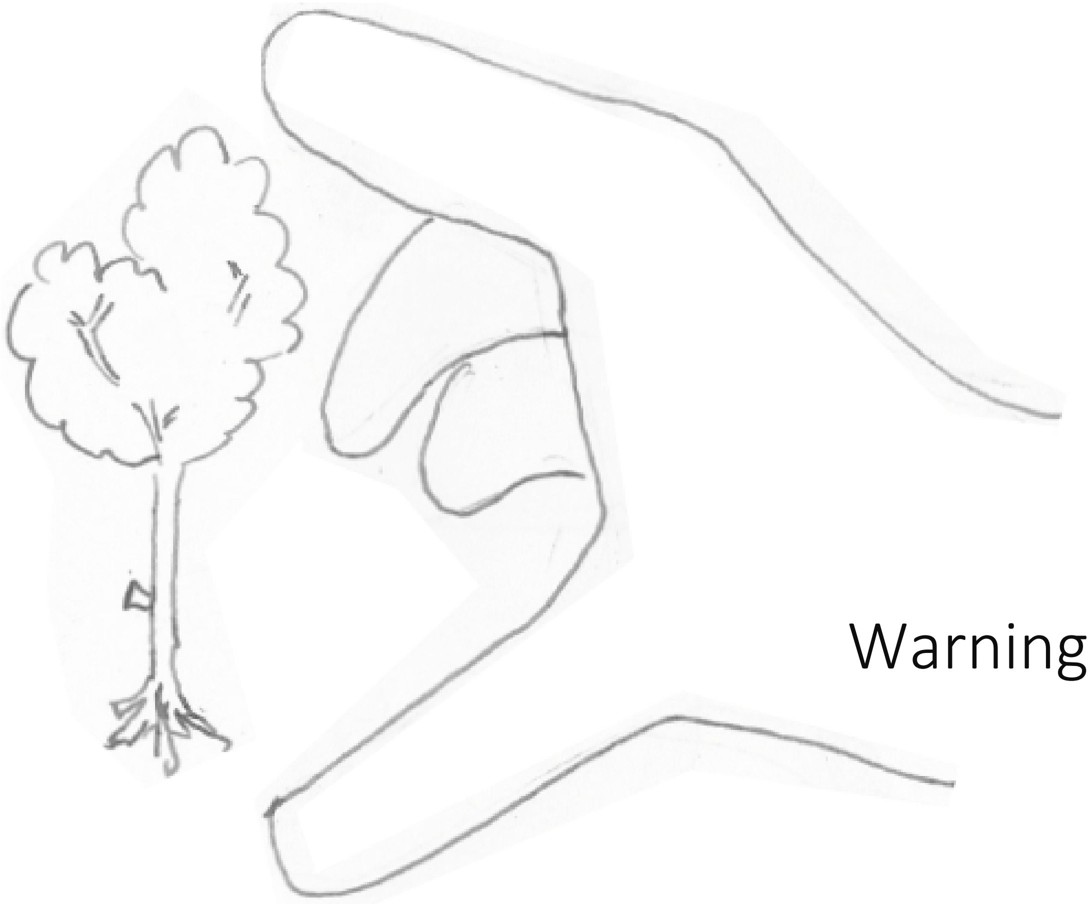
Summarizing Performance Data
Confidence Intervals



Important

Easy to Difficult

Warning: some mathematical content

Jean-Yves Le Boudec, February 2019

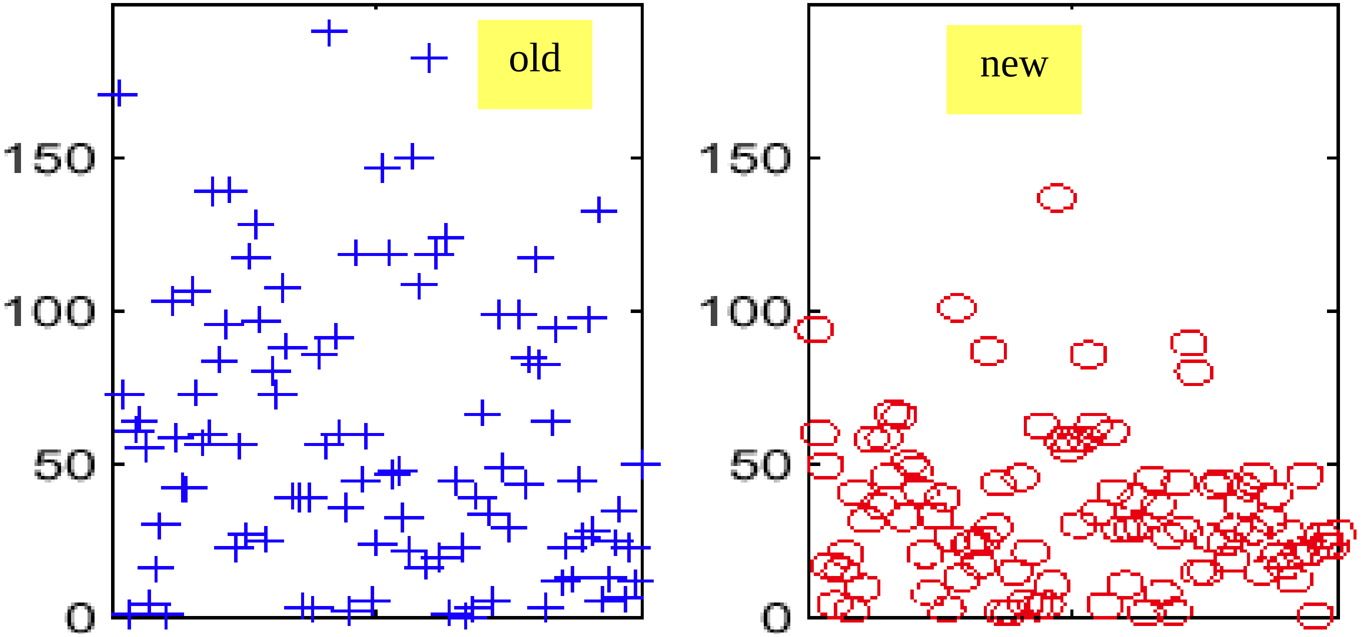
Contents

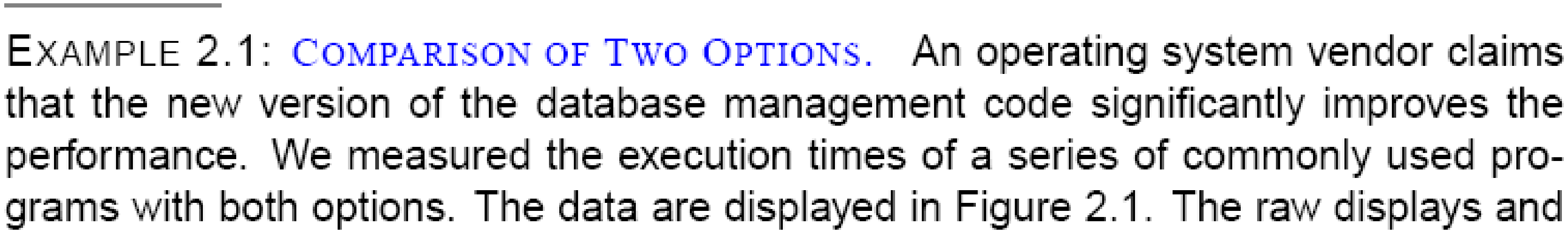
1. Summarized data
2. Confidence Intervals
3. Independence Assumption
4. Prediction Intervals
5. Which Summarization to Use ?

1 Summarizing Performance Data

How do you quantify:

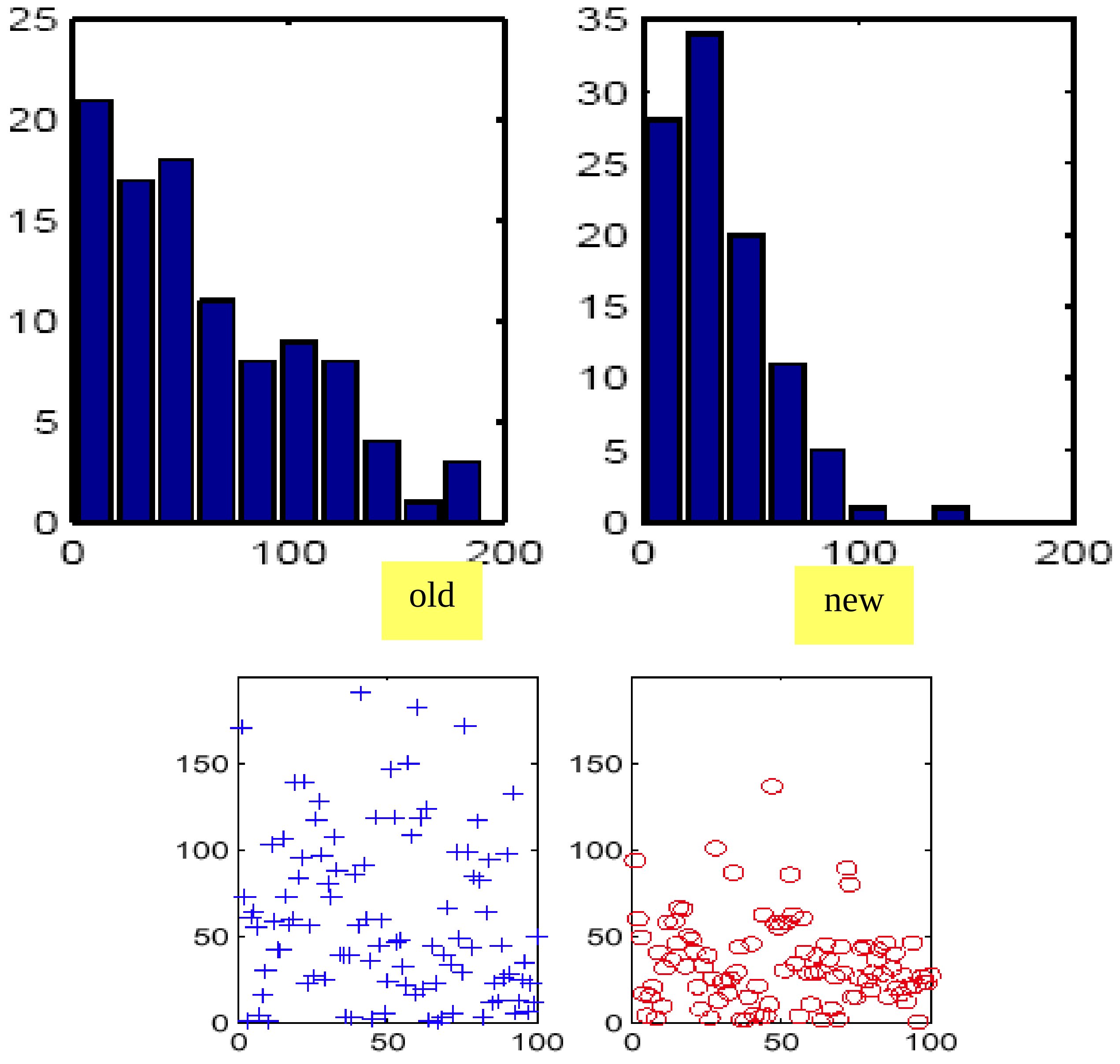
- Central value
- Dispersion (Variability)





EXAMPLE 2.1: COMPARISON OF TWO OPTIONS. An operating system vendor claims that the new version of the database management code significantly improves the performance. We measured the execution times of a series of commonly used programs with both options. The data are displayed in Figure 2.1. The raw displays and

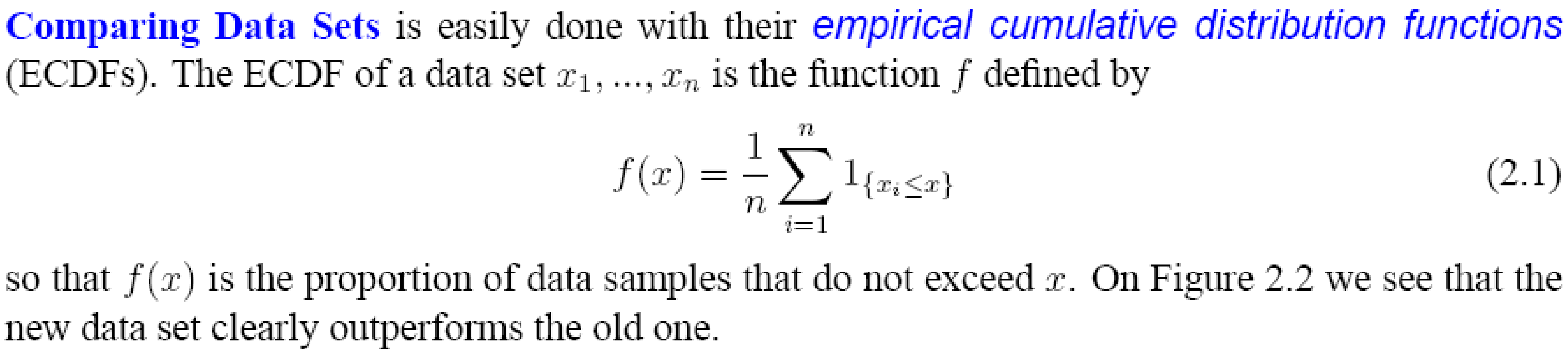
Histogram is one answer



old

new

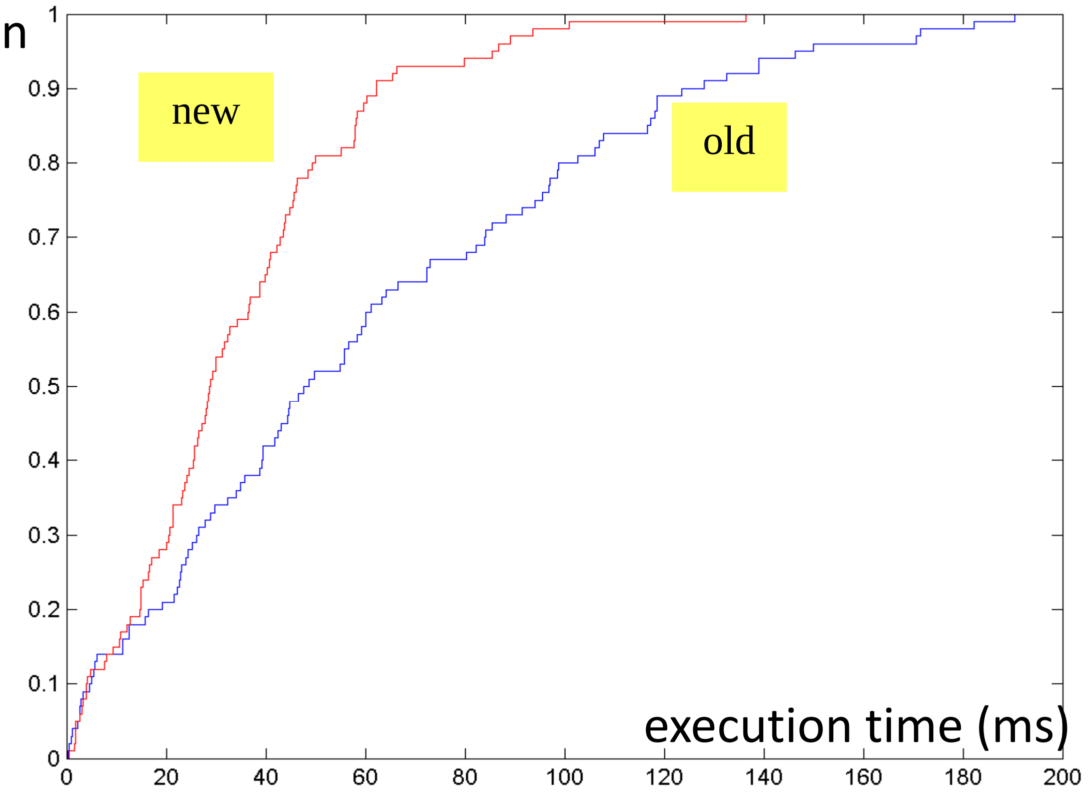
ECDF allow easy comparison



Comparing Data Sets is easily done with their *empirical cumulative distribution functions* (ECDFs). The ECDF of a data set $x_1, \ldots, x_n$ is the function f defined by

