assignment2

Data description

The babyboom dataset contains the time of birth, sex, and birth weight for 44 babies born in one 24-hour period at a hospital in Brisbane, Australia.

Format

A data frame with 44 observations on the following 4 variables.

clock.time - Time on clock

gender - a factor with levels girl boy

wt - weight in grams of child

running.time - minutes after midnight of birth

```r
library(UsingR)
```

```
## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units

##
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':
##
##     cancer
```

```r
library("quantmod")
```

```
## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: TTR
## Version 0.4-0 included new data defaults. See ?getSymbols.
```

```
##
## Attaching package: 'quantmod'
```

```
## The following object is masked from 'package:Hmisc':
##
##     Lag
```

```
babyweight = na.omit(babyboom$wt)
babygender = na.omit(babyboom$gender)
summary(babyweight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1745    3142    3404    3276    3572    4162
```

No NA´s in the weight colume.

Q: Obtain a 97% conffidence interval for the population mean.

```
t.test(babyweight,conf.level=0.97)
```

```
##
##  One Sample t-test
##
## data:  babyweight
## t = 41.153, df = 43, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 97 percent confidence interval:
##  3097.284 3454.625
## sample estimates:
## mean of x
##  3275.955
```

Here we can see that with a probability of 97% that all babys born in the hospital in a time of 24h weight between 3097.284 - 3454.625 grams.

Q: Perform a t-test on whether the population mean is equal to the sample median. Clearly state the null and alternative hypotheses, provide the p-value.

This will be testet with an confidence level of 97% like for the first example.

H0: The true mean of the population is equal to the sample median "3404" off my dataset.

H1: The true mean of the population is higher or lower then the sample median of my dataset

```
t.test(babyweight, mu = 3404, conf.level=0.97)
```

```
##
##  One Sample t-test
##
## data:  babyweight
## t = -1.6085, df = 43, p-value = 0.115
## alternative hypothesis: true mean is not equal to 3404
## 97 percent confidence interval:
##  3097.284 3454.625
## sample estimates:
## mean of x
##  3275.955
```

Here we can see, that my p-value ist 0.115. This is higher then 0.03 which means that my null hypottheses can not be rejected. It must also be stated that the alternative hypotheses is not wrong.

Q: Obtain a 95% conffidence interval for the population standard deviation.

```
n <- sd(babyweight)
conf.level<- .95
z <- qt((1+conf.level)/2, df = n-1)
se<- sd(babyweight)/sqrt(n)
ci<-z*se
n <- sd(babyweight)
n-ci
```

```
## [1] 482.8909
```

```
n+ci
```

```
## [1] 573.1741
```

```
sd(babyweight)
```

sample

```
## [1] 528.0325
```

The 95% conffidence interval for the population standard deviation is 482.8909 - 573.1741.The sd() shows that the standard deviation is 528.0325.

Q: Find some dataset with a categorical variable. For that variable, compute the proportion of some level. Obtain a 99% conffidence interval for that proportion.

```
babygender=table(babygender)
sumAll = (babygender[names(babygender)=="boy"]+babygender[names(babygender)=="girl"])
sumBoy = babygender[names(babygender)=="boy"]
prop.test(x = sumBoy, n = sumAll, conf.level = 0.99, alt="two.sided")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sumBoy out of sumAll, null probability 0.5
## X-squared = 1.1136, df = 1, p-value = 0.2913
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.3901305 0.7665441
## sample estimates:
##         p
## 0.5909091
```

With a probability of 99% the conffidence intervall for this propotion is 0.3901305 - 0.7665441.

Q: Perform a hypothesis test on whether the population proportion is equal to 1/2. Clearly state the null and alternative hypotheses, provide the p-value.

H0: Half of the population is female while the other halfe is born male.

