Домашна работа №3

на
Денис Тихогло
Софтуерно инженерство, 3 курс, 3 група, ф.н. 855256

- **Задача 1. .**

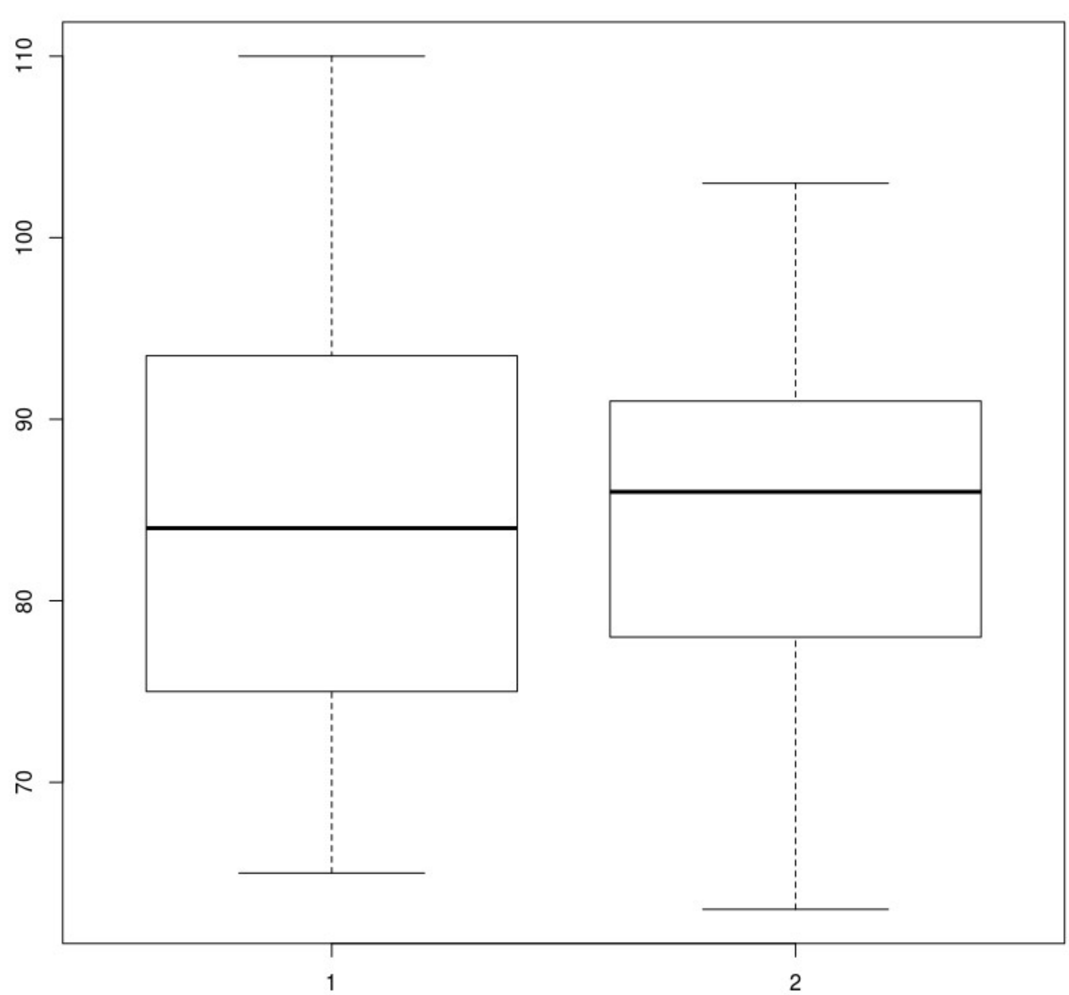
  а) Формулирайте хипотеза за проверка на твърдението, че резултатите от измерванията на експерта се различават от резултатите от измерването на машината.

  б) Каква е критичната област при ниво на съгласие $\alpha = 0.1$?

- **в) Използвайте подходяща функция в R, за да тествате хипотезата. Какъв е резултатът? Интерпретирайте p–стойността, стойността на тестовата статистика и доверителния интервал.**

Първо, което трябва да направим, е да се убедим, че даннтие в колоните `Machine` и `Expert` са приблизително нормално разпределени

```
> boxplot(Machine, Expert)
```



Данните са дискретни, но общата форма на boxplot-ове личи да е приблизително нормална.
Ще се възползваме от сдвоен t-test, тъй като сравняваме два метода на измерение (с помощ на eксперта и с помощ на машината.
Нека $X_{diff,i} = X_{machine,i} - X_{expert,i}$ са разлики между съответните стойности на всяка двойка измервания, а μ_{diff} е очакваното средно на тези разлики.
Ще формираме хипотезите:

$H_0 : \mu_{diff} = 0$

$H_A : \mu_{diff} \neq 0$

Ще приемем твърдеието, че резултатите от измерванията на експерта се различават от резултатите от измерването на машината, ако ще отхвърлим нулевата хипотеза.