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i \le x\}} \tag{2.1}$$

so that $f(x)$ is the proportion of data samples that do not exceed x. On Figure 2.2 we see that the new data set clearly outperforms the old one.



$$CDF(x) = \int_0^x f_X(y)dy$$
$$= P(X \le x)$$

Summarized Measures

Median, Quantiles

- Median : the data point in the middle:

if n is odd, median = $x_{\left(\frac{n+1}{2}\right)}$ else $\frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right)$

- Quartiles
- p-quantiles

Mean and standard deviation

- Mean $m = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Standard deviation

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \text{ or } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$$

- Coefficient of variation = $\frac{s}{m}$ (scale-free)

Which are correct interpretations of standard deviation ? ($\boldsymbol{m}$ = mean, $\boldsymbol{s}$ = standard deviation)

A. With 95% probability, a new data sample lies in the interval m±1.96s

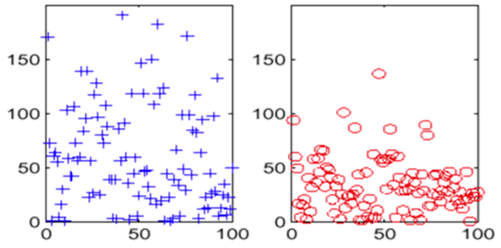
B. With 99.9% probability, a new data sample lies in the interval $m \pm 3.30s$

C. A and B

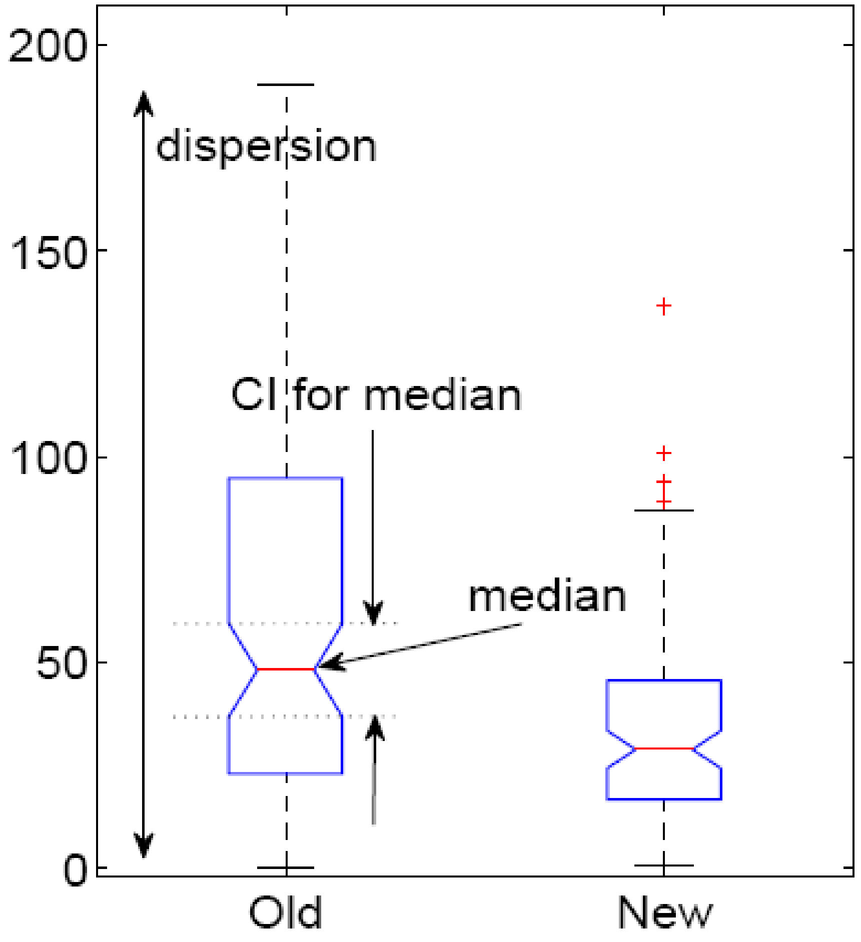
D. None

E. I don’t know

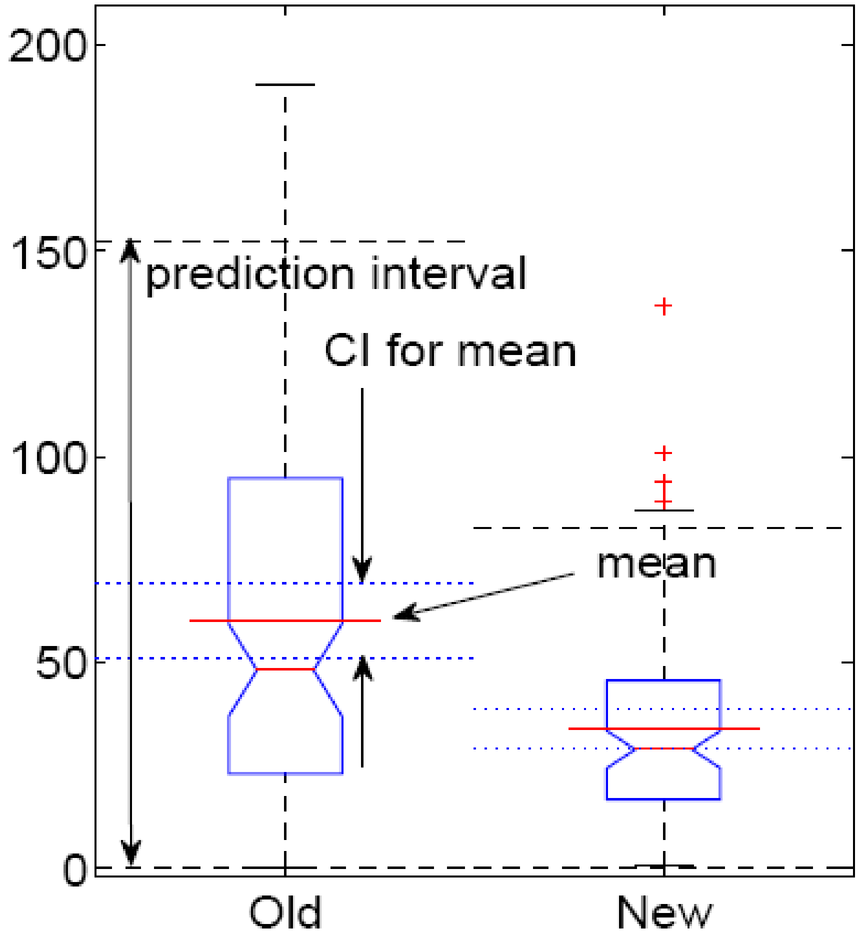
Example



quantiles



mean and standard deviation



Box plots

Other summarizations commonly used: Fairness indices

Jain’s fairness index

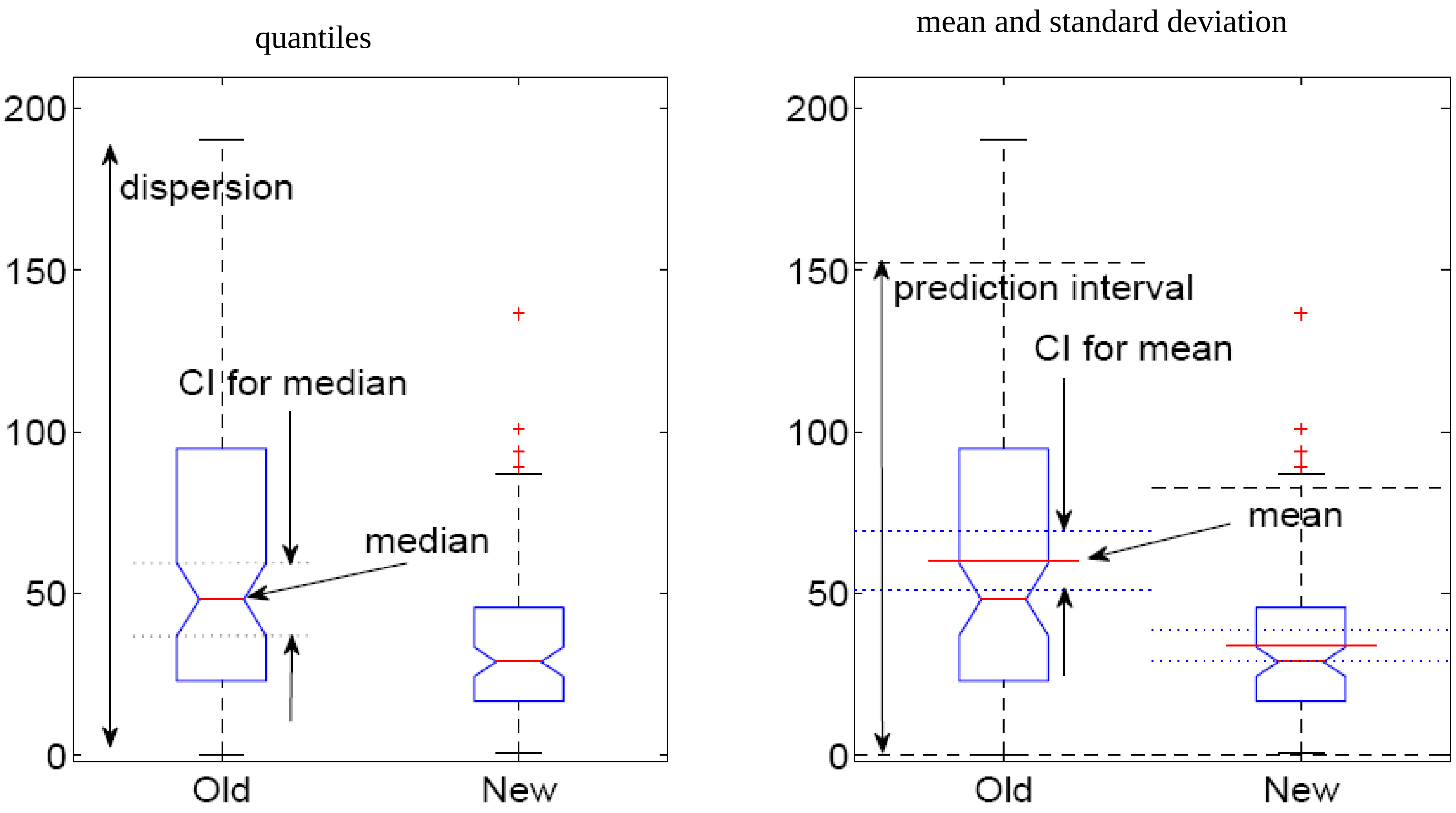
Lorenz curve gap and Gini index

see lecture notes for more details

2. Confidence Interval

Do not confuse with *prediction interval*

Confidence interval quantifies *uncertainty* about an estimation



quantiles

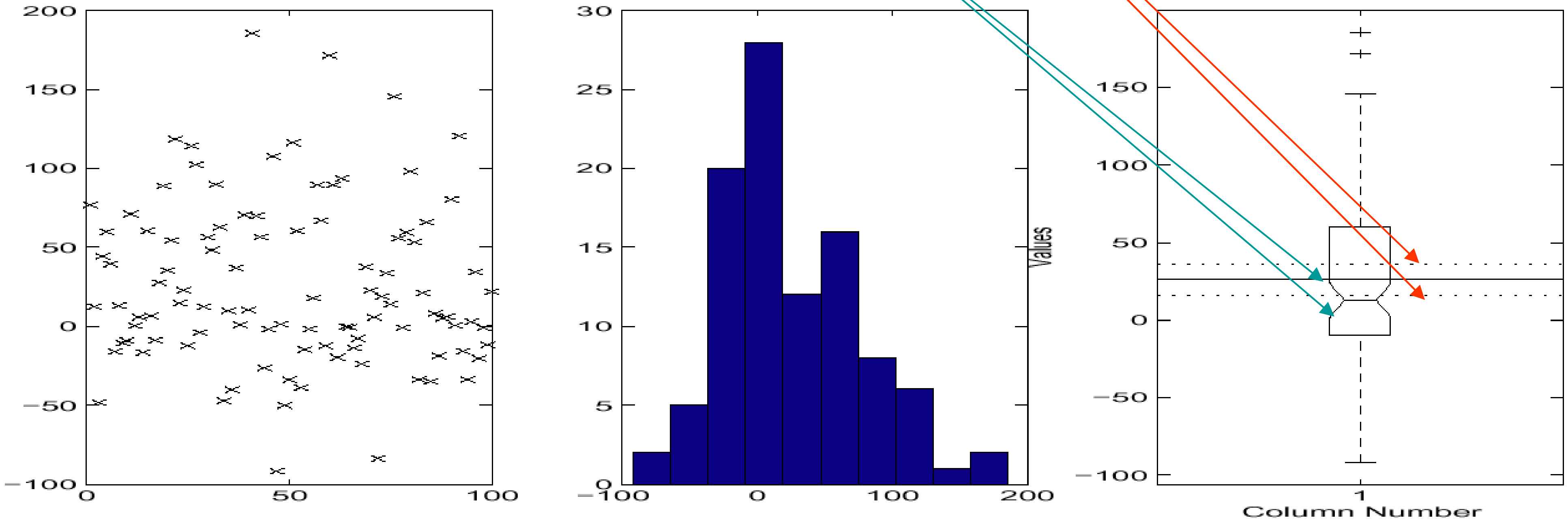
mean and standard deviation

Confidence Intervals for Mean of Difference

Mean reduction = 26.1 ± 10.2

0 is outside the confidence intervals for mean and for median

Confidence interval for median



How are Confidence Intervals computed ?

This is simple if we can assume that the data comes from an iid model

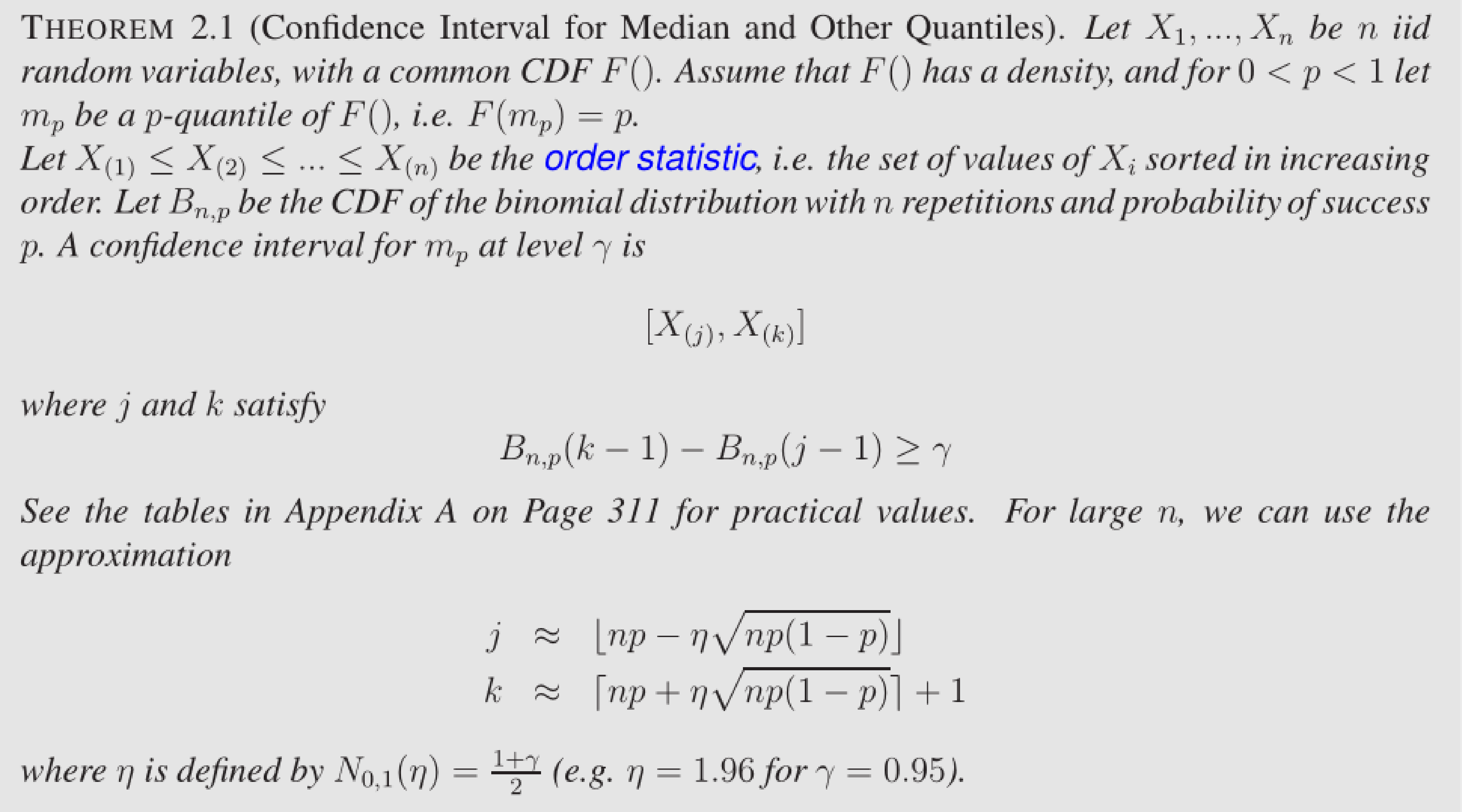
iid = Independent Identically Distributed

What is a confidence interval, really ?

