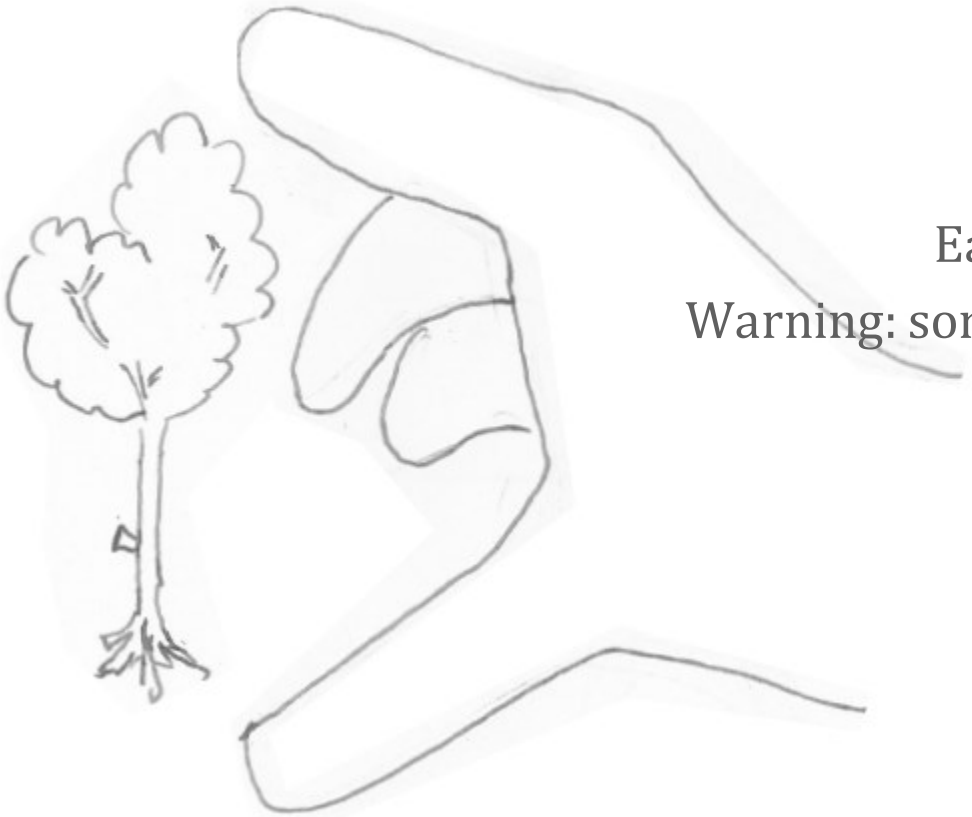
Summarizing Performance Data
Confidence Intervals



Important

Easy to Difficult

Warning: some mathematical content

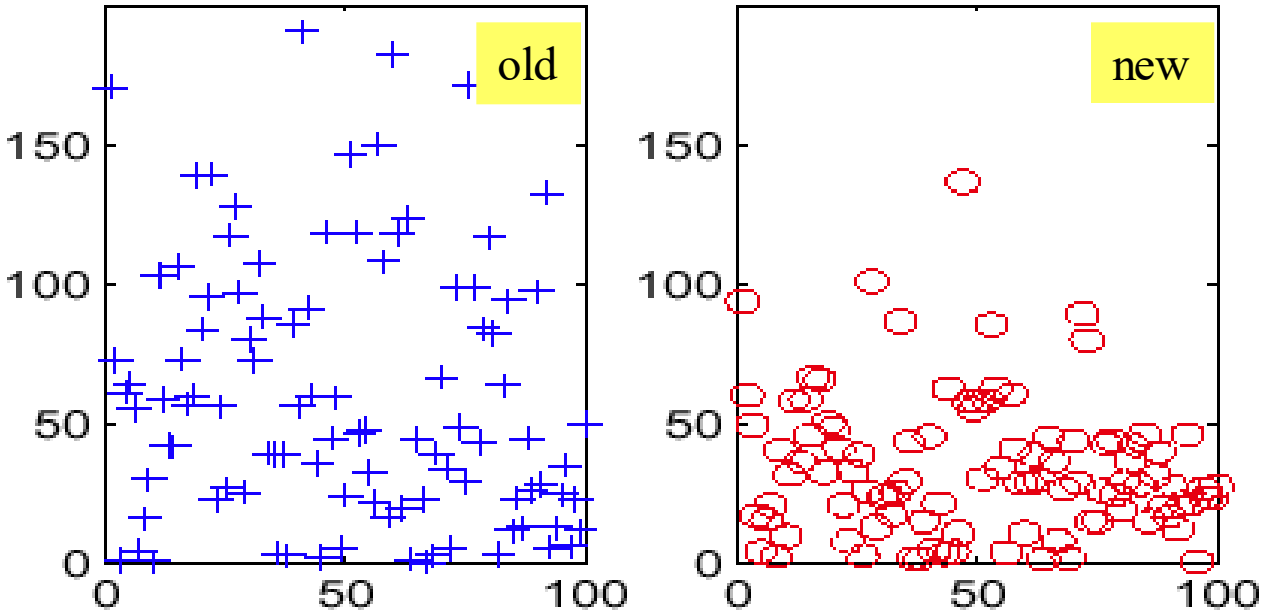
Slide set based on Jean-Yves Leboudec's teaching material, see http://perfeval.epfl.ch/
with changes and additions by Michele Zorzi

What is Performance Evaluation ?

- Characterizing quantitatively the service provided by a system (e.g., computer or communication)
  - Throughput, delay, energy consumption, memory, resources, …
- Purpose(s) of performance evaluation
  - Compare competitive solutions
  - Provide dimensioning guidelines
  - Test design in realistic conditions
  - Identify performance problems and study behaviors and trends
- Tools for PE: analysis, simulation, experimentation
- Importance of carefully defining load, metrics and goals
- Importance of understanding factors and patterns

1 Summarizing Performance Data

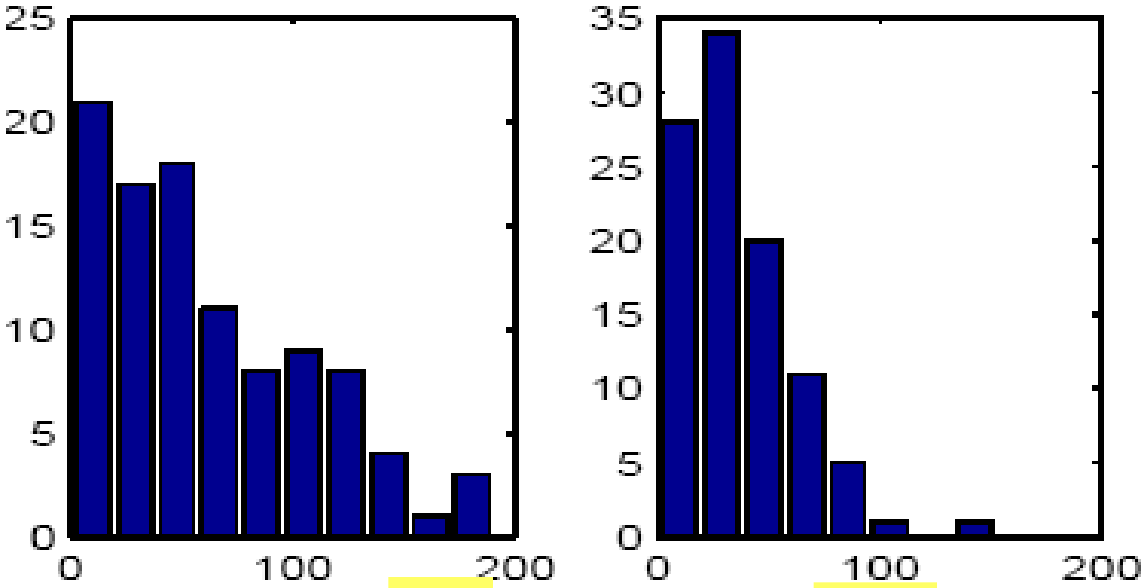
- How do you quantify:
  - Central value of the data (often the best estimate)
  - Dispersion (accuracy of the estimate)

