Bus 273: Statistical Analysis For Business

Harald Schmidbauer

İSTANBUL BİLGİ ÜNİVERSİTESİ

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 4:

Location, Variation, and Shape of a Distribution

4.1 Location and Variation: Two Aspects of a Distribution

Two basic properties of the distribution of a metric variable:

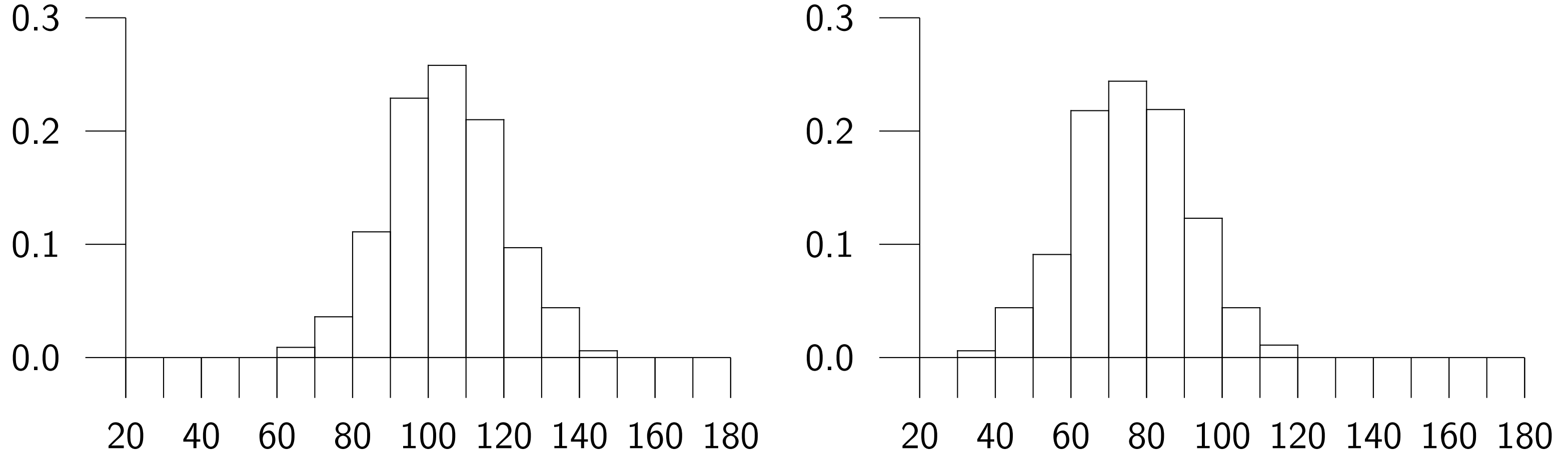
- its location:
  Where are the observations (the data)?

- its variation or dispersion:
  How scattered are the observations (the data)?

The main goal of the present chapter is to show how these can be **measured**.

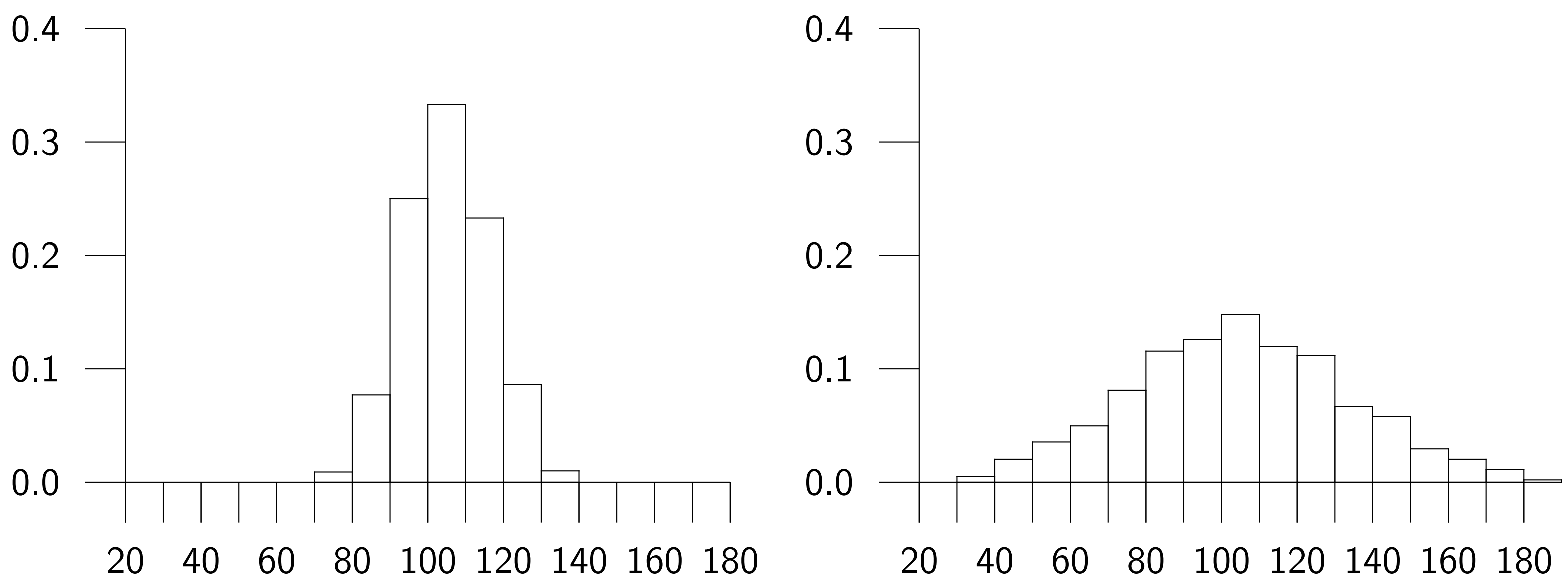
4.1 Location and Variation: Two Aspects of a Distribution

