Metody probabilistyczne i statystyka, 2021
informatyka algorytmiczna, WIiT PWr

5-Statistics - Introduction

Sampling a population

- **population of units**
- **each unit has some properties /numerical values**

Investigation:

- **Approach 1: take the whole population and analyze**
- **Approach 2:**
  - **take only a (random) sample,**
  - **analyze sample**
  - **conclude that the whole population has the same properties**

Examples:

- **democracy in ancient Athens**
- **pharmacy, medical research**
- **system testing**
- **jury in US courts**

Statistics

By *statistics* we mean any function f of the sample

Examples:

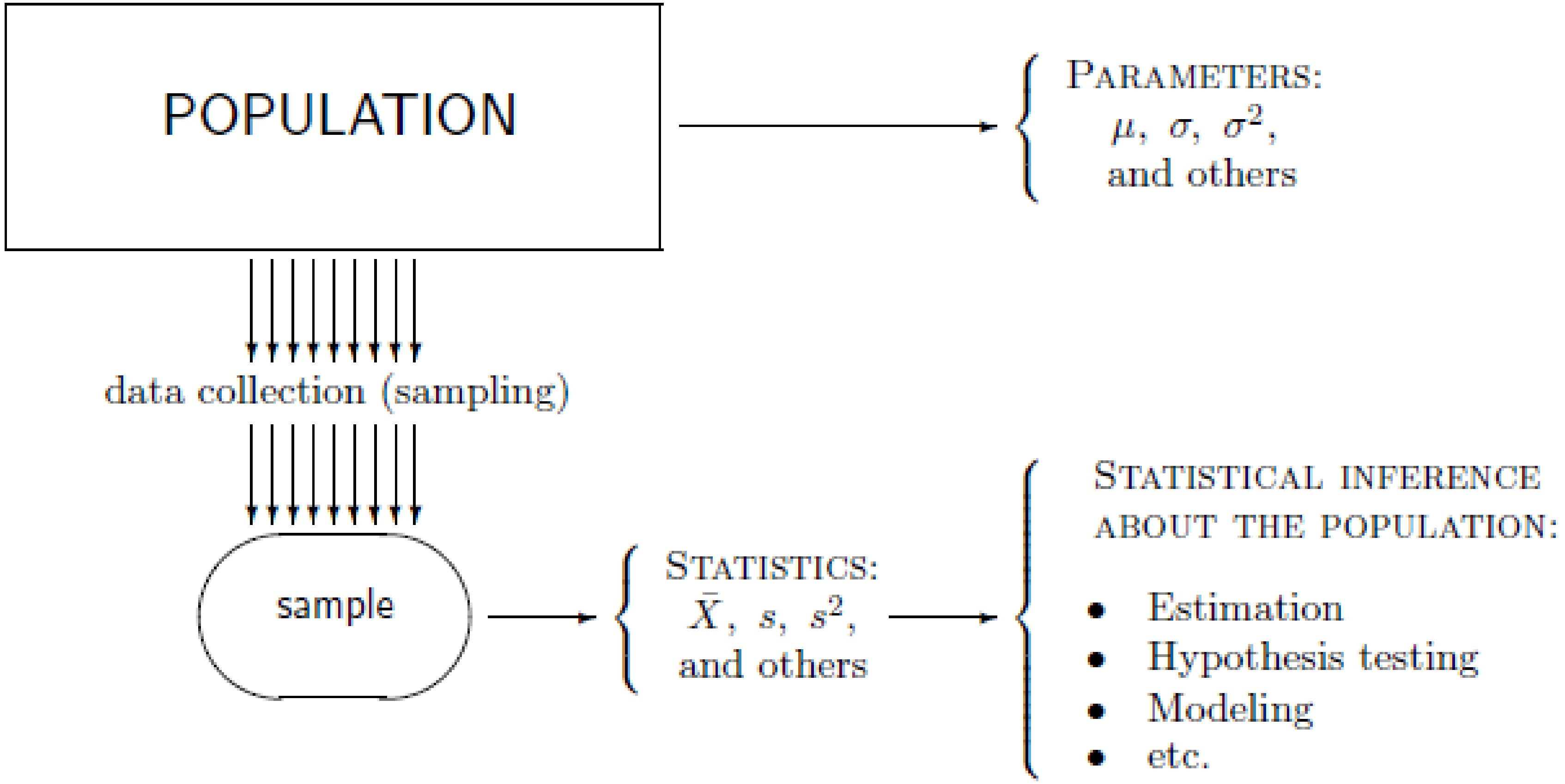
- **mean (average value)**
- **variance of the sample**
- **median value**
- **smallest value**
- **…**

