Hipotezės su 2 imtimis

Matas Amšiejus

5/10/2021

1)

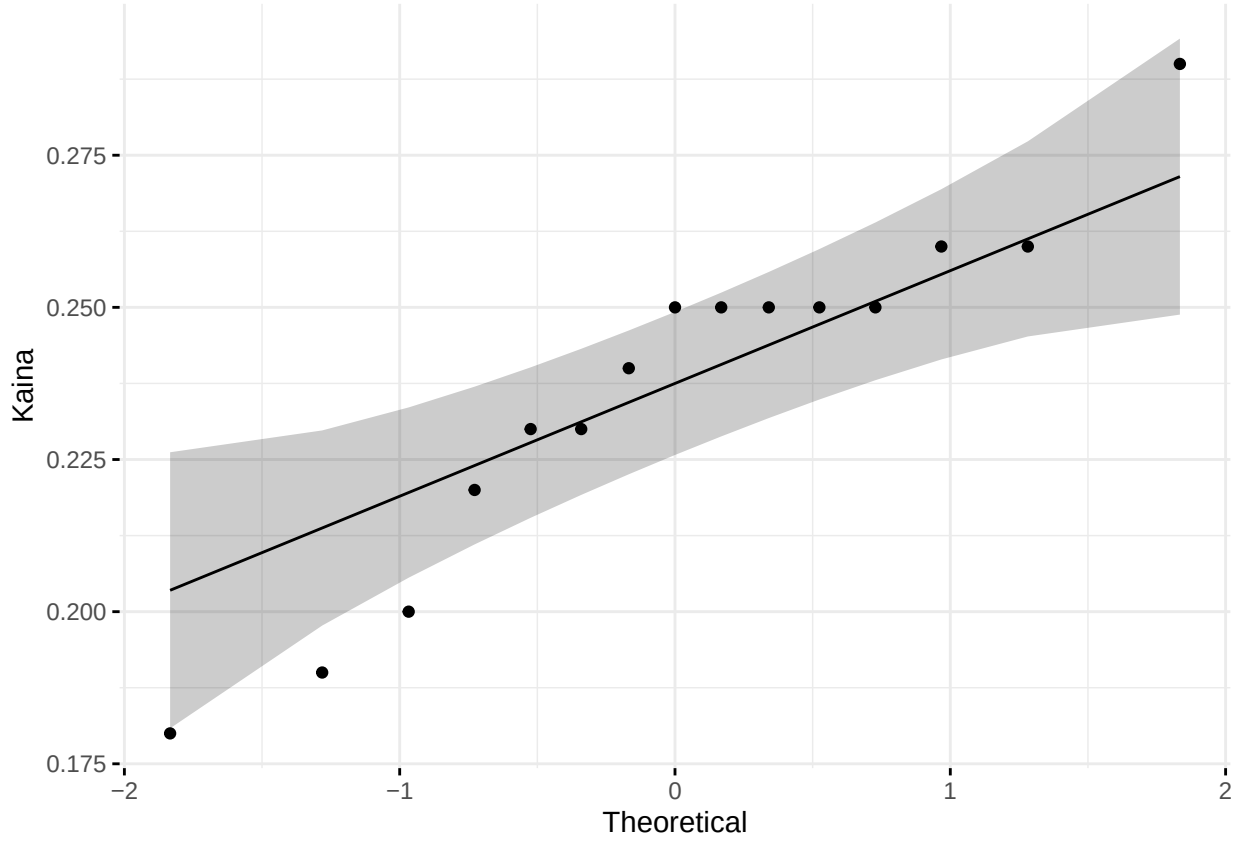
Tyrimo tikslas – nustatyti, ar vidutinė duomenų rinkinio analizės kaina, naudojant statistinį paketą A yra tokia pati kaip naudojant paketą B. Lentelėje pateikta 15 duomenų rinkinių, kurie buvo analizuoti statistiniu paketu A (kintamasis X) ir statistiniu paketu B (kintamasis Y), analizės kaina.

```
Xi<-c(0.26,0.24,0.26,0.22,0.25,0.23,0.18,0.25,0.19,0.25,0.29,0.25,0.23,0.2,0.25)
Yi<-c(0.29,0.32,0.24,0.33,0.28,0.27,0.25,0.26,0.33,0.33,0.33,0.28,0.30,0.24,0.32)
data1<-data.frame(Xi,Yi)
```

