Descriptive statistics

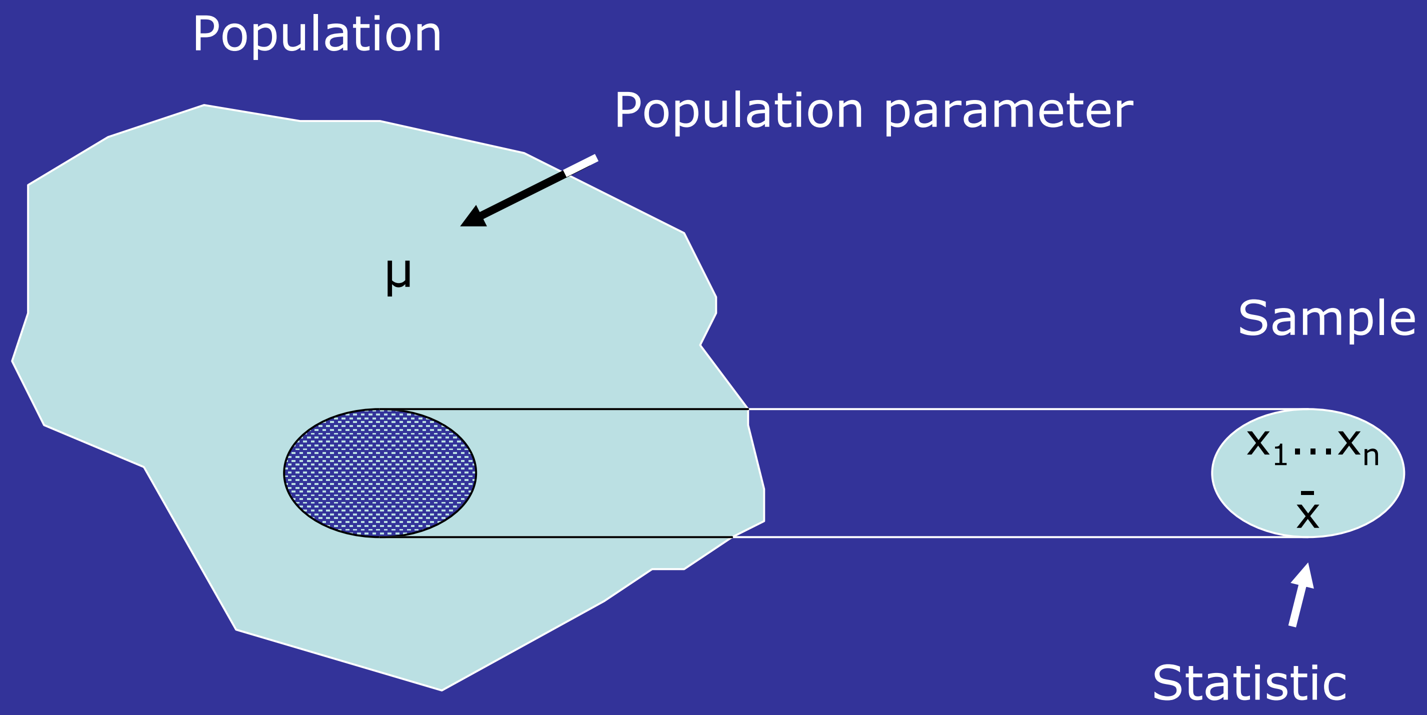
András Keszei

1st Budapest Clinical Epidemiology Course – organized jointly with the 15th Budapest Nephrology School

Biostatistics

- Descriptive statistics
  - summarizing a collection of data in a clear and understandable way

- Inferential statistics
  - conclusions extending beyond the data
  - inferring from sample what the population might look like



Population

- Aggregate of individuals or items from which the sample is taken
  - Population of Pest county aged 18 years or older
  - Patients undergone lithotripsy between 1996-1999 in Ontario
  - Internal medicine departments in Hungary

Sample

- subset of the population

Data

Variables

Observations

	Variable 1.	Variable 2.	Variable …
Patient 1.	10	1	
Patient 2.	15	4	
Patient 3.	10	9	
Patient …			

Data

Variables

Observations

	Variable 1.	Variable …	weight
~~Observation 1.~~	10		17
~~Observation 2.~~	15		21
~~Observation 3.~~	10		10
~~Observation …~~			

Data

Változók

Megfigyelések

	Patient ID.	Variable 1.	Variable 2.
Observation 1.	1	1	17
Observation 2.	1	4	21
Observation 3.	2	9	10
Observation 4.	2	8	11

Types of data

- **Qualitative** data
  - categories, groups
  - grouped into mutually exclusive categories
    - blood type, tumour stage
- **Quantitative**, numerical data
  - *discrete* or *continuous*
    - *number of children in the family, blood pressure*

Levels of measurement

- Nominal
  - Categories
  - No order
    - Gender, Social Security Number
- Ordinal
  - Categories in order
    - tumour stage, level of education, Likert scale
- Interval
  - meaningful intervals, no zero point
    - temperature
- Ratio
  - intervals and ratios between measurements
  - true zero point
    - blood pressure, weight, temperature in Kelvin

Levels of measurement

Hierarchy

+

information

- Nominal
- Ordinal
- Interval
- Ratio

Levels of measurement

Colour

- black, brown, red, orange, yellow, green, blue, violet, gray, and white

Data summary

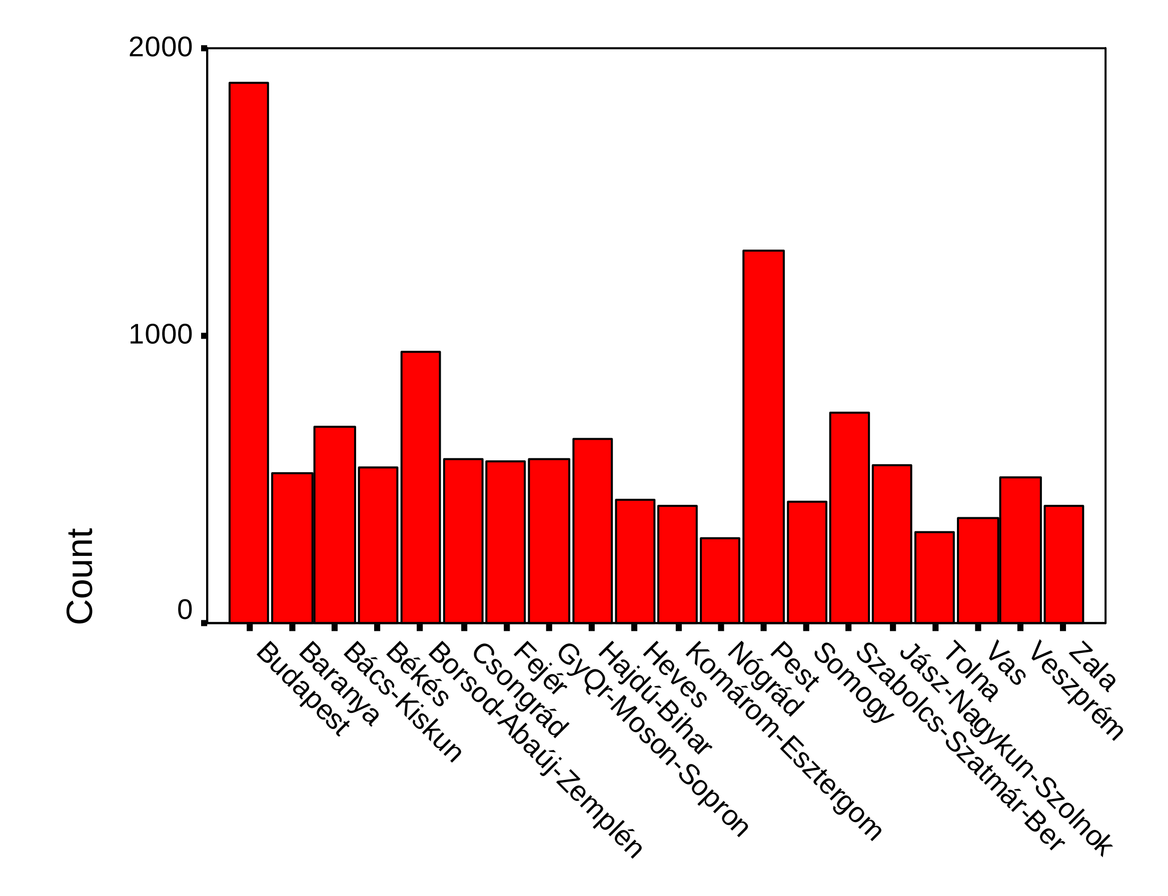
- Visualizing data
  - Graphs
  - Tables
- Describing with numbers
  - central tendency
  - variability
  - distribution

Graphical representation

- Bar chart
- Pie chart
- Box plot
- Pareto chart
- Histogram
- Run chart
- Frequency polygon
- Stem-and-leaf plot
- Scatter plot
- Pictograph
- Violin plot

