CHAPTER 1: INTRODUCTION (BASIC DEFINITIONS AND CONCEPTS)

POPULATION:

A SPECIFIC, WELL DEFINED GROUP (USUALLY QUITE LARGE) OF OBJECTS (PEOPLE, COMPANIES, AFRICAN ELEPHANTS, ETC.) WITH SOME PROPERTIES (ATTRIBUTES) CALLED **VARIABLES** (AGE, SALARY - NUMBER OF EMPLOYEES, YEARLY SALES - WEIGHT, DAILY FOOD CONSUMPTION) OF SOME INTEREST TO US.

TO SIMPLIFY THINGS, WE ASSUME THAT POPULATIONS ARE FIXED (UNCHANGING - 'FROZEN IN TIME').

THIS CORRESPONDING INFORMATION CAN BE PRESENTED IN A MATRIX FORM - ONE ROW FOR EACH OBJECT (CASE, ITEM, INDIVIDUAL), ONE COLUMN FOR EACH VARIABLE - COLLECTIVELY KNOWN AS **DATA**, E.G.:

	AGE	GENDER	SALARY
1	42	M	37000
2	35	F	29000
3	21	M	21000
4	18	M	20000
5	49	F	31000
!	!	!	!
147	52	M	43000

TO STUDY A POPULATION, WE HAVE TO FIRST GATHER THE APPROPRIATE DATA (EITHER FROM THE COMPLETE POPULATION - THIS IS CALLED **CENSUS**, OR FROM A REPRESENTATIVE PORTION OF IT - CALLED A **SAMPLE**), AND LEARN TO ORGANIZE IT AND PRESENT IN A MEANINGFUL MANNER (**DESCRIPTIVE STATISTICS** - SUMMARIES, GRAPHS, ETC).

OFTEN, WE ALSO WANT TO ANSWER ONE OR MORE SPECIFIC ISSUES, SUCH AS: DOES SALARY INCREASE WITH AGE, AND BY HOW MUCH.

TYPICALLY, GETTING A COMPLETE INFORMATION ABOUT A POPULATION (A CENSUS) IS QUITE PROHIBITIVE, AND WE HAVE NO CHOICE BUT TO RESORT TO **RANDOM SAMPLING**, I.E. SELECTING, RANDOMLY, ONLY A RELATIVELY SMALL SUBGROUP.

DATA (A SET OF OBSERVATIONS) IS THEN COLLECTED FOR ONE SUCH SAMPLE, AND ANALYZED IN FULL DETAIL TO ANSWER THE ISSUES THAT INTEREST US (E.G. YES, SALARY INCREASES WITH AGE, BY ABOUT $472 PER YEAR). THIS CLEARLY RESULTS IN ONLY **ESTIMATES** INSTEAD OF THE EXACT (POPULATIONS) VALUES (EACH SAMPLE YIELDING A SLIGHTLY DIFFERENT VALUE). WE WOULD STILL LIKE TO MAKE SOME CONCLUSION ABOUT THE WHOLE POPULATION (THIS IS CALLED STATISTICAL **INFERENCE**). HOW RELIABLE IS THIS PROCESS, AND HOW LARGE AN ERROR ARE WE POTENTIALLY MAKING (I.E. COULD THE ACTUAL, POPULATION VALUE BE AS LARGE AS $500 PER YEAR)?

THERE ARE SEVERAL SAMPLING TECHNIQUES ONE CAN EMPLOY:

- **SIMPLE** SAMPLING: ALL INDIVIDUALS OF A POPULATION MUST HAVE AN EQUAL CHANCE TO BE SELECTED, WITH NO DUPLICATION (WITHOUT REPLACEMENT). THIS SAMPLING IS ASSUMED THROUGHOUT THE COURSE.

- **SYSTEMATIC** SAMPLING: EACH 100^{TH} INDIVIDUAL IS SELECTED. SIMILAR TO SIMPLE SAMPLING, BUT MORE DANGER OF **BIAS**ING.

- **STRATIFIED** SAMPLING: THE POPULATION OF ALL COLLEGE STUDENTS IS DIVIDED INTO 4 SUB-POPULATIONS (STRATA) BY YEAR (FRESHMEN, ...) - THESE ARE THEN SAMPLED SEPARATELY.

< **CLUSTER** SAMPLING: THE POPULATION OF A PROVINCE IS DIVIDED INTO MUNICIPALITIES (CLUSTERS) WHICH DIFFER (UNLIKE STRATA) ONLY BY SIZE. WE THEN RANDOMLY SELECT A FEW OF THESE TO BE INCLUDED, WITH EACH OF ITS MEMBERS, IN OUR SAMPLE.

