TESTS OF FIT

—

Ariane FERREIRA / Philippe CASTAGLIOLA

Ecole des Mines de Nantes & IRCCyN, Nantes, France

aprosa@emn.fr

Definition

- $X_1, \ldots, X_n$ is a sample of n independent r.v. Let $F_X(x)$ be the unknown cdf of the X_j's.
- Let $F(x)$ be a cdf.
- A *test of fit* allows to decide, from a set n observations $x_1, \ldots, x_n$, if $F(x)$ is the cdf of the X_j's, i.e.

$$H_0 : F_X(x) = F(x)$$
$$H_1 : F_X(x) \neq F(x)$$

Definition

- In practice, it is never possible to guess what the real cdf of the X_j's is.
- Decide $H_1 : F_X(x) \neq F(x)$ allows to reject the fit between the data and the cdf $F(x)$.
- Decide $H_0 : F_X(x) = F(x)$ allows to conclude only that $F(x)$ seems to be a good cdf for the X_j's.
- In that case, there is perhaps another cdf (what else ?) which fits better with the data.
- Famous quote "All models are wrong, but some are useful".

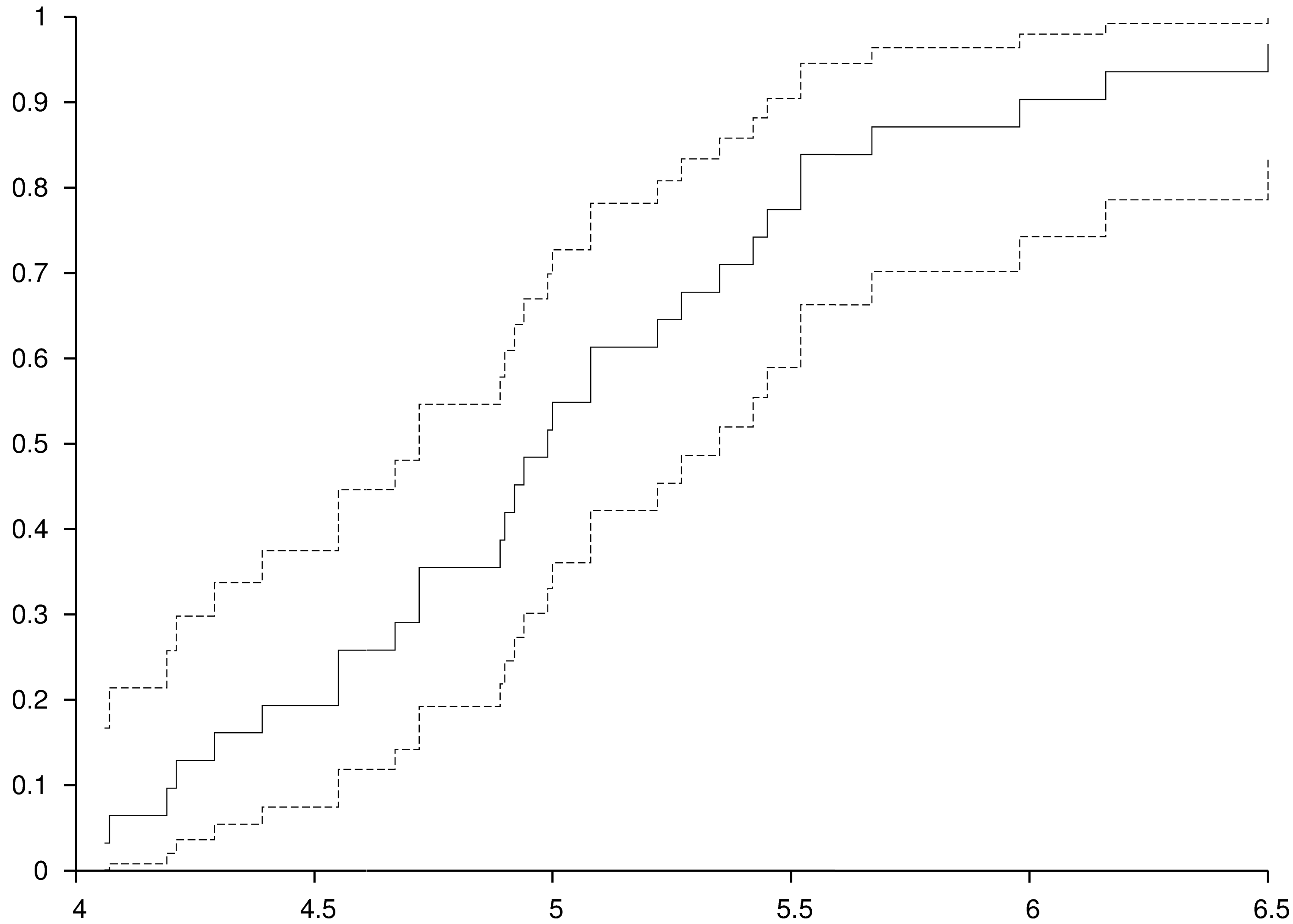
Non parametrical estimation of a cdf

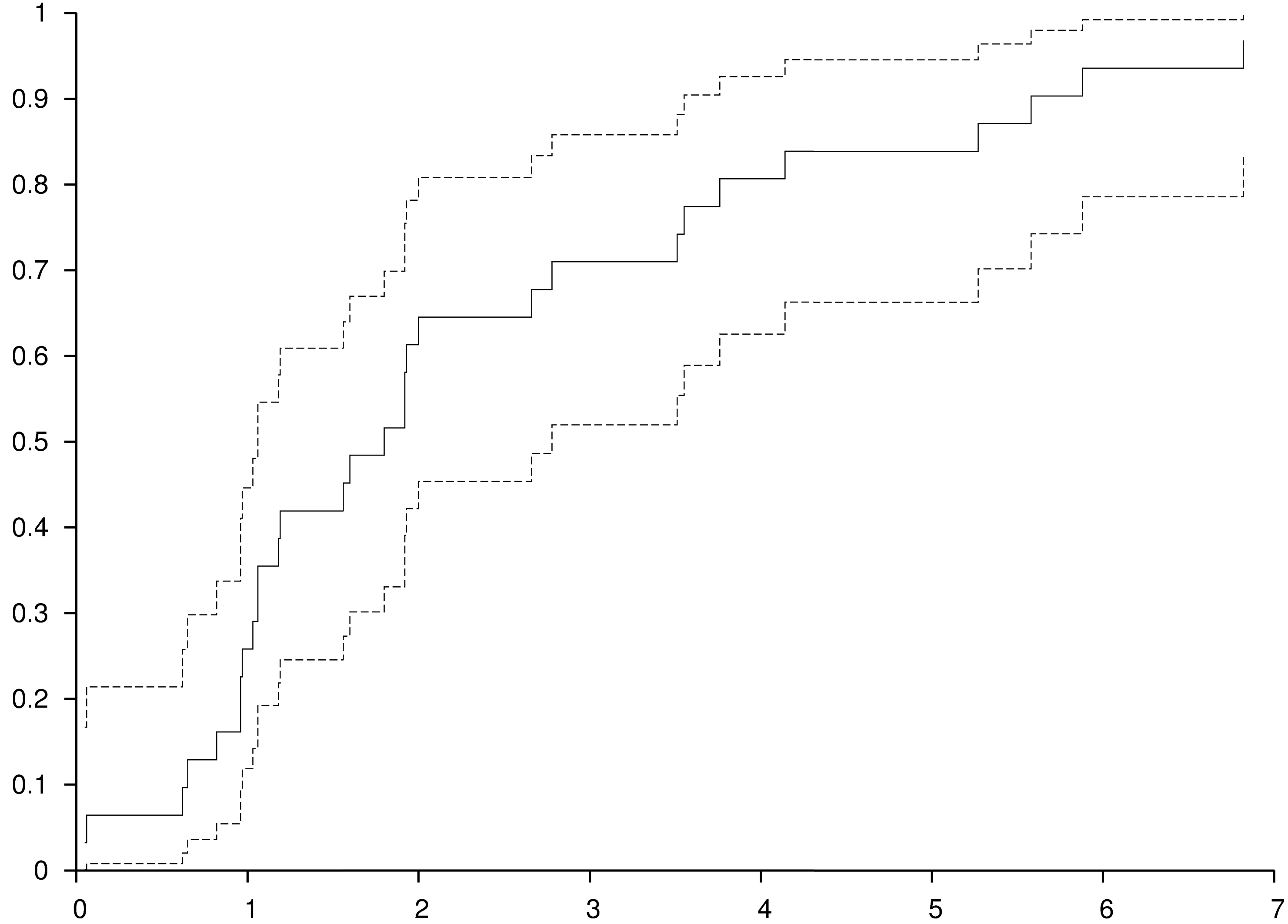
- Let $x_1, \ldots, x_n$ be a set of n observations and let $x_{(1)}, \ldots, x_{(n)}$ be the corresponding ordered set.
- Non parametrical estimator for $F(x)$

