Universidad del Cauca

ERASMUS UNIVERSITEIT ROTTERDAM

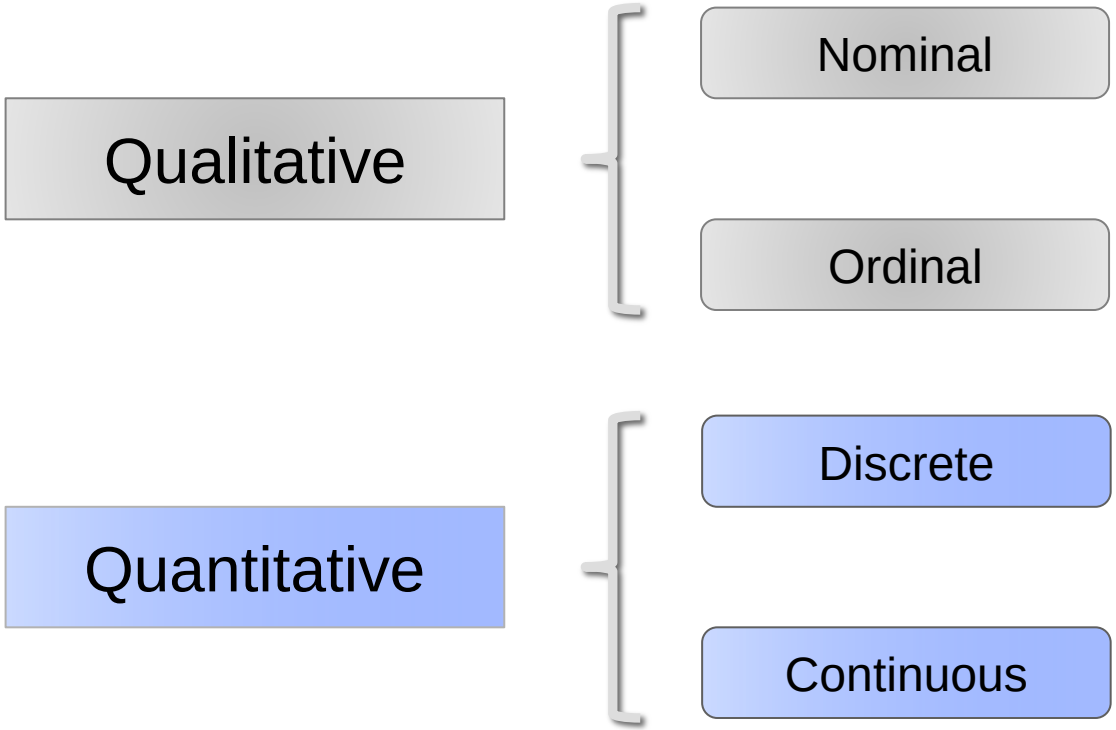
GRAPHS AND DATA DESCRIPTION

Jose Andres Calvache MD MSc PhD

Department of Anesthesiology, Universidad del Cauca, Colombia

Department of Anesthesiology, Easmus University Medical Centre, Rotterdam, The Netherlands

Variable


Qualitative
Nominal
Ordinal
Quantitative
Discrete
Continuous

Qualitative variables (n, %)

- Absolute frequencies (n) and proportions (%).
- Sample (X_1, X_2,.. X_n).
- Each cathegory *j*(1,2…*k*).

$$p_j = \frac{n_j \times 100\%}{n} \; (j = 1, \ldots, k)$$

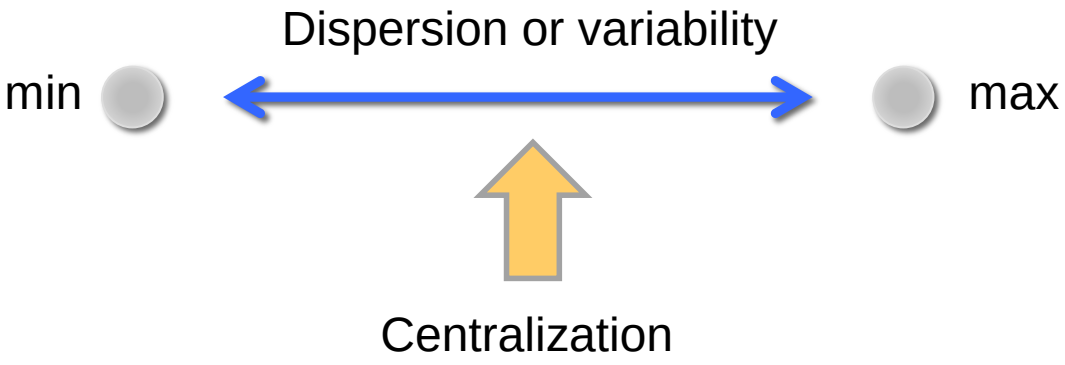
Pain level	**Frequency (n)**	**%**
None	**10**	20
Mild	15	30
Moderate	15	30
Severe	10	20

$$\frac{\mathbf{10} \text{ x } 100\%}{50}$$

Quantitative variables

- Glycemic values of 20 medical students (mg/dL)

110	90	70	45	85	110
112	114	108	48	140	100
113	100	87	95	70	109

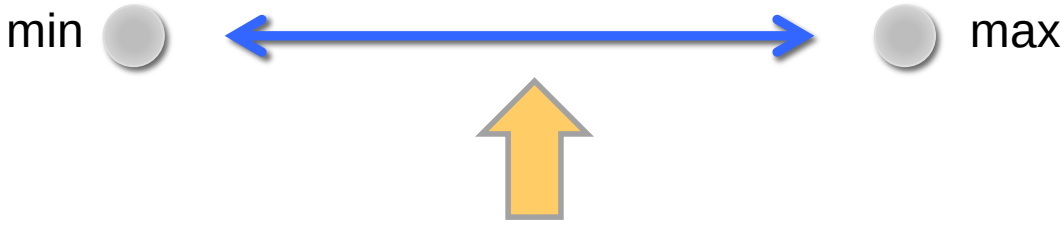


Range

Interquartile range

Standard deviation

Dispersion or variability



Centralization

Mean / average

Median

Mode

Mean ($\bar{\chi}$)

- Arithmetic mean or average
- Sample (X_1, X_2,… X_n)

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - Centralization statistic
  - Sensitive to extreme values of data
  - Gravity center of data

Standard deviation (*SD*)

- Lets consider a sample (X_1, X_2,… X_n) with mean ($\bar{\chi}$)

$$SD = \sqrt{\frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}} = s$$

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

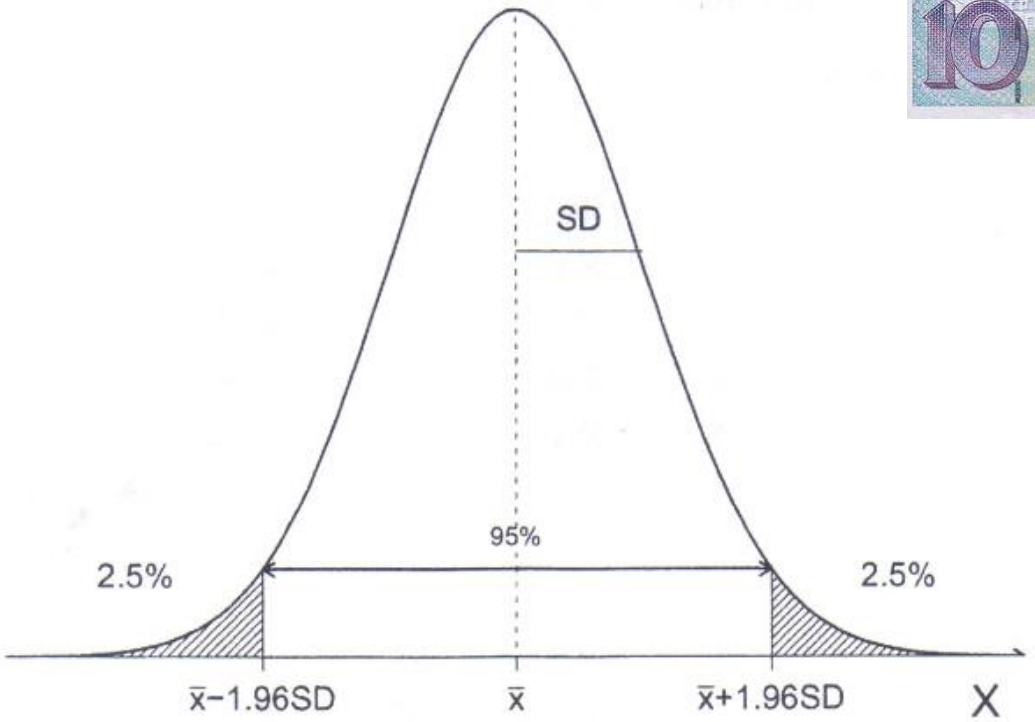
- Measure of variability : $s \geq 0$
- Sensitive to extreme values
- Num: Sum of squares → $\sum(x - \bar{x})^2$
- s^2: variance

Characteristics of the 355 participating doctors

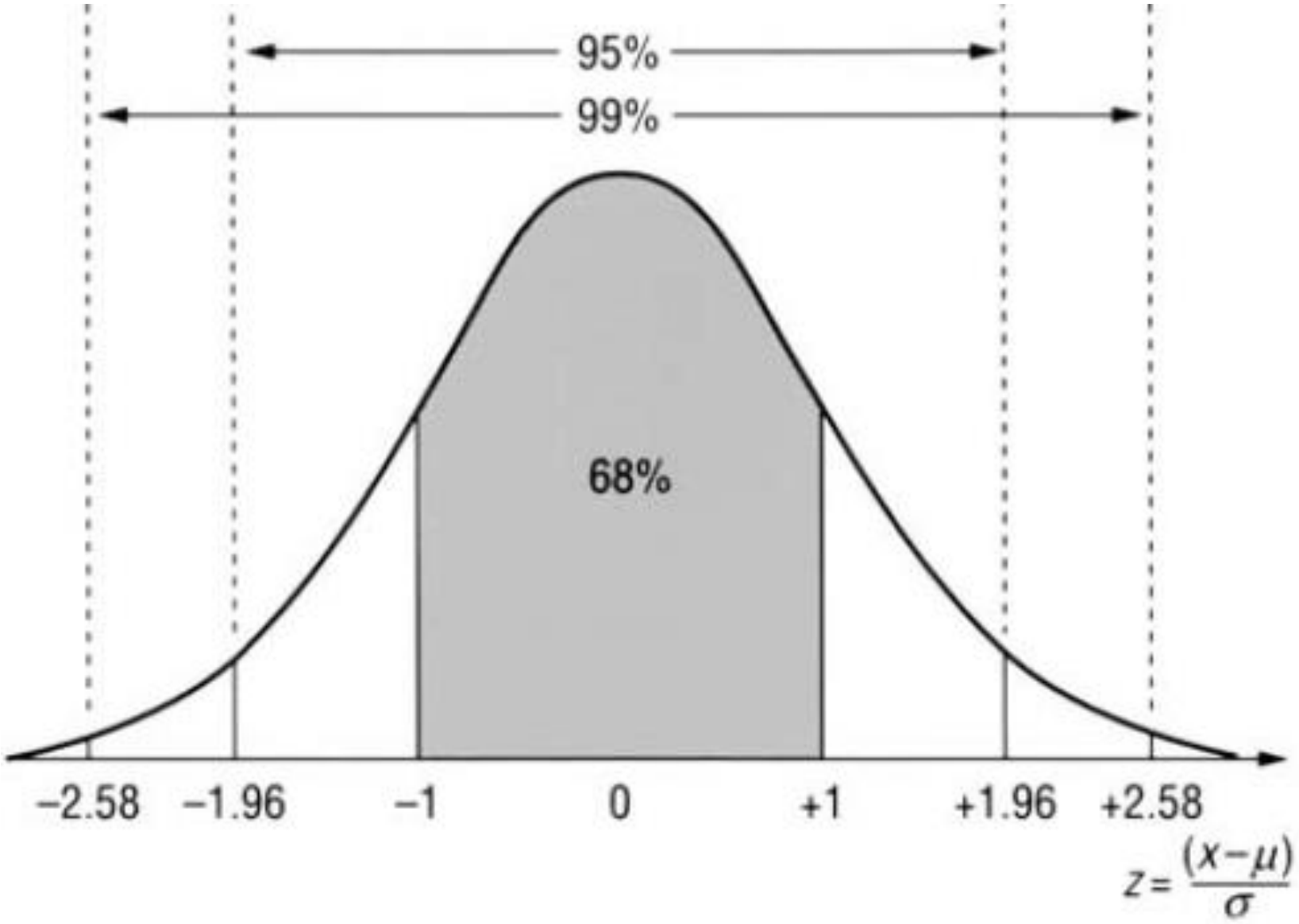
Characteristics	Mean ± *SD*	Frequency	%
Age (years)	43.8 ± 10.		
Sex			
Male		295	83
Female		60	17
Certification			
Yes		326	93
No		26	7
Professional experience (years)	18.0 ± 10.5		
Size of practice (patients per week)			
≤ 30		27	8
31–60		70	20
61–90		84	24
> 90		171	48

