SPC en staartkansen: kritische kanttekeningen en alternatieve technieken

Alessandro Di Bucchianico
John H.J. Einmahl
Technische Universiteit Eindhoven
Faculteit Wiskunde en Informatica

- theoretisch interessant, maar praktisch volslagen onbelangrijk

Eigenschap 1

- praktisch belangrijk, maar theoretisch triviaal

Eigenschap 2

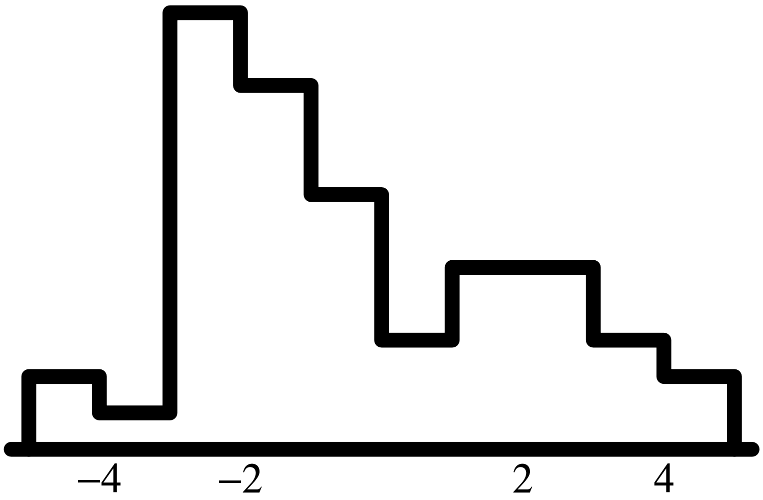
Deze lezing heeft beide eigenschappen **niet**.

Inhoud

- Dichtheidsschatters
- Toetsen op normaliteit
- Extreme quantielen schatten
- Capability indices

Histogrammen

Een van de veelgebruikte hulpmiddelen binnen SPC (”de Magnificent Seven”) is het histogram. Het histogram wordt gepropageerd om een indruk te krijgen van de vorm van de verdeling (i.h.b. voor normaliteitscontrole).



Histogrammen nader bekeken

Gegeven: waarnemingen $X_1, \ldots, X_n$.

Een histogram met klassen I_i, $i = 1, \ldots, B$, is eigenlijk een schatter voor de kansdichtheid f (op een multiplicatieve factor na):

$$\widehat{f}(x) = \frac{\text{aantal waarnemingen in klasse van}\, x}{n\,|I_{i(x)}|}$$

Dichtheidsschatters

Algemener kan men dichtheden schatten met behulp van *kernschatters*:

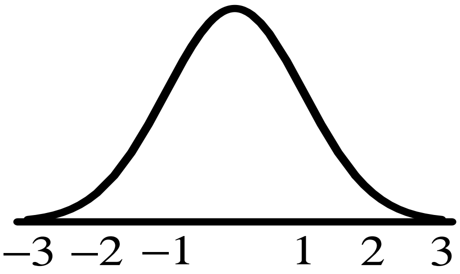
$$\widehat{f_h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

waarbij de *kern* K voldoet aan

$$K \geq 0 \text{ en } \int_{-\infty}^{\infty} K(x)\, dx = 1.$$

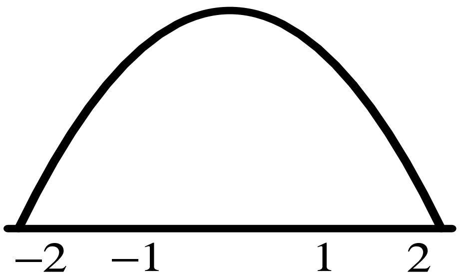
Veelgebruikte kernen zijn:

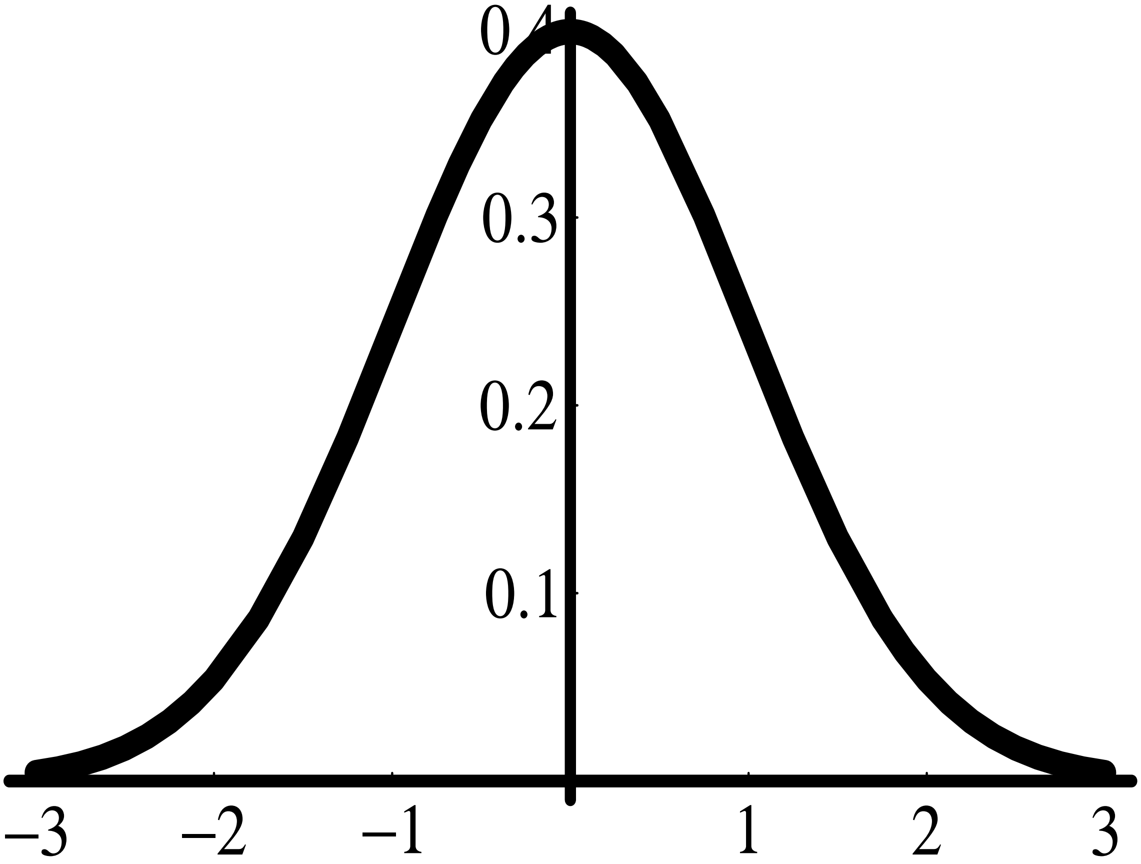
- de standaardnormale kern: $\dfrac{1}{\sqrt{2\pi}}\,e^{-x^2/2}$