Bar chart

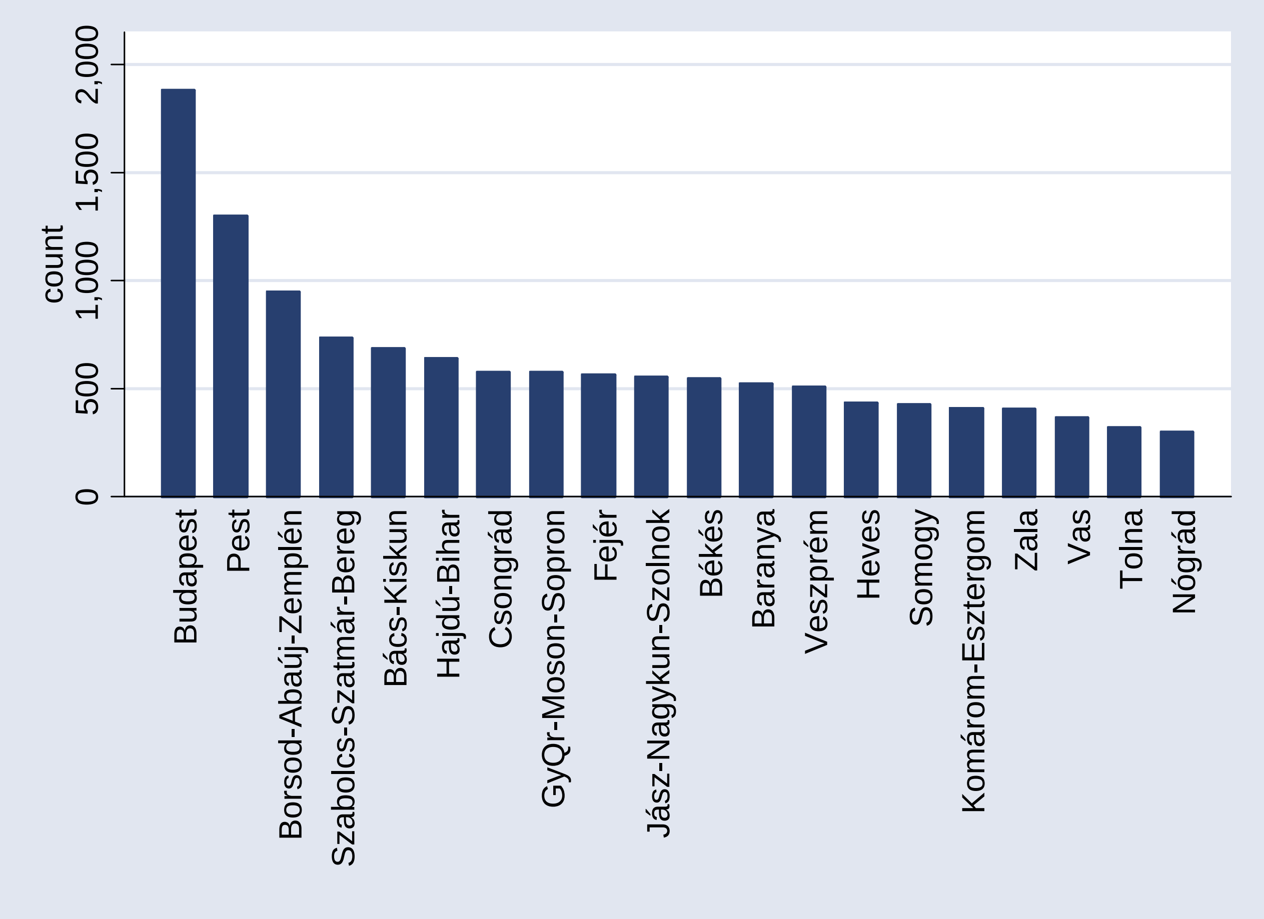
One categorical variable



n=12662

Bar chart

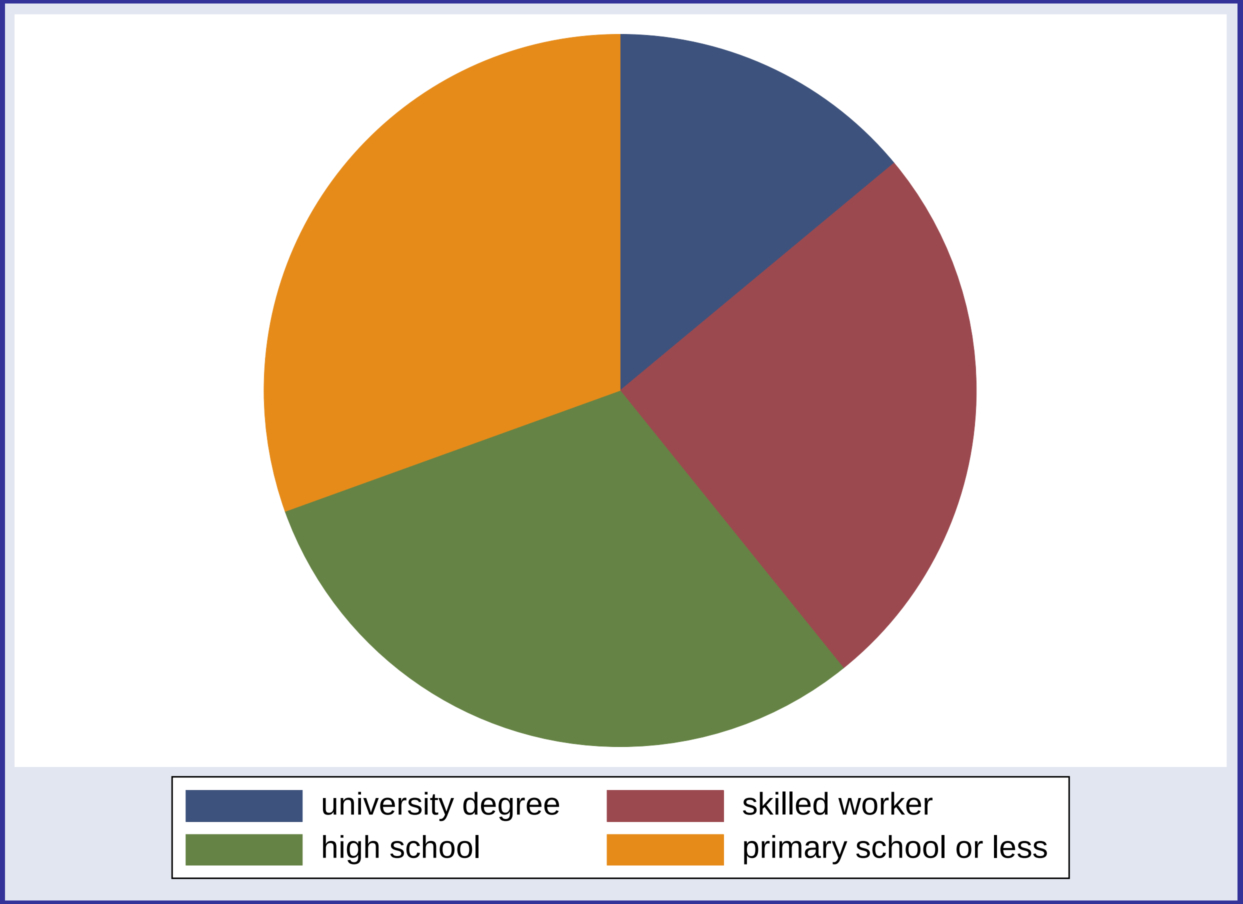
One categorical variable



n=12662

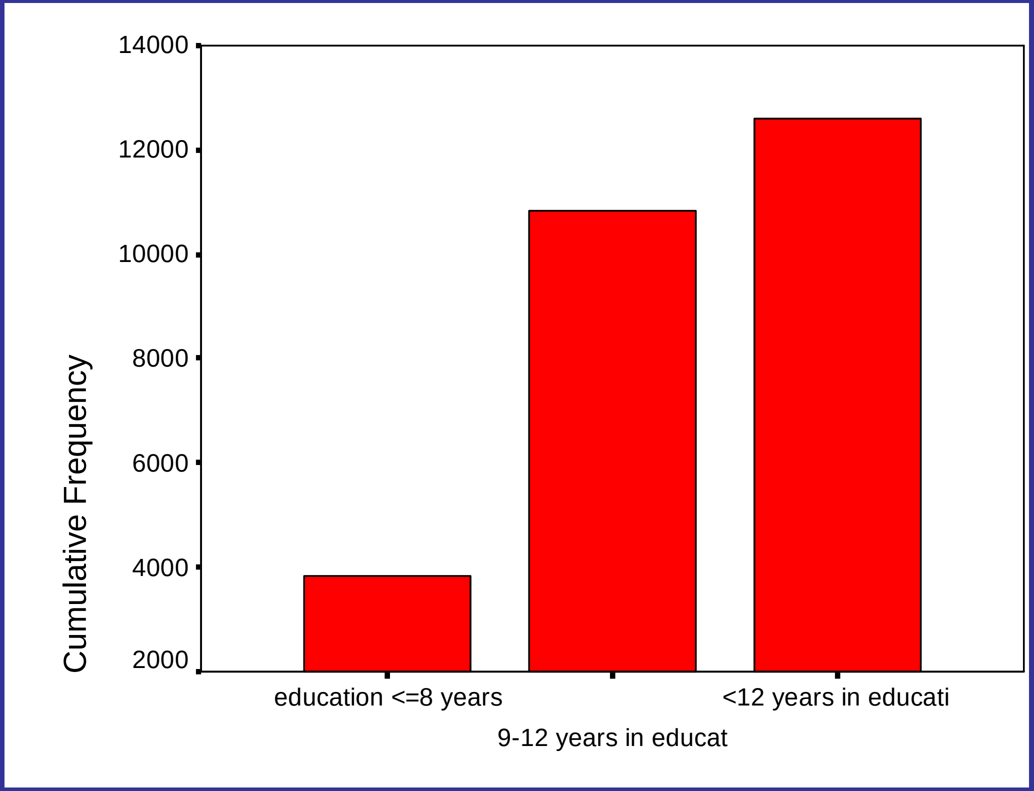
Pie chart

Categorical variable



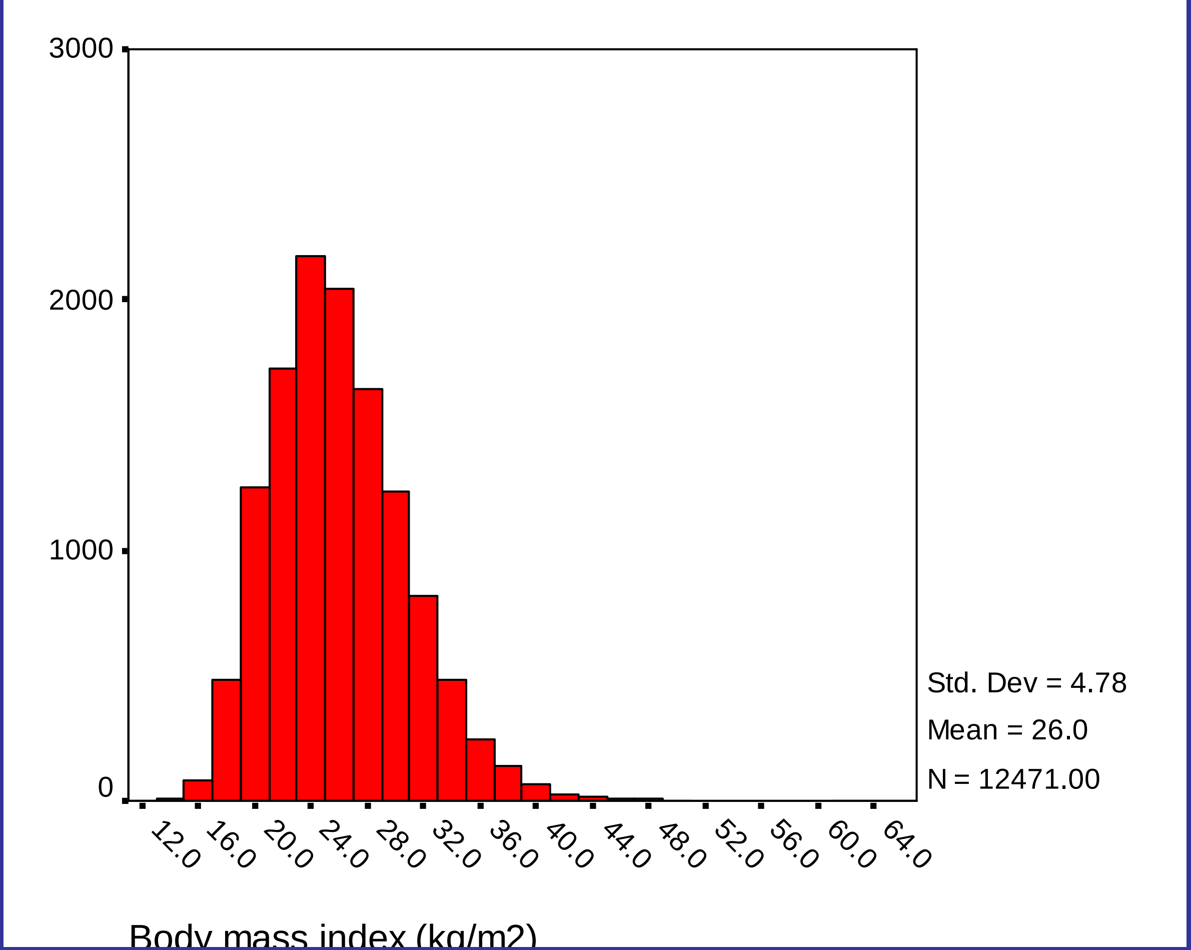
Bar chart

Categorical variable



Histogram

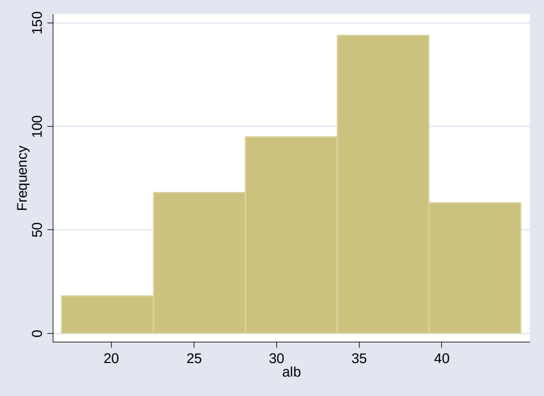
continuous variable

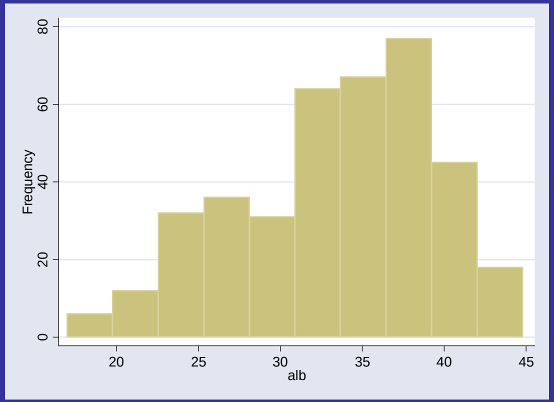


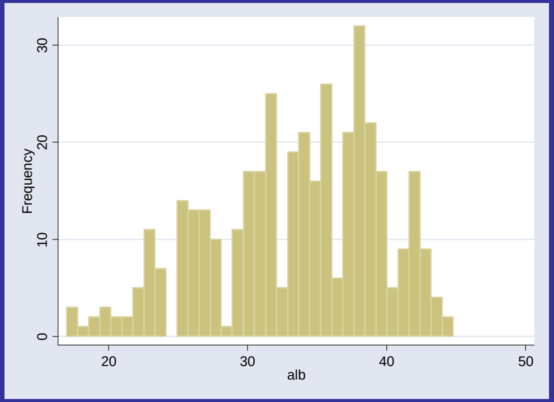
Histogram

Class interval width

$$\text{Width} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Number of class intervals}}$$





