ESTYMACJA WYSOKICH KWANTYLI

Ryszard Zieliński, IMPAN Warszawa

XL Konferencja Zastosowań Matematyki
Zakopane-Kościelisko 30.VIII - 6.IX.2011

Problem szacowania wysokich kwantyli pojawia się w różnych zastosowaniach (ekonomia, finanse, VaR, ekologia).

Problem szacowania wysokich kwantyli pojawia się w różnych zastosowaniach (ekonomia, finanse, VaR, ekologia).
Chodzi o kwantyle np. rzędu 0.99, rzędu 0.999, lub nawet wyższego

Problem szacowania wysokich kwantyli pojawia się w różnych zastosowaniach (ekonomia, finanse, VaR, ekologia).
Chodzi o kwantyle np. rzędu 0.99, rzędu 0.999, lub nawet wyższego

Później sprecyzuję pojęcie „WYSOKI RZĄD”

Problem szacowania wysokich kwantyli pojawia się w różnych zastosowaniach (ekonomia, finanse, VaR, ekologia).
Chodzi o kwantyle np. rzędu 0.99, rzędu 0.999, lub nawet wyższego

Później sprecyzuję pojęcie „WYSOKI RZĄD”

Dla danego zjawiska, taki kwantyl interpretowany jest jako próg, który może być przekroczony z małym prawdopodobieństwem, np. 0.01 lub 0.001.

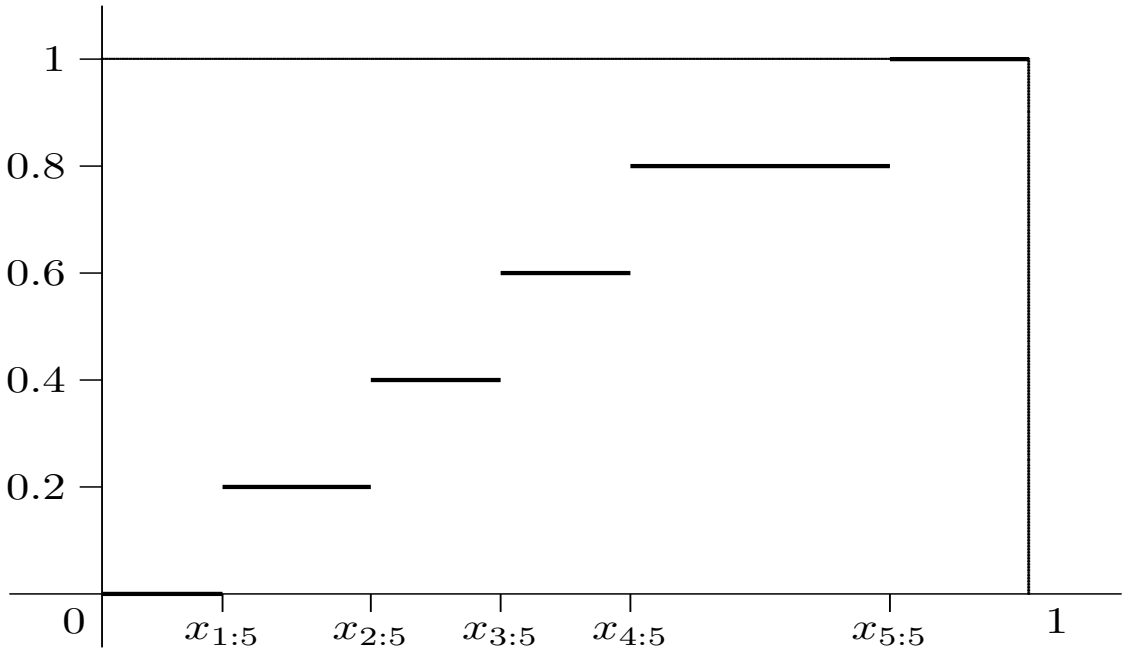
Problem szacowania wysokich kwantyli pojawia się w różnych zastosowaniach (ekonomia, finanse, VaR, ekologia).
Chodzi o kwantyle np. rzędu 0.99, rzędu 0.999, lub nawet wyższego

Później sprecyzuję pojęcie „WYSOKI RZĄD”

