Statistika za jezikoslovno istraživanje

Damir Ćavar
Sveučilište u Zadru
14. travnja 2010.

Domaći

- Jednadžba u R-u, za x niz vjerojatnosti (tj. vrijednosti između 0 i 1, i sum(x)=1:

$$-\sum_{i=1}^{n} p(x_i) log_b p(x_i)$$

Domaći

- U R-u:
  - -sum(x * log2(x))
- Primjer:
  - x <- c(0.1, 0.3, 0.11, 0.3, 0.08, 0.02, 0.09)
  - -sum(x * log2(x))
  - 1.724659

Promjene razmjera podataka

Promjene razmjera

- Ako imamo rezultate:
  - 1, 3, 4, 5, 7
- Kolika je aritmetička sredina?
- Koja je srednja vrijednost?
- Koja je standardna devijacija?

Promjene razmjera

- Za rezultate u R-u: `x <- c(1, 3, 4, 5, 7)`
  - Aritmetička sredina: 4
    - R: `mean(x)`
  - Srednja vrijednost: 4
    - R: `median(x)`
  - Standardna devijacija: 2
    - R: `sqrt(sum((x-mean(x))^2)/length(x))`

R funkcije

- Definiramo funkciju za standardnu devijaciju u R-u:

```
sdd <- function (x) { sqrt(sum
(((x)-mean((x)))^2)/length(x)) }
```

- Pozivamo funkciju:

```
sdd(x)
```

Promjene razmjera

- Što ako mjenjamo rezultate, npr. dodajemo 4 ili oduzimamo 4 od svake vrijednosti u rezultatima?
  - u R-u:

```
mean(x-4)

median(x-4)

sdd(x-4)
```

Promjene razmjera

- Dodavanje ili oduzmanje jedne konstantne vrijednosti (npr. 4) svakoj vrijednosti
  - diže ili smanjuje aritmetičku sredinu i srednju vrijednost za vrijednost te konstante
  - standardna devijacija ostaje ista
- Histogram se samo miče desno ili lijevo.

Promjene razmjera

- Što će se desiti ako umnožavamo ili dijelimo sve vrijednosti s pozitivnim konstantnim brojevima?

```
mean(x*3)

median(x*3)

sdd(x*3)
```

Promjene razmjera

- Množenje svih vrijednosti s pozitivnom konstantnom vrijednosti (npr. 3):
  - Umnožava aritmetičku sredinu i srednju vrijednost s tom konstantom
  - Standardna devijacija se isto umnožava s tom konstantom
- Histogram se rasteže.

Promjene razmjera

- Množenje svih vrijednosti s negativnom konstantnom vrijednosti (npr. -1)
    - Umnožava aritmetičku sredinu i srednju vrijednost za vrijednost te konstante
    - Standardna devijacija se isto umnožava za vrijednost te konstante
- Histogram se rasteže (za vrijednosti veće od -1), ali se redoslijed vrijednosti odražava obratni redoslijed vrijednosti.

Promjene razmjera

- Ako odbijemo od svakog rezultata aritmetičku sredinu i dijelimo kroz standardnu devijaciju, što to znači na kraju za rezultirajuće mjere?

```
(x - mean(x))/sdd(x)
```

- Koje vrijednosti dobijemo za:
  - aritmetičku sredinu
  - standardnu devijaciju

Promjene razmjera

- Histogram distribucije se pomiče lijevo i centrira na 0
  - `mean(y)`: 0
- Standardna devijacija se ne mijenja ako dodajemo ili oduzimamo konstantnu vrijednost od svakog rezultata:
  - `sdd(x-mean(x))`: 2

Promjene razmjera

- Ako dijelimo svaku vrijednost kroz standardnu devijaciju distribucije, mijenjamo gustoću, smanjujemo raspršenost:

  `(x-mean(x))/sdd(x)`

  `-1.5 -0.5  0.0  0.5  1.5`

- Tako da je `sdd(x)` na kraju uvjek `1`
  - Zašto?

