

Bus 273: Statistical Analysis For Business

Harald Schmidbauer



About These Slides

- The present slides are not self-contained; they need to be explained and discussed. This will be done in the lectures.
- Even though being a “work in progress” and subject to revision, the slides constitute copyrighted material.
If you want to reproduce or copy anything from the slides, please ask:

Harald Schmidbauer **harald** at **hs-stat** dot **com**
Angi Rösch **angi** at **angi-stat** dot **com**

- The slides were produced using \LaTeX and R (the R project; website: www.R-project.org) on a GNU/Linux system.
- R files used for this course are available upon request.



Chapter 3: Displaying Univariate Data



3.1 Frequency Distributions

The notion of frequency distribution.

- So far, we were concerned with the structure of data.
- To obtain insight into a heap of data, we need to look at the frequency distribution of the variable in question.
- The question “How often is each value taken on?” leads to the notion of frequency distribution.



3.1 Frequency Distributions

Frequencies.

Let our observations of a variable X be given as

$$x_1, x_2, \dots, x_n.$$

Further, let

$$a_1, a_2, \dots, a_k$$

be the values which appear among the observations.

Then:

$$h(a_j) = \# \text{ observations equal to } a_j: \quad \textbf{absolute frequency of } a_j$$

$$f(a_j) = h(a_j)/n: \quad \textbf{relative frequency of } a_j$$

A list of the a_j , together with their frequencies, is called the (empirical) distribution of X .



3.1 Frequency Distributions

Example:

Observations of $X = \textit{gender}$ from Example 1:

m, m, m, m, f, f, f, f, f, f, m, m, m, m, m, m, f, f, f, f, f, f, f, m, f,
m, f, f, m, m, f, m, f, m, m, m, f, m, m, f, f, f, f, f, f, m, m, m, f, f,
f, m, m

Distribution of X among the 54 students:

a_j	h_j	f_j
$a_1 = f$	29	29/54
$a_2 = m$	25	25/54
Σ	54	1



3.1 Frequency Distributions

Which graphical display is appropriate?

This depends on the scaling of the variable. Among others:

- categorical variable: pie chart, bar chart
- discrete metric variable: bar chart
- continuous metric variable: stemplot, histogram



3.1 Frequency Distributions

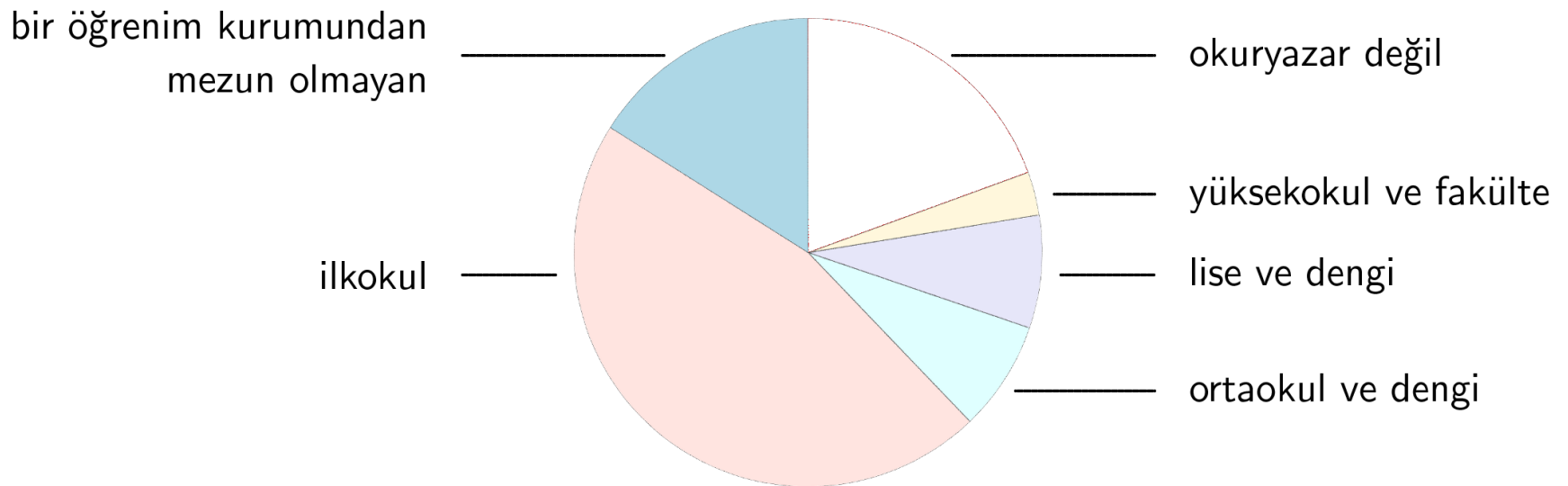
Example: Educational attainment in Turkey, 1990.

Category	h_i	f_i
1: okuryazar değil	9.56	0.195
2: bir öğrenim kurumundan mezun olmayan	7.84	0.160
3: ilkokul	22.68	0.462
4: ortaokul ve dengi	3.72	0.076
5: lise ve dengi	3.82	0.078
6: yüksekokul ve fakülte	1.50	0.030
Σ	49.14	1.000



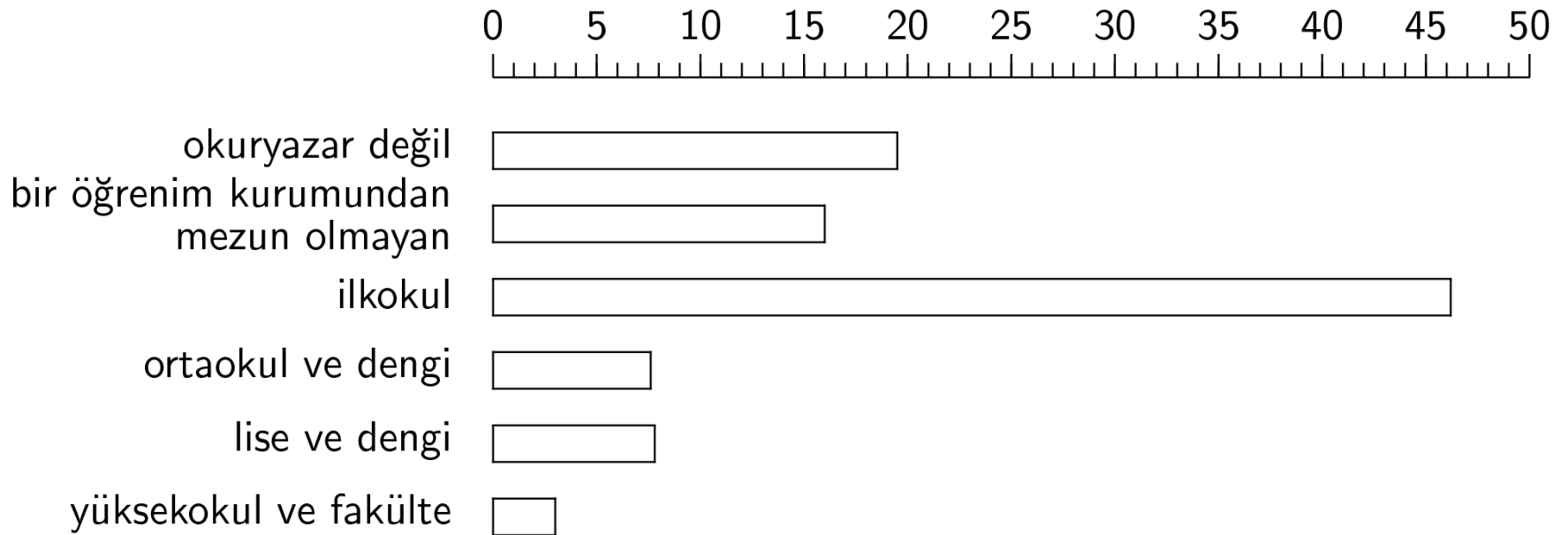
3.1 Frequency Distributions

Example: Educational attainment in Turkey, 1990.



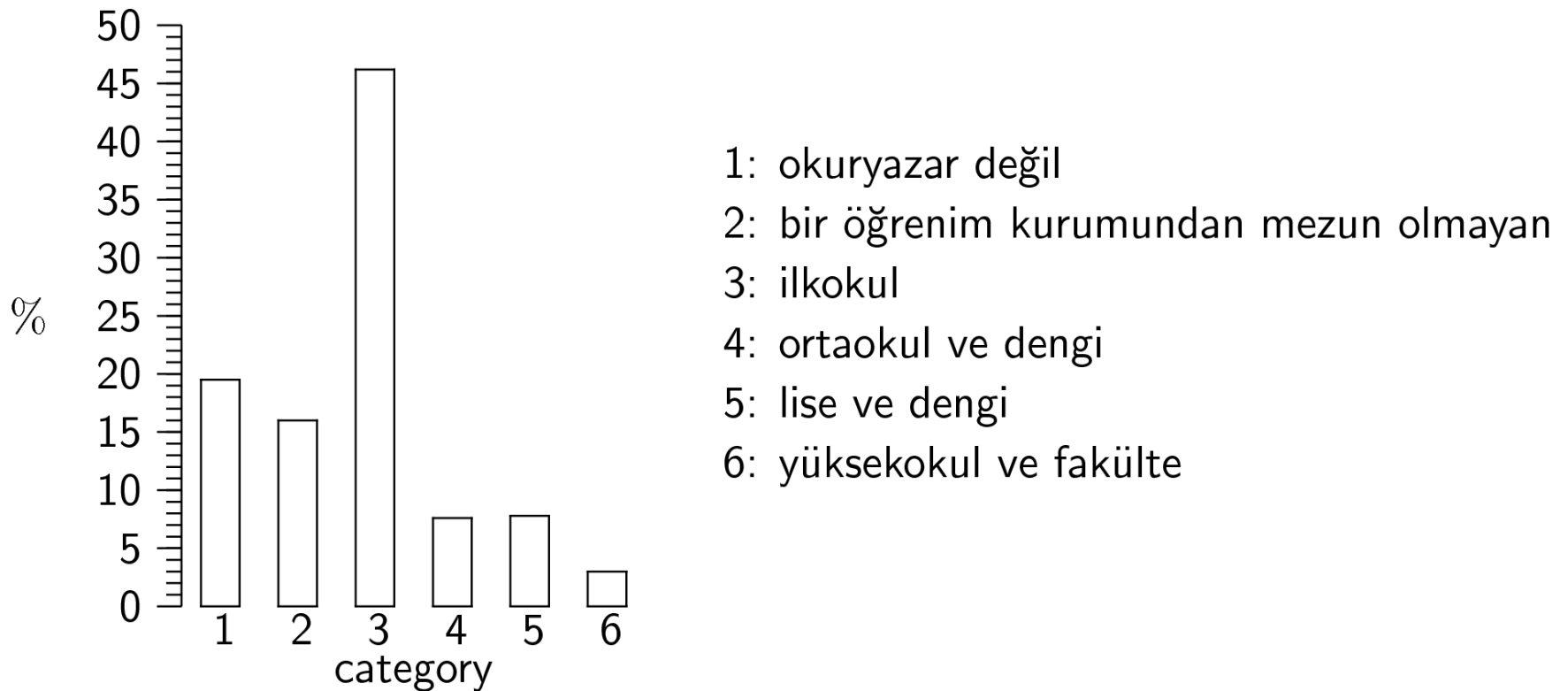
3.1 Frequency Distributions

Example: Educational attainment in Turkey, 1990.



3.1 Frequency Distributions

Example: Educational attainment in Turkey, 1990.



3.1 Frequency Distributions

Example: The number of goals scored in 170 matches of Beşiktaş İstanbul.

Raw data:

4, 8, 2, 3, 2, 3, 4, 2, 2, 1, 2, 5, 5, 4, 0, 2, 6, 2, 2, 3, 1, 4, 4, 5, 4, 4, 10, 3,
3, 1, 3, 2, 0, 0, 2, 7, 6, 2, 1, 2, 5, 2, 4, 1, 0, 3, 4, 3, 3, 5, 4, 4, 0, 3, 0, 1, 7,
4, 3, 3, 5, 6, 4, 2, 7, 5, 4, 2, 3, 4, 3, 0, 3, 3, 3, 1, 0, 5, 1, 3, 2, 8, 4, 6, 3, 2,
4, 2, 1, 4, 1, 5, 5, 3, 1, 3, 2, 5, 4, 2, 1, 0, 5, 3, 2, 1, 6, 2, 3, 4, 5, 1, 2, 3,
2, 2, 2, 2, 0, 1, 3, 2, 2, 3, 3, 2, 3, 3, 2, 1, 0, 3, 2, 3, 3, 4, 1, 4, 2, 6, 1, 4,
3, 0, 2, 5, 1, 1, 4, 1, 3, 3, 2, 4, 2, 2, 3, 3, 3, 2, 1, 3, 4, 4, 2, 4, 6, 6, 4, 5

What can we do with this dataset?



3.1 Frequency Distributions

Example: The number of goals scored in 170 matches of Beşiktaş İstanbul.

The empirical distribution is:

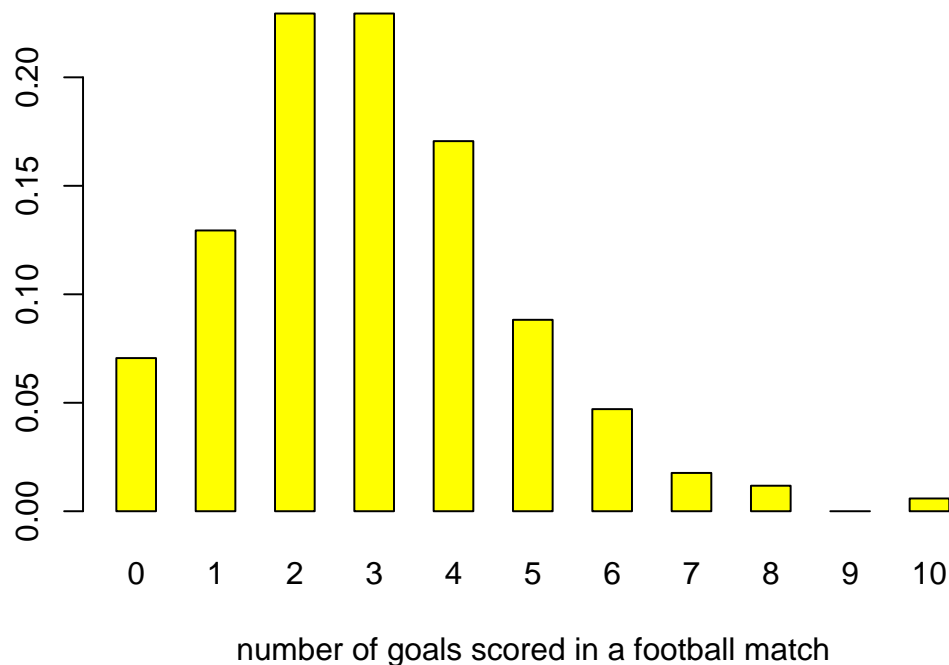
i	0	1	2	3	4	5	6	7	8	9	10
h_i	12	22	39	39	29	15	8	3	2	0	1
f_i	$\frac{12}{170}$	$\frac{22}{170}$	$\frac{39}{170}$	$\frac{39}{170}$	$\frac{29}{170}$	$\frac{15}{170}$	$\frac{8}{170}$	$\frac{3}{170}$	$\frac{2}{170}$	$\frac{0}{170}$	$\frac{1}{170}$



3.1 Frequency Distributions

Example: The number of goals scored in 170 matches of Beşiktaş İstanbul.

A bar chart of the distribution:



3.1 Frequency Distributions

Example: The number of goals scored in 170 matches of Beşiktaş İstanbul.

How can we compute the average number \bar{x} of goals per match?

- Use the observations x_1, \dots, x_{170} themselves:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{170} (4 + 8 + \dots + 5) = 2.96$$

- Use the distribution of the observations:

$$\bar{x} = \sum_i i \cdot f_i = 0 \cdot \frac{12}{170} + 1 \cdot \frac{22}{170} + \dots + 10 \cdot \frac{1}{170} = 2.96$$



3.2 Stemplots and Histograms

Pie charts and bar charts are not suitable to display the distribution of a continuous metric variable, such as:

X = total expenditure of a customer at a supermarket
(when shopping once),

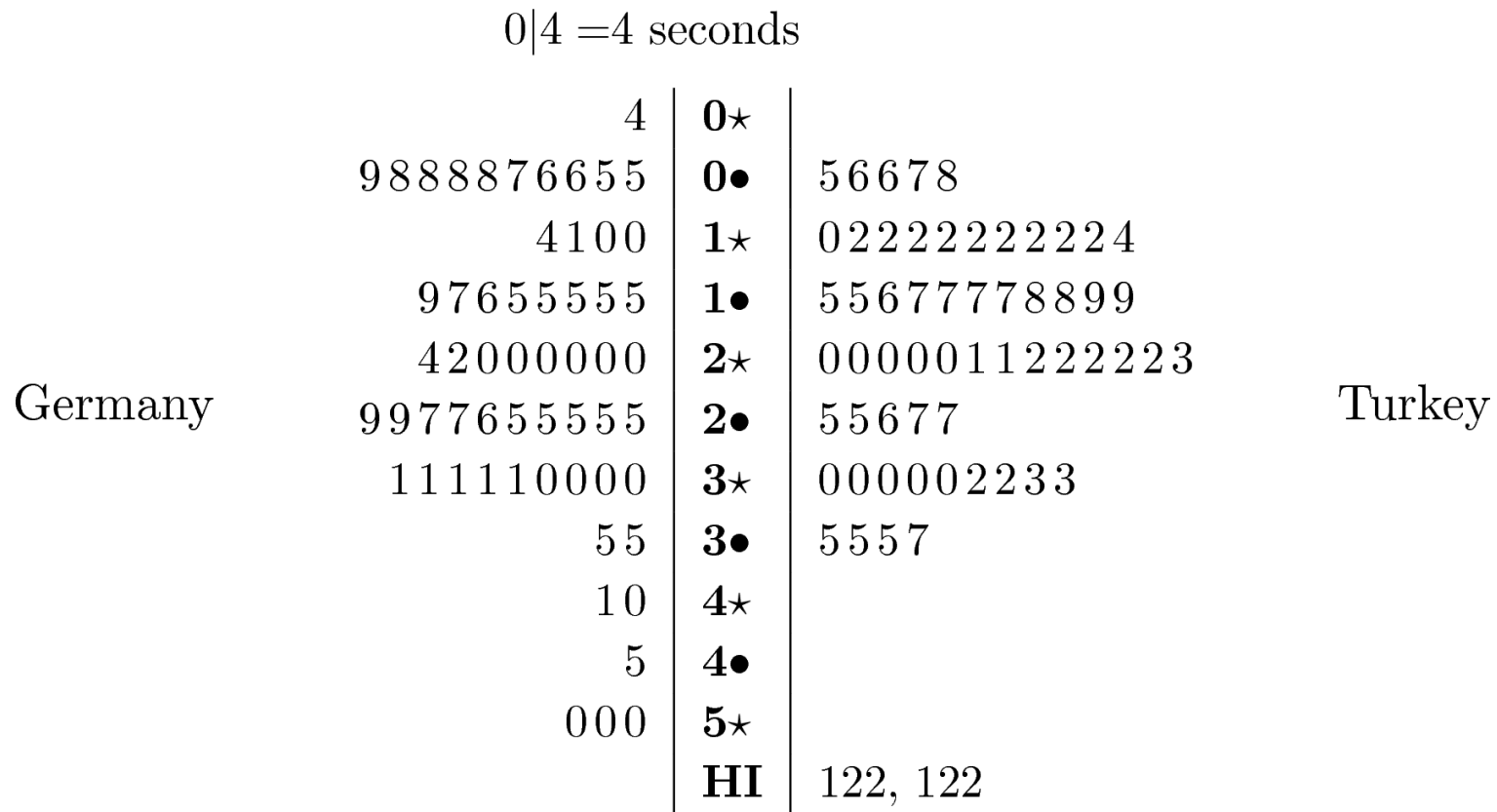
X = duration of a TV commercial.

For this, we need a stemplot (if the number of observations is not too big) or a histogram.



3.2 Stemplots and Histograms

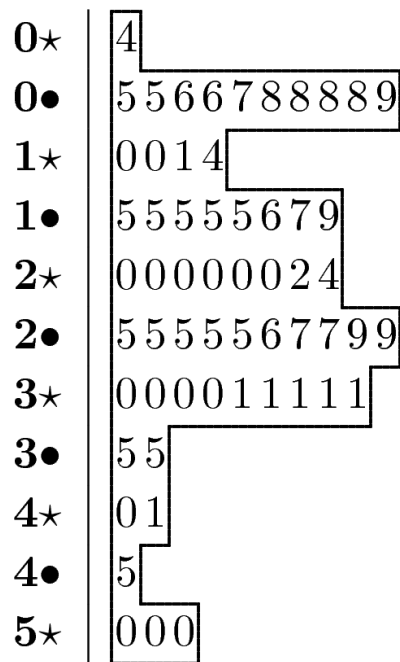
Example: Duration of a TV commercial.



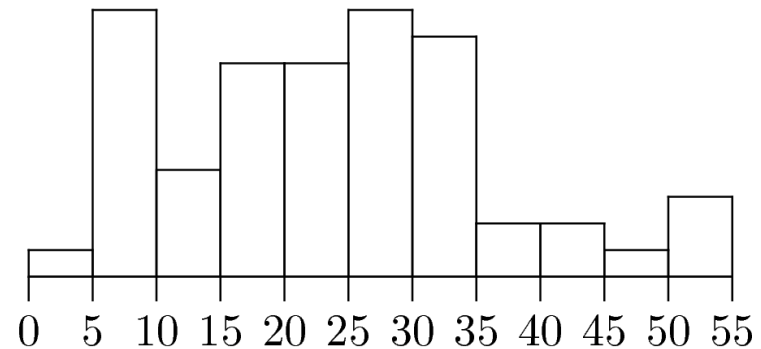
3.2 Stemplots and Histograms

Example: Duration of a TV commercial.

How to obtain a histogram.



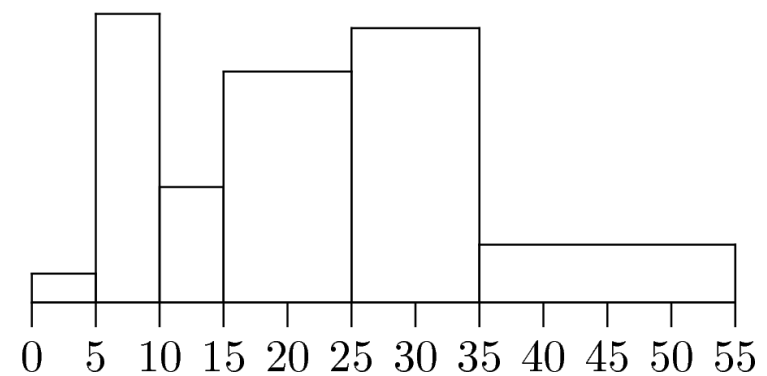
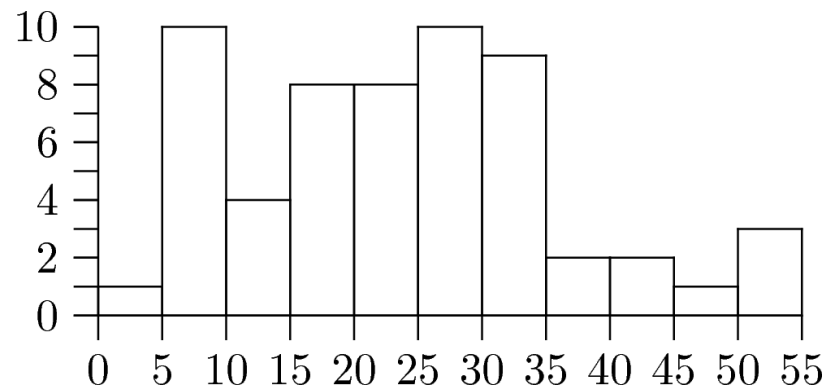
turn it →



3.2 Stemplots and Histograms

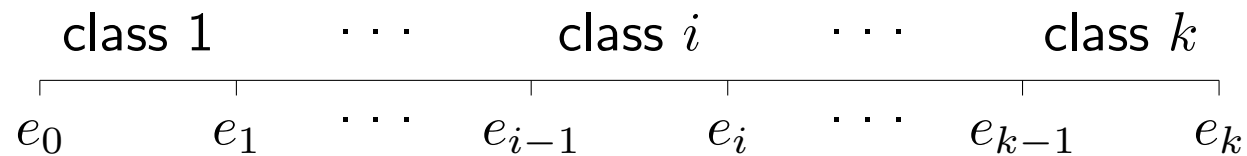
Example: Duration of a TV commercial.

Two versions of a histogram — detailed and less detailed.



3.2 Stemplots and Histograms

Construction of a histogram.



number of classes: k

for each class i ($i = 1, \dots, k$):

class limit: lower: e_{i-1}

 upper: e_i

class width: $d_i = e_i - e_{i-1}$

A **histogram** consists of rectangles over each class with **area** proportional to the number of observations in each class.



3.2 Stemplots and Histograms

Construction of a histogram.

Let:

- h_i = number of observations in class i
- H_i = height of rectangle over class i

Then:

$$H_i \cdot d_i \propto h_i, \quad \text{or:} \quad H_i = \alpha \cdot \frac{h_i}{d_i}$$



3.2 Stemplots and Histograms

Example: Total expenditure of customers in a supermarket.

The total expenditure of 508 customers (in euros) was recorded.

Raw data: 10.07, 22.61, 14.48, . . . , 28.68

Ordered raw data: 0.59, 0.72, 0.74, . . . , 75.54

How do we get a first insight into this dataset?



3.2 Stemplots and Histograms

Example: Total expenditure of customers in a supermarket.

A (not very detailed) histogram can be obtained like this:

i	interval	h_i	d_i	$\alpha \cdot h_i/d_i$
1	[0, 10)	216	10	21.60α
2	[10, 30)	233	20	11.65α
3	[30, 80]	59	50	1.18α

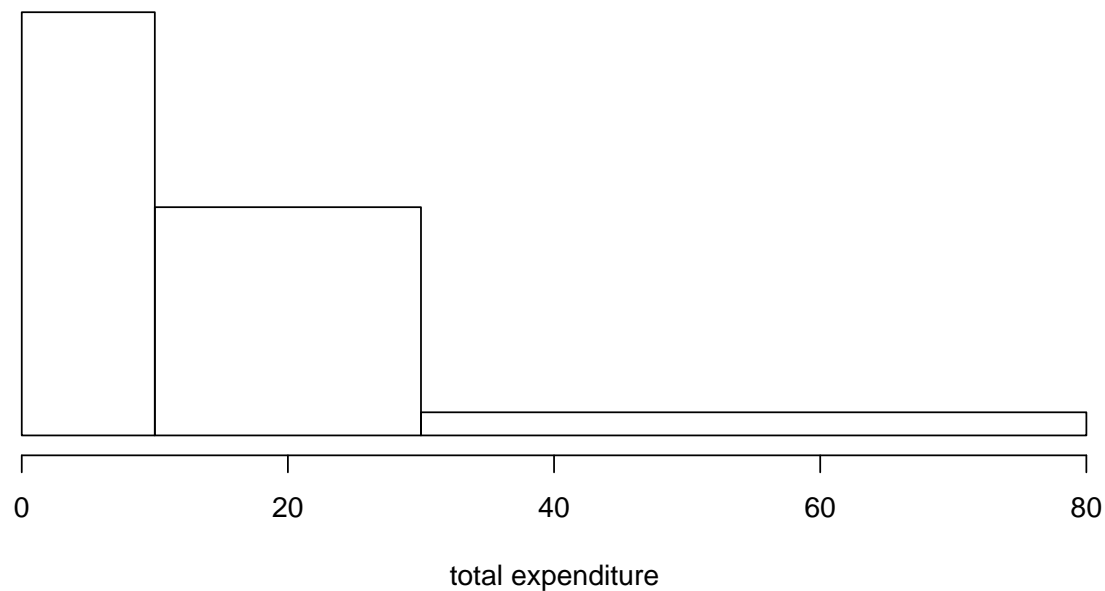
For a nice picture on an A4 page, choose $\alpha = 0.3\text{cm}$ (for example).



3.2 Stemplots and Histograms

Example: Total expenditure of customers in a supermarket.

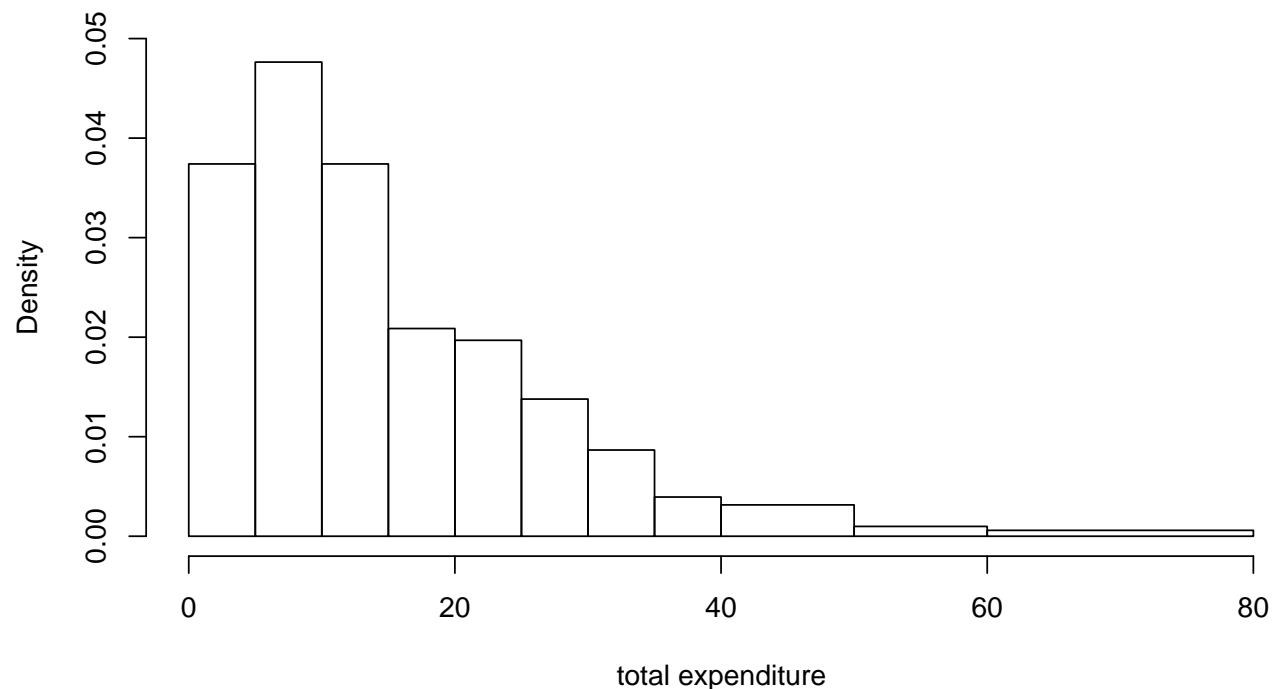
The resulting histogram is:



3.2 Stemplots and Histograms

Example: Total expenditure of customers in a supermarket.

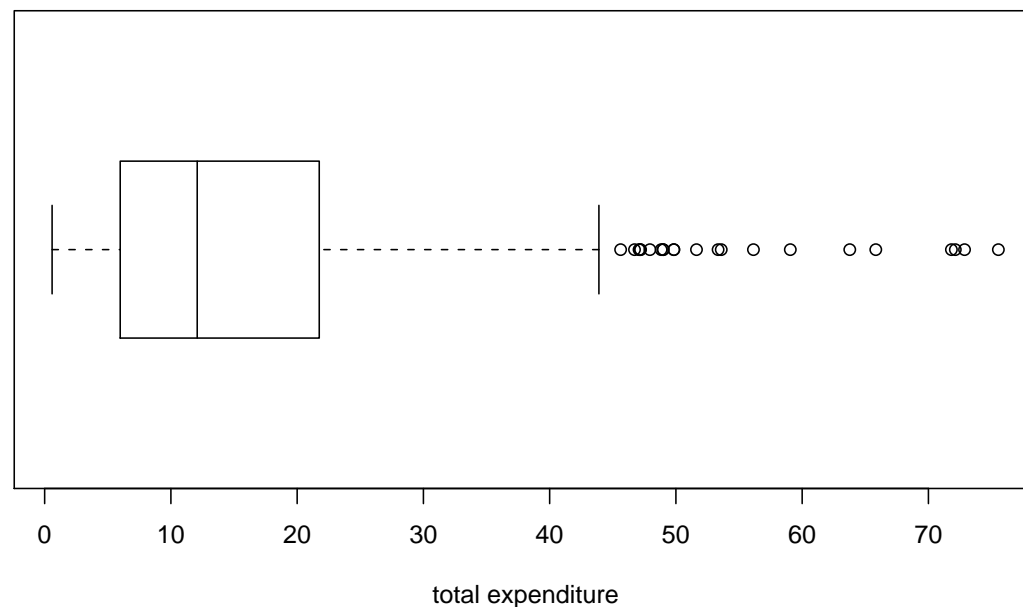
A more detailed histogram is:



3.2 Stemplots and Histograms

Example: Total expenditure of customers in a supermarket.

A boxplot is another elegant way to display the distribution:



3.3 The Shape of a Distribution

The shape of a distribution.

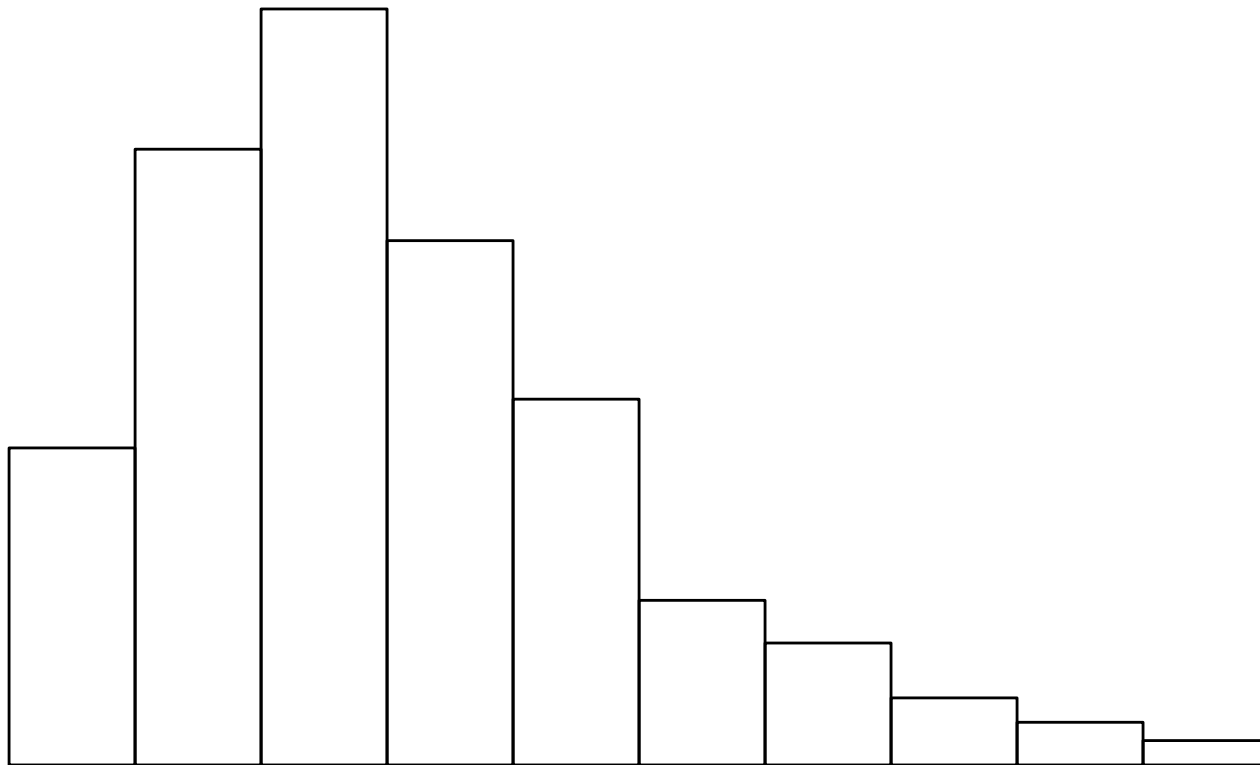
The shape of a distribution is often an interesting clue. We distinguish distributions with respect to:

- skewness
- kurtosis



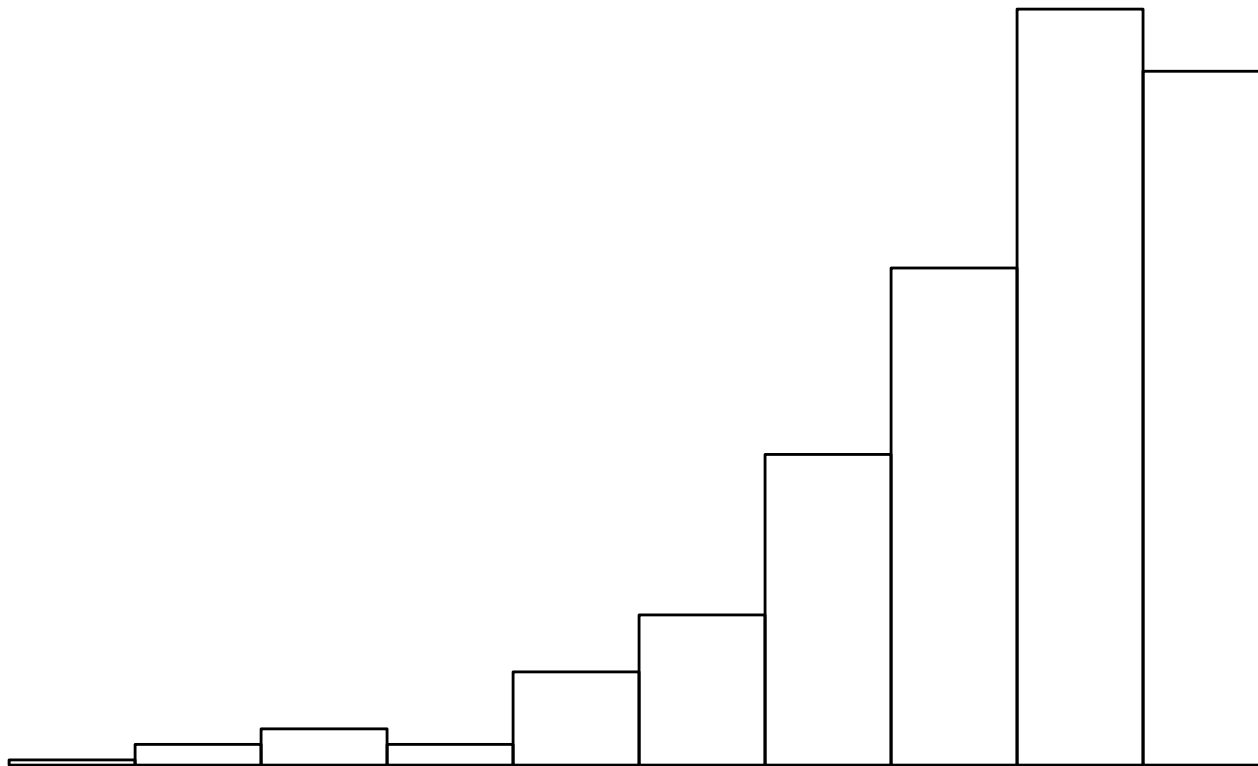
3.3 The Shape of a Distribution

A right-skewed distribution:



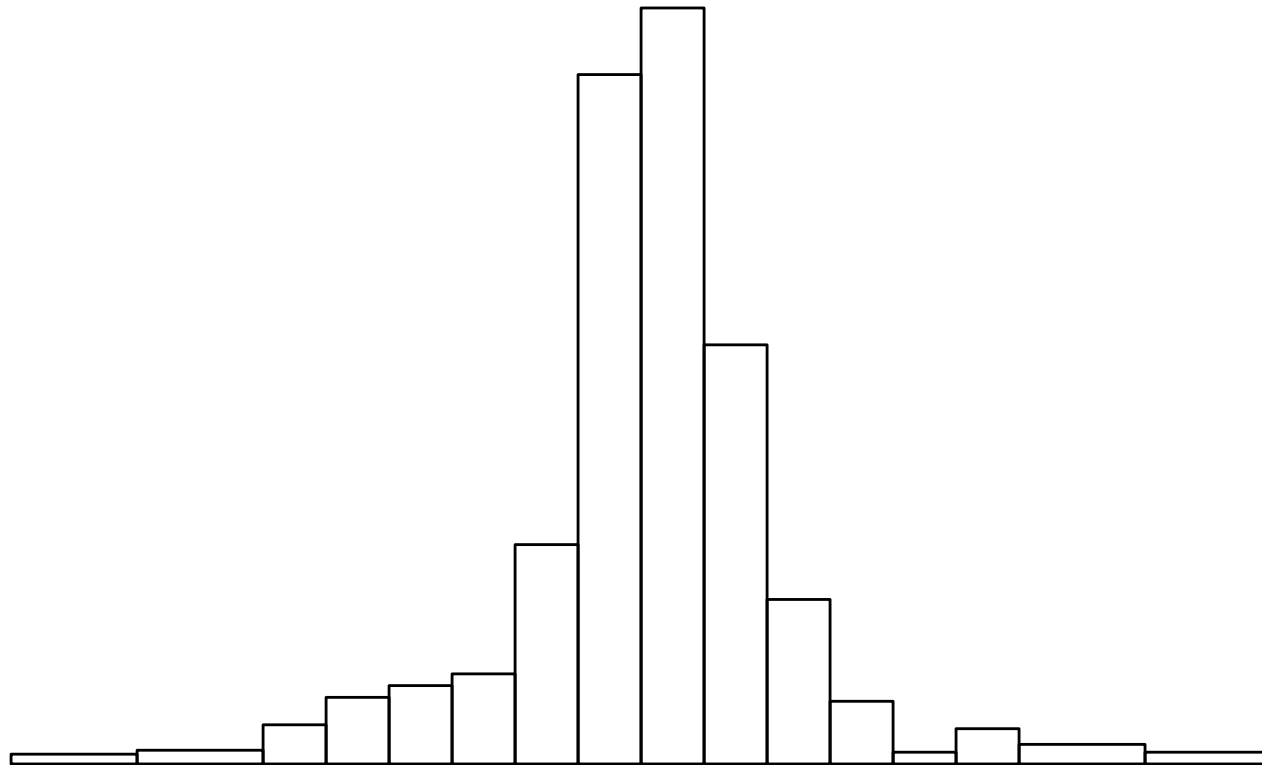
3.3 The Shape of a Distribution

A left-skewed distribution:



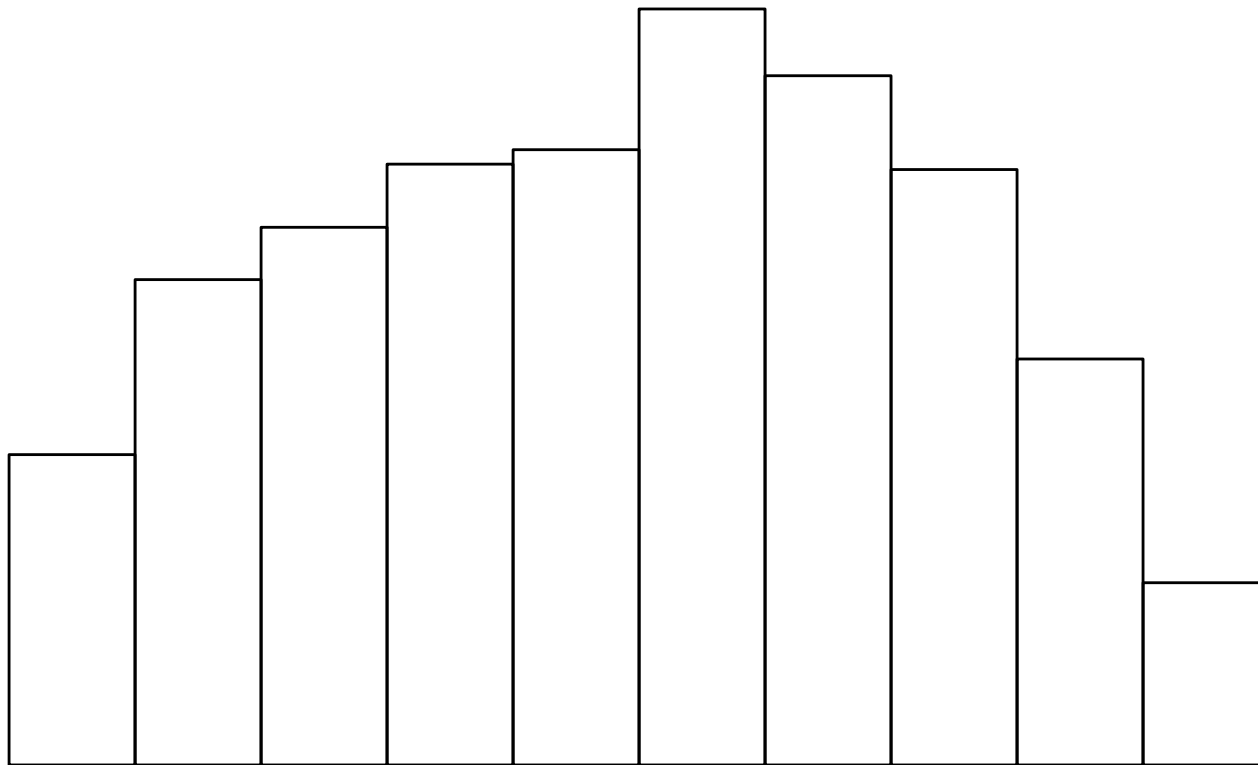
3.3 The Shape of a Distribution

A leptokurtic distribution:



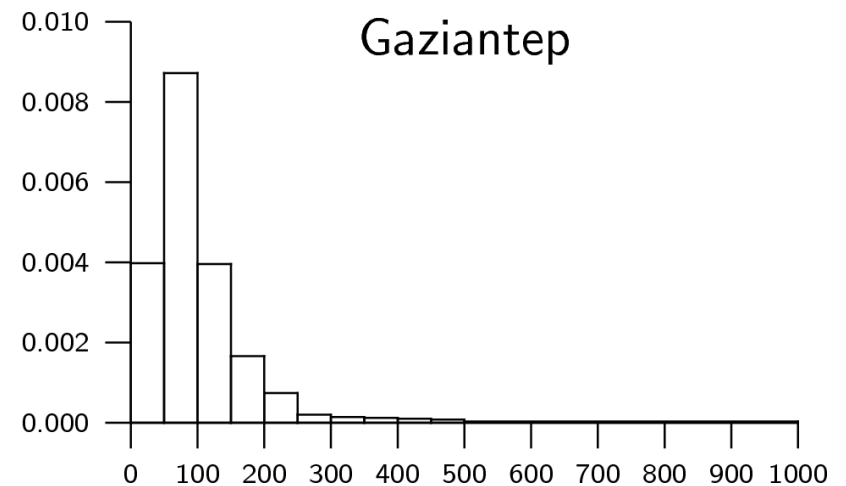
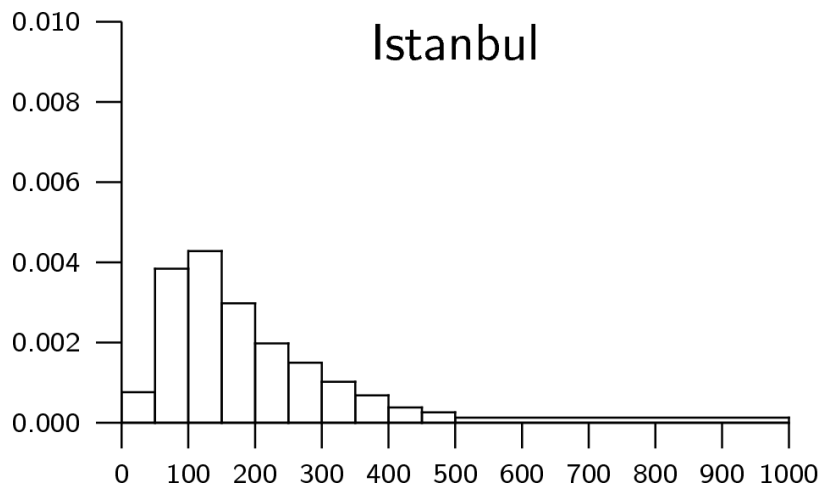
3.3 The Shape of a Distribution

A platy-kurtic distribution:



3.3 The Shape of a Distribution

Example: Household income. (1994; mill. TL)



3.3 The Shape of a Distribution

Example: Monthly, weekly, daily returns on the Dow-Jones Industrial Average, 1995-01 through 2005-10.

