Problem Set 6

Adam Ten Hoeve

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)

mu\_0 = 5000  
mu\_1 = 5300  
sigma = 500  
n = 16  
  
# Calculate the Z-statistic by standardizing the normal  
Z = (mu\_1 - mu\_0) / sqrt(sigma^2 / n)  
Z

## [1] 2.4

# We know that the new batteries are at least as powerful than the old batteries  
# so we want to test if they're significantly better than the old batteries: mu\_1 > mu\_0  
# To test this, we can use a right tailed test, which is 1-CDF(Z)  
1-pnorm(Z)

## [1] 0.008197536

# For a p-value of 0.01, what is the xbar?  
# Z' = CDF(1 - 0.01)  
# Z' = (xbar - mu\_0) / sqrt(sigma^2 / n)  
# => xbar = (z \* sigma / sqrt(n)) + mu\_0  
p = 0.01  
Z\_prime = qnorm(1 - p)  
xbar\_16 = (Z\_prime \* sqrt(sigma^2 / n)) + mu\_0  
xbar\_16

## [1] 5290.793

cat("When n=16: Z-stat=", Z, "with p-value:", 1-pnorm(Z), "\n")

## When n=16: Z-stat= 2.4 with p-value: 0.008197536

cat("When n=16, for a p-value of 0.01, we need xbar=", xbar\_16)

## When n=16, for a p-value of 0.01, we need xbar= 5290.793

# Repeat when n=49  
# Calculate the Z-Statistic  
n = 49  
Z\_49 = (mu\_1 - mu\_0) / sqrt(sigma^2 / n)  
Z\_49

## [1] 4.2

# Calculate the P-Value for a 1-sided test  
1 - pnorm(Z\_49)

## [1] 1.334575e-05

# Find the xbar when p=0.01  
xbar\_49 = (Z\_prime\*sqrt(sigma^2/n)) + mu\_0  
xbar\_49

## [1] 5166.168

cat("When n=49: Z-stat=", Z\_49, "with p-value:", 1-pnorm(Z\_49), "\n")

## When n=49: Z-stat= 4.2 with p-value: 1.334575e-05

cat("When n=49, for a p-value of 0.01, we need xbar=", xbar\_49)

## When n=49, for a p-value of 0.01, we need xbar= 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)

# Calculate the p-val for a 1-sided T-test using the theoretical xbars.  
# When n=16  
prob\_16 = 1 - pnorm(xbar\_16, mean=mu\_1, sd=sqrt(sigma^2 / n))  
  
# When n=49  
prob\_49 = 1 - pnorm(xbar\_49, mean=mu\_1, sd=sqrt(sigma^2 / n))  
  
cat("When n=16, the probability that the p-value <= 0.01 is", prob\_16, "\n")

## When n=16, the probability that the p-value <= 0.01 is 0.5512781

cat("When n=49, the probability that the p-value <= 0.01 is", prob\_49, "\n")

## When n=49, the probability that the p-value <= 0.01 is 0.9695108

### 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)

# We saw from part b that n=16 resulted in prob<0.8 and n=49 resulted  
# in prob>0.8, so let's iteratre through all n values between them to find  
# which is the smallest that results in prob>=0.8.  
n = 16:49  
xbars = Z\_prime \* sqrt(sigma^2 / n) + mu\_0  
probs = 1 - pnorm(xbars, mean=mu\_1, sd=sqrt(sigma^2 / n))  
  
# Find the first n which has a prob>=0.8  
n.index = which(probs >= 0.8)[1]  
n[n.index]

## [1] 28

cat("The smallest sample size that has a probability of at least a 0.8 for rejecting the null is", n[n.index], "with probability", probs[n.index])

## The smallest sample size that has a probability of at least a 0.8 for rejecting the null is 28 with probability 0.8019352

## 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)

set.seed(12345)  
  
mat<-matrix(rnorm(10000\*15,5,2),ncol=15)  
  
# Initialize the variables  
mu = 5  
var = 2  
n = 15  
# Find the critical values for a T-Dist with 95% confidence  
t.val = qt(0.975, df=n-1)  
  
# Calculate the sample means and variances for each sample  
xbars = rowMeans(mat)  
sample.vars = apply(mat, 1, var)  
  
# Calculate the CI for each sample  
LHS = xbars - t.val\*sqrt(sample.vars / n)  
RHS = xbars + t.val\*sqrt(sample.vars / n)  
df = data.frame(LHS=LHS, RHS=RHS)  
  
# Find which CIs contained the true mean  
contains.mean = (LHS < mu) & (mu < RHS)  
# Find the proportion that contain the mean  
cat("The proportion of sampled CIs that contained the true mean is: ", sum(contains.mean) / length(contains.mean))

## The proportion of sampled CIs that contained the true mean is: 0.9473

The confidence of the confidence interval tells us how likely our interval will contain the true mean. That means that if we have a single sample and calculate the confidence interval, then there is still a chance that the true mean is not contained in the interval.