Sprendimas

Vadinasi mums duotos 2 imtys, kurios yra susijusios (tiriami tie patys duomenys, tik kitais paketais). Vadinasi imtys (kainos) priklausomos. Galime patikrinti ar iš tiesų dydis pasiskirstęs pagal normalųjį skirstinį:

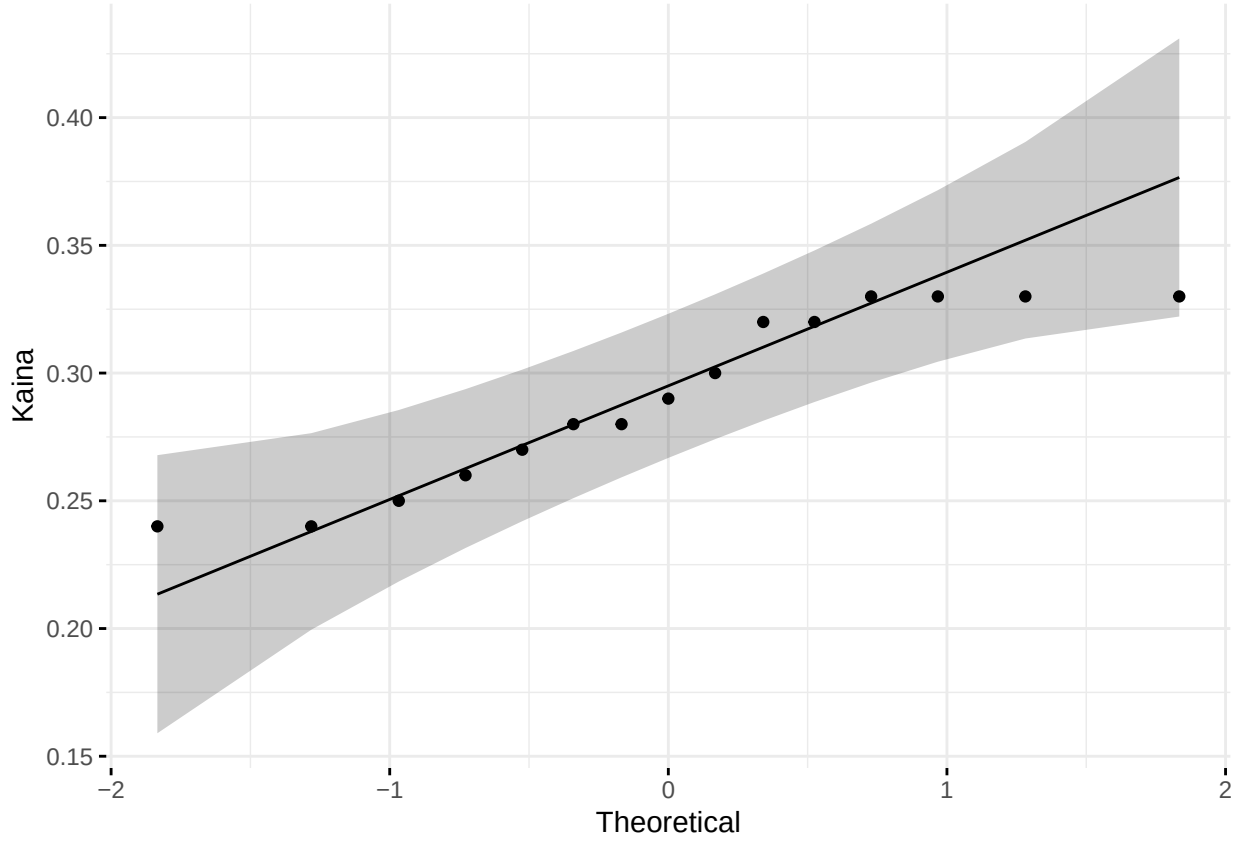
```
ggqqplot(data1$Xi,ylab = "Kaina", ggtheme = theme_minimal())
```



```
shapiro.test(data1$Xi)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data1$Xi
## W = 0.92903, p-value = 0.2639
```

```
ggqqplot(data1$Yi,ylab = "Kaina", ggtheme = theme_minimal())
```




```
shapiro.test(data1$Yi)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data1$Yi
## W = 0.88489, p-value = 0.05616
```

