MAT 271E: PROBABILITY AND STATISTICS

PROF. DR. CANAN SARICAM

WEEK 2

DATA SETS, DATA ACQUISITION.
DATA PATTERNS: ARRAYS, STEM AND LEAF DISPLAYS, DOT PLOTS, CUMULATIVE DISTRIBUTIONS
FREQUENCY DISTRIBUTION TABLES

DATA PATTERNS

Simple displays of data

3. Dot diagram

- The dot diagram is a very useful plot for displaying **a smaIl body of data**, say, up to about 20 observations. This plot allows us quickly and easily see the **location or central tendency in the data and the spread or variability.**

- A dot plot is constructed by plotting a dot for each observation above its value on a line scale. **Two data sets can be readily compared by means** of dot plots that are aligned on a common scale.

- When the number of observations is moderately large, other graphical displays may be more useful.

DATA PATTERNS

Simple displays of data

3. Dot diagram

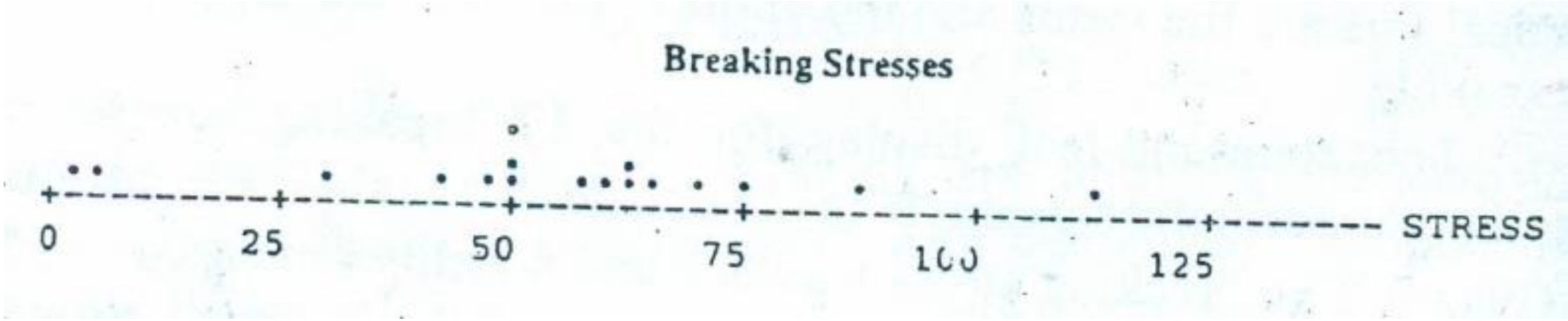
1, 43, 51, 59, 88, 113, 47, 62, 71, 31, 63, 51, 4, 66, 58, 50 ,75

DATA PATTERNS

Simple displays of data

3. Dot diagram

1, 43, 51, 59, 88, 113, 47, 62, 71, 31, 63, 51, 4, 66, 58, 50 ,75



DATA PATTERNS

Simple displays of data

4. Cumulative distributions

- A cumulative distribution is the **rank-ordered array of observations on a quantitative variable** in a data set.

- The cumulative distribution is constructed by assigning ranks to the arrayed observations. The cumulative distribution of the observations for a quantitative variable **describes the number of observations above or below given value.**

DATA PATTERNS

Simple displays of data

4. Cumulative distributions

EXAMPLE:

The array of breaking stresses in the semiconductor failure example and their assigned ranks are as follows:

1, 4, 31, 43, 47, 50, 51, 51, 58, 59, 62, 63, 66, 71, 75, 88, 113

DATA PATTERNS

Simple displays of data

4. Cumulative distributions

EXAMPLE:

The array of breaking stresses in the semiconductor failure example and their assigned ranks are as follows:

Observation:	1	4	31	43	47	50	51	51	58	59	62	63	66	71	75	88	113
Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

DATA PATTERNS

Simple displays of data

4. Cumulative distributions

- The smallest observation (1) is assigned rank 1, the second smallest observation (4) is assigned rank 2, and so on for the whole data array.
- If there are several observations with the same value, such as 51, they receive individual consecutive ranks. This array, with the corresponding ranks, is called the cumulative distribution.
- We can generally see from the rank of an observation in the cumulative distribution how many observations in data set are smaller or larger. Thus, since value 58 is 9th in the array, there are eight breaking stresses in the data set that are smaller than 58 milligrams, and there are eight that are larger.
- When several observations **have the same value, as for the value 51, the rank of each such observation cannot be interpreted** in this direct fashion.