### 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)

set.seed(12345)  
  
mat<-matrix(rnorm(10000\*15,5,2),ncol=15)  
  
# Round the sample to the first decimal place  
mat=round(mat, digits=1)  
  
# Initialize the variables  
mu = 5  
var = 2  
n = 15  
# Find the critical values for a T-Dist with 95% confidence  
t.val = qt(0.975, df=n-1)  
  
# Calculate the sample means and variances for each sample  
xbars = rowMeans(mat)  
sample.vars = apply(mat, 1, var)  
  
# Calculate the CI for each sample  
LHS = xbars - t.val\*sqrt(sample.vars / n)  
RHS = xbars + t.val\*sqrt(sample.vars / n)  
  
# Find which CIs contained the true mean  
contains.mean = (LHS < mu) & (mu < RHS)  
# Find the proportion that contain the mean  
cat("The proportion of sampled CIs, when rounded to the nearest 0.1, that contained the true mean is: ", sum(contains.mean) / length(contains.mean))

## The proportion of sampled CIs, when rounded to the nearest 0.1, that contained the true mean is: 0.9474

The proportion of confidence intervals that contain the true mean is basically the same. When we rounded, the proportion of CI that contained the true mean 0.0001 more than when we didn’t round.

### 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)

set.seed(12345)  
  
mat<-matrix(rnorm(10000\*15,5,2),ncol=15)  
  
# Round the sample to the nearest 0.5  
mat= ceiling(mat\*2) / 2  
  
# Initialize the variables  
mu = 5  
var = 2  
n = 15  
# Find the critical values for a T-Dist with 95% confidence  
t.val = qt(0.975, df=n-1)  
  
# Calculate the sample means and variances for each sample  
xbars = rowMeans(mat)  
sample.vars = apply(mat, 1, var)  
  
# Calculate the CI for each sample  
LHS = xbars - t.val\*sqrt(sample.vars / n)  
RHS = xbars + t.val\*sqrt(sample.vars / n)  
  
# Find which CIs contained the true mean  
contains.mean = (LHS < mu) & (mu < RHS)  
# Find the proportion that contain the mean  
cat("The proportion of sampled CIs, when rounded to the nearest 0.5, that contained the true mean is: ", sum(contains.mean) / length(contains.mean))

## The proportion of sampled CIs, when rounded to the nearest 0.5, that contained the true mean is: 0.9193

When we round to the nearest , the proportion of Confidence Intervals is smaller than when we didn’t round, by about 0.3.

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.

set.seed(12345)  
  
mat<-matrix(rnorm(10000\*15,5,2),ncol=15)  
  
# Round the sample to the nearest 0.5  
mat= ceiling(mat\*2) / 2  
  
# Initialize the variables  
mu = 5  
var = 2  
n = 15  
# Find the critical values for a T-Dist with 95% confidence  
t.val = qt(0.975, df=n-1)  
  
# Calculate the sample means and variances for each sample  
xbars = rowMeans(mat)  
sample.vars = apply(mat, 1, var)  
  
# Calculate the CI for each sample  
LHS = xbars - t.val\*sqrt(sample.vars / n)  
RHS = xbars + t.val\*sqrt(sample.vars / n)  
  
# Find which CIs contained the true mean  
contains.mean = (LHS < mu) & (mu < RHS)  
# Find the proportion that contain the mean  
sum(contains.mean) / length(contains.mean)

## [1] 0.9193

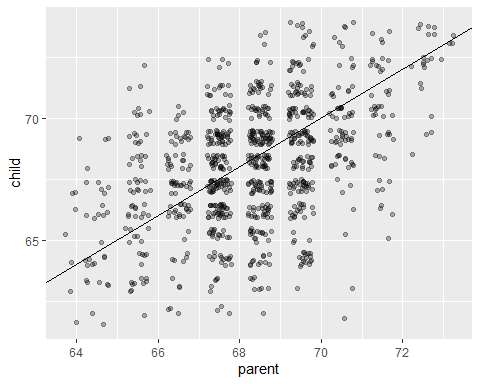
## 3. (Galton data)

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”.

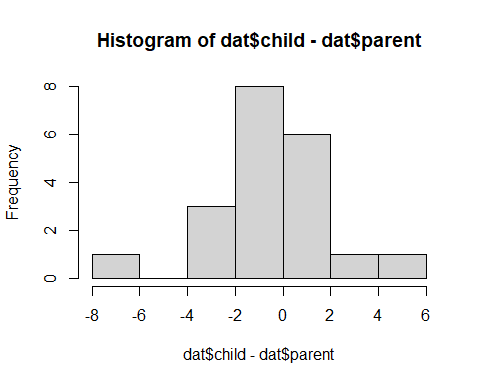
### 3a.

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)

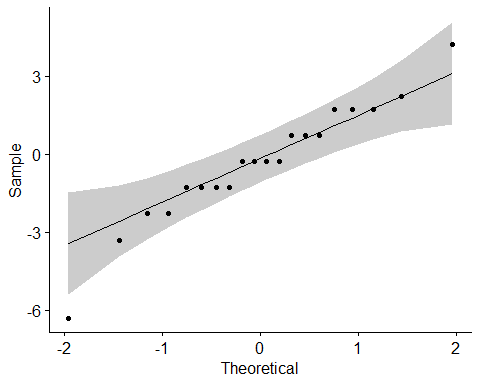
data("Galton")  
ggplot(data=Galton,aes(x=parent,y=child))+geom\_jitter(height=.3,width=.3,alpha=.3)+  
 geom\_abline(slope=1,intercept=0)



set.seed(234567)  
ind<-sample(1:nrow(Galton),20)  
dat<-Galton[ind,]  
  
# Check if difference appears to be normally distributed visually  
hist(dat$child - dat$parent)



# Check using a QQPlot  
ggqqplot(dat$child-dat$parent)



### 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)

t.test(dat$child-dat$parent)

##   
## One Sample t-test  
##   
## data: dat$child - dat$parent  
## 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

From the t-test of the difference between child and parent, we get a p-value of . Assuming a 95% confidence level, this is much larger than our target p-value of . Therefor, we fail to reject the null and can assert that the true mean is still .