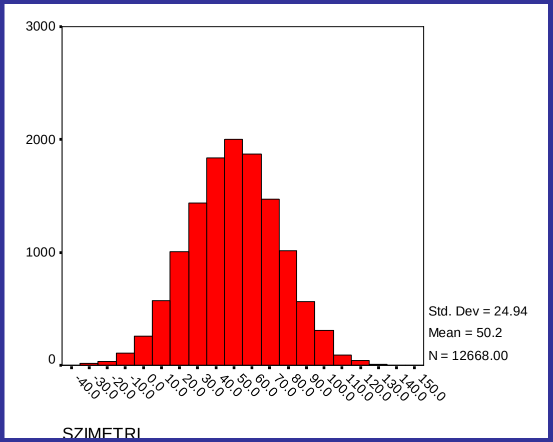


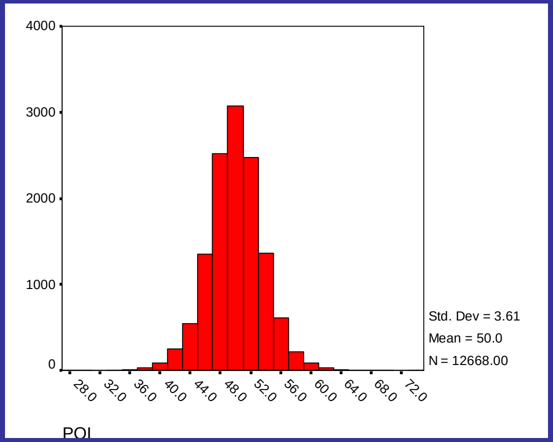
$$5.6 = \frac{48.4 - 17}{5}$$

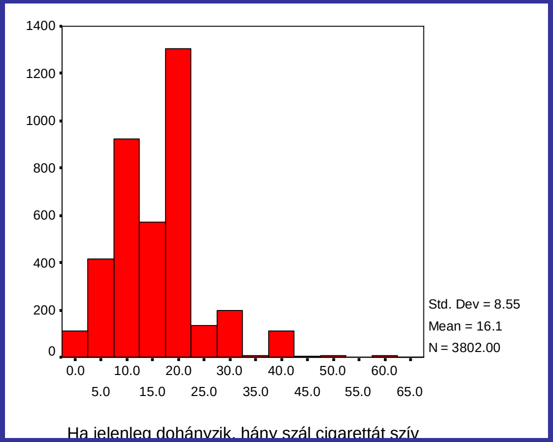
$$2.8 = \frac{48.4 - 17}{10}$$

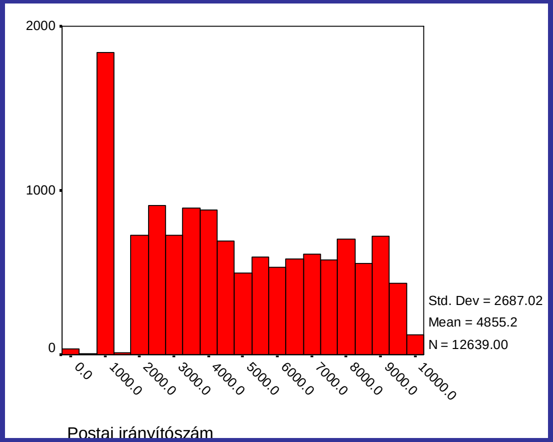
$$1.8 = \frac{48.4 - 17}{35}$$

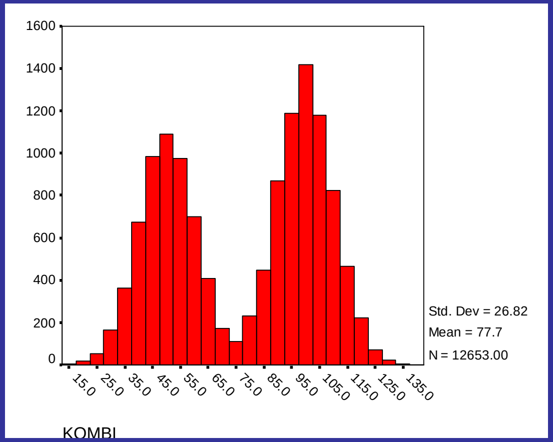
Distribution

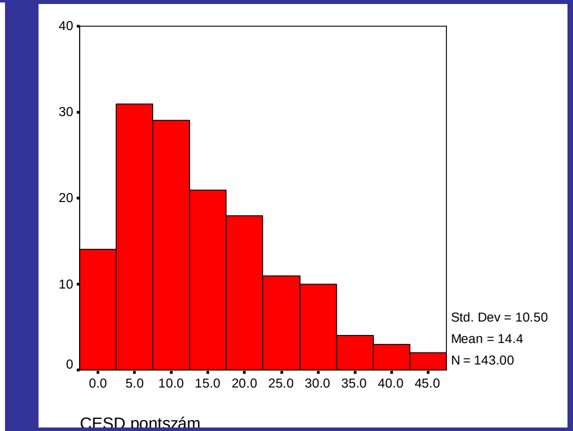




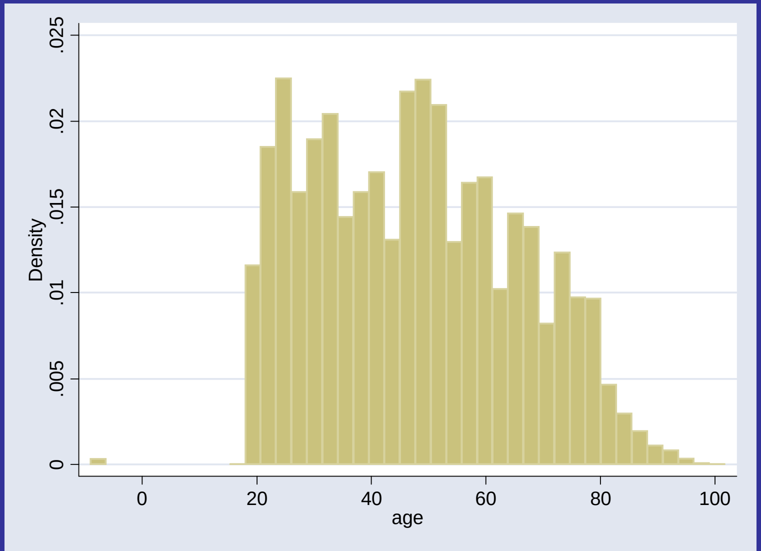


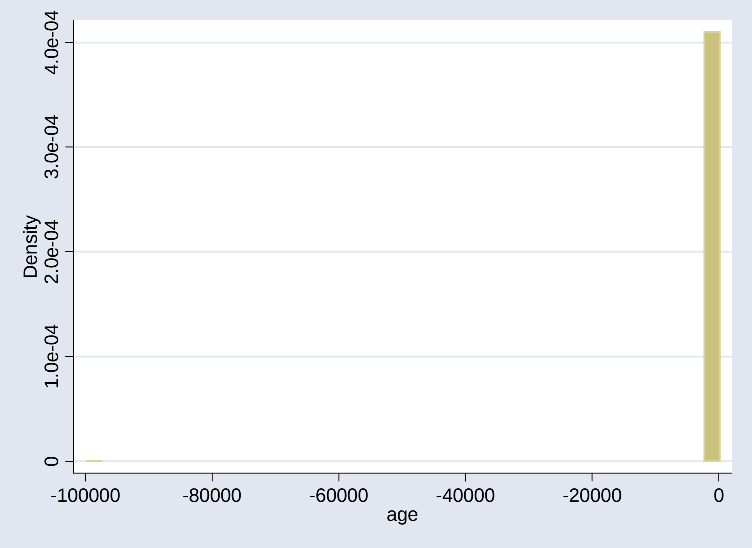


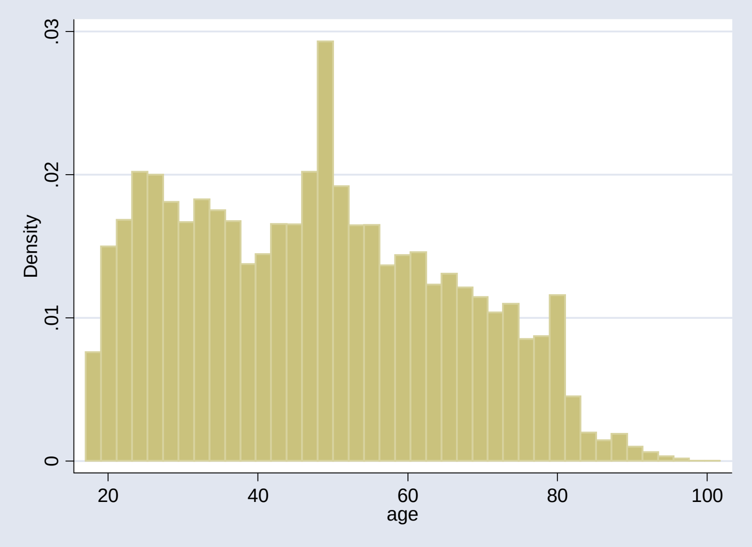




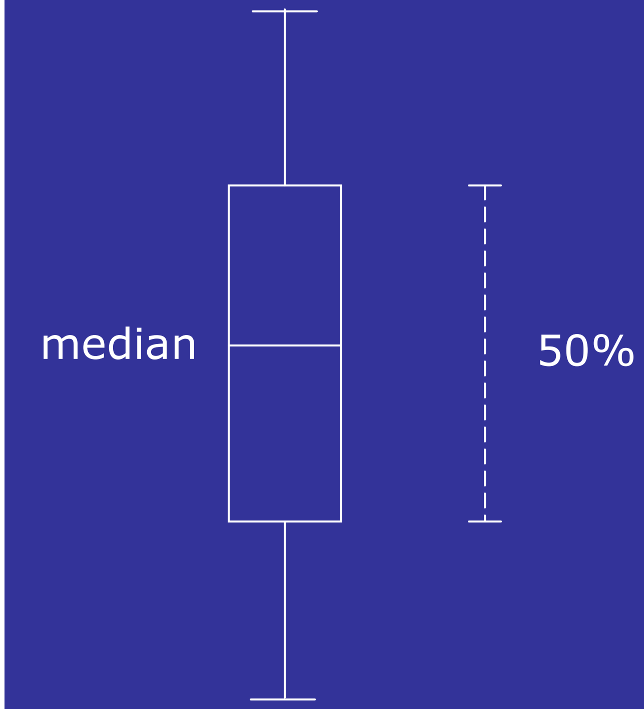
Visualization





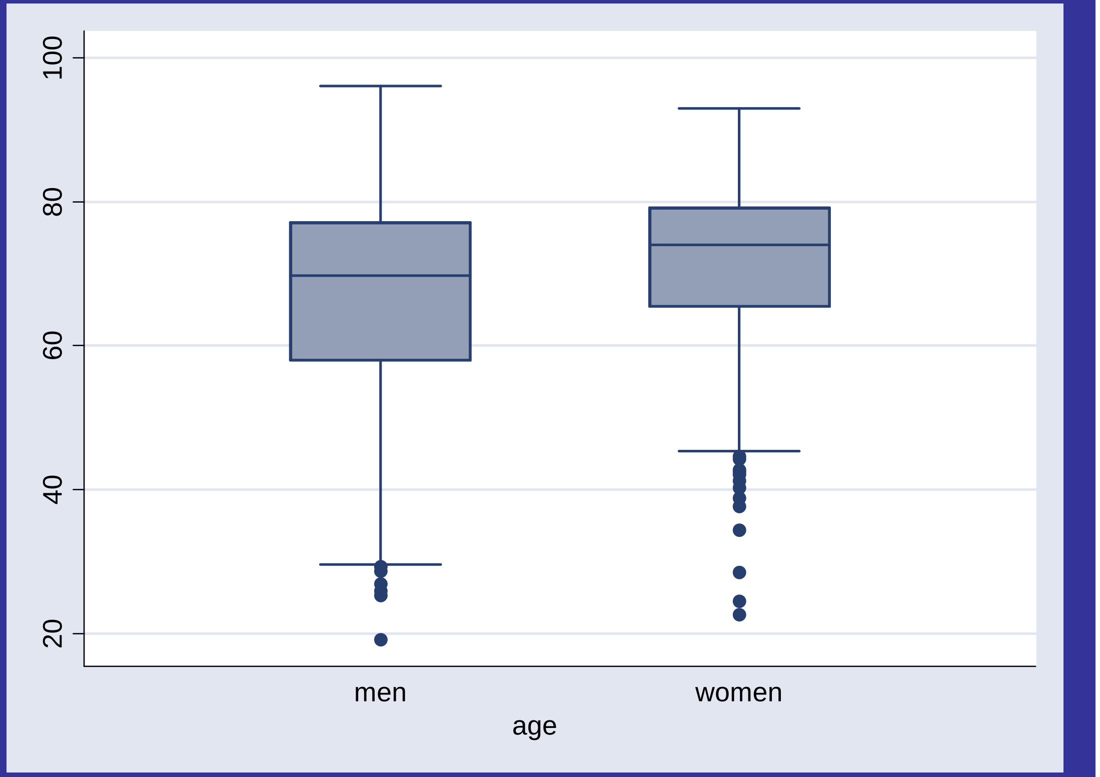


Box plot

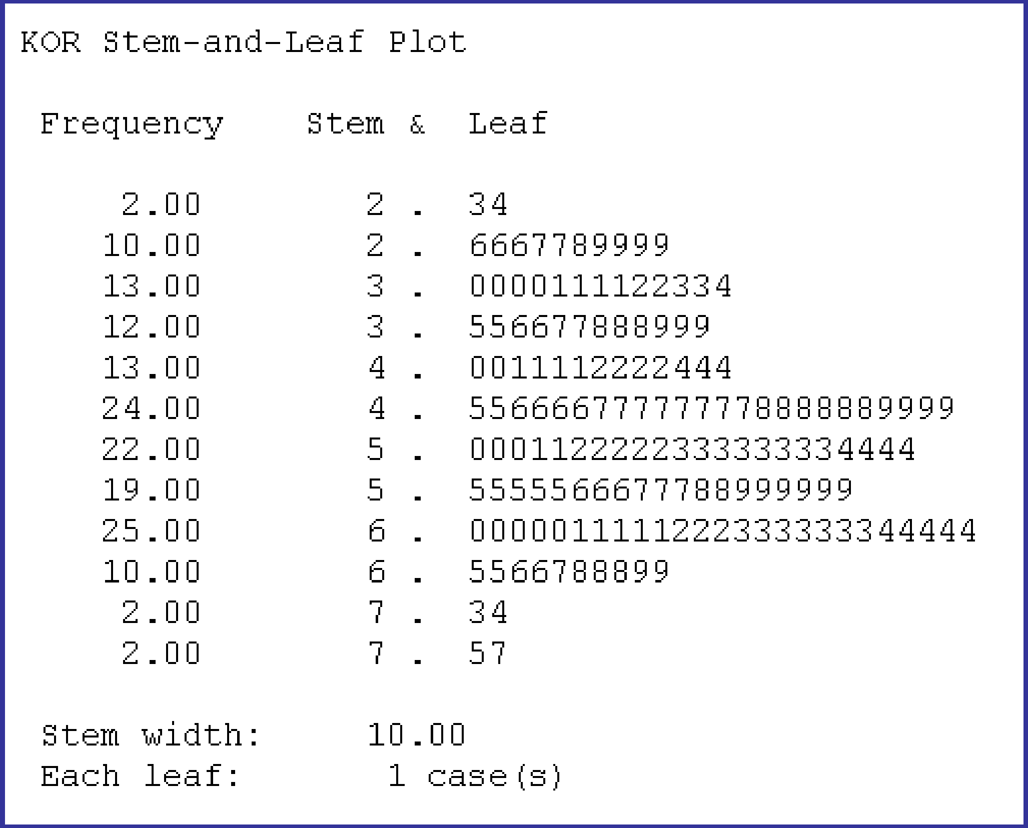


median

50%

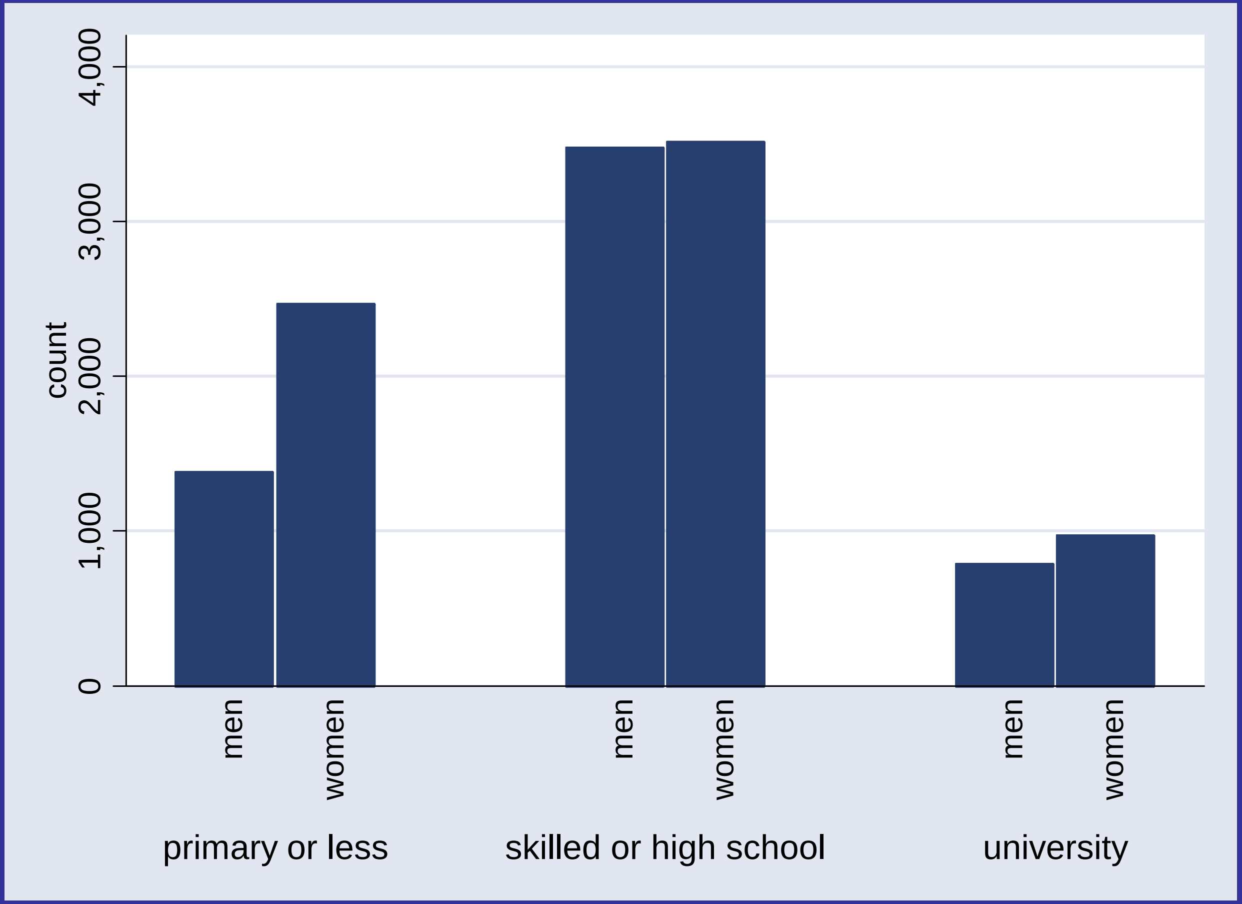


Stem-and leaf plot

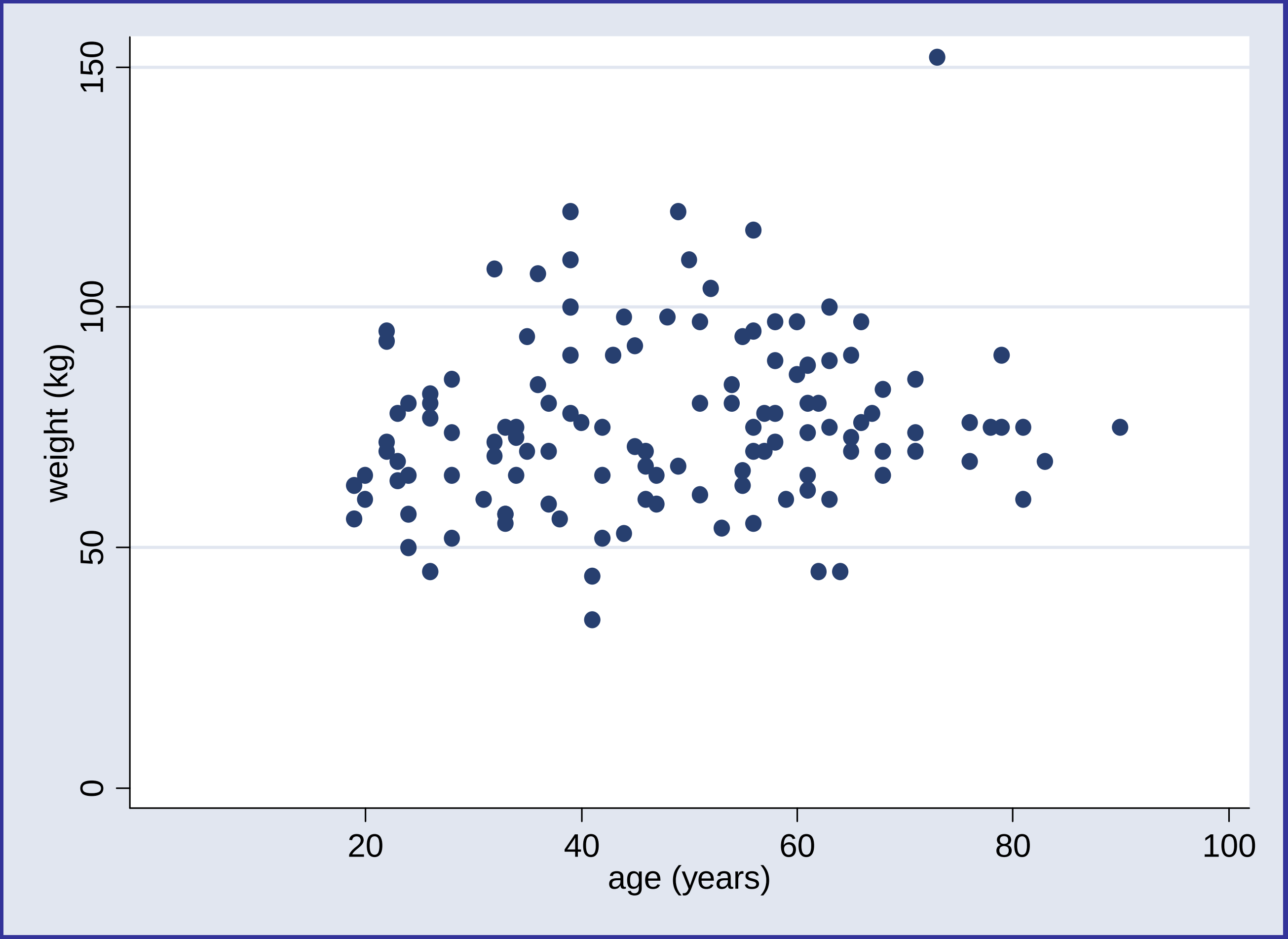


```
KOR Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     2.00        2 .  34