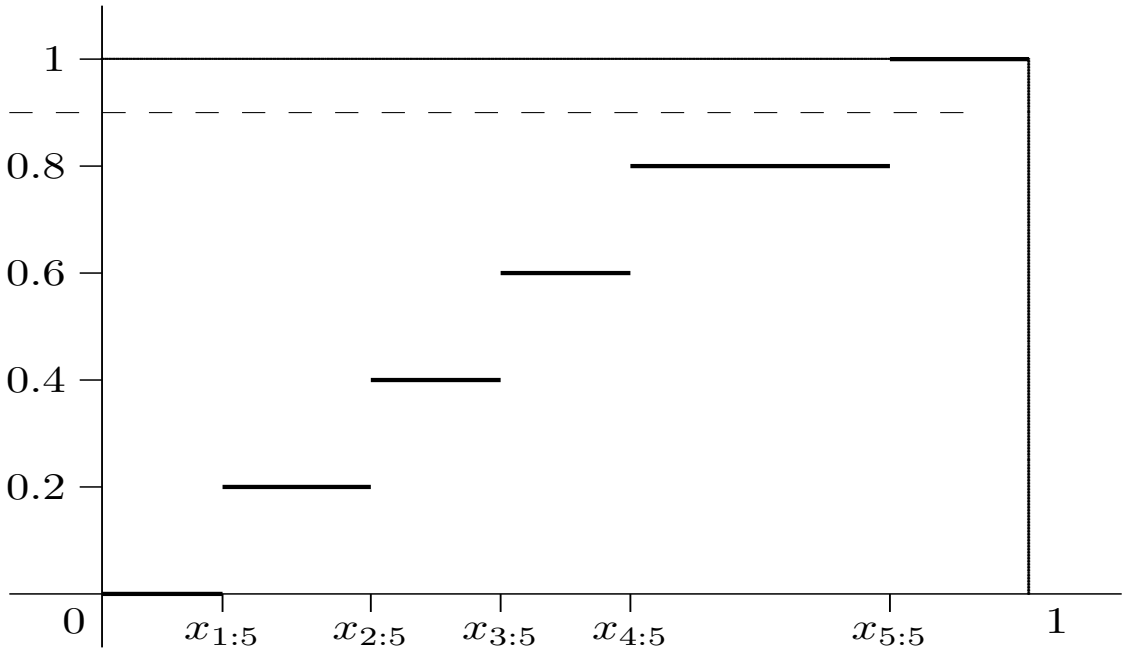
Dla danego zjawiska, taki kwantyl interpretowany jest jako próg, który może być przekroczony z małym prawdopodobieństwem, np. 0.01 lub 0.001.

Nasze źródło informacji (jedyne?): obserwacje historyczne danego lub podobnego zjawiska. ⊣

Cała informacja z obserwacji, którymi dysponujemy, jest zawarta w dystrybuancie empirycznej:

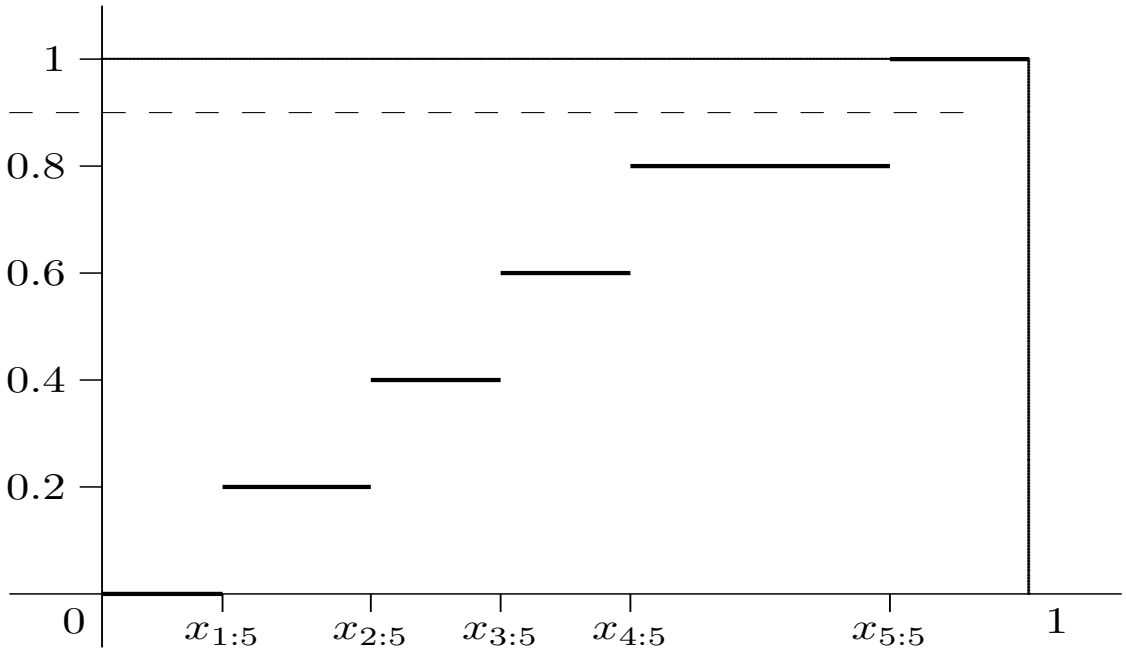


Cała informacja z obserwacji, którymi dysponujemy, jest zawarta w dystrybuancie empirycznej:



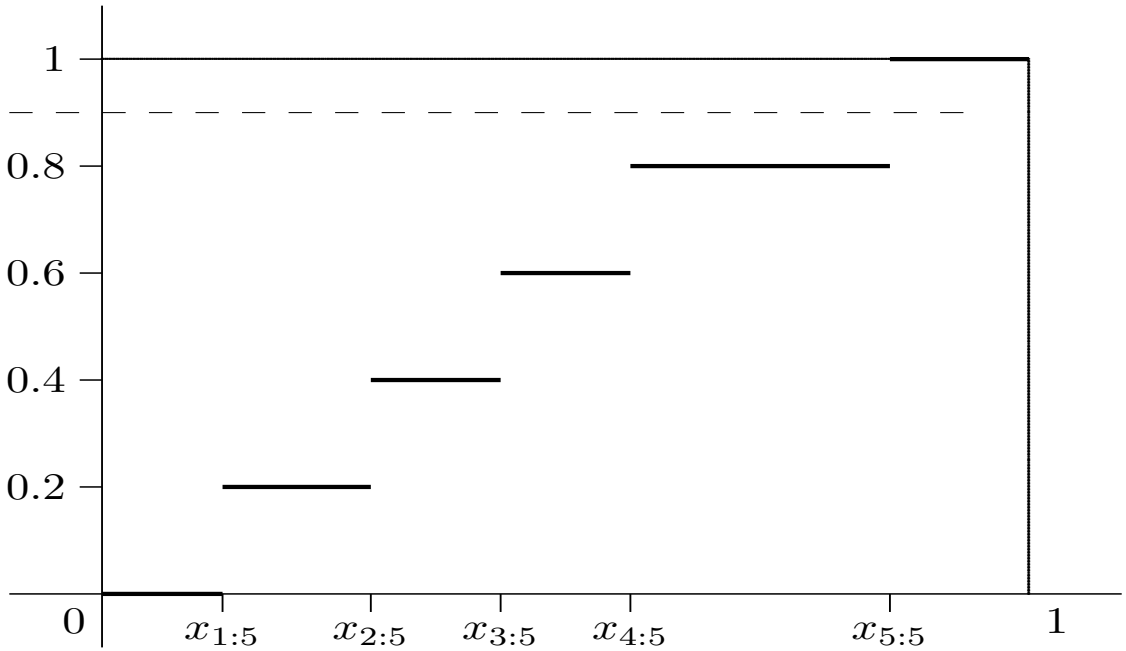
Przykład: *oszacować kwantyl rzędu* 0.9

Cała informacja z obserwacji, którymi dysponujemy, jest zawarta w dystrybuancie empirycznej:



Przykład: *oszacować kwantyl rzędu* 0.9
Ekstrapolacja EDF? Wygładzanie i ekstrapolacja?

Cała informacja z obserwacji, którymi dysponujemy, jest zawarta w dystrybuancie empirycznej:



Przykład: *oszacować kwantyl rzędu* 0.9
Ekstrapolacja EDF? Wygładzanie i ekstrapolacja? **Model?** ⊣

Typowe podejście: ekstrapolacja

Hill, B.M. (1975), A simple general approach to inference about the tail of a distribution. *The Annals of Statistics* 3, 5, 1163–1174

$\dashv$

Nonparametric Analysis of Univariate Heavy-Tailed Data

Research and Practice

Natalia Markovich
Institute of Control Sciences,
Russian Academy of Sciences,
Moscow, Russia

BICENTENNIAL 1807 WILEY 2007 BICENTENNIAL
John Wiley & Sons, Ltd

Markovich (2008):

Markovich (2008):

The lack of information beyond the range of the sample creates the main problem in the estimation of high quantiles.

Markovich (2008):

The lack of information beyond the range of the sample creates the main problem in the estimation of high quantiles. Since $F_n(X_{n:n}) = 1$, it is impossible to estimate the quantiles without knowledge of the behavior of F at infinity.

Markovich (2008):

The lack of information beyond the range of the sample creates the main problem in the estimation of high quantiles. Since $F_n(X_{n:n}) = 1$, it is impossible to estimate the quantiles without knowledge of the behavior of F at infinity. The main idea behind all estimators for high quantiles is to select first some auxiliary pilot estimate inside the range of the sample (one can use one of the order statistics close to the boundary as a pilot estimate) and to move this pilot estimate to the right.

Markovich (2008):

The lack of information beyond the range of the sample creates the main problem in the estimation of high quantiles. Since $F_n(X_{n:n}) = 1$, it is impossible to estimate the quantiles without knowledge of the behavior of F at infinity. The main idea behind all estimators for high quantiles is to select first some auxiliary pilot estimate inside the range of the sample (one can use one of the order statistics close to the boundary as a pilot estimate) and to move this pilot estimate to the right.
Obviously, in order to extrapolate the pilot estimate beyond the sample range, one needs to use some model of the tail of the distribution.

Markovich (2008):

The lack of information beyond the range of the sample creates the main problem in the estimation of high quantiles. Since $F_n(X_{n:n}) = 1$, it is impossible to estimate the quantiles without knowledge of the behavior of F at infinity. The main idea behind all estimators for high quantiles is to select first some auxiliary pilot estimate inside the range of the sample (one can use one of the order statistics close to the boundary as a pilot estimate) and to move this pilot estimate to the right.
Obviously, in order to extrapolate the pilot estimate beyond the sample range, one needs to use some model of the tail of the distribution. Such models are not available in many applications.

Markovich (2008):

The lack of information beyond the range of the sample creates the main problem in the estimation of high quantiles. Since $F_n(X_{n:n}) = 1$, it is impossible to estimate the quantiles without knowledge of the behavior of F at infinity. The main idea behind all estimators for high quantiles is to select first some auxiliary pilot estimate inside the range of the sample (one can use one of the order statistics close to the boundary as a pilot estimate) and to move this pilot estimate to the right.
Obviously, in order to extrapolate the pilot estimate beyond the sample range, one needs to use some model of the tail of the distribution. Such models are not available in many applications. Therefore, the asymptotic tail models based on the distribution of the largest order statistic are usually used. ⊣

Przykład 1 (p - rząd estymowanego kwantyla)

In the POT (*Picks Over Threshold*) estimator, the GPD (*Generalized Pareto Distribution*) is used as a distribution of excesses over SOME high threshold u:

$$\begin{aligned} CDF(x) &= 1 - \left(1 + \gamma \frac{x-u}{\sigma}\right)^{-1/\gamma}, \quad \gamma \neq 0 \\ &= 1 - exp\{-(x-u)/\sigma\}, \quad \gamma = 0 \end{aligned}$$

Estymator:

$$x_p^{POT} = u + \frac{\hat{\sigma}}{\hat{\gamma}} \left(\left(\frac{p}{1 - F_n(u)} \right)^{-\hat{\gamma}} - 1 \right),$$

where $\hat{\sigma}$ and $\hat{\gamma}$ are SOME estimates of the parameters $\dashv$

Przykład 2.

