Ch.8 INTRO. TO STATISTICS

8.1 DESCRIPTIVE STATISTICS

BASIC CONCEPTS:

- population vs sample
- target population
- parameter (↙ population) vs statistic (↘ sample)
- descriptive (↙ study sample only) vs inferential (↘ study population based on sample) statistics
- Data refers to sample in most cases
  - but ↳ census data = from population
- quantitative vs qualitative data } types of variables
- discrete vs continuous

· we dont collect data in this course

SYMBOLS:

	POPULATION PARAM:	SAMPLE STATISTICS:
mean:	μ	$\bar{x}$
med:	$\tilde{\mu}$	$\tilde{x}$
variance:	σ^2	s^2
SD:	σ	s
Proportion (%):	p	$\hat{p}$
Size:	N	n

WHEN WE MAKE AN INFERENCE BASED ON A SAMPLE, DIFFERENT SAMPLES LEAD TO DIFFERENT CONCLUSIONS,
HOW CAN WE SAMPLE TO MAKE CONCLUSION VALID?

A. You need a random sample or it will be invalid

8.2 SUMMARIZING DATA NUMERICALLY 12:01

↳ we compute #'s from the data

↳ compute data then graph it

HOW TO → CALCULATOR FOR MEAN & MEDIAN

1) enter data in L1 (STAT → ENTER)
2) STAT → CALC → ENTER → PICK L1 → CALCULATE ENTER

MEASURE of CENTER aka the TYPICAL VALUE

MEAN: average

MEDIAN: the one in the middle after ordered from smallest to largest.
take the average of the two in the middle if there are even #'s of observations
→ median means half of the observations are above & the other half is below

THEY ARE NOT THE SAME! they can be very close

- use median when data is skewed, median is resistant to outliers

MEAN IS SENSITIVE TO OUTLIERS, IT IS MISLEADING WHEN DATA IS SKEWED
↑ easier to deal w/ mathematically, use this when data not skewed

EXAMPLE 1:

IN A CERTAIN CLASS OF 13 STUDENTS, 10 SHOWED UP THE FIRST EXAM, WHILE 3 BLEW IT OFF:

HERE ARE THE GRADES IN ORDER:

0 0 0 55 68 78 79 81 84 87 93 94 98

WHAT IS THE MEDIAN?

↳ INCLUDING ALL STUDENTS: 79

↳ IGNORE STUDENTS WHO SLEPT IN: 82

WHAT IS THE AVERAGE? (MEAN)

↳ INCLUDING ALL STUDENTS: ≈ 62.8462

↳ IGNORE STUDENTS WHO SLEPT IN: 81.7

CONCLUSION:

- MEAN < MEDIAN w/ 3 0's
- MEAN CLOSE TO MED. W/O 3 0's

EXAMPLE 2 (1 CONTINUED):

SUPPOSE ONE STUDENT GOT 980 INSTEAD OF 98, HOW WOULD IT AFFECT MEAN N MEDIUM?

- MEAN: WOULD BE MUCH LARGER
- MEDIAN: NOT AFFECTED

- GENERALLY SPEAKING -

OUTLIERS	NO OUTLIERS
MEDIAN	MEAN
IQR	SD

MEASURE of VARIATION / SPREAD

VARIANCE S^2: $\frac{1}{n-1}\sum_{n=1}^{n}(x_i-\bar{x})^2$ (deviation)

- WHY IS IT SQUARED? DEVIATION CAN BE +/-, $\sum_{i=0}^{n}(x_i-\bar{x})$ can = 0
- WHAT HAPPENS TO THE UNIT? IT GETS SQUARED
- we don't do abs. because its hard to do w/ math, square is easier

STANDARD DEVIATION SD: $= s = \sqrt{\text{variance}} = \sqrt{S^2}$ = HOW "SPREAD OUT" THE DATA IS; HOW MUCH THE VARIANCE IS; AVG. DISTANCE FROM MEAN

EXAMPLE 3:

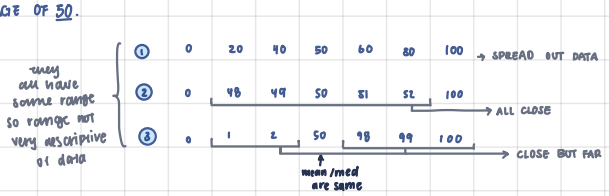
EACH OF THE FOLLOWING LISTS HAS AN AVERAGE OF 50.

OF WHICH IS THE SD GREATEST?

↳ 3

SMALLEST?