Keista, nes pagal kvantilių grafiką atrodė, jog Xi nepasisiskirstęs pagal N, o Yi - pasiskirstęs, bet Shapiro test rodo kitaip (Yi pasiskirstęs, bet ant ribos)
Dėl įdomumo patikrinkime koreliaciją tarp imčių:

```
cor.test(data1$Xi,data1$Yi)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data1$Xi and data1$Yi
```

```
## t = 0.78307, df = 13, p-value = 0.4476
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3366248  0.6534549
## sample estimates:
##       cor
## 0.2122361
```

Išties gauname, kad su pasikliovimo lygmeniu $\alpha = 0.05$, o $p = 0.4479$ egzisutuoja koreliacija tarp imčių. Tada

```
t.test(data1$Xi,data1$Yi,paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data1$Xi and data1$Yi
## t = -5.2962, df = 14, p-value = 0.0001129
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07680474 -0.03252859
## sample estimates:
## mean of the differences
##             -0.05466667
```

Gauname, kad p-value $p < \alpha$, kur $\alpha = 0.05$, todėl hipotezę atmetame, nes vidurkiai statistiškai reišmingai skiriasi. Jei tiksliai suprantu, antras yra 0.0547 didesnis už pirmą.

Taip pat dėl įdomumo galime ieškoti skirtumų ir tada pritaikyti t.test vienai imčiai. Taigi randame skirtumus:

```
Zi<-Xi-Yi;Zi
```

```
##  [1] -0.03 -0.08  0.02 -0.11 -0.03 -0.04 -0.07 -0.01 -0.14 -0.08 -0.04 -0.03
## [13] -0.07 -0.04 -0.07
```

Dabar atnaujinsime hipotezę, t.y. $H_0 : \mu_z = 0$, $H_1 : \mu \neq 0$. Tikriname su t.test:

```
t.test(Zi)
```

```
##
##  One Sample t-test
##
## data:  Zi
## t = -5.2962, df = 14, p-value = 0.0001129
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.07680474 -0.03252859
## sample estimates:
##   mean of x
## -0.05466667
```

2)