Promjene razmjera

- Ako dijelimo vrijednosti rezultata kroz neki konstantni broj, npr. `sdd(x)`, posljedice za rezultirajuću standardnu devijaciju su da se i ona dijeli kroz tu konstantu.
- I:

```
sdd( (x-mean(x)/sdd(x) )
```

  - je isto kao:

```
sdd(x-mean(x)) / sdd(x)
```

  - Za svaki broj n (≠ 0), n/n = 1

Promjene razmjera

- Pretvaranje distribucije u standardne mjere (tkzv. z-vrijednosti u literaturi)
  - dimenzije imaju standardne mjere:
    - aritmetička sredina 0
    - standardna devijacija 1
  - Ne znači da sve distribucije izgeladju iste, da su uopće normalne distribucije!

Standardne mjera

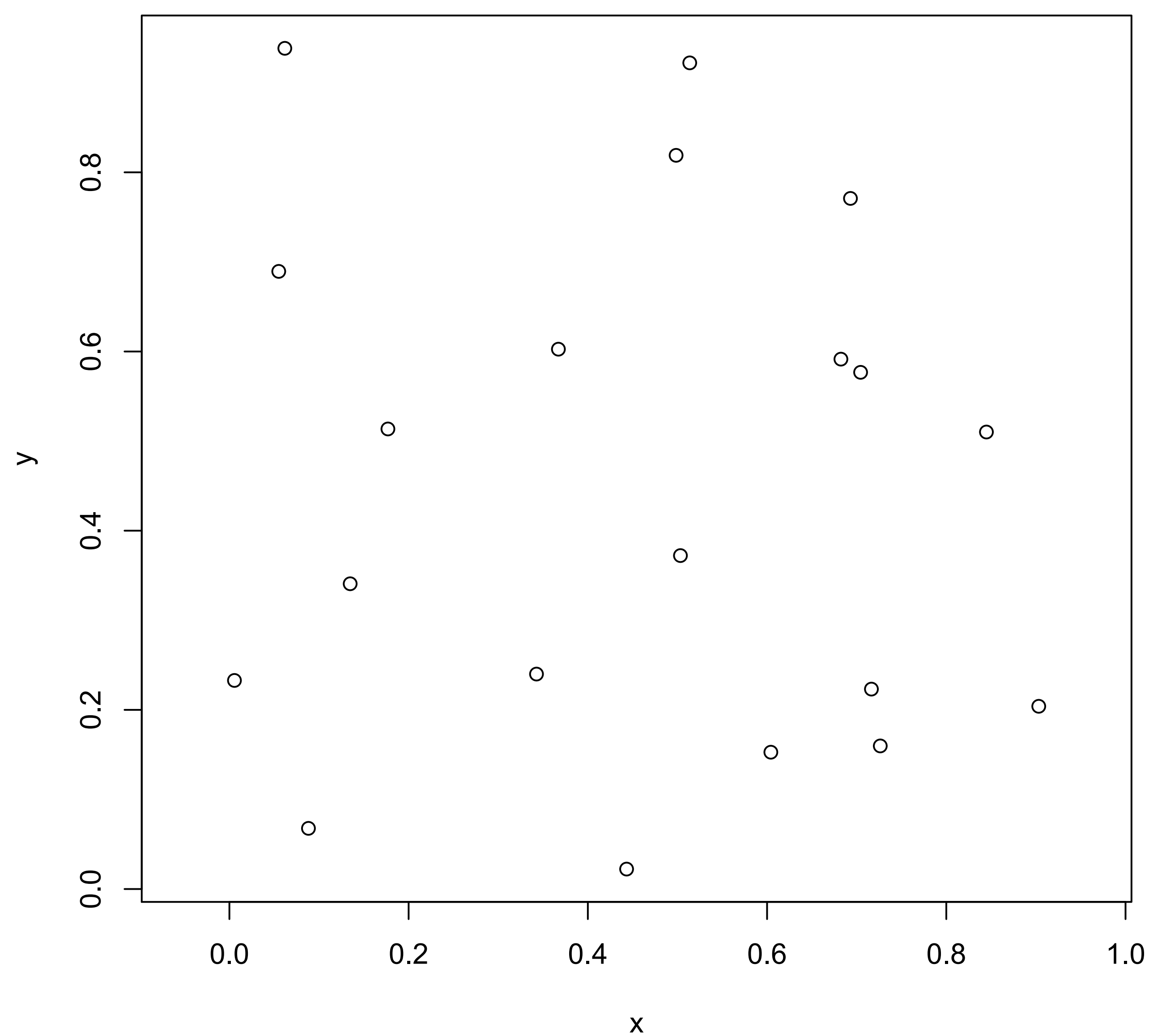
- Standardna mjera nam daje odgovor na pitanje:
  - Koliko standardnih devijacija je svaka mjera odmaknuta od aritmetičke sredine?

