Problem Set 4

C. Durso

February 25, 2019

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 work should be based on the data's being in the same folder as the .Rmd file. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

1. (The Trial of the Pyx) The London Mint, by the 1200's and through the 1800's, conducted quality control tests on the minted gold coins, both for purity and weight. The standards made allowance for variability in the weight of the minted coins. The standards for weight were in terms of a target weight T and a "remedy" R. A sample of n coins produced during a period would be weighed together as part of the Trial of the Pyx. The weight was required to be at least $nT - nR$. At one point in the history of the mint, the remedy was determined so that 95% of coins in a calibration set would have weights greater than or equal to the mean of the set minus R. http://www.stat.wisc.edu/sites/default/files/TR442_0.pdf The questions below model the variability of the weights of the minted coins as independently Normally distributed. The goal is to use variance calculations and variable transformations to understand how difficult the Trial was and by how much the Master of the Mint could reduce the mean weight of the coins and still be confident of passing the Trial. As a hint, there is some suspicion that the Master of the Mint was expected to supplement the salary of the position in this way. This is an example with medieval roots of the role of statistical analysis in process control and the impact of using unsuitable measures to assess whether the process is in control. (5 points each)

   a. Assuming that the size of the calibration set is m and the weights of coins are independent identically distributed samples $(x_1, x_2, ...x_m)$ from Normal distribution with mean μ and standard deviation σ, what is the distribution of the random variable $X_i - \bar{X}$ where $\bar{X} = \frac{1}{m}\sum_{j=1}^{m} X_j$? You may use the fact that the sum of independent Normal distributions is Normal, but note that X_i and $\bar{X}$ are not independent.

We replace the expression of $\bar{X}$: $X_i - \frac{1}{m}\sum\limits_{j=1}^{m} X_j$

Since these two random variables are not independent, we "extract" the element in common to make them independent : $(1-\frac{1}{m})X_i - \frac{1}{m}\sum\limits_{j=1,j\neq i}^{m} X_j$

We can deduce that the mean is : $\mu_i = \left((1-\frac{1}{m}) + (\frac{m-1}{m^2})\right)\mu = 0$

We obtain : $\sigma_i = (1-\frac{1}{m})^2\sigma^2 + \frac{m-1}{m^2}\sigma^2 = \frac{(m-1)(m)}{m^2}\sigma^2$

Therefore we have : $X_i - \bar{X} \sim Norm(\mu = 0, sd = \frac{(m-1)}{m}\sigma^2)$

   b. If m is sufficiently large then $\bar{X}$ can be taken to be essentially equal to μ. What value of R satisfies the condition that $P\left(X_i > \mu - R\right) = .95$?

$P\left(X_i > \mu - R\right) = P\left(\frac{X_i-\mu}{\sigma_i} > \frac{\mu-R}{\sigma_i}\right) = P\left(Z > \frac{R}{\sigma_i}\right)$

```
R=-qnorm(0.05)
R
```

```
## [1] 1.644854
```

```r
# Or
R=qnorm(0.95, 0, 1)
R
```

```
## [1] 1.644854
```

```
c. Setting $T=0.28$, $R=0.002$, $n=10,000$, and $\sigma=0.001$, what is the probability of the sample's
```

```r
t<-0.28
R<-0.002
sigma<-0.001
n<-10000
mu<-.27805

# This is the probability of failure with the distribution transformed to the standard normal (not requ
pnorm(sqrt(n)*(t-mu-R)/sigma)
```

```
## [1] 2.866516e-07
```

```r
# This uses the paramenters given :
pnorm(n*t-n*R,sd = sqrt(n)*sigma, mean = n*mu)
```

```
## [1] 2.866516e-07
```

```
d. Given a fixed value of $R$ and a random sample of $n$ coins, what value of $\mu$ in terms of $T$, $R
The distribution of the weight $W$ is $Normal\left(n\mu,n\sigma^2\right)$. This requires $P\left(\frac{
```

```r
a<-qnorm(.0001)
t-R-a*sigma/sqrt(n)
```

```
## [1] 0.2780372
```