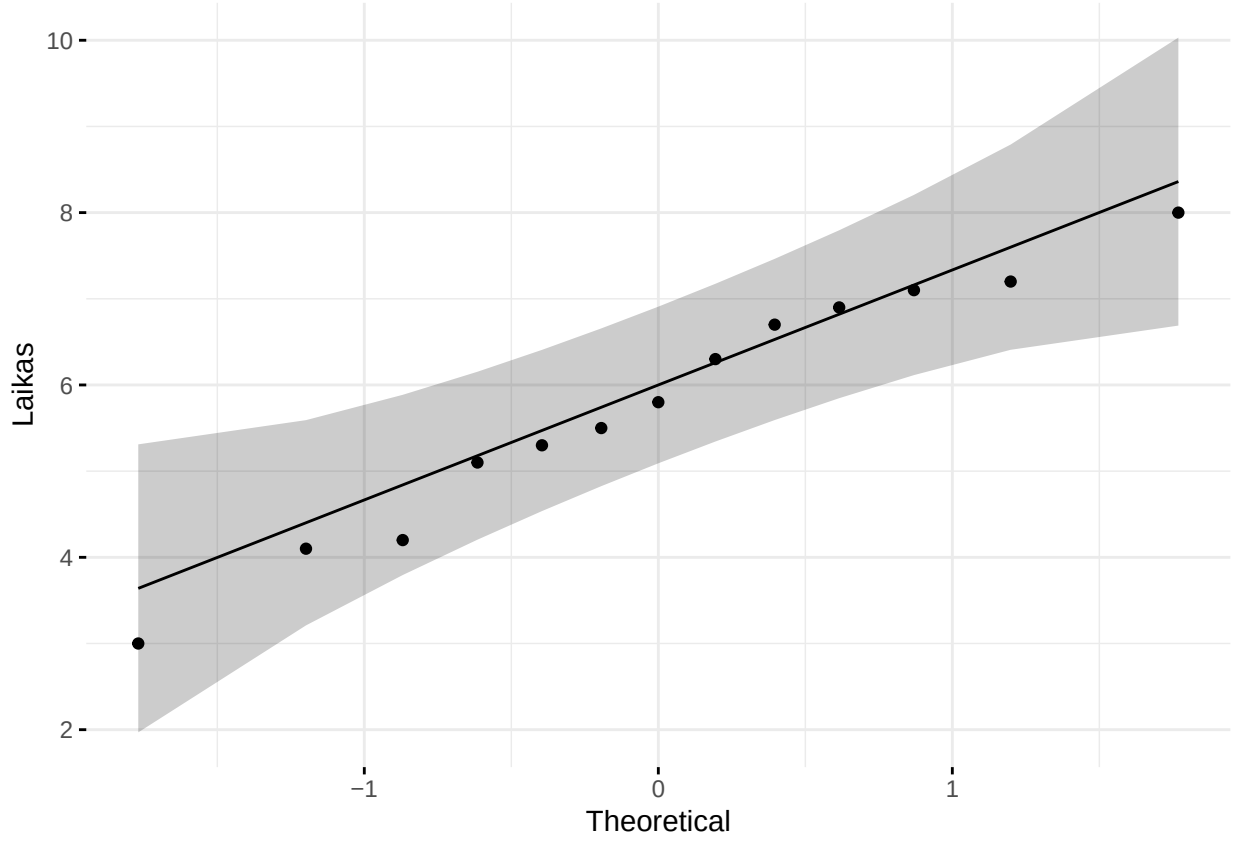
Tyrimo tikslas palyginti laiką, kuris reikalingas patikrinti laidų sujungimus ir izoliaciją dviejų tipų srovės pertraukikliuose. Pirmoji populiacija susideda iš visų vakuumo tipo srovės pertraukiklių, o antroji populiacija iš visų oro-magnetinių srovės pertraukiklių. Buvo atrinktos paprastos atsitiktinės imtys iš kiekvienos populiacijos ir ištirtas kiekvienas į imtį patekęs gaminys (matuotas laikas).

```r
vak<-c(3.0,5.3,6.9,4.1,8.0,6.7,6.3,7.1,4.2,7.2,5.1,5.5,5.8)
oroMag<-c(7.1,9.3,8.2,10.4,9.1,8.7,12.1,10.7,10.6,10.5,11.3,11.5)
```

