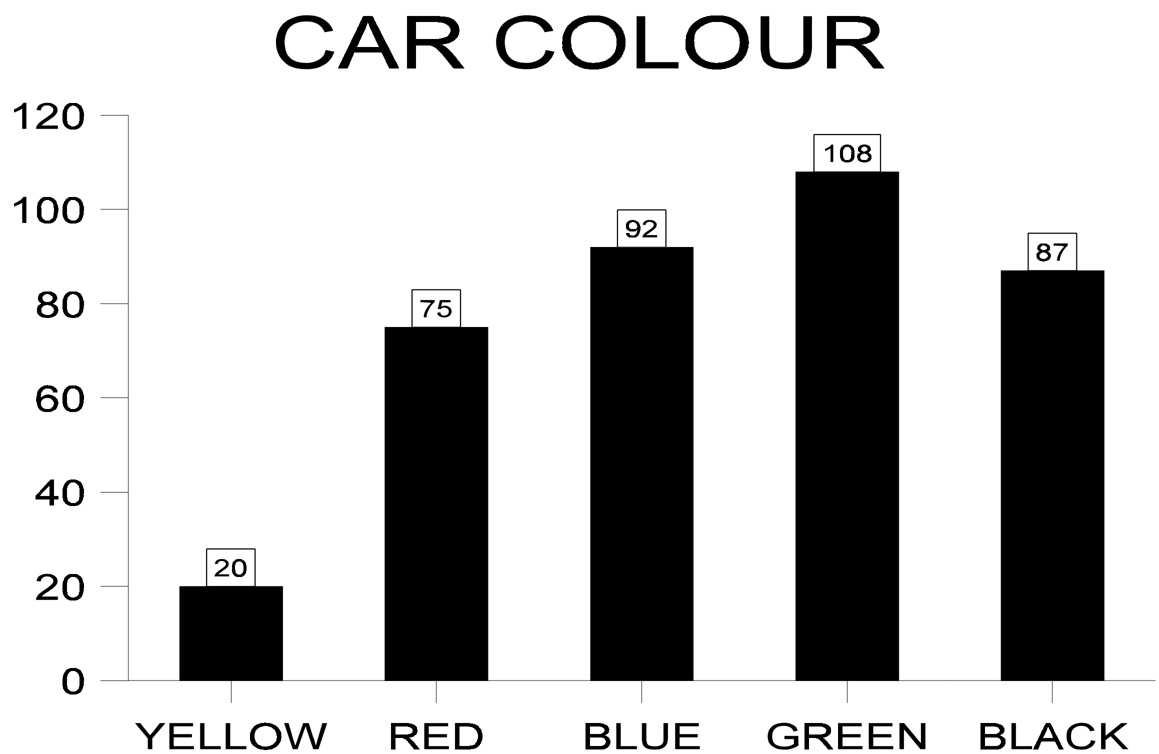
CHAPTER 2: ORGANIZING DATA (TABLES AND GRAPHS)

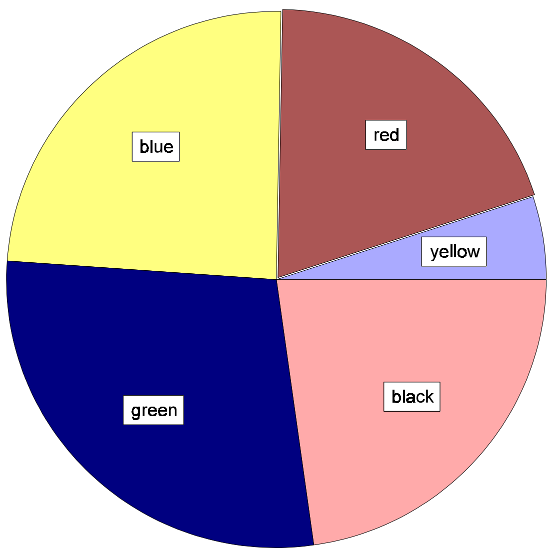
BAR GRAPHS AND **CIRCLE** (OR **PIE**) **GRAPHS** ARE USUALLY USED WITH NOMINAL (QUALITATIVE) TYPE VARIABLES HAVING SEVERAL POSSIBLE VALUES



IF THE BARS ARE RE-ARRANGED FROM THE LARGEST TO THE SMALLEST, WE GET **PARETO CHART**

PIE CHART

$$\frac{20}{20+75+92+108+87}\times 360 = 18.8^{o}$$



TIME PLOTS ARE USEFUL WITH SEQUENTIAL (TIME-RELATED) DATA, E.G. CAR SALES DURING THE PAST 7 YEARS.

