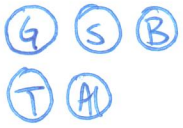
DATA TYPES

NOMINAL.



ORDINAL.

INTERVAL

RATIO.

"RANK"

← SCALAR →

← DISCRETE →

← CONTINUOUS →

- buckets

ordering

- regular, meaningful distances b/n each point.
- arbitrary zero.

---

preference

performance: speed, errors.

‖ across people.

DESCRIPTIVE STATISTICS.

CENTRAL TENDENCY.

MEAN.

MEDIAN

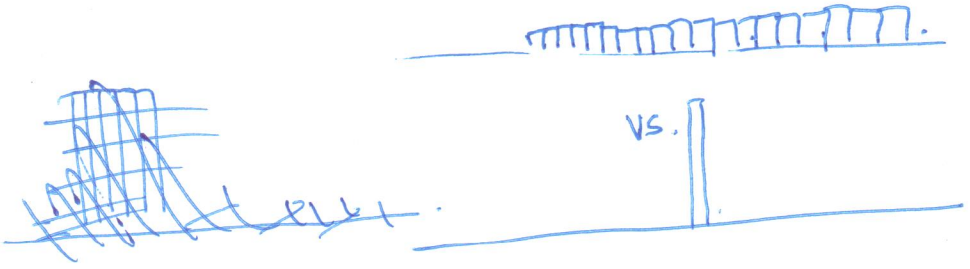
MODE.

MEAN: $\frac{\Sigma x_i}{n}$.

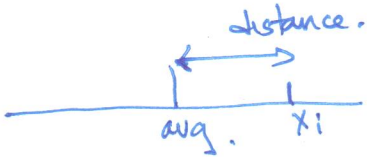
MEDIAN: protection from extreme outliers.

MODE: useful for nominal data.

DISPERSION.



COMMON MEASURE: RANGE.
VARIANCE



$$\sigma^2 = \frac{\sum (\bar{x} - x_i)^2}{n}$$

↖ squared values.

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

↖ std dev.

~~STANDARD ERROR.~~

WHY INFERENTIAL STATS.

MIKE 5'3"

~~MIKE 6'2~~

LARRY 5'5"

~~LARRY~~

"LARRY IS TALLER THAN MIKE" —
- comment about the sample.

NED 5'6"

GAIL 5'9"
HEATHER 6'1"
IVY 5'11"

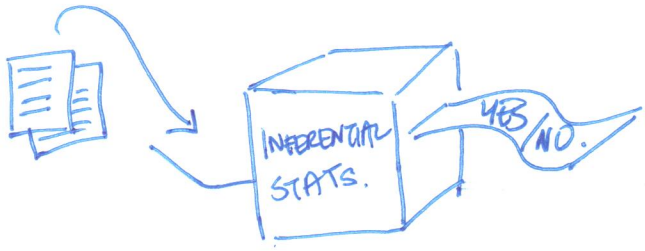
AVG ~5'4⅔" ~5'11"

"Men are shorter than women"
- comment about the populatio

- ☐ common sense.
- ☐ I can manipulate the sample. (sampling bias)
- ☐ small group size.
- ☐ these are not representative.

IDEAL.

0. HAVE A HYPOTHESIS.
1. COLLECT DATA
2. INSERT DATA INTO MACHINERY
   DATA.



3. GET YES / NO.

INFERENCE ⇒ BEST / INFORMED GUESS.

NULL-HYPOTHESIS ~~STATISTICAL~~ SIGNIFICANCE TESTING.

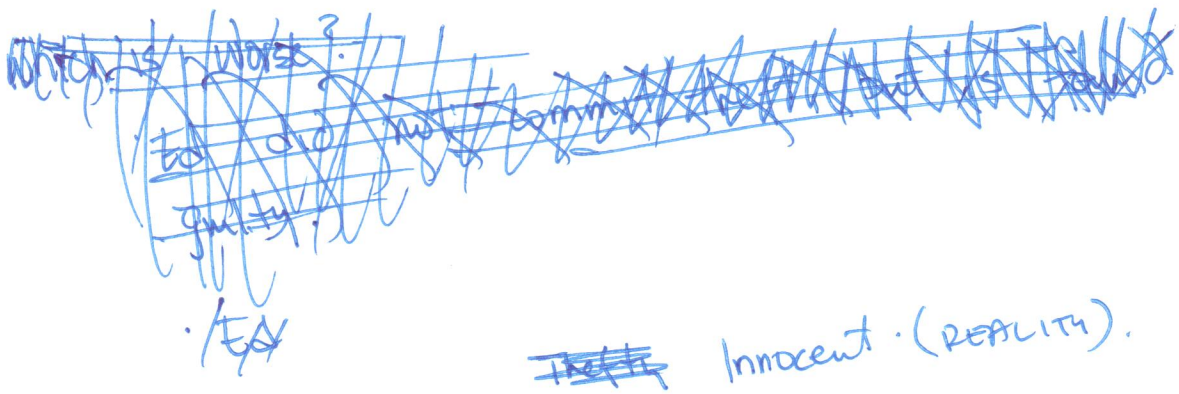
PROCEDURE

1. Generate a null hypothesis (H_0)
2. Generate an alternate hypothesis (H_a)
3. Select α value. (Probability of a Type I error)
   ↳ Reject H_0 when H_0 is true
4. Run test, get p-value.
   If $p < \alpha$ reject H_0.
   else, cannot reject H_0.

* Nothing magical going on here. *.

TYPE I & TYPE II ERRORS

Honest Ed is accused of theft.



Innocent. (REALITY).

VERDICT. (INNOCENCE)	TRUE	FALSE
TRUE	Ed did not commit theft & he is innocent.	TYPE II.
FALSE	TYPE I	Ed committed theft & is found guilty.

H_0 = Ed is INNOCENT.

Normally, we say TYPE I errors are worse than TYPE II errors.

TYPE I ERROR = ~~ACC~~ REJECTING H_0 WHEN H_0 IS TRUE

"GUILTABILITY"

Ex: New touch screen ATMs vs. REGULAR BUTTON ATM.

H_0: People make same # of errors w TS as RB

H_0: People can use @ same speed TS as RB.

⇒ Expensive to change over.

Probability (Type I error) = $P(\text{rejecting } H_0 \mid H_0 \text{ is true}) = \alpha$.

$\alpha = 0.05$ or $\alpha = 0.01$. (Convention)

STATS TEST.

t-test

ANOVA. $\Rightarrow$ p-value.

⋮

p-value: likelihood of getting this data if H_0 were true.

NOT

likelihood H_0 is true/false.

If $p < \alpha$, reject H_0. $\Rightarrow$ "SIGNIFICANT DIFFERENCE"

else, cannot reject H_0.

MYTHS OF NULL HYPOTHESIS TESTING

(1) P is NOT probability of H_0 being true.

It is likelihood of getting this, or more extreme data if H_0 is true.

H_0 true is binary ⇒ YES OR NO.

(2) P-VALUE DOESN'T SAY "MORE/LESS" SIGNIFICANT, NOT BORDERLINE SIGNIFICANT.

P-value is what it is

α is set in advance. ⇒ no cheating!

(3) α IS ARBITRARY.

Nothing magical.

(4) SIGNIFICANCE ≠ IMPORTANT

↳ interpretation is important.