$$\hat{F}(x_{(k)}) = \frac{k}{n+1}$$

- For the sample median, we have $\hat{F}(\tilde{x}) = 1/2$.
- Confidence interval

$$\begin{aligned}
(\hat{F}(x_{(k)}))_L &= \left(1 + \frac{n-k+2}{kF_F^{-1}\{\alpha/2, 2k, 2(n-k+2)\}}\right)^{-1} \\
(\hat{F}(x_{(k)}))_U &= \left(1 + \frac{n-k+1}{(k+1)F_F^{-1}\{1-\alpha/2, 2(k+1), 2(n-k+1)\}}\right)^{-1}
\end{aligned}$$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
4
4.5
5
5.5
6
6.5


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7

The Q-Plot

- This method can be applied when the cdf $F(x)$ verifies

$$g\{F(x)\} = a + bh(x)$$

- $g()$ and $h()$ are two functions.
- a and b are two parameters.
- In order to test the fit of the observations $x_{(1)}, \ldots, x_{(n)}$ with the cdf $F(x)$, we just need to plot the points

$$\left\{h(x_{(k)}), g\left(\frac{k}{n+1}\right)\right\}$$

and to check that those points are approximatively on a straight line.

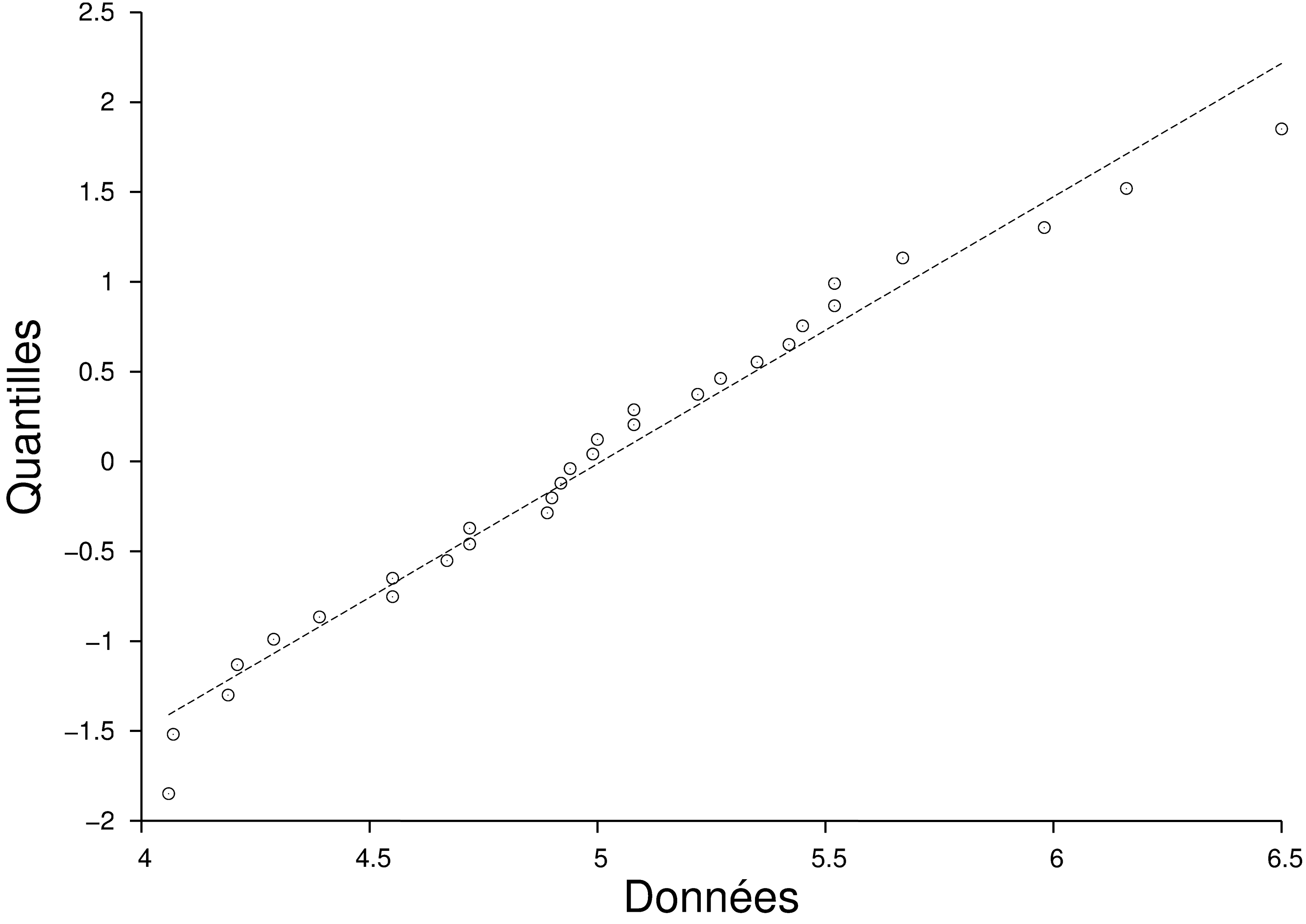
The Q-Plot

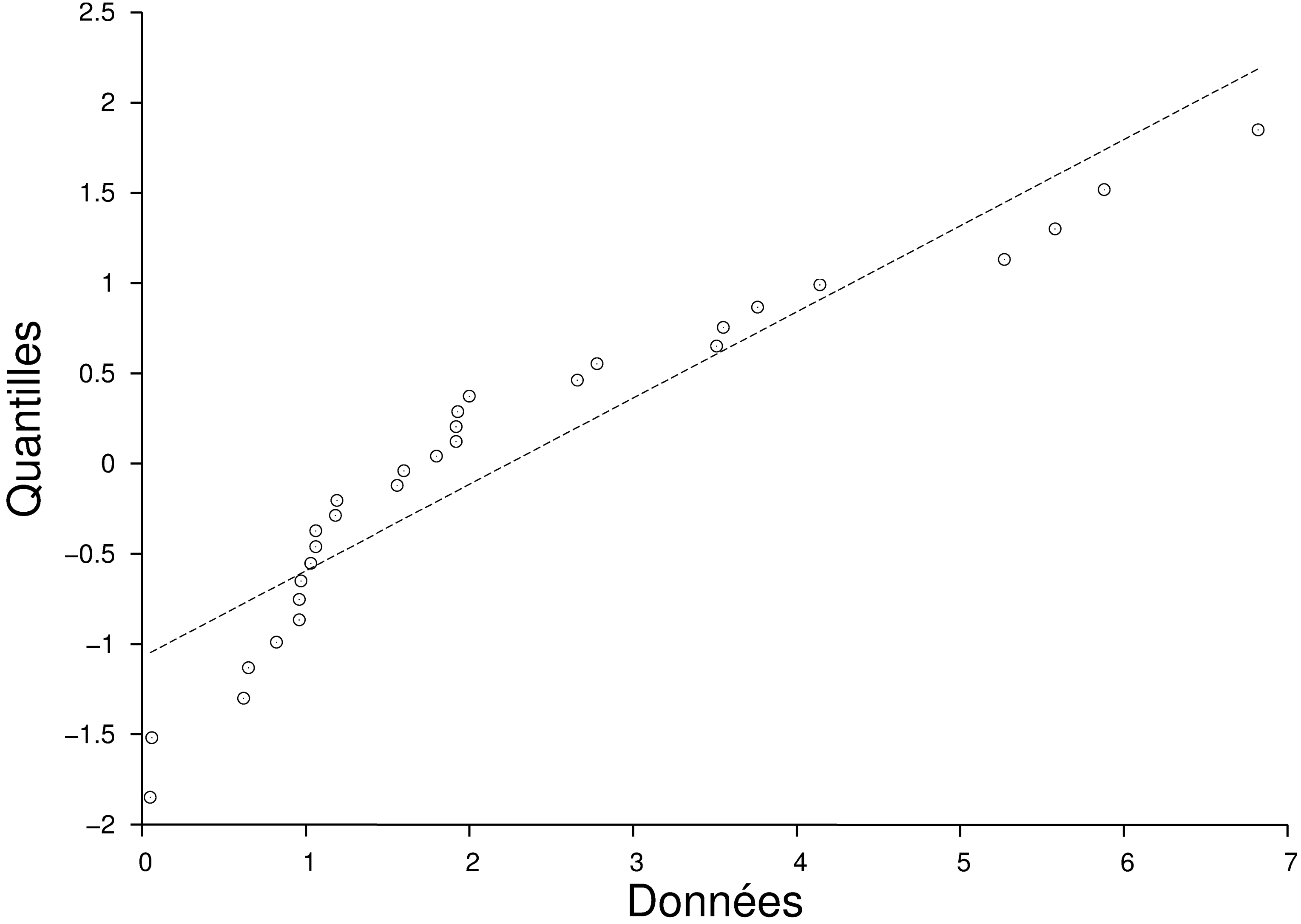
Normal $$\left\{x_{(k)}, \Phi^{-1}\left(\frac{k}{n+1}\right)\right\}$$

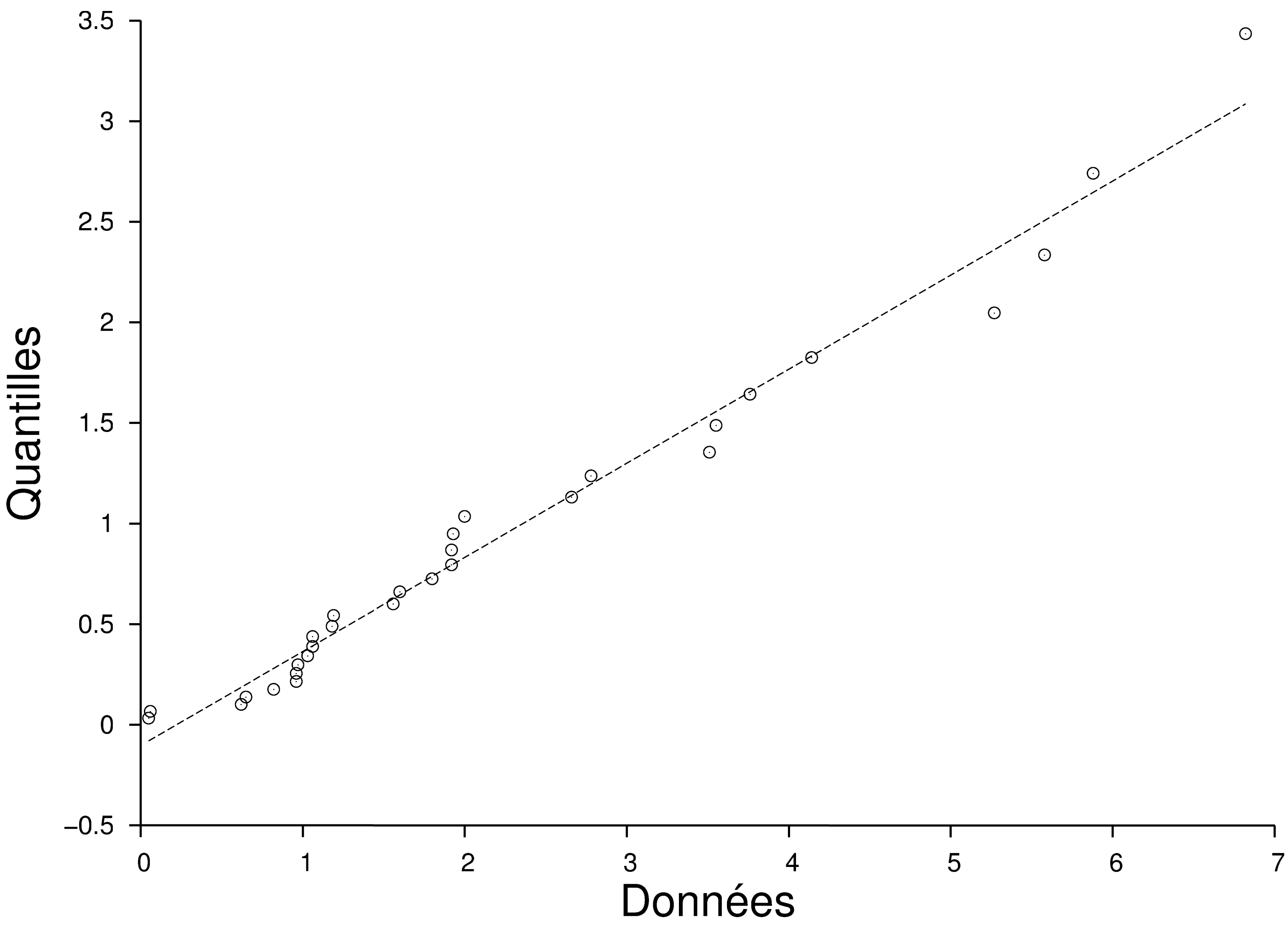
Exponential $$\left\{x_{(k)}, -\ln\left(1-\frac{k}{n+1}\right)\right\}$$

Lognormal $$\left\{\ln(x_{(k)}), \Phi^{-1}\left(\frac{k}{n+1}\right)\right\}$$

Weibull $$\left\{\ln(x_{(k)}), \ln\left[-\ln\left(1-\frac{k}{n+1}\right)\right]\right\}$$


Quantilles
Données
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
4
4.5
5
5.5
6
6.5


Quantilles
Données
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
1
2
3
4
5
6
7