Estimators

Θ = f(whole population)   population parameter

$\hat{\Theta}$ = estimator of Θ computed over the sample

Overall picture



Errors

- **Sampling errors – based on the fact that we have only a small sample**

- **Non-sampling error - faulty choice of a sample**

Non-sampling errors

examples of poor sampling leading to misleading results

1) **Comparing the number of Covid related deaths in 2021 for vaccinated versus non-vaccinated persons**

Arguments from the Ministry:

X cases of fully vaccinated patients
Y cases of non- vaccinated patients
“X<<Y so statistics shows that vaccine works”

Non-sampling errors

examples of poor sampling

2) Comparing the number of Covid related deaths in January 2022 for vaccinated versus non-vaccinated persons (fraction of vaccinated people almost constant)

“25% of death cases for fully vaccinated patients, so a vaccine reduces the probability of death 3 times”

Wrong: comparing apples with peaches

Example of professional approach

See e.g. reports of the Washington State health authority

Compare patients splitting them into groups depending on crucial characteristics such as

- **age**
- **health condition**
- **…**

then comparisons within each homogenous group

Mean

Sample mean $\bar{X}$ is the arithmetic average,

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

Bias

An estimator $\hat{\theta}$ is unbiased for a parameter θ if its expectation equals the parameter,

$$\mathbf{E}(\hat{\theta}) = \theta$$

for all possible values of θ.

Bias of $\hat{\theta}$ is defined as $\text{Bias}(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)$.

For the mean value:

$$\mathbf{E}(\bar{X}) = \mathbf{E}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\mathbf{E}X_1 + \ldots + \mathbf{E}X_n}{n} = \frac{n\mu}{n} = \mu.$$

Consistency

The estimator $\hat{\theta}$ is consistent if

$$P\left\{|\hat{\theta}-\theta| > \varepsilon\right\} \to 0 \text{ as } n \to \infty$$

Consistency of mean estimator

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\mathrm{Var}X_1 + \ldots + \mathrm{Var}X_n}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

$$\boldsymbol{P}\left\{|\bar{X} - \mu| > \varepsilon\right\} \leq \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2/n}{\varepsilon^2} \to 0,$$

Chebyshev inequality

Asymptotic normality

By Central Limit Theorem, the following random variable converges to the Standard Normal random variable:

$$Z = \frac{\bar{X} - \mathrm{E}\bar{X}}{\mathrm{Std}\bar{X}} = \frac{\bar{X} - \mu}{\sigma\sqrt{n}}$$

Sample median

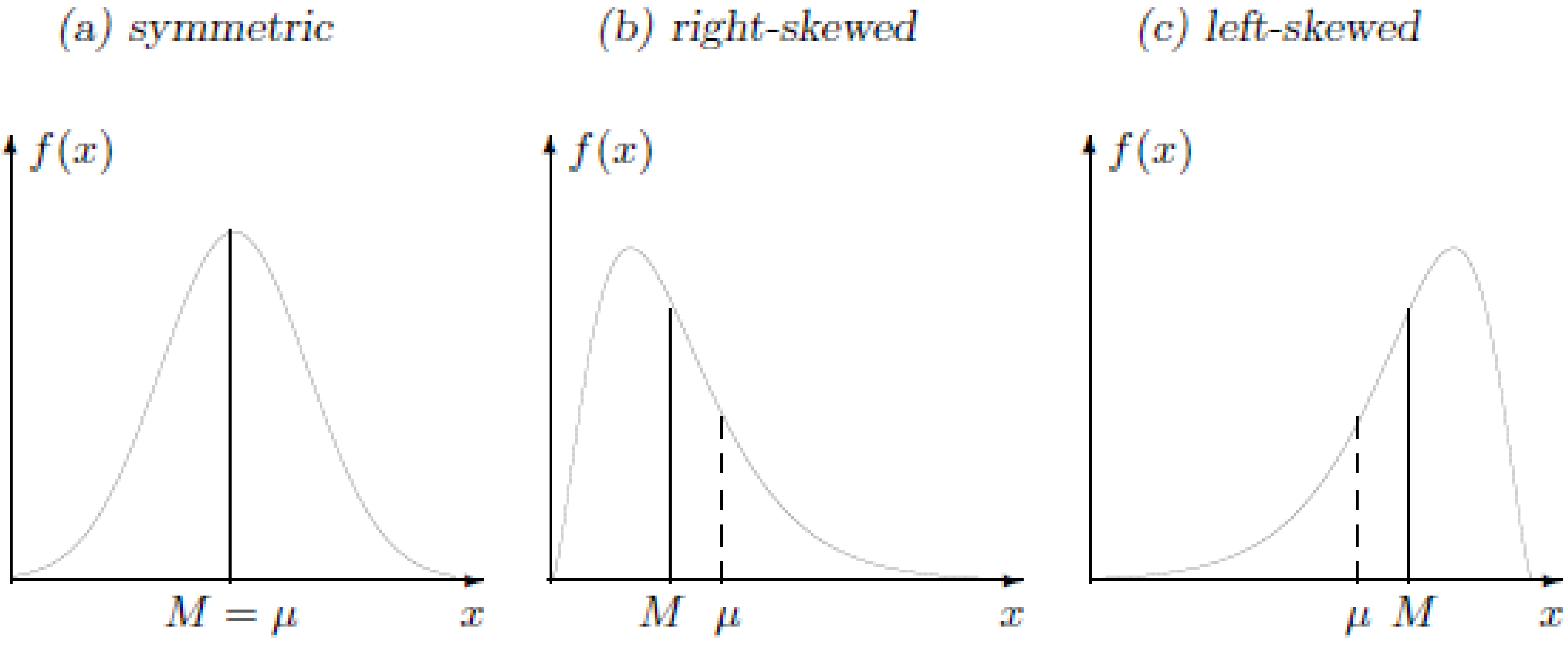
Sample median $\hat{M}$ is a number that is exceeded by at most a half of observations and is preceded by at most a half of observations.