    10.00        2 .  6667789999
    13.00        3 .  0000111122334
    12.00        3 .  556677888999
    13.00        4 .  0011112222444
    24.00        4 .  556666777777788888889999
    22.00        5 .  0001122222233333334444
    19.00        5 .  5555566677788999999
    25.00        6 .  0000011111222333333344444
    10.00        6 .  5566788899
     2.00        7 .  34
     2.00        7 .  57

 Stem width:     10.00
 Each leaf:       1 case(s)
```

Bar chart

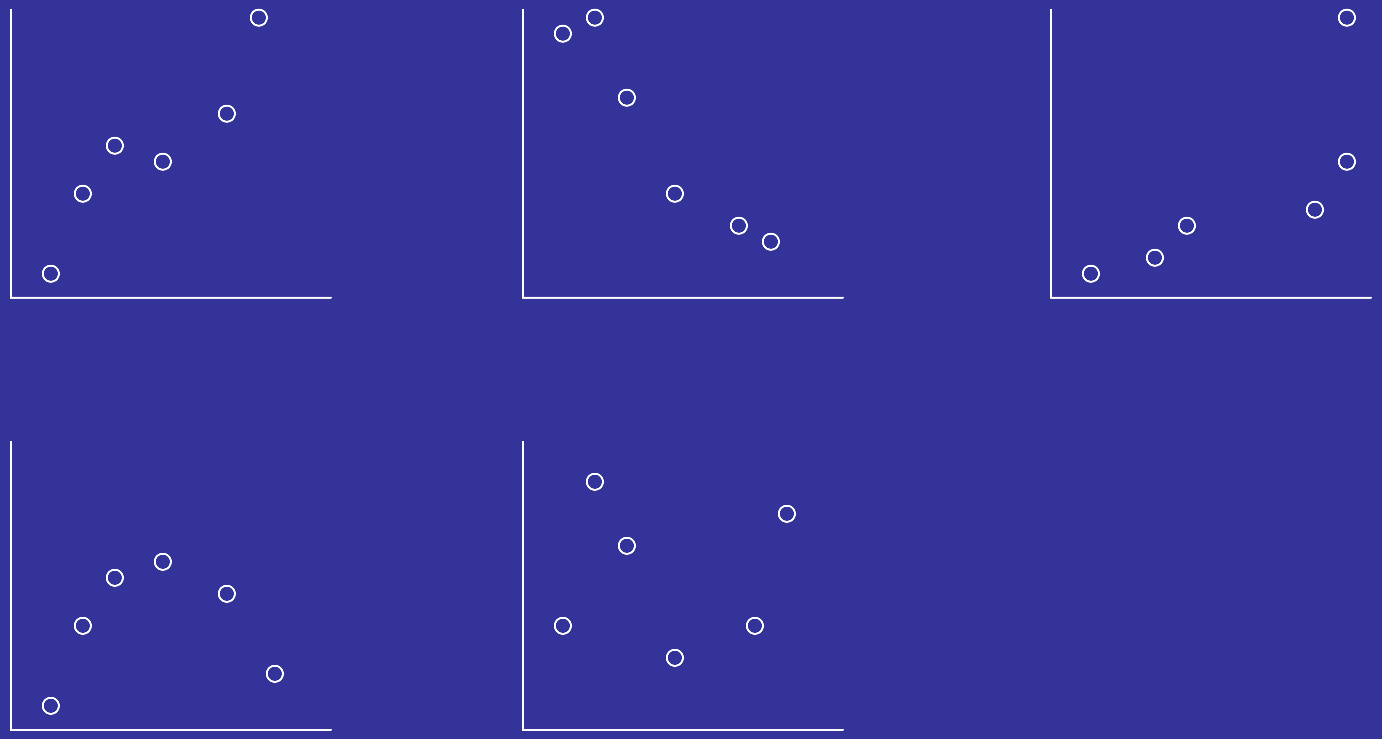

count
0
1,000
2,000
3,000
4,000
men
women
men
women
men
women
primary or less
skilled or high school
university

Scatter plot



n=126

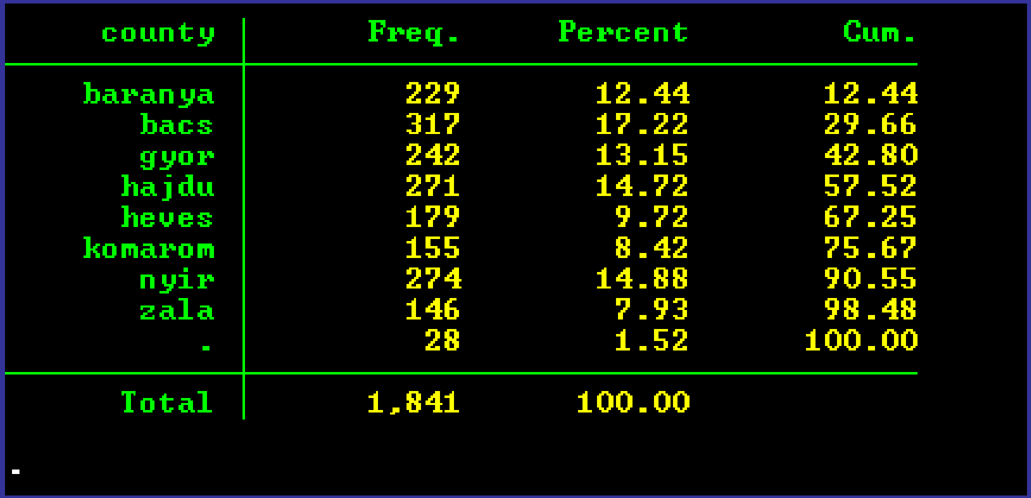
Scatter plot



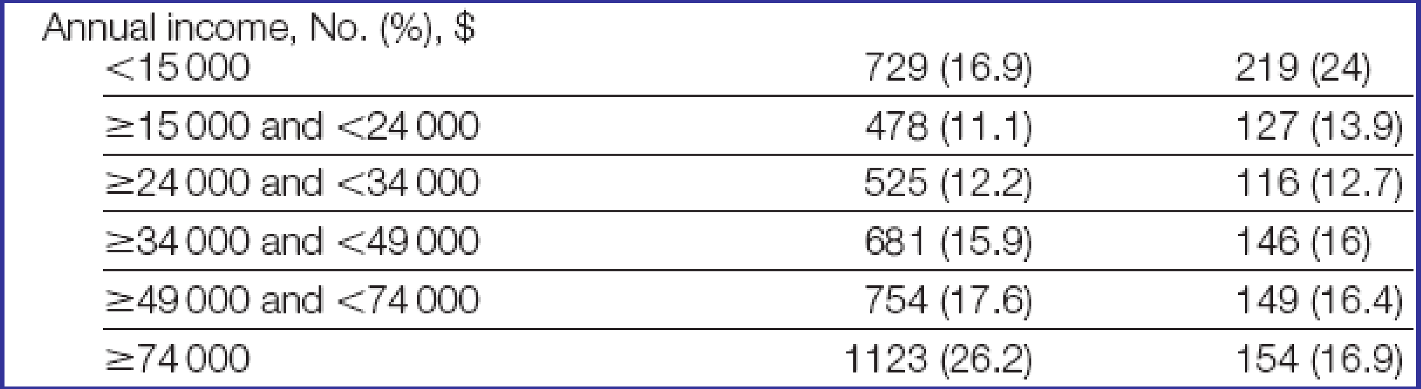
Tables

- Purpose
  - To record numbers for later reference