1. Assume the data we have obtained is generated by a simulator, which has drawn the data from a distribution with CDF $F(\,)$
2. The distribution $F()$ has a true median $m_{0.5}$ i.e. $F(m_{0.5}) = 0.5$, which we want to estimate.
3. We have obtained $X_1, \dots, X_n$; a *point estimate* $\widehat{m}$ of $m_{0.5}$ is derived using the formula given earlier*;* $\widehat{m}$ depends on the data and is also random (hopefully not too much)
4. A *confidence interval* for $m_{0.5}$ at confidence level 95% is an interval $[u(X_1, \dots, X_n), v(X_1, \dots, X_n)]$ such that

$$P\big(u(X_1, \dots, X_n) \le m_{0.5} \le v(X_1, \dots, X_n)\big) \ge 0.95$$

The formula for CI for median



THEOREM 2.1 (Confidence Interval for Median and Other Quantiles). *Let $X_1, ..., X_n$ be n iid random variables, with a common CDF $F()$. Assume that $F()$ has a density, and for $0 < p < 1$ let m_p be a p-quantile of $F()$, i.e. $F(m_p) = p$.*

Let $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ be the order statistic, i.e. the set of values of X_i sorted in increasing order. Let $B_{n,p}$ be the CDF of the binomial distribution with n repetitions and probability of success p. A confidence interval for m_p at level γ is

$$[X_{(j)}, X_{(k)}]$$

where j and k satisfy

$$B_{n,p}(k-1) - B_{n,p}(j-1) \geq \gamma$$

See the tables in Appendix A on Page 311 for practical values. For large n, we can use the approximation

$$\begin{aligned} j &\approx \lfloor np - \eta\sqrt{np(1-p)} \rfloor \\ k &\approx \lceil np + \eta\sqrt{np(1-p)} \rceil + 1 \end{aligned}$$

where η is defined by $N_{0,1}(\eta) = \frac{1+\gamma}{2}$ (e.g. $\eta = 1.96$ for $\gamma = 0.95$).

Using the formula (level 95%)

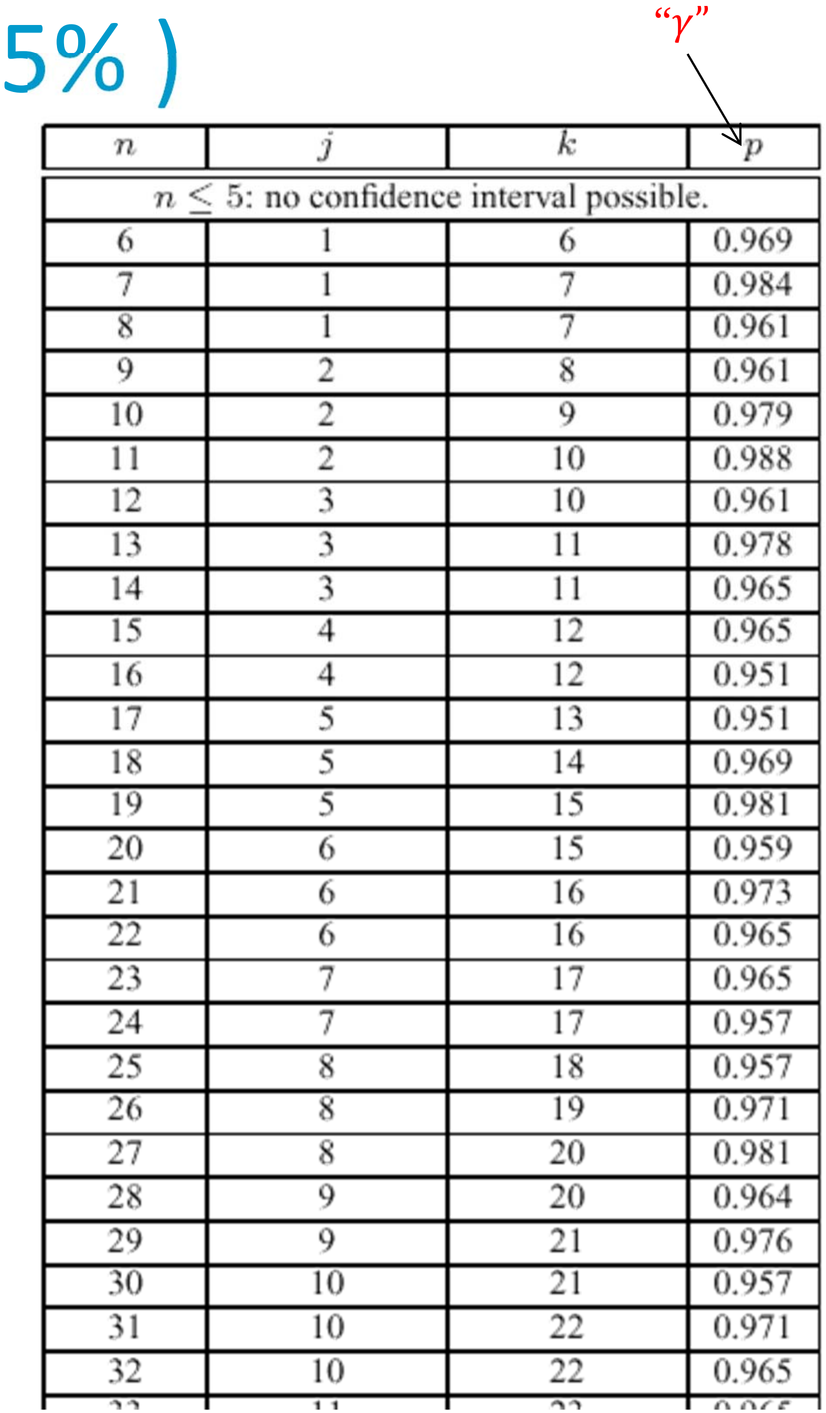
$$n = 10$$

We have 10 values $X_1, \dots, X_{10}$

$X_{(1)} \le X_{(2)} \le \cdots \le X_{(9)} \le X_{(10)}$

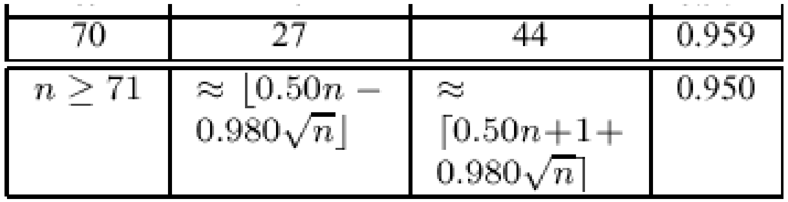
$[X_{(2)}\,;X_{(9)}]$

Simple !



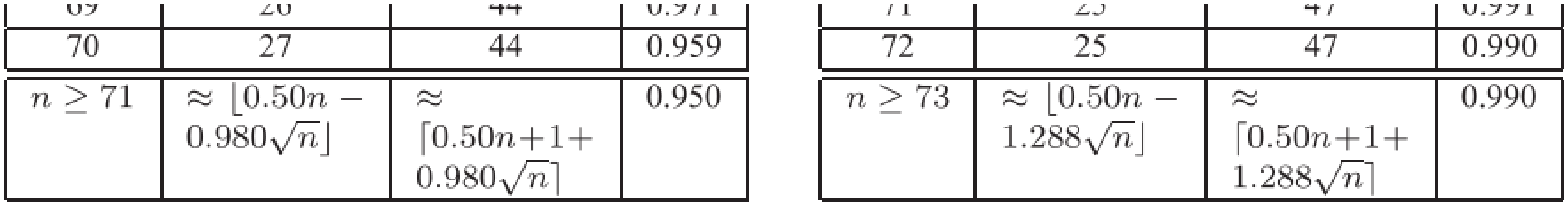
“γ”

n	j	k	p
$n \lesssim 5$: no confidence interval possible.			
6	1	6	0.969
7	1	7	0.984
8	1	7	0.961
9	2	8	0.961
10	2	9	0.979
11	2	10	0.988
12	3	10	0.961
13	3	11	0.978
14	3	11	0.965
15	4	12	0.965
16	4	12	0.951
17	5	13	0.951
18	5	14	0.969
19	5	15	0.981
20	6	15	0.959
21	6	16	0.973
22	6	16	0.965
23	7	17	0.965
24	7	17	0.957
25	8	18	0.957
26	8	19	0.971
27	8	20	0.981
28	9	20	0.964
29	9	21	0.976
30	10	21	0.957
31	10	22	0.971
32	10	22	0.965



70	27	44	0.959
$n \geq 71$	$\approx \lfloor 0.50n - 0.980\sqrt{n} \rfloor$	$\approx \lceil 0.50n+1+ 0.980\sqrt{n} \rceil$	0.950

Example n $= \mathbf{100}$, confidence interval for median



69	26	44	0.971	71	25	47	0.991
70	27	44	0.959	72	25	47	0.990
$n \geq 71$	$\approx \lfloor 0.50n - 0.980\sqrt{n} \rfloor$	$\approx \lceil 0.50n+1+ 0.980\sqrt{n} \rceil$	0.950	$n \geq 73$	$\approx \lfloor 0.50n - 1.288\sqrt{n} \rfloor$	$\approx \lceil 0.50n+1+ 1.288\sqrt{n} \rceil$	0.990

Table A.1: Quantile $q = 50\%$, Confidence Levels $\gamma = 95\%$ (left) and 0.99% (right)

The median estimate is $\frac{X_{(50)}+X_{(51)}}{2}$

At confidence level 95%

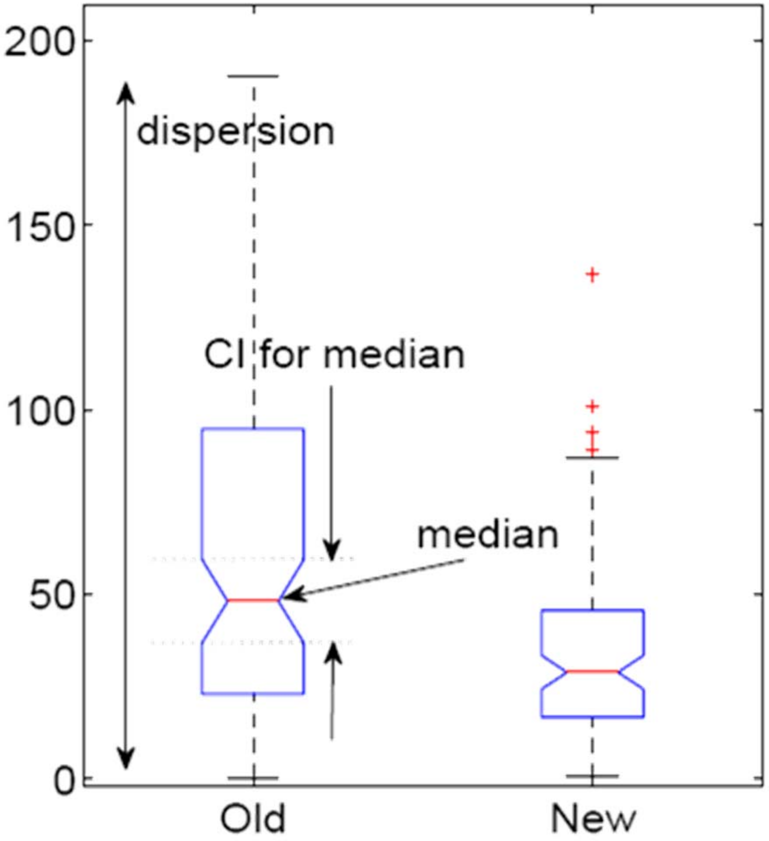
$$j = \lfloor 50 - 9.8 \rfloor = 40$$
$$k = \lceil 51 + 9.8 \rceil = 61$$

a confidence interval for the median is $[X_{(40)}; X_{(61)}]$

At confidence level 99%

$$j = \lfloor 50 - 12.8 \rfloor = 37$$
$$k = \lceil 51 + 12.8 \rceil = 64$$

a confidence interval for the median is $[X_{(37)}; X_{(64)}]$



Idea of the proof of Theorem 2.1

Say $n = 10$, we want to compute $P\left(X_{(2)} \leq m_{0.5} < X_{(9)}\right)$

Consider the game, done by an oracle who knows $m_{0.5}$:

for each data, if $X_i \leq m_{0.5}$ declare success, else declare failure

Let N be the (random) number of successes

$$(N = k) \Leftrightarrow X_{(k)} \leq m_{0.5} < X_{(k+1)}$$

Idea of the proof of Theorem 2.1

Say $n = 10$, we want to compute $P\left(X_{(2)} \leq m_{0.5} < X_{(9)}\right)$

Consider the game, done by an oracle who knows $m_{0.5}$:

for each data, if $X_i \leq m_{0.5}$ declare success, else declare failure

Let N be the (random) number of successes

$$(N = k) \Leftrightarrow \; X_{(k)} \leq m_{0.5} < X_{(k+1)}$$

Number of successes	Is $\left(X_{(2)} \leq m_{0.5}\right)$ true ?
$N = 1$	
$N = 2$	
$N = 3$	
…	…
$N = 10$	

Idea of the proof of Theorem 2.1

Say $n = 10$, we want to compute $P\left(X_{(2)} \leq m_{0.5} < X_{(9)}\right)$

Consider the game, done by an oracle who knows $m_{0.5}$:

for each data, if $X_i \leq m_{0.5}$ declare success, else declare failure

Let N be the (random) number of successes

$$(N = k) \Leftrightarrow X_{(k)} \leq m_{0.5} < X_{(k+1)}$$

Number of successes	Is $\left(m_{0.5} < X_{(9)}\right)$ true ?
$N = 1$	
…	…
$N = 8$	
$N = 9$	
$N = 10$	

Idea of the proof of Theorem 2.1

Say $n = 10$, we want to compute $P\left(X_{(2)} \leq m_{0.5} < X_{(9)}\right)$
Consider the game, done by an oracle who knows $m_{0.5}$:
for each data, if $X_i \leq m_{0.5}$ declare success, else declare failure
Let N be the (random) number of successes

$$(N = k) \Leftrightarrow \; X_{(k)} \leq m_{0.5} < X_{(k+1)}$$

The event $\left(X_{(2)} \leq m_{0.5}\right)$ means : $2 \leq N$

The event $\left(m_{0.5} < X_{(9)}\right)$ means : $N \leq 8$

The event $\left(X_{(2)} \leq m_{0.5} < X_{(9)}\right)$ means : $2 \leq N \leq 8$

The distribution of N is Binomial$(n = 10, p = 0.5)$

$$P\left(X_{(2)} \leq m_{0.5} < X_{(9)}\right) = P(2 \leq N \leq 8) = P(1 < N \leq 8)$$
$$= P(N \leq 8) - P(N \leq 1) = B_{10,0.5}(8) - B_{10,0.5}(1)$$

$P\left(X_{(2)} \leq m_{0.5} < X_{(9)}\right) = P\left(X_{(2)} \leq m_{0.5} \leq X_{(9)}\right)$ if the distribution of X_i has a density

Confidence Interval for Mean

This is the most commonly used confidence interval

But requires some assumptions to hold, may be misleading if they do not hold

There is no exact theorem as for median and quantiles, but there are asymptotic results and a heuristic.