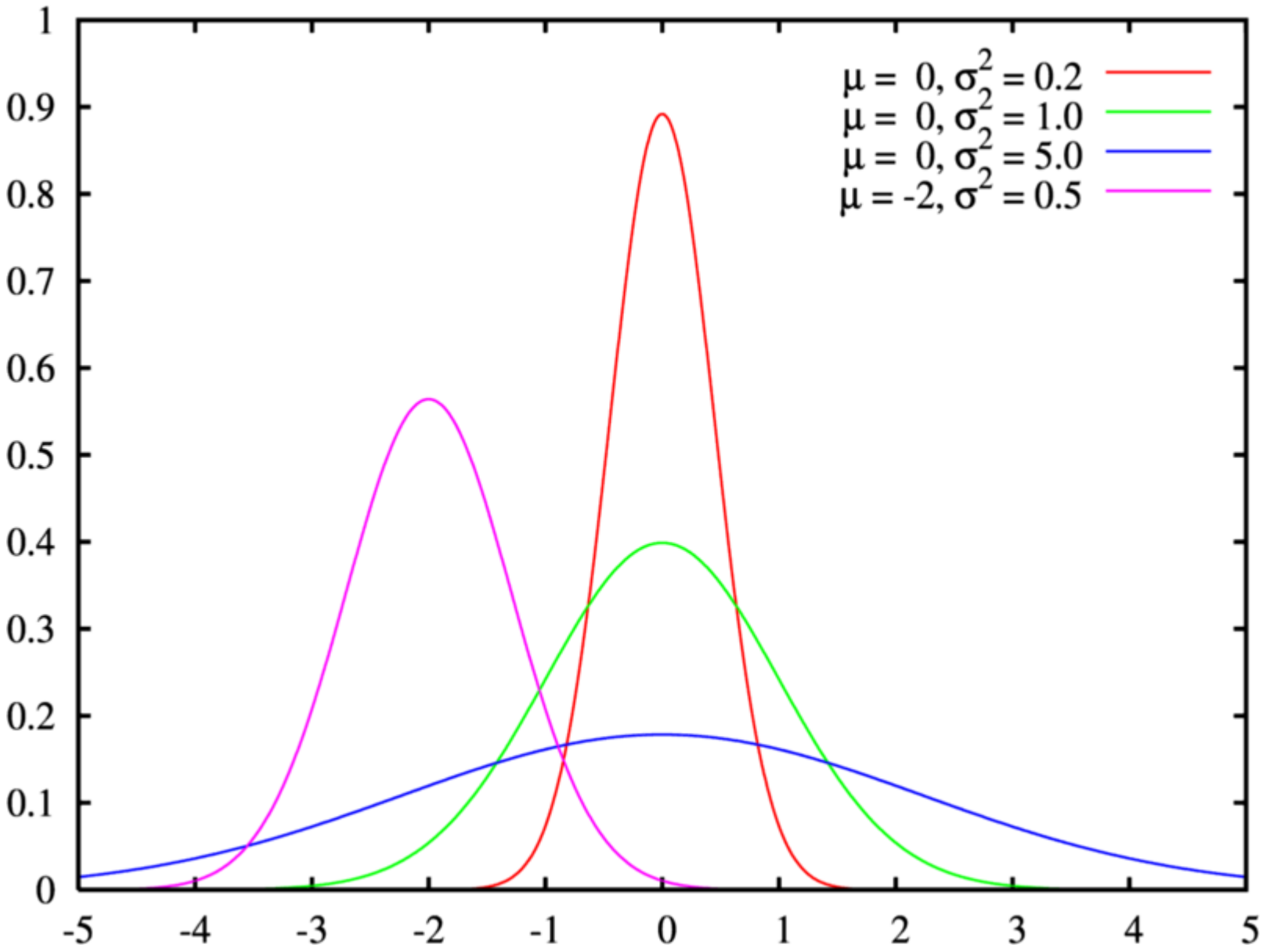
Normal distribution

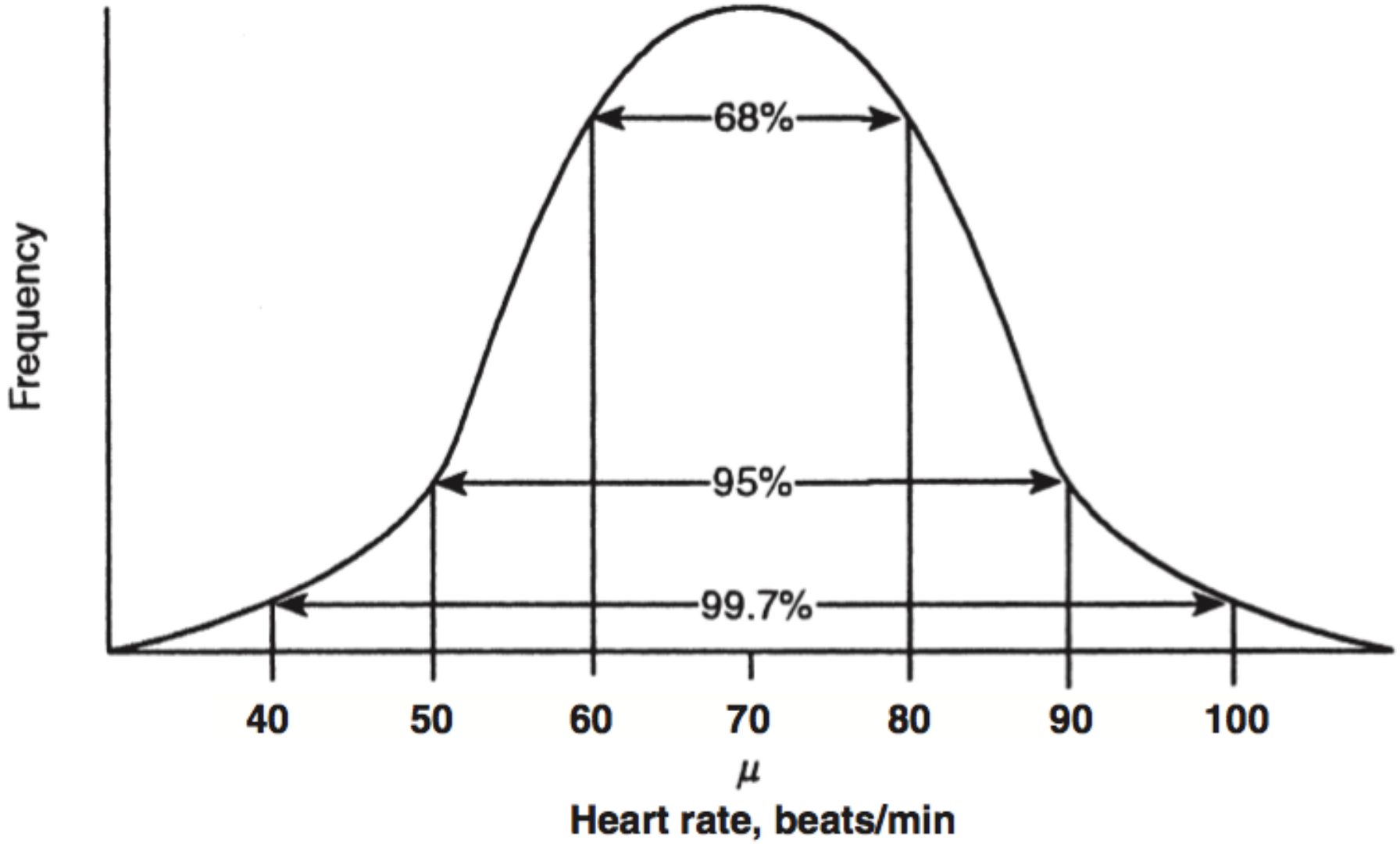
1777 – 1855
Carl Friedrich Gauss



- Two parameters: mean and SD
- Simetry around its mean

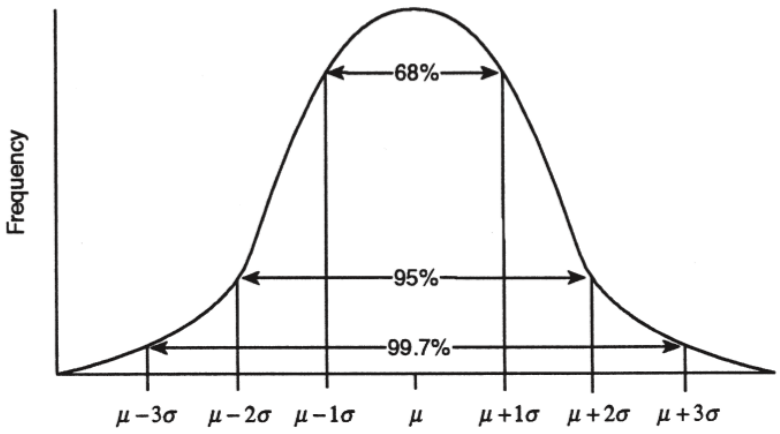

95%
99%
68%
−2.58
−1.96
−1
0
+1
+1.96
+2.58
$z = \frac{(x-\mu)}{\sigma}$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
μ = 0, σ2 = 0.2
μ = 0, σ2 = 1.0
μ = 0, σ2 = 5.0
μ = -2, σ2 = 0.5


Frequency
68%
95%
99.7%
40
50
60
70
80
90
100
μ
Heart rate, beats/min

SD and normal distribution

- Mean ± 1 SD contains 68% of the data
- Mean ± 2 SD contains 95% of the data
- Mean ± 3 SD contains 99,7% of the data



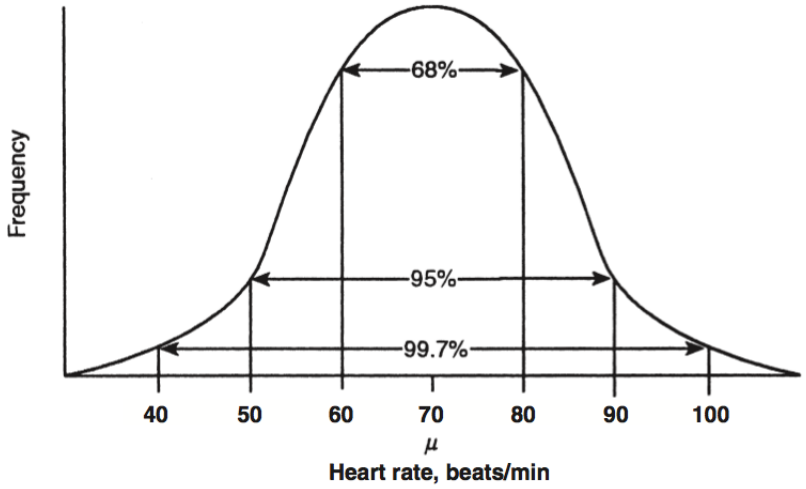
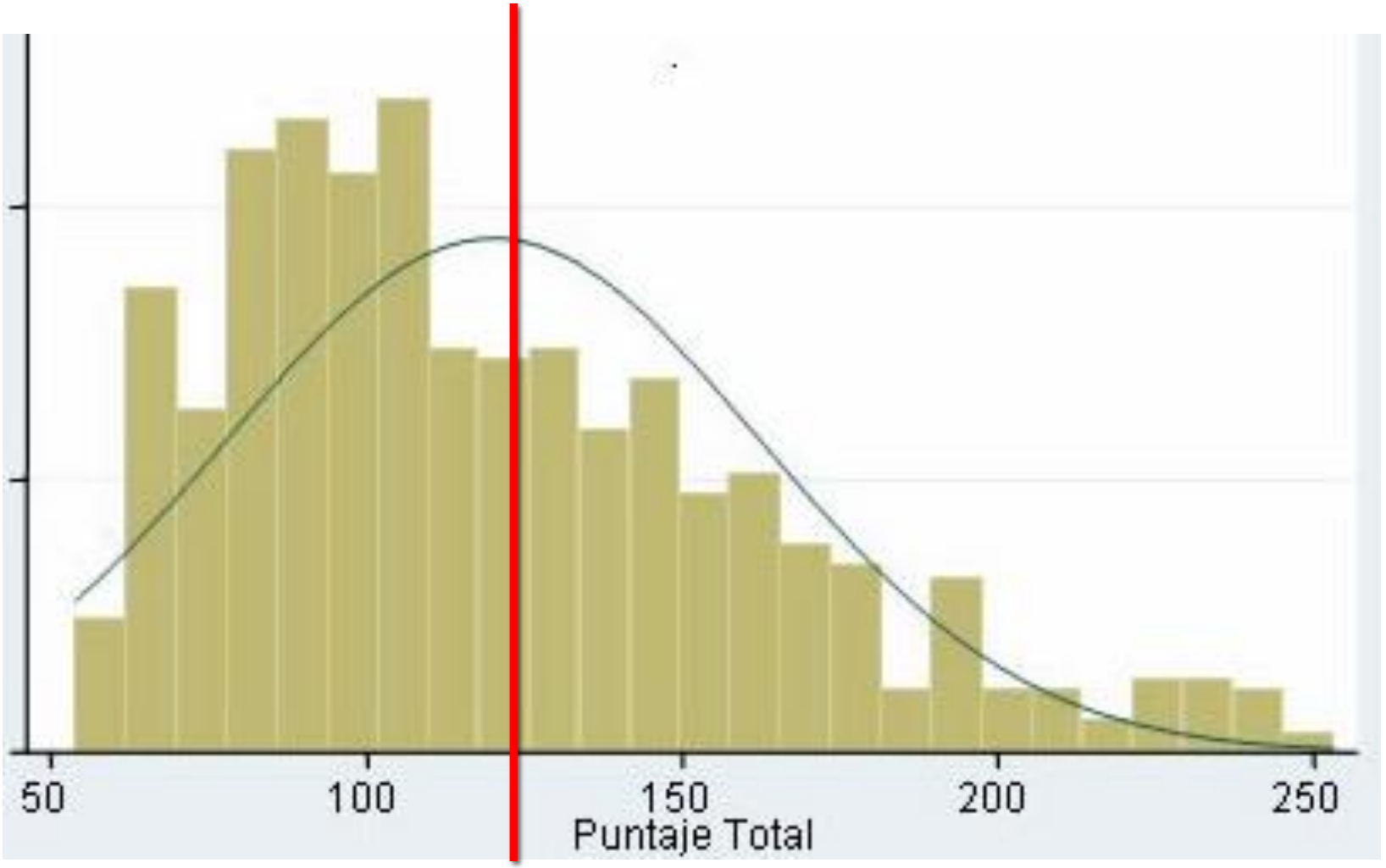


Table 1

Demographics and peri-operative data.

	Placebo $n = 34$	Morphine 100 µg $n = 34$
Age (year)	51 ± 8 (38–69)	51 ± 8 (38–66)
Weight (kg)	71 ± 12	70 ± 9
Length (cm)	166 ± 6	166 ± 7
Diagnosis (*n*) myoma uteri	24	23
sarcoma/carcinoma uteri	10	11
Duration of surgery (min)	95 ± 27	102 ± 29
Range	(42–163)	(60–170)
Blood loss during surgery (ml)	307 ± 246	314 ± 221
Range	(0–1050)	(0–800)

Values are expressed as mean, SD (range), and number of patients (*n*).



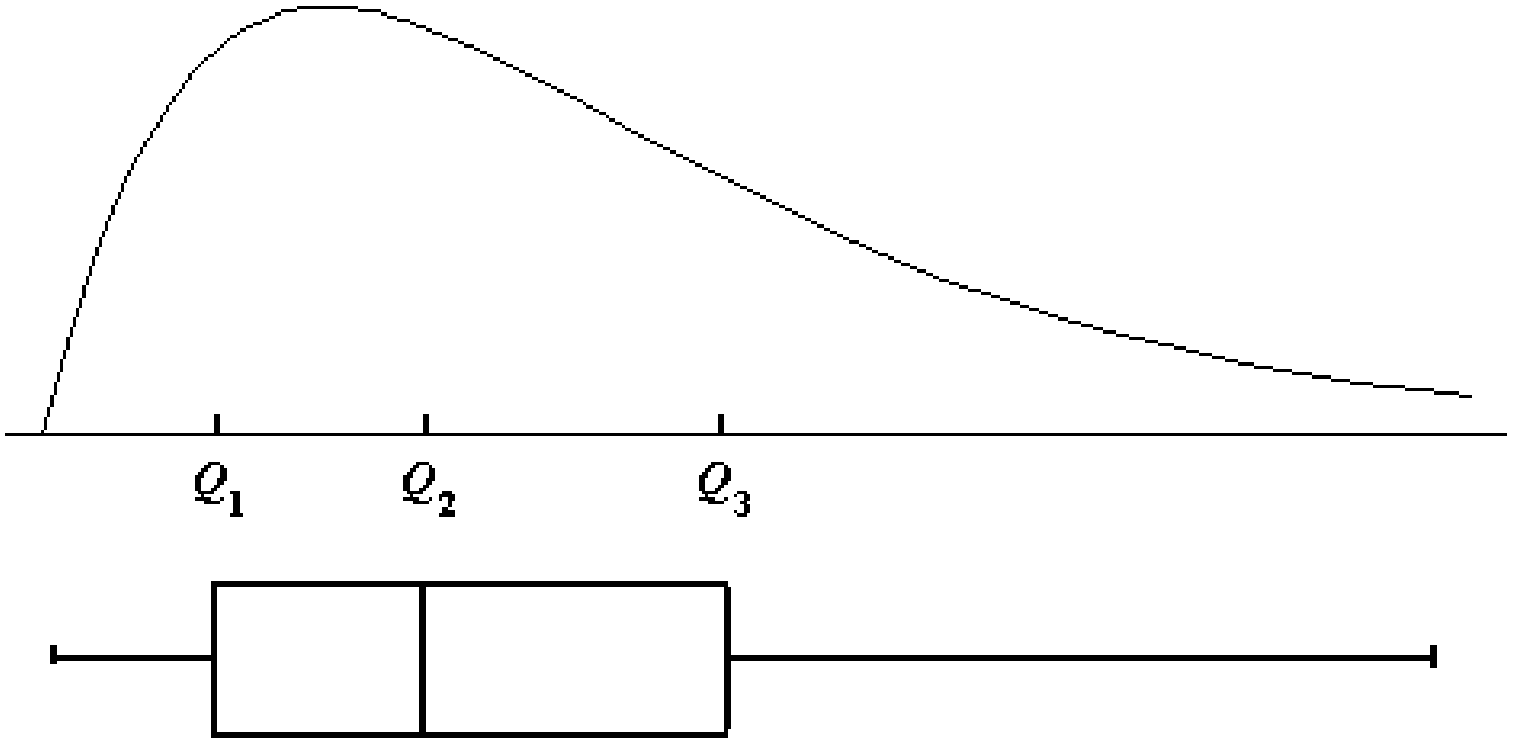
$\bar{x}$

Median and quartiles

- Asymmetric distribution
- Non parametric distribution

- Median: centralization
- Quartiles: variability
- Interquartile range (IQR): variability

- 10 percentile
- P_{10}: left 10% of data below this boundary and 90% above.