Population median

Each M such that:

$$\begin{cases} P\{X > M\} & \leq & 0.5 \\ P\{X < M\} & \leq & 0.5 \end{cases}$$

Examples



(a) symmetric
(b) right-skewed
(c) left-skewed
$f(x)$
$f(x)$
$f(x)$
$M = \mu$
x
M μ
x
μ M
x

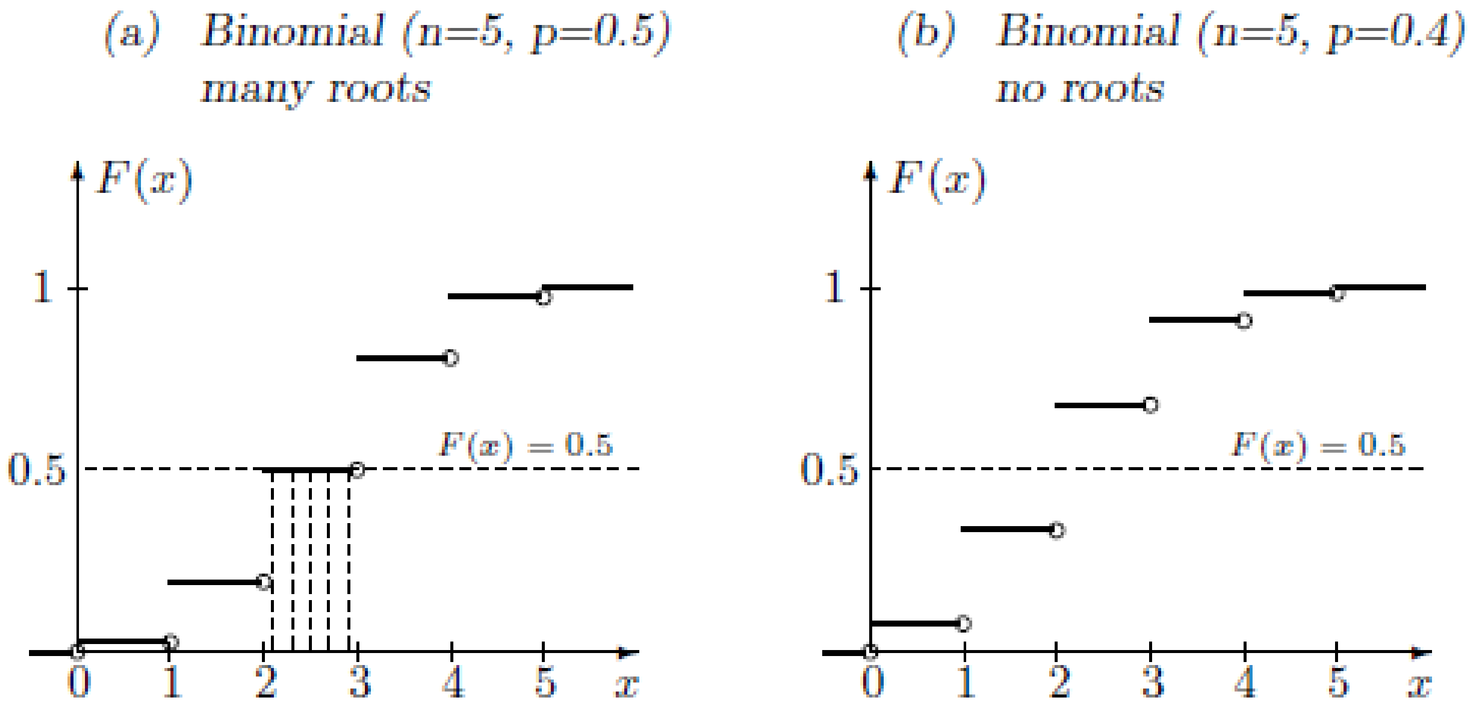
Example: exponential distribution

$$F(x) = 1 - e^{-\lambda x} \quad \text{for} \quad x > 0.$$

$$F(M) = 1 - e^{-\lambda x} = 0.5$$

$$M = \frac{\ln 2}{\lambda} = \frac{0.6931}{\lambda}.$$

Examples for discrete binomial distribution



Quantyle (Kwantyl)

A p-quantile of a population is such a number x that solves equations

$$\begin{cases} P\{X < x\} & \leq & p \\ P\{X > x\} & \leq & 1-p \end{cases}$$

Percentile (Percentyl)

A γ-percentile is (0.01γ)-quantile.

Kwartyl

Q1=25percentile
Q2=50percentile
Q3=75percentile

Sample variance

For a sample $(X_1, X_2, \ldots, X_n)$, a sample variance is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 .$$

Alternative formula for sample variance

$$\boxed{s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}.}$$

$$\sum (X_i - \bar{X})^2 = \sum X_i^2 - 2\bar{X}\sum X_i + \sum \bar{X}^2 = \sum X_i^2 - 2\bar{X}(n\bar{X}) + n\bar{X}^2 = \sum X_i^2 - n\bar{X}^2.$$

Unbiasedness of s

assume $\mu = \mathbf{E}(X) = 0.$

$$\mathbf{E}X_i^2 = \mathrm{Var}X_i = \sigma^2,$$

$$\mathbf{E}\bar{X}^2 = \mathrm{Var}\bar{X} = \sigma^2/n.$$

$$\mathbf{E}s^2 = \frac{\mathbf{E}\sum X_i^2 - n\,\mathbf{E}\bar{X}^2}{n-1} = \frac{n\sigma^2 - \sigma^2}{n-1} = \sigma^2.$$

If mean value is non-zero:

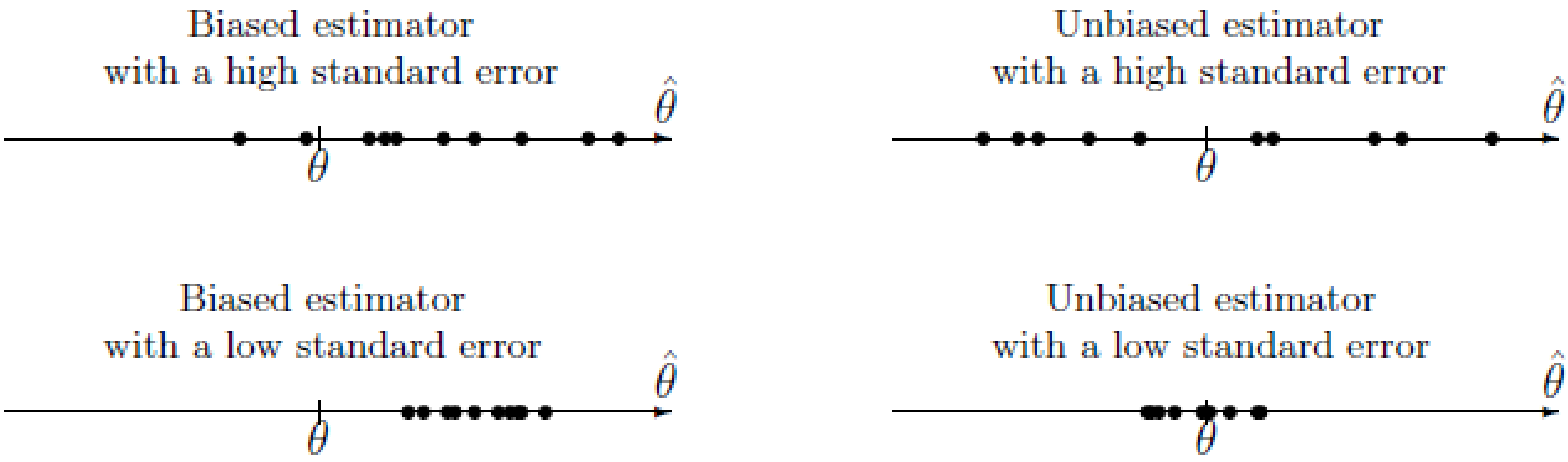
let $Y_i = X_i - \mu.$

$$s_Y^2 = \frac{\sum \left(Y_i - \bar{Y}\right)^2}{n-1} = \frac{\sum \left(X_i - \mu - (\bar{X} - \mu)\right)^2}{n-1} = \frac{\sum \left(X_i - \bar{X}\right)^2}{n-1} = s_X^2 .$$

$$\mathbf{E}(s_X^2) = \mathbf{E}(s_Y^2) = \sigma_Y^2 = \sigma_X^2 .$$

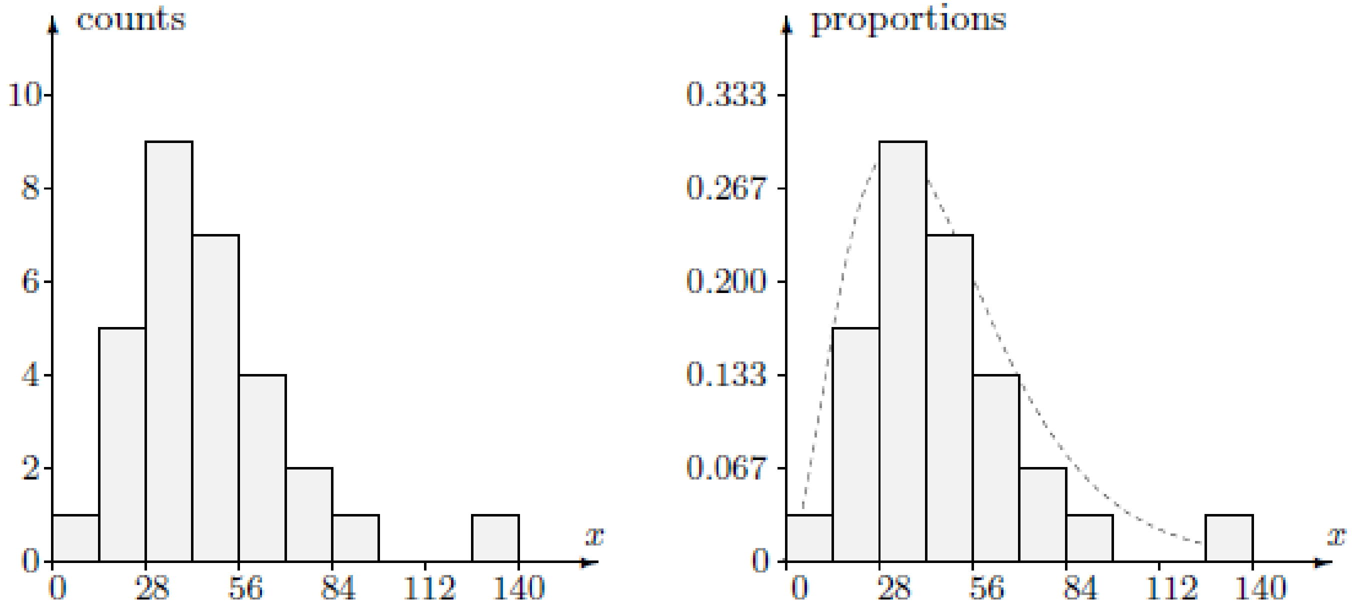
Standard error of an estimator

Standard error of an estimator $\hat{\theta}$ is its standard deviation, $\sigma(\hat{\theta}) = \text{Std}(\hat{\theta})$.