Korelacijski koeficijent

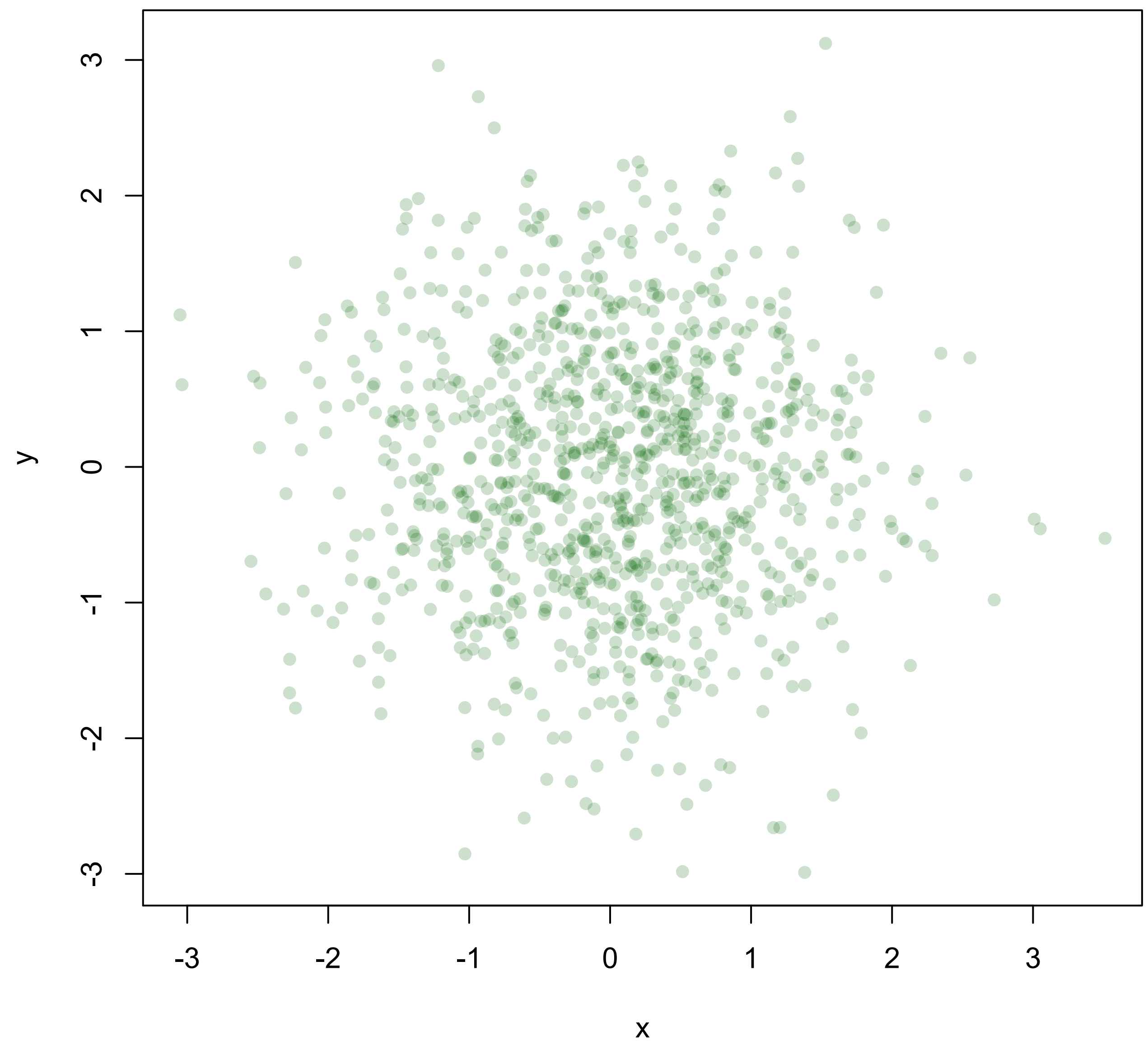
Korelacijski koeficijent

- Do sada:
  - Centar distribucije: `mean(x)`
  - Simetrija: `median(x)`, `mode(x)` u relaciji s `mean(x)`
  - Raspršenost: `sdd(x)`
- Ako imamo dvije varijable, kao u primjerima kada smo računali Chi2 vrijednosti, što želimo znati ili pronaći?

