CHAPTER 3:
MAIN CHARACTERISTICS OF QUANTITATIVE DATA

SUPPOSE WE WANT TO REDUCE (SUMMARIZE) THE INFORMATION OF A LONG LIST OF VALUES OF A NUMERIC-TYPE VARIABLE (VISUALIZE ITS DOTPLOT) TO A HANDFUL OF BASIC CHARACTERISTICS, WHAT SHOULD THESE BE?

IF WE ARE ALLOWED ONLY ONE QUANTITY TO CHARACTERIZE THE DATA, WE WOULD CLEARLY WANT TO KNOW THE LOCATION OF ITS CENTER. TO DO THIS, WE HAVE TWO CHOICES:

THE **MEAN** (AVERAGE) VALUE IS THE USUAL ARITHMETIC AVERAGE OF ALL VALUES (SUM THEM AND DIVIDE BY n) - GRAPHICALLY, THIS REPRESENTS THE 'CENTER OF MASS'.

THE **MEDIAN** IS THE MIDDLE VALUE, PROVIDED WE FIRST ARRANGE THE NUMBERS FROM THE SMALLEST TO THE LARGEST.

EXAMPLE: 2 2 3 4 4 6 7 7 7 (NINE VALUES, AFTER RE-ARRANGING) YIELDS 4 (THE FIFTH VALUE).

WHEN *n* IS EVEN, THERE ARE TWO VALUES 'IN THE MIDDLE', SO WE TAKE THEIR AVERAGE: 2 2 3 4 4 6 7 7 7 8 (TEN VALUES), YIELDS (4+6)/2=5.

THERE IS YET ANOTHER (LESS USEFUL) RELATED CHARACTERISTIC: A **MODE** (REPRESENTING THE MOST 'LIKELY' OR TYPICAL OBSERVATION) IS THE VALUE WHICH OCCURRED MOST OFTEN (7 IN THE PREVIOUS EXAMPLE).

SEVERAL DIFFICULTIES MAY ARISE:

- ALL VALUES MAY BE UNIQUE (E.G. WHEN SALARIES ARE QUOTED TO THE LAST PENNY).
- THERE MAY BE MORE THAN ONE MODE
- HOW TO DEAL WITH TABULATED (**GROUPED**) DATA

LUCKILY, MODE IS NEVER NEEDED BEYOND THIS CHAPTER.

WHAT IS IMPORTANT TO REALIZE IS THAT EACH OF THESE DEFINITIONS (MEAN, MEDIAN AND MODE) CAN BE APPLIED EITHER TO THE POPULATION AS A WHOLE, OR TO A RELATED SAMPLE (WE THEN REFER TO THE POPULATION MEAN μ, AND THE SAMPLE MEAN $\bar{x}$, RESPECTIVELY).

HERE, WE USUALLY DEAL WITH SAMPLES (IMPLYING THAT THE RESULTS ARE, TO SOME EXTENT, RANDOM), WITH THE OBJECTIVE OF **ESTIMATING** THE CORRESPONDING POPULATION QUANTITIES (WHOSE EXACT VALUES, EVEN THOUGH WELL DEFINED AND NON-RANDOM, REMAINS UNKNOWN TO US).

IF WE ARE NOW ALLOWED ONE EXTRA CHARACTERISTIC OF A SET OF VALUES, WE WOULD CLEARLY WANT TO MEASURE THE SPREAD (WIDTH, VARIATION, THE $\pm$) OF THIS DATA.

SIMILARLY TO WHAT WE DID TO DEFINE THE DATA'S 'CENTER', WE HAVE THE FOLLOWING CHOICES:

- **STANDARD DEVIATION**, COMPUTED AS FOLLOWS:

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$$

- **INTERQUARTILE RANGE** Q_3 - Q_1 , WHERE Q_1 IS THE LOWER QUARTILE (25% OF ALL VALUES ARE SMALLER, 75% BIGGER), AND Q_3 IS THE UPPER QUARTILE (REVERSE).

