Problem Set 6 Ethan Engel

C. Durso

These questions were rendered in R markdown through RStudio (<https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>, <http://rmarkdown.rstudio.com> ).

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .doc, .docx, or .pdf document. Your solution document should have your answers to the questions and should display the requested plots.

## 1. (Z-test power calculation)

Power analysis at the phase of the design of an experiment often involves calculating the number of cases needed to reject the null hypothesis at a specified level, such as , with a particular probability, such as , given a specific alternative model. This allows researchers to set the sample size to have a reasonable probability of rejecting the null hypothesis if their estimate of the true model is approximately correct.

Suppose the capacity of the old version of a battery is known to be milliampere hours. You believe that the capacity of the new version is milliampere hours. The method you use for measurement of capacity is known to have the distribution where is the true capacity. Each measurement is an independent sample from .

There is consensus that the capacity of the new version is at least as large as the old version, so a 1-sided test of significance is acceptable.

### 1a.

Let the null model be that the true capacity is milliampere hours (mAh) and that measurements are Normally distributed around this with standard deviation equal to mAh . Let the actual capacity be mAh. Suppose you make independent measurements, calculate the Z-statistic, and perform a 1-sided Z-test of the null hypothesis that sample comes from the null model. How large must the sample mean be for the p-value to satisfy ? Please repeat for (10 points)

z1 <- (5300-5000)/(500/sqrt(16))  
z1

## [1] 2.4

1-pnorm(5300, mean=5000, sd=125)

## [1] 0.008197536

minXbar1 <- qnorm(.99)\*125+5000  
minXbar1

## [1] 5290.793

z2 <- (5300-5000)/(500/sqrt(49))  
z2

## [1] 4.2

1-pnorm(5300, mean=5000, sd=(500/7) )

## [1] 1.334575e-05

minXbar2 <- qnorm(.99)\*(500/7)+5000  
minXbar2

## [1] 5166.168

### 1b.

Again, let the null model be that the true capacity is milliampere hours (mAh) and that measurements are Normally distributed around this with standard deviation equal to mAh . Let the actual capacity be mAh. Suppose you make n independent measurements, calculate the Z-statistic, and perform a 1-sided Z-test of the null hypothesis that sample comes from the null model. What is the probability that the p-value for will satisfy for in ? You may use your result from 1a. (5 points)

*In either case, whether n=16 or n=49, or any value in between, a sample mean of 5300 milliampere hours is sufficient to reject the null hypothesis at the alpha=.01 level. There is compelling statistical evidence, in either case, that the actual mu is greater than 5000 milliampere hours. The probability in that interval is 1.*

### 1c.

What is the smallest sample size that results in a probability of at least 0.8 of rejecting the null hypothesis? You may solve this by calculating the probability of rejecting the null hypothesis for a range of sample sizes.(10 points)

*In this situation, for mu null=5000 milliampere hours and xbar=5300, the lowest sample size which produces a p value<.2 is n=2. These is is illustrated with calculations below. When p<.2, that means that the probability of rejecting the null hypothesis is >.8.*

n1pvalue <- 1-pnorm(5300, mean=5000, sd=500)   
n1pvalue

## [1] 0.2742531

sd2 <- 500/sqrt(2)  
n2pvalue <- 1-pnorm(5300, mean=5000, sd=sd2)  
n2pvalue

## [1] 0.198072

sd3 <- 500/sqrt(3)  
n3pvalue <- 1-pnorm(5300, mean=5000, sd=sd3)   
n3pvalue

## [1] 0.1493488

## 2. (Confidence Intervals)

Consider the problem of estimating a parameter for a sample from a distribution known to be in a parametrized family. A 95% confidence interval, for example, for the parameter of interest is an interval calculated in such a way that the interval calculated from sample has a probability of .95 of having the true value of the parameter, say, satisfy .

The purpose of this exercise is to illustrate the concept of a confidence interval. You will simulate multiple Normally distributed data sets, calculate their 95% confidence intervals based on the Student’s t-distribution, and compare the observed coverage proportion to the theoretical proportion.