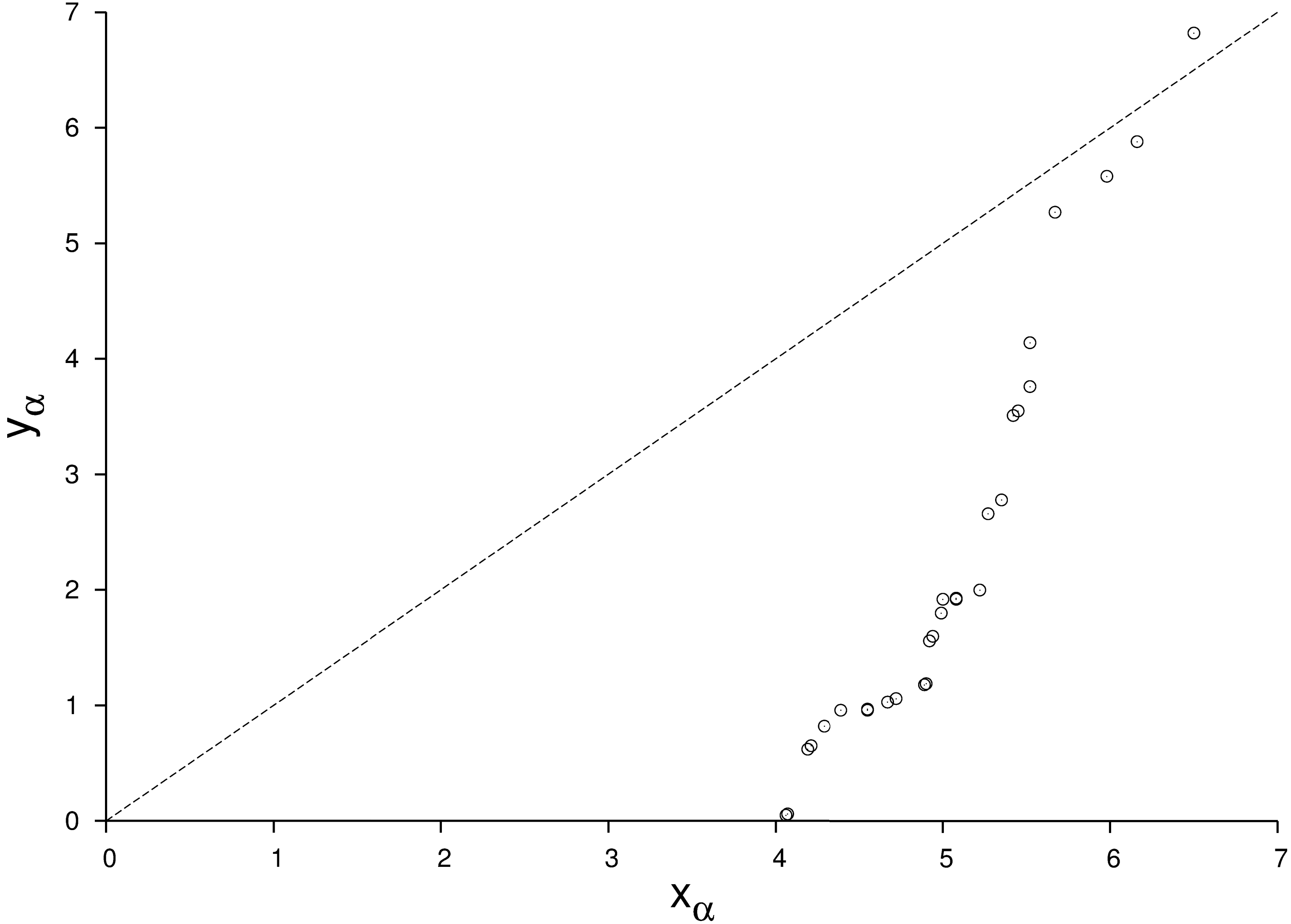

3.5
3
2.5
2
1.5
1
0.5
0
−0.5
Quantilles
0
1
2
3
4
5
6
7
Données

The QQ-Plot

- This graphical method can be used in order to check if two sets of observations $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ have the same unknown cdf $H_0 : F_X(x) = F_Y(y)$, or not $H_1 : F_X(x) \neq F_Y(y)$.
- Plot the points

$$\{(x_{\alpha_1}, y_{\alpha_1}), \ldots, (x_{\alpha_p}, y_{\alpha_p})\}$$

- If these points are approximatively on the line $y = x$, then we decide $H_0 : F_X(x) = F_Y(y)$.
- For the α_i's, we can choose $\{1/(p+1), \ldots, p/(p+1)\}$, with $p = \min(m, n)$.

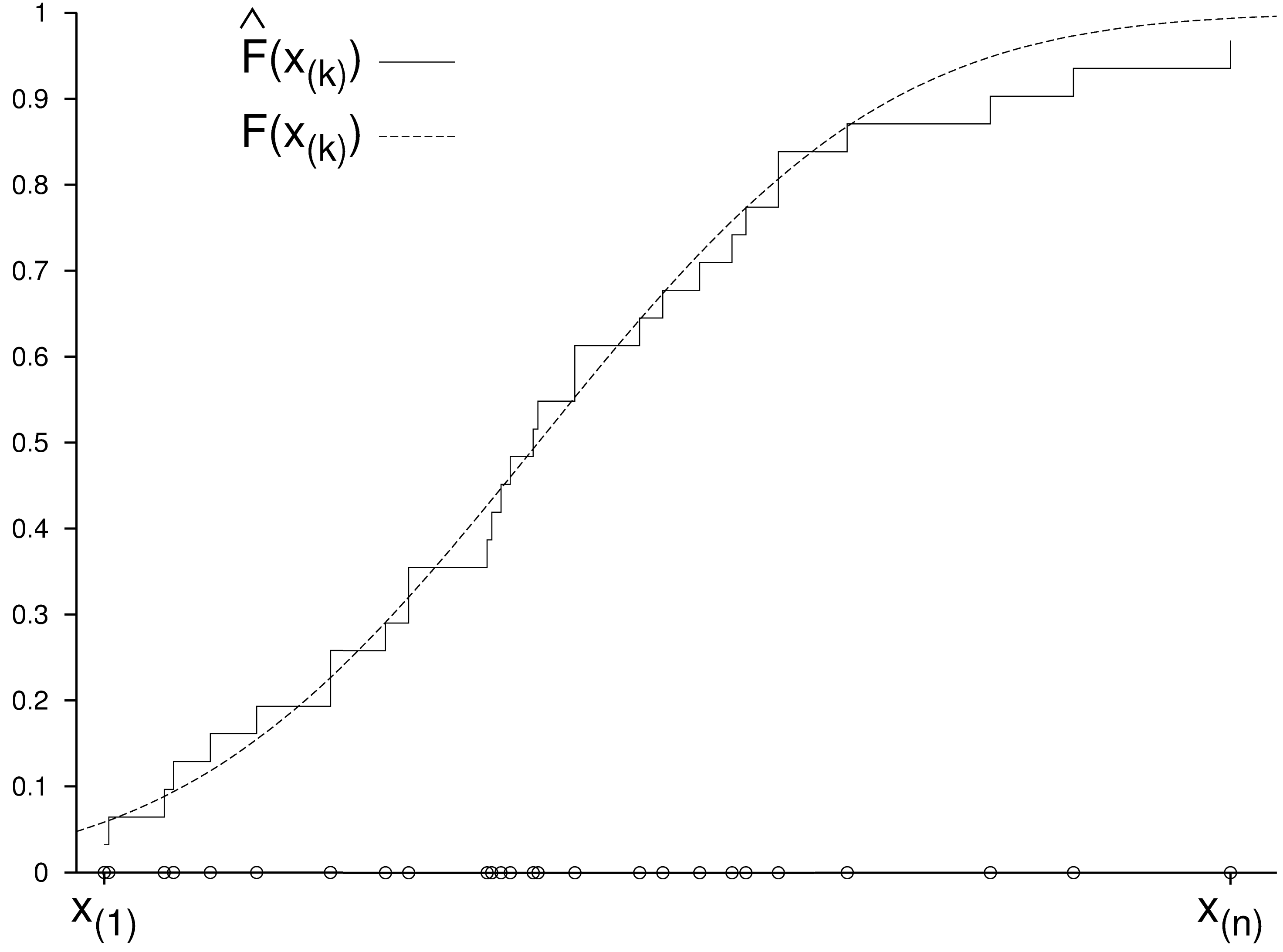

y_α
x_α
0
1
2
3
4
5
6
7

The Kolmogorov-Smirnov test

- For each $x_{(k)}$ compute $\hat{F}(x_{(k)})$.
- For each $x_{(k)}$ compute $F(x_{(k)})$.
- Compute the statistic D_n

