

Normalna prilagoditev frekvenčne porazdelitve

A. Blejec

17. december 2012

Kazalo

1 Problem	1
2 Priprava tabele	2
3 Povprečje in standardni odklon	4
4 Standardizirane vrednosti	5
5 Grafični prikazi	6
5.1 Frekvenčni poligon	6
5.2 Kumulativni frekvenčni poligon	8
5.3 Kumulativni poligon v verjetnostni skali	9

Povzetek

Izračunali bomo pričakovane frekvence za frekvenčno porazdelitev.

1 Problem

Frekvenčna porazdelitev

Denimo, da imamo frekvenčno porazdelitev, ki je podana s tabelo:

Dolžina	f
60 - 62	3
63 - 65	16
66 - 68	40
69 - 71	25
72 - 76	6

Predvidevamo, da je porazdelitev normalna. Frekvenčni porazdelitvi bi radi prilagodili pričakovane frekvence.

2 Priprava tabele

Za izračun pričakovanih frekvenc potrebujem oceno za povprečje in standardni odklon. S pomočjo teh ocen lahko izračunamo pričakovane frekvence za posamezne razrede. Za opis tabele si lahko pomagamo z oznakami razredov in širinami razredov. Iz teh vrednosti lahko izračunamo vse druge opise razredov: spodnje meje, zgornje meje, sredine in zgornje oznake razredov.

Opis tabele

Vpišimo spodnje oznake razredov in širine razredov

```
> so <- c( 60, 63, 66, 69, 72)
> so
[1] 60 63 66 69 72
> (d <- c(3, 3, 3, 3, 5))
[1] 3 3 3 3 5
```

in frekvence v razredih

```
> (f <- c(3, 16, 40, 25, 6))
[1] 3 16 40 25 6
```

Vaša naloga

1. Oglejte si strani s pomočjo za funkciji `rep()` in `seq()`.

```
> ?rep
> ?seq
```

2. Preizkusite nekaj primerov iz razdelka *Examples*

Širine razredov bi lahko podali na drugačne načine

Funkciji `rep()` in `seq()`

Če bi bile vse širine enake, bi lahko pripravili vektor širin s funkcijo `rep`:

```
> (u <- rep(3, 5))
[1] 3 3 3 3 3
```

Zadnji razred pa je širši, zato ga lahko popravim:

```
> u[5] <- 5
> u
[1] 3 3 3 3 5
```

Ali pa:

```
> u <- c(rep(3, 4), 5)
> u
[1] 3 3 3 3 5
```

Meje razredov

Spodnje in zgornje meje razredov

```
> (xmin <- so-0.5)
[1] 59.5 62.5 65.5 68.5 71.5
> (xmax <- xmin + d)
[1] 62.5 65.5 68.5 71.5 76.5
```

Zgornje oznake bi bile

```
> (zo <- xmax-0.5)
[1] 62 65 68 71 76
```

Sredine razredov in frekvence

Sredine razredov

```
> (x <- (xmin+xmax)/2)
[1] 61 64 67 70 74
```

Združimo vse skupaj v stolpce

Sestavimo tabelo

```
> tbl <- data.frame(xmin, xmax, so, zo, d, x, f)
> tbl
  xmin xmax so zo d  x  f
1 59.5 62.5 60 62 3 61  3
2 62.5 65.5 63 65 3 64 16
3 65.5 68.5 66 68 3 67 40
4 68.5 71.5 69 71 3 70 25
5 71.5 76.5 72 76 5 74  6
```

Izvlecimo informacije o drugem in tretjem razredu

```
> tbl[2:3, ]
  xmin xmax so zo d  x  f
2 62.5 65.5 63 65 3 64 16
3 65.5 68.5 66 68 3 67 40
```

Vaša naloga

1. Izvlecite podatek o frekvenci v drugem razredu
2. Kateri razredi imajo frekvenco večjo od 20?
3. Izračunajte gostoto frekvence (shranite v npr. `g`)
4. Zaokrožite gostote na dve decimalki (funkcija `round()`)
5. Setavite tabelo, v kateri bodo spodnje in zgornje oznake, sredine razredov in frekvence

3 Povprečje in standardni odklon

Za izračun vsot je zelo pripravna funkcija `sum()`.