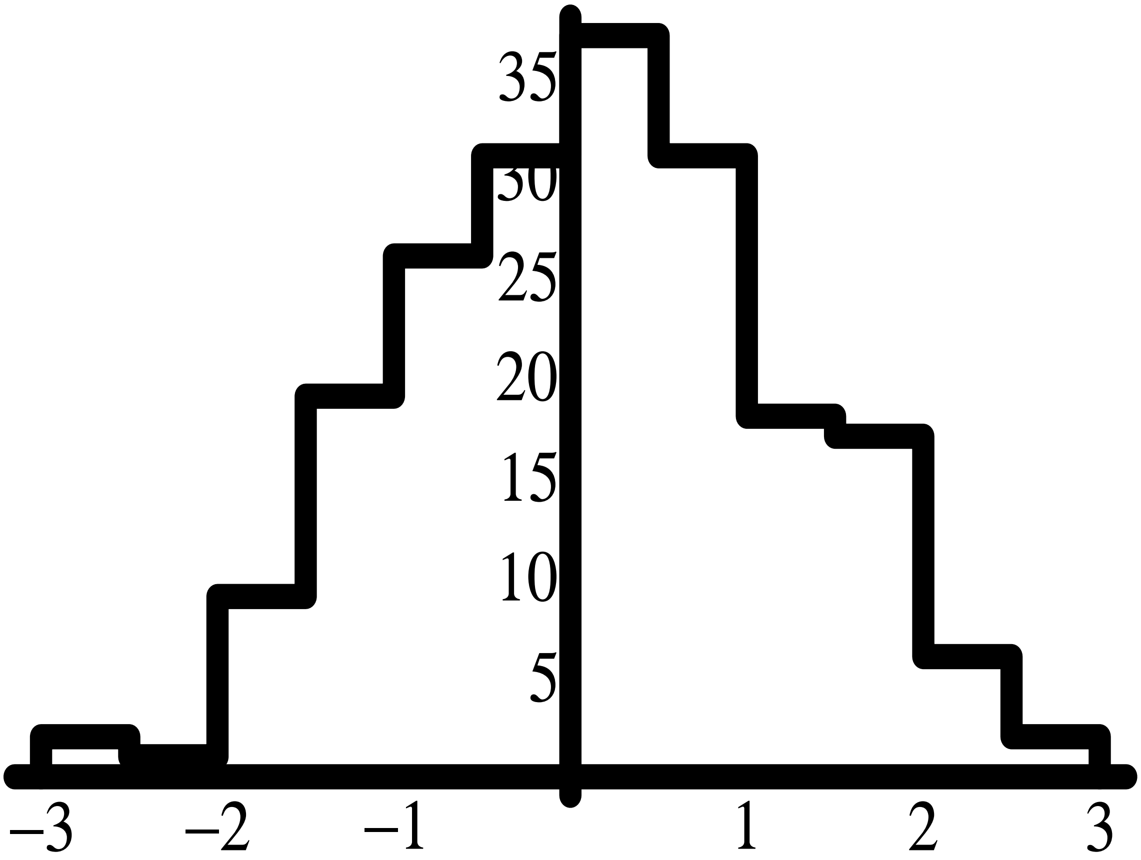
- de Epanechnikov kern: $\dfrac{3}{4}\left(1-\dfrac{1}{5}x^2\right)/\sqrt{5}$

