Data, Descriptive Statistics and Statistical Graphics

Peoples intelligence is about 40 times greater than their college achievement since the average IQ is about 100 and the average GPA is about 2.5.

1 Frequency distributions

Once data have been collected, they must be presented in such a way that any important pattern becomes apparent. The concept of distribution is one of fundamental concepts in statistics. Data are summarized in such a way that frequencies of individual observations are given.

Speed of Light. Light travels very fast. It takes about 8 minutes to reach us from the Sun and over four years to reach us from the closest star outside the solar system. Radio and radar waves also travel at the speed of light, and an accurate value of that speed is important to communicate with astronauts and orbiting satellites. Because of its nature it is very hard to measure the speed of light. The first reasonably accurate measurements of the speed of light were made a little over 100 years ago by A. Michelson and S. Newcomb. Table below contains 66 transformed[1] measurements made by Newcomb between July and September 1882.

28	22	36	26	28	28
26	24	32	30	27	24
33	21	36	32	31	25
24	25	28	36	27	32
34	30	25	26	26	25
-44	23	21	30	33	29
27	29	28	22	26	27
16	31	29	36	32	28
40	19	37	23	32	29
-2	24	25	27	24	16
29	20	28	27	39	23

```
>light_scan()
 1:  28  22  36  26  28  28  26  24  32  30  27  24  33  21  36  32  31  25  24
20:  25  28  36  27  32  34  30  25  26  26  25 -44  23  21  30  33  29  27  29
39:  28  22  26  27  16  31  29  36  32  28  40  19  37  23  32  29  -2  24  25
58:  27  24  16  29  20  28  27  39  23
67:

> barplot(light)
```

[1]The entry 28, for instance, corresponds to the actual measurement of 0.000024828 second. That is the time needed for light to travel approximately 4.65 milles.

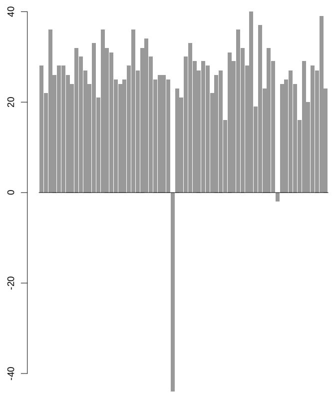

Figure 1: `barplot(light)`

```
> hist(light)
```

1.1 Normal distribution

```
> eda.shape(light)
```

2 Numerical characteristics of data

There are three kinds of statisticians: those who can count and those who can not.

2.1 Stem and leaf displays

2.2 Measures of location

Accordingly to the fittest style of lofty, mean, or lowly. - Milton

After a set of data has been collected it must be summarized (condensed, organized, categorized) for purpose of further analysis. The most important measures are measures of location or central tendency.

Mean. (From old French: meien = occupying a middle position) If $x_1, x_2, \ldots, x_n$ is a sample the mean is defined as