Q_1
Q_2
Q_3

Median

- Sample (X_1, X_2,… X_n)
- Reorder data from lowest to highest

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

- $X_{(1)}$: minimun; y $X_{(n)}$: maximum. Between them «range»

$$x_{\left(\frac{n+1}{2}\right)} \quad \left| \quad \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \right.$$

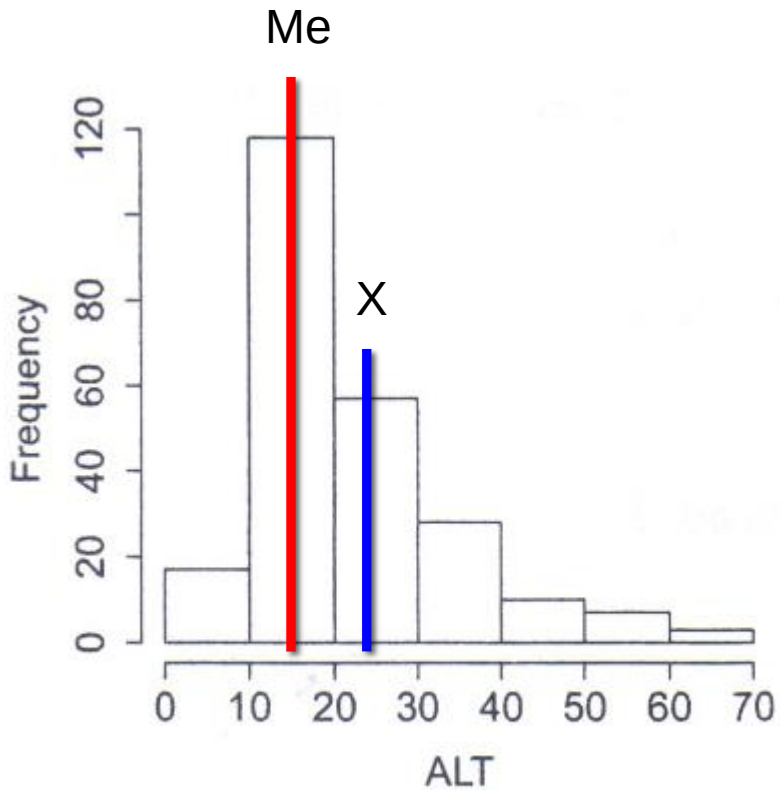
- It is not affected by the extreme data. Robust
- Useful in asymmetric distributions
- $P_{(50)}$ = Median = Quartile 2 (Q2)

Data from 240 medical students (ALT)

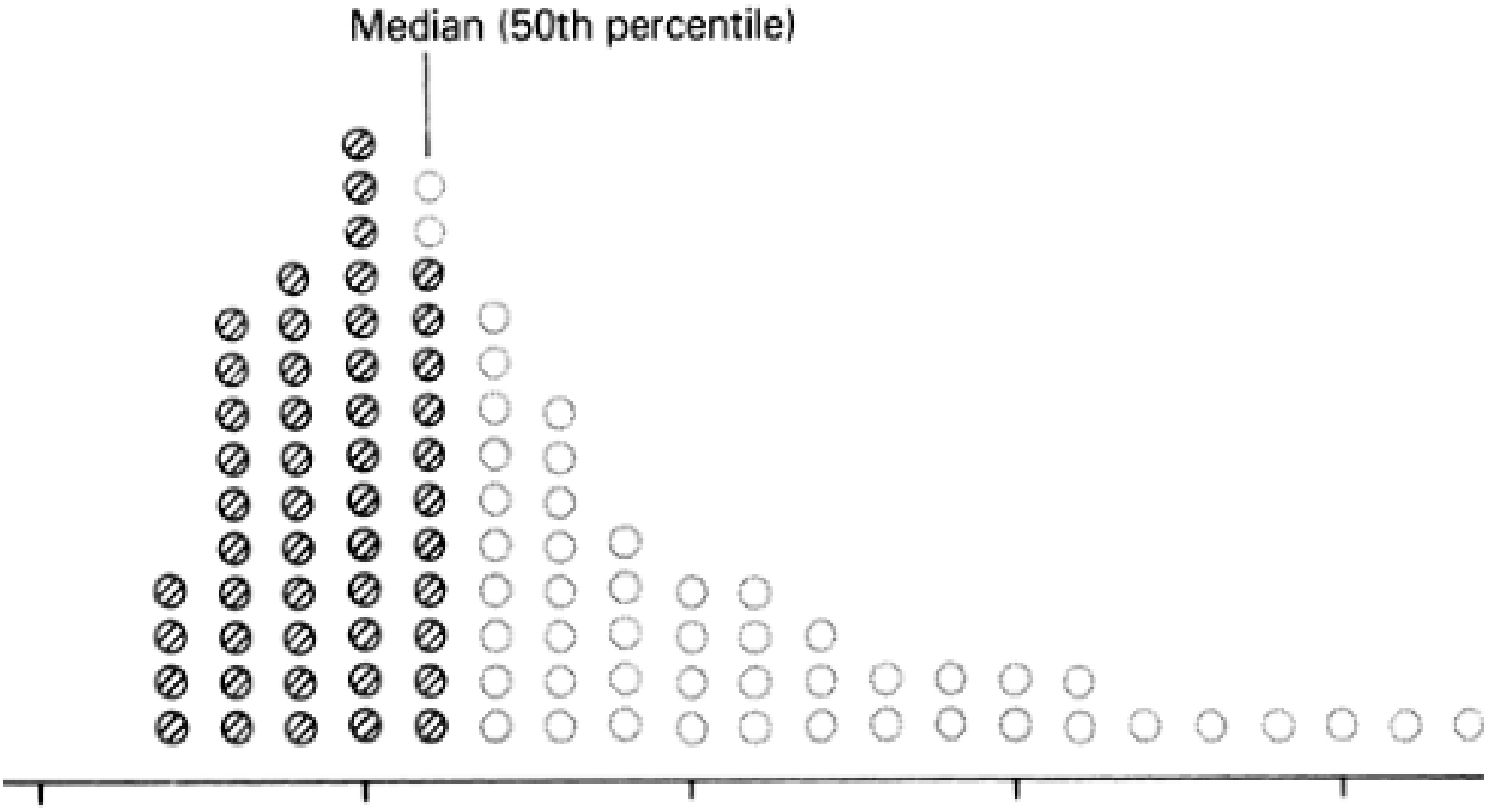
5	6	6	6	7	8	8	8	8	8	9	9	9	10	10	10	10	11	11	11
11	11	11	11	11	11	11	11	12	12	12	12	12	12	12	12	12	12	12	12
12	13	13	13	13	13	13	13	13	13	13	13	13	13	14	14	14	14	14	14
14	14	14	14	14	14	14	15	15	15	15	15	15	15	15	15	15	16	16	16
16	16	16	16	16	16	16	16	17	17	17	17	17	17	17	17	17	18	18	18
18	18	18	18	18	18	18	19	19	19	19	19	19	19	19	19	19	19	19	**19**
20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	21	21	21	21	21
21	21	21	21	21	21	22	22	22	22	22	22	22	22	23	23	23	23	23	24
24	24	25	25	25	25	25	25	25	25	25	25	25	26	26	26	26	26	27	28
28	28	28	28	28	29	29	30	30	30	30	30	31	31	31	31	31	32	33	34
34	35	35	36	36	36	36	36	36	37	37	37	38	38	39	39	39	40	40	40
41	42	45	45	46	47	47	48	48	49	51	51	51	53	54	55	55	62	65	69

$$\text{Median} = \frac{x_{(120)} + x_{(121)}}{2} = \frac{19+20}{2} = 19.5 \text{ IU/L}$$

Data from 240 medical students (ALT)



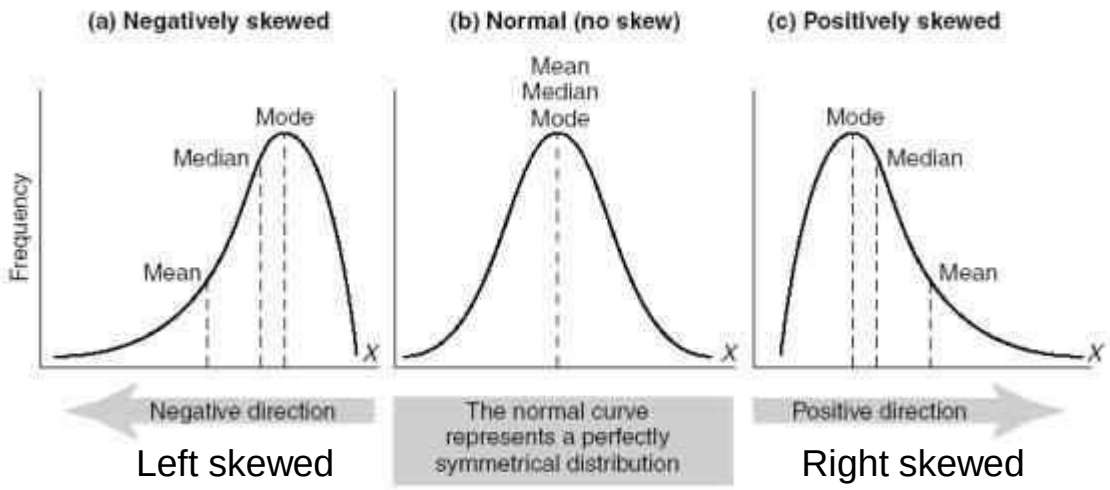
22.5 ± 11.86


Median (50th percentile)

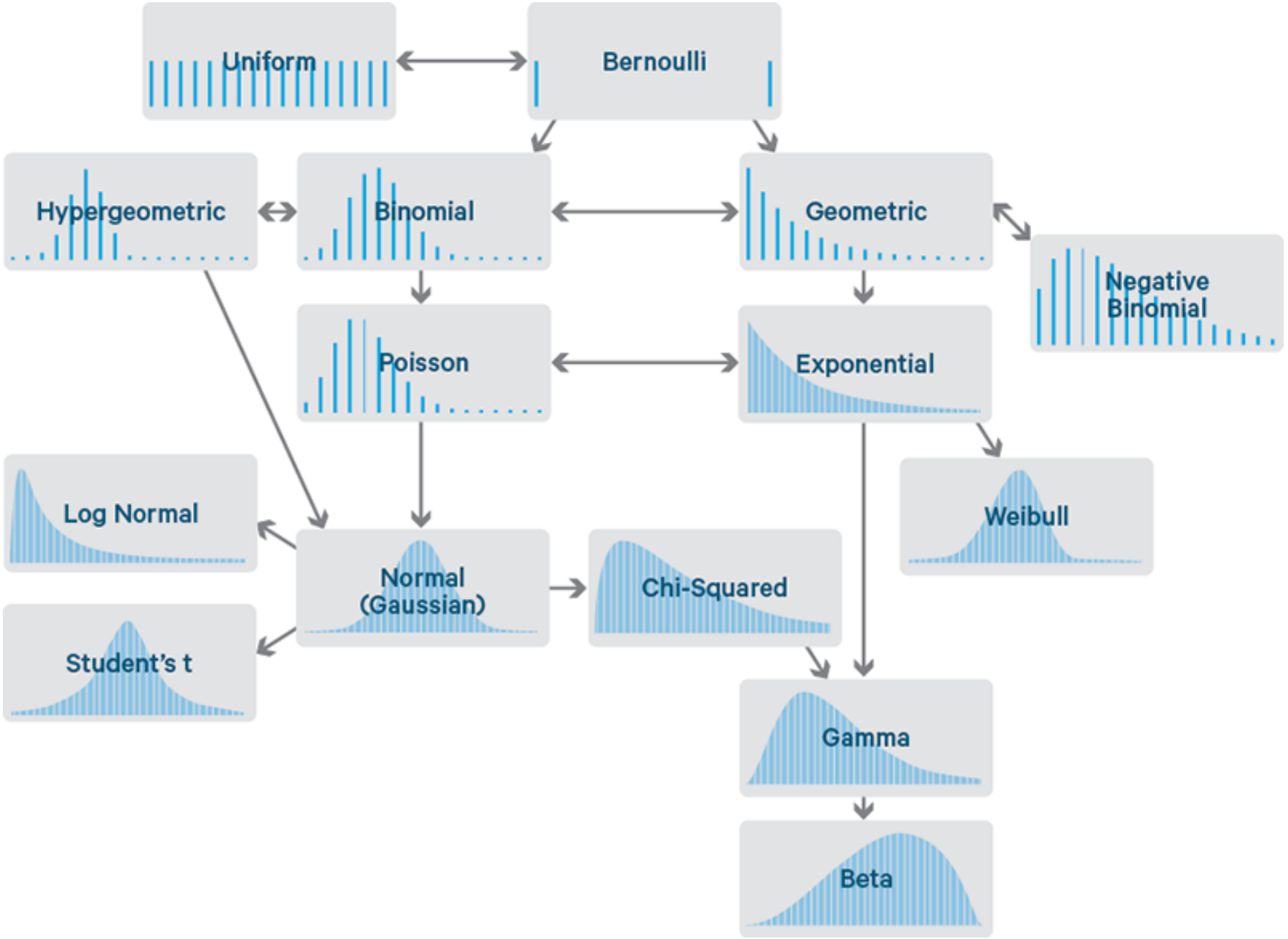
Median and IQR

- Useful in asymmetric distributions
- Robust parameters
- IQR contains 50% of data
- Compounds of the boxplot

Distribution of the data



- Mean > Median : Right skewed distribution
- Mean < Median : Left skewed distribution