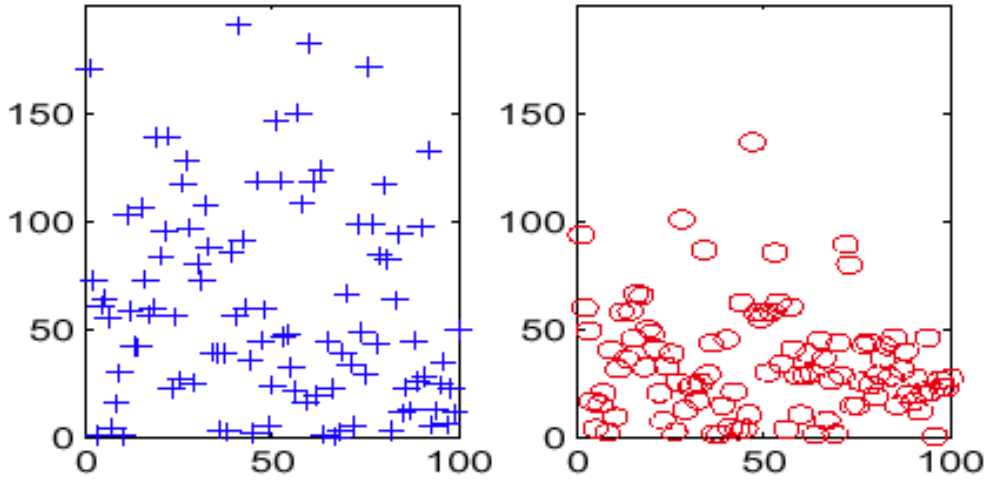


EXAMPLE 2.1: COMPARISON OF TWO OPTIONS. An operating system vendor claims that the new version of the database management code significantly improves the performance. We measured the execution times of a series of commonly used programs with both options. The data are displayed in Figure 2.1. The raw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

- Note: the load pattern is the same in the two cases (*paired experiment*)

Histogram is one answer (empirical pdf)



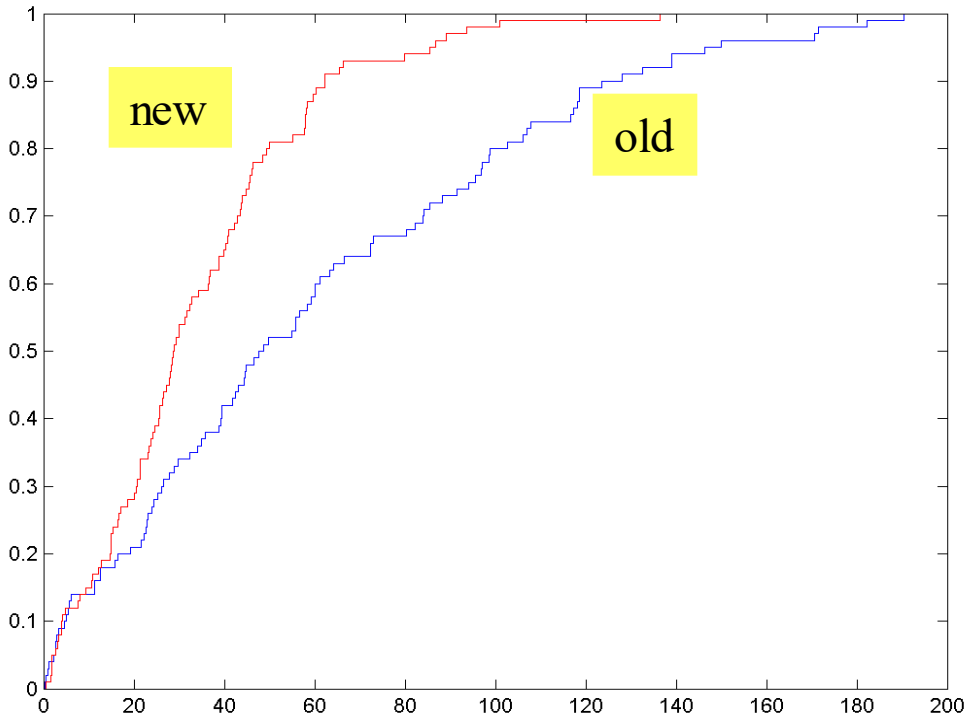


ECDFs allow easy comparison

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

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

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.

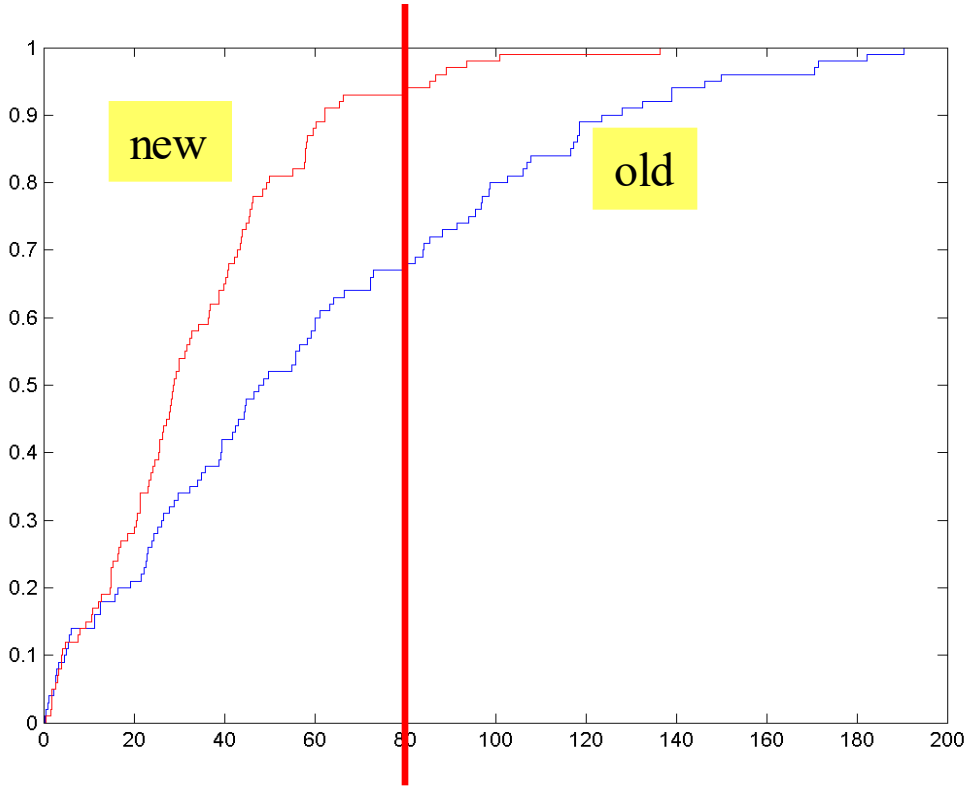


ECDFs allow easy comparison

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

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

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.

