Chuvilina Anna, group MBD171

Assignment on Confidence Intervals and Hypothesis Testing R

1. Get a univariate dataset from sources 1 or 2 and briefly describe it. Using these data,

I have chosen data frame “alaska.pipeline” from UsingR. This data frame contains the following columns: field.defect (depth of defect as measured in field), lab.defect (depth of defect as measured in lab), batch (one of 6 batches). Length of data frame is 107.

```
 al1 = alaska.pipeline$field.defect
 al2 = alaska.pipeline$lab.defectq
 al3 = alaska.pipeline$batch
 length(al1)
[1] 107
 length(al2)
[1] 107
 length(al3)
[1] 107
```

(a) Obtain a 97% confidence interval for the population mean.

```
al2 = alaska.pipeline$lab.defect

t.test(al2, conf.level = 0.97 )

        One Sample t-test

data:  al2
t = 16.812, df = 106, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
97 percent confidence interval:
 33.98341 44.21472
sample estimates:
mean of x
 39.09907
```

We can see that with probability 97% depth of defect as measured in lab lying in interval [33.98341, 44.21472].

(b) 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.

We have two hypotheses:
H0: The true mean of the population is equal to the sample median (38) of data set al2
H1: The true mean of the population isn't equal (higher or lower) to the sample median (38) of data set al2

```
> t.test(al2, mu = median(al2), conf.level = 0.95)

        One Sample t-test

data:  al2
t = 0.47259, df = 106, p-value = 0.6375
alternative hypothesis: true mean is not equal to 38
95 percent confidence interval:
 34.48830 43.70983
sample estimates:
mean of x
 39.09907
```

We can see that p-value is 0.6375. P-value is higher than 0.03, which means that null hypotheses can't be rejected. We can't say exactly that the observed results are the result of pure chance or manipulation of variables.

(c) Obtain a 95% confidence interval for the population standard deviation.

```
len = length(al2)
sdAl2 = sd(al2)
erAl2 = qt((1 + 0.95)/2, df = len -1)*sdAl2/sqrt(len)
erAl2
[1] 4.610767
sdAl2 - erAl2
[1] 19.44563
sdAl2 + erAl2
[1] 28.66717
 sd(al2)
[1] 24.0564
```

We can see that interval of 95% confidencel is [19.44563, 28.66717]. Standard deviation is 24.0564.

(d) Find some dataset with a categorical variable. For that variable, compute the proportion of some level. Obtain a 99% confidence interval for that proportion.

I have chosen data frame “cats” from MASS. This data frame contains the following columns: Sex (M or F), Bwt (body weight in kg) and Hwt (heart weight in g). I bring information about heart weight for univariate data. Length of data frame is 144.

```
a = table(cats$Sex)
all = (a[names(a) == "F"] + a[names(a) == "M"])
allF = a[names(a) == "F"]
prop.test(x = allF, n = all, conf.level = 0.99, alt = "two.sided")

        1-sample proportions test with continuity correction

data:  allF out of all, null probability 0.5
X-squared = 16.674, df = 1, p-value = 4.439e-05
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.2322728 0.4363131
sample estimates:
       p
0.3263889
```

The 99% confidence interval for that proportion is [0.2322728, 0.4363131].

(e) 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.

We have two hypotheses:
H0: Half of the cats is female or half of the cats is male.
H1: More than half of the cats is female or more than half of the cats is male.

```
> prop.test(x = allF, n = all, conf.level = 0.99)

        1-sample proportions test with continuity correction

data:  allF out of all, null probability 0.5
X-squared = 16.674, df = 1, p-value = 4.439e-05
alternative hypothesis: true p is not equal to 0.5
```

99 percent confidence interval:
 0.2322728 0.4363131
sample estimates:
        p
0.3263889

The p-value is 4.439e-05 (p-value < 0.01). That means I can reject null hypotheses and there are differences between the population.