    - constants
    - normal laboratory values
    - population census
  - To communicate a message
    - scientific report

Frequency tables



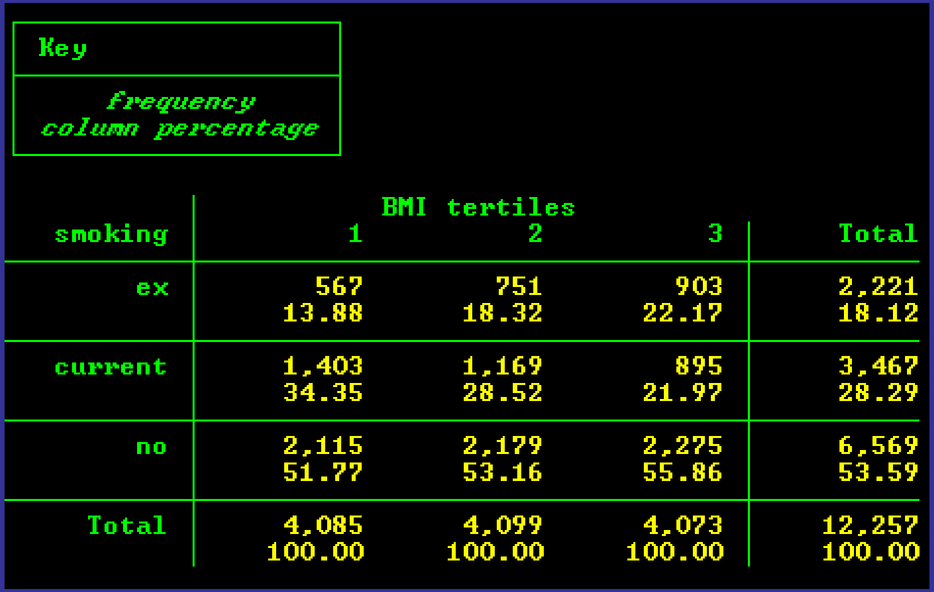
```
     county |      Freq.     Percent        Cum.
------------+-----------------------------------
    baranya |        229       12.44       12.44
       bacs |        317       17.22       29.66
       gyor |        242       13.15       42.80
      hajdu |        271       14.72       57.52
      heves |        179        9.72       67.25
    komarom |        155        8.42       75.67
       nyir |        274       14.88       90.55
       zala |        146        7.93       98.48
          . |         28        1.52      100.00
------------+-----------------------------------
      Total |      1,841      100.00

.
```



