DS606_HW5_JagdishChhabria

Jagdish Chhabria

March 21, 2019

Chapter 5 - Inference for Numerical Data Practice: 5.5, 5.13, 5.19, 5.31, 5.45 Graded: 5.6, 5.14, 5.20, 5.32, 5.48

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.

The sample mean should be mid-way between the confidence interval. So it's value is 71. The margin of error is 6 based on the information provided. A 90% confidence interval implies 5% area in each tail. So the t-value would be 1.710882 as shown below. The sample standard deviation is then 17.53.

```
t.val<-qt(0.05, 24, lower.tail = TRUE)

(sd.sample<-(6/t.val)*sqrt(25))
```

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

5.14

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? Based on an assumed normal distribution for the SAT scores, if the margin of error is 25 points and the percentile corresponding to a 5% area (two-sided confidence interval of 90%) is 1.645, then the sample size can be derived as shown below.

```
(n=((1.645*250)/25)^2)
```

```
## [1] 270.6025
```

So Raina should collect a sample of at least 271 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. Since Luke wants a higher confidence interval, it means a 0.05% area in each of the 2 . This would imply a higher sample size since the z-value would be higher, while the population standard deviation and the margin of error remains the same.

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

```
(n=((2.575*250)/25)^2)
```

```
## [1] 663.0625
```

So Luke needs a sample of 664 students.

5.20

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?

```
require(openintro)
```

```
## Loading required package: openintro
```

```
## Please visit openintro.org for free statistics materials
```

```
##
## Attaching package: 'openintro'
```

```
## The following objects are masked from 'package:datasets':
##
##     cars, trees
```

```
summary(hsb2$read)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   44.00   50.00   52.23   60.00   76.00
```

```
summary(hsb2$write)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31.00   45.75   54.00   52.77   60.00   67.00
```

```
(diff.mean=mean(hsb2$read)-mean(hsb2$write))
```

```
## [1] -0.545
```

```
diff.mean
```

```
## [1] -0.545
```

There does not seem to be a clear difference between the average reading and writing scores.

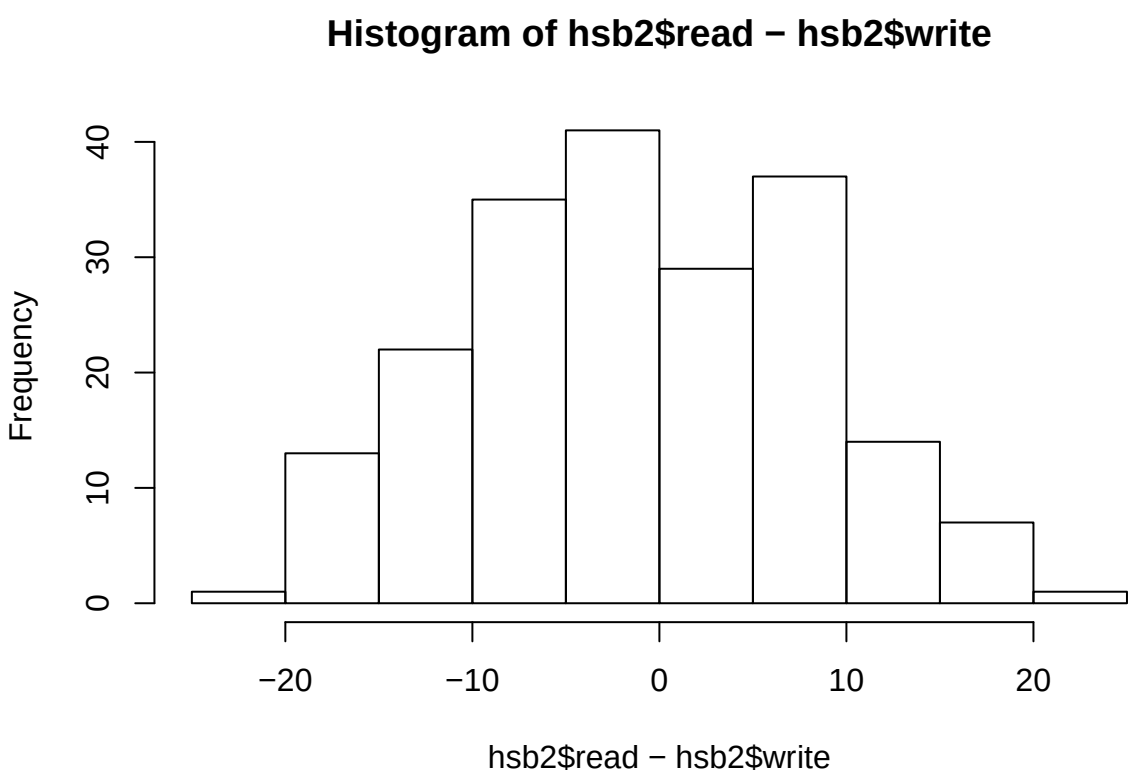
(b) Are the reading and writing scores of each student independent of each other? The reading and writing scores of each student should be positively correlated because typically they would move in tandem based on the student's ability.