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

I have chosen data about who often memorize information from advertising: men or women. 18 men from 35 remembered information and 36 from 45 women remembered information too.

```
prop.test(x = c(18, 35), n = c(36, 45), conf.level = 0.99)

        2-sample test for equality of proportions with continuity correction

data:  c(18, 35) out of c(36, 45)
X-squared = 5.6499, df = 1, p-value = 0.01746
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.57028385  0.01472829
sample estimates:
   prop 1    prop 2
0.5000000 0.7777778
```

The 99% confidence interval for the difference between proportions is [-0.57028385, 0.01472829]. This interval contains zero therefore we can't say about equality of proportion.

(g) Perform an appropriate hypothesis test for the difference between proportions. Draw a conclusion.

We have two hypotheses:
H0: Proportion of men which remember information from advertising is the same proportion of women.
H1: Proportion of men which remember information from advertising isn't the same (higher or lower) proportion of women.

```
prop.test(x = c(18, 35), n = c(36, 45), conf.level = 0.99)

        2-sample test for equality of proportions with continuity correction

data:  c(18, 35) out of c(36, 45)
X-squared = 5.6499, df = 1, p-value = 0.01746
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.57028385  0.01472829
sample estimates:
   prop 1    prop 2
0.5000000 0.7777778
```

The p-value is 0.01746. That means we can't reject null hypotheses. We can say about equality proportion of women and men with 99% probability, but zero in confidence interval can say about difference between proportions.

(h) Do the F test for two population variances. State the null and alternative hypothesis.

We have two hypotheses about compare variance of cats' body weight and variance of cats' heart weight:
H0: The variance of the body weight is equal the variance of the heart weight.
H1: The variance of the body weight isn't equal (higher or lower) the variance of the heart weight.

```
body = cats$Bwt
heart = cats$Hwt
var.test(body, heart)

        F test to compare two variances

data:  body and heart
F = 0.039734, num df = 143, denom df = 143, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.02859126 0.05521991
sample estimates:
ratio of variances
         0.0397342
```

The p-value is 2.2e-16 (p-value < 0.05). That means we can reject null hypotheses and there are differences between the population.

2. Pick up time series of monthly (yearly, daily,..., up to you) log returns on three securities or commodities from data sources 1. Use getSymbols command to download the data and:

I have chosen data frame “MSFT” (Microsoft) from Yahoo Finance. This data frame has data from 1. 03. 1986 to nowadays. I use monthly data. Length of data frame is 516.

```
msft = getSymbols('MSFT', src ='yahoo', auto.assign = FALSE)
msft1 = na.omit(msft1)
msftmonth = to.monthly(msft1)
length(msftmonth)
[1] 516
tail(msftmonth)
         msft1.Open msft1.High msft1.Low msft1.Close
апр 2017      65.55      68.46     64.95       68.46
май 2017      69.41      70.41     67.48       69.84
июн 2017      70.10      72.52     68.49       68.93
июл 2017      68.17      74.22     68.17       72.70
авг 2017      72.58      74.77     71.41       74.77
сен 2017      73.94      75.44     73.40       74.41
```

(a) Perform the Jarque-Bera for normality. State clearly the null and alternative hypothesis.

We have two hypotheses:
H0: The distribution of business software is normal of all time period in months.
H1: The distribution of business software isn’t normal.

```
msftmonth = as.numeric(msftmonth)
jarque.bera.test(msftmonth)

        Jarque Bera Test

data:  msftmonth
X-squared = 99.565, df = 2, p-value < 2.2e-16

ajb.norm.test(msftmonth)
```