Two distributions with different locations:



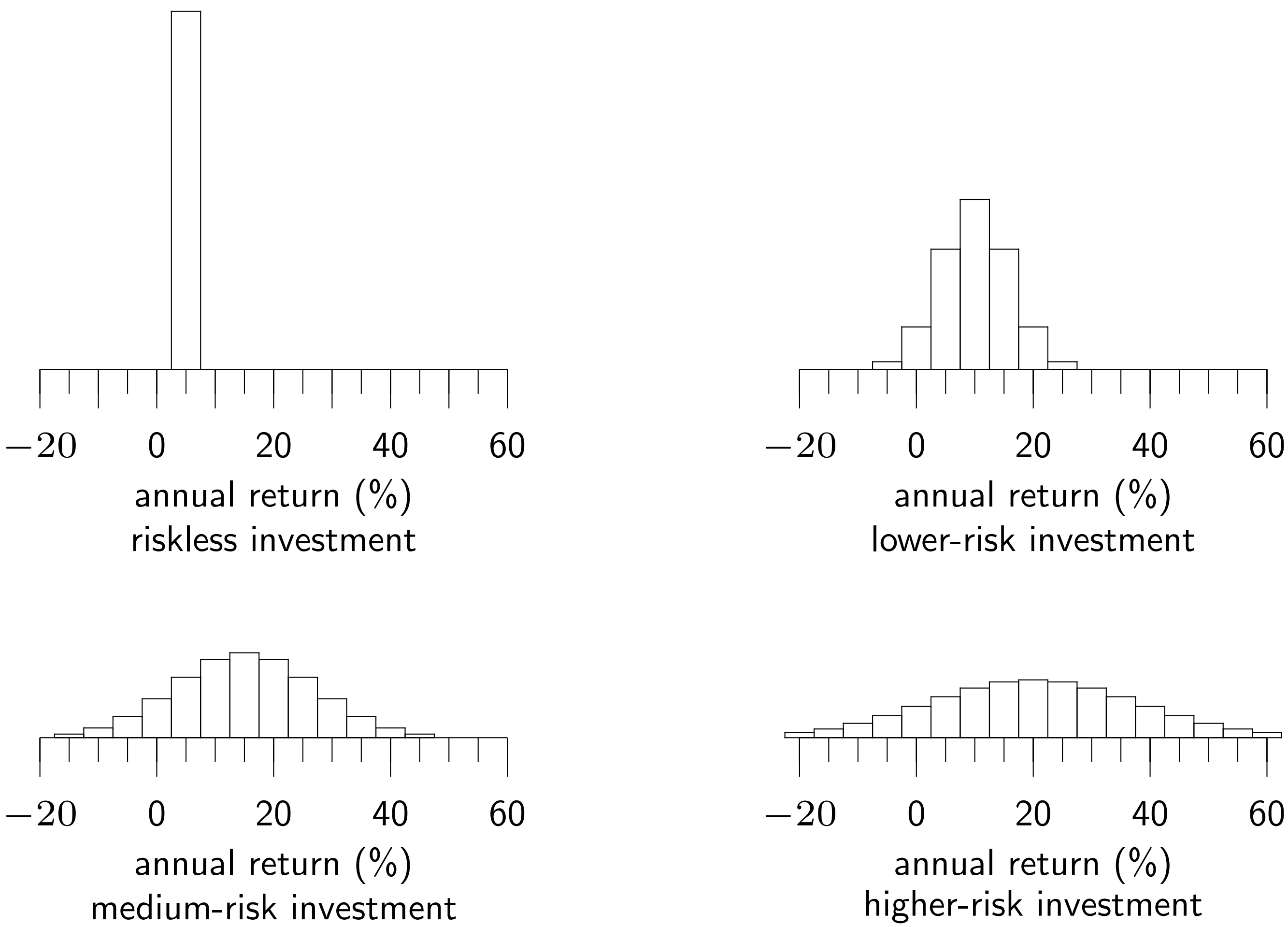
4.1 Location and Variation: Two Aspects of a Distribution

Two distributions with different variations:



4.1 Location and Variation: Two Aspects of a Distribution

A financial context:



4.2 Averages

Measuring the location, using the mode:

- The value with the highest frequency of a distribution is called the **mode** of the distribution.

For a continuous metric variable:

- The class with the highest frequency h_i/d_i is called the **modal class**.