In Weissman (1978) the estimator

$$x_p^w = X_{n-k,n}\left(\frac{k+1}{(n+1)p}\right)^{\hat{\gamma}}, \quad k = 1, \ldots, n$$

is obtained for the Pareto tail model:

$$CDF(x) = 1 - \exp\{-x^{-1/\gamma}\}, \quad \gamma > 0, x > 0$$

Przykład 2.

In Weissman (1978) the estimator

$$x_p^w = X_{n-k,n}\left(\frac{k+1}{(n+1)p}\right)^{\hat{\gamma}}, \quad k = 1, \ldots, n$$

is obtained for the Pareto tail model:

$$CDF(x) = 1 - \exp\{-x^{-1/\gamma}\}, \quad \gamma > 0, x > 0$$

(*asymptotyczny rozkład maksimum, rozkład Frécheta*) ⊣

Przykład 3 (Markovich and Krieger (2002)

$$x_p^c = X_{n-k,n}\left(-0.5 + \sqrt{0.25 + \frac{pnc(\hat{\gamma})}{k}}\right)^{-\hat{\gamma}}$$

gdzie

$$c(\gamma) = 1 + X_{n-k,n}^{-1/\gamma} + X_{n-k,n}^{-2/\gamma}$$

Przykład 3 (Markovich and Krieger (2002)

„one can expect that the statistic

$$x_p^c = X_{n-k,n}\left(-0.5 + \sqrt{0.25 + \frac{pnc(\hat{\gamma})}{k}}\right)^{-\hat{\gamma}}$$

gdzie

$$c(\gamma) = 1 + X_{n-k,n}^{-1/\gamma} + X_{n-k,n}^{-2/\gamma}$$

approximates x_p" ⊣

Kwantyl Pareto: $0.99 - 100, \quad 0.999 - 1000$

Table 6.2 Tolerant 90% confidence intervals of estimates x_p^w and x_p^c for heavy-tailed distributions; 500 samples of $n = 100$ observations each.

PDF	$(1-p)\cdot 100\%$	x_p^c		x_p^w	
		mean(x_p^c) (StDev(x_p^c))	Confidence interval	mean(x_p^w) (StDev(x_p^w))	Confidence interval
Pareto $\gamma = 1$	99	75.814 (46.949)	(−7.567, 159.195)	117.957 (90.903)	(−43.487, 279.401)
		MSE = $2.789 \cdot 10^3$		MSE = $8.586 \cdot 10^3$	
	99.9	963.094 ($1.553 \cdot 10^3$)	($-1.795 \cdot 10^3$, $3.721 \cdot 10^3$)	$1.616 \cdot 10^3$ ($2.661 \cdot 10^3$)	($-3.11 \cdot 10^3$, $6.342 \cdot 10^3$)
		MSE = $2.413 \cdot 10^6$		MSE = $7.460 \cdot 10^6$	

Kwantyl Pareto: $0.99 - 100, \quad 0.999 - 1000$

174 ESTIMATION OF HIGH QUANTILES

Table 6.3 Tolerant 90% confidence intervals of estimates x_p^w and x_p^c for heavy-tailed distributions: 500 samples of $n = 1000$ observations each.

PDF	$(1-p)$-100%	x_p^c		x_p^w	
		mean(x_p^c) (StDev(x_p^c))	Confidence interval	mean(x_p^w) (StDev(x_p^w))	Confidence interval
Pareto $\gamma = 1$	99	80.452 (16.272)	(51.533, 109.351)	101.93 (22.841)	(61.364, 142.496)
		MSE = 646.902		MSE = 525.436	
	99.9	791.071 (290.593)	(274.978, 1307)	$1.051 \cdot 10^3$ (410.541)	(321.879, 1780)
		MSE = $1.281 \cdot 10^5$		MSE = $1.711 \cdot 10^5$	

Mój model:

$\mathcal{F}$ – rodzina wszystkich rozkładów
z ciągłymi i ściśle rosnącymi dystrybuantami

Mój model:

$\mathcal{F}$ – rodzina wszystkich rozkładów
z ciągłymi i ściśle rosnącymi dystrybuantami