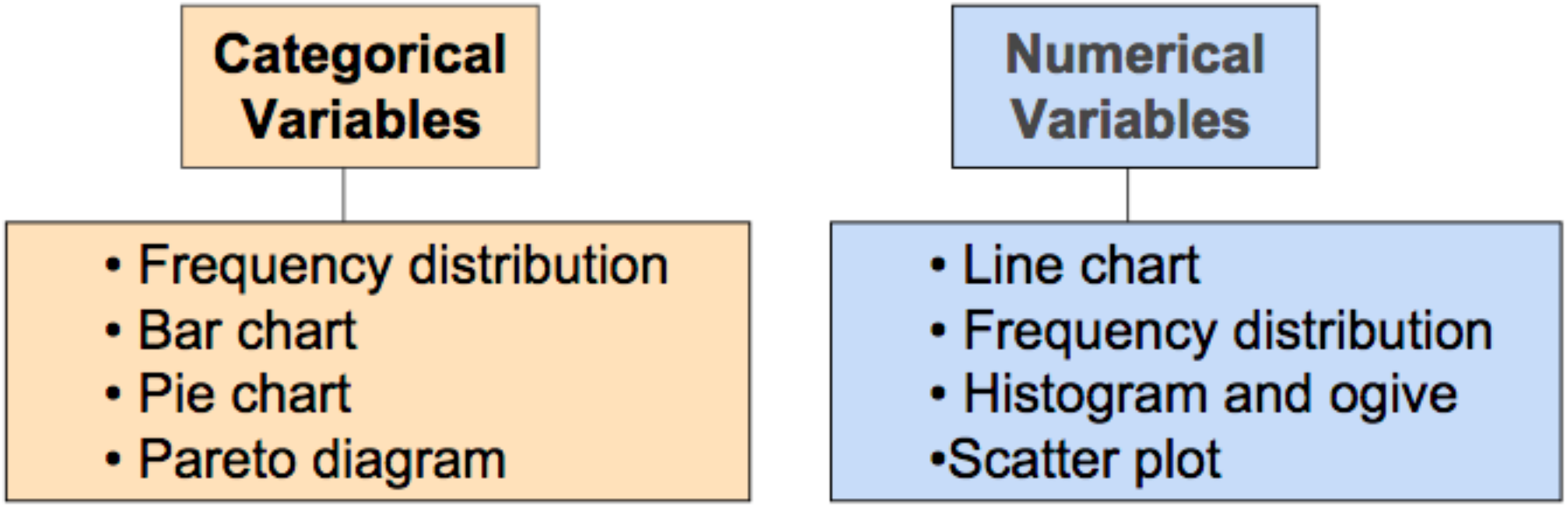

Uniform
Bernoulli
Hypergeometric
Binomial
Geometric
Negative Binomial
Poisson
Exponential
Log Normal
Weibull
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta

Graphs in statistics

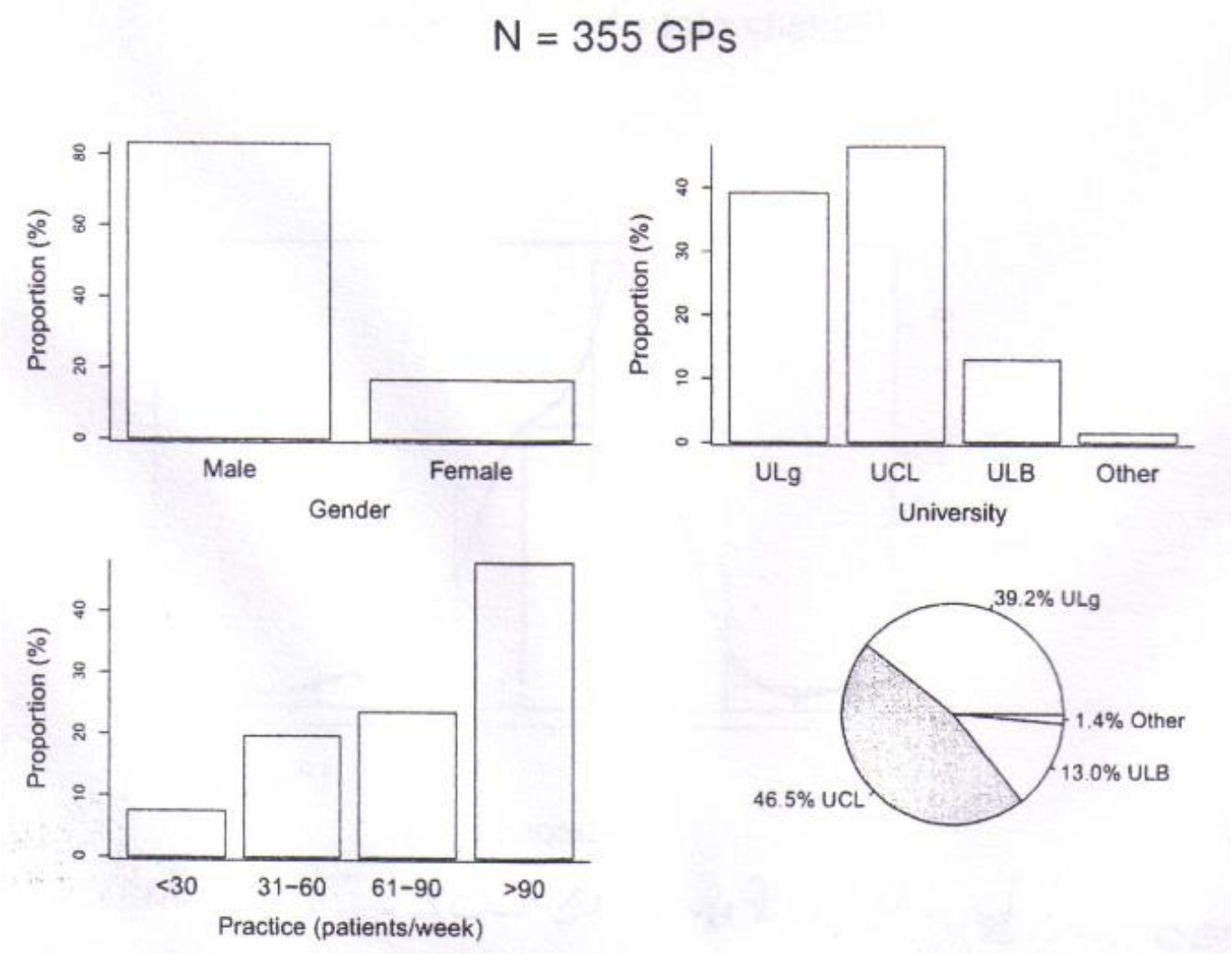
- Fisher (1890-1962)
- Type of graphs depends on type of data to summarize
- Very useful to check and present the data
- First step during analysis

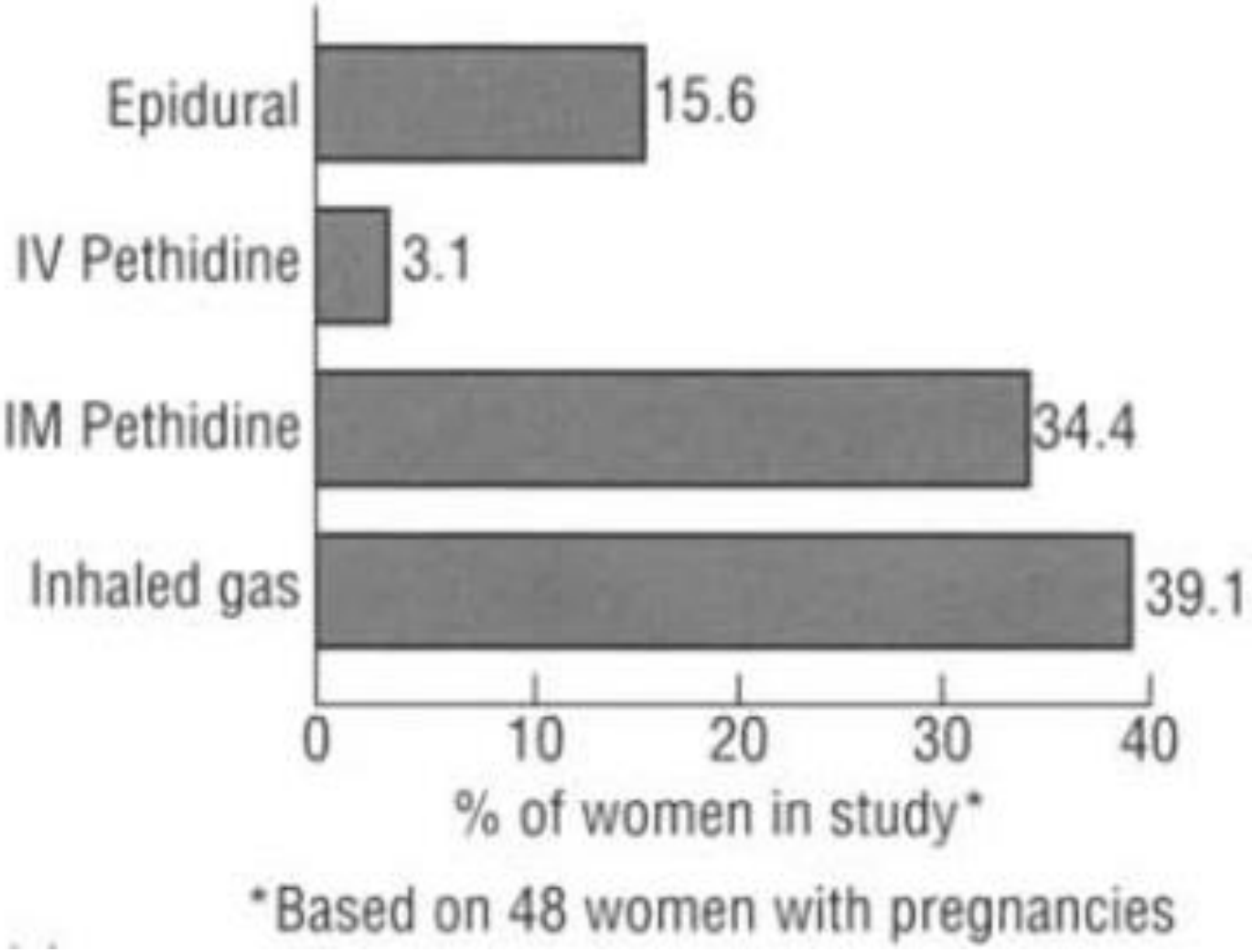
1. Qualitative: Bar graph, Pie graph
2. Ordinal: Bar graph, (order)
3. Discrete: Bar graph
4. Continuous: Histogram, density plot, frequency plot, box plot, scatter plot.

Graphs in statistics



Qualitative




Epidural
15.6
IV Pethidine
3.1
IM Pethidine
34.4
Inhaled gas
39.1
0
10
20
30
40
% of women in study*
*Based on 48 women with pregnancies

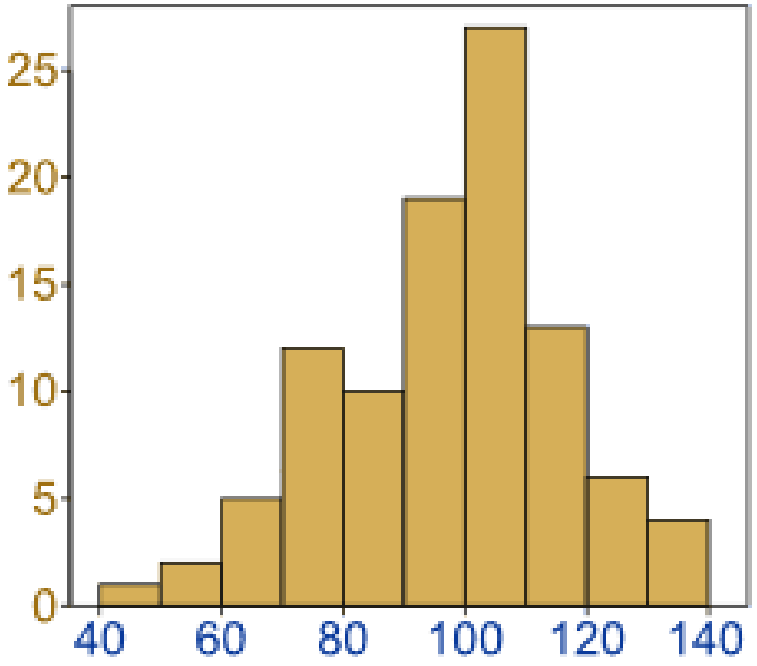
Quantitative data

- Discrete / continuous
- Absolute or relative frequency

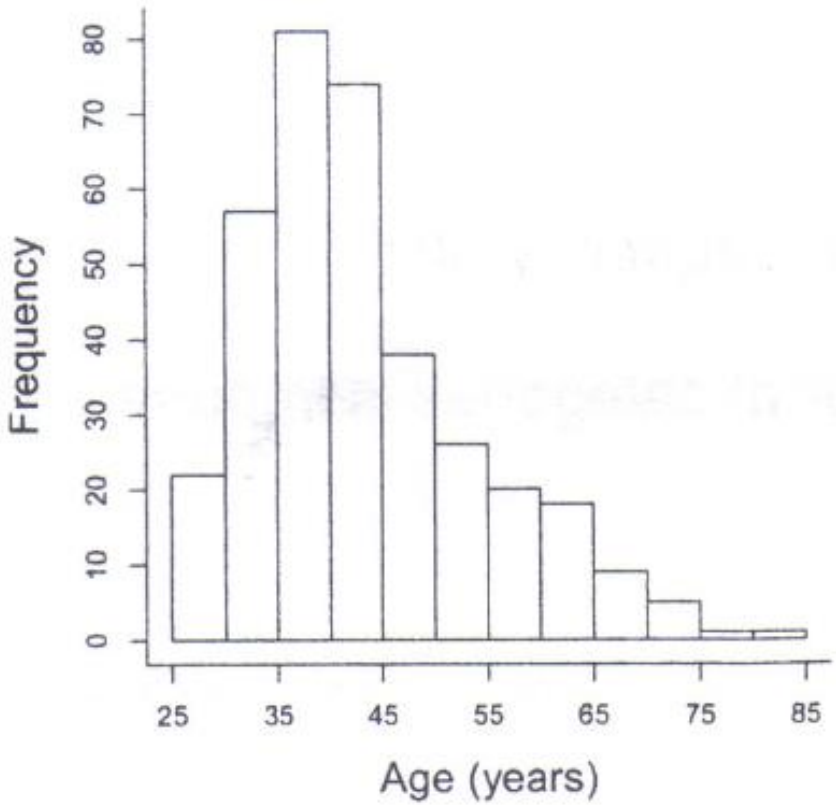
- Histogram

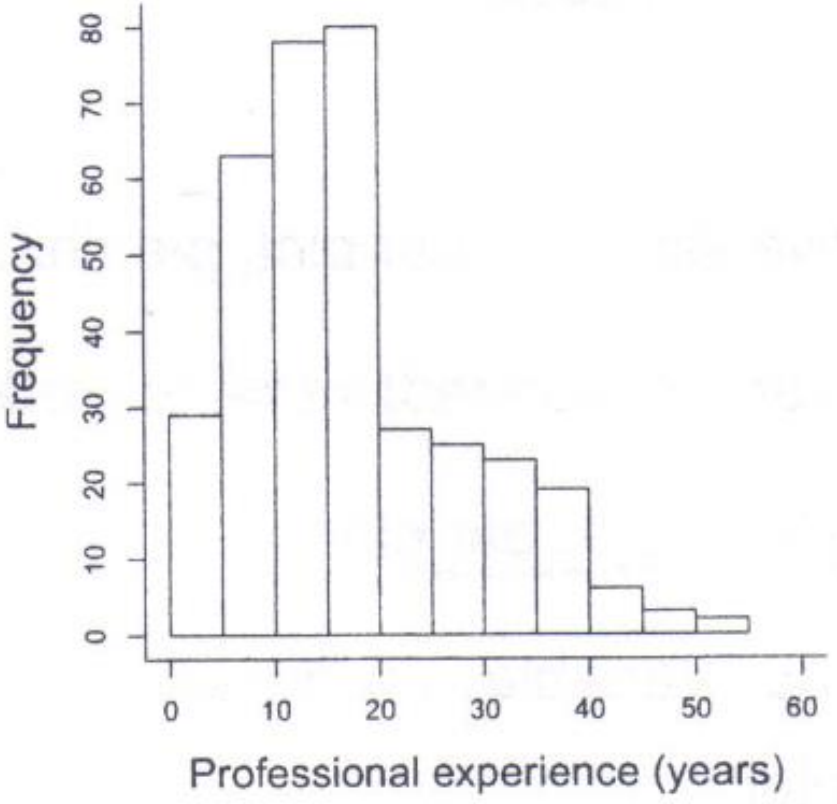
110	90	70	45	85	110
112	114	108	48	140	100
113	100	87	95	70	109

