Statistics

Lecture 2

Descriptive Statistics: Qualitative and Quantitative Data Items

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

David Bartl
Statistics
INM/BASTA

Outline of the lecture

- Bar chart
- Histogram
- Measures of central tendency (arithmetic mean, mode, median)
- Measures of variability (range, variance, coefficient of variation)
- Measures of data concentration (skewness, kurtosis)

Statistics

The purpose of statistics is to present data in a comprehensive form.

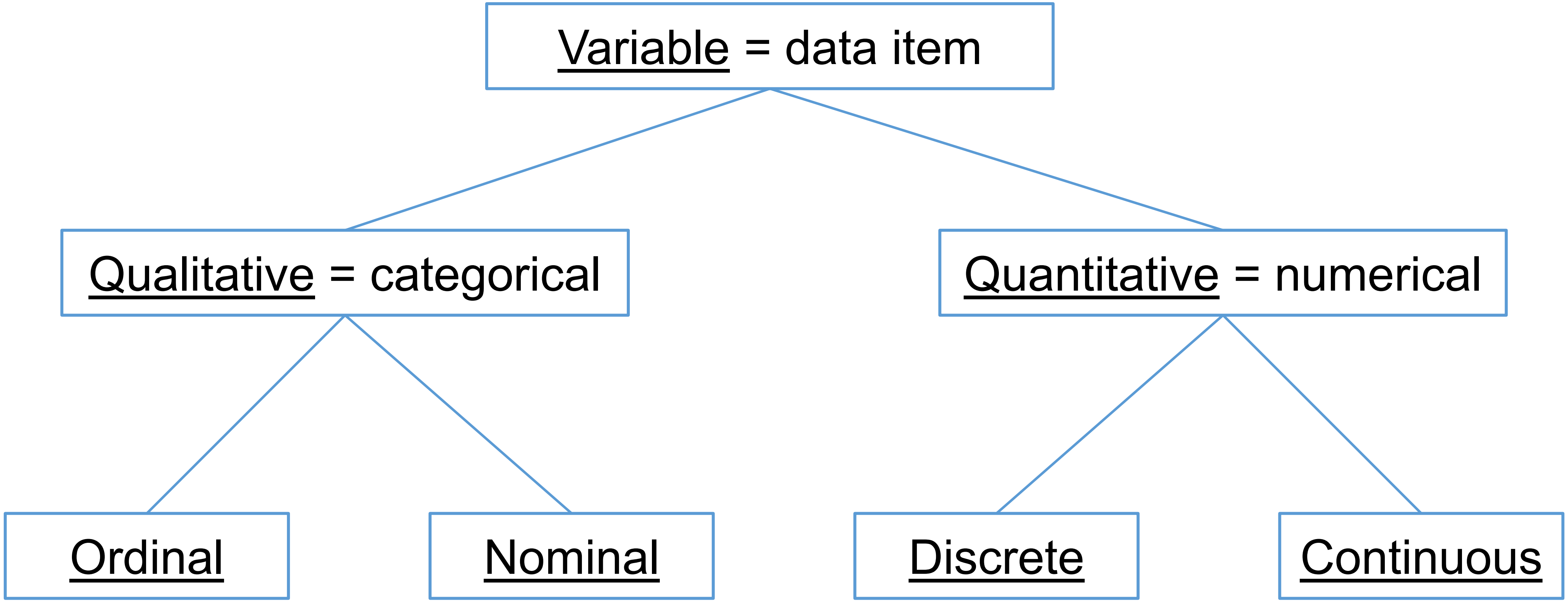
The goal is to analyse the information and reveal relations hidden in the data.

There are two approaches:

- Descriptive statistics (categorization, characteristics)
  → we shall deal with it now

- Inductive statistics (assumptions about the origin of the data, probability distributions)
  → we shall deal with it later

Data items = Variables



Variable = data item

Qualitative = categorical

Quantitative = numerical

Ordinal

Nominal

Discrete

Continuous

Example: Employees (a sample of the Dataset)

ID	Gender	Age	Marital Status	Education	Position	Salary per Year	Evaluation
5060	M	65	divorced	secondary	worker	258800	4
1030	M	60	divorced	university	manager	630000	2
3049	M	60	married	primary	operator	436600	5
5047	M	60	widowed	primary+vocational	worker	240600	3
5061	M	60	widowed	primary+vocational	worker	241800	1
5087	M	60	widowed	secondary	worker	239500	—
5133	F	60	married	secondary	worker	241100	4
5177	F	60	widowed	secondary	worker	239600	4
3030	F	58	widowed	primary	operator	422600	1
3014	F	56	widowed	university	operator	303600	3
5012	F	56	widowed	primary+vocational	worker	223100	4
5056	M	56	divorced	primary	worker	225200	5
5101	M	56	unmarried	primary+vocational	worker	224600	4
5106	M	56	married	primary+vocational	worker	226100	7
5146	F	56	married	primary+vocational	worker	224900	3
5153	M	56	divorced	secondary	worker	224500	4
5189	M	56	married	primary+vocational	worker	224600	1
5196	M	56	widowed	primary+vocational	worker	222800	3
1031	M	55	married	university	manager	429000	—
5016	M	55	divorced	secondary	administrative officer	259000	5
5021	F	55	married	primary+vocational	worker	220200	—
5062	F	55	widowed	primary+vocational	worker	221400	5
5107	M	55	divorced	primary+vocational	worker	220500	4
5154	F	55	widowed	primary+vocational	worker	219200	5
5195	M	55	married	primary+vocational	worker	219400	6

Methods to present data in a comprehensive form

For **qualitative** (categorical) data items:

- Bar chart of frequencies
- Mode ($\hat{x}$)

For **ordinal qualitative** (categorical) data items:

- Median ($\tilde{x}$)

Methods to present data in a comprehensive form

For **quantitative** (numerical) data items:

- Histogram of frequencies
- Mode $(\hat{x})$
- Median $(\tilde{x})$
- Sample mean $(\bar{x})$
- Sample standard variance (s^2) and sample standard deviation $(s = \sqrt{s^2})$
- Sample coefficient of variation $(\mathrm{cv} = \frac{s}{|\bar{x}|})$

Bar chart & Histogram of frequencies

Bar chart

- used for qualitative [categorical (nominal or ordinal)] data items
- can also be used for discrete numerical data items
- presents the frequencies of each category by the height of a rectangular bar (the height is proportional to the frequency)
- there are gaps between the bars, i.e. the bars are not adjacent

Bar chart of frequencies for qualitative data items

Example: The dataset of the employees.