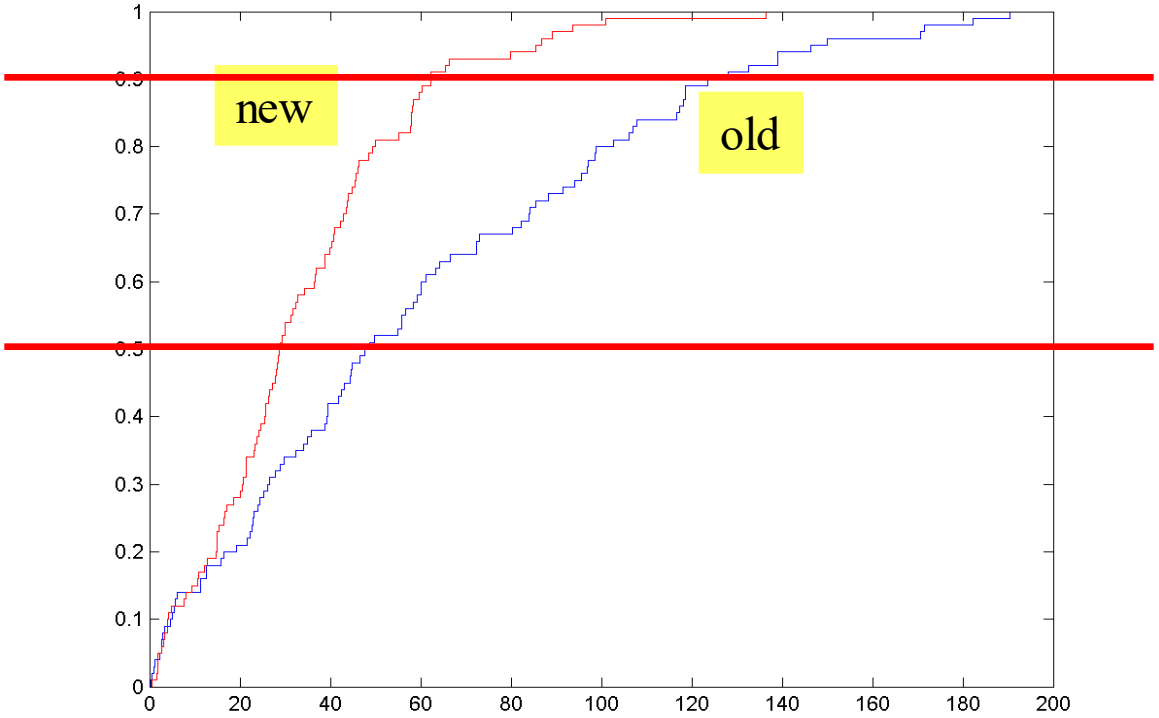


ECDFs allow easy comparison

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

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

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.



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
  - P-quantiles (see http://www.stat.wisc.edu/~wardrop/courses/chap12.pdf)

- (Sample) Mean and standard deviation (computed from data)
  - 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 $m \pm 1.96s$
  - Also Chebyshev's inequality

Theorem 8.3. Chebyshev Inequality. *For an arbitrary random variable* Y *and constant* $c > 0$,