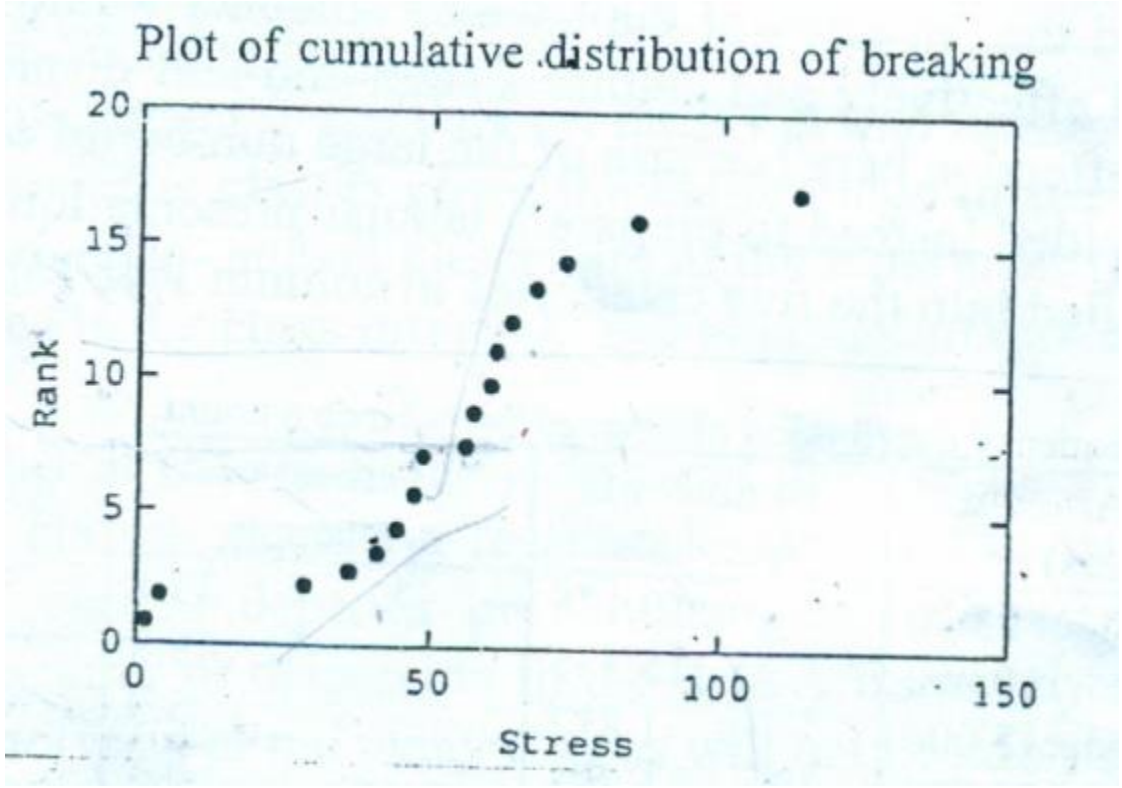
DATA PATTERNS

Simple displays of data

Observation:	1	4	31	43	47	50	51	51	58	59	62	63	66	71	75	88	113
Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

4. Cumulative distributions

- A plot of the cumulative distribution provides a useful graphic display. As shown below, the smallest observation 1 and rank 1, and the plotted at point (1, 1) on the graph.
- The second smallest observation 4 and rank 2, and it is plotted at point (4,2).
- The other observations are similarly plotted.
- Most of the observations are concentrated in the range from 40 to 80 milligrams.



DATA PATTERNS

Simple displays of data

4. Cumulative distributions

The type of cumulative distribution we have considered, in. which the data are arrayed in ascending order and rank 1 is given to the smallest observation, is called a **less-than cumulative distribution.**

Occasionally, the data are arrayed in descending order and rank 1 is assigned to the largest observation. Such distributions are called **more-than cumulative distributions.**

DATA PATTERNS

Simple displays of data

4. Cumulative distributions

Sometimes interest lies in the percentage of observations in the data array that are above or below given values. For that purpose, we express the rank as a percent of the total number of observations. in our example, ranks, 1,2, ... ,17 would be replaced by 100 (1/17) = 5.9, 100 (2/17) = 11.8, ...,100 (17117) = 100, respectively.

We shall refer to the cumulative in this case as a cumulative percent distribution.

FREQUENCY DISTRIBUTIONS

- When the number of observations on a quantitative variable is large, classifying the observations into a tabular distribution is helpful for studying the data pattern.
- A frequency distribution is a more compact summary of data than a stem-and-leaf diagram.
- A **frequency distribution** is the classification of the elements of a data set by a quantitative variable.
- It has basically two characteristics, classes and frequencies,which are shown with the coloumns in frequency distribution tables.To construct a frequency distribution, we must divide the range of the data into intervals, which are usually called classes cells, or bins. We have to calculate the frequency of each class.

FREQUENCY DISTRIBUTIONS

EXAMPLE:

A bank has **30,794** savings accounts. The balances of these accounts are to be studied to assist management in revising the service charge schedule for savings account transactions.

Table 2.1. Frequency distribution of savings by balance amount

Balance Amount (dollars)	Number of Accounts	Percent of Accounts
0 - under 5,000	10,196	33.1
5,000 - under10,000	15,335	49.8
10,000 - under15,000	1,812	5.9
15,000 - under20,000	1,798	5.8
20,000 - under25,000	1,653	5.4
Total	30,794	100.0
		(30,794)

FREQUENCY DISTRIBUTIONS

- The frequency distribution in Table 2.1 has five categories or classes.

- The class frequency is the number. of data points that fall into a particular class. Thus, the frequency of the **class 0-under 5,000 is 10,196**.

Table 2.1. Frequency distribution of savings by balance amount

Balance Amount (dollars)	Number of Accounts	Percent of Accounts
0 - under 5,000	10,196	33.1
5,000 - under10,000	15,335	49.8
10,000 - under15,000	1,812	5.9
15,000 - under20,000	1,798	5.8
20,000 - under25,000	1,653	5.4
Total	30,794	100.0
		(30,794)

FREQUENCY DISTRIBUTIONS

- When the frequency of each class is expressed as a percentage of the total number of elements in a data set, it is called a **percent frequency or relative frequency** , and the resulting distribution is called a percent frequency distribution or relative frequency distribution.
- A percent frequency distribution should show the total number of elements in the data set, as in Table 2.1 in which the total number of accounts **(30,794)** is shown in the parentheses under the **100.0** percent total. Otherwise, the user does not know whether the percent frequencies are derived from a small or a large number of observations.