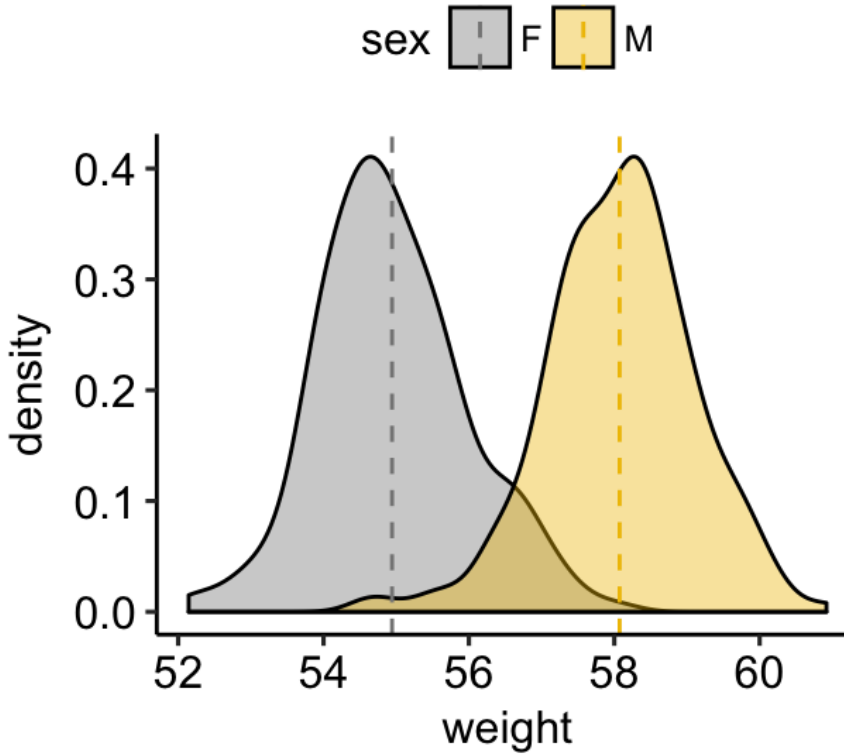
n = 20

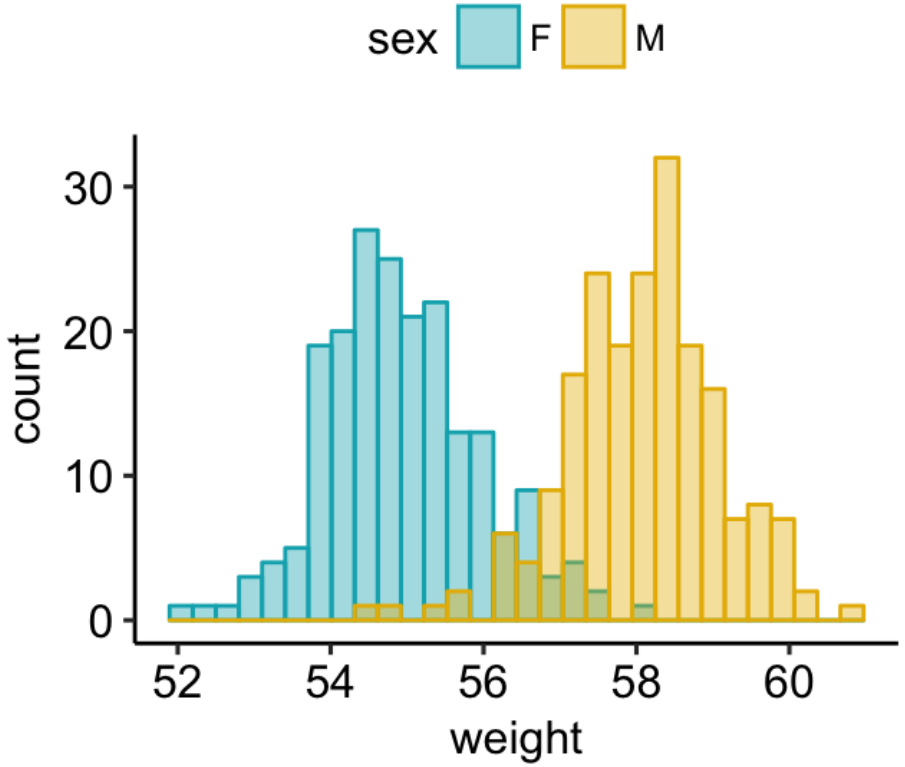


Histograms - N = 355 GPs





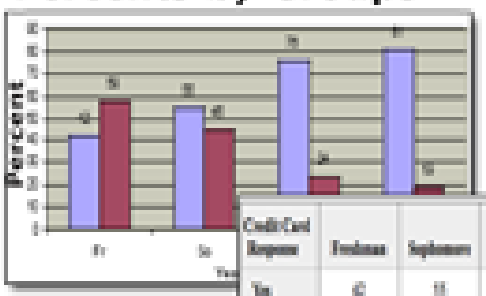

sex
F
M
density
0.4
0.3
0.2
0.1
0.0
52
54
56
58
60
weight


sex
F
M
count
30
20
10
0
52
54
56
58
60
weight

Relationships between Variables

Two Categorical Variables

Bar Graph of Percents by Groups

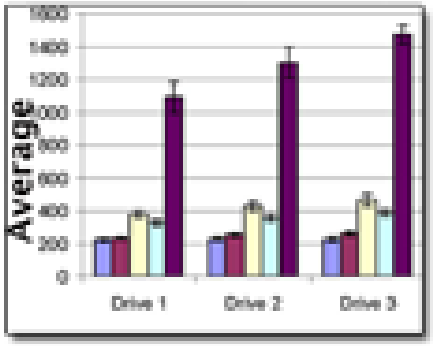


Table

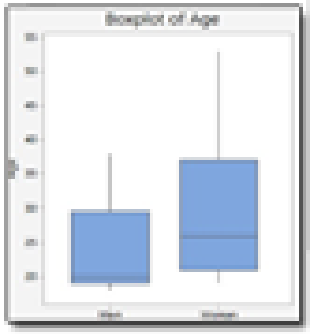
Credit Card Response	Freshman	Sophomore	Junior	Senior	Total
Yes	42	53	76	83	254
No	58	47	24	17	146
Total	100	100	100	100	400