H1: More then half are born female or more then half are born male.

```r
prop.test(x = (sumAll/2), n = sumAll, conf.level = 0.99, alt="two.sided")
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  (sumAll/2) out of sumAll, null probability 0.5
## X-squared = 0, df = 1, p-value = 1
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.3190069 0.6809931
## sample estimates:
##   p
## 0.5
```

Here we can see that the p-value is 1 which means we cant reject H0.

Q: Generate the (imaginary) data for calculating the conffidence intervals between proportions of two populations (in fact, you need just four numbers). Describe your imaginary data. Obtain a 99% conffidence interval for the diffierence between proportions.

Q: Perform an appropriate hypothesis test for the difference between proportions. Draw a conclusion.

I have my Data from before and added my imaginary data which are that 10 out of 44. We can say that the first data set was from hospital 1 and the imaginary from hospital 2.

H0: The proportions in the two hospitals are the same

H1: The proportions in the two hospitals are different

```r
prop.test(x = c((sumAll/2),10), n = c(sumAll,44), conf.level = 0.99, alt="two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c((sumAll/2), 10) out of c(sumAll, 44)
## X-squared = 5.942, df = 1, p-value = 0.01478
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.003338778  0.548793323
## sample estimates:
##    prop 1    prop 2
## 0.5000000 0.2272727
```

My 99% coffindence intervall for those two dataset are -0.003338778 - 0.548793323. This means that we are not sure if the proportions are qual because the 0 is also in the intervall. The p-value is 0.01478 so we cant reject the Null hypothesis but there is still the zero included in the intervall

Both hypo testing and conf interval reveal
the same: two proportions may be equal.
Your answer is correct except for "still"

Q: Do the F test for two population variances. State the null and alternative hypothesis.

H0: The variances of "babyweight" and "babyboom$clock.time" are equal

H1: The variances of "babyweight" and "babyboom$clock.time" differ

```
var.test(babyweight, babyboom$clock.time)
```

```
##
##  F test to compare two variances
##
## data:  babyweight and babyboom$clock.time
## F = 0.58444, num df = 43, denom df = 43, p-value = 0.08177
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3188964 1.0710841
## sample estimates:
## ratio of variances
##          0.5844355
```

The p-value for this test is 0.08177 this means that it is more then 0.05 which says that we do not reject out H0.

Q: Perform the Jarque-Bera for normality. State clearly the null and alternative hypothesis.

U.S. / Euro Foreign Exchange Rate. Data goes from 1999-01-01 to 2017-09-01 in monthly steps.

H0: The distribution of the exchange rate is uniform over month of years

H1: The distribution of calls is not uniform

```
EXUSEU=getSymbols('EXUSEU', src='FRED', auto.assign=FALSE)
```

```
## 'getSymbols' currently uses auto.assign=TRUE by default, but will
## use auto.assign=FALSE in 0.5-0. You will still be able to use
## 'loadSymbols' to automatically load data. getOption("getSymbols.env")
## and getOption("getSymbols.auto.assign") will still be checked for
## alternate defaults.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.warning4.0"=FALSE). See ?getSymbols for details.
```

```
EXUSEU=na.omit(EXUSEU)
chisq.test(EXUSEU, p=rep(1/length(EXUSEU),length(EXUSEU)))
```

```
## Warning in chisq.test(EXUSEU, p = rep(1/length(EXUSEU), length(EXUSEU))):
## Chi-squared approximation may be incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  EXUSEU
## X-squared = 5.5865, df = 224, p-value = 1
```