Table 2.1. Frequency distribution of savings by balance amount

Balance Amount (dollars)	Number of Accounts	Percent of Accounts
0 - under 5,000	10,196	33.1
5,000 - under10,000	15,335	49.8
10,000 - under15,000	1,812	5.9
15,000 - under20,000	1,798	5.8
20,000 - under25,000	1,653	5.4
Total	30,794	100.0
		(30,794)

CONSTRUCTION OF FREQUENCY DISTRIBUTIONS

Rules

- Every element in the data set must fall into one and only one class of the system.
- The classes cover the entire range of the data values and do not overlap

Steps

- Find the number of class
- Find the class width
- Identify class limits
- Count the frequency of each class

CONSTRUCTION OF FREQUENCY DISTRIBUTIONS

- **Number of Classes.** Some judgement must be used in selecting the number of classes so that a reasonable display can be developed.
  - The number of classes depends on the number of observations and the amount of scatter or dispersion in the data. A frequency distribution that uses either too few or too many classes will not be informative.

CONSTRUCTION OF FREQUENCY DISTRIBUTIONS

- ✓ **The larger the number of classes** in a frequency distribution, the more detail is shown. If the number of classes is too large, though, the table loses its effectiveness for summarizing the data.

- ✓ **Too few classes**, on the other hand, condense the information so much as to leave little insight into the pattern of the distribution.

CONSTRUCTION OF FREQUENCY DISTRIBUTIONS

- The best number of classes in a frequency distribution often needs to be determined by experimentation.
- We usually find that between 5 and 20 classes is satisfactory in most cases and that the number of classes should increase with the number of observations.
- Choosing the number of classes approximately equal to the square root of the number of observations often, works well in practice.
  - Number of classes = $\sqrt{Number\ of\ observations}$

CONSTRUCTION OF FREQUENCY DISTRIBUTIONS

Width of Classes. – When all classes are of the same size, the common size of all classes is the **width of class** or **class interval.** The choice of the class width or class interval is related to the determination of the number of classes. it is generally best if all the classes have the same width.

- Sometimes, one must use unequal class intervals. If the classes are not equally wide, it is often difficult to tell whether differences in class frequencies result mainly from differences in the concentration of observations or from differences in the class widths.

- Open-end classes may be necessary at the upper end or at the lower end of the distribution, and occasionally at both ends. An open-end class is one that has only one limit, either upper or lower.

CONSTRUCTION OF FREQUENCY DISTRIBUTIONS

Class limits. Still another issue that must be considered in constructing a frequency distribution is the choice of class limits.

- Calculations from a frequency distribution often use the midpoint of each class to represent all the observations in the class. The **midpoint** of a class is the value halfway between the two class limits. The midpoint of any class of any class can also be expressed as the **class mark.**

- It is usually assumed that the midpoint of each class is approximately equal to the arithmetic average of the observations falling in that class. Often, this objective can be accomplished without special concern about the class limits.

CONSTRUCTION OF FREQUENCY DISTRIBUTIONS

In expressing class limits, one should be careful to be unambiguous.

- For instance, the limits 300 - 400 and 400 - 500 are not clear because one cannot be sure in which class 400 is included. Stating limits as 300-399, 400-499 is clear when the data are expressed in units.
- The limits 300 - under 400, 400 - under 500 are clear.
- Besides, the classses can be expressed with class midpoints. The midpoint of the first class is (300 + 399) / 2 = 349.50, or for most practical purposes, 350 unit. (for discrete data set)
- However, without additional information, it may not be possible to determine the midpoints accurately. If no additional information is provided, we shall assume that the midpoint of a class to be the arithmetic average of the two limits.
- Thus, the midpoint of the class 300 - under 400 would be taken to be (300 + 400)/ 2 = 350. (for continuous data set)

EXAMPLE 1 – Frequency distribution

The compressive strengths in -pounds per square inch (psi) of a new aluminum-lithium alloy material for aircraft structural elements.

Compressive Strength of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

EXAMPLE 1 – Frequency distribution

- The data were recorded in the order of testing, and in this format they do not convey much information about compressive strength.
- Question such as "What percent of the specimens fail below 120 psi?“ are not easy to answer.
- Because there are many observations, constructing a dot diagram of these data would be relatively inefficient; more effective displays are available for large data sets.
- Let’s construct a frequency distribution.

EXAMPLE – Frequency distribution

- Since the data set contains 80 observations, and since $\sqrt{80} \simeq 9$ we suspect that about eight to nine classes will provide a satisfactory frequency distribution.
- The largest and the smallest data values are 245 and 76 respectively, so the classes must cover a range of at least 245 -76 = 169 units on the psi scale.
- If we want to the lower limit for the first class to begin slightly below the smallest data value and the upper limit for the last class to be slightly above the largest data value, then we might start the frequency distribution at 70 and end it at 250. This is an interval or range of 180 psi units.
- Nine classes, each of width 20 psi (180/9=20), gives a reasonable frequency distribution.

