Summarizing Performance Data
Confidence Intervals

Important

Easy to Difficult

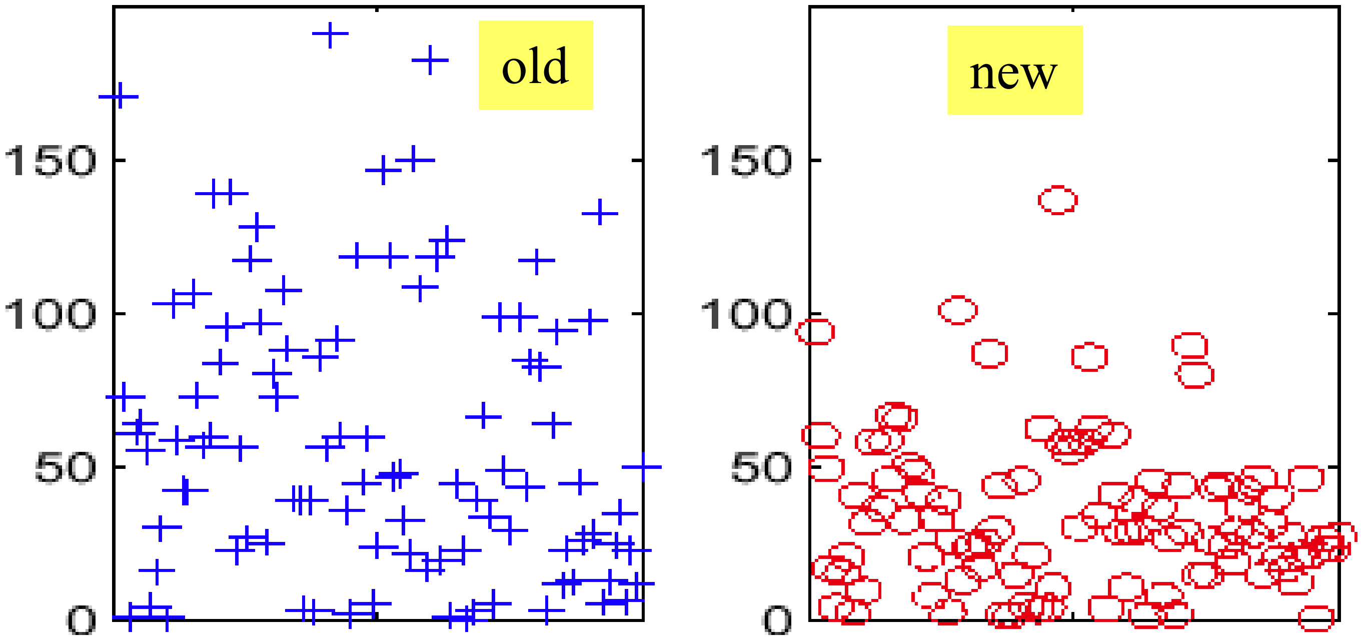
Warning: some mathematical content

Contents

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

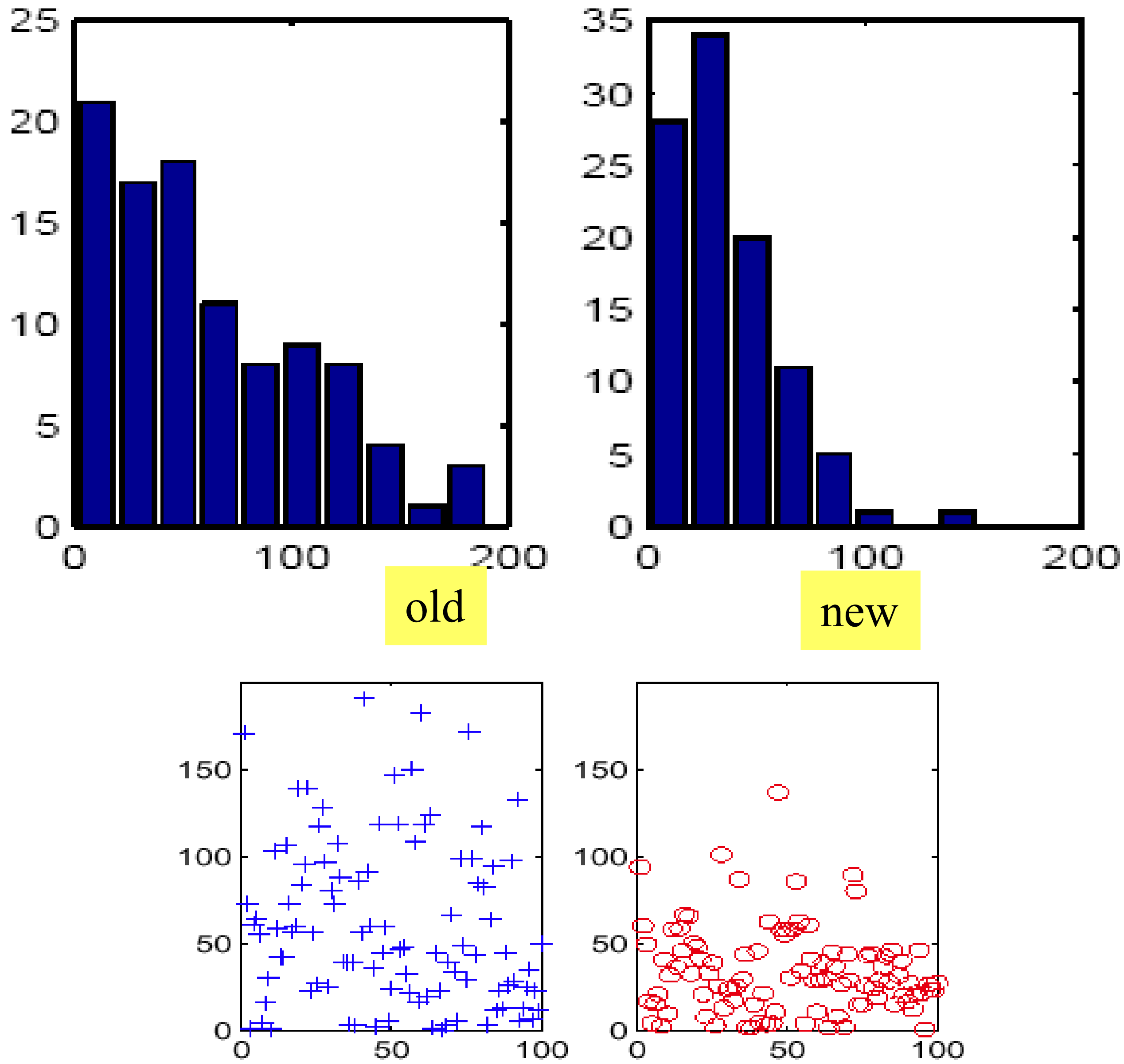
1 Summarizing Performance Data

- 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



ECDF allow easy comparison

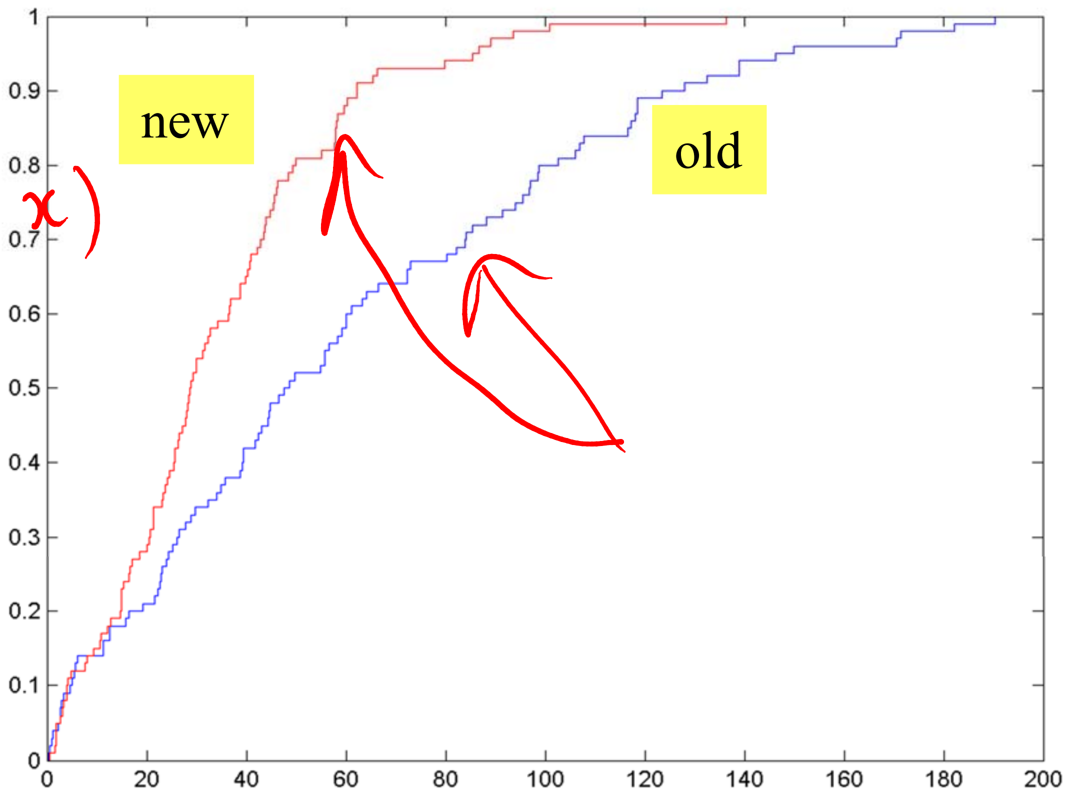
Comparing Data Sets is easily done with their *empirical cumulative distribution functions* (ECDFs). The ECDF of a data set $x_1, \ldots, x_n$ is the function f defined by

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

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

$F(x) = \mathbb{P}(X \le x)$



Summarized Measures

- Median, Quantiles
  - Median — If n is odd, the median is $x_{(\frac{n+1}{2})}$, else $\frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$
  - Quartiles — 25%, 75%
  - P-quantiles

