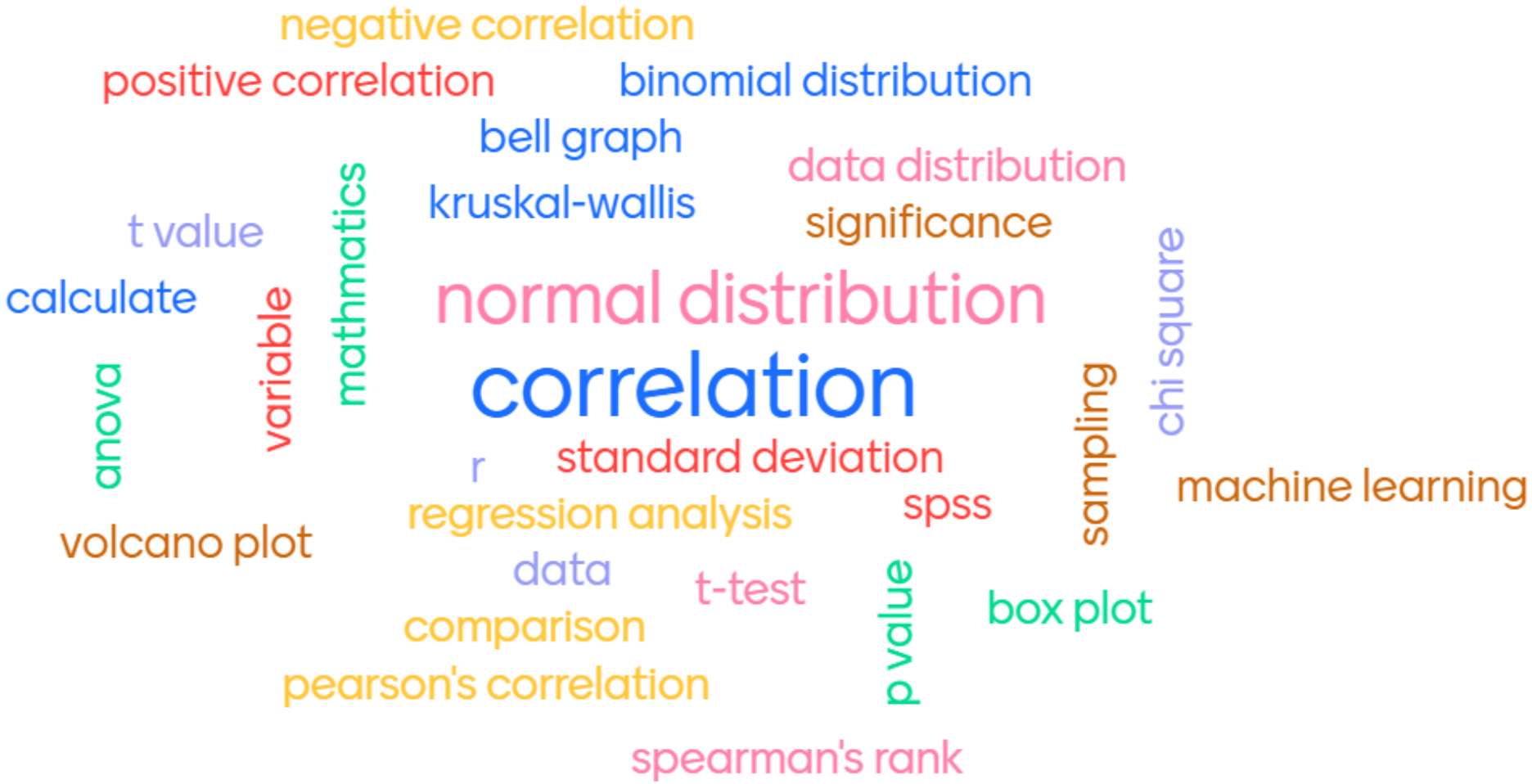
Introduction to Statistics

Dr. Xiangyu Li
xiangyu.li@scilifelab.se
Postdoctoral researcher
KTH Royal Institure of Technology