$$P[|Y - \mu_Y| \geq c] \leq \frac{\text{Var}[Y]}{c^2}$$

Example

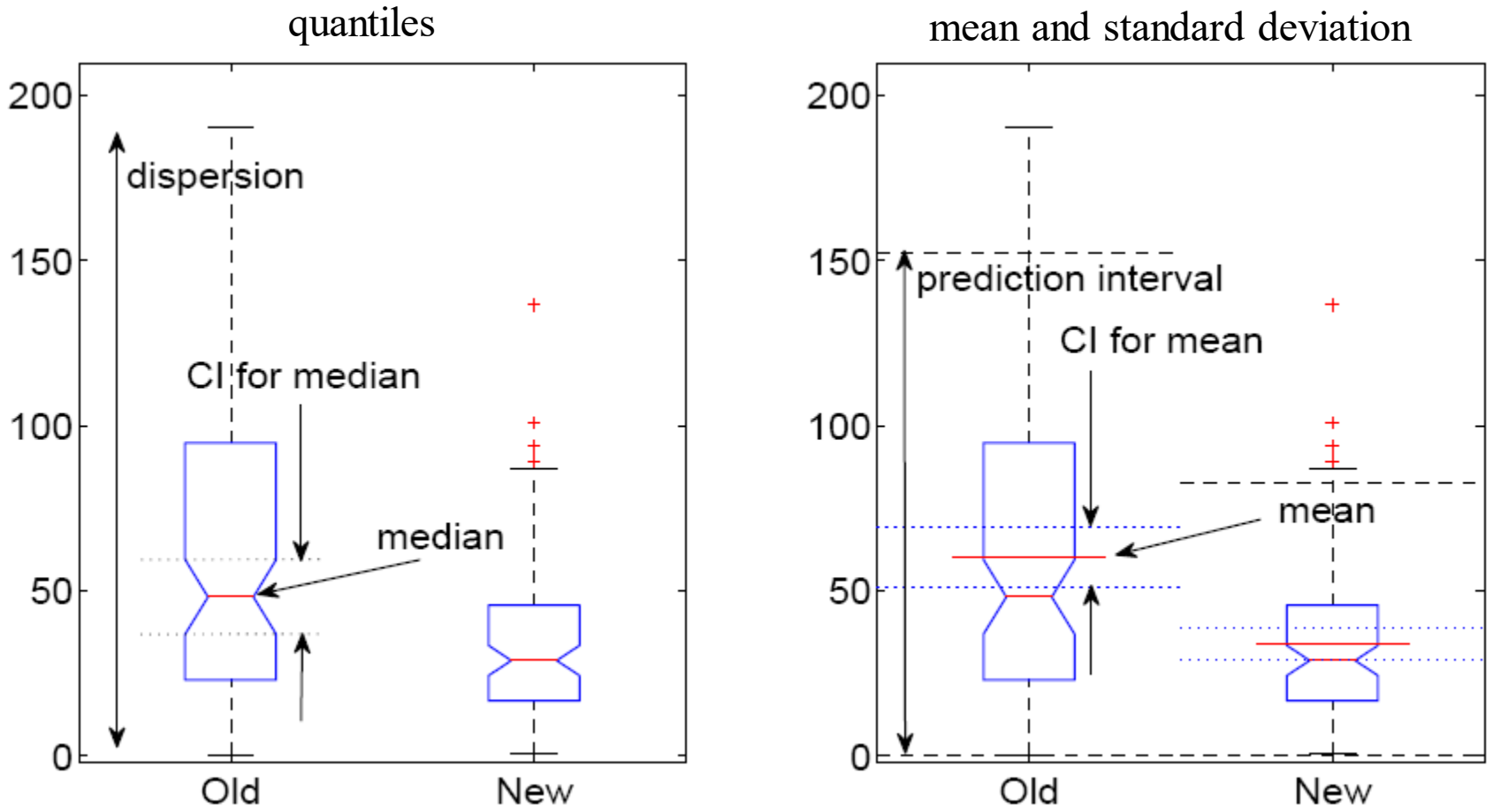


Figure 2.3: Box Plots for the data for Example 2.1. Left: standard box plot commonly used by statisticians showing median (notch) and quartiles (top and bottom of boxes); "dispersion" is an ad-hoc measure, defined here as 1.5 times the inter-quartile distance; the notch width shows the confidence interval for the median. Right: same, overlaid with quantities commonly used in signal processing: mean, confidence interval for the mean ($=$ mean $\pm 1.96\sigma/\sqrt{n}$, where σ is the standard deviation and n is the number of samples) and prediction interval ($=$ mean $\pm 1.96\sigma$).

Coefficient of Variation Summarizes Variability

- Scale free (invariant to change of scale)
- Second order

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

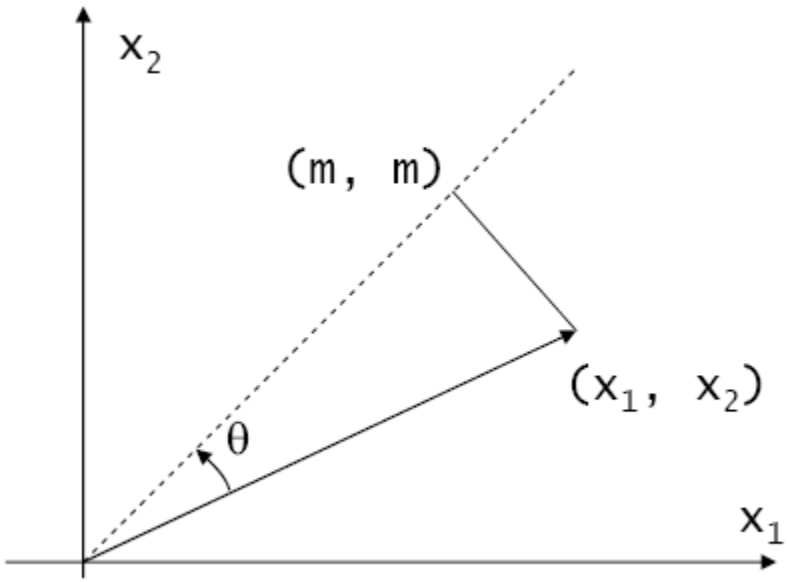
where m is the mean and s the standard deviation

- For any data set with n samples, we have

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

- Does not exist if infinite variance (heavy tailed r.v.)
- Exponential distribution: CoV =1
- What does CoV = 0 mean ?

Jain's Fairness Index is an Alternative to CoV



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

- 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

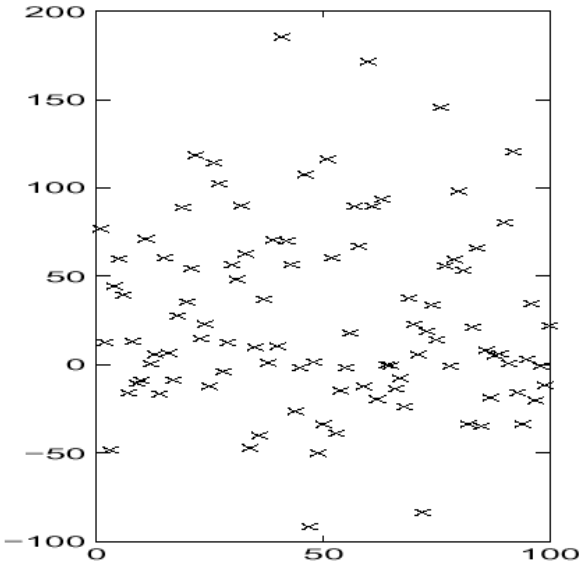
$$\text{JFI} = \frac{1}{1 + \text{CoV}^2}$$

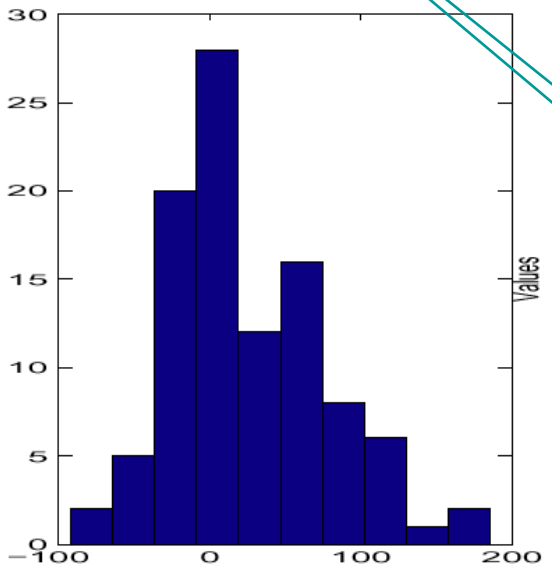
2. Confidence Intervals

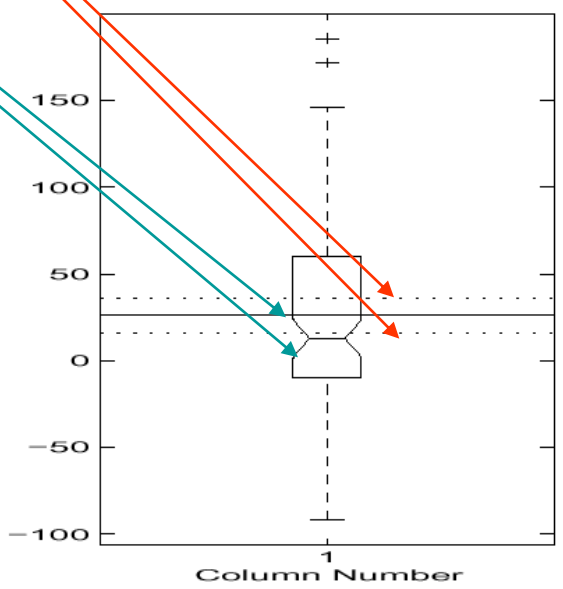
- Any measured performance metric is a random variable, and is therefore only an estimate of the real value
  - We need to quantify the reliability of such estimate
  - The accuracy of an estimate is measured with the confidence intervals
  - There is a confidence interval for every summarized quantity
- Confidence interval is defined by a confidence level (e.g., 95%)
  - It is an interval which contains the true value with that probability
- No simulation result is meaningful without confidence