< **CONVENIENCE** SAMPLING: INTERVIEWING A GROUP OF PEOPLE WE MEET WALKING ON A STREET.

STRATIFIED AND CLUSTER SAMPLING ARE MORE DIFFICULT TO ANALYZE STATISTICALLY (WE WILL NOT DO IT HERE), CONVENIENCE SAMPLING IS VIRTUALLY IMPOSSIBLE TO ANALYZE PROPERLY.

WE THEN COLLECT OUR DATA BY:

- SIMPLY ASKING A FEW QUESTIONS - THIS IS CALLED A **SURVEY**

- **OBSERVATION** (OR MEASUREMENT -

TAKING THE ANIMAL'S WEIGHT) - NOT CHANGING THE OBJECT OF STUDY IN ANY WAY

- **EXPERIMENT** (E.G. TO ESTABLISH BY HOW MUCH A SPECIFIC FERTILIZER INCREASES YIELD PER ACRE, WE TRY IT ON A FEW RANDOMLY SELECTED FIELDS; WE ADMINISTER A SPECIFIC DRUG TO A GROUP OF PATIENTS) - THUS CHANGING THEIR CONDITION.

FINALLY, OUR DATA CAN ALSO BE GENERATED BY A COMPUTER **SIMULATION** (E.G. RATHER THAN ROLLING A DIE 6000 TIMES, WE ASK THE COMPUTER TO DO THIS FOR US, USING RANDOM NUMBERS WHICH FOLLOW THE SAME PROBABILITY RULES AS AN ACTUAL EXPERIMENT).

VARIABLES CAN BE OF SEVERAL TYPES, DEPENDING WHETHER THEIR VALUES ARE MEASURED ON THE SCALE (LEVEL) WHICH IS:

< **NOMINAL** - THE 'VALUES' CANNOT BE COMPARED IN SIZE (I.E. THERE IS NO 'BIGGER' OR 'SMALLER'), E.G. COLOR OR NATIONALITY, REGARDLESS HOW IT IS CODED.

< **ORDINAL** - THE VALUES ARE 'ORDERABLE' (I.E. BETTER OR WORSE),

E.G. FINAL MARK (A, B, C, D, F).

< **INTERVAL** - THE VALUES WOULD NORMALLY BE NUMERIC, WITH A MEANINGFUL DIFFERENCE BETWEEN TWO NUMBER (E.G. TEMPERATURE IN DEGREES), BUT AN ARBITRARY (I.E. MEANINGLESS) ORIGIN (ZERO) - THAT'S WHY WE CAN HAVE NEGATIVE TEMPERATURES.

< **RATIO** - THE VALUES ARE POSITIVE NUMBERS, HAVING MEANINGFUL RATIOS (NOT JUST DIFFERENCES), E.G. WEIGHT (10 LB. IS TWICE AS HEAVY AS 5 LB.).

THE MAIN AND MOST IMPORTANT DISTINCTION IS BETWEEN THE FIRST TYPE (ORDINAL, ALSO CALLED **QUALITATIVE** OR CATEGORICAL), AND THE REMAINING THREE (**QUANTITATIVE**), WHERE THE BOUNDARIES ARE A BIT FUZZIER.

NOTE THAT A NOMINAL-SCALE VARIABLE IS SOMETIMES **CODED** USING NUMBERS -

WE MUST NOT BE MISLED BY THIS.

FOR QUANTITATIVE-TYPE VARIABLES, ANOTHER IMPORTANT DISTINCTION IS WHETHER THE SCALE IS **DISCRETE** (THIS USUALLY MEANS THAT ONLY INTEGER VALUES CAN BE OBSERVED, E.G. NUMBER OF CHILDREN IN A FAMILY), OR **CONTINUOUS** (ANY REAL VALUE IS PERMISSIBLE - WE USUALLY NEED SEVERAL DIGITS TO RECORD IT ACCURATELY), E.G. WEIGHT (IN PRACTICE, WE MAY STILL **ROUND** IT **OFF** TO AN INTEGER).

DURING MOST OF THIS COURSE, OUR DATA WILL CONSIST OF ONE VARIABLE ONLY (AND EVEN WHEN THERE ARE MORE VARIABLES, WE STUDY THEM ONE AT A TIME). IN THAT CASE, RATHER THAN DISPLAYING IT A SINGLE COLUMN OF DATA, ONE CAN USE A SIMPLE LIST, E.G. A SAMPLE OF NUMBER OF CHILDREN PER FAMILY:

2, 0, 1, 1, 3, 4, 1, 2, 2, 0, ..., 2