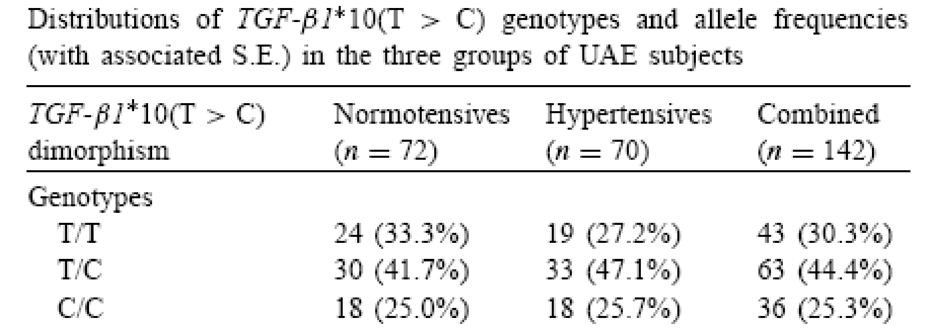
Annual income, No. (%), $		
<15 000	729 (16.9)	219 (24)
≥15 000 and <24 000	478 (11.1)	127 (13.9)
≥24 000 and <34 000	525 (12.2)	116 (12.7)
≥34 000 and <49 000	681 (15.9)	146 (16)
≥49 000 and <74 000	754 (17.6)	149 (16.4)
≥74 000	1123 (26.2)	154 (16.9)

Sherita HG et al. 2008

Contingency tables



```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |             BMI tertiles
   smoking |         1          2          3 |     Total
-----------+---------------------------------+----------
        ex |       567        751        903 |     2,221
           |     13.88      18.32      22.17 |     18.12
-----------+---------------------------------+----------
   current |     1,403      1,169        895 |     3,467
           |     34.35      28.52      21.97 |     28.29
-----------+---------------------------------+----------
        no |     2,115      2,179      2,275 |     6,569
           |     51.77      53.16      55.86 |     53.59
-----------+---------------------------------+----------
     Total |     4,085      4,099      4,073 |    12,257
           |    100.00     100.00     100.00 |    100.00
```



Distributions of *TGF-β1**10(T > C) genotypes and allele frequencies (with associated S.E.) in the three groups of UAE subjects

*TGF-β1**10(T > C) dimorphism	Normotensives (*n* = 72)	Hypertensives (*n* = 70)	Combined (*n* = 142)
Genotypes			
T/T	24 (33.3%)	19 (27.2%)	43 (30.3%)
T/C	30 (41.7%)	33 (47.1%)	63 (44.4%)
C/C	18 (25.0%)	18 (25.7%)	36 (25.3%)

Frossard PM et al. 2001

Contingency tables

Mean Base-Line Laparoscopic Scores for Stage III and IV Endometriosis Patients (Includes Mean, SE.)

Treatment	**Stage III**	**Stage IV**
Nafarelin (800 mg)	25.6 (1.9)	73.8 (2.8)
Nafarelin (400 mg)	26.9 (1.8)	59.0 (2.8)
Danazol (800 mg)	24.6 (1.9)	55.1 (2.9)

Henzl MR et al. 1988

Tabular presentation

Death Rates in Proportions for High Death Rate Operations by Anesthetic Risk Levels

Anesthetic Risk Code	Halothane	Nitrous Oxide	Cyclopropane	Ether	Other
Unknown	0.11369	0.08682	0.08147	0.06148	0.09957
Risk 1	0.02454	0.02452	0.01634	0.01355	0.03358
Risk 2	0.05471	0.06893	0.04941	0.03812	0.05859
Risk 3	0.12471	0.16599	0.18187	0.11453	0.15306
Risk 4	0.15892	0.23140	0.18582	0.17919	0.35531
Risk 5	0.04665	0.06759	0.05725	0.04898	0.07606
Risk 6	0.22143	0.12996	0.17615	0.16008	0.17741
Risk 7	0.44164	0.43689	0.36689	0.62121	0.43348

Bunker JP et al. 1969

Tabular presentation

- Present marginal averages for visual focus
- Order rows and columns in logical order
- Use similar order if when there are multiple similar tables
- Line up numbers to be compared vertically
- Round to two significant digits

Tabular presentation

Death Rates in Percentages for High Death Rate Operations by Anesthetics Versus Anesthetic Risk Levels

	Anesthetic Risk Code								
Anesthetic Group	**1**	**2**	**5**	**Un-known**	**3**	**6**	**4**	**7**	**Weighted average**
Other	3	6	8	10	15	18	36	43	11.7
Nitrous oxide	2	7	7	9	17	13	23	44	10.3
Cyclo-propane	2	5	6	8	18	18	19	37	9.8
Halothane	2	5	5	11	12	22	16	44	8.7
Ether	1	4	5	6	11	16	18	62	6.1
Weighted average	2.2	5.5	57	9.6	14.6	17.4	20.6	42.4	9.3

Bunker JP et al. 1969

Describing with numbers

- Central tendency
  - mean
  - median
  - mode

Arithmetic average

- Commonly used to describe continuous variables and discrete variables with several categories
- Due to its convenient properties many statistical techniques employ the mean
- Suitable for data with symmetric distribution
- It is not appropriate for data measured on a nominal or ordinal scale
- Sensitive for extreme values

$$\overline{X} = \frac{\sum X}{n}$$

Arithmetic average

Age (years)	No. of individuals f	Interval Midpoint X	Relative frequency f/n	X(f/n)
14-16	9	15	0.09	1.35
16-18	13	17	0.13	2.21
18-20	24	19	0.24	4.56
20-22	38	21	0.38	7.98
22-24	16	23	0.16	3.68
	Total: 100		1.00	19.78

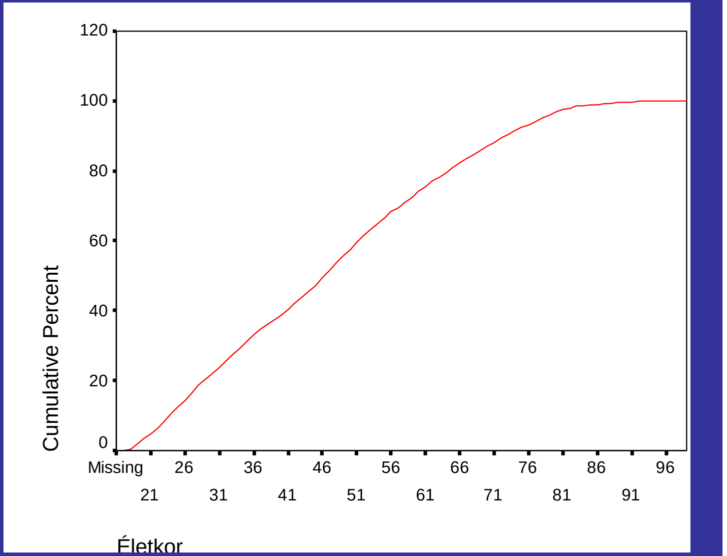
$$\overline{X} = \sum X(f/n)$$

Median

- Central observation when all the data are arranged in increasing sequence
- Can be used with asymmetrically distributed data
- Not influenced by extreme values
- Ordinal, interval or ratio scale

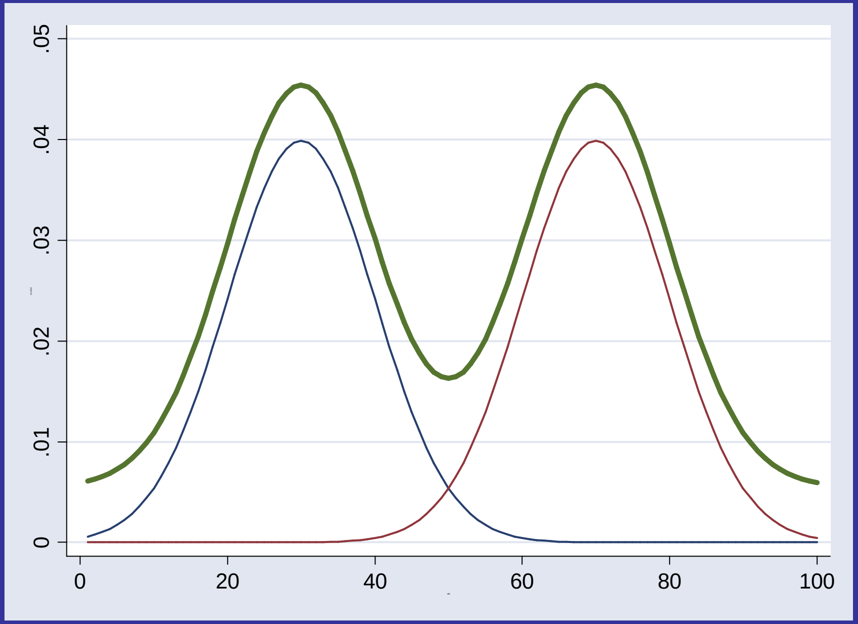
- Data: {2, 4, 11, 15, 16, 21,30}

m M



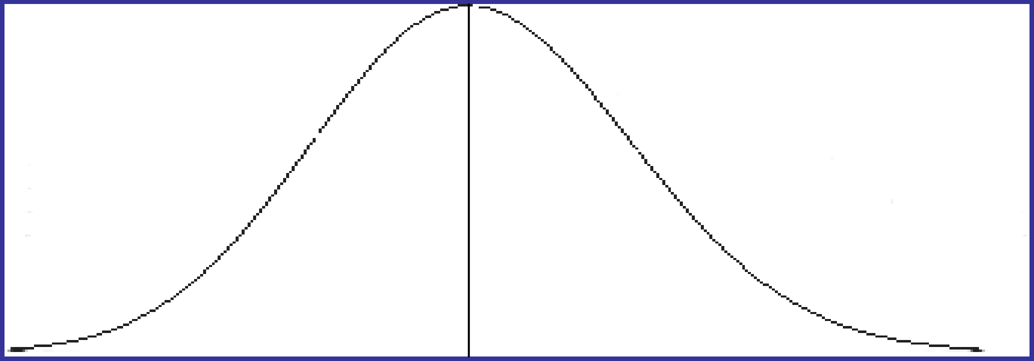
Mode

- The most frequently occurring value
- when data are grouped, the mode is represented by the midpoint of the interval having the greatest class frequency
- Data can show bimodal distribution