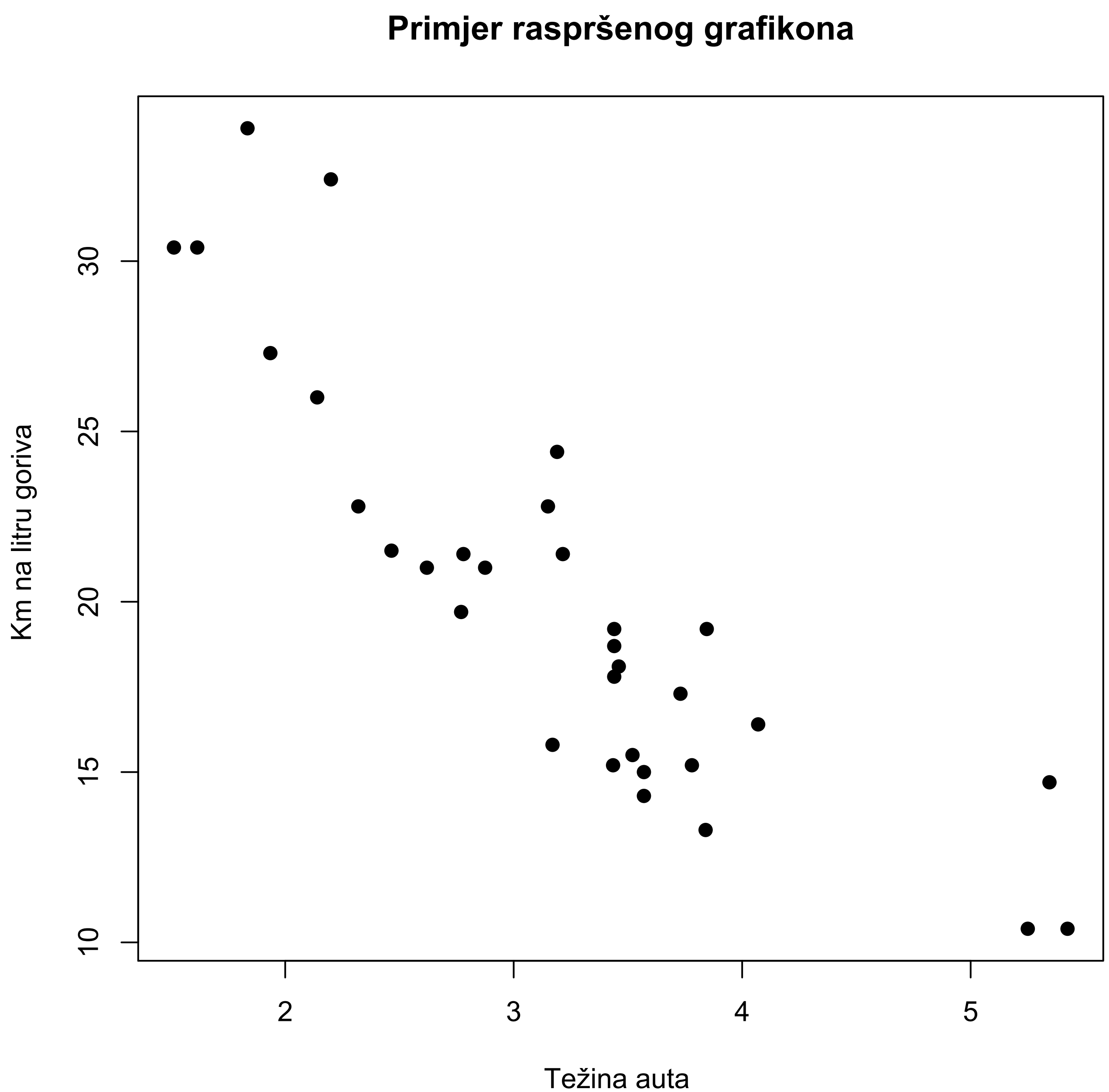
Raspršeni grafikon



Raspršeni grafikon

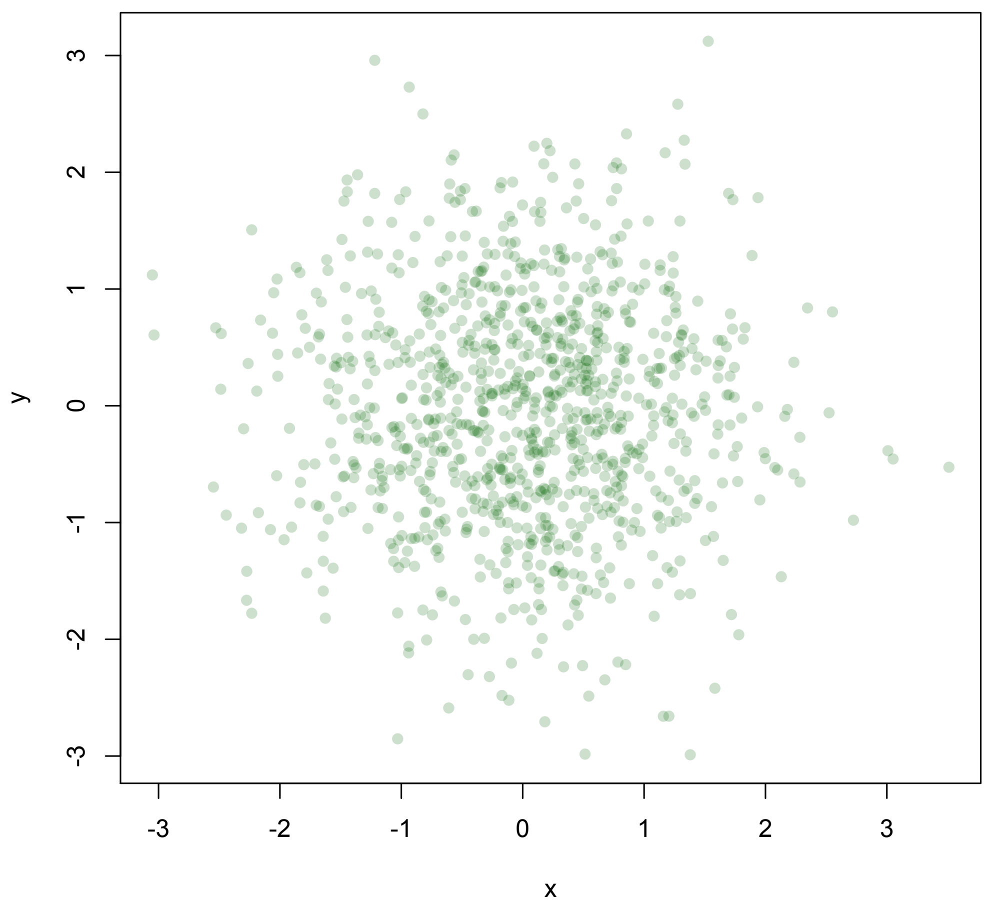


