1 Lab 1: Significance tests: proportions, correlations, means

1.1 initialize R

Create the folder `Rlab1` inside the PSY710 folder located in your home directory. Then launch R and enter the following commands:

```r
setwd("~/PSY710/Rlab1") # set working directory
source(url("http://psycserv.mcmaster.ca/bennett/psy710/labs/labs2019/L1_19_init.R") ) # load data
```

1.2 Comparing proportions

1. The United States Census Bureau reported that poverty rate in the United States rose from 11.3 percent in 2000 to 11.7 percent in 2001. The data for 2000 came from the census. For the purpose of answering this question, we will assume that the data for 2001 come from a survey of 50,000 randomly selected individuals. Is the increase in poverty rate from 2000 to 2001 statistically significant? We will use the `prop.test` command to answer this question:

```r
sample.size <- 50000;
(observed.frequency <- 0.117 * sample.size); # number of poverty cases

## [1] 5850

null.probability <- 0.113;
prop.test(x=observed.frequency,n=sample.size,p=null.probability,alternative="greater")
```

```
## 1-sample proportions test with continuity correction
##
## data: observed.frequency out of sample.size, null probability null.probability
## X-squared = 7.942, df = 1, p-value = 0.00242
## alternative hypothesis: true p is greater than 0.113
## 95 percent confidence interval:
## 0.114646 1.000000
## sample estimates:
## p
## 0.117
```

We are interested in determining if the poverty rate has increased. Therefore, our null hypothesis is that the poverty rate is 11.3\% or lower. Given this null hypothesis, the highest expected value for the number of individuals in poverty is $0.113 \times 50000 = 5650$. The results of `prop.test` indicate that when the null hypothesis is true, the probability of observing at least 5850 (out of 50000) individuals in poverty, which corresponds to a poverty rate of at least $p = 5850/50000 = 0.117$, is $p = 0.0024$. In other words, if the null hypothesis is true, then our observed poverty rate is unusually high.
2. Repeat the previous test, only this time use `prop.test` to examine the idea that the poverty rate has *changed* rather than *increased*.

```r
prop.test(x=observed.frequency,n=sample.size,p=null.probability,alternative="two.sided")
```

##
## 1-sample proportions test with continuity correction
##
## data: observed.frequency out of sample.size, null probability null.probability
## X-squared = 7.942, df = 1, p-value = 0.00483
## alternative hypothesis: true p is not equal to 0.113
## 95 percent confidence interval:
## 0.114202 0.119857
## sample estimates:
## p
## 0.117
```

3. This question shows how `prop.test` can be used to compare two proportions. Assume that the Census Bureau estimated that the poverty rate in 2002 was 12.1% based on a random sample of 60,000 individuals. We can use `prop.test` to determine if the poverty rates in 2001 and 2002 differed significantly:

```r
p.sample <- c(0.121,0.117); # sample proportions
n.sample <- c(60000,50000) # sample sizes
(counts <- p.sample * n.sample) # poverty cases in each sample

## [1] 7260 5850

prop.test(x=counts,n=n.sample,alternative="two.sided")
```

##
## 2-sample test for equality of proportions with continuity correction
##
## data: counts out of n.sample
## X-squared = 4.119, df = 1, p-value = 0.0424
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## 0.000141502 0.007858498
## sample estimates:
## prop 1 prop 2
## 0.121 0.117
```

The null hypothesis is that the poverty rates are the same in the two samples, and the alternative hypothesis is that the poverty rates differ. The p value, 0.042, means that – when the null hypothesis is true – there is a 4.2% chance of observing a difference between samples (positive or negative) at least as large as the one we observed. How would you test the proposition that the poverty rate *increased* from 2001 to 2002? What would be the null and alternative hypotheses in that case?

```r
p.sample <- c(0.121,0.117); # sample proportions
n.sample <- c(60000,50000) # sample sizes
(counts <- p.sample * n.sample) # poverty cases in each sample

## [1] 7260 5850

prop.test(x=counts,n=n.sample,alternative="greater")
```
4. A new drug therapy is tested. Of 50 patients in the study, 40 had no recurrence of their illness after 18 months. With no drug therapy, the expected percentage of no recurrence would be 75%. Do the data support the hypothesis that the drug was effective?

```r
prop.test(x=40,n=50,p=0.75,alternative="greater")
```