- Mean and standard deviation
  - Mean

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

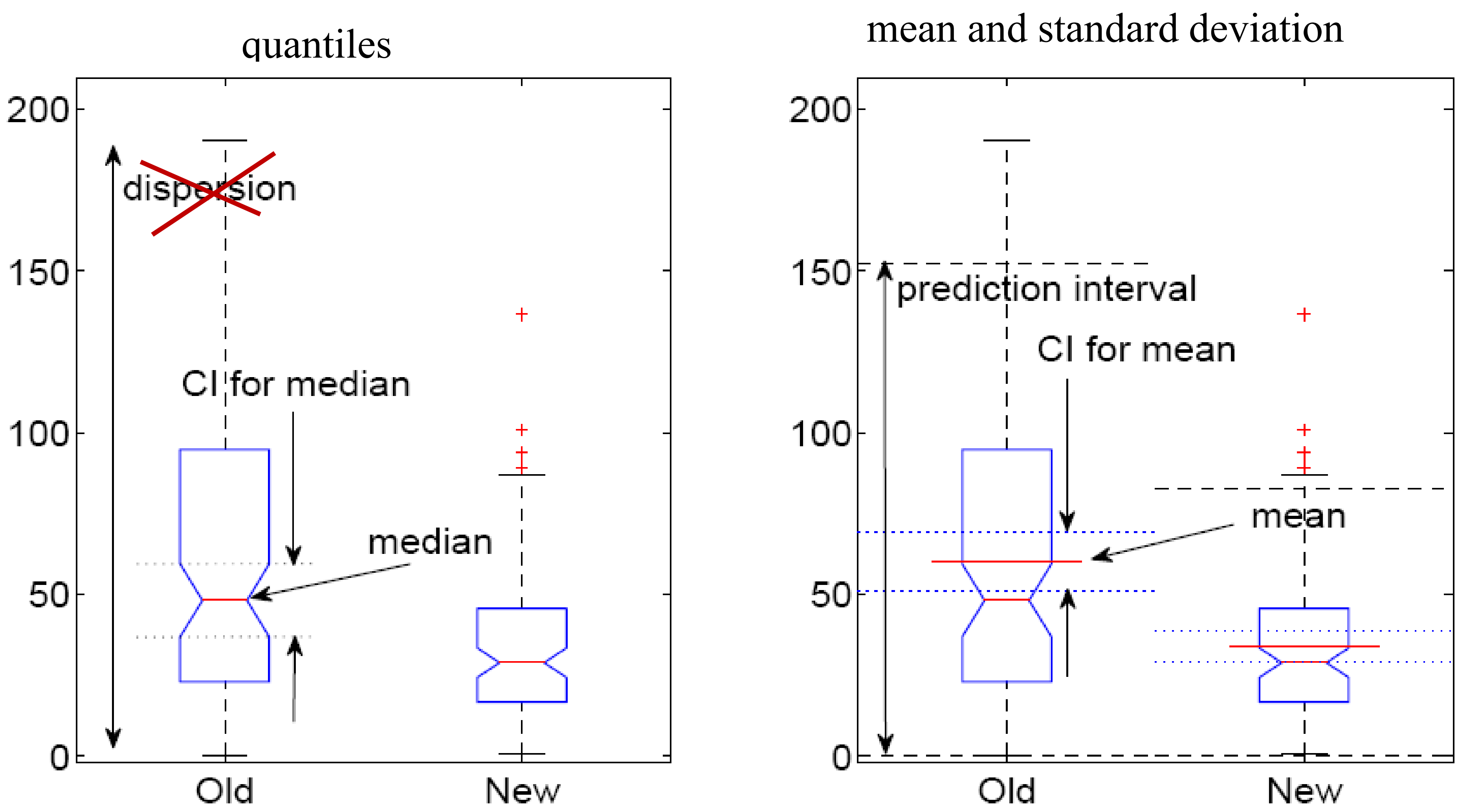
  - Standard deviation

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

  - What is the interpretation of standard deviation ?
  - A: if data is normally distributed, with 95% probability, a new data sample lies in the interval $m \pm 1.96s$

NORMAL DISTRIBUTION
95% proba

Example


quantiles
200
150
100
50
0
dispersion
CI for median
median
Old
New
mean and standard deviation
200
150
100
50
0
prediction interval
CI for mean
mean
Old
New

Coefficient of Variation Summarizes Variability

- Scale free
- Second order variability

$$\text{CoV} = \frac{s}{m}$$

: m is the mean and s the standard deviation.

- For a data set with n samples

$$0 \leq \text{CoV} \leq \sqrt{n-1}$$

- Exponential distribution: CoV =1
- What does CoV = 0 mean ?

Lorenz Curve Gap is an Alternative to CoV

- Alternative to CoV: First-order variability

$$\text{MAD} = \frac{1}{n}\sum_{m=1}^{n}|x_i - m|$$

Mean Absolute Deviation

$$\text{gap} = \frac{\text{MAD}}{2m}$$

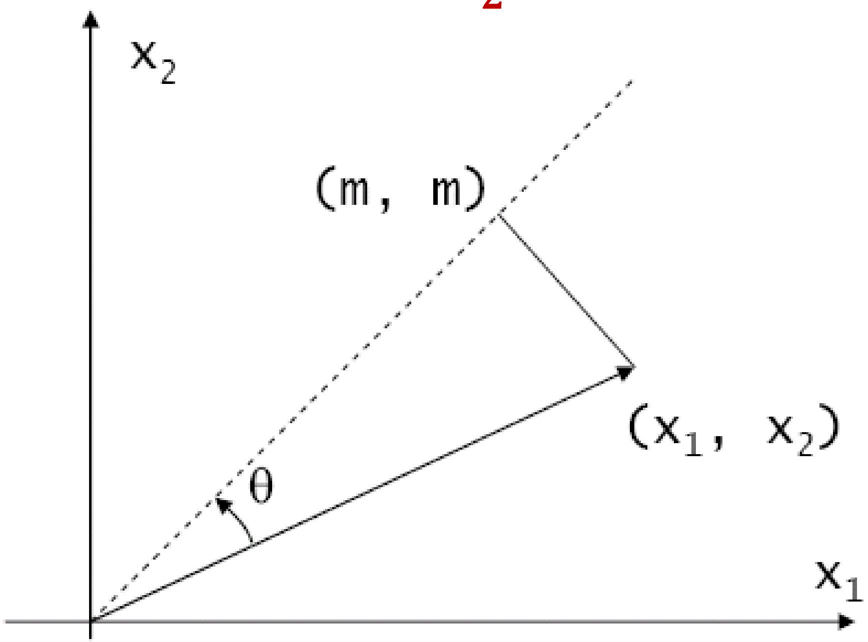
- For a data set with n samples

$$0 \leq \text{gap} \leq 1 - \frac{1}{n}$$

- Scale free, index of *unfairness*

Jain's Fairness Index is an Alternative to CoV

$$m = \frac{x_1 + x_2}{2}$$



: Jain's fairness index is $\cos^2 \theta$.

For $n = 2$

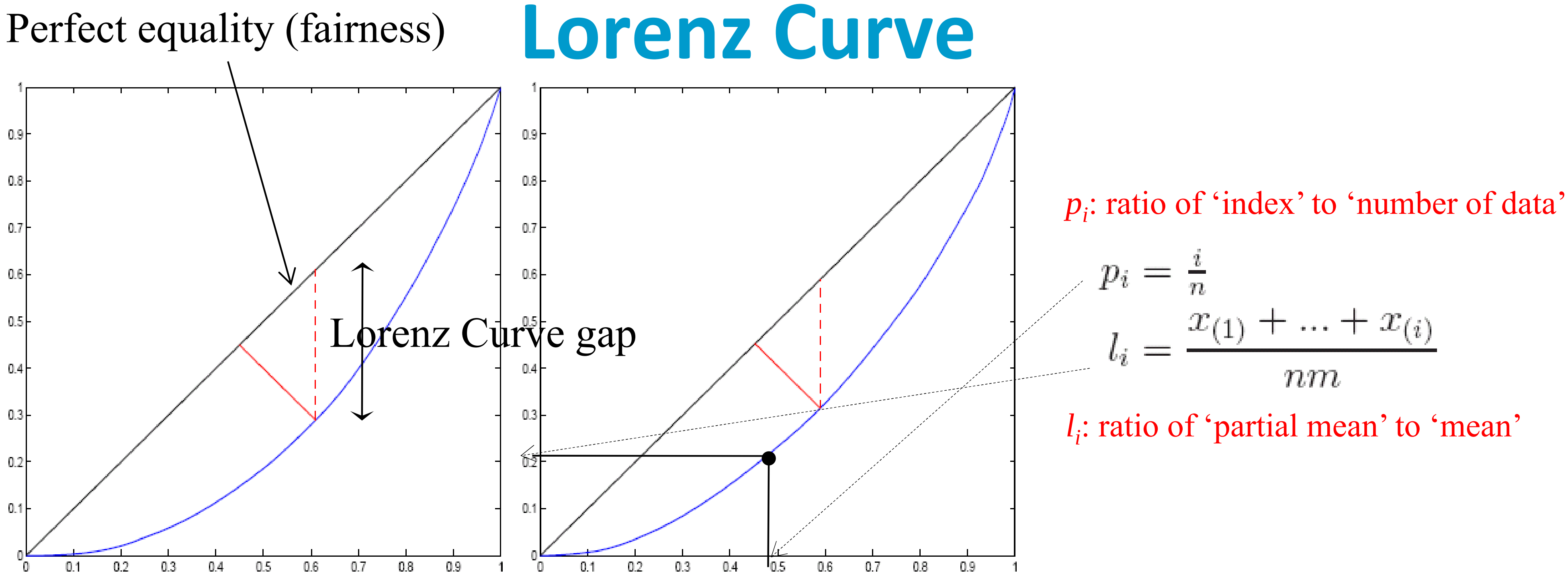
- Quantifies fairness of x;

$$\text{JFI} = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2}$$

- Ranges from
  - 1: all x_i equal
  - 1/n: maximum unfairness
- Fairness and variability are two sides of the same coin

$$\text{JFI} = \frac{1}{1 + \text{CoV}^2}$$

Lorenz Curve

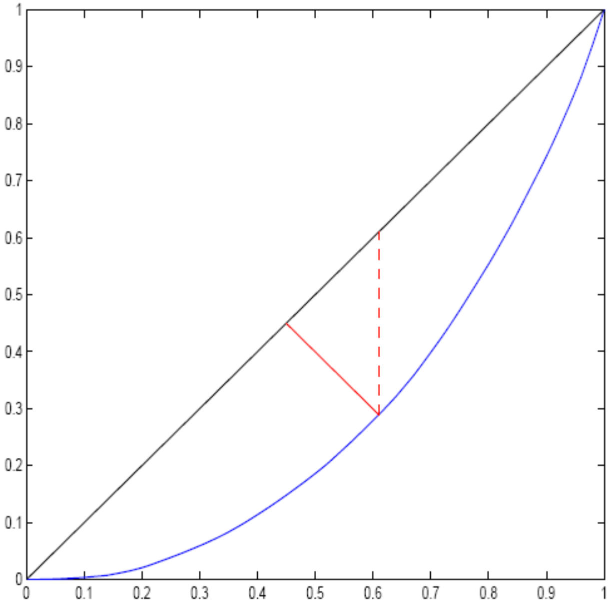


(a) Execution times in Figure 3.2, old code. CV=0.779; JFI=0.622; gap=0.321; gini=0.434; gini-approx=0.430

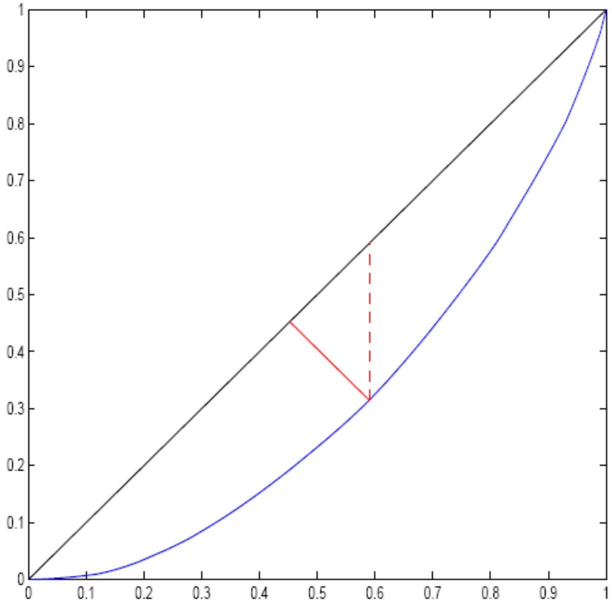
(b) Execution times in Figure 3.2, new code. CV=0.720; JFI=0.658; gap=0.275; gini=0.386; gini-approx=0.375

LORENZ CURVE The *Lorenz Curve* is defined as follows. A point (p, ℓ) on the curve, with $p, \ell \in [0, 1]$, means that the bottom fraction p of the distribution contributes to a fraction ℓ of the total $\sum_{i=1}^{n} x_i$.