Kwantyl wysokiego rzędu (wysoki kwantyl)? $\dashv$

Optymalny estymator kwantyla w modelu $\mathcal{F}$

Klasa $\mathcal{T}$ estymatorów ekwiwariantnych

$T \in \mathcal{T}$ wtedy i tylko wtedy, gdy dla każdego ściśle rosnącego przekształcenia prostej g,

$$T(g(x_1), g(x_2), \ldots, g(x_n)) = g\left(T(x_1, x_2, \ldots, x_n)\right)$$

Optymalny estymator kwantyla w modelu $\mathcal{F}$

Klasa $\mathcal{T}$ estymatorów ekwiwariantnych

$T \in \mathcal{T}$ wtedy i tylko wtedy, gdy dla każdego ściśle rosnącego przekształcenia prostej g,

$$T(g(x_1), g(x_2), \ldots, g(x_n)) = g\left(T(x_1, x_2, \ldots, x_n)\right)$$

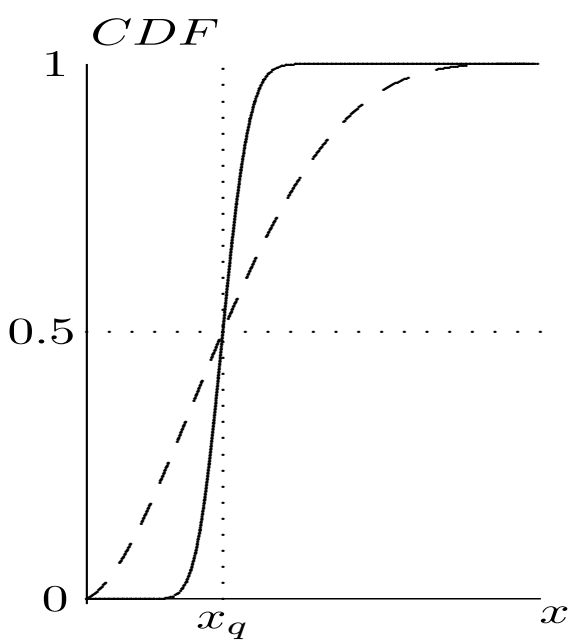
TWIERDZENIE. T jest estymatorem ekwiwariantnym wtedy i tylko wtedy, gdy jest postaci $T = X_{J:n}$, gdzie J jest losowym indeksem o rozkładzie $P\{J = j\} = \lambda_j,\ j = 1, 2, \ldots, n.$ $\dashv$

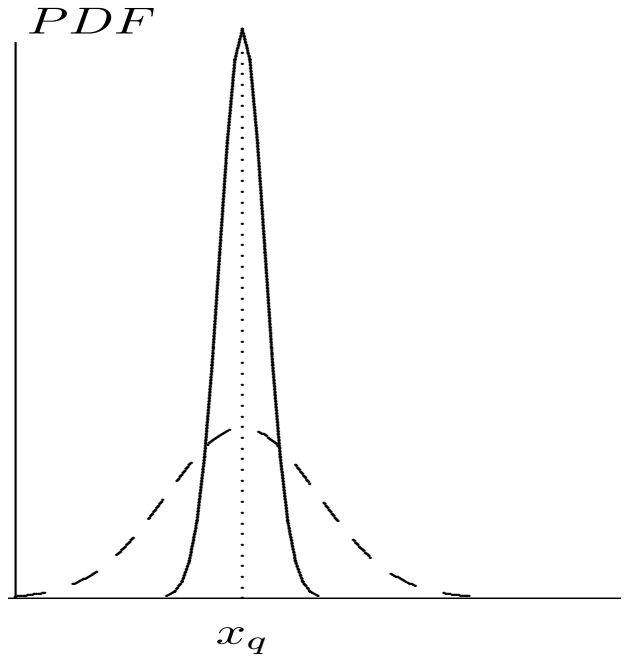
Kryterium

Kryterium

medianowo nieobciążony estymator
o maksymalnej koncentracji wokół estymowanego kwantyla

⊣





Estimate of x_q with solid cdf and pdf
is more concentrated median-unbiased estimator of x_q
than that with dashed pdf

Medianowo nieobciążony estymator kwantyla rzędu q, z próby $X_1, X_2, \ldots, X_n$ o liczności n, istnieje wtedy i tylko wtedy, gdy

$$1 - (1/2)^{1/n} \leqslant q \leqslant (1/2)^{1/n}$$

Medianowo nieobciążony estymator kwantyla rzędu q, z próby $X_1, X_2, \ldots, X_n$ o liczności n, istnieje wtedy i tylko wtedy, gdy

$$1 - (1/2)^{1/n} \leqslant q \leqslant (1/2)^{1/n}$$

Definicja. *Dla danego n, x_q jest kwantylem wysokiego rzędu, gdy $q > q(n) = (1/2)^{1/n}$.*

Medianowo nieobciążony estymator kwantyla rzędu q, z próby $X_1, X_2, \ldots, X_n$ o liczności n, istnieje wtedy i tylko wtedy, gdy

$$1 - (1/2)^{1/n} \leqslant q \leqslant (1/2)^{1/n}$$

Definicja. *Dla danego n, x_q jest kwantylem wysokiego rzędu, gdy $q > q(n) = (1/2)^{1/n}$.*

Definicja. *Dla danego q, x_q jest kwantylem wysokiego rzędu, gdy $n < n(q) = -log\, 2/log\, q$.*