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

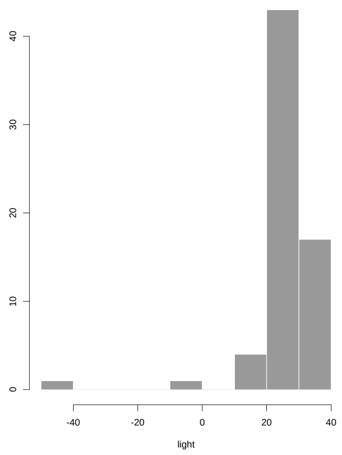

Figure 2: `hist(light)`

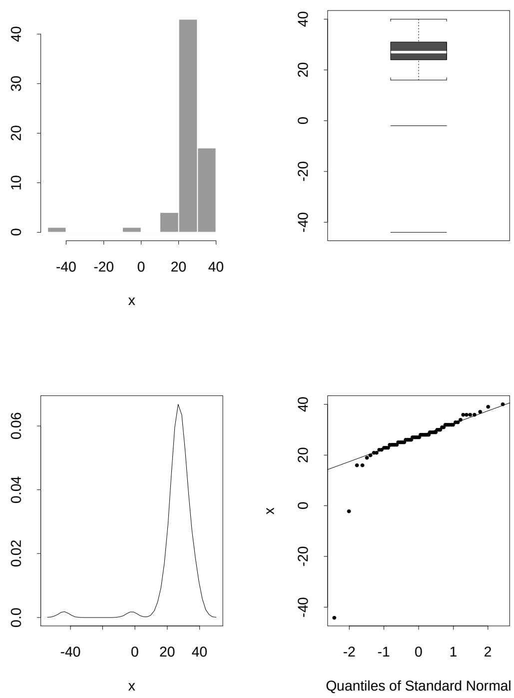

Figure 3: `eda.shape`

If the sample is composite

$$\bar{x} = \frac{x_1 \cdot f_1 + \ldots + x_k \cdot f_k}{f_1 + \ldots f_k} = \frac{\sum_{i=1}^k x_i f_i}{\sum_{i=1}^k f_i}.$$

Mode. The most frequent (fashionable[2]) observation in the sample (if such exists) is the mode of the sample. If the sample is composite then the observation x_i corresponding to the largest frequency f_i is the mode. Mode may not be unique. If there are tow modes, the sample is bimodal, three modes - trimodal, etc.

The mode is very easy to obtain by inspection. However, it is rather unreliable measure of central tendency since it depends on grouping, outliers, etc.

Median. (Latin: *medianus* = middle) To define the median we first need a definition of order statistics.

If the sample $x_1, x_2, \ldots, x_n$ is ordered increasingly then the new, ordered sample $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is called *order statistic.* The indices in $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are put in parentheses to suggest that x_i is generally not the same as $x_{(i)}$. From the definition $x_{(1)}$ is the minimum and $x_{(n)}$ is the maximum of the sample.

If the sample size n is an odd number ($n = 2k + 1$) then there is one middle element in the ordered sample, $x_{(\frac{n+1}{2})}$. The median is the middle element in the order statistic. But if n is even there are two middle elements. Their average is the median.

8. **Hours spent studying per week.**[3] All time around campus you hear people bragging about how they never study. It is felt, given the academic challenges of Duke, this could not be entirely true, especially when you consider how competitive Duke students are. There is a great fear of being labeled as "loser" or "geek" if person studies excessively. We assume that our 20 randomly selected students were honest about how much (in average) they really studied per week.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Hours	12	10	16	5	11	10	10	8	9	9	12	13	6	11	9	11	14	14	10	3

Find the location measures (the mean, 5%-trimmed mean, mode, and the median) for the hours spent studying per week.

```
> ho_scan()
1: 12 10 16 5 11 10 10 8 9 9 12 13 6 11 9 11 14 14 10 3
21:
> mean(ho)
[1] 10.15
> mean(ho, trim=.05)
[1] 10.22222
```

The ordered sample is:

$$3 5 6 8 9 9 9 10 10 10 10 11 11 11 12 12 13 14 14 16$$

[2]mode(fr) = fashion

[3]From students' project in STA 110 Spring 1993.

The median is the average of 10th and 11th order statistics. Or in Splus:

```
> median(ho)
[1] 10
```

In the composite form the hours of stydying can be represented as a frequency table

Hours	3	5	6	8	9	10	11	12	13	14	16
frequency	1	1	1	1	3	4	3	2	1	2	1

Mo=10 since the frequency of 10 the largest (4).

Small World The following information is for Problems 1 through 5. Travers and Milgram [4] conducted an experiment to find out how many intermediaries were needed to reach one person in the United States from another person through a chain of personal acquaintances. Among 296 people, 196 from Nebraska and 100 from Boston, 64 chains were completed. Each person was asked to mail a packet to the "target" person if the sender knew the target person on a first-name basis, or to the person the sender knew on a first-name bases and believed most likely to so know the target person. Table 1 gives the distribution of the lengths of the completed chains. 1. Make a cumulative frequency table from Table

1.

