Quantiles and related stuff

(quantiles are everywhere)

Arthur Charpentier

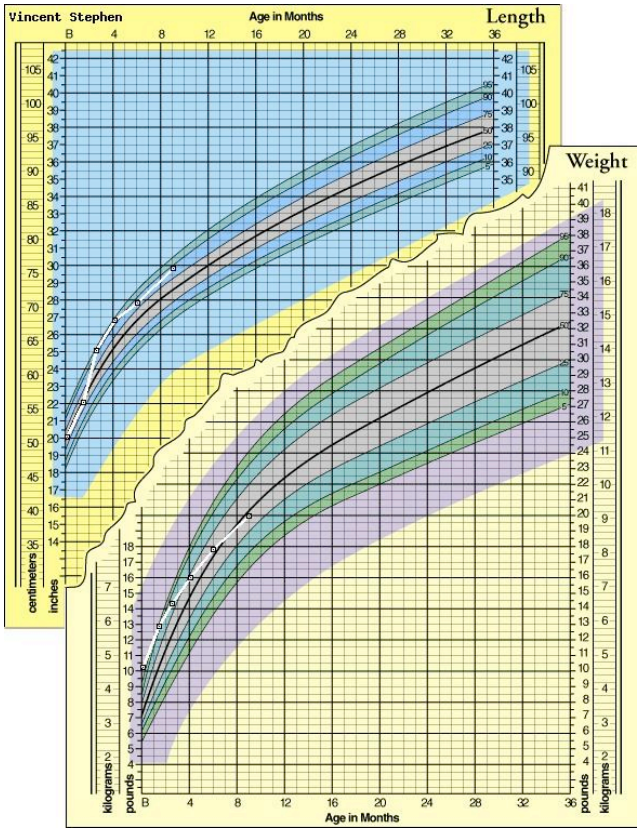
UQAM, Quantact

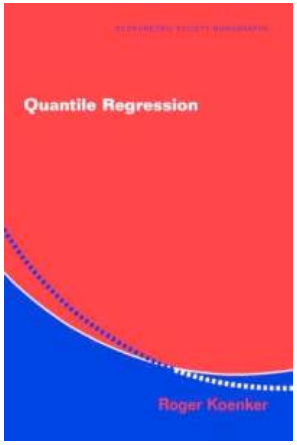
http://freakonometrics.blog.free.fr

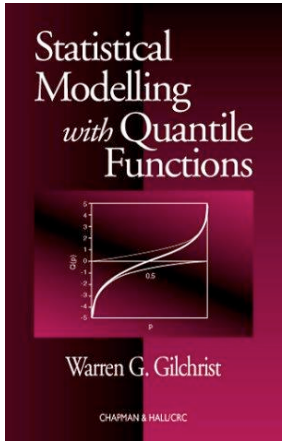
@freakonometrics

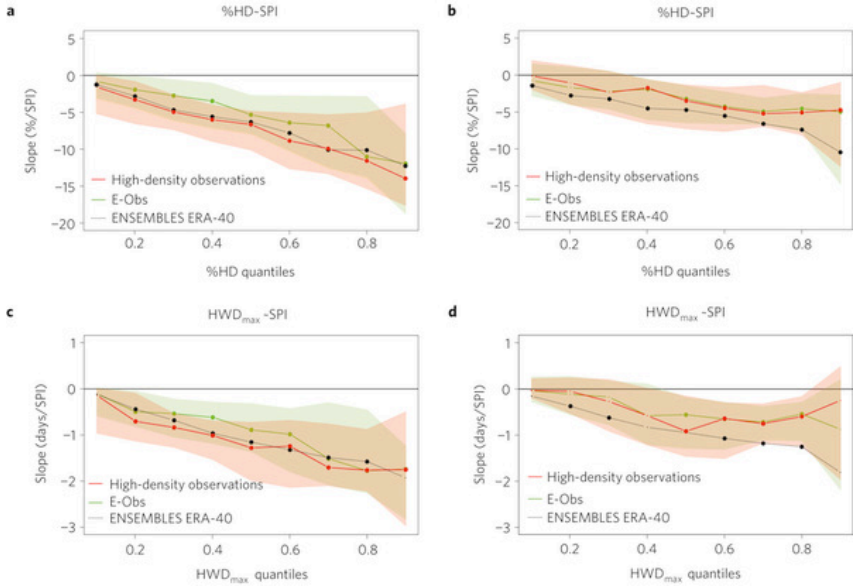
charpentier.arthur@uqam.ca

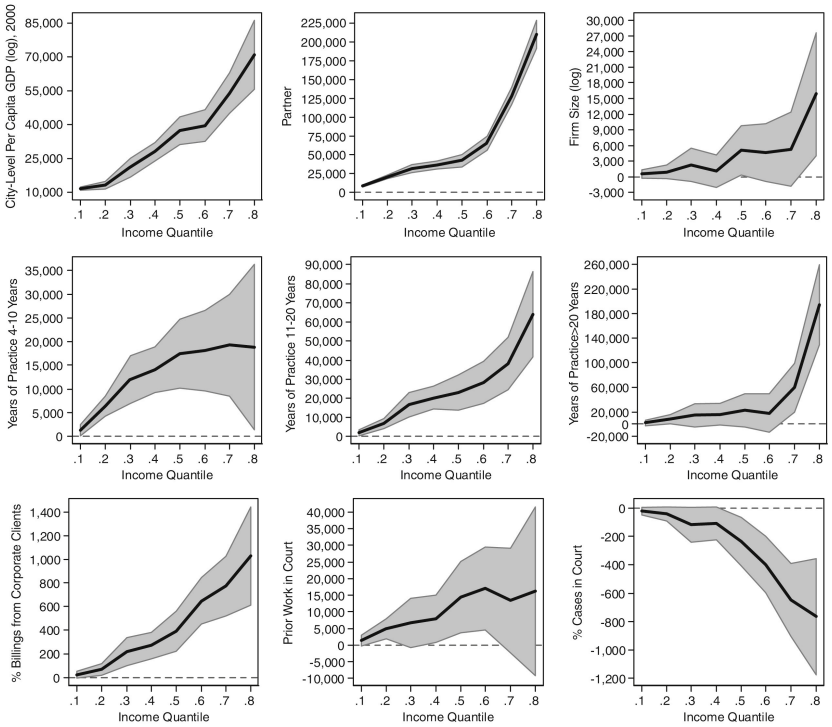
MontRéal User Group


Vincent Stephen
Age in Months
Length
Weight
centimeters
inches
kilograms
pounds
Age in Months


Quantile Regression
Roger Koenker

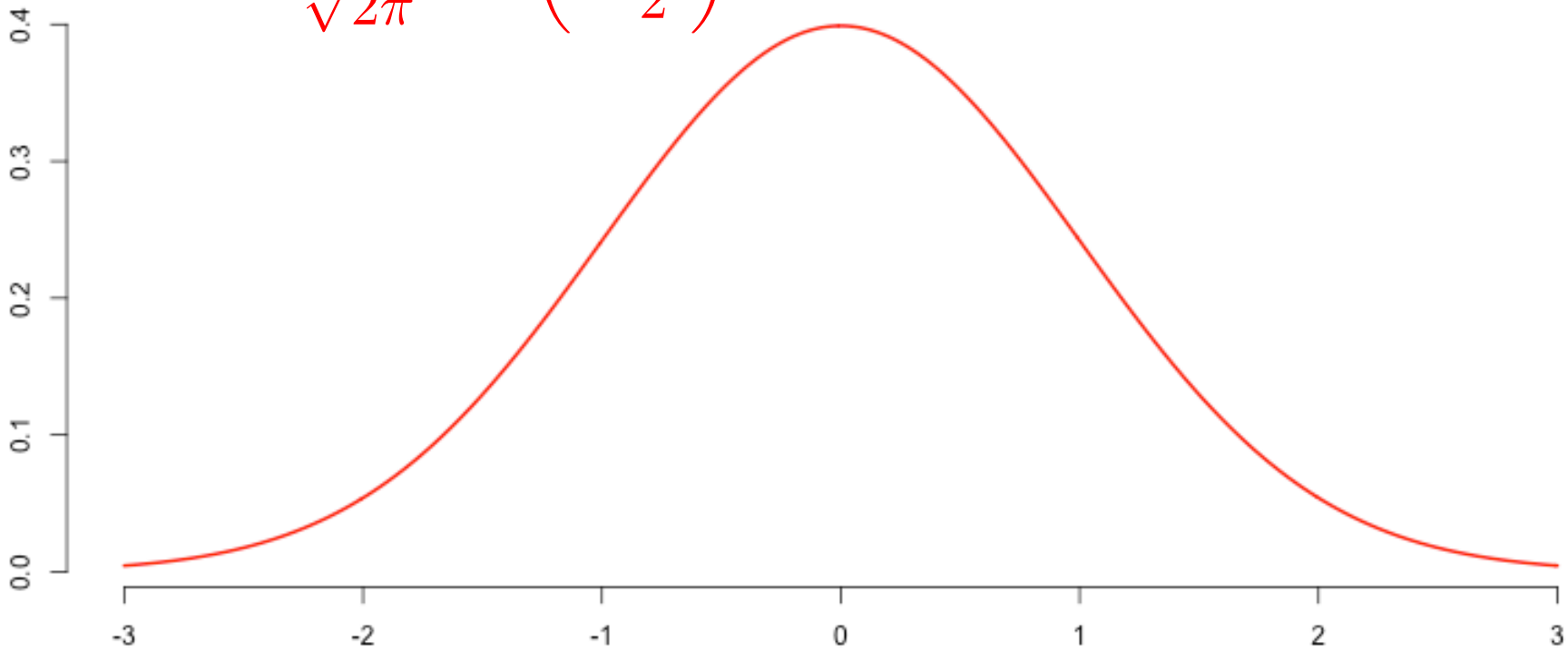

Statistical Modelling with Quantile Functions
Warren G. Gilchrist
CHAPMAN & HALL/CRC


a
%HD-SPI
Slope (%/SPI)
High-density observations
E-Obs
ENSEMBLES ERA-40
%HD quantiles
b
%HD-SPI
Slope (%/SPI)
High-density observations
E-Obs
ENSEMBLES ERA-40
%HD quantiles
c
HWDmax -SPI
Slope (days/SPI)
High-density observations
E-Obs
ENSEMBLES ERA-40
HWDmax quantiles
d
HWDmax -SPI
Slope (days/SPI)
High-density observations
E-Obs
ENSEMBLES ERA-40
HWDmax quantiles


City-Level Per Capita GDP (log), 2000
Partner
Firm Size (log)
Years of Practice 4-10 Years
Years of Practice 11-20 Years
Years of Practice>20 Years
% Billings from Corporate Clients
Prior Work in Court
% Cases in Court
Income Quantile

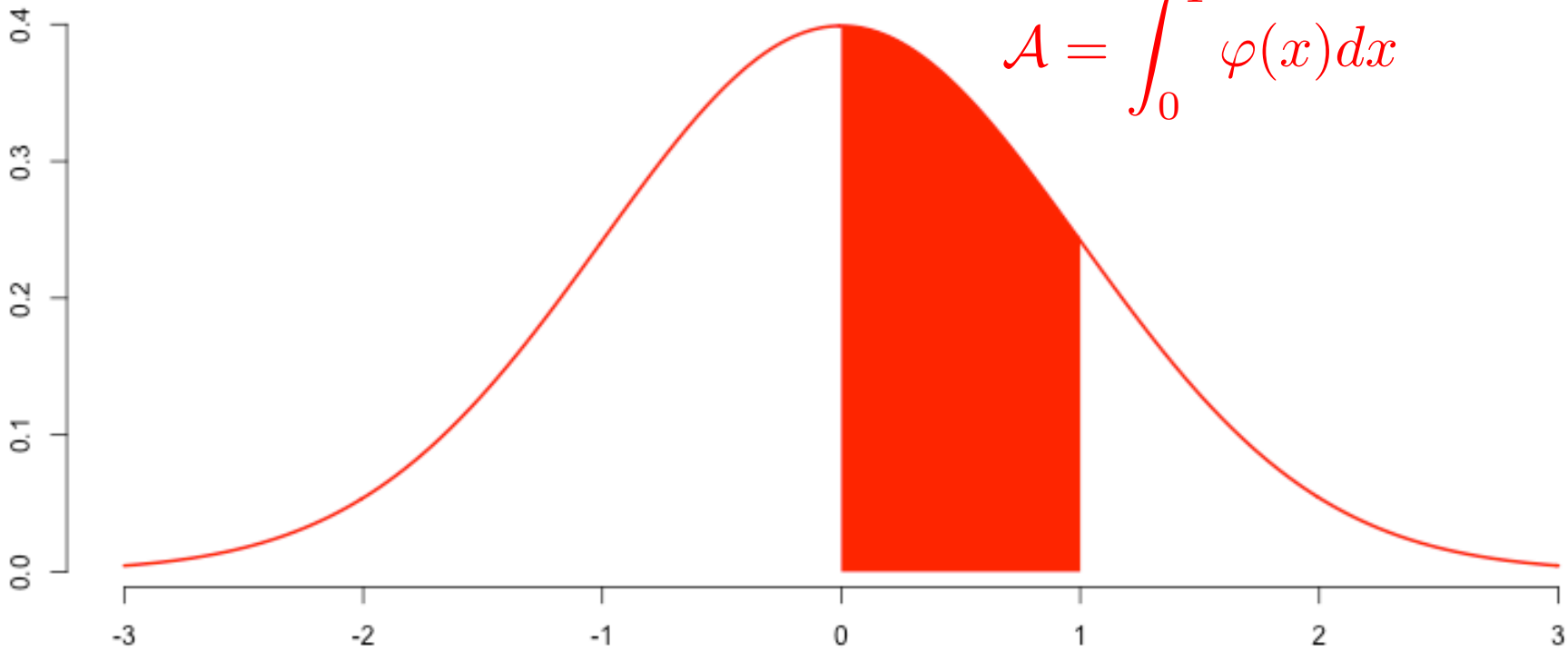
Quantiles of the $\mathcal{N}(0,1)$ distribution

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$




```
> dnorm(x,mean=0,sd=1)
```

Quantiles of the $\mathcal{N}(0,1)$ distribution

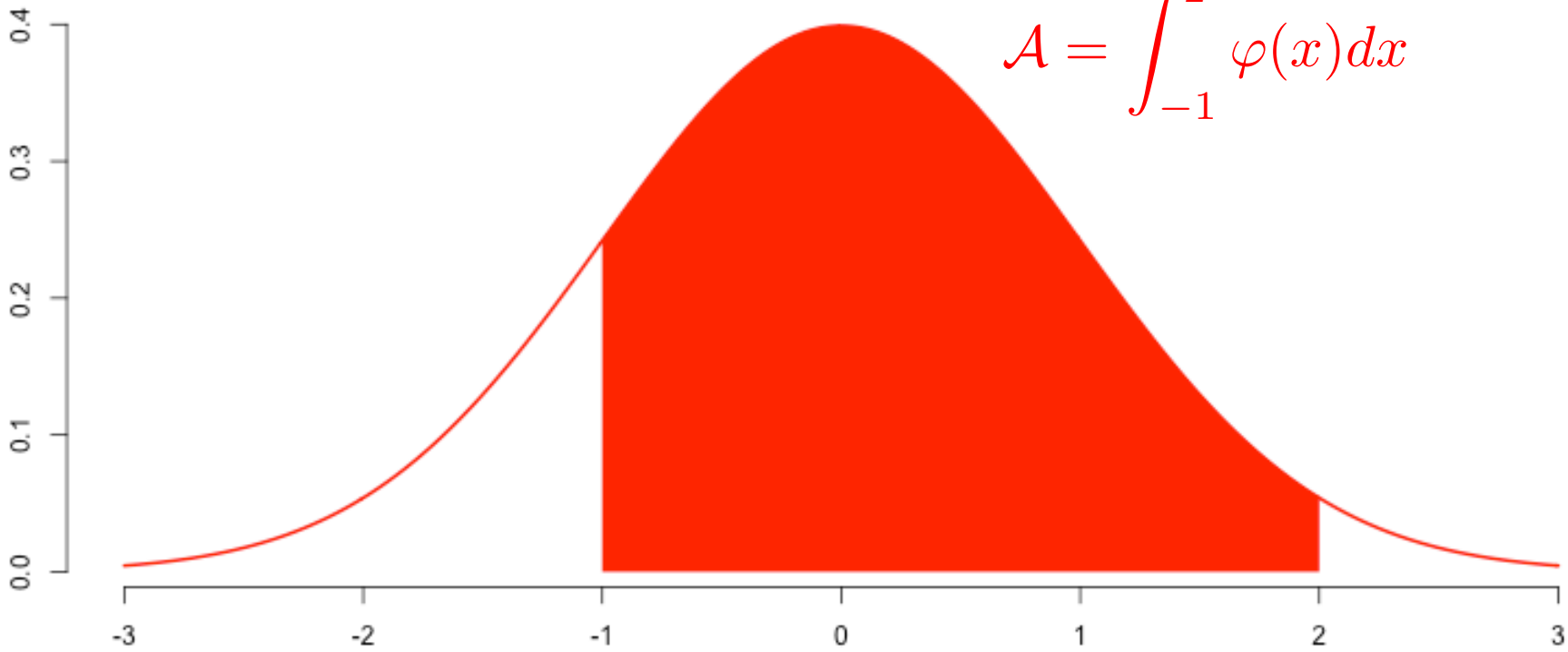


$$\mathcal{A} = \int_0^1 \varphi(x)dx$$

$$\mathcal{A} = \mathbb{P}(X \in [0,1])$$

where $X \sim \mathcal{N}(0,1)$

Quantiles of the $\mathcal{N}(0,1)$ distribution

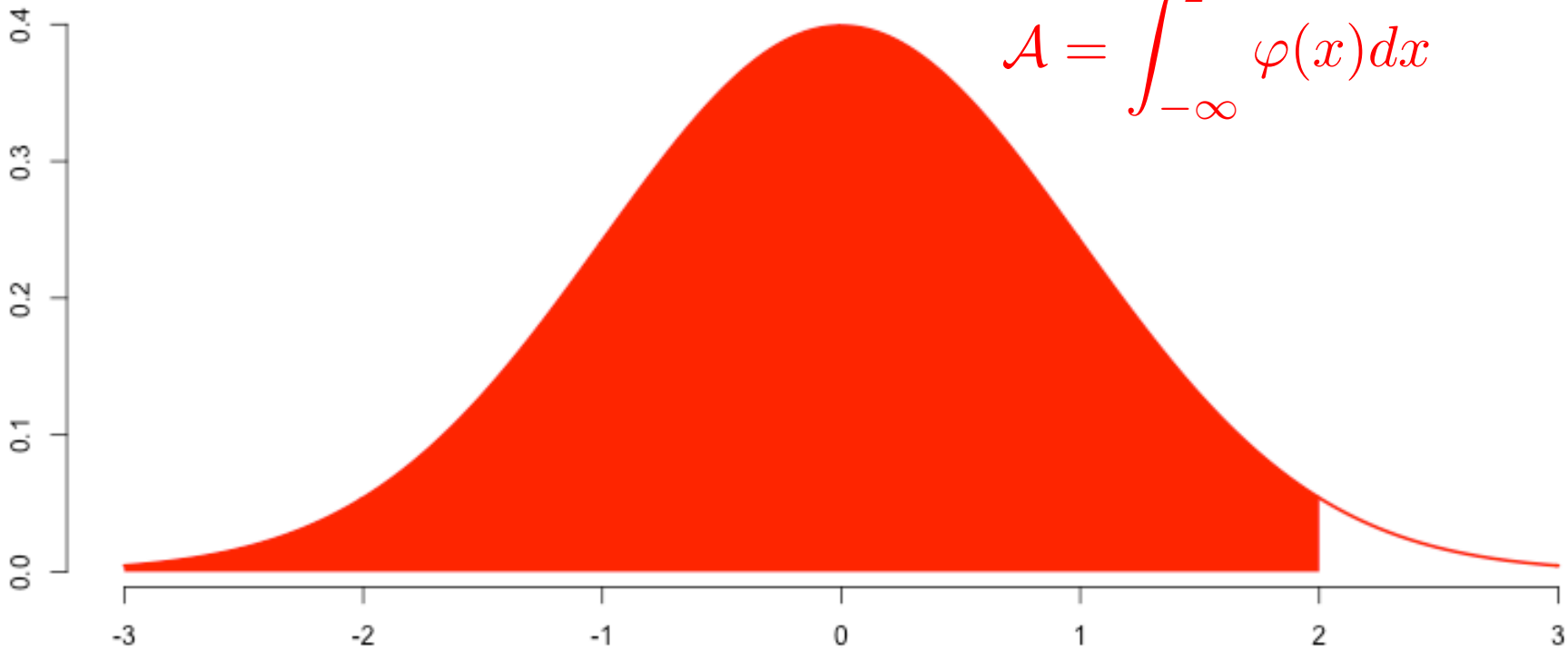


$$\mathcal{A} = \int_{-1}^{2} \varphi(x)dx$$

$$\mathcal{A} = \mathbb{P}(X \in [-1, 2])$$

where $X \sim \mathcal{N}(0,1)$

Quantiles of the $\mathcal{N}(0,1)$ distribution



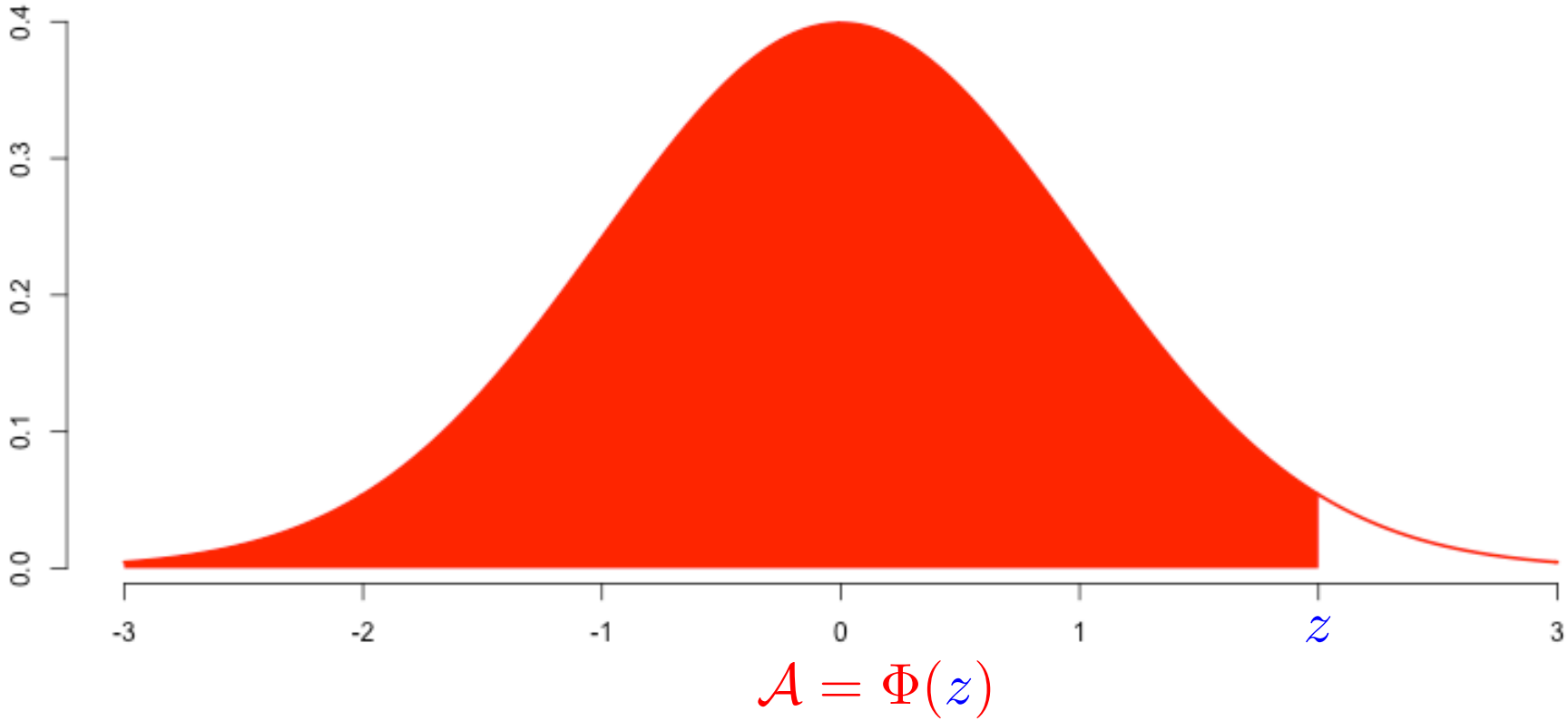
$$\mathcal{A} = \int_{-\infty}^{2} \varphi(x)dx$$

$$\mathcal{A} = \mathbb{P}(X \in (-\infty, 2]) = \mathbb{P}(X \leq 2)$$

where $X \sim \mathcal{N}(0,1)$

Quantiles of the $\mathcal{N}(0,1)$ distribution

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

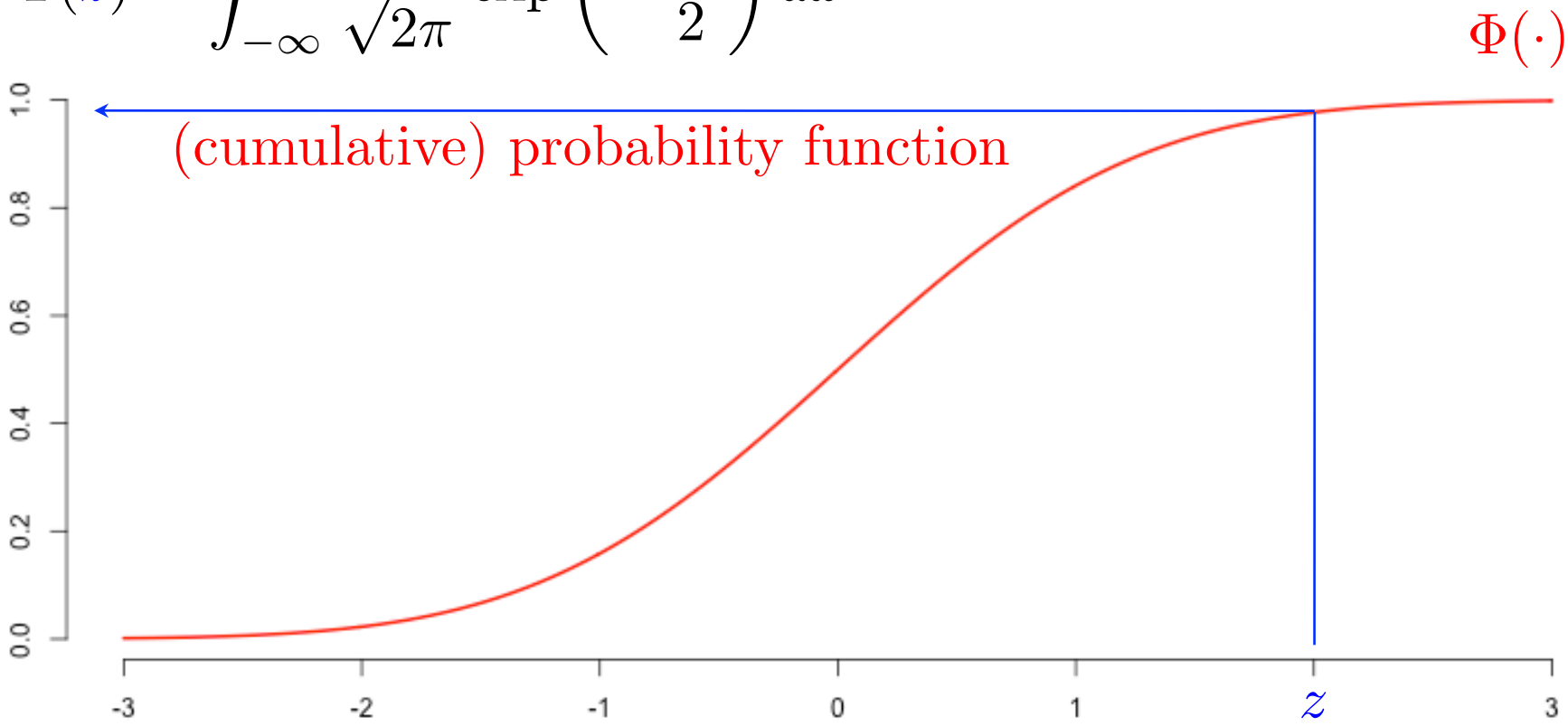


$\mathcal{A} = \Phi(z)$

cumulative distribution function c.d.f.

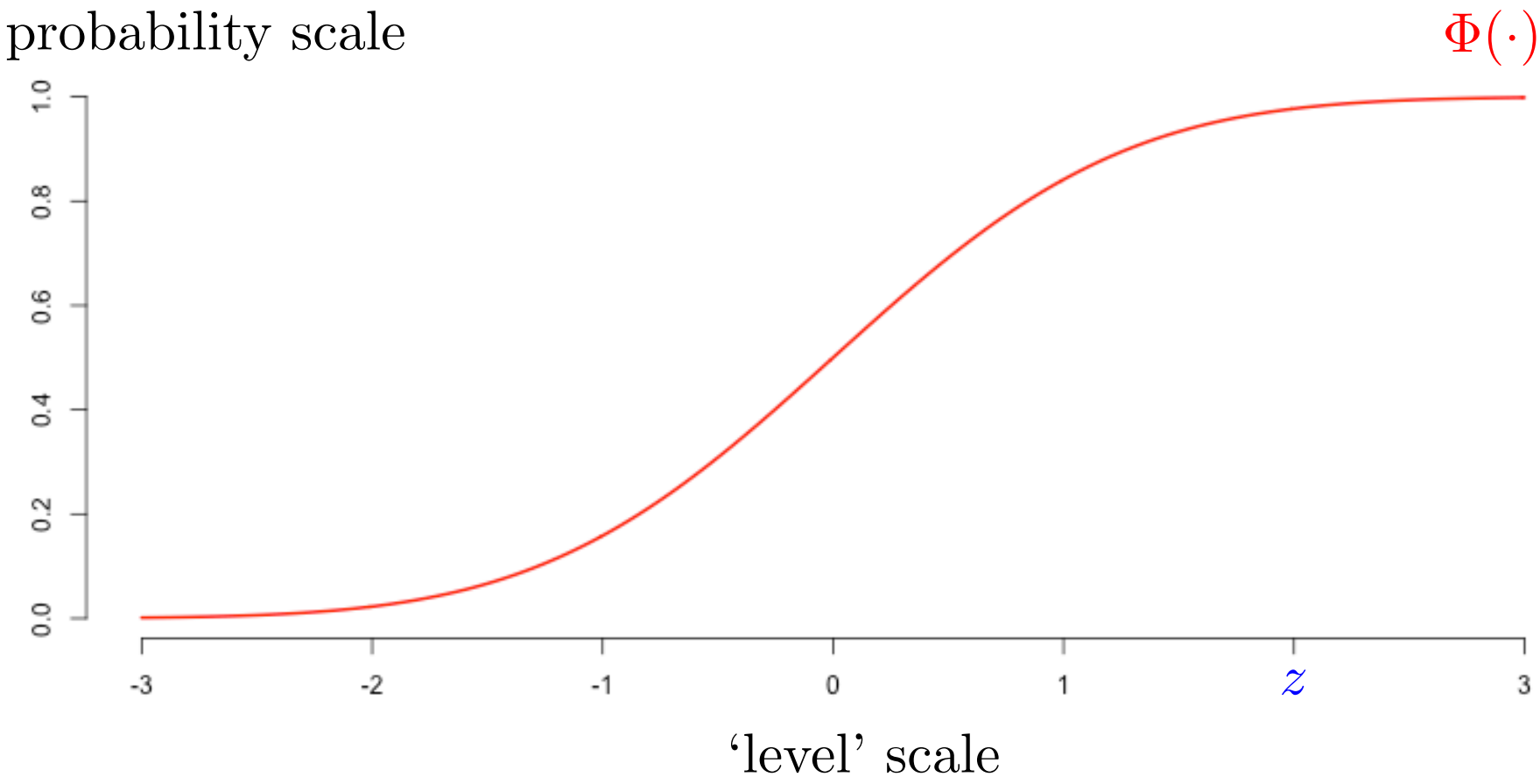
Quantiles of the $\mathcal{N}(0,1)$ distribution

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

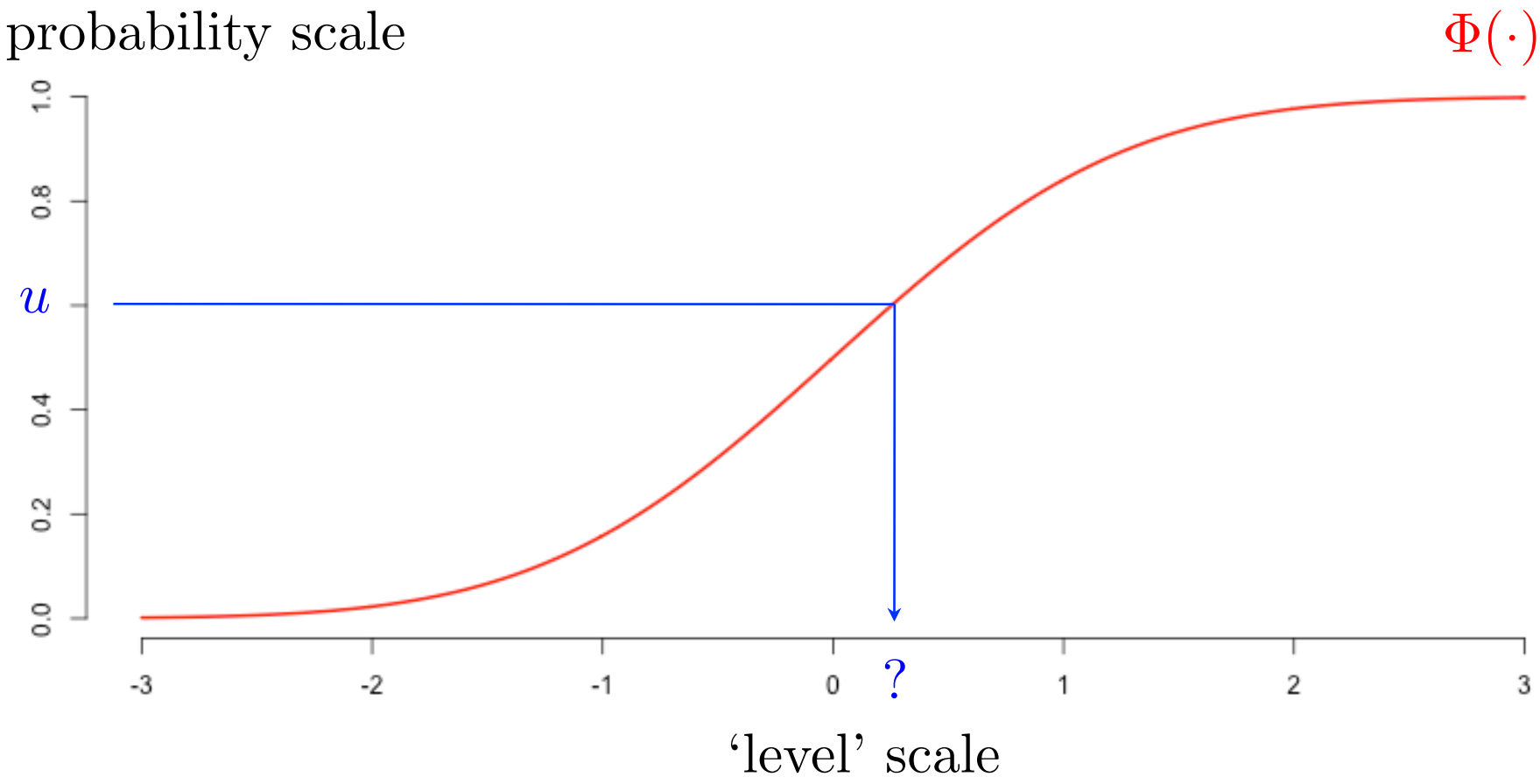



```
> pnorm(z,mean=0,sd=1)
```

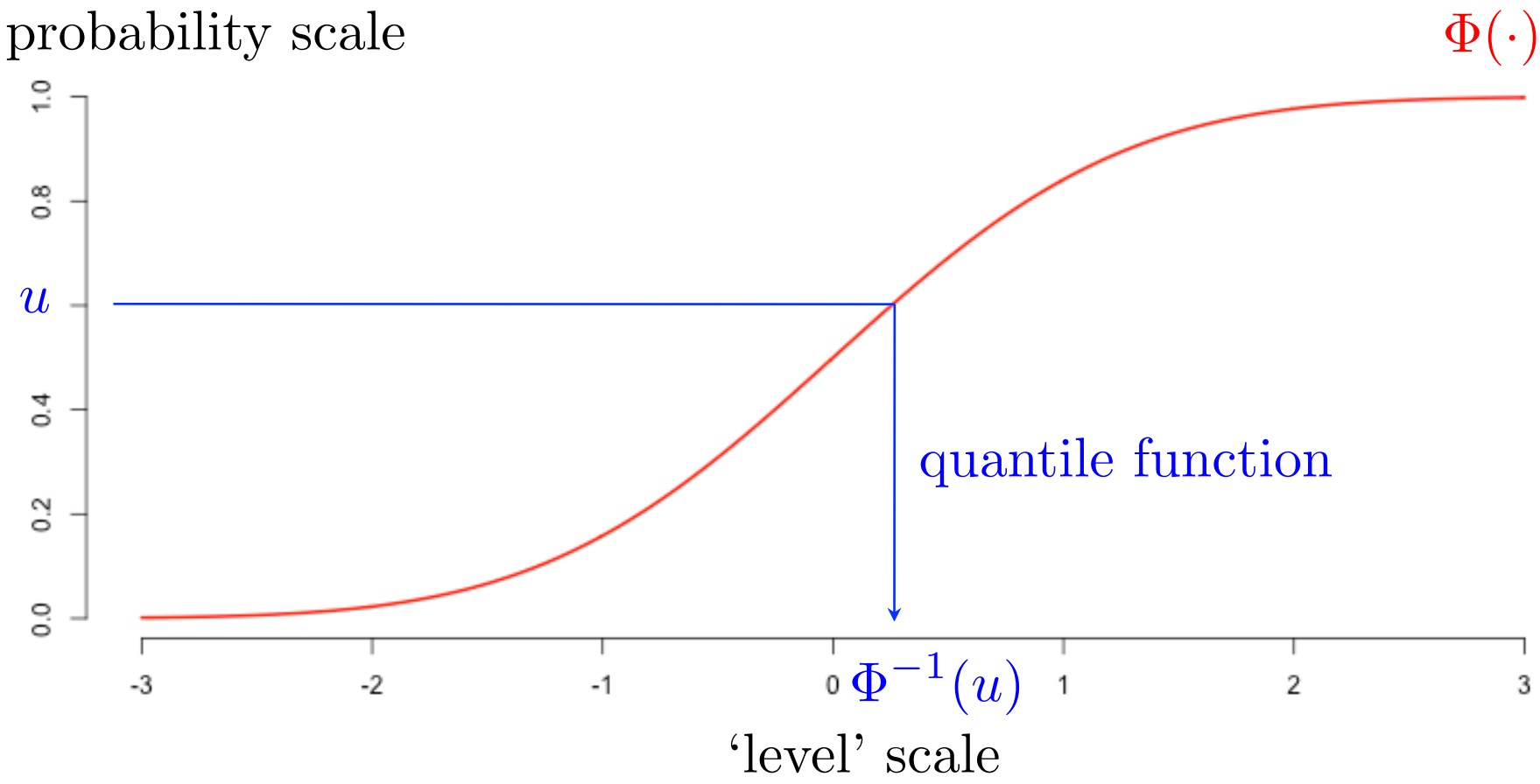
Quantiles of the $\mathcal{N}(0,1)$ distribution

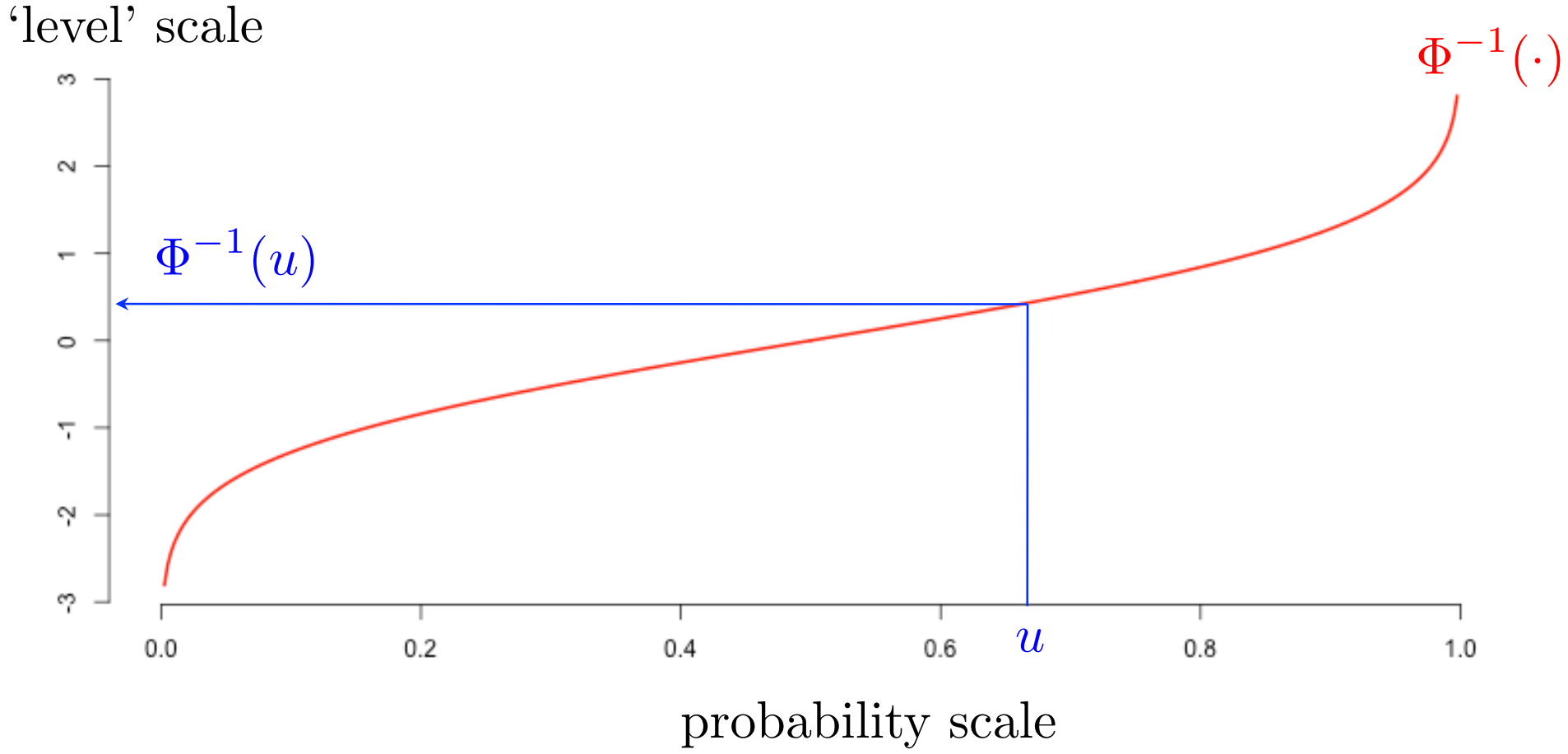


Quantiles of the $\mathcal{N}(0,1)$ distribution



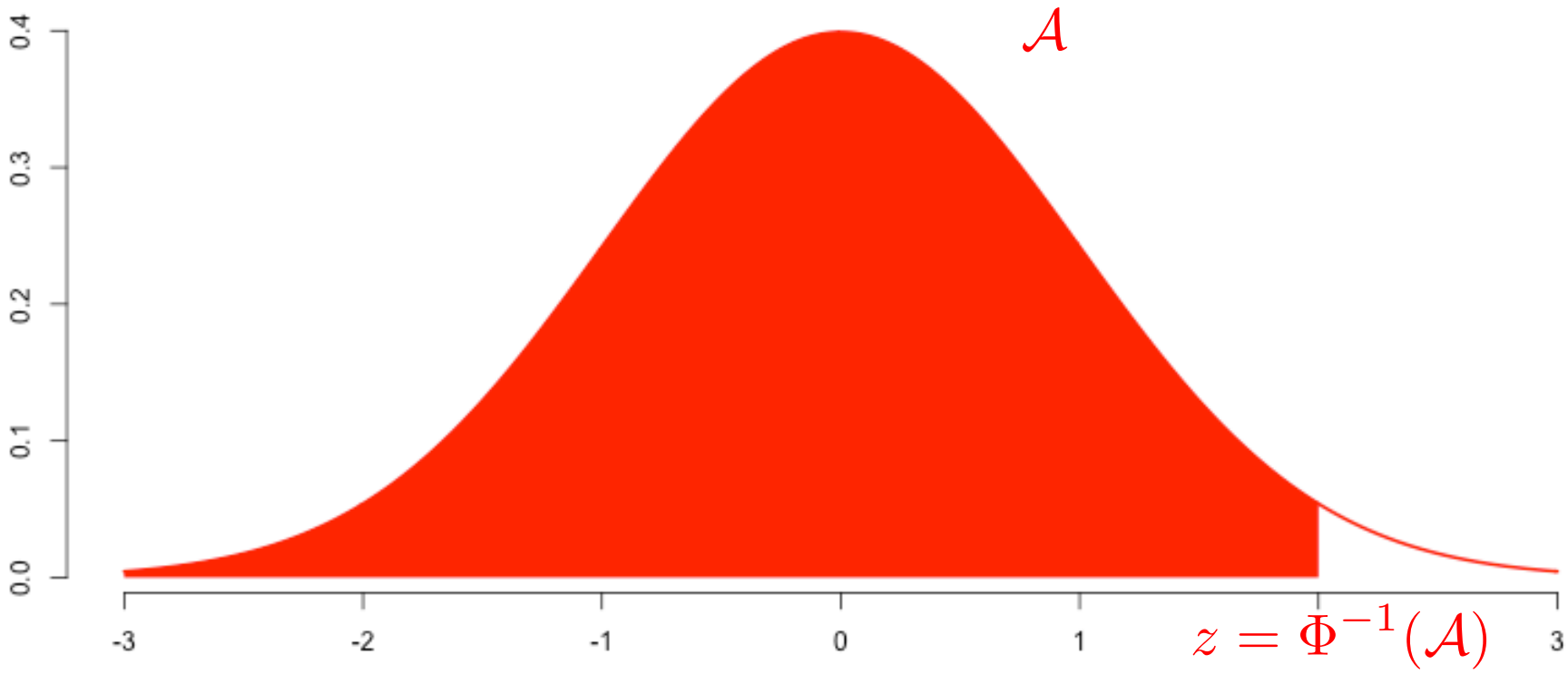
Quantiles of the $\mathcal{N}(0,1)$ distribution