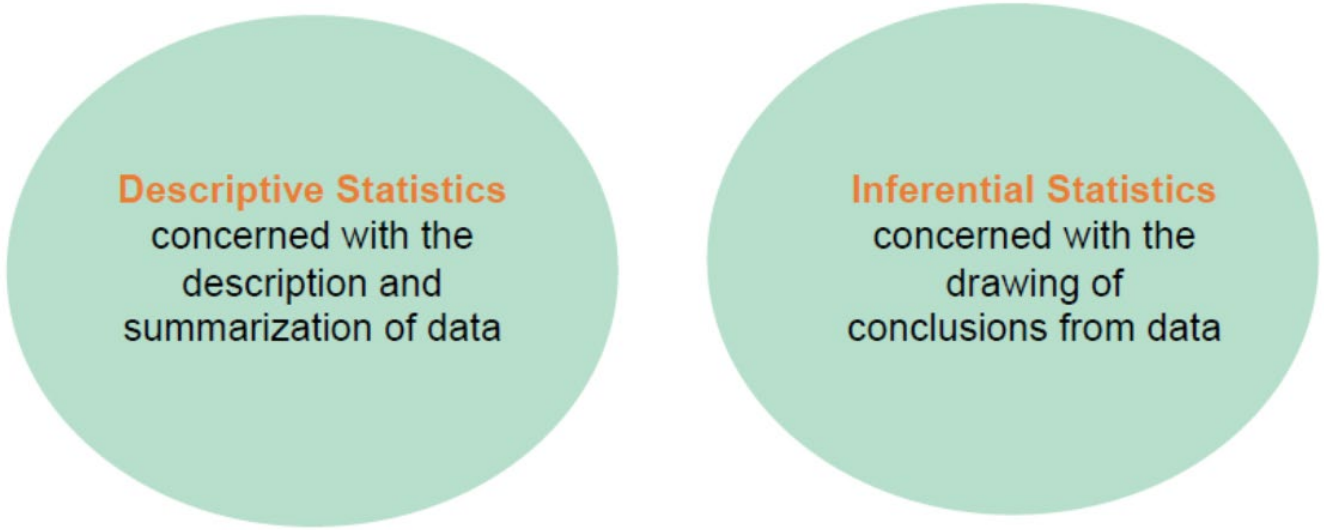
Go to **www.menti.com** and use the code **87 54 41 1**

Could you please provide two words associated with Statistics, Health Statistics or Biostatistics?

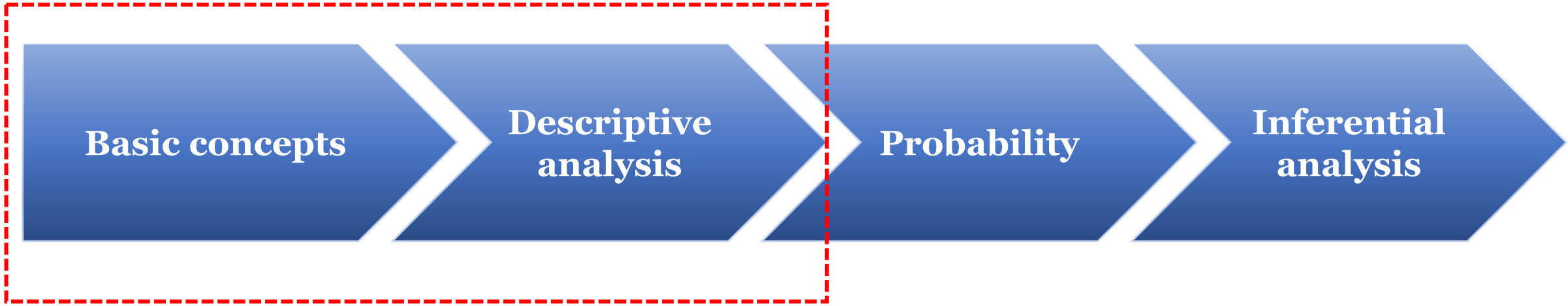

negative correlation
positive correlation
binomial distribution
bell graph
data distribution
kruskal-wallis
significance
t value
mathmatics
calculate
normal distribution
chi square
variable
correlation
anova
sampling
r
standard deviation
machine learning
regression analysis
spss
volcano plot
data
t-test
p value
box plot
comparison
pearson's correlation
spearman's rank

Statistics

- Statistics is the art of learning from data.