5. Ginkgo biloba extract has been claimed to have a variety of beneficial effects on health and cognition. For example, some have claimed that ginkgo biloba can cure and/or prevent acute mountain sickness (AMS). To test this idea a study took 44 healthy individuals to the Himalayas: have received the extract (80 mg twice per day) and half received placebos. AMS was assessed in each group. The study found evidence of AMS in 18 of the 22 subjects in the placebo group and 3 of the 22 subjects in the ginkgo biloba group. Evaluate the null hypothesis that the proportion of incidence of AMS was the same in the two groups.

```r
p.sample <- c(3/22,18/22);
n.sample <- c(22,22);
ams.counts <- c(3,18)
prop.test(ams.counts,n.sample,alternative="less")
```
1.3 Evaluating correlations

In this section we will use the mtcars data set to calculate correlations and perform a linear regression analysis. To read about the data set, type `?mtcars` at the command line prompt. Then use the following commands to load and inspect the data:

```r
tmp <- mtcars # load data set into R
myData <- tmp[,c("mpg","disp","hp","wt")]
  # save these 4 columns
class(myData)
## [1] "data.frame"
```

```r
summary(myData)
```

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>10.4</td>
<td>15.4</td>
<td>19.2</td>
<td>20.1</td>
<td>22.8</td>
<td>33.9</td>
</tr>
<tr>
<td>disp</td>
<td>71.1</td>
<td>120.8</td>
<td>196.3</td>
<td>230.7</td>
<td>326.0</td>
<td>472.0</td>
</tr>
<tr>
<td>hp</td>
<td>52.0</td>
<td>96.5</td>
<td>123.0</td>
<td>146.7</td>
<td>180.0</td>
<td>335.0</td>
</tr>
<tr>
<td>wt</td>
<td>1.51</td>
<td>2.58</td>
<td>3.33</td>
<td>3.22</td>
<td>3.61</td>
<td>5.42</td>
</tr>
</tbody>
</table>

Next we plot mileage (miles per gallon) vs. weight. The following commands all create the plot shown in Figure 1a.

```r
plot(x=myData$wt,y=myData$mpg,type='p',xlab="weight (1000 lbs)",ylab="MPG")
with(myData,plot(x=wt,y=mpg,type='p',xlab="weight (1000 lbs)",ylab="MPG"))
plot(mpg~wt,data=myData,type='p',xlab="weight (1000 lbs)",ylab="MPG")
```

Next we can use linear regression to compute the best-fitting (least-squares) line to the data:

```r
lin.reg.model <- lm(mpg~wt,data=myData)
summary(lin.reg.model)
```

```
## Call:
## lm(formula = mpg ~ wt, data = myData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -4.543 -2.365 -0.125 1.410 6.873
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.285 1.878 19.86 < 2e-16 ***
## wt -5.344 0.559 -9.56 1.3e-10 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.05 on 30 degrees of freedom
## Multiple R-squared: 0.753, Adjusted R-squared: 0.745
## F-statistic: 91.4 on 1 and 30 DF, p-value: 1.29e-10
```

The regression table indicates the the slope of the line, -5.34, differs significantly from zero. Also, \( R^2 = 0.75 \), which means that the regression accounts for 75\% of the variance in mileage, which is statistically significant \( F(1,30) = 91.38, p < 0.001 \). Finally, we can add the regression line to the scatter plot using `abline` to create Figure 1b. You can see that the line fits the data reasonably well, which is consistent with the quantitative tests presented in the regression table.
Finally, we can compute the Pearson (r) and Spearman (rho) correlations between mileage and weight:

```r
with(myData, cor.test(x=wt, y=mpg, method="pearson"))
```

```r
##
## Pearson's product-moment correlation
##
## data: wt and mpg
## t = -9.559, df = 30, p-value = 1.29e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.933826  -0.744087
## sample estimates:
##   cor
##  -0.867659
```

```r
with(myData, cor.test(x=wt, y=mpg, method="spearman", exact=F))
```

```r
##
## Spearman's rank correlation rho
##
## data: wt and mpg
## S = 10290, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##   rho
##  -0.886422
```

**Task:** The data frame `corDataSet` contains two quantitative variables named `x` and `y`. Evaluate the null hypothesis that `x` and `y` are not correlated.

```r
with(corDataSet, cor.test(x, y)) # Pearson r
```

```r
##
## Pearson's product-moment correlation
##
## data: x and y
## t = 2.196, df = 73, p-value = 0.0313
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0232458  0.4504127
## sample estimates:
##   cor
##  0.248895
```

```r
summary(lm(y~x, data=corDataSet)) # linear regression
```

```r
##
## Call:
## lm(formula = y ~ x, data = corDataSet)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2.318 -0.517 -0.050  0.629  3.777
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
```
Figure 1: a) Plot of mileage (miles per US gallon) against car weight. b) Same as (a) except that the best-fitting (least-squares) line has been drawn through the data. Data come from R’s `mtcars` data set.

```r
with(corDataSet, cor.test(x, y, method = "spearman")) # Spearman rho
```

```r
tmp <- corDataSet[corDataSet$x < 5, ] # remove outlier
with(tmp, cor.test(x, y))
```

```r
with(corDataSet, cor.test(x, y))
```
1.4 Evaluating means (t tests)