EXAMPLE 1 – Frequency distribution

Frequency Distribution for the Compressive Strength Data in Table

Class Interval (psi)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
$70 \leq x < 90$	||	2	0.0250	0.0250
$90 \leq x < 110$	|||	3	0.0375	0.0625
$110 \leq x < 130$	𝍸 |	6	0.0750	0.1375
$130 \leq x < 150$	𝍸 𝍸 ||||	14	0.1750	0.3125
$150 \leq x < 170$	𝍸 𝍸 𝍸 𝍸 ||	22	0.2750	0.5875
$170 \leq x < 190$	𝍸 𝍸 𝍸 ||	17	0.2125	0.8000
$190 \leq x < 210$	𝍸 𝍸	10	0.1250	0.9250
$210 \leq x < 230$	||||	4	0.0500	0.9750
$230 \leq x < 250$	||	2	0.0250	1.0000

EXAMPLE 1 – Frequency distribution

The forth column contains a relative frequency distribution.

The relative frequencies are found by dividing the observed frequency in each class by the total number of observations.

Frequency Distribution for the Compressive Strength Data in Table

Class Interval (psi)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
$70 \le x < 90$	||	2	0.0250	0.0250
$90 \le x < 110$	|||	3	0.0375	0.0625
$110 \le x < 130$	𝍸 |	6	0.0750	0.1375
$130 \le x < 150$	𝍸 𝍸 |||	14	0.1750	0.3125
$150 \le x < 170$	𝍸 𝍸 𝍸 𝍸 ||	22	0.2750	0.5875
$170 \le x < 190$	𝍸 𝍸 𝍸 ||	17	0.2125	0.8000
$190 \le x < 210$	𝍸 𝍸	10	0.1250	0.9250
$210 \le x < 230$	|||	4	0.0500	0.9750
$230 \le x < 250$	||	2	0.0250	1.0000

= 2 / 80

2 is frequency of the class

80 is the total number of observations.

0.0250 + 0.0375

0.0250 + 0.0375 + 0.0750

CUMULATIVE FREQUENCY DISTRIBUTION

- A cumulative form of a frequency distribution is called a **cumulative frequency distribution**.

- It is constructed by summing the class frequencies from the lowest class up to and including the class of current interest. Such a distribution provides an effective summary of the number or percentage of observations that are above or below given values.

- The last column of the **EXAMPLE 1** expresses the relative frequencies on a cumulative basis. For example, it is very easy to see that most of the specimens have compressive strengths between 130 and 190 psi and that 97.5 percent of the specimens fail below 230 psi.

EXAMPLE 1 – Frequency distribution

The forth column contains a relative frequency distribution.

The relative frequencies are found by dividing the observed frequency in each class by the total number of observations.

Frequency Distribution for the Compressive Strength Data in Table

Class Interval (psi)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
$70 \leq x < 90$	||	2	0.0250	0.0250
$90 \leq x < 110$	|||	3	0.0375	0.0625
$110 \leq x < 130$	𝍸 |	6	0.0750	0.1375
$130 \leq x < 150$	𝍸 𝍸 |||	14	0.1750	0.3125
$150 \leq x < 170$	𝍸 𝍸 𝍸 𝍸 ||	22	0.2750	0.5875
$170 \leq x < 190$	𝍸 𝍸 𝍸 ||	17	0.2125	0.8000
$190 \leq x < 210$	𝍸 𝍸	10	0.1250	0.9250
$210 \leq x < 230$	|||	4	0.0500	0.9750
$230 \leq x < 250$	||	2	0.0250	1.0000

= 2 / 80

2 is frequency of the class

80 is the total number of observations.

0.0250 + 0.0375

0.0250 + 0.0375 + 0.0750

EXAMPLE 2 – Frequency distribution

The following data represent the record high temperatures for each of the 50 states.

Construct a frequency distribution for the data.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

EXAMPLE 2 – Frequency distribution

Since the data set contains 50 observations, and since $\sqrt{50} \simeq 7$ we suspect that about seven to eigth classes will provide a satisfactory frequency distribution.

The highest record is 134 and the lowest record is 100, the difference is 134-100=34, and the class width can be defined as 5 ($34 / 7 \simeq 5$).

Class interval	Frequency	Relative frequency	Cumulative relative frequency
100 – under 105	2	0.04	0.04
105 – under 110	8	0.16	0.20
110 – under 115	18	0.36	0.56
115 – under 120	13	0.26	0.82
120 – under 125	7	0.14	0.96
125 – under 130	1	0.02	0.98
130 – under 135	1	0.02	1

GRAPHIC PRESENTATION OF FREQUENCY DISTRIBUTIONS

- Graphic displays of frequency distributions facilitate presentation and analysis.
- Two common methods of graphic display are the **histogram** and the **frequency polygon**.
- A histogram is a rectangular graph of a frequency distribution. A frequency polygon is a line graph of a frequency distribution.