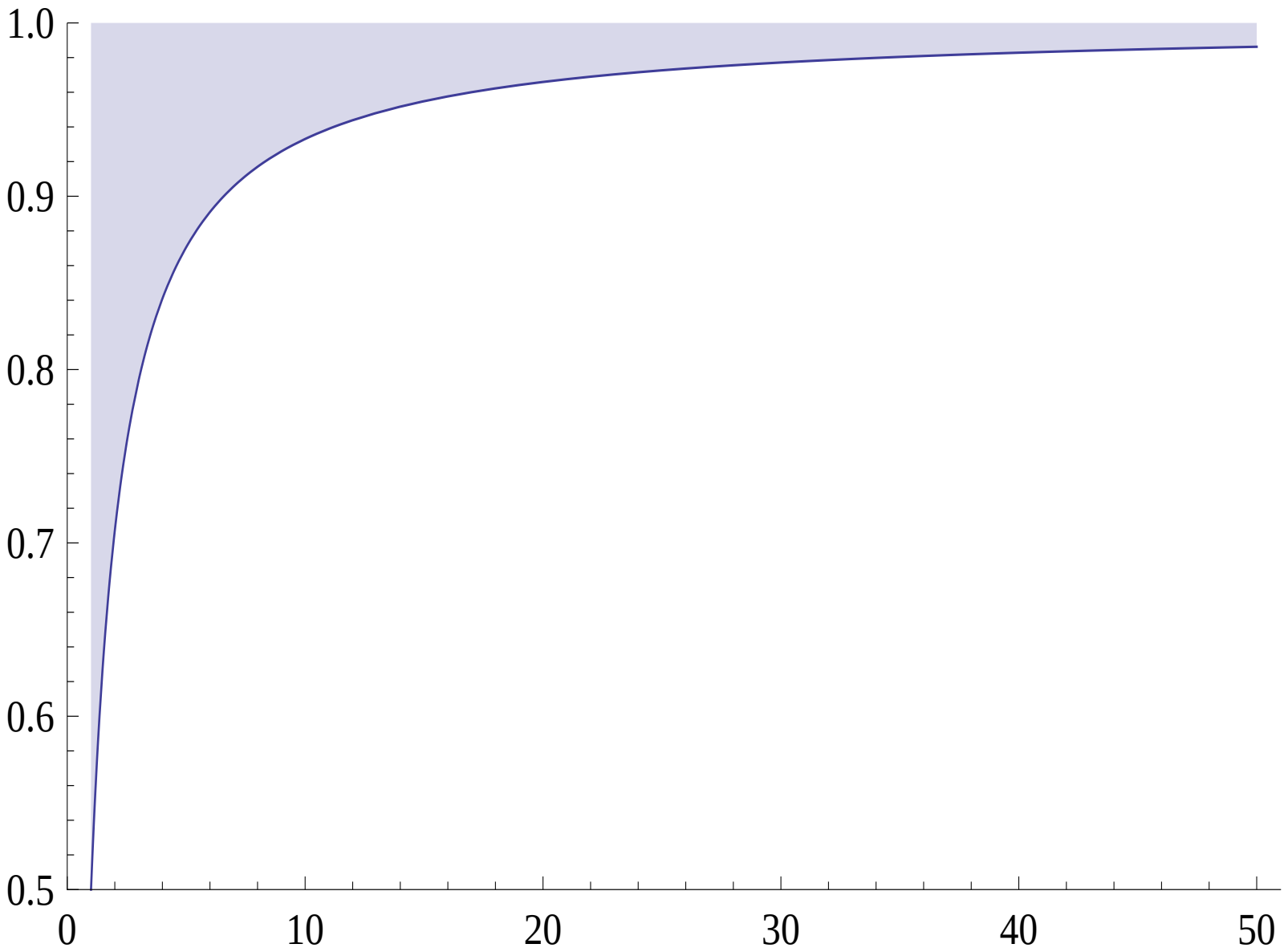
⊣Tabl

Table 1

n	5	10	20	50	100
$q(n)$	0.8706	0.9331	0.9660	0.9863	0.9931
n	200	500	1000	2000	5000
$q(n)$	0.9966	0.9987	0.9993	0.9997	0.99986