$$D_n = \max_{k=1\dots n} |\hat{F}(x_{(k)}) - F(x_{(k)})|$$

- If $D_n > D_n^{-1}(1-\alpha, n)$ then $H_0 : F_X(x) = F(x)$ will be rejected.


$\hat{F}(x_{(k)})$
$F(x_{(k)})$
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
$x_{(1)}$
$x_{(n)}$

	$1-\alpha$				
n	0.8	0.85	0.9	0.95	0.99
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.823
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.361
20	0.231	0.246	0.264	0.294	0.352
25	0.21	0.22	0.24	0.264	0.32
30	0.19	0.20	0.22	0.242	0.29
35	0.18	0.19	0.21	0.23	0.27
∞	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

The Anderson-Darling test

- Compute $\hat{\mu}$ and S.
- Compute the probabilities $z_{(k)} = \Phi\left(\frac{x_{(k)} - \hat{\mu}}{S}\right)$
- Compute the statistics

$$
\begin{aligned}
A &= -\frac{1}{n}\left\{\sum_{k=1}^{n}(2k-1)\{\ln z_{(k)} + \ln(1 - z_{(n+1-k)})\}\right\} - n \\
A^* &= (1 + 0.75/n + 2.25/n^2)A
\end{aligned}
$$

- If $A^* > A^{-1}(1-\alpha)$ then reject the hypothesis of normality.

$1-\alpha$	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995
$A^{-1}(1-\alpha)$	0.472	0.509	0.561	0.631	0.752	0.873	1.035	1.159

The normal Kolmogorov-Smirnov test

- Compute $\hat{\mu}$ and S.
- Compute probabilities $z_{(k)} = \Phi\left(\frac{x_{(k)} - \hat{\mu}}{S}\right)$
- Compute the statistics

$$\begin{aligned} D &= \max_{k=1\ldots n} \{k/n - z_{(k)}, z_{(k)} - (k-1)/n\} \\ D^* &= (\sqrt{n} + 0.85/\sqrt{n} - 0.01)D \end{aligned}$$

- If $D^* > D^{-1}(1-\alpha)$ then reject the hypothesis of normality.

$1-\alpha$	0.85	0.90	0.95	0.975	0.99
$D^{-1}(1-\alpha)$	0.775	0.819	0.895	0.955	1.035

The Skewness and Kurtosis tests

- Compute $\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j$
- Compute $\hat{\mu_k} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{\mu})^k$ for $k = 2, 3, 4$.
- Compute $\hat{\gamma_3} = \frac{\hat{\mu_3}}{\hat{\mu_2}^{3/2}}$ and $\hat{\gamma_4} = \frac{\hat{\mu_4}}{\hat{\mu_2}^{2}} - 3$.
- Compute $V(\hat{\gamma_3})$ and $V(\hat{\gamma_4})$

$$
\begin{aligned}
V(\hat{\gamma_3}) &= \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \\
V(\hat{\gamma_4}) &= \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}
\end{aligned}
$$

The Skewness and Kurtosis tests

- Compute the standardized variables

$$\hat{\delta}_3 = \frac{\hat{\gamma}_3}{\sqrt{V(\hat{\gamma}_3)}} \qquad \text{and} \qquad \hat{\delta}_4 = \frac{\hat{\gamma}_4}{\sqrt{V(\hat{\gamma}_4)}}$$

- The hypothesis of normality is rejected if
  - $2\Phi(-|\widehat{\delta_3}|) \leq \alpha$ or,
  - $2\Phi(-|\widehat{\delta_4}|) \leq \alpha$ or,
  - $1 - F_{\chi^2}(\widehat{\delta_3}^2 + \widehat{\delta_4}^2, 2) \leq \alpha$.

Mardia's skewness & kurtosis test

- We have n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^p$ and we want to check if these data have a multivariable normal $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. We have to compute
  - an estimation $\bar{\mathbf{x}}$ of $\boldsymbol{\mu}$ and $\mathbf{S}$ of $\boldsymbol{\Sigma}$

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \qquad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

  - an estimation $\widehat{\delta_3}$ of the Mardia's skewness coefficient

$$\widehat{\delta_3} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{(\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})\}^3$$

  - an estimation $\widehat{\delta_4}$ of the Mardia's kurtosis coefficient

$$\widehat{\delta_4} = \frac{1}{n}\sum_{i=1}^{n}\{(\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\}^2$$

Mardia's skewness & kurtosis test

- Assuming n is large enought, the hypothesis of multivariable normality must be rejected if

$$\begin{array}{rcl} 1 - F_{\chi^2}\left\{\dfrac{n\widehat{\delta_3}}{6}, \dfrac{p(p+1)(p+2)}{6}\right\} & \leq & \alpha \\ & \text{or} & \\ 2\Phi\left\{-\dfrac{|\widehat{\delta_4} - p(p+2)|}{\sqrt{\dfrac{8p(p+2)}{n}}}\right\} & \leq & \alpha \end{array}$$

Multivariable Normal Q-Plot

- We have n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^p$ and we want to check if these data have a multivariable normal distribution.
- If $\mathbf{x}_i$ is a multivariable normal point

$$y_i = (\mathbf{x}_i - \bar{\mathbf{x}})^T S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

