CAMBRIDGE

Cambridge University Press
978-1-316-60700-8 - A First Course in Statistics
Robert Loveday
Excerpt
More information

1

Frequency Distributions

1. The formation of a frequency distribution. Suppose there are thirty-six tomato plants in a garden each bearing three trusses of fruit and that we go round the plants counting the number of tomatoes per truss. The results of our survey might be recorded as in table 1A.

TABLE 1A

No. of tomatoes per truss		Frequency
8	1	1
9	1	1
10	11	2
11	~~1111~~	5
12	~~1111~~ 1111	9
13	~~1111~~ ~~1111~~ ~~1111~~	15
14	~~1111~~ ~~1111~~ ~~1111~~ 1111	19
15	~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~	20
16	~~1111~~ ~~1111~~ ~~1111~~ 1	16
17	~~1111~~ ~~1111~~	10
18	~~1111~~	5
19	111	3
20	1	1
21		0
22	1	1
	Total number of trusses examined	108

It should be noted that a small stroke is placed in the appropriate row of the middle column as each truss is counted and that every fifth stroke is drawn diagonally across the preceding four strokes to form small groups of five. This enables us to see at a glance the total for each row when we enter it in the right-hand column giving the *frequency* with which each particular number of tomatoes per truss occurs. Table 1A is an example of a *frequency distribution.*

2. The histogram. It is customary to represent a frequency distribution diagrammatically in the form of a *histogram* as shown in fig. 1.

3. The mode. The size of truss which occurs most frequently is 15, and this is called the *mode.*

4. A frequency distribution by grouping. Let us next suppose that, while gathering in the potato crop, we weigh in lb. and oz. the quantity of potatoes obtained from each of 100 roots and that our results are shown in table 1 B.

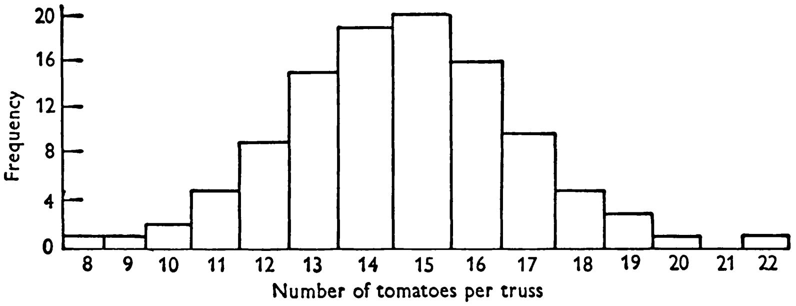

Fig. 1. Histogram.

TABLE IB

Weight of potatoes per root in lb.		Frequency
Under 3	卌 1111	9
3–	卌 卌 卌 卌 11	22
6–	卌 卌 卌 卌 卌 111	28
9–	卌 卌 卌 卌 1	21
12–	卌 卌 卌 11	17
15–18	111	3
	Total	100

In this case the weights in the left-hand column are grouped into classes of 3 lb. intervals the first row showing the number of roots yielding under 3 lb. of potatoes, the second row showing the number of roots yielding from 3 to 6 lb. but not including 6 lb. and so on. The histogram for this frequency distribution would be as shown in fig. 2.

5. Modal class. The 6–9 lb. class in the distribution shown in fig. 2 is called the *modal class*. In the first example the size of truss occurring most frequently was called the mode. In this example, however, it is quite probable that no two of the 100 roots have exactly the same weight, and there may be no actual weight which occurs most frequently. The name *modal class* is therefore given to the *class* which contains the greatest number of members, that is, to the 6–9 lb. class.

6. Discrete and continuous variation. It should be noted that the number of tomatoes per truss varies *discretely*, whilst the weight of potatoes per root varies *continuously*. The weight of a root of potatoes might have any value between 0 and 18 lb. It need not be a whole number. It might lie, for example, somewhere between 6 and 7 lb., while the number of tomatoes must be 6 or 7.

7. The frequency polygon. In fig. 2 the polygon *ABCDEF*, whose sides are the straight lines joining the mid-points of the tops of the rectangles of the histogram, is called the *frequency polygon*. Several frequency polygons can be drawn on one diagram, thus enabling us to compare frequency distributions. An illustration of this is given in §8.