Outline


Basic concepts
Descriptive analysis
Probability
Inferential analysis

Basic concepts

- Data type
- Population vs sample
- Relation between variables

Basic concepts

- Data type

Data type

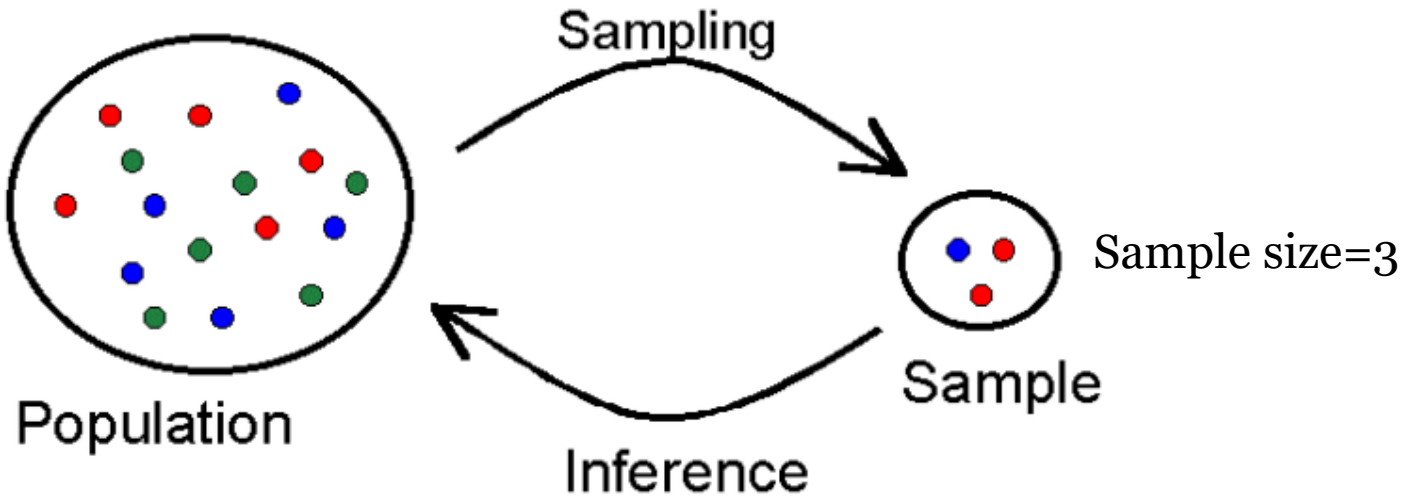
- Categorical variables
  - Nominal variables (e.g. gender, color)
  - Ordinal variables (e.g. drug dose, tumor grade)
- Quantitative variables
  - Discrete variables (e.g. NO. Of patients)
  - Continuous variables (e.g. weight, height)

Practice

- Plasma glucose level √ quantitative and continuous variable
- Drug sensitive or resistant √ categorical and nominal variable
- Recurrence risk: low, medium and high risk √ categorical and ordinal variable
- Body mass index (BMI, Kg/M^2) √ quantitative and continuous variable
- Underweight, normal weight, overweight, obesity √ categorical and ordinal variable
- Age √ Can be discrete or continuous variable

Population vs sample

- Population: a complete set of subjects with specific characteristics
- Sample: a representative sub-set from the population



Relation between variables

Independent variables

- The occurrence of one variable provides no information about the occurrence of the other variable (e.g. the heights of two strangers)

Dependent variables

- The two viarables are somehow linked by a systematic relationship (e.g. calorie uptake and weight)

Why we need descriptive statistics?

Patients (100)

Health (100)

PARENT_SAMPLE_NAME	cysteine	spermidine	serine	12,13-DiHOME	kynurenate	alpha-ketoglutarate	nicotinamide	3-hydroxyisobutyrate	methylglutarate
KGCO-25376	0,857326239	1,845740666	0,542809507	1,991993205	NA	1,027888812	1,119401497	1,352410605	0,761073143
KGCO-25380	0,906967699	1,076327293	0,770303482	1,788673309	0,735033424	0,996881452	1,179609058	1,33691452	0,623795752
KGCO-25384	0,76845518	1,582165422	0,669438525	2,782947183	NA	0,850709681	1,140420156	1,940250587	0,665442734
KGCO-25388	0,535618001	1,544328335	0,712492821	1,851849338	0,758999752	0,74721195	1,065009204	2,411561056	0,593900523
KGCO-25392	0,613320799	1,502149489	0,819490661	1,66533952	0,765602047	0,81347602	0,763426586	4,893474982	1,444356031
KGCO-25396	0,562837291	1,753463805	0,885579202	1,403103506	NA	0,790687537	0,775830911	2,746082638	0,658310663
KGCO-24884	1,198664279	1,628336651	0,51708489	2,384351027	1,172398923	0,840664841	0,732317231	0,971508255	0,977973973
KGCO-24888	2,153153285	1,619464313	28,18731197	2,31229031	0,682055772	0,667252637	0,687474199	0,881584329	0,915029654
KGCO-24892	1,646600993	0,910076367	36,49076951	2,880662755	0,582246557	0,795859409	0,579672291	0,506382181	1,898493801
KGCO-24896	1,220015193	1,193398887	42,46265778	1,787710143	0,523220866	0,713979847	0,515808683	0,405188569	0,622717226
KGCO-24900	1,455790173	1,032921517	53,85303531	1,169397544	0,649361001	0,734274989	0,741858384	1,045008659	0,425322346
KGCO-24904	1,149233413	2,066975739	56,66209501	1,24492891	0,696260095	0,957998959	0,580852878	0,763935558	0,636186865
KGCO-25277	0,666423877	2,207428368	0,202355229	1,225450725	0,923215691	0,956357867	1,312653063	1,009932522	1,134755649
KGCO-25280	0,879895741	1,146286769	9,487693228	0,701423427	NA	0,924896591	0,802541763	1,328971685	0,897815686
KGCO-25283	0,433679893	1,381635125	12,60796533	0,93932539	NA	1,002605444	0,846188918	0,895461136	0,692820492
KGCO-25286	0,628958953	1,21794906	14,00379674	1,026644671	NA	0,742875876	0,815076024	1,089332842	0,825476056
KGCO-25289	0,528562122	0,995569633	12,93792403	1,863028204	NA	0,899979924	0,807291287	1,577088438	0,689008234
KGCO-25292	0,635146548	1,868576656	7,65279441	1,308398189	NA	0,769176735	0,634547245	2,206776753	0,743138576
KGCO-24740	1,302035099	1,753805735	1,069246725	0,979825224	0,782715385	1,512837171	0,994250819	1,248968695	1,506066508
KGCO-24744	2,017572688	1,259924487	47,45754681	0,518179904	0,937111664	1,210167754	1,087632162	1,106689374	1,718138646
KGCO-24748	1,336408724	1,650317651	61,20346762	0,77393003	NA	1,155089246	1,089808057	0,934788675	1,266492212
KGCO-24752	0,842868405	1,780214098	58,19777741	0,856189663	0,6879072	1,096446424	1,01896123	1,044763305	1,062916483