  $(-\sqrt{5} \leq x \leq \sqrt{5})$

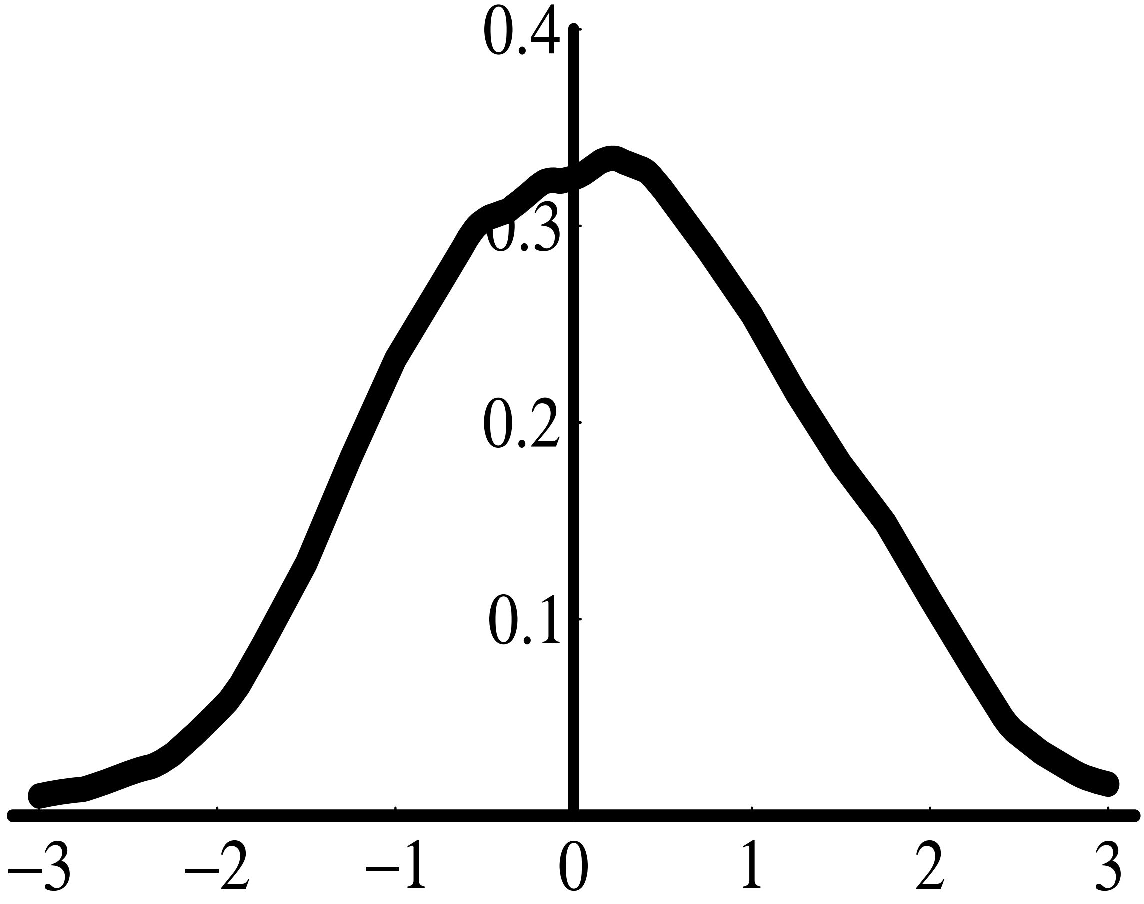




$N(0,1)$



$N(0,1), n = 200$, binlengte $= 0.5$



$N(0,1), n = 200, h = 0.34$, kern = Epanechnikov

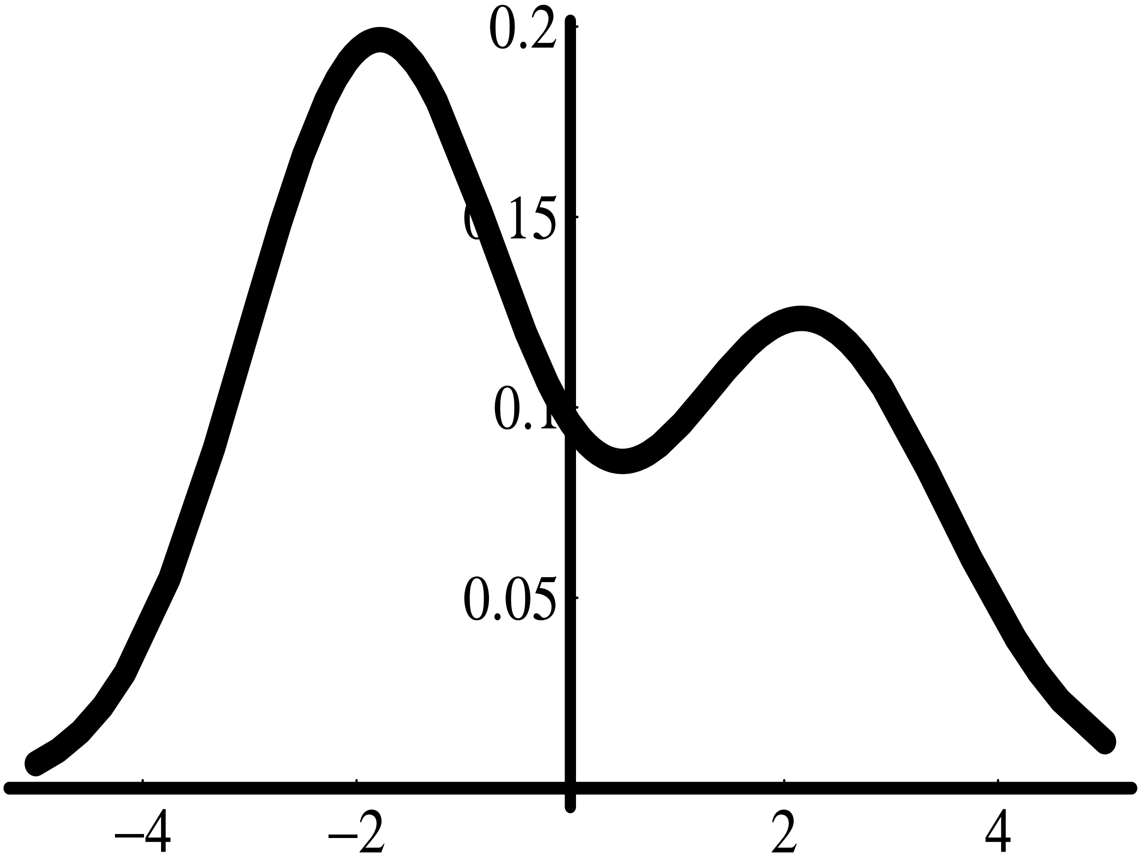
De keuze van de kern is van ondergeschikt belang. Van groot belang is het kiezen van de bandbreedte h.