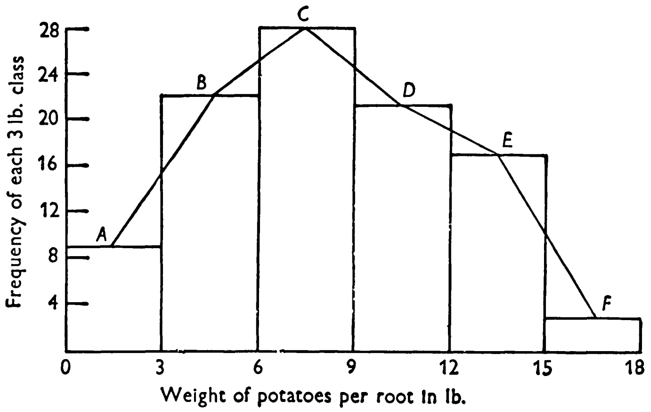

Fig. 2. Histogram and frequency polygon.

8. Example. An analysis of the number of words per sentence in the first hundred sentences of

(*a*) *Pride and Prejudice*, by Jane Austen,

(*b*) *The Cathedral*, by Hugh Walpole,

gives the frequency table as shown on p. 4.

On one sheet of graph paper draw two frequency polygons, setting them out in such a way that you can use them to compare the two distributions given in the table.

Discuss the use of this analysis to demonstrate that one of the novels is more difficult to read than the other. (Northern.)

The two frequency polygons shown in fig. 3 indicate quite clearly that Hugh Walpole's sentences are longer than Jane Austen's, the modal group of *The Cathedral* being 15–19 words per sentence, that of *Pride and Prejudice* 5–19 words per sentence. This might demonstrate that *The Cathedral* is more difficult to read than *Pride and Prejudice* were it not for the following important factors which are not shown in the graphs:

(i) Difficulty of vocabulary.

(ii) Does the author get the reader's interest?

No. of words per sentence	No. of sentences (*a*) *Pride and Prejudice*	(*b*) *The Cathedral*
0–4	6	2
5–9	33	12
10–14	22	14
15–19	15	19
20–24	10	18
25–29	4	6
30–34	2	5
35–39	3	6
40–44	2	6
45–49	2	4
50–54	0	1
55–59	0	1
60–64	0	2
65–69	0	0
70–74	0	1
75–79	1	0
80–84	0	0
85–89	0	0
90–94	0	1
95–99	0	0
100–104	0	2

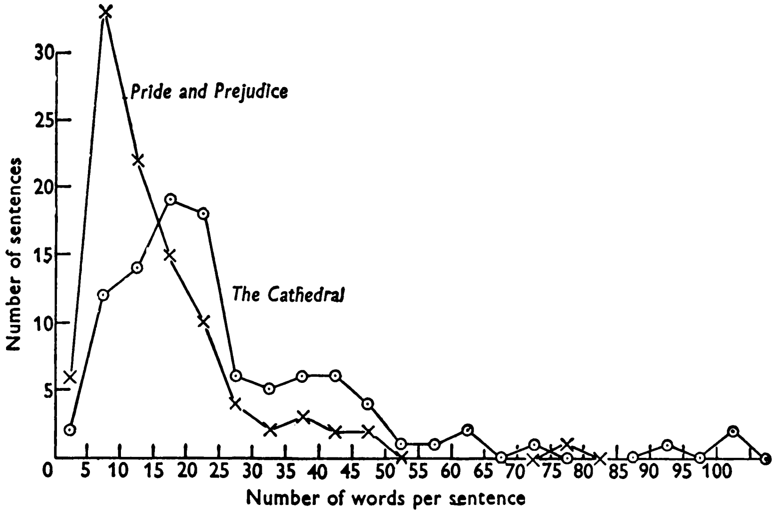

Fig. 3. Two frequency polygons in one diagram.

(iii) The first hundred sentences form a very small sample of the whole book.

(iv) They may not be typical of the whole book.

This example shows that the greatest care must be exercised in the drawing of conclusions from statistical data.

9. Experiments with dice. The histograms or frequency polygons obtained from the measurements of natural phenomena generally approximate in shape to certain well-defined curves called *frequency curves*. It is an amusing and useful exercise for the beginner to establish the shapes of these curves by throwing dice.

10. Throwing three dice. Normal distributions. It can be shown mathematically that, if three dice are cast, the probability of getting three sixes is 1/216. This means that there is 1 chance in 216 of scoring a total of 18. Further, the probability of getting two sixes and one five is 3/216. This means that there are 3 chances in 216 of scoring a total of 17. The following table is a complete list of the chances for the various possible total scores:

Total score	3	4	5	6	7	8	9	10
Chances in 216 of getting the score	1	3	6	10	15	21	25	27
Total score	11	12	13	14	15	16	17	18
Chances in 216 of getting the score	27	25	21	15	10	6	3	1