...

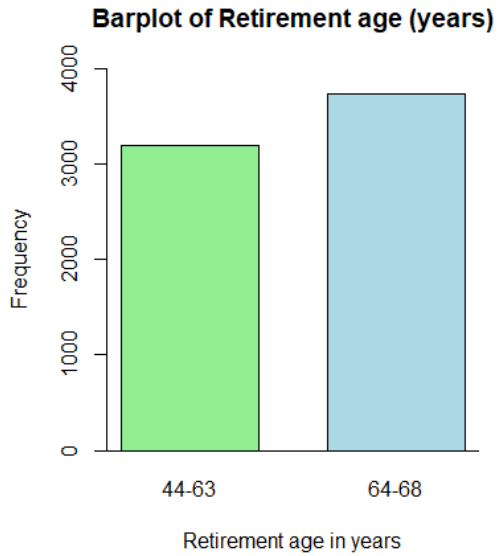
...

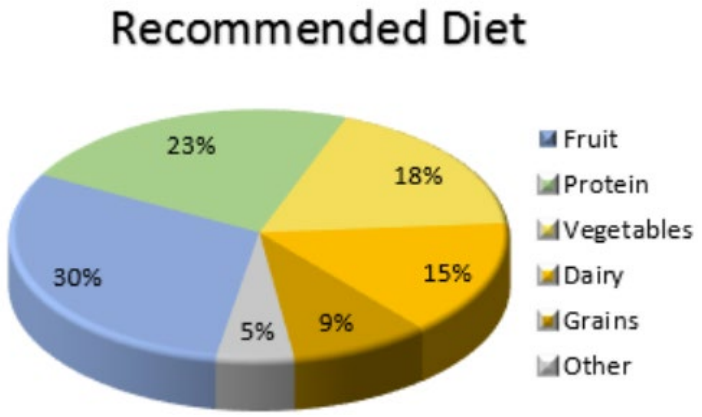
THE HUMAN PROTEIN ATLAS SciLifeLab sysmedicine

