MIE-SPI 2014 – Homework No. 2

Team leader: Timur Tatarshaov (tatartim)
Date: 04.05.2014

1. One-sample t-test for the mean:

1.I. Two-sided t-test

R commands according to the instructions:

```
n = 20;alpha = 0.01;
x = rnorm(n, mean=10.5, sd=1.3);

hypothesisTest = t.test(x, mu=10, conf.level = 1-alpha);
print(hypothesisTest);
```

- Output:

```
    One Sample t-test

data:  x
t = 2.1675, df = 19, p-value = 0.0431
alternative hypothesis: true mean is not equal to 10
99 percent confidence interval:
  9.76995 11.66833
sample estimates:
mean of x
 10.71914
```

1.I.a. Confidence interval

Two-sided 99% confidence interval for the expectation μ can be computed using

```
n <- length(x);
criticalValue <- qt(alpha/2,19, lower.tail=FALSE);
avgX <- mean(x);
stdDev <- sd(x);
sqrtn <- sqrt(n);
intv <- criticalValue*(stdDev/sqrtn);

confidence <- c(avgX - intv, avgX + intv);
```

- Output

```
[1] 10.71914
[1]  9.76995 11.66833
```

- We can see that the 99% confidence interval (9.76995 11.66833) **agrees with the output** of the previous t.test command.

1.I.b. Hypothesis test using a confidence interval

Based on the 99% confidence interval (9.76995 11.66833) for the mean μ we can test H_0: $\mu = 10$ against the two-sided alernative H_A: $\mu \neq 10$ by checking if $\mu = 10$ is in the confidence interval

- In our case $\mu = 10$ is in the confidence interval therefore we **accept the hypothesis H_0**
- The probability that this decision is incorrect is **alpha.**

1.I.c. Hypothesis test using the test statistic *T*

We will compute the value of the test statistic T and the corresponding critical value using

```
tStat <- abs(mean(x) - 10)/(sqrt(var(x))/sqrt(20));
criticalValue <- qt(alpha/2,19, lower.tail=FALSE);
tStat
criticalValue
```

- Output

```
[1] 2.167545
[1] 2.860935
```

- The test statistic *t* = 2.167545 **agrees with the output** of the previous t.test command.
- The absolute value of the t statistic |*t*| = 2.167545 is inside interval of (-2.860935, +2.860935), therefore we cam **accept the hypothesis H_0.**
- This decision **confirms** the above test result using the confidence interval.

1.II. One-sided t-test

R commands according to the instructions:

```
n = 20;
alpha = 0.01;
x = rnorm(n, mean=10.5, sd=1.3);
hypothesisTest = t.test(x, mu=10, alternative = "greater", conf.level = 1-alpha);
print(hypothesisTest);
n <- length(x);
criticalValue <- qt(alpha, 19, lower.tail = FALSE);
avgX <- mean(x);
stdDev <- sd(x);
sqrtn <- sqrt(n);
intv <- criticalValue*(stdDev/sqrtn);
confidence <- c(avgX - intv, +Inf);
testT <- abs(mean(x) - 10)/(sqrt(var(x))/sqrt(20));
```

- Output

```
One Sample t-test

data:  x
t = 1.7898, df = 19, p-value = 0.04472
alternative hypothesis: true mean is greater than 10
99 percent confidence interval:
 9.77381      Inf
sample estimates:
mean of x
 10.53997
> testT
[1] 1.789765
> confidence
[1] 9.77381      Inf
```

Greater alternative is chosen because **original mean** of normal distribution is **greater** then value in test. T statistic values and confidence interval **fit**.