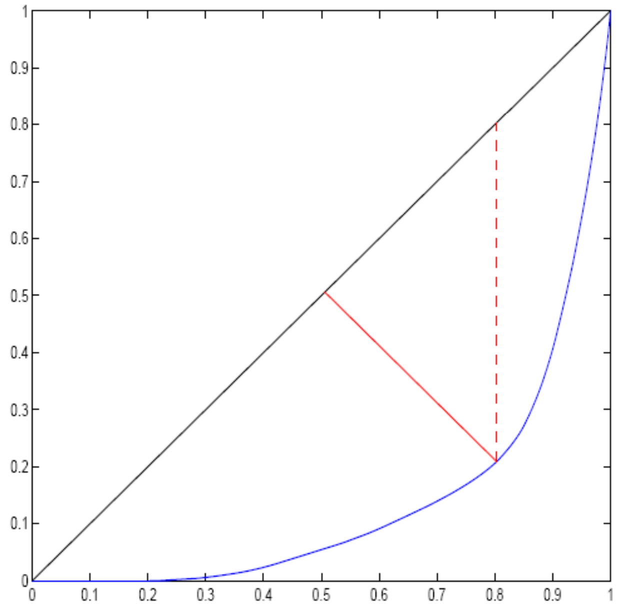
- Old code, new code: is JFI larger ? Gap ?
- Gini's index is also used; Def: 2 x area between diagonal and Lorenz curve
  - More or less equivalent to Lorenz curve gap



(a) Execution times in Figure 2.3, old code



(b) Execution times in Figure 2.3, new code



(c) Ethernet Byte Counts (x_n is the byte length of the nth packet of an Ethernet trace [55])

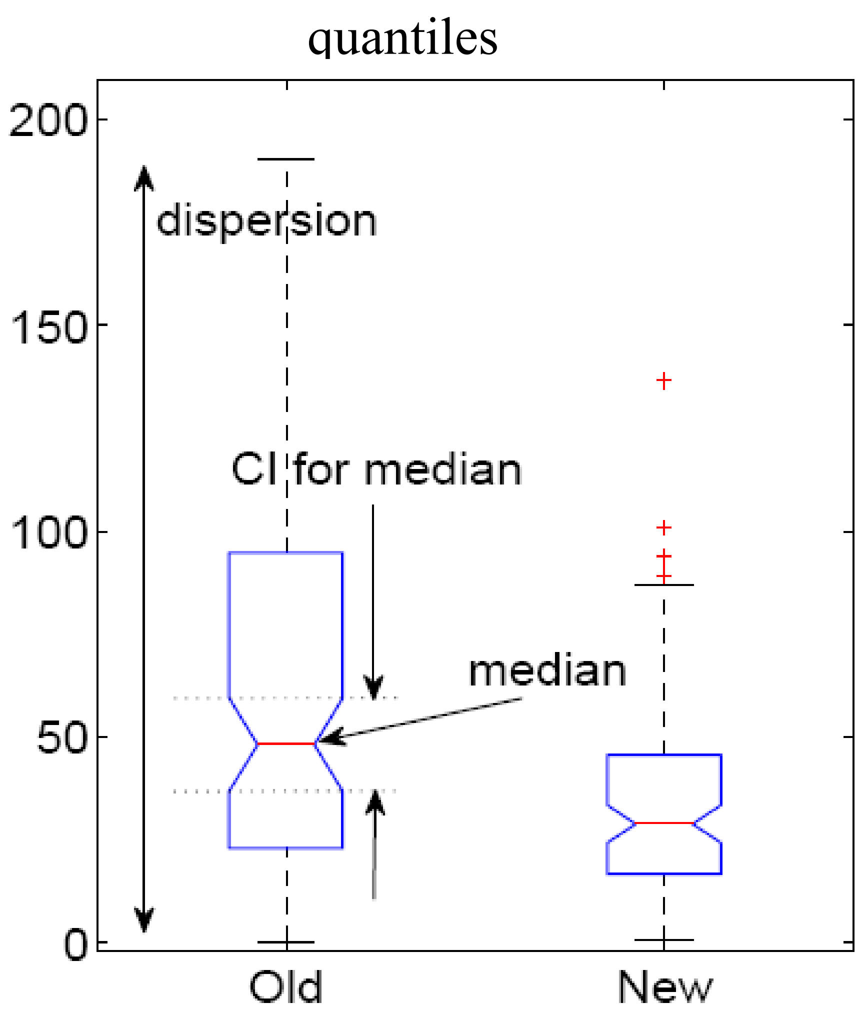
	CoV	JFI	gap	Gini	Gini-approx
Figure 2.3, old code	0.779	0.622	0.321	0.434	0.430
Figure 2.3, new code	0.720	0.658	0.275	0.386	0.375
Ethernet Byte Counts	1.84	0.228	0.594	0.730	0.715

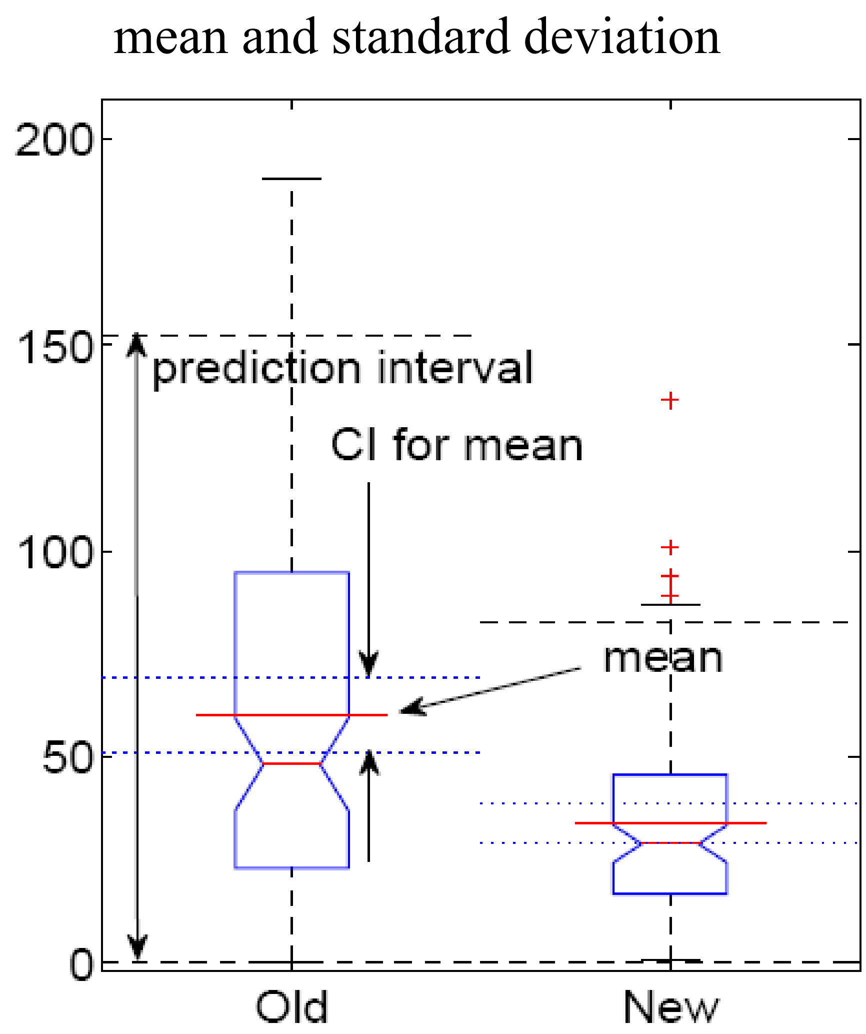
Which Summarization Should One Use ?

- There are (too) many synthetic indices to choose from
  - Traditional measures *in engineering* are standard deviation, mean and CoV
  - Traditional measures *in computer science* are mean and JFI
    - JFI is equivalent to CoV
  - In economy, gap and Gini's index (a variant of Lorenz curve gap)
  - Statisticians like medians and quantiles (robust to statistical assumptions)
- We will come back to the issue after discussing confidence intervals

2. Confidence Interval

- Do not confuse with *prediction interval*
- Quantifies *uncertainty* about an estimation


quantiles
200
150
100
50
0
dispersion
CI for median
median
Old
New


mean and standard deviation
200
150
100
50
0
prediction interval
CI for mean
mean
Old
New

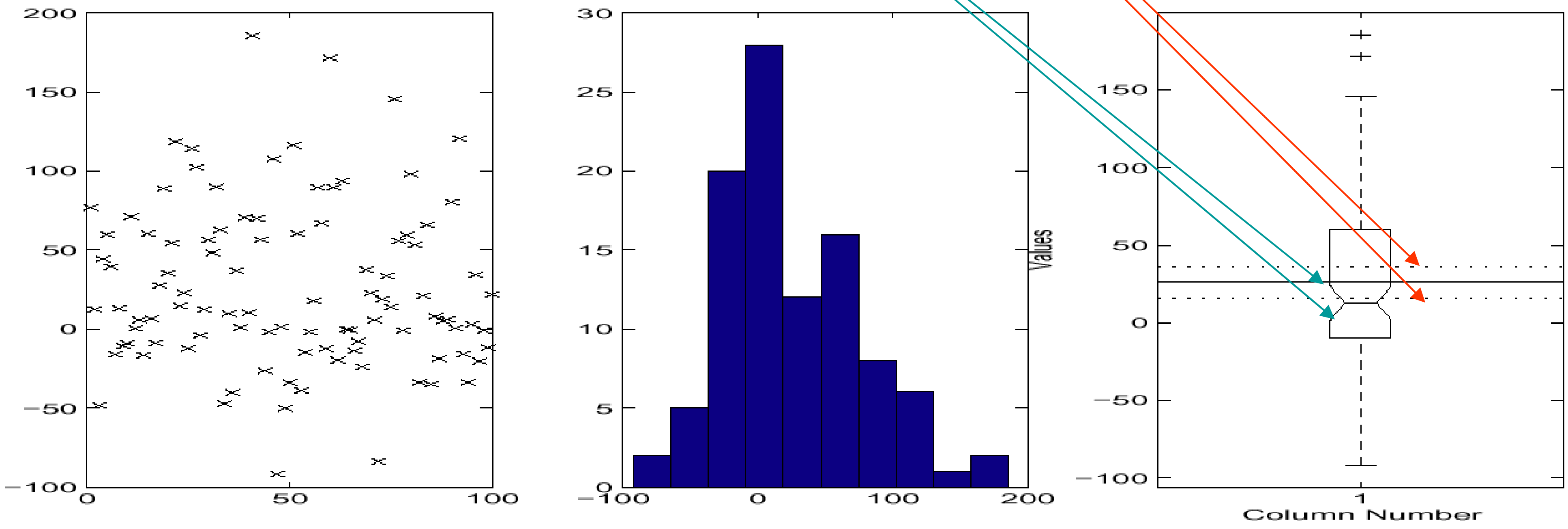
Confidence Intervals for Mean of Difference

Datasets come from the same database transaction sequences:
Paired Experiment

- Mean reduction = 26.1 ± 10.2

  0 is outside the confidence intervals for mean and for median

- Confidence interval for median



Computing Confidence Intervals

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

  Independent Identically Distributed

CI for median

- Is the simplest of all
- Robust: always true provided iid assumption holds

DEFINITION 2.2.1. *A* **confidence interval** *at level* γ *for the fixed but unknown parameter* m *is an interval* $(u(X_1, ..., X_n), v(X_1, .., X_n))$ *such that*

$$\mathbb{P}\left(u(X_1, ..., X_n) < m < v(X_1, ..., X_n)\right) \geq \gamma \tag{2.2}$$

In other words, the interval is constructed from the data, such that with at least 95% probability (for $\gamma = 0.95$) the true value of m falls in it. Note that **it is the confidence interval that is random, not the unknown parameter** m.

While *u*() and *v*() are random, the true value of *m* is deterministic.

m_p is a threshold (and one of the data) which divides the data into bottom p*100% and the rest.

THEOREM 2.2.1 (Confidence Interval for Median and Other Quantiles). *Let $X_1, \ldots, X_n$ 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. $F(m_p) = p$.* **p=0.50 for median.**
Let $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ 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 Section A for practical values. For large n, we can use the approximation

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

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

∴ **true** p-quantile m_p → Pr (one sampled datum < m_p) = p.

∴ $m_p \in [X_{(j)}, X_{(k)}]$ implies!
$X_{(j)} \leq m_p$ → At **least** j samples satisfy $X_i < m_p$
$m_p \leq X_{(k)}$ → At **most** k-1 samples satisfy $X_i < m_p$

The probability of intersection of the above event sets must be ≥γ

The above derivation is a bit tricky but can be understood easily by noting that the two event sets must be maximized. Note also the power of "order statistic": $F()$ is not used at all.