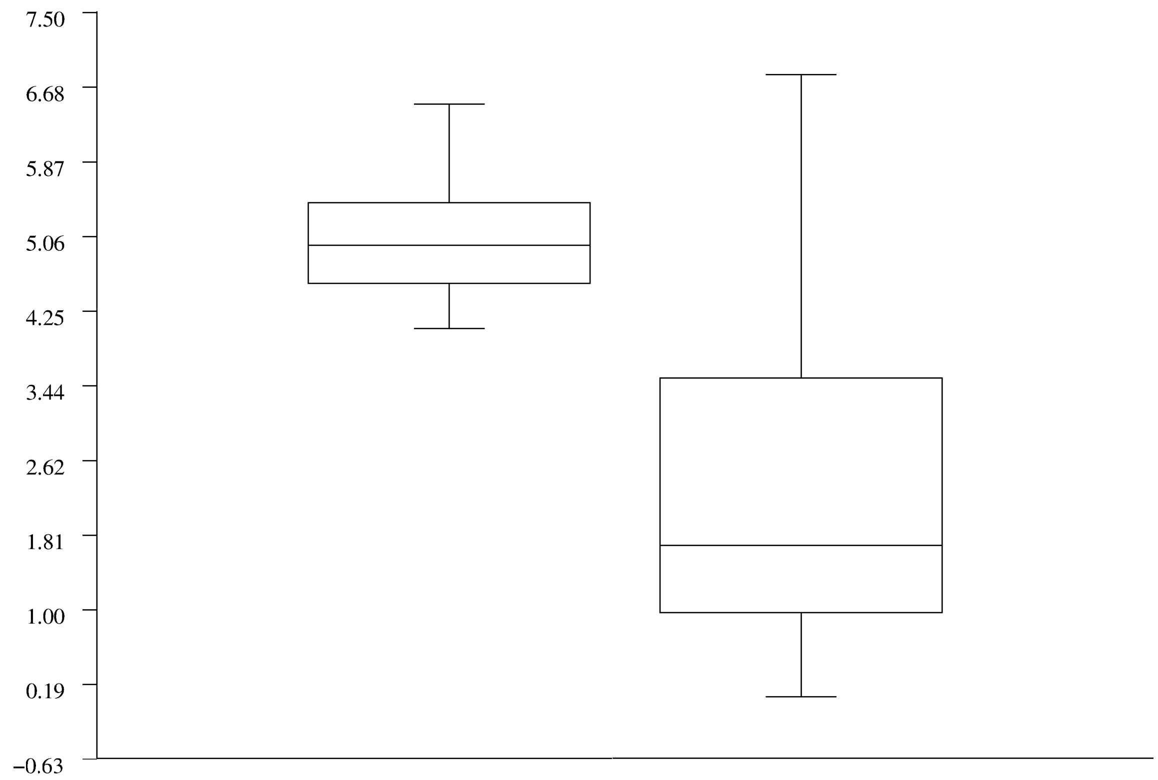
is a chi-square random variable with p degree of freedom.

- In order to test the fit of the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with a multivariable normal distribution we just need to plot the points

$$\left\{ y_{(i)}, F_{\chi^2}^{-1}\left(\frac{i}{n+1}, p\right) \right\}$$

and to check that those points are approximatively on a straight line.

The Box-Plot



The Box-Plot

- Compute $\hat{X}_{0.25}$, $\tilde{X} = \hat{X}_{0.5}$ and $\hat{X}_{0.75}$.
- Compute the interquartile range $IQR = \hat{X}_{0.75} - \hat{X}_{0.25}$.
- For the smallest values

$$\begin{aligned} X_1^{\min} &= \min(X_k | \hat{X}_{0.25} - 1.5IQR \leq X_k < \hat{X}_{0.25}) \\ X_2^{\min} &= \min(X_k | \hat{X}_{0.25} - 3IQR \leq X_k < \hat{X}_{0.25} - 1.5IQR) \end{aligned}$$

- For the largest values

$$\begin{aligned} X_1^{\max} &= \max(X_k | \hat{X}_{0.75} < X_k \leq \hat{X}_{0.75} + 1.5IQR) \\ X_2^{\max} &= \max(X_k | \hat{X}_{0.75} + 1.5IQR < X_k \leq \hat{X}_{0.75} + 3IQR) \end{aligned}$$

The Box-Plot

- A vertical box with sides bounded by $\hat{X}_{0.25}$ and $\hat{X}_{0.75}$.
- An horizontal line corresponding to the sample median $\tilde{X}$.
- A vertical line from point $X_1^{\min}$ to point $\hat{X}_{0.25}$.
- A vertical line from point $\hat{X}_{0.75}$ to point $X_1^{\max}$.
- The points X_k (if they exist) such that $X_2^{\min} \leq X_k < X_1^{\min}$ or $X_1^{\max} < X_k \leq X_2^{\max}$ are plotted with a "$\circ$" and have to be considered as potential outliers.
- The points X_k (if they exist) such that $X_k < X_2^{\min}$ or $X_2^{\max} < X_k$ are plotted with a "$\times$" and have to be considered seriously as potential outliers.

Non parametrical estimation of a pdf

- Kernel estimator

$$\hat{f}(x) \simeq \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

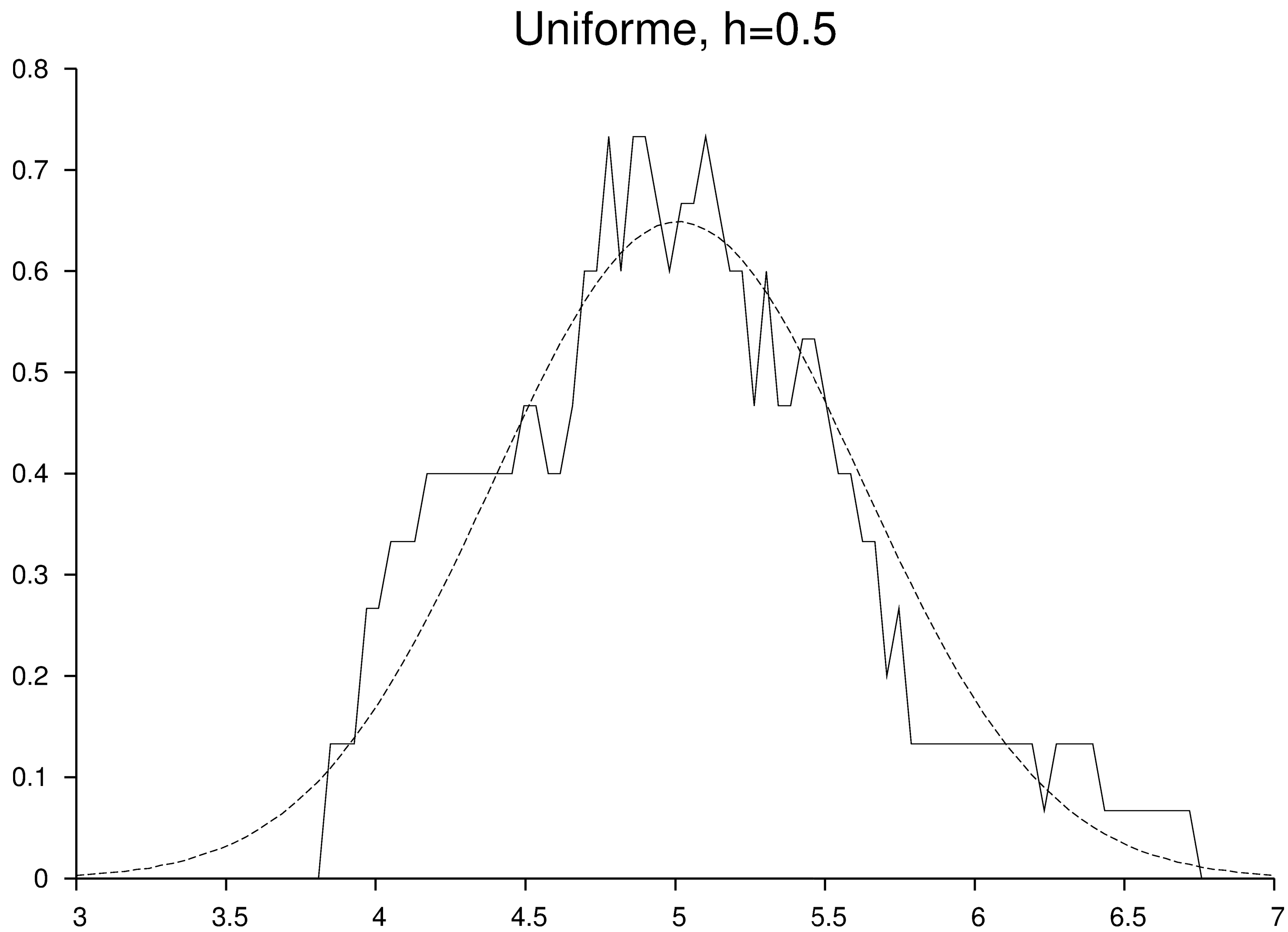
- $K(u)$ is the *Kernel function*.