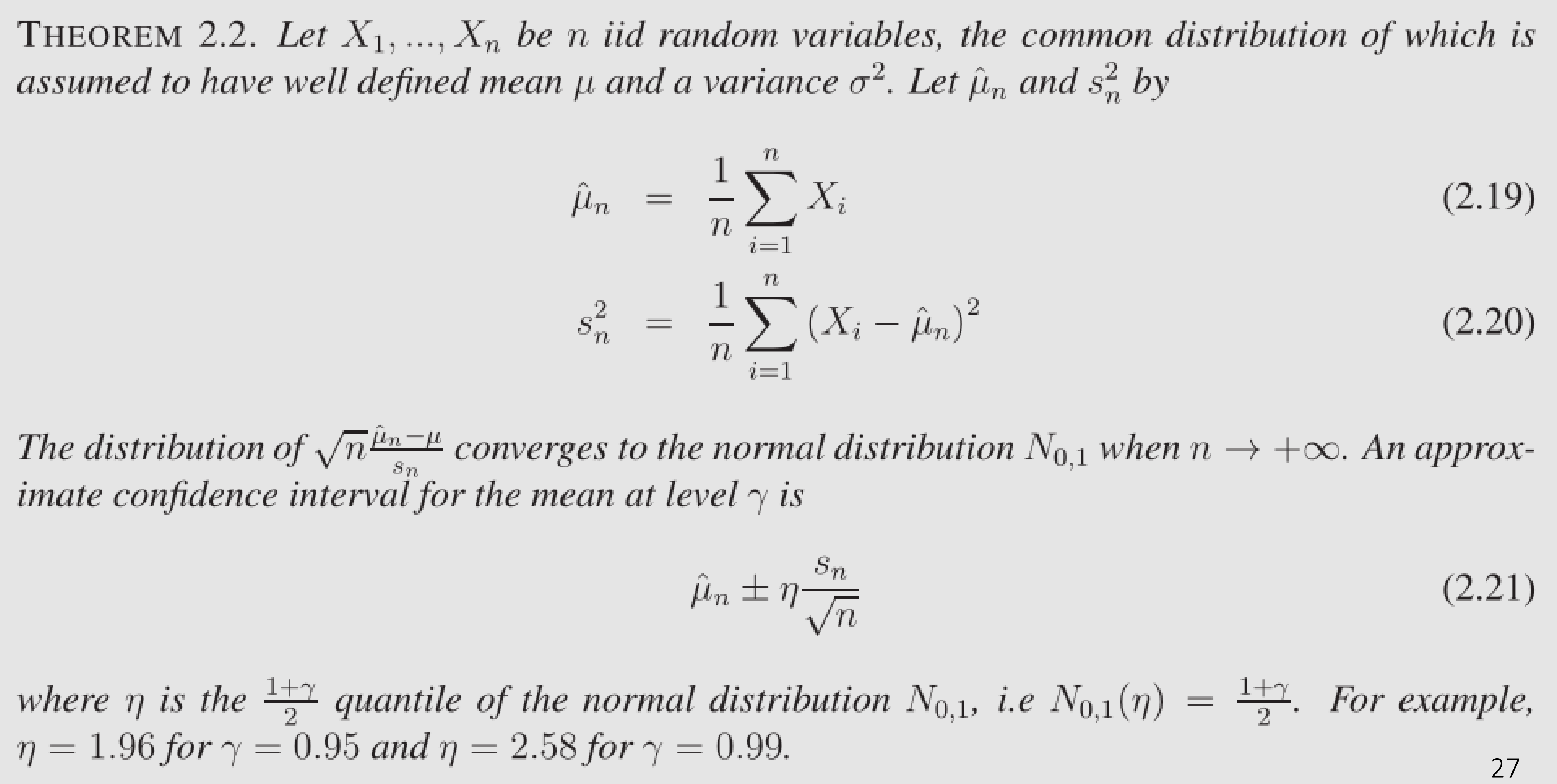
Computing a confidence interval for the mean

1. Assume the data we have obtained is generated by a simulator, which has drawn the data from a distribution with CDF $F(\quad)$
2. The distribution $F()$ has a true mean m : $\int_{-\infty}^{+\infty} x dF(x) = m$
3. We have obtained $X_1, \dots, X_n$; we *estimate* m using some formula $\widehat{m}(X_1, \dots, X_n)$; our *point estimate* $\widehat{m}$ depends on the data and is also random (hopefully not too much)
4. A *confidence interval* for m at confidence level 95% is an interval $[u(X_1, \dots, X_n), v(X_1, \dots, X_n)]$ such that

$$P\big(u(X_1, \dots, X_n) \leq m \leq v(X_1, \dots, X_n)\big) \geq 0.95$$

CI for mean, asymptotic case

If central limit theorem holds, i.e the sample mean is approx. normal (in practice: *n* is large and distribution is not “wild”- not heavy tailed -- see chapter 3)



THEOREM 2.2. *Let $X_1, \ldots, X_n$ be n iid random variables, the common distribution of which is assumed to have well defined mean μ and a variance σ^2. Let $\hat{\mu}_n$ and s_n^2 by*

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.19}$$

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \tag{2.20}$$

The distribution of $\sqrt{n}\frac{\hat{\mu}_n - \mu}{s_n}$ converges to the normal distribution $N_{0,1}$ when $n \to +\infty$. An approximate confidence interval for the mean at level γ is

$$\hat{\mu}_n \pm \eta \frac{s_n}{\sqrt{n}} \tag{2.21}$$

where η is the $\frac{1+\gamma}{2}$ quantile of the normal distribution $N_{0,1}$, i.e $N_{0,1}(\eta) = \frac{1+\gamma}{2}$. For example, $\eta = 1.96$ for $\gamma = 0.95$ and $\eta = 2.58$ for $\gamma = 0.99$.

Example

95% confidence level

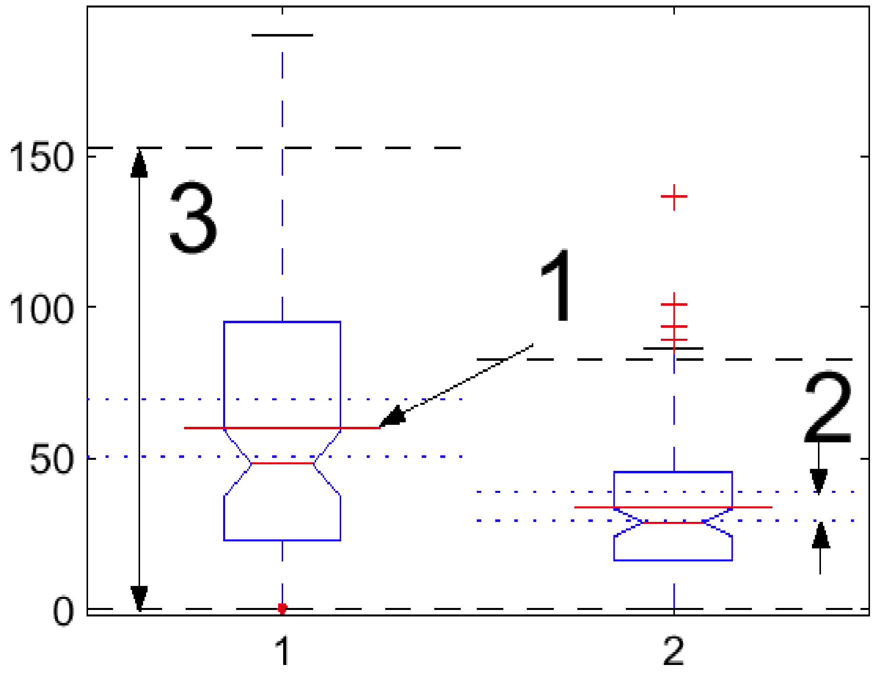
CI for mean: $\mu \pm 1.96 \frac{s}{\sqrt{n}}$ where

μ = sample mean

s = estimate of standard deviation

amplitude of CI decreases in $1/\sqrt{n}$

compare to prediction interval



Idea of Proof

$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_n$, $\mathrm{E}(\hat{\mu}_n) = \mu$, $\mathrm{var}(\hat{\mu}_n) = \frac{1}{n}\sigma^2$

central limit theorem : approximately, for large n

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_n \approx N\left(\mu, \frac{1}{n}\sigma^2\right)$$

$$s_n^2 \approx \sigma^2$$

$$\frac{\hat{\mu}_n - \mu}{\frac{1}{\sqrt{n}} s_n} \approx N(0,1)$$

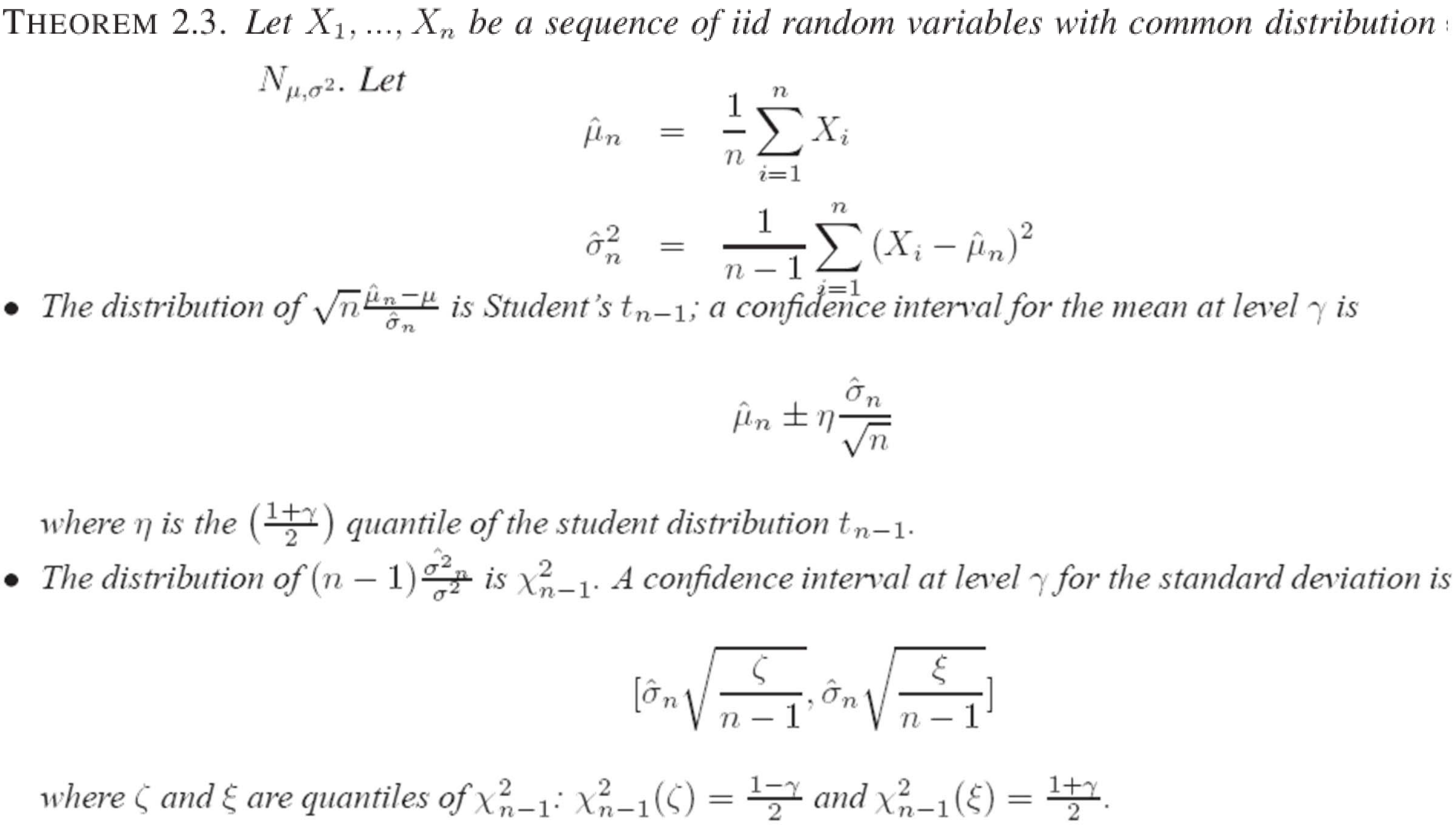
with proba 95%, a standard normal variable is in $[-1.96;\ 1.96]$, thus

$$-1.96 \leq \frac{\hat{\mu}_n - \mu}{\frac{1}{\sqrt{n}} s_n} \leq 1.96$$

CI for Mean, Normal Case

Assume data comes from an iid + *normal* distribution

Not so frequent in reality; Useful for very small data samples ($n < 30$)



THEOREM 2.3. *Let $X_1, \ldots, X_n$ be a sequence of iid random variables with common distribution N_{μ,σ^2}. Let*

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2$$

- *The distribution of $\sqrt{n}\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n}$ is Student's t_{n-1}; a confidence interval for the mean at level γ is*

$$\hat{\mu}_n \pm \eta \frac{\hat{\sigma}_n}{\sqrt{n}}$$

where η is the $\left(\frac{1+\gamma}{2}\right)$ quantile of the student distribution t_{n-1}.

- *The distribution of $(n-1)\frac{\hat{\sigma}^2_n}{\sigma^2}$ is χ^2_{n-1}. A confidence interval at level γ for the standard deviation is*

$$[\hat{\sigma}_n\sqrt{\frac{\zeta}{n-1}}, \hat{\sigma}_n\sqrt{\frac{\xi}{n-1}}]$$

where ζ and ξ are quantiles of χ^2_{n-1}: $\chi^2_{n-1}(\zeta) = \frac{1-\gamma}{2}$ and $\chi^2_{n-1}(\xi) = \frac{1+\gamma}{2}$.