The problem of outliers

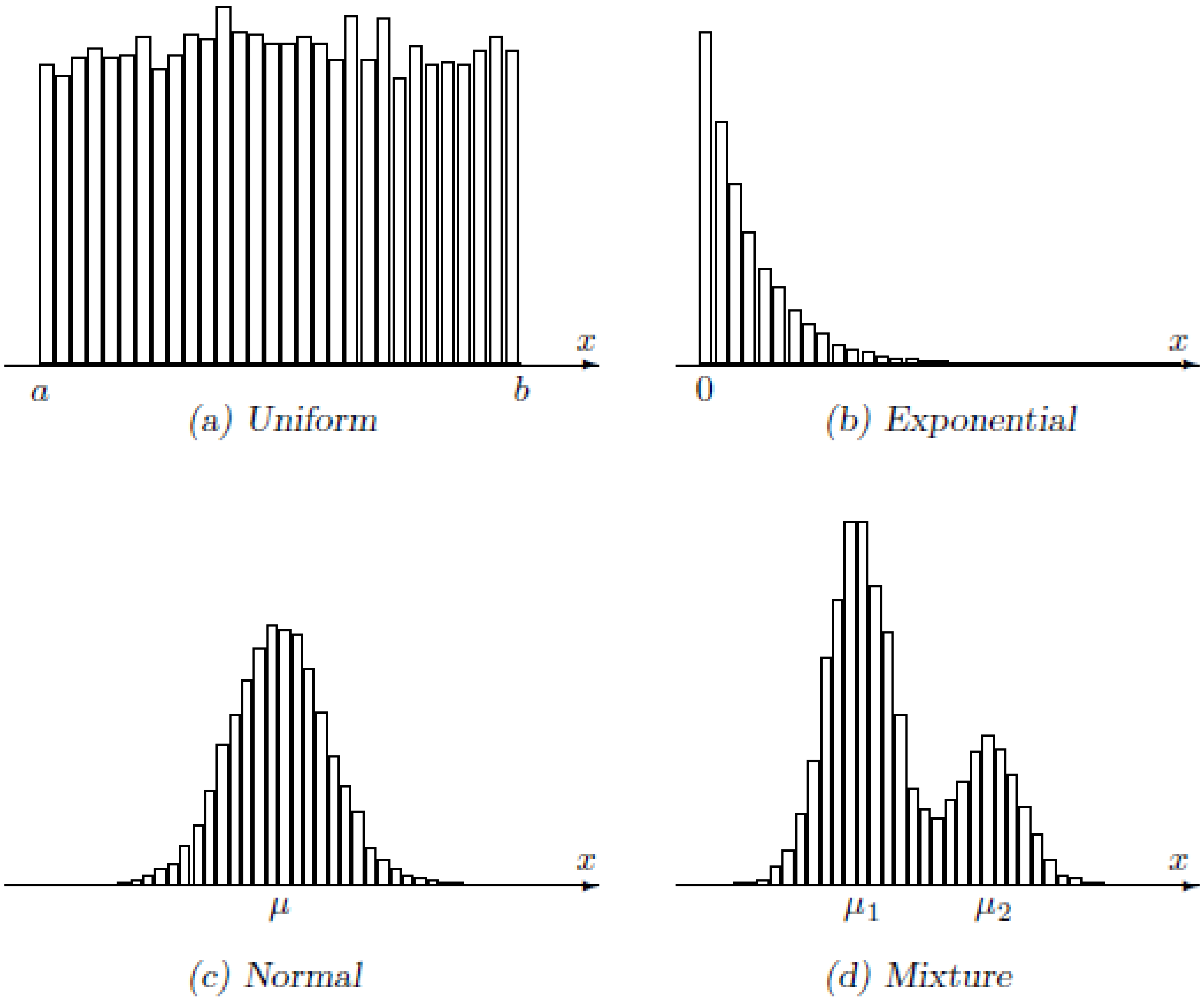
Visualizing a sample – histogram



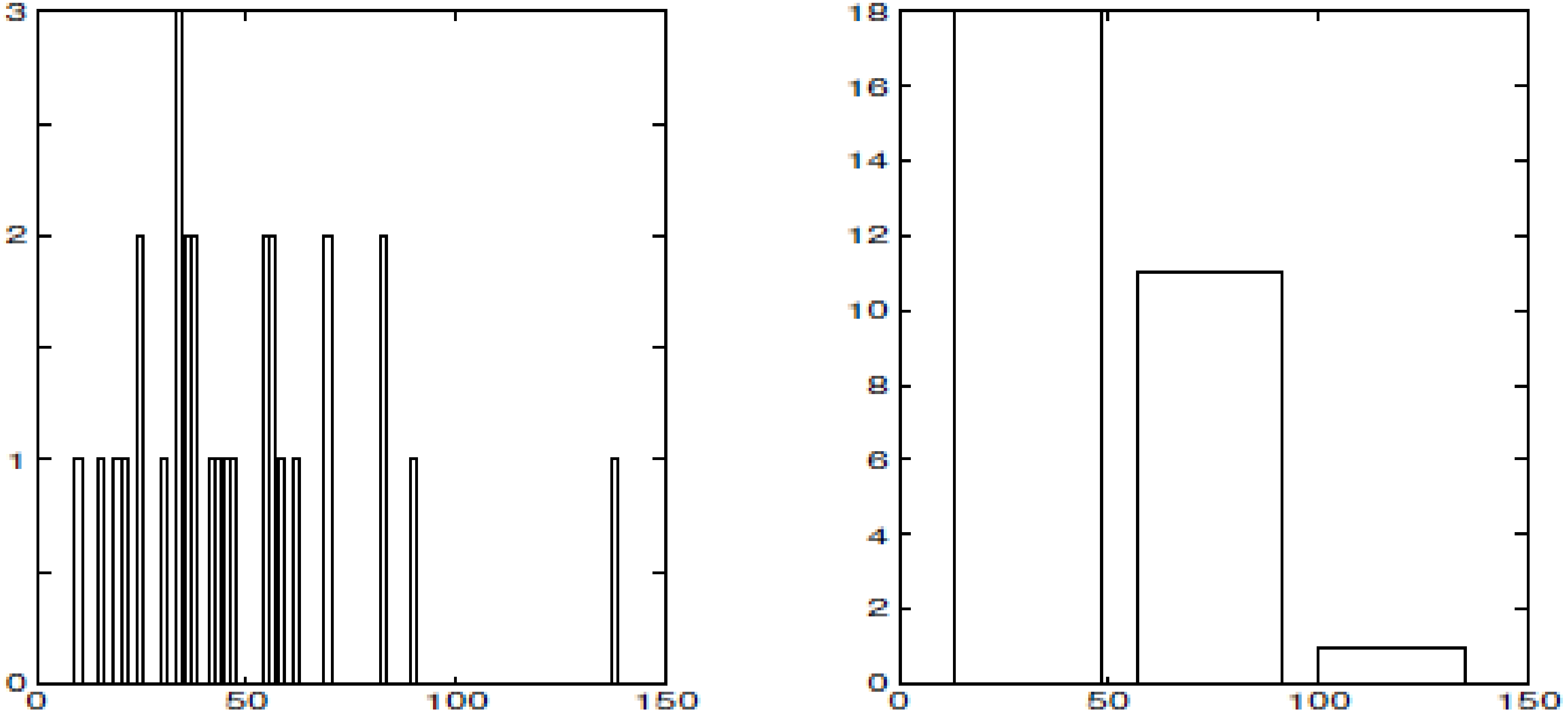
(a) Frequency histogram

(b) Relative frequency histogram

A few cases



Wrong choice of bin size



Stem+leaf

Sample values:
0.9, 1.5, 1.9, 2.2, 2.4, 2.5,