One Measurement, One Categorical Variable

Bar Graph of Proportions by Groups

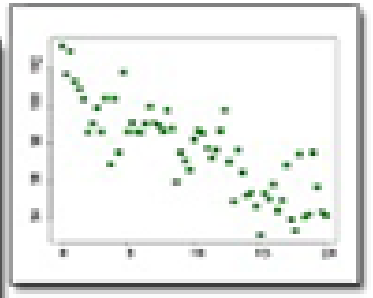


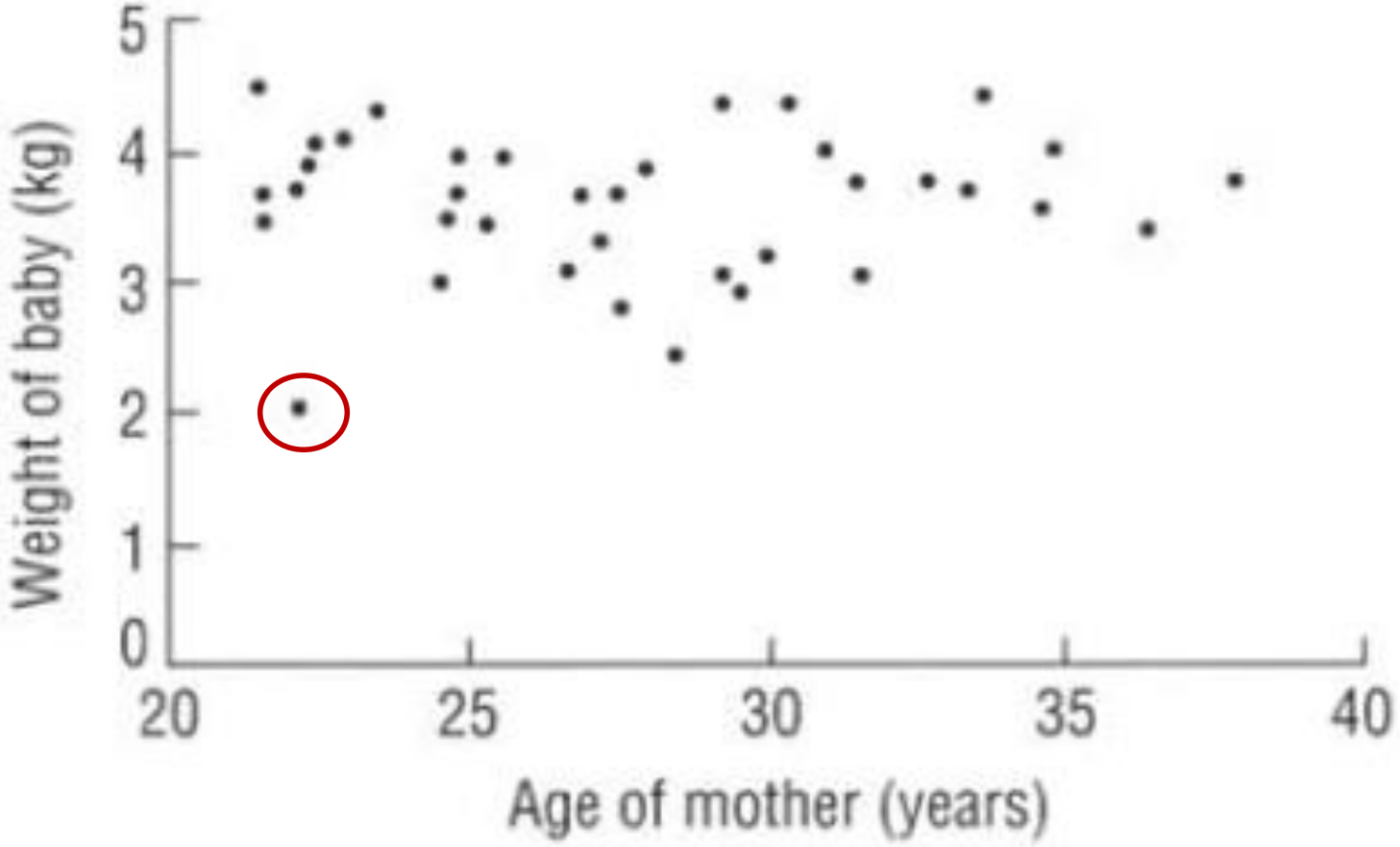
Side by side Boxplot

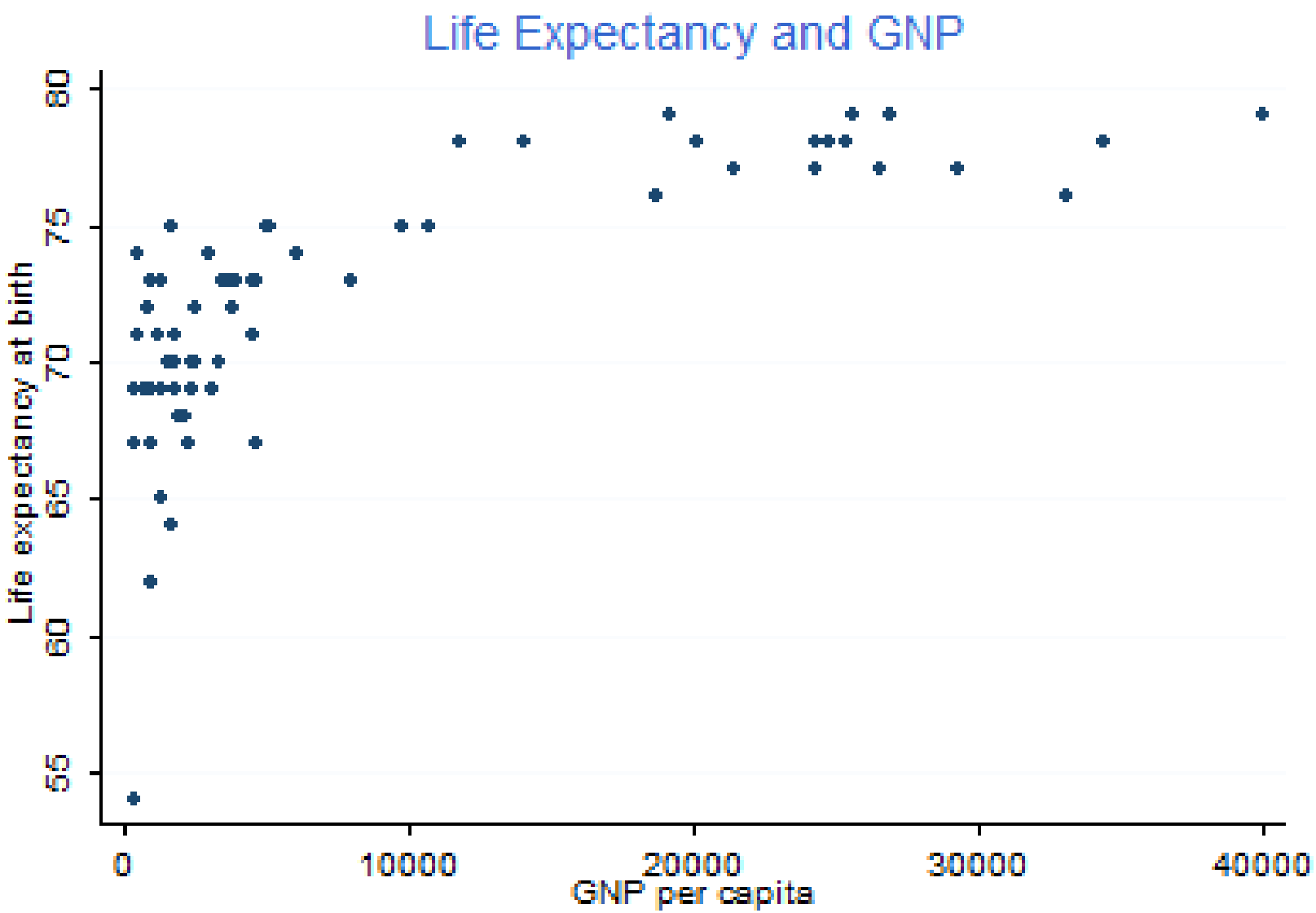


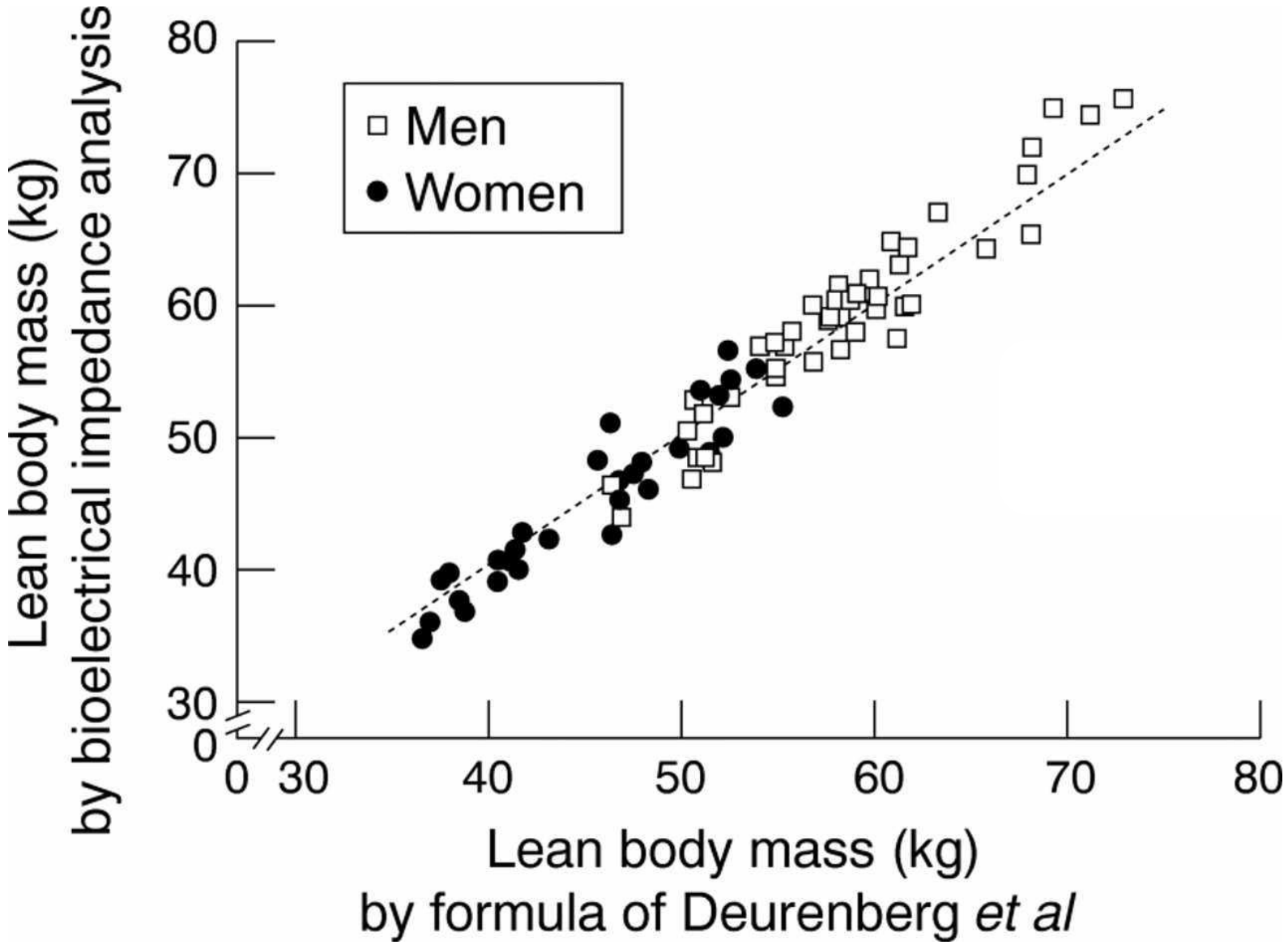
Two Measurement Variables

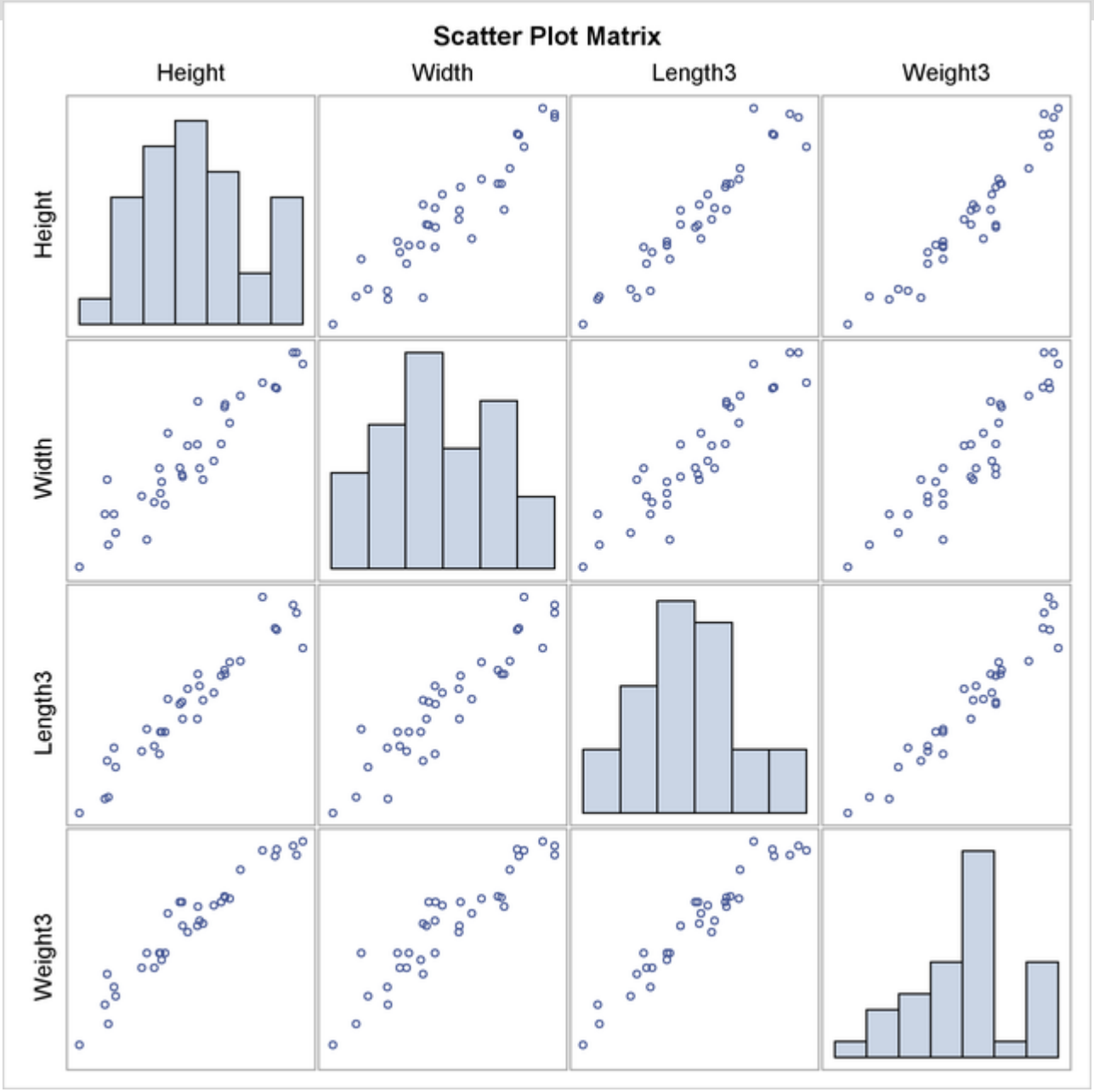
Scatterplot



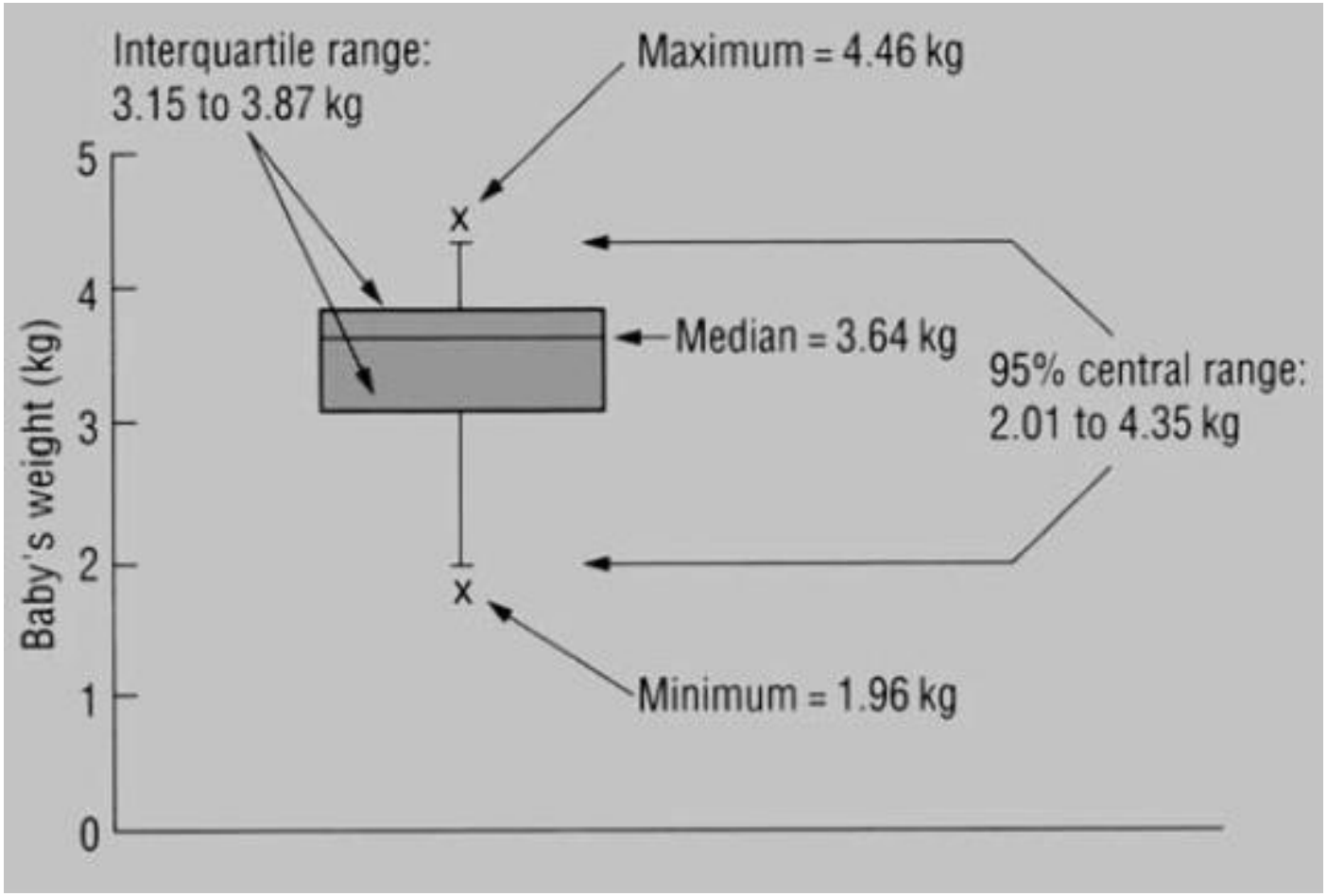

Weight of baby (kg)
5
4
3
2
1
0
20
25
30
35
40
Age of mother (years)

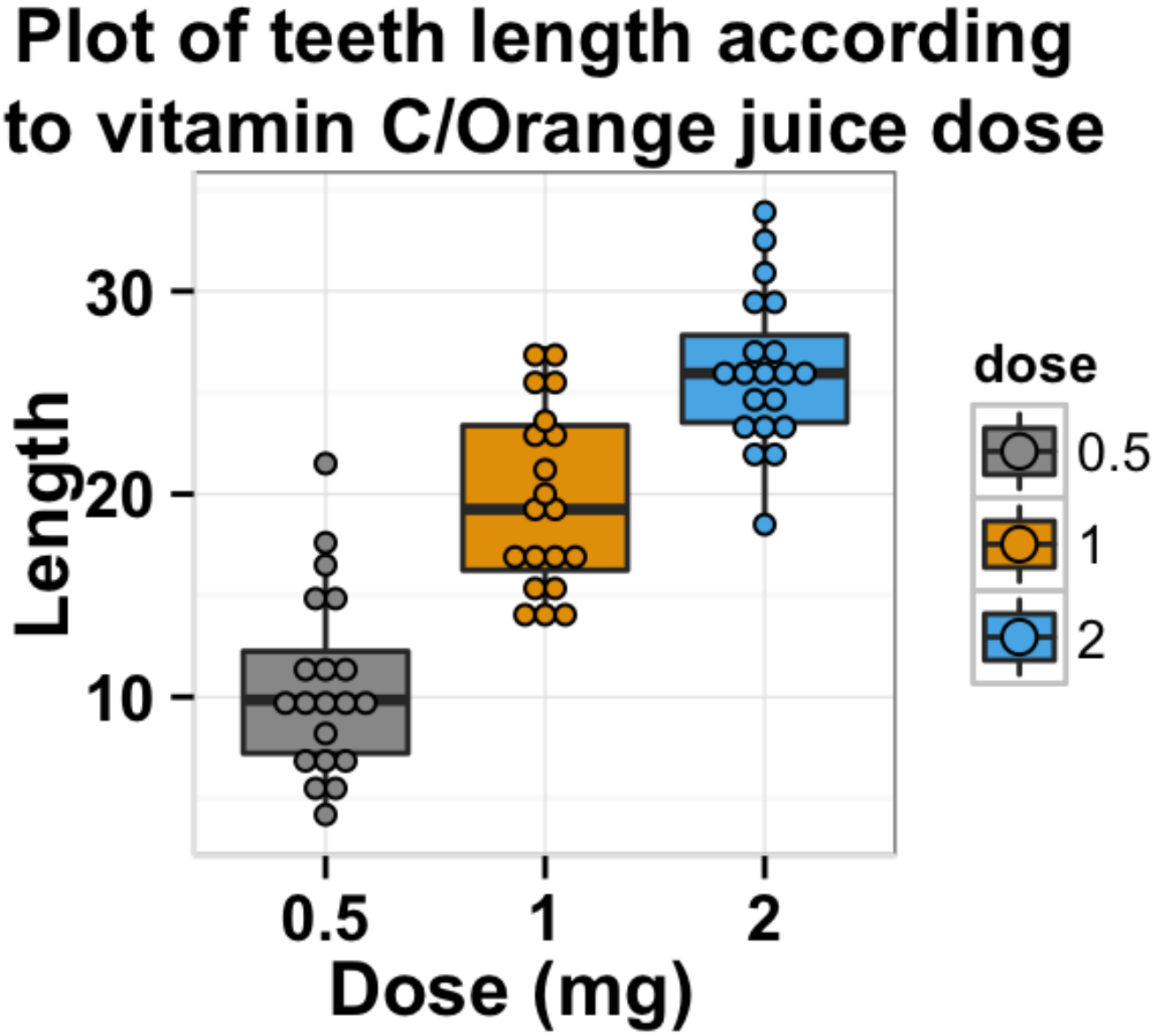

Life Expectancy and GNP
Life expectancy at birth
55
60
65
70
75
80
0
10000
20000
30000
40000
GNP per capita


Lean body mass (kg)
by bioelectrical impedance analysis
80
70
60
50
40
30
0
Men
Women
0
30
40
50
60
70
80
Lean body mass (kg)
by formula of Deurenberg *et al*