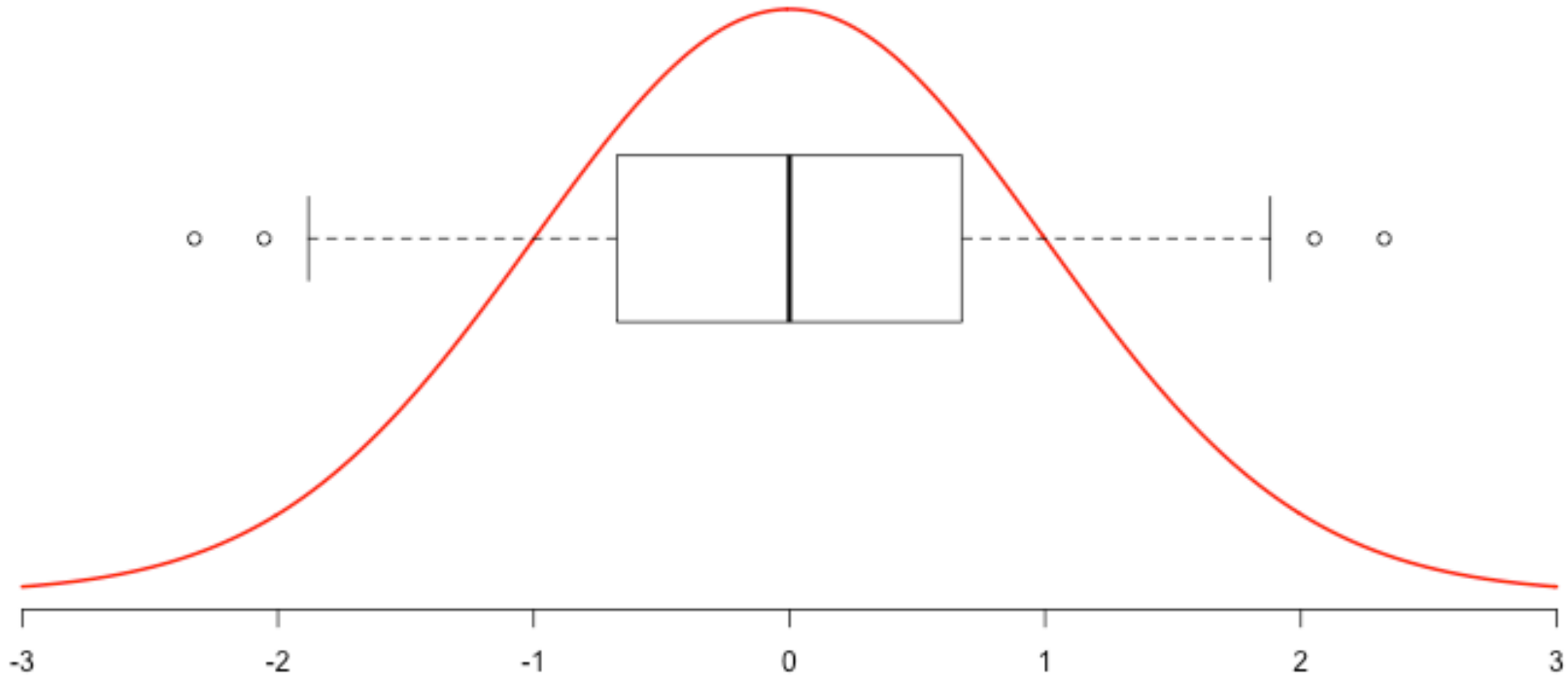



```
> qnorm(u,mean=0,sd=1)
```

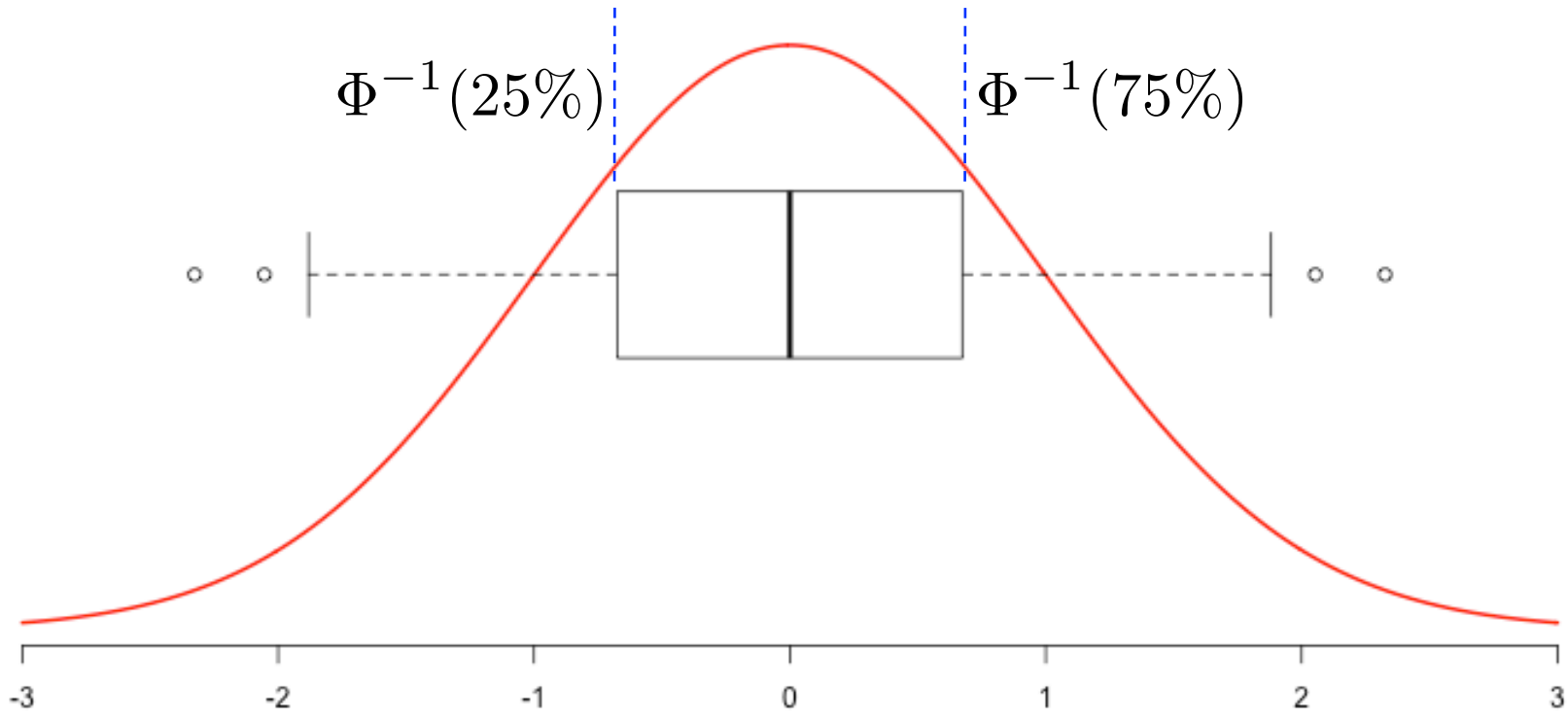
Quantiles of the $\mathcal{N}(0,1)$ distribution



Quantiles of the $\mathcal{N}(0,1)$ distribution

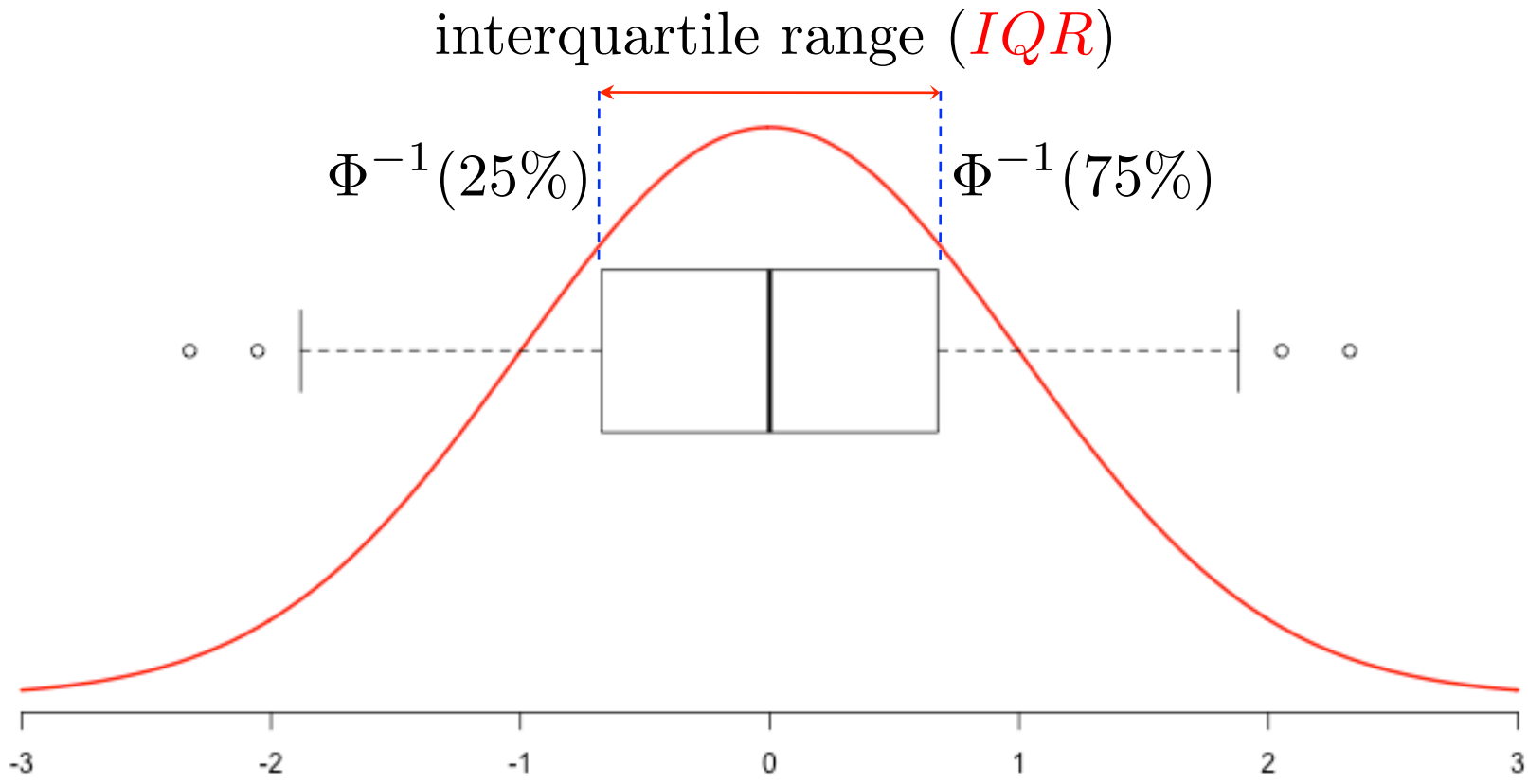


Quantiles of the $\mathcal{N}(0,1)$ distribution



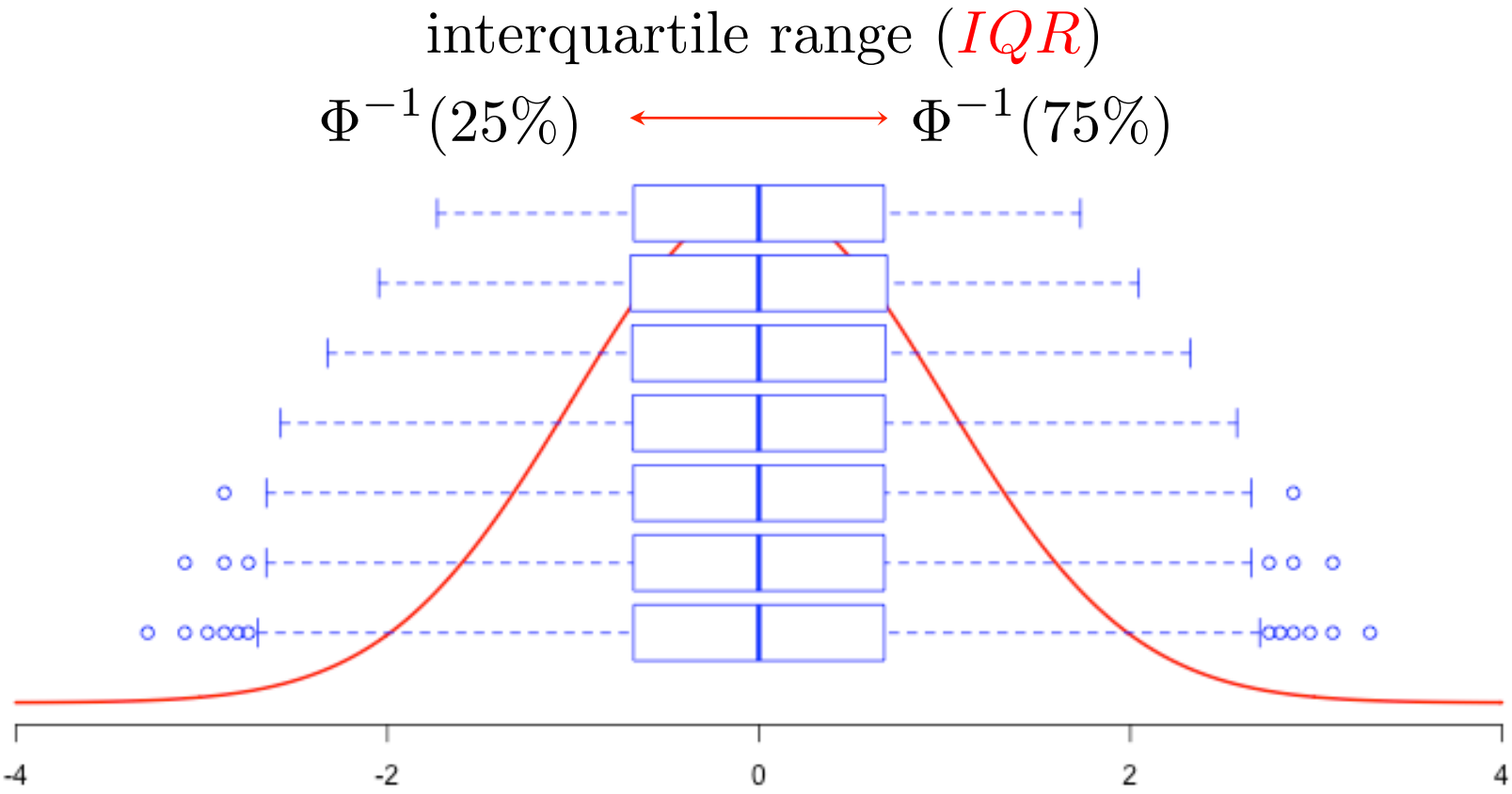
the ends of the whiskers can represent several possible alternative values

Quantiles of the $\mathcal{N}(0,1)$ distribution



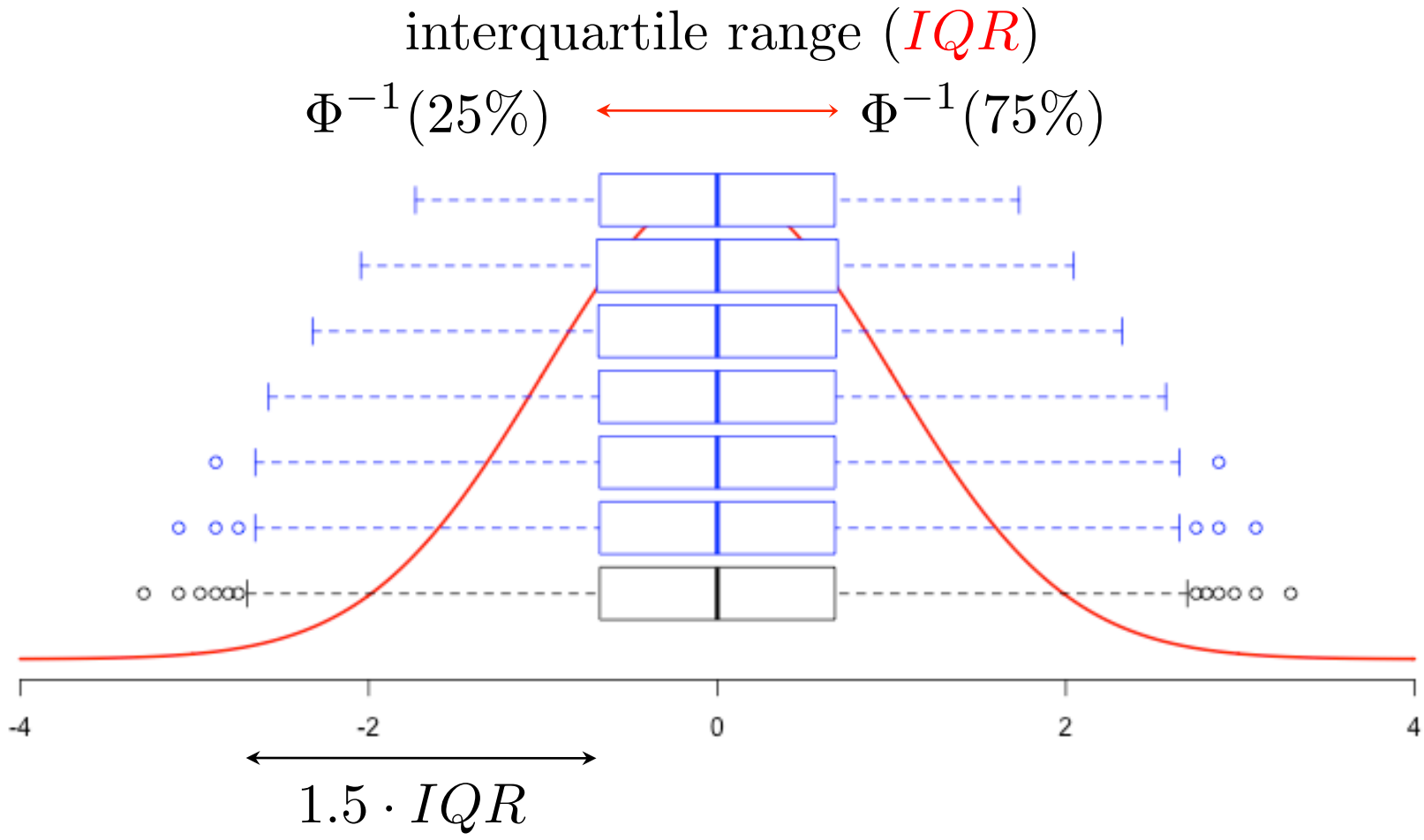
the ends of the whiskers can represent several possible alternative values

Boxplots with R



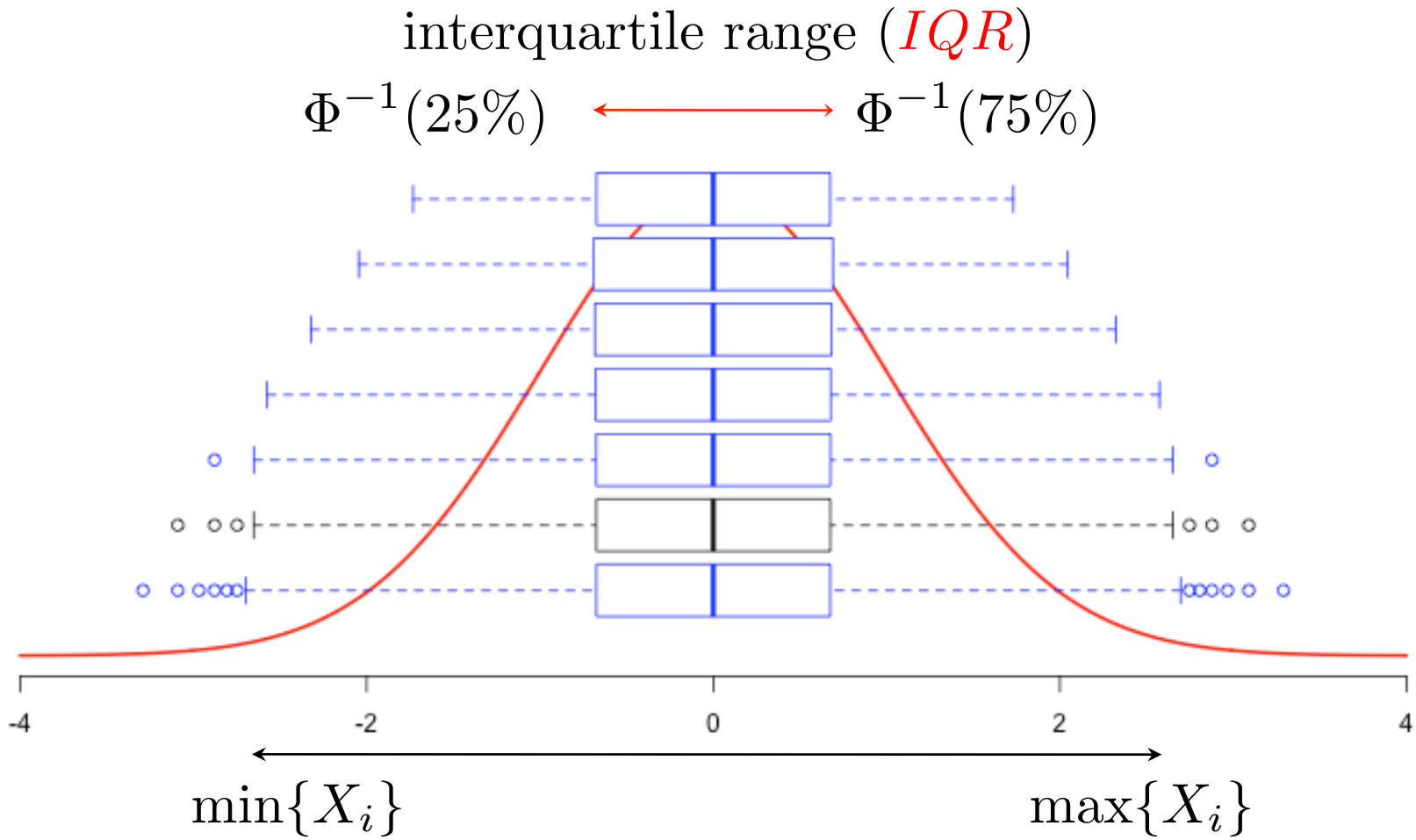

```
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R



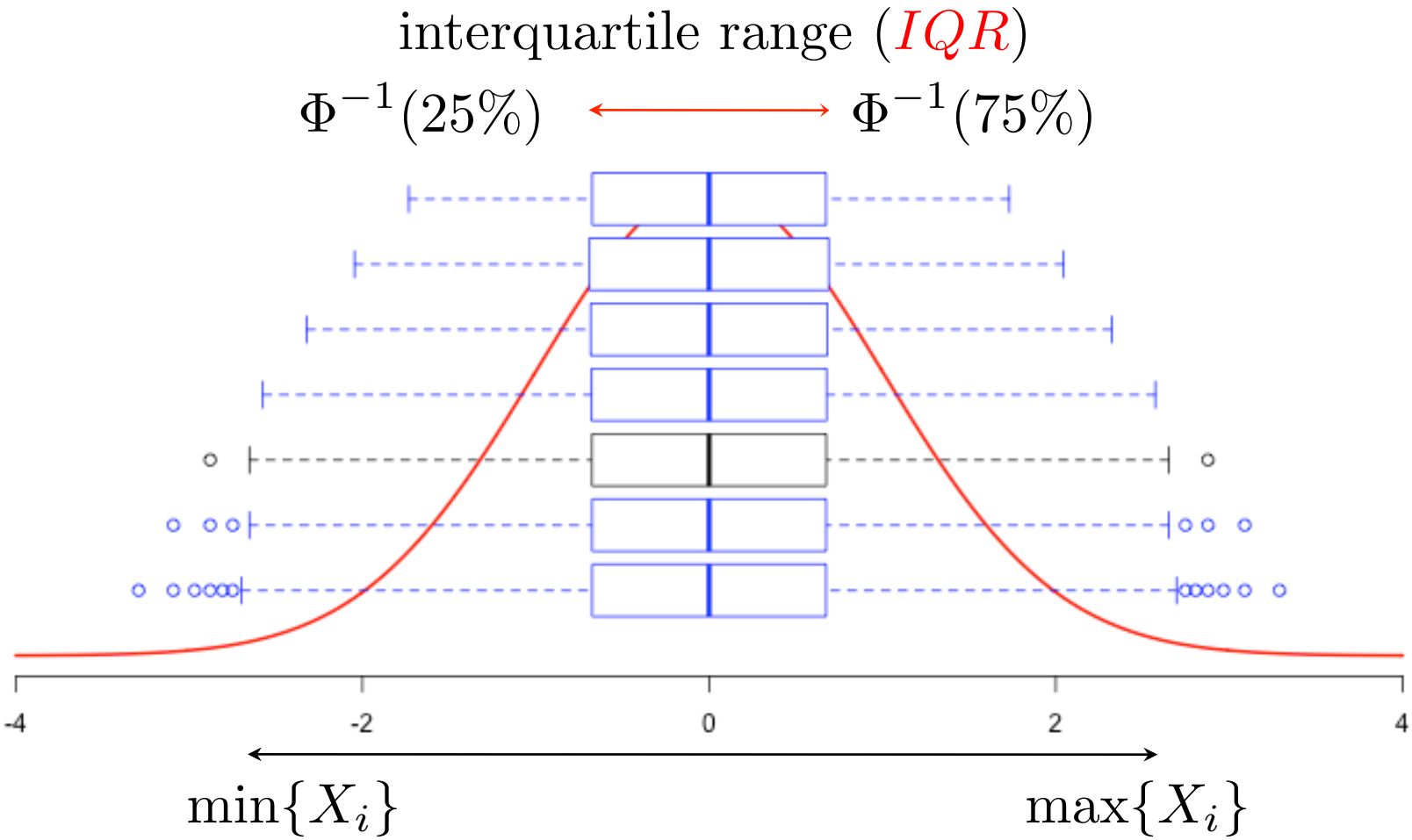

```
> n=2000
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R



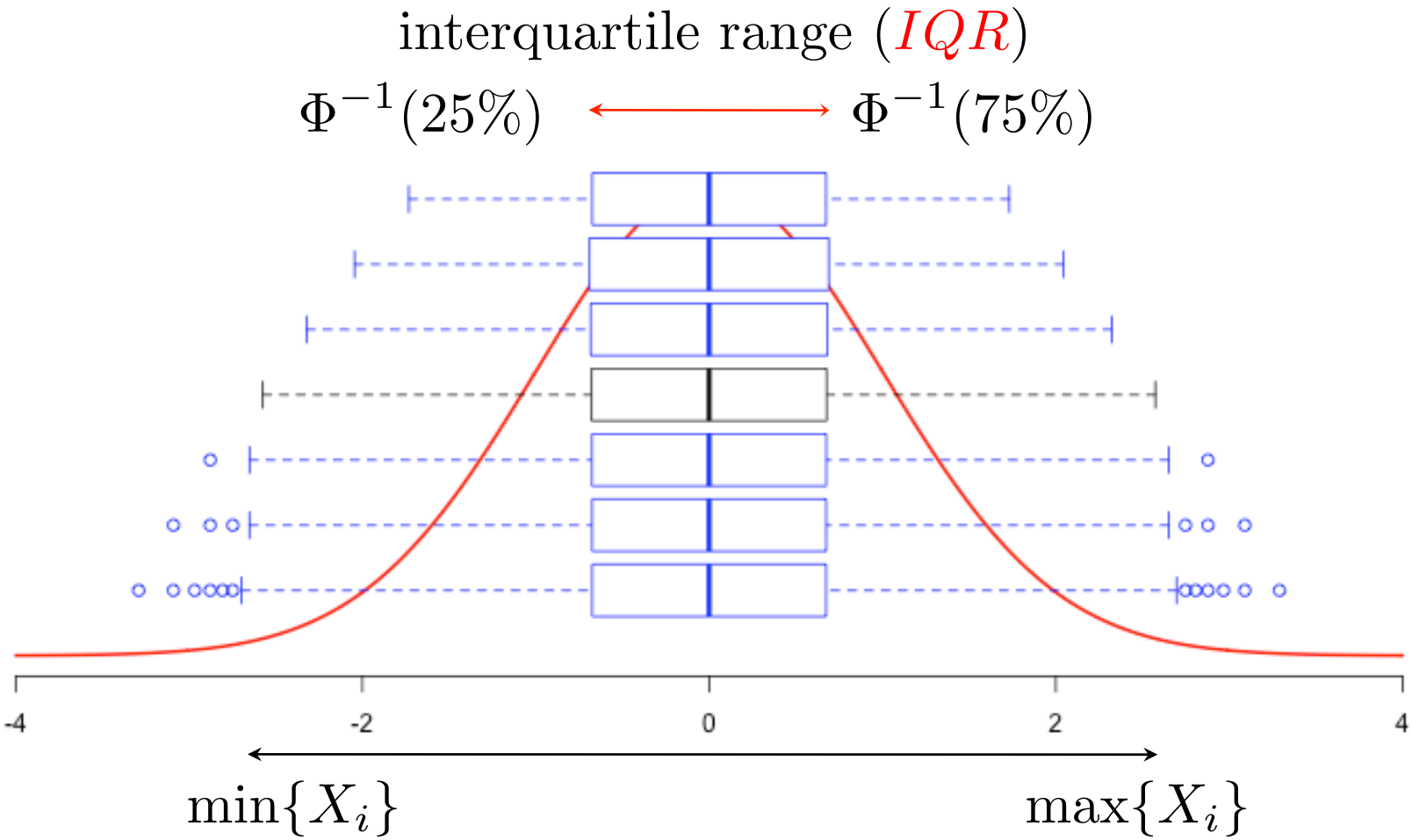

```
> n=1000
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R



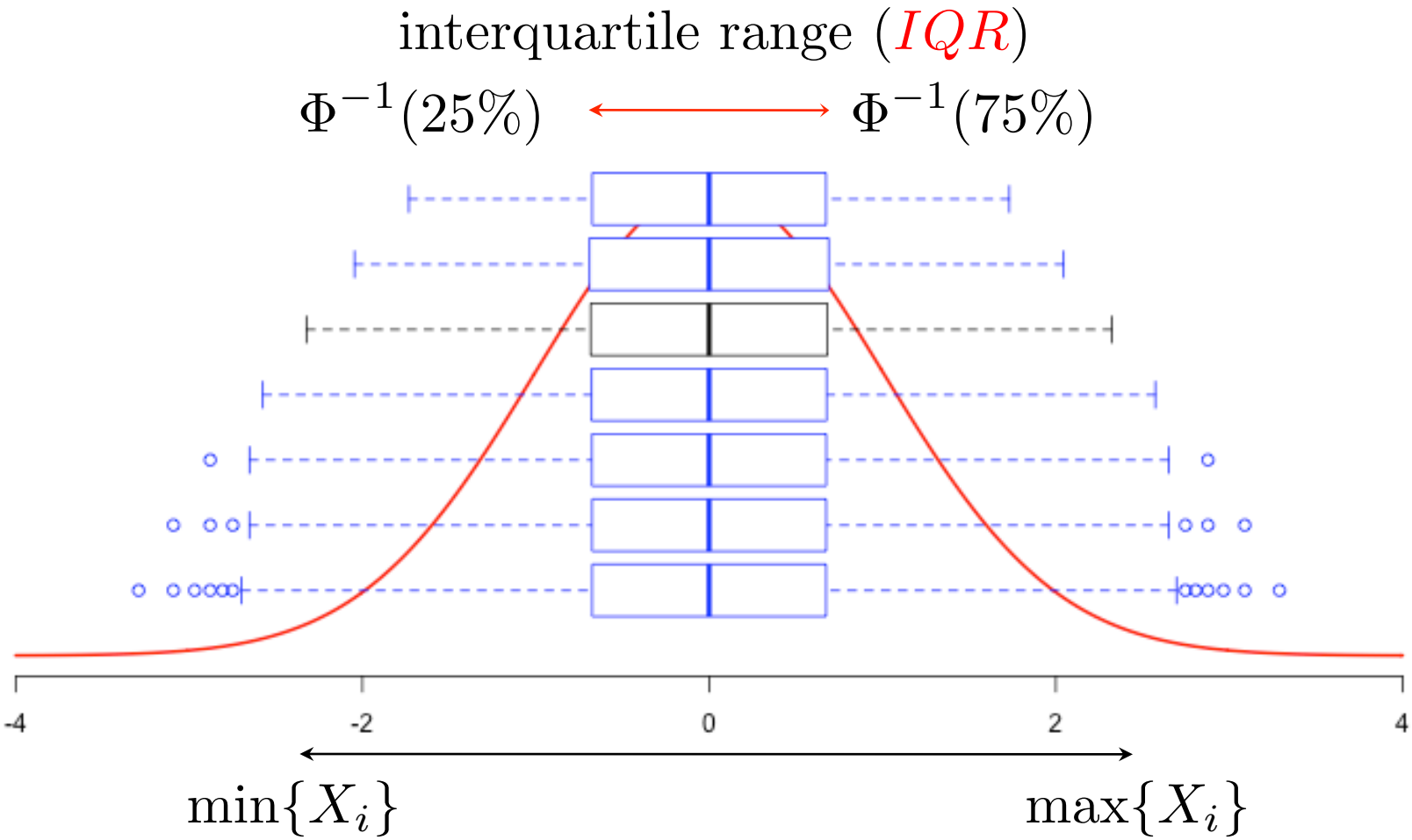

```
> n=500
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R



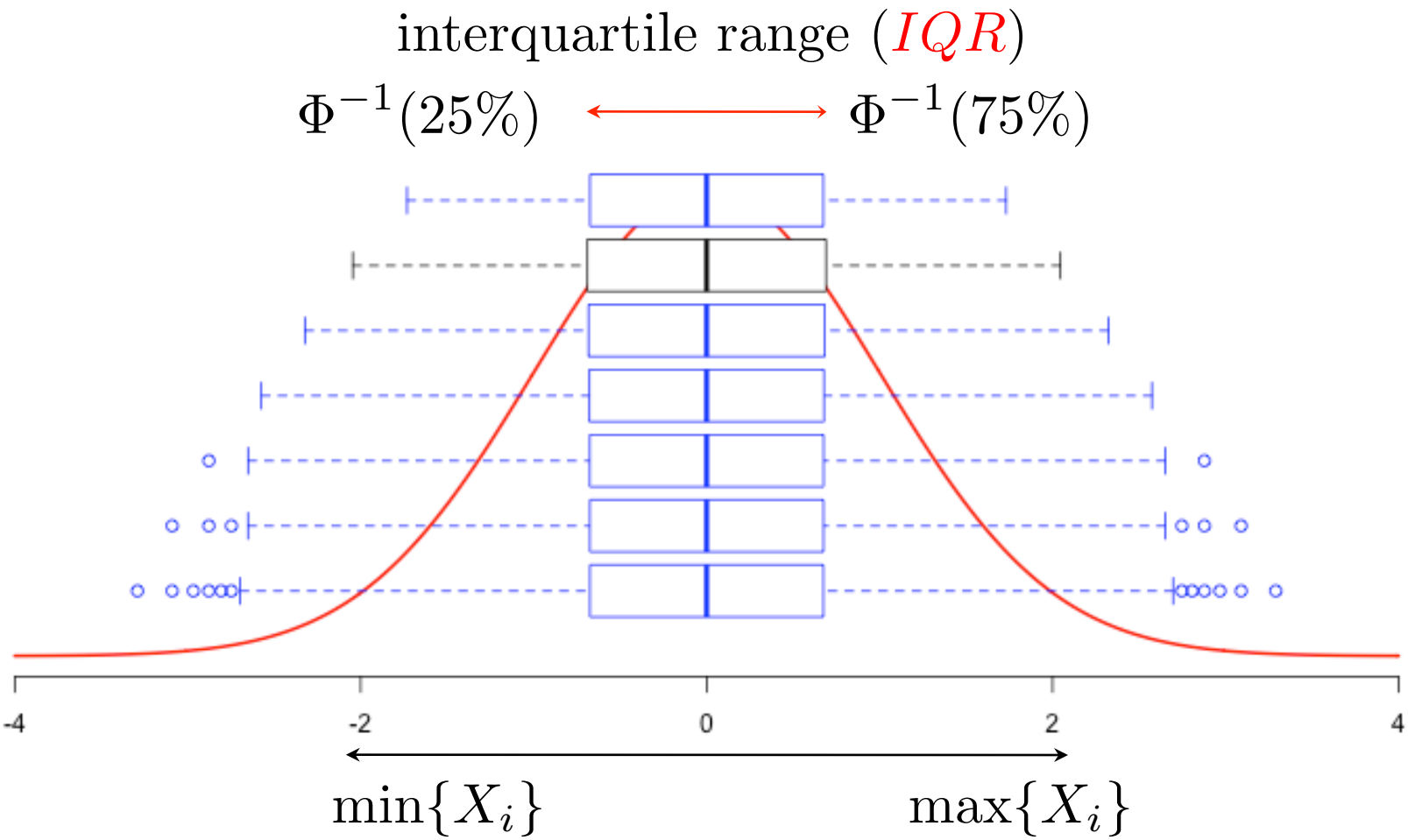

```
> n=200
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R



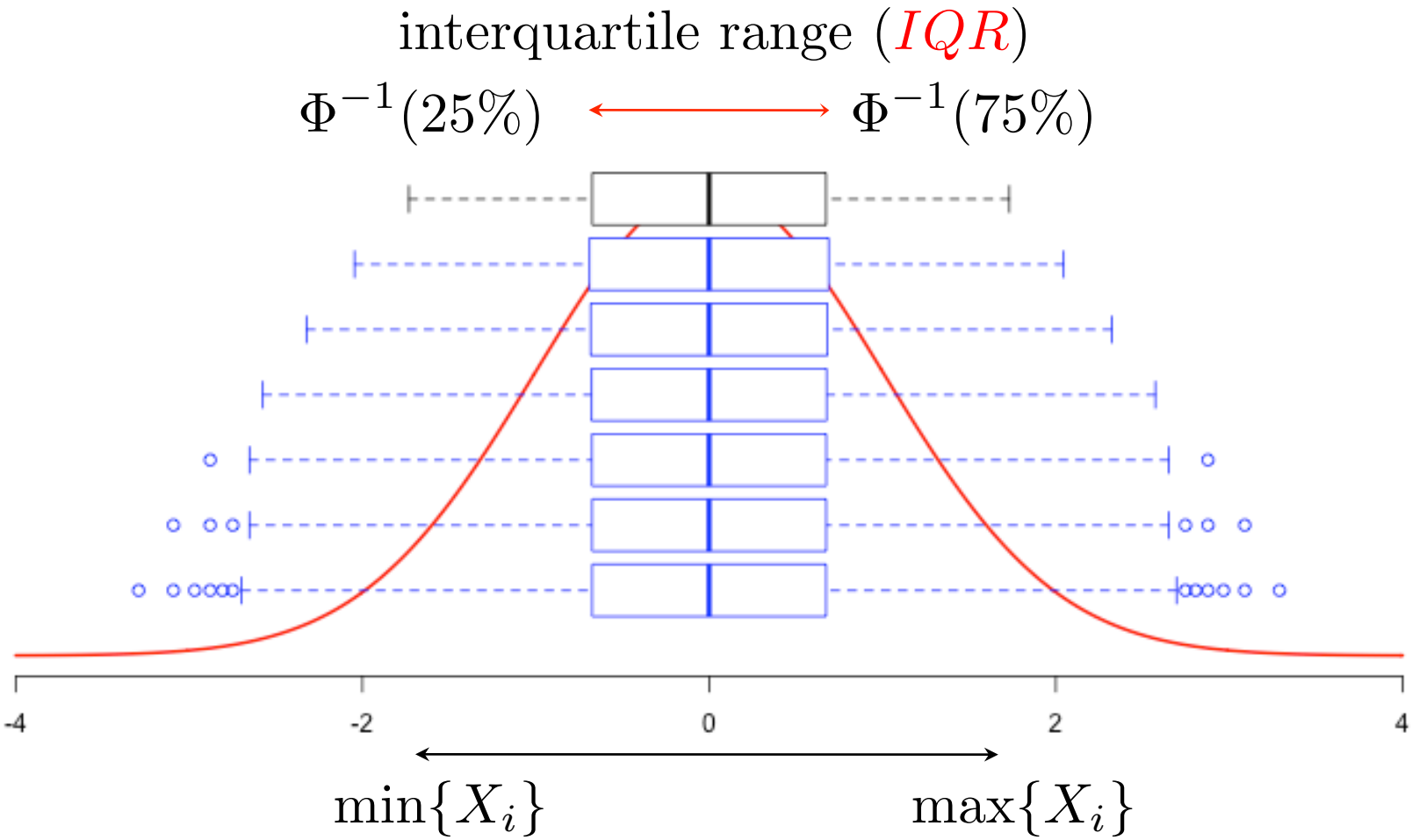

```
> n=100
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with 




```
> n=50
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R

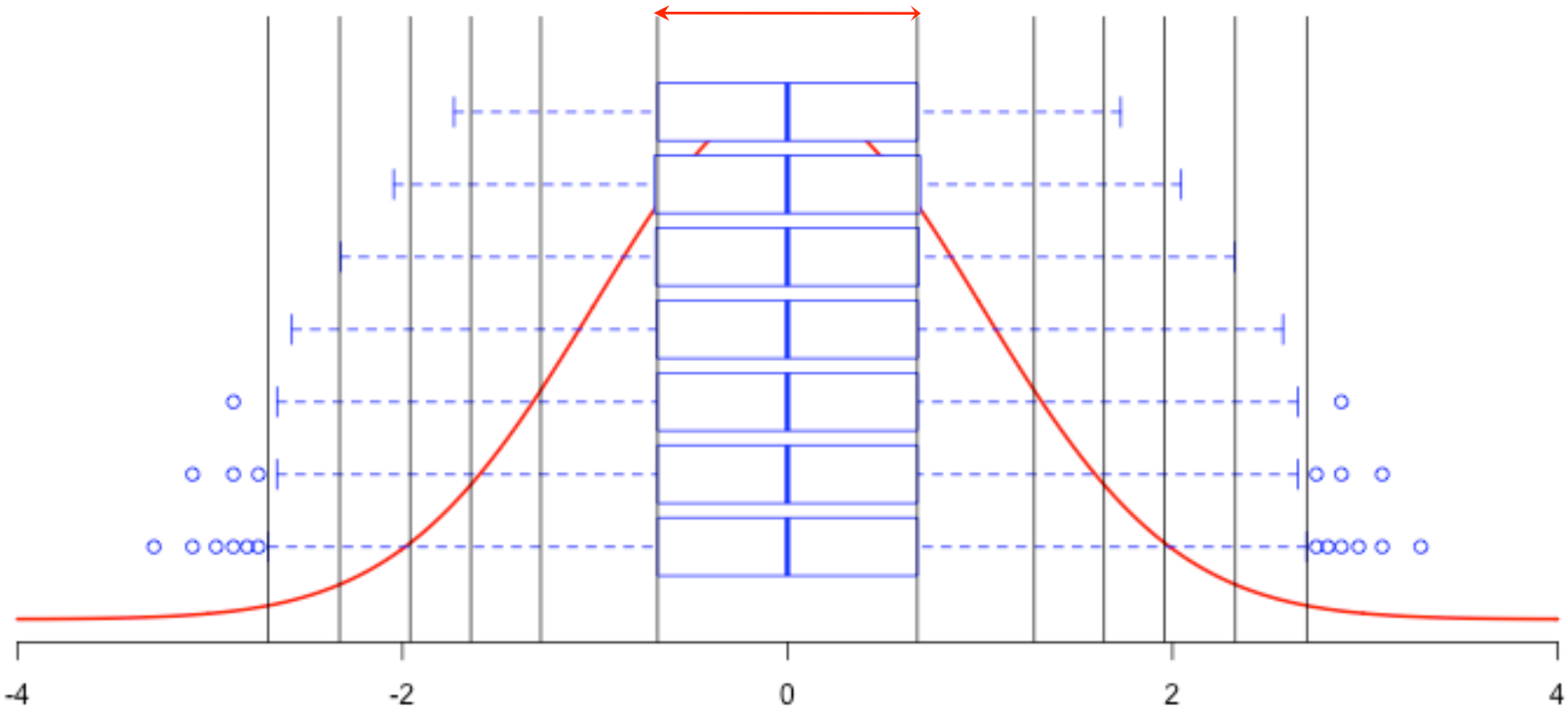



```
> n=20
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R

interquartile range (IQR)

$\Phi^{-1}(25\%)$ $\Phi^{-1}(75\%)$

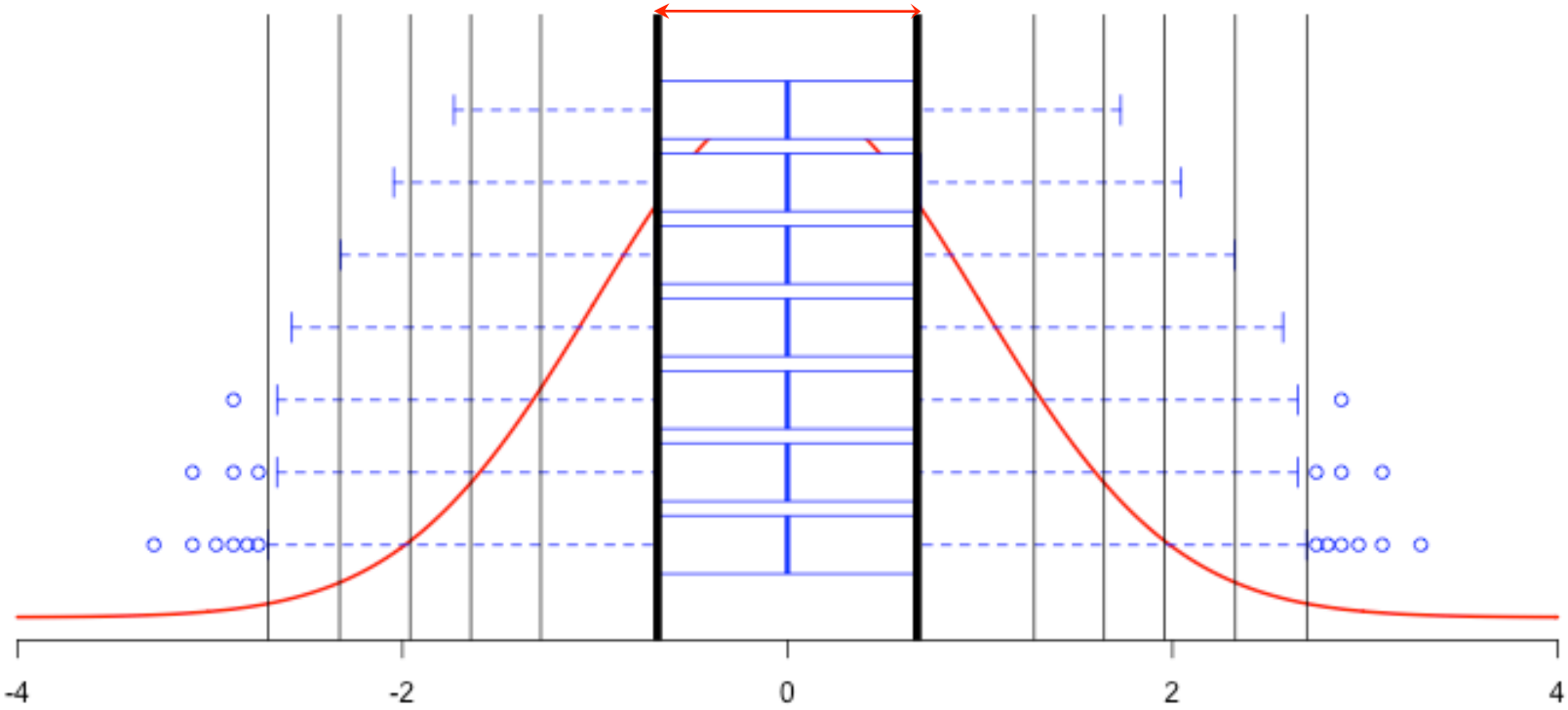



```
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R

interquartile range (IQR)

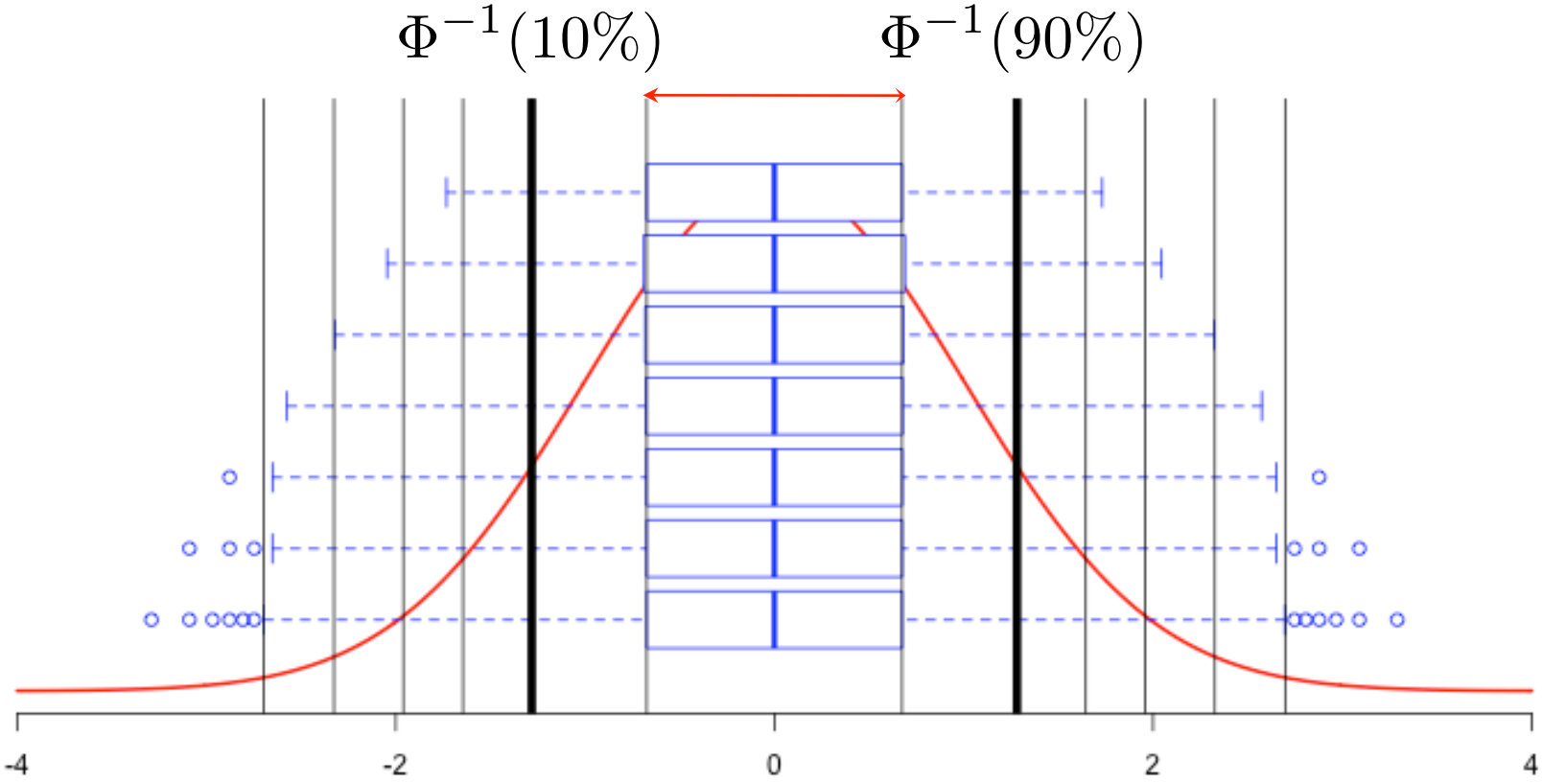
$\Phi^{-1}(25\%)$ $\Phi^{-1}(75\%)$