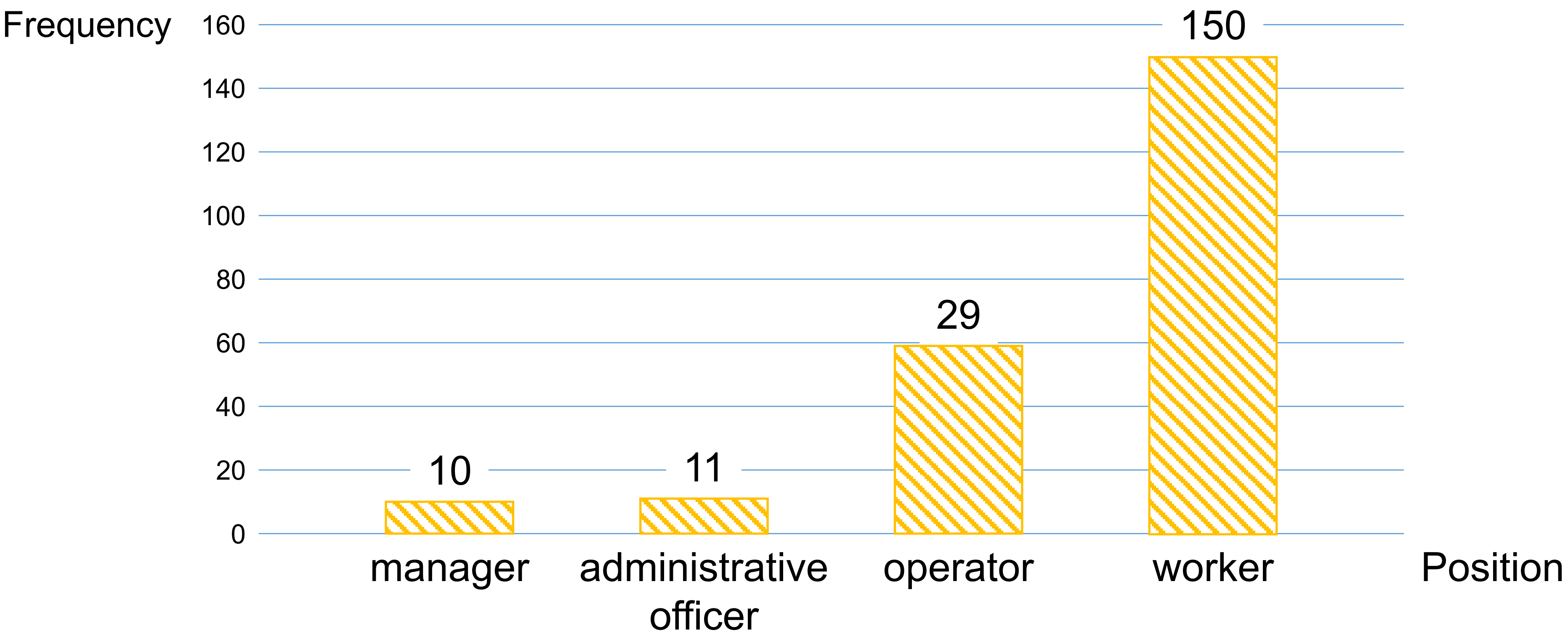
We examine now the nominal data item “Position”:

Table:

Position	Frequency (number)	Relative frequency
manager	□ 10	□ □ 5.0 %
administrative officer	□ 11	□ □ 5.5%
operator	□ 29	□ 14.5%
worker	150	□ 75.0%
TOTAL	200	100.0 %

Bar chart of frequencies for qualitative data items

Bar chart – the frequencies (numbers) of the nominal data item “Position”:



Bar chart for ordinal qualitative data items

Example: The dataset of the employees.

We examine now the ordinal data item “Evaluation”:

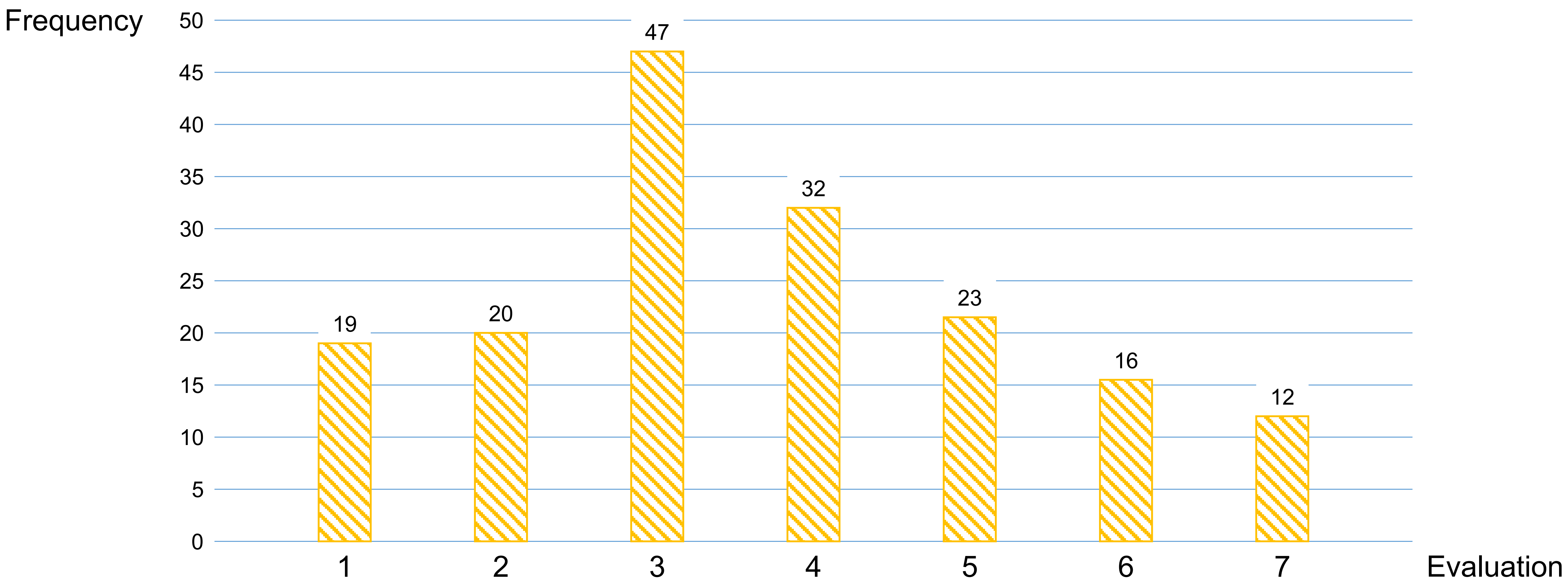
Table:

(rounded to 3 decimal places)

Evaluation	Frequency (number)	Relative frequency
□ □ □ 1— very bad	□ 19	□ 11.243 %
□ □ □ 2— bad	□ 20	□ 11.834 %
□ □ □ 3— rather bad	□ 47	□ 27.811 %
□ □ □ 4— acceptable	□ 32	□ 18.935 %
□ □ □ 5— quite good	□ 23	□ 13.609 %
□ □ □ 6— good	□ 16	□ □ 9.467 %
□ □ □ 7— very good	□ 12	□ □ 7.101 %
TOTAL	169	100.000 %

Bar chart of frequencies for qualitative data items

Bar chart – the frequencies (numbers) of the ordinal data item “Evaluation”:



Measures of central tendency of the data item

Mode $\hat{x}$ is the most frequent value in the population.

In our first example, when $x \in$ “Position”, the mode is $\hat{x}$ = worker

In our second example, when $x \in$ “Evaluation”, the mode is $\hat{x}$ = 3 = “rather bad”

Measures of central tendency of the data item

Assume that the variable (data item) is ordinal or numerical and

— either the number of the data units in the population is odd

(and the variable may be ordinal qualitative or numerical quantitative)

— or the number of the data units in the population is even,

but the variable must be numerical

Median $\tilde{x}$ is the “middle value”.

Measures of central tendency of the data item

To find the **median** $\tilde{x}$,

— sort the all the data units according to the data item in the ascending order:

$$x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n$$

— if n is odd, then

$$\tilde{x} = x_{\left((n+1)/_2\right)}$$

e.g., if $n = 9$, then $\tilde{x} = x_5$

— if n is even, then

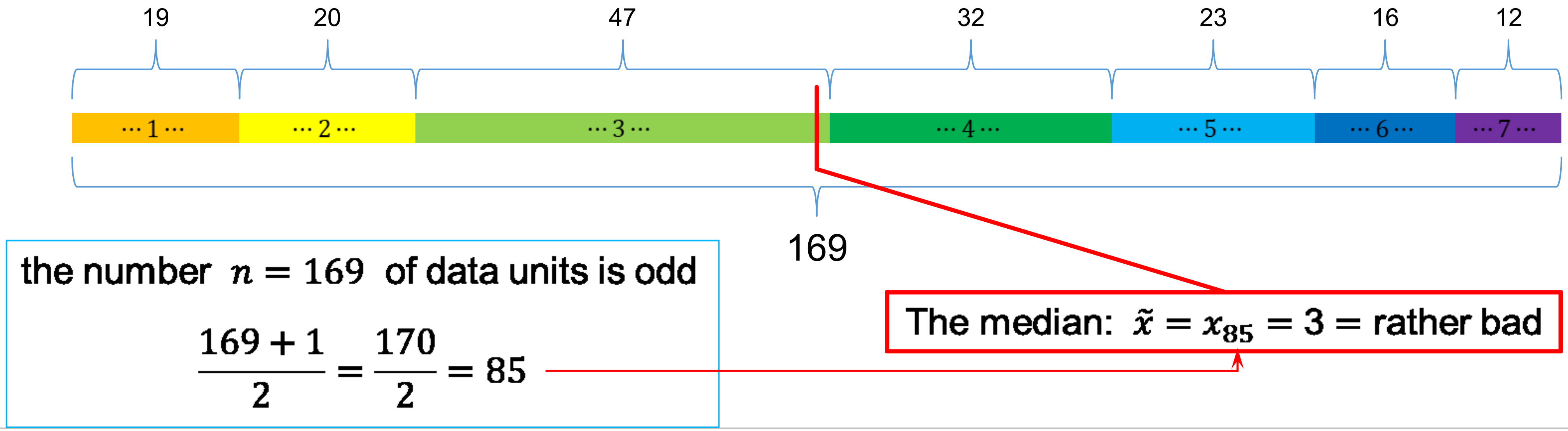
$$\tilde{x} = \frac{x_{(n/_2)} + x_{\left((n+2)/_2\right)}}{2}$$

e.g., if $n = 10$, then $\tilde{x} = \dfrac{x_5 + x_6}{2}$

Measures of central tendency of the data item

Median $\tilde{x}$ is the "middle value".

In our second example, when $x \in$ "Evaluation", the <u>median is found as follows</u>:



the number $n = 169$ of data units is odd

$$\frac{169+1}{2} = \frac{170}{2} = 85$$

The median: $\tilde{x} = x_{85} = 3$ = rather bad

Bar chart & Histogram of frequencies

Histogram

— used for quantitative [numerical (continuous or discrete)] data items

— the numerical range of the variable is divided into disjoint adjacent intervals

— usually of the type $(a, b]$ where $a < b$ are finite numbers

— the height of the bar =

(the number of the data items in the interval) + (the width of the interval)

— there are no gaps between the adjacent intervals, i.e. the bars touch each other

Histogram of frequencies for quantitative data items

Example: The dataset of the employees.