2. (Confidence Intervals) The purpose of this exercise is to illustrate the concept of a confidence interval. Consider the problem of estimating a parameter for a sample from a distribution known to be in a parametrized family. The 95% confidence interval, for example, for the parameter of interest is an interval calculated in such a way that the interval (a_X, b_X) calculated from sample $\boldsymbol{X}$ has a probability of .95 of having the true value of the parameter, c say, satisfy $c \in (a_X, b_X)$. The code below generates 10,000 samples of size 15 from the Normal distribution with $\mu = 5$ and $\sigma = 2$. For what proportion of these samples does the 95% confidence interval for $\mu$ from the Student's t test include the true value of $\mu$? (The Student's t-test is implemented in R as "t.test". You may use this.) 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.)(10 points)

```r
set.seed(12345)
ci.check<-function(x,mu){
  int.ci<-t.test(x)$conf.int
  return(mu>=int.ci[1]&mu<=int.ci[2])
}
mat<-matrix(rnorm(10000*15,5,2),ncol=15)
mean(apply(mat,1,ci.check,mu=5))
```

```
## [1] 0.9473
```

The mean of each row was in the 95% confidence interval that we obtained from the t-test, supporting the accuracy of the confidence interval.

3. Please repeat the analysis in question 2 with same set of samples, but with the values rounded to 1 decimal place. What change do you observe? (10 points)

```
mat.1<-round(mat,1)
mean(apply(mat.1,1,ci.check,mu=5))
```