```
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R

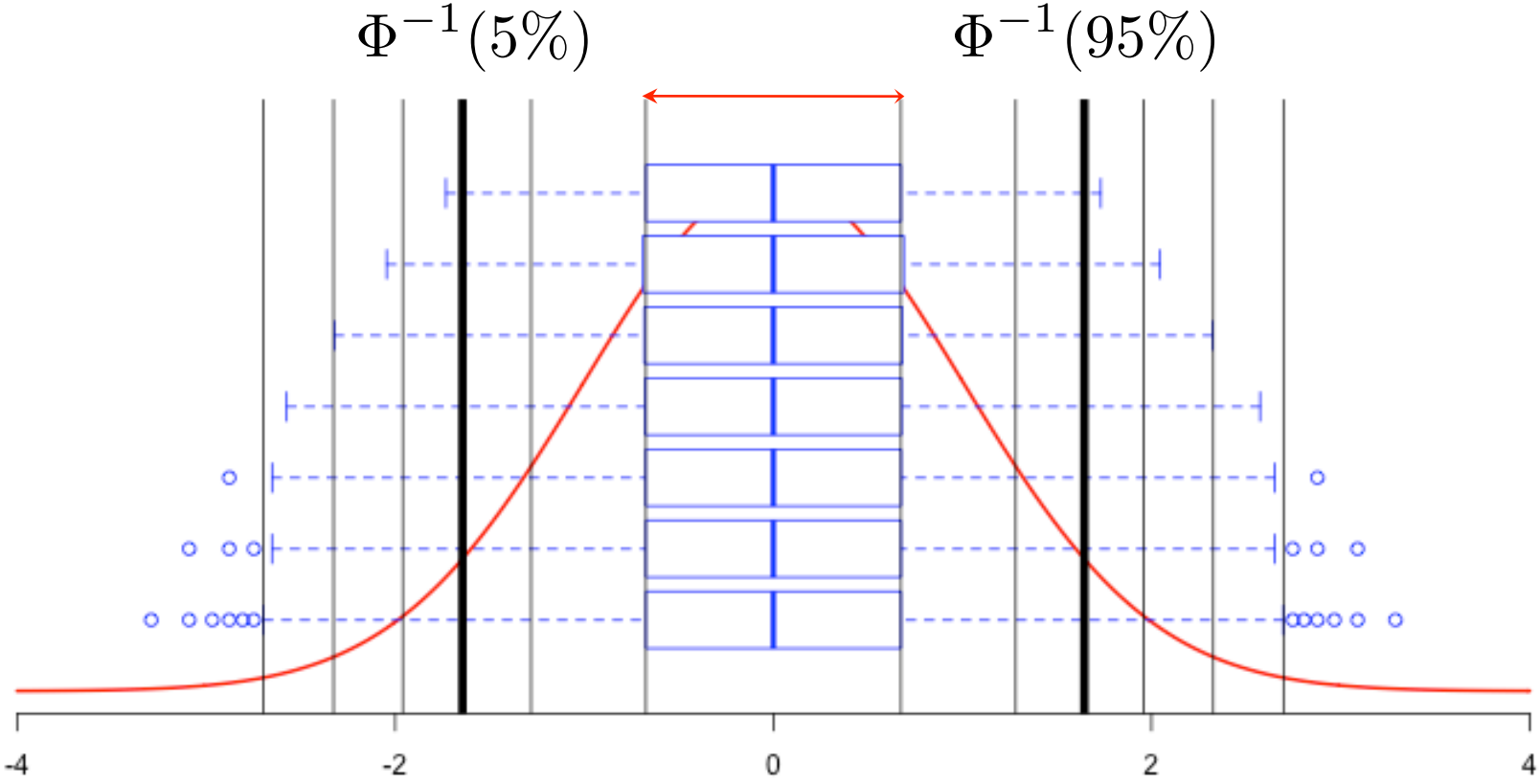
interquartile range (IQR)




```
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with R

interquartile range (IQR)

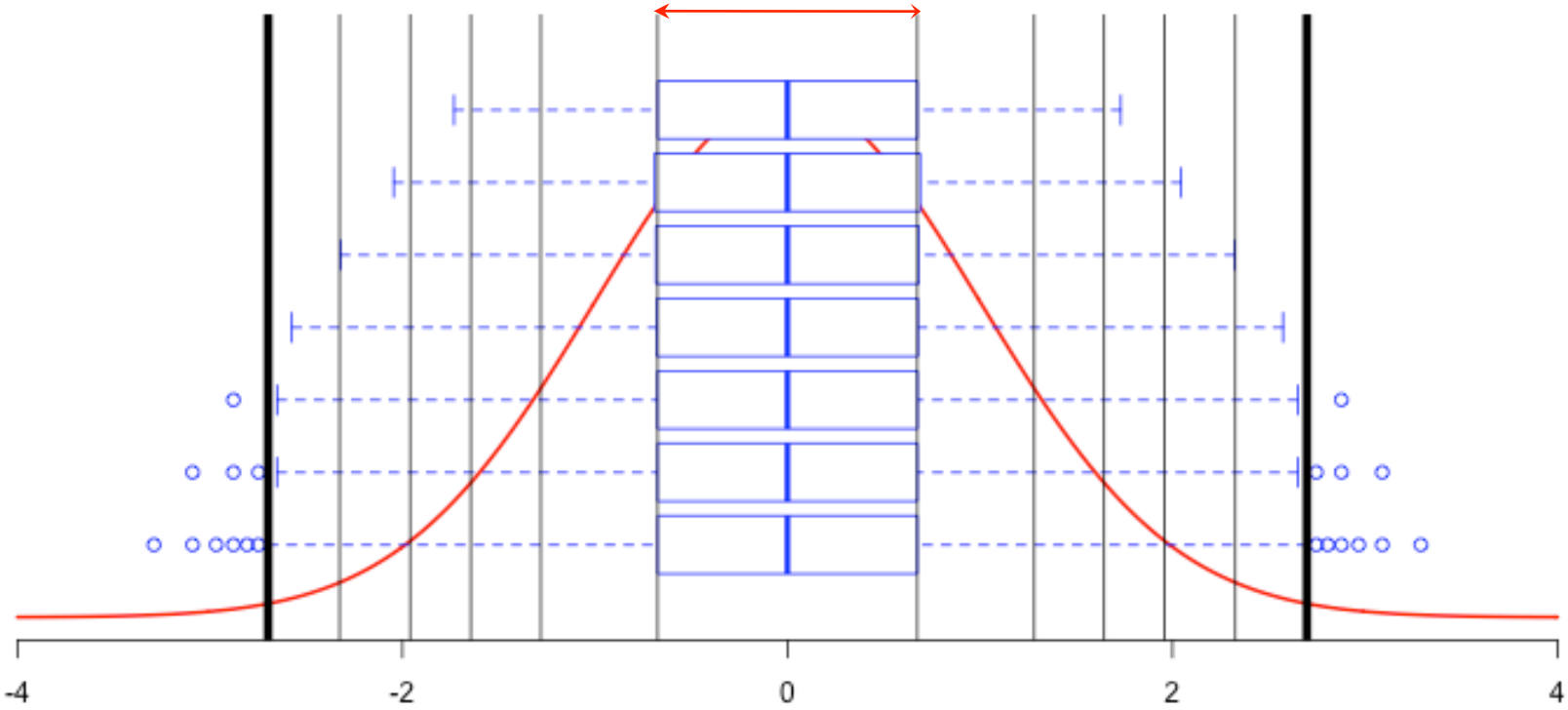



```
> boxplot(qnorm(seq(0,1,length=n)))
```

Boxplots with 

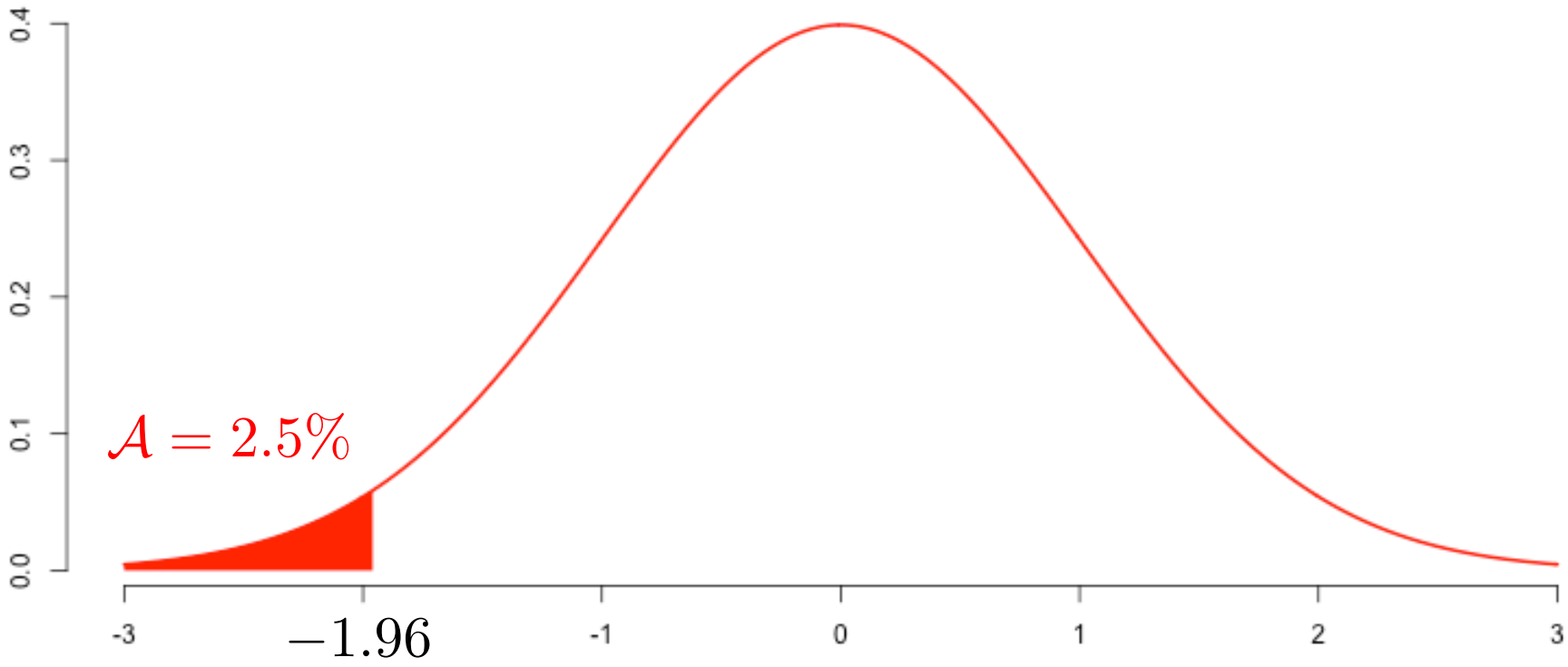
interquartile range (IQR)

$\Phi^{-1}(25\%) - 1.5 \cdot IQR$ $\Phi^{-1}(75\%) + 1.5 \cdot IQR$

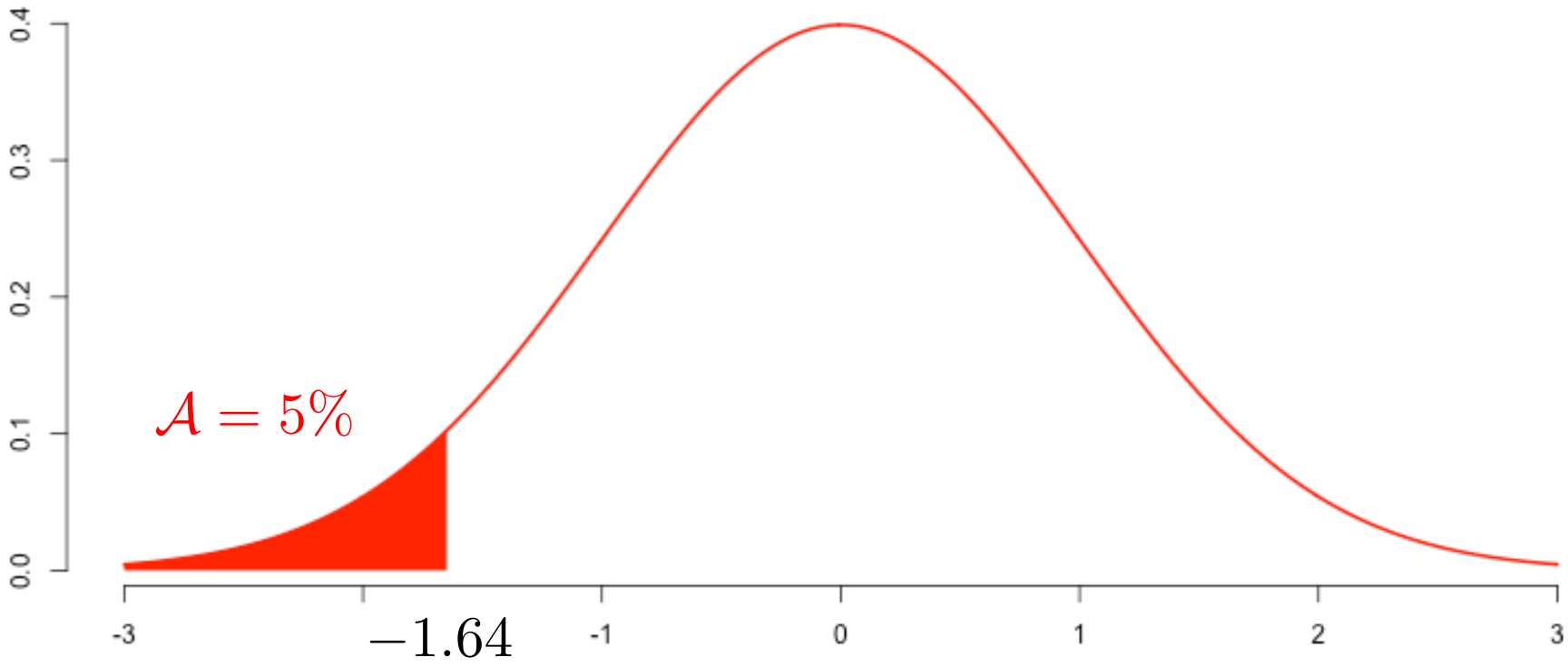



```
> boxplot(qnorm(seq(0,1,length=n)))
```

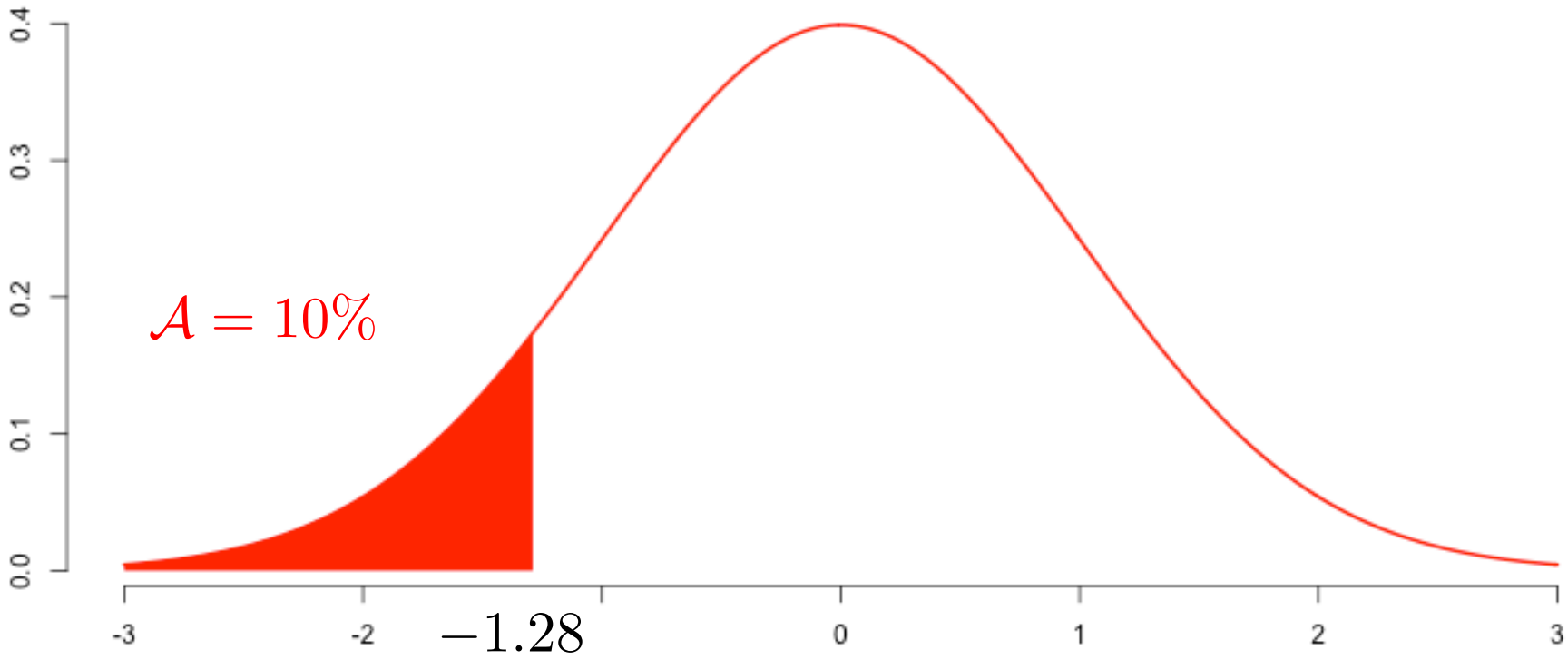
Quantiles of the $\mathcal{N}(0,1)$ distribution



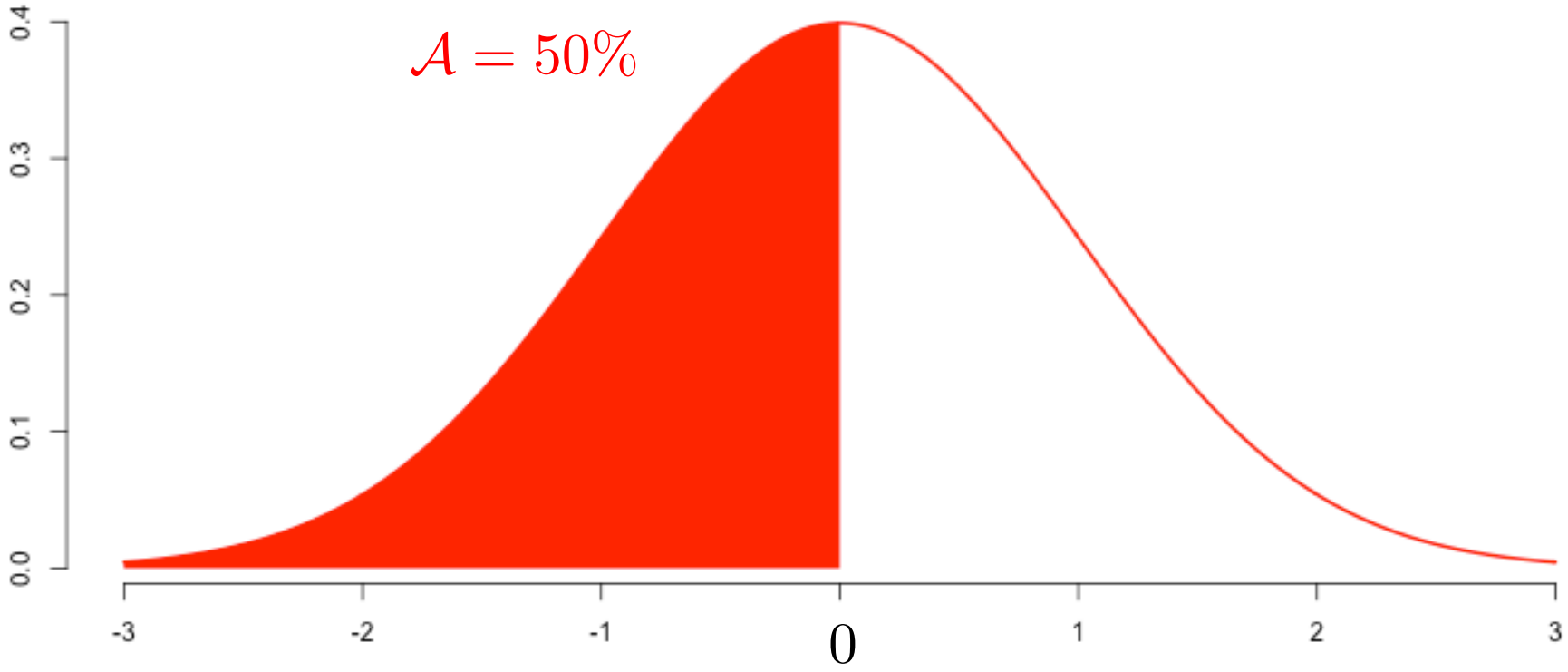
Quantiles of the $\mathcal{N}(0,1)$ distribution



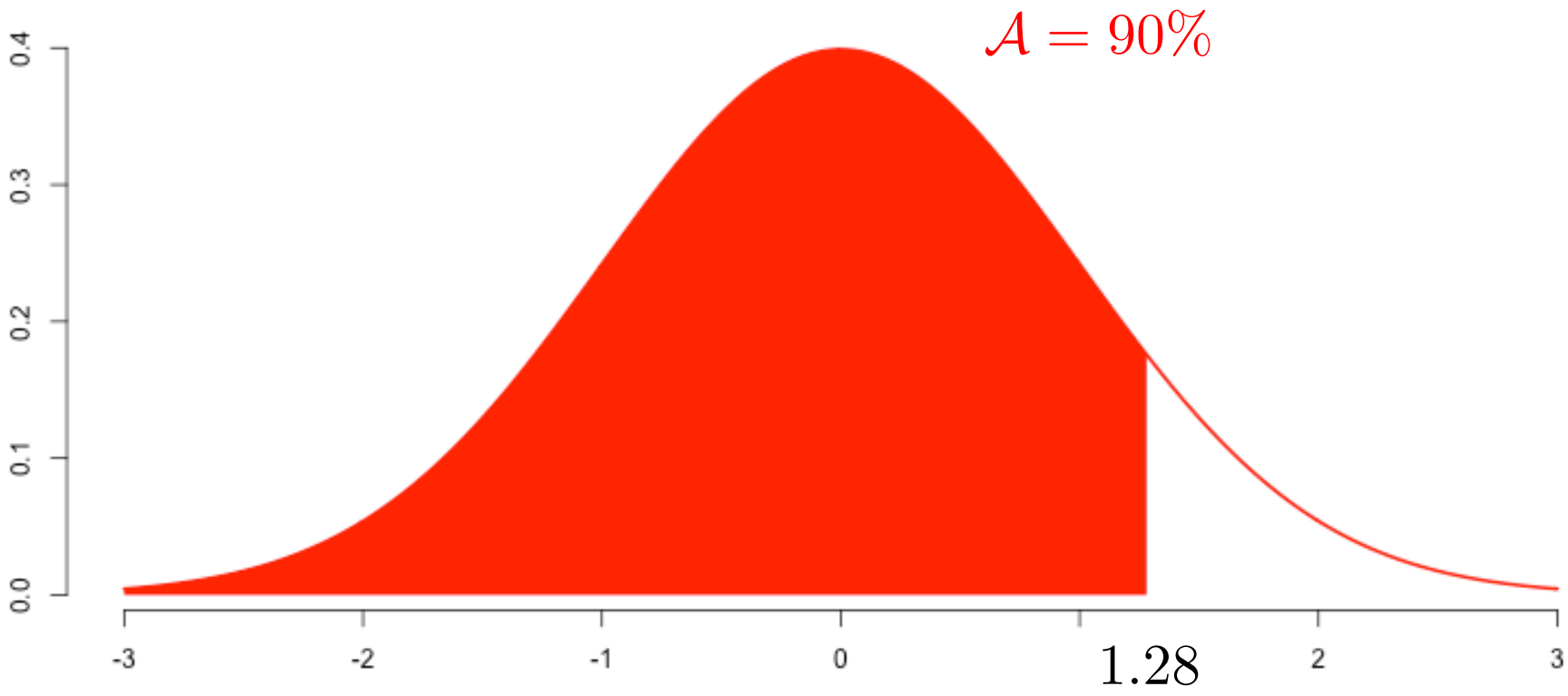
Quantiles of the $\mathcal{N}(0,1)$ distribution



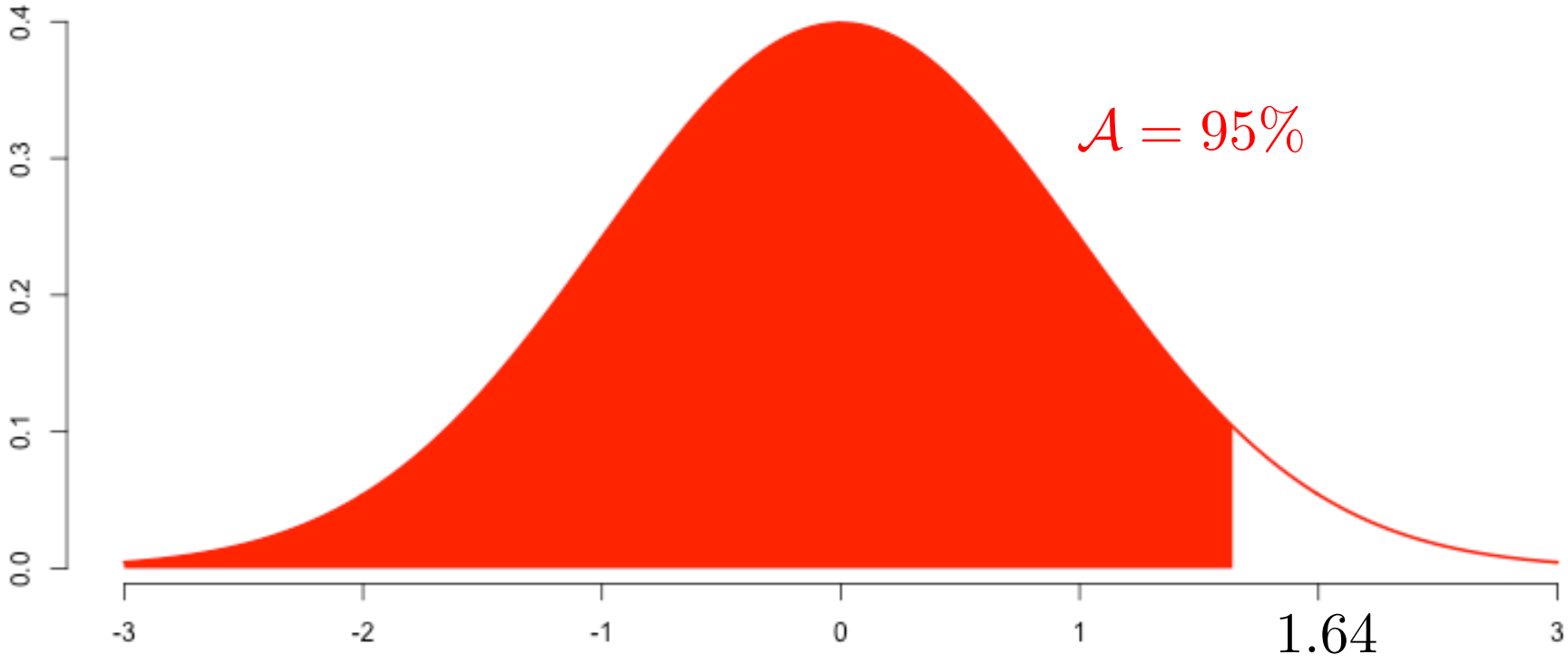
Quantiles of the $\mathcal{N}(0,1)$ distribution



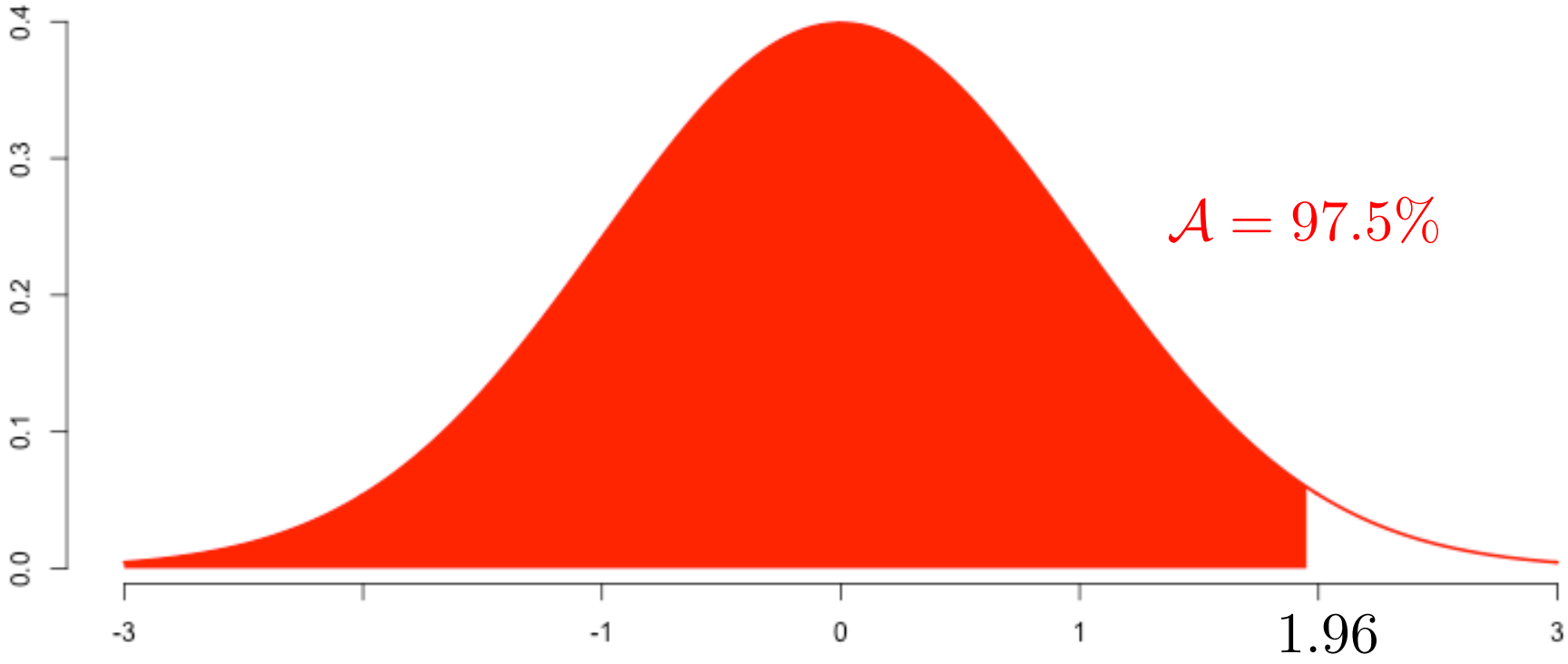
Quantiles of the $\mathcal{N}(0,1)$ distribution



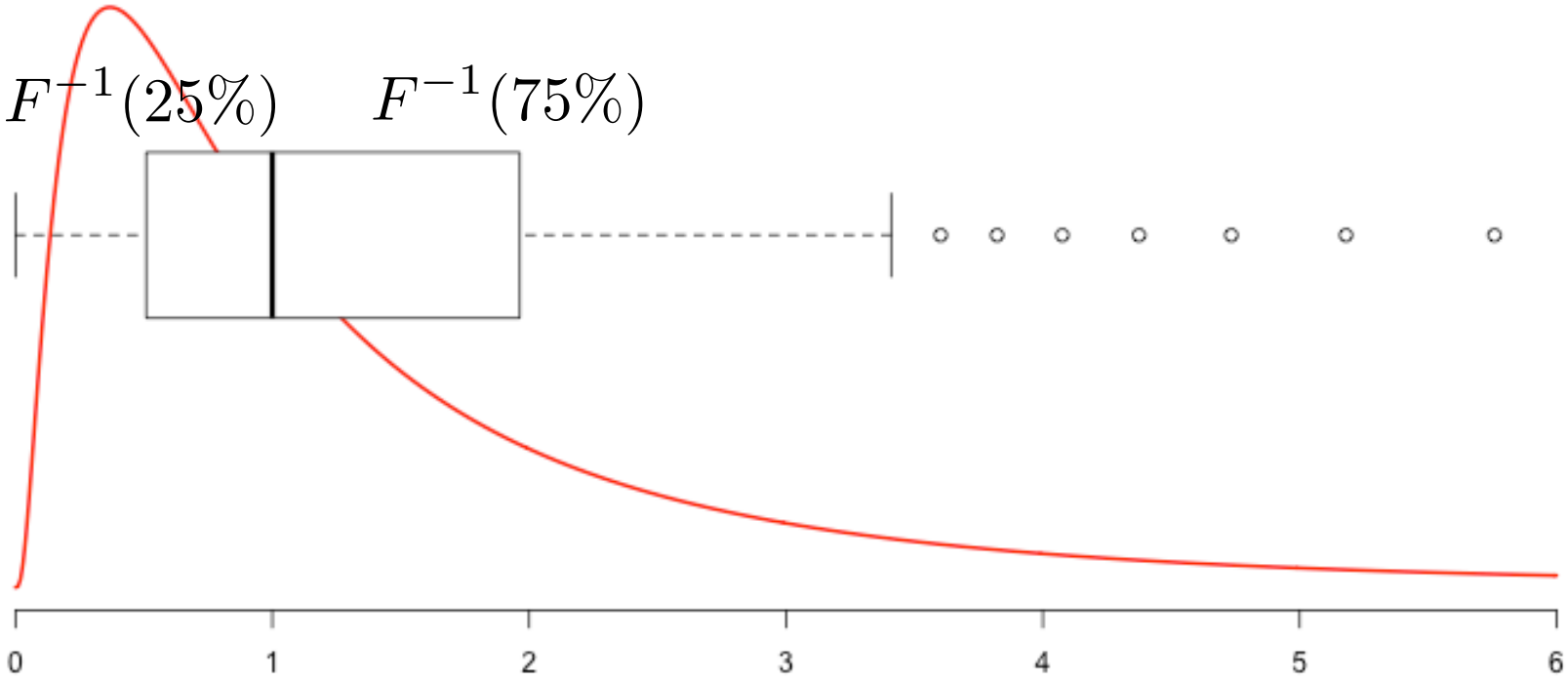
Quantiles of the $\mathcal{N}(0,1)$ distribution



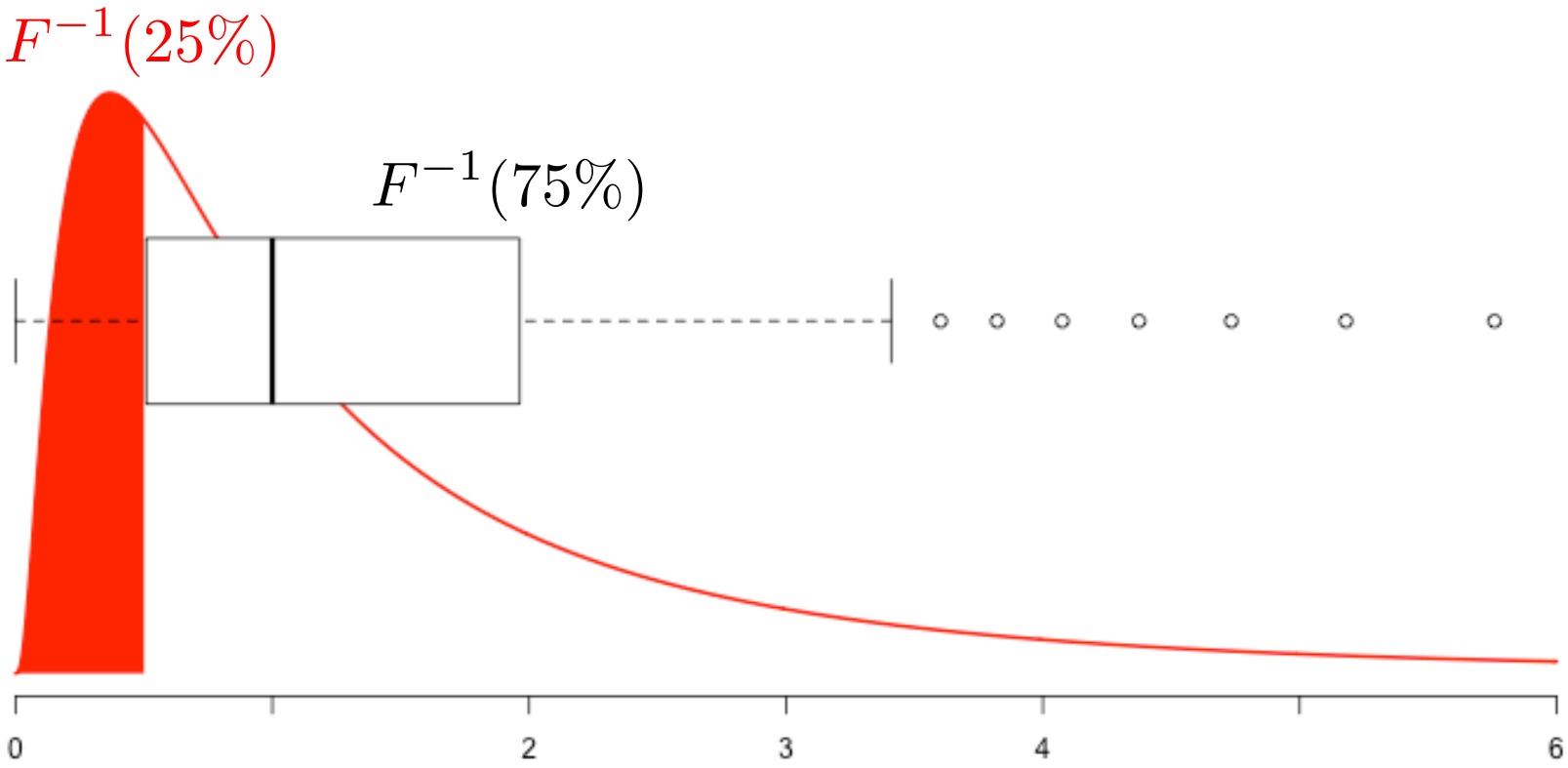
Quantiles of the $\mathcal{N}(0,1)$ distribution



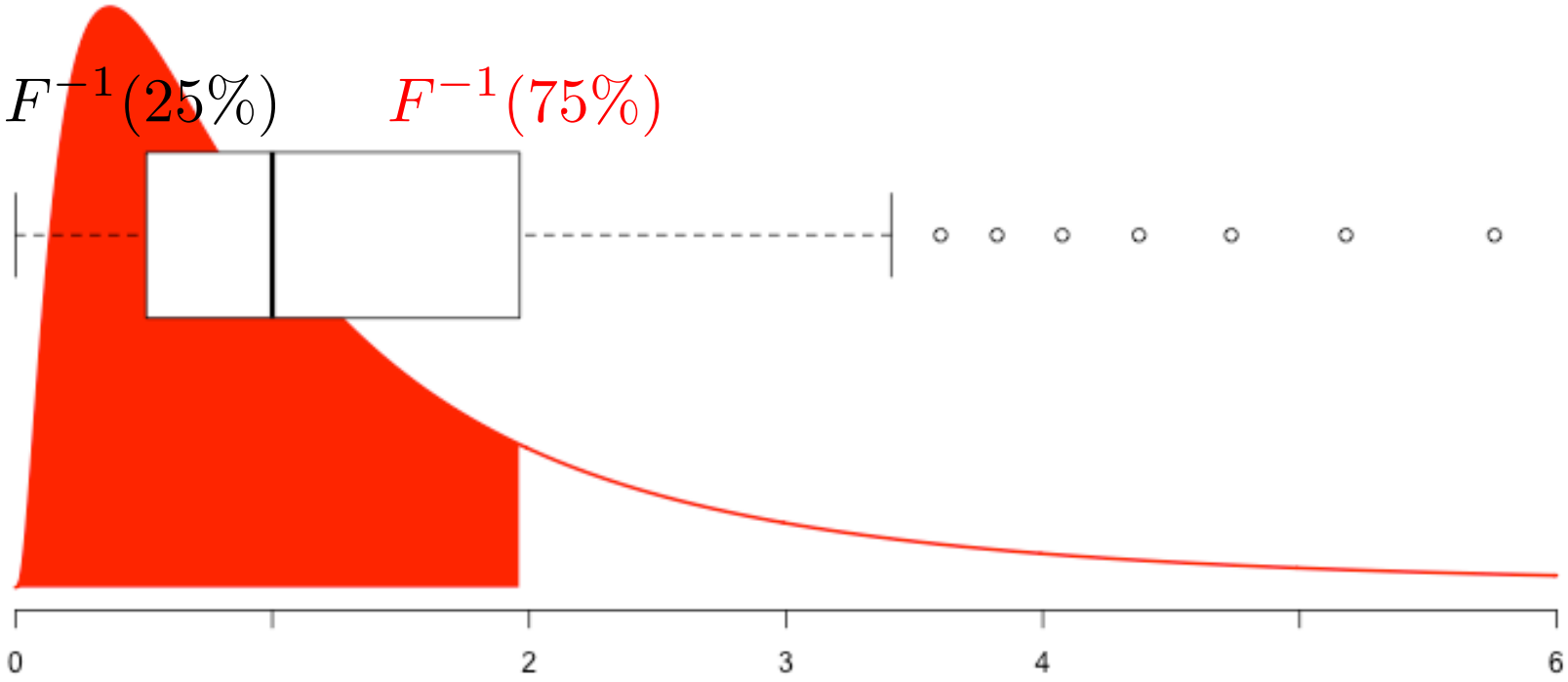
Quantiles and boxplot



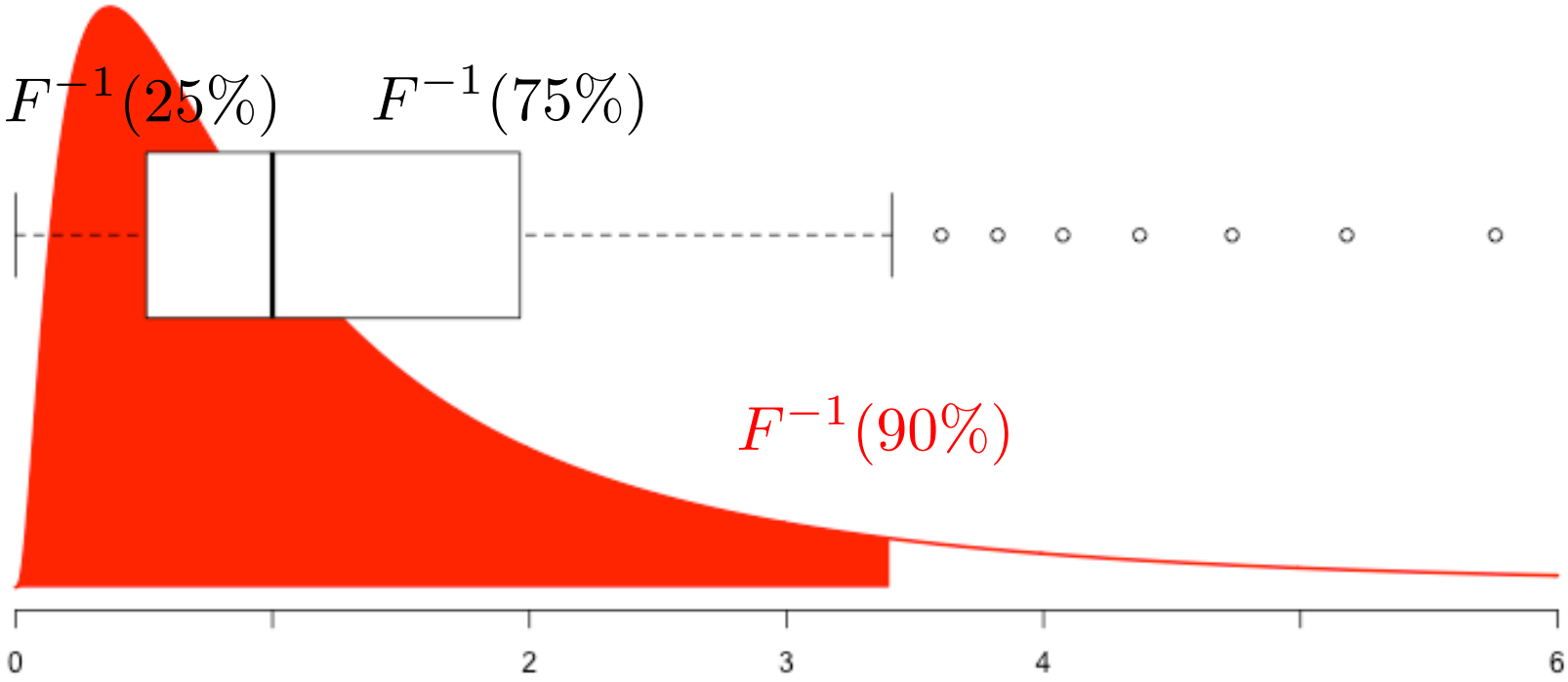
Quantiles and boxplot



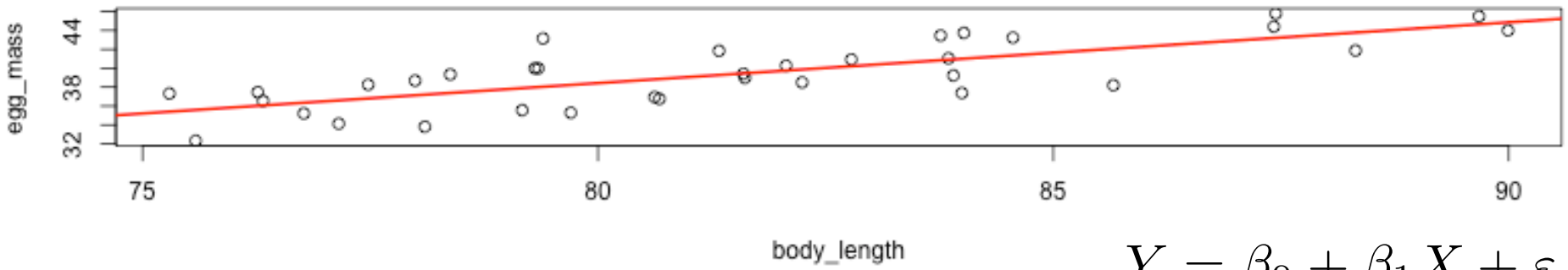
Quantiles and boxplot



Quantiles and boxplot



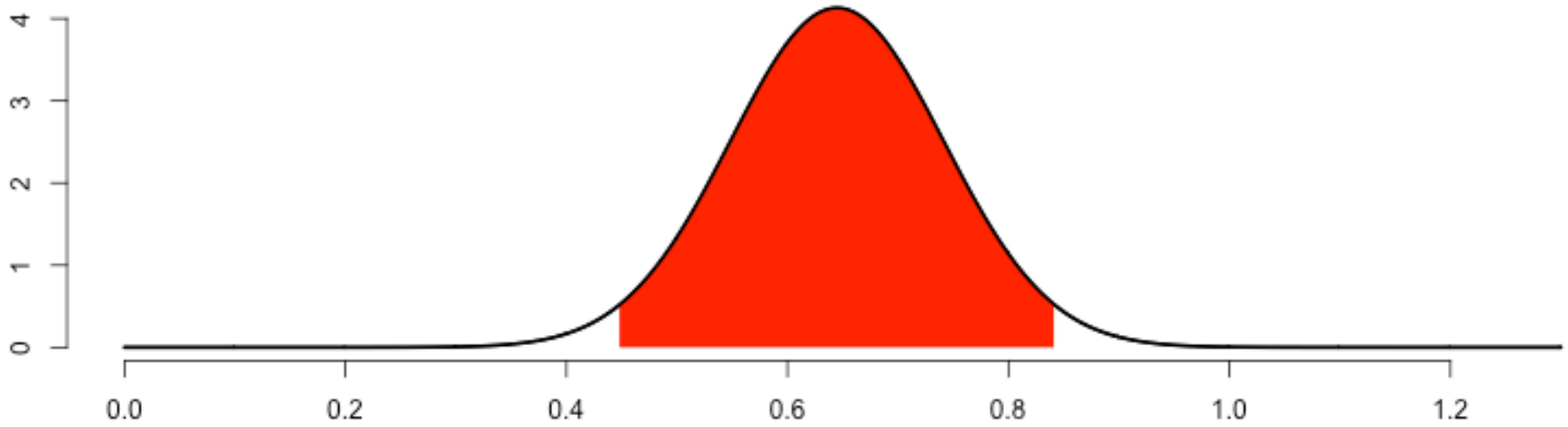
Confidence intervals and confidence regions



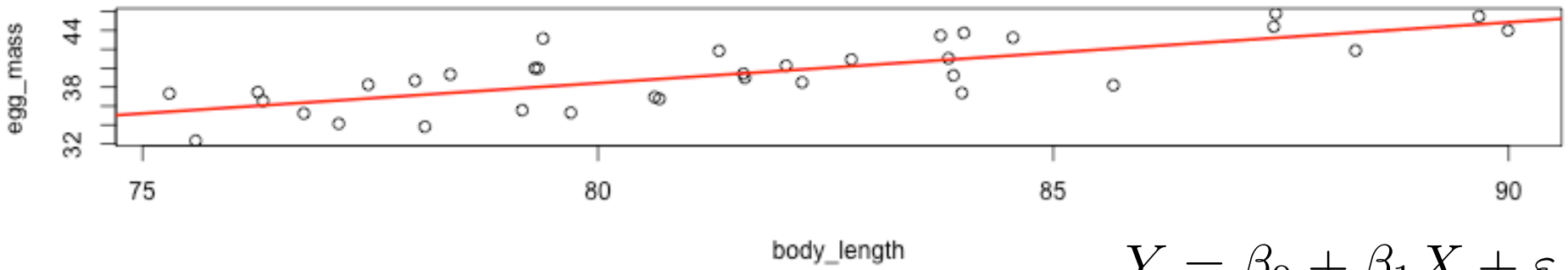
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