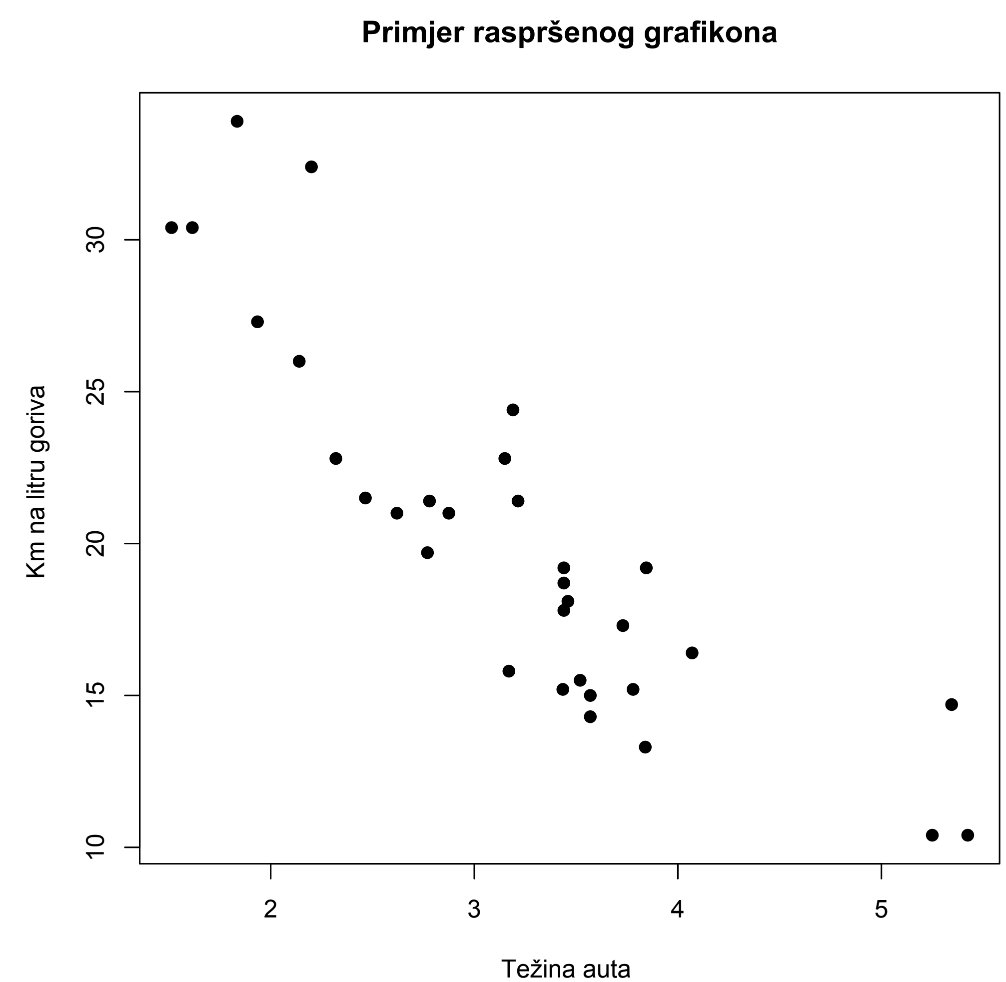
Raspršeni grafikon



Raspršeni grafikon