- (FULL) **RANGE**: THE DIFFERENCE BETWEEN THE LARGEST AND SMALLEST OBSERVATION.

NOTE: THE Q_1 AND Q_3 QUARTILES ARE FOUND BY TAKING THE MID POINT VALUE (THE WAY WE CONSTRUCTED THE MEDIAN) OF THE LOWER (Q_1) AND UPPER (Q_3) HALF OF THE DATA, E.G. 3 3 5 6 8 10 10 13 15 17 21 YIELDS $Q_1 = 5$ AND $Q_3 = 15$, OR 0 0 1 1 1 1 2 2 2 3 3 4 5 5 5 7 8 8 8 9 YIELDS $Q_1 = 1$ AND $Q_3 = 6$.

OF THE THREE, THE MOST IMPORTANT IS THE STANDARD DEVIATION, WHICH WE NEED TO DISCUSS IN MORE DETAIL:

OUR DEFINITION IS THAT OF THE SAMPLE STANDARD DEVIATION; THE POPULATION STANDARD DEVIATION (WHOSE VALUE WE ARE TRYING TO ESTIMATE, BASED ON s) IS DENOTED σ AND ITS DEFINITION HAS N (INSTEAD OF n-1) IN THE DENOMINATOR.

THE QUANTITY UNDER THE SQUARE ROOT IS CALLED **VARIANCE** (AGAIN, WE HAVE TWO VERSIONS: SAMPLE VARIANCE AND POPULATION VARIANCE).

THE NUMERATOR OF THE FORMULA (DENOTED SS_x - SUM OF SQUARES OF THE DEVIATIONS FROM THE MEAN) CAN BE MORE CONVENIENTLY COMPUTED BY

$$\Sigma x_i^2 - \frac{\left(\Sigma x_i\right)^2}{n}$$ (NO MESSING UP WITH $\overline{x}$).

THE RATIO $s/\overline{x}$ (USUALLY EXPRESSED IN PERCENT, I.E. MULTIPLIED BY 100) IS CALLED THE **COEFFICIENT OF VARIATION**. IT IS MEANINGFUL ONLY FOR RATIO-SCALE (I.E. NON-NEGATIVE) VARIABLES.

EXAMPLE: 7 2 9 13 4 5 9 4 10 6 8

FIRST, WE COMPUTE THE SUM: 7 + 2 + 9 + ... + 8 = 77 AND THE SIMPLE (ORDINARY) SUM OF SQUARES: $7^2 + 2^2 + 9^2 + ... + 8^2 =$ 641. THEN, RATHER EASILY, $\overline{x}$ = 77 / 11 = 7 AND

$$s = \sqrt{\frac{641 - 77^2/11}{11-1}} = 3.194$$ (VARIANCE = 10.2)

ONE CAN EASILY VERIFY THAT $SS_x = 641 - 77^2/11 = 102$ AGREES WITH $(7-7)^2 + (2-7)^2 + ... + (8-7)^2$

THE COEFFICIENT OF VARIATION IS THUS $3.197 / 7 \times 100 = 45.6\ \%$

TO COMPUTE THE MEDIAN AND INTER-QUARTILE RANGE (IQR), WE MUST FIRST **SORT** (ARRANGE) THE DATA FROM THE SMALLEST TO THE LARGEST:

2 4 4 5 6 7 8 9 9 10 13

THE MEDIAN THE SIXTH (11+1) /2 SMALLEST (LARGEST) VALUE, NAMELY 7.

THE INTER QUARTILE RANGE IS THE DIFFERENCE BETWEEN THE THIRD (4) AND NINTH (9) OBSERVATION, NAMELY 5.

