CENG 222
Statistical Methods for Computer Engineering

Week 8

Chapter 8

Introduction to Statistics

Outline

- Population and sample, parameters and statistics
- Simple descriptive statistics
- Graphical statistics

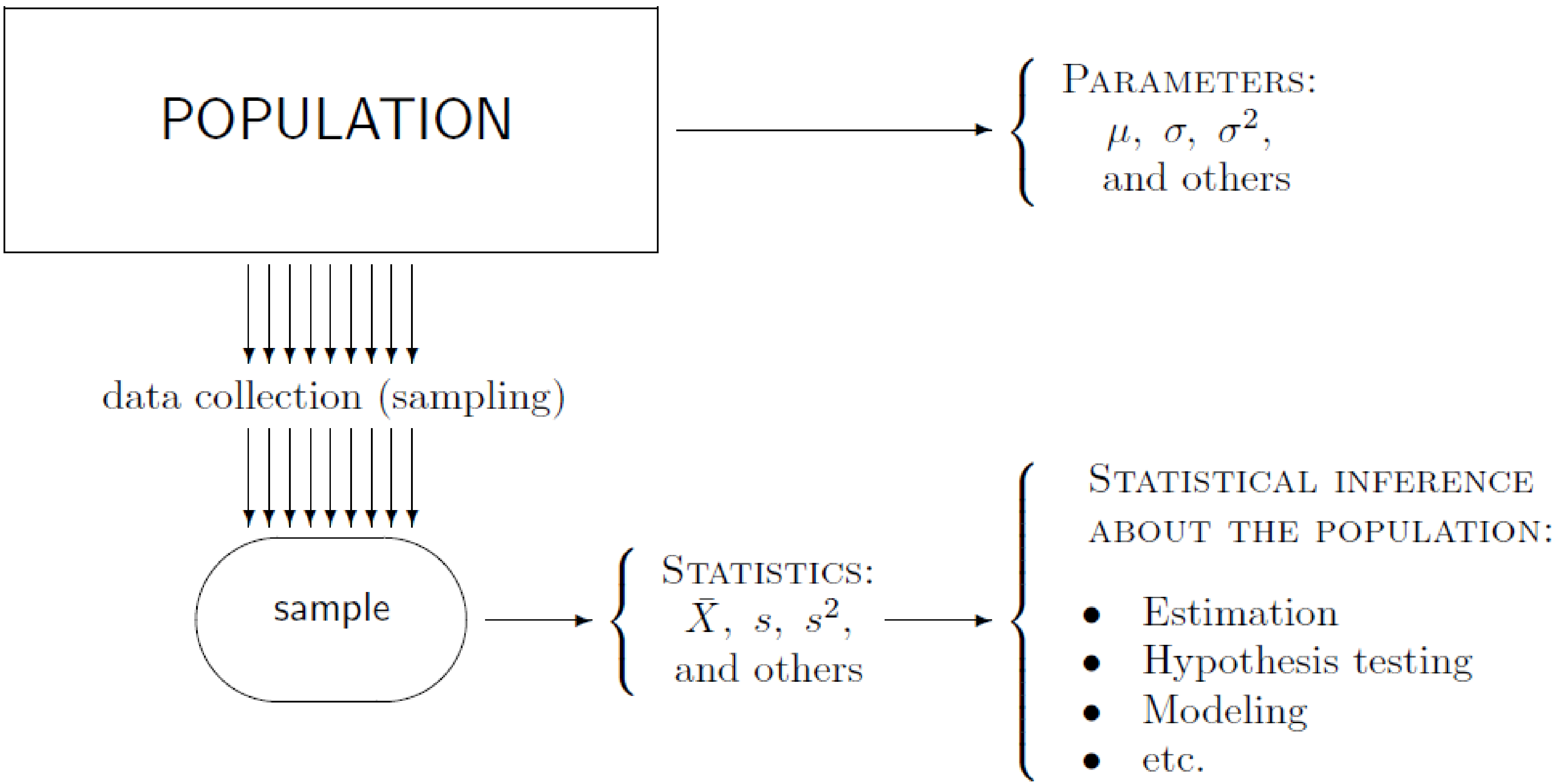
Statistics

- Focus on:
  - Data collection
  - Data analysis
    - Visualization
    - Estimation of distribution parameters
    - Finding correlations
    - Assessing the reliability of the estimates
    - Testing statements about the parameters

Terminology and Notation

- Population
  - Set of all possible sources of a random variable
- Parameter
  - Any numerical characteristic of a population
- Sample
  - A set of observed sources from the population
- Statistic
  - Any function of a sample
- θ: population parameter, $\hat{\theta}$: estimator of θ calculated using a sample

Population and Sample



Sampling

- Need to be careful when selecting samples from the population
  - Biases
  - Dependencies
- In general, any sample will be an approximation to the whole population; however, if sampling is done correctly, as the number of samples increases the approximation error should decrease.