Een te grote h leidt tot overdreven gladheid; een te kleine h leidt tot te veel fluctuaties.

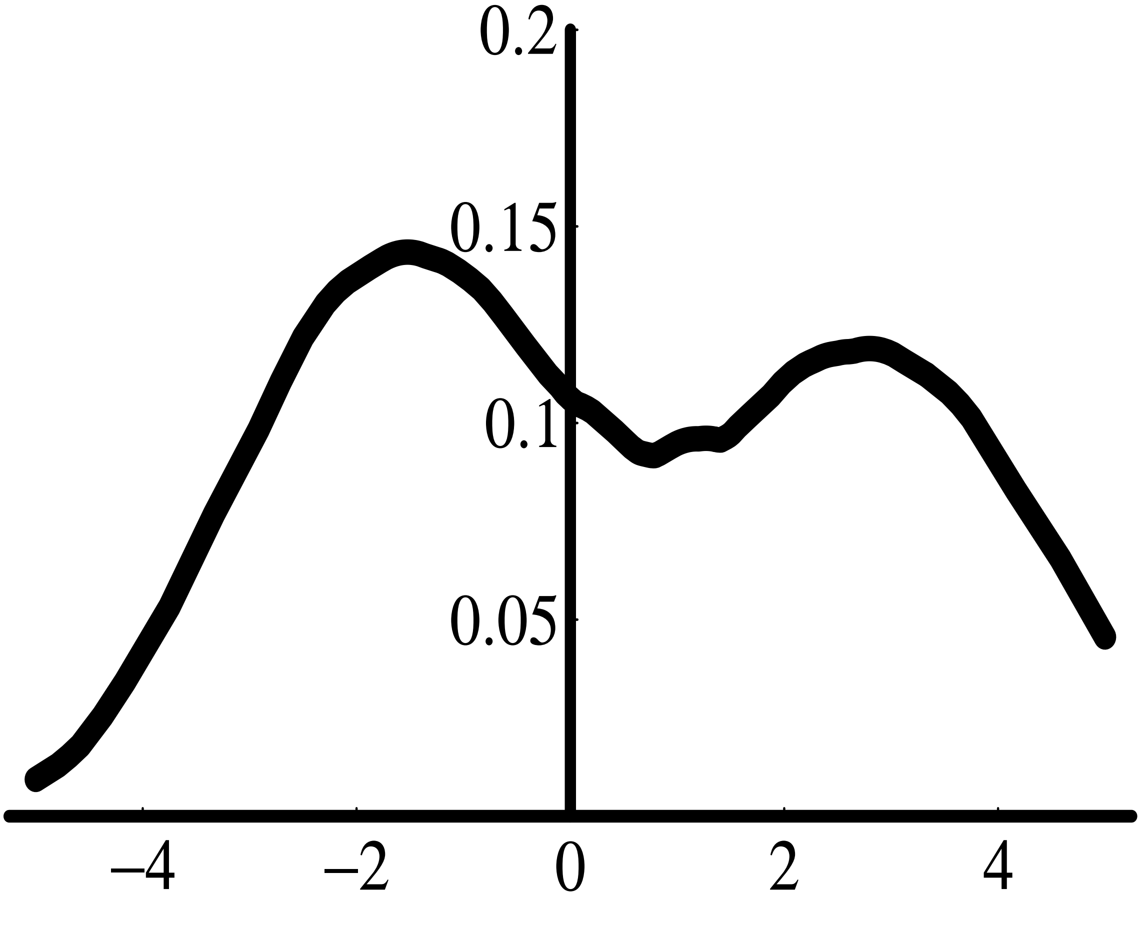
Vuistregel voor zowel standaardnormale als Epanechnikovkern (vgl. Silverman (1986)):