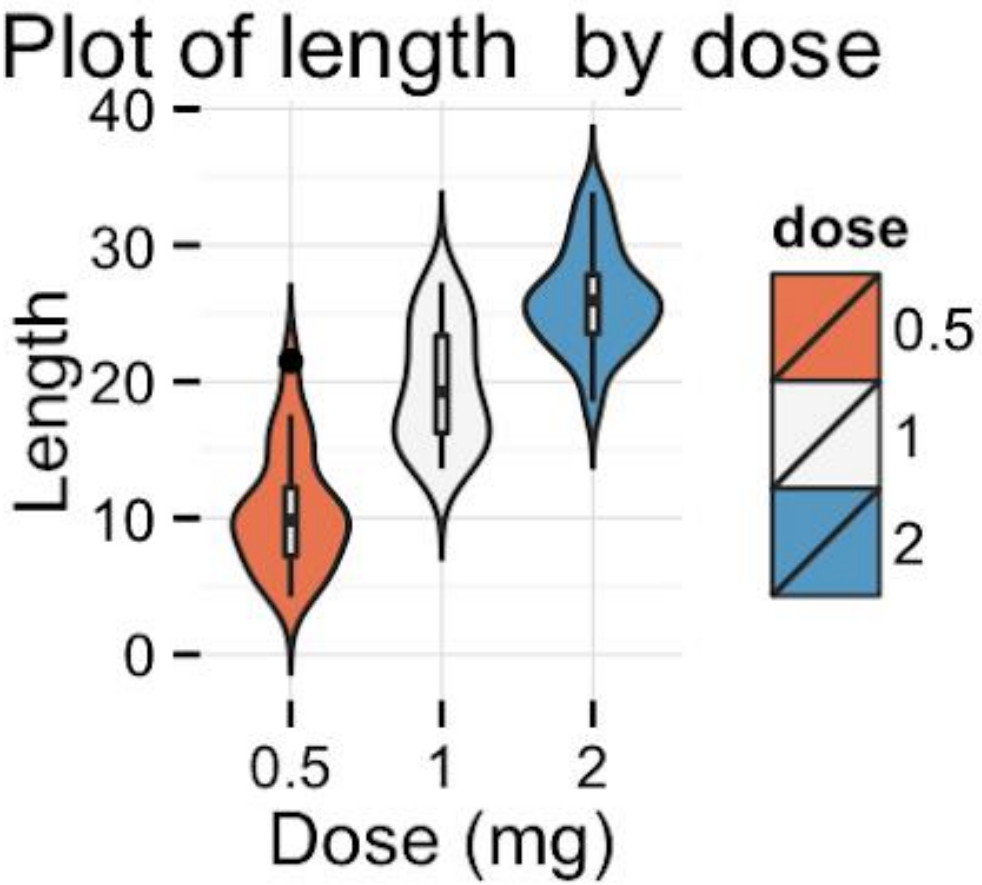

Scatter Plot Matrix
Height
Width
Length3
Weight3
Height
Width
Length3
Weight3

Boxplot

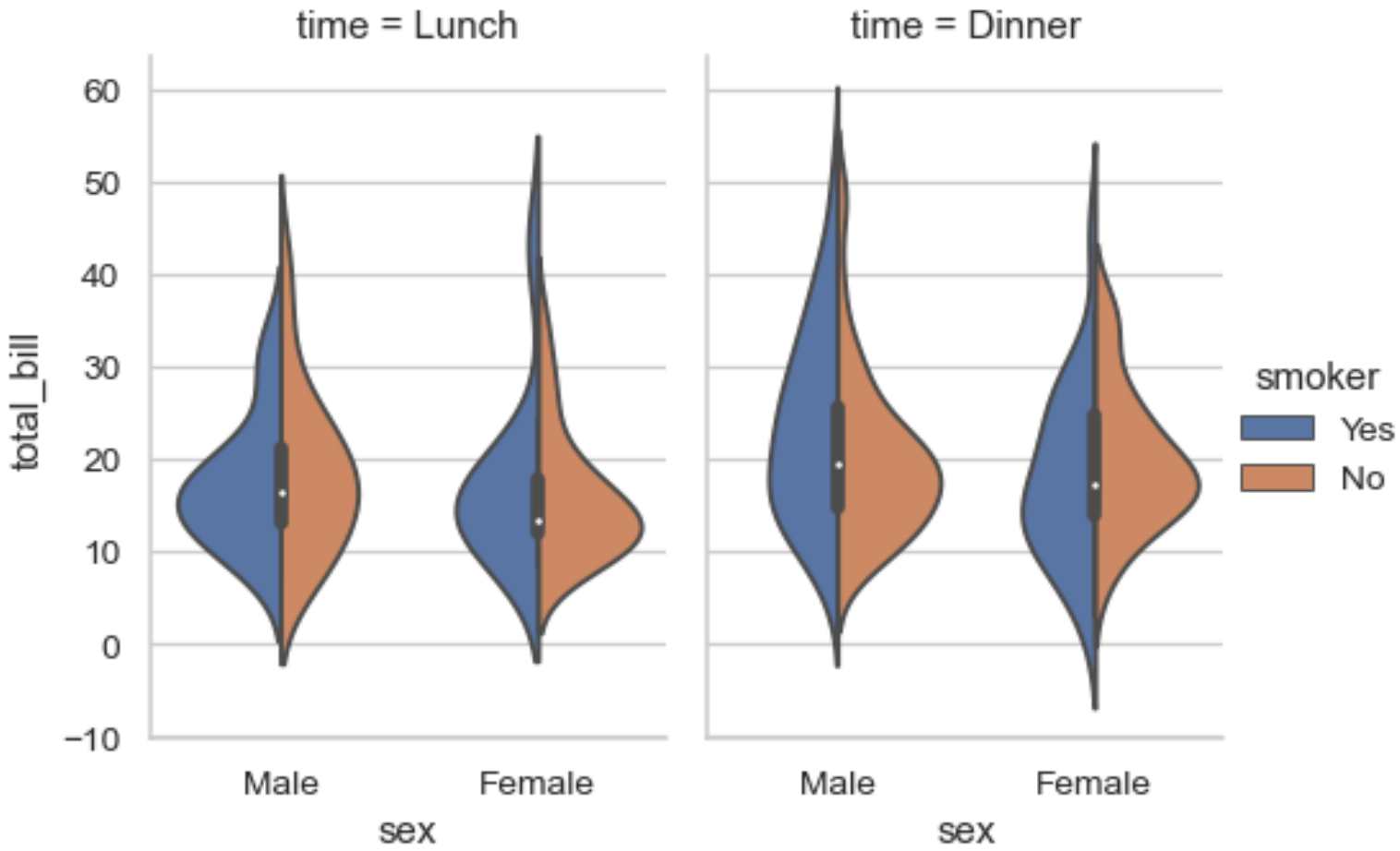

Interquartile range:
3.15 to 3.87 kg
Maximum = 4.46 kg
Median = 3.64 kg
95% central range:
2.01 to 4.35 kg
Minimum = 1.96 kg
Baby's weight (kg)
5
4
3
2
1
0


Plot of teeth length according to vitamin C/Orange juice dose
Length
30
20
10
0.5
1
2
Dose (mg)
dose
0.5
1
2