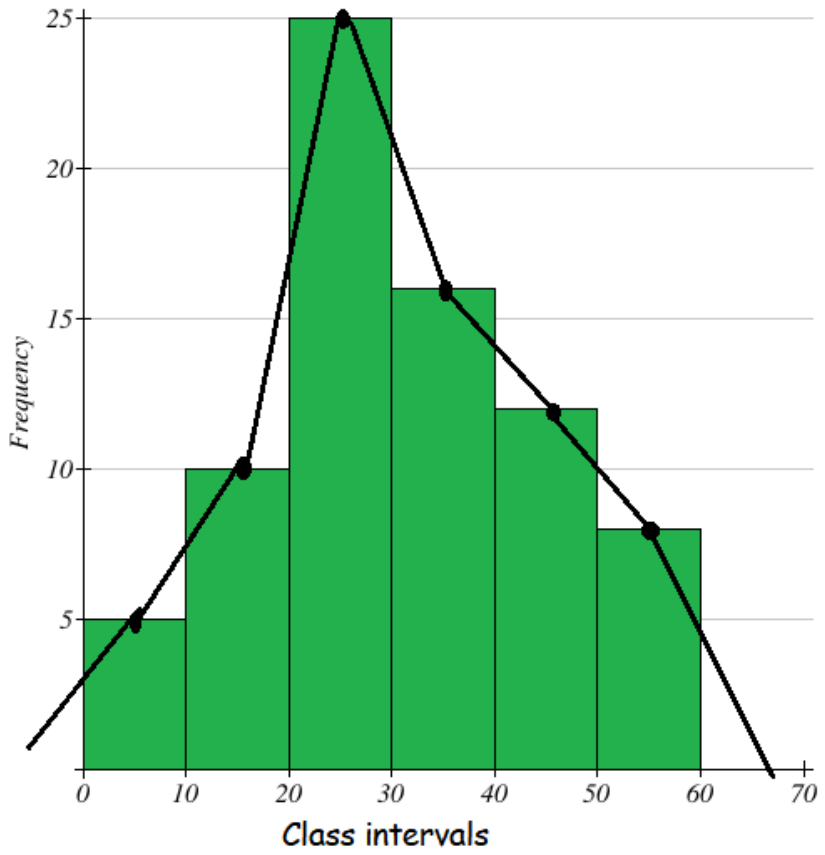
- To draw a histogram, use the horizontal axis to represent the measurement scale, and draw the limits of the classes. The vertical axis represents the frequency (or relative frequency) scale.

GRAPHIC PRESENTATION OF FREQUENCY DISTRIBUTIONS

- In the **histogram**, adjoining rectangulars are drawn with a width spanning the class interval and a height corresponding to the frequency (or percent frequency) of that class.

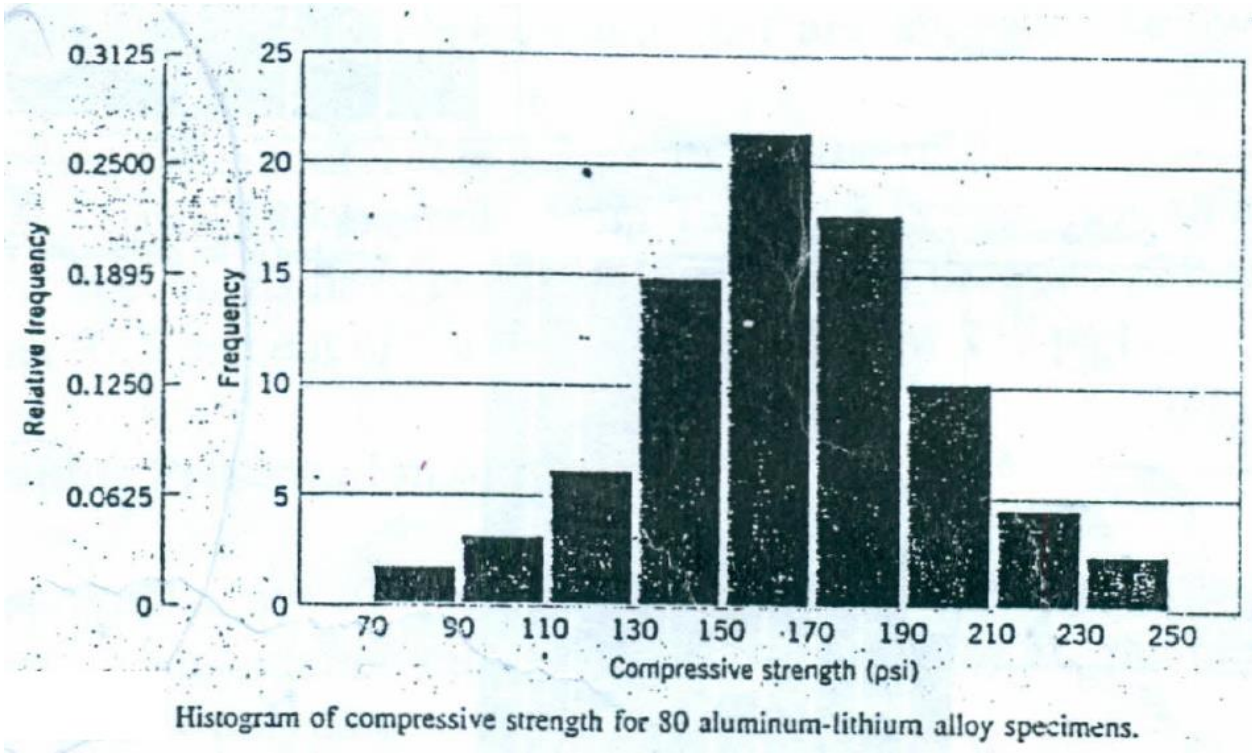
- In the **frequency polygon**, each class frequency (or percent frequency) is plotted corresponding to the midpoint of that class. The plotted points are then connected by straight lines.

