1. Nacrtati histogram i boxplot za sledeće podatke. Prokomentarisati rezultate.

(-0.8, 0.9, 0.6, -1.2, 1.6, -1.5, 0, 0, 1.1, -1.1,-2.2, 0.4, -0.1, -0.8, -0.5, -0.5, -0.1, -1.6, 0.3,-1.6, -1.4, -0.3, 1.5, -0.9, 0)

Rešenje:

Kako bi nam bilo lakše, prvo ćemo soritrati podatke, odnosno odrediti varijacioni niz:

varijacioni niz:

(-2.2, -1.6, -1.6, -1.5, -1.4, -1.2, -1.1, -0.9, -0.8, -0.8, -0.5, -0.5, -0.3, -0.1, -0.1, 0.0, 0.0, 0.0, 0.3, 0.4, 0.6, 0.9, 1.1, 1.5, 1.6)

Za crtanje histograma nam je potreban obim uzorka, minimalni element i uzoračkI raspon:

$$n = 25, \quad R = x_{(n)} - x_{(1)} = 1{,}6 - (-2{,}2) = 3{,}8$$

Odredimo broj intervala nad kojima ćemo crtati histogram:

$$k = \lceil \log_2 n \rceil + 1 = 6$$

Minimalni element uzorka je -2.2, a za levi kraj prvog intervala ćemo izabrati tačku koja je malo manja od toga, pa neka to bude - 2.35. Dužina intervala h se dobija kao:

$$h = \frac{R}{k} = 0{,}63 \approx 0{,}7$$ (ovo ćemo uzeti)

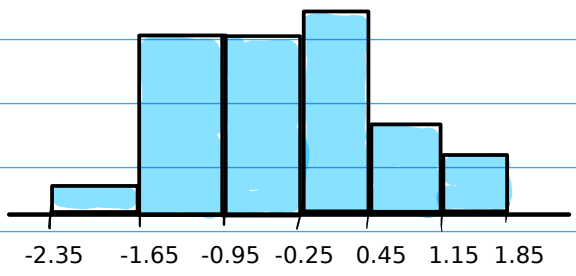
Intervali nad kojima crtamo stupce histograma su:

$$(-2{,}35, -2{,}35 + 0{,}7], (-2{,}35+0{,}7, -2{,}35+2\cdot 0{,}7], \ldots, (-2{,}35 + 5\cdot 0{,}7, -2{,}35+6\cdot 0{,}7]$$
$$= (-2{,}35, -1{,}65], (-1{,}65, -0{,}95], (-0{,}95, -025], (-0{,}25, 0{,}45], (0{,}45, 1{,}15], (1{,}15, 1{,}85]$$

u ovaj interval je upala 1 tačka | -||- 6 tačaka | -||- 6 tačaka | -||- 7 tačaka | -||- 3 tačke | -||- 2 tačke

Sada možemo nacrtati histogram frekvencija:

visine ovih stubaca su redom 1,6,6,7,3 i 2



Ovakav oblik histograma nas asocira na neku simetričnu raspodelu, tj. raspodelu čija je gustina simetrična oko neke tačke. U takve raspodele možemo ubrojati normalnu raspodelu, studentovu, Košijevu raspodelu... Kako često radimo baš sa normalnom raspodelom, aprokismiraćemo gustinu baš gustinom normalne raspodele sa parametrima koje ćemo oceniti iz podataka. Videli smo da parametre normalne raspodele možemo oceniti kao:

$$\hat{m} = \overline{X_n}, \quad \hat{\sigma}^2 = \overline{S}^2$$

Funkcija u R-u koja može izračunati uzoračku sredinu je mean(), a funkcija koja računa popravljenu uzoračku disperziju je var(). Dakle, ako želimo disperziju da ocenimo baš uzoračkom, a ne popravljenom disperzijom, pomnožićemo var() sa n/(n-1).

Za crtanje boxplota su nam potrebni uzorački kvartili. Rekli smo da se oni mogu dobiti na više načina, pa ćemo predstaviti jedan često korišćen.

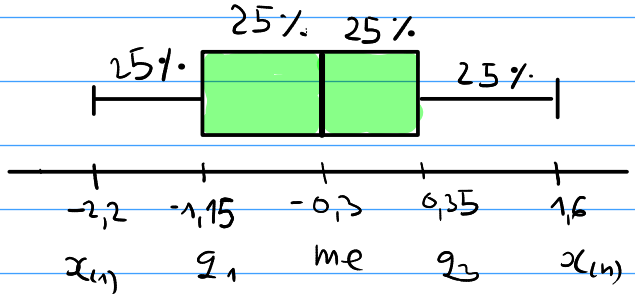
Drugi uzorački kvartil je uzoračka medijana i nju znamo da odredimo. Kako je obim uzorka neparan, uzoračka medijana je srednji element uzorka, a to je -0.3.

Prvi uzorački kvartil deli podatke tako da je 25% njih levo od prvog kvartila, a 75% desno. Zato prvi uzorački kvartil možemo dobiti kao medijanu leve polovine uzorka, tj. kao medijanu prvih 12 tačaka. Treći uzorački kvartil slično dobijamo kao medijanu desne polovine tačaka, tj. kao medijanu poslednjih 12 tačaka. Dakle:

$me = -0,3$ $\quad q_1 = \frac{x_{(6)} + x_{(7)}}{2} = -1,15$ $\quad q_3 = \frac{x_{(19)} + x_{(20)}}{2} = 0,35$

$IQR = q_3 - q_1 = 1,5$ $\quad q_3 + 1,5\,IQR = 2,6$ $\quad q_1 - 1,5\,IQR = -3,4$

Kako je najmanja tačka uzorka veća od -3.4 i najveća tačka uzorka manja od 2.6, vidimo da u ovom uzorku nema autlajera. Dakle, granice boxplota će se nalaziti u $x_{(1)} = -2,2$ i $x_{(n)} = 1,6$



U svakom od ova 4 dela se nalazi 25% tačaka iz uzorka

Ako je kutija približno jednako udaljena od krajeva boxplota i ako su rastojanja između krajeva i kvartila približno jednaka, boxplot nam ukazuje na simetričnu raspodelu, u šta smo i posumnjali uvidom u oblik histograma.
Histogram i boxplot se zajedno koriste u vizuelizaciji podataka, jer nadopunjuju nedostatke ovog drugog. Histogram se značajno može promeniti ako se u uzorak dodaju autlajeri, dok će boxplot neznatno promeniti oblik. Međutim, sa boxplota ne možemo oceniti gustinu raspodele iz koje dolaze podaci.