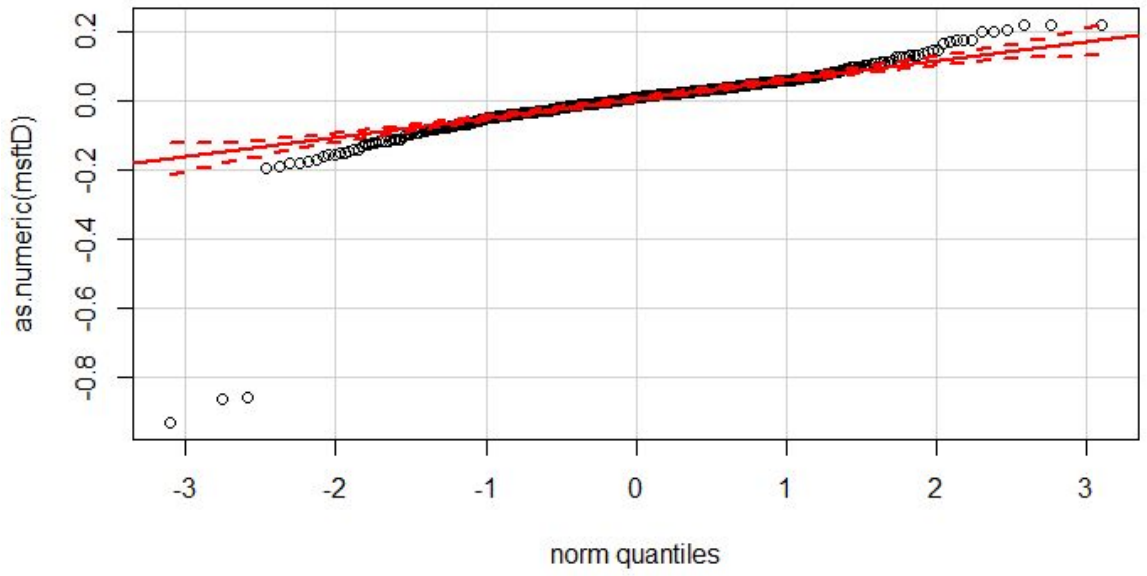
Adjusted Jarque-Bera test for normality

data: msftmonth
AJB = 100.9, p-value < 2.2e-16

The p-value is 2.2e-16 (p-value < 0.05). We can reject null hypotheses, the distribution isn't normal.

(b) 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 confidence interval. See examples in [2], Section 1.6.

```
msftD = diff(log(msftmonth))
qq.plot(msftD)
```



We can observe heavy-tailed distribution on qqplot.

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

```
tabCat = table(cats$Bwt, cats$Sex)
tabCat
```

```
     F  M
  2    3  2
  2.1  9  1
  2.2  6  8
  2.3 12  1
  2.4  4  5
  2.5  2  8
  2.6  3  6
  2.7  3  9
  2.8  0  7
  2.9  3  5
  3    2  9
  3.1  0  6
  3.2  0  6
  3.3  0  5
  3.4  0  5
  3.5  0  5
  3.6  0  4
  3.7  0  1
  3.8  0  2
  3.9  0  2
chisq.test(tabCat)

        Pearson's Chi-squared test

data:  tabCat
X-squared = 61.969, df = 19, p-value = 1.881e-06
```

The p-value is 1.881e-06, therefore body weight of cats and their gender are dependent.

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

I have chosen data frame “homeprice” from UsingR. This data frame contains the following columns: list (list price of home (in thousands)), sale (actual sale price), full (Number of full bathrooms), half (number of half bathrooms), bedrooms (number of bedrooms), rooms (total number of rooms), neighborhood (Subjective assessment of neighborhood on scale of 1-5). Length of data frame is 29.

H0: There is dependence between list price of home and actual sale price.
H1: There is no dependence between list price of home and actual sale price.

```
s = rbind(homeprice$list, homeprice$sale)
 s
      [,1]  [,2] [,3] [,4] [,5]  [,6]  [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16]
[1,]  80.0 151.4  310  295  339 337.5 228.7  245  339    43 279.0   599   119   289
249   178
[2,] 117.7 151.0  300  275  340 337.5 215.0  239  345    48 262.5   613   119   305
249   170
     [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26] [,27] [,28] [,29]
[1,]   159   289   488   376   249   275 275.0   459   219   359   379   189   173
[2,]   153   291   450   370   245   275 272.5   459   230   360   370   185   185
 chisq.test(s)

        Pearson's Chi-squared test

data:  s
X-squared = 12.646, df = 28, p-value = 0.9943
```

The p-value is 0.9943, therefore we can observe dependence between list price of home and actual sale price.