WK 10 - 1

Ch.8 INTRO. TO STATISTICS

8.1 DESCRIPTIVE STATISTICS

BASIC CONCEPTS:

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

· 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 <u>VALID</u>?

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 N MEDIAN

1) enter data in L1 (STATS → 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 <u>even</u> #'s of observations

→ median means half of the observations are above + the other half is below

THEY ARE <u>NOT</u> 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: $\approx$ 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 dont 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

they all have same range so range not very descriptive of data

① 0 20 40 50 60 80 100 → SPREAD OUT DATA

② 0 48 49 50 51 52 100 → ALL CLOSE

③ 0 1 2 50 98 99 100 → CLOSE BUT FAR

mean/med are same

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 dont 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

↓ RESISTANT TO OUTLIERS, USE FOR SKEWED DATA

↓ SENSITIVE TO OUTLIERS, USE FOR NORMAL DATA

IQR OF EXAMPLE 1:

median

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

not included in Q_1 + Q_3

Q_1 lower: med = 27.5

Q_3 UPPER: med = 90

IQR = 90 - 27.5 = 62.5

EXTRA QUESTIONS (MAY BE ON QUIZZES)

★ og: 0 10 20 30 40 50 60

MEAN: 30 SD: 21.60247

og +10: 10 20 30 40 50 60 70

MEAN: 40 SD: 21.60247

og * 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 69 60 78 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

0 : XXXX

10's digit = STEM, 1's digit = LEAF

WOMEN DATA:

REPEATING STEM: 6 APPEARS AS STEM TWICE BECAUSE TOO MANY

5: 999
6: 0001112244
6: 557778
7: 0

MEN DATA:

6: 004 < 5
6: 567889 ≥ 5
7: 00002244 < 5
7: 589 ≥ 5

also repeating stem

TRUNCATION:

USED IF OUR DATA SET IS LIKE
720 723 731
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	
4422111000	6	004
877755	6	567889
0	7	00002244
	7	589

- BOTH SYMMETRIC
- PRETTY MUCH SAME VARIATION

HOW TO → CALCULATOR

BOX WHISKER PLOT

1) ENTER DATA TO L_1
2) 2ND + Y= , turn on plot 1 + turn off plot 2 + 3
   - ↳ UNDER PLOT 1 SELECT BOX-WHISKER ICON (THE ONE WITH OUTLIERS)
   - ↳ SELECT L_1 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 $1.5 I_{qr}$

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

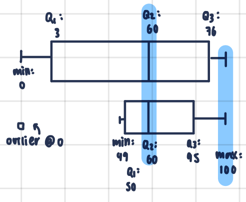
wk 10-2

PARACLLE BOX PLOT

USED 4 COMPARISON

EXAMPLE:

① 0 49 50 51 54 60 74 75 76 78 100
② 0 1 3 15 20 60 89 91 95 99 100 →



same med, max
different variation

FREQUENCY DISTRIBUTION

• HAS 3 COLUMNS

↳ 1st col: ranges of the data (classes)

↳ 2nd col: the count (frequency) for each class

↳ 3rd col: the percentage (relative frequency)

∘ gets frequency from histogram

EXAMPLE:

	CLASS	FREQUENCY	RELATIVE FREQUENCY
[9, 34)	9 - < 34	7	
[34, 59)	34 - < 59	15	
	59 - < 84	6	
	84 - < 109	1	
	109 - < 134	0	
	134 - < 159	1	

↑ end point belong to class on the right

HISTOGRAM

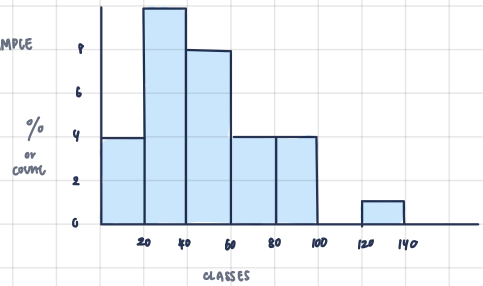
: BAR CHART W/O GAPS

∘ helps to see variation

↳ HEIGHT OF BAR = % OR COUNT

↳ END POINT CONVENTION: because two bars are connected to each other, the end points belong to the class on the right.

FROM CLASS SAMPLE DATA:



LOWER LIMITS OF FIRST CLASS
↑
∘ starting # + width of histogram will be given

↳ smaller classes = more bars

• HOW MANY CLASSES TO USE?

↳ STOP ONCE LARGEST OBSERVATION IS COVERED

HOW TO → CALCULATOR

HISTOGRAM

1) ENTER DATA TO L_n

2) WINDOW, Xmin = smallest # in data set

Xmax = largest class bound

Xscl = given

Ymin = -1

Ymax = WHEN YOU CAN SEE WHOLE HISTOGRAM

Yscl = 1

3) UNDER PLOT 1, CHOOSE HISTOGRAM, enter L_n

→ GRAPH

4) TRACE

WHEN TO USE:

HISTOGRAM

∘ LARGE, > 30 data points

BOX

∘ smaller data sets

STEM-LEAF

∘ BIGGER, > 20 data points