Descriptive statistics

Categorical variables

- Frequency
- Percentage
- Mode



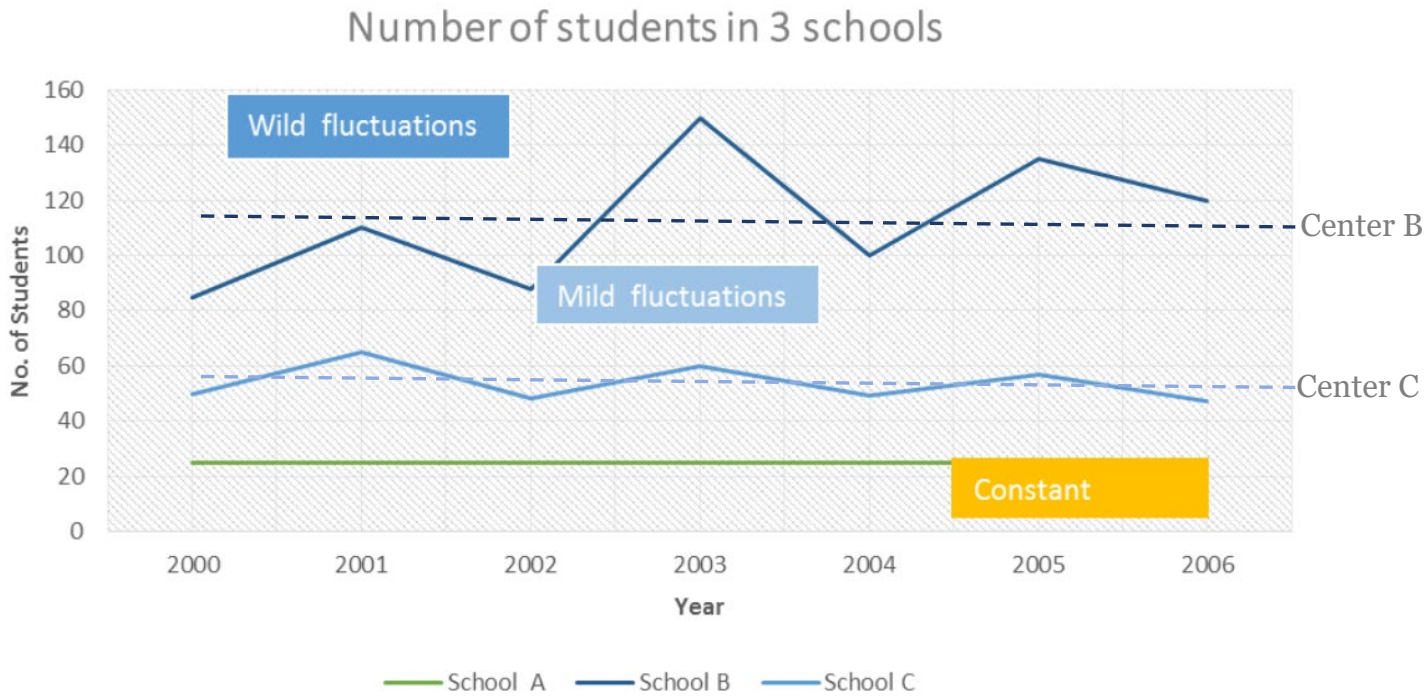


Characteristic	n (%)
All	6938 (100.0)
Retirement age (years)	
44–63	3196 (46.1)
64–68	3742 (53.9)
Educational level	
Compulsory school	2149 (31.0)
Upper secondary school	3054 (44.0)
University education	1735 (25.0)
Married	
Yes	4881 (70.4)
No	2057 (29.7)
Income (SEK/year)	
<250 000	2390 (34.5)
≥250 000	4548 (65.6)

Descriptive statistics

Quantitative variables

- Variability
- Centrality



Descriptive statistics

Quantitative variables

- Centrality
  - Mean: Sum of all values divided by number of values
  - Median: The middle value when all values are ranked
  - Mode: The value occurring most often

Heigh of 9 individuals:

147 – 151 – 151 – 152 – 153 – 154 – 155 – 156 – 159

Mean= 1378/9=153.11

Median= 153

Mode= 151

When the sample size is even:

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5

Descriptive statistics

- Mean or median?

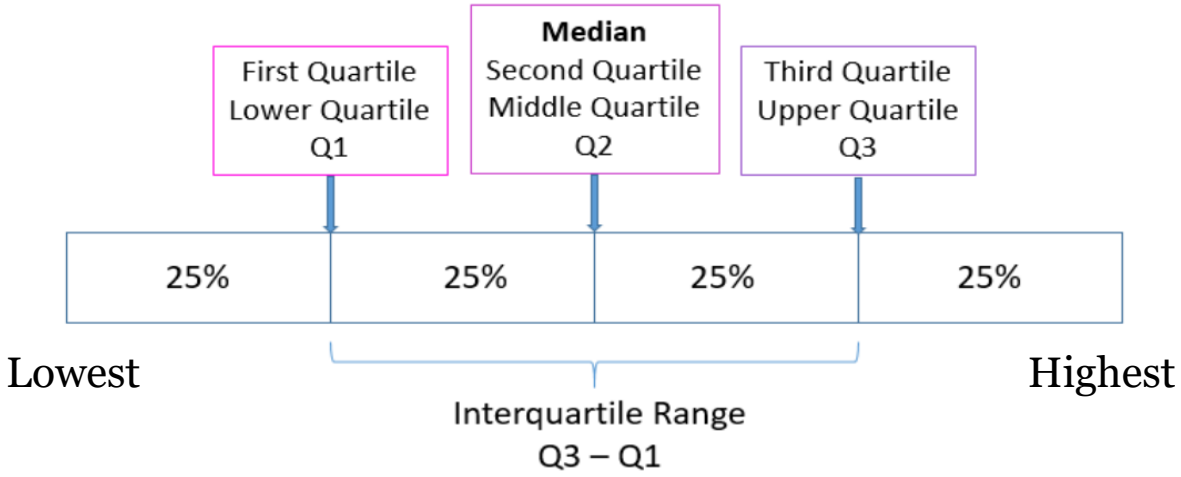
Salary (K SEK) of 10 individuals									
20	21	24	20	24	23	24	23	25	100