2. Paired and two-sample t-tests for comparison of means

2.I. Paired t-test

2.I.a. Analyze the output and test H0

R commands according to the instructions:

```
n = 20;
alpha = 0.01
x = rnorm(n, mean=10, sd=1)
error = rnorm(n, mean=0.5, sd=0.8306624)
y = x + error

t.test(x, y=y, paired = TRUE, alternative = "less", conf.level = 1-alpha);
```

- Output

```
Paired t-test

data:  x and y
t = -2.9009, df = 19, p-value = 0.004581
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
       -Inf -0.0778925
sample estimates:
```

```
mean of the differences
             -0.6252375
```

Mean in differences is inside confidence interval. Probability that decision is incorrect is alpha(0.01).

2.I.b. Compute the differences diff = x - y and test the null hypothesis H0

R commands according to the instructions:

```
diff <- x - y;
t.test(diff, mu=0, alternative="less", conf.level = 1-alpha);
```

- Output

```
One Sample t-test

data:  diff
t = -2.9009, df = 19, p-value = 0.004581
alternative hypothesis: true mean is less than 0
99 percent confidence interval:
       -Inf -0.0778925
sample estimates:
 mean of x
-0.6252375
```

Less alternative has been chosen, because **error** is mostly **positive** and **diff** is mostly **negative** and we are suppose to show that it less than null hypothesis which is 0.

Interval and mean fit, that means that our choose of alternative was correct. And mean **is in the confidence interval**.

2.II. Two-sample t-test

R commands according to the instructions:

```
n1 = 20;
n2 = 25;
alpha = 0.01
x=rnorm(n1, mean=10, sd=1.3)
y=rnorm(n2, mean=11.25, sd=1.3)

t.test(x, y=y, paired = FALSE, var.equal = TRUE, conf.level = 1-alpha)
```

- Output

```
Two Sample t-test

data:  x and y
t = -1.4014, df = 43, p-value = 0.1683
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -1.5204419  0.4801825
sample estimates:
mean of x mean of y
 10.60806  11.12819
```

Difference in means is **inside** confidence interval, we can accept null hypothesis.

2.II.a. Modify the command for testing of the null hypothesis H0: muX = muY against one-sided alternative HA: muX < muY

R commands according to the instructions:

```
n1 = 20;
n2 = 25;
alpha = 0.01
x=rnorm(n1, mean=10, sd=1.3)
y=rnorm(n2, mean=11.25, sd=1.3)

t.test(x, y=y, paired = FALSE, alternative = "less", var.equal = TRUE, conf.level = 1-
alpha);
```

- Output

```
Two Sample t-test

data:  x and y
t = -3.2143, df = 43, p-value = 0.001241
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
     -Inf -0.3604372
sample estimates:
mean of x mean of y
 9.949165 11.400945
```

Test command changed according to requirements.

2.II.b. Using formulas presented in the lectures, compute the test statistic 't' and the degrees of freedom 'df'

R commands according to the instructions:

```
sx2 <- var(x);
sy2 <- var(y);
meanX <- mean(x);
meanY <- mean(y);
degOfFreedom <- n1 + n2 - 2;

sxy <- sqrt ( ((n1-1)*sx2 + (n2-1)*sy2)/degOfFreedom )
t <- (meanX - meanY)/(sxy*sqrt(1/n1+1/n2));

pValue <- pt(t, degOfFreedom);

degOfFreedom;
pValue;
t;
```

- Output

```
> degOfFreedom;
[1] 43

> pValue;
[1] 0.001240904

> t;
[1] -3.214264
```

Values calculated manually according to formulas **match** values got by test integrated in R.

2.III. Obtain a two-sample t-test

R commands according to the instructions:

```
n1 = 20;
n2 = 25;
alpha = 0.01
x=rnorm(n1, mean=10, sd=1.3)
y=rnorm(n2, mean=11.28, sd=1.2)

t.test(x, y=y, paired = FALSE, var.equal = FALSE, conf.level = 1-alpha)
t.test(x, y=y, paired = FALSE, alternative="less", var.equal = FALSE, conf.level = 1-
alpha)

sx2 <- var(x);
sy2 <- var(y);
meanX <- mean(x);
meanY <- mean(y);
degOfFreedom <- (sx2/n1 + sy2/n2)^2/( (sx2/n1)^2/(n1-1) + (sy2/n2)^2/(n2-1) );


sxy <- sqrt (sx2/n1 + sy2/n2)
t <- (meanX - meanY)/sxy;
pValue <- pt(t, degOfFreedom);

degOfFreedom;
pValue;
t;
```