Example with $\boldsymbol{n = 100}$

Theorem 2.2.2
(Large $\boldsymbol{n}$)

CI for mean: $\hat{\mu} \pm 0.196\, s$

$$s = \frac{1}{100}\sum(x_i - \hat{\mu})^2$$

$$\hat{\mu} = \frac{1}{100}\sum x_i$$

Sample variance,
= maximum likelihood estimator of variance

Theorem 2.2.3 (Normal case)

CI for mean: $\hat{\mu} \pm 0.198\, \hat{\sigma}$

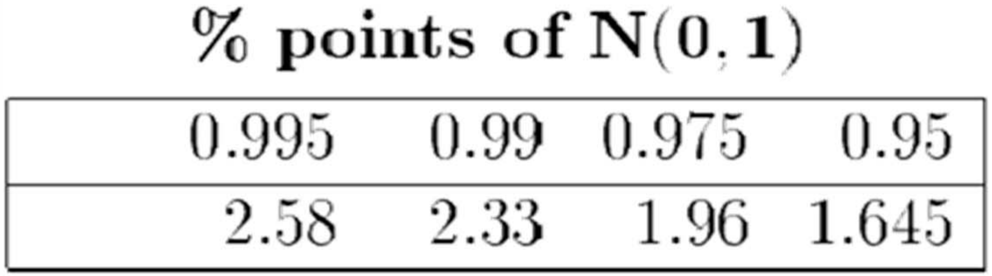
$$\hat{\sigma} = \frac{1}{99}\sum(x_i - \hat{\mu})^2$$

$$\hat{\mu} = \frac{1}{100}\sum x_i$$

Unbiased estimator of variance

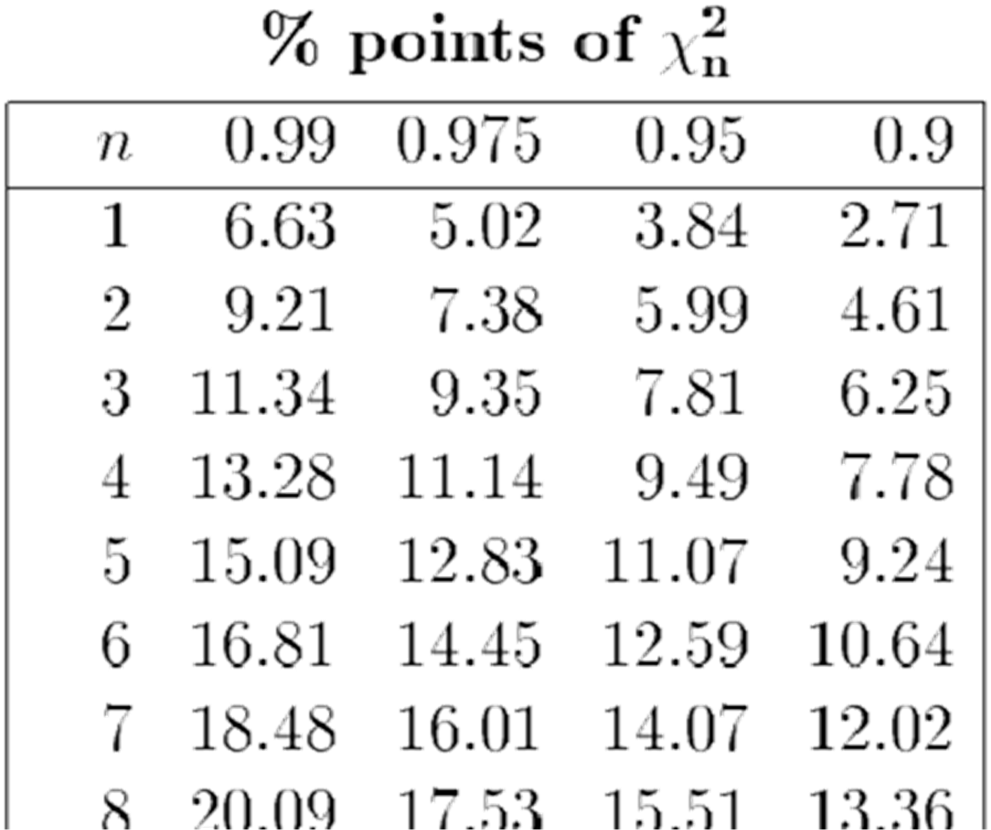
In practice both are the same if $n \geq 30$
But Theorem 2.2.2 does not require data to be normal

Tables in [Weber-Tables]



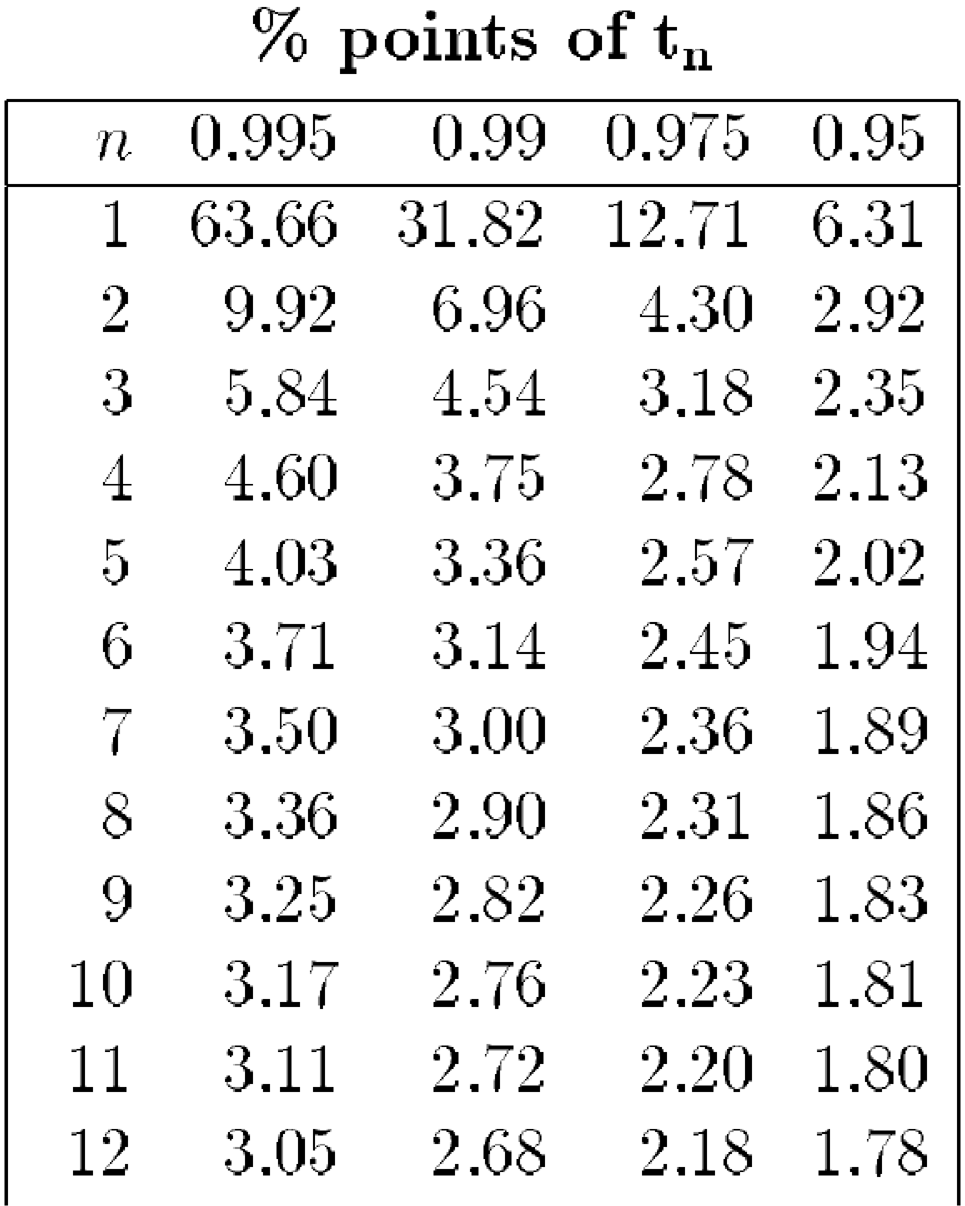
% points of $\mathbf{N(0,1)}$

0.995	0.99	0.975	0.95
2.58	2.33	1.96	1.645



% points of χ^2_n

n	0.99	0.975	0.95	0.9
1	6.63	5.02	3.84	2.71
2	9.21	7.38	5.99	4.61
3	11.34	9.35	7.81	6.25
4	13.28	11.14	9.49	7.78
5	15.09	12.83	11.07	9.24
6	16.81	14.45	12.59	10.64
7	18.48	16.01	14.07	12.02
8	20.09	17.53	15.51	13.36



% points of $\mathbf{t_n}$

n	0.995	0.99	0.975	0.95
1	63.66	31.82	12.71	6.31
2	9.92	6.96	4.30	2.92
3	5.84	4.54	3.18	2.35
4	4.60	3.75	2.78	2.13
5	4.03	3.36	2.57	2.02
6	3.71	3.14	2.45	1.94
7	3.50	3.00	2.36	1.89
8	3.36	2.90	2.31	1.86
9	3.25	2.82	2.26	1.83
10	3.17	2.76	2.23	1.81
11	3.11	2.72	2.20	1.80
12	3.05	2.68	2.18	1.78

We test a system 10’000 time for failures and find 200 failures: give a 95% confidence interval for the failure probability p.

Let $X_i = 0$ or 1 (failure / success); $E(X_i) = p$

So we are estimating the mean. The asymptotic theory applies (no heavy tail)

$$\mu_n = 0.02$$

$$s_n^2 = \frac{1}{n} \sum_{i=1 \dots n} X_i^2 - \mu_n^2 = \frac{1}{n} \sum_{i=1 \dots n} X_i^2 - \mu_n^2 = \mu_n - \mu_n^2$$

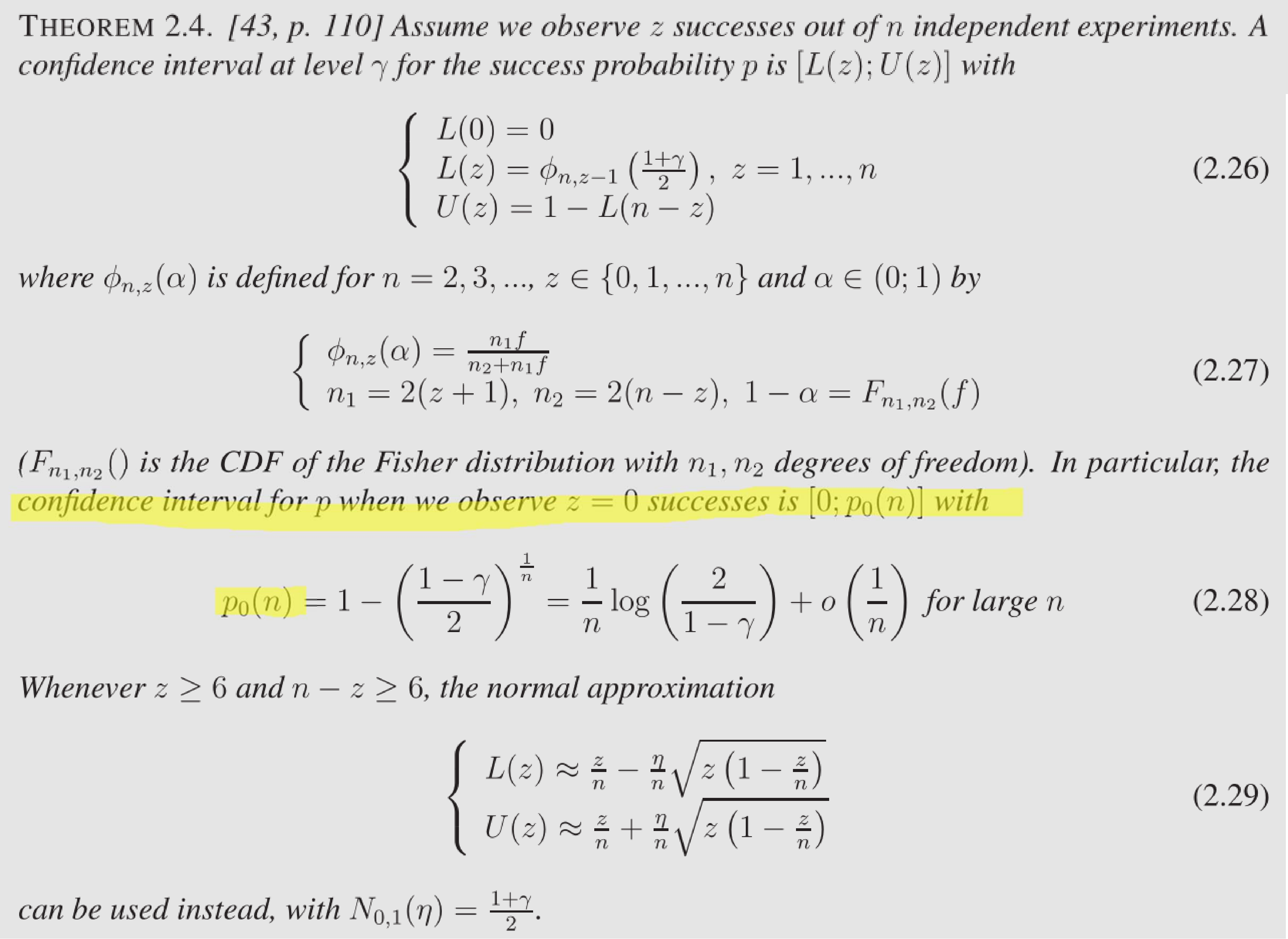
$$= \mu_n(1 - \mu_n) = 0.02 \times 0.98 \approx 0.02$$

$$s_n = \sqrt{0.02} \approx 0.14$$

Confidence Interval: $\mu_n \pm \frac{\eta s_n}{\sqrt{10000}} = 0.02 \pm 0.003$ at level 0.95

We test a system 10 time for failures and find 0 failure: give a 95% confidence interval for the failure probability p.

A. [0 ; 0]
B. [0 ; 0.1]
C. [0 ; 0.11]
D. [0 ; 0.21]
E. [0; 0.31]
F. [0; 1]
G. I do not know



THEOREM 2.4. *[43, p. 110] Assume we observe z successes out of n independent experiments. A confidence interval at level γ for the success probability p is $[L(z); U(z)]$ with*

$$\begin{cases} L(0) = 0 \\ L(z) = \phi_{n,z-1}\left(\frac{1+\gamma}{2}\right), \ \ z = 1, ..., n \\ U(z) = 1 - L(n-z) \end{cases} \tag{2.26}$$

where $\phi_{n,z}(\alpha)$ is defined for $n = 2, 3, ..., z \in \{0, 1, ..., n\}$ and $\alpha \in (0; 1)$ by

$$\begin{cases} \phi_{n,z}(\alpha) = \frac{n_1 f}{n_2 + n_1 f} \\ n_1 = 2(z+1), \ n_2 = 2(n-z), \ 1 - \alpha = F_{n_1,n_2}(f) \end{cases} \tag{2.27}$$

($F_{n_1,n_2}()$ is the CDF of the Fisher distribution with n_1, n_2 degrees of freedom). In particular, the confidence interval for p when we observe $z = 0$ successes is $[0; p_0(n)]$ with

$$p_0(n) = 1 - \left(\frac{1-\gamma}{2}\right)^{\frac{1}{n}} = \frac{1}{n}\log\left(\frac{2}{1-\gamma}\right) + o\left(\frac{1}{n}\right) \text{ for large } n \tag{2.28}$$

Whenever $z \geq 6$ and $n - z \geq 6$, the normal approximation

$$\begin{cases} L(z) \approx \frac{z}{n} - \frac{\eta}{n}\sqrt{z\left(1 - \frac{z}{n}\right)} \\ U(z) \approx \frac{z}{n} + \frac{\eta}{n}\sqrt{z\left(1 - \frac{z}{n}\right)} \end{cases} \tag{2.29}$$

can be used instead, with $N_{0,1}(\eta) = \frac{1+\gamma}{2}$.

When we observe no event in n experiments, a confidence interval for the probability of occurrence is $[0, p_0(n)]$ with

$$p_0(n) = 1 - \left(\frac{1-\gamma}{2}\right)^{\frac{1}{n}}$$

For $\gamma = 0.95$ and $n \geq 20,\ p_0(n) \approx \frac{3.689}{n}$

We test a system 10 time for failures and find 0 failure: give a 95% confidence interval for the failure probability p.