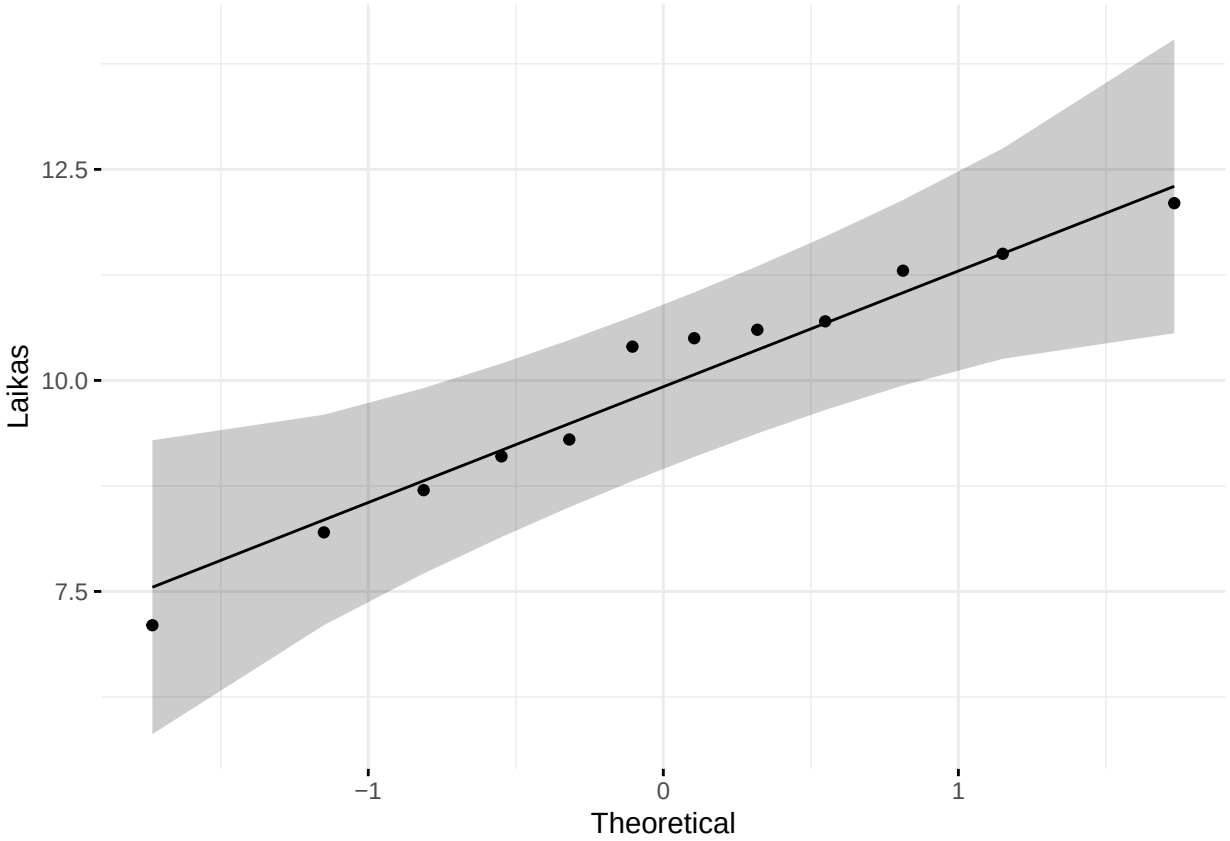
Sprendimas

Pagal sąlygą atrodo, kad buvo paimtos 2 nepriklausomos imtys, tuo labiau, nes imčių dydžiai nevienodi. Tikrinkime, ar dydžiai pasiskirstę pagal N skirstinį:

```r
ggqqplot(vak,ylab = "Laikas", ggtheme = theme_minimal())
```




```r
ggqqplot(oroMag,ylab = "Laikas", ggtheme = theme_minimal())
```



Iš grafikų atrodo, kad abi imtys išties pasiskirsčiusios pagal N skirstinį. Dabar tikrinkime, ar dispersijos lygios:

```
var.test(vak,oroMag)
```

```
##
##  F test to compare two variances
##
## data:  vak and oroMag
## F = 0.94117, num df = 12, denom df = 11, p-value = 0.9131
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2744255 3.1260897
## sample estimates:
## ratio of variances
##          0.9411734
```

Gauname, kad statistiškai dispersijos nesiskiria su p-value $0.9131 > 0.05 = \alpha$. Tada keliame hipotezę ir alternatyvą, bet kadangi tikslas palyginti laiką, manau, kad turėtume ieškoti, kad kažkuris didesnis. Todėl pirma galime tiesiog patikrinti vidurkius:

```
vid_vak<-mean(vak);vid_vak
```

```
## [1] 5.784615
```

```
vid_oro<-mean(oroMag);vid_oro
```

```
## [1] 9.958333
```

Antro vidurkis didesnis, todėl tikrinsime hipotezę:
$H_0 : \mu_{vak} \leq \mu_{oro}$
$H_1 : \mu_{vak} > \mu_{oro}$

```
t.test(vak, oroMag, alternative = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  vak and oroMag
## t = -7.1353, df = 23, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -5.17623      Inf
## sample estimates:
## mean of x mean of y
##  5.784615  9.958333
```

Gauname, kad pirmos imties vidrukis statistiškai reikšmingai skiriasi (yra mažesnis) negu antros.