Then you will repeat this with data sets produced by rounding Normally distributed data. This examines how rounding data affects the coverage of confidence intervals based on Student’s t. That is, can these 95% confidence intervals for rounded data be interpreted as using a method that has a .95 probability of including the true mean?

### 2a.

The code below generates 10,000 samples of size 15 from the Normal distribution with and . For what proportion of these samples does the 95% confidence interval for from the Student’s t test include the the true value of ? (The Student’s t-test is implemented in R as “t.test”. This is used here to calculate the 95% confidence interval.) How does this relate to the concept of confidence interval? (You may find that rerunning the test with different seeds helps you address this question.) (5 points)

*When repeatedly generating 95% confidence intervals, we expect, in the long run, for the true mu to be captured 95% of the time. In the sample of 10,000 here, the actual mu was captured in 94.73% of the simulations.*

set.seed(12345)  
  
mat<-matrix(rnorm(10000\*15,5,2),ncol=15)  
  
pvalue <- rep(0,nrow(mat))  
  
   
for (i in 1:10000){  
 pvalue[i] <- t.test(mat[i,1:15],mu=5)$p.value  
}  
  
(10000-sum(pvalue<.05))/10000

## [1] 0.9473

### 2b.

Please repeat the analysis in question 2a with same set of samples, but with the values rounded to 1 decimal place. What change do you observe? (5 points)

*The proportion of simulations in which the actual mu was captured increased slightly, to 94.74%*

mat.1<-round(mat,1)  
  
pvalue2 <- rep(0,nrow(mat.1))  
  
   
for (i in 1:10000){  
 pvalue2[i] <- t.test(mat.1[i,1:15],mu=5)$p.value  
}  
  
(10000-sum(pvalue2<.05))/10000

## [1] 0.9474

### 2c.

Please repeat the analysis in question 2 with same set of samples, but with the values rounded to the nearest . What change do you observe? (10 points)

*The proportion of simulations in which the actual mu was captured again increased slightly, this time to 94.75%*

mat.5<-round(mat\*2,0)/2  
  
pvalue3 <- rep(0,nrow(mat.5))  
  
   
for (i in 1:10000){  
 pvalue3[i] <- t.test(mat.5[i,1:15],mu=5)$p.value  
}  
  
(10000-sum(pvalue3<.05))/10000

## [1] 0.9475

If you’re curious, the code below shows that, in this case, the rounding has little effect on the length of the confidence intervals, though the difference is statistically significant.

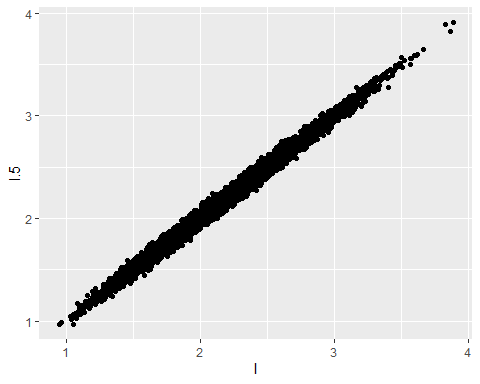
ci.length<-function(x,mu){  
 int.ci<-t.test(x)$conf.int  
 return(int.ci[2]-int.ci[1])  
}  
  
l.5<-(apply(mat.5,1,ci.length,mu=5))  
l<-(apply(mat,1,ci.length,mu=5))  
  
mean(l.5)

## [1] 2.180342

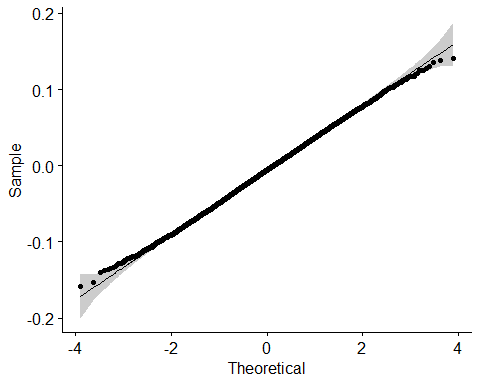
mean(l)

## [1] 2.174353

qplot(l,l.5)



ggqqplot(l-l.5)



t.test(l-l.5)