Използваме t-разпределението, защото броят наблюдения е по-малък от 30. То ще е със 14 степени на свобода, защото наблюдаваме общо 15 двойки наблюдения.

Ще намерим критичната област на тестовата статистика при ниво на съгласие $\alpha = 0.1$:

```
> qt(0.05, df=14)
> qt(0.95, df=14)
```

Получаваме стойности -1.76131 и 1.76131. Следователно, търсената критична област е $(-\infty, -1.76131) \cup (1.76131, +\infty)$

Тестовата статистика ще има вида $\frac{\overline{X_{diff}} - 0}{\frac{s_{diff}}{\sqrt{n}}} \sim t_{14}$, където $\overline{X_{diff}} = \frac{\sum_{i=1}^{15} X_{diff,i}}{n}$ е

извадковото средно на съответните разлики, а $s_{diff} = \sqrt{\left(\frac{\sum_{i=1}^{15} (X_{diff,i} - \overline{X_{diff}})^2}{n-1}\right)}$ е стандартното отклонение на разликите.

За намирането на стойността на тестовата статистика и p-value за дадената извадка, а също и за доверитлния интарвал за $\overline{X_{diff}}$, ще използвам функция t.test с параметрите paired=TRUE и conf.level=0.9:

```
> t.test(Machine, Expert, paired=TRUE, conf.level=0.9)
```

Отговорът, който получаваме, е

```
	Paired t-test

data:  Machine and Expert
```

```
t = 0.68162, df = 14, p-value = 0.5066
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -1.584017  3.584017
sample estimates:
mean of the differences
                      1
```

Даденият резултат означава, че нулевата хипотеза не трябва да бъде отхвърлена, по няколко причини.

Една от причините, която ни води да направим такова заключение, е голямо значение на p-value. P-value е вероятността да наблюдаваме значения на тестовата статикика, които по същия или по-силен начин подкрепят алтернативната хипотеза, при допускане, че приемаме нулевата хипотеза. Щом p-value = 0.5066 > α = 0.1, не отхвърляме нулевата хипотеза.

Доверителният интервал (90 percent confidence interval: -1.584017 3.584017) показва, за какви стойности на очакваното средно на разлики (μ_{diff}) нулевата хипотеза няма да бъде отхвърлена. Значенията на границите отговарят на стойностите ($\overline{X_{diff}}+\frac{s_{diff}}{\sqrt{n}}t_{0.05,df=14}$, $\overline{X_{diff}}+\frac{s_{diff}}{\sqrt{n}}t_{0.95,df=14}$). Нулевата хипотеза предполага, че $\mu_{diff} = 0$. Нула попада в даденият доверителен интервал, следователно не отхвърляме нулевата хипотеза.

t = 0.68162 е изчислената стойност на тестовата статистика. Значението влиза в интервала (-1.76131; 1.76131), следователно не отхвърлям нулевата хипотеза.

- **Данните brightness от пакета UsingR съдържат стойностите на яркост на 966 звезди.**

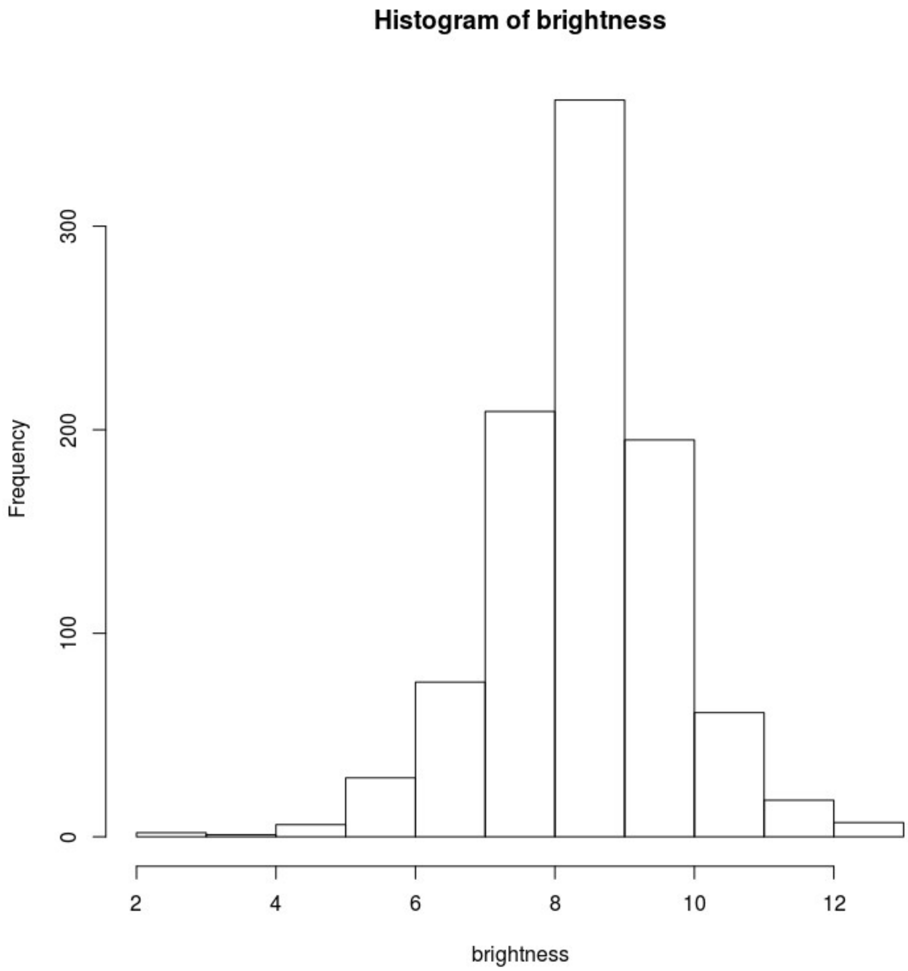
  а) Удачно ли е да се използва тест за средната стойност на яркостта и какво разпределение бихте използвали? Обосновете отговора си;