3)

Tikrinama hipotezė, kad impulso atpažinimo paklaidos dispersija nepriklauso nuo jo intensyvumo. Buvo atlikti du nepriklausomi eksperimentai. Impulsas, kurio intensyvumas lygus 10 sąlyginių vienetų, buvo įvertintas: 9 9 8 10 12 13 10 10 vienetų; impulsas, kurio intensyvumas 20 sąlyginių vienetų, buvo įvertintas 15 16 17 23 22 20 21 24 27

```
d10<-c(9,9,8,10,12,13,10,10)
d20<-c(15,16,17,23,22,20,21,24,27)
```

Sprendimas Iš sąlygos gauname hipotezę:
$H_0 : \sigma_1^2 = \sigma_2^2$;
$H_1 : \sigma_1^2 \neq \sigma_2^2$.

```
var.test(d10,d20)
```

```
##
##  F test to compare two variances
##
## data:  d10 and d20
## F = 0.1709, num df = 7, denom df = 8, p-value = 0.03098
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.03773834 0.83729929
## sample estimates:
## ratio of variances
##          0.1709004
```

Tagi gauname, kad p-value$= 0.031 < 0.05 = \alpha$, todėl hipotezę, kad dispersijos nesiskiria priklausomai nuo impulso intensyvumo, atmetame.

4)

Tiriant nutekamųjų vandenų poveikį ežero vandeniui, nitratų koncentracija vandenyje matuojama rankiniu metodu. Siūlomas naujas automatinis metodas. Jeigu yra stipri teigiama

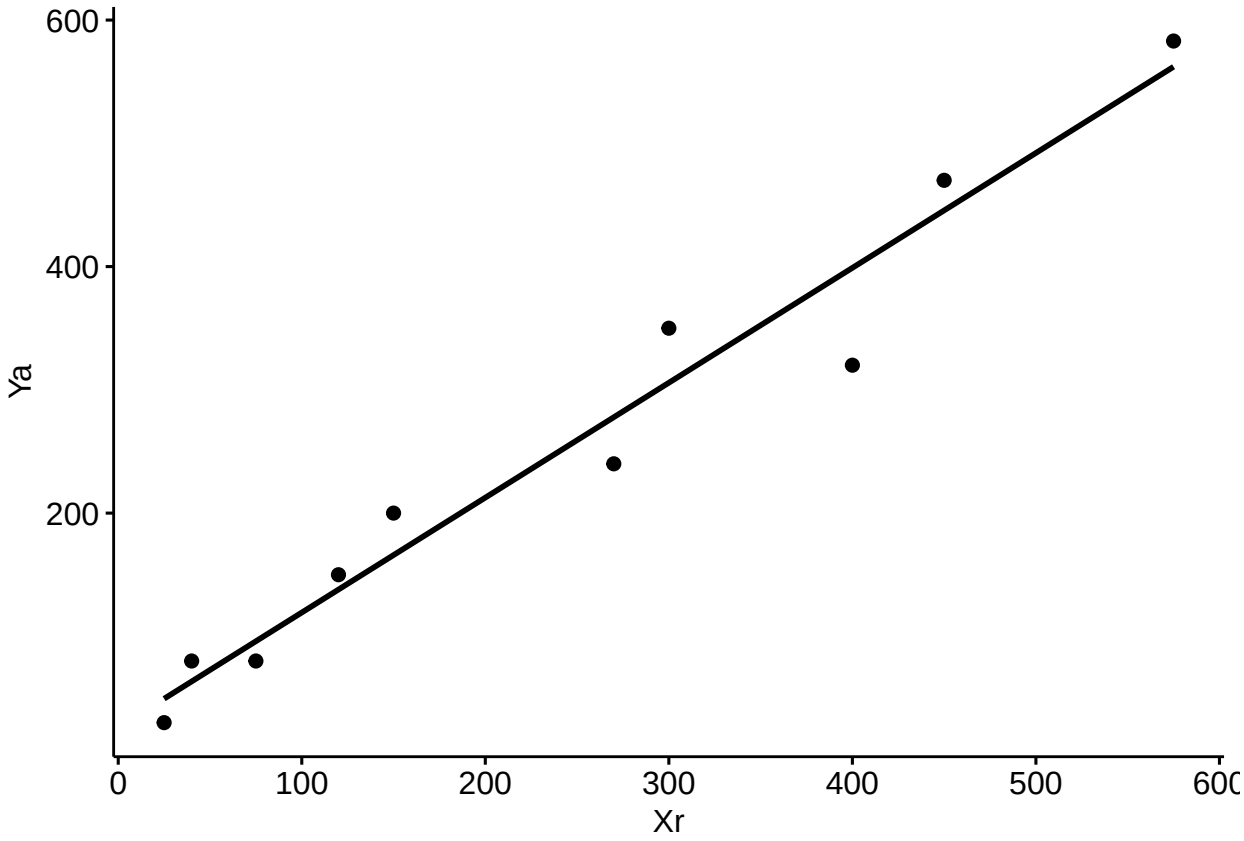
koreliacija tarp matavimų, gautų šiais metodais, tai rankinis metodas, bus keičiamas nauju automatiniu. Duomenys:

```
Xr<-c(25,40,120,75,150,300,270,400,450,575)
Ya<-c(30,80,150,80,200,350,240,320,470,583)
```