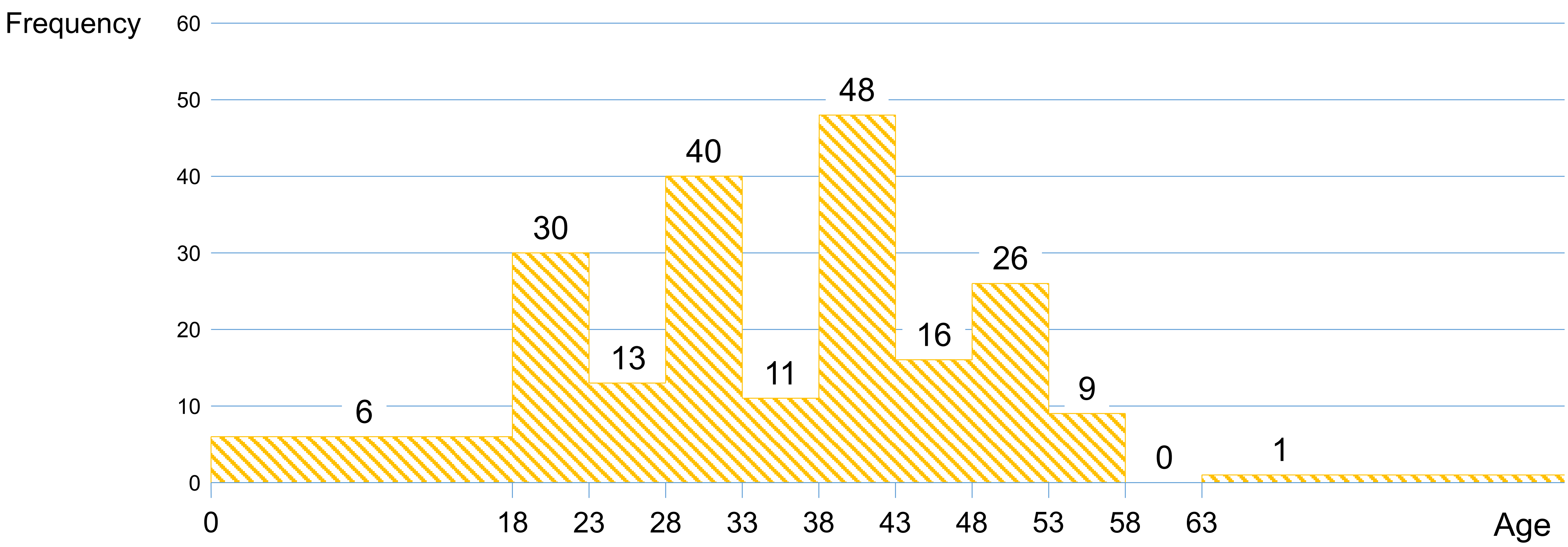
We examine now the numerical data item “Age” considered as a continuous value:

Table:

Age interval	Frequency (number)	Cumulative frequency	Relative frequency	Cumulative relative frequency
□ 0 <$x \le$ 18	□ 6	□ □ 6	□ 3.0 %	□ □ 3.0 %
18 < $x \le$ 23	30	□ 36	15.0 %	□ 18.0 %
23 < $x \le$ 28	13	□ 49	□ 6.5 %	□ 24.5 %
28 < $x \le$ 33	40	□ 89	20.0 %	□ 44.5 %
33 < $x \le$ 38	11	100	□ 5.5 %	□ 50.0 %
38 < $x \le$ 43	48	148	24.0 %	□ 74.0 %
43 < $x \le$ 48	16	164	□ 8.0 %	□ 82.0 %
48 < $x \le$ 53	26	190	13.0 %	□ 95.0 %
53 < $x \le$ 58	□ 9	199	□ 4.5 %	□ 99.5 %
58 < $x \le$ 63	□ 0	199	□ 0.0 %	□ 99.5 %
63 < x	□ 1	200	□ 0.5 %	100.0 %

Histogram of frequencies for continuous data items

Histogram – the frequencies (numbers) of the continuous data item “Age”:



Histogram of frequencies for continuous data items

If the (ordinary) histogram is used to display relative frequencies, then

— if the variable is continuous,

the histogram gives an estimate of the underlying probability density

— if the variable is discrete,

the histogram gives an estimate of the underlying probability distribution

If the cumulative histogram is used to display the cumulative relative frequencies and the variable is continuous, then the cumulative histogram gives an estimate of the cumulative distributive function.

Histogram of frequencies for continuous data items

- In the bar chart, each category has its own bar.

- If the variable is continuous,

  how to determine the number of intervals in the histogram?

→ If the intervals should be of the same length and a length ℓ is suggested,

then the number of the intervals should be

$$N = \left\lceil \frac{\max x - \min x}{\ell} \right\rceil$$

The suggested number of the intervals in the histogram

Denote: n – the number of observations in the sample
N – the suggested number of the intervals in the histogram

Sturges' formula:

$$N = \lceil \log_2 n \rceil + 1 \approx \lceil 3.3 \times \log_{10} n \rceil + 1$$

Square root:

$$N = \lceil \sqrt{n} \rceil$$

Rice University et al. rule:

$$N = \lceil 2\sqrt[3]{n} \rceil$$

Etc.

Histogram of frequencies for quantitative data items

Example: The dataset of the employees.

We examine now the numerical data item “Salary per year” considered as a continuous value.

There are $n = 200$ employees (data units).

Hence the number of the intervals in the histogram suggested by Sturges’ rule is

$$N = \lceil \log_2 200 \rceil + 1 = \lceil 7.643856\ldots \rceil + 1 = 8 + 1 = 9$$

Histogram of frequencies for quantitative data items

The minimum salary in the dataset is 71 000.
The maximum salary in the dataset is 657 000.

The suggested number of the intervals in the histogram is $N = 9$.

We have

$$\frac{657000 - 71000}{9} = \frac{586000}{9} = 65111\frac{1}{9} = 65111.111\ldots$$

We round the result to 70 000.