A BIT MORE INVOLVED IS THE ISSUE PLOTS RELATED TO **GROUPED DATA**.

LET'S ASSUME THAT WE HAVE ONLY ONE QUANTITATIVE (NUMERICAL) VARIABLE OF THE CONTINUOUS TYPE, SUCH AS YEARLY INCOME, IN $1000), E.G. 17, 28, 42, ..., 31, SUMMARIZED IN A **FREQUENCY TABLE**, E.G.

SALARY (IN $1000)	FREQUENCY
10-19	7
20-29	12
....	....
110-119	4

EACH ROW CORRESPONDS TO A **CLASS**.
THE NUMBERS IN THE LEFT COLUMN ARE CALLED **CLASS LIMITS** (LOWER AND UPPER - NOTE THAT THEY LEAVE GAPS), THE **MIDPOINT** VALUES (14.5, 24.5, ...) ARE CLASS **MARKS**.

CLASS LIMITS CAN BE EXTENDED TO **CLASS BOUNDARIES**, THUS: (9.5-19.5, 19.5-29.5, ...) WHICH 'MEET' (NO GAPS) AND GIVE THE RANGE OF <u>ACTUAL</u> SALARIES CONTRIBUTING TO EACH CLASS (THESE ARE SOMETIMES INSERTED AS AN EXTRA COLUMN).

RELATIVE FREQUENCIES ARE COMPUTED BY DIVIDING EACH OF THE REGULAR FREQUENCIES f BY THEIR TOTAL (THE SAMPLE **SIZE** n).

SIMILARLY, WE CAN ALSO COMPUTE **CUMULATIVE** (RUNNING SUM) **FREQUENCIES**, AND CUMULATIVE **RELATIVE** FREQUENCIES.

CLASS LIMITS SHOULD BE CHOSEN SENSIBLY (WHEN IT IS UP TO US) - TO END UP WITH 5-15 CLASSES - BASED ON THE SMALLEST AND LARGEST OBSERVATION.

GRAPHICALLY, THE INFORMATION OF A FREQUENCY TABLE CAN BE DISPLAYED IN A **HISTOGRAM** (A KIND OF BAR GRAPH), THE BARS CENTERED ON THE CLASS MARKS, EXTENDING FROM ONE BOUNDARY TO THE NEXT (NO GAPS) - SEE FIG. 2-8.

HISTOGRAMS CAN BE OF DIFFERENT SHAPES, YOUR TEXTBOOK MENTIONS FIVE POSSIBILITIES: (APPROXIMATELY) SYMMETRICAL, LEFT OR RIGHT SKEWED, (APPROXIMATELY) UNIFORM, BIMODAL.

ALTERNATELY (TO HISTOGRAM), CONNECTING THE CLASS MARK (x) - FREQUENCY (y) POINTS BY STRAIGHT-LINE SEGMENTS TO GET THE **FREQUENCY POLYGON** (WE USUALLY ADD AN EXTRA ZERO-FREQUENCY CLASS ON EITHER SIDE)- FIG. 2-17.

PLOTTING THE CUMULATIVE (RELATIVE) FREQUENCIES IN THIS MANNER (HERE, x MUST BE THE CORRESPONDING UPPER CLASS BOUNDARY), WE GET A SO CALLED **OGIVE** - FIG. 2-18, 2-20.

BASED ON THE LATTER, WE CAN EASILY ESTIMATE THE PERCENTAGE OF PEOPLE WITH A SALARY LESS THAN, SAY $27,000 PER YEAR.

AND REVERSE, I.E. ESTIMATE THE SALARY WHICH 25% OF PEOPLE EXCEED.