GRAPHIC PRESENTATION OF FREQUENCY DISTRIBUTIONS



GRAPHIC PRESENTATION OF FREQUENCY DISTRIBUTIONS

Class Interval (psi)	Tally	Frequency
$70 \leq x < 90$	𝍸	2
$90 \leq x < 110$	𝍸	3
$110 \leq x < 130$	𝍸	6
$130 \leq x < 150$	𝍸 𝍸	14
$150 \leq x < 170$	𝍸 𝍸 𝍸	22
$170 \leq x < 190$	𝍸 𝍸	17
$190 \leq x < 210$	𝍸	10
$210 \leq x < 230$	𝍸	4
$230 \leq x < 250$	𝍸	2



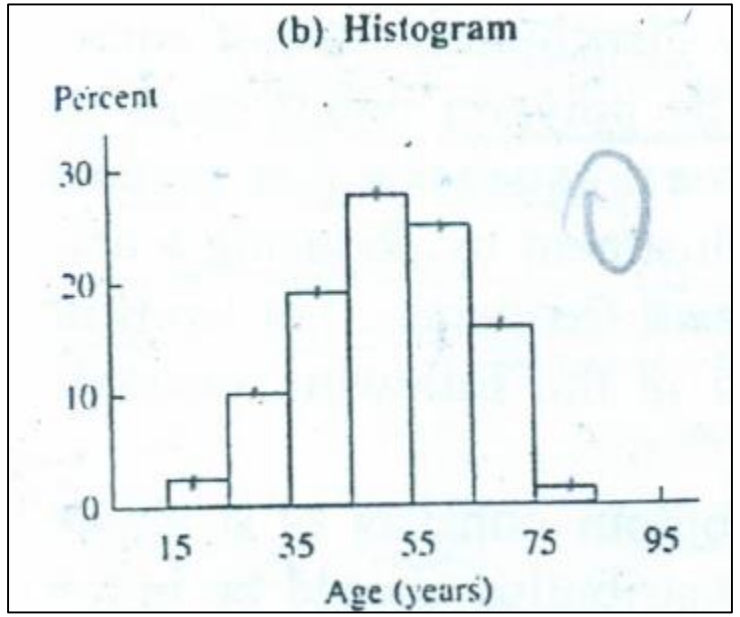
Histogram of compressive strength for 80 aluminum-lithium alloy specimens.

Histogram for Example 1

GRAPHIC PRESENTATION OF FREQUENCY DISTRIBUTIONS

(a) Percent Frequency Distribution

Age (years)	Percent of Farm Operators
15–under 25	2
25–under 35	10
35–under 45	19
45–under 55	27
55–under 65	25
65–under 75	16
75–under 85	1
Total	100
	(27,303)



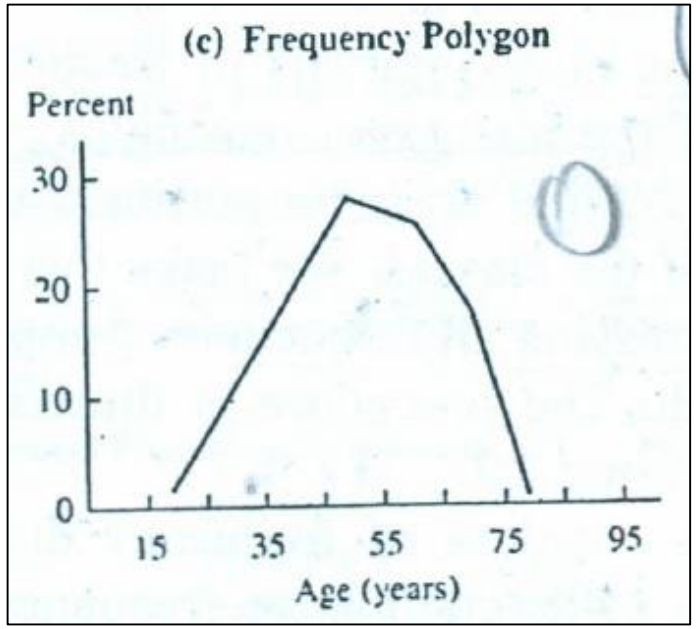


Figure 2.5a presents a percent frequency distribution of ages of 27,303 farm operators classified in equal, 10-year intervals. The age data were obtained from a recent survey of farm operators conducted by a farm equipment manufacturer. Figures 2.5b and 2.5c show a histogram and frequency polygon, respectively, of this frequency distribution.