$$h = 0.9 A n^{-1/5},$$

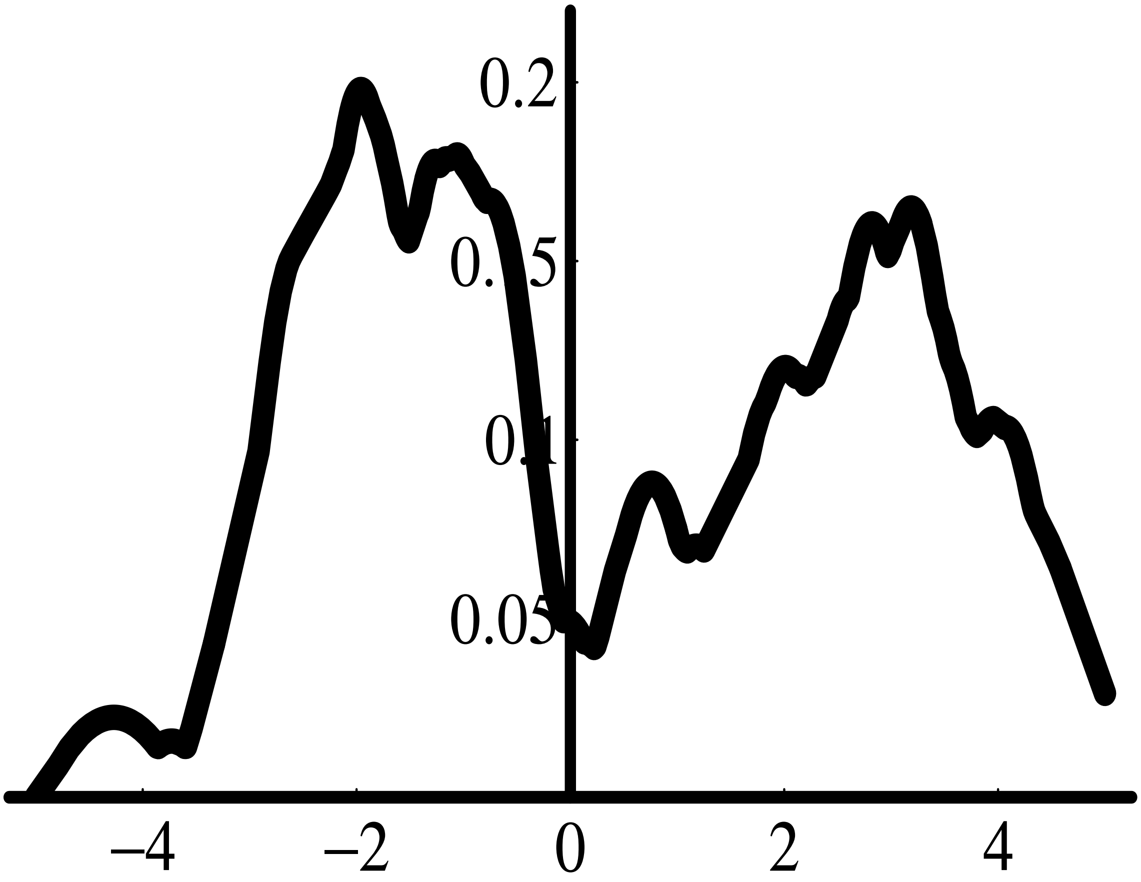
waarbij $A = \min\{s, IQR/1.34\}$.


0.2
0.15
0.1
0.05
–4
–2
2
4

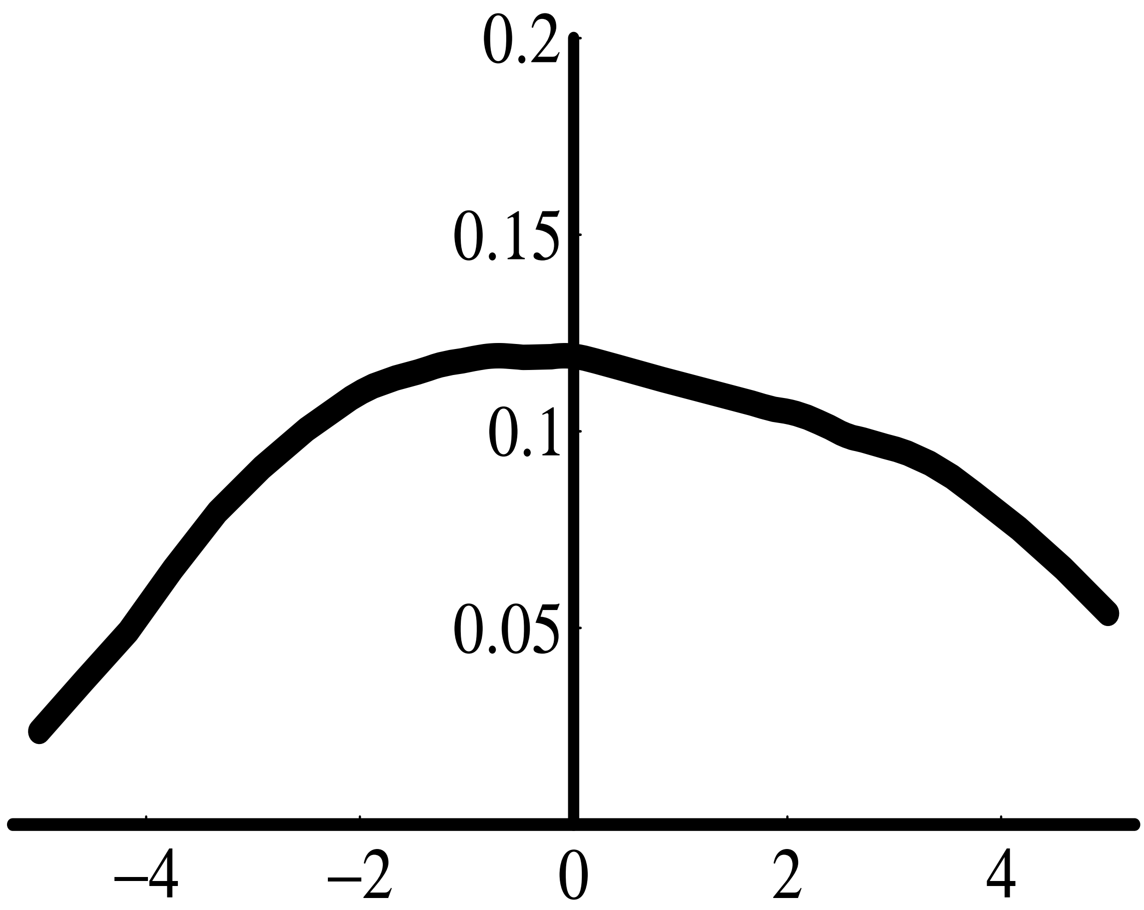
Epanechnikov, h = 1.0
(vuistregel), n=50



Epanechnikov, h = 0.3, n=50



Epanechnikov, h = 1.5, n=50



Toetsen op normaliteit

$F_{\mu,\sigma^2}(x)$ is verdelingsfunctie van $N(\mu,\sigma^2)$.