- Output

```
Welch Two Sample t-test

data:  x and y
t = -3.2139, df = 41.542, p-value = 0.001267
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
       -Inf -0.2996283
sample estimates:
mean of x mean of y
 10.15098  11.36326

> degOfFreedom;
[1] 41.54238

> pValue;
[1] 0.001266738

> t;
[1] -3.213865
```

Values calculated manually according to formulas **match** values got by test integrated in R.

3. Practical applications of t-tests

3.II. Illustrate the practical use of t-tests

3.II.a. No task here

3.II.b. No task here

3.II.c. Examine the speed of your computer

R commands according to the instructions:

```
sequenceLength = 2500000;
x = runif(sequenceLength, 0, 100);
print(system.time(sort(x))[1]);
```

- Output

```
user.self
       0.303
```

3.III. Generate L*40 random numerical sequences of equal length and measure the durations of their sorting

R commands according to the instructions:

```
sampleSize = L*40;
time1 = time2 = numeric(sampleSize); # Declare an array
for(i in 1:sampleSize){
  x = runif(sequenceLength, 0, 100);
  time1[i] = system.time(x1 <- sort(x, method = "quick"),  gcFirst = TRUE)[1];
  time2[i] = system.time(x2 <- sort(x, method = "shell"), gcFirst = TRUE)[1];
}
time1;
time2;
```

- Output

```
Output is too long. Quick in average is 0.212, shell is 0.312
```

3.III.b. At the level alpha = K/100 test if the measured data provide statistical evidence for our working hypothesis from the previous step

R commands according to the instructions:

```
alpha <- K/100;
t.test(time1, y=time2, paired = TRUE, alternative = "less", conf.level = 1-alpha)
```

- Output

```
Paired t-test

data:  time1 and time2
t = -146.8284, df = 399, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1001007
sample estimates:
mean of the differences
             -0.1012375
```

3.III.c. Describe in detail how and why you selected the null hypothesis H0 and the alternative HA

Null hypothesis is equality in means and alternative is less for the reasons described below.

3.III.d. Justify in detail which t-test you have used and why

Paired t-test with less alternative has been used because distributions have **same size** and values in first set is **smaller** than in second.

3.IV. Repeat the previous task for separate measurements

R commands according to the instructions:

```
sequenceLength = 2500000;


sampleSize = L*40;
time1 = time2 = numeric(sampleSize); # Declare an array
for(i in 1:sampleSize){
  x = runif(sequenceLength, 0, 100); # Generate the sequence to be sorted
  time1[i] = system.time(x1 <- sort(x, method = "quick"),  gcFirst = TRUE)[1];
}
sampleSize2 = L*35;
for(i in 1:sampleSize2){
  x = runif(sequenceLength, 0, 100); # Generate the sequence to be sorted
  time2[i] = system.time(x2 <- sort(x, method = "shell"), gcFirst = TRUE)[1];
}

alpha <- K/100;
t.test(time1, y=time2, paired = FALSE, alternative="less", var.equal = FALSE, conf.level
= 1-alpha)
```

- Output

```
Welch Two Sample t-test

data:  time1 and time2
t = -11.263, df = 404.422, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.04988581
sample estimates:
mean of x mean of y
0.2133325 0.2717725
```

3.V. Compare the results of both experiments

In both cases **null hypothesis accepted** as far as mean (in first case) and difference in means (in second case) **inside confidence interval**. So we couldn't show that "quick" sort is faster than "shell" sort. But in case of paired test it was really close to the border, **not directly** it shows that first sort is faster though.