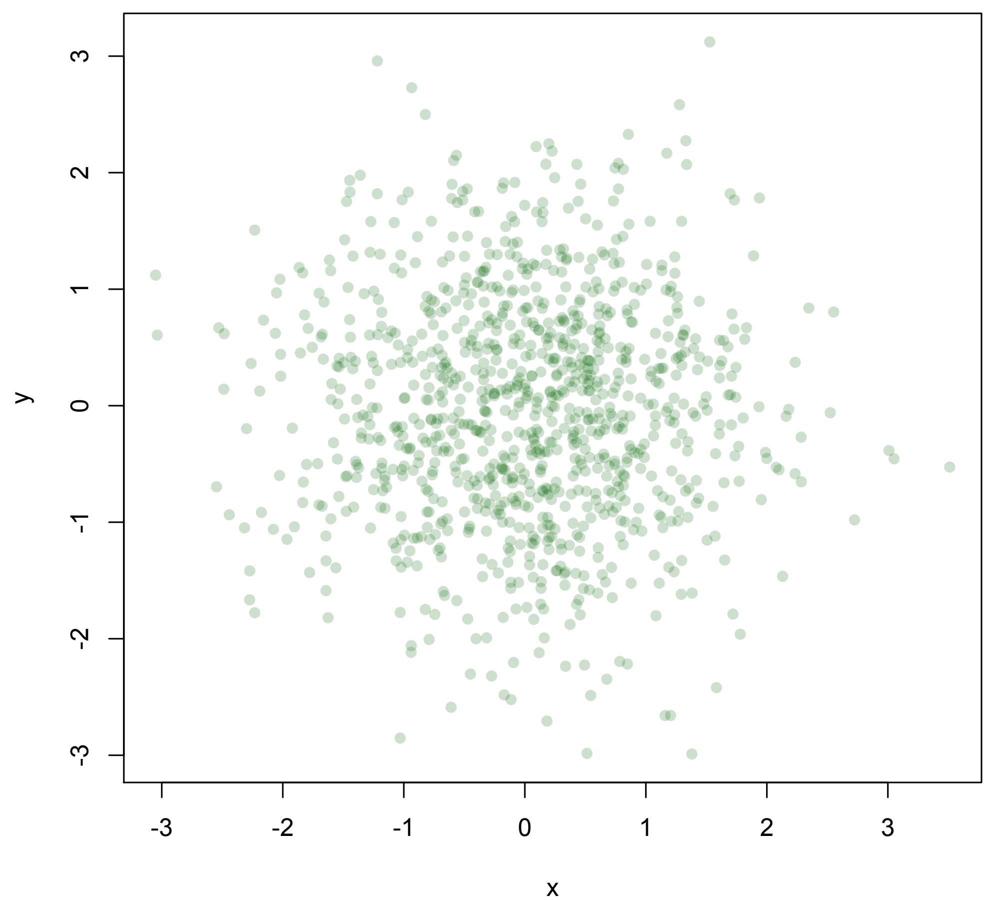
- Što nam pokazuje raspršeni grafikon, ako uspoređujemo prvi i drugi primjer?