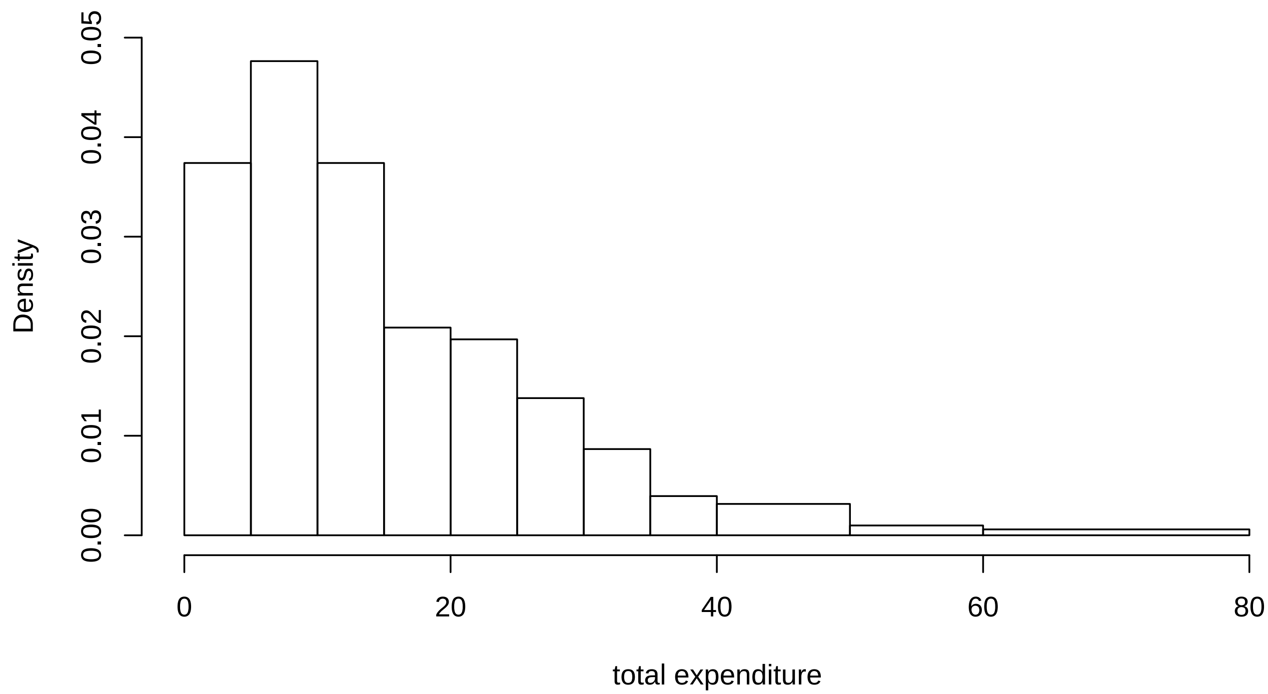
4.2 Averages

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	$\leftarrow$ mode
4:	ortaokul ve dengi	3.72	0.076	
5:	lise ve dengi	3.82	0.078	
6:	yüksekokul ve fakülte	1.50	0.030	
	$\sum$	49.14	1.000	

4.2 Averages

Example: Total expenditure of customers in a supermarket.



The modal class is [5, 10].

4.2 Averages

Measuring the location, using the arithmetic mean:

- if the observations $x_1, \ldots, x_n$ are given:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- if a distribution $f_1, \ldots, f_k$ is given:

$$\bar{x} = \sum_{i=1}^{k} f_i \cdot a_i$$

(Here, f_i is the relative frequency of value a_i.)

4.2 Averages

Example:

X = number of goals scored in a match of Beşiktaş İstanbul

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

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

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

- Use the distribution of the observations:

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

4.2 Averages

Example:

X = expenditure (in euros) of a customer in a supermarket

Data from 508 customers: 10.07, 22.61, 14.48, . . . , 28.68

The arithmetic mean is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{508}(10.07 + \ldots + 28.68) = 15.43$$

4.2 Averages

Properties of the arithmetic mean.

- Linearity: Let X, Y, Z be metric variables.
  - If $Y = aX + b$, then $\bar{y} = a\bar{x} + b$.
  - If $Z = X + Y$, then $\bar{z} = \bar{x} + \bar{y}$.

- Minimization property:
  The arithmetic mean $\bar{x}$ minimizes the function

$$a \mapsto \sum_{i=1}^{n} (x_i - a)^2.$$

4.2 Averages

If data are given as a histogram.

Approximate formula for the arithmetic mean:

$$\bar{x} \approx \sum_{i=1}^{k} x_i \cdot f_i,$$

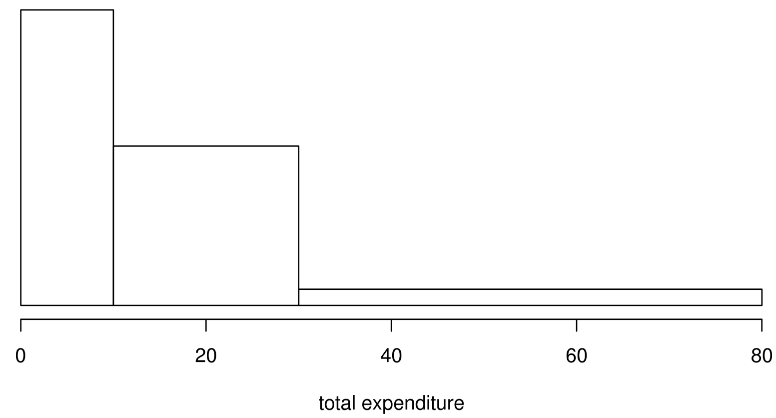
where

$$x_i = \text{center of class } i,$$
$$f_i = \text{relative frequency of class } i.$$

4.2 Averages

Example: Total expenditure of customers in a supermarket.

