Stochastiek 2

Inleiding in de Mathematische Statistiek

Herhaling H.1

Mathematische Statistiek

- We beschouwen de beschikbare data als realisatie(s) van een stochastische grootheid X. (Vaak een vector van de vorm $X = (X_1, \ldots, X_n)$.) Oftewel: we vatten de data op als *uitkomsten van een kansexperiment*.

Mathematische Statistiek

- We beschouwen de beschikbare data als realisatie(s) van een stochastische grootheid X. (Vaak een vector van de vorm $X = (X_1, \ldots, X_n)$.) Oftewel: we vatten de data op als *uitkomsten van een kansexperiment*.
- De variabiliteit in de data wordt beschreven met behulp van kansverdelingen.
- Statistisch model: collectie van kansverdelingen die de mogelijke variabiliteit in de data beschrijven.
- Doel van een statistische procedure: het identificeren van de verdeling die de data het "beste" beschrijft.

Algemene setup

- Data: stochastische grootheid X. Vaak $X = (X_1, \ldots, X_n)$.
- Model: collectie kansverdelingen $\{P_\theta : \theta \in \Theta\}$ die de mogelijke verdelingen van de data beschrijven. Terminologie: θ: parameter, Θ: parameter ruimte.

Algemene setup

- Data: stochastische grootheid X. Vaak $X = (X_1, \ldots, X_n)$.
- Model: collectie kansverdelingen $\{P_\theta : \theta \in \Theta\}$ die de mogelijke verdelingen van de data beschrijven. Terminologie: θ: parameter, Θ: parameter ruimte.

- Doel: schatten van θ, toetsen van hypotheses over θ, construëren van betrouwbaarheidsgebieden voor θ.

H.2 Verdelingsonderzoek

Achtergrond

Bij het ontwikkelen van schatters, toetsen, etc. zullen we meestal het statistisch model als gegeven beschouwen. We maken dan aannames over de verdeling van de data.

Om het gebruik van bepaalde statistische methoden in de praktijk te rechtvaardigen is het daarom belangrijk dat we die aannames verifiëren.

In data analyse toepassingen vaak een eerste stap, voordat meer "precieze" methoden worden gebruikt.

Grafische Methoden

- Histogram: visualiseert de verdeling van een dataset als staafdiagram.
- Boxplot: visualiseert verdeling compacter, handig om snel verschil in verdeling tussen verschillende datasets te detecteren.
- QQ-plot: visuele check of dataset uit een concrete *locatie-schaalfamilie* komt.
- Scatter plot: geeft indruk van samenhang tussen variabelen.
- Auto-correlatieplot: geeft indruk van de correlatie tussen variabelen.

Histogram

Idee: Heb $x_1, \ldots, x_n \in \mathbb{R}$. Vat op als steekproef uit een verdeling op $\mathbb{R}$. Visualiseer die verdeling.

Constructie:

- Kies $a_0 < a_1 < \cdots < a_m$ en beschouw de corresponderende partitie van $\mathbb{R}$. (Intervallen $(a_{j-1}, a_j]$ heten de bins van het histogram.)
- Definieer $h : \mathbb{R} \to [0, \infty)$, constant op iedere bin $(a_{j-1}, a_j]$:

$$h(x) = \begin{cases} \frac{1}{n(a_j - a_{j-1})} \sum_{i=1}^{n} 1_{(a_{j-1}, a_j]}(x_i) & \text{als } x \in (a_j, a_{j-1}] \\ 0 & \text{anders.} \end{cases}$$

Als bins even lang zijn: waarde van h op de bin $(a_{j-1}, a_j]$ is proportioneel met het aantal waarnemingen in de bin.