(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?

H0: read.avg - write.avg = 0 Ha: read.avg - write.avg <> 0

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

```
hist(hsb2$read-hsb2$write)
```



The underlying distribution seems to be nearly normal. The sample size is greater than 30, but smaller than 10% of overall student population. Scores of individual students should be independent of each other.

(e) The average observed difference in scores is ¯xread???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?

```
std.dev.diff=8.887
std.err.diff=8.887/sqrt(200)
test.stat=1.96
(margin.err.diff=test.stat*std.err.diff)
```

```
## [1] 1.231675
```

```
(lower.diff.95=diff.mean-margin.err.diff)
```

```
## [1] -1.776675
```

```
(upper.diff.95=diff.mean+margin.err.diff)
```

```
## [1] 0.6866754
```

Since the confidence interval contains the null value, it cannot be rejected. We accept the null hypothesis that the difference in average reading and writing scores is 0.

```
z.stat=((0-diff.mean)/std.err.diff)
z.stat
```

```
## [1] 0.867274
```

The corresponding p-value for this is 0.1949, which translates to about 0.39 on a two-tailed basis. Since this

is greater than the significance value, the null hypothesis cannot be rejected.

(f) What type of error might we have made? Explain what the error means in the context of the application. The error that could have been made here is a Type II error i.e. accepting the null hypothesis when it is false. In this context, it means accepting that there is no difference in the mean reading and writing scores, when in fact there is.

(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. No, I would not expect that (in fact it has been proven to not include 0 above), since 0 is the null value and that would mean accepting the null hypothesis.

5.32

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. Automatic Manual Mean 16.12 19.85 SD 3.58 4.51 n 26 26

Since the sample size is less than 30, we use the t-distribution.

```
mean.auto=16.12
mean.manual=19.85
std.dev.auto=3.58
std.dev.manual=4.51
num.cars=26
#inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,         #alternative = "tw
(diff.mean.mpg=mean.auto-mean.manual)
```

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

```
(std.error.mpg=sqrt((((std.dev.auto)^2)/num.cars)+(((std.dev.manual)^2)/num.cars)))
```

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

```
(t.stat.mpg=diff.mean.mpg/std.error.mpg)
```

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

```
df.mpg=num.cars-1
(p.value.mpg=pt(t.stat.mpg,df.mpg))
```

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

Since the p-value is less than 5%, we reject the null hypothesis that the fuel efficiency is the same.

5.48

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. 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

```
mean.less.hs=38.67
std.dev.less.hs=15.81
num.less.hs=121

mean.hs=39.6
std.dev.hs=14.97
num.hs=546

mean.jr.coll=41.39
std.dev.jr.coll=18.1
num.jr.coll=97

mean.bach=42.55
std.dev.bach=13.62
num.bach=253

mean.grad=40.85
std.dev.grad=15.51
num.grad=155

mean.total=40.45
std.dev.total=15.17
num.total=1172

df.groups=5-1
df.total=1172-1
df.err=df.total-df.groups

sse=267382

(ssg=num.less.hs*(mean.less.hs-mean.total)^2 + num.hs*(mean.hs-mean.total)^2 + num.jr.coll*(mean.jr.coll
```

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

```
(sst=sse+ssg)
```

```
## [1] 269386.1
```

```
(msg=ssg/df.groups)
```

```
## [1] 501.0251
```

```
(mse=sse/df.err)
```

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

```
(f.value=msg/mse)
```

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

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

H0: mean.less.hs=mean.hs=mean.jr.coll=mean.bach=mean.grad Ha: At least one of the means is different from the rest

(b) Check conditions and describe any assumptions you must make to proceed with the test. The assumption is that the people in the different educational attainment groups are independent. This seems like a

reasonable assumption given the sample size.

(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 4 2004.101 501.54 2.186745 0.0682 Residuals 1167 267,382 229.1191 Total 1171 269386.1

(d) What is the conclusion of the test? Given that the p-value is higher than the critical value 0f 0.05, we accept the null hypothesis that all the means are same.