```
> model=lm(egg_mass ~ body_length)
> confint(model)
                    2.5 %    97.5 %
(Intercept)    -29.1863825  2.880614
body_length      0.4483009  0.841390
```

$$\mathbb{P}(\beta_1 \in CI) = 95\%$$



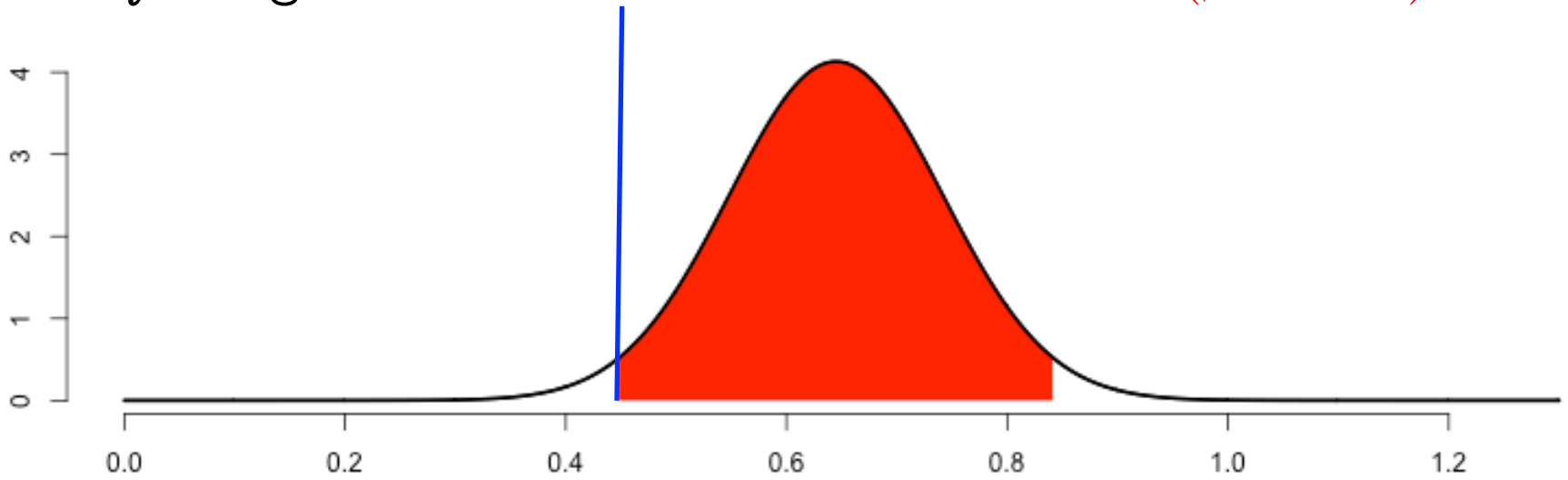
Confidence intervals and confidence regions



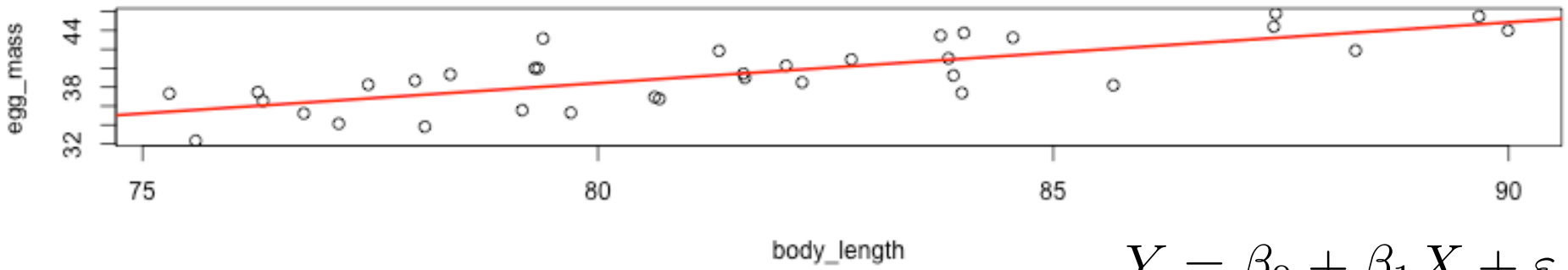
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

```
> model=lm(egg_mass ~ body_length)
> confint(model)
                    2.5 %    97.5 %
(Intercept)    -29.1863825  2.880614
body_length      0.4483009  0.841390
```

$$\mathbb{P}(\beta_1 \in CI) = 95\%$$



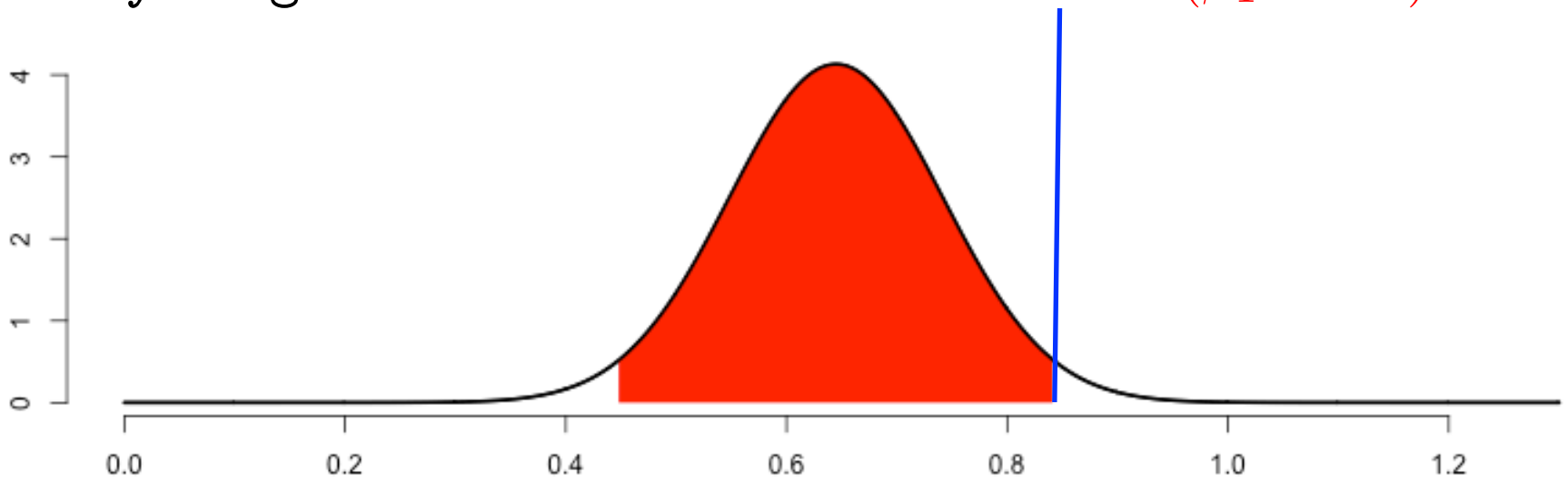
Confidence intervals and confidence regions



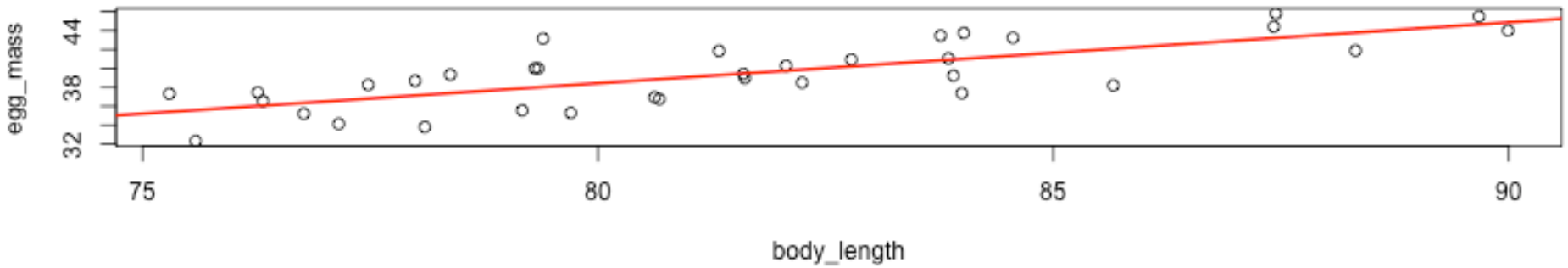
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

```
> model=lm(egg_mass ~ body_length)
> confint(model)
                    2.5 %    97.5 %
(Intercept)   -29.1863825  2.880614
body_length     0.4483009  0.841390
```

$$\mathbb{P}(\beta_1 \in CI) = 95\%$$

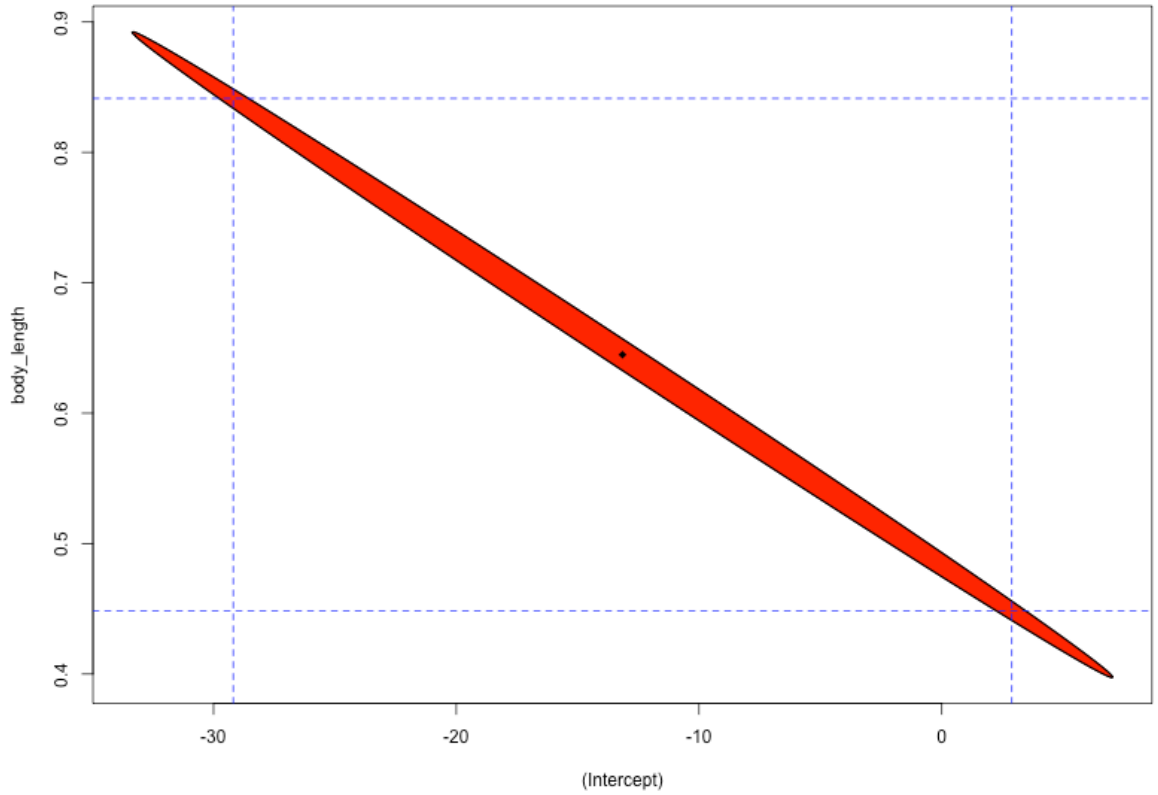


Confidence intervals and confidence regions



$\mathbb{P}((\beta_0, \beta_1) \in CR) = 95\%$

```
> ellipse(model,1:2)
```

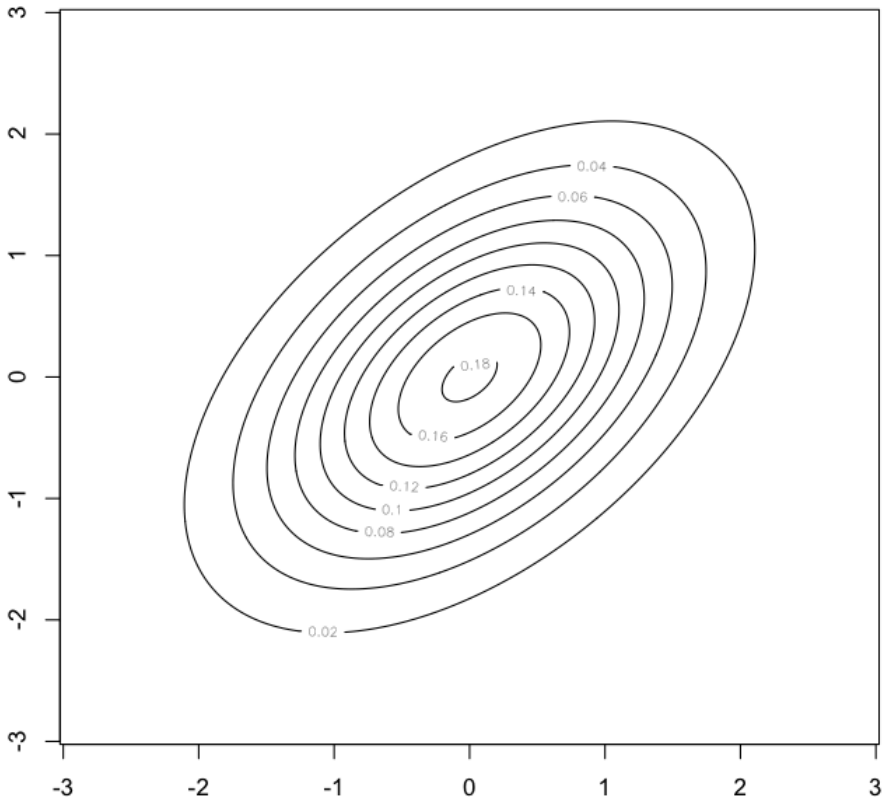


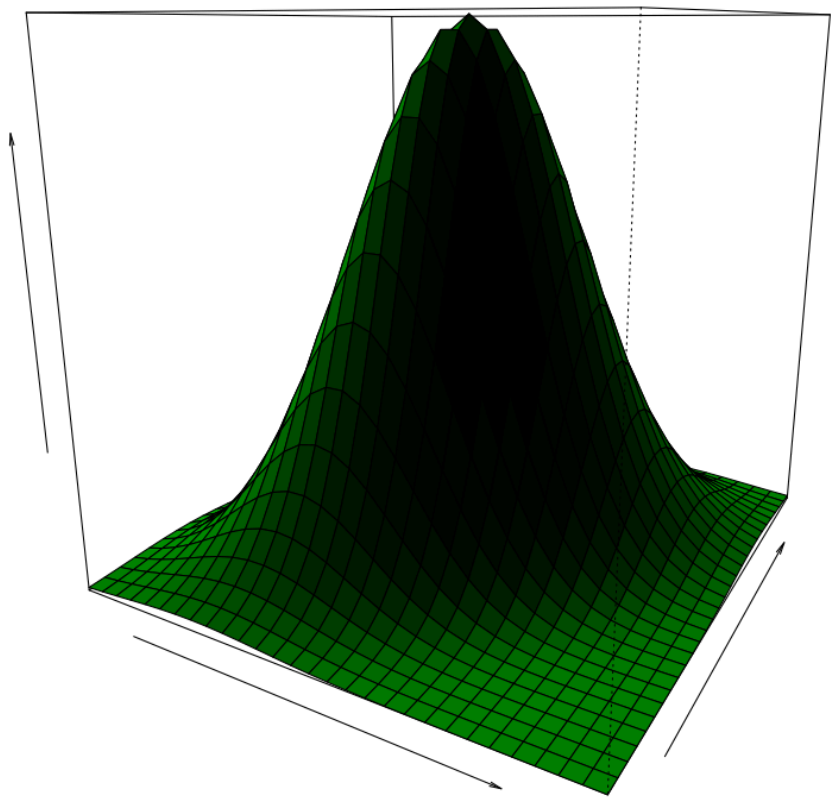
Why ellipses of confidence ?

$$\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \Sigma\right)$$

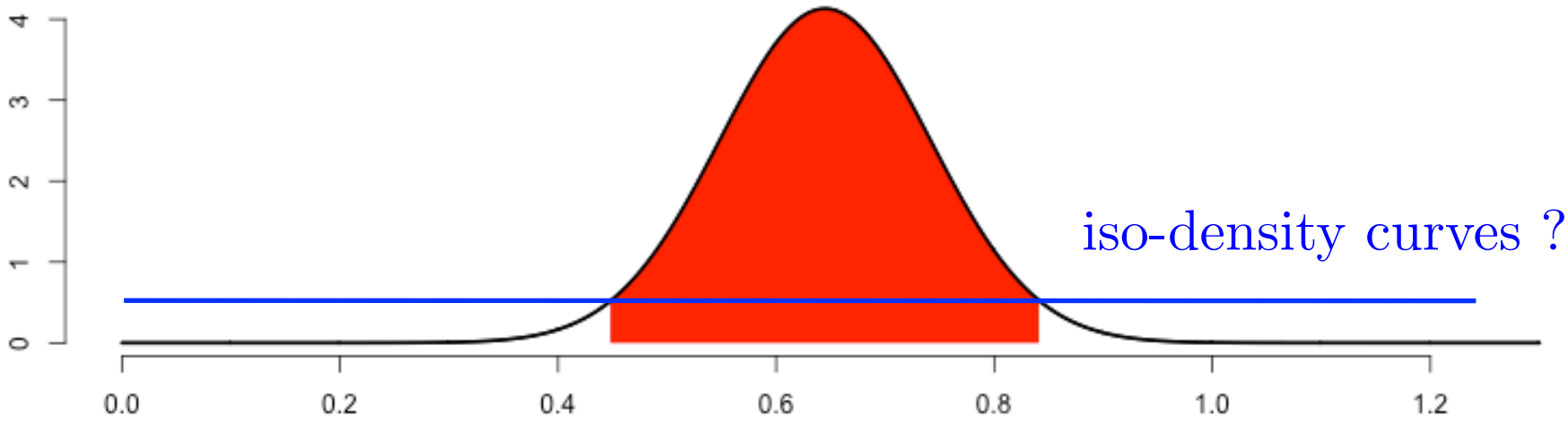
iso-density curves

$$\phi(\boldsymbol{x}) = (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$





Why ellipses of confidence ?

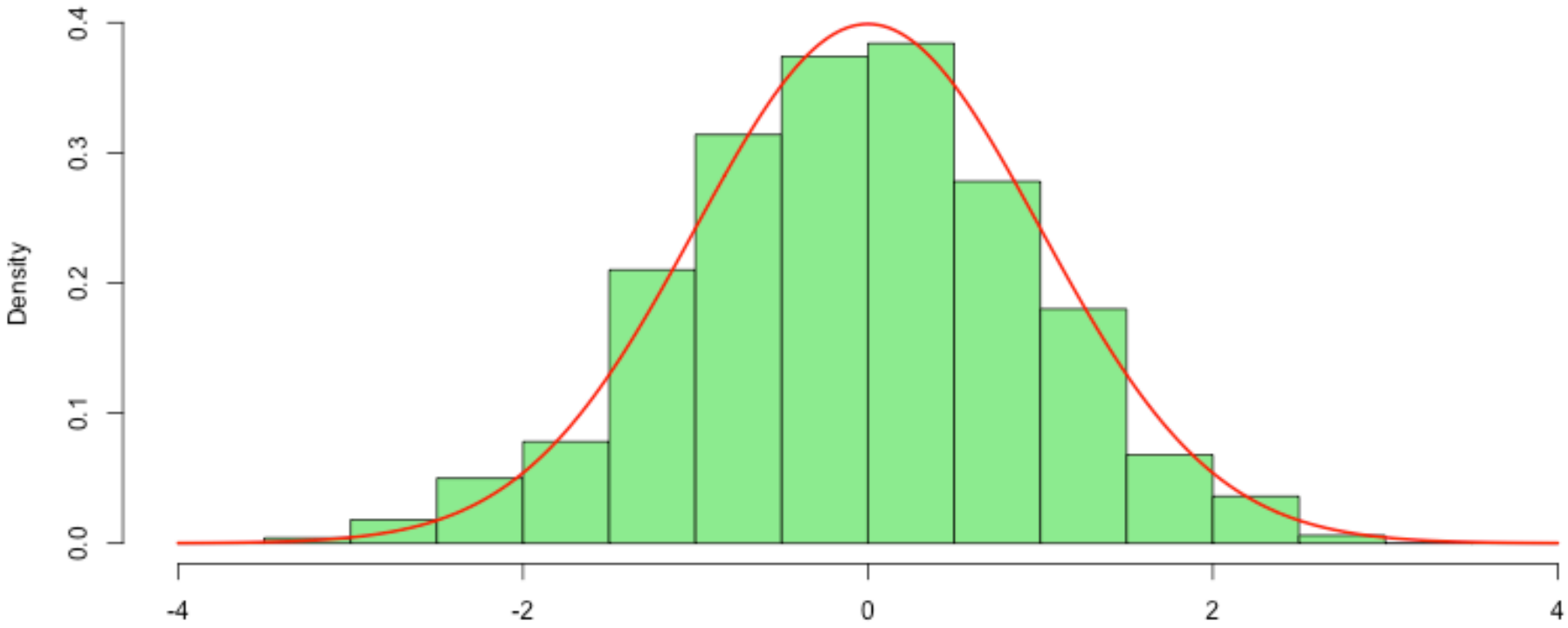


works fine (only) with symmetric distributions

Quantiles and monte carlo simulations

Proposition: If $U \sim \mathcal{U}([0,1])$, then $X = F^{-1}(U) \sim F$

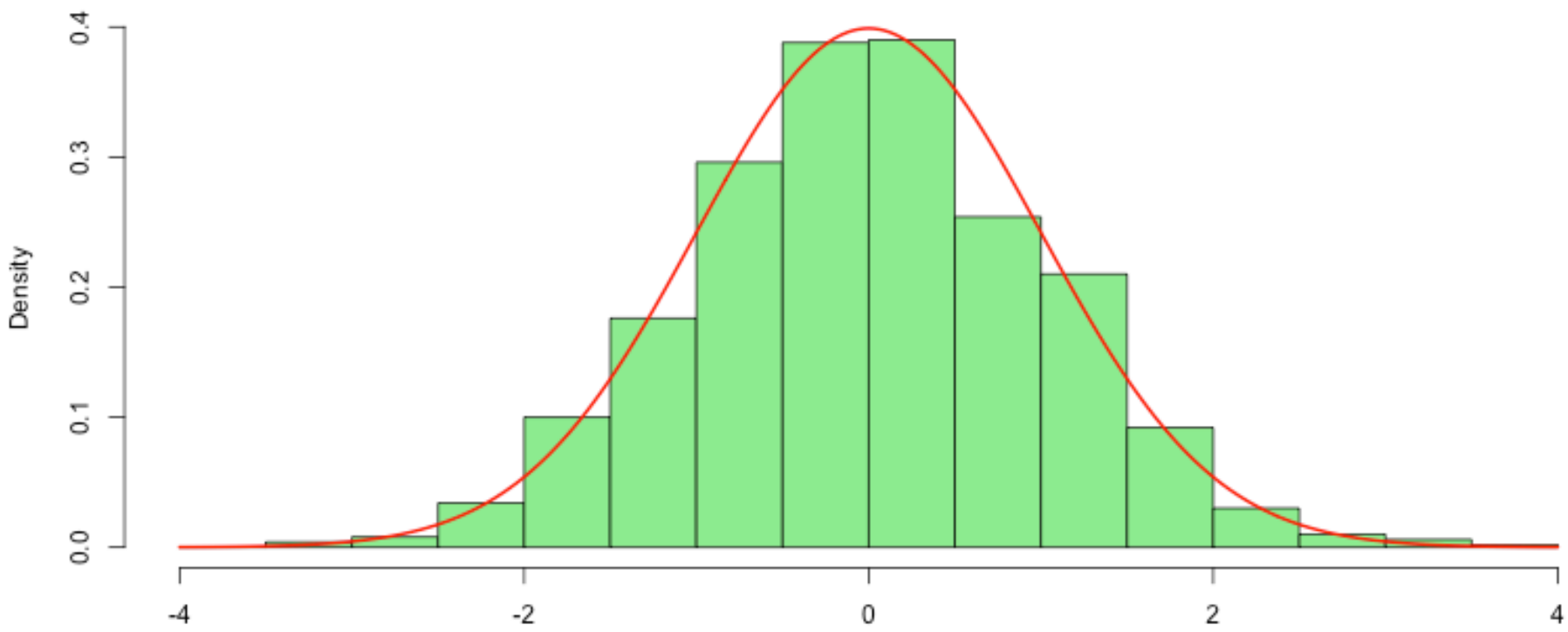
```
> hist(qnorm(runif(1000),proba=TRUE)
```



Quantiles and monte carlo simulations

Proposition: If $U \sim \mathcal{U}([0,1])$, then $X = F^{-1}(U) \sim F$

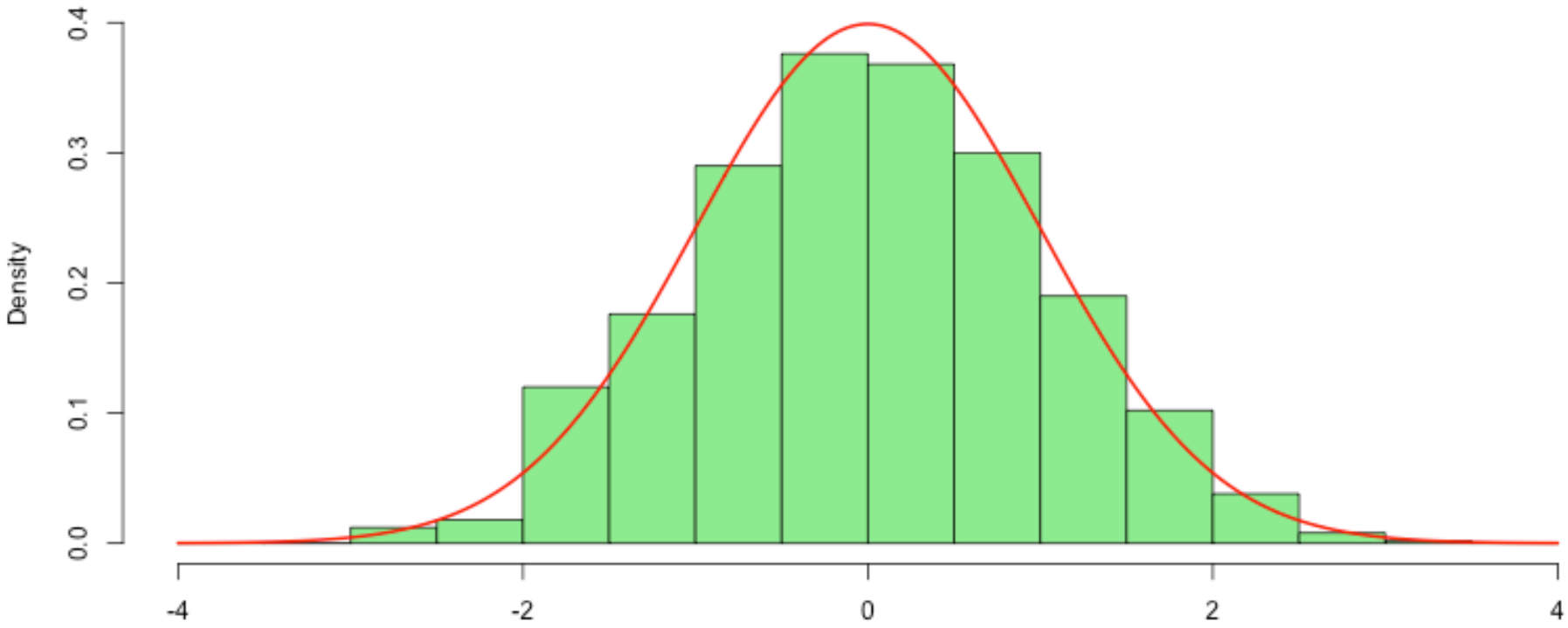
```
> hist(qnorm(runif(1000),proba=TRUE)
```



Quantiles and monte carlo simulations

Proposition: If $U \sim \mathcal{U}([0,1])$, then $X = F^{-1}(U) \sim F$

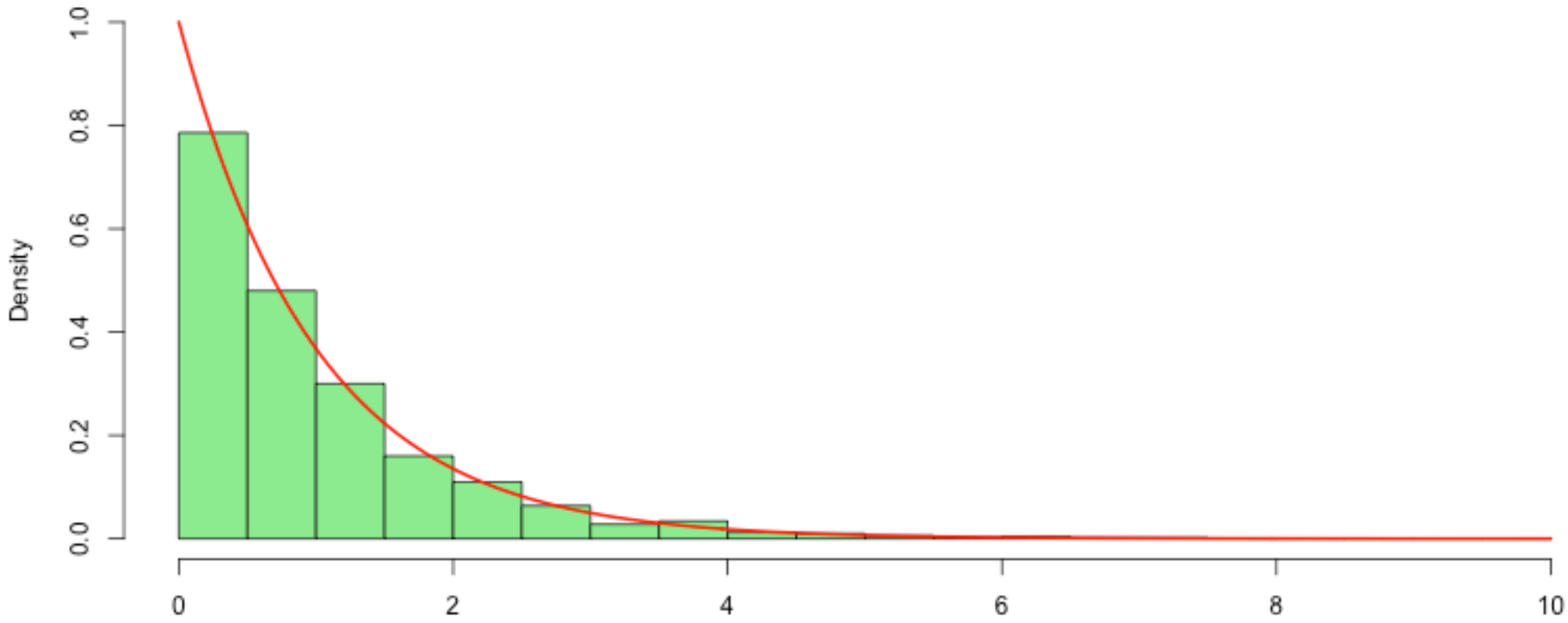
```
> hist(qnorm(runif(1000),proba=TRUE)
```



Quantiles and monte carlo simulations

Proposition: If $U \sim \mathcal{U}([0,1])$, then $X = F^{-1}(U) \sim F$

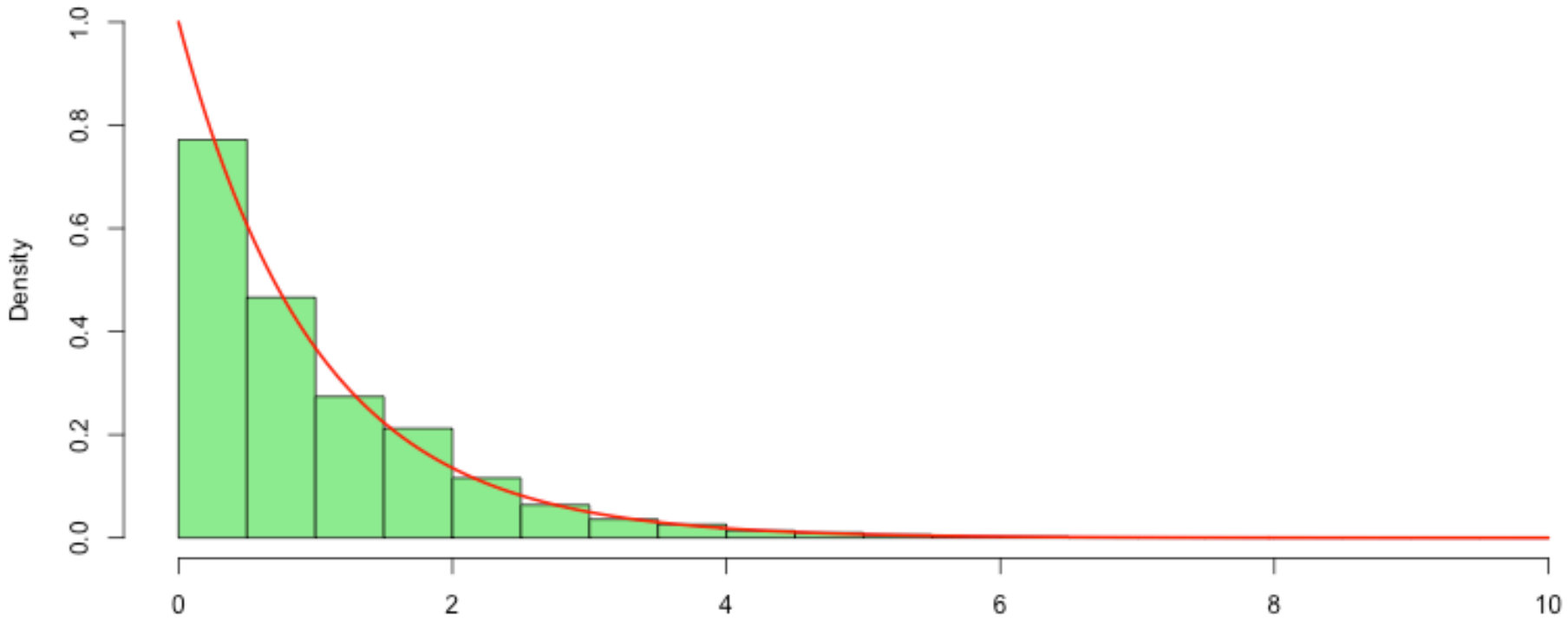
```
> hist(qexp(runif(1000),proba=TRUE)
```



Quantiles and monte carlo simulations

Proposition: If $U \sim \mathcal{U}([0,1])$, then $X = F^{-1}(U) \sim F$

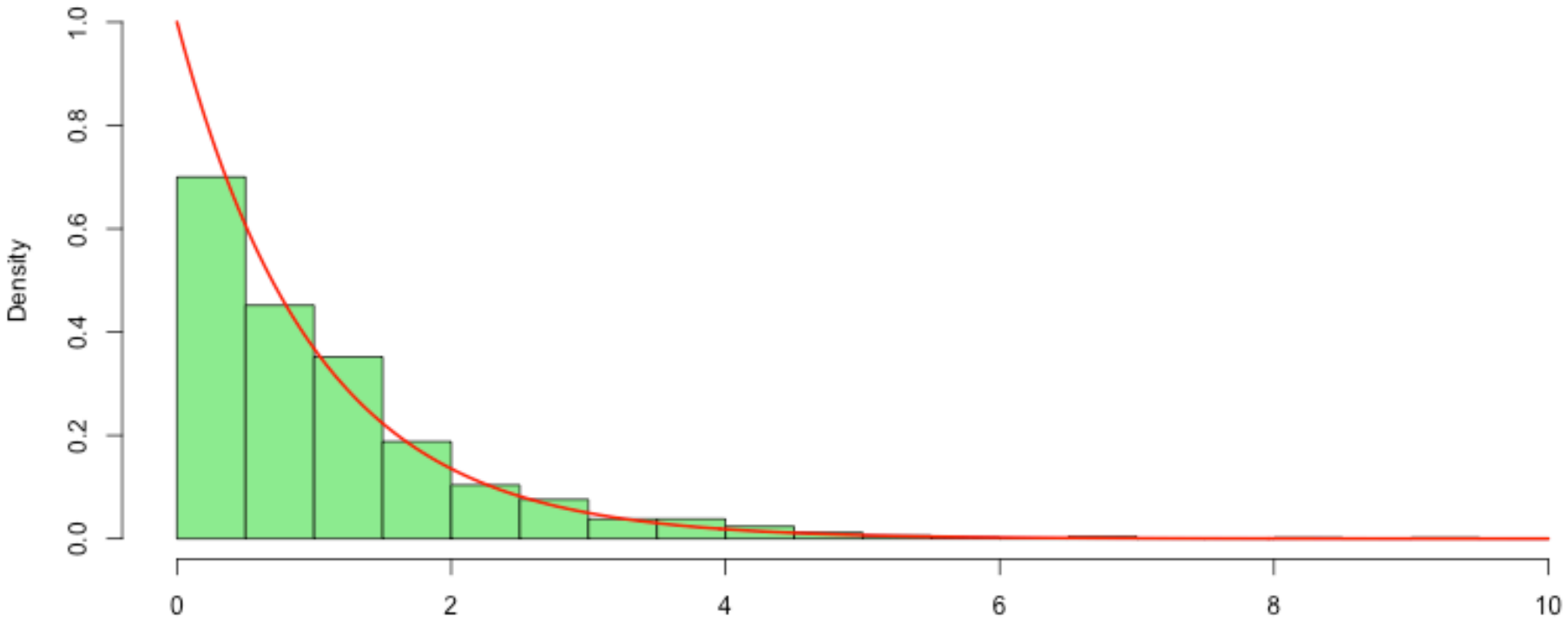
```
> hist(qexp(runif(1000),proba=TRUE)
```



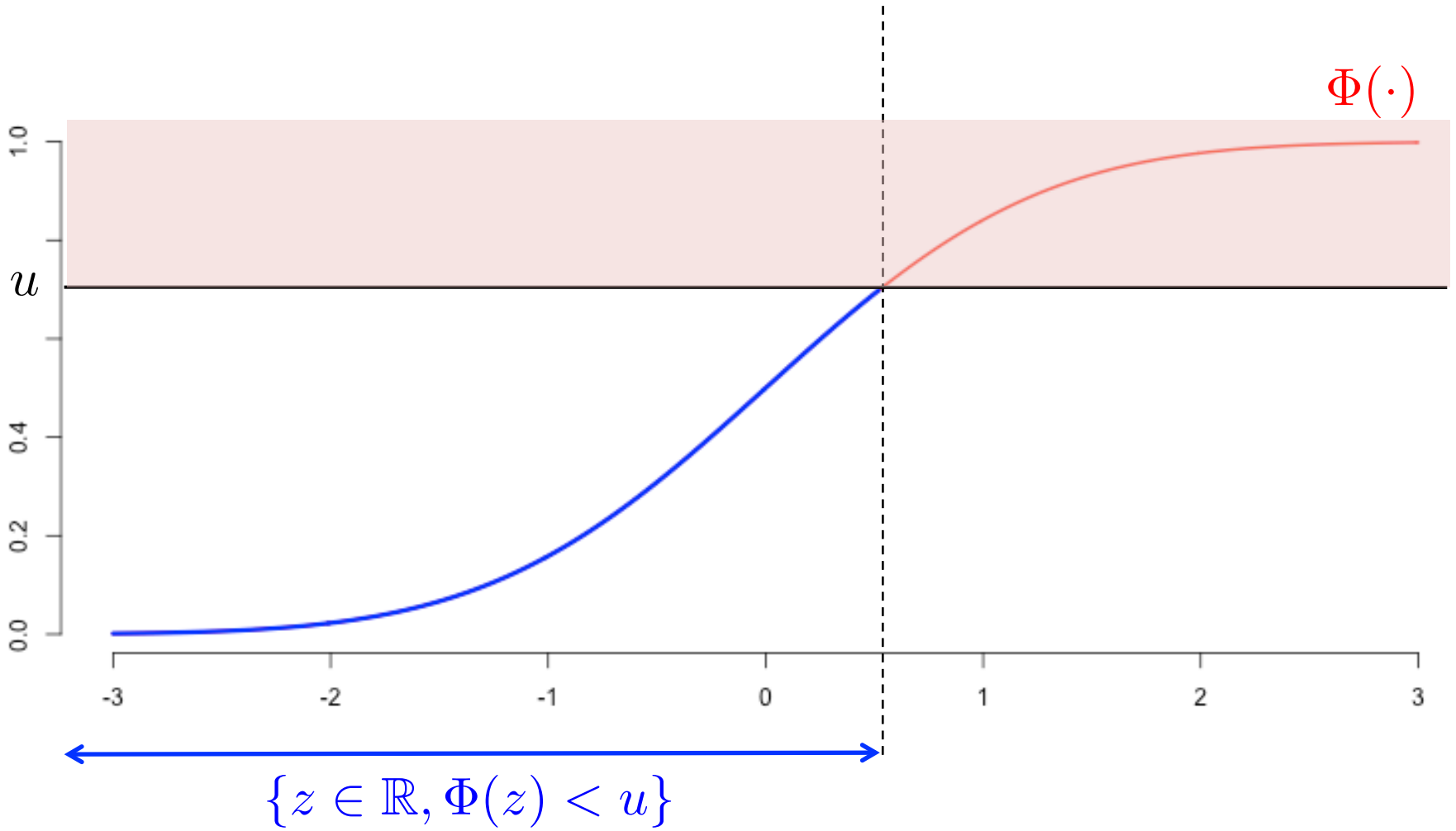
Quantiles and monte carlo simulations

Proposition: If $U \sim \mathcal{U}([0,1])$, then $X = F^{-1}(U) \sim F$

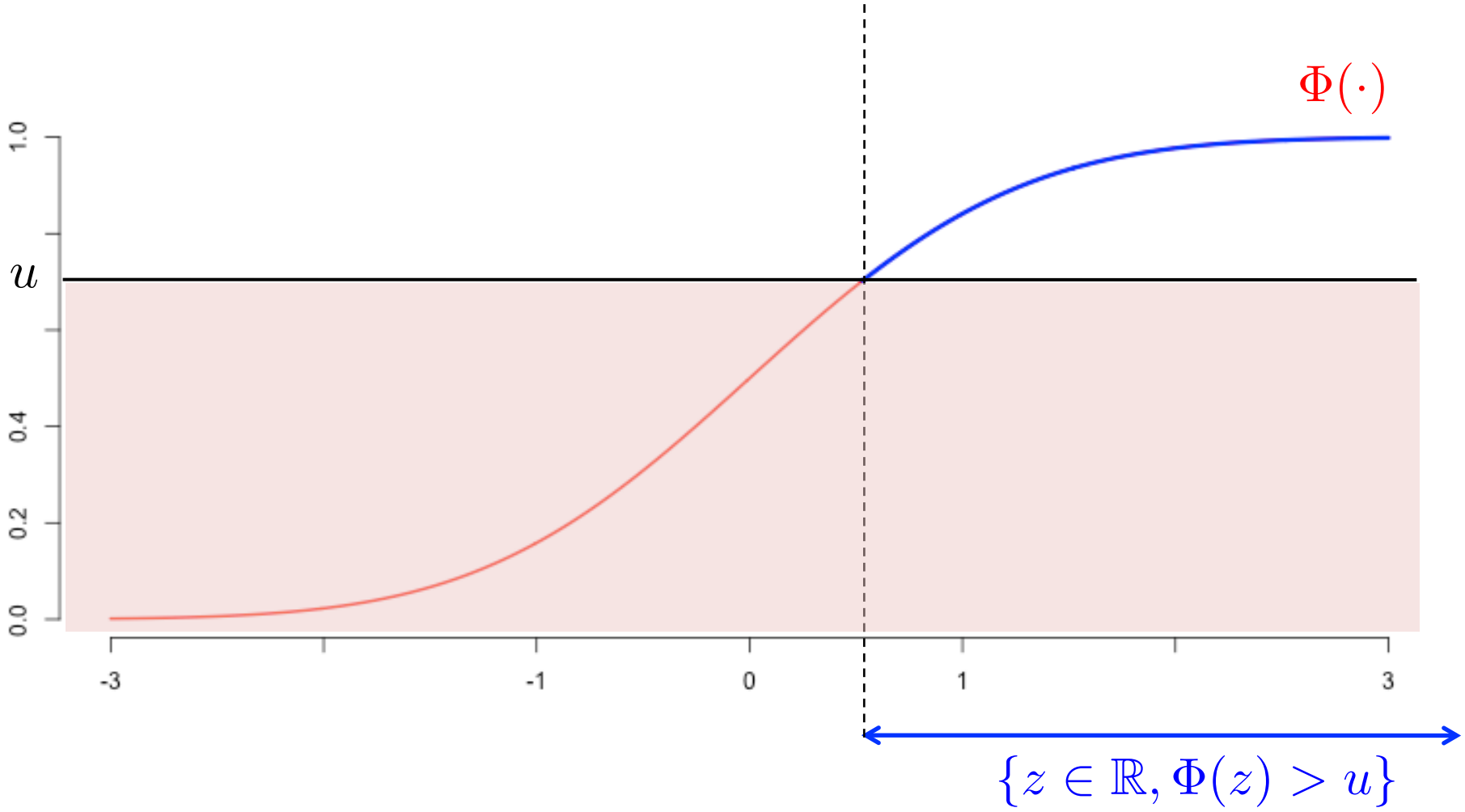
```
> hist(qexp(runif(1000),proba=TRUE)
```



$$\Phi^{-1}(u) = \sup\{z \in \mathbb{R}, \Phi(z) < u\}$$



$$\Phi^{-1}(u) = \inf\{z \in \mathbb{R}, \Phi(z) > u\}$$

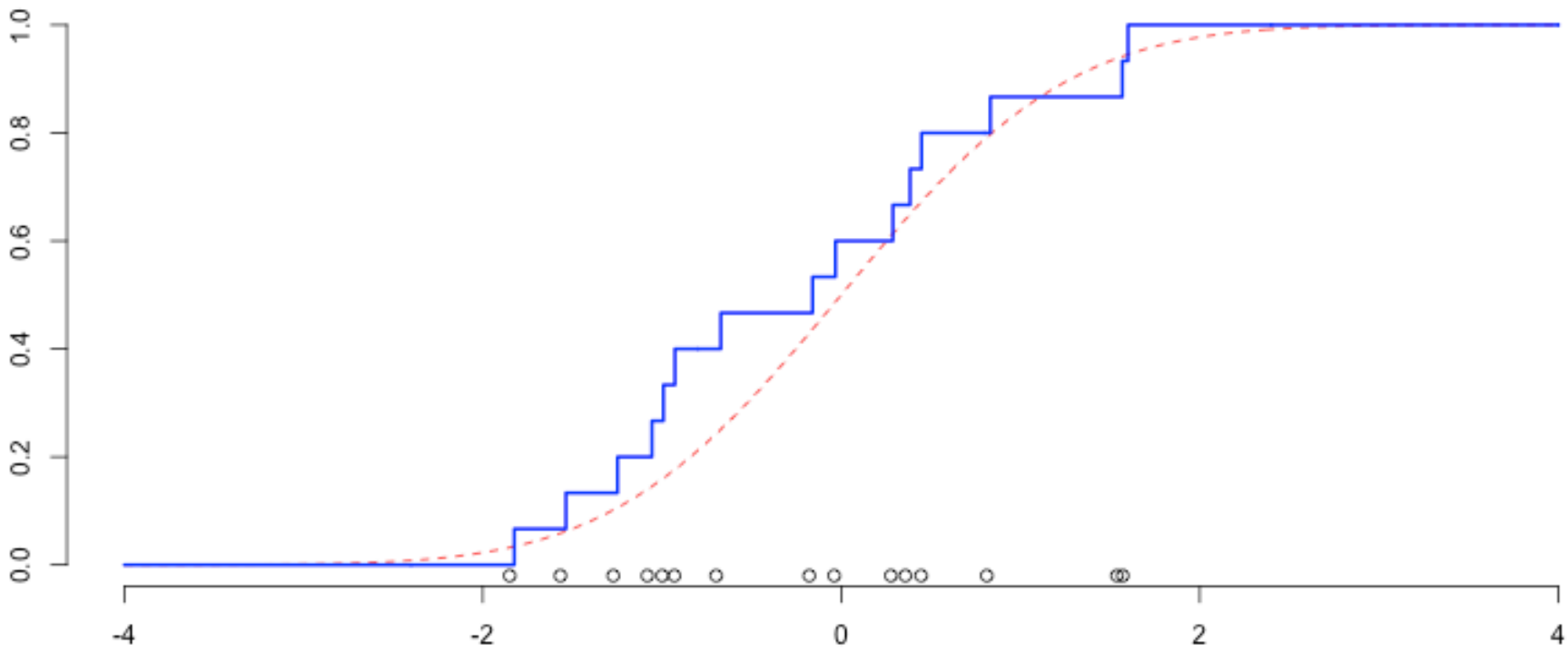


Quantiles and monte carlo simulations

Given a sample $\{X_1, \cdots, X_n\}$ define the e.c.f. as

$$\widehat{F}_n(x) = \widehat{\mathbb{P}}(X \leq x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$$

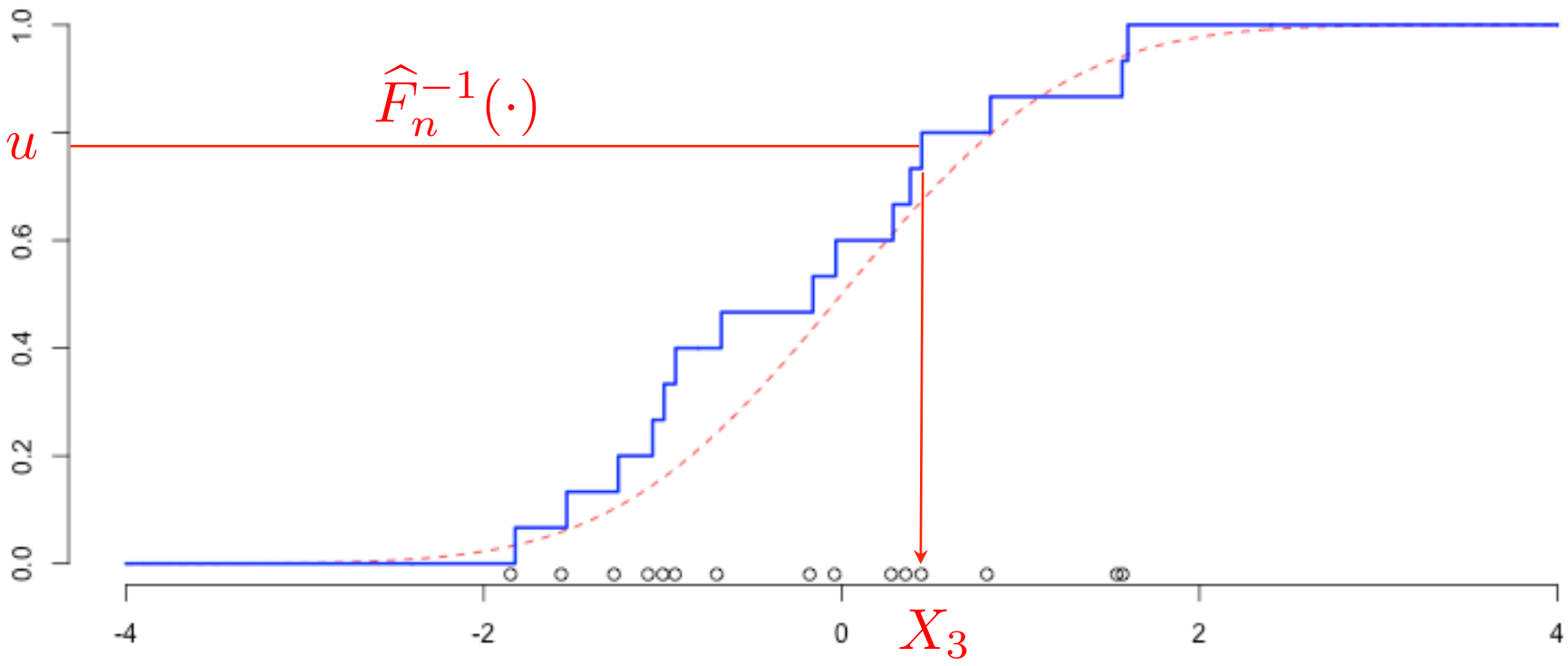
```
> Femp=function(x) mean(X<=x)
```



Quantiles and monte carlo simulations

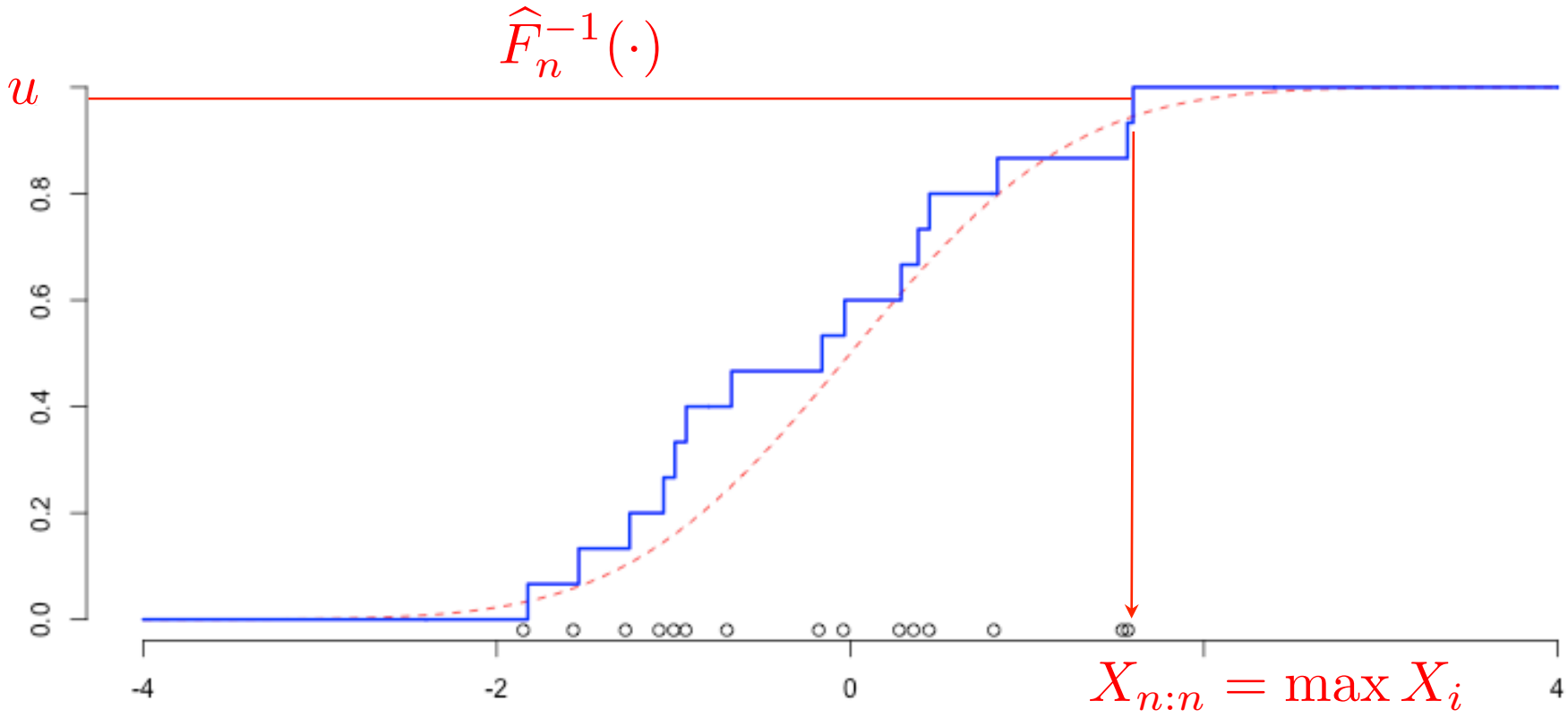
Given a sample $\{X_1, \cdots, X_n\}$ define the e.c.f. as

$$\widehat{F}_n(x) = \widehat{\mathbb{P}}(X \leq x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$$



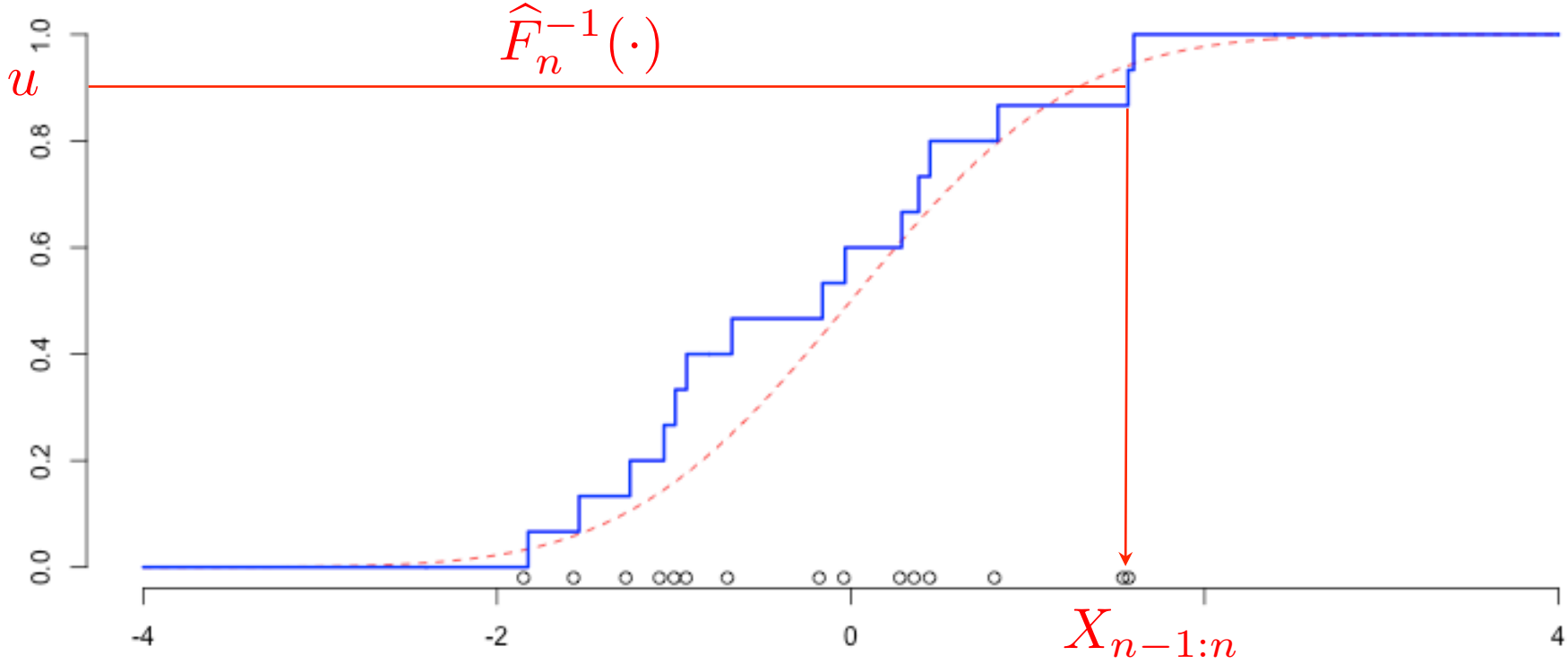
Quantiles and monte carlo simulations

If $u \in \left(\frac{n-1}{n}, 1\right]$ then $\widehat{F}_n^{-1}(u) = X_{n:n} = \max\{X_i\}$



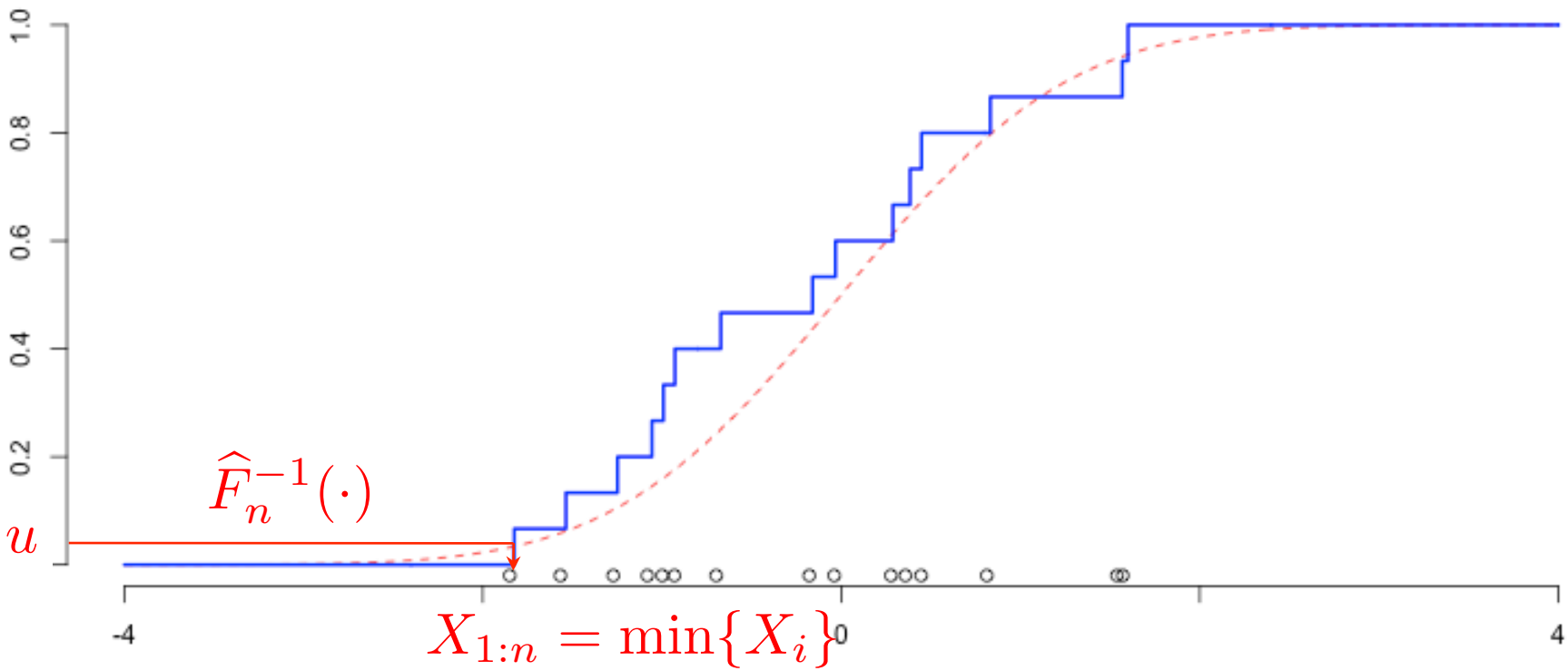
Quantiles and monte carlo simulations

If $u \in \left(\frac{n-2}{n}, \frac{n-1}{n} \right]$ then $\widehat{F}_n^{-1}(u) = X_{n-1:n}$



Quantiles and monte carlo simulations

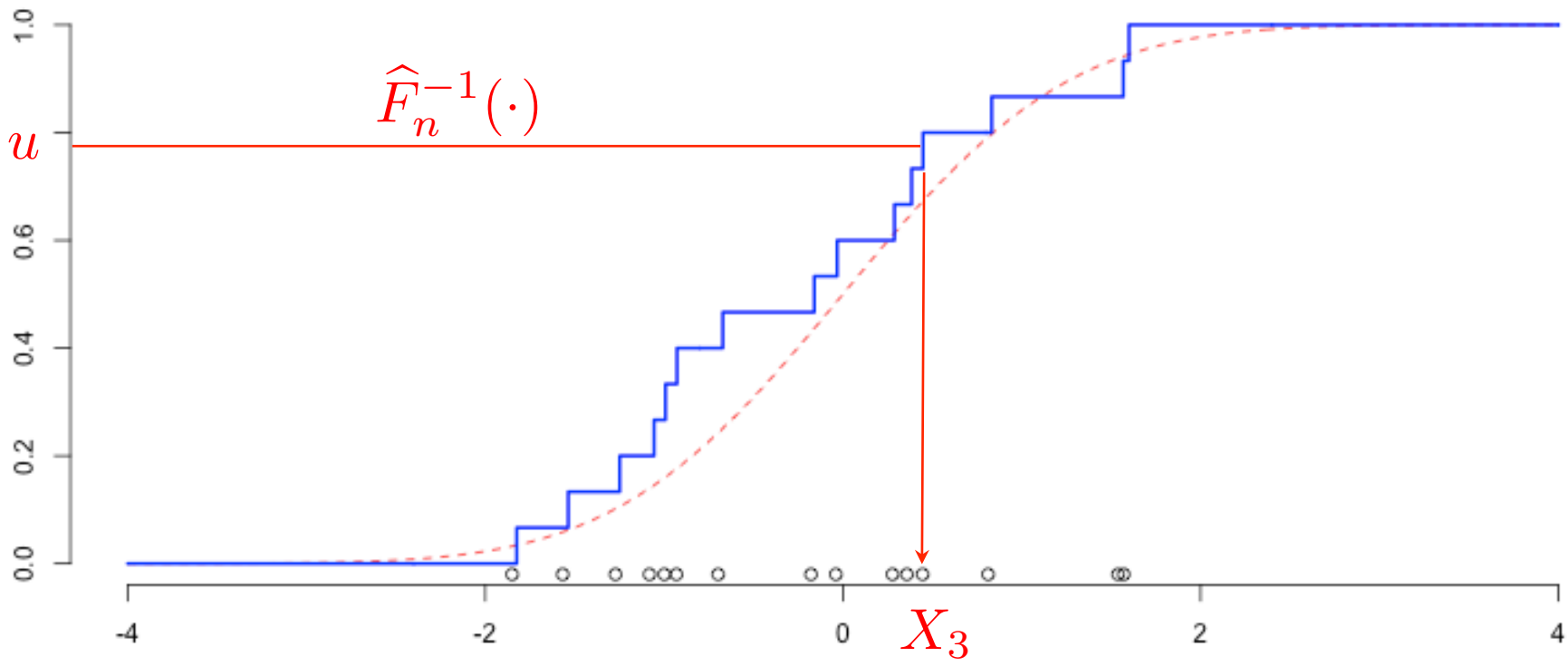
If $u \in \left(0, \frac{1}{n}\right]$ then $\widehat{F}_n^{-1}(u) = X_{1:n} = \min\{X_i\}$



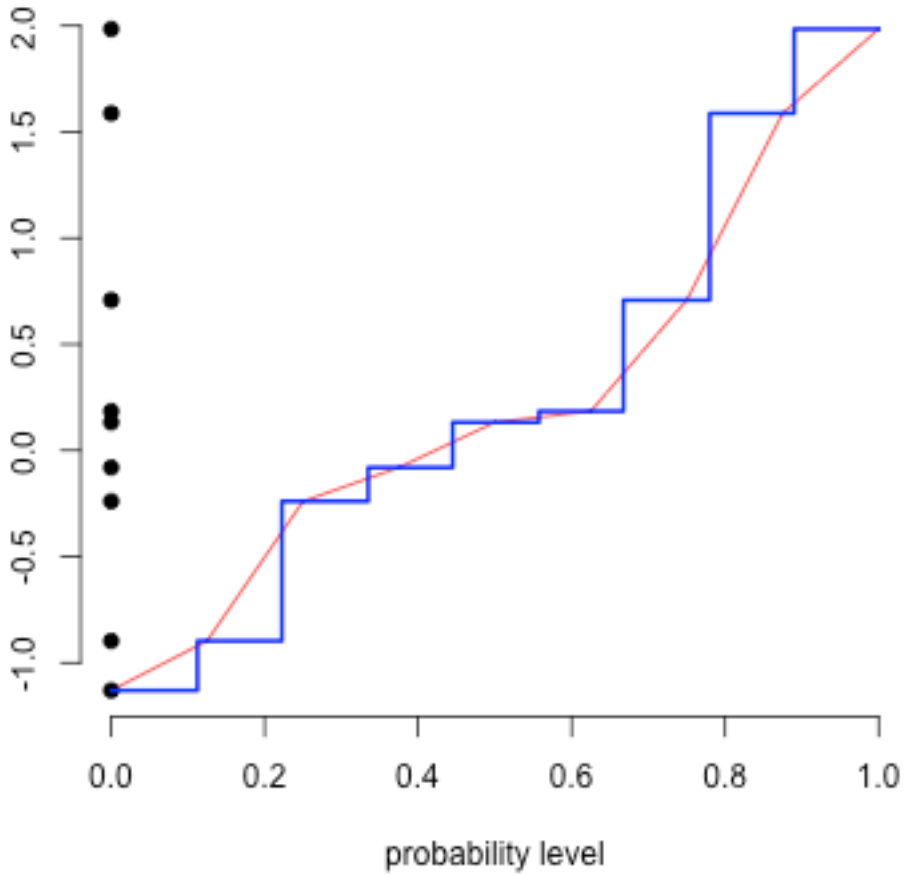
Quantiles and monte carlo simulations