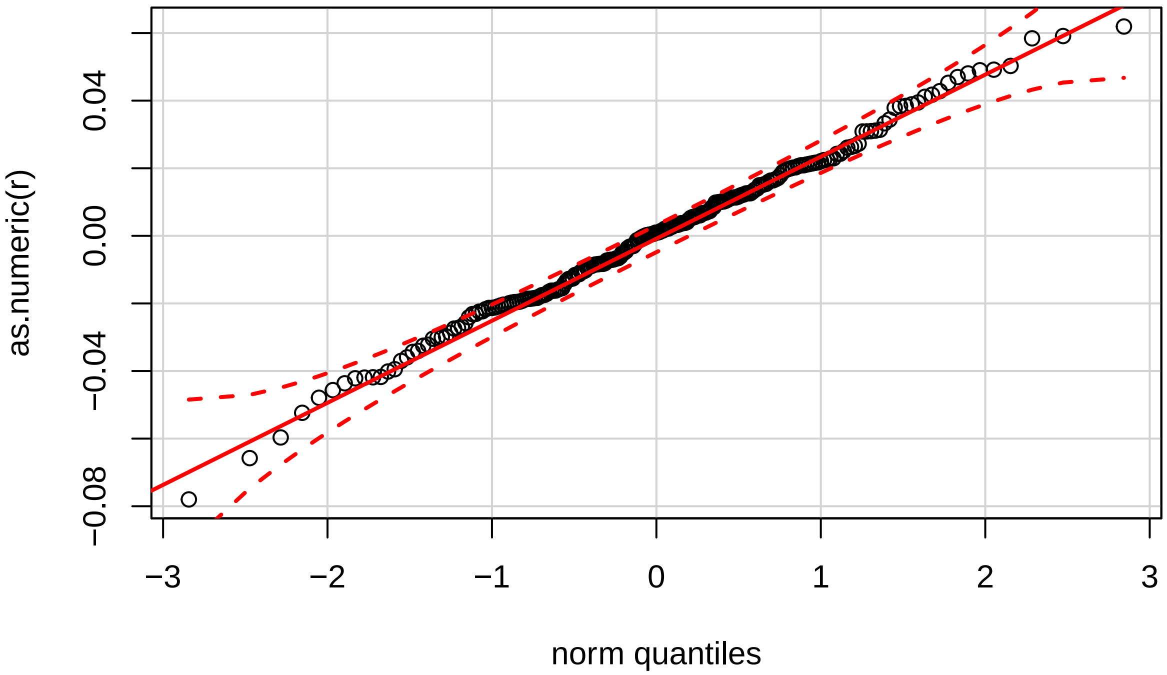
This is not Jarque-Bera test.... Use command normalTest(data,method="jb")

The p-value for this test is 1 so we cant reject H0.

Q: Check whether the (univariate) empirical distribution of log returns for each stock is normal by examining the QQ-plot. Use the command qq.plot() from car package instead of the built-in function. Discuss whether the observations are within the conffidence interval.

```
library("car")
r=diff(log(EXUSEU))
qq.plot(as.numeric(r))
```

```
## Warning: 'qq.plot' is deprecated.
## Use 'qqPlot' instead.
## See help("Deprecated") and help("car-deprecated").
```



In the qq-plot we cann see that the observations are really in the conffidence intervals.

Q: Use a built-in set from 2 to perform the x^2-test for homogeneity. Describe the data and discuss the result. See lecture slides and Section 9.1.2 of [1].

For this we are taking the smokers and their mainly used hand to test for homogeneity.

```
library(MASS)
tbl = table(survey$Smoke, survey$W.Hnd)
tbl
```

```
##
##           Left Right
##   Heavy    1    10
##   Never   13   175
##   Occas    3    16
##   Regul    1    16
```

```
chisq.test(tbl)
```

```
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  tbl
## X-squared = 2.0307, df = 3, p-value = 0.5661
```

The p-value is 0.5661 ,hence it is not possible to verify different hand preferences between the regularity of smokers.

Q: Get a two-way contingency table from sources 3. Conduct a x^2-test for association (independence) between the variables. See lecture slides and Section 9.2 of [1]

```
tbl2 = table(survey$Smoke, survey$Sex)
tbl2
```

```
## 
##         Female Male
##   Heavy      5    6
##   Never     99   89
##   Occas      9   10
##   Regul      5   12
```

```
chisq.test(tbl2)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  tbl2
## X-squared = 3.5536, df = 3, p-value = 0.3139
```

The p-value for this 0.3139. This means that the smoking habit is independet to the sex.