The variable iq contains IQ scores from 20 individuals. The following commands compute summary statistics of the data:

```r
summary(iq)
#> Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>  76.4   90.3   93.9   95.3  101.4  113.7
```

```r
IQR(iq)
#> [1] 11.0588
```

```r
sd(iq)
#> [1] 10
```

The boxplot commands (shown below) create the plot in Figure 3. A box plot is a useful graphical summary of a data distribution and we will be using them a lot in this course. From this plot we can determine that the median is approximately 94, the inter-quartile range is about 12 (i.e., 102-90), and that the extreme scores are approximately 75 and 115.
boxplot(iq,ylab="IQ")
axis(side=2,at=seq(80,110,5),labels=F,tcl=-0.25)
axis(side=2,at=seq(80,110,1),labels=F,tcl=-0.125)
abline(h=median(iq),lty=3)

Figure 3: Boxplot of data in iq.

Finally, we use a $t$ test to examine the idea that the IQ scores were drawn from a population with a mean $\mu = 100$.

```r
> t.test(iq,mu=100)

# One Sample t-test

## data: iq
## t = -2.093, df = 19, p-value = 0.05
## alternative hypothesis: true mean is not equal to 100
## 95 percent confidence interval:
## 90.6397 100.0000
## sample estimates:
## mean of x
## 95.3199
```

1. What are the null and alternative hypotheses that are begin evaluated by this $t$ test?
2. What is your conclusion regarding the null hypothesis?
3. What is the 95% confidence interval?
4. Suppose we replicated this IQ study, using the same methods, sample size, etc. Given the result of this \( t \) test, what do you think is the probability that the replication would also reject the null hypothesis?

\[
n \leftarrow 20 \quad \# \text{sample size} \\
m \leftarrow \text{mean}(\text{iq}) \quad \# \text{mean of population} \\
sd \leftarrow \text{sd}(\text{iq}) \quad \# \text{sd of population} \\
B \leftarrow 10000 \quad \# \text{number of simulated experiments} \\
p.\text{val} \leftarrow \text{rep}(0,B) \quad \# \text{array to store } p \text{ values} \\
t.\text{val} \leftarrow \text{rep}(0,B) \quad \# \text{array to store } t \text{ values} \\
\]

```r
for(kk in 1:B){
    simDat <- rnorm(n,m,sd);  # draw random sample
    t.results <- t.test(simDat,mu=100,alternative="two.sided")
    t.val[kk] <- t.results$statistic  # store t value
    p.val[kk] <- t.results$p.value  # store p value
}
p.sig <- (p.val < .05)  # determine statistical significance
sum(p.sig)/length(p.sig)  # proportion of significant p values
## [1] 0.5087
```

5. Next, consider the case where the IQ scores are selected from a population with a mean of 100. In other words, the null hypothesis (\( \mu = 100 \)) is true. If we were to conduct our IQ experiment many times, how do you think our \( t \) and \( p \) values would be distributed? To answer this question, re-do the simulation used in the previous answer but set the value of \( m \) to 100, and then plot the \( t \) and \( p \) values in separate histograms.

\[
n \leftarrow 20 \quad \# \text{sample size} \\
m \leftarrow 100 \quad \# \text{mean of population is 100} \\
sd \leftarrow \text{sd}(\text{iq}) \quad \# \text{sd of population} \\
B \leftarrow 10000 \quad \# \text{number of simulated experiments} \\
p.\text{val} \leftarrow \text{rep}(0,B) \quad \# \text{array to store } p \text{ values} \\
t.\text{val} \leftarrow \text{rep}(0,B) \quad \# \text{array to store } t \text{ values} \\
\]

```r
for(kk in 1:B){
    simDat <- rnorm(n,m,sd);  # draw random sample
    t.results <- t.test(simDat,mu=100,alternative="two.sided")
    t.val[kk] <- t.results$statistic  # store t value
    p.val[kk] <- t.results$p.value  # store p value
}
p.sig <- (p.val < .05)  # determine statistical significance
sum(p.sig)/length(p.sig)  # proportion of significant p values
## [1] 0.0514
```
Figure 4: Histogram of $t$ values.

Figure 5: Histogram of $p$ values.