Binomial distribution $B_{n,p}()$ is the distribution of the sum of n **Bernoulli** trials with probability p.

Confidence Interval for Median, level 95%

- n = 31

j	k	$\mathbb{P}\left(X_{(j)} < m_{0.5} < X_{(k)}\right)$
9	21	0.959
10	22	0.971
11	23	0.959

- n = 32

j	k	$\mathbb{P}\left(X_{(j)} < m_{0.5} < X_{(k)}\right)$
10	22	0.965
11	23	0.965

n	j	k	γ
$n \le 5$: 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
$n \ge 71$	$\approx \lfloor 0.50n - 0.980\sqrt{n} \rfloor$	$\approx \lceil 0.50n + 1 + 0.980\sqrt{n} \rceil$	0.950

j and k are chosen s.t. $\gamma \ge 0.95$.

Example n = 100, confidence interval for median

69	26	44	0.971
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

71	25	47	0.991
72	25	47	0.990
$n \geq 73$	$\approx \lfloor 0.50n - 1.288\sqrt{n} \rfloor$	$\approx \lceil 0.50n+1+ 1.288\sqrt{n} \rceil$	0.990

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

- The median estimate is $\frac{X_{(50)}+X_{(51)}}{2}$
- Confidence level 95%

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

  a confidence interval for the median is

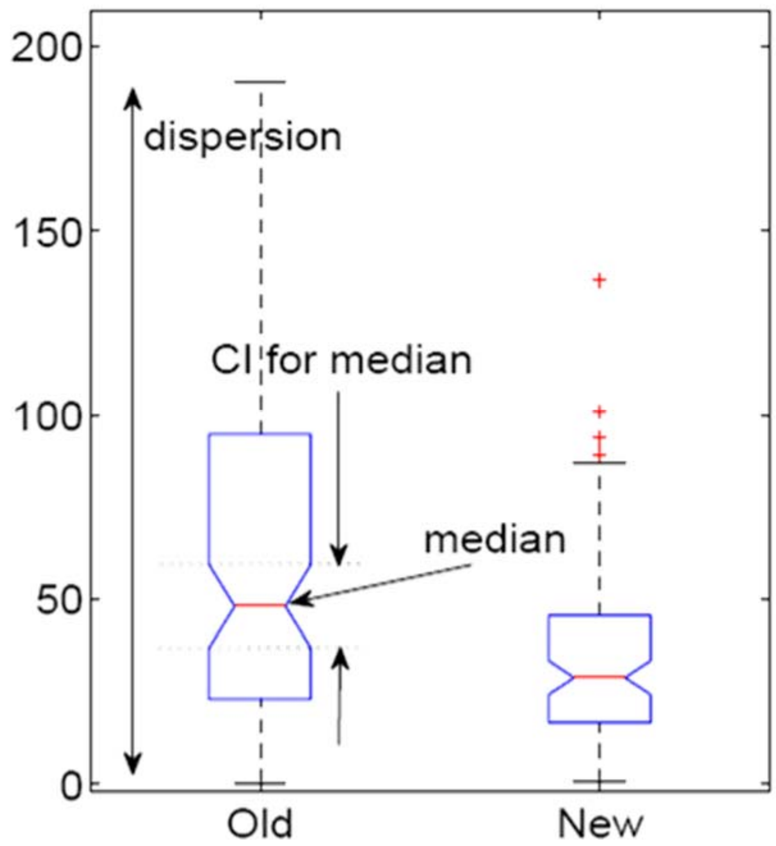
  $$[X_{(40)}; X_{(61)}]$$

- Confidence level 99%

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

  a confidence interval for the media is

  $$[X_{(37)}; X_{(64)}]$$



CI for mean and Standard Deviation

- This is another method, most commonly used method…
- But requires some ***additional*** assumptions to hold, may be misleading if they do not hold

CI for mean, asymptotic case

- If central limit theorem holds
  (in practice: *n* is large and distribution is not "wild")

finite variance
finite mean

THEOREM 2.2.2. *Let $X_1, ..., X_n$ 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 $\gamma = 0.99$.

∵ a normal distribution is symmetric.

Example

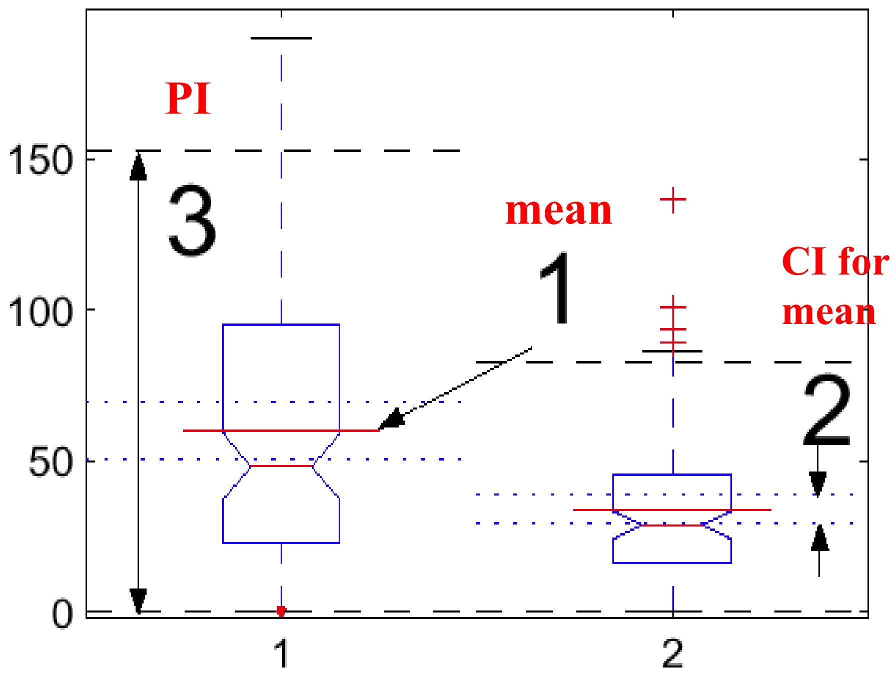
- n =100 ; 95% confidence level

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

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

  compare to prediction interval

Box Plot Representation



Normal Case

- Assume data comes from an iid + *normal* distribution
  Useful for very small data samples (n <30)

THEOREM 2.2.3. *Let $X_1, ..., X_n$ be a sequence of iid random variables with common distribution N_{μ,σ^2}*

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

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

CI for mean at level γ $\quad \hat{\mu}_n \pm \eta \frac{\hat{\sigma}_n}{\sqrt{n}}$

No More An Approximation for Normal Case

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

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

CI for std at level γ $\quad [\hat{\sigma}_n\sqrt{\frac{\zeta}{n-1}}, \hat{\sigma}_n\sqrt{\frac{\xi}{n-1}}]$

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

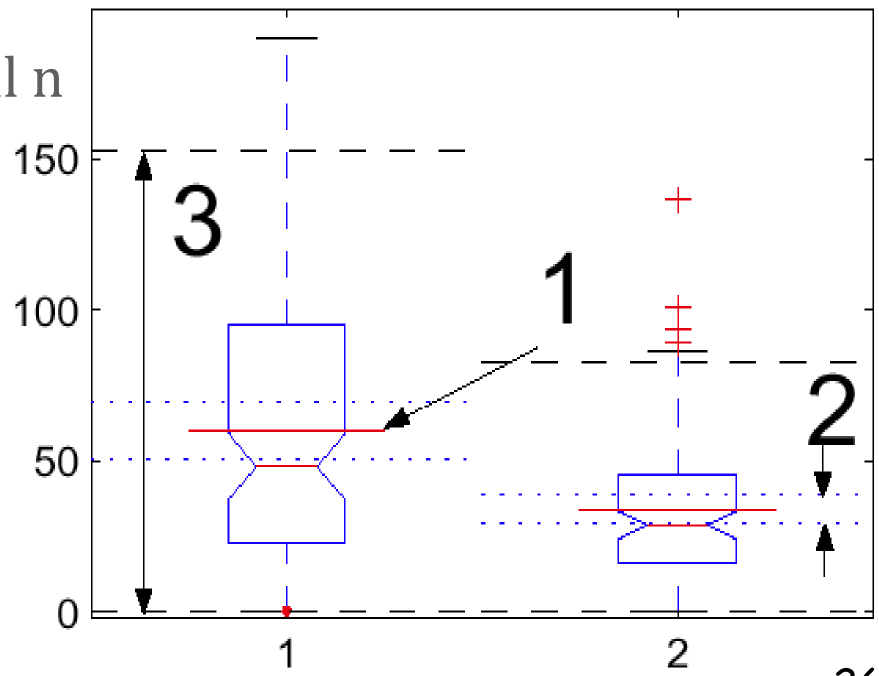
- n =100 ; 95% confidence level

  CI for mean: $[\hat{\mu} - 0.198\hat{\sigma}, \hat{\mu} + 0.198\hat{\sigma}]$

  CI for standard deviation: $[0.86\hat{\sigma}, 1.14\hat{\sigma}]$

- same as before except
  - $\hat{\sigma}_n$ instead of s_n
  - 1.98 for n=100 instead of 1.96 for all n

- In practice both (normal case and large n asymptotic) are the same if n > 30
- But large n asymptotic does not require normal assumption



Tables in [Weber-Tables]

% points of 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 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

Standard Deviation: n or n-1 ?

The estimators of the variance $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2$ and $s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2$ differ by the factor $\frac{1}{n}$ versus $\frac{1}{n-1}$. The factor $\frac{1}{n-1}$ may seem unnatural, but it is required for Theorem 2.2.3 to hold exactly. The factor $\frac{1}{n}$ appears naturally from the theory of maximum likelihood estimation (Section B.1). In practice, it is not required to have an extreme accuracy for the estimator of σ^2 (since it is a second order parameter); thus using $\frac{1}{n-1}$ or $\frac{1}{n}$ makes little difference. Both $\hat{\sigma}_n$ and s_n are called *sample standard deviation*.

We use 1/(n-1) instead of 1/n when the data is normal.

Bootstrap Percentile Method

- A heuristic that is robust (requires only iid assumption)
  - But be careful with heavy tail, see next
- but tends to underestimate CI
- Simple to implement with a computer
- Idea: use the empirical distribution in place of the theoretical (unknown) distribution

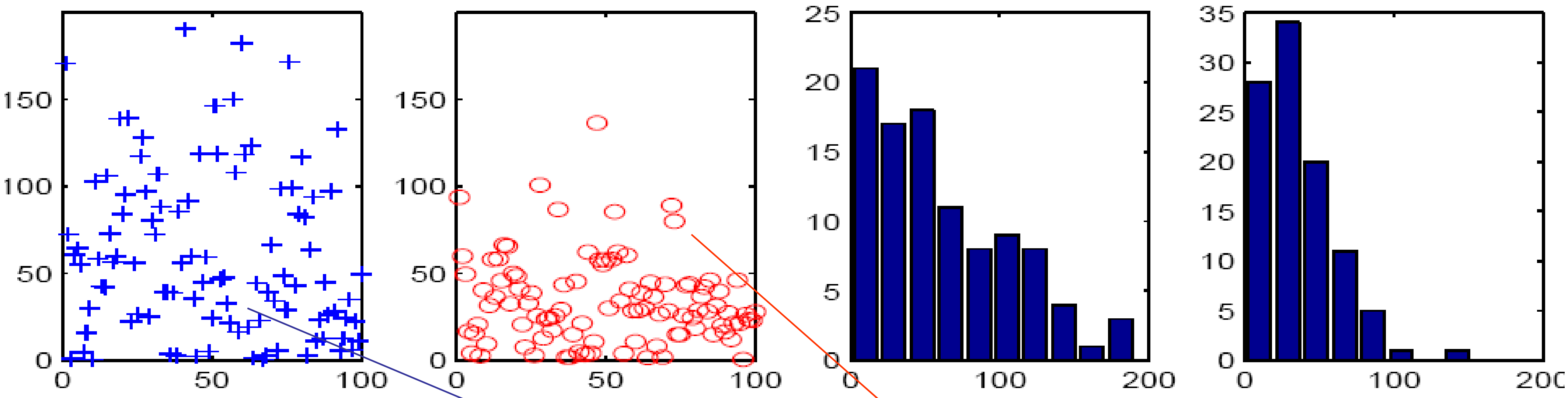
Assumption: empirical distribution can substitute for the real distribution.