THE (FULL) RANGE EQUALS: 13 ! 2 = 11

CHEBYSHEV'S THEOREM (INEQUALITY) IS MORE THEORETICAL THAN PRACTICAL: FOR ANY k > 1, THE **PROPORTION** OF OBSERVATIONS INSIDE THE $\bar{x} \pm s \cdot k$ INTERVAL IS **AT LEAST** $1 - 1 / k^2$

AND REVERSE: THE PROPORTION OF OBSERVATIONS OUTSIDE THIS INTERVAL (INCLUDING BOUNDARIES) IS **SMALLER** THAN $1/k^2$ IE. LESS THAN 1/4 FOR k=2, LESS THAN 1/9 FOR k=3, LESS THAN 1/16 FOR k=4 ETC. (TYPICALLY, THESE ARE ONLY 5%, ½% AND NONE, RESPECTIVELY).

GROUPED-DATA MODIFICATION:

THE MEAN IS THEN COMPUTED BY $\frac{\Sigma x_i \cdot f_i}{n}$,

AND SS_X BY $\Sigma x_i^2 \cdot f_i - \frac{(\Sigma x_i \cdot f_i)^2}{n}$, WHERE

x_i ARE NOW THE CLASS MARKS.

NOTE THAT THIS WAY OF COMPUTING THE MEAN IS DONE BY THE SO CALLED **WEIGHTED AVERAGING** (CLASS FREQUENCIES BEING THE WEIGHTS).

EXAMPLE: USING THE TABLE OF SECTION 3.3 PROBLEM 5

LIMITS	18-24	25-34	35-44	45-54	55-64	65-80
f_i	78	75	48	33	33	33
x_i	21	29.5	39.5	49.5	59.5	72.5

WE FIRST COMPUTE THE TOTAL FREQUENCY n =78 + 75 + ... + 33 = 300,
THEN THE $x_i f_i$ SUM: 21×78 + 29.5×75 + ... + 72.5×33 = 11736
AND THE CORRESPONDING $\Sigma x_i^2 f_i =$ 21²×78 + 29.5²×75 + ... + 72.5²×33 = 545702
BASED ON THESE THREE, $\bar{x}$ = 11736 /300 = 39.12 AND

$$s=\sqrt{\frac{545702-11736^2\ /\ 300}{300-1}}=17.018$$ Y

$$\text{CV}=\frac{17.018}{39.12}\times 100=43.5\,\%$$

ONE CAN ALSO VERIFY THAT THE NUMERATOR (UNDER THE SQUARE ROOT) REPRESENTS A SHORTCUT COMPUTATION OF

$$\Sigma(x_i-\bar{x})^2 f_i$$

THE MEDIAN IS WHAT WE USED TO CALL THE 50th PERCENTILE (WE ALSO KNOW HOW TO FIND IT GRAPHICALLY).

NOTE THAT THE MEAN, STANDARD DEVIATION, ETC. OF GROUPED DATA SHOULD BE QUITE CLOSE (BUT NOT EXACTLY IDENTICAL) TO THE MEAN, ... OF THE CORRESPONDING RAW DATA.

THE MEDIAN AND QUARTILES (OF THE ORIGINAL, 'RAW' DATA) ARE IMPORTANT FOR CONSTRUCTING THE SO CALLED

BOX AND WHISKER PLOTS:

THIS IS A BOX WITH TWO 'WHISKERS' PLACED NEXT TO A (USUALLY VERTICAL) SCALE. THE BOTTOM (TOP) LINE OF THE BOX IS PLACED AT THE LOWER (UPPER)

QUARTILE, THERE IS ALSO AN EXTRA HORIZONTAL BAR INSIDE THE BOX INDICATING THE MEDIAN (THE BOX' WIDTH IS ARBITRARY). THE WHISKERS EXTEND TO THE LOWEST AND HIGHEST VALUE IN THE DATA.

OVERALL, WE GET A NICE GRAPHICAL SUMMARY OF THE DATA'S IMPORTANT FEATURES.

OBSERVATIONS FURTHER FROM THE 'BOX' THAN 1.5 OF ITS LENGTH (IQR) ARE CONSIDERED (AND MARKED, BY MINITAB) AS **OUTLIERS**.