Histogram - 2

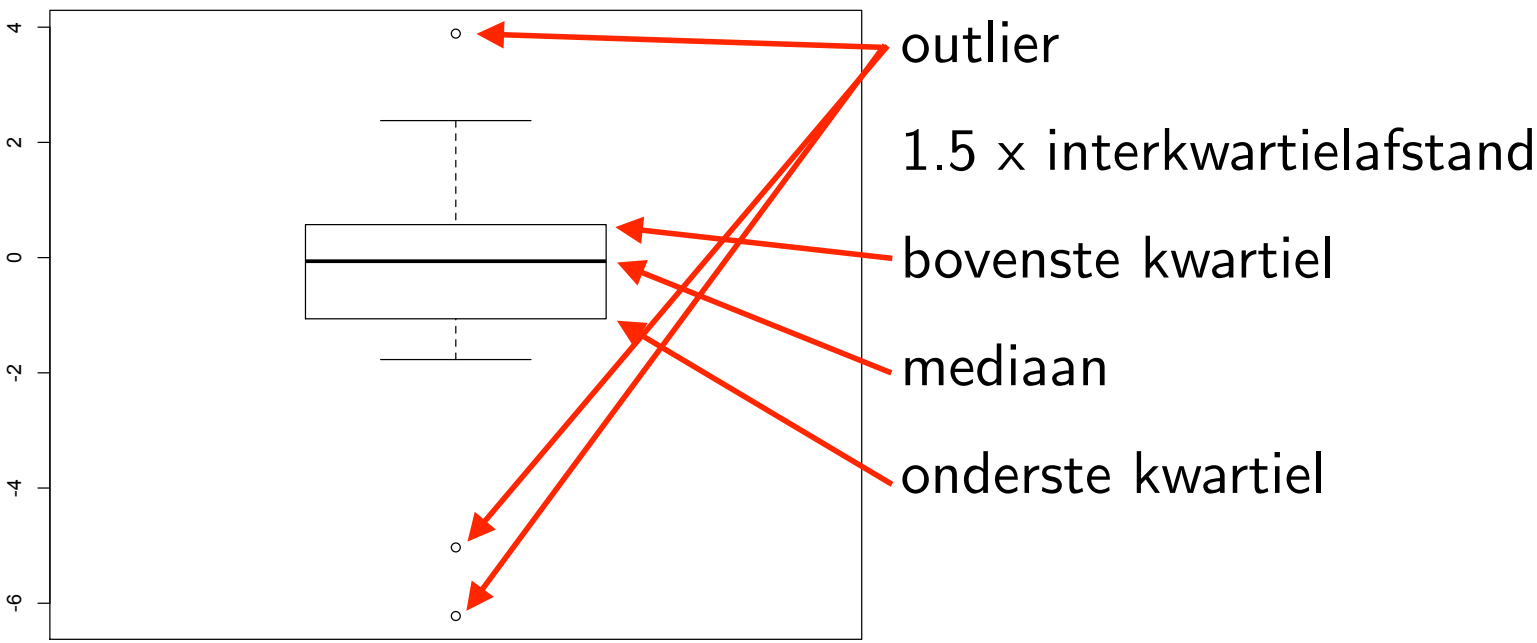
Stel $X_1, \ldots, X_n$, onafh, dichtheid f. Zij h het corresponderende histogram voor partitie $a_0 < a_1 < \cdots < a_m$. Dan voor $x \in (a_j, a_{j-1}]$

$$\mathbb{E}h(x) = \mathbb{E}\frac{1}{n(a_j - a_{j-1})}\sum_{i=1}^{n} 1_{(a_{j-1},a_j]}(X_i) = \frac{1}{a_j - a_{j-1}}\int_{a_{j-1}}^{a_j} f(y)\,dy.$$

Dus als $X_1, \ldots, X_n$ een steekproef is uit f, n is niet te klein, en de partitie is fijn genoeg, dan geeft h een indruk van de vorm van f.

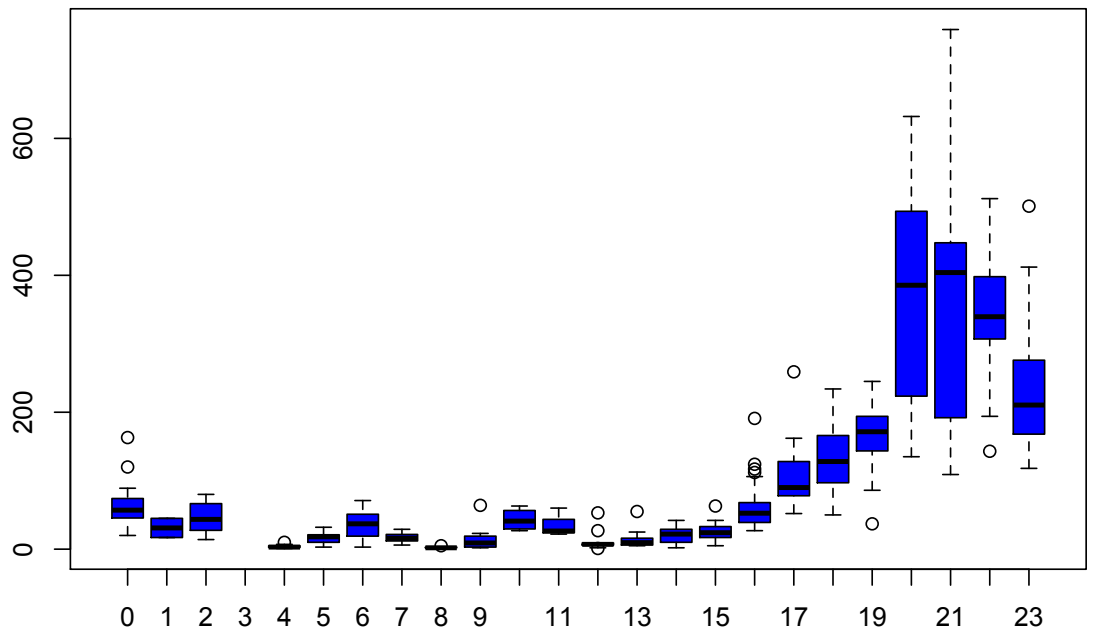
Boxplot

Idee: Grove visualisatie van verdeling van data $x_1, \ldots, x_n \in \mathbb{R}$.



Boxplot - 2

Vooral handig om verschillen in verdeling snel te zien.



QQ-plot

Idee: visueel nagaan of dataset $x_1, \ldots, x_n \in \mathbb{R}$ uit een gegeven *locatie-schaalfamilie* komt.