Povprečje

$$n = \sum f_k$$

```
> f
[1] 3 16 40 25 6
> (n <- sum(f))
[1] 90
```

$$\mu = \sum f_k \cdot x_k / n$$

```
> sum(f*x)
[1] 6081
> (povp <- sum(f*x)/n)
[1] 67.56667
```

Standardni odklon

$$\sigma^2 = \sum f_k \cdot (x_k - \mu)^2 / n = \sum (f_k \cdot x_k^2) / n - \mu^2$$

```
> sum(f*(x-povp)^2) / n
[1] 8.245556
> (v <- sum(f*x^2) / n - povp^2)
[1] 8.245556
```

Standardni odklon

```
> (sd <- sqrt(v))
[1] 2.871508
```

4 Standardizirane vrednosti

Standardizirajmo meje

Standardizirane meje

$$z = \frac{x - \mu}{\sigma}$$

```
> zmin <- (xmin - povp) / sd
> zmax <- (xmax - povp) / sd
> data.frame(xmin, xmax,
+ zmin=round(zmin, 2), zmax=round(zmax, 2))
  xmin xmax  zmin  zmax
1 59.5 62.5 -2.81 -1.76
2 62.5 65.5 -1.76 -0.72
3 65.5 68.5 -0.72  0.33
4 68.5 71.5  0.33  1.37
5 71.5 76.5  1.37  3.11
```

Pričakovane verjetnosti

$$P(a < X \leq b) = H(z_b) - H(z_a) = \Phi(z_b) - \Phi(z_a)$$

```
> p <- pnorm(zmax) - pnorm(zmin)
> #
> data.frame(zmin=round(zmin, 2),
+ zmax=round(zmax, 2), p=round(p, 4))
  zmin  zmax      p
1 -2.81 -1.76 0.0363
2 -1.76 -0.72 0.1970
3 -0.72  0.33 0.3916
4  0.33  1.37 0.2872
5  1.37  3.11 0.0844
```

Pričakovane frekvence

$$e_k = n \cdot p_k$$

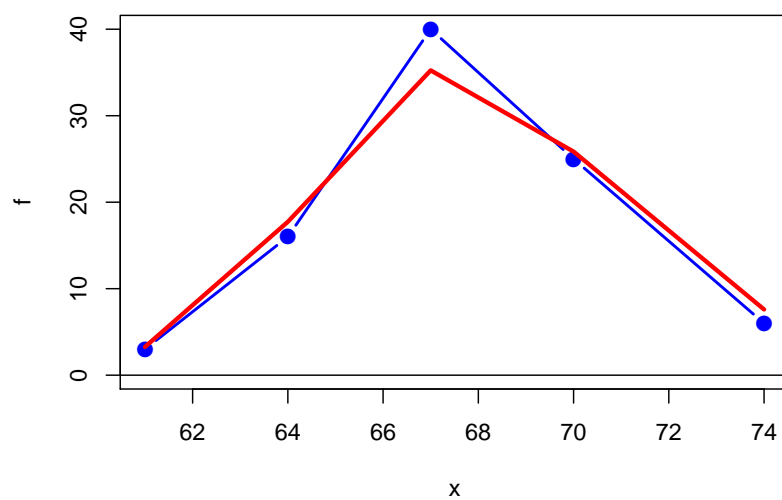
```
> e <- n * p
> data.frame(xmin, xmax, x, f,
+ e=round(e, 2), delta=round(f-e, 2))
  xmin xmax  x  f      e delta
1 59.5 62.5 61  3  3.27 -0.27
2 62.5 65.5 64 16 17.73 -1.73
3 65.5 68.5 67 40 35.24  4.76
4 68.5 71.5 70 25 25.85 -0.85
5 71.5 76.5 74  6  7.60 -1.60
```