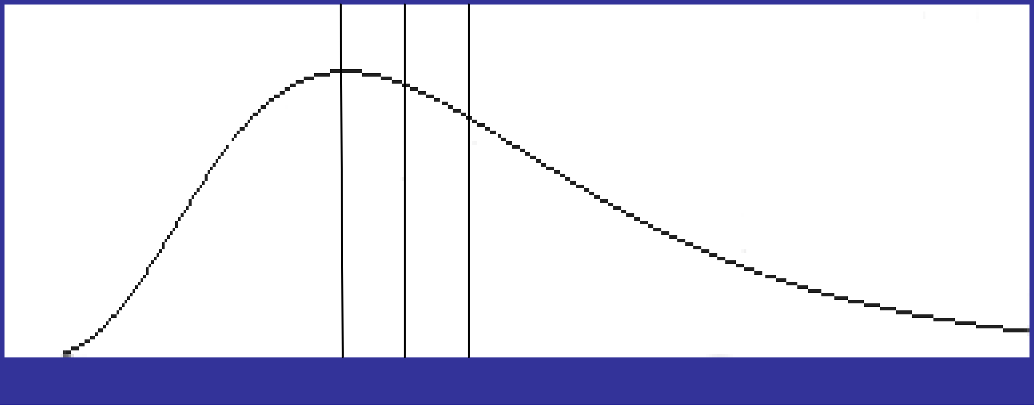


Central tendency

Mean, median, and mode coincide

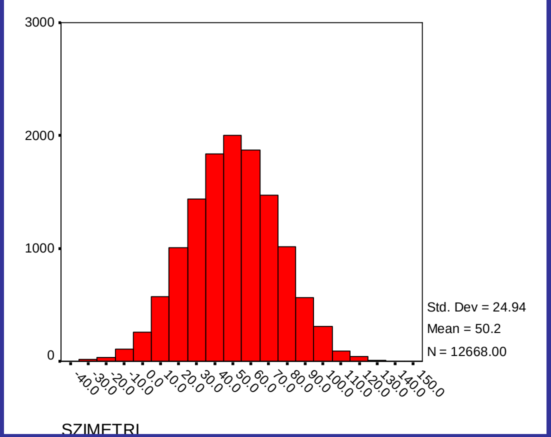


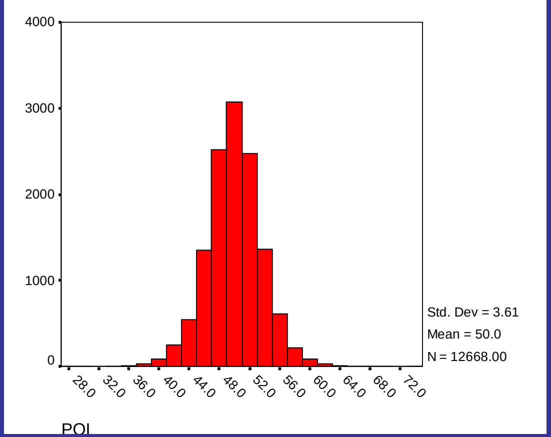
Mean and median will lie to the same side of mode



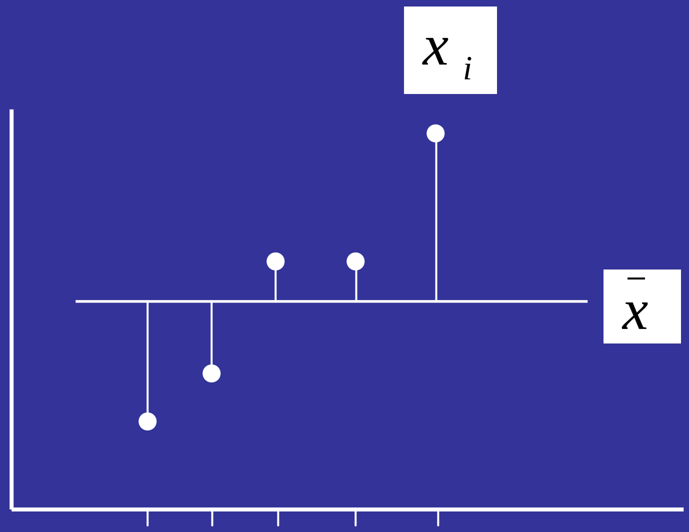
Variability

- {1, 3, 5, 7, 9 } mean:5
- {4, 5, 5, 5, 6 } mean:5





Sum of squared differences



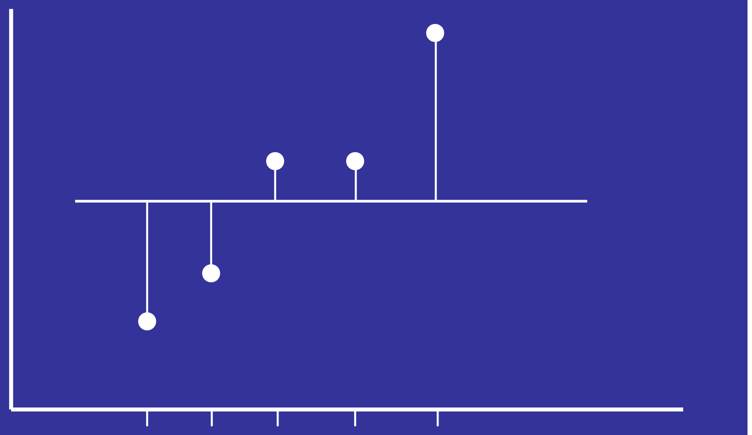
$$\sum (x_i - \bar{x})^2$$

Variability

- – variance (s^2), standard deviation (sd)

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

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

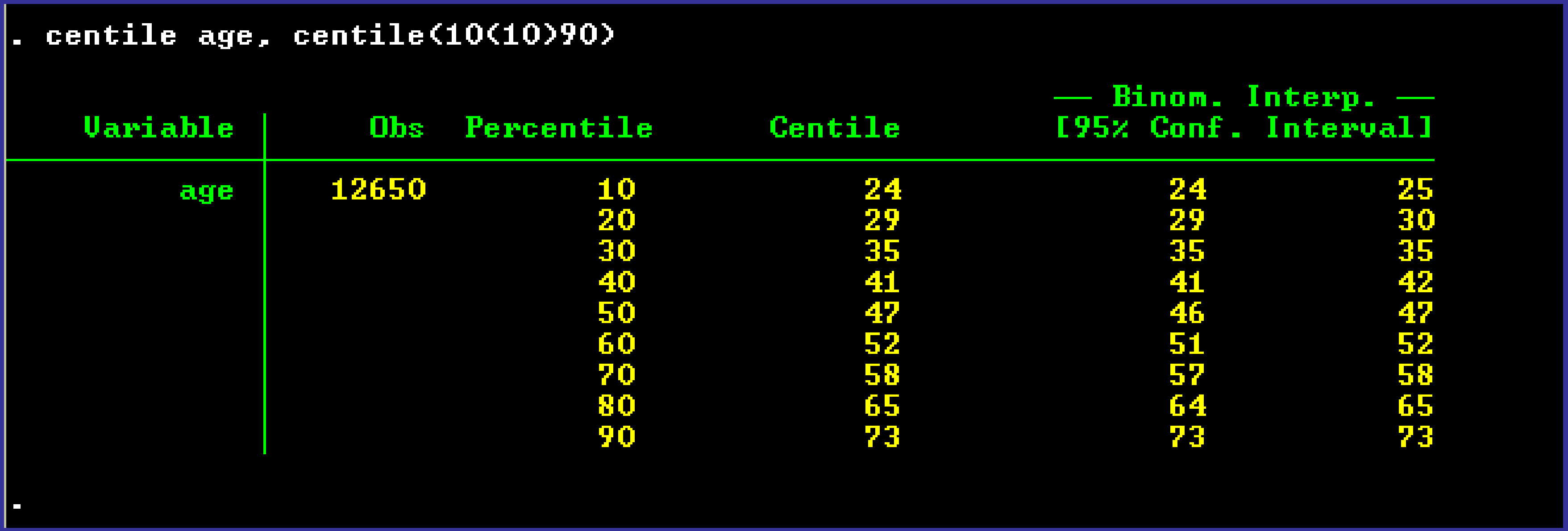


Variability

- Range
  - Highest value minus the lowest value
  - does not take into account values in between

Data: {1, ...?............. 919 }

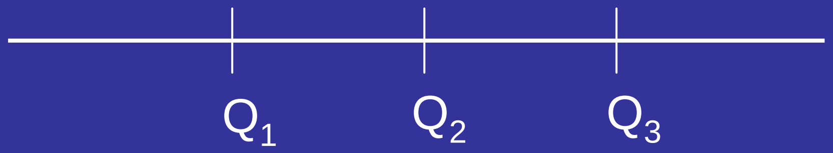
Percentiles



```
. centile age, centile(10(10)90)

                                                       -- Binom. Interp. --
    Variable |     Obs   Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
         age |   12650          10            24              24          25
             |                  20            29              29          30
             |                  30            35              35          35
             |                  40            41              41          42
             |                  50            47              46          47
             |                  60            52              51          52
             |                  70            58              57          58
             |                  80            65              64          65
             |                  90            73              73          73

.
```

