Statistics 101:

"Academic discipline dealing with all aspects of data."

learning from

quantification

Perspectives:

- art of summarizing data
  ↳ make data comprehensible

- science of uncertainty
  ↳ most information in the world is uncertain

- Science of decisions
  ↳ ultimate goal of statistics

- science of variation
  ↳ central tendency and spread

- art of forecasting

- Science of measurement and data collection.

Foundations of data

Source of data

- Designed data — "artificially collected"
   (surveys, studies etc)
- Organic data
   (process generated)

For both, data needs to be i.i.d
"independent", "identically distributed".
→ more on this later!

Question: What is the <u>source</u> of NHANES data?

Types of data:

- Some data is not numeric!
   For instance race or gender
- Just as we have data types in programming languages, we have different types here.

- Weight — numeric, continuous

- | # of kids —— numeric, discrete

- | Age group (child, adult, elder) —— categorical, ordered
   (0-, 1-, 2-)

- Gender —— categorical, (unordered)

Practical Note:

Gender represented as: (M / F)

or: (0 / 1) → But still unordered!

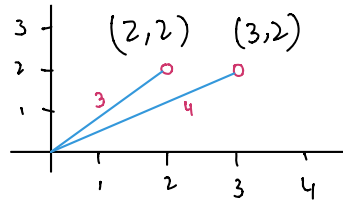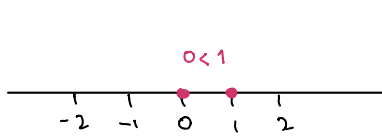or: $\begin{bmatrix} 1 \\ 0 \end{bmatrix} / \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ → now unordered!

"one-hot vector representation"

B: 0
W: 1
H: 2

$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ → 2
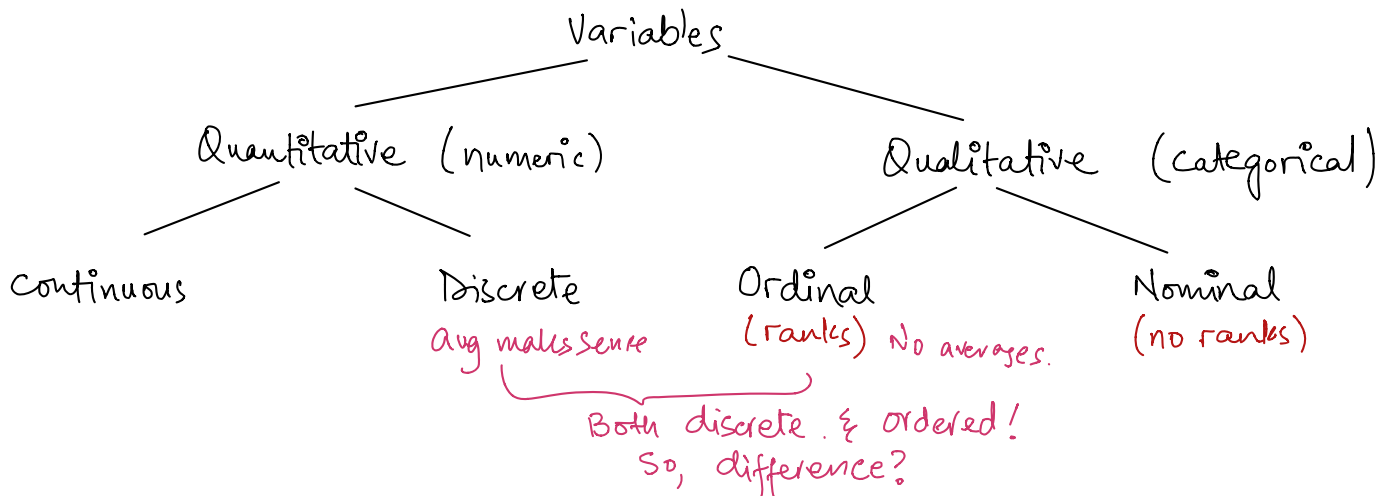
B W H

