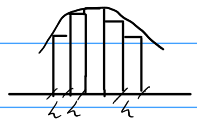
Хистограм - прва процена густине



идеја је да функцију густине апроксимирамо 'степенастим' функцијама

↳ 'део по део' константне функције

$$P\{x \le X \le x+h\} \approx f(x)\cdot h \Rightarrow f(x) \approx \frac{P\{x \le X \le x+h\}}{h}$$

(мало)

Ако смо извукли ПСУ $(x_1, x_2, \dots, x_n)$, $P\{x \le X \le x+h\}$ можемо оценити као $\frac{\text{број тачака које су упале у } [x, x+h]}{n}$.

Густину сваке тачке $t \in [x, x+h]$ тако оценимо са

$$\hat{f}(t) = \frac{\text{број тачака које су упале у } [x, x+h]}{n \cdot h}$$

Проблем: како изабрати h, интервале на којима оцењујемо f, колико тих интервала да изаберемо?

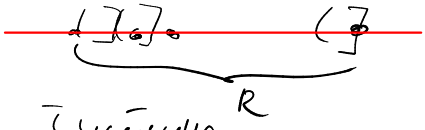
Постоје смернице за избор ових параметара, важно је напоменути да се облик хистограма може знатно променити ако само мало променимо нпр. h, ако не памтимо $[\,]$, већ $[\,)$ или $(\,]$ итд.

Дакле, хистограм треба схватити као првобитну процену густине расподеле из које су дошли наши подаци.

Постоји прецизнија оцена густине, а то је оцена језгром.

За радознале: Silverman: Density estimation for statistics and data analysis

$h \cdot k \approx R$



R

Поступак за цртање хистограма густине

1. изабрати број интервала k
   често се користи Sturges-ово правило: $k = \lceil \log_2 n \rceil + 1$
   $\lceil \cdot \rceil$ - заокруживање на већи цео број
2. одредити узорачки распон $R = x_{(n)} - x_{(1)}$
3. h тако да је тачно да $h \cdot k \approx R$
4. Одређујемо интервале над којима цртамо хистограм:
   (обично су то полуотворени интервали)
   $(x_0, x_0+h], (x_0+h, x_0+2h], \ldots, (x_0 + (k-1)h, x_0 + kh]$,
   x_0 је мало мање од $x_{(1)}$

- Sturges-ово правило је настало из претпоставке да подаци долазе из нормалне расподеле, а затим апроксимирамо нормалне расподеле биномном.
- Није теоријски поткрепљено, али у пракси даје прихватљиве резултате
- За радознале: Scott's rule, Freedman and Diaconis's rule

правила за одређивање h која имају теоријску основу

Нешто лакши за цртање је хистограм фреквенција:
у хистограму фреквенција над сваким интервалом
$(x_0 + ih, x_0 + (i+1)h]$
повучемо дуж на висини која је једнака броју тачака из узорка које су у њему биле.

Boxplot - визуелни приказ расподеле података по квартилима

p-ти квантил расподеле: $t \in \mathbb{R}$ т.д. $F(t) = P\{X \leq t\} = p$, $0 < p < 1$

Пример: медијана расподеле је 0,5-ти или 50% квантил
1. квартил := 25% квантил
3. квартил := 75% квантил

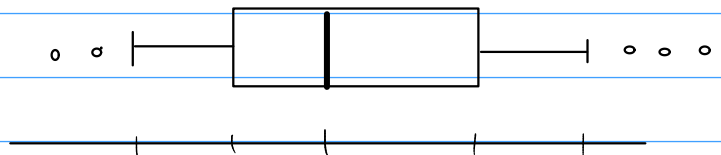
Статистике којима оцењујемо квантиле расподеле се зову **узорачки квантили**

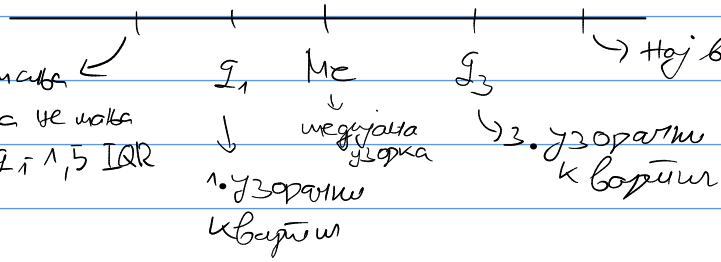
Постоји 9 дефиниција узорачких квантила, све су јако сличне

Интуитивно, 1. узорачки квартил је број од којег је 25% тачака из узорка мање, а 3. узорачки квартил је број од кога је 75% тачака из узорка мање, односно од којег је 25% тачака из узорка веће.

Како изгледа boxplot?

$IQR = Q_3 - Q_1$ ← интерквартилно растојање





Тачке које су $Q_1 - 1{,}5\,IQR$ и $Q_3 + 1{,}5\,IQR$ зовемо **аутлајерима**

Интуитивно, аутлајери су тачке које одступају од осталих података, а ово је само један начин на који се могу дефинисати.