##   
## One Sample t-test  
##   
## data: l - l.5  
## t = -14.146, df = 9999, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.006818716 -0.005158966  
## sample estimates:  
## mean of x   
## -0.005988841

## 3. (Galton data)

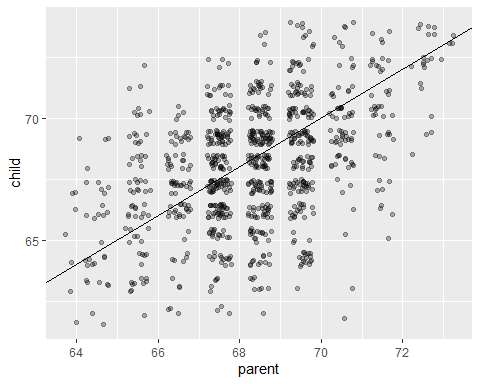
Francis Galton, a contemporary and a relative of Charles Darwin, made some groundbreaking analyses of characteristics of human populations. He also held some racist views and espoused eugenics, inventing the word. His biography is interesting reading, but not required for this problem. The “Galton” data set in the “HistData” package has his measurements of a population of parents and adult offspring. Details are available in the help for “Galton”.

### 3a.

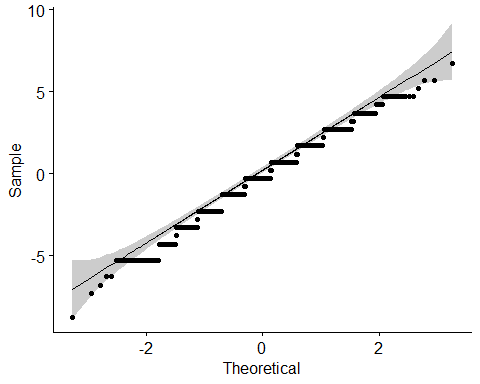
Consider the data frame “dat” below. Perform a visual check of whether the value of “child” minus the value of “parent” could be considered Normally distributed, allowing for the rounding of heights. The function “ggqqplot” in the “ggpubr” package may help. (5 points)

*If we could adjust the values to truly represent the continuity of height differences without the rounding, then, arguably we would have the symmetry and unimodality that we would need for the presumption of normality.*

data("Galton")  
ggplot(data=Galton,aes(x=parent,y=child))+geom\_jitter(height=.3,width=.3,alpha=.3)+  
 geom\_abline(slope=1,intercept=0)



ggqqplot(Galton$child-Galton$parent)



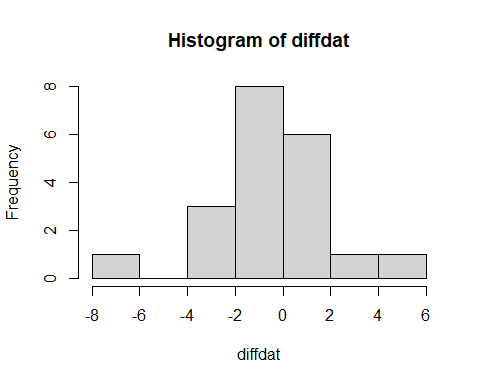
set.seed(234567)  
ind<-sample(1:nrow(Galton),20)  
dat<-Galton[ind,]

### 3b.

Test the hypothesis that this sample of child-parent is drawn from a Normal distribution with mean equal to 0 using Student’s t. Please provide your interpretation of the results. In your interpretation, please address the question of the extent to which the data satisfy the hypotheses for the t-test. (5 points)

*The t-test generated a p value of .5, which is too high to reject the null hypothesis at any realistic alpha level. So there is not evidence to dispute the null hypothesis, that mu=0. Zero is also well within the 95% confidence interval. There is the one caveat that that data did not perfectly conform to the normality in the ggqqplot. However, when we generate a histogram of this sample, it appears to be unimodal and symmetric enough to presume normality.*

diffdat <- c(dat$child-dat$parent)  
hist(diffdat)



t.test(diffdat)

##   
## One Sample t-test  
##   
## data: diffdat  
## t = -0.68763, df = 19, p-value = 0.5  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -1.415345 0.715345  
## sample estimates:  
## mean of x   
## -0.35