3.0, 3.4, 3.5, 3.5, 3.6, 3.6, 3.7,
3.8
…
8.2, 8.2, 8.9, 13.9

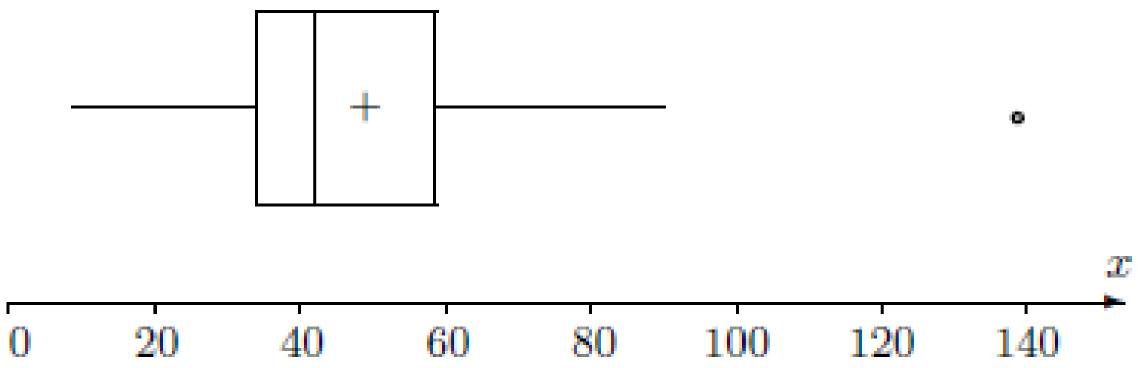
0	9							
1	5	9						
2	2	4	5					
3	0	4	5	5	6	6	7	8
4	2	3	6	8				
5	4	5	6	6	9			
6	2	9						
7	0							
8	2	2	9					
9								
10								
11								
12								
13	9							