We conclude thus that the length of each interval in the histogram should be

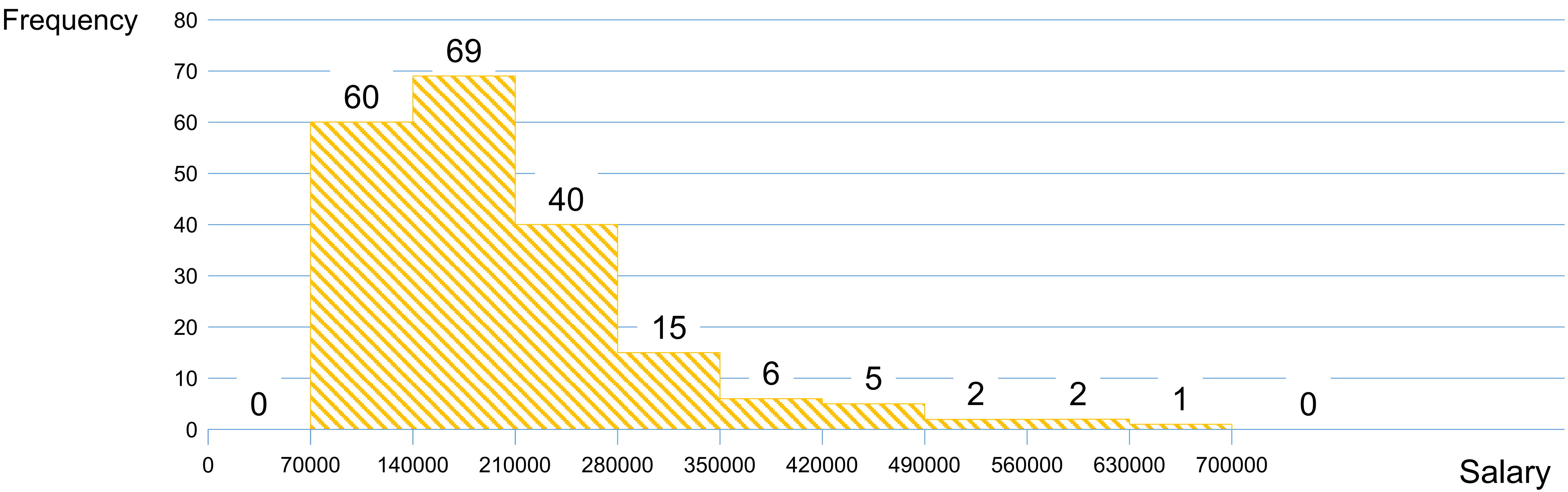
$$\ell = 70000$$

Histogram of frequencies for quantitative data items

Salary interval	Frequency (number)	Cumulative frequency	Relative frequency	Cumulative relative frequency
0 < x ≤ 70 000	0	0	0.0 %	0.0 %
70 000 < x ≤ 140 000	60	60	30.0 %	30.0 %
140 000 < x ≤ 210 000	69	129	34.5 %	64.5 %
210 000 < x ≤ 280 000	40	169	20.0 %	84.5 %
280 000 < x ≤ 350 000	15	184	7.5 %	92.0 %
350 000 < x ≤ 420 000	6	190	3.0 %	95.0 %
420 000 < x ≤ 490 000	5	195	2.5 %	97.5 %
490 000 < x ≤ 560 000	2	197	1.0 %	98.5 %
560 000 < x ≤ 630 000	2	199	1.0 %	99.5 %
630 000 < x ≤ 700 000	1	200	0.5 %	100.0 %
700 000 < x	0	200	0.0 %	100.0 %

Histogram of frequencies for continuous data items

Histogram – the frequencies (numbers) of the continuous data item “Salary”:



Measures of central tendency

Assume that a variable (data item) is numerical, i.e. quantitative, discrete or continuous. We then consider several measures of central tendency of the variable:

- Arithmetic mean
- Mode
- Median

Population & Sample

Assume that we have a set (i.e. a “**population**”) of values of some phenomenon, which we observe / measure / study / deal with. In practice, this set may be very very large (e.g. some data item, the data units being all the people living on the Earth), thus unknown to us. Another example might be the set of all results of some experiment, yet the instances which we have not done yet.

Assume however, that the set exists (in theory at least) and that the set is finite (for simplicity).

Population & Sample

Let

$$\Omega = \{1, 2, 3, \dots, N\}$$

be the underlying set of all data units. We assume for simplicity that the set Ω is finite and that N is the number of its elements.

Now, we assume that the values of the variable (data item)

$$x_1, x_2, x_3, \dots, x_N$$

exist (in theory at least).

We assume that $x_1, x_2, x_3, \dots, x_N \in \mathbb{R}$, i.e. the values are real numbers.

Population & Sample

All the values

$$x_1, x_2, x_3, \ldots, x_N$$

which exist (in theory at least), are called the **population.**

Notice that we may not know the whole population for various reasons (we cannot do all the measurements since the population is very large; the values include the results of future experiments, which have not been done yet).

Population & Sample

Our method of examination of the population $x_1, x_2, x_3, \dots, x_N$ consists in **the selection of a sample**

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

out of the population.

Notice that the same letter ("x") is used to denote the values of both the population and the sample. No misunderstanding occurs because it is always clear from the context whether we mean the population or the sample.

Notice also <u>the notation:</u> N = the number of elements of the **population**

n = the number of elements of the **sample**

Arithmetic mean

Population arithmetic mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Sample arithmetic mean:

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

Arithmetic mean

Weighted population mean:

Assuming that weights $w_1, w_2, w_3, \ldots, w_N$ (positive real numbers)

of the values $x_1, x_2, x_3, \ldots, x_N$ are given, the weighted population mean is

$$\mu_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

Arithmetic mean

Notice the notation:

Greek letters denote quantities relating to the population:

μ = the **population** mean (theoretical, may not be known exactly)

μ_w = the weighted **population** mean (theoretical, may not be known exactly)

Latin letters denote quantities relating to the sample:

$\bar{x}$ = the **sample** mean (the result of measurements really done)

Median & Mode

Median $\tilde{x}$ is the “middle value” such that

- one half of the values is ≤ the median
- one half of the values is ≥ the median

Mode $\hat{x}$ is the “most frequent value” such that

- the probability distribution attains a local maximum at the mode
- there may be more than one mode:
    - unimodal probability distribution (one mode)
    - bimodal probability distribution (two modes)
    - etc.

The median and the mode are defined both for the population and for the sample. The definition is the same.