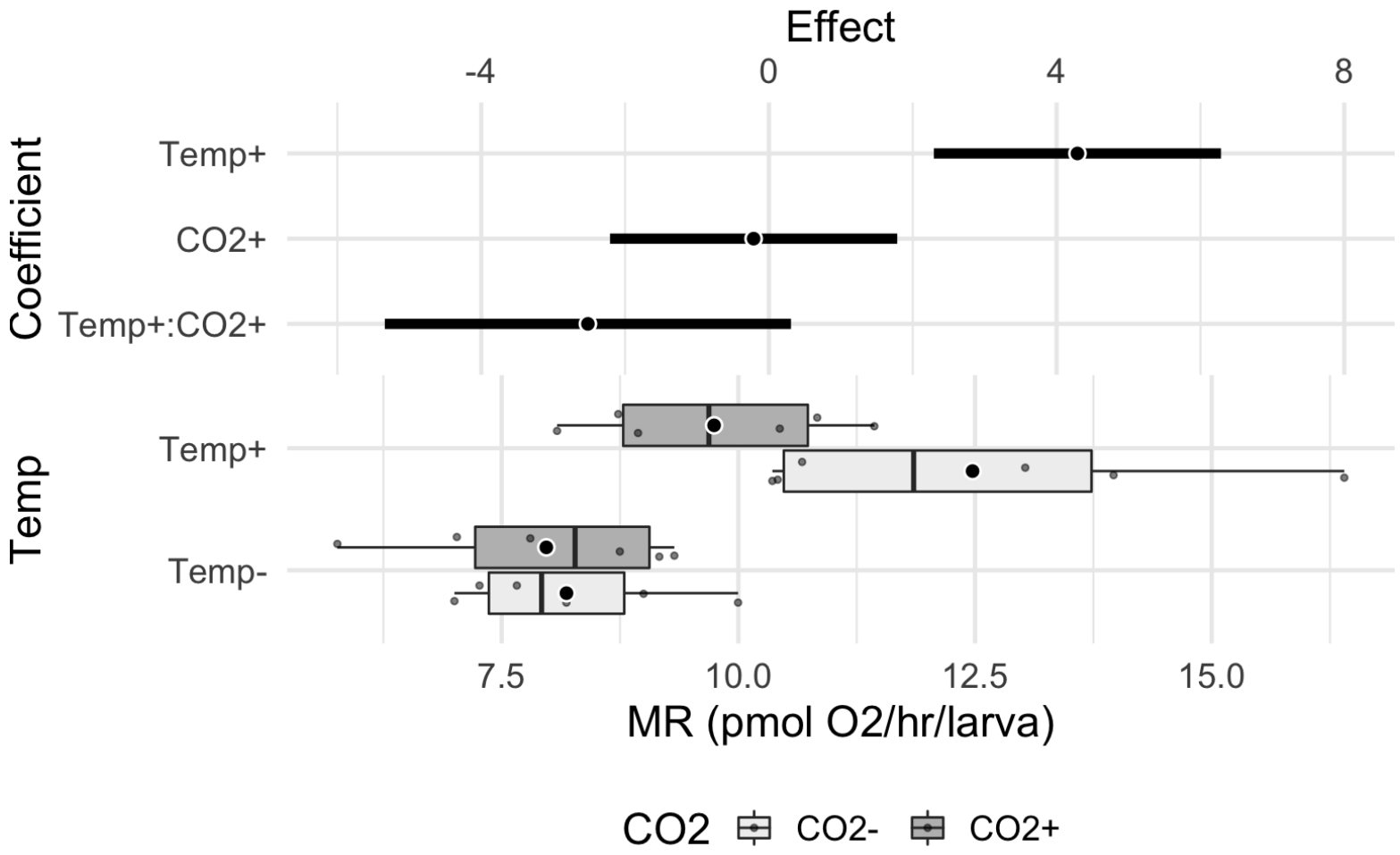
Violin plot 1



Violin plot 2

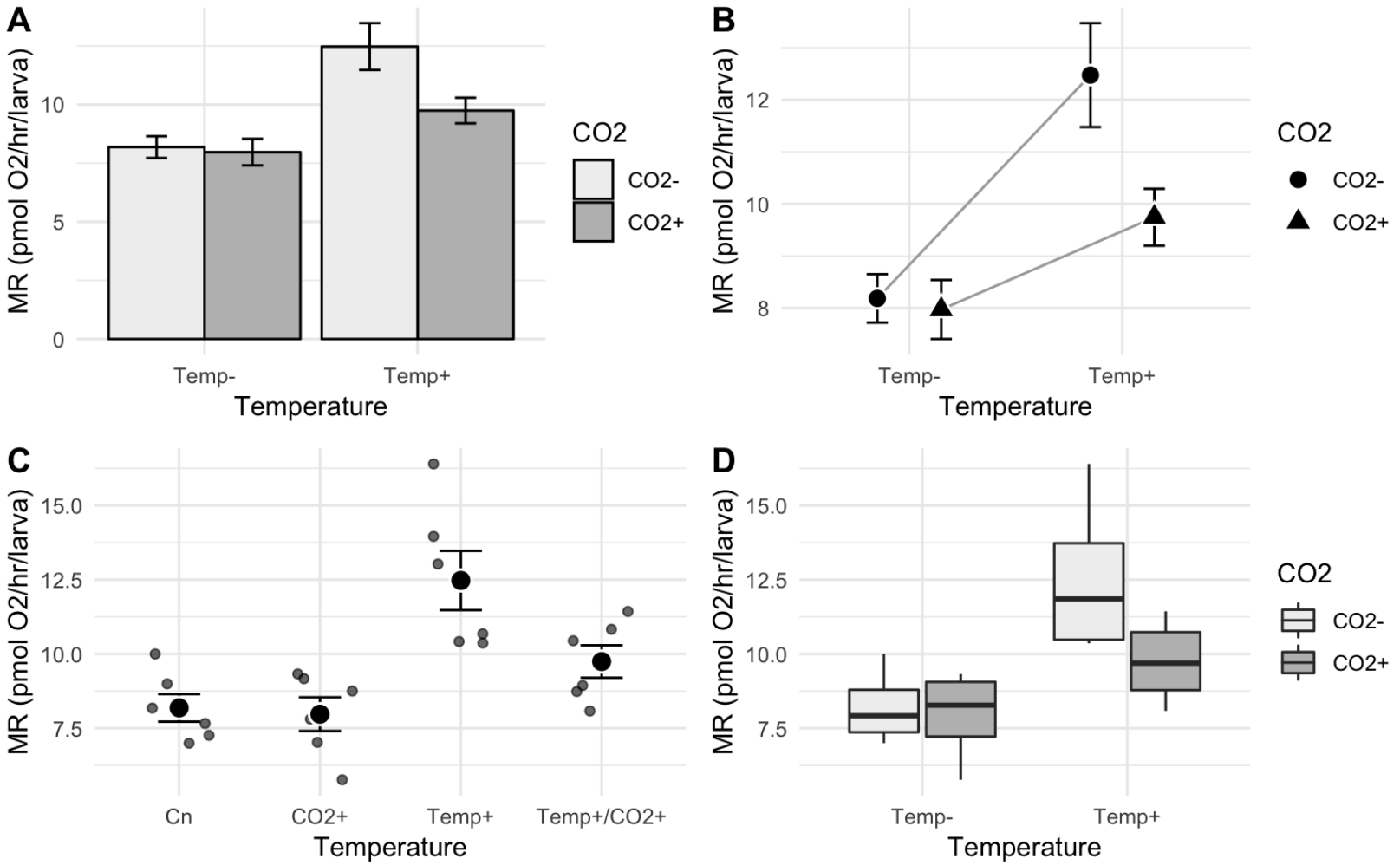


Harrell Plot

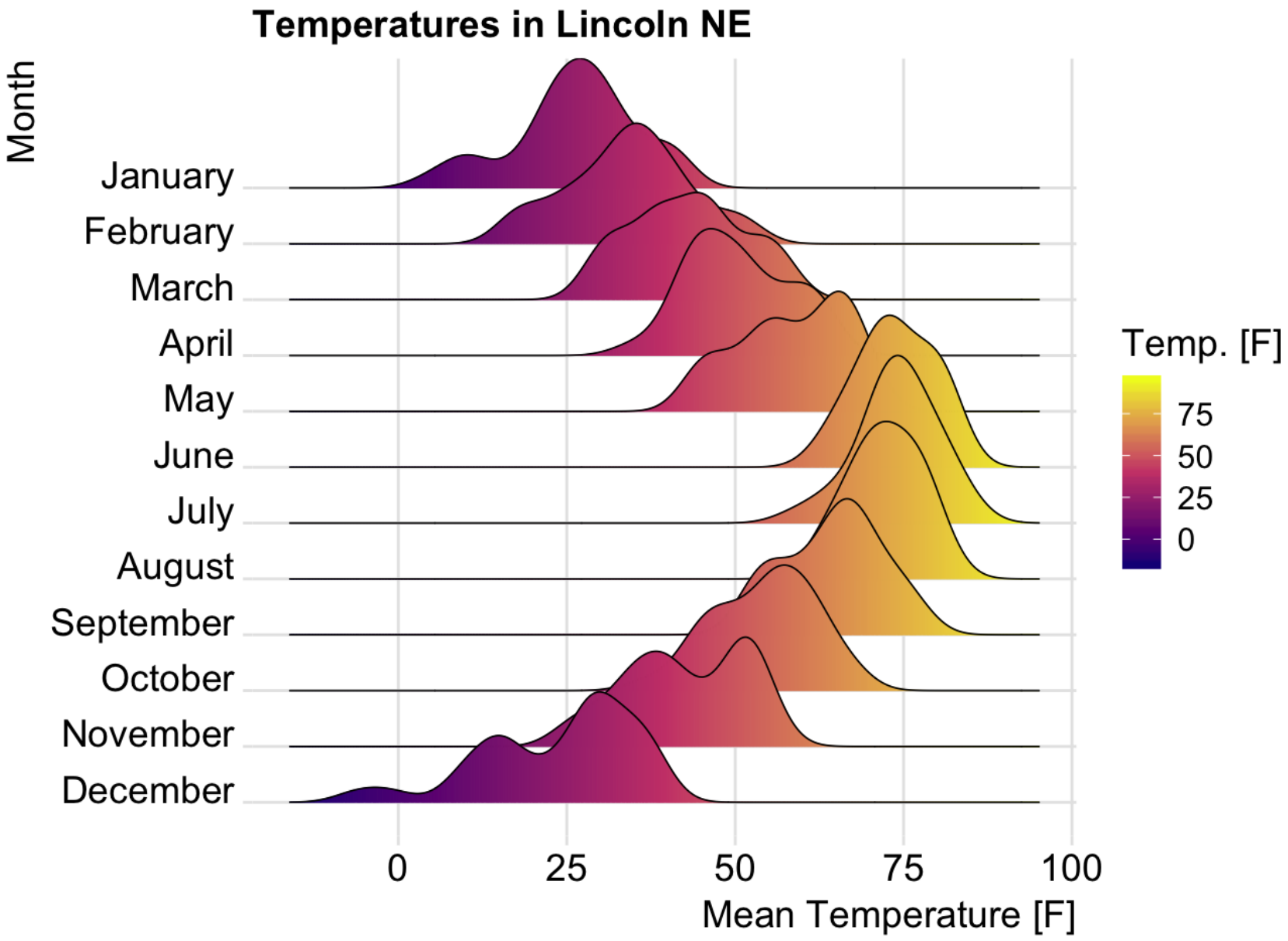


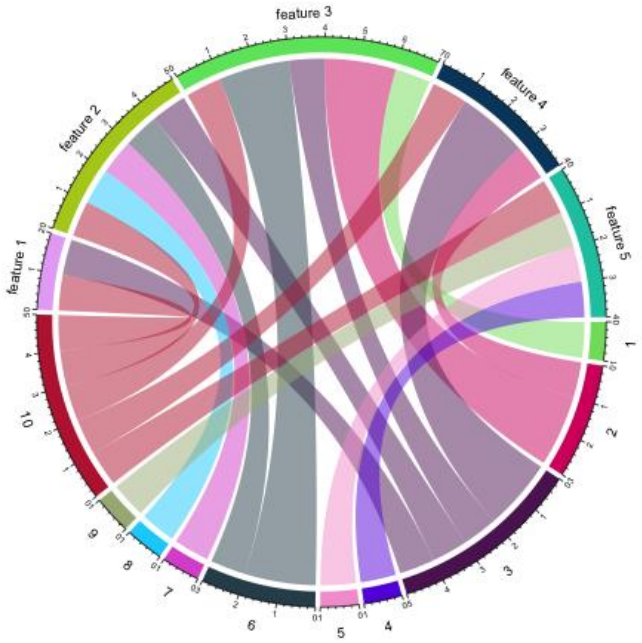
Harrell, Frank E. 2014. “Principles of Graph Construction.”

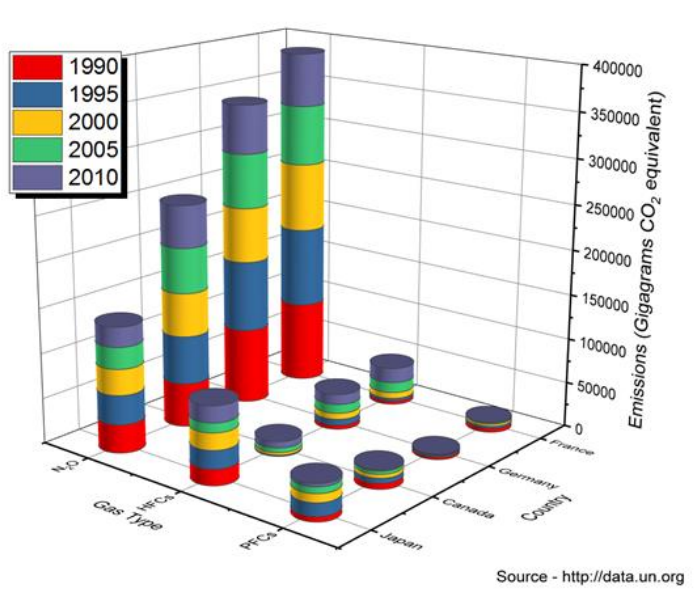
Harrell plots vs. common alternatives

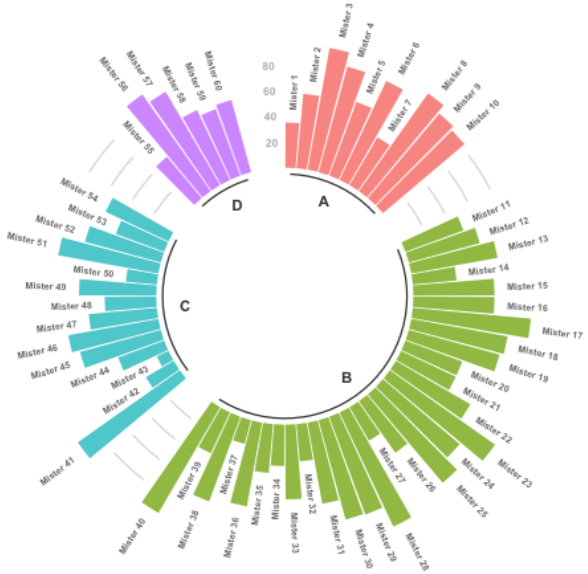


https://www.middleprofessor.com/files/quasipubs/harrell_plot_intro.html#introduction

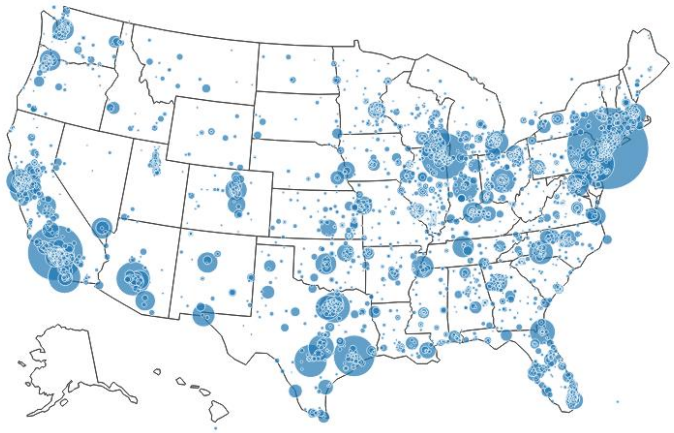

Temperatures in Lincoln NE
Month
January
February
March
April
May
June
July
August
September
October
November
December
0
25
50
75
100
Mean Temperature [F]
Temp. [F]
75
50
25
0







2014 U.S. City Populations

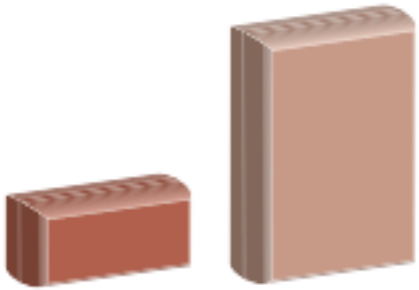


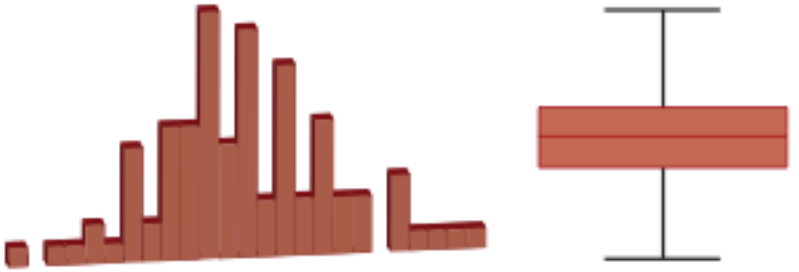
Tufte's Views on Graphical Excellence

"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency.

- Graphical displays should
  - Show the data
  - Induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
  - Avoid distorting what the data have to say
  - Present many numbers in a small space
  - Make large data sets coherent
  - Encourage the eye to compare different pieces of data
  - Reveal the data at several levels of detail, from a broad overview to the fine structure
  - Serve a reasonably clear purpose: description, exploration, tabulation, or decoration
  - Be closely integrated with the statistical and verbal descriptions of a data set.

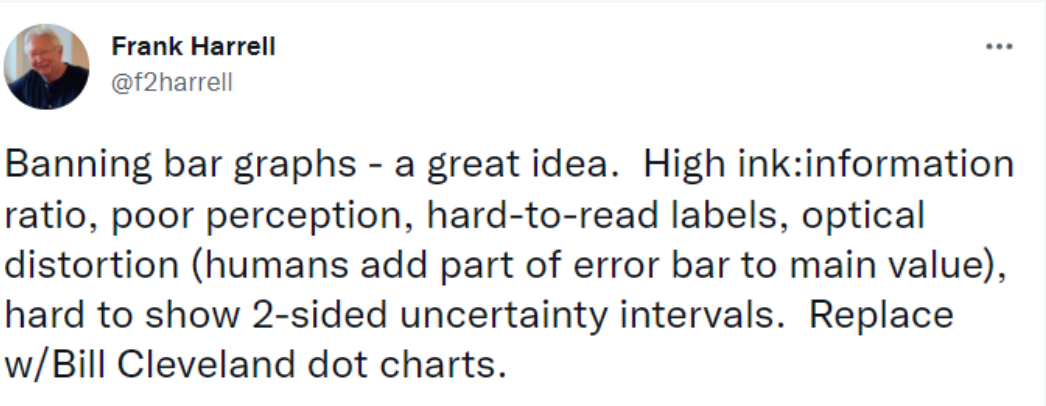
Variable

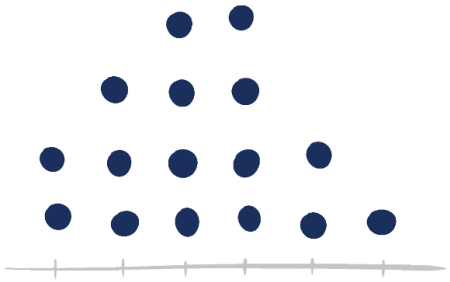
- Categorical
  - Nominal
    - Ex. gender
    - 
    - Descriptives: absolute and relative frequencies, mode
  - Ordinal
    - Ex. educational level, NYHA classification
    - 
    - Descriptives: absolute and relative frequencies, median, interquartile range, minimum, maximum, percentile
- Numerical
  - Continuous
    - Ex. weight, systolic blood pressure
  - Discrete
    - Ex. number of visits



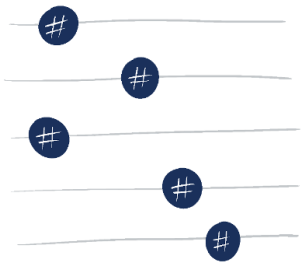
Descriptives: mean, median, SD, minimum, maximum, percentile

Banning bar graphs

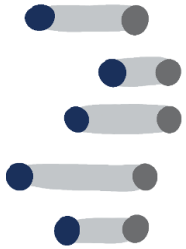




DOT PLOT

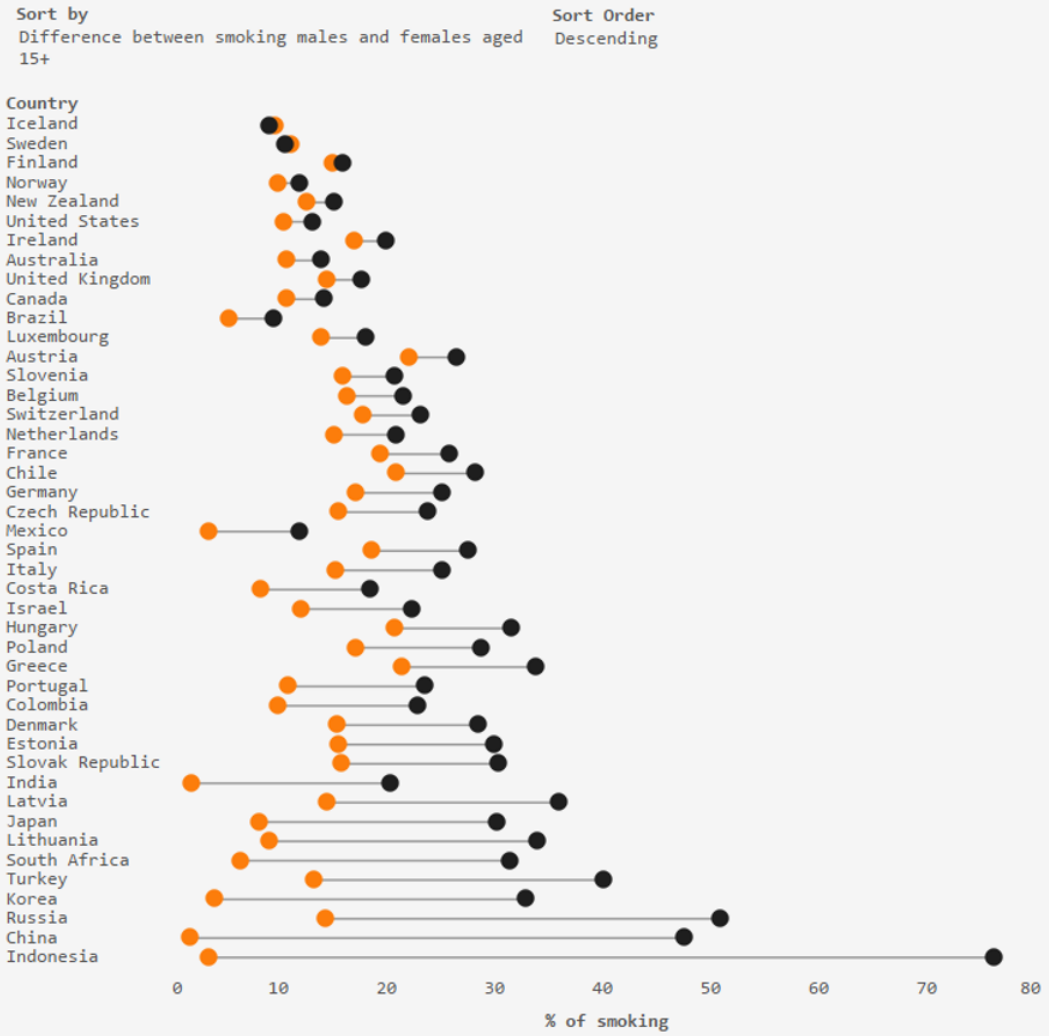


CLEVELAND DOT PLOT



CONNECTED DOT PLOT

In Iceland and Sweden there are more smoking women than men aged 15+



Source: OECD Health Data 2018
Author: Anna H. Dzikowska @AnnDzikowskaViz

Descriptive statistics / Describing data

- To look slowly the data
- To discover trends
- To show relationships

- To what ?
  - To find the best approach to show it
  - To increase the comprehension of the data
  - To give utility to the data