Confidence Intervals for Mean of Difference

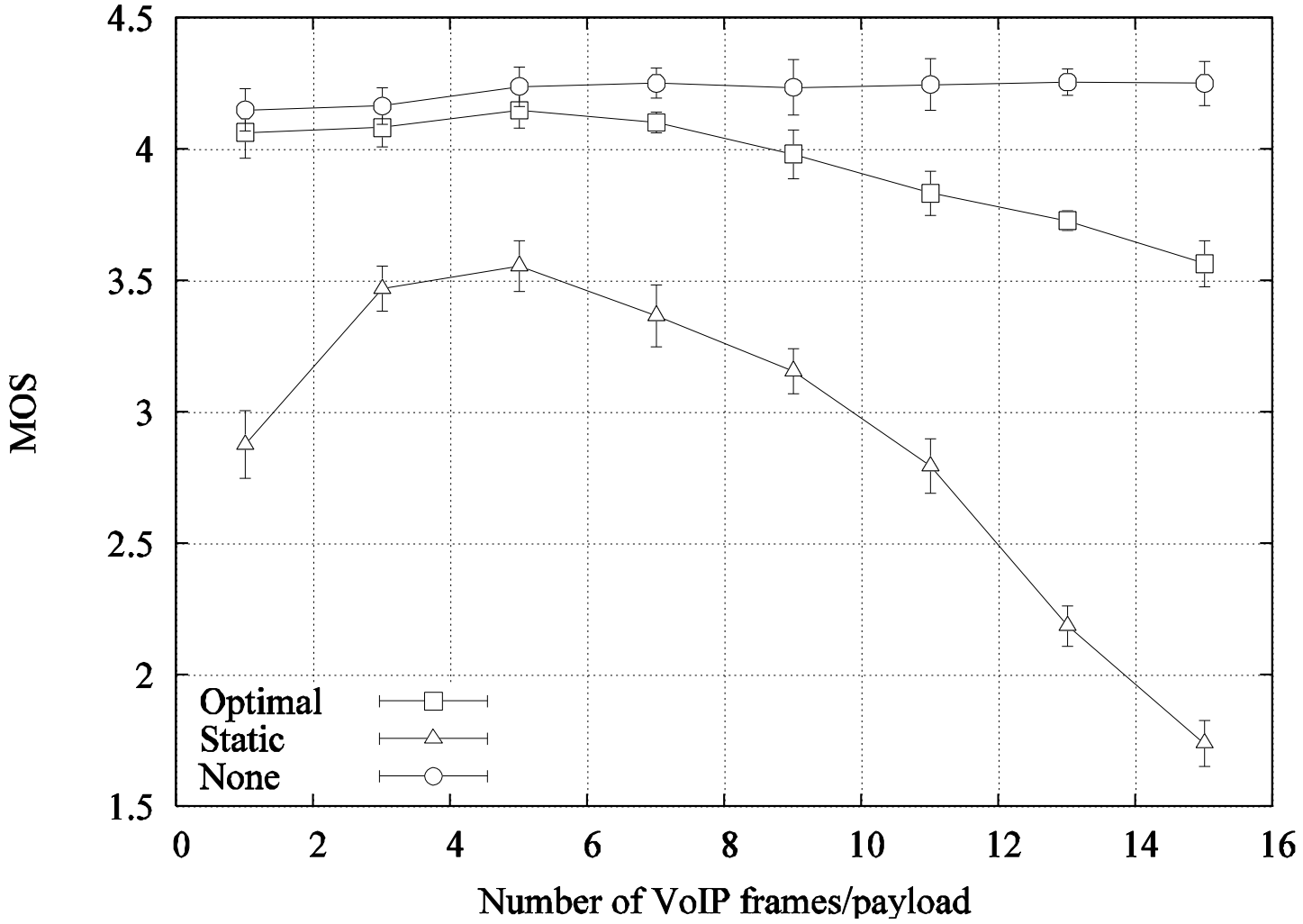
- Continuation of the previous example
- Mean reduction = 26.1 ± 10.2
  - Good gain, but large uncertainty
  - 0 is outside the confidence intervals for mean and for median: we can conclude that there is a positive gain
- Confidence interval for median







Confidence intervals in performance results



Computing Confidence Intervals

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.

- How to compute the confidence intervals depends on the quantity being studied and on the assumptions we can make about the data
- It is fairly simple in general if we can assume that the data comes from an Independent and Identically Distributed (iid) model
  - We will discuss the case of dependent data later
- Also, we assume that the data follows a well-defined (though unknown) F(x)
- Note that the confidence interval is not unique in general

CI for median (or percentiles)

- The simplest of all. Median is defined as middle data point (average of the two middle points if the data set contains an even number of data points)
- Robust: always true **provided** iid assumption holds (critical)

THEOREM 2.2.1 (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 Section A 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 $\gamma = 0.95$).

CI for mean and Standard Deviation

- This is another, most commonly used method
- It refers to mean and variance instead of percentiles
- It requires some assumptions to hold, **in addition** to iid
  - Typically, gaussian data or finite variance and large n
  - may be misleading if they do not hold, need to check
- Unlike for median and quantiles, there is no exact theorem in this case, but the results are asymptotic and/or heuristic.

