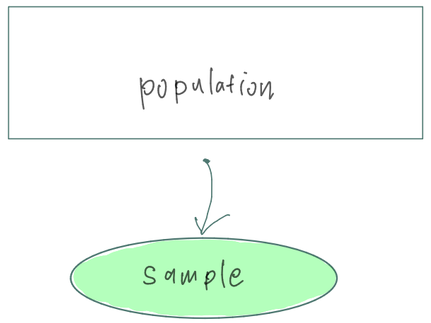
2023S1

Module 1: Exploring Data

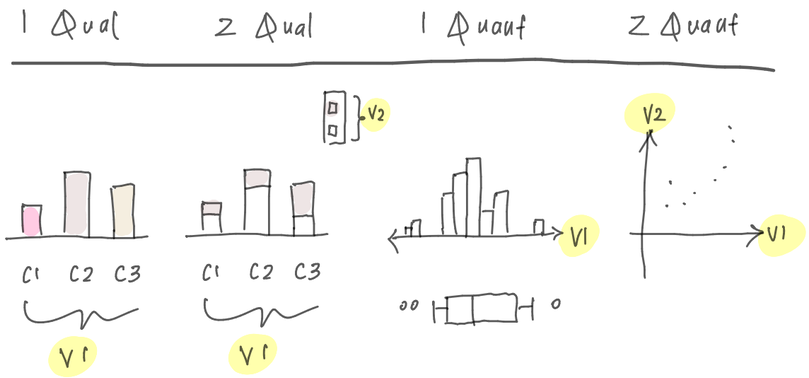


T1 : Design of Experiments

T2 : Data & Graphical Summaries

T3 : Numerical Summaries

LO3: Produce, interpret & compare graphical & numerical summaries using base R & ggplot.



Graphical Summaries

Numerical Summaries

most common category

(mean, SD)

(median, IQR)

$\hat{y} = a + bx$

$(\bar{x}, \bar{y}, SD_x, SD_y, r)$

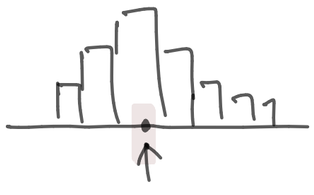
Module 2

↑

Why? · simple but informative!
· good for communication & comparisons

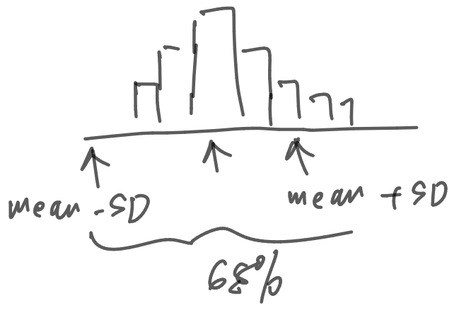
Main features — max & min
— centre (mean, median)
— spread (SD, range, IQR)

Mean = average
= sum / size
= balancing point
uses all data.



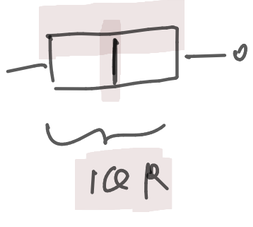
Median = middle point
odd / unique
even — average of 2 middle points

SD = standard deviation
= RMS of (gaps from the mean)
= $\sqrt{\text{mean of } (\text{gaps from mean})^2}$



pop SD
vs
sample SD

IQR = interquartile range
= $Q_3 - Q_1$
= 75% percentile
- 25% percentile



Notation (2nd year)

$$\begin{cases} \text{data} & x_1, x_2 \cdots x_n \\ \text{ranked} & x_{(1)}\ x_{(2)} \cdots x_{(n)} \\ \text{sum} & \sum_{i=1}^{n} x_i \end{cases}$$

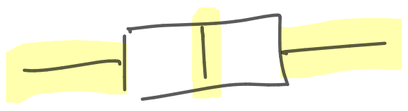
RMS
= root mean square
= $\sqrt{\text{mean of } (\text{data})^2}$

Q: Which summary do we use?

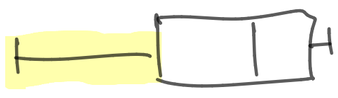
- context
- (mean, SD) vs (median, IQR) — robust
- sometimes none!

bimodal data 

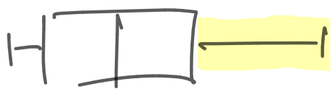
- pop SD = $\sqrt{\frac{n-1}{n}}$ sample SD



symmetric
mean ≈ median



left skewed
mean < median



right skewed
mean > median

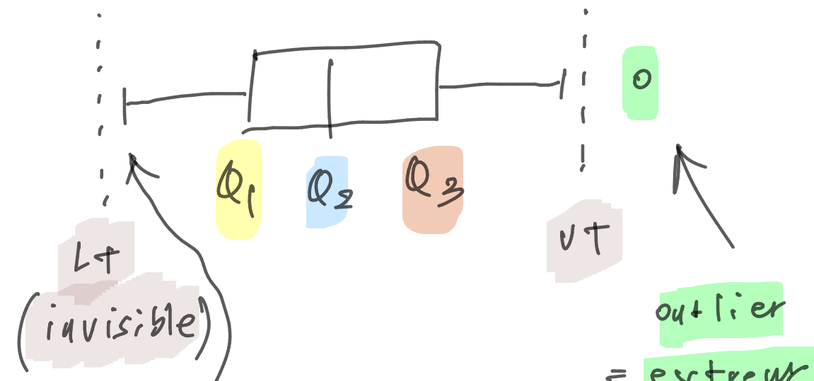
Other Summaries

SU = standard units
= $\frac{\text{data point} - \text{mean}}{\text{SD}}$

CoV = coefficient of variation
= $\frac{\text{SD}}{\text{mean}}$

Q: How to draw a boxplot?

Data → Q_1, Q_2, Q_3 (IQR) → LT = $Q_1 - 1.5$ IQR; UT = $Q_3 + 1.5$ IQR



nearest data point within threshold