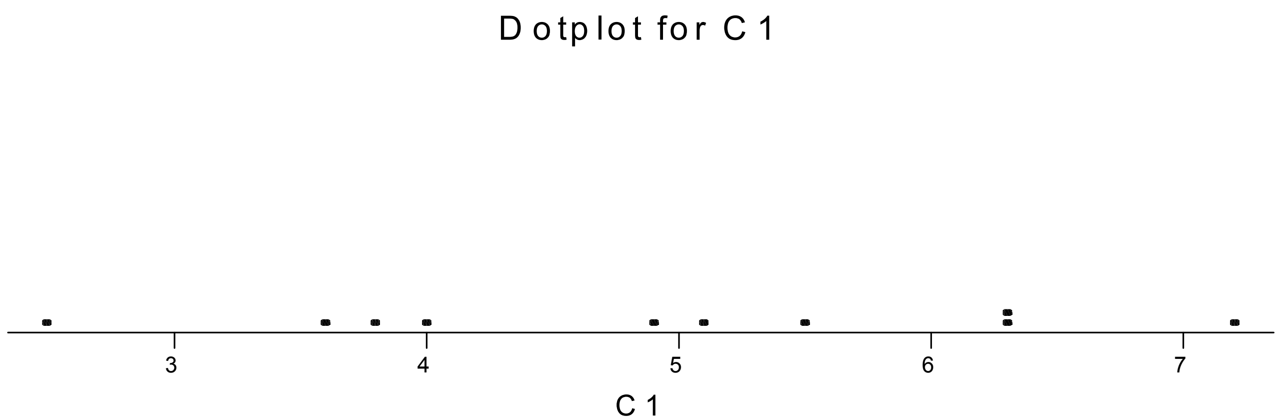
STEM-AND-LEAF DISPLAY IS SIMILAR TO (FASTER AND LESS 'FORMAL' THEN) A HISTOGRAM.

WE START WITH THE ORIGINAL DATA LIST, DESIGNATE THE LAST DIGIT (OR LAST TWO DIGITS) AS 'LEAF' (THE REST ARE 'STEM') AND TALLY THE DATA IN A MANNER OF FIG. 2-24.

EXAMPLE: 3.9 5.4 5.5 4.2 3.6 3.0 4.5 5.9 4.4 2.9 4.1 5.4 5.1 2.0 4.6 4.2 3.5 4.7 5.9 5.5 2.8 6.0 2.6 2.4 3.2 4.9 4.9 5.6 4.9 2.5