Mean=30.4

Median=(24+23)/2=23.5

Quantitative variables

- Variability
  - Range: the distance between the highest and lowest values
  - Interquartile range: First, we rank the data and divide it into four equal-sized sub-groups. Then, we get three cut points (lowest to highest :Q1, Q2, Q3). Q3-Q1 is the interquartile range.



Height (cm) of 10 individuals

147 – 151 – 151 – 152 – 153 – 154 – 155 – 156 – 159 – 162

Min Q1 Q3 Max

Range = [147, 162] = 15

Interquartile range (IQR) = [151, 156] = 5

Descriptive statistics

Quantitative variables

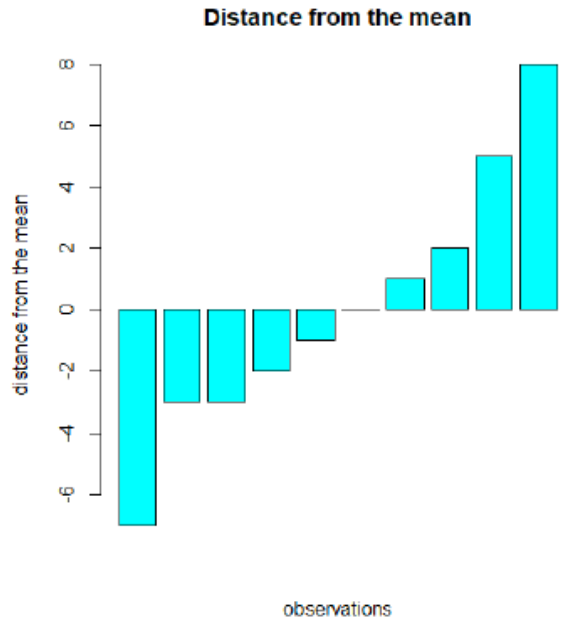
- Variability----measurement of dipersion
  - Variance: the average of the squared differences from the mean value

Height (cm) of 10 individuals

x	x-μ
147	-7
152	-2
155	+1
156	+2
151	-3
153	-1
151	-3
159	+5
162	+8
154	0
Σ	0

where x is the height (cm)
μ the mean, corresponding to 154
∑ the summation

$$\frac{\Sigma(x-\mu)}{N} = \frac{0}{10} = 0$$



Descriptive statistics

Quantitative variables

- Variability----measurement of dipersion
  - Variance: the average of the squared differences from the mean value

Height (cm) of 10 individuals

x	$(x-\mu)$	$(x-\mu)^2$
147	-7	49
152	-2	4
155	+1	1
156	+2	4
151	-3	9
153	-1	1
151	-3	9
159	+5	25
162	+8	64
154	0	0
Σ	0	166

where x is the height (cm)
μ the mean, corresponding to 154
∑ the summation

$$\frac{\Sigma(x-\mu)^2}{N} = 16.6$$ Variance

$$\sqrt{\frac{\Sigma(x-\mu)^2}{N}} = 4.1$$ Standard deviation:

square root of the variance.

Example for descriptive statistics

Table 3. Clinical characteristics of the six subjects involved in serine supplementation study.