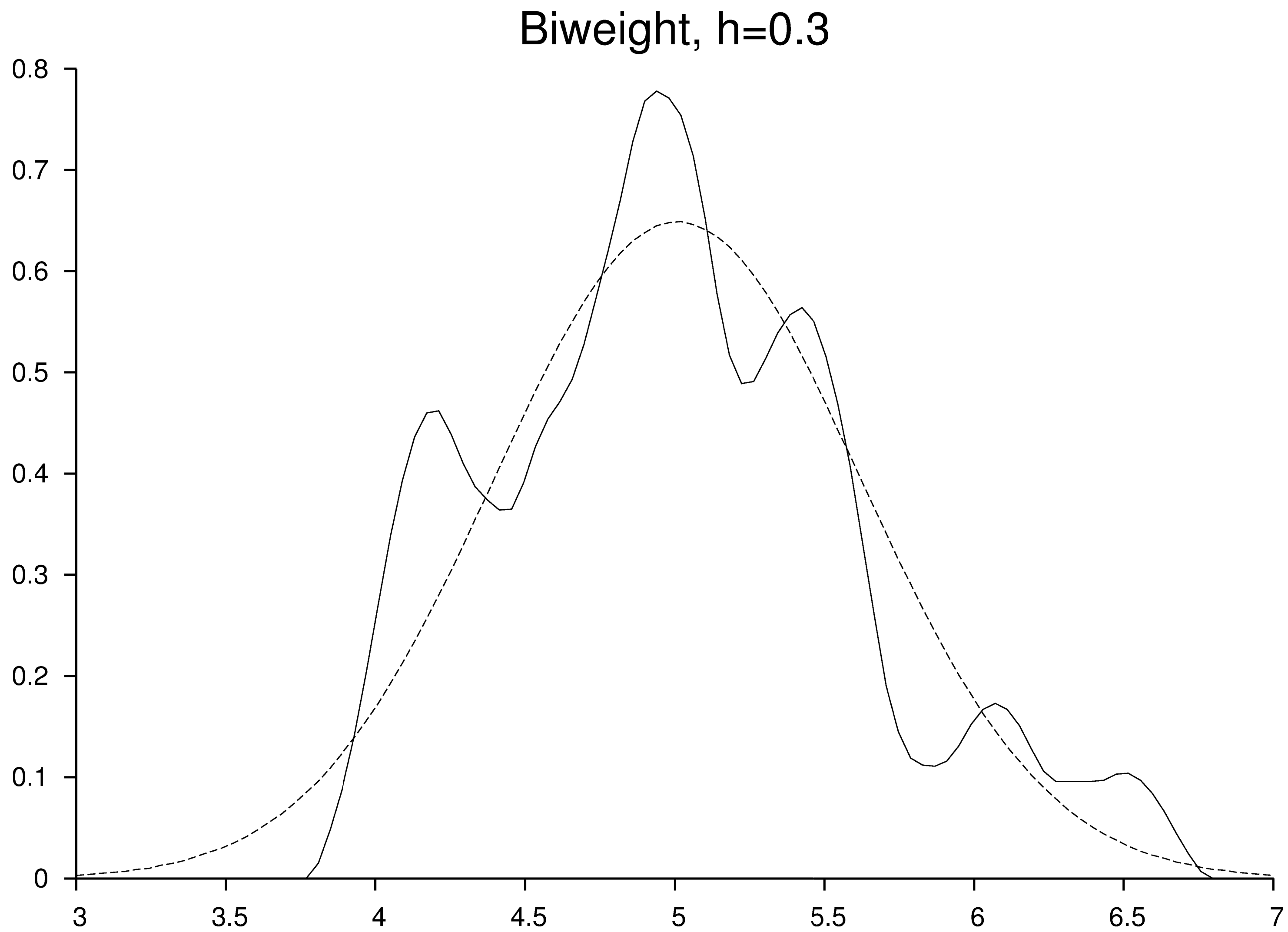
$$K(u) = \begin{cases} 0 & \text{if } |u| > 1/2 \\ 1 & \text{if } |u| \leq 1/2 \end{cases}$$

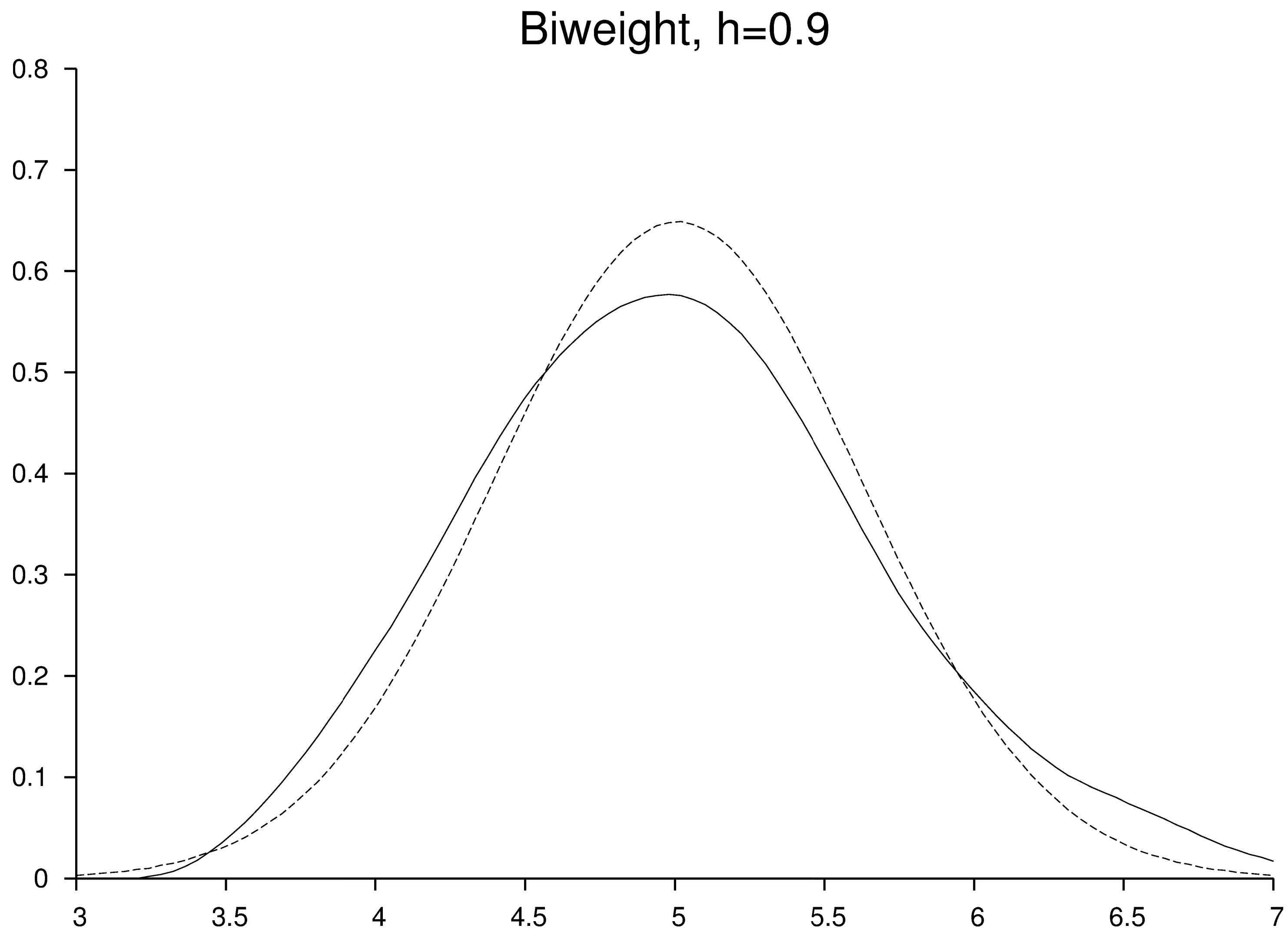
- h is the *bandwidth*.
  - if the value of h is too small the estimation of $f(x)$ is not smooth enough.
  - if the value of h is too large the estimation of $f(x)$ is too smooth (lost of information).