Sample Mean / Median / Mode in Excel

In Excel, use the functions:

=**AVERAGEA**()	to calculate the sample arithmetic mean
=**MEDIAN**()	to find the sample median
=**MODE.SNGL**()	to find one of the sample modes
=**MODE.MULT**()	to find many of the sample modes (matrix function, press “Ctrl-Shift-Enter”)
=MODE()	to find one of the sample modes (the same as =MODE.SNGL(), deprecated)

Example: Employees (a sample of the Dataset)

ID	Gender	Age	Marital Status	Education	Position	Salary per Year	Evaluation
5060	M	65	divorced	secondary	worker	258800	4
1030	M	60	divorced	university	manager	630000	2
3049	M	60	married	primary	operator	436600	5
5047	M	60	widowed	primary+vocational	worker	240600	3
5061	M	60	widowed	primary+vocational	worker	241800	1
5087	M	60	widowed	secondary	worker	239500	—
5133	F	60	married	secondary	worker	241100	4
5177	F	60	widowed	secondary	worker	239600	4
3030	F	58	widowed	primary	operator	422600	1
3014	F	56	widowed	university	operator	303600	3
5012	F	56	widowed	primary+vocational	worker	223100	4
5056	M	56	divorced	primary	worker	225200	5
5101	M	56	unmarried	primary+vocational	worker	224600	4
5106	M	56	married	primary+vocational	worker	226100	7
5146	F	56	married	primary+vocational	worker	224900	3
5153	M	56	divorced	secondary	worker	224500	4
5189	M	56	married	primary+vocational	worker	224600	1
5196	M	56	widowed	primary+vocational	worker	222800	3
1031	M	55	married	university	manager	429000	—
5016	M	55	divorced	secondary	administrative officer	259000	5
5021	F	55	married	primary+vocational	worker	220200	—
5062	F	55	widowed	primary+vocational	worker	221400	5
5107	M	55	divorced	primary+vocational	worker	220500	4
5154	F	55	widowed	primary+vocational	worker	219200	5
5195	M	55	married	primary+vocational	worker	219400	6

sample

Example: Employees — data item “Age”

The true (population) values:

Population size:

$$N = 200$$

Population mean (Age):

$$\mu = 39.9$$

Population median:

$$\tilde{x} = 42$$

Population mode:

$$\hat{x} = 45$$

The estimated (sample) values:

Sample size:

$$n = 8$$

Sample mean (Age):

$$\bar{x} = 60.625$$

Sample median:

$$\tilde{x} = 60$$

Sample mode:

$$\hat{x} = 60$$

The measures of the central tendency

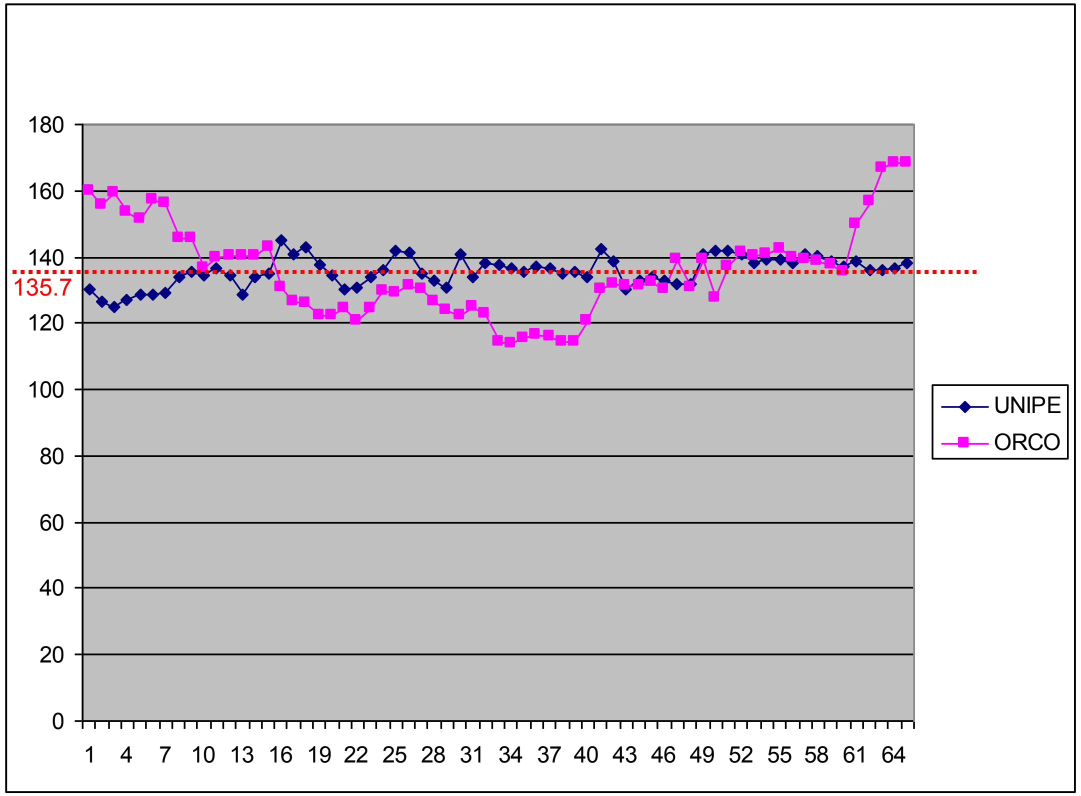
Which of the measures of the central tendency are the best?

Consider the next example – monthly salaries in 2001 and 2002:

Employee	Salary in 2001	Salary in 2002
A	10	25
B	10	25
C	10	25
D	20	20
E	20	20
F	20	20
G	20	20
H	20	20
I	20	20
J	20	20
K	20	20
L	20	20
M	20	20
N	50	50
O	50	50
MEAN	**22**	**25**
MEDIAN	**20**	**20**
MODE	**20**	**20**

Measures of variability

To appreciate the measures of variability, consider the next example – the prices of two stocks (ORCO and UNIPE) during a period of time:



The average price of both stocks is the same:

$$\bar{x}_{\text{UNIPE}} = \bar{x}_{\text{ORCO}} = 135.7$$

But we feel that UNIPE is more stable, while ORCO is more volatile.

Measures of variability

Assume that a variable (data item) is numerical, i.e. quantitative, discrete or continuous. We then consider several measures of variability of the variable:

- Range
- Variance (dispersion)
- Coefficient of variation

Range

Population range:

$$R = \max_{i=1,\dots,N} x_i - \min_{i=1,\dots,N} x_i$$

Sample range:

$$R = \max_{i=1,\dots,n} x_i - \min_{i=1,\dots,n} x_i$$

Variance (dispersion)

Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Sample variance:

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

Variance (dispersion)

Notice that once the population $x_1, x_2, x_3, \dots, x_N$ of the values is fixed, then the population mean μ and the population variance σ^2 are given, i.e. these theoretical values are fixed (though not known exactly sometimes).