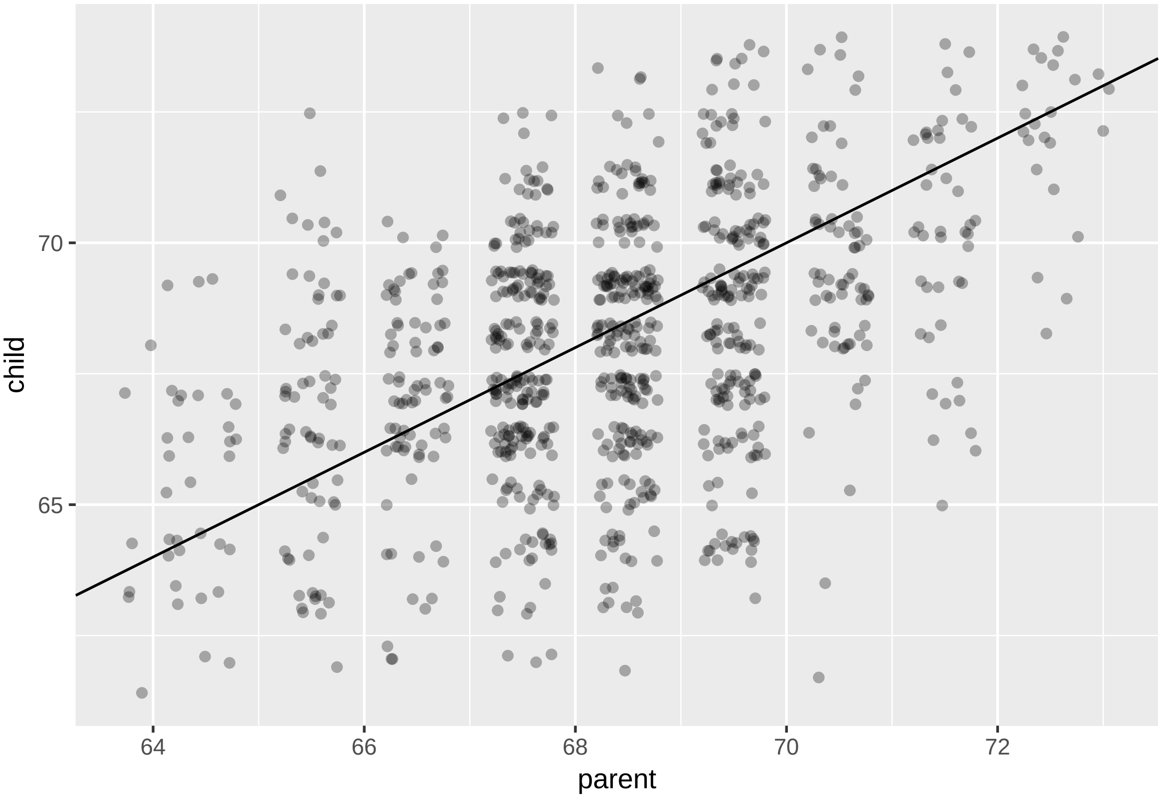
```
## [1] 0.9474
```

There is virtually no change.

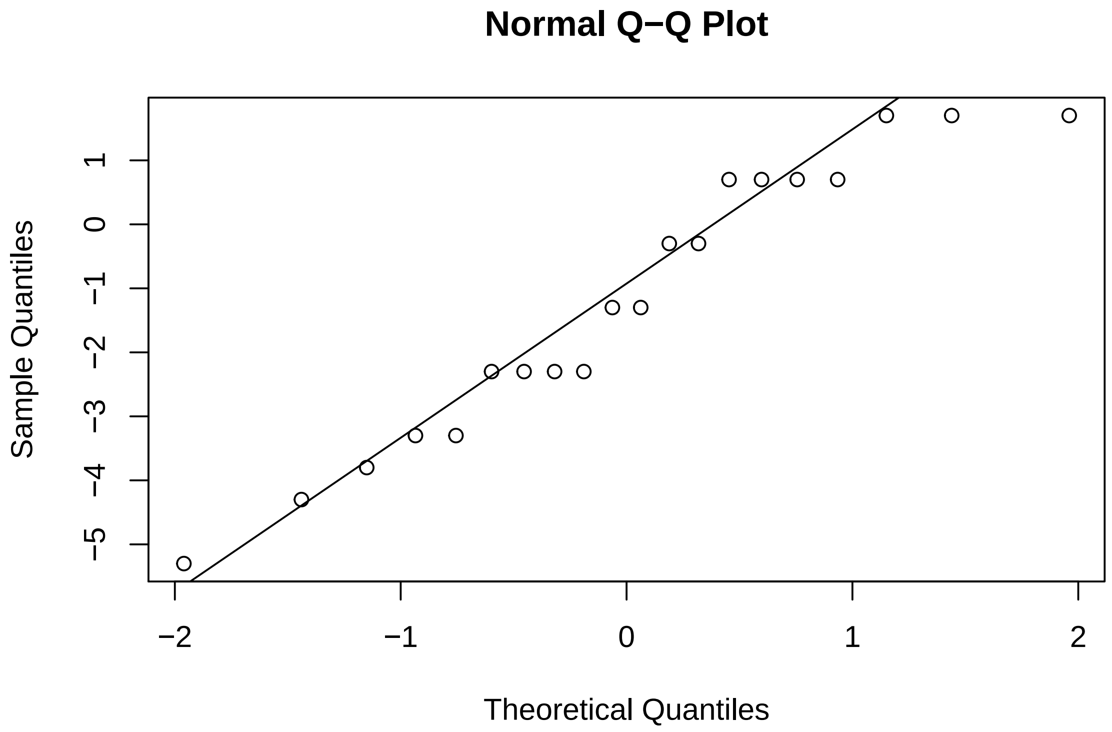
4. (Galton data) Francis Galton, a contemporary and a relative of Charles Darwin, made some groundbreaking analyses of characteristics of human populations. He also held some racist views and espoused eugenics, inventing the word. His biography is interesting reading, but not required for this problem. The "Galton" data set in the "HistData" package has his measurements. Details are available in the help for "Galton". (5 points each)

   a. 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.

```
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,]
diff<-dat$child-dat$parent
qqnorm(diff)
qqline(diff)
```



```r
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.4.4
```