Fig. 4*a* shows the frequency curve obtained by drawing a graph from the above table. If the student actually throws three dice a large number of times and records the scores as shown in table 1A he will obtain a frequency distribution which will give a histogram or frequency polygon approximating in shape to the *normal frequency curve*. The author supervised the throwing of three dice 4320 (i.e. 20×216) times by twenty boys and obtained the following distribution, the frequency polygon of which approximates very closely in shape to fig. 4*a*:

Score	3	4	5	6	7	8	9	10
Frequency	19	52	93	194	285	411	473	510
Score	11	12	13	14	15	16	17	18
Frequency	541	520	438	333	218	149	68	16

11. Throwing six dice. Skewed distributions. If six dice are cast, it can be shown by the binomial theorem that the probability of getting no sixes is approximately 72/216, while that of getting one six is approximately 87/216. These and other values have been used in the drawing of fig. 4*b* which shows the shape of curve associated with a *skewed* distribution. When the greater part of the curve exists to the right of the mode it is said to be skewed positively. The student will also meet cases of

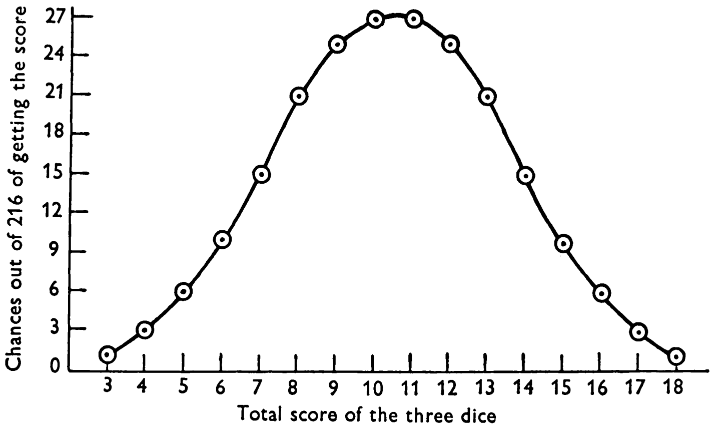

Fig. 4*a*. Approximately normal distribution. (A truly normal curve is rather more sharply peaked rising to 29 instead of 27. It will be fully discussed in 'A Second Course in Statistics'.)

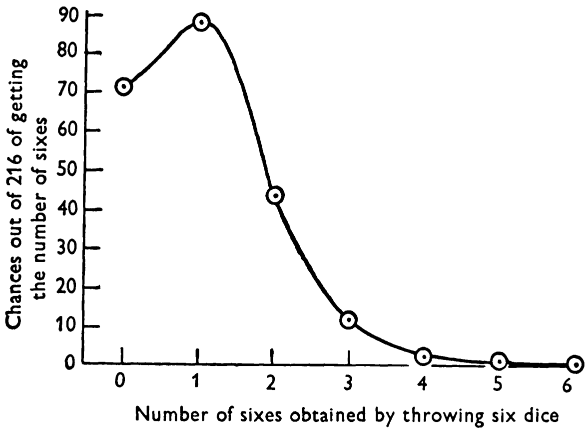

Fig. 4*b*. Skewed distribution (skewed positively).

negative skewness in which the greater part of the curve is to the left of the mode. *Skewness* is, therefore, a term which distinguishes distributions from *normal* distributions. The latter are symmetrical about the mode.

The following distribution was obtained in an actual experiment:

No. of sixes	0	1	2	3	4	5	6
Frequency	77	84	34	19	2	0	0

The student will find the frequency polygon obtained from this distribution closely resembles fig. 4*b*, but he should carry out a similar experiment himself.

12. Throwing one die. Rectangular distributions. If a single die is cast, each number has an equal chance of appearing uppermost. Thus the probability for each number is 1/6. This can be stated quite simply as 1 chance in 6, but in fig. 4*c* it has been interpreted as 36 chances in 216 in order to keep the totals the same in figs. 4*a*, *b* and *c*. Thus the

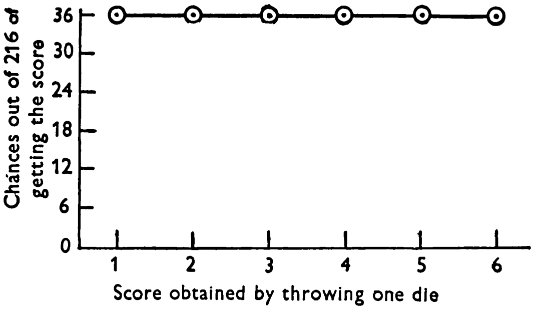

Fig. 4*c*. Rectangular distribution.

frequency curve in this case is a *straight line parallel to the abscissa of the graph* and the distribution is described as *rectangular*. The following table gives the results of an actual experiment:

Score	1	2	3	4	5	6
Frequency	94	96	115	140	105	98

The student should carry out a similar experiment for himself. In these experiments it is assumed, of course, that the dice are true (i.e. that they are not loaded).