$\Phi(x)$ is verdelingsfunctie van $N(0,1)$.

$F_{\mu,\sigma^2}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{-\mu}{\sigma} + \frac{1}{\sigma}\,x\right)$.

$$\widehat{F}_n(x) = \frac{\#\{i \mid X_i \le x\}}{n}$$

(empirische verdelingsfunctie).

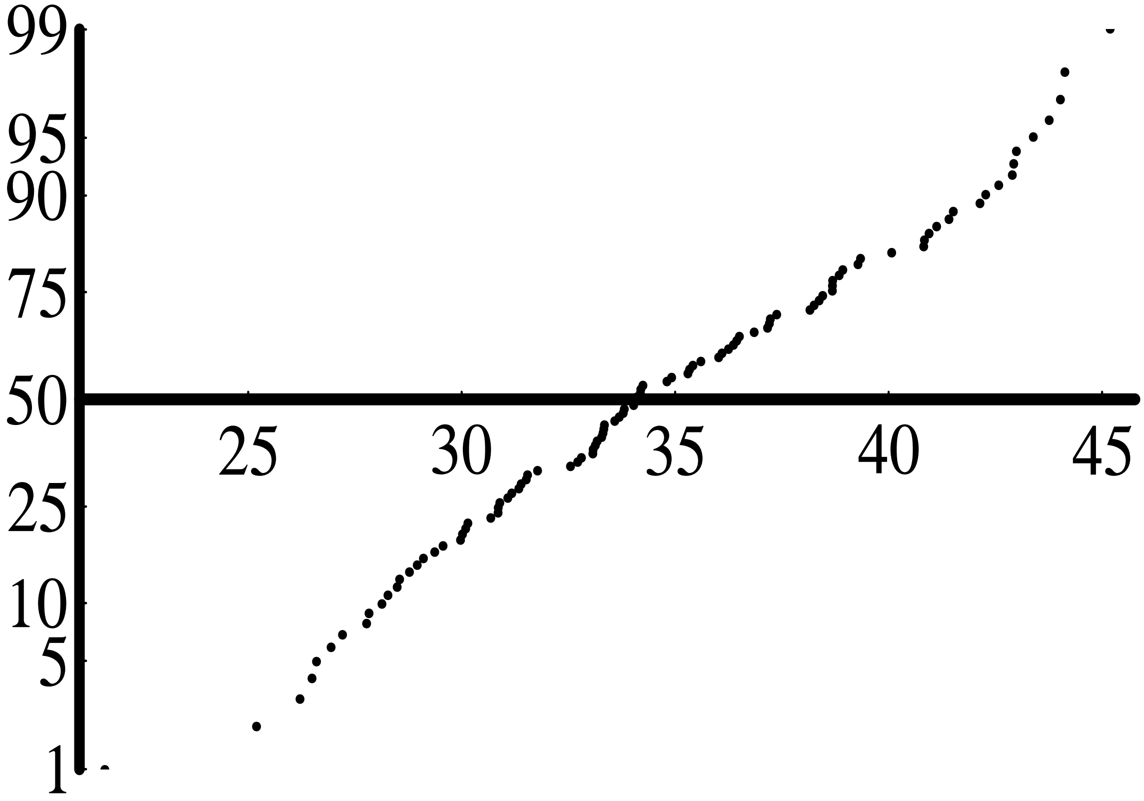
$\Phi^{-1}\left(\widehat{F}_n(x)\right) \approx \Phi^{-1}\left(F_{\mu,\sigma^2}(x)\right) = \frac{-\mu}{\sigma} + \frac{1}{\sigma}\,x$.

Grafische controle op normaliteit:

plot $\left(X_{(i)}, \Phi^{-1}\left(\widehat{F}_n(X_{(i)})\right)\right) = \left(X_{(i)}, \Phi^{-1}\left(\frac{i}{n}\right)\right)$.

Dit heet een *normal probability plot* (zie bijv. Castillo 1988).

Voorbeeld $(N(35, 5^2), n = 100)$



Rechte lijn?

De normal probability plot is (afgezien van een schaling van de y-as) gebaseerd op de lijn $y = \frac{-\mu}{\sigma} + \frac{1}{\sigma}x$.

De toetsen van Shapiro-Wilk en Shapiro-Francia zijn goede toetsen op normaliteit gebaseerd op regressietechnieken door deze punten (na verwisselen van de $x-$ en y-as). Zie D'Agostino en Stephens (1986) voor verdere details.

Niet-specifieke toetsen zoals de toets van Kolmogorov-Smirnov zijn *minder geschikt* voor toetsen op normaliteit.

Schatten van hoge (of lage) quantielen

$$P(X > x_p) = p$$

Schat x_p.