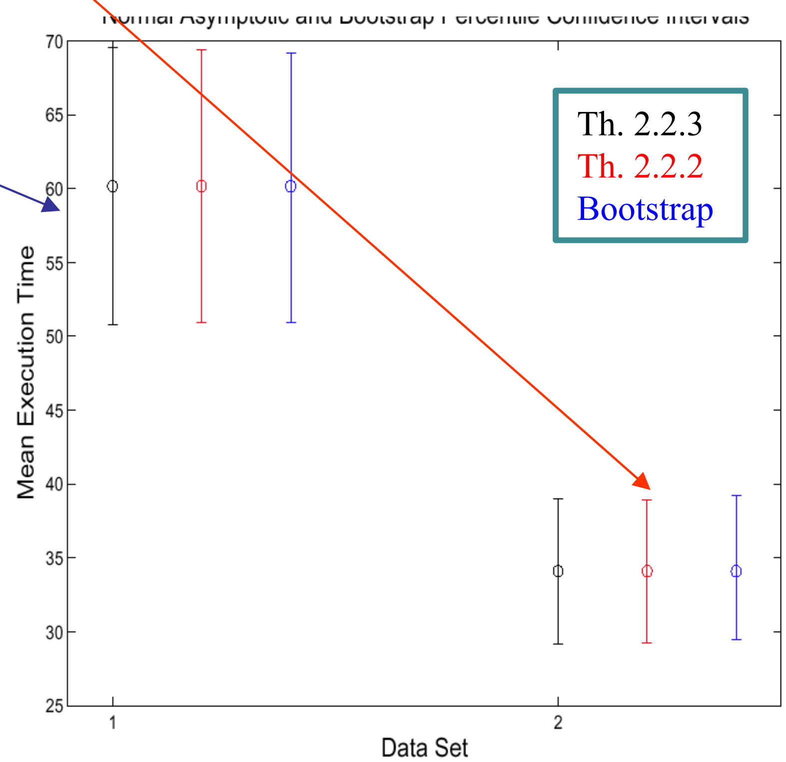
- For example, with confidence level = 95%:
  - the data set is S= $\{x_1, \ldots, x_n\}$
  - Do r=1 to r=999
    - (replay experiment) Draw *n* bootstrap replicates *with replacement* from S
    - Compute sample mean T_r

If x_1 is drawn this time, you draw next time from the entire data set $\{x_1, \ldots, x_n\}$.

  - Bootstrap percentile estimate is $(T_{(25)}, T_{(975)})$

Example: Compiler Options





- Does data look normal ?
  - No
- Theorems 2.2.2 and 2.2.3 give same result (n >30)
- Chapter 2.2.4 (Bootstrap) gives same result
  - => Asymptotic assumption valid

Confidence Interval for Fairness Index

- Use bootstrap if data is iid

interval (in this context $t(\vec{x})$ is called a *statistic*). For example, if the statistic of interest is the Lorenz curve gap, then by Section 2.1.3:

$$t(\vec{x}) = \frac{1}{2\sum_{i=1}^{n} x_i} \sum_{j=1}^{n} \left| x_j - \frac{1}{n}\sum_{i=1}^{n} x_i \right|$$

$\because$ gap $= \frac{\text{MAD}}{2m}$

```
R = ⌈2 r_0/(1 − γ)⌉ − 1                                ▷ For example r_0 = 25, γ = 0.95, R = 999
for r = 1 : R do
    draw n numbers with replacement from the list (x_1, ..., x_n) and call them X_1^r, ..., X_n^r
    let T^r = t(X⃗^r)
end for
(T_(1), ..., T_(R)) = sort (T^1, ..., T^R)
Prediction interval is [T_(r_0) ; T_(R+1−r_0)]
```

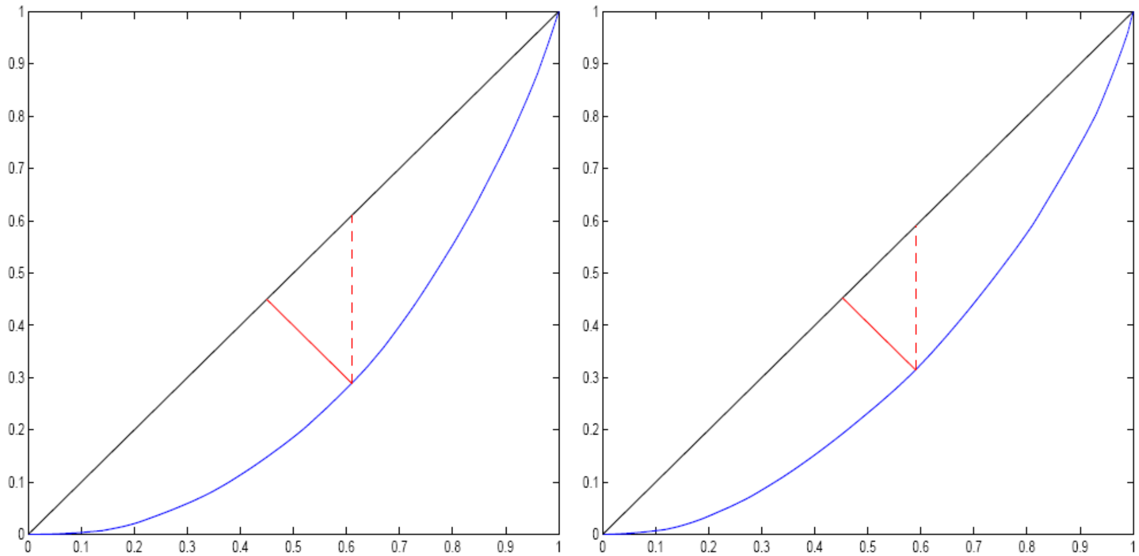
Typically 999~4999

Re-sampled data?

Prediction interval: variability of samples/data themselves

Confidence interval: variability of statistics of samples/data

To put it simply,
compute the statistic sufficient number of times *R* by "draw *n* numbers with replacement"!



(a) Execution times in Figure 2.3, old code (b) Execution times in Figure 2.3, new code

EXAMPLE 2.3: CONFIDENCE INTERVALS FOR FAIRNESS INDICES. The confidence intervals for the left two cases on Figure 2.5 were obtained with the Bootstrap, with a confidence level of 0.99, i.e. with $R = 4999$ bootstrap replicates (left and right: confidence interval; center: value of index computed in Figure 2.5).

	Jain's Fairness Index			Lorenz Curve Gap		
Old Code	0.5385	0.6223	0.7057	0.2631	0.3209	0.3809
New Code	0.5673	0.6584	0.7530	0.2222	0.2754	0.3311

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)

Theorem 2.2.2: Anyway (whether X_i is discrete r.v. or not), the normalized mean converges to a normal distribution.

$$\mu_n = 0.02$$

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

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

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

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.

1. [0 ; 0]
2. [0 ; 0.1]
3. [0 ; 0.11]
4. [0 ; 0.21]
5. [0; 0.31]

Confidence Interval for Success Probability

- Problem statement: want to estimate proba of failure; observe n outcomes; no failure; confidence interval ? ➔ Theorem 2.2.2 says [0,0]
- Example: we test a system 10 time for failures and find 0 failure: give a 95% confidence interval for the failure probability p.
- Is this a confidence interval for the mean ? (explain why)
- The general theory does not give good results when mean is very small

If *n* is extremely large, you will still be able to apply the general theory, Theorem 2.2.2.

Exploiting the fact that the data X_i is the outcome of a Bernoulli experiment, we have the theorem in the next page.

Just as "normality" was exploited for extension of Theorem 2.2.2 to Theorem 2.2.3, "Bernoullian" is used from Theorem 2.2.2 to Theorem 2.2.4.

THEOREM 2.2.4. *[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*

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

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

no success

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

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 $N_{0,1}(\eta) = \frac{1+\gamma}{2}$.

For $\gamma = 0.95$, Eq.(2.28) gives $p_0(n) \approx \frac{3.689}{n}$ and this is accurate with less than 10% relative error for $n \geq 20$ already.

If we manage to test a system more than 20 times with *no success* at all, the following simple formula can be used:

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

EXAMPLE: SENSOR LOSS RATIO. We measure environmental data with a sensor network. There is reliable error detection, i.e. there is a coding system which declares whether a measurement is correct or not. In a calibration experiment with 10 independent replications, the system declares that all measurements are correct. What can we say about the probability p of finding an incorrect measurement ?

Apply Eq.(2.28): we can say, with 95% confidence, that $p \leq 30.8\%$.

Theorem 2.2.4.

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 $N_{0,1}(\eta) = \frac{1+\gamma}{2}$.

Theorem 2.2.4.

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

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

Take Home Message

- Confidence interval for median (or other quantiles) is easy to get from the Binomial distribution
  - Requires iid
  - No other assumption
- Confidence interval for the mean
  - Requires iid
  - And
    - Either if data sample is **normal** and n is small
    - Or data sample is **not wild** and ***n* is large enough**

- The bootstrap is more robust and more **general** but is more than a simple formula to apply (NB: Even bootstrap highly depends on no. of sample data)
- Confidence interval for success probability requires special attention when success or failure is **rare**
- If the data is **not normal** and the **size of data is very small**, use “median” approach rather than risking accuracy of confidence interval of “mean” approach.

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 samples randomly with replacement
  - Simulation: independent replications (with random seeds)
  - Else: we do not know – in some cases we will have methods for time series

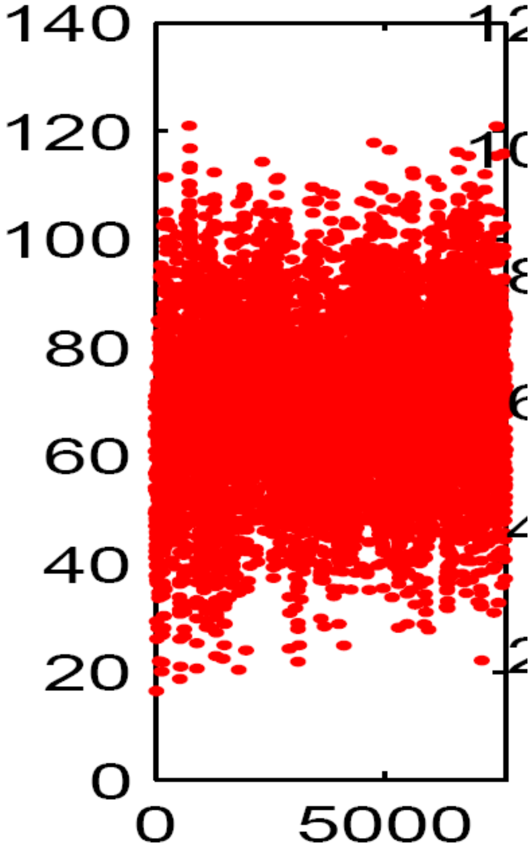
What does independence mean ?

$$\mathbb{P}(X_i \in A \mid X_1 = x_1, ..., X_{i-1} = x_{i-1}) = \mathbb{P}(X_i \in A) \qquad (2.30)$$

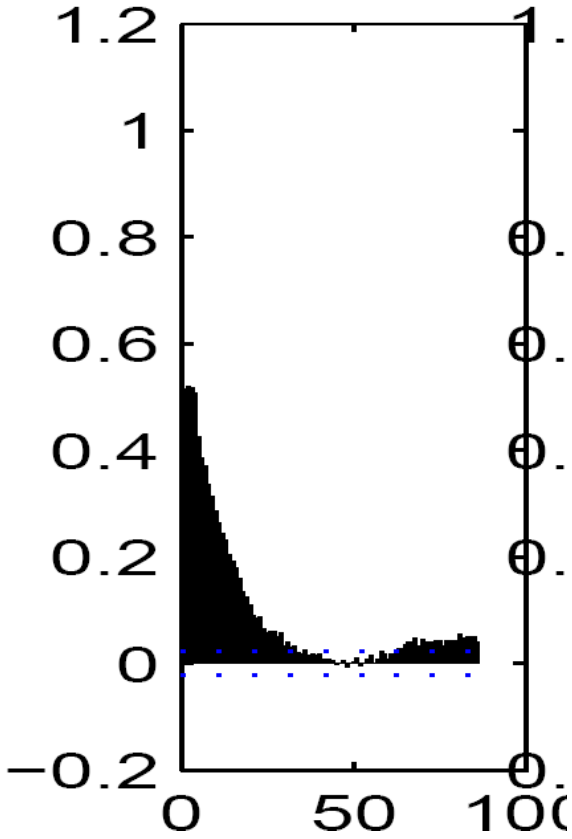
i.e. **if we know the distribution** $F(x)$, observing $X_1, ..., X_{i-1}$ does not give more information about X_i.