13. Exercises.

1. The following are the populations, in hundreds, of 80 towns and villages in England and Wales, taken at random from a hotel directory (towns of over 50,000 inhabitants are omitted):

11	200	31	7	11	15	17	55
72	6	72	13	52	377	12	9
15	16	18	5	139	18	30	3
8	6	30	142	159	46	171	149
15	131	43	70	89	1	30	25
3	5	2	86	4	24	6	52
23	18	1	31	52	48	160	12
26	407	52	38	28	5	58	124
2	99	40	70	19	26	222	120
319	127	3	51	150	39	13	10

CAMBRIDGE

Cambridge University Press
978-1-316-60700-8 - A First Course in Statistics
Robert Loveday
Excerpt
More information

Form a frequency distribution and construct the histogram. The following grouping is appropriate: 0–29, 30–59, 60–89, etc. State which is the modal group. (Northern.)

Note that the mode is 52, since it occurs in the list 4 times and no other number occurs more than 3 times.

2. *Cambridge University Boat Club*

Weights of crews, in lb. (to nearest lb.), in the Lent Bumping Races, 1956

First Division

	Bow	2	3	4	5	6	7	Str.	Cox
Jesus 1	148	168	168	174	172	185	174	174	132
1st and 3rd Trinity 1	160	171	178	157	178	183	168	168	133
L.M.B.C. 1	161	162	171	173	156	181	168	155	128
Peterhouse 1	171	168	176	169	196	175	157	171	125
Pembroke 1	146	152	160	170	190	190	169	173	133
Emmanuel 1	145	169	165	174	199	182	178	170	122
Clare 1	154	144	178	194	168	187	144	154	124
Trinity Hall 1	163	165	168	171	184	172	169	184	136
King's 1	161	146	170	173	166	161	172	160	126
Corpus Christi 1	147	160	161	153	174	187	152	150	157
Magdalene 1	183	179	174	175	209	190	196	175	140
Jesus 2	173	164	178	169	209	164	154	145	134
Queens' 1	150	171	176	182	178	176	175	171	116
Christ's 1	148	168	176	178	183	186	170	159	138
Caius 1	161	165	173	182	186	176	177	202	118
St Catharine's 1	143	169	174	171	181	167	168	146	148

Second Division

	Bow	2	3	4	5	6	7	Str.	Cox
1st and 3rd Trinity 2	146	147	164	168	170	158	154	154	126
St Catharine's 2	156	164	178	148	182	158	186	154	136
Selwyn 1	164	160	159	193	188	158	164	185	127
L.M.B.C. 2	158	163	164	179	178	172	177	163	128
Downing 1	141	157	154	185	191	178	168	166	124
Sidney Sussex 1	152	154	189	176	177	189	187	189	130
Clare 2	178	147	170	151	189	170	182	156	112
Trinity Hall 2	161	150	160	178	173	175	161	162	124
Jesus 3	158	163	159	168	164	178	172	161	132
Fitzwilliam 1	156	181	203	208	169	190	151	155	123
Pembroke 2	162	156	177	166	195	175	153	147	136
Peterhouse 2	150	140	157	184	220	169	140	171	126
1st and 3rd Trinity 3	154	160	162	174	159	178	161	146	128
Emmanuel 2	150	136	149	185	150	224	164	152	134
Caius 2	154	158	161	178	171	166	174	147	125
King's 2	152	156	173	167	159	163	170	136	134

Form a frequency distribution for the above table which shows the weights of 288 men and construct the histogram. The following grouping is appropriate: $109\frac{1}{2}$–$119\frac{1}{2}$, $119\frac{1}{2}$–$129\frac{1}{2}$, $129\frac{1}{2}$–$139\frac{1}{2}$, etc., to $219\frac{1}{2}$–$229\frac{1}{2}$. The cox's constitute a different *weight population* to the rest of the crews, and it will be noticed that, if the weights of the coxes are excluded the frequency polygon is *approximately* normal in shape. The inclusion of the coxes causes it to be *skewed negatively.*

Which is the modal class?