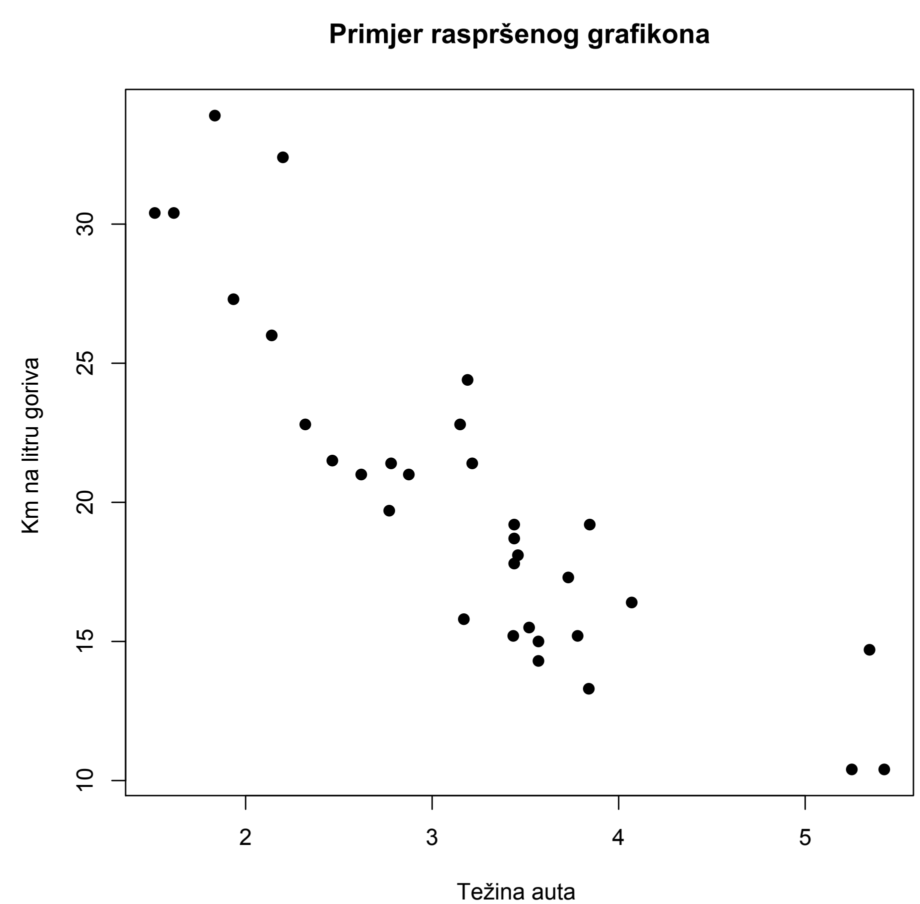
Raspršeni grafikon

- U prvom primjeru dvije varijable ne izgledaju ovisne, ne koreliraju, ili koreliraju minimalno:



Raspršeni grafikon

- U drugom primjeru dvije varijable izgledaju ovisne, koreliraju, ili koreliraju značajno:



Raspršeni grafikon

- Oblik:
  - krug: nema korelacije
  - oval: moguća umjerena korelacija
  - linija: jaka korelacija

Korelacijski koeficijent

- Korelacijski koeficijent `r`
  - mjeri jakost jačine linearne relacije između dvaju varijabli.
  - Koliko blizu su svi rezultati jednoj liniji?

Korelacijski koeficijent

- Korelacijski koeficijent `r`
    - r nema mjere
    - r je uvijek između -1 i 1, tj. ako r = -1 ili r = 1: rezultati se nalaze na jednoj liniji, s negativnim ili pozitivnim nagibom
        - ako r = -1:
            - što veće X vrijednosti, to manje Y vrijednosti
        - ako r = 1:
            - što veće X vrijednosti, to veće Y vrijednosti

Korelacijski koeficijent

- Ako r=0:
    - Nema korelacije između varijabli.
    - Ne može se naći nijedna idealna linija kroz rezultate.
        - rezultati se slažu u krug
        - rezultati se slažu u neki simetrički oblik (npr. četverokut, trokut), bez idealne linije

Korelacijski koeficijent

- Vrijednosti od r između 0 i 1 ili -1:
  - Što god bliže 1 ili -1, to veća korelacija između varijabli
- Kako izračunati r?