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
$\sum$		508		



An approximation of the arithmetic mean is then:

$$\bar{x} \approx \sum_{i=1}^{k} x_i \cdot f_i = 5 \cdot \frac{216}{508} + 20 \cdot \frac{233}{508} + 55 \cdot \frac{59}{508} = 17.69$$

4.2 Averages

The median.

Definition: Any value which divides the ordered set of observations into two equal parts is called a median.

Example. Compare the median of two datasets:

$$\underbrace{\underbrace{19 \quad 19 \quad 20 \quad 20 \quad 21}_{x_{\text{med}}=20} \quad 60}_{x_{\text{med}}=20}$$

The median is the same for both datasets. The median is "outlier-insensitive".

4.2 Averages

Example:

X = expenditure (in euros) of a customer in a supermarket

Data from 508 customers: 10.07, 22.61, 14.48, . . . , 28.68. — The ordered dataset is:

0.59,	. . . ,	12.05,	12.12,	. . . ,	75.54
[1]		[254]	[255]		[508]

The median is:

$$x_{\text{med}} = \frac{1}{2}(12.05 + 12.12) = 12.085$$

In words: Half of the customers spent less than 12.08 euros.

4.2 Averages

Example:

Educational attainment in New York, 2000.

code	category	number	percent	cum.
1	Less than 9th grade	689,368	11.20%	11.20%
2	Some high school, no diploma	910,155	14.79%	25.99%
3	High school graduate	1,487,728	24.17%	50.16%
4	Some college, no degree	943,044	15.32%	65.48%
5	Associate degree	329,580	5.36%	70.84%
6	Bachelor's degree	1,018,915	16.56%	87.40%
7	Graduate or professional degree	775,733	12.60%	100.00%
	Total Population Age 25+	6,154,523	100.00%	

What is the median of this distribution?

4.2 Averages

Properties of the median.

- Linearity: Let X, Y, Z be metric variables.
  - If $Y = aX + b$, then $y_{\text{med}} = ax_{\text{med}} + b$.
  - $Z = X + Y$ does not imply $z_{\text{med}} = x_{\text{med}} + y_{\text{med}}$.
- Minimization property:
  The median x_{med} minimizes the function

$$a \mapsto \sum_{i=1}^{n} |x_i - a|.$$

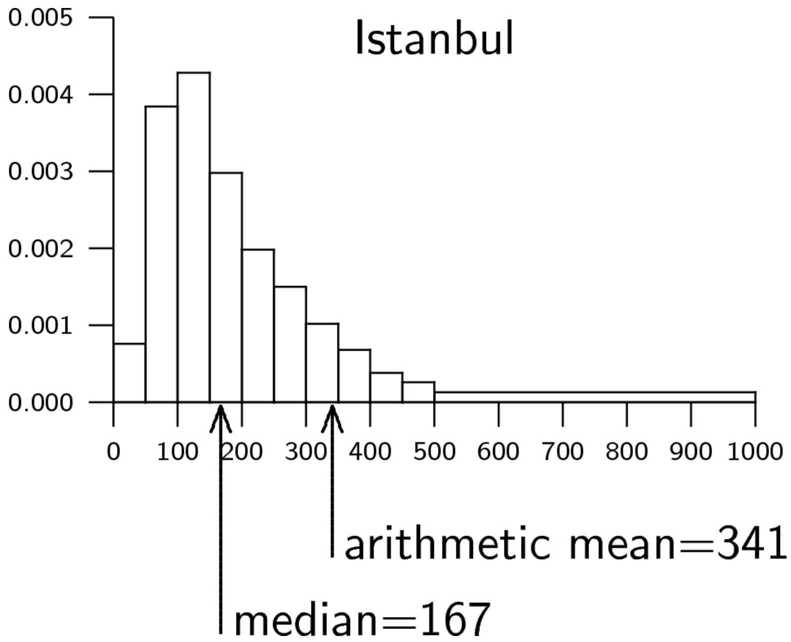
4.2 Averages

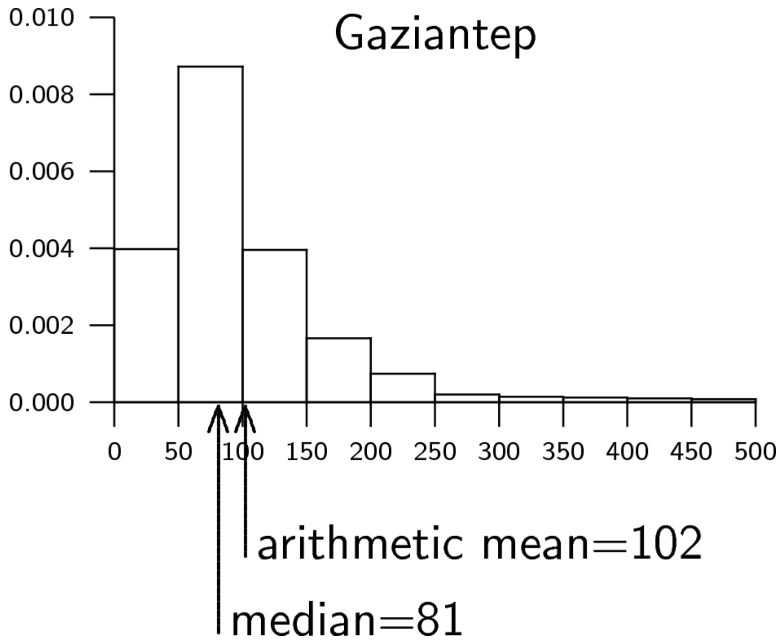
Arithmetic mean and median: a comparison.

