Week 2

Monday, 2 August 2021 23:36

BIOSTATISTICS

CONTINUOUS DATA MEASURES

NUMERICAL WAYS TO SUMMARIZE CHARACTERISTICS OF SAMPLE

- MEASURE OF TENDENCY
- MEASURE OF VARIABILITY
- MEASURE OF LOCATION IN DISTRIBUTION
- COMPARING DISTRIBUTIONS
- DIFFERENCE IN SAMPLE SIZES

- USEFUL SUMMARY STATISTICS

CENTRAL TENDENCY	MEDIAN, MEAN
VARIABILITY	STANDARD DEVIATION
LOCATION	PERCENTILES

$$\text{MEAN} = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}$$

$\bar{x}$ SAMPLE NOT μ POPULATION MEAN

μ CAN ONLY BE ESTIMATED BY $\bar{X}$

MEAN IN SMALL SAMPLE IS SENSITIVE TO EXTREME VALUE JUST ONE VALUE CAN VARY IT BY A LOT.

MEDIAN — MIDDLE VALUE IN ORDERED SET OF CONTINUOUS DATA

50% PERCENTILE, II QUANTILE
LESS SENSITIVE TO EXTREMES

ONLY AFFECTED BY RELATIVE POSITIONS OF THE VALUES

$$\text{MEDIAN} = \begin{cases} x_{(N/2)} & \frac{N}{2} \in \mathbb{N} \\ \frac{x_{\lfloor N/2 \rfloor} + x_{\lfloor N/2+1 \rfloor}}{2} & \frac{N}{2} \notin \mathbb{N} \end{cases}$$

$$\text{VARIANCE} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})}{N-1} = S^2$$

MINIMAL IMPACT
IMPORTANT TO CORRECT SINCE WE DON'T KNOW μ

CUMULATIVE SQUARE DISTANCE AVERAGED

STANDARD DEVIATION $S = \sqrt{S^2}$

HOW FAR ON AVERAGE EVERY SINGLE OBSERVATION IS FROM THE SAMPLE MEAN

$S = 0$ NO VARIABILITY $X_1 = X_2 = \cdots = X_N$

S^2 BEST ESTIMATE OF σ^2

PERCENTILES $\gamma \in [0, 1]$

γTH PERCENTILE THAT HAS γ% OF DATA LESS OR EQUAL

AND $1-\gamma$% MORE THAN VALUE

$$\psi_\gamma = \begin{cases} \chi_{(m\gamma)} & \text{IF } m\gamma \in \mathbb{N} \\ \frac{\chi_{\lfloor m\gamma \rfloor} + \chi_{\lfloor m\gamma+1 \rfloor}}{2} & \text{IF } m\gamma \notin \mathbb{N} \end{cases}$$

BEST ESTIMATES OF UNKNOWN UNDERLYING POPULATION VALUES

$\overline{X_N} \longrightarrow \mu$

$S_N^2 \longrightarrow \sigma^2$

$N \longrightarrow \infty$

• VISUAL DISPLAYS

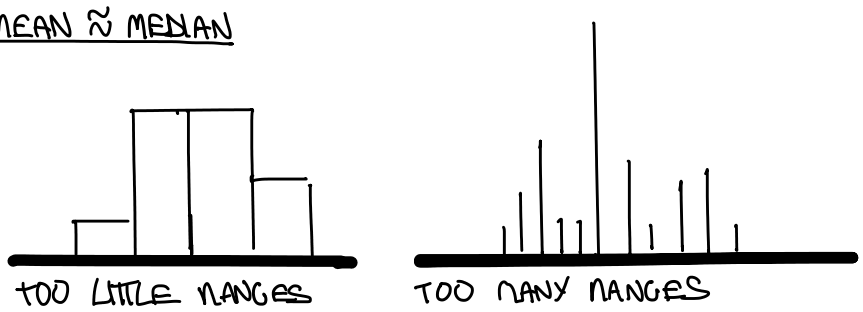
HISTOGRAM, BOXPLOT } VISUALIZING DISTRIBUTION

USEFUL TO COMPARE DATA FROM DIFFERENT SAMPLES

COUNTING # OF DATA IN A PREDEFINED RANGE

$\overline{X} \approx m$ NO SPECIAL TENDENCY (SYMMETRY AROUND CENTER)

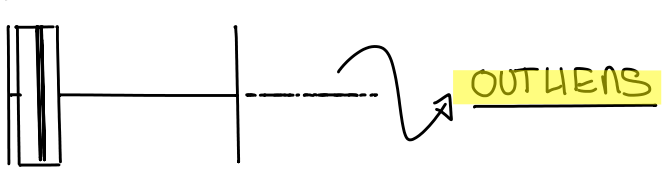
MEAN ≈ MEDIAN



TOO LITTLE RANGES TOO MANY RANGES

SKEWED $\overline{X} > m$

(RIGHT) POSITIVE SKEW



OUTLIERS

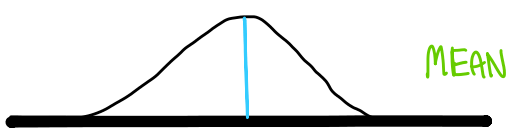
$IQR = Q_3 - Q_1$

INTER QUARTILE RANGE

$MAX = Q_3 + \frac{3}{2} IQR$

$MIN = Q_1 - \frac{3}{2} IQR$

SYMMETRIC, BELL-SHAPED



MEAN ≈ MEDIAN ≈ MODE

THE ROLE OF SAMPLE SIZE

RANDOM SAMPLE MIMICS CHARACTERISTICS

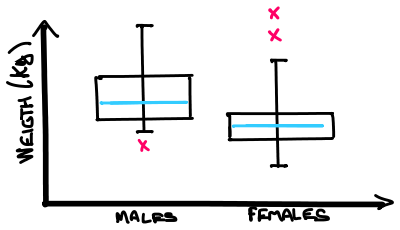
OF THE LARGER POPULATION

LARGER SAMPLE ⟶ MINIMIZING VARIANCE

COMPARING CONTINOUS DISTRIBUTIONS

VISUALLY

SIDE BY SIDE BOXPLOTS ARE EASIER TO INTERPRET THAN HISTOGRAMS



NUMERICALLY

$\bar{x}_{MALES} - \bar{x}_{FEMALES} =$ AVERAGE SHIFT

STUDING THE DIFFERENCE IN MEANS

$$\delta = \mu_1 - \mu_2 \approx \overline{X_1} - \overline{X_2}$$

ON AVERAGE 1 DIFFERS FROM 2 BY δ (KG)

SKEWED DISTRIBUTIONS ARE DIFFICULT TO ANALYZE EASILY

"STUDYING THE SHIFT OF MASS"

IF MORE THAN ONE GROUP WE CAN PICK ONE AS THE 'REFERENCE'

IMPORTANT

μ, δ PARAMETERS OF POPULATION CAN'T WE DERIVED DIRECTLY.

$\overline{X_N}, S_N$ ARE CALCULATED BASED ON THE CHOOSEN SAMPLE (IMPERFECT REPRESENTATIONS)

$\overline{X_N}, S_N, m$ CHANGE DEPENDING ON SAMPLE

THERE IS NO REAL WAY TO PREDICT HOW THEY WILL CHANGE GIVEN A DIFFERENT SAMPLE

S_N DECREASES IF N INCREASES ⟵
(ONLY WORKS IF $N_0 < 30$)