```
> Vectorize(Femp)(runif(100))
> sample(X,size=100,replace=TRUE)
```

see bootstrap techniques



Empirical quantiles ?




```
> set.seed(2); X=rnorm(9)
> Q=quantile(X,u,type=1)
```

$$\alpha X_{[np]:n} + (1-\alpha) X_{[np]+1:n}$$

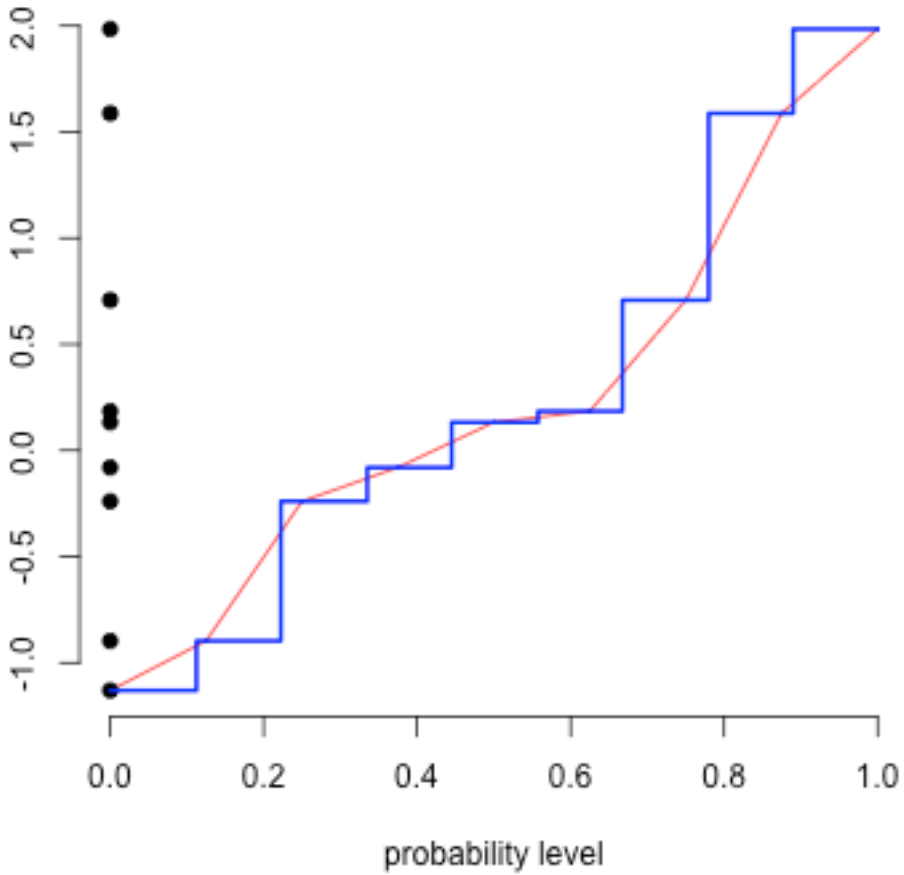
where

$$X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$$

Inverse of empirical distribution function

Hyndman, R. J. & Fan, Y. (1996) Sample quantiles in statistical packages *American Statistician* **50** 361–365

Empirical quantiles ?




```
> set.seed(2); X=rnorm(9)
> Q=quantile(X,u,type=2)
```

$$\alpha X_{[np]:n} + (1-\alpha) X_{[np]+1:n}$$

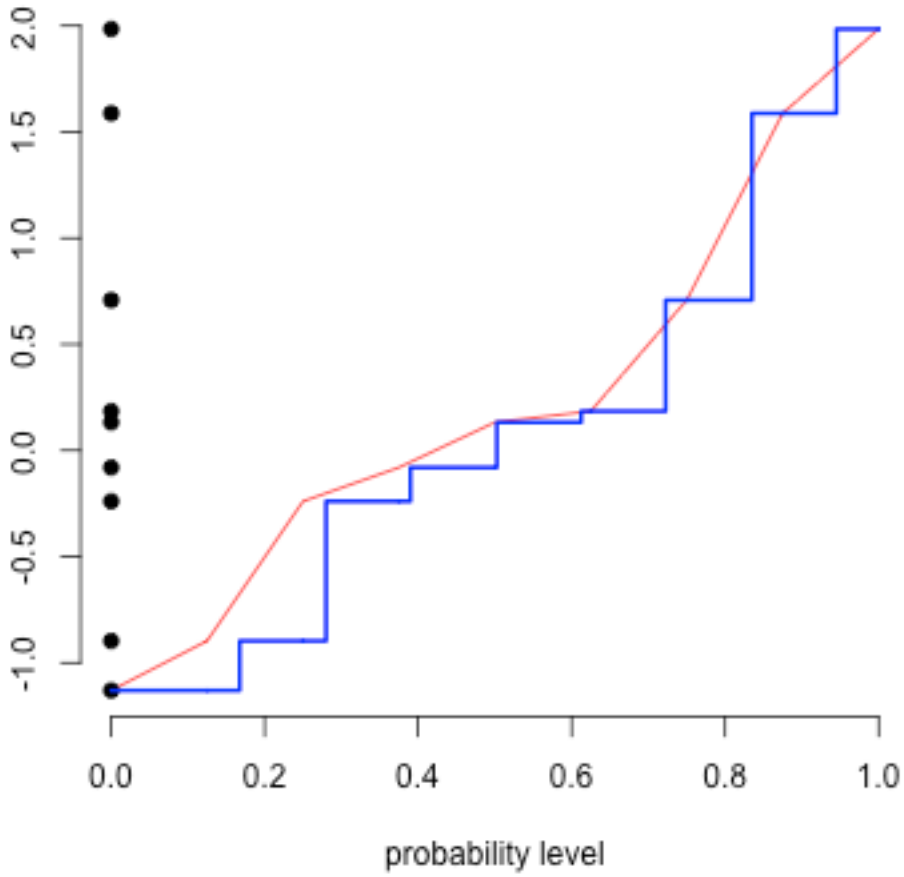
where

$$X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$$

Inverse of empirical distribution function

Hyndman, R. J. & Fan, Y. (1996) Sample quantiles in statistical packages *American Statistician* **50** 361–365

Empirical quantiles ?




```
> set.seed(2); X=rnorm(9)
> Q=quantile(X,u,type=3)
```

$$\alpha X_{[np]:n} + (1-\alpha) X_{[np]+1:n}$$

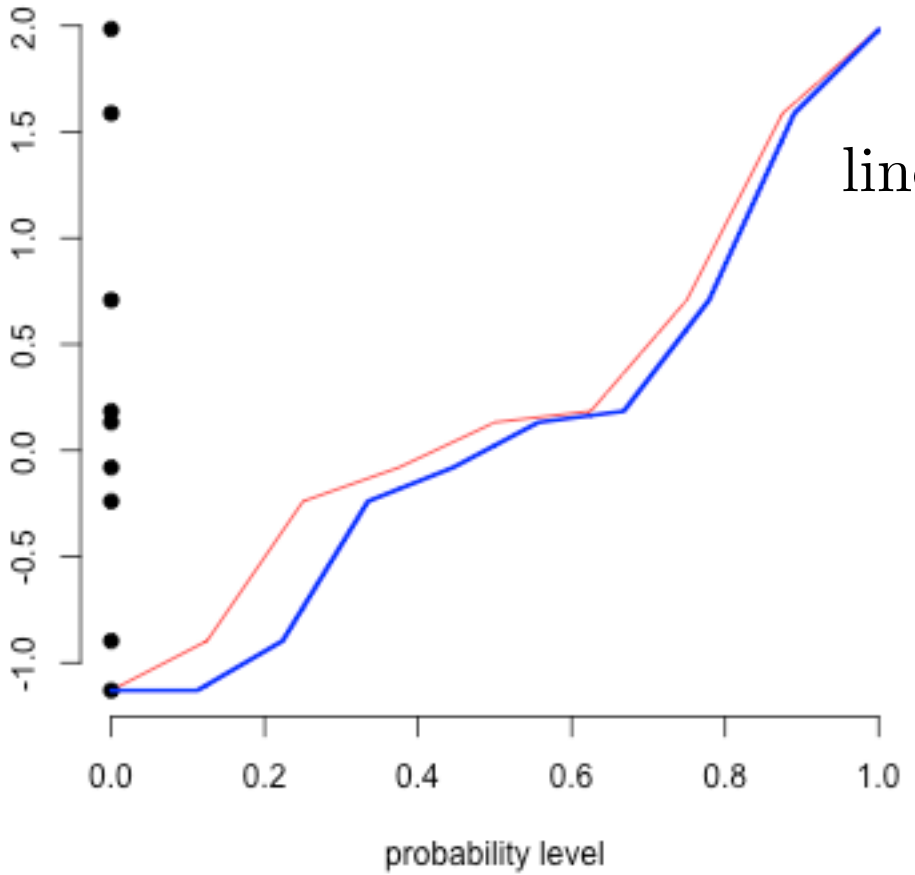
where

$$X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$$

Nearest even order statistic

Hyndman, R. J. & Fan, Y. (1996) Sample quantiles in statistical packages *American Statistician* **50** 361–365

Empirical quantiles ?




```
> set.seed(2); X=rnorm(9)
> Q=quantile(X,u,type=4)
```

linear interpolation between points

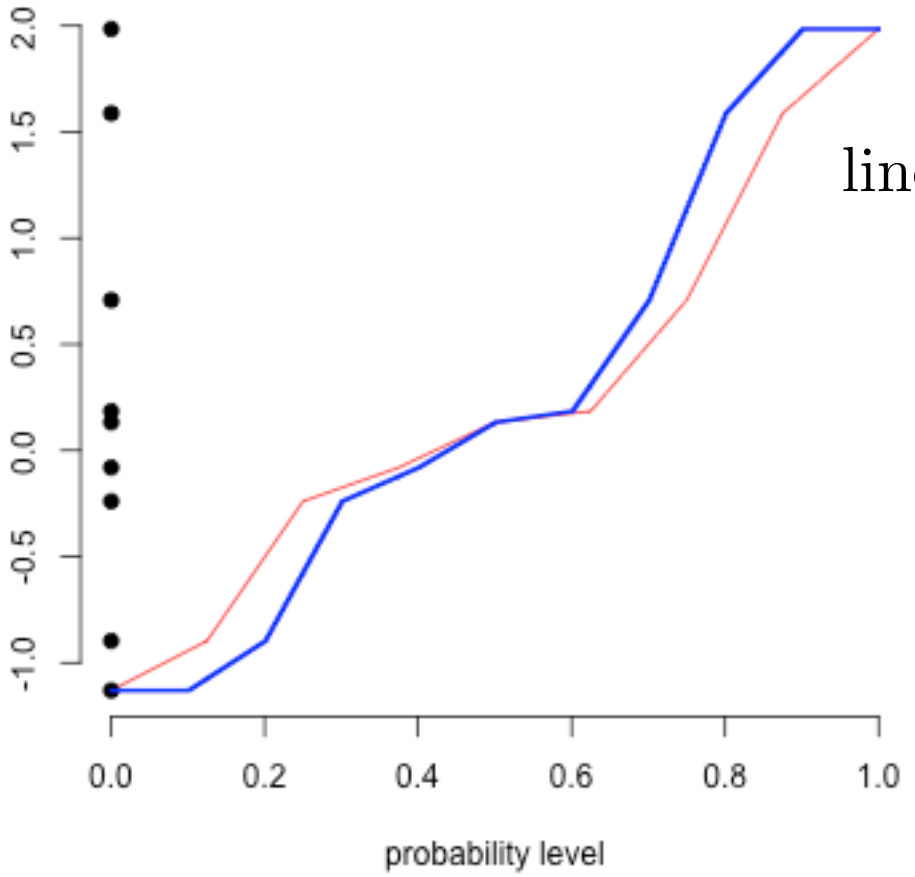
$$\{(p(k), X_{k:n}); k = 1, \cdots, n\}$$

$$p(k) = \frac{k}{n}$$

Linear interpolation of the empirical cdf

Hyndman, R. J. & Fan, Y. (1996) Sample quantiles in statistical packages *American Statistician* **50** 361–365

Empirical quantiles ?




```
> set.seed(2); X=rnorm(9)
> Q=quantile(X,u,type=5)
```

linear interpolation between points

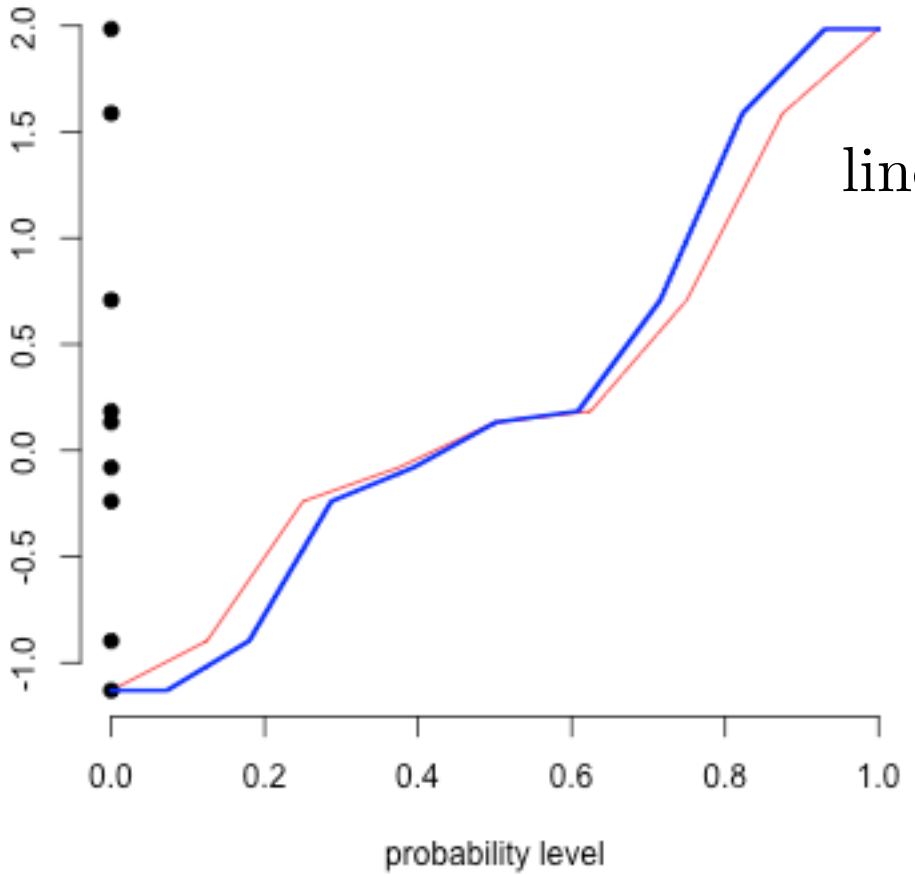
$$\{(p(k), X_{k:n}); k = 1, \cdots, n\}$$

$$p(k) = \frac{k - 1/2}{n}$$

Piecewise linear function, knots are the values midway through steps

Hyndman, R. J. & Fan, Y. (1996) Sample quantiles in statistical packages *American Statistician* **50** 361–365

Empirical quantiles ?




```
> set.seed(2); X=rnorm(9)
> Q=quantile(X,u,type=6)
```

linear interpolation between points

$$\{(p(k), X_{k:n}); k = 1, \cdots, n\}$$

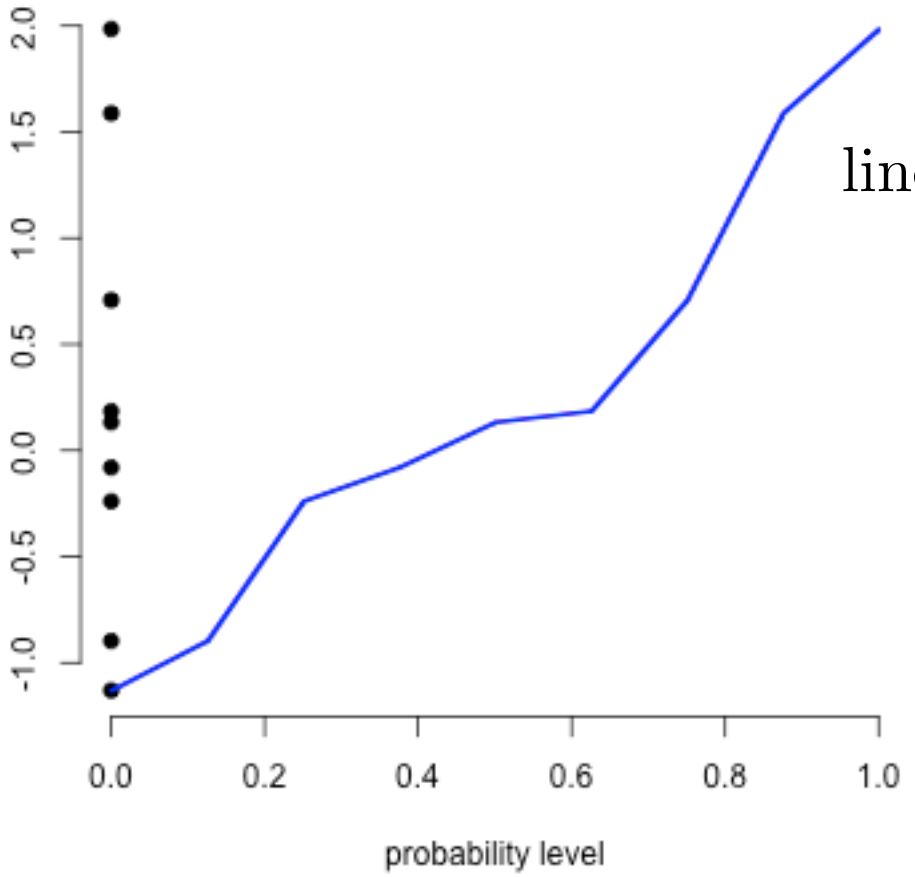
$$p(k) = \frac{k}{n+1}$$

Minitab QUALITY. ANALYSIS. RESULTS.

SPSS

Hyndman, R. J. & Fan, Y. (1996) Sample quantiles in statistical packages *American Statistician* **50** 361–365

Empirical quantiles ?




```
> set.seed(2); X=rnorm(9)
> Q=quantile(X,u,type=7)
```

linear interpolation between points

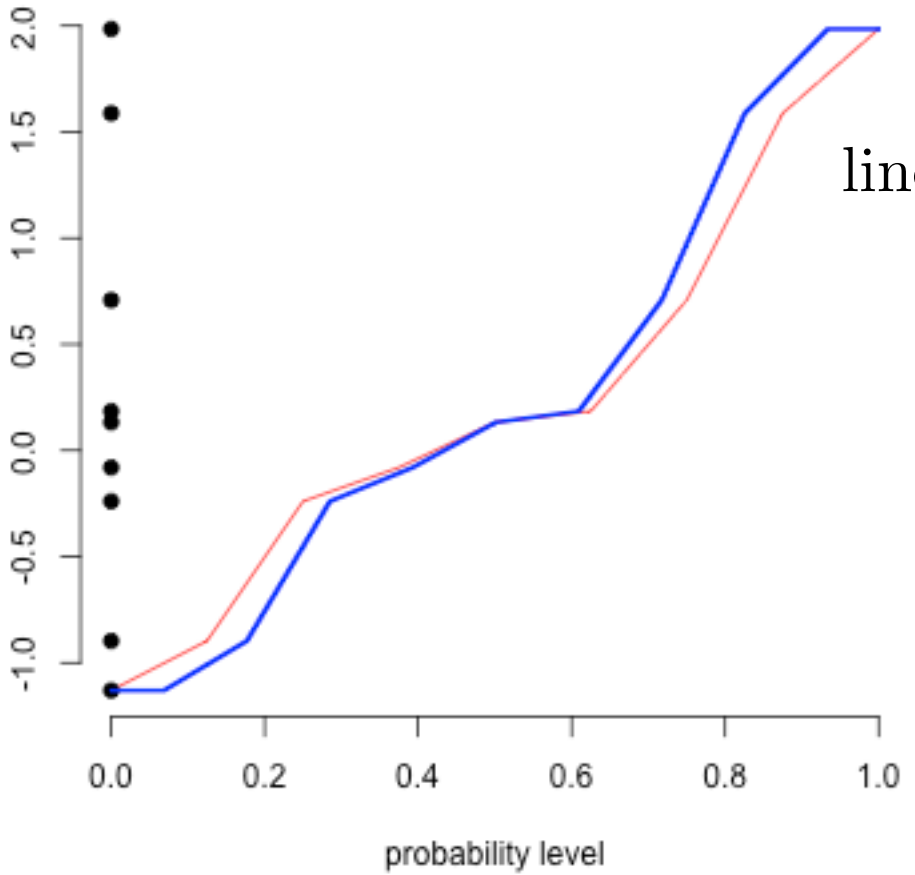
$$\{(p(k), X_{k:n}); k = 1, \cdots, n\}$$

$$p(k) = \frac{k-1}{n-1}$$

S-PLUS 6

Hyndman, R. J. & Fan, Y. (1996) Sample quantiles in statistical packages *American Statistician* **50** 361–365

Empirical quantiles ?