- Both measure the location of a distribution.
- Required scaling of the variable:
  - arithmetic mean: metric variable
  - median: rank or metric variable
- The arithmetic mean is more outlier-sensitive than the median.
- For a right-skewed distribution, the arithmetic mean is always larger than the median.

4.2 Averages

Example: Household income 1994.





4.2 Averages

The geometric mean.

Definition: Let $x_1, \ldots, x_n \geq 0$ be real numbers.

$$\bar{x}_{\mathrm{G}} := \sqrt[n]{\prod_{i=1}^{n} x_i}$$

is called the geometric mean of the numbers $x_1, \ldots, x_n$.

4.2 Averages

Example: Increase in WPI numbers.

The increase in wholesale price indices in Turkey was:

1996	1997	1998	1999
84.9%	91.0%	54.3%	62.9%

Average annual increase in WPI for the period December 1995 through December 1999:

$$r = \sqrt[4]{1.849 \cdot 1.910 \cdot 1.543 \cdot 1.629} - 1 = 72.6\%.$$

4.3 The Variation of a Distribution

Measuring the variation, using the variance s^2:

- if the observations $x_1, \ldots, x_n$ are given:

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

- if a distribution $f_1, \ldots, f_k$ is given:

$$s^2 = \sum_{i=1}^{k} f_i \cdot (a_i - \bar{x})^2 = \sum_{i=1}^{k} f_i \cdot a_i^2 - \bar{x}^2$$

(Here, f_i is the relative frequency of value a_i.)

4.3 The Variation of a Distribution

Measuring the variation, using the standard deviation s:

- if the observations $x_1, \ldots, x_n$ are given:

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

- if a distribution $f_1, \ldots, f_k$ is given:

$$s = \sqrt{\sum_{i=1}^{k} f_i \cdot (a_i - \bar{x})^2}$$

4.3 The Variation of a Distribution

Example:

X = number of goals scored in a match of Beşiktaş İstanbul

How can we compute the variance s^2 of the number of goals per match?

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

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{170}\left(4^2 + \ldots + 5^2\right) - 2.96^2 = 3.21$$

- Use the distribution of the observations:

$$s^2 = \sum_i i^2 \cdot f_i - \bar{x}^2 = 0^2 \cdot \frac{12}{170} + \ldots + 10^2 \cdot \frac{1}{170} - 2.96^2 = 3.21$$

4.3 The Variation of a Distribution

Example:

X = expenditure (in euros) of a customer in a supermarket

The variance is:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{508}\left(10.07^2 + \ldots + 28.68^2\right) - 15.43^2 = 166.96 \text{ [euros}^2\text{]}$$

The standard deviation is:

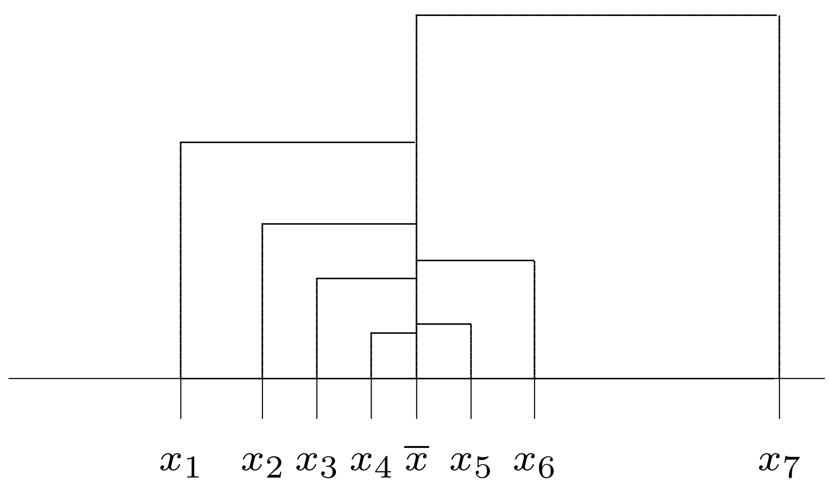
$$s = \sqrt{s^2} = \sqrt{166.96} = 12.92 \text{ [euros]}$$

4.3 The Variation of a Distribution

Properties of variance and standard deviation.

