Metody probabilistyczne i statystyka, 2022
informatyka algorytmiczna, WIiT PWr

5-Statistics, Introduction

Chapter 8 in texbook

Sampling a population

Population: $u_1, u_2, \ldots$
A (numerical) **property** $F(u_i)$ for each u_i

Question: **how F behaves in the population**

Approach 1: take the whole population and analyze

Approach 2:

- take only a (random) sample,
- analyze sample
- attempt to say something about the whole population

Examples:

- **pharmacy, medical research**

- **system testing**

- **jury in US courts**

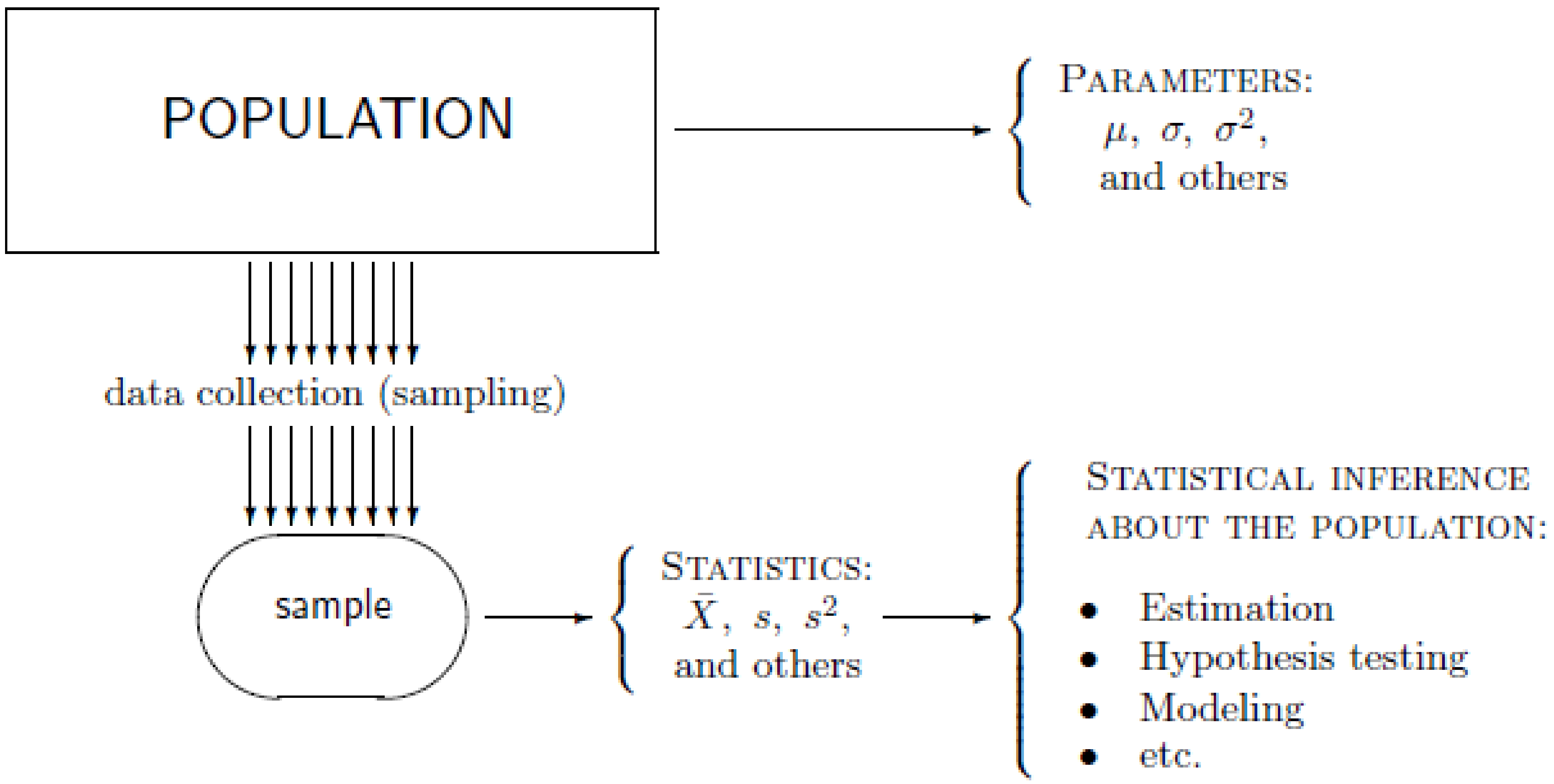
Statistics

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

Examples:

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

Statistics should be useful (not every f is useful)


POPULATION
PARAMETERS:
μ, σ, σ^2,
and others
data collection (sampling)
sample
STATISTICS:
$\bar{X}$, s, s^2,
and others
STATISTICAL INFERENCE ABOUT THE POPULATION:
Estimation
Hypothesis testing
Modeling
etc.

Estimators

Θ = f(whole population)   population parameter

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

$\hat{\Theta}$:= F(sample)

Errors

- **Sampling errors:** due to the fact that we see only a small sample and not the whole population

- **Non-sampling error:** faulty way of choosing a sample

Non-sampling errors: Example of poor sampling

asking for political preferences on Facebook and projection on the whole population

Example of professional approach

e.g. COVID reports of 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

Important statistics: 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 $\text{Bias}(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)$. $= E(\hat{\theta}) - \theta$

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 (zgodny) if

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

(n is the sample size)

Consistency of mean estimator

Recall that

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

So:

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

Chebyshev inequality

Asymptotic normality

By Central Limit Theorem, the random variable

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

converges to the Standard Normal random variable:

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.

Example:

Sample values:
2345, 3248, 3356, 6788, 12122

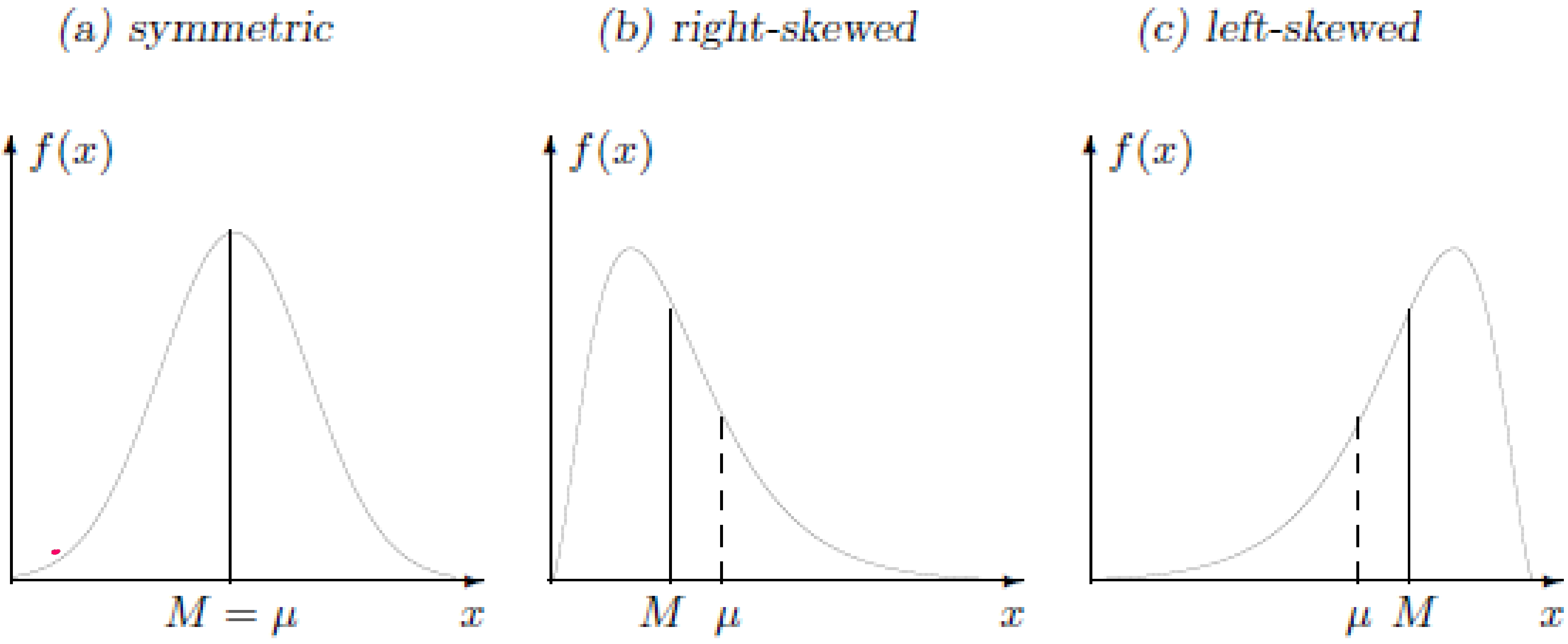
Median: 3356
Mean: 5571.8

Population median

Each M such that:

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

Examples



Example: exponential distribution

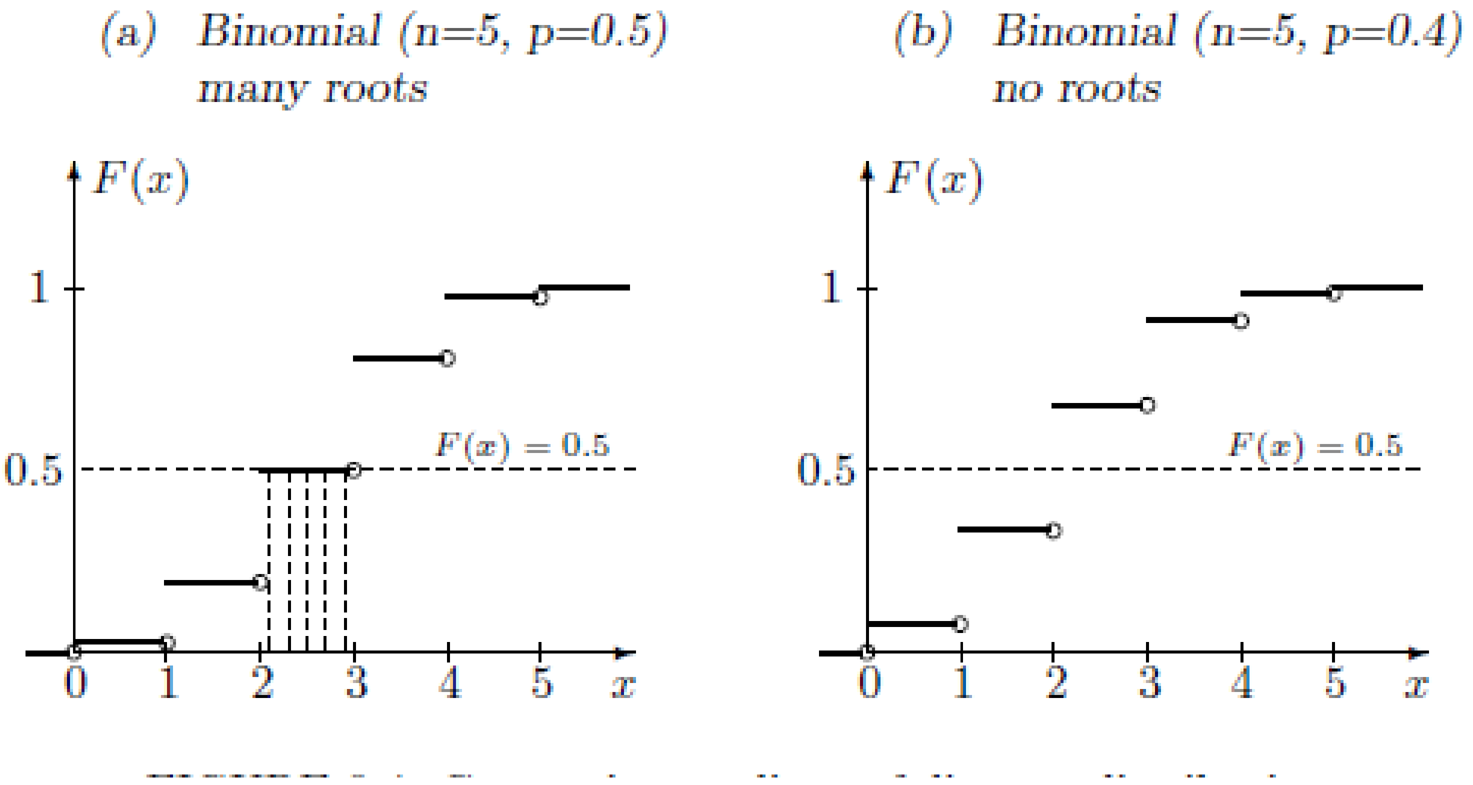
$$F(x) = 1 - e^{-\lambda x} \text{ for } x > 0.$$

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

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

recall that E(x)=1/λ

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

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

Alternative formula for sample variance

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

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

Estimator s is not biased!

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:

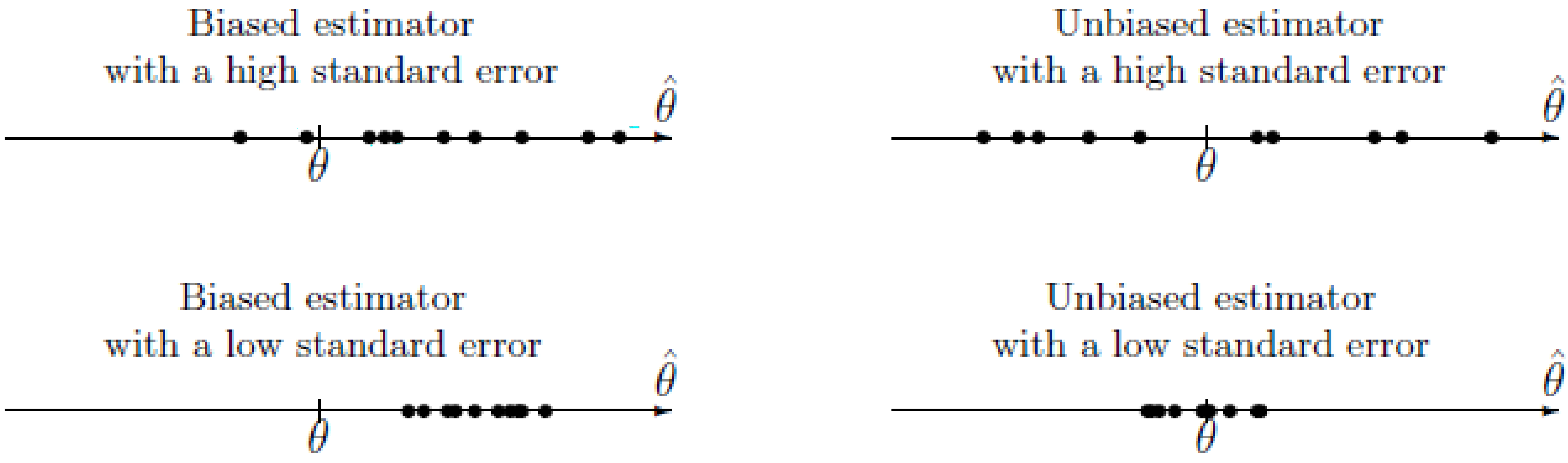
let $Y_i = X_i - \mu.$

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

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

Standard error of an estimator

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



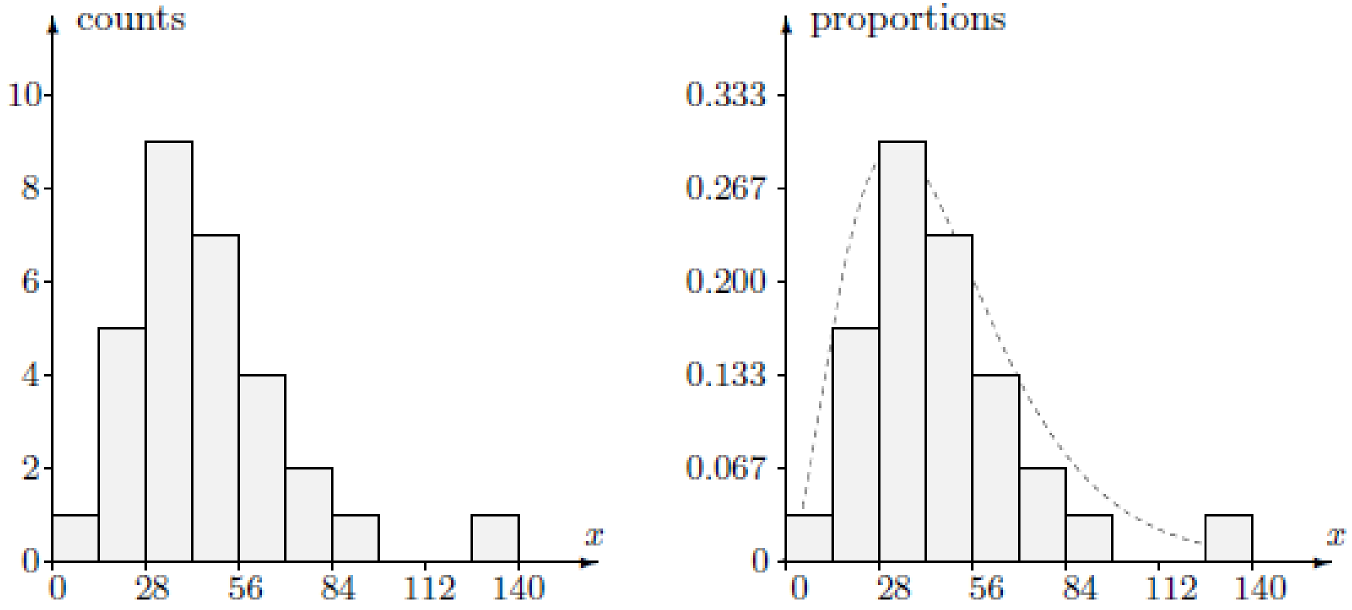
Standard error of an estimator

standard error: concerns a sample and an estimator

the standard deviation for the population is something different

The problem of outliers

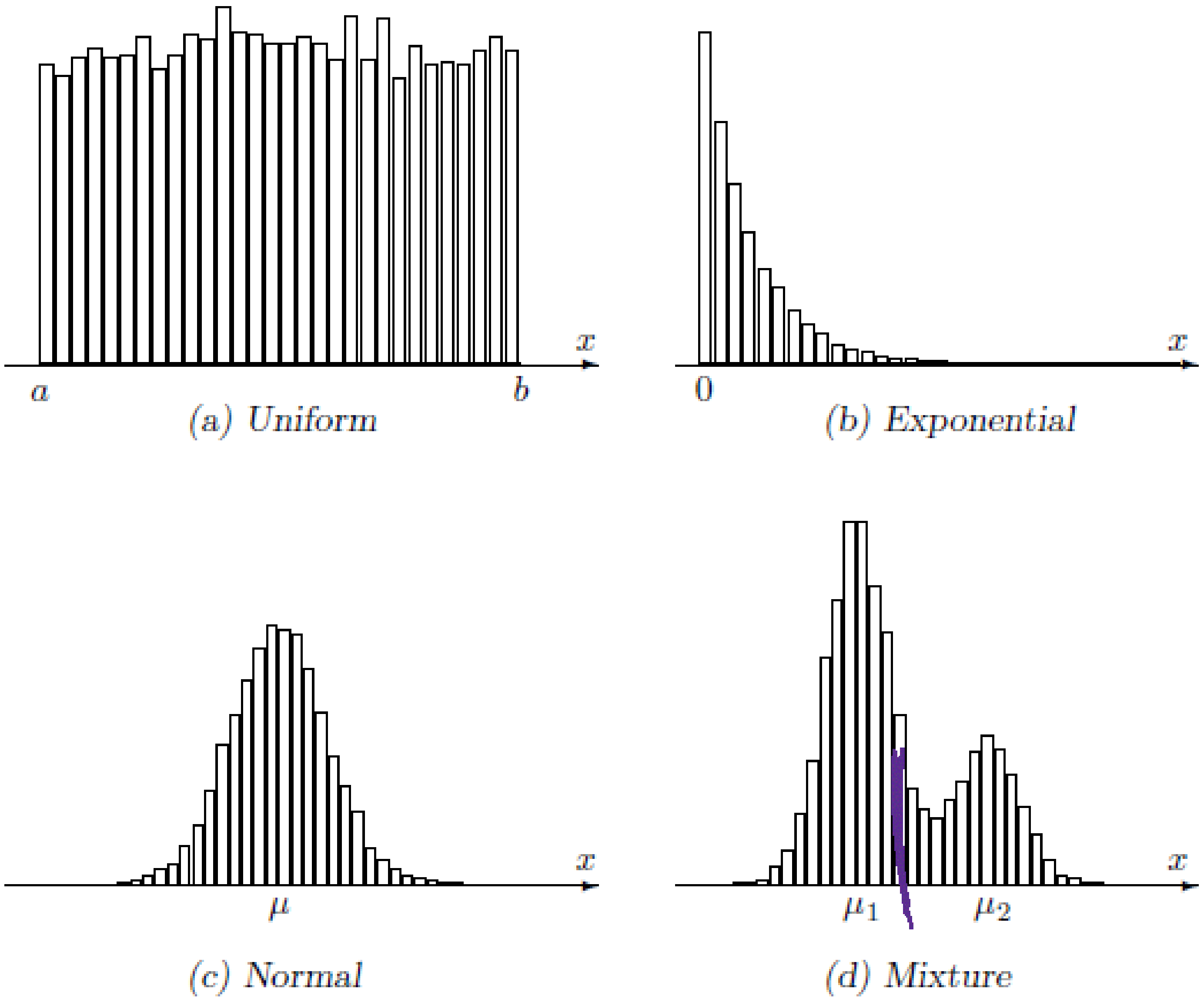
Visualizing a sample – histogram



(a) Frequency histogram

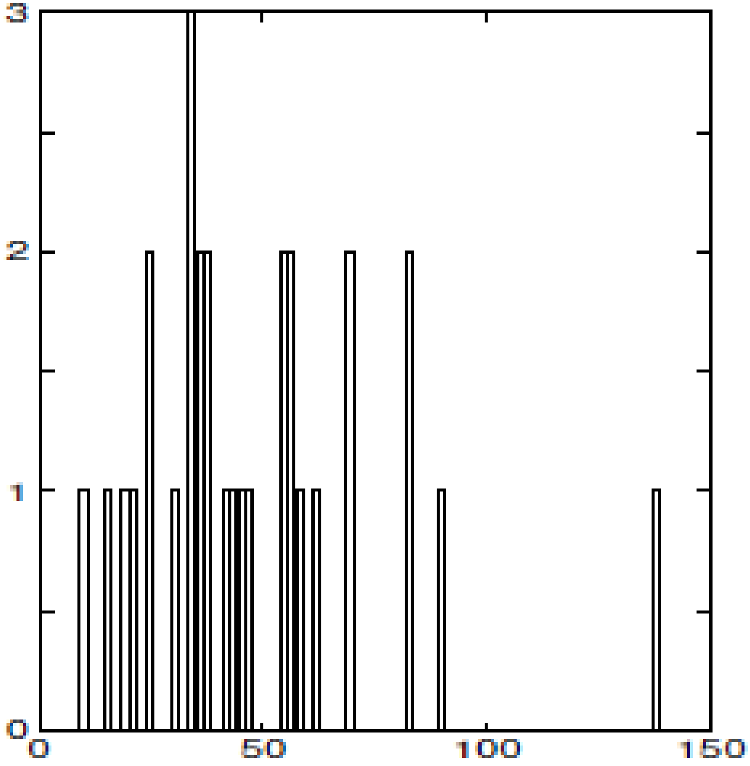
(b) Relative frequency histogram

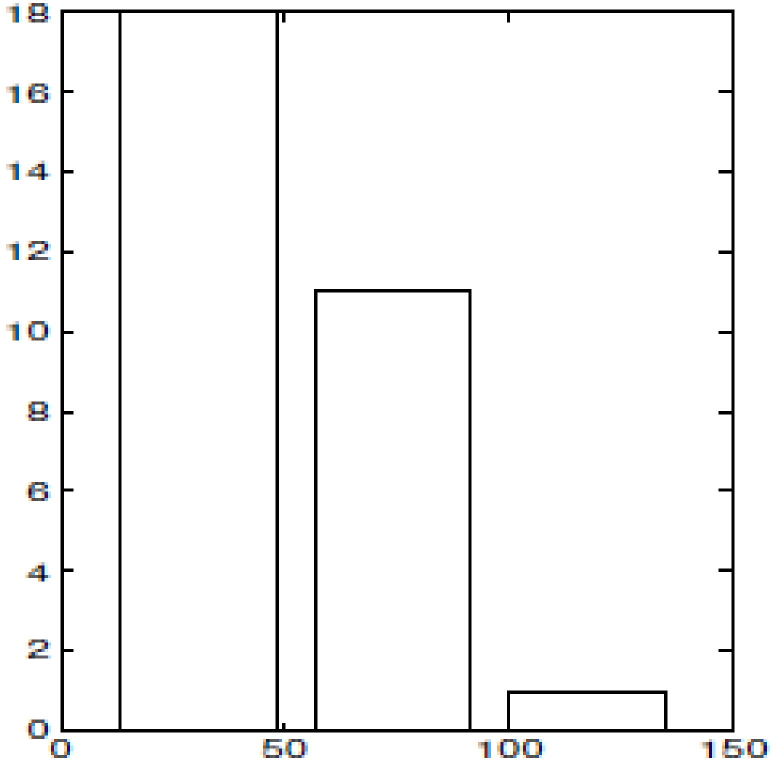
A few cases



x
a
b
(a) Uniform
x
0
(b) Exponential
x
μ
(c) Normal
x
μ_1
μ_2
(d) Mixture

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

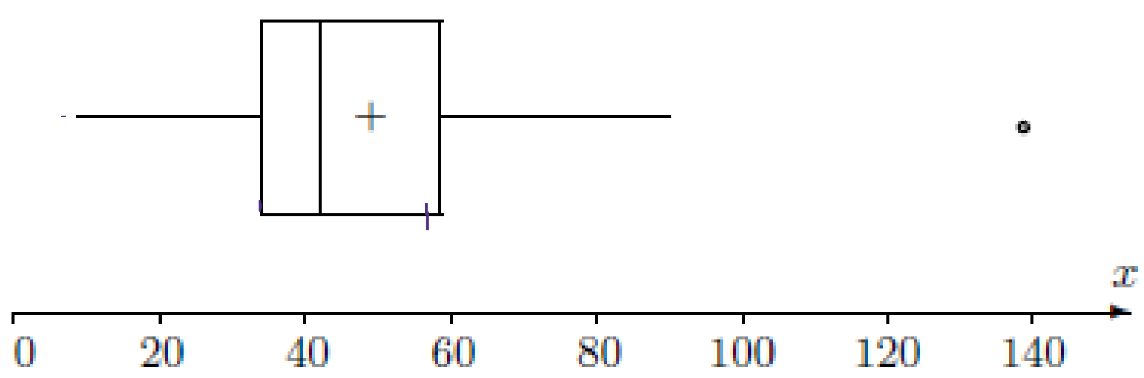
0	9
1	5 9
2	2 4 5
3	0 4 5 5 6 6 7 8
4	2 3 6 8
5	4 5 6 6 9
6	2 9
7	0
8	2 2 9
9	
10	
11	
12	
13	9

Stem+leaf for two samples

samples of Y							samples of X		
						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

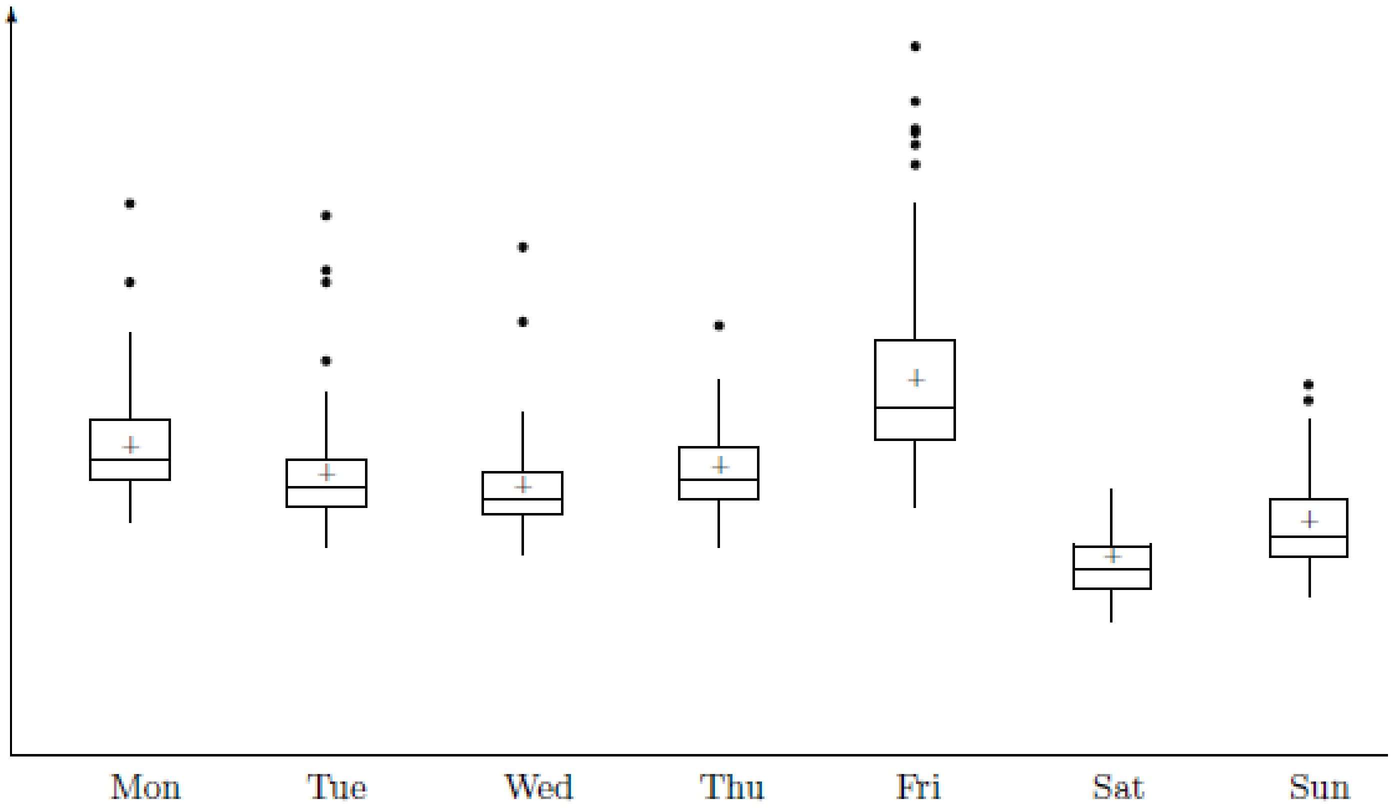


FIGURE 8.10: *Parallel boxplots of internet traffic.*