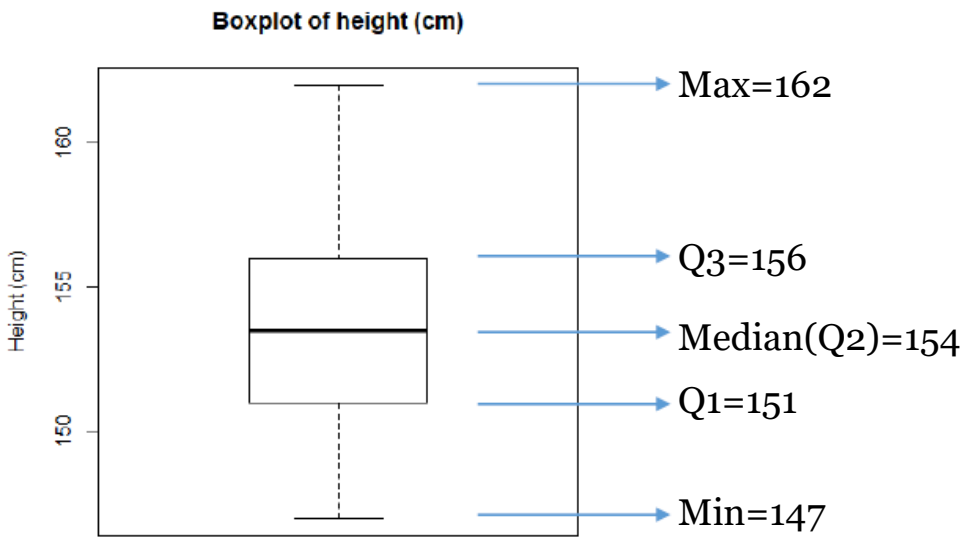
Clinical variable	Baseline (n = 6)	After serine (n = 6)	P-value
Liver fat (%)	26.8 ± 6.0	20.4 ± 7.0	**<0.05**
Age (years)	56.7 ± 5.2	56.7 ± 5.2	–
Weight (kg)	103.0 ± 14.3	103.0 ± 13.9	–
Body mass index (BMI) (kg/m^2)	32.5 ± 2.70	32.5 ± 2.60	–
Alanine aminotransferase (ALT) (U/l)	50.8 ± 15.2	37.6 ± 5.3	**<0.05**
Aspartate aminotransferase (AST) (U/l)	34.5 ± 8.10	27.4 ± 8.4	**<0.05**
Alkaline phosphatase (ALP) (U/l)	76.3 ± 17.2	71.3 ± 17.9	**<0.05**
γ-glutamyl transferase (GT) (U/l)	63.8 ± 12.9	62.3 ± 16.3	0.30
Fasting plasma glucose (mmol/l)	6.57 ± 1.41	6.33 ± 1.41	0.25
Fasting plasma insulin (FPI) (pmol/l)	46.3 ± 33.8	34.7 ± 25.2	0.23
HOMA-IR	2.15 ± 1.85	1.54 ± 1.49	0.18
LDL cholesterol (mmol/l)	3.68 ± 0.80	3.85 ± 0.94	0.50
HDL cholesterol (mmol/l)	1.00 ± 0.21	1.02 ± 0.18	0.30
Plasma triglycerides (TG) (mmol/l)	6.90 ± 6.65	3.63 ± 1.81	0.13
Total cholesterol (mmol/l)	6.23 ± 1.49	5.85 ± 1.15	0.18
Bilirubin (μmol/l)	7.33 ± 4.11	6.48 ± 3.94	0.13

Data are presented as means ± SD. P-value (calculated using Student's t-test) indicates the significance level of difference before and after the oral supplementation of serine. Bold text indicate significantly different values.