  б) Постройте 93% доверителен интервал за средната яркост.

Ще проверим, дали е удачно използването на теста за средната стойност.

От централната гранична теорема следва, че разпределението трябва да прилича на нормалното (извадката е голяма). За проверка ще построим хистограмата:

```
> hist(brightness)
```



Разпределението на данните прилича на нормално и не е прекалено изкривено, затова можем да използваме тестът за средната стойност.

Трябва да изберем между `z-test` (нормално разпределение) и `t-test` (t-разпределение). `Z-test` е по-удачен за големите извадки, като дадената. Но точното значение на стандартното отклонение е неизвстно. Затова ще използвам `t-test` със 965 степени на свобода.

Тестовата статистика е $\dfrac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{965}$

Двустранен 93% доверителен интервал за средната яркост е

$\left(\overline{X}+\frac{s}{\sqrt{n}}t_{0.035, df=965} \quad , \quad \overline{X}+\frac{s}{\sqrt{n}}t_{0.965, df=965} \right)$

Ще го изчислим, използвайки стандартните функции на R:

```
> n = 966
> x_bar = mean(brightness)
> s = sd(brightness)
> t_0.035 = qt(0.035, df = n-1)
> t_0.965 = qt(1-0.035, df = n-1)
> L1 = x_bar + s*t_0.035/sqrt(n)
> L2 = x_bar + s*t_0.965/sqrt(n)
```

Получаваме отговорът: L1 = 8.342209 и L2 = 8.493277

Можем да проверим правилността на изчисленията с помощта на функция t.test:

```
> t.test(brightness, conf.level = 0.93)
```

Получаваме същите граници за интервала:

```
...
93 percent confidence interval:
 8.342209 8.493277
...
```

- **Задача 3. От запитани 1337 души, 1000 отговарят, че използват интернет всеки ден. Използвайте това, за да тествате твърдението, че повече от 70% от хората използват интернет всеки ден, при ниво на съгласие α = 0.05**

Нека X е случайнна велиина, равна на брой хора, използващи интернет всеки ден.

Хипотезите ще имат вид:

H_0 : p ≥ 70 (повече от 70% от хората използват интернет всеки ден)

H_A : p < 70 (по-малко от 70% от хората използват интернет всеки ден)

Ще използвам функция `prop.test` за тестване на нулевата хипотеза:

```
> n = 1337
```
– общ брой наблюдения

```
> x = 1000
```
– брой успешни наблюдения

```
> p = 0.7
```
– хипотезата относно популационната вероятност

```
> prop.test(x, n, p, alternative="less")
```
– алтернативата е едностранна ($p < 70$), затова параматърът `alternative` е `"less"`

Полученият резултат говори, че не трябва да отхвърляме нулевата хипотеза.

```
	1-sample proportions test with continuity correction

data:  x out of n, null probability p
X-squared = 14.407, df = 1, p-value = 0.9999
alternative hypothesis: true p is less than 0.7
95 percent confidence interval:
 0.0000000 0.7673247
sample estimates:
        p
0.7479432
```

Заключението следва от това, че p-value е много по-голямо от нивото на съгласие ($0.9999 > 0.05$). Също така, нулевата хипотеза, че $p \geq 70$ ($p = 70$), попада в доверителния интервал: $0.7 \in (0, 0.7673247)$