CI for mean, asymptotic case

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

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*

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

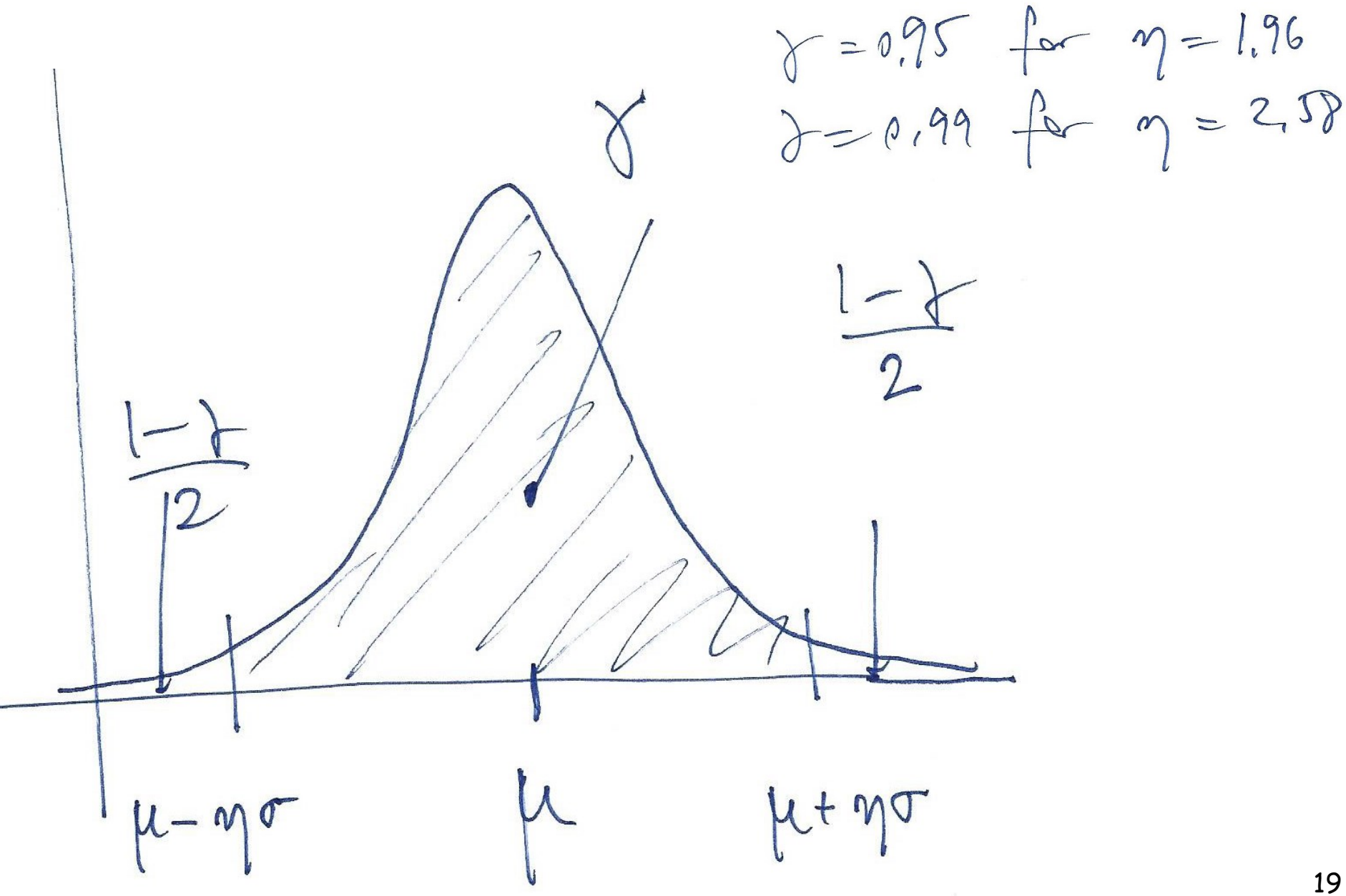
$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \tag{2.20}$$

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 $\gamma = 0.99$.

Gaussian quantiles



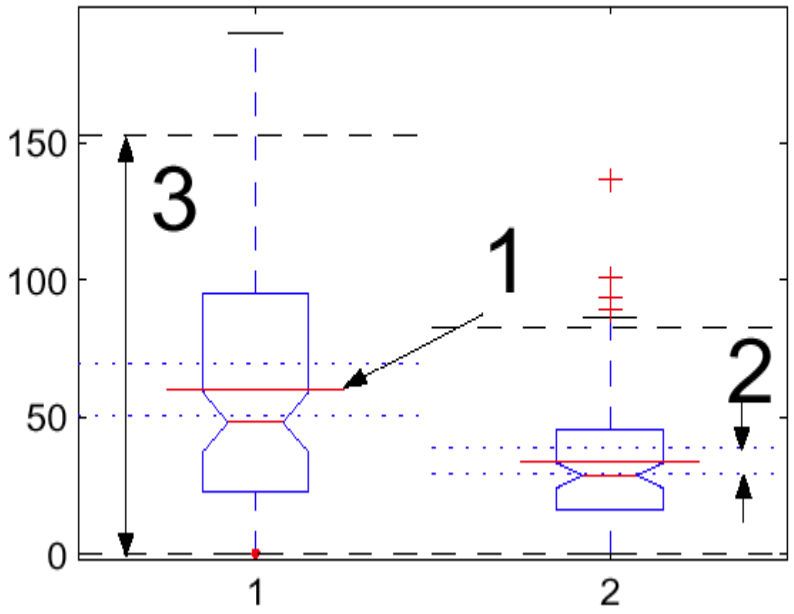
Example

- 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



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

We test a system 10'000 times 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)

$$\mu_n = 0.02$$

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

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

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

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

Confidence Interval for Success Probability

- Problem statement: want to estimate proba of failure; observe n outcomes; no failure; confidence interval ?
- Example: we test a system 10 times for failures and find 0 failures: 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

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*

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

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

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.

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

Check on the web: "rule of three"

Also read the article at http://www.pmean.com/01/zeroevents.html

---

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 $p \leq 30.8\%$.

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

- 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 statistic T_r as a function of these replicates
  - Bootstrap percentile estimate is $(T_{(25)}, T_{(975)})$

- See Ross, pp. 126-133

Confidence Interval for statistical indices

- Use of 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:

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

1: $R = \lceil 2\, r_0/(1-\gamma) \rceil - 1$ ▷ For example $r_0 = 25$, $\gamma = 0.95$, $R = 999$
2: **for** $r = 1 : R$ **do**
3: draw n numbers with replacement from the list $(x_1, ..., x_n)$ and call them $X_1^r, ..., X_n^r$
4: let $T^r = t(\vec{X}^r)$
5: **end for**
6: $(T_{(1)}, ..., T_{(R)}) = \text{sort}\,(T^1, ..., T^R)$
7: Prediction interval is $[T_{(r0)}\ ;\ T_{(R+1-r0)}]$

- Try to prove it (hint: start from the proof of Theorem 2.4.1 – will see it later)
- The method can be used for any choice of the statistic

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
- Confidence interval for success probability requires special attention when success or failure is rare
- We need to *verify* the assumptions