```
> set.seed(2); X=rnorm(9)
> Q=quantile(X,u,type=9)
```

linear interpolation between points

$$\{(p(k), X_{k:n}); k = 1, \cdots, n\}$$

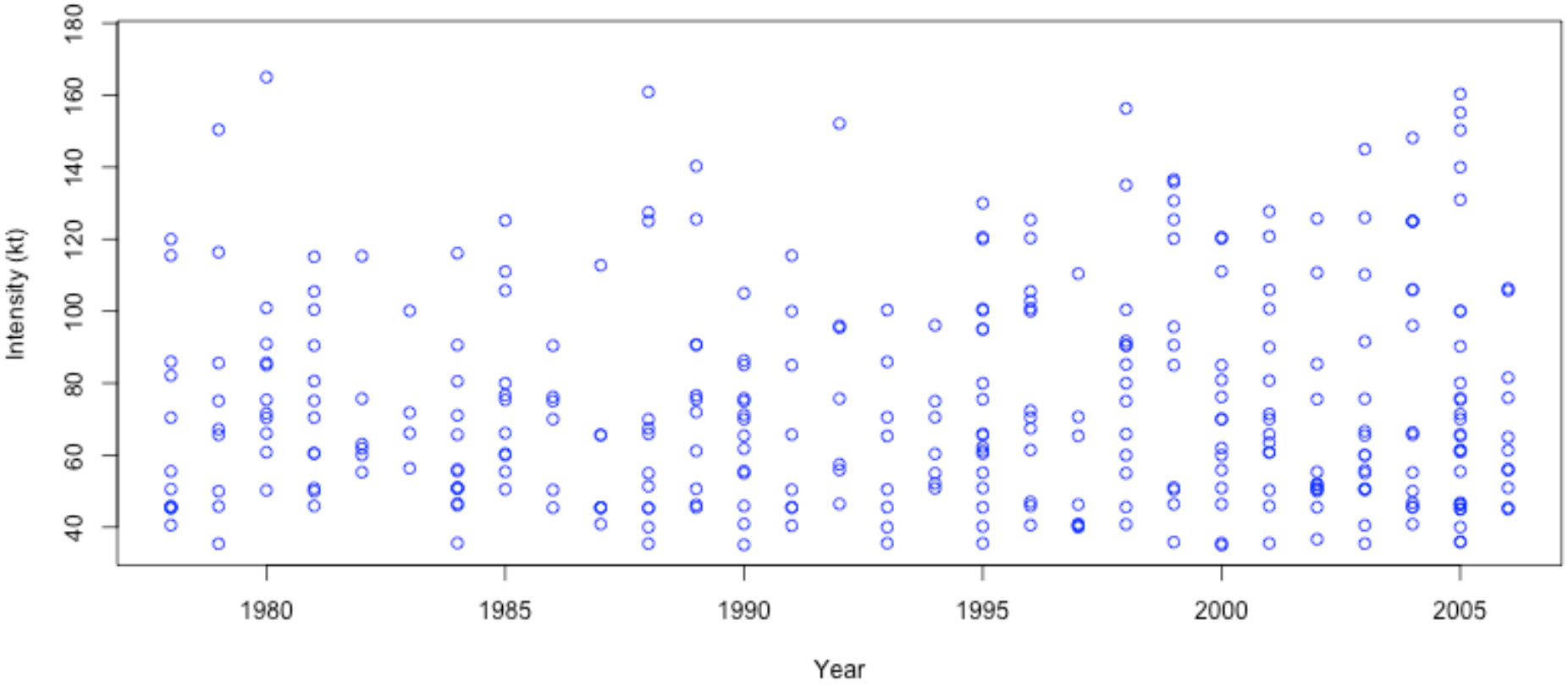
$$p(k) = \frac{k - 3/8}{n + 1/4}$$

Approximately median-unbiased (when Gaussian)

Hyndman, R. J. & Fan, Y. (1996) Sample quantiles in statistical packages *American Statistician* **50** 361–365

Quantile regression, a motivation

```
> StormMax=read.table(
+"http://freakonometrics.free.fr/extremedatasince1899.csv"
+header=TRUE,sep=",")
```

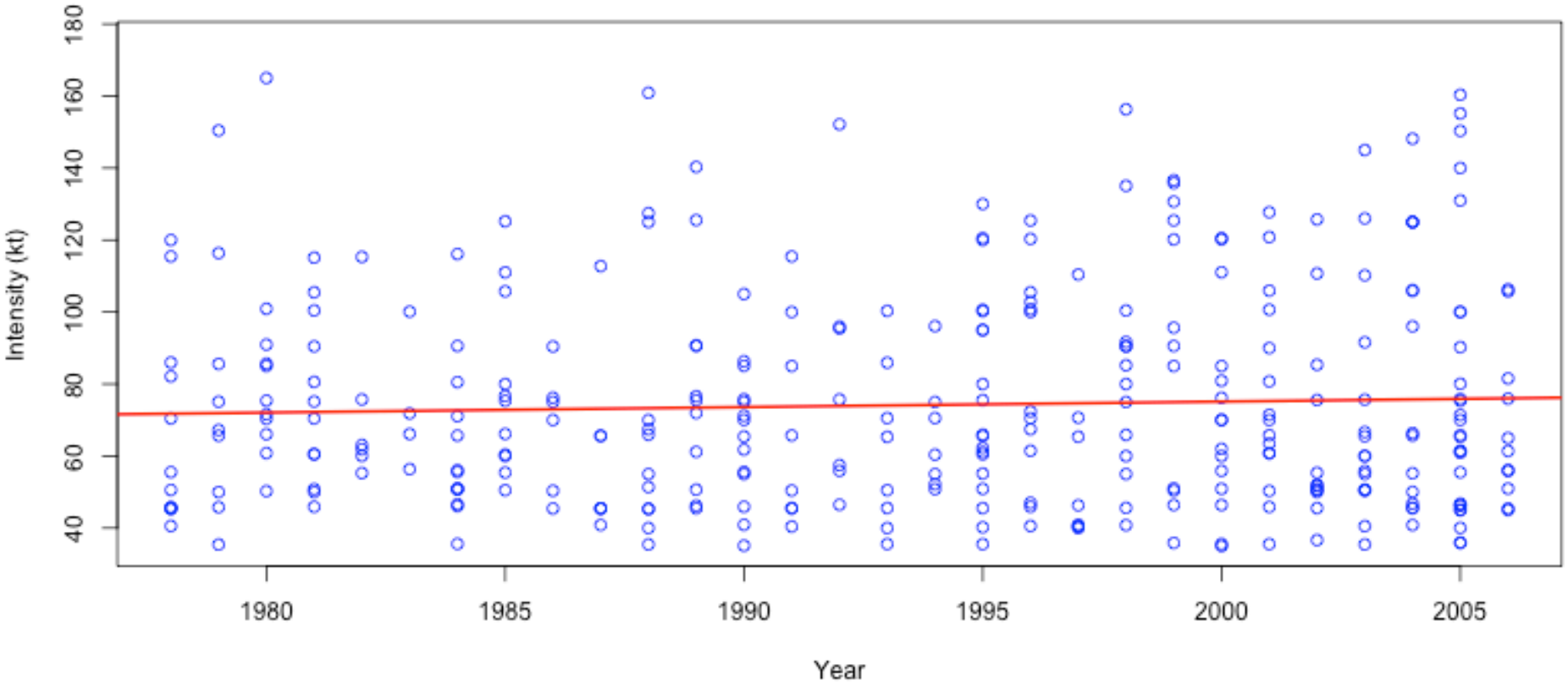


Jagger, T.H. & Elsner, J.B. (2008) Modeling tropical cyclone intensity with quantile regression. *Int. J. Climatol.* doi:10.1002/joc.1804

Elsner, J.B., Kossin, J.P. & Jagger, T.H. (2008) The increasing intensity of the strongest tropical cyclones. *Nature* **455**, 92-95.

Quantile regression, a motivation

```
> abline(lm(Wmax Yr),lwd=2,col="red")
```

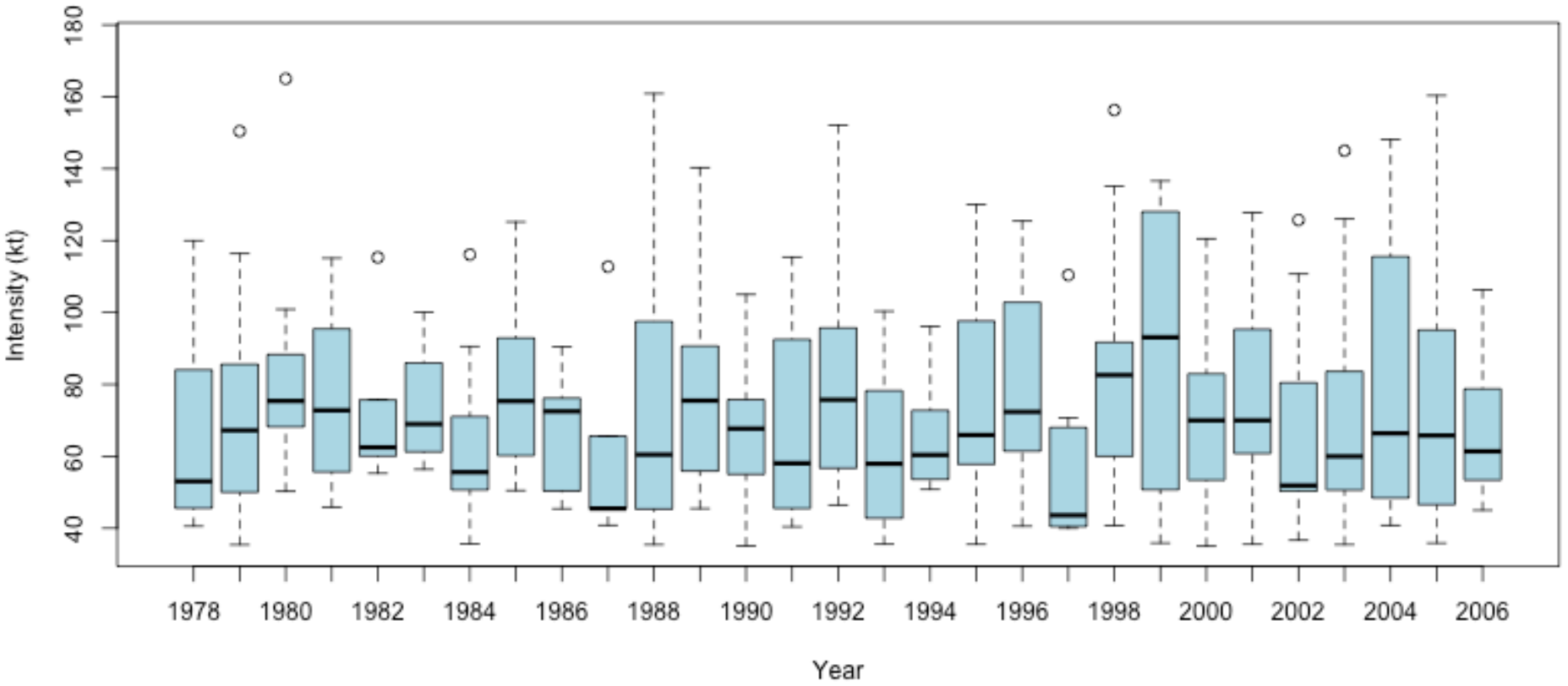


Jagger, T.H. & Elsner, J.B. (2008) Modeling tropical cyclone intensity with quantile regression. *Int. J. Climatol.* doi:10.1002/joc.1804

Elsner, J.B., Kossin, J.P. & Jagger, T.H. (2008) The increasing intensity of the strongest tropical cyclones. *Nature* **455**, 92-95.

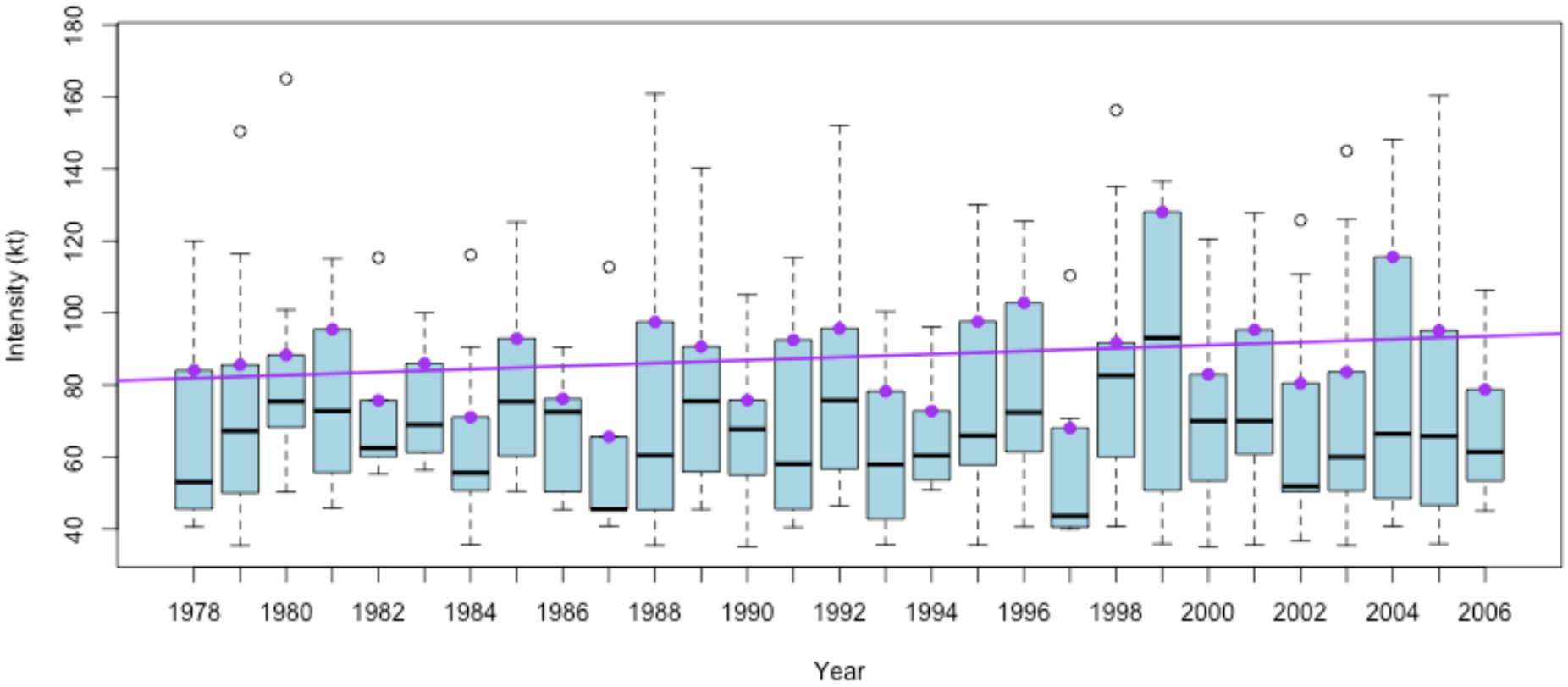
Regression on quantiles

```
> bp=boxplot(Wmax as.factor(Yr))
```



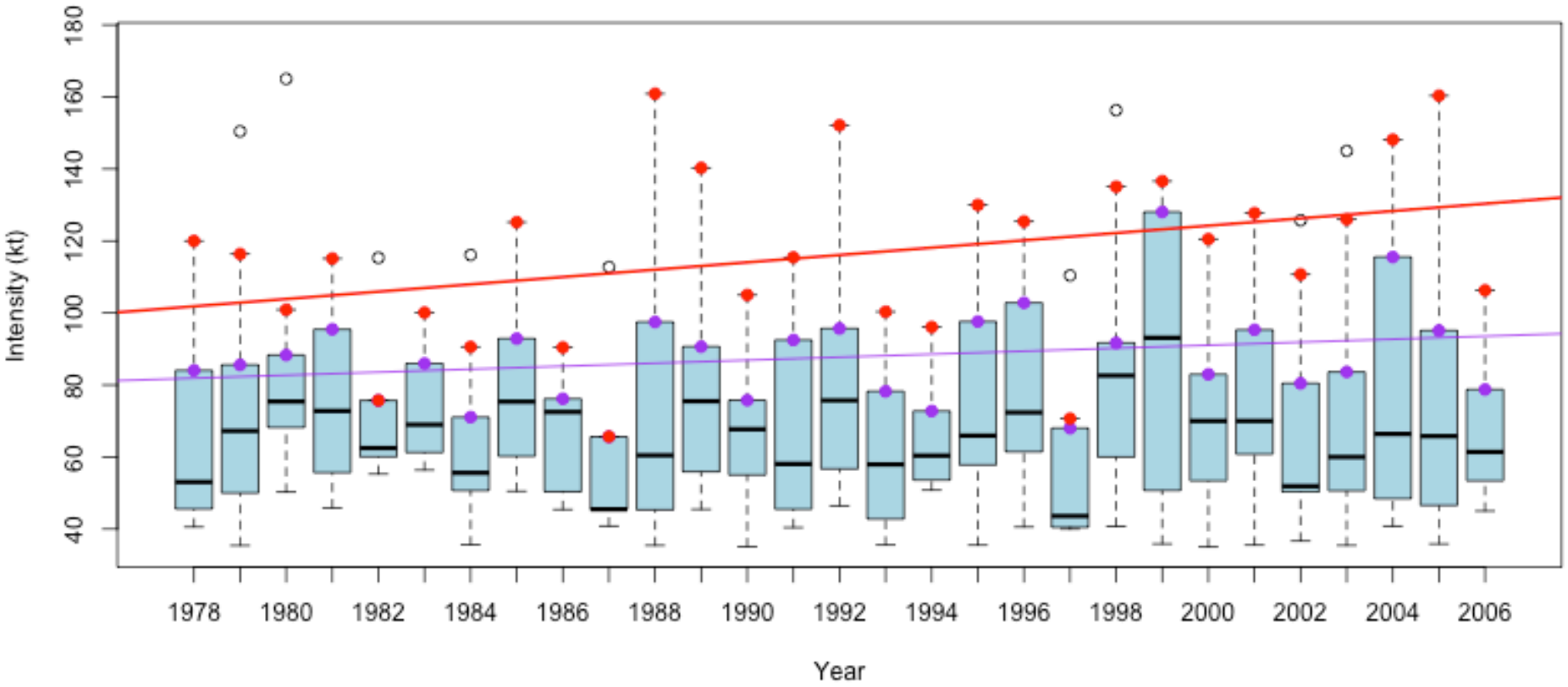
Regression on quantiles

```
> bp=boxplot(Wmax as.factor(Yr))
> x=1:29; abline(lm(bp$stats[4,]~x))
```



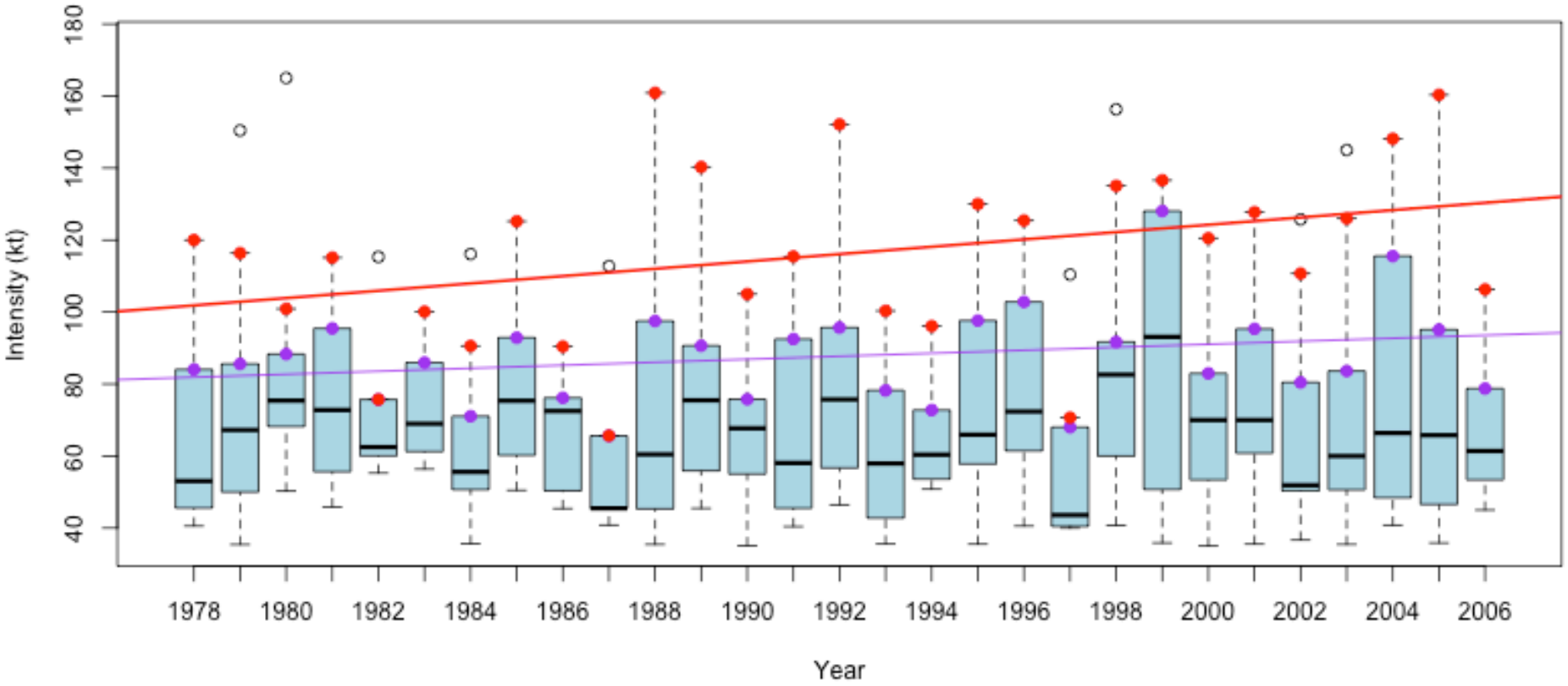
Regression on quantiles

```
> bp=boxplot(Wmax as.factor(Yr))
> x=1:29; abline(lm(bp$stats[4,]~x))
```



Regression on quantiles

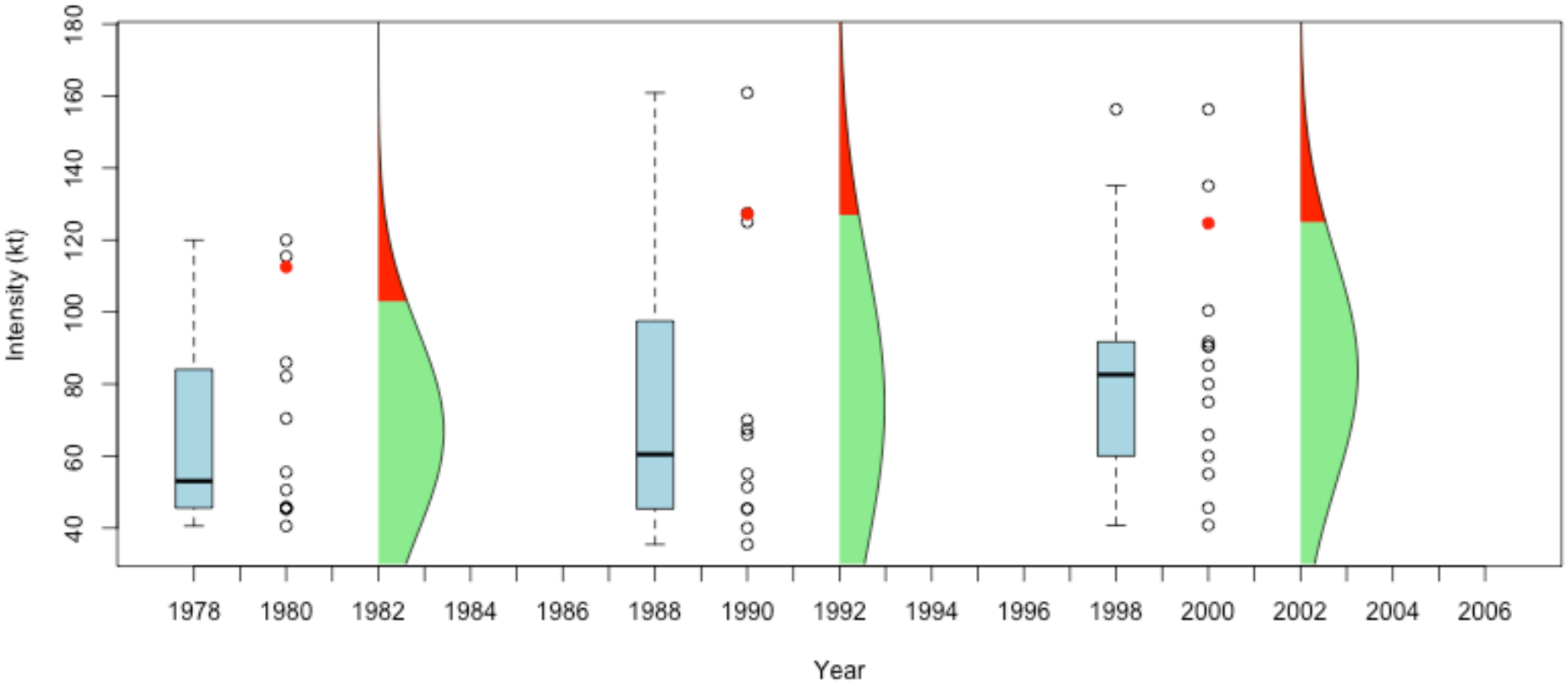
```
> bp=boxplot(Wmax as.factor(Yr))
> x=1:29; abline(lm(bp$stats[4,]~x))
> x=1:29; abline(lm(bp$stats[5,]~x))
```



Regression on quantiles

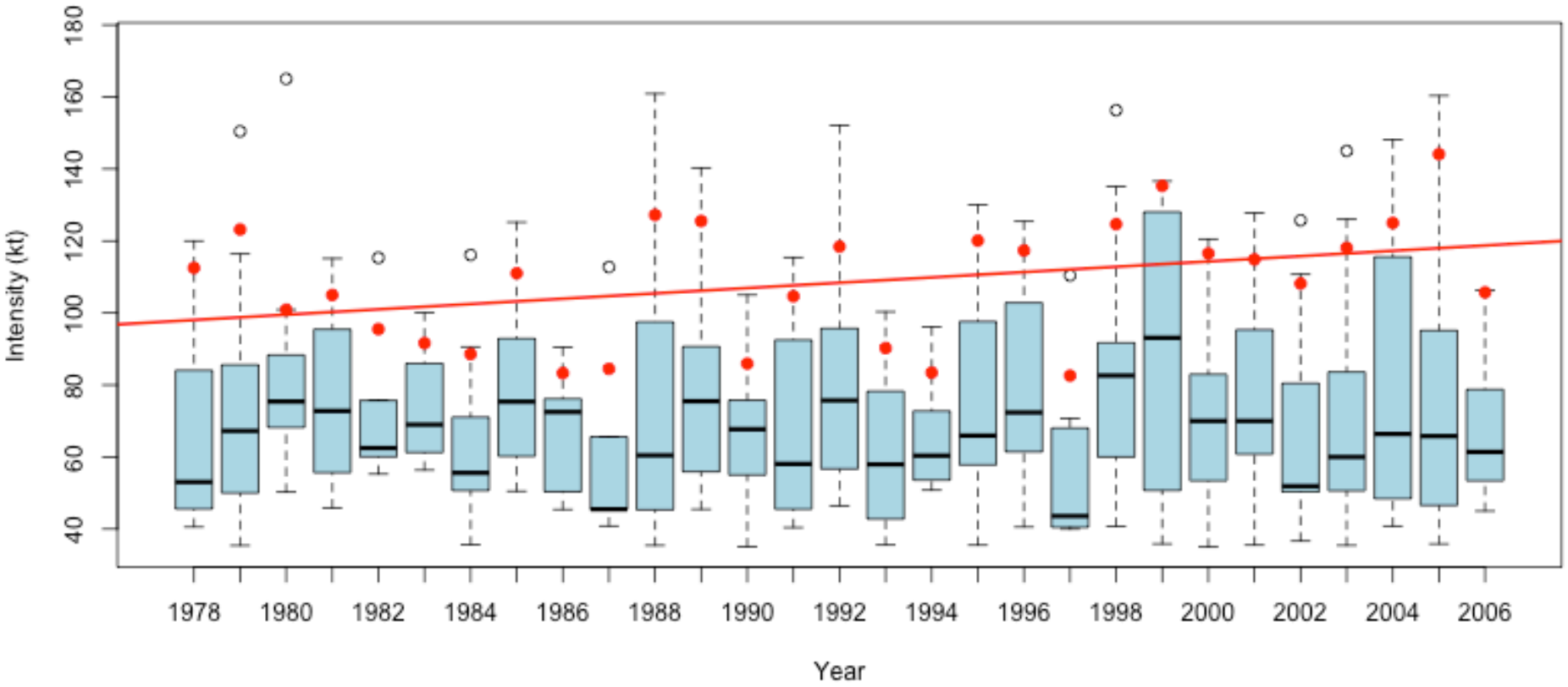
Empirical quantile `> quantile(Wmax[Yr==1988],.9)`

Gaussian quantile `> qnorm(.9,mean(Wmax[Yr==1988]),sd(Wmax[Yr==1988]))`



Regression on quantiles

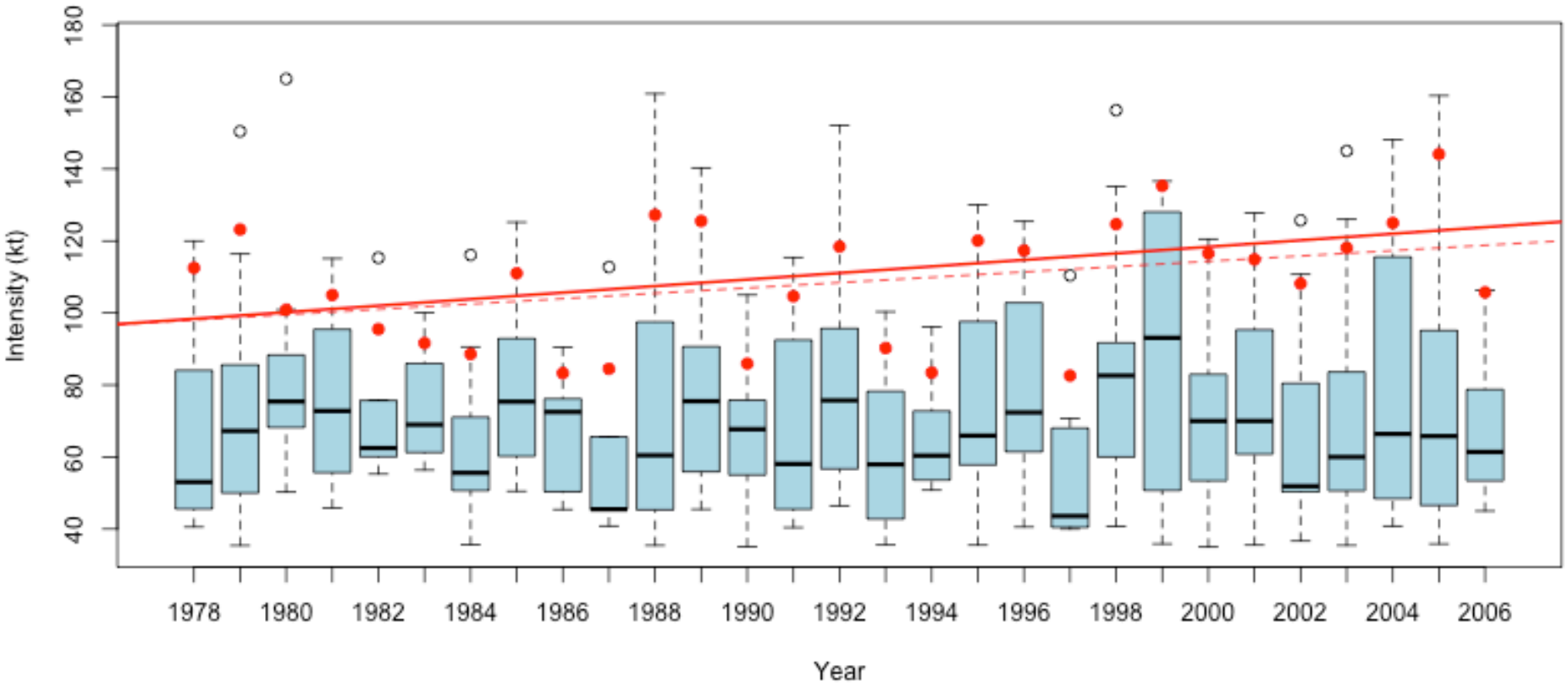
```
> Q=function(p=.9) as.vector(by(Wmax,as.factor(Yr),
+ function(x) quantile(x,p)))
```




```
> abline(lm(Q()~u))
```

Regression on quantiles

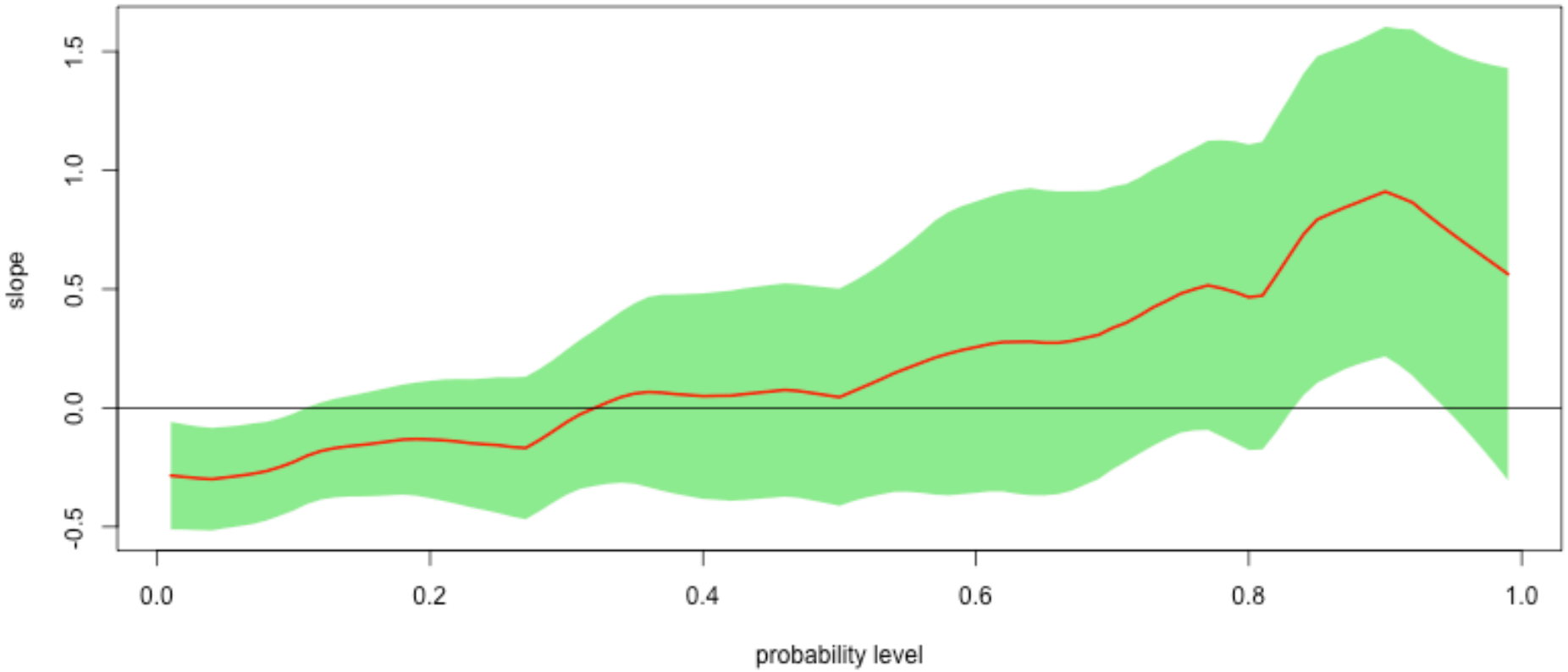
```
> Q=function(p=.9) as.vector(by(Wmax,as.factor(Yr),
+ function(x) quantile(x,p)))
```




```
> abline(lm(Q()~u,weights=table(Yr)))
```

Regression on quantiles

```
> slope=function(p){
+ return((lm(Q(p)~u,weights=table(Yr))$coefficients[2]))
+ }
```



From quantiles to quantile regression

quantiles: $F(Q(u)) = u$ i.e. $Q(u) = F^{-1}(u)$

quantiles regression: $F(Q(u|X = x)|X = x) = u$

how can we estimate $Q(u|X = x)$?

Quantiles and optimization

Proposition: $F^{-1}(u) = \underset{Q}{\operatorname{argmin}} \{\mathbb{E}[(X-Q)\cdot(u-\mathbf{1}(X<u))]\}$

$$F^{-1}(u) = \min_Q \left\{ (u-1)\int_{-\infty}^{u} (x-Q) f(x)dx + u\int_{Q}^{\infty} (x-Q) f(x)dx \right\}$$

Proof: The first order condition of this optimization problem is

$$(1-u)\underbrace{\int_{-\infty}^{Q} f(x)dx}_{u} - u\underbrace{\int_{Q}^{\infty} f(x)dx}_{(1-u)} = 0$$

which is valid since

$$(1-u)u = u(1-u)$$

Quantiles and optimization

Proposition: $F^{-1}(u) = \underset{Q}{\text{argmin}} \{\mathbb{E}[(X - Q) \cdot (u - \mathbf{1}(X < u))]\}$

Remark: If $u = \dfrac{1}{2}$ this equation if simply

$$m = F^{-1}\left(\frac{1}{2}\right) = \underset{Q}{\text{argmin}} \{\mathbb{E}(|X - Q|)\}$$

$$\text{median}(\mathcal{X}) = \underset{Q}{\text{argmin}} \left\{ \sum_{i=1}^{n} |X_i - Q| \right\}$$

Remark: The expected value is the solution of

$$\mathbb{E}(X) = \underset{Q}{\text{argmin}} \left\{ \mathbb{E}\left[(X - Q)^2\right] \right\}$$

$$\overline{X} = \underset{Q}{\text{argmin}} \left\{ \sum_{i=1}^{n} (X_i - Q)^2 \right\}$$

Quantiles and optimization

$$\text{median}(\mathcal{X}) = \underset{Q}{\text{argmin}} \left\{ \sum_{i=1}^{n} |X_i - Q| \right\}$$

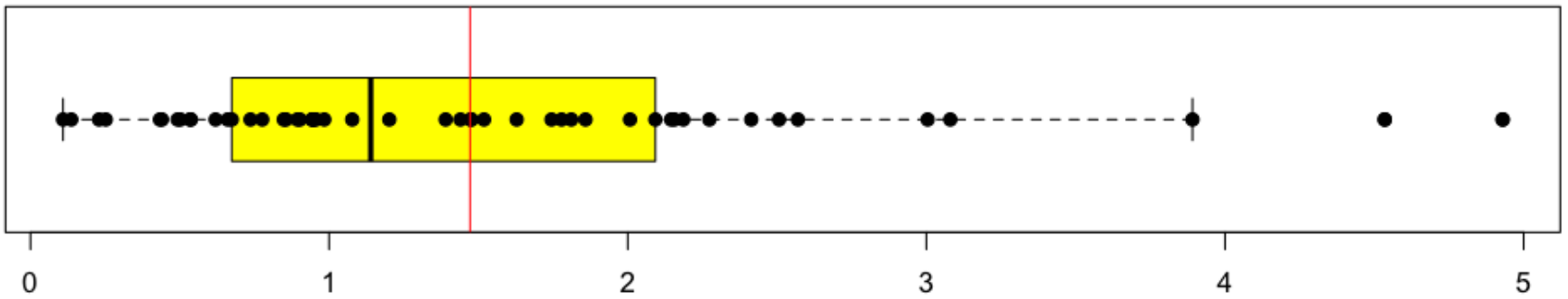
Boscovitch (1755, 1757)

Laplace (1793)

$$\overline{X} = \underset{Q}{\text{argmin}} \left\{ \sum_{i=1}^{n} (X_i - Q)^2 \right\}$$

Gauss (1809)

median($\mathcal{X}$) $\overline{X}$

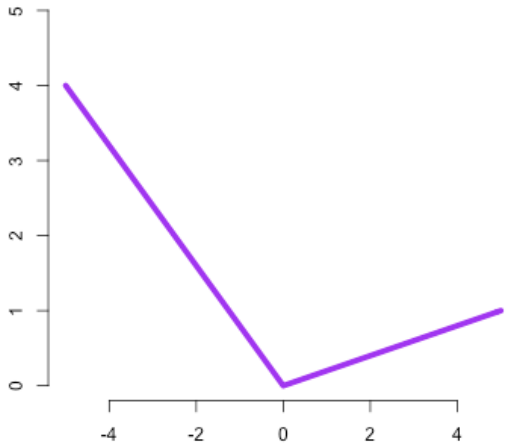


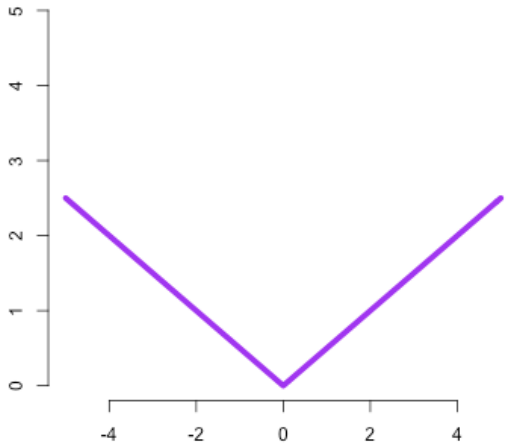
Quantiles and optimization

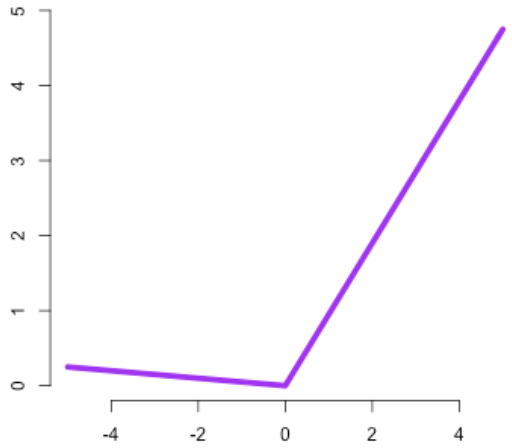
$$\text{Let } \mathcal{R}_u(x) = x \cdot (u - \mathbf{1}(x < 0)) = \begin{cases} x \cdot (u-1) \text{ pour } x < 0 \\ x \cdot u \text{ pour } x > 0 \end{cases}$$

$$\text{then } F^{-1}(u) = \underset{Q}{\text{argmin}} \left\{ \mathbb{E}\left[\mathcal{R}_u(X - Q) \right] \right\}$$

$$\textbf{Proposition: } \widehat{F}_n^{-1}(u) = \underset{Q}{\text{argmin}} \left\{ \sum_{i=1}^{n} \mathcal{R}_u(X_i - Q) \right\}$$



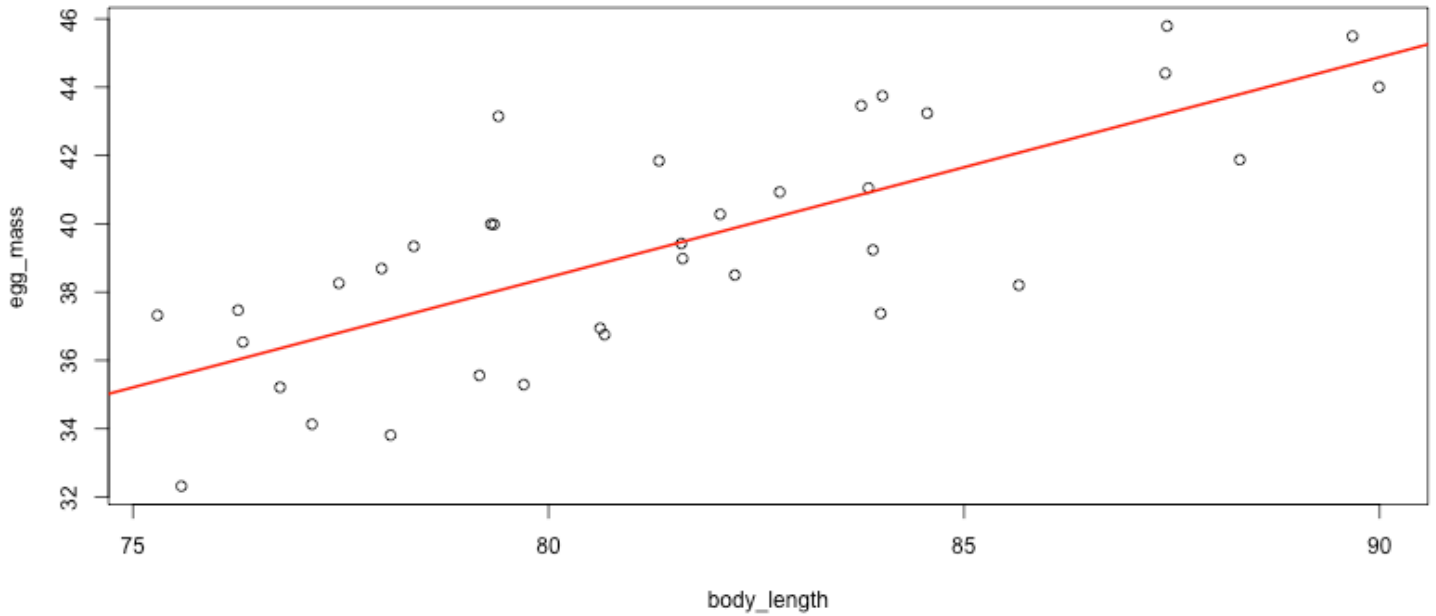




$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathcal{R}\left[Y_i - (\alpha + \beta X_i)\right] \right\}$$

$$\mathcal{R}(x) = x^2$$

least squares

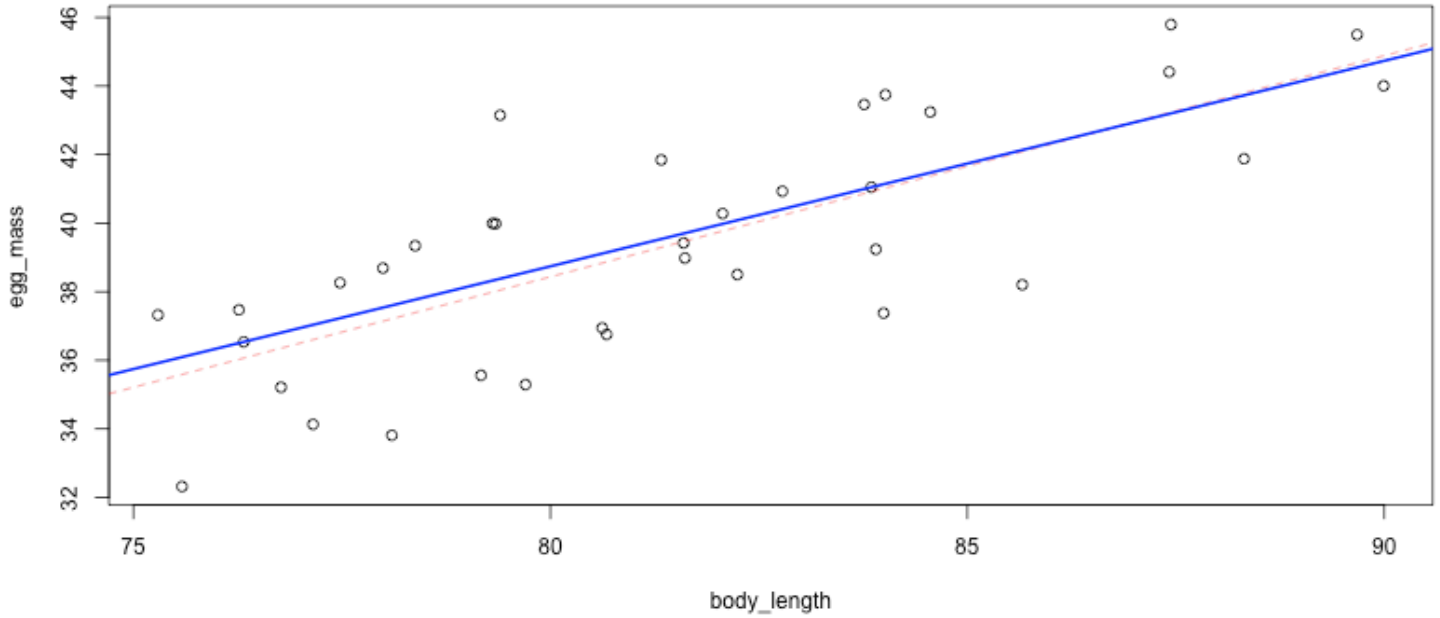