5 Grafični prikazi

5.1 Frekvenčni poligon

Frekvenčni poligon

```
> plot(x, f, type="b", ylim=c(0, max(e, f)),  
+ pch=16, cex=1.5, col="blue", lwd=2)  
> abline(h=0)  
> lines(x, e, col=2, lwd=3)
```



Razširitev frekvenčne tabele

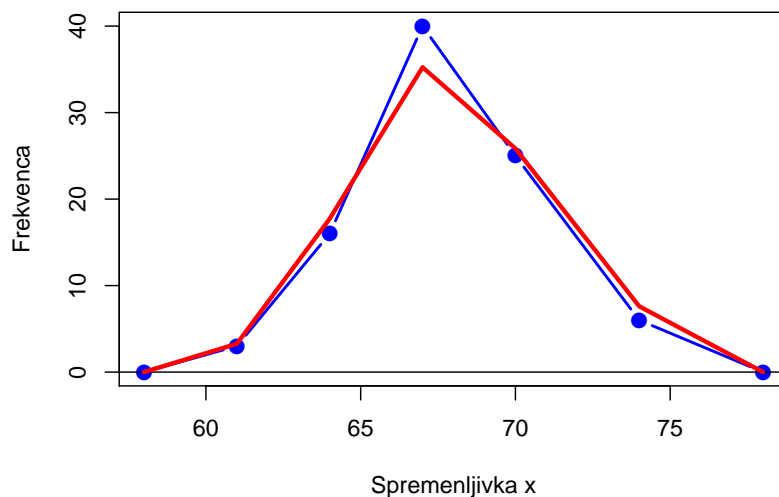
Razširimo porazdelitev z robnima razredoma

```
> m <- length(x)
> f1 <- c(0, f, 0)
> e1 <- c(0, e, 0)
> x1 <- c(x[1]-(x[2]-x[1]), x, x[m]+(x[m]-x[m-1]))
> data.frame(x=x1, f=f1, e=round(e1, 1))
```

	x	f	e
1	58	0	0.0
2	61	3	3.3
3	64	16	17.7
4	67	40	35.2
5	70	25	25.8
6	74	6	7.6
7	78	0	0.0

Dopolnjen frekvenčni poligon

```
> plot(x1, f1, type="b", ylim=c(0, max(e1, f1)),
+ pch=16, cex=1.5, col="blue", lwd=2,
+ xlab="Spremenljivka x", ylab="Frekvenca")
> abline(h=0)
> lines(x1, e1, col=2, lwd=3)
```