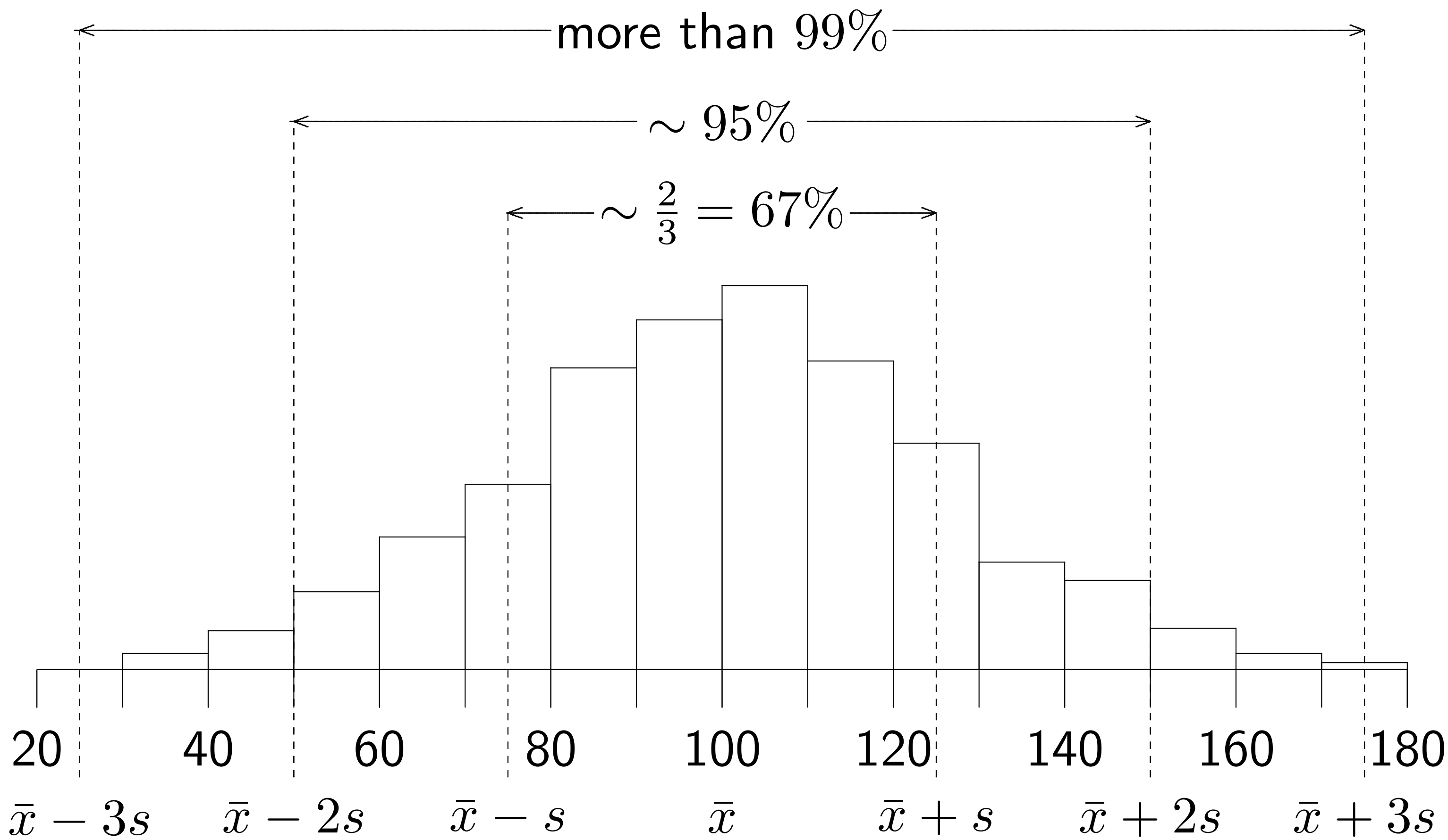
- If $Y = aX + b$, then
  $\text{var}(Y) = a^2\ \text{var}(X)$ and $\text{sd}(Y) = |a|\ \text{sd}(X)$.

- Outlier-sensitivity:



4.3 The Variation of a Distribution

The sigma-rules: A visual approach.



4.3 The Variation of a Distribution

The sigma-rules.

If the observations are approximately normally distributed:

- One-sigma-rule: About two thirds, or 67%, of all observations will be in the interval $[\bar{x} - s, \bar{x} + s]$; in words: About two thirds of the observations will be within a distance of one standard deviation from the arithmetic mean.

- Two-sigma-rule: About 95% of all observations will be in the interval $[\bar{x} - 2s, \bar{x} + 2s]$.