↳ 2



BY CALCULATOR:

- SD ≈ 38.157
- SD ≈ 28.896 HAS MOST CLOSE TO 50 #'s
- SD ≈ 49.007 ALL #'S FAR FROM 50

RANGE: largest - smallest (we don't use this, we use IQR)

IQR (INTERQUARTILE RANGE): $Q_3 - Q_1$

- Q_3 = third quartile; median of the upper half after data ordered
- Q_1 = first quartile; median of the lower half after data ordered
- Q_2 = median

IQR VS SD

IQR	SD
RESISTANT TO OUTLIERS	SENSITIVE TO OURLIERS
USE FOR SKEWED DATA	USE FOR NORMAL DATA

IQR OF EXAMPLE 1:

IQR = 90 - 27.5 = 62.5

EXTRA QUESTIONS (MAY BE ON QUIZZES)

★ eg. 0 10 20 30 40 50 60

MEAN: 30 SD: 21.60247

eg +10: 10 20 30 40 50 60 70

MEAN: 40 SD: 21.60247

eg * 2: 20 40 60 80 100 120

MEAN: 60 SD: 43.2049

- IF A CONSTANT C IS ADDED TO EACH DATA VALUE, WHAT HAPPENS TO MEAN n STANDARD DEVIATION?
  - mean will increase by C
  - SD will stay the same
- IF EACH DATA IS MULTIPLIED BY A CONSTANT C, WHAT HAPPENS TO MEAN n STANDARD DEVIATION?
  - mean gets multiplied by C
  - SD is multiplied by C as well

★ Qualitative data (categorical data) cannot have MEAN OR SD, but can have percentage. (AKA STATISTICS)

8.3 SUMMARIZING DATA GRAPHICALLY

- FREQUENCY DISTRIBUTION N HISTOGRAM
- FIVE NUMBER SUMMARY N BOX-WHISKER PLOT, PARALLEL PLOTS
- STEM-LEAF PLOT, SIDE-BY-SIDE STEM-LEAF PLOTS

WHAT INFO DO WE GET FROM GRAPH?

- center
- variation
- distribution shape
  - ↳ symmetrical, left skewed, right skewed

EXAMPLE 4: **STEM-LEAF PLOT**

TWO SAMPLES OF WOMEN AND MEN WERE COLLECTED & THEIR HEIGHTS WERE MEASURED.

WOMEN: 62 64 61 70 67 59 67 62 68 61
65 59 61 64 59 67 60 60 65 60

MEN: 68 70 65 74 70 60 67 60 98 72
69 64 68 72 70 70 79 74 75 66

1) ENTER INTO STATCRUNCH → STEM-LEAF
(TI-84 DON'T HAVE IT)

2) DO IT MANUALLY

D : XXX
↑ 10's digit = STEM
↑ 1's digit = LEAF

WOMEN DATA:

REPEATING STEM: 6 APPEARS AS STEM TWICE BECAUSE TOO MANY

5: 999
6: 000111224 4
6: 55 7778
7: 0

MEN DATA:

6: 004 < 5
6: 567889 ≥ 5 (also repeating stem)
7: 0000224 4 < 5
7: 589 ≥ 5

TRUNCATION:

USED IF OUR DATA SET IS LIKE
720 713 731
681 684 627
↳ we take 1st 2 digits

7 : 223
6 : 2
6 : 88
↑ 100's ↑ 10's

EXAMPLE 4: **SIDE-BY-SIDE**

BY USING THE SAME STEM, RIGHT = WOMEN, LEFT = MEN

WOMEN		MEN
999	5	
44 22 11 000	6	004
877755	6	567889
0	7	0000 2244
	7	589

- BOTH SYMMETRIC
- PRETTY MUCH SAME VARIATION

HOW TO → **CALCULATOR** BOX WHISKER PLOT

1) ENTER DATA TO L1
2) 2ND + Y=, turn on plot 1 + turn off plot 2 + 3
   ↳ UNDER PLOT 1 SELECT BOX-WHISKER ICON (THE ONE WITH OUTLIERS)
   ↳ SELECT L1 AS LIST
3) TRACE TO SEE MIN, MAX, Q_1, Q_2, Q_3, OUTLIERS

FIVE NUMBER SUMMARY: min, Q_1, Q_2, Q_3, max

BOX-WHISKER: Q_1, Q_2, Q_3 form the box, min + max are end of the whiskers

OUTLIER: any points above Q_3 + below Q_1 by 1.5iqr

IF OUTLIERS EXISTS, THEN WE MARK THEM W/ A CIRCLE, + THE WHISKER ENDS AT THE SMALLEST OR LARGEST DATA POINTS ARE NOT OUTLIERS.

PARALLEL BOX PLOT

EXAMPLE:

0	49	50	51	54	60	74	75	76	98	100
0	1	3	15	20	60	89	91	95	99	100