Išvadas formuluokite naudodami Pearson koreliacijos koeficientą.
Pastaba. Pearson koeficientas naudojamas, kai yra tiesinis ryšys tarp kintamųjų, jei sąryšis netiesinis jo naudoti negalime. Prieš skaičiuojant Pearson koreliacijos koeficientą reikia patikrinti ar ryšys tiesinis, pvz. nubraižyti duomenų sklaidos diagramą **Sprendimas**
Pirma reikia patikrinti, ar ryšys yra tiesinis:

```
data4<-data.frame(Xr,Ya) #ggscatter reikia tureti data.frame pasidarius
ggscatter(data = data4, x="Xr", y="Ya", add = "reg.line")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Matome, kad ryšys tarp kintamųjų yra tiesinis. Vadinasi tereikia patikrinti, ar egzistuoja stipri teigiama koreliacija su Pearson koreliacijos koeficientu. Mūsų hipotezė: $H_0 : kor \geq 0$(manau, kad reikia, jog būtų > 0.5)
$H_1 : kor < 0$

```
cor.test(Xr, Ya, alternative="less", method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Xr and Ya
```

```
## t = 13.195, df = 8, p-value = 1
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000  0.9935435
## sample estimates:
##       cor
## 0.9777899
```

Matome, kad koreliacijos koficientas $= 0.9778$ su p-value$= 1 > 0.05 = \alpha$. Todėl galime tvirtinti, kad verta keisti rankinį metodą į naują automatinį.

5)

Lentelėse pateikti duomenys apie Li2O ir Rb2O kiekį (0.01 procentais) mikroklinuose iš albitinių pegmatitų ir albitinių-spodumeninių pegmatitų.