GRAPHIC PRESENTATION OF FREQUENCY DISTRIBUTIONS

(a) Percent Frequency Distribution

Age (years)	Percent of Farm Operators
15–under 25	2
25–under 35	10
35–under 45	19
45–under 55	27
55–under 65	25
65–under 75	16
75–under 85	1
Total	100

(b) Cumulative Percent Frequency Distribution

Less than This Age (years)	Cumulative Percent of Farm Operators
15	0
25	2
35	12
45	31
55	58
65	83
75	99
85	100

(c) Plot of Cumulative Percent Frequency Distribution

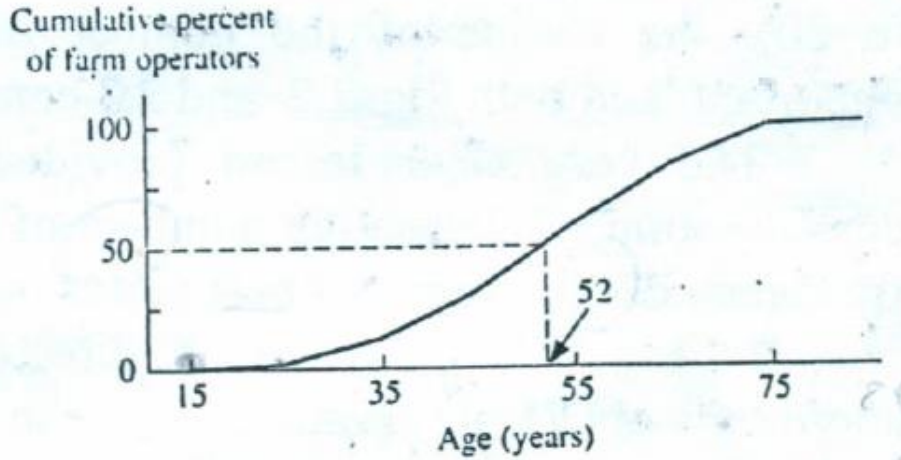


Figure 2.6a repeats the percent frequency distribution of the ages of farm operators and Fig 2.6b and 2.6c shows the cumulative percent frequency distribution and a plot of it, respectively. Here, the cumulative frequency is plotted against the corresponding limit.

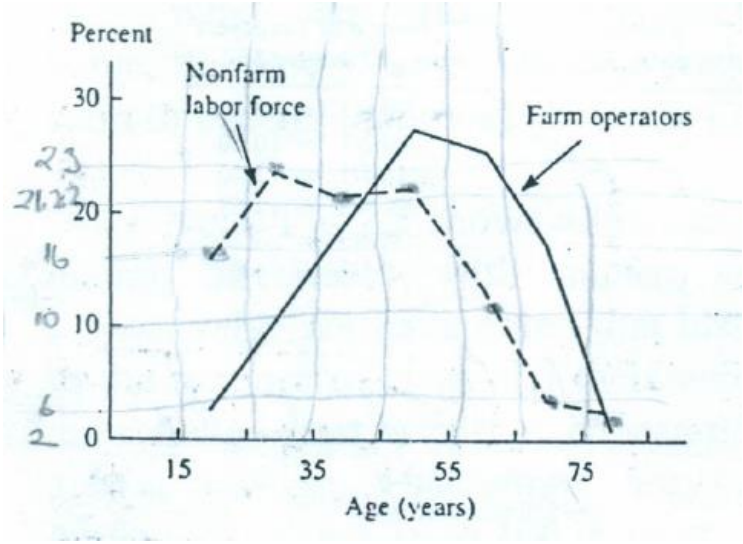
COMPARISON OF FREQUENCY POLYGONS

Two or more frequency polygons can be compared readily if

(1) they have the same class intervals and

(2) they have the same type of frequency expression. (if both expressed in total frequency or in percentage form

Plots of different cumulative frequency distributions can be compared readily on the same graph even if the distributions have different class intervals, as long as the total frequencies are the same or percent frequencies are used.



DISPLAYS OF QUALITATIVE DATA

- A **qualitative distribution** is the classification of the elements of a data set by a qualitative variable .
- The general principles of classification discussed for frequency distributions apply equally to the qualitative distributions .
- The number of classes utilized should be small enough to provide an effective summary, yet large enough to avoid losing essential information.
- When using categorical data, the classes should be drawn to have equal width.

DISPLAYS OF QUALITATIVE DATA

EXAMPLE:

Figure 2.10 presents a histogram showing the production of transport aircraft by the Boeing Company in 1985. Notice that the 737was the most popular model, followed by the 757, 747, 767,and 707.

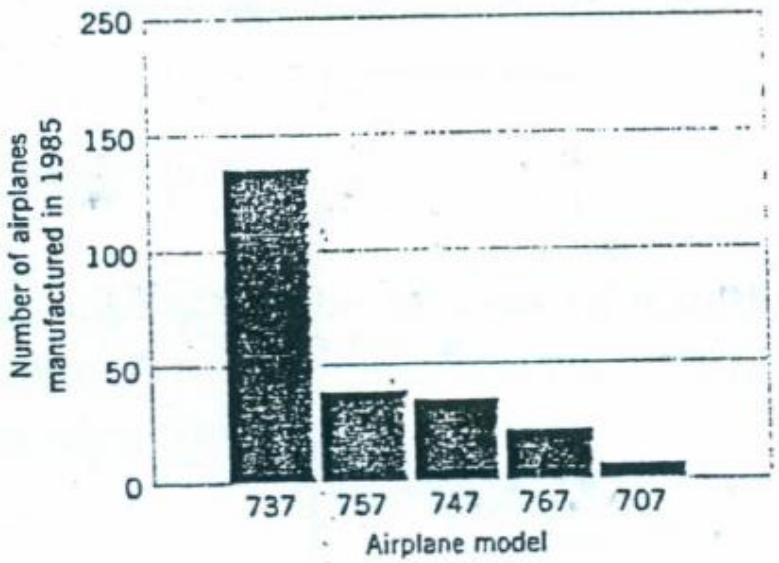


Figure 2.10 Airplane production in 1985. (*Source:* Boeing Company.)