Other possible kernels

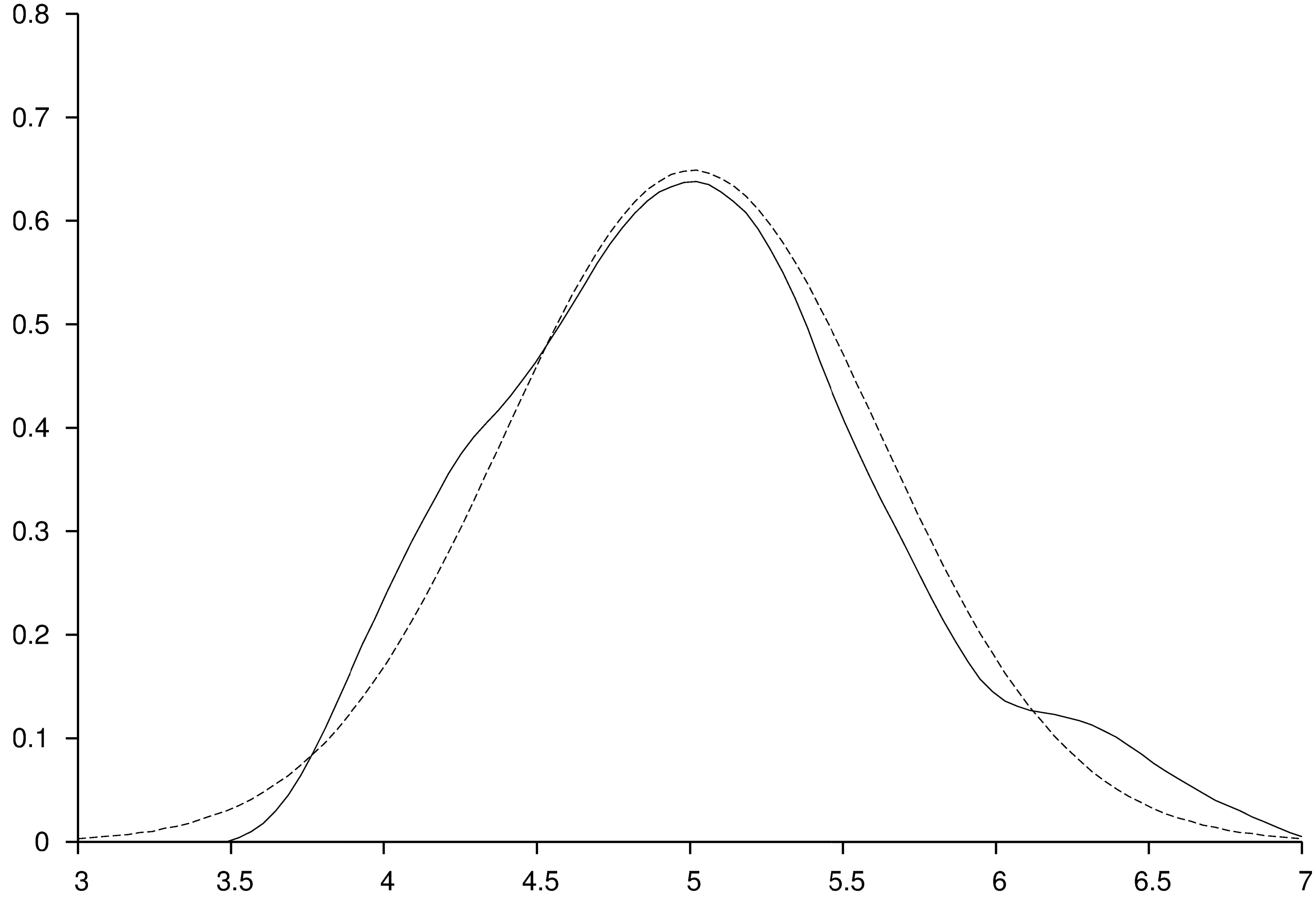
- $K(u) = 1 - |u|$, $|u| \leq 1$ (triangular).
- $K(u) = \frac{3}{4}(1 - u^2)$, $|u| \leq 1$ (epanechnikov).
- $K(u) = \frac{15}{16}(1 - u^2)^2$, $|u| \leq 1$ (biweight).
- $K(u) = \frac{35}{32}(1 - u^2)^3$, $|u| \leq 1$ (triweight).
- $K(u) = \frac{e^{-u^2/2}}{\sqrt{2\pi}}$, (normal).
- $K(u) = \frac{1}{\pi(1 + u^2)}$, (laplace).
- $K(u) = \frac{2\sin^2(u/2)}{\pi u^2}$.


Uniforme, h=0.5
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3
3.5
4
4.5
5
5.5
6
6.5
7


Biweight, h=0.3
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3
3.5
4
4.5
5
5.5
6
6.5
7


Biweight, h=0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3
3.5
4
4.5
5
5.5
6
6.5
7

Biweight, h=0.6



The χ^2 test of fit

- We have a sample of n discrete observations. Among these observations
  - n_1 take the value x_1,

    $\vdots$
  - n_k take the value x_k
- We denote $f_i = f(x_i) = P(X = x_i)$. The statistic of the test is

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - nf_i)^2}{nf_i}$$

- We reject $H_0 : f_X(x) = f(x)$ if

$$1 - F_{\chi^2}(\chi^2, k - e - 1) \leq \alpha$$

where e is the number of estimated parameters.

Contingency table

- n individuals are divided into two factors A and B having respectively r and s modalities.
- Example: n individuals are divided in function of their
  - appurtenance to a political part (modalities $A_1, \ldots, A_r$).
  - appurtenance to an union (modalities $B_1, \ldots, B_s$).
- We have a *contingency table* with r rows and s columns.
  - $n_{i,j}$ is the number of individuals in the couple (A_i, B_j).
  - $n_{i,.} = \sum_{j=1}^{s} n_{i,j}$ is the number of individuals in the modality A_i.
  - $n_{.,j} = \sum_{i=1}^{r} n_{i,j}$ is the number of individuals in the modality B_j.

Contingency table

	B_1	$\cdots$	B_j	$\cdots$	B_s	
A_1	$n_{1,1}$	$\cdots$	$n_{1,j}$	$\cdots$	$n_{1,s}$	$n_{1,.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	$n_{i,1}$	$\cdots$	$n_{i,j}$	$\cdots$	$n_{i,s}$	$n_{i,.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	$n_{r,1}$	$\cdots$	$n_{r,j}$	$\cdots$	$n_{r,s}$	$n_{r,.}$
	$n_{.,1}$	$\cdots$	$n_{.,j}$	$\cdots$	$n_{.,s}$	n

The χ^2 test of independence

- The statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{i,j} - \frac{n_{i,.}n_{.,j}}{n}\right)^2}{\frac{n_{i,.}n_{.,j}}{n}}$$

- The independence of factors A and B is rejected if

$$1 - F_{\chi^2}\{\chi^2, (r-1)(s-1)\} \leq \alpha$$