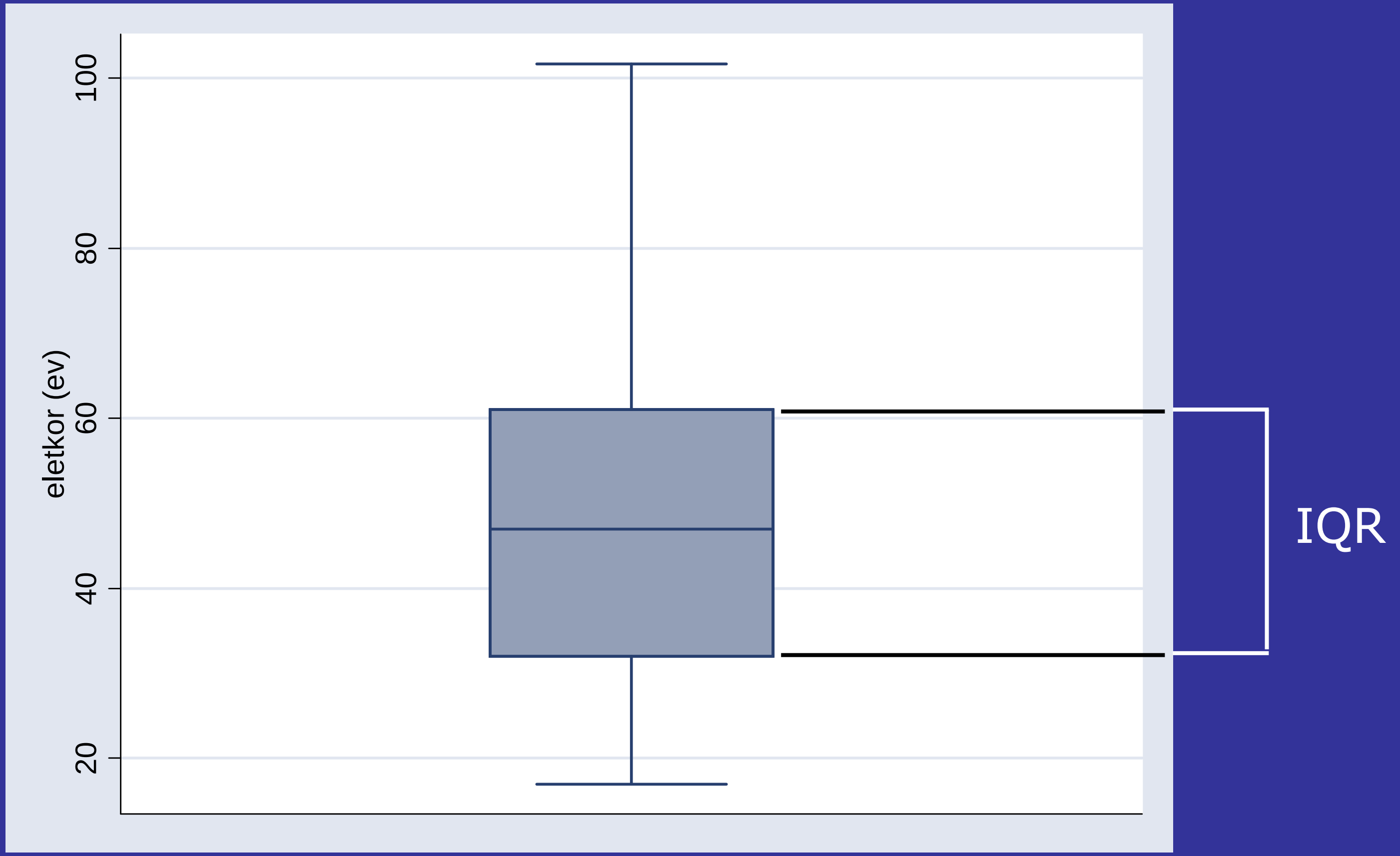
Distribution

Describe data using percentile



- quartile
- percentilis
- interquartile range (IQR): Q_3-Q_1

Interquartile range



Descriptive statistics

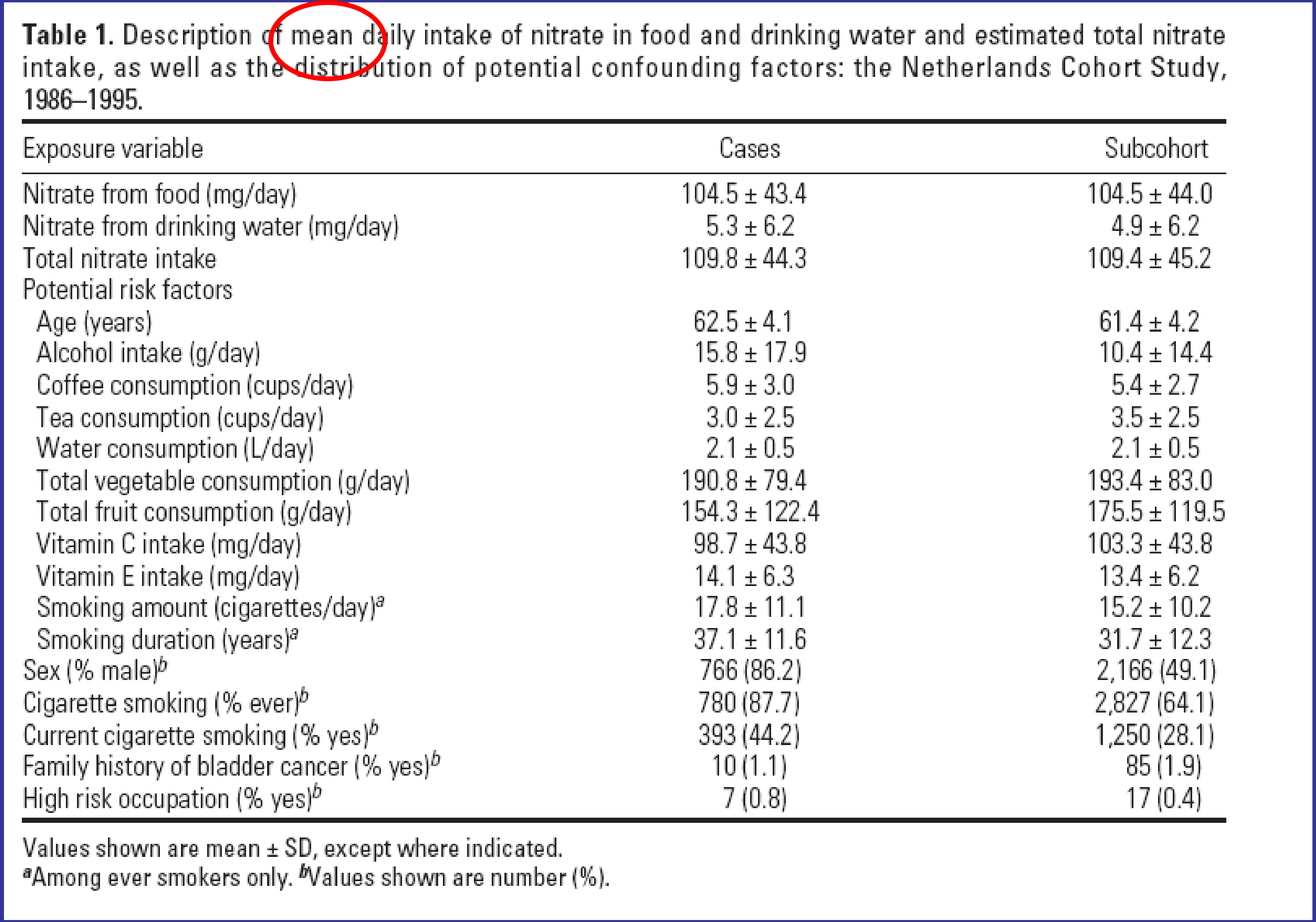


Table 1. Description of mean daily intake of nitrate in food and drinking water and estimated total nitrate intake, as well as the distribution of potential confounding factors: the Netherlands Cohort Study, 1986–1995.

Exposure variable	Cases	Subcohort
Nitrate from food (mg/day)	104.5 ± 43.4	104.5 ± 44.0
Nitrate from drinking water (mg/day)	5.3 ± 6.2	4.9 ± 6.2
Total nitrate intake	109.8 ± 44.3	109.4 ± 45.2
Potential risk factors		
Age (years)	62.5 ± 4.1	61.4 ± 4.2
Alcohol intake (g/day)	15.8 ± 17.9	10.4 ± 14.4
Coffee consumption (cups/day)	5.9 ± 3.0	5.4 ± 2.7
Tea consumption (cups/day)	3.0 ± 2.5	3.5 ± 2.5
Water consumption (L/day)	2.1 ± 0.5	2.1 ± 0.5
Total vegetable consumption (g/day)	190.8 ± 79.4	193.4 ± 83.0
Total fruit consumption (g/day)	154.3 ± 122.4	175.5 ± 119.5
Vitamin C intake (mg/day)	98.7 ± 43.8	103.3 ± 43.8
Vitamin E intake (mg/day)	14.1 ± 6.3	13.4 ± 6.2
Smoking amount (cigarettes/day)[a]	17.8 ± 11.1	15.2 ± 10.2
Smoking duration (years)[a]	37.1 ± 11.6	31.7 ± 12.3
Sex (% male)[b]	766 (86.2)	2,166 (49.1)
Cigarette smoking (% ever)[b]	780 (87.7)	2,827 (64.1)
Current cigarette smoking (% yes)[b]	393 (44.2)	1,250 (28.1)
Family history of bladder cancer (% yes)[b]	10 (1.1)	85 (1.9)
High risk occupation (% yes)[b]	7 (0.8)	17 (0.4)

Values shown are mean ± SD, except where indicated.
[a] Among ever smokers only. [b] Values shown are number (%).

Zeegers MP. et al. 2006

Descriptive statistics

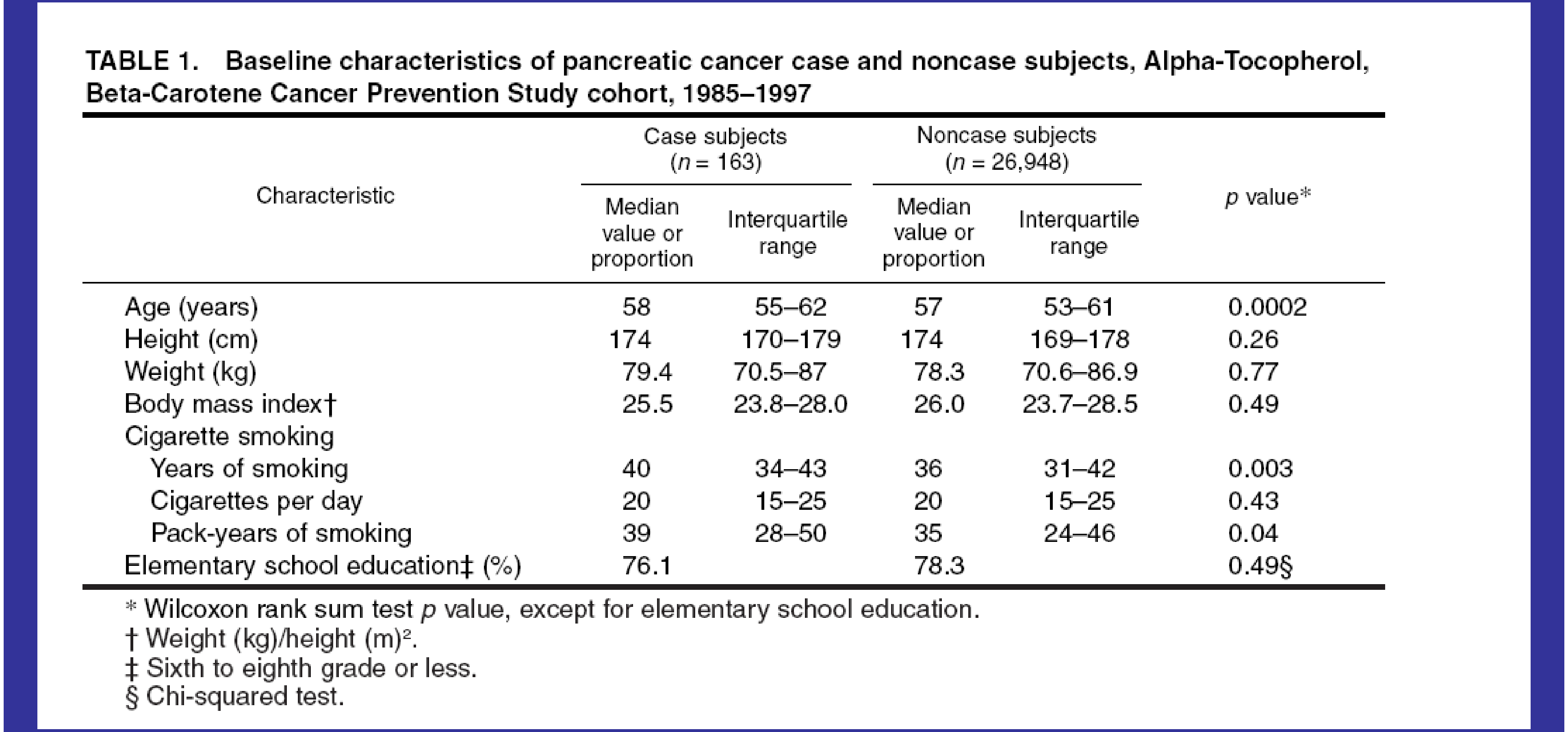


TABLE 1. Baseline characteristics of pancreatic cancer case and noncase subjects, Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study cohort, 1985–1997

Characteristic	Case subjects (*n* = 163)		Noncase subjects (*n* = 26,948)		*p* value*
	Median value or proportion	Interquartile range	Median value or proportion	Interquartile range	
Age (years)	58	55–62	57	53–61	0.0002
Height (cm)	174	170–179	174	169–178	0.26
Weight (kg)	79.4	70.5–87	78.3	70.6–86.9	0.77
Body mass index†	25.5	23.8–28.0	26.0	23.7–28.5	0.49
Cigarette smoking					
Years of smoking	40	34–43	36	31–42	0.003
Cigarettes per day	20	15–25	20	15–25	0.43
Pack-years of smoking	39	28–50	35	24–46	0.04
Elementary school education‡ (%)	76.1		78.3		0.49§

* Wilcoxon rank sum test *p* value, except for elementary school education.

† Weight (kg)/height (m)^2.

‡ Sixth to eighth grade or less.

§ Chi-squared test.

Stolzenberg-Solomon RZ. et al. 2002

References

- Tufte ER. The visual display of quantitative information. Cheshire, Conn.: Graphic Press, 1983
- Ehrenberg AC. The problem of numeracy. Am Statisticain 1981;35:67-71.
- Bailer III JC. and Mosteller F. Medical Uses of Statistics 2nd edition Boston, Massachusetts: NEJM Books, 1992