2. Make a cumulative frequency graph.

3. Use the cumulative frequency graph to estimate the median length of the chains of acquaintances required to get from one person to another in the nation. Find Q_1, Q_3, and the IQR.

4. Had it been possible to persuade the people who interrupted the chains to continue, and if all chains had gone to completion, do you think the median would have been larger or smaller? Explain why. (Fact: The median stopping point for interrupted chains was 2.)

5. Considering the evidence in Table 1, do you think the median length of the chains of acquaintances to reach the president of the United States would be longer or shorter than the median length of chains to reach a random person? Explain why.

Length	1	2	3	4	5	6	7	8	9	10	11	12	≥ 13	Total
Frequency	0	0	2	3	8	14	8	16	6	2	2	3	0	64

[4]Source: J. Travers and S. Milgram (1969). An experimental study of "the small world problem." *Sociometry* 32: 435.

2.3 Measures of spread

2.4 Coeficient of Variation

Definition. Coeficient of variation is the quantity

$$CV = \frac{s}{\bar{X}}100\%$$

and expresses the variability in units of the mean. The assumption is that the mean is nonnegative.

Two places that CV proves useful.

(1) (in a survey, for example) which one of two measures varies more in the population, where the specific value for variance is dependent on the measurement scale used. It might be nice, for example, to determine whether UK voters (whose parties have somewhat more distinct policy positions than in the US) have as a result wider variation in their evaluations of the parties than in the US. Problem is the British election survey takes evaluations scored 0-10, while the US National Election Survey gets evaluations scored 0-100. Using CV would allow easy comparisons of the amount of variation without worrying about the different scales.

(2) You and I both teach sections of the same class, but we make up our own finals. To compare the effectiveness of our final exam designs at creating maximum variance in exam scores (an oft-unmentioned goal of exam design) we can compare CV, which is not related to how many points we have on the exam, to the mean performance of our students, or to the mean difficulty of the exam. It might be related to the homogeneity of our classes, but to test exams like this we would need CV.

Cost of electrical insulation. [5] To reduce the cost of electrical insulation, a substitute insulating material was extensively tested for its electrical resistance. The first 50 measurements are given in Table 2. Reading from left to right beginning at the top row gives the order in which measurements were made. Use these data to make (a) a histogram, and (b) a stem-and-leaf diagram.

50.5	43.5	43.5	39.8	42.9	44.3	44.9	42.9	39.8	39.3
36.5	37.6	33.0	36.9	34.6	52.0	51.0	46.4	51.0	54.5
46.4	47.2	48.1	45.7	44.1	40.7	45.7	51.9	47.3	46.4
46.4	49.0	47.9	48.5	47.0	46.0	41.4	44.1	41.8	47.9
47.0	43.4	49.0	57.5	47.4	50.0	49.0	42.6	41.7	38.5

Antelope Bands. E. T. Seton, in The Arctic Prairies, listed the numbers of antelopes in

[5]Source: W. A. Shewhart (1931). *Economic Control of Quality of Manufactured Product*, p. 20. Princeton. N.J.: D. Van Nostrand.

26 bands seen along the Canadian Pacific Railroad in Alberta, within a stretch of 70 miles, as follows:

8,	4,	7,	18,	3,	9,	14,	1,	6,	12,	2,	8,	10
1,	3,	6,	4,	18,	4,	25,	4,	34,	6,	5,	16,	14

(a) Construct a stem-and-leaf diagram for these data.

(b) Use the diagram of part (a) to answer the following questions:

(i) What is the spread of the measurements?
(ii) What measurements seem most popular?
(iii) Estimate the center of the distribution.
(iv) Does the distribution separate into isolated groups?

Safety failures. The accompanying table contains one year's data on the number of safety failures during standby and operations at 17 U.S. nuclear power plants. If we viewed these 17 cases as a random sample