3. The following is a record of marks obtained by a group of boys in an examination:

25	63	82	71	12	57	63	38	17	23	0	96	81	72	35	76	44
54	19	70	45	70	44	43	18	93	2	15	60	71	82	3	61	64
25	42	40	70	63	62	83	18	27	58	50	52	53	89	90	50	53
23	81	70	58	31	32	28	19	23	72	58	37	33	30	20	62	71
48	63	62	59	38	35	37	46	81	73	75	63	65	53	47	52	38

Tabulate as a frequency distribution, grouping in intervals of 10 marks 0–9, 10–19, 20–29, etc. Plot the result as a histogram and determine the modal class.

By examining the original marks determine the mode. (London.)

4. Make out a table similar to the following and use it to take a census of the sizes of caps worn by the boys of your school:

	Size of cap										
Form	$6\frac{1}{4}$	$6\frac{3}{8}$	$6\frac{1}{2}$	$6\frac{5}{8}$	$6\frac{3}{4}$	$6\frac{7}{8}$	7	$7\frac{1}{8}$	$7\frac{1}{4}$	$7\frac{3}{8}$	$7\frac{1}{2}$
I*a*	1	4	8	7	6	3	1	.	.	.	.
I*b*	.	3	5	7	8	6	0	1	.	.	.
II*a*	.	2	6	9	8	3	1	1	.	.	.
etc.	.	.	.	.	.	.	.	.	.	.	.
Totals for each size for the whole school	1	9	19	23	22	12	2	2	.	.	.

Draw the histogram and state the modal size of cap.

Explain briefly how the school outfitter could make use of your census.

5. If you can obtain permission to use the class registers, tabulate a frequency distribution of the ages of the pupils in your school as follows:

Age in years	10–	11–	12–	13–	14–	15–	16–	17–	18 and over
No. of pupils									

Draw the frequency polygon. You may find that it is approximately rectangular up to 16 years of age.

6. A company which manufactures tubes for television receivers conducted a test of a sample batch of 1000 tubes and recorded the number of faults in each tube in the following table:

No. of faults	0	1	2	3	4	5	6
Frequency	620	260	88	20	8	2	2

(London.)

Plot the histogram and you will find it is an example of what is called *the reverse-J shape.*

7. The following table gives the distribution of 1000 families according to the number of children:

No. of children in family	0	1	2	3	4	5	6	7
No. of families	25	306	402	200	53	8	4	2

(London.)

Draw the histogram and note that it is skewed positively.

14. Determination of the mode. The following table shows the frequency distribution of the marks of 800 candidates in an examination:

Marks	1–10	11–20	21–30	31–40	41–50
No. of candidates	10	40	80	140	170
Marks	51–60	61–70	71–80	81–90	91–100
No. of candidates	130	100	70	40	20

Construct the histogram and find the mode. (London.)

The histogram is shown in fig. 5. Note that the bases of the rectangles extend from $\frac{1}{2}$ to $10\frac{1}{2}$, $10\frac{1}{2}$ to $20\frac{1}{2}$, etc., so as to *enclose* the ranges 1–10,

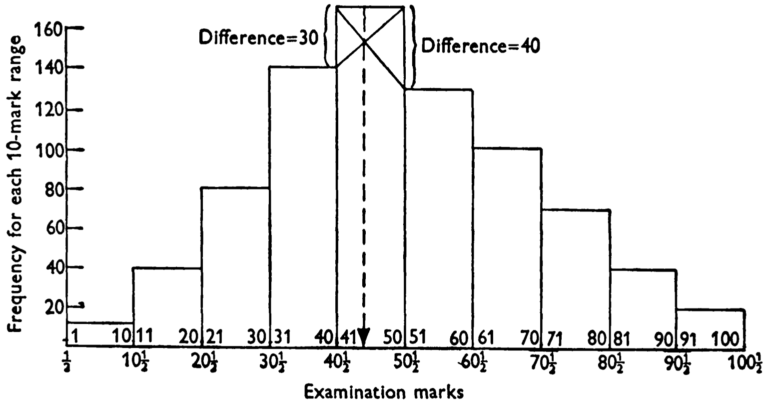

Fig. 5

11–20, etc. The modal class is 41–50. It is represented by the rectangle with base extending from $40\frac{1}{2}$ to $50\frac{1}{2}$. It contains 30 more candidates than the class below it and 40 more than the class above it. We therefore argue that the mode is *likely* to divide the modal class in the ratio 30:40. Hence we *estimate* the mode as

$$40\tfrac{1}{2}+\tfrac{3}{7}\text{ of }10 \quad\text{or}\quad 50\tfrac{1}{2}-\tfrac{4}{7}\text{ of }10,$$

and we write

$$\text{mode}=44{\cdot}8,$$

which indicates that, if the marks are integers, the mode is likely to be 44 or 45.

Fig. 5 shows a geometrical method of determining the mode using the properties of similar triangles.