Note the importance of the "if" statement in the last sentence: remove it and the sentence is no longer true. To understand why, consider a sample $x_1, ..., x_n$ for which we assume to know that it is generated from a sequence of iid random variables $X_1, ..., X_n$ with normal distribution but with unknown parameter (μ, σ^2). If we observe for example that the average of $x_1, ..., x_{n-1}$ is 100 and all values are between 0 and 200, then we can think that it is very likely that x_n is also in the interval $[0, 200]$ and that it is unlikely that x_n exceeds 1000. Though the sequence is iid, we did gain information about the next element of the sequence having observed the past. There is no contradiction: if we know that the parameters of the random generator are $\mu = 100$ and $\sigma^2 = 10$ then observing $x_1, ..., x_{n-1}$ gives us no information about x_n.

Example



data



ACF

- Pretend data is iid: CI for mean is [69; 69.8]
- Is this biased ?

What happens if data is not iid ?

- If data is positively correlated
  - Neighbouring values look similar
  - Frequent in measurements (particularly if data are sampled over fine time scale)
  - CI is underestimated: there is less information in the (non-iid) data than one thinks

You must be less confident.

4. Prediction Interval

- CI for mean or median summarize
  - Central value (a scalar function of data)+ uncertainty about it
- Prediction interval summarizes variability of data

DEFINITION 2.4.1. *Let* $X_1, ..., X_n, X_{n+1}$ *be a sequence of random variables. A prediction interval at level* γ *is an interval of the form* $[u(X_1, ..., X_n), v(X_1, ..., X_n)]$ *such that*

$$\mathbb{P}\left(u(X_1, ..., X_n) \leq X_{n+1} \leq v(X_1, ..., X_n)\right) \geq \gamma \tag{2.31}$$

↑ Instead of central values, i.e., mean & median

Prediction Interval based on Order Statistic

- Assume data comes from an iid model
- Simplest and most robust result (**not well known**, though):

THEOREM 2.4.1 (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 $1 \le j \le k \le n$:*

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

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

For example, with $n = 999$, a prediction interval at level 0.95 ($\alpha = 0.05$) is $[X_{(25)}, X_{(975)}]$. This theorem is similar to the bootstrap result in Section 2.2.4, but is exact and much simpler.

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

Normal case

THEOREM 2.4.2 (Normal IID Case). *Let $X_1, \ldots, X_n, X_{n+1}$ be an iid sequence with common distribution N_{μ,σ^2}. Let $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ be as in Theorem 2.2.3. 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 based on Mean

- If data is **not normal**, there is no general result – bootstrap can be used
  - Self-evident because, in two-number (mean, std) summarization, the variability depends on std and the distribution type as well.

- If data is assumed normal, how do CI for mean and Prediction Interval based on mean compare ?

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% $= \mu \pm 1.96\, s$

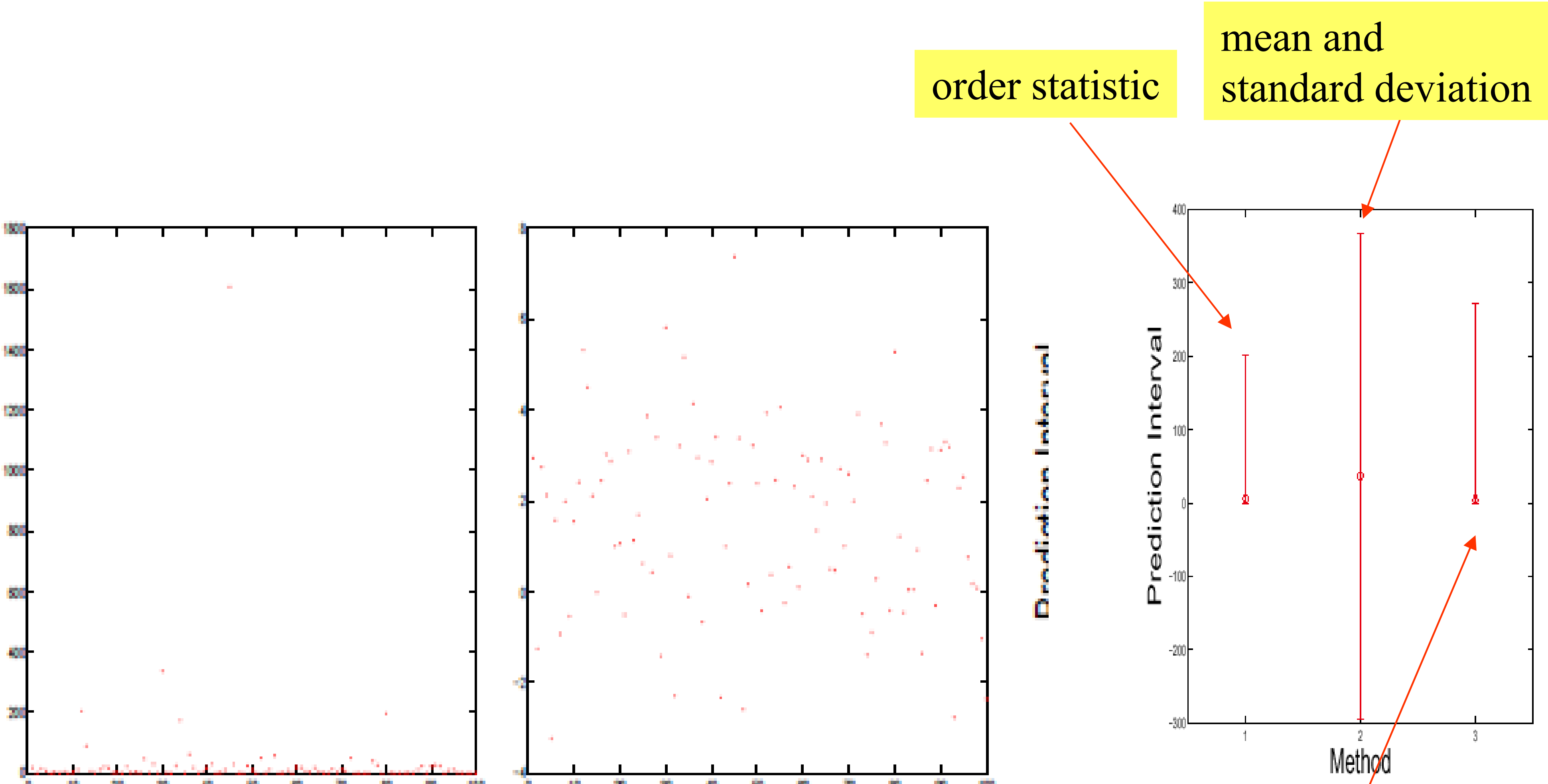
Re-Scaling

- 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. Put it another way, is it *normalizable* through re-scaling?
  - Ex: log of the data instead of the data
- A generic transformation used in statistics is the *Box-Cox* transformation:

$$b_s(x) = \begin{cases} \frac{x^s - 1}{s} & , s \neq 0 \\ \ln x & , s = 0 \end{cases}$$

  - Continuous in s
    s=0 : log
    s=-1: 1/x
    s=1: identity

Prediction Intervals for File Transfer Times



(a) (Data)

(b) (Log of data)

(c) (Prediction Intervals)

↑ **Not normal apparently**

mean and
standard deviation
on rescaled (log) data

Which Summarization Should I Use ?

- Two issues
  - Robustness to outliers (i.e., significantly bigger/smaller values)
    - e.g., what if data is not normal?
    - e.g., what if some data are extremely large?
  - Compactness (i.e., how do you want to summarize them in your paper?)

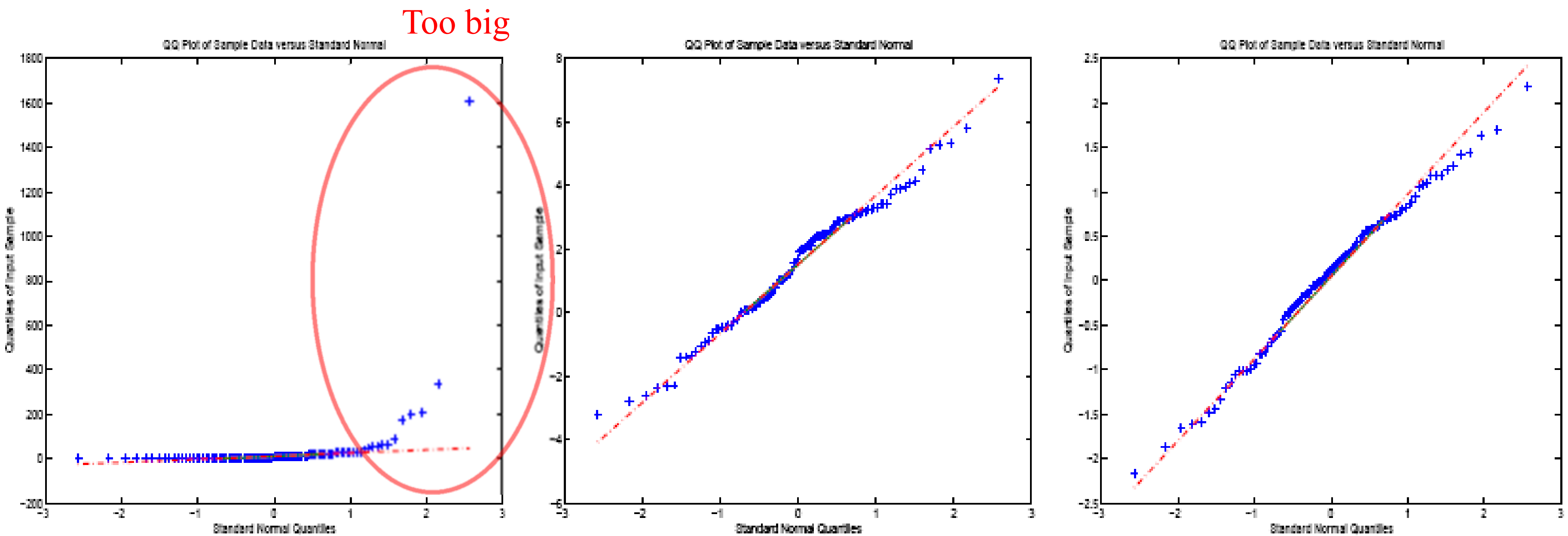
QQplot is common tool for verifying assumption

- **Normal** Qqplot
  - X-axis: standard **normal** quantiles $$x_i := F^{-1}\left(\frac{i}{n+1}\right)$$
    ↑ Inverse of normal CDF
  - Y-axis: Ordered statistic of sample: $$X_{(1)} \leq X_{(2)} \leq \ldots.$$