Normale verdeling: $X \sim N(\mu, \sigma^2)$

$$P(X > x_p) = P(Z > \frac{x_p - \mu}{\sigma}) = p$$

$$1 - \Phi\left(\frac{x_p - \mu}{\sigma}\right) = p$$

$$x_p = \mu + \sigma\,\Phi^{-1}(1-p)$$

$$\begin{array}{rl} p = 0.00135 & \Phi^{-1}(1-p) = 3 \\ x_p = \mu + 3\sigma & \widehat{x}_p = \overline{x} + 3s \\ x_{1-p} = \mu - 3\sigma & \widehat{x}_{1-p} = \overline{x} - 3s \end{array}$$

$$\widehat{x}_p - \widehat{x}_{1-p} = \overline{x} + 3s - (\overline{x} - 3s) = 6s$$

Schatten (vervolg)

Als de verdeling niet normaal is (in de staart), dan kan $\widehat{x}_p$ een (heel) slechte schatter van x_p zijn.

Het is dan beter, als de steekproefomvang voldoende groot is, een ruimer toepasbare schatter te gebruiken uit de moderne extreme-waardentheorie . Deze schatter 'werkt' onder zwakke voorwaarden die niet parametrisch van aard zijn.

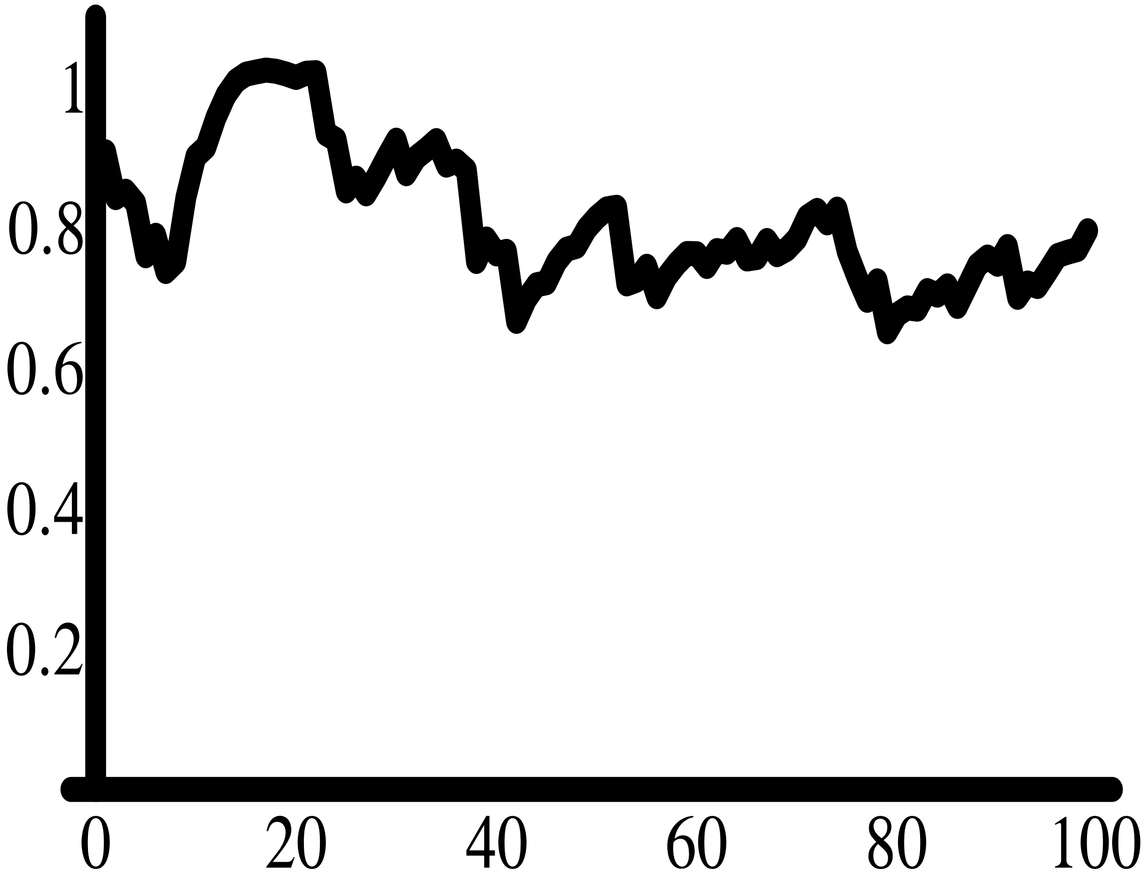
Voorbeeld

Symmetrische (om 0) verdeling met dikkere staarten dan normale verdeling.