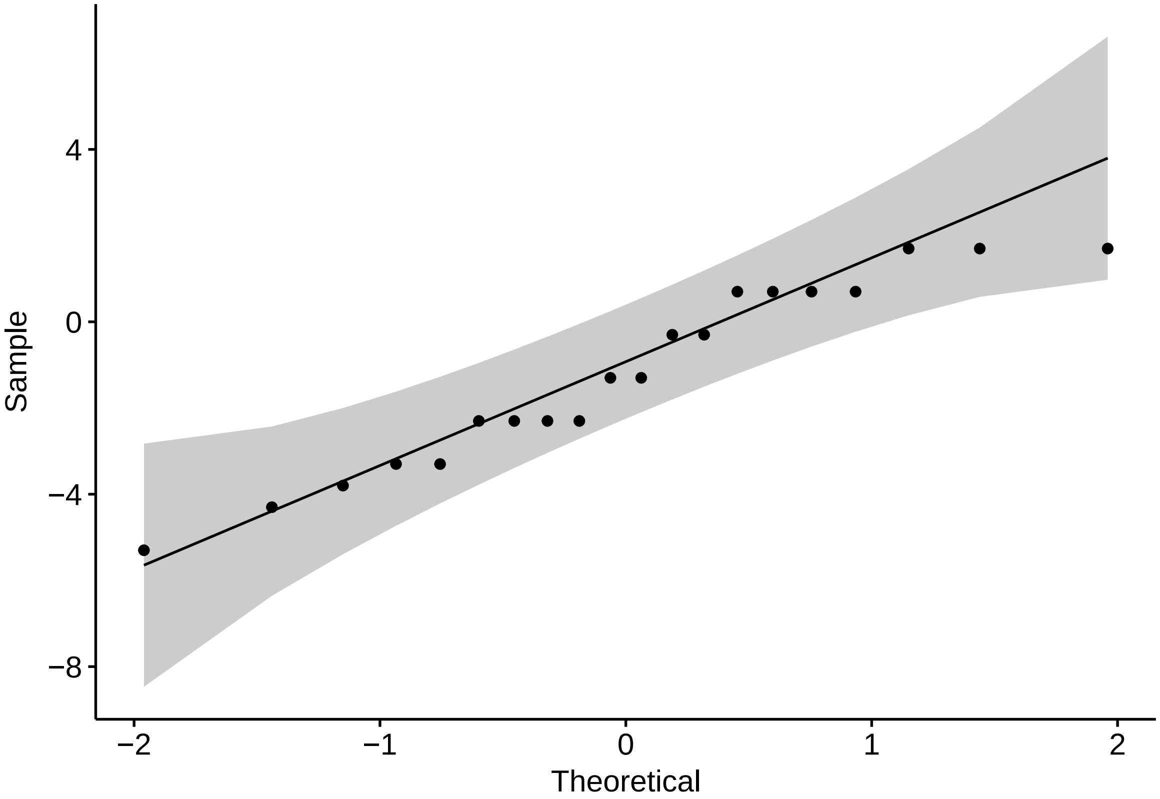
```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
ggqqplot(diff)
```



We can see from the plot that our data points lie more or less on the line and stays within the shaded grey area indicating that it is normally distributed.

b. Is Student's t-test suitable for these data and the null hypothesis that the mean of the difference between the child's height and the parents' combined height is 0? Please discuss.

Performaing a t-test requires normality of the data, which our qqplot showed to be the case. As for the number of samples used, t-test works best for sample sizes greater than 30 but it is not required.