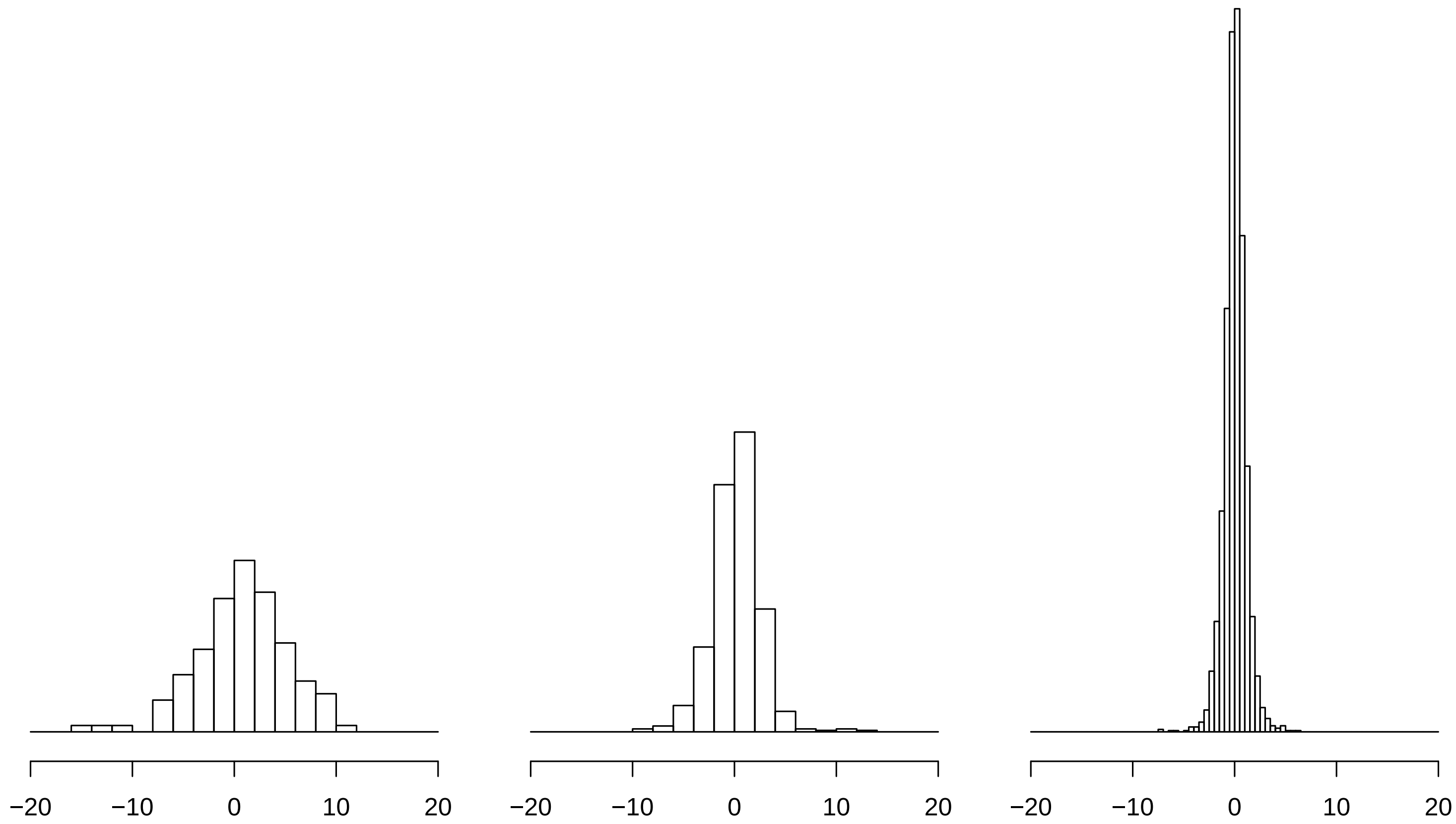
- Three-sigma-rule: Almost all observations (more than 99%) will be in the interval $[\bar{x} - 3s, \bar{x} + 3s]$

4.3 The Variation of a Distribution

The normal distribution.

4.3 The Variation of a Distribution

Example: Returns on the Dow-Jones Industrial Average, 1995-01 through 2005-10.



4.3 The Variation of a Distribution

Example: Returns on the Dow-Jones Industrial Average.

	monthly	weekly	daily
number of observations	129	560	2700
arithmetic mean, $\bar{r}$	0.89	0.21	0.04
variance, s^2	20.01	5.70	1.21
standard deviation, s	4.47	2.39	1.10
skewness, γ_1	-0.54	0.16	-0.11
kurtosis, γ_2	0.84	3.12	4.08
$[\bar{r} - s, \bar{r} + s]$	[-3.58 , 5.36]	[-2.18 , 2.59]	[-1.06 , 1.14]
observed	89	416	2031
expected	86	375	1809
$[\bar{r} - 2s, \bar{r} + 2s]$	[-8.06 , 9.84]	[-4.57 , 4.98]	[-2.15 , 2.24]
observed	124	535	2569
expected	123	532	2565
$[\bar{r} - 3s, \bar{r} + 3s]$	[-12.53 , 14.31]	[-6.96 , 7.37]	[-3.25 , 3.34]
observed	128	552	2668
expected	128	554	2673

4.4 The Shape of a Distribution

Shape parameters: The skewness.

The skewness is defined as

$$\gamma_1 = \frac{1}{n} \sum_i \left(\frac{x_i - \bar{x}}{s} \right)^3$$

If. . .	the distribution is. . .
$\gamma_1 = 0$	. . . symmetric
$\gamma_1 > 0$	. . . right-skewed
$\gamma_1 < 0$	. . . left-skewed

4.4 The Shape of a Distribution

Shape parameters: The kurtosis.

The kurtosis is defined as

$$\gamma_2 = \frac{1}{n} \sum_i \left(\frac{x_i - \bar{x}}{s} \right)^4 - 3$$

If. . .	the distribution is. . .
$\gamma_2 = 0$	. . . meso-kurtic
$\gamma_2 > 0$	. . . lepto-kurtic
$\gamma_2 < 0$	. . . platy-kurtic

4.4 The Shape of a Distribution

Example:

$X =$ expenditure (in euros) of a customer in a supermarket

Data from 508 customers: 10.07, 22.61, 14.48, . . . , 28.68.