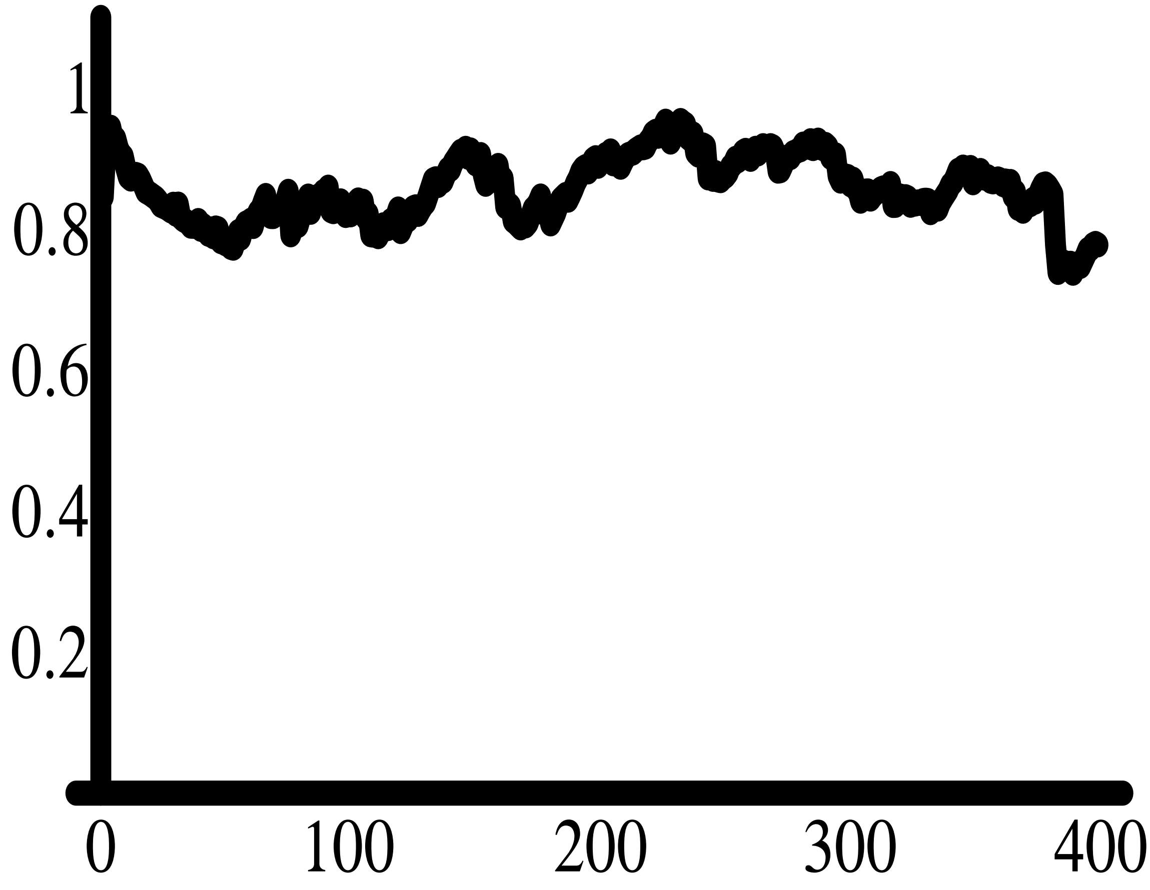
99.865 percentiel $= 0.807$, $\mu = 0$, $\sigma = \frac{1}{6}$

De quantielschatter gebaseerd op de normale verdeling schat dus niet 0.807, maar $\mu + 3\sigma = \frac{1}{2}$. De momentenquantielschatter (Dekkers et al. (1989)) schat wel de echte waarde 0.807 en is dus (zie hierna) te prefereren. Bij deze schatter is het, net als bij dichtheidsschatten, belangrijk hoeveel 'upper order statistics' meegenomen worden.

Dit heeft belangrijke consequenties voor het schatten van C_p's en C_{pk}'s (zie hierna en ook Kotz en Johnson (1993)). De schatter gebaseerd op de normale verdeling overschat drastisch.



Momentenquantielschatter, $n = 500$



Momentenquantielschatter, $n = 2000$

Capability indices

C_p en C_{pk} kunnen drastisch veranderen bij monotone transformaties van de data.

Denk bijv. aan pH, of aan doorloopsnelheid vs. doorlooptijd.

Maar: $P(X > \text{USL}) = P(T(X) > T(\text{USL}))$ indien T strikt stijgend is.

Het is *in principe* veel directer en beter om $P(X > \text{USL}) + P(X < \text{LSL})$ te schatten.

C_{pk} voor en na transformatie

$X \sim N(0,1), \text{USL} = 6,\ \text{LSL} = -6, C_{pk} = C_p = 2$

$Y = e^X$ is lognormaal verdeeld

Definitie: C_{pk} in het niet-normale geval:

$$C_{pu} = \frac{\text{USL} - \text{centrum}}{Q_u - \text{centrum}}$$

$Q_u = 99.865$ percentiel, centrum is verwachting (gemiddelde) of mediaan

$$C_{pl} = \frac{\text{centrum} - \text{LSL}}{\text{centrum} - Q_l}$$

$Q_l = 0.135$ percentiel

$$C_{pk} = \min\{C_{pl}, C_{pu}\}$$

Na transformatie

Met *verwachting*: $\mu_Y = e^{\frac{1}{2}}$

$$C_{pu} = \frac{e^6 - e^{\frac{1}{2}}}{e^3 - e^{\frac{1}{2}}} = 21.8$$

$$C_{pl} = \frac{e^{\frac{1}{2}} - e^{-6}}{e^{\frac{1}{2}} - e^{-3}} = 1.03$$

Met *mediaan*: $\text{med}(Y) = e^0 = 1$

$$C_{pu} = \frac{e^6 - 1}{e^3 - 1} = 21.1$$

$$C_{pl} = \frac{1 - e^{-6}}{1 - e^{-3}} = 1.05$$

In beide gevallen: $C_{pk} \approx 1$.

Literatuur

- E. Castillo, Extreme Value Theory in Engineering, Academic Press, 1988.

- R.B. D'Agostino en M.A. Stephens, Goodness-of-fit Techniques, Marcel Dekker, 1986.

- A.L.M. Dekkers, J.H.J. Einmahl en L. de Haan, A moment estimator for the index of an extreme-value distribution, Annals of Statistics 17 (1989), 1833-1855.

- L. Devroye en L. Györfi, Nonparametric density estimation. The L_1 view, John Wiley, 1985.

- S. Kotz en N.L. Johnson, Process Capability Indices, Chapman and Hall, 1993.

- B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, 1986.