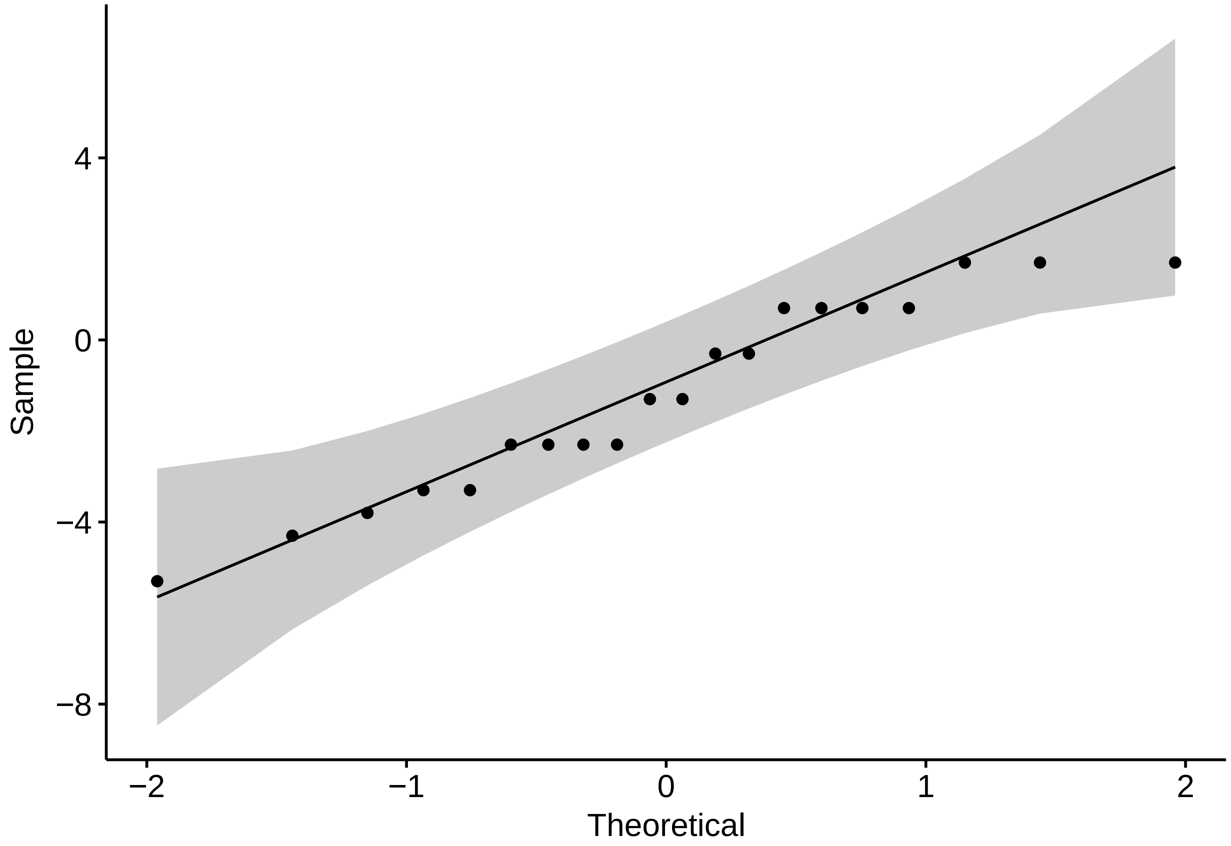
~~~
c. Test the hypothesis that this sample of child-parent is drawn from a Normal distribution with mean e

```r
t.test(diff)
```

```
##
##  One Sample t-test
##
## data:  diff
## t = -2.5436, df = 19, p-value = 0.01983
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.2330052 -0.2169948
## sample estimates:
## mean of x
##    -1.225
```
By simply observing the p-value, very low, we can reject the null hypothesis since this signifies that g
However, by analyzing the 95% confidence interval obtained and the fact that 0 was not in it, you could
~~~

```
set.seed(234567)
ind<-sample(1:nrow(Galton),20)
dat<-Galton[ind,]
df <- data.frame(child <- dat$child,parent <- dat$parent,child_minus_parent <- child - parent)

ggqqplot(df$child_minus_parent)
```


Sample
Theoretical