- 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

QQplot : Quantile-Quantile (Ordered Data) Comparison

QQPlots of File Transfer Times

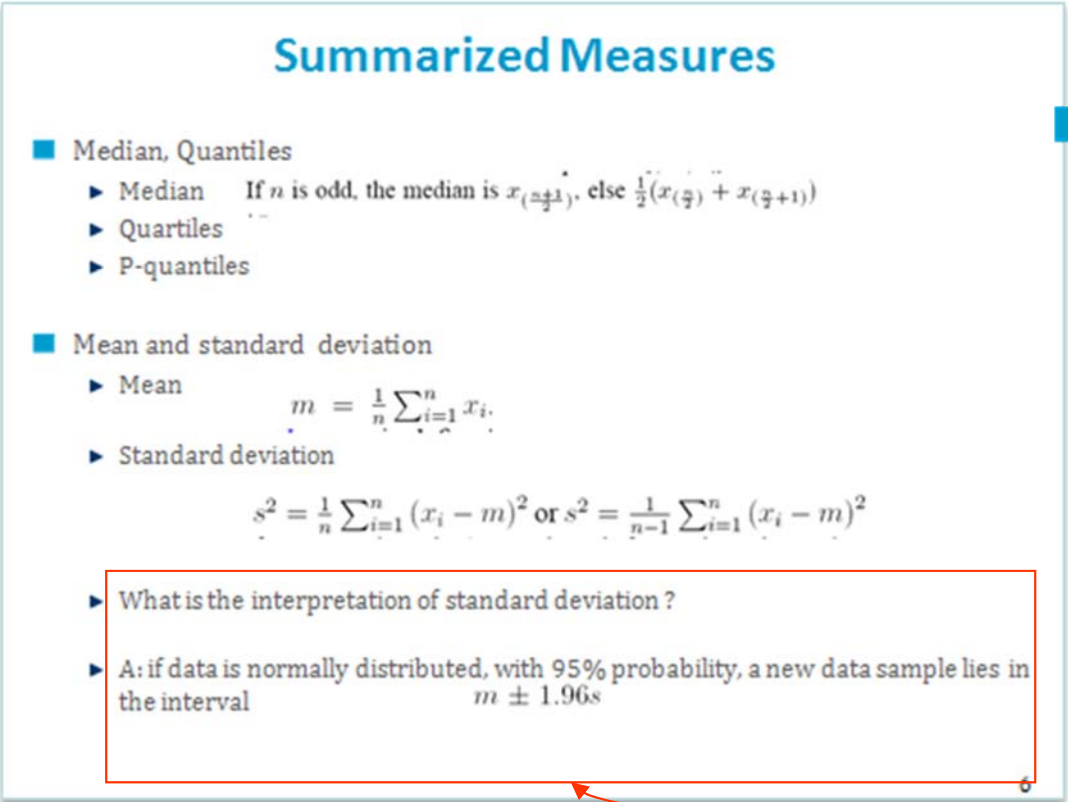


(a) (QQ-plot) (b) (QQ-plot of log of data) (c) (normal sample)

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.

Handy tool for checking normality

Take Home Message



- The interpretation of σ as measure of variability is meaningful **if the data is normal** (or close to normal). Else, it is misleading. The data should be best re-scaled.

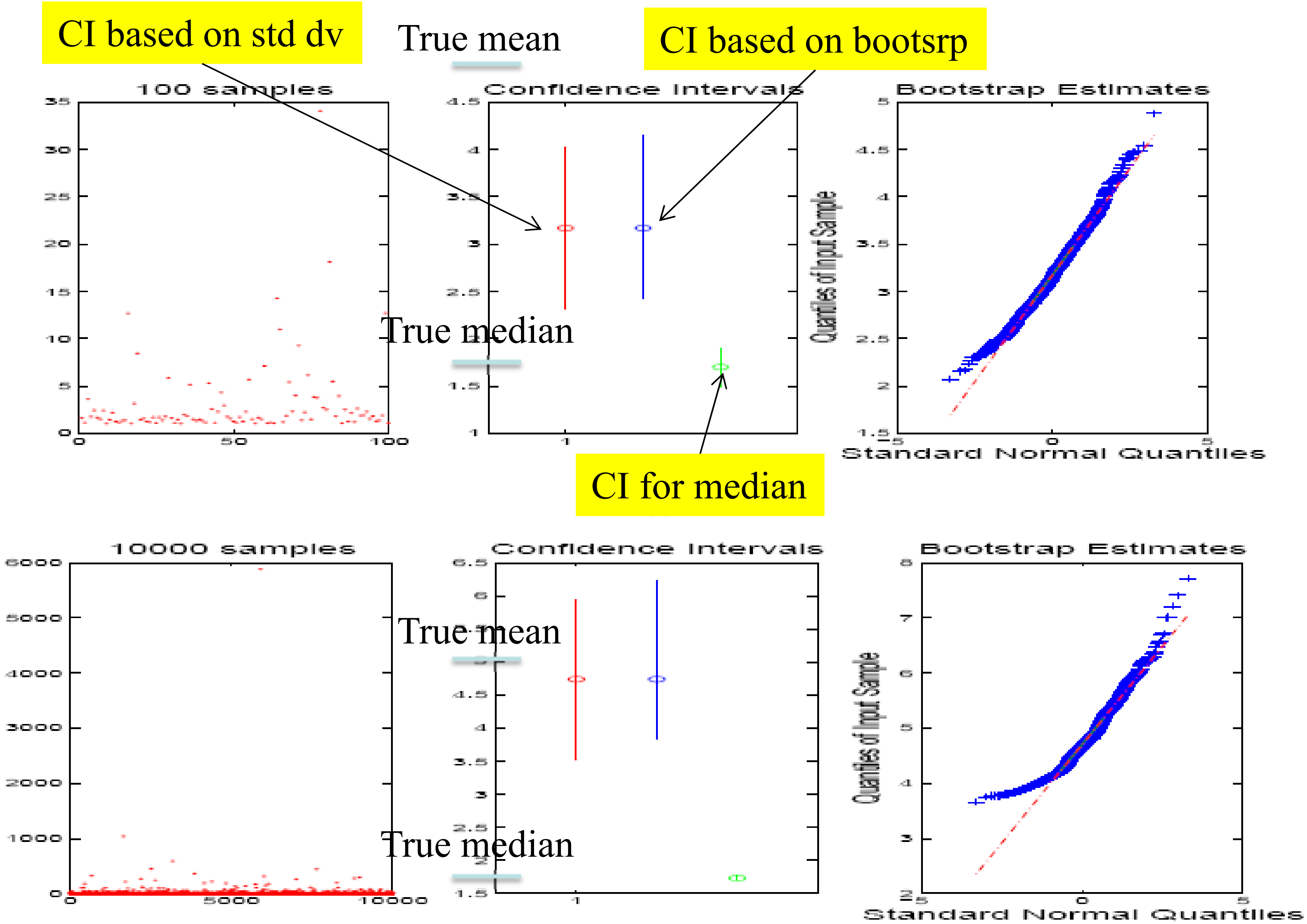
5. Which Summarization to Use ?

- Issues
  - Robustness to outliers
  - Distribution assumptions

Example of Outlier:

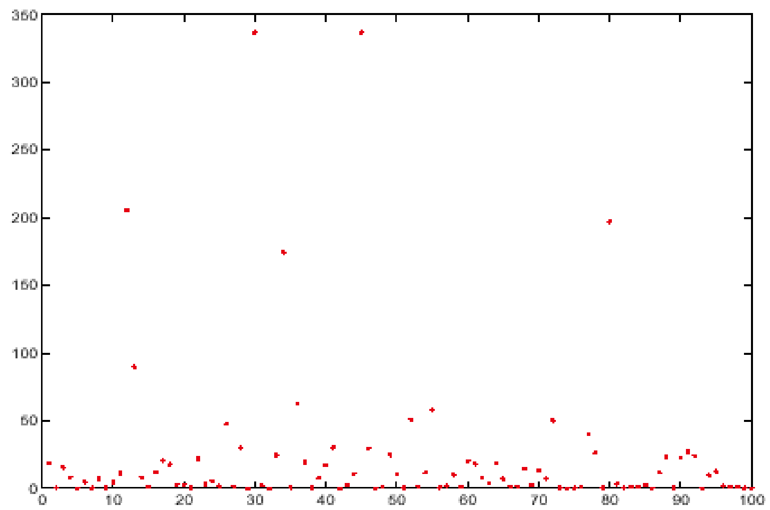
You are measuring the seismic intensity of a very weak earthquake in a lab. All of a sudden, a friend of yours slams the door of the lab and you get extremely strong seismic intensity on the seismometer.

A Distribution with Infinite Variance

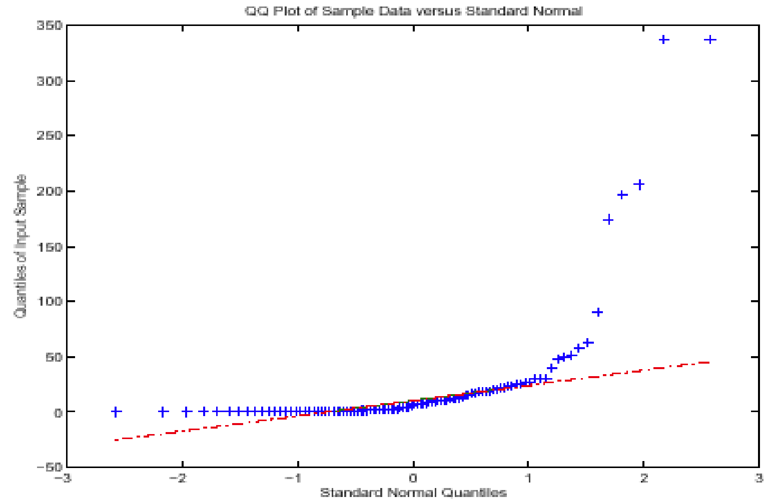


"True median" lies within the CI for median even for 100 samples.
→ "Median" is more robust for infinite variance distributions.