Stem+leaf for two samples

						0	3	4	
					5	1	0	6	9
			1	1	8	2			
0	2	3	5	5	9	3	8		
		1	3	8	8	4	6		
					4	5			
						6	1	6	7
						7	8		

Box plot example

$\bar{X} = 48.2333;\ \min X_i = 9,\ \hat{Q}_1 = 34,\ \hat{M} = 42.5,\ \hat{Q}_3 = 59,\ \max X_i = 139.$



Example of usage

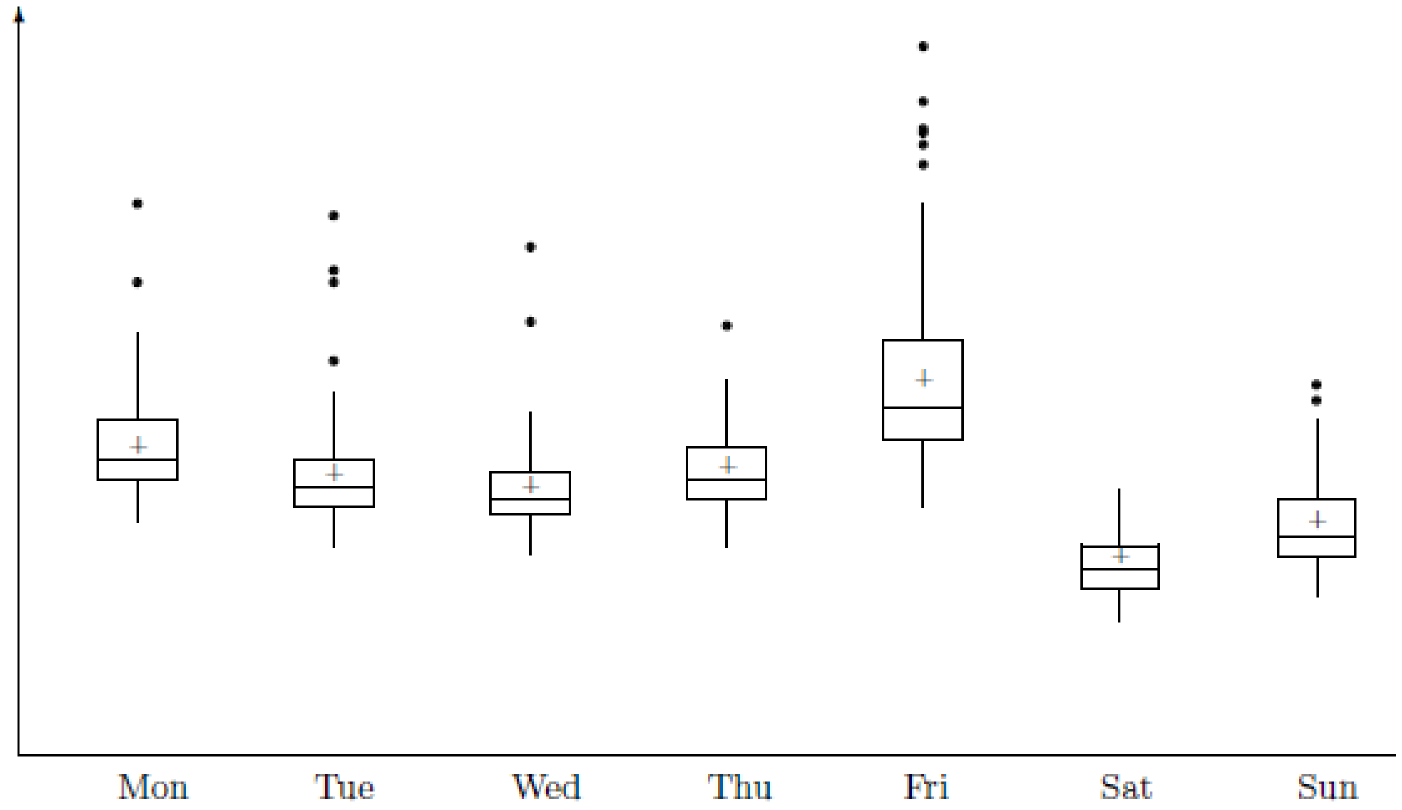


FIGURE 8.10: *Parallel boxplots of internet traffic.*