A. [0 ; 0]

B. [0 ; 0.1]

C. [0 ; 0.11]

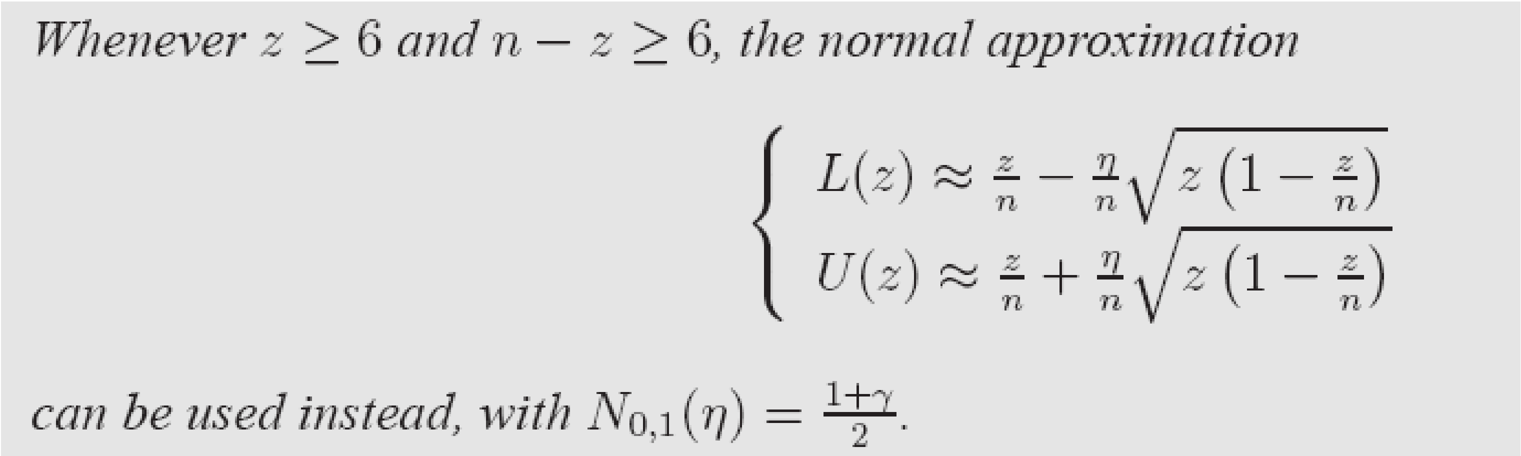
D. [0 ; 0.21]

E. [0; 0.31]

F. [0; 1]

G. I do not know

We test a system 10’000 time for failures and find 200 failures: give a 95% confidence interval for the failure probability p.



Whenever $z \geq 6$ and $n - z \geq 6$, the normal approximation

$$\begin{cases} L(z) \approx \frac{z}{n} - \frac{\eta}{n}\sqrt{z\left(1 - \frac{z}{n}\right)} \\ U(z) \approx \frac{z}{n} + \frac{\eta}{n}\sqrt{z\left(1 - \frac{z}{n}\right)} \end{cases}$$

can be used instead, with $N_{0,1}(\eta) = \frac{1+\gamma}{2}$.

Apply formula 2.29 ($z = 200 \geq 6$ and $n - z \geq 6$)

$$0.02 \pm \frac{1.96}{10000}\sqrt{200(1-0.02)} \approx 0.02 \pm \frac{1.96}{10000} 10\sqrt{2}$$
$$\approx 0.02 \pm 0.003$$

i.e. in this case it is the same as the classical formula for a confidence interval for the mean (Theorem 2.2.2).

Take Home Message

Confidence interval for median (or other quantiles) is easy to get from the Binomial distribution
Requires iid-ness, No other assumption

Confidence interval for the mean
requires iid-ness and

- ▶ Either if data sample is normal
- ▶ Or data sample is not wild (heavy-tailed) and *n* is large enough

Confidence interval for success probability requires special attention when success or failure is rare

3. The Independence Assumption

Confidence Intervals require that we can assume that the data comes from an iid model

Independent Identically Distributed

How do I know if this is true ?

- Controlled experiments: draw factors randomly with replacement
- Simulation: independent replications (with random seeds)

- Else: we do not know – in some cases we can test it using auto-correlation plots (Chapter 5)

What happens if data is not iid ?

If data is positively correlated

- Neighboring values look similar
- Frequent in measurements
- Confidence Interval is underestimated: there is less information in the data than one thinks

Doing 10 positively correlated measurements gives less information than doing 10 independent measurements

What is true when $X_1, \ldots X_n$ are independent ?

A. Observing $X_1, \ldots, X_{n-1}$ gives no information about X_n

B. Observing $X_1, \ldots, X_{n-1}$ gives no information about X_n when we know the distribution of X_n

C. $P(X_n = x_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1})$ is a function of x_n only

D. A and B

E. A and C

F. B and C

G. All

H. None

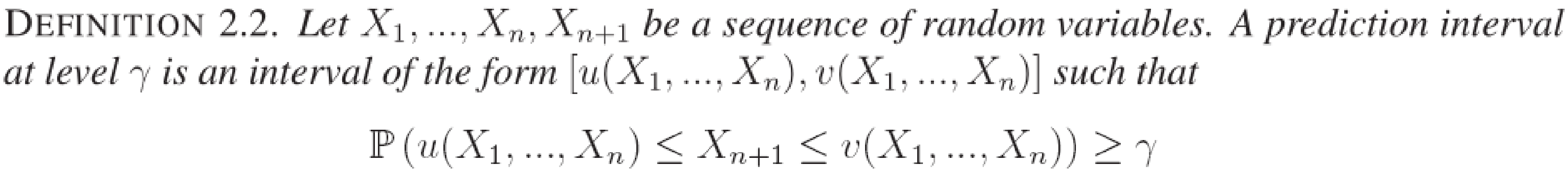
I. I don't know

4. Prediction Interval

CI for mean or median summarize

- Central value + uncertainty about it

Prediction interval is often used to summarize variability of data



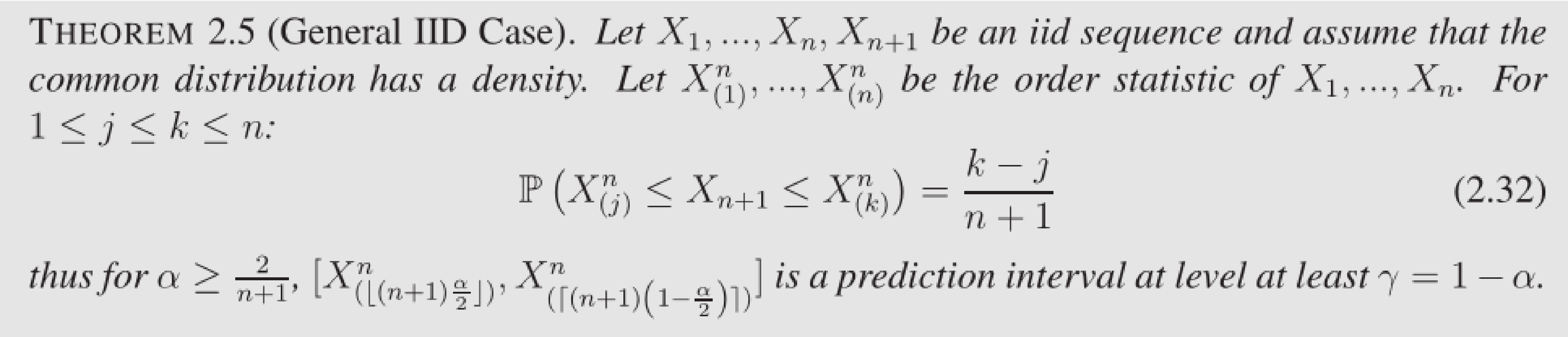
DEFINITION 2.2. *Let $X_1, \ldots, X_n, X_{n+1}$ be a sequence of random variables. A prediction interval at level γ is an interval of the form $[u(X_1, \ldots, X_n), v(X_1, \ldots, X_n)]$ such that*

$$\mathbb{P}\left(u(X_1, \ldots, X_n) \leq X_{n+1} \leq v(X_1, \ldots, X_n)\right) \geq \gamma$$

Prediction Interval based on Order Statistic

Assume data comes from an iid model

Simplest and most robust result (not well known, though):



THEOREM 2.5 (General IID Case). *Let $X_1, ..., X_n, X_{n+1}$ be an iid sequence and assume that the common distribution has a density. Let $X^n_{(1)}, ..., X^n_{(n)}$ be the order statistic of $X_1, ..., X_n$. For $1 \leq j \leq k \leq n$:*

$$\mathbb{P}\left(X^n_{(j)} \leq X_{n+1} \leq X^n_{(k)}\right) = \frac{k-j}{n+1} \tag{2.32}$$

thus for $\alpha \geq \frac{2}{n+1}$, $\left[X^n_{(\lfloor (n+1)\frac{\alpha}{2} \rfloor)}, X^n_{(\lceil (n+1)\left(1-\frac{\alpha}{2}\right) \rceil)}\right]$ is a prediction interval at level at least $\gamma = 1 - \alpha$.

For example: $n = 999$; a prediction interval at level 0.95 is

$$[X_{(25)} \,; X_{(975)}]$$

Prediction Interval for small *n*

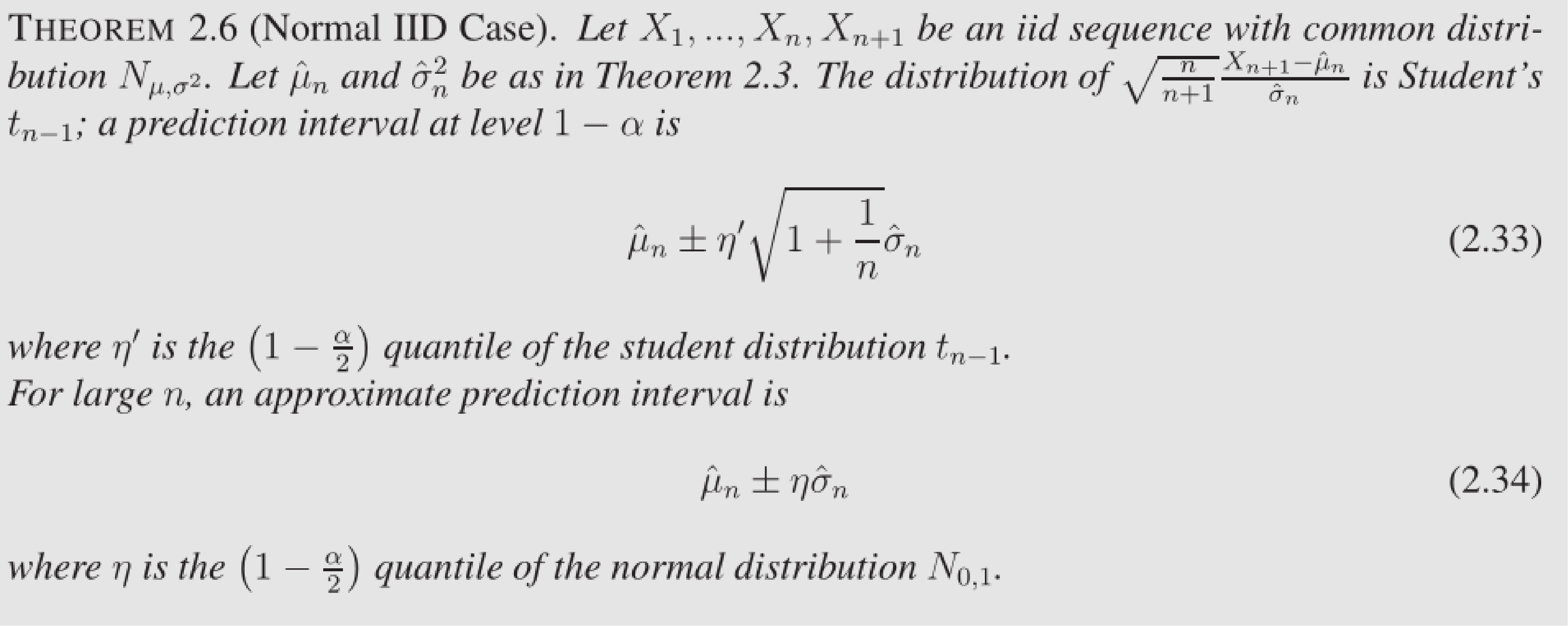
For $n = 39, [x_{\min}, x_{\max}]$ is a prediction interval at level 95%

For $n < 39$ there is no prediction interval at level 95% with this method

- But there is one at level 90% for $n > 18$
- For $n = 10$ we have a prediction interval $[x_{\min}, x_{\max}]$ at level 81%

Prediction Interval based on Mean

Very often used – but see later



THEOREM 2.6 (Normal IID Case). *Let $X_1, ..., X_n, X_{n+1}$ be an iid sequence with common distribution N_{μ,σ^2}. Let $\hat{\mu}_n$ and $\hat{\sigma}^2_n$ be as in Theorem 2.3. The distribution of $\sqrt{\frac{n}{n+1}}\frac{X_{n+1}-\hat{\mu}_n}{\hat{\sigma}_n}$ is Student's t_{n-1}; a prediction interval at level $1-\alpha$ is*

$$\hat{\mu}_n \pm \eta' \sqrt{1+\frac{1}{n}}\hat{\sigma}_n \tag{2.33}$$

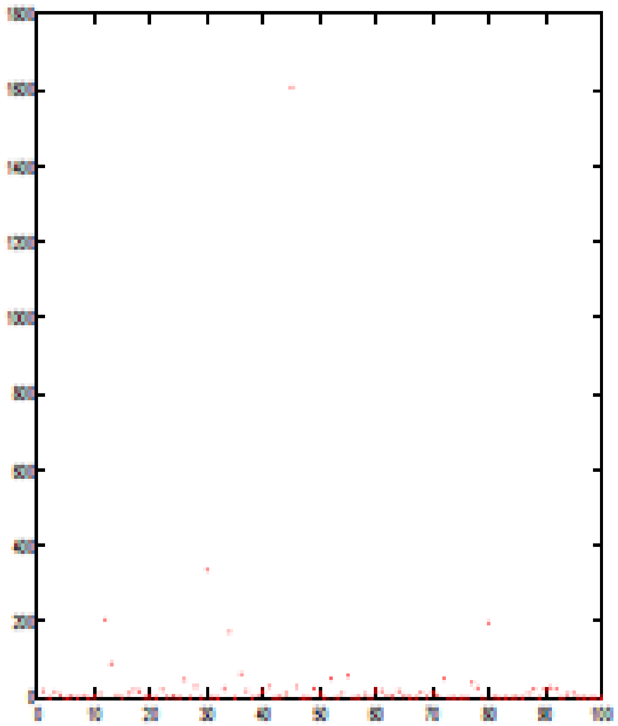
where η' is the $\left(1-\frac{\alpha}{2}\right)$ quantile of the student distribution t_{n-1}.
For large n, an approximate prediction interval is

$$\hat{\mu}_n \pm \eta \hat{\sigma}_n \tag{2.34}$$

where η is the $\left(1-\frac{\alpha}{2}\right)$ quantile of the normal distribution $N_{0,1}$.