3. The Independence Assumption

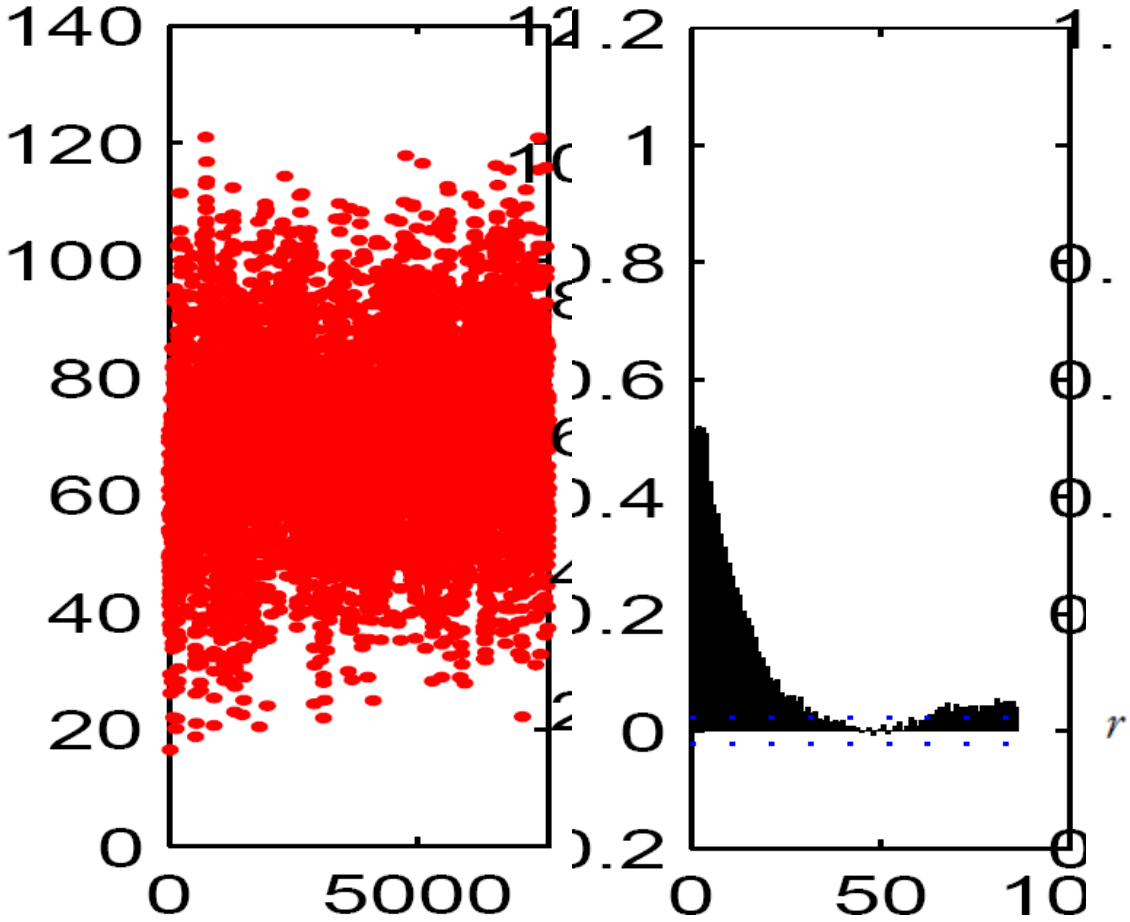
- Confidence Intervals require that we can assume that the data comes from an iid model

  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 will have methods for time series

Example



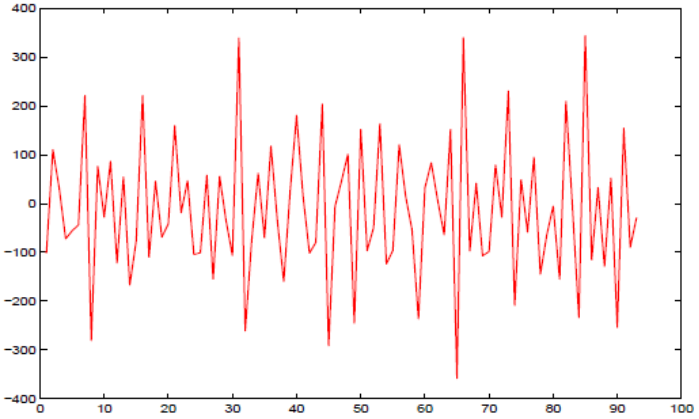
data ACF

- Pretend data is iid: CI for mean is [69; 69.8]
- Is this biased ?
- How to check for correlation? Look at the ACF (see black curve)

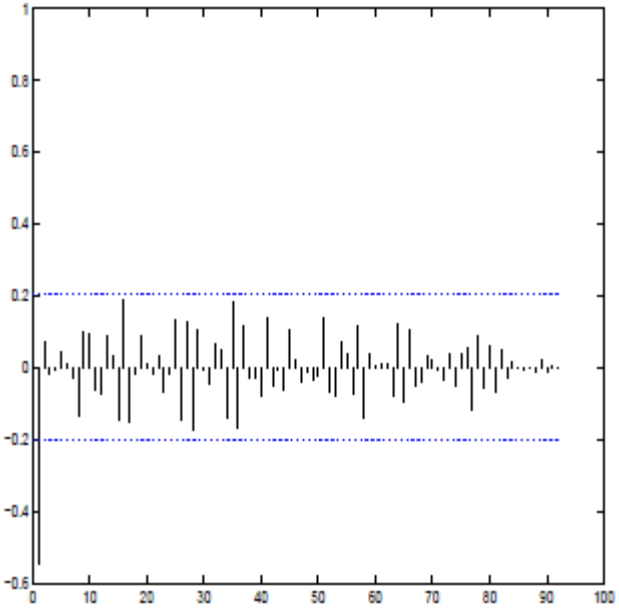
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