  - Konvertiranje svih vrijednosti u standardnu mjeru
  - Izračunati umnožak tih standardnih mjera za obe varijable
  - Izračunati srednju vrijednost umnoška

Korelacijski koeficijent

- Primjer u R-u:

```
x = c(2, 4, 5, 6, 8)
y = c(5, 2, 4, 8, 6)
mean(x) = mean(y) = 5
sdd(x) = sdd(y) = 2
```

- standardizirana mjera u R-u:

```
(x-mean(x))/sdd(x)
(y-mean(y))/sdd(y)
```

Korelacijski koeficijent

- Definiramo funkciju za standardizirane mjere u R-u:

```
stm <- function(x) { (x-mean
(x))/sdd(x) }
```

- Pozivamo funkciju:

```
stm(x)

-1.5 -0.5  0.0  0.5  1.5
```

Korelacijski koeficijent

- Primjer u R-u:
  - Umnožak standardnih mjera za x i y:
    - psm <- stm(x) * stm(y)
  - Srednja vrijednost tog umnoška:
    - mean(psm)
  - Korelacija u našem primjeru: 0.45
- Zaključak: niska pozitivna korelacija

Korelacijski koeficijent

- Definiramo funkciju za korelacijski koeficijent:

```
mojkk <- function(x, y)
{ mean(stm(x) * stm(y)) }
```

  - ili koristimo jednostavno postojeću funkciju u R-u:

```
cor(x, y)
```

Domaći

- Varijabla X

```
x <- c(3, 4, 8, 4, 2, 1, 0, 6)
```

- Varijabla Y

```
y <- c(1, 2, 4, 2, 2, 0, 1, 4)
```

- Izračunajte:
  - Standardnu mjeru za sve rezultate varijabli
  - Korelacijski koeficijent za X i Y

Dodatni domaći

- Prevedite sljedeću jednadžbu u R:

$$-\sum_{i=1}^{n} p(x_i)\frac{log_b p(x_i)}{log_b(n)}$$