Definitie:

- Voor een verdelingsfunctie F, $a \in \mathbb{R}$ en $b > 0$,

$$F_{a,b}(x) = F((x - a)/b).$$

- $\{F_{a,b} : a \in \mathbb{R}, b > 0\}$ heet de locatie-schaalfamilie bij F. (achtergrond: als $X \sim F$, dan $a + bX \sim F_{a,b}$)

Voorbeeld: de locatie-schaalfamilie van de $N(0, 1)$ verdeling is de collectie $\{N(a, b^2) : a \in \mathbb{R}, b > 0\}$ van alle (niet-gedegenereerde) normale verdelingen.

QQ-plot - 2

QQ-plots zijn gebaseerd op kwantielen (Engels: Quantiles).

Definitie: Zij F een verdelingsfunctie en $\alpha \in (0,1)$. Het α-kwantiel van F is het punt

$$F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}.$$

Feit 1: voor de locatie-schaal familie van F geldt, voor alle $\alpha \in (0,1)$,

$$F^{-1}_{a,b}(\alpha) = a + bF^{-1}(\alpha).$$

Bewijs: opgave 2.2!

QQ-plot - 3

De punten $\{(F^{-1}(\alpha), F_{a,b}^{-1}(\alpha)) : \alpha \in (0,1)\} \subset \mathbb{R}^2$ liggen dus op een rechte lijn!

Definitie: Zij $x_1, \ldots, x_n \in \mathbb{R}$ een rij verschillende getallen. De corresponderende geordende rij $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ (van dezelfde getallen) is de rij van ordestatistieken.

Feit 2: als $X_1, \ldots, X_n$ o.o., $\sim F$, dan

$$\mathbb{E}F(X_{(i)}) = \frac{i}{n+1}, \quad i = 1, \ldots, n.$$

Bewijs: opgave 2.8!

QQ-plot - 4

Zij nu $X_1, \ldots, X_n$ o.o., $\sim F_{a,b}$. Dan:

Feit 2:	$i/(n+1) \approx F_{a,b}(X_{(i)})$,
en dus:	$F^{-1}_{a,b}(i/(n+1)) \approx X_{(i)}$,
en dus (Feit 1):	$a + bF^{-1}(i/(n+1)) \approx X_{(i)}$.

Conclusie: de punten $\{(F^{-1}(i/(n+1)), X_{(i)}), i = 1, \ldots, n\}$ liggen ongeveer op de lijn $a + bx = y$.

Definitie: De QQ-plot van de punten $x_1, \ldots, x_n$ t.o.v. de verdelingsfunctie F is een plot van de punten $\{(F^{-1}(i/(n+1)), x_{(i)}), i = 1, \ldots, n\}$.

Scatter plot

Idee: observeer paren $(x_1, y_2), \ldots, (x_n, y_n)$. Zet x-jes en y-tjes tegen elkaar uit om te zien of er een mooi verband is.

Wordt vaak gebruikt om te zien of er (bij benadering) een lineair verband is.

Auto-correlatie plot

Idee: visualisatie van correlatie tussen observaties $x_1, \ldots, x_n$.

Herinner:

- correlatie tussen stochasten X en Y (in L^2) is

$$\rho = \frac{\mathbb{C}\mathrm{ov}(X, Y)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}} = \frac{\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)}{\sqrt{\mathbb{E}(X - \mathbb{E}X)^2\mathbb{E}(Y - \mathbb{E}Y)^2}}.$$

- Als X en Y onafhankelijk zijn, dan is $\mathbb{C}\mathrm{ov}(X, Y) = 0$. Maar niet andersom!

Auto-correlatie plot - 2

Als $X_1, \ldots, X_n$ onafhankelijk zijn, is dus

$$\mathbb{C}\mathrm{ov}(X_i, X_{i+h}) = 0, \quad \forall h \geq 1, \forall i \text{ zdd } i, i+h \in \{1, \ldots n\}.$$

Voor elke $h \geq 0$ observeren we een aantal paren (X_i, X_{i+h}) en kunnen we de correlatie benaderen door het corresponderende steekproefgemiddelde (wet van grote aantallen)

$$r_X(h) = \frac{\sum_{i=1}^{n-h}(X_i - \bar{X})(X_{i+h} - \bar{X})}{(n-h)S_X^2}, \quad h = 0, \ldots, n-1,$$

waarbij

$$S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

Dit is de steekproef auto-correlatie van orde h.

Auto-correlatie plot - 3

Definitie: de auto-correlatieplot van $x_1, \ldots, x_n$ is een plot van de punten $\{(h, r_x(h)), h = 0, 1, \ldots, n\}$ (of een deel daarvan).

De auto-correlatieplot wordt wel gebruikt om visueel te onderzoeken of voldaan is aan de aanname van onafhankelijkheid van $X_1, \ldots, X_n$. Omdat correlatie 0 geen onafhankelijkheid impliceert, is het raadzaam om in zulke gevallen niet alleen de steekproef auto-correlaties van X_i's de plotten, maar ook die van de X_i^2's, X_i^3's etc.