```
> abline(lm(eggmass ~bodylength),col="red",lwd=2)
```

$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathcal{R}\left[Y_i - (\alpha + \beta X_i)\right] \right\}$$

$\mathcal{R}(x) = x^2$ least squares

$\mathcal{R}(x) = |x|$ least absolute value



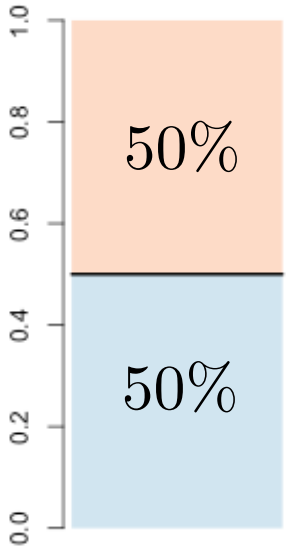

```
> abline(rq(eggmass ~bodylength,tau=.5),col="blue",lwd=2)
```

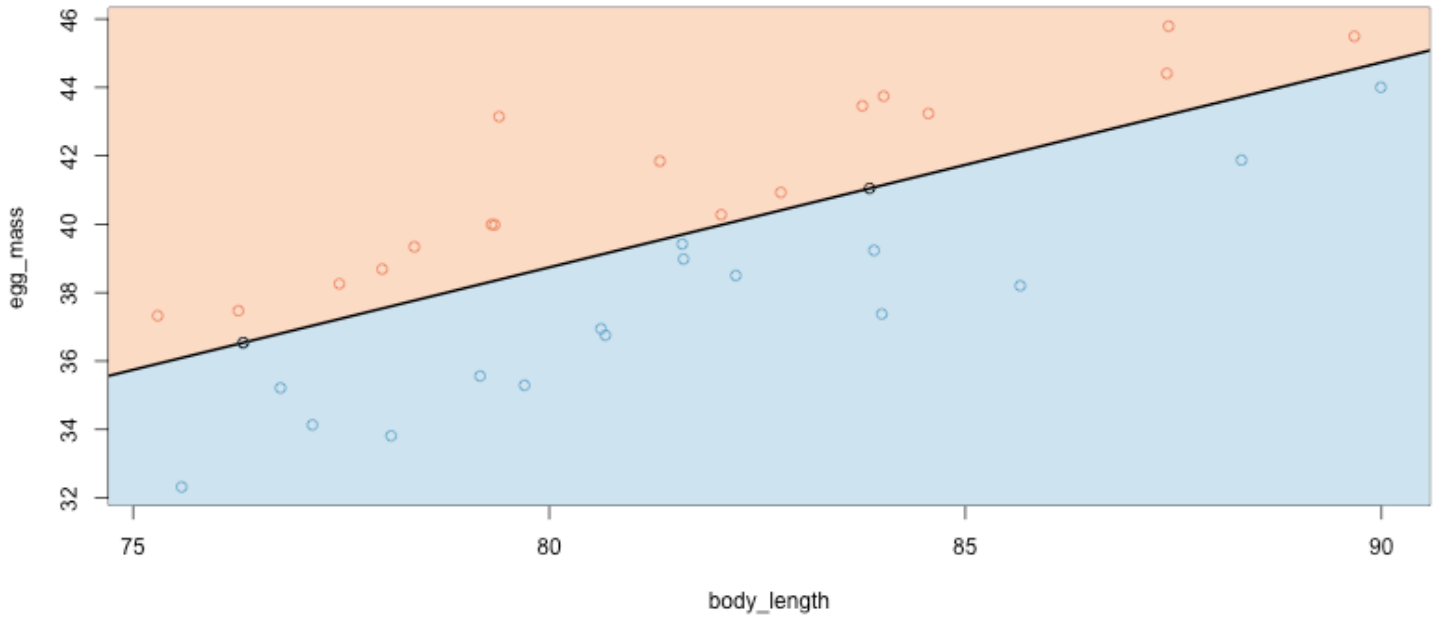
$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathcal{R}\left[Y_i - (\alpha + \beta X_i) \right] \right\}$$

$\mathcal{R}(x) = x^2$ least squares

$\mathcal{R}(x) = |x|$ least absolute value

$\mathcal{R}(x) = \mathcal{R}_u(x)$ quantile regression





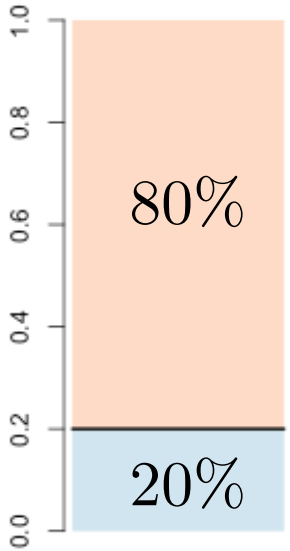

```
> abline(rq(eggmass ~bodylength,tau=.5),lwd=2)
```

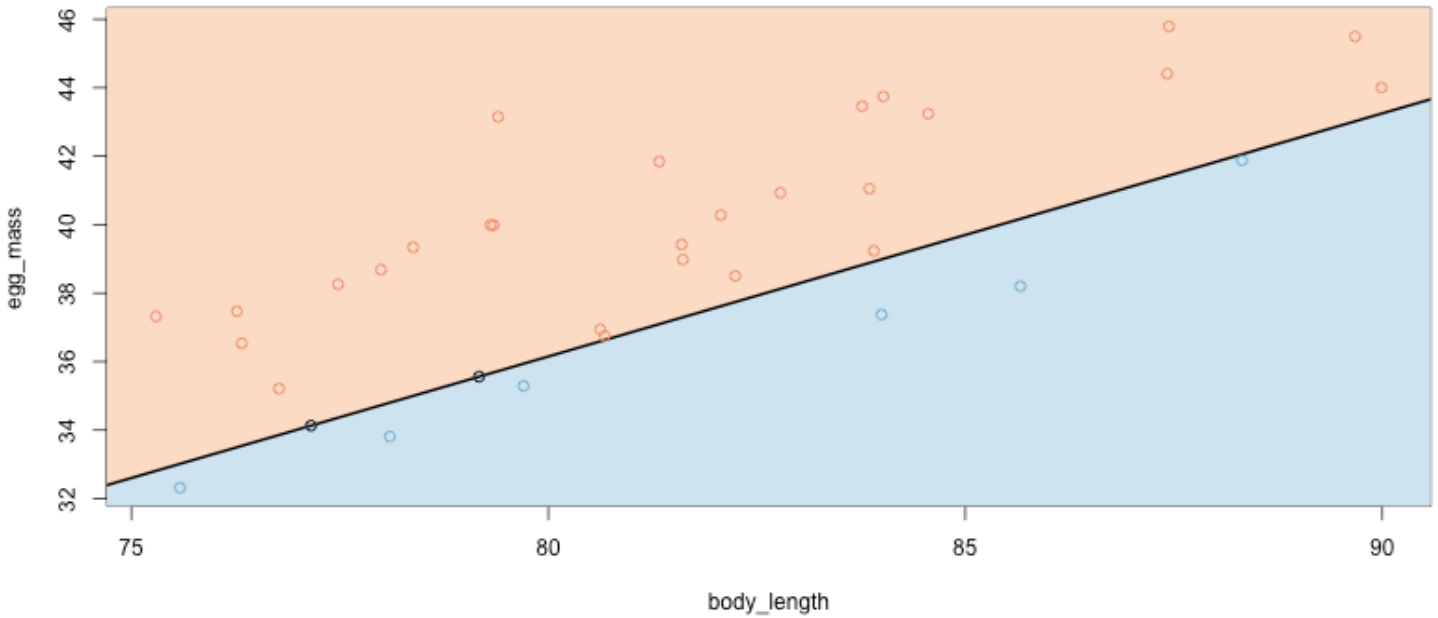
$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathcal{R}\left[Y_i - (\alpha + \beta X_i)\right] \right\}$$

$\mathcal{R}(x) = x^2$ least squares

$\mathcal{R}(x) = |x|$ least absolute value

$\mathcal{R}(x) = \mathcal{R}_u(x)$ quantile regression





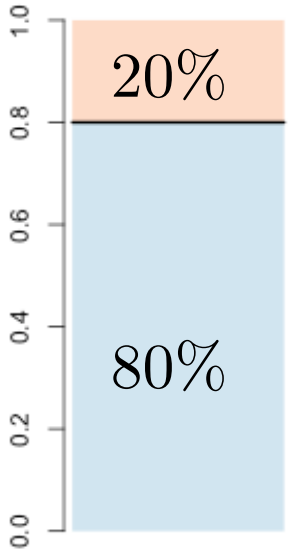

```
> abline(rq(eggmass ~bodylength,tau=.2),lwd=2)
```

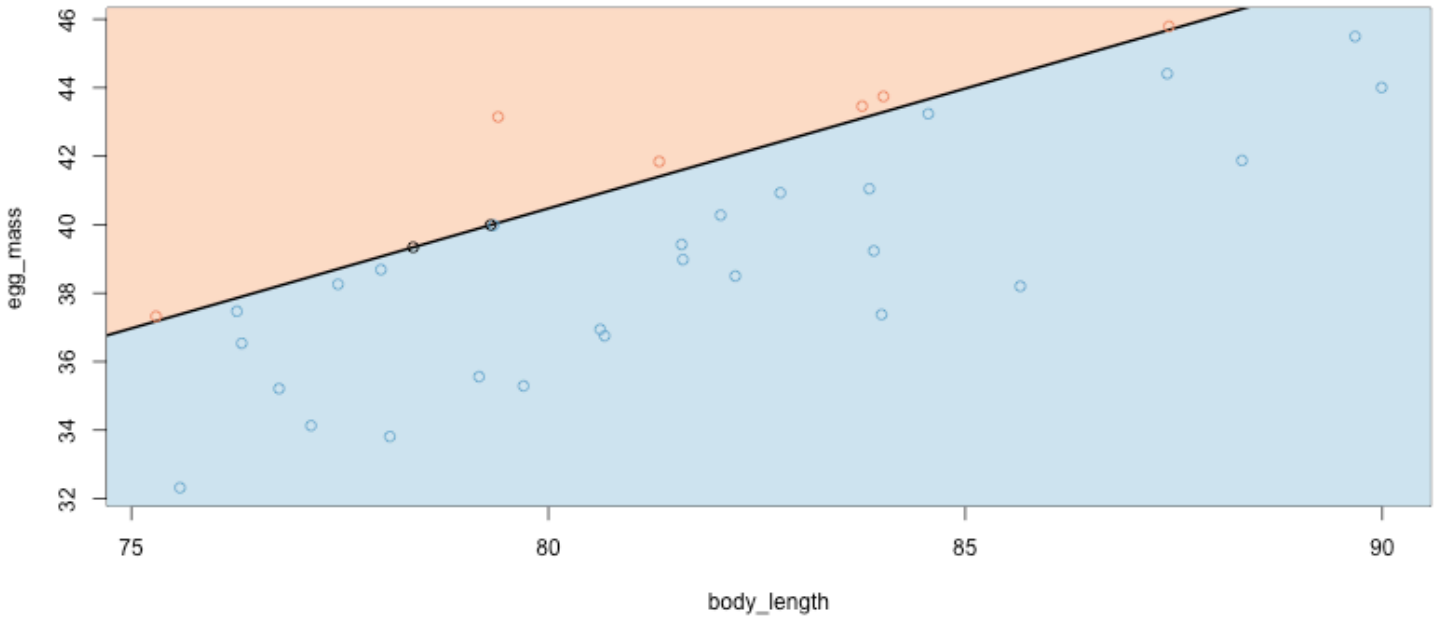
$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathcal{R}\left[Y_i - (\alpha + \beta X_i)\right] \right\}$$

$\mathcal{R}(x) = x^2$ least squares

$\mathcal{R}(x) = |x|$ least absolute value

$\mathcal{R}(x) = \mathcal{R}_u(x)$ quantile regression






```
> abline(rq(eggmass ~bodylength,tau=.8),lwd=2)
```

Regression and inference issues

$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathcal{R}\left[Y_i - (\alpha + \beta X_i)\right] \right\}$$

$$\min_{(\alpha,\beta)} \{h(\alpha,\beta)\}$$

first order conditions

$$\left.\frac{\partial}{\partial\alpha} h(\alpha,\beta)\right|_{(\alpha^\star,\beta^\star)} = \left.\frac{\partial}{\partial\beta} h(\alpha,\beta)\right|_{(\alpha^\star,\beta^\star)} = 0$$

If $\mathcal{R}(x) = x^2$, we obtain

$$\sum_{i=1}^{n} \underbrace{Y_i - [\alpha^\star + \beta^\star X_i]}_{\varepsilon_i^\star} = \sum_{i=1}^{n} X_i \left(Y_i - [\alpha^\star + \beta^\star X_i]\right) = 0$$

Regression and inference issues

$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathcal{R}\left[Y_i - (\alpha + \beta X_i)\right] \right\}$$

$$\min_{(\alpha,\beta)} \{h(\alpha,\beta)\}$$

first order conditions (if $h(\cdot)$ is differentiable)

$$\left.\frac{\partial}{\partial\alpha} h(\alpha,\beta)\right|_{(\alpha^\star,\beta^\star)} = \left.\frac{\partial}{\partial\beta} h(\alpha,\beta)\right|_{(\alpha^\star,\beta^\star)} = 0$$

If $\mathcal{R}(x) = |x|$, ...

Weighted least squares

$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathcal{R}\left[Y_i - (\alpha + \beta X_i)\right] \right\}$$

can be written

$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \omega_i [Y_i - (\alpha + \beta X_i)]^2 \right\}$$

if

$$\omega_i = \frac{\mathcal{R}[Y_i - (\alpha + \beta X_i)]}{[Y_i - (\alpha + \beta X_i)]^2}$$

Example: for least absolute-value regression, $R(x) = |x|$

$$\omega_i = \frac{1}{|Y_i - (\alpha + \beta X_i)|} = \frac{1}{|\varepsilon_i|}$$

Iterated weighted least squares

1. run a standard OLS regression $\varepsilon_i^{(0)} = Y_i - (\widehat{\alpha} + \widehat{\beta} X_i)$

2. use weights $\omega_i^{(0)} = \dfrac{1}{|\varepsilon_i^{(0)}|}$ and run a WLS regression

3. obtain *new* residuals $\varepsilon_i^{(1)}$

4. use weights $\omega_i^{(1)} = \dfrac{1}{|\varepsilon_i^{(1)}|}$ and run a WLS regression

5. obtain *new* residuals $\varepsilon_i^{(2)}$

6. use weights $\omega_i^{(2)} = \dfrac{1}{|\varepsilon_i^{(2)}|}$ and run a WLS regression

... etc.

Iterated weighted least squares

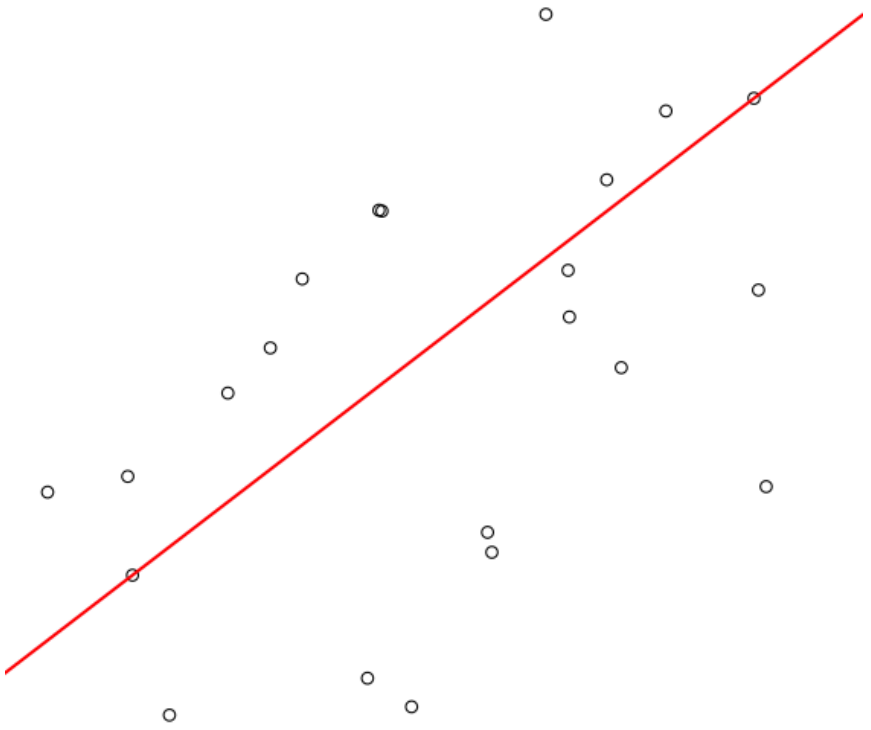
```
> REG0=lm(egg_mass~body_length)
> as.vector(REG0$coefficients)
[1] -13.1528841 0.6448454
> E=residuals(REG0)
> REG1=lm(egg_mass body_length,weight=1/abs(E))
> as.vector(REG1$coefficients)
[1] -12.348908 0.635714
> E=residuals(REG1)
> REG2=lm(egg_mass body_length,weight=1/abs(E))
> as.vector(REG2$coefficients)
[1] -11.5065491 0.6262763
```

Iterated weighted least squares

```
> library(quantreg)
> rq(egg_mass~body_length,tau=.5)
Call:
rq(formula = egg_mass~body_length, tau = 0.5)

Coefficients:
(Intercept) body_length
 -9.2309764   0.5996611

Degrees of freedom:  35 total; 33 residual
```

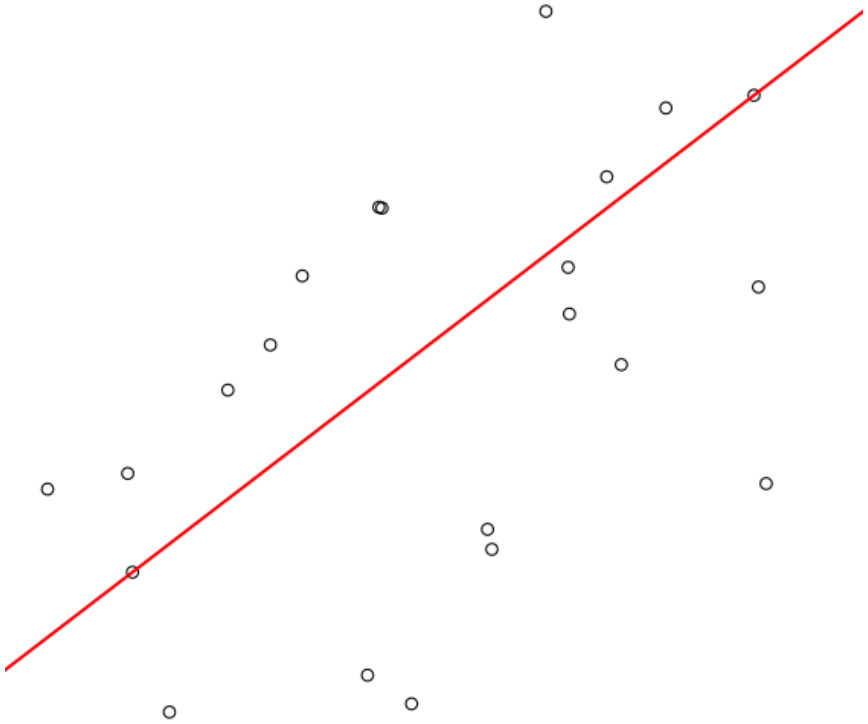
Iterated weighted least squares

```
> REG=lm(egg_mass~body_length)
> for(i in 1:200){
+ E=residuals(REG)
+ REG=lm(egg_mass~body_length,weight=1/abs(E))
+ print(as.vector(REG$coefficients))
+ }
[1] -12.348908 0.635714
[1] -11.5065491 0.6262763
[1] -10.6658488 0.6167386
[1] -9.7799162 0.6062327
[1] -9.275325 0.600209
[1] -9.233893 0.599709
[1] -9.235921 0.599729
[1] -9.2353140 0.5997189
[1] -9.2342389 0.5997041
[1] -9.2332920 0.5996915
[1] -9.2325789 0.5996821
[1] -9.2320725 0.5996754
[1] -9.2317219 0.5996708
[1] -9.2314820 0.5996677
```



Iterated weighted least squares

```
[1] -9.2309764 0.5996611
[1] -9.2309764 0.5996611
[1] 36.53687       NA
[1] 12.0522875 0.3208028
[1] 5.634130 0.404895
[1] -0.09161996 0.47991507
[1] -2.7181723 0.5143288
[1] -5.3679027 0.5490462
[1] -5.4819086 0.5505399
[1] -5.653411 0.552787
[1] -5.8988873 0.5560032
[1] -6.2298597 0.5603397
[1] -6.6474831 0.5658115
[1] -7.1397961 0.5722619
[1] -7.682229 0.579369
[1] 36.53687       NA
[1] 12.0522434 0.3208033
[1] 5.6342425 0.4048935
[1] -0.09155391 0.47991421
```

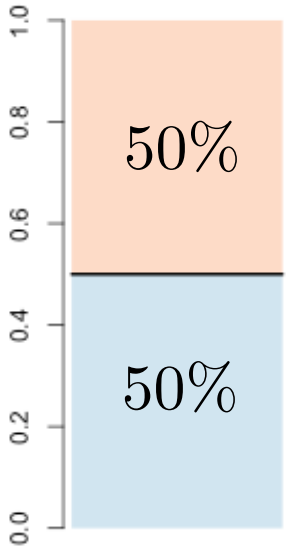


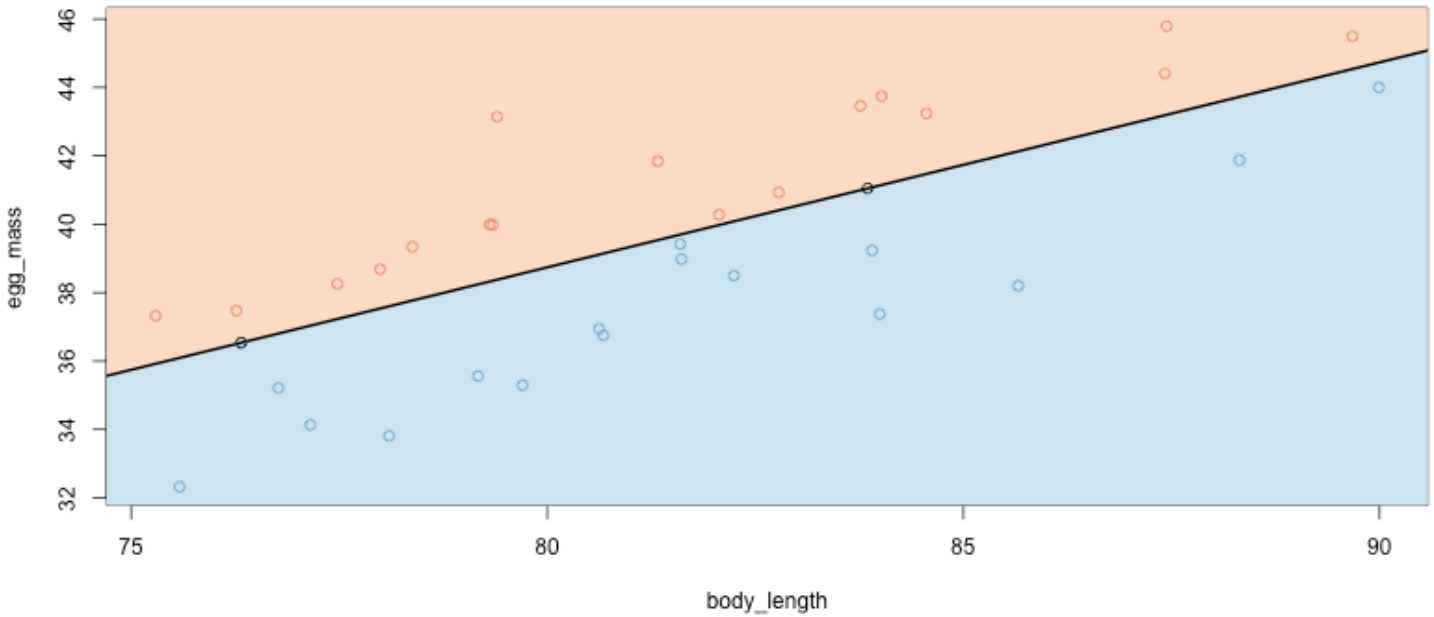
Iterated weighted least squares

```
> poids=function(e) pmin(1/abs(e),10000)
> REG=lm(egg_mass~body_length)
> for(i in 1:200){
+ E=residuals(REG)
+ REG=lm(egg_mass~body_length,weight=poids(E))
+ }
[1] -12.348908 0.635714
[1] -11.5065491 0.6262763
[1] -10.6658488 0.6167386
[1] -9.7799162 0.6062327
[1] -9.275325 0.600209
[1] -9.233893 0.599709
[1] -9.235921 0.599729
[1] -9.2353140 0.5997189
[1] -9.2341642 0.5997033
[1] -9.2330243 0.5996883
[1] -9.2322484 0.5996782
[1] -9.2317216 0.5996712
```

Quantile regression on the salmon dataset

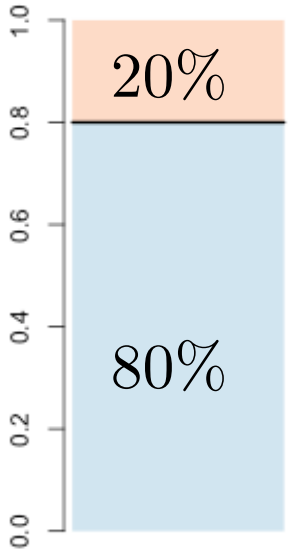
```
> abline(rq(eggmass ~bodylength,tau=.5),lwd=2)
```

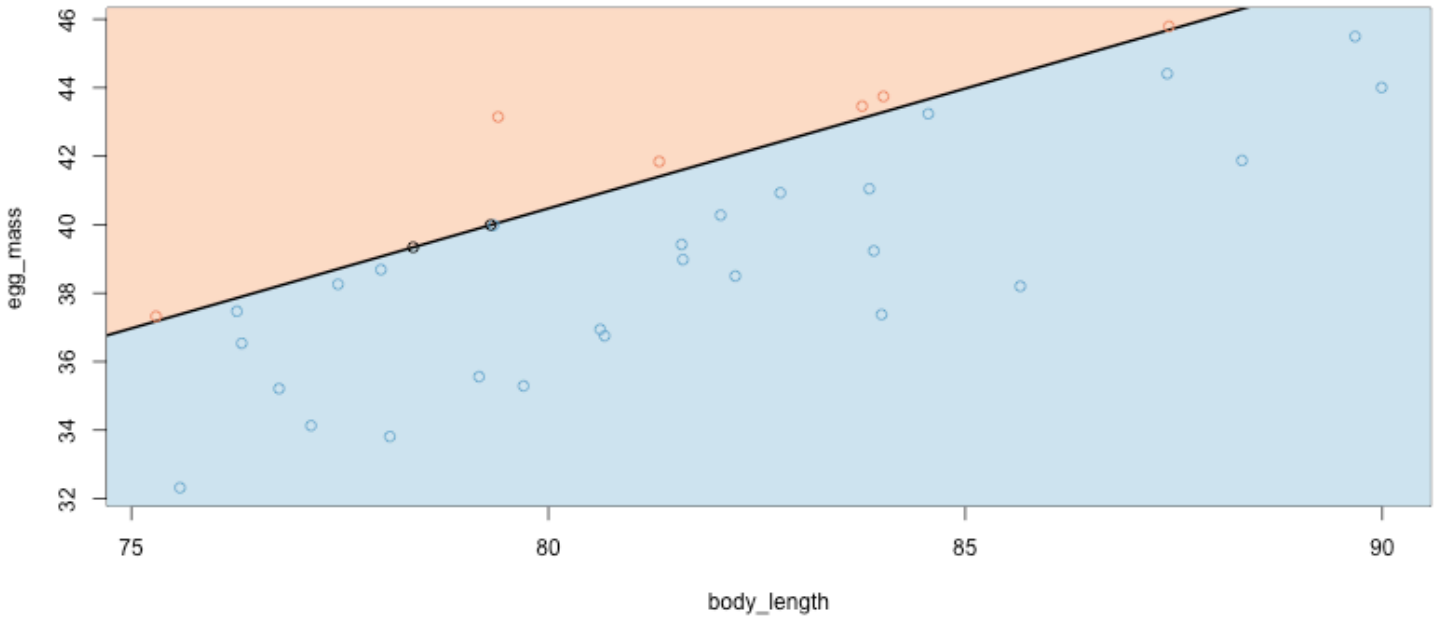




Quantile regression on the salmon dataset

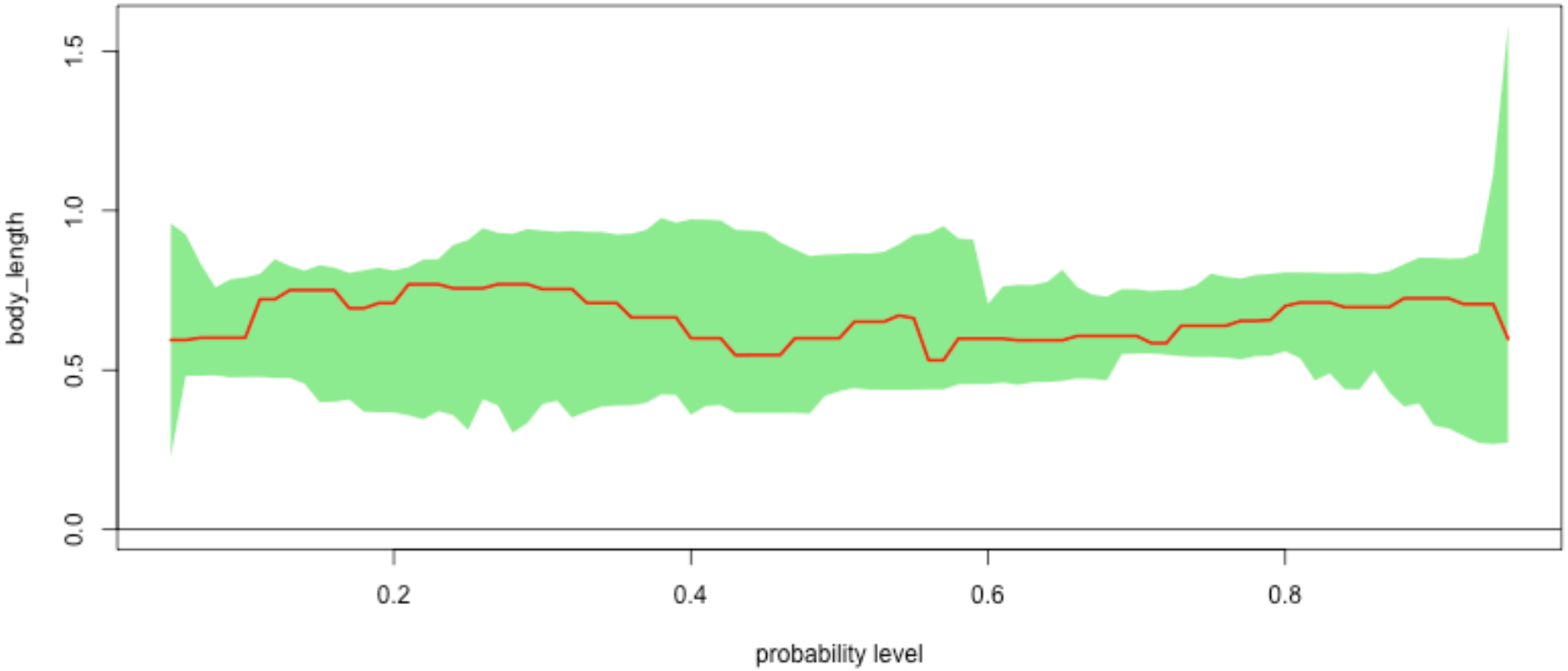
```
> abline(rq(eggmass ~bodylength,tau=.8),lwd=2)
```



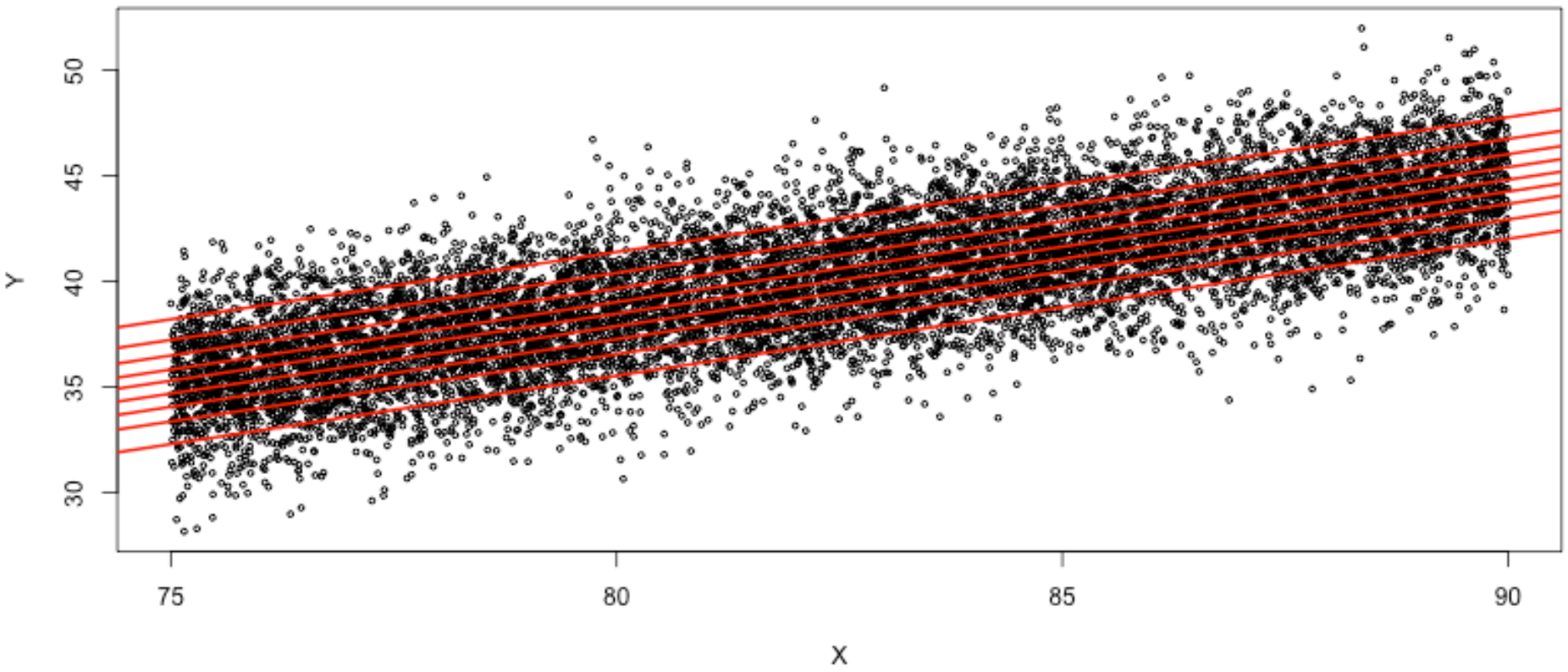


Quantile regression on the salmon dataset

```
> u=seq(.05,.95,by=.01)
> coefsup=function(u) summary(rq(egg_mass~body_length,tau=u))$coefficients[,3]
> coefinf=function(u) summary(rq(egg_mass~body_length,tau=u))$coefficients[,2]
> CS=Vectorize(coefsup)(u)
> CI=Vectorize(coefinf)(u)
> polygon(c(u,rev(u)),c(CS[k,],rev(CI[k,])),col="light green",border=NA)
```

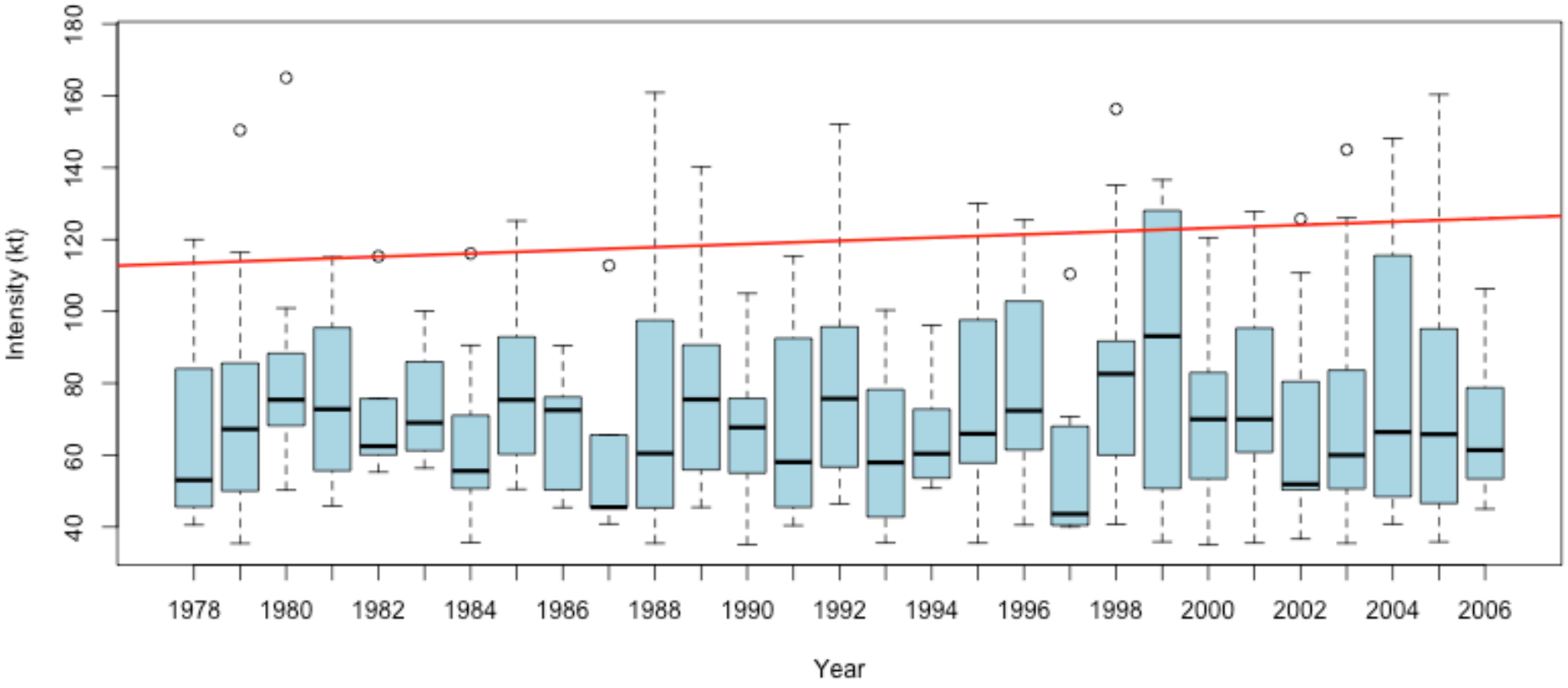


Quantile regression in a Gaussian model



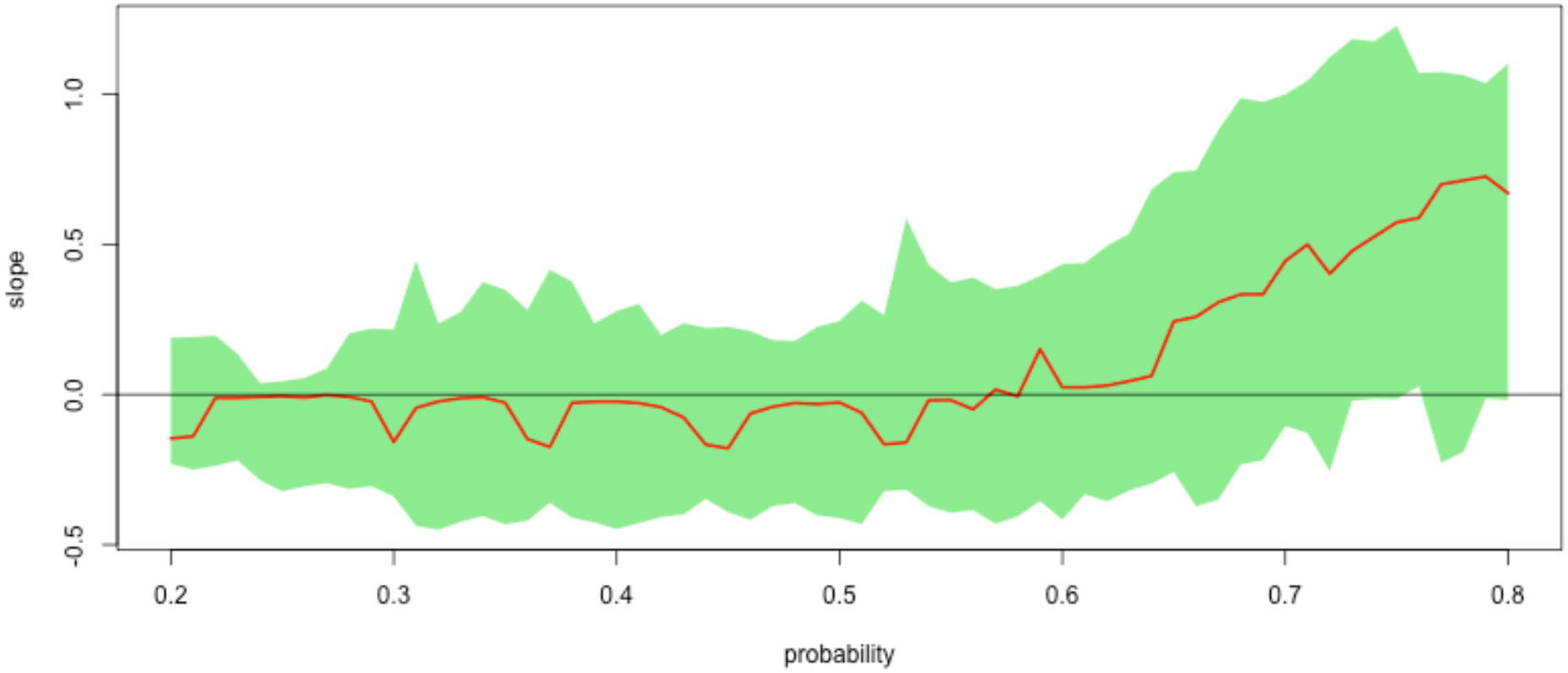
Quantile regression and hurricanes

```
> rq(Wmax~Yr,tau=.9)
```

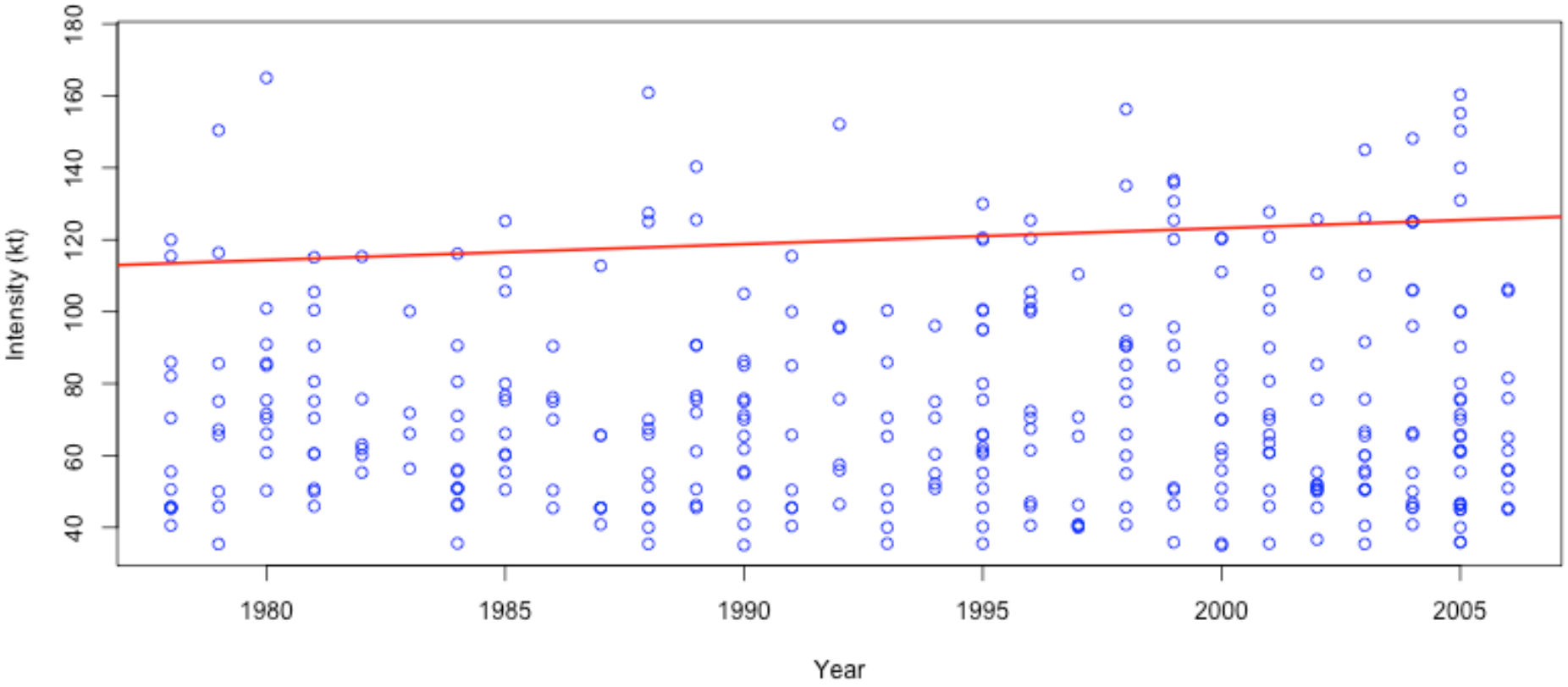


Quantile regression and hurricanes

```
> u=seq(.05,.95,by=.01)
> coefsup=function(u) summary(rq(Wmax~Yr,tau=u))$coefficients[,3]
> coefinf=function(u) summary(rq(Wmax~Yr,tau=u))$coefficients[,2]
> CS=Vectorize(coefsup)(u)
> CI=Vectorize(coefinf)(u)
> polygon(c(u,rev(u)),c(CS[k,],rev(CI[k,])),col="light green",border=NA)
```



Quantile regression and hurricanes



Quantile regression and obesity studies

Beyerlein, A. von Kries, R., Ness, A.R. & Ong, K.K. (2011). Genetic Markers of Obesity Risk: Stronger Associations with Body Composition in Overweight Compared to Normal-Weight Children. *PLoS ONE* **6** (4): e19057

Beyerlein, A. Toschke, A.M. & von Kries (2008) Breastfeeding and Childhood Obesity: Shift of the Entire BMI Distribution or Only the Upper Parts? *Obesity* **16** 12, 2730–2733

Determinants of Infant Birthweight

Abreveya, J. (2001) The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, **26**, 247-257.

Determinants of Infant Birthweight

Abreveya, J. (2001) The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, **26**, 247-257.