If the sample $x_1, x_2, \dots, x_n$ of the values is selected from the population randomly (select an element randomly n-times; the same element may be chosen repeatedly several times), then the resulting values of the

sample mean $\bar{x}$ and sample variance s^2

are random variables too!!!

Variance (dispersion)

Calculating

- the expected value $\mathrm{E}[\bar{x}]$ of the sample mean and
- the expected value $\mathrm{E}[s^2]$ of the sample variance,

we obtain that

$$\mathrm{E}[\bar{x}] = \mu \qquad \text{and} \qquad \mathrm{E}[s^2] = \sigma^2$$

Variance (dispersion)

That is,

- taking a sample of randomly selected n elements of the population (¡ where one element of the population may be present several times in the sample !)
- calculating the sample mean $\bar{x}$ and the sample variance s^2,
- repeating the above process infinitely many times, and
- calculating **the average value of the sample mean and the average value of the sample variance,**

we obtain precisely

the population mean μ and the population variance σ^2

Variance (dispersion)

<u>Conclusion:</u> We often do <u>not</u> know the exact values μ and σ^2 in practice.

However, if we take a sample of n elements selected randomly with repetition (i.e. an element can be selected several times) from the population and calculate the sample mean $\bar{x}$ and the sample variance s^2, then we have

$$\bar{x} \approx \mu \qquad \text{and} \qquad s^2 \approx \sigma^2$$

i.e. the sample mean $\bar{x}$ and the sample variance s^2 **are good estimates** of the unknown population mean μ and population variance σ^2.

→ **<u>That is why</u>** we divide by $(n-1)$ in the sample variance s^2, not by n.

Standard deviation

Population standard deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Sample standard deviation:

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

Variance (dispersion) & Standard deviation

Notice the notation:

Greek letters denote quantities relating to the population:

σ^2 = the **population** variance (theoretical, may not be known exactly)

σ = the **population** standard deviation

Latin letters denote quantities relating to the sample:

s^2 = the **sample** variance (the result of measurements really done)

s = the **sample** standard deviation

Coefficient of variation

Coefficient of variation:

$$V = \frac{\sigma}{|\mu|}$$

Sample coefficient of variation:

$$v = \frac{s}{|\bar{x}|}$$

Example

We have

$$\bar{x}_{\text{UNIPE}} = 135.7 \qquad \text{and} \qquad s_{\text{UNIPE}} = 2.09$$

hence

$$v_{\text{UNIPE}} = \frac{s_{\text{UNIPE}}}{|\bar{x}_{\text{UNIPE}}|} = \frac{2.09}{135.7} = 0.0154$$

We have

$$\bar{x}_{\text{ORCO}} = 135.7 \qquad \text{and} \qquad s_{\text{ORCO}} = 3.72$$

hence

$$v_{\text{ORCO}} = \frac{s_{\text{ORCO}}}{|\bar{x}_{\text{ORCO}}|} = \frac{3.72}{135.7} = 0.0274$$

Example

We have

$$v_{\mathrm{UNIPE}} = 1.54\ \% \qquad \text{and} \qquad v_{\mathrm{ORCO}} = 2.74\ \%$$

We conclude (by the above fact) that

the UNIPE bonds are less risky than the ORCO bonds.

In this particular example,

the UNIPE bonds are about 1.8× less risky than the ORCO bonds.

Sample Variance / Standard deviation

In Excel, use the functions:

=**VARA**() to calculate the sample variance

=**STDEVA**() to calculate the sample standard deviation

=VAR.S() to calculate the sample variance (skipping text values)

=VAR() to calculate the sample variance (skipping text values) (the same as =VAR.S(), deprecated)

Population Variance / Standard deviation

In Excel, use the functions:

=**VARPA**()	to calculate the population variance
=**STDEVPA**()	to calculate the population standard deviation
=VAR.P()	to calculate the population variance (skipping text values)

Measures of data concentration

Assume that a variable (data item) is numerical, i.e. quantitative, discrete or continuous. We then consider several measures of data concentration of the variable:

- Skewness
- Kurtosis

Skewness: Pearson’s moment coefficient of skewness

Population **skewness:**

$$\mathrm{Skew}(X) = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^3}{\sigma^3}$$

Sample **skewness:**

$$\mathrm{Skew}(X) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{s^3}$$

Skewness: Properties and interpretation

Pearson’s moment coefficient of skewness

$$\mathrm{Skew}(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^3}{\sigma^3} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^3$$

is a sum of the third powers of the fractions $\frac{x_i-\mu}{\sigma}$.

If the fraction is “small”, i.e. $\left|\frac{x_i-\mu}{\sigma}\right| < 1$, then its third power is yet smaller,

almost vanishes, $\left|\frac{x_i-\mu}{\sigma}\right|^3 < \left|\frac{x_i-\mu}{\sigma}\right| < 1$, i.e. is not counted much in the sum.

Skewness: Properties and interpretation

Pearson's moment coefficient of skewness

$$\mathrm{Skew}(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^3}{\sigma^3} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^3$$

is a sum of the third powers of the fractions $\frac{x_i-\mu}{\sigma}$.

If the fraction is "large", i.e. $\left|\frac{x_i-\mu}{\sigma}\right| \geq 1$, then its third power is also large,

$\left|\frac{x_i-\mu}{\sigma}\right|^3 \geq \left|\frac{x_i-\mu}{\sigma}\right| \geq 1$, i.e. is counted in the sum properly.

Skewness: Properties and interpretation

Pearson's moment coefficient of skewness

$$\mathrm{Skew}(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^3}{\sigma^3} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^3$$

can be positive or zero or negative.

- $\mathrm{Skew}(X) < 0$ — the majority of the values is left to the mean
- $\mathrm{Skew}(X) = 0$ — the values are distributed ≈ symmetrically around the mean
- $\mathrm{Skew}(X) > 0$ — the majority of the values is right to the mean

Large positive or negative value — there are "outliers", i.e.
values far away from the mean

Skewness in Excel

In Excel, use the functions:

=**SKEW.P**()	to calculate the population skewness
=**SKEW**()	to calculate the sample skewness

Skewness in Excel

Notice that we have defined sample skewness as the Pearson moment coefficient

$$\mathrm{Skew}(X) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{s^3}$$

cf. the function =**SKEW.P()** in Excel.