Prediction interval at level 95% = $\mu \pm 1.96\ s$

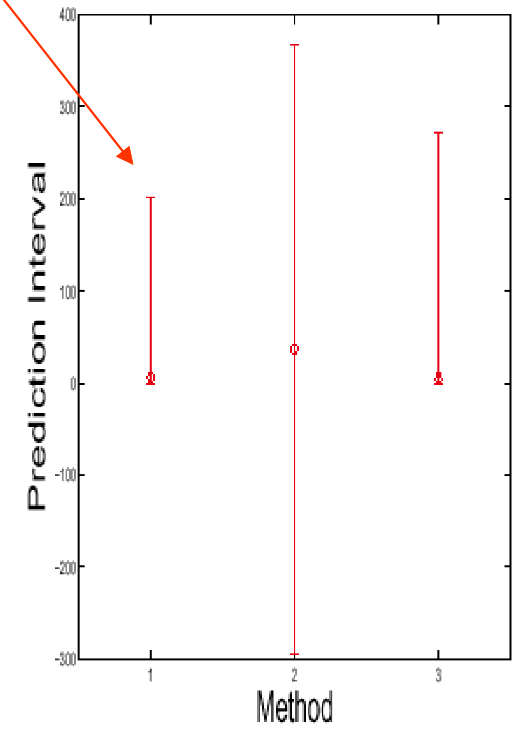
Prediction Intervals for File Transfer Times



(a) (Data)

order statistic
thm 2.4.1

mean and
standard deviation



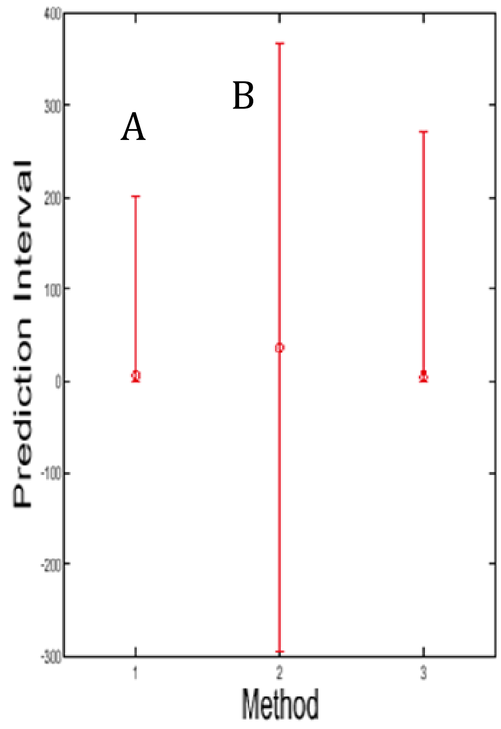
(c) (Prediction Intervals)

Which one is a "valid" prediction interval ?

A. The left one
B. The middle one
C. Both
D. I don't know

order statistic thm 2.4.1

mean and standard deviation



(c) (Prediction Intervals)

Re-Scaling

Many results are simple if the data is normal, or close to it (i.e. not wild). An important question to ask is: can I change the *scale* of my data to have it look more normal.

- Ex: log of the data instead of the data

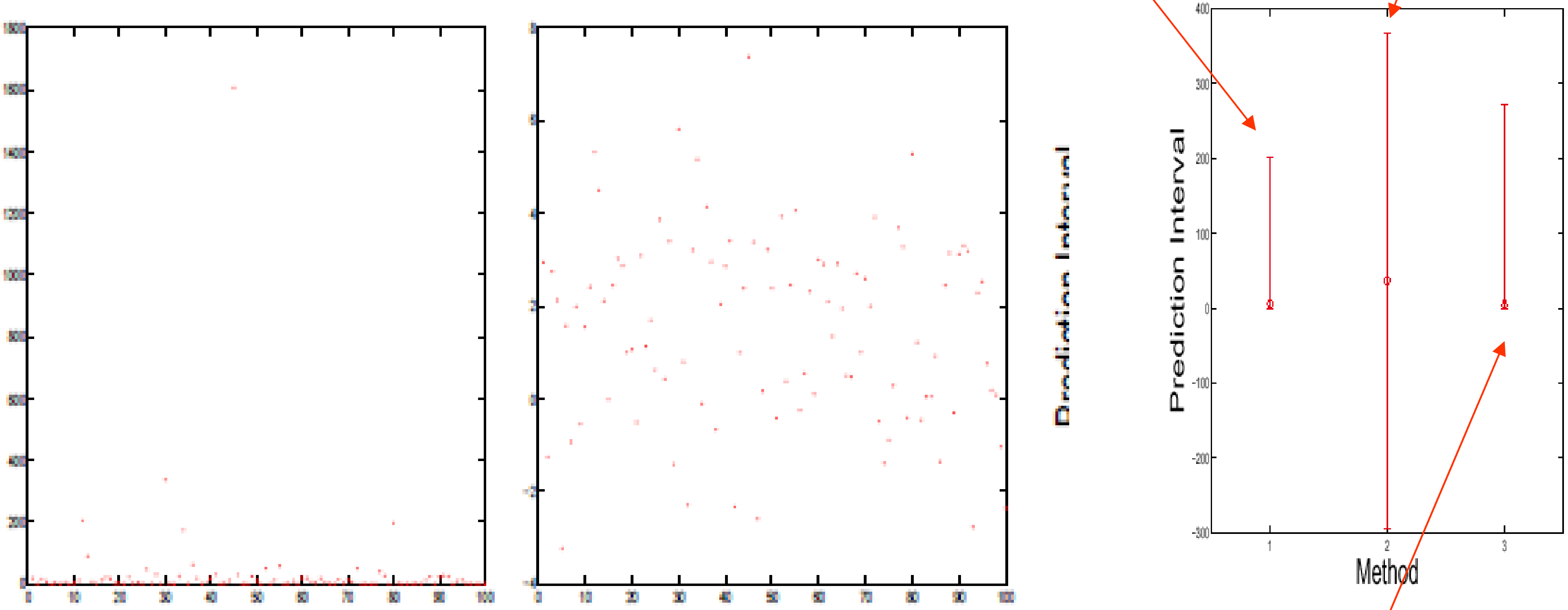
A generic transformation used in statistics is the *Box-Cox* transformation:

$$b_{s(x)} = \begin{cases} \dfrac{x^s - 1}{s}, & s \neq 0 \\ \log(x), & s = 0 \end{cases}$$

- Continuous in s

$$\begin{aligned} s &= 0: \ \log \\ s &= -1: \ 1/x \\ s &= 1: \ \text{identity} \end{aligned}$$

Prediction Intervals for File Transfer Times



(a) (Data)

(b) (Log of data)

(c) (Prediction Intervals)

order statistic thm 2.4.1

mean and standard deviation

mean and standard deviation on rescaled data

How is the prediction interval $[\boldsymbol{U},\boldsymbol{V}]$computed when re-scaling ?

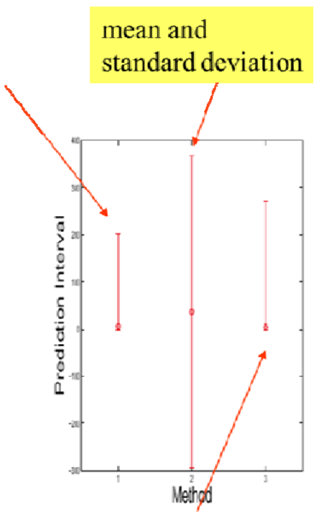
A. $U = \mu - 1.96\,\sigma,\ V = \mu + 1.96\,\sigma,$
μ is the mean, σ is the std-deviation of the log of the data

B. $U = e^{u}, V = e^{v}$ with
$u = \mu - 1.96\,\sigma, v = \mu + 1.96\,\sigma,$
μ is the mean, σ is the std-deviation of the data

C. $U = e^{u}, V = e^{v}$ with
$u = \mu - 1.96\,\sigma, v = \mu + 1.96\,\sigma,$
μ is the mean, σ is the std-deviation of the log of the data

D. None of these

E. I don't know



mean and standard deviation

(c) (Prediction Intervals)

mean and standard deviation on rescaled data

QQplot is common tool for verifying normal assumption

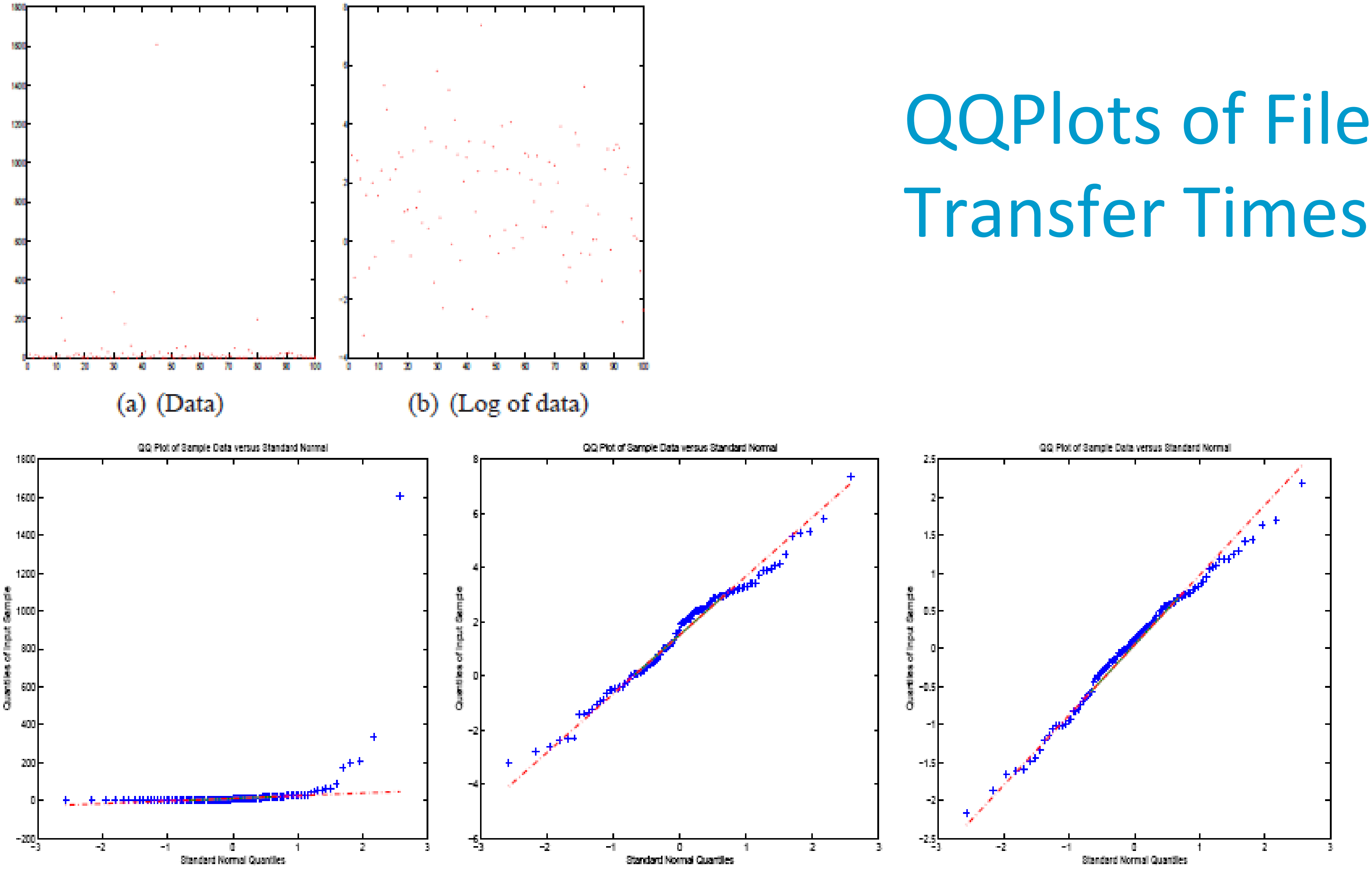
Normal Qqplot

- X-axis: standard normal quantiles $x_i = F^{-1}\left(\frac{i}{n+1}\right)$
- Y-axis: Ordered statistic of sample: $X_{(1)}, X_{(2)}, \dots, X_{(n)}$

If data comes from a normal distribution, qqplot is close to a straight line (except for end points)

- Visual inspection is often enough
- If not possible or doubtful, we will use tests later (Chapter 4)

QQPlots of File Transfer Times



(a) (Data) (b) (Log of data)

(a) (QQ-plot) (b) (QQ-plot of log of data) (c) (normal sample)

Figure 2.13: Normal qqplots of file transfer times in Figure 2.12 and of an artificially generated sample from the normal distribution with the same number of points. The former plot shows large deviation from normality, the second does not.

Prediction Interval
versus
Confidence Interval

If data is assumed normal

μ = estimated mean
s^2 = estimated variance

Confidence interval for mean at level 95 % $= \mu \pm \frac{1.96}{\sqrt{n}}\ s$

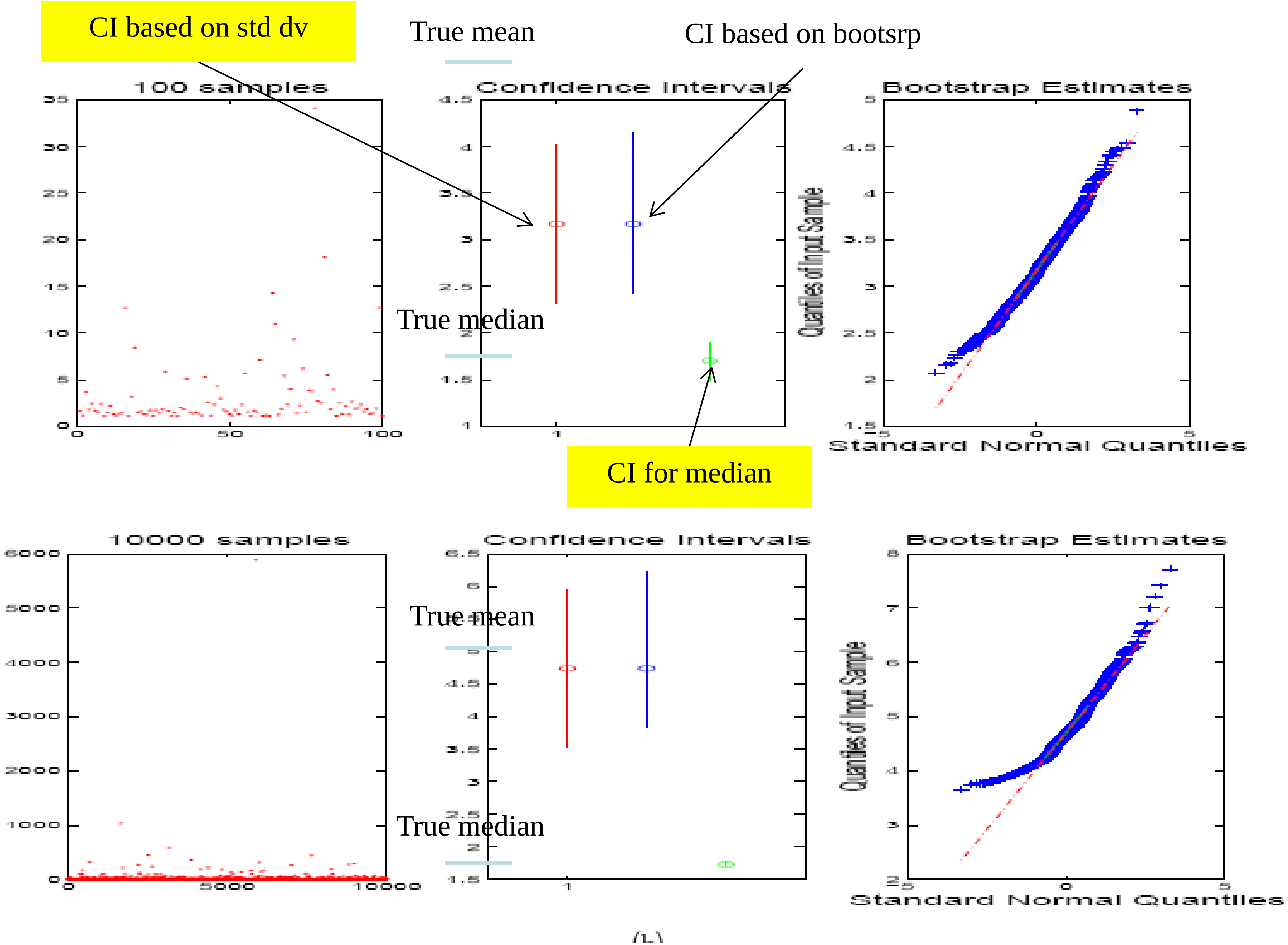
Prediction interval at level 95% $= \mu \pm 1.96\ s$