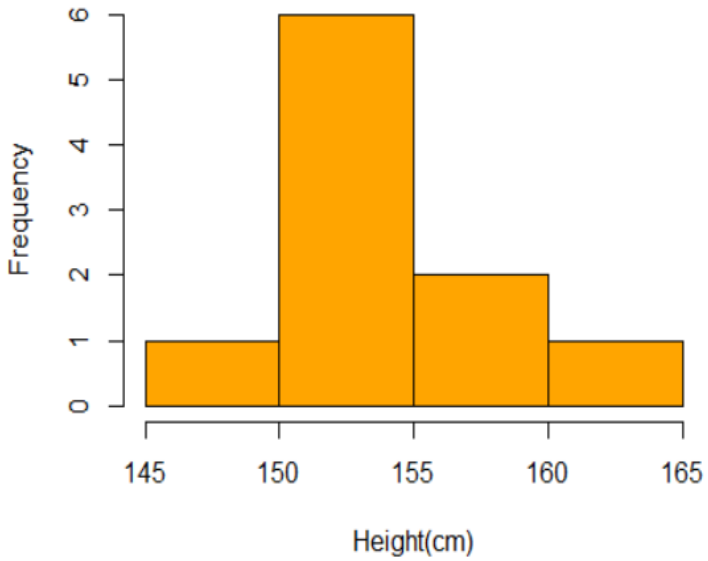
Data visualization

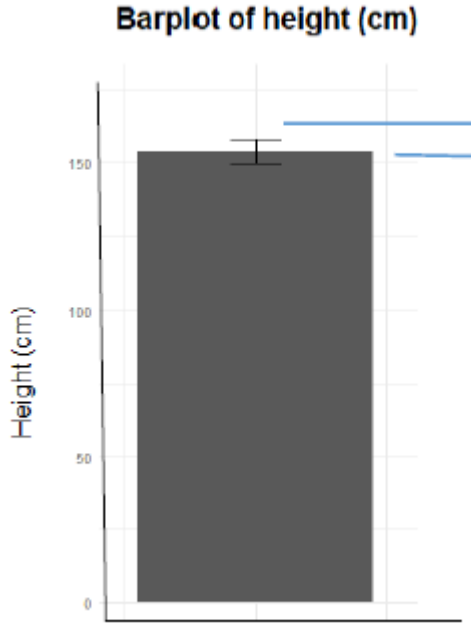
Height (cm) of 10 individuals

147 – 151 – 151 – 152 – 153 – 154 – 155 – 156 – 159 – 162



Histogram of height



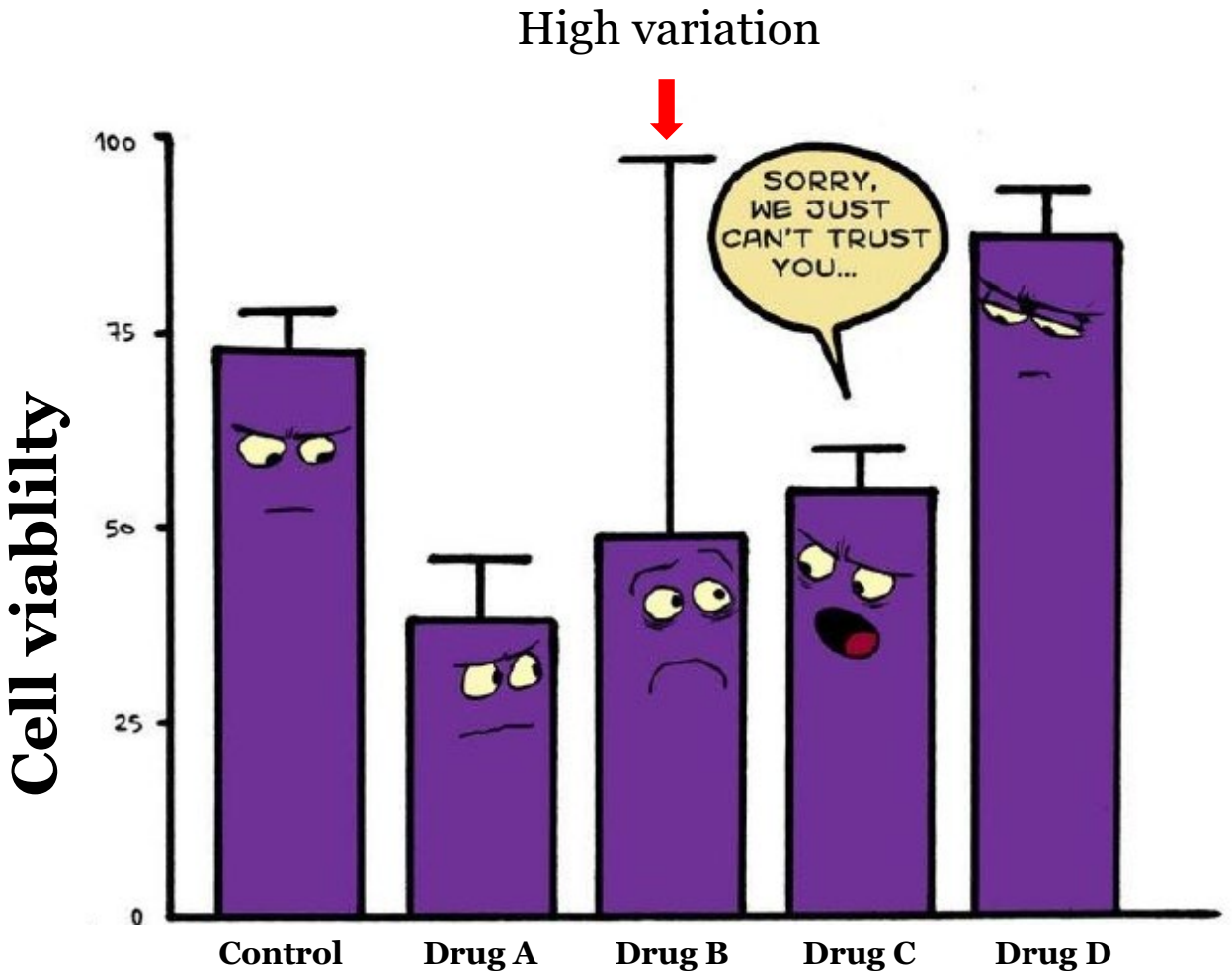


Mean±SD

Group A

R tips: boxplot(), hist(), barplot(), ggplot2

Data visualization



Thank you & questions?