charleyferrari_week5hw

Charley Ferrari

October 24, 2015

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.

We can solve backwards from knowing what a confidence interval is:

$$CI = \bar{x} \pm t^*_{24} SE$$

$$65 = \bar{x} - t^*_{24} SE$$

```
xbar <- mean(65,77)

moe <- xbar-65

t <- qt(0.975,24)

SE <- (xbar-65)/t

xbar
```

```
## [1] 65
```

```
moe
```

```
## [1] 0
```

$$SE = \sqrt{\frac{\sigma}{n}}$$

```
n <- 25

sd <- n*SE^2

sd
```

```
## [1] 0
```

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?

$$MOE = t \times \sqrt{\frac{\sigma}{n}}$$

After trying to calculate this with z, it seems Raina would be able to acheive her margin of error using less than 30 observations. Since the t statistic changes depending on the number of observations, we can't just solve this algebraically. I'll calculate a few versions of the margin of error:

```
n <- seq(2,25,by=1)
t <- qt(0.95,n-1)
sd <- 250
se <- sqrt(sd/n)
moe <- t*se

moe
```

```
##  [1] 70.589888 26.655699 18.604972 15.074433 13.007081 11.612723 10.591016
##  [8]  9.800679  9.165565  8.640569  8.197055  7.815848  7.483559  7.190519
## [15]  6.929540  6.695158  6.483131  6.290110  6.113408  5.950843  5.800623
## [22]  5.661257  5.531497  5.410284
```

It looks like the margin of error can be acheived with a sample of 4 (although it probably wouldn't be safe to use this small a sample size.)

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?

There is a clear difference according to the box plots. The average reading score is around 50, while the average writing score is around 55.

b. Are the reading and writing scores of each student independent of each other? They are not, because this data is paired. Each student has a reading and a writing score.

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?

H_0: $\mu_{reading} = \mu_{writing}$, or $\mu_{reading} - \mu_{writing} = 0$ H_a: $\mu_{reading} \neq \mu_{writing}$ or $\mu_{reading} - \mu_{writing} \neq 0$

d. Check the conditions required to complete this test.

Our sample size is large enough at 200, and the differences appears to be relatively normally distributed. We should be able to use z-scores and the normal distribution to perform our inference, but we can use the t-distribution just to be sure.

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?

We're using the second form of the hypotheses I stated in part c.

$$T = \frac{\bar{x} - 0}{\sigma/\sqrt{n}}$$

```
xbar <- -0.545
sd <- 8.887
n <- 200
t <- xbar/(sd/sqrt(n))

pt(t,n-1)
```

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

The p-value isn't below 0.05, so we can't reject the null hypothesis that the averages of the reading and writing scores are the same.

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

We might have made a type II error in this case, since we have failed to reject the null hypothesis. This is a false negative, failing to detect a difference that is present.

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.

I would expect the confidence interval to include 0 because we failed to reject the null hypothesis. Finding a p-value above 0.05 is the same as finding a value within the 95% confidence interval.

Fuel efficiency of manual and automatic cars, Part I

Each year the EPA releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuuel 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.

H_0: $\bar{x}_{auto} - \bar{x}_{manual} = 0$ H_a: $\bar{x}_{auto} - \bar{x}_{manual} \neq 0$

```
xbarauto <- 16.12
xbarman <- 19.85
sdauto <- 3.58
sdman <- 4.51
nauto <- 26
nman <- 26
se <- sqrt(sdauto^2/nauto + sdman^2/nman)
tstat <- (xbarauto-xbarman)/se

pt(tstat,nauto-1)*2
```

```
## [1] 0.002883615
```

the p-value of our t-stat (multiplied by 2 since this is a two-tailed test) is less than 0.05, so we can reject the null hypothesis. It would appear that the average fuel economy of automatic transmissions is significantly less than the average fuel economy of manual 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. 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.

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

H_0: $\mu_{<hs} = \mu_{hs} = \mu_{jc} = \mu_b = \mu_g$ H_a: at least one of these means are different.

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

The observations must be independent. This is a representative survey of US residents that's below 10% of the population, so as long as this data is truly a random sample this condition should hold.

The data must also be approximately normal. The data appears normal enough from the boxplots, and for a sample size of around 1000 this should be usable.

The data must also have comparable standard deviations. This condition also appears to be satisfied looking at the table, with the potential outlier of junior college with a standard deviation of 18.1.

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

Lets work our way backwards. First, lets calculate our two degrees of freedom:

```
k <- 5
n <- 1172
df1 <- k-1
df2 <- n-k
df1
```

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

```
df2
```

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

Now that we know the degrees of freedom, we can use the qf function on the Pr(>F) to get the F-statistic:

```
prf <- 0.0682
fstat <- qf(1-prf,df1,df2)
fstat
```

```
## [1] 2.188931
```

The f-stat is equal to MSG/MSE. We have MST, so we can use that to get MSE:

```
MSG <- 501.54
MSE <- MSG/fstat
```

$MSG = \frac{1}{df_g} SSG$, so we can just multiply MSG by df_g to get SSG:

```
SSG <- df1*MSG
SSE <- df2*MSE
SSG
```

```
## [1] 2006.16
```

```
SSE
```

```
## [1] 267389.5
```

Just to test it out, I calculated SSE in this way as well. It seems to check out!

SST = SSG + SSE, and $df_{total} = df_g + df_e$

```
SST <- SSG + SSE
dft <- df1 + df2
SST
```

```
## [1] 269395.6
```

```
dft
```

```
## [1] 1171
```

d. What is the conclusion of this test?

the p-value is somewhat greater than 0.05, so we do not reject the null hypothesis that the five means are equal to eachother.