5. Which Summarization to Use ?

Issues

- Robustness to outliers
- Distribution assumptions

A Distribution with Infinite Variance



CI based on std dv

True mean

CI based on bootsrp

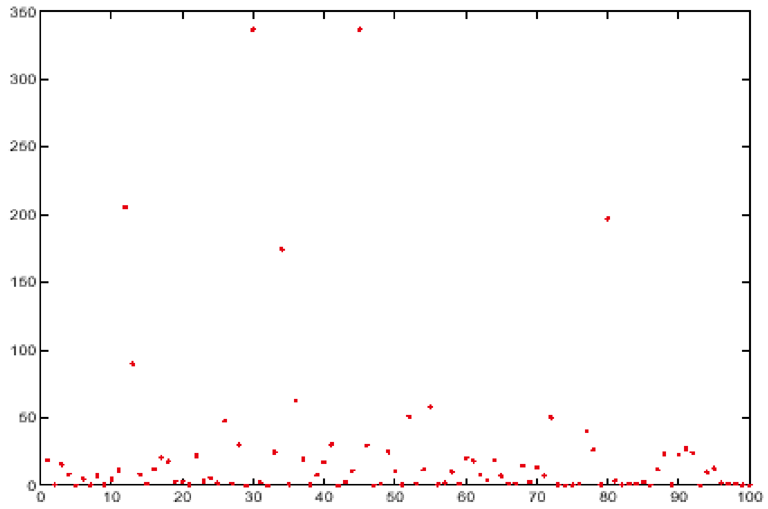
True median

CI for median

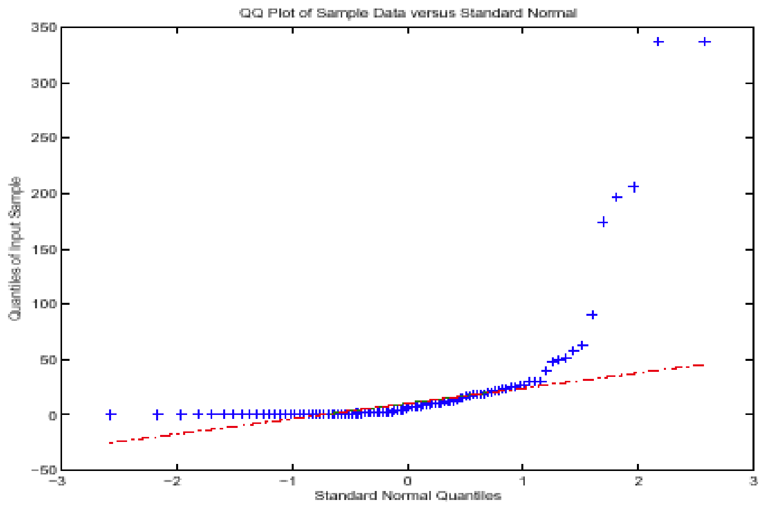
True mean

True median

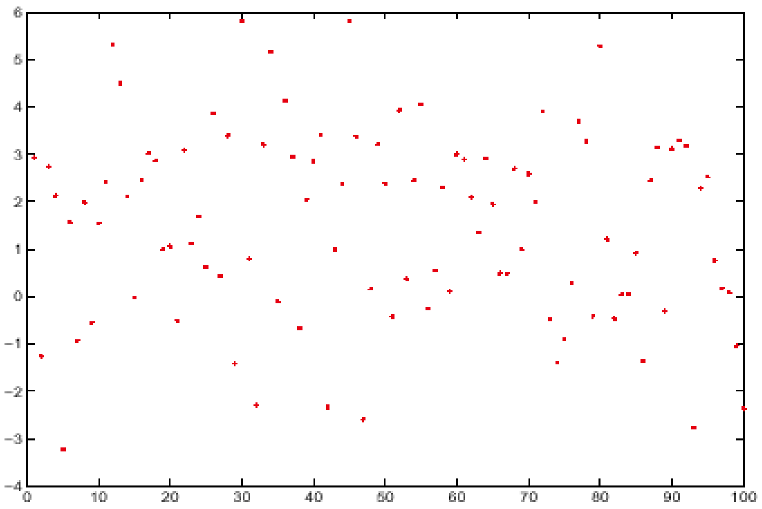
Outlier in File Transfer Time



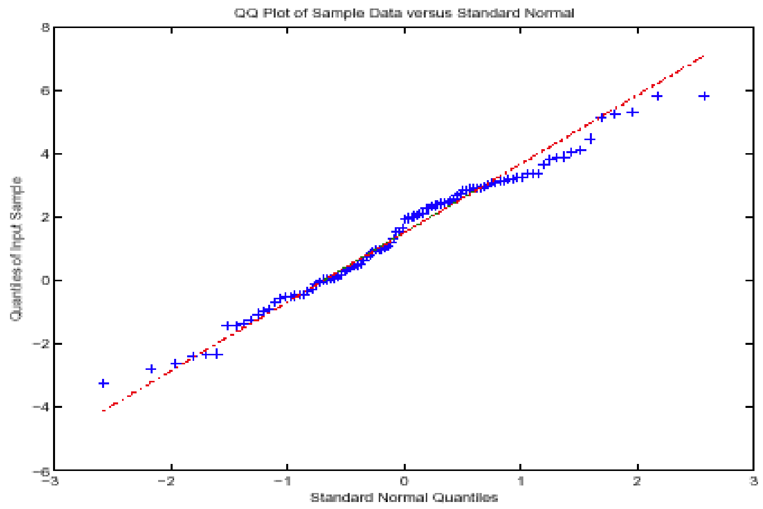
(a) (Data without outlier)



(b) (QQ-plot of (a))

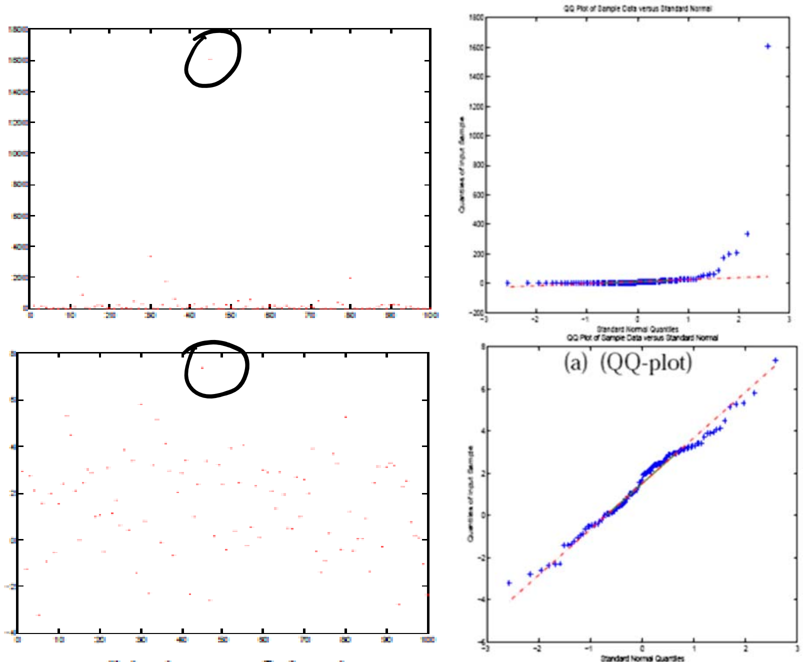


(c) (Log of data without outlier)



(d) (QQ-plot of (c))

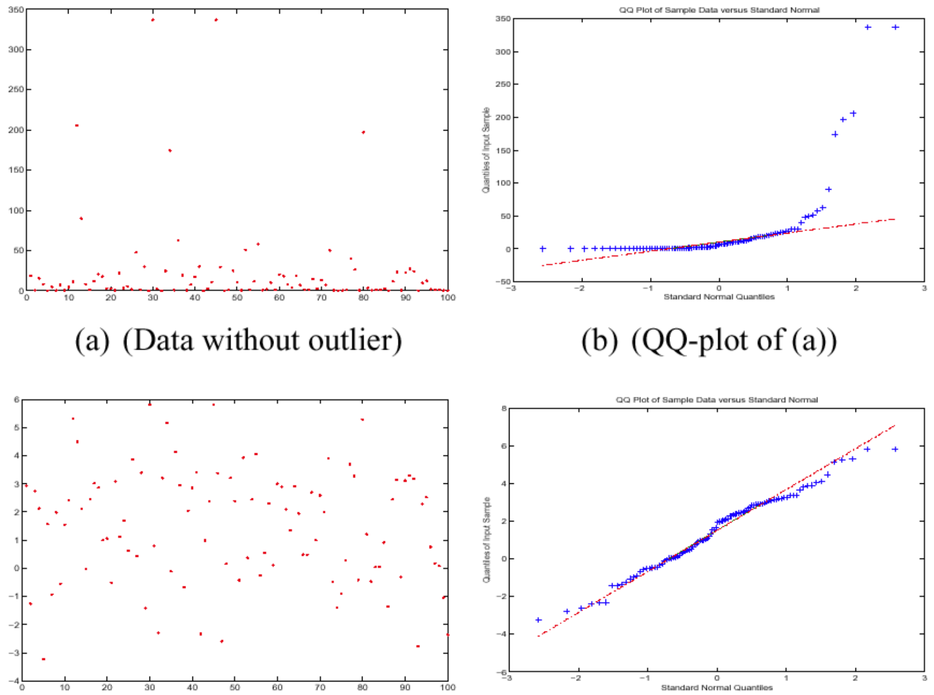
Outlier in File Transfer Time



(a) (QQ-plot)

(b) (Log of data)

(b) (QQ-plot of log of data)



(a) (Data without outlier)

(b) (QQ-plot of (a))

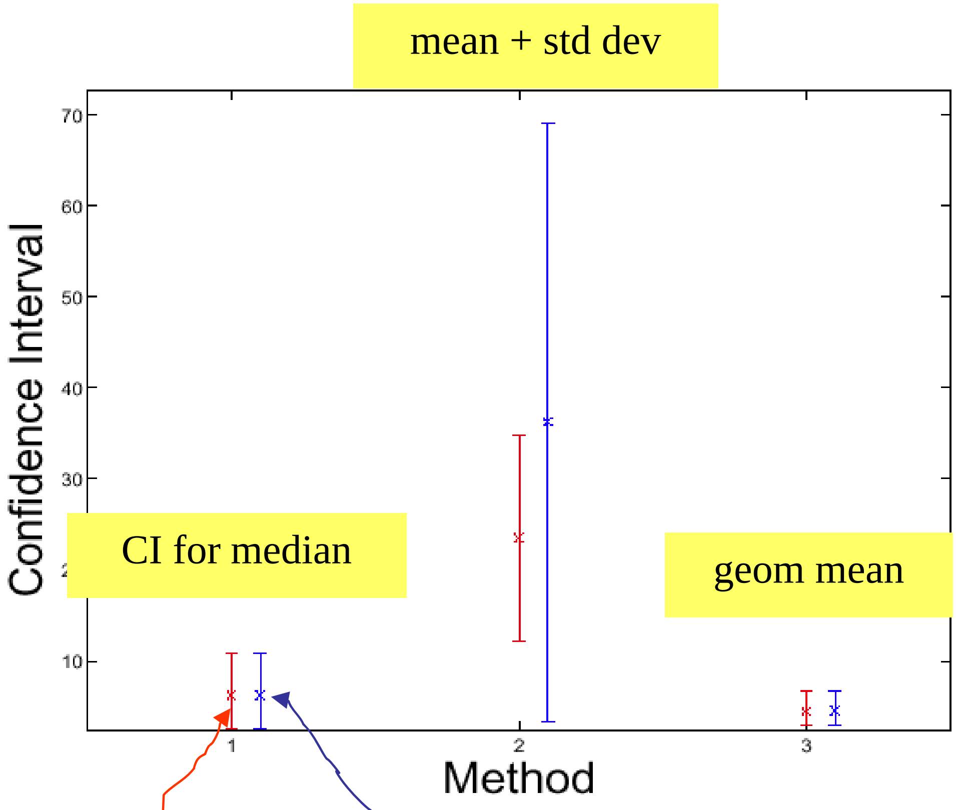
(c) (Log of data without outlier)

(d) (QQ-plot of (c))

Original data, showing one outlier

Outlier removed

Robustness of Conf/Prediction Intervals

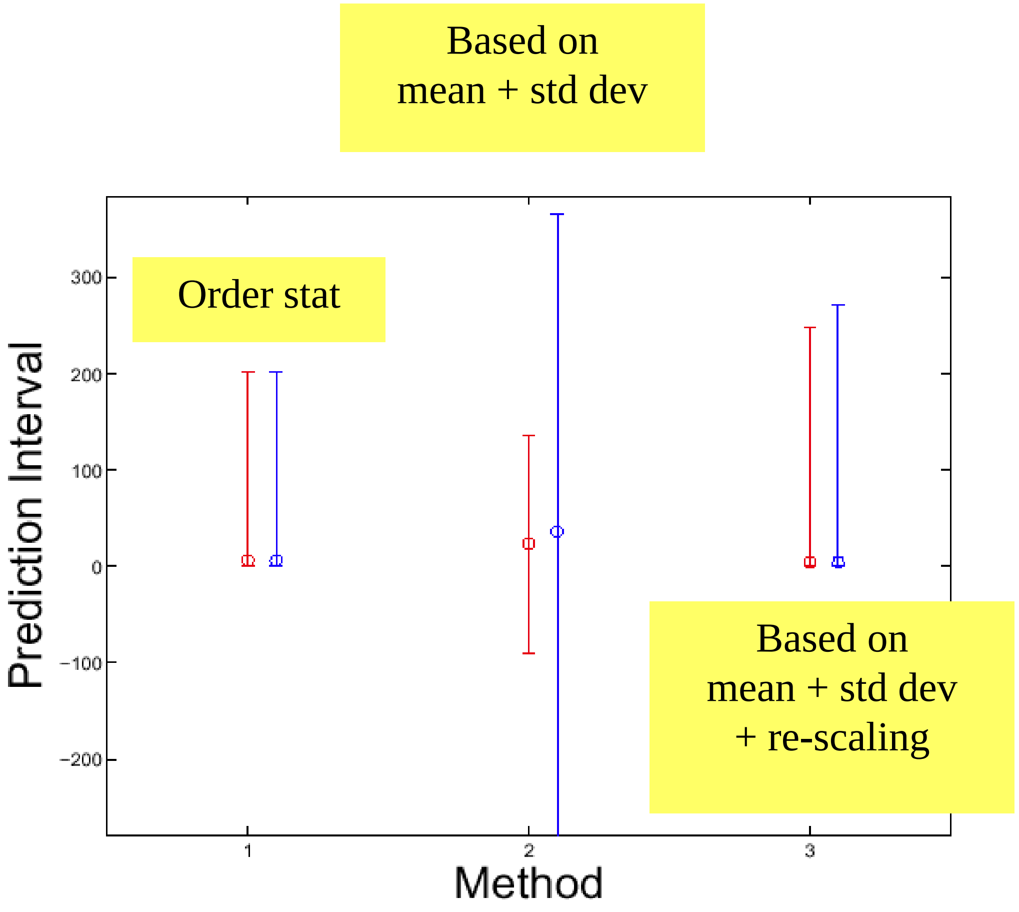


mean + std dev

CI for median

geom mean

(e) (Confidence Intervals)



Based on
mean + std dev

Order stat

Based on
mean + std dev
+ re-scaling

(f) (Prediction Intervals)

Outlier removed
Outlier present

Robustness

Methods based on quantiles and order statistics are robust to outliers and do not make any distributional assumption other than iid

Methods based on mean and standard deviation are sensitive to outliers and make assumptions about distributions
--- use with care

Since methods based on mean and standard deviation are less robust, why are they used at all ?

A. By intellectual laziness
B. Because they are easier to compute
C. They are more compact
D. I don’t know

Take-Home Message

Use methods that you understand

Mean and standard deviation make sense when data sets are not wild

- Close to normal, or not heavy tailed and large data sample

Use quantiles and order statistics if you have the choice

Rescale if needed