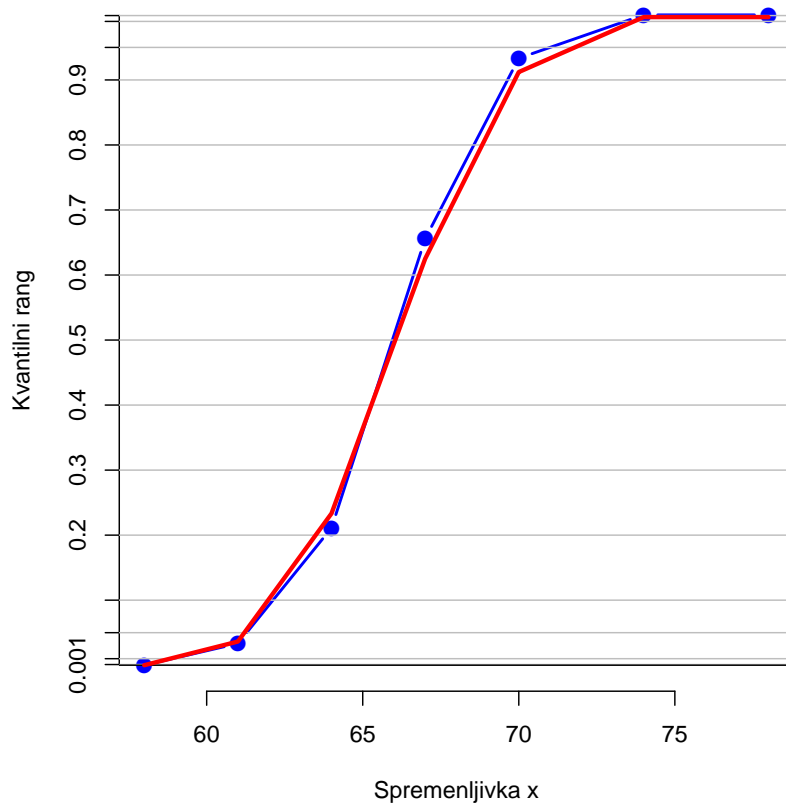


5.2 Kumulativni frekvenčni poligon

Kumulativni frekvenčni poligon

```
> par(mar=c(5, 4, 0, 2))
> plot(x1, cumsum(f1/n), type="b", ylim=c(0, 1),
+ pch=16, cex=1.5, col="blue", lwd=2,
+ xlab="Spremenljivka x", ylab="Kvantilni rang", axes=FALSE)
> lbl <- c(0.001, 0.01, 0.05, seq(0.1, 0.9, .1), 0.95, 0.99, 0.999)
> axis(2, at=(lbl), labels=lbl)
> axis(1)
> abline(h=(lbl), col=8)
> abline(h=0)
> lines(x1, cumsum(e1/n), col=2, lwd=3)
```

Kumulativni poligon

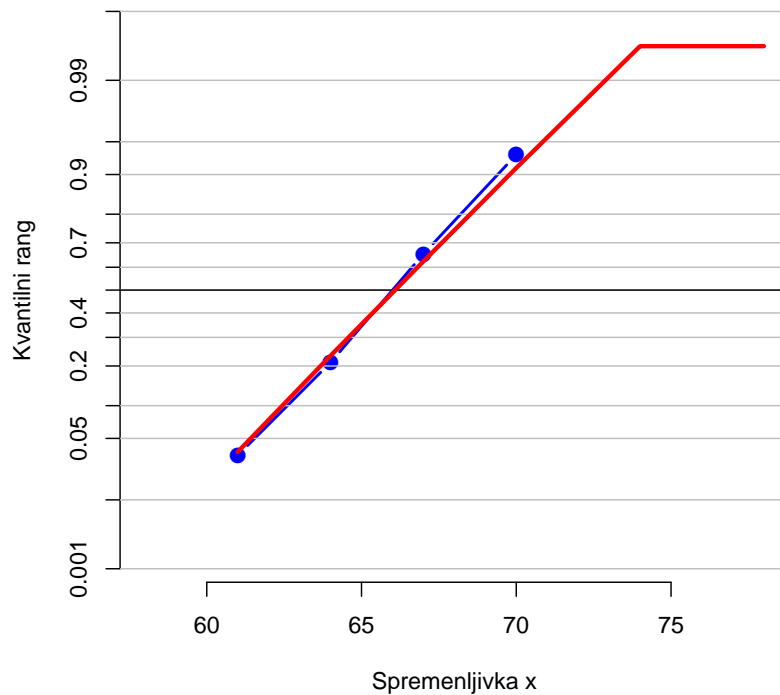


5.3 Kumulativni poligon v verjetnostni skali

Kumulativni verjetnostni poligon

```
> plot(x1, qnorm(cumsum(f1/n)), type="b", ylim=c(-3, 3),  
+ pch=16, cex=1.5, col="blue", lwd=2,  
+ xlab="Spremenljivka x", ylab="Kvantilni rang", axes=FALSE)  
> lbl <- c(0.001, 0.01, 0.05, seq(0.1, 0.9, .1), 0.95, 0.99, 0.999)  
> axis(2, at=qnorm(lbl), labels=lbl)  
> axis(1)  
> abline(h=qnorm(lbl), col=8)  
> abline(h=0)  
> lines(x1, qnorm(cumsum(e1/n)), col=2, lwd=3)
```

Kumulativni verjetnostni poligon



Kakšno je ujemanje kumulativnih frekvenc?

Ujemanje kumulativnih frekvenc

```
> par(mar=c(5, 4, 0, 2))  
> plot(cumsum(e1), cumsum(f1), type="b", lwd=3, col=4)  
> abline(c(0, 1), col=2)
```