DISPLAYS OF QUALITATIVE DATA

EXAMPLE:

An analyst for a large retailer wished to study the occupational profile of the company`s credit card customers and to compare it with the occupational profile of the total labor force as reported in a recent government publication. The analyst expected that the comparison would provide insights useful for credit policy, advertising, and billing.

Occupation	Number of customers	Percent of customers	Percent of total labor force
Managerial	38,835	29.5	9.2
Professional and technical	31,262	23.7	9.9
Service and recreation	14,011	10.6	10.9
Clerical and sales	12,797	9.7	20.7
Crafts and production workers	11,090	8.4	24.2
Laborers and unskilled workers	1,577	1.2	5.0
Others	22,273	16.9	20.1
Total	131,845	100.0 (131,845)	100.0

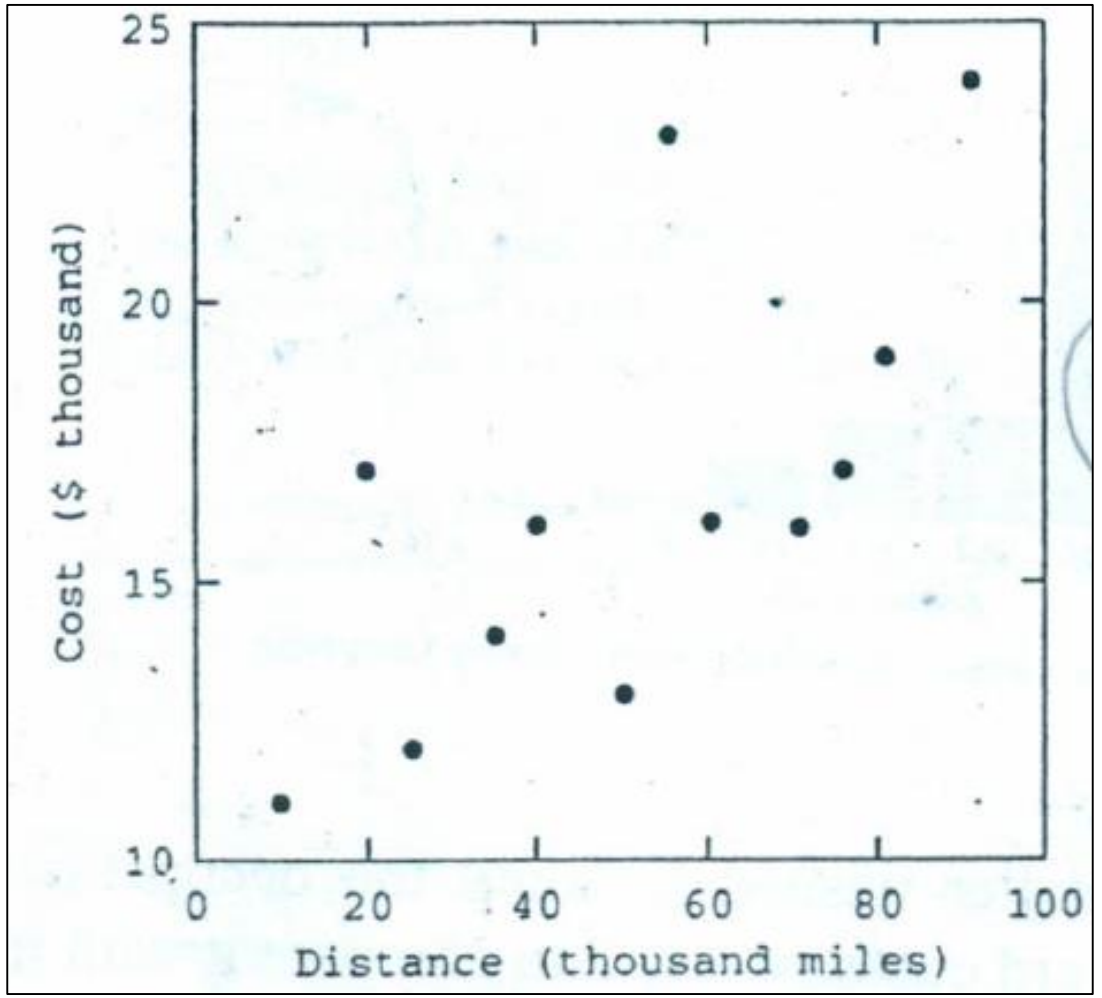
DISPLAYS OF BIVARIATE DATA

- The major objectives for analysing bivariate data is to gain insights into the nature of the relationship between the two variables.

- Bivariate data sets may be based on two. quantitative variables, two qualitative variables or one variable of each type.

DISPLAYS OF BIVARIATE DATA

- Scatter Plots. In a scatter plot, the observations for the two variables for each element in the data set are plotted in a two-dimensional graph.
- The pattern of points indicates whether a relationship exists between the two variables and if so, the nature of the relationship.



DISPLAYS OF BIVARIATE DATA

The figure presents a time series plot of the capital expenditures of a major oil company during 1981-1990. The points in the plot, corresponding to consecutive years, have been joined by straight lines to' show more clearly the time pattern of the annual capital expenditures.

The graph shows that capital expenditures experienced a generally upward trend over the decade. Then the board of the company can discuss the capital expenditures plans for the next year.