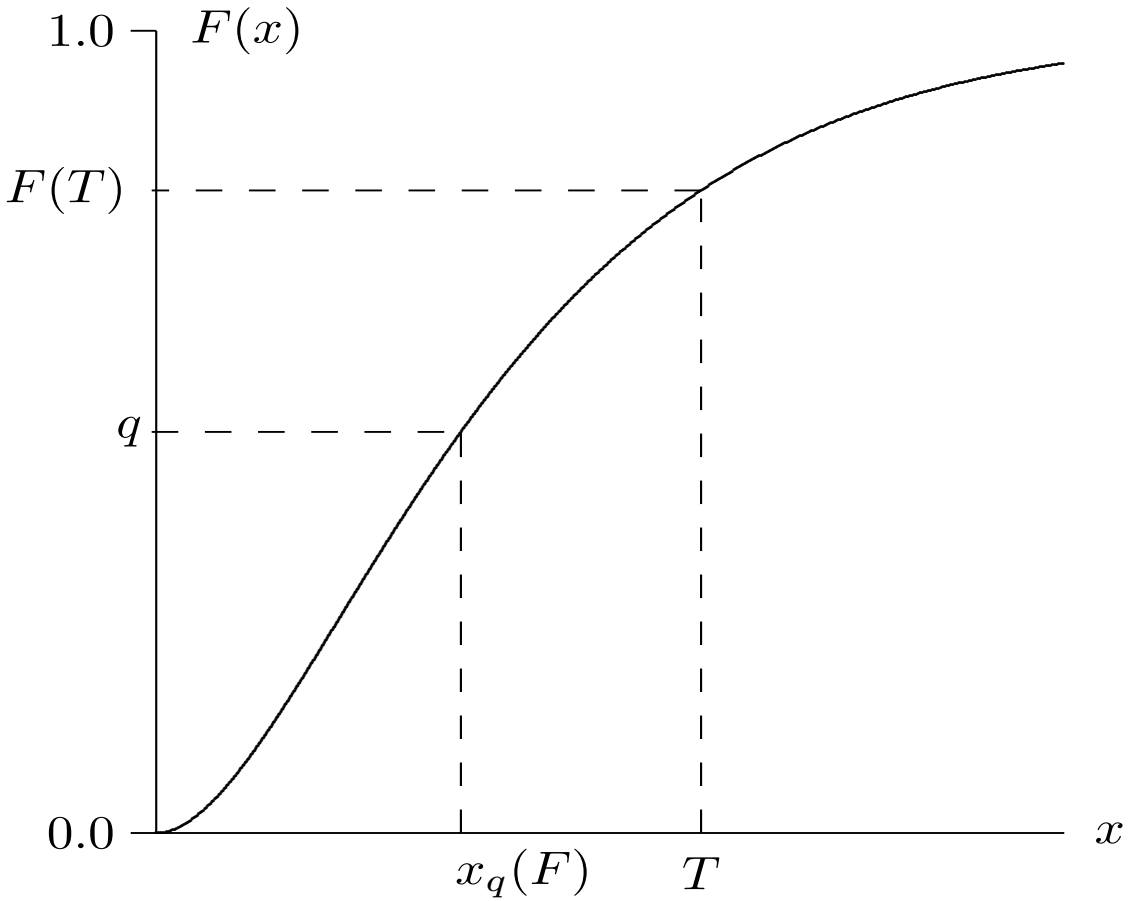
Table 2

q	0.9	0.95	0.99	0.999	0.9999	0.99999
$n(q)$	7	14	69	693	6932	69315

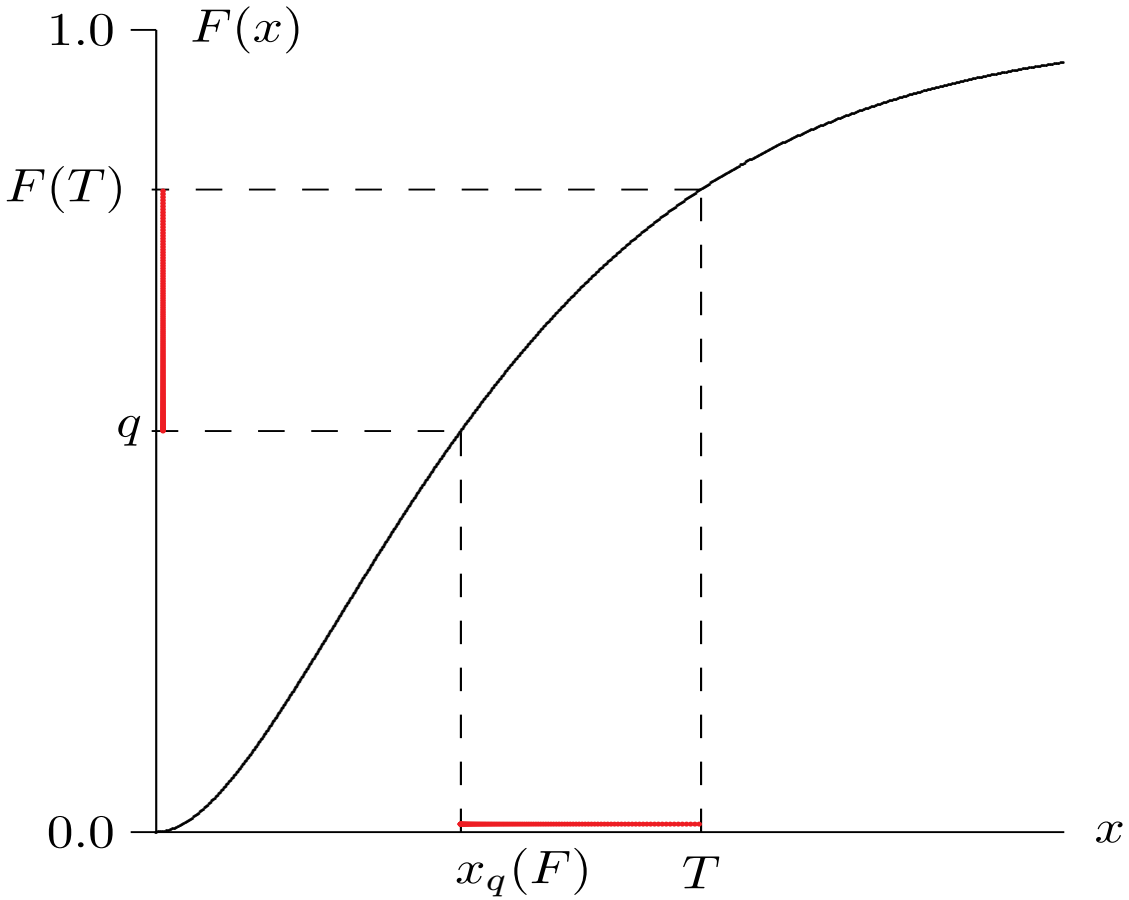

1.0
0.9
0.8
0.7
0.6
0.5
0
10
20
30
40
50

SZACOWANIE WYSOKICH KWANTYLI

F-przekształcenie:



F-przekształcenie:



Jeżeli T jest estymatorem kwantyla x_q rzędu q, to $F(T)$ jest estymatorem liczby q

T jest F-nieobciążonym estymatorem kwantyla x_q, jeżeli

$$E_F\left(F(T)\right) = q, \quad \text{dla każdego} \quad F \in \mathcal{F}$$

T jest F-nieobciążonym estymatorem kwantyla x_q, jeżeli

$$E_F\left(F(T)\right) = q, \quad \text{dla każdego} \quad F \in \mathcal{F}$$

Średniokwadratowy F-błąd estymatora T wyraża się wzorem

$$E_F\left(F(T) - q\right)^2$$

$\dashv$

TWIERDZENIE. Dla wysokiego kwantyla, estymator F-nieobciążony nie istnieje.

DOWÓD. Mamy

$$E_F F(X_{J:n}) = \sum_{j=1}^{n} \lambda_j E U_{j:n} = \frac{1}{n+1} \sum_{j=1}^{n} j\lambda_j$$

Równanie

$$\frac{1}{n+1} \sum_{j=1}^{n} j\lambda_j = q$$

ma rozwiązanie wtedy i tylko wtedy, gdy $1/(n+1) \leqslant q \leqslant n/(n+1)$. Kwantyl jest wysoki, gdy $q > n(q)$, ale $n(q) = (1/2)^{1/n} > n/(n+1)$ dla $n > 1$. QED

$\dashv$

TWIERDZENIE. Dla wysokiego kwantyla, estymatorem o jednostajnie minimalnym F-błędzie średniokwadratowym jest $X_{n:n}$.

$\dashv$

DOWÓD.

Mamy

$$
\begin{aligned}
FMSE_n(q) &= E_F\Big(F(X_{J:n}) - q\Big)^2 = E(U_{J:n} - q)^2 \\
&= \sum_{j=1}^{n} \lambda_j E(U_{j:n} - q)^2 \\
&= \sum_{j=1}^{n} \frac{\lambda_j \Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} \int_0^1 (x-q)^2 x^{j-1}(1-x)^{n-j}\,dx \\
&= \frac{1}{(n+1)(n+2)} \sum_{j=1}^{n} j\big(j+1-2(n+2)q\big)\lambda_j + q^2
\end{aligned}
$$

Rozkładem $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_n^*)$, który minimalizuje ten błąd, jest rozkład z $\lambda_{j*} = 1$, $\lambda_j = 0, j \neq j^*$, gdzie j^* minimalizuje $j\big(j+1-2(n+2)q\big)$, jednostajnie względem $q > q(n)$, czyli $j^* = n$.

QED $\dashv$

Średniokwadratowy F-błąd optymalnego estymatora wyraża się wzorem:

$$FMSE_n(q) = \frac{n\Big(n+1-2(n+2)q\Big)}{(n+1)(n+2)} + q^2, \quad q \geqslant q(n)$$

Dla danego n, mamy $FMSE_n \nearrow 1 - [n(n+3)]/[(n+1)(n+2)]$ gdy $q \nearrow 1$,

Ponadto $FMSE_n \searrow 0$, gdy $n \nearrow +\infty$, jednostajnie względem $q \geqslant q(n)$

$\dashv$

Udowodniliśmy

TWIERDZENIE. Dla wysokiego kwantyla, estymatorem o jednostajnie minimalnym F-błędzie średniokwadratowym jest $X_{n:n}$.

Udowodniliśmy

TWIERDZENIE. Dla wysokiego kwantyla, estymatorem o jednostajnie minimalnym F-błędzie średniokwadratowym jest $X_{n:n}$.

Czy prawdziwe jest takie twierdzenie dla ryzyka przy dowolnej wypukłej funkcji strat?

Udowodniliśmy

TWIERDZENIE. Dla wysokiego kwantyla, estymatorem o jednostajnie minimalnym F-błędzie średniokwadratowym jest $X_{n:n}$.

Czy prawdziwe jest takie twierdzenie dla ryzyka przy dowolnej wypukłej funkcji strat? Myślę, że tak.

Udowodniliśmy

TWIERDZENIE. Dla wysokiego kwantyla, estymatorem o jednostajnie minimalnym F-błędzie średniokwadratowym jest $X_{n:n}$.

Czy prawdziwe jest takie twierdzenie dla ryzyka przy dowolnej wypukłej funkcji strat? Myślę, że tak.

A może dla funkcji LINEX ? ⊣

Błąd estymacji dużych kwantyli w modelu nieparametrycznym $\mathcal{F}$ jest mierzony w terminach $(F(T) - q)$.
Ocena tego błędu w terminach $(T - x_q(F))$ jest w tym modelu niemożliwa (?), chyba że wprowadzimy jakieś precyzyjne warunki na zachowanie się ogonów rozkładów.
Interesująco w tym kontekście wygląda problem estymacji wysokich kwantyli w mniejszych modelach

$$\mathcal{F}_1 = \mathcal{F} \cap \{F : \int_0^1 |F^{-1}(t)|dt < \infty\}$$

lub

$$\mathcal{F}_2 = \mathcal{F} \cap \{F : \int_0^1 (F^{-1}(t))^2 dt < \infty\}$$

$\dashv$