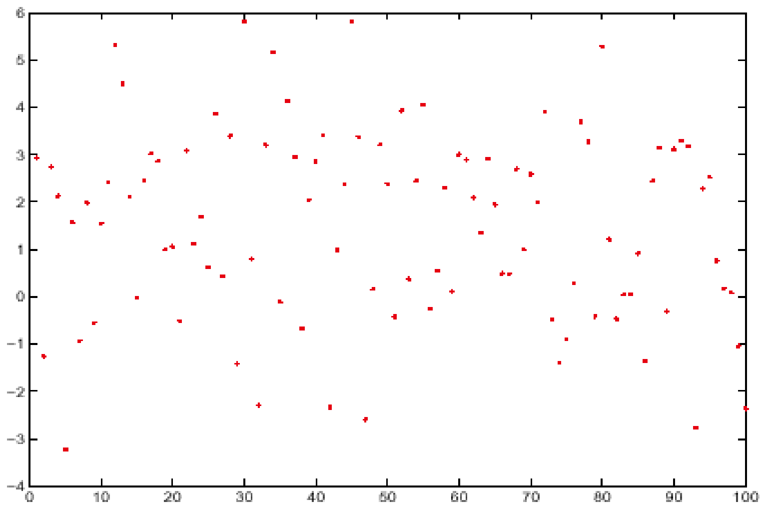
Outlier in File Transfer Time



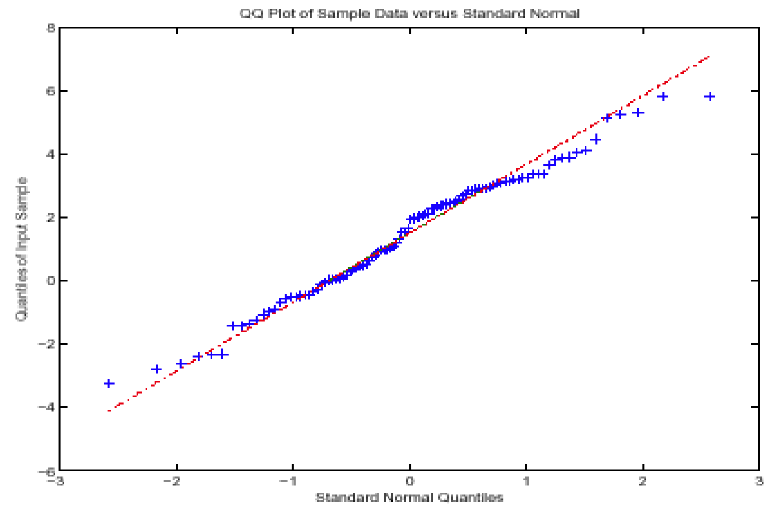
(a) (Data without outlier)



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

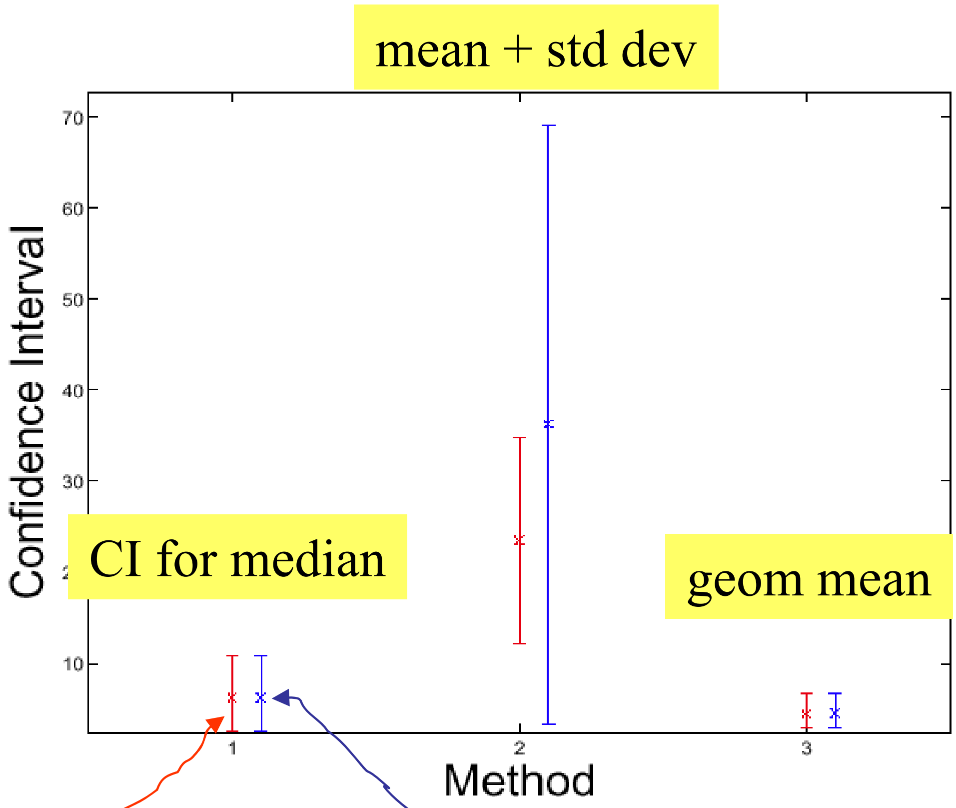


(c) (Log of data without outlier)

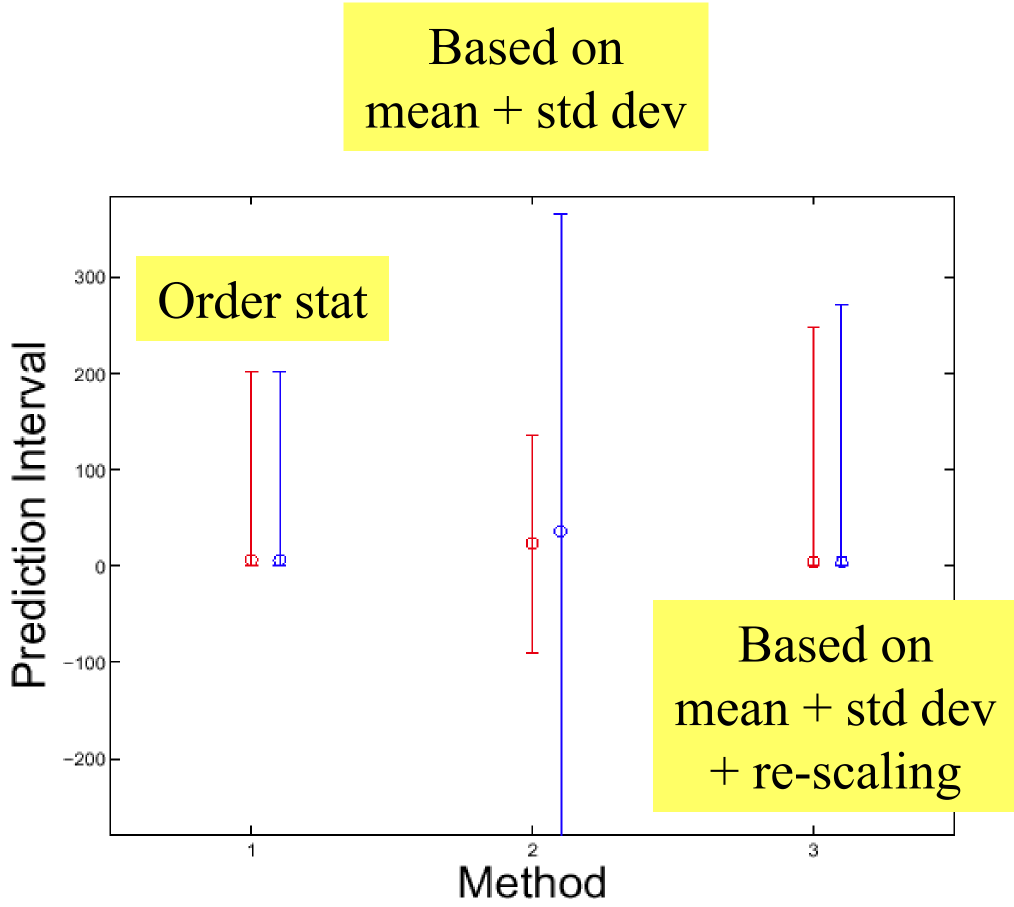


(d) (QQ-plot of (c))

Robustness of Conf/Prediction Intervals



(e) (Confidence Intervals)



(f) (Prediction Intervals)

Outlier removed
Outlier present

Normalization aside,
order statistics based methods are much more robust to outliers.

Fairness Indices with *different orders*

	Index	Lower Bound, CI	Index	Upper Bound, CI
Without Outlier	JFI	0.1012	0.1477	0.3079
	gap	0.4681	0.5930	0.6903
With Outlier	JFI	0.0293	0.0462	0.3419
	gap	0.4691	0.6858	0.8116

Table 3.2: Fairness indices with and without outlier.

- Confidence Intervals obtained by Bootstrap

- JFI is very dependent on one outlier
  - As expected, since JFI is essentially CoV, i.e. standard deviation

- Gap is sensitive, but less
  - Does not use squaring ; why ? ➔ Lower-order statistics are less sensitive

Compactness

- If **normal** assumption (or, for CI; asymptotic regime) holds, μ and σ are more compact
  - two values give both: CIs at all levels, prediction intervals
  - Derived indices: CoV, JFI

- In contrast, CIs for median does not give information on variability (PI)
  - PI has to be computed through an **additional** procedure.

- Prediction interval based on order statistic is robust (and, IMHO, best)
  - **Use order statistic for prediction intervals**

Take-Home Message

- Understand methods before using them.
- Mean and standard deviation make sense when data sets are not wild.
  - Close to normal, or not heavy tailed and large data sample
    - For example, certain Weibull distributions are close to a normal one.
- For non-norml case, use quantiles and order statistics.
- Sometimes, you need to rescale.

Questions

QUESTION 2.8.1. *Compare (1) the confidence interval for the median of a sample of n data values, at level 95% and (2) a prediction interval at level at least 95%, for $n = 9, 39, 99$.* [9]

[8]From the tables in Chapter A and Theorem 2.4.1 we obtain: (confidence interval for median, prediction interval): $n = 9$: $[x_{(2)}, x_{(9)}]$, impossible; $n = 39$: $[x_{(13)}, x_{(27)}]$, $[x_{(1)}, x_{(39)}]$; $n = 99$: $[x_{(39)}, x_{(61)}]$, $[x_{(2)}, x_{(97)}]$. The confidence interval is always smaller than the prediction interval.

QUESTION 2.8.2. *Call $L = \min\{X_1, X_2\}$ and $U = \max\{X_1, X_2\}$. We do an experiment and find $L = 7.4$, $U = 8.0$. Say which of the following statements is correct: (θ is the median of the distribution). (1) the probability of the event $\{L \leq \theta \leq U\}$ is 0.5 (2) the probability of the event $\{7.4 \leq \theta \leq 8.0\}$ is 0.5* [10]

[9]In the classical (non-Bayesian) framework, (1) is correct and (2) is wrong. There is nothing random in the event $\{7.4 \leq \theta \leq 8.0\}$, since θ is a fixed (though unknown) parameter. The probability of this event is either 0 or 1, here it happens to be 1. Be careful with the ambiguity of a statement such as "the probability that θ lies between L and U is 0.5". In case of doubt, come back to the roots: the probability of an event can be interpreted as the ideal proportion of simulations that would produce the event.

QUESTION 2.8.3. *How do we expect a 90% confidence interval to compare to a 95% one ? Check this on the tables in Section A.* [11]

[10]It should be smaller. If we take more risk we can accept a smaller interval. We can check that the values of j [resp. k] in the tables confidence intervals at level $\gamma = 0.95$ are larger [resp. smaller] than at confidence level $\gamma = 0.99$.

Questions

QUESTION 2.8.4. *A data set has 70 points. Give the formulae for confidence intervals at level 0.95 for the median and the mean* [12]

[11]Median: from the table in Section A $[x_{(27)}, x_{(44)}]$. Mean: from Theorem 2.2.2: $\hat{\mu} \pm 0.2343S$ where $\hat{\mu}$ is the sample mean and S the sample standard deviation. The latter is assuming the normal approximation holds, and should be verified by either a qqplot or the bootstrap.

QUESTION 2.8.5. *A data set has 70 points. Give formulae for a prediction intervals at level 95%*

[12]From Theorem 2.4.1: $[\min_i x_i, \max_i x_i]$.

Questions

QUESTION 2.8.6. *A data set $x_1, \dots x_n$ is such that $y_i = \ln x_i$ looks normal. We obtain a confidence interval $[\ell, u]$ for the mean of y_i. Can we obtain a confidence interval for the mean of x_i by a transformation of $[\ell, u]$?* [14]

[13]No, we know that $[e^\ell, e^u]$ is a confidence interval for the geometric mean, not the mean of x_i. In fact x_i comes from a log-normal distribution, whose mean is $e^{\mu + \frac{\sigma^2}{2}}$ where μ is the mean of the distribution of y_i, and σ^2 its variance.

QUESTION 2.8.7. *Assume a set of measurements is corrupted by an error term that is normal, but positively correlated. If we would compute a confidence interval for the mean using the iid hypothesis, would the confidence interval be too small or too large ?* [15]

[14]Too small: we underestimate the error. This phenomenon is known in physics under the term *personal equation*: if the errors are linked to the experimenter, they are positively correlated.

Confusing term: *log-normal* distribution is the distribution of an *exponential* of a normal random variable.

Read!

- To make a good start of this course, please read Chapter 2.
- If it is affordable, also read Chapter 1.

- If you have no knowledge in Markov chain, read Chapter 7.6 before the next lecture.