- 0.0682
F_statistic <- qf( 1 - Prf, df , dfResidual)

# F-statistic = MSG/MSE

MSG <- 501.54
MSE <- MSG / F_statistic

# MSG = 1 / df * SSG

SSG <- df * MSG
SSE <- 267382

# SST = SSG + SSE, and df_Total = df + dfResidual

SST <- SSG + SSE
dft <- df + dfResidual
```

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
degree	4	2006.16	501.54	2.188984	0.0682
Residuals	1167	267,382	229.12		
Total	1171	269388.16			

(d) What is the conclusion of the test?

Answer

Since the p-value is high, we cannot reject the Null hypothesis and conclude that there are no differences in the group.