```
> base=read.table(
+ "http://freakonometrics.free.fr/natality2005.txt",
+ header=TRUE,sep";")
> head(base)
```

	WEIGHT	SEX	MARRIED	RACEM	RACEF	EDUCM	SMOKER	WEIGHTGAIN	BIRTHRECORD	WHITEM
1	3714	M	1	4	1	12	FALSE	46	1	FALSE
2	4026	M	1	1	1	14	FALSE	43	7	TRUE
3	3515	F	1	3	2	12	FALSE	62	3	FALSE
4	3260	F	1	3	3	08	FALSE	9	1	FALSE
5	4252	M	1	3	1	14	FALSE	30	3	FALSE
6	3515	F	1	3	9	15	TRUE	16	2	FALSE

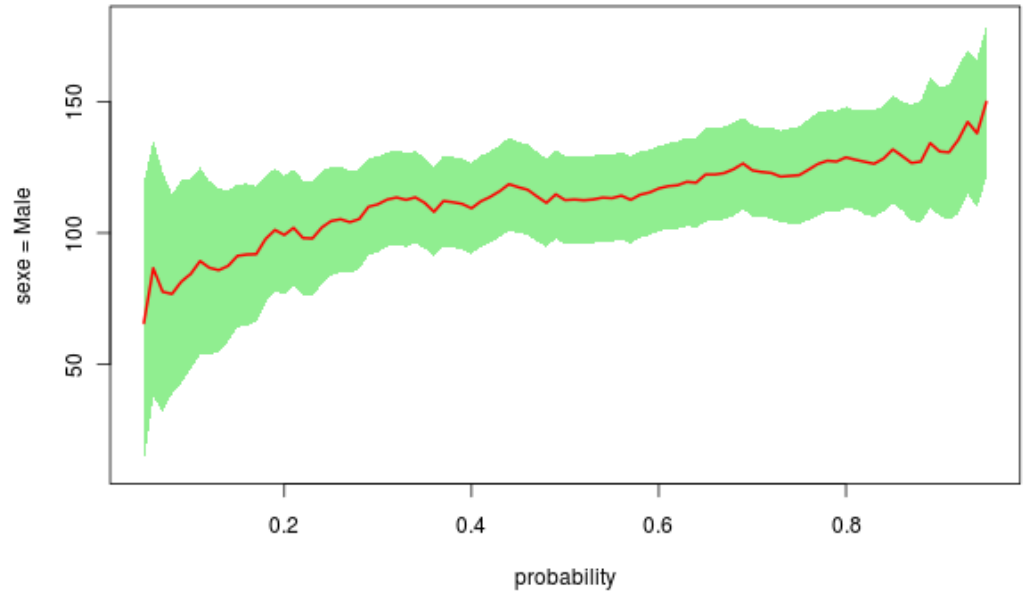
	BLACKM	AMERINDIANM	ASIANM	WHITEF	BLACKF	AMERINDIANF	ASIANF	COLLEGE	AGE
1	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	22
2	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	37
3	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	23
4	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	20
5	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	29
6	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	28

Determinants of Infant Birthweight

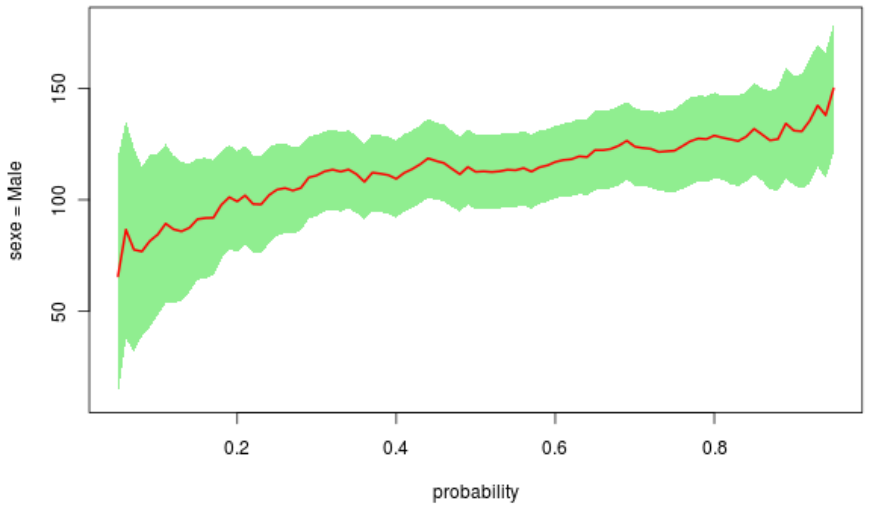
```
> summary(rq(WEIGHT∼SEX+SMOKER+WEIGHTGAIN+
+ BIRTHRECORD+AGE+ BLACKM+ BLACKF+COLLEGE,data=base,tau=.8))

Call:  rq(formula = WEIGHT   SEX + SMOKER + WEIGHTGAIN + BIRTHRECORD +
 AGE + BLACKM + BLACKF + COLLEGE, tau = 0.8, data = base)

tau:  [1] 0.8

Coefficients:
               Value  Std.  Error    t value    Pr(>|t|)
(Intercept) 3702.70107   54.10183   68.43947     0.00000
SEXM         128.07236   20.09695    6.37272     0.00000
SMOKERTRUE  -186.02610   28.54802   -6.51625     0.00000
WEIGHTGAIN     2.35469    0.59457    3.96033     0.00008
BIRTHRECORD   28.07473    5.75009    4.88249     0.00000
AGE           -0.07592    2.05313   -0.03698     0.97050
BLACKMTRUE  -292.16370   52.21090   -5.59584     0.00000
BLACKFTRUE  -134.62278   33.24923   -4.04890     0.00005
COLLEGETRUE   10.40688   22.75930    0.45726     0.64751
```

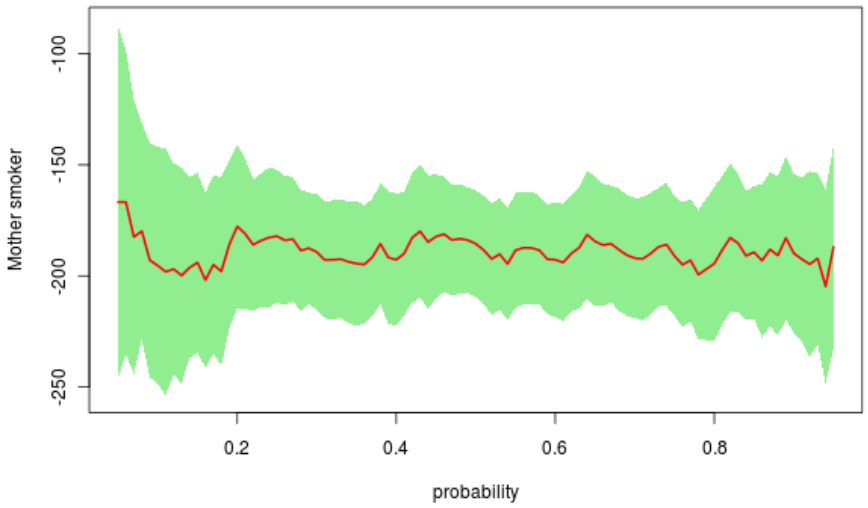
Determinants of Infant Birthweight

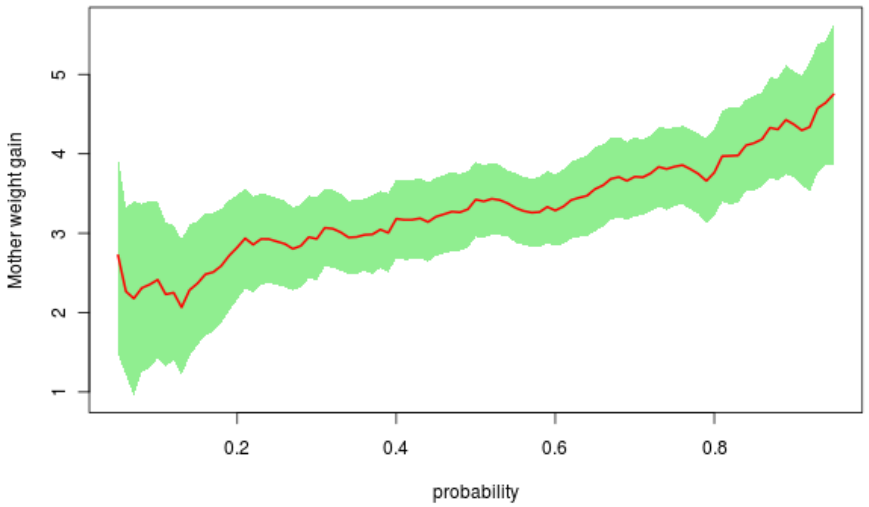
```
> base=base[1:5000,]
> u=seq(.05,.95,by=.01)
> coefstd=function(u) summary(rq(WEIGHT~SEX+SMOKER+WEIGHTGAIN+
+ BIRTHRECORD+AGE+ BLACKM+ BLACKF+COLLEGE,data=base,tau=u))$coefficients[,2]
> coefest=function(u) summary(rq(WEIGHT~SEX+SMOKER+WEIGHTGAIN+
+ BIRTHRECORD+AGE+ BLACKM+ BLACKF+COLLEGE,data=base,tau=u))$coefficients[,1]
> CS=Vectorize(coefsup)(u)
> CE=Vectorize(coefest)(u)
> k=2
> plot(u,CE[k,])
> polygon(c(u,rev(u)),c(CE[k,]+1.96*CS[k,],rev(CE[k,]-1.96*CS[k,])))
```

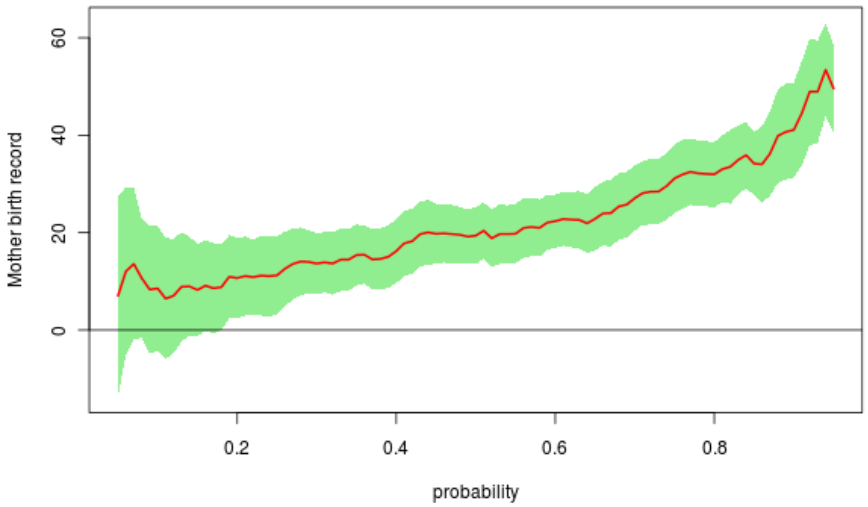


Determinants of Infant Birthweight

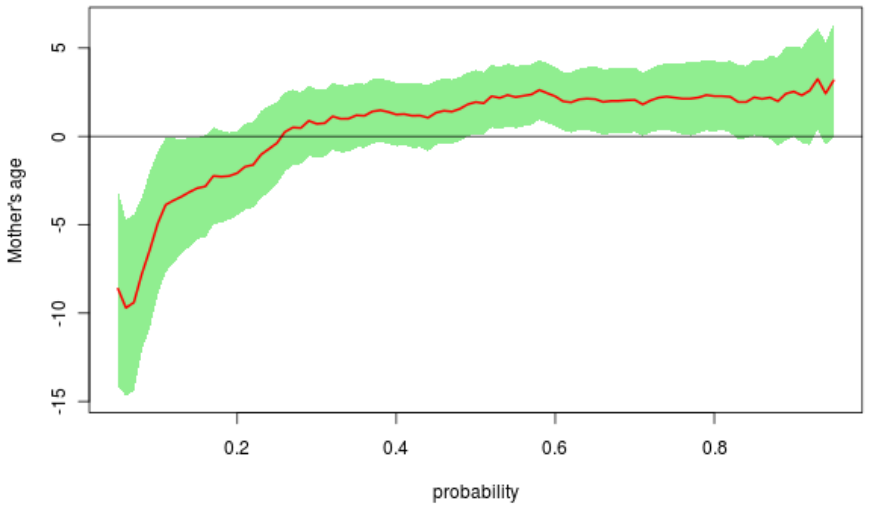


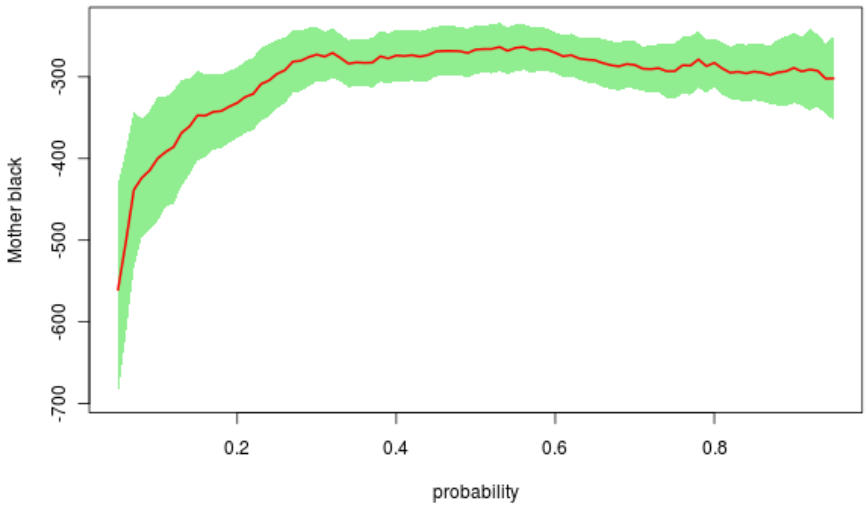


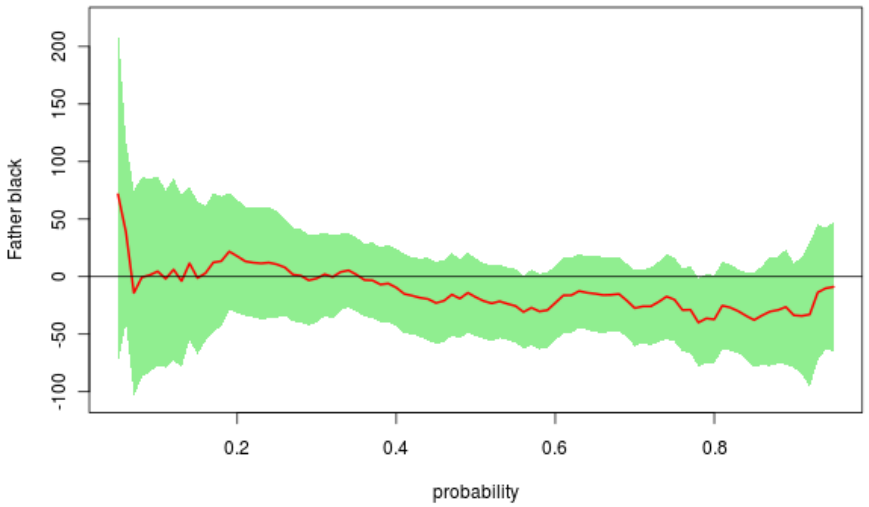


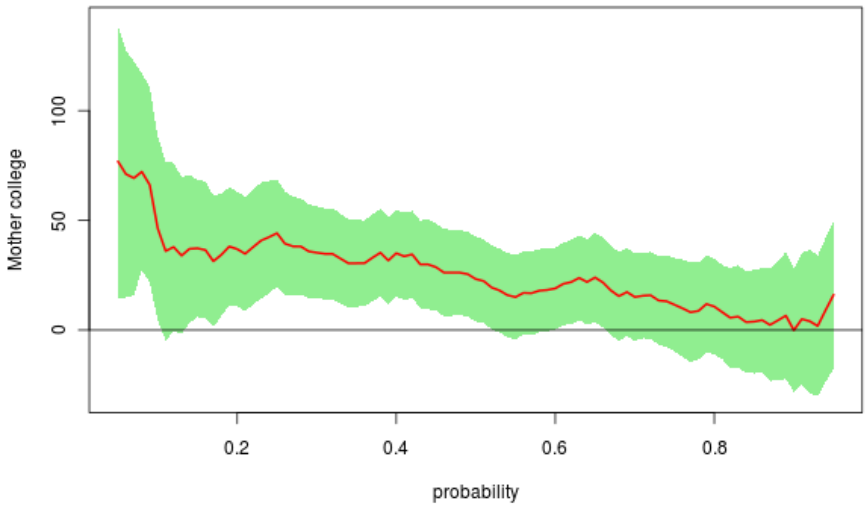


Determinants of Infant Birthweight



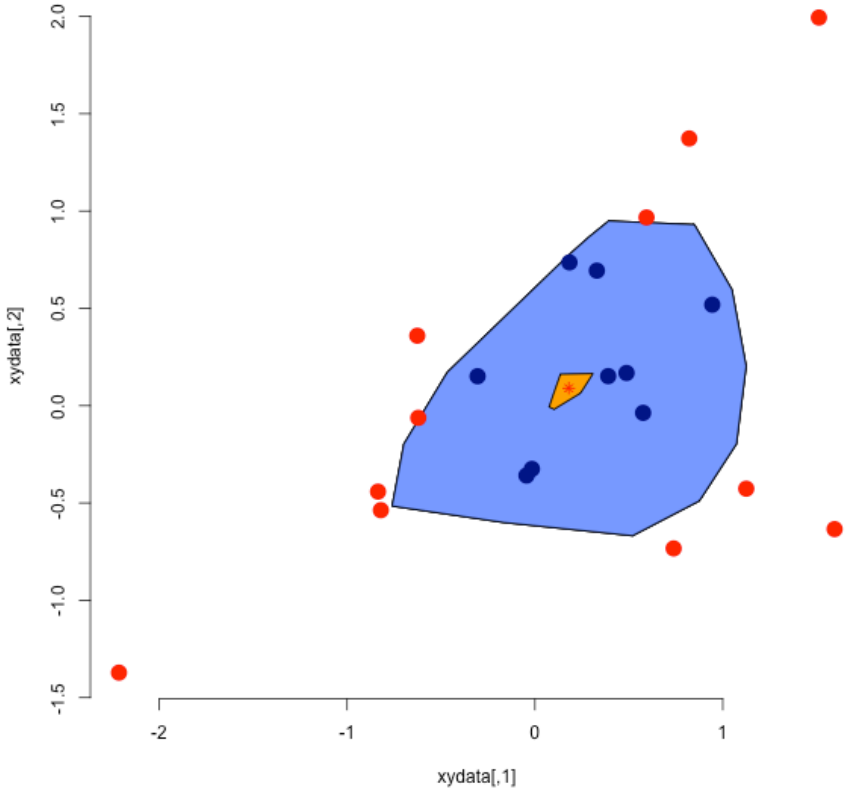


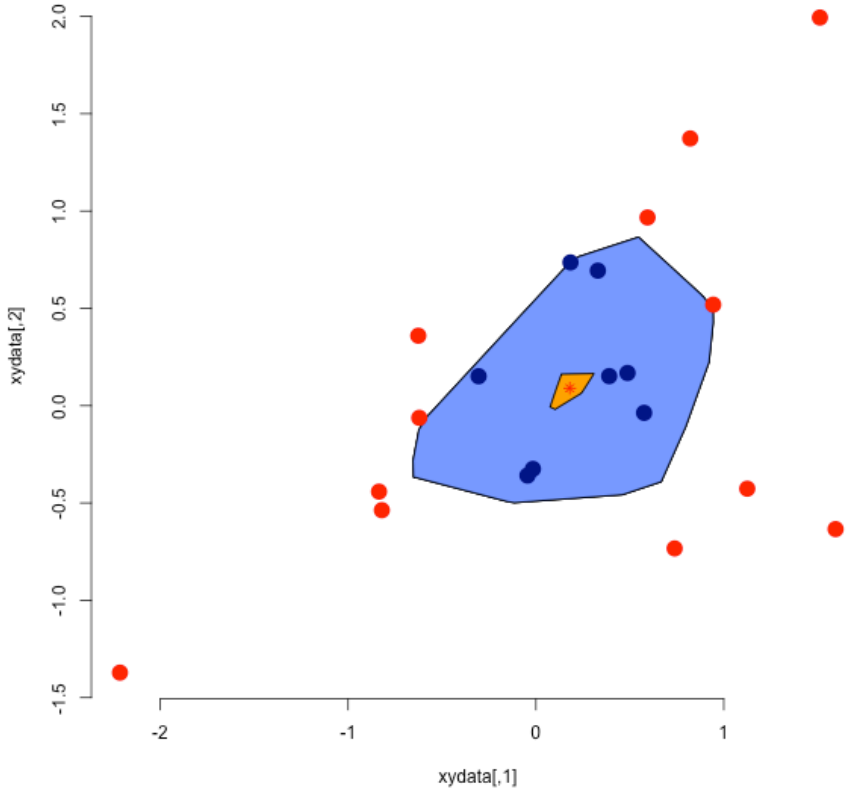




From boxplots to bagplots

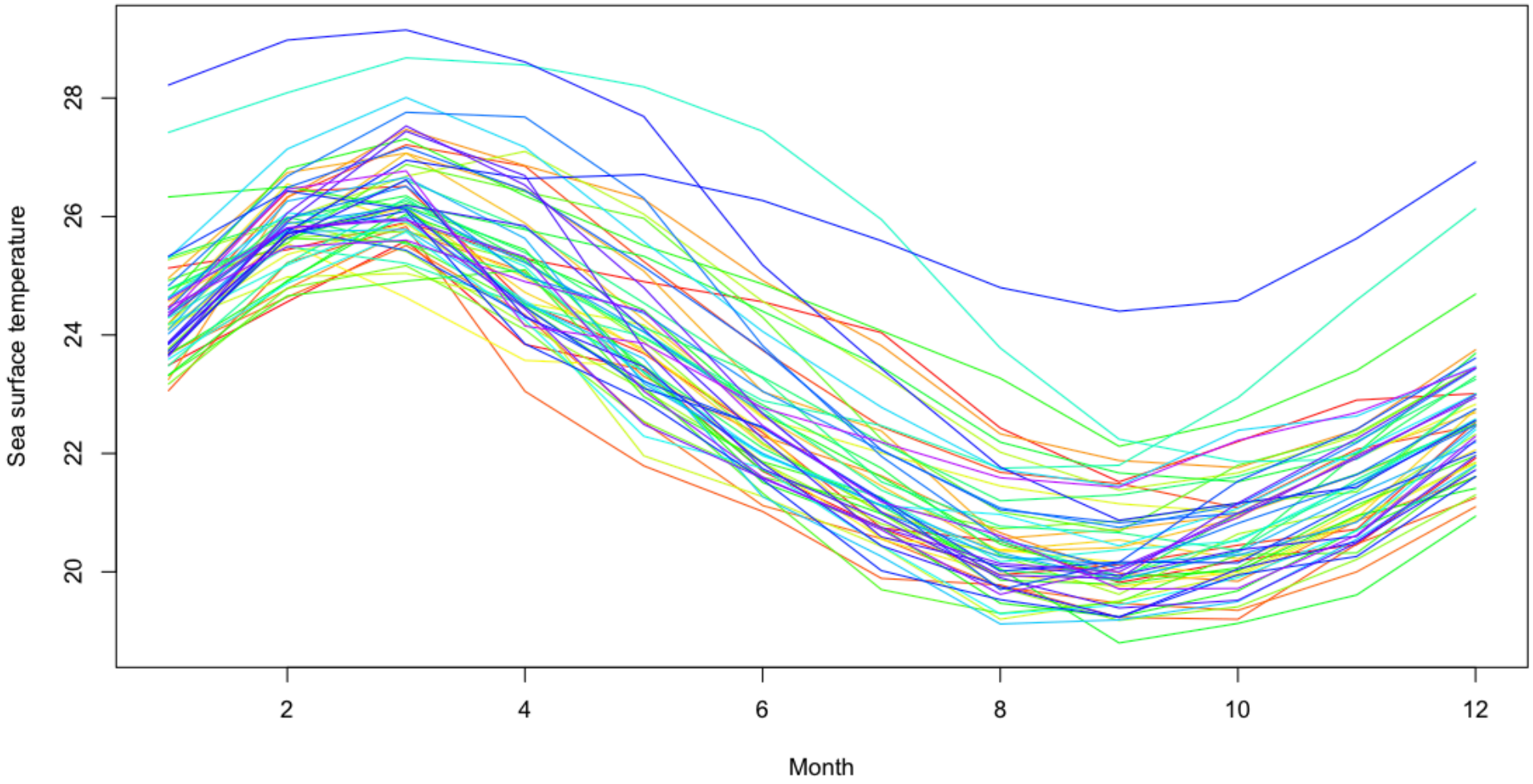
```
> library(mnormt)
> Z=rmnorm(20,mean=c(0,0),varcov=matrix(c(1,.6,.6,1),2,2))
> source("http://www.wiwi.uni-bielefeld.de/ wolf/software/R-wtools/bagplot/bagplot.R")
> bagplot(Z,factor=1,dkmethod=1)
> bagplot(Z,factor=1,dkmethod=2)
```





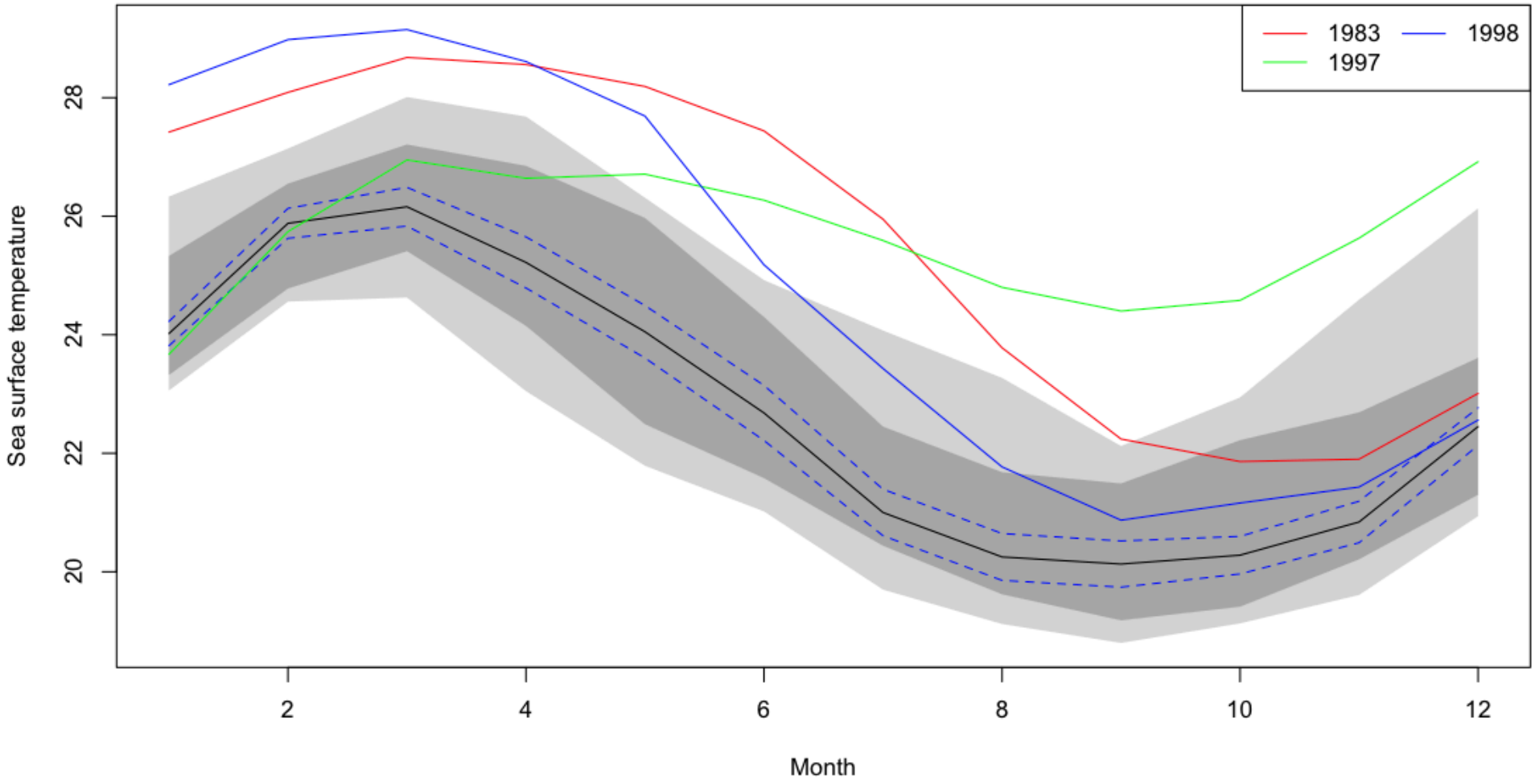
Outliers and curves

```
> library(rainbow))
> plot(ElNino))
```



Outliers and curves

```
> fboxplot(data=ElNino, plot.type="functional", type="bag")
```



Outliers and curves

```
> ElNino
Sliced functional time series
 y:  Sea surface temperature
 x:  Month
> ElNino$y
    1950  1951  1952  1953  1954  1955  1956  1957  1958  1959  1960  1961  196
1  23.49 25.13 24.47 24.19 23.06 23.70 23.67 23.25 24.96 24.09 24.57 24.45 24.1
2  24.56 25.44 26.42 26.33 25.22 24.64 24.78 26.32 26.74 25.68 25.72 26.55 25.5
3  25.59 25.90 26.51 27.21 25.83 25.51 25.76 27.47 27.07 27.07 26.15 25.91 24.6
4  23.84 25.29 24.54 26.85 23.05 24.34 25.06 26.86 26.45 25.89 24.72 24.97 23.5
5  23.41 24.90 23.69 25.24 21.79 22.53 23.30 26.29 25.08 24.26 23.76 23.71 23.4
6  21.70 24.56 22.35 23.74 21.02 21.12 22.29 24.92 23.04 22.63 22.15 22.39 21.8
7  20.74 24.04 20.74 22.45 19.89 20.56 21.62 23.82 22.38 21.45 21.01 20.55 20.5
8  20.52 22.43 19.95 21.68 19.78 19.75 20.66 22.33 20.56 20.34 20.37 19.96 20.3
9  19.82 21.52 20.12 21.49 19.23 19.47 19.98 21.88 20.72 20.41 20.54 19.75 20.1
10 20.19 22.20 20.45 21.08 19.20 19.35 19.84 21.76 20.96 20.91 20.22 20.00 19.9
11 20.46 22.90 20.72 22.07 20.47 20.00 20.86 22.40 21.64 21.99 20.91 21.07 20.8
12 21.92 23.01 22.53 22.47 21.25 21.10 21.60 23.75 22.47 22.69 22.74 22.06 21.8
```

$X_{i,t}$

Outliers and curves

```
> ElNino
Sliced functional time series
 y:  Sea surface temperature
 x:  Month
> ElNino$y
```

$X_{i,t}$

year

	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	196
1	23.49	25.13	24.47	24.19	23.06	23.70	23.67	23.25	24.96	24.09	24.57	24.45	24.1
2	24.56	25.44	26.42	26.33	25.22	24.64	24.78	26.32	26.74	25.68	25.72	26.55	25.5
3	25.59	25.90	26.51	27.21	25.83	25.51	25.76	27.47	27.07	27.07	26.15	25.91	24.6
4	23.84	25.29	24.54	26.85	23.05	24.34	25.06	26.86	26.45	25.89	24.72	24.97	23.5
5	23.41	24.90	23.69	25.24	21.79	22.53	23.30	26.29	25.08	24.26	23.76	23.71	23.4
6	21.70	24.56	22.35	23.74	21.02	21.12	22.29	24.92	23.04	22.63	22.15	22.39	21.8
7	20.74	24.04	20.74	22.45	19.89	20.56	21.62	23.82	22.38	21.45	21.01	20.55	20.5
8	20.52	22.43	19.95	21.68	19.78	19.75	20.66	22.33	20.56	20.34	20.37	19.96	20.3
9	19.82	21.52	20.12	21.49	19.23	19.47	19.98	21.88	20.72	20.41	20.54	19.75	20.1
10	20.19	22.20	20.45	21.08	19.20	19.35	19.84	21.76	20.96	20.91	20.22	20.00	19.9
11	20.46	22.90	20.72	22.07	20.47	20.00	20.86	22.40	21.64	21.99	20.91	21.07	20.8
12	21.92	23.01	22.53	22.47	21.25	21.10	21.60	23.75	22.47	22.69	22.74	22.06	21.8

Outliers and curves

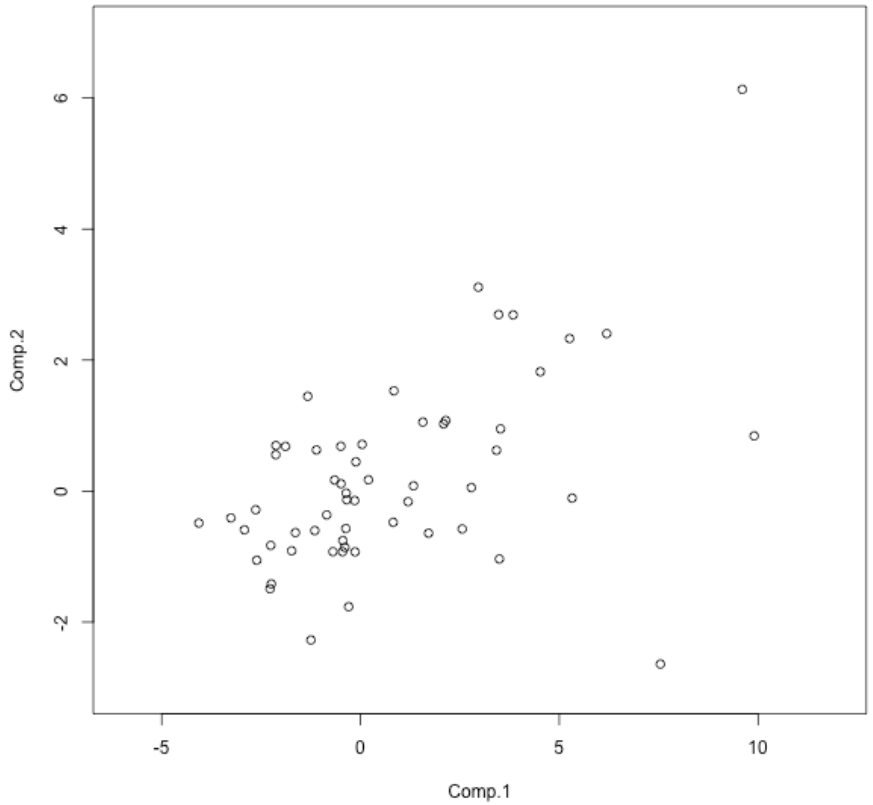
```
> ElNino
Sliced functional time series
 y:  Sea surface temperature
 x:  Month
> ElNino$y
    1950  1951  1952  1953  1954  1955  1956  1957  1958  1959  1960  1961  196
1  23.49 25.13 24.47 24.19 23.06 23.70 23.67 23.25 24.96 24.09 24.57 24.45 24.1
2  24.56 25.44 26.42 26.33 25.22 24.64 24.78 26.32 26.74 25.68 25.72 26.55 25.5
3  25.59 25.90 26.51 27.21 25.83 25.51 25.76 27.47 27.07 27.07 26.15 25.91 24.6
4  23.84 25.29 24.54 26.85 23.05 24.34 25.06 26.86 26.45 25.89 24.72 24.97 23.5
5  23.41 24.90 23.69 25.24 21.79 22.53 23.30 26.29 25.08 24.26 23.76 23.71 23.4
6  21.70 24.56 22.35 23.74 21.02 21.12 22.29 24.92 23.04 22.63 22.15 22.39 21.8
7  20.74 24.04 20.74 22.45 19.89 20.56 21.62 23.82 22.38 21.45 21.01 20.55 20.5
8  20.52 22.43 19.95 21.68 19.78 19.75 20.66 22.33 20.56 20.34 20.37 19.96 20.3
9  19.82 21.52 20.12 21.49 19.23 19.47 19.98 21.88 20.72 20.41 20.54 19.75 20.1
10 20.19 22.20 20.45 21.08 19.20 19.35 19.84 21.76 20.96 20.91 20.22 20.00 19.9
11 20.46 22.90 20.72 22.07 20.47 20.00 20.86 22.40 21.64 21.99 20.91 21.07 20.8
12 21.92 23.01 22.53 22.47 21.25 21.10 21.60 23.75 22.47 22.69 22.74 22.06 21.8
```

$X_{i,t}$

Outliers and curves

principal component analysis on $\boldsymbol{X} = (X_{i,t})$

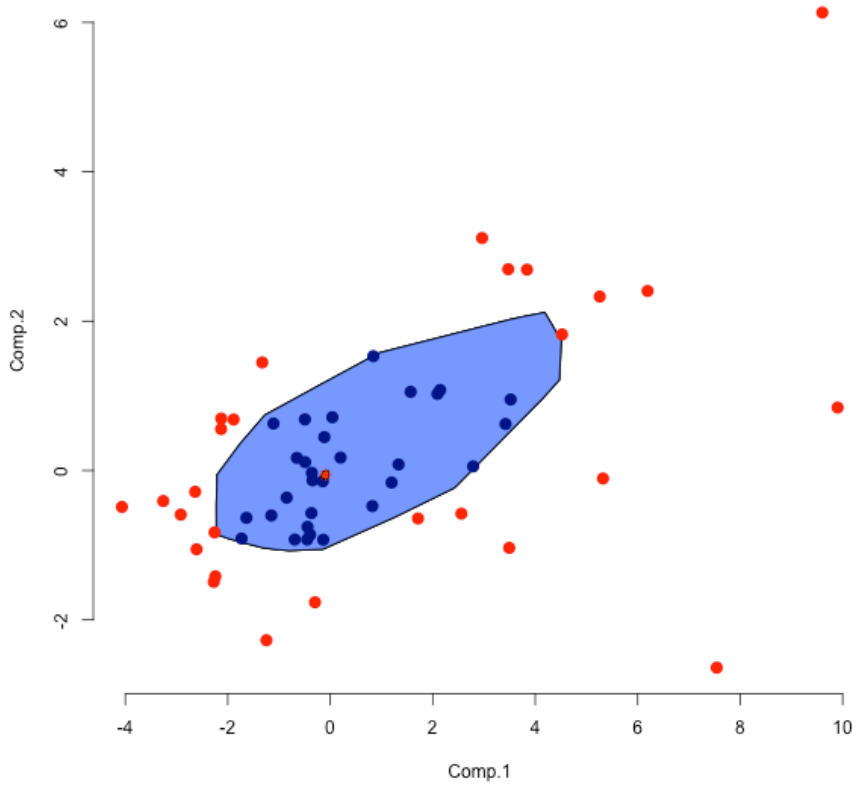
```
> (PCA=PCAproj(t(ElNino$y), center = median))
Call:
PCAproj(x = t(ElNino$y), center = median)

Standard deviations:
  Comp.1   Comp.2
2.712685 1.170426

 12 variables and 57 observations.
> plot(PCA$scores)
```



Outliers and curves

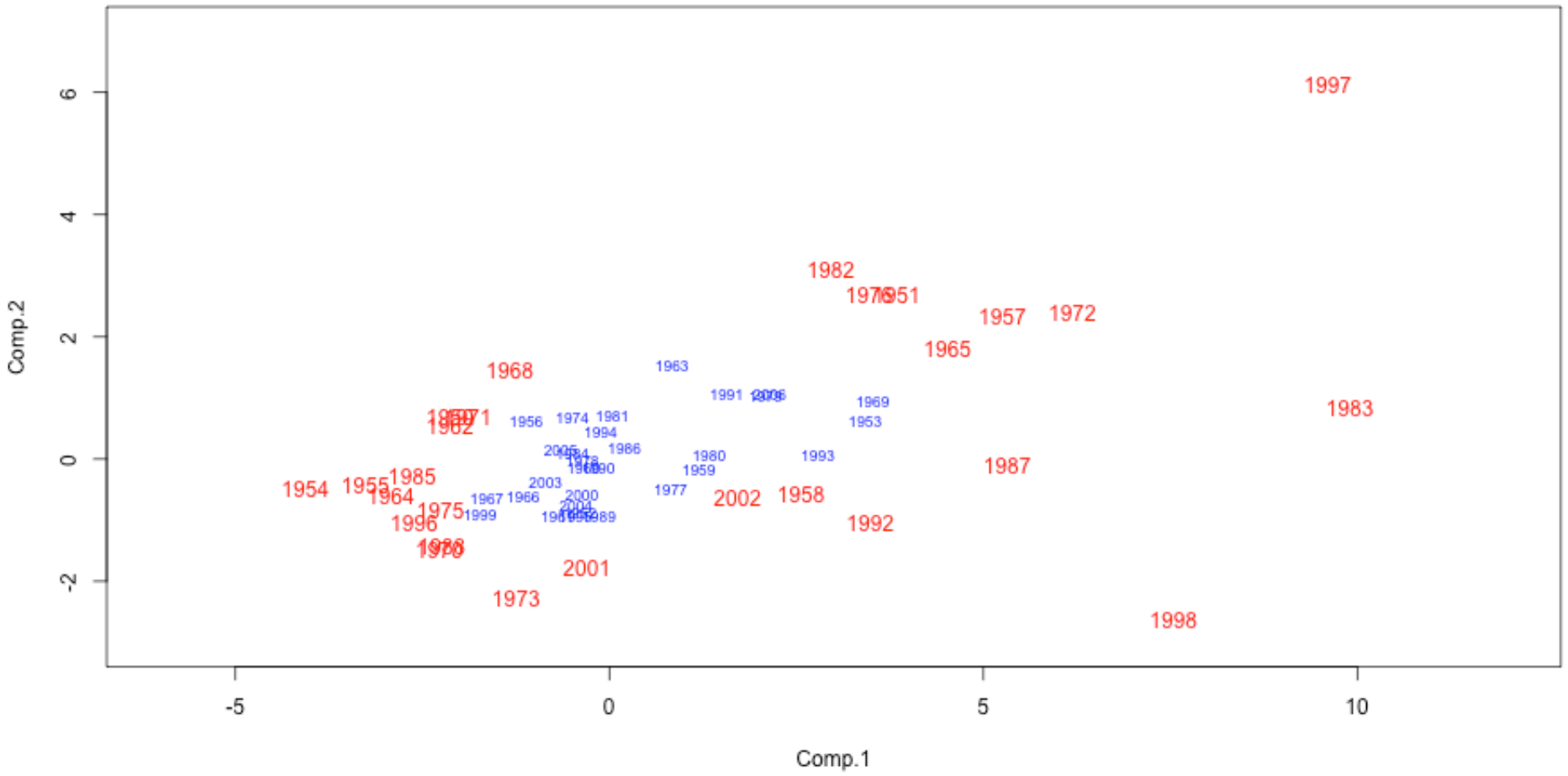
principal component analysis on $\boldsymbol{X} = (X_{i,t})$

```
> (PCA=PCAproj(t(ElNino$y), center = median))
Call:
PCAproj(x = t(ElNino$y), center = median)

Standard deviations:
  Comp.1   Comp.2
2.712685 1.170426

 12 variables and 57 observations.
> plot(PCA$scores)
> bagplot(PCA$scores,
+ factor=1,dkmethod=1)
```



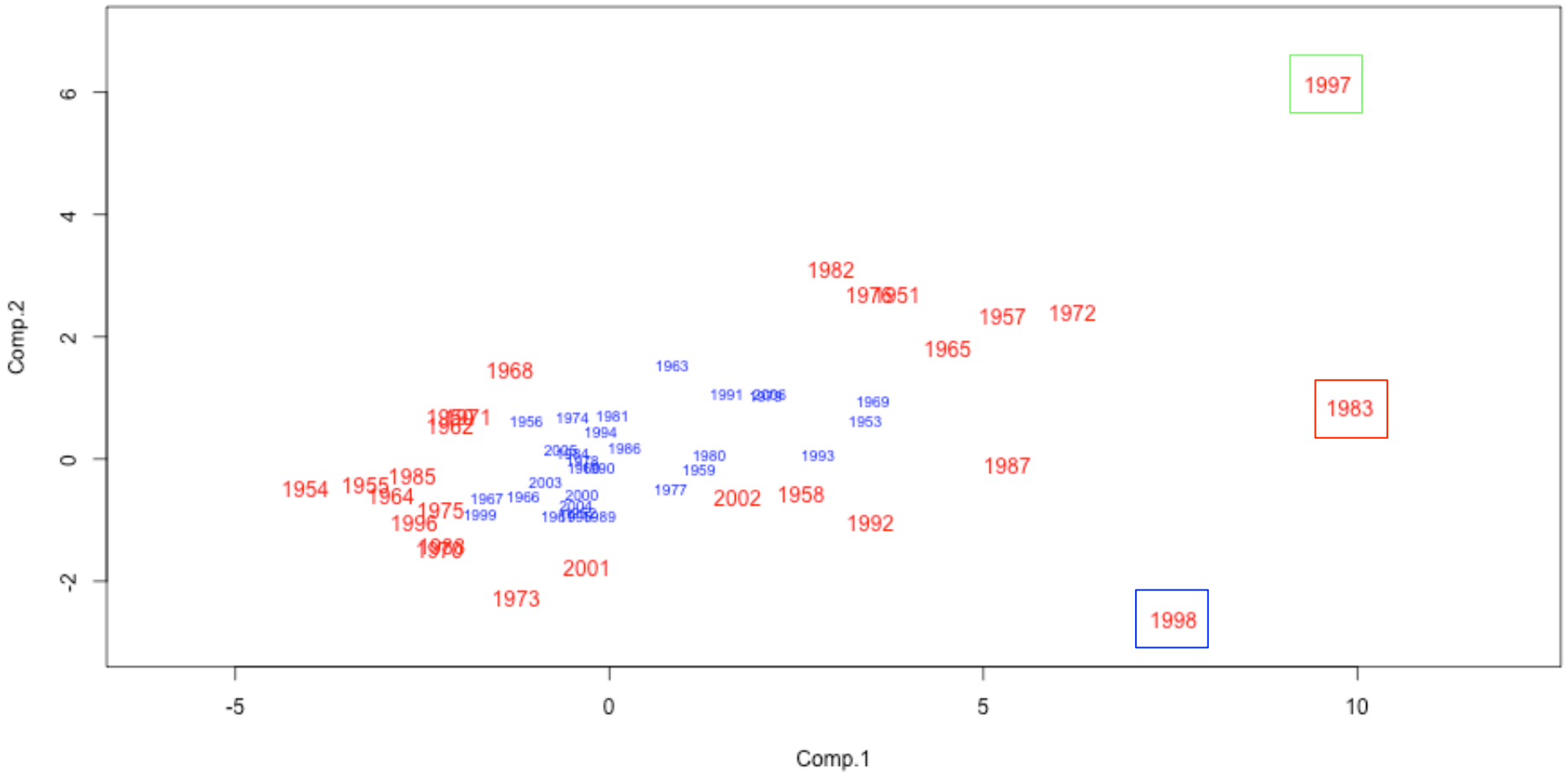
Outliers and curves

```
> bp=bagplot(PCA$scores,factor=1,dkmethod=1,cex=1.5)
> plot(PCA$scores,xlim=c(-6,12),ylim=c(-3,7),col="white")
> text(bp$pxy.bag[,1],bp$pxy.bag[,2],as.numeric(rownames(bp$pxy.bag))+1949,col="blue",cex=.7
> text(bp$pxy.outlier[,1],bp$pxy.outlier[,2],as.numeric(rownames(bp$pxy.outlier))+1949,col='
```



Outliers and curves

```
> bp=bagplot(PCA$scores,factor=1,dkmethod=1,cex=1.5)
> plot(PCA$scores,xlim=c(-6,12),ylim=c(-3,7),col="white")
> text(bp$pxy.bag[,1],bp$pxy.bag[,2],as.numeric(rownames(bp$pxy.bag))+1949,col="blue",cex=.7
> text(bp$pxy.outlier[,1],bp$pxy.outlier[,2],as.numeric(rownames(bp$pxy.outlier))+1949,col='
```



Outliers and curves

```
> fboxplot(data=ElNino, plot.type="functional", type="bag")
```