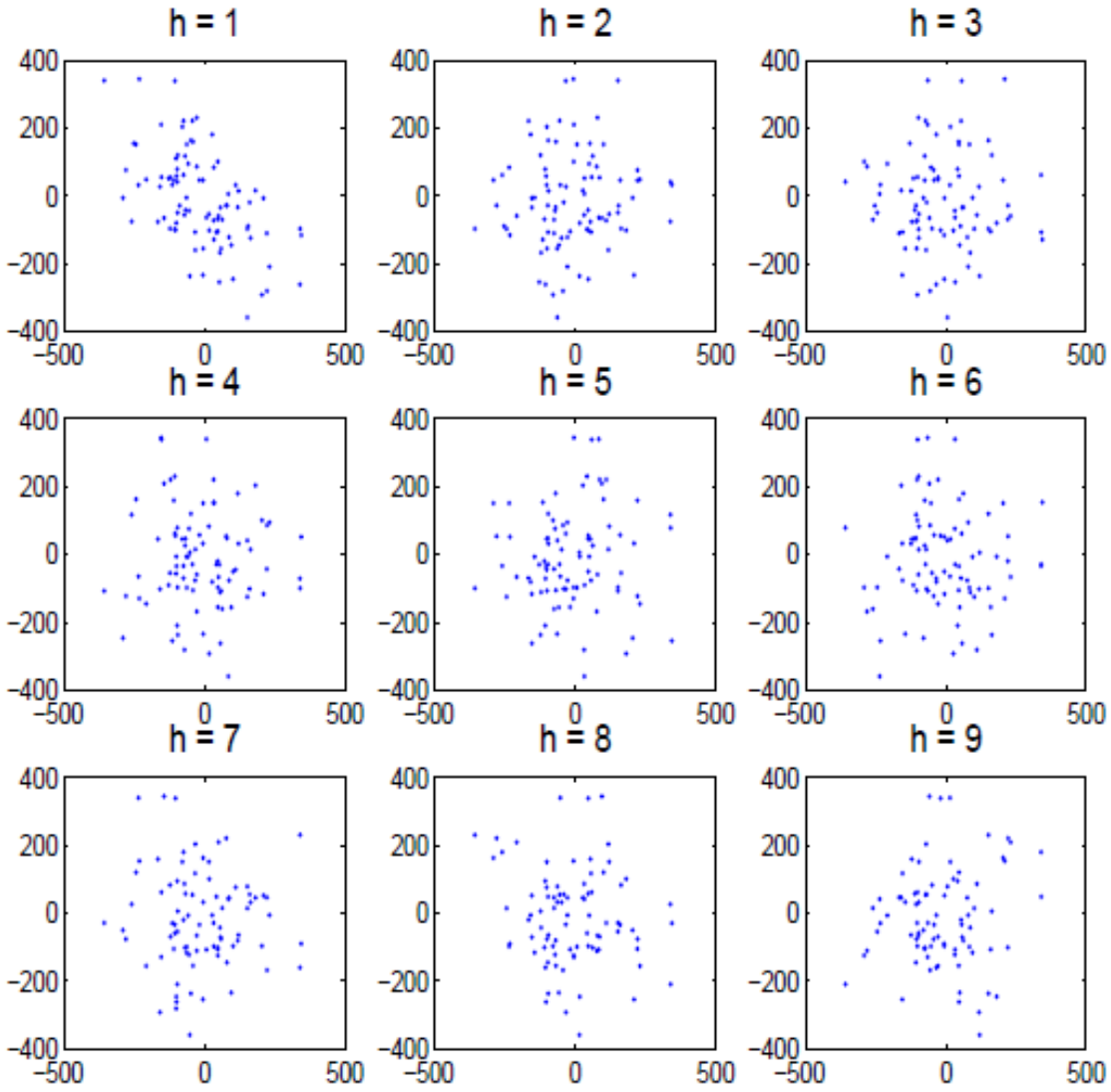
Lag plot



(a) Data



(c) Autocorrelation



(d) Lag Plots

- Lag plots show correlation between shifted versions of the same data sequence

What happens if data is not iid ?

- If data is positively correlated
  - Neighboring values look similar
  - Frequent in measurements
  - CI is underestimated: there is less information in the data than one thinks

- Possible solution:
  - Subsample the data so that the correlation is broken
    - Periodic vs. Random sampling
    - data looks more similar to iid but there is less data
    - For certain data (called long-range dependent) this does not work
  - Try to model the correlation explicitly (e.g., see Example 2.4 in the book)

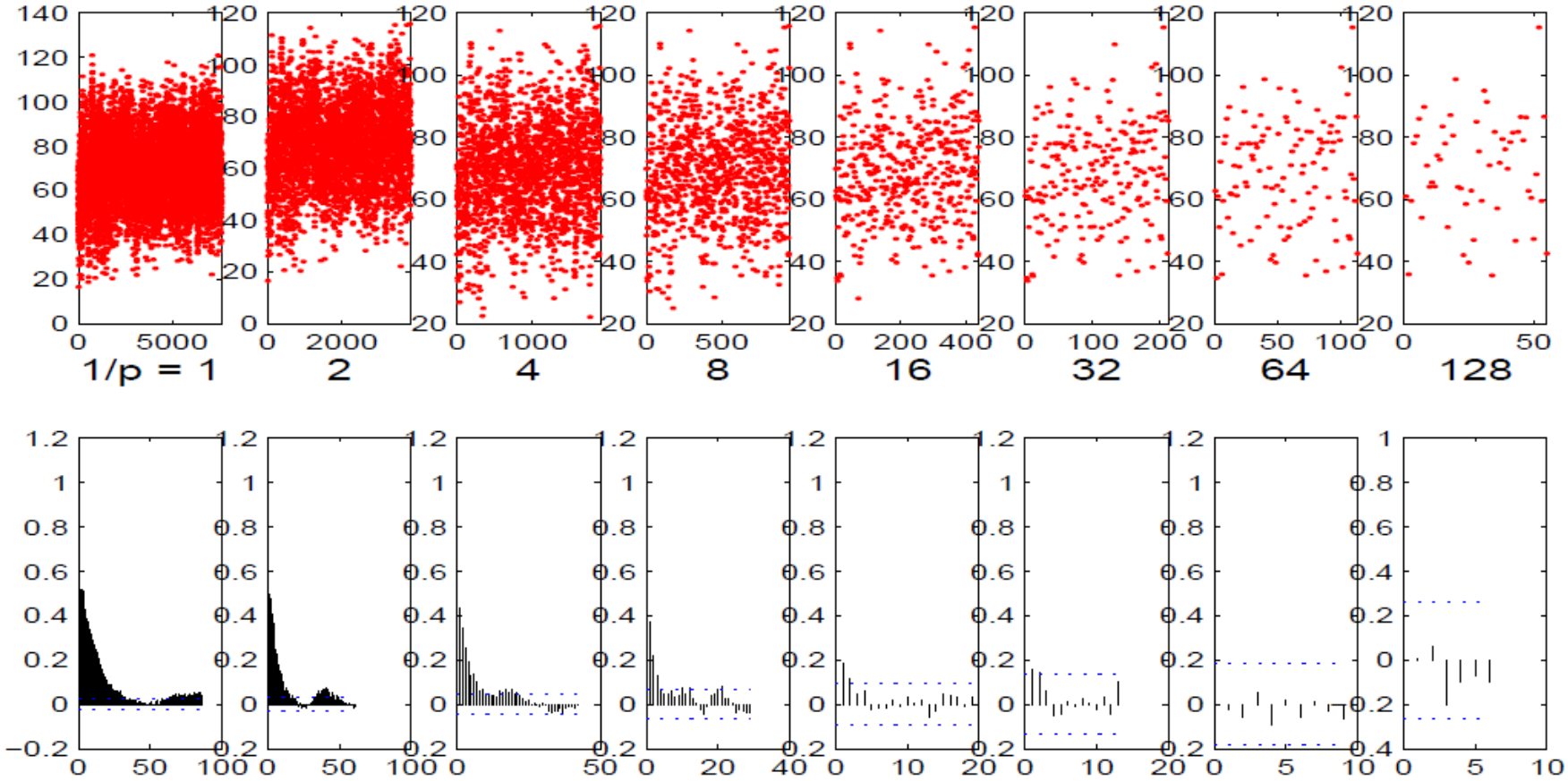


Figure 2.9: Execution times for $n = 7632$ requests (top left) and autocorrelation function (bottom left), and for the data sub-sampled with probability $p = 1/2$ to $1/2^7 = 1/128$. The data appears stationary and roughly normal so the auto-correlation function can be used to test independence. The original data is positively correlated, but the sub-sampled data looses correlation when the sampling probability is $p = 1/64$. The turning point test for the subsampled data with $p = 1/64$ has a p-value of 0.52648, thus at confidence level 0.95 we accept the null hypothesis, namely, the data is iid. The sub-sampled data has 116 points, and the confidence interval obtained from this for the median of the sub-sampled data is $[66.7, 75.2]$ (using Theorem 2.2.1). Compare with the confidence interval that would be obtained if we would (wrongly) assume the data to be iid : $[69.0, 69.8]$. The iid assumption underestimates the confidence interval because the data is positively correlated.

Further discussion

- Monte-Carlo vs. Event-driven simulations
  - Effect of correlation
  - What is a run in this case?
  - How to compute CI?
- When to stop a simulation?
  - Rigorous approach from theory
  - Some empirical techniques
- How to improve CI?
  - Run more samples: obvious but not always possible
  - Variance reduction techniques: tricks to improve accuracy