$0 < 1$

$-2 \quad -1 \quad 0 \quad 1 \quad 2$

$(2,2) \quad (3,2)$

3

4

$3 < 4$

$(2,2) < (3,2)$

un comparable
not

Variables

Quantitative (numeric) — Qualitative (categorical)

Continuous — Discrete — Ordinal — Nominal

Discrete: *Avg makes sense*

Ordinal (ranks) *No averages.*

Nominal (no ranks)

Both discrete & ordered!
So, difference?
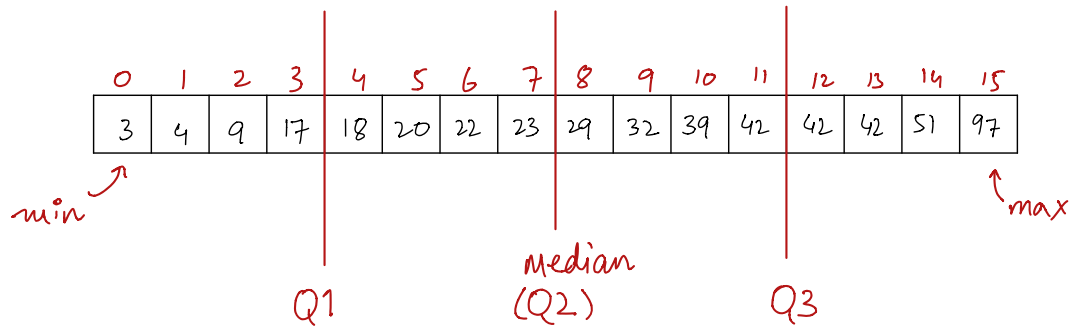
Let's see these in NHANES

— Categorical Data:
— Frequency tables — SQL "group by" & count()
— Bar charts
— Pie charts *eww*

— Quantitative data
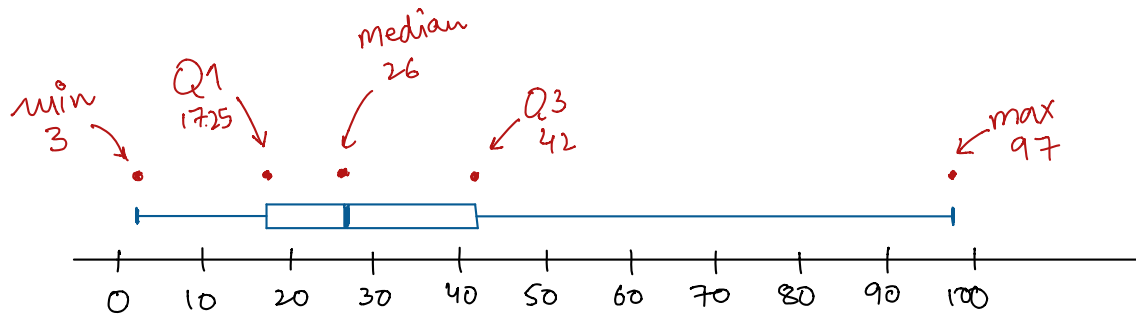— "Average" — Arithmetic mean

Problem — outliers

— Geometric mean
— Harmonic mean

— Median
— Mode

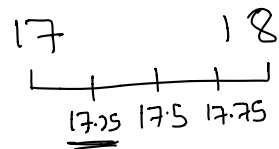| 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	4	9	17	18	20	22	23	29	32	39	42	42	42	51	97

min

max

Q1   Median (Q2)   Q3

Magnitude of values is not clear!

median
26

Q1
17.25

Q3
42

min
3

max
97

0   10   20   30   40   50   60   70   80   90   100

Box Plot.

Computing quantiles:

0.25 — Q1

$$\left[0.25 \, (16 + 1) \right]^{th} \text{ value.}$$

$$4.25^{th} \text{ entry} \Rightarrow$$

17   18

17.25   17.5   17.75

0.25 Quantile = 17.25