To calculate the sample skewness, cf. the function =**SKEW()**, Excel uses the adjusted Fisher-Pearson standardized moment coefficient

$$\mathrm{Skew}(X) = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{s^3}$$

Kurtosis: Pearson's moment coefficient of kurtosis

Population kurtosis:

$$\mathrm{Kurt}(X) = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^4}{\sigma^4}$$

Sample kurtosis:

$$\mathrm{Kurt}(X) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{s^4}$$

Kurtosis: Properties and interpretation

Pearson's moment coefficient of kurtosis

$$\mathrm{Kurt}(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^4}{\sigma^4} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^4$$

is a sum of the fourth powers of the fractions $\frac{x_i-\mu}{\sigma}$.

If the fraction is "small", i.e. $\left|\frac{x_i-\mu}{\sigma}\right| < 1$, then its fourth power is yet smaller,

almost vanishes, $\left|\frac{x_i-\mu}{\sigma}\right|^4 < \left|\frac{x_i-\mu}{\sigma}\right| < 1$, i.e. is not counted much in the sum.

Kurtosis: Properties and interpretation

Pearson’s moment coefficient of kurtosis

$$\mathrm{Kurt}(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^4}{\sigma^4} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^4$$

is a sum of the fourth powers of the fractions $\frac{x_i-\mu}{\sigma}$.

If the fraction is “large”, i.e. $\left|\frac{x_i-\mu}{\sigma}\right| \geq 1$, then its fourth power is also large,

$\left|\frac{x_i-\mu}{\sigma}\right|^4 \geq \left|\frac{x_i-\mu}{\sigma}\right| \geq 1$, i.e. is counted in the sum properly.

Kurtosis: Properties and interpretation

Pearson’s moment coefficient of kurtosis

$$\mathrm{Kurt}(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^4}{\sigma^4} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^4$$

can be positive or zero.

- $\mathrm{Kurt}(X) \geq 0$ is small — the values are concentrated ≈ around the mean
- $\mathrm{Kurt}(X) > 0$ is large — there are “outliers”, i.e. values far away from the mean

The Skewness & Kurtosis describe the shape of the distribution of the values (i.e. the shape of the histogram).

Excess kurtosis

The kurtosis of the Gaussian normal distribution is = 3.

That is why, the number 3 is sometimes subtracted to obtain the **population excess kurtosis:**

$$\mathrm{ExKurt}(X) = \mathrm{Kurt}(X) - 3 = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^4}{\sigma^4} - 3$$

Kurtosis in Excel

In Excel, use the function:

=**KURT**() to calculate the **sample excess kurtosis**

Kurtosis in Excel

Notice that we would define the sample excess kurtosis by using the Pearson moment coefficient

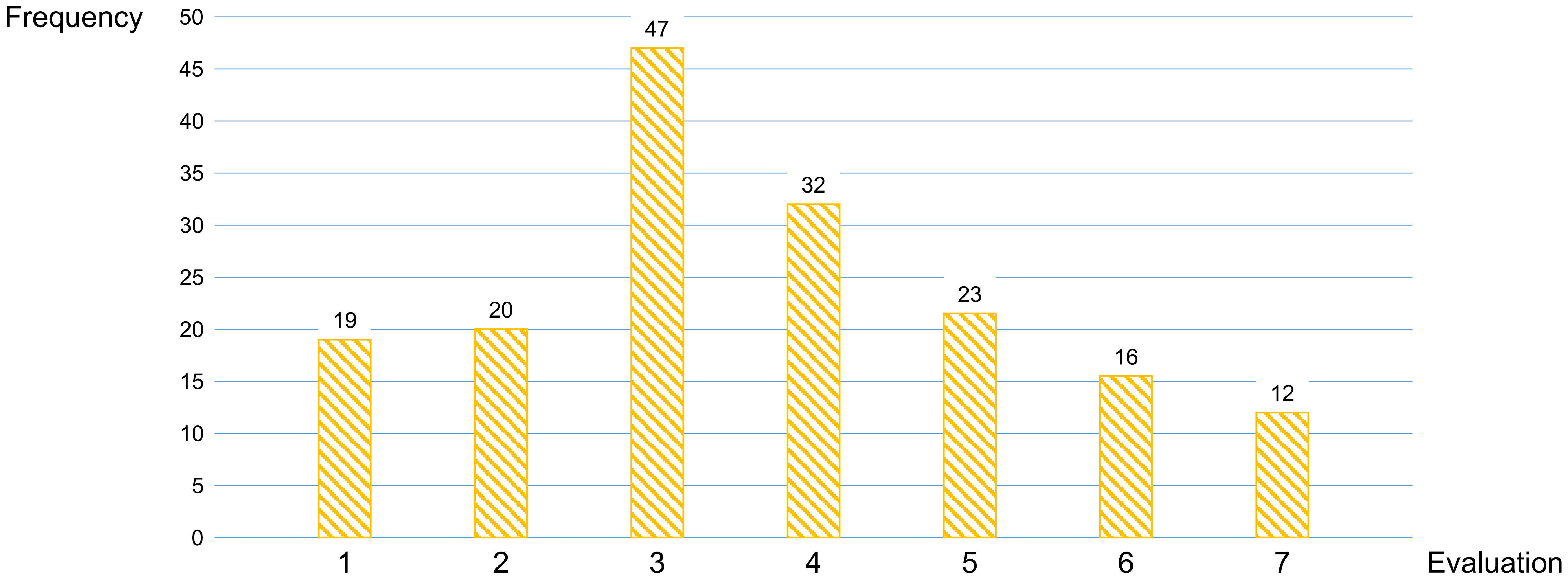
$$\mathrm{ExKurt}(X) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{s^4} - 3$$

To calculate the sample kurtosis, the function **=KURT()** in Excel uses the formula

$$\mathrm{ExKurt}(X) = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{s^4} - 3\,\frac{(n-1)^2}{(n-2)(n-3)}$$

Example

Population mean $\mu = 3.686\ldots$ Median $\tilde{x} = 3$ Population skewness: $\mathrm{Skew} = 0.26\ldots$
Population variation $\sigma^2 = 2.8188\ldots$ Mode $\hat{x} = 3$ Population excess kurtosis: $\mathrm{ExKurt} = -0.66\ldots$


Frequency
Evaluation