```
# pirma lentele
Li20<-c(0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,1.5,0.7,3.5,1.2,2.5,2.5,0.5,0.5,0.5,0.8,0.2,0.2)
Rb20<-c(22,18,18,22,22,53,38,15,84,7,90,80,80,88,25,40,40,40,27,19)
albit<-data.frame(Li20,Rb20)
# antra lentele
Li20<-c(0.5,0.8,1.3,2.5,1.5,0.6,0.5,1.1,0.5,1.0,0.7,1.0,0.2,1.4,1.2,2.4,1.2,0.5,0.5,0.6)
Rb20<-c(43,40,46,42,43,38,40,46,39,50,50,49,50,46,51,51,43,51,43,48)
spodum<-data.frame(Li20,Rb20)
```

a)Esant 5 procentų reikšmingumo lygmeniui, patikrinkite hipotezę, kad Rb2O kiekio dispersija abiejų rūšių pegmatituose nesiskiria.
b)Esant 5 procentų reikšmingumo lygmeniui, patikrinkite hipotezę, kad Li2O kiekis abiejų rūšių pegmatituose nesiskiria.
c)Nubrėžkite Li2O ir ir Rb2O kiekio albitiniuose pegmantintuose sklaidos diagramą. Esant 1 % reikšmingumo lygmeniui, patikrinkite hipotezę, kad abiejų medžiagų kiekiai albitiniuose pegmatituose nepriklauso vienas nuo kito.
Sprendimas
a)
Tereikia patikrinti, ar dispersijos yra lygios, tai mūsų
$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$

```
var.test(albit$Rb20,spodum$Rb20)
```

```
##
##  F test to compare two variances
##
## data:  albit$Rb20 and spodum$Rb20
## F = 39.686, num df = 19, denom df = 19, p-value = 4.015e-11
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   15.70823 100.26489
## sample estimates:
## ratio of variances
##           39.68606
```

Gauname, kad hipotezę atmetame su p-value$< \alpha = 0.05$, nes dispersijos statistiškai skiriasi.
b)
Reikia patikrinti, kad kiekis nesiskiria $\rightarrow$ tikrinsime vidurkį (o gal sumą reikėtų?). Pirma tikriname, ar dydžiai pasiskirstę pagal normalųjį skirstinį:

```
shapiro.test(albit$Li20)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  albit$Li20
## W = 0.67813, p-value = 2.135e-05
```

```
shapiro.test(spodum$Li20)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  spodum$Li20
## W = 0.86393, p-value = 0.009208
```

Gauname, kad abu dydžiai nėra pasiskirstę pagal N skirstinį, tačiau tęsiame darbą. Tikriname, ar dispersijos yra lygios:

```
var.test(albit$Li20,spodum$Li20)
```

```
##
##  F test to compare two variances
##
## data:  albit$Li20 and spodum$Li20
## F = 2.1151, num df = 19, denom df = 19, p-value = 0.1111
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8371926 5.3437622
## sample estimates:
## ratio of variances
##           2.115126
```

Gauname, kad su p-value$= 0.1111 > 0.05 = \alpha$ dispersijos statisstiškai reikšmingai nesiskiria viena nuo kitos. Tai:

```
t.test(albit$Li20, spodum$Li20, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  albit$Li20 and spodum$Li20
## t = -0.28934, df = 38, p-value = 0.7739
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5597678  0.4197678
## sample estimates:
## mean of x mean of y
##      0.93      1.00
```

Gauname, kad su p-value$= 0.7739 > 0.05 = \alpha$ mūsų vidurkiai statistiškai reikšmingai nesiskiria, t.y. kiekis vienodas.

c) Patikrinkime, ar abiejų medžiagų kiekiai nepriklausomi albitiniuose pegmatituose ($H_0 : kor = 0$). Taigi

```
cor.test(albit$Li20, albit$Rb20)
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  albit$Li20 and albit$Rb20
## t = 6.3301, df = 18, p-value = 5.773e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6137844 0.9309800
## sample estimates:
##       cor
## 0.8306818
```

```
cor.test(spodum$Li20, spodum$Rb20)
```

```
##
##  Pearson's product-moment correlation
##
## data:  spodum$Li20 and spodum$Rb20
## t = 0.50924, df = 18, p-value = 0.6168
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3413495  0.5335558
## sample estimates:
##       cor
## 0.1191728
```

Tai galima teigti, kad pirmajame *dalyke* koreliacija labai didelė, p-value$< 0.01 = \alpha$, todėl galime teigti, kad priklausoma. Antrajmame koreliacija nedidelė, p-value$= 0.6168 > 0.01 = \alpha$, todėl priimame hipotezę.