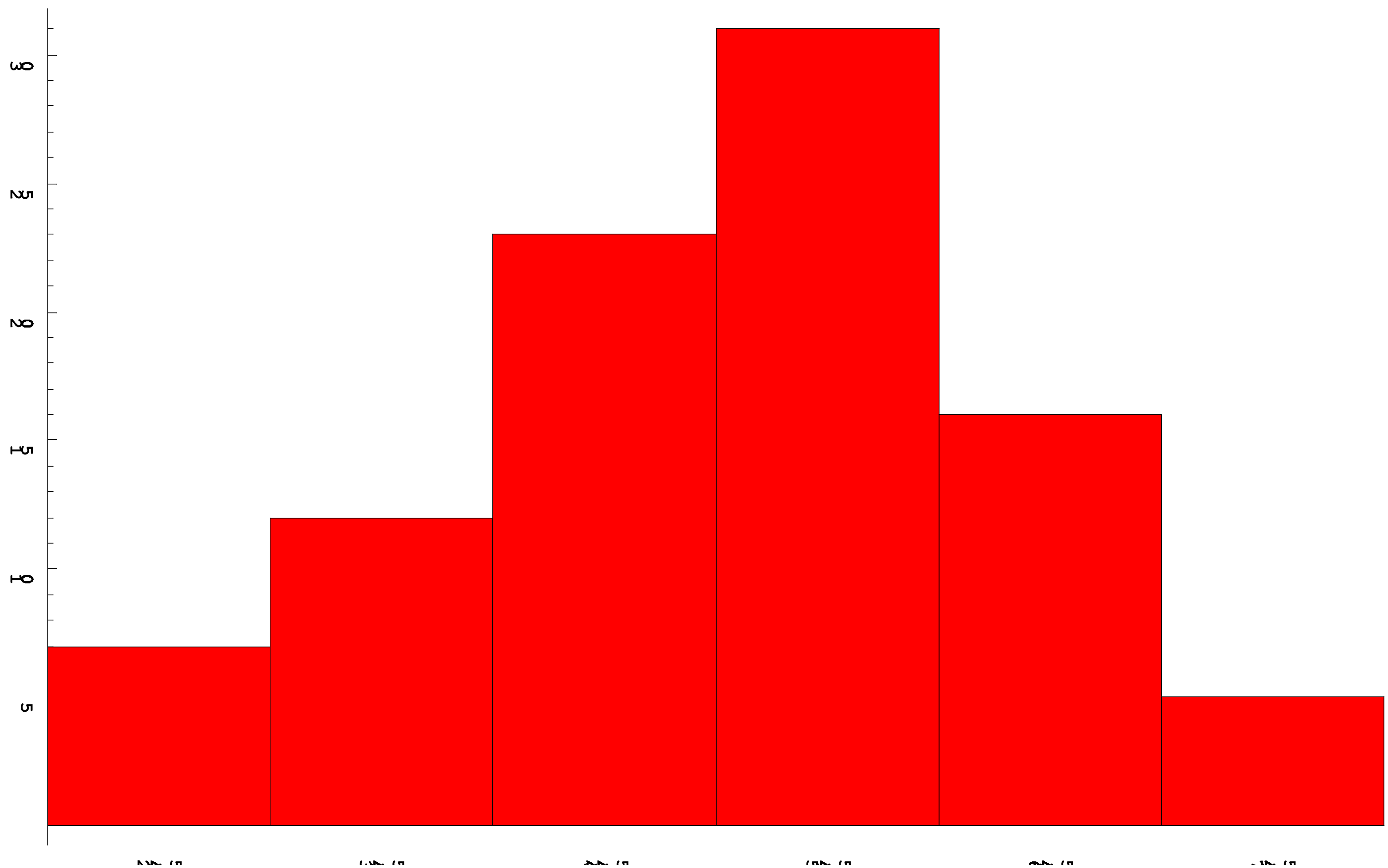
2. 9 0 8 6 4 5
3. 9 6 0 5 2
4. 2 5 4 1 6 2 7 9 9 9
5. 4 5 9 4 1 9 5 6
6. 0

DOT PLOT IS A SIMPLE GRAPHICAL REPRESENTATION OF THE INDIVIDUAL OBSERVATIONS: 3.8, 2.5, 6.3, 7.2, 4.9, 6.3, 5.5, 5.1, 3.6, 4.0

