# DATA-606
# Hw5
# Dhananjay Kumar

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.

Solution:

First lets calculate the t statistic

Confidence level 90%, which means alpha level, is 0.10 and df = n-1 = 24

Using qt(1-0.10/2, 24)

T = 1.71

**Sample mean** = 77 + 65/2 = **71**

**Sample standard deviation(s.d):** 77 = 71 + 1.71 * s.d/5

s.d = **17.54**

**Margin of error** = t * standard error

= 1.71 * 17.54/sqrt25

= **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.

**Solution:** a)

Margin error = 25

Using R lets calculate the Z score for alpha 0.10

qnorm(1-0.10/2)

Z = 1.64

Standard error (S.E) = margin error/ z

= 25/1.64

= 15.24

Sample size = (standard deviation/S.E)^2

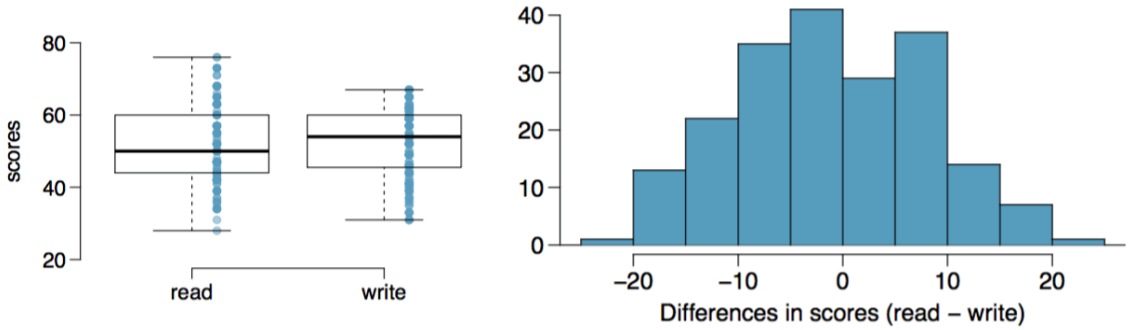
= 270

b) As Luke wants 99 % confidence interval a large sample size would be useful. As the sample size increases the precision of the results also increases

c) Calculating Z score using R:

```
z = qt(1-0.01/2) = 2.57
S.E = 25/2.57 = 9.73
N = (250/9.73)^2
  = 660
```

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.

**Solution: a)** The difference between the average scores of reading and writing are not clearly visible. The difference distribution is though seems normally distributed.

**b)** The reading and writing score might be paired for each student and might not be independent.

**c)** H(null) = There is no difference in the average scores of reading and writing (mu writing = mu reading)

Ha = There is difference in the average reading and writing scores (mu

writing ≠ mu reading)

**d)** A t-test would be needed for this data.

Condition1: The sample should be normally distributed and large, n=200. This condition is satisfied from the histogram.

Condition2: The sample should be random and presumably from less than 10% of the population data

**e)** Average difference = - 0.545

Standard deviation = 8.887

t = -0.545/(8.887/sqrt(200))

= -0.867

df = 199

p value using R:

2*pt(-abs(t),df=199) = 0.387

p >0.05

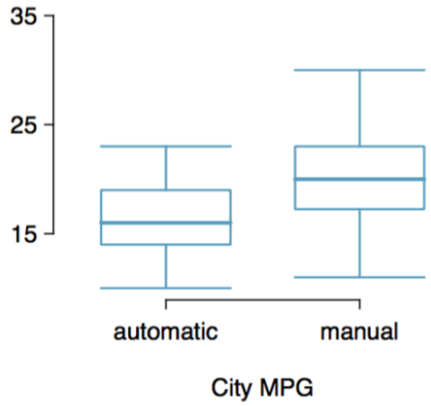
At alpha 0.05 we fail to reject the null hypothesis. This means there is not enough evidence of difference in average reading and writing score

**f)** We might have type 2 error, failing to reject the null hypothesis when in actual the alternate hypothesis is true.

**g)** Yes we may expect the confidence interval to include 0 because the average difference in reading and writing score is -0.545, as it is a two-tailed test.

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 |




**Solution:** This is a 2 sample independent t test

H(null): mu automatic = mu manual

H(a) = mu automatic ≠ mu manual

Standard error = sqrt(s1^2/n1 + s2^2/n2)

Sqrt(3.58^2/25 + 4.5^2/25) = 1.150

T = 19.85-16.12/1.150 = 3.243

P value for this t value is calculated using R

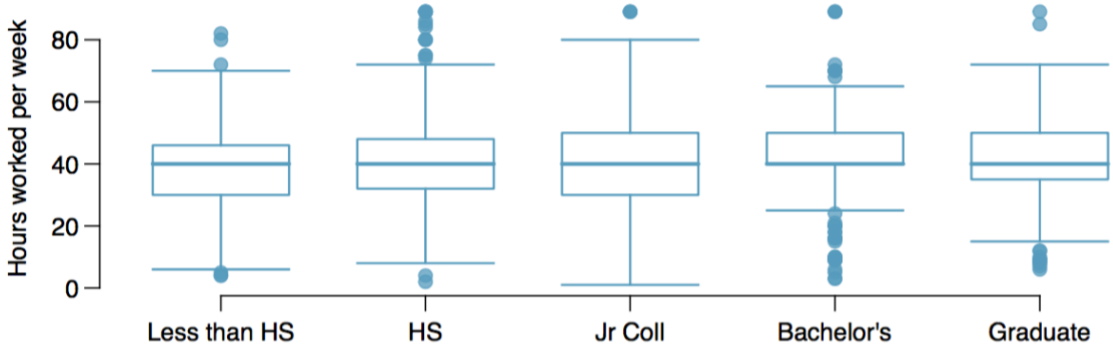
2*pt(-abs(T),df=25) = 0.003

p value < 0.05

We reject the null hypothesis. This means there is strong evidence of fuel efficiency of automatic and manual cars.

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?

**Solution: a)**

Let, mu_1 = less than HS group
Mu_2 = HS group
Mu_3 = Jr Coll group
Mu_4 = Bachelor's group
Mu_5 = graduate group

H (null): mu_1 = mu_2 = mu_3 = mu_4 = mu_5
H (a) = at least one of them is true (mu_1 ≠ mu_2 ≠ mu_3 ≠ mu_4 ≠ mu_5)

**b) Assumptions:**
1) The observations are independent in each and across the group. We assume the data collected in this survey are independent
2) The data is normally distributed in each group. The box plot doesn't support the normality and besides the data is skewed
3) Equal variability across the group. There is similarity of variance across the group.

**c) R code**

```
 n <- 1172
l <- 5

dfG <- l - 1
dfG
dfE <- n-l
dfE

meanT <- 40.45
dfData <- 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( dfData$n * (dfData$mean - meanT)^2 )
# Compute the MSG
MSG <- (1 / dfG) * SSG
# Compute the F statistic
F <- 501.54 / 229.12
F
```

## Meansq Residuals:229.1
## Sumsquare total: 269388.2

## F value = 2.188984

| Anova | Df | Sum sq | Mean sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Degree | **4** | **2006.16** | 501.54 | **2.1889** | 0.0682 |
| Residuals | **1167** | 267382 | **229.12** | | |
| Total | **1171** | **269388.16** | | | |

**d)** The p-value = 0.068 > 0.05, we fail to reject the null hypothesis. We conclude there is not a significant difference between the groups.