Simple random sampling

- Data points are collected from the population independently of each other
- All data points are equally likely to be sampled
- iid: independent, identically distributed samples

Descriptive Statistics

- Mean
- Median
- Quantiles and quartiles
- Variance, standard deviation, and interquartile range
- Each statistic is a random variable, because different samples will result in different statistics
  - A statistic is a random variable with *sampling distribution*

Mean

- $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$
- Sample mean is unbiased, consistent, and asymptotically Normal.
- **Unbiasedness:** If the expectation of an estimator is equal to the estimated parameter, the estimator is called unbiased.
  - $\mathbf{E}(\hat{\theta}) = \theta$
  - $\text{Bias}(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)$

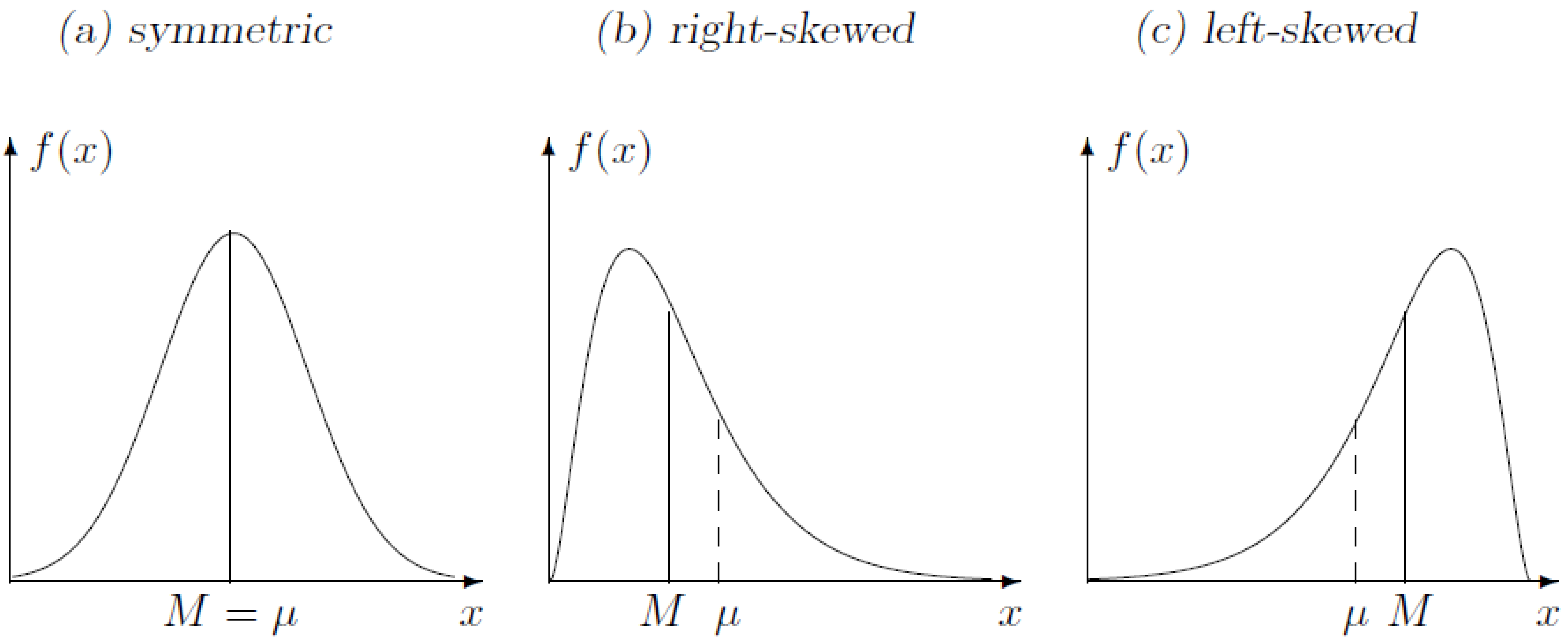
Consistency

- If the sampling error converges to 0 as the sample size increases, the estimator is called consistent
- $P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \to 0$ as $n \to \infty$

Median

- Sample mean is sensitive to "outliers".
  - Outlier: extreme observation
- Median is the "central" value
- Sample median $\widehat{M}$ 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 M is a number that is exceeded with probability no greater than 0.5 and is preceded with probability no greater than 0.5.

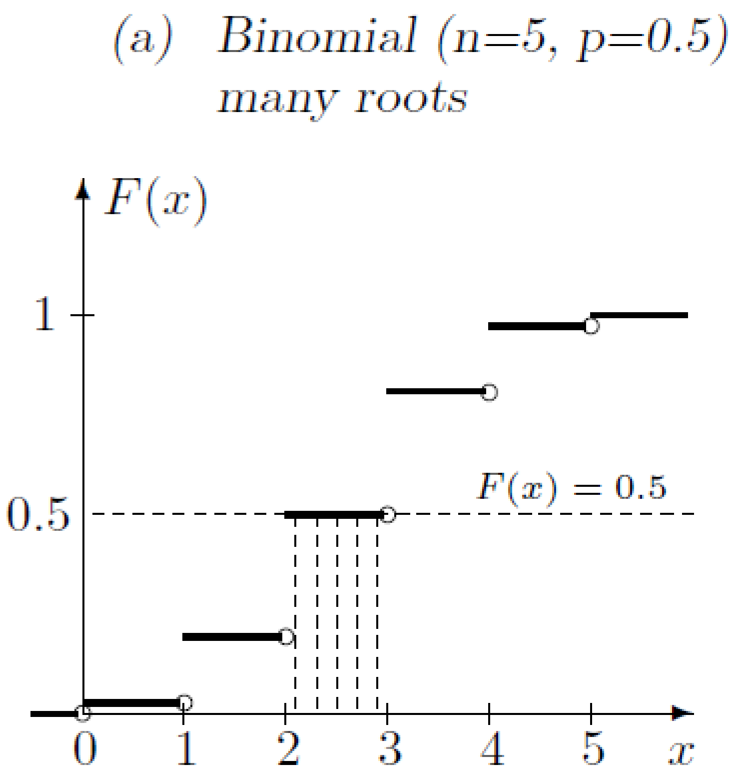
Mean vs. Median

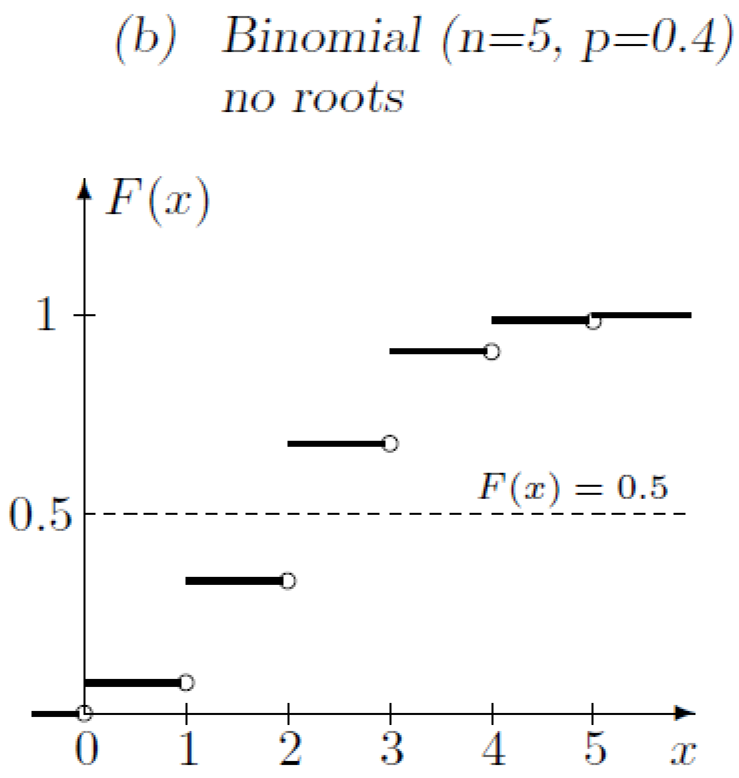


Population median

- Solve for $F(M) = 0.5$
- Example: exponential
- $F(M) = 1 - e^{-\lambda M} = 0.5$
- $\rightarrow M = \frac{\ln 2}{\lambda} = \frac{0.6931}{\lambda}$
- μ was 1/λ → larger than M → right skewed

Population median for discrete distributions





Sample median

- Just sort the samples
  - If n is odd, median is the unique middle element
  - If n is even, median is any point between the two middle elements

Quantiles, percentiles, quartiles

- Generalization of the notion of the median ($F(M)=0.5$) to arbitrary values
- p-quantile is a number x that satisfies $F(x)=p$
- q-percentile is $0.01q$-quantile
- First, second, and third quartiles are the 25^{th}, 50^{th}, and 75^{th} percentiles.
  - They split a population or a sample into 4 equal size parts.
- Median is the 0.5-quantile, the 50^{th}-percentile, and the 2^{nd} quartile.

Notation

$$
\begin{aligned}
q_p &= \text{population } p\text{-quantile} \\
\hat{q}_p &= \text{sample } p\text{-quantile, estimator of } q_p \\
\\
\pi_\gamma &= \text{population } \gamma\text{-percentile} \\
\hat{\pi}_\gamma &= \text{sample } \gamma\text{-percentile, estimator of } \pi_\gamma \\
\\
Q_1,\ Q_2,\ Q_3 &= \text{population quartiles} \\
\hat{Q}_1,\ \hat{Q}_2,\ \hat{Q}_3 &= \text{sample quartiles, estimators of } Q_1,\ Q_2,\ \text{ and } Q_3 \\
\\
M &= \text{population median} \\
\hat{M} &= \text{sample median, estimator of } M
\end{aligned}
$$

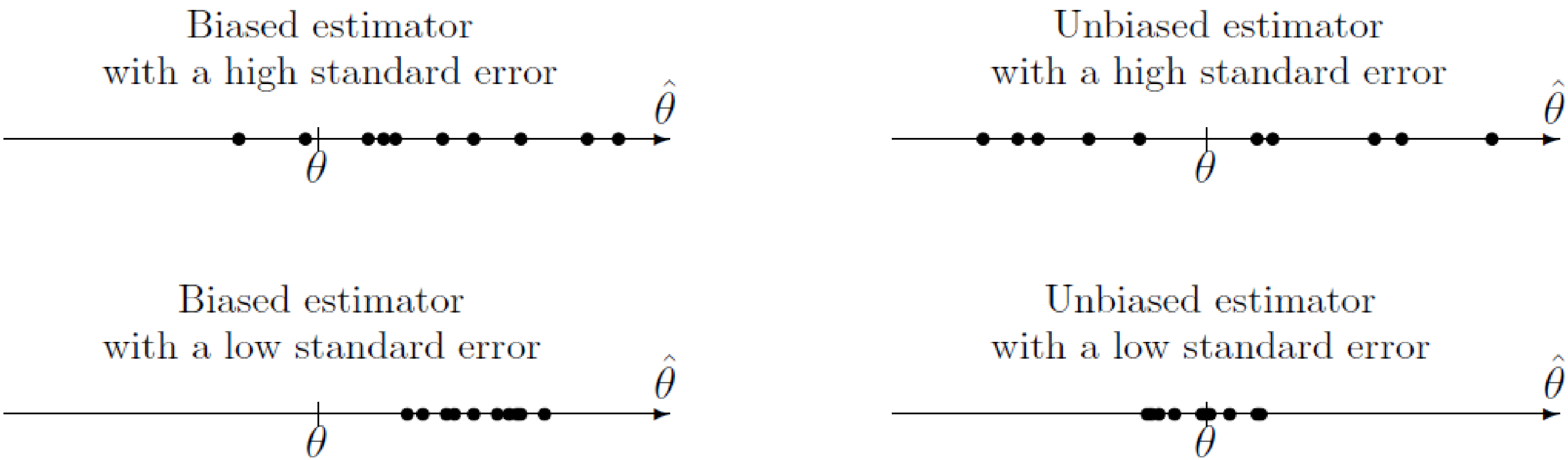
Example 8.15

- Deciding on warranty duration for computer with lifetimes that follow a Gamma distribution with $\alpha=60$ and $\lambda=5$ years^{-1}.
  - The company wants to ensure that only 10% of the customers use the warranty

Sample variance

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- $1/n$-1 needed for an unbiased estimator
- This estimator is also consistent and asymptotically Normal

Standard errors of estimates



Outliers and Interquartile Range

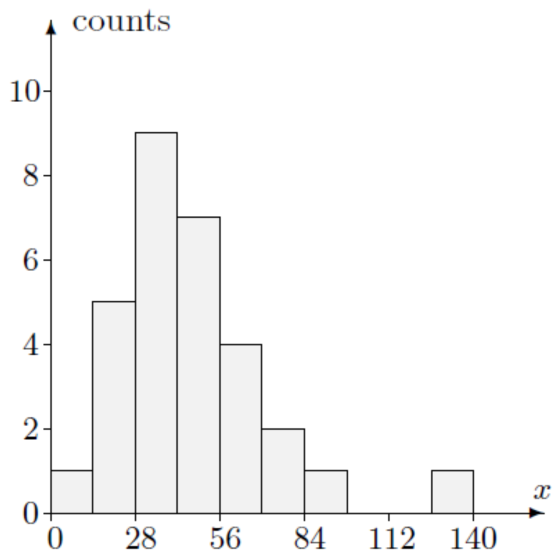
- Q_3-Q_1 is called the interquartile range, IQR.
- Usually, data that lie below 1.5IQR below Q_1 and data that lie above 1.5IQR above Q_3 are called outliers

Graphical statistics

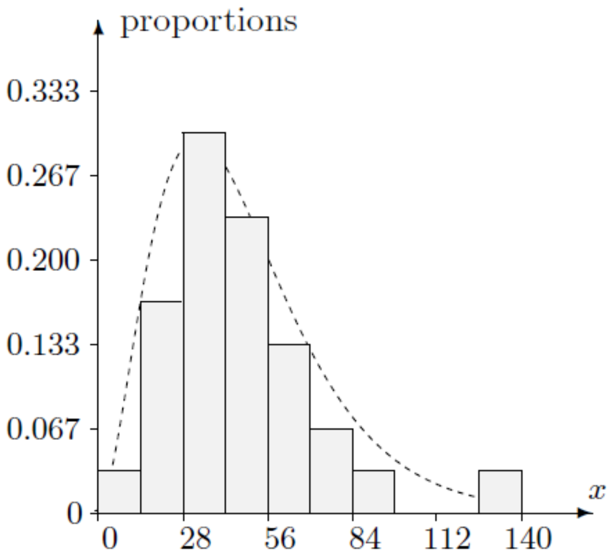
- Histograms
- Stem-and-leaf plots
- Box plots
- Scatter plots
- Time plots

Histograms

- Shows the shape of the pmf or pdf
- Split range of data into equal "bins" and count how many observations fall into each bin.

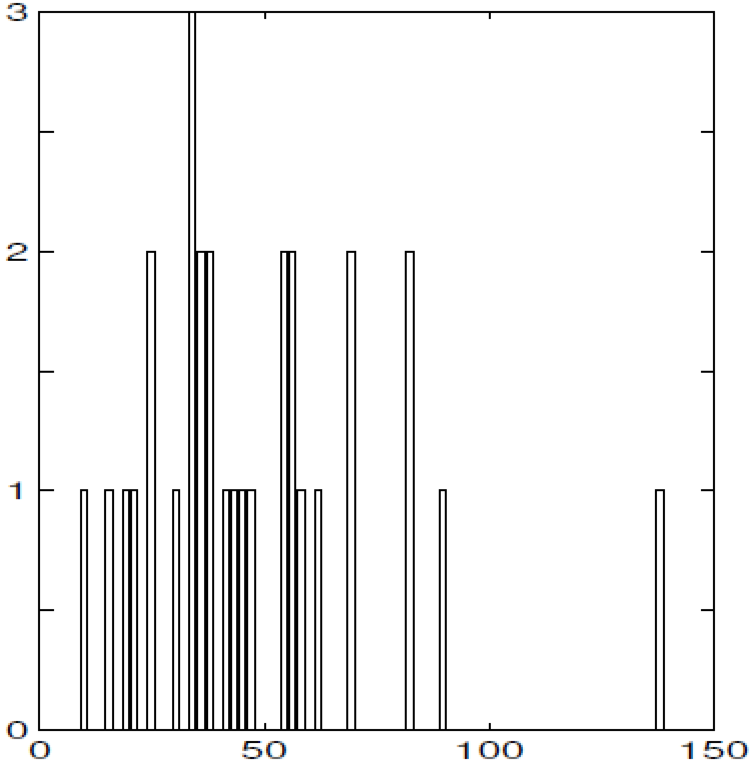


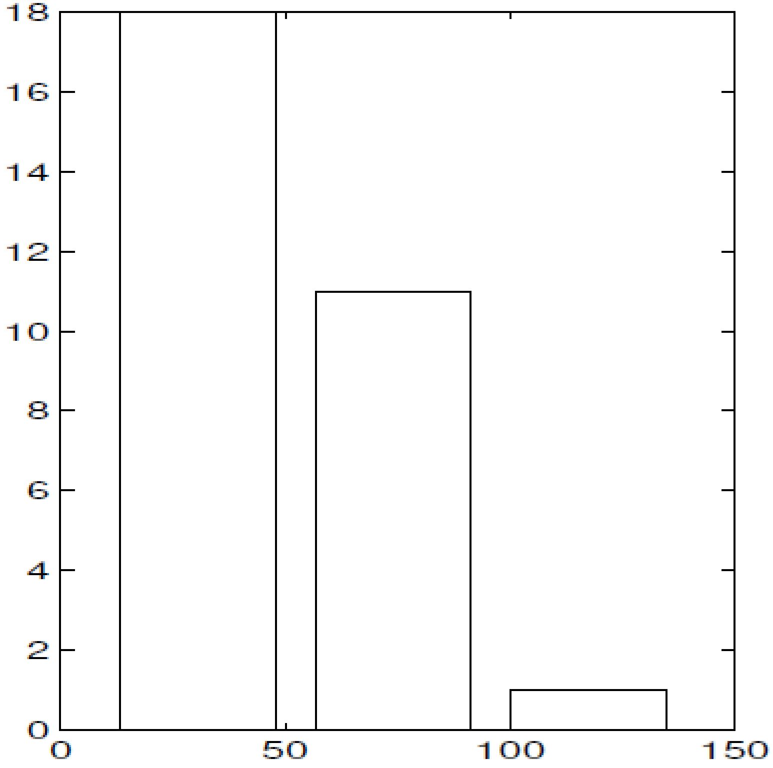
(a) Frequency histogram



(b) Relative frequency histogram

Non-appropriate bin sizes





Stem-and-leaf plots

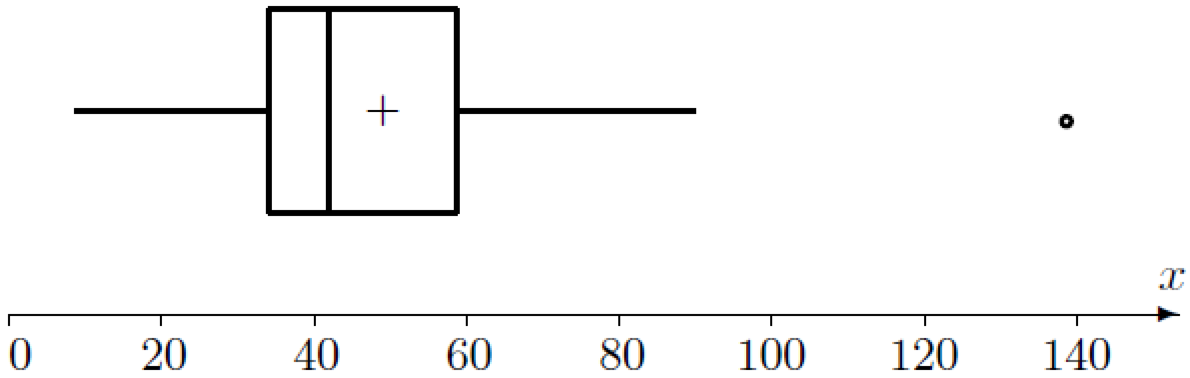
- Similar to histograms but also show the distribution within a column

LEAF UNIT = 1

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

Boxplot

- A box is drawn between the first and third quartiles. Median is shown within the box. Smallest and largest observations (excluding outliers) are shown outside the box as extended whiskers



Parallel Boxplots