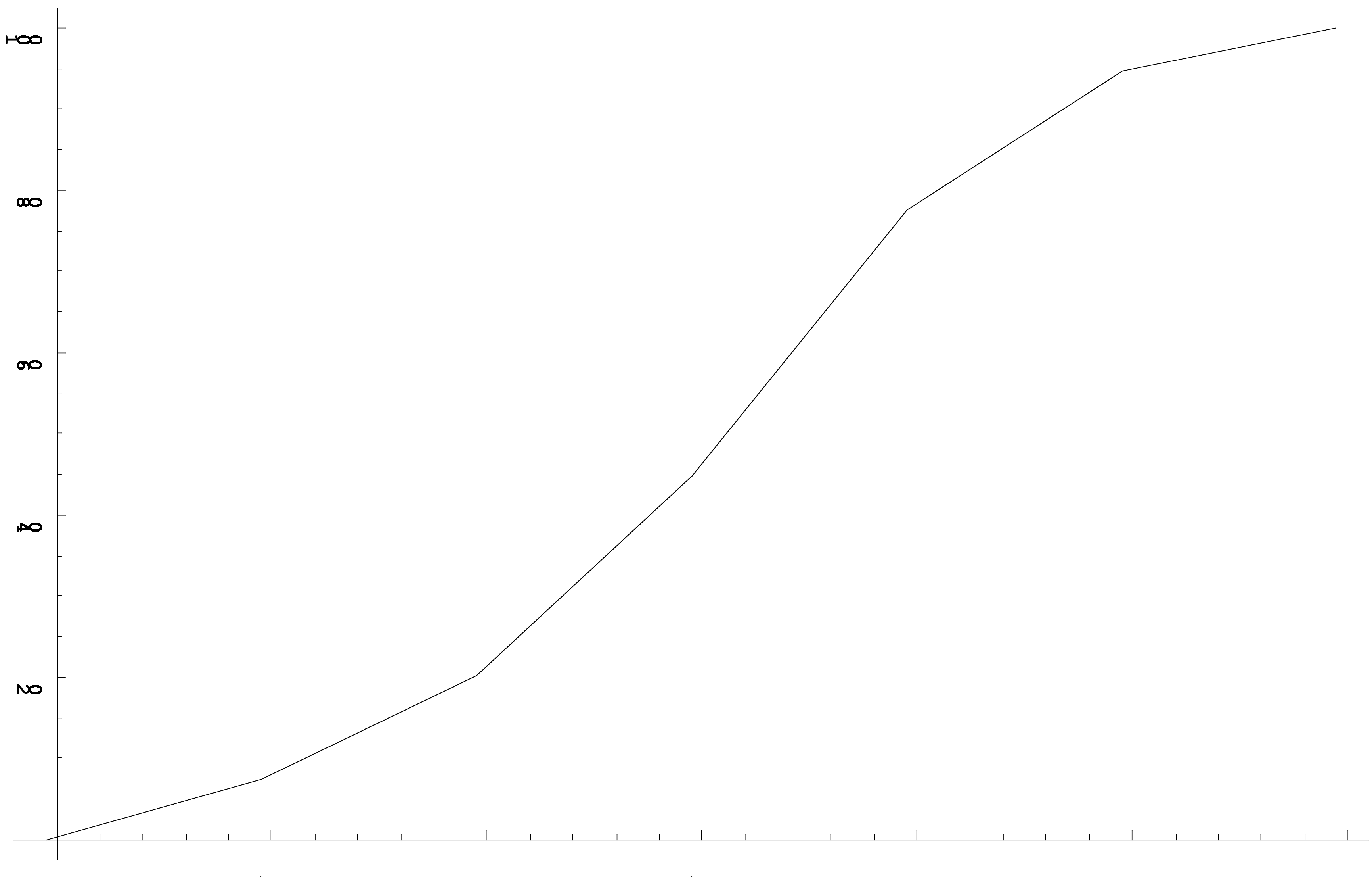


Limits	Frequency	Boundaries	Midpoint	Relative f	Cumulat. f	Rel. cum. f
20 - 29	7	19.5 -29.5	24.5	7/94=7.4%	7	7.4%
30 - 39	12	29.5 - 39.5	34.5	12.8%	19	20.2%
40 - 49	23	39.5 - 49.5	44.5	24.5%	42	44.7%
50 - 59	31	49.5 - 59.5	54.5	33.0%	73	77.7%
60 - 69	16	59.5 - 69.5	64.5	17.0%	89	94.7%
70 - 79	5	69.5 - 79.5	74.5	5.3%	94	100.0%

SALARIES OF 94 EMPLOYEES (OF A SPECIFIC COMPANY), ROUNDED OFF TO THE NEAREST THOUSAND, HAVE BEEN TALLIED TO YIELD THE ABOVE FREQUENCY TABLE



(FREQUENCY) HISTOGRAM



(RELATIVE CUMULATIVE FREQUENCY) OGIVE

FIND THE PERCENTAGE OF EMPLOYEES WHOSE SALARY IS LESS THAN 55000:

$$44.7 + 33 \times \frac{55 - 49.5}{59.5 - 49.5} = 66.15\%$$

80% OF ALL EMPLOYEES HAVE A SALARY SMALLER THAN .... ? (THIS IS CALLED THE 80^{th} PERCENTILE):

$$59.5 + 10 \times \frac{80 - 77.7}{17} = \$60850$$