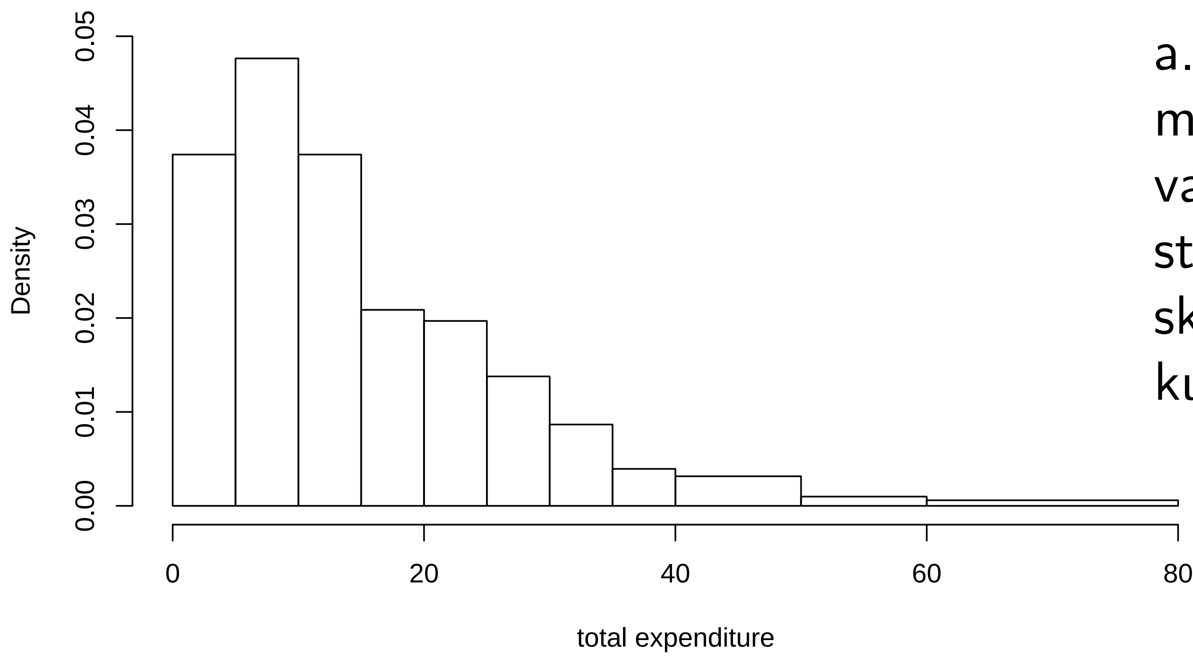
Skewness and kurtosis are:

$$\gamma_1 = 1.66, \qquad \gamma_2 = 3.48$$

This indicates a right-skewed, lepto-kurtic distribution.

4.4 The Shape of a Distribution

Example: Total expenditure of customers in a supermarket.



a.m.:	15.43
median:	12.09
variance:	166.96
st.dev.:	12.92
skewness:	1.658
kurtosis:	3.484

4.4 The Shape of a Distribution

Example of a service process: Customers of a copy-shop.

Interarrival times. . .

. . . before 2 p.m.:

(#)		
(11)	**0**⋆	00001122334
(6)	**0**•	678889
(6)	**1**⋆	033344
(2)	**1**•	57
(2)	**2**⋆	44
(1)	**2**•	7
(1)	**3**⋆	3
	3•	
	4⋆	
(1)	**4**•	5
(1)	**5**⋆	2
	5•	
(31)		

. . . after 2 p.m.:

(#)		
(24)	**0**⋆	000011111112222223334444
(12)	**0**•	566667778899
(4)	**1**⋆	0113
(1)	**1**•	8
(3)	**2**⋆	002
	2•	
	3⋆	
(1)	**3**•	5
	4⋆	
	4•	
	5⋆	
	5•	
(45)		

1|0=10 minutes

4.4 The Shape of a Distribution

Example of a service process: Customers of a copy-shop.

Service times:

(#)		
(10)	**0**⋆	1122333444
(15)	**0**•	555567888889999
(12)	**1**⋆	000011223444
(16)	**1**•	5555555556666778
(7)	**2**⋆	0002234
(3)	**2**•	557
(2)	**3**⋆	14
(3)	**3**•	789
(1)	**4**⋆	0
	4•	
(1)	**5**⋆	2
(1)	**5**•	6
(71)		

4.4 The Shape of a Distribution

Example of a service process: Customers of a copy-shop.

parameter	interarrival times... before 2 p.m.	after 2 p.m.	service times
arithmetic mean $\bar{x}$	12.63	6.90	15.75
variance s^2	163.02	49.53	129.23
skewness γ_1	1.55	2.01	1.48
kurtosis γ_2	2.02	4.55	2.35
minimum x_{min}	0.50	0.50	1.50
lower quartile $\tilde{x}_{0.25}$	2.50	2.50	8.50
median x_{med}	8.50	4.50	14.50
upper quartile $\tilde{x}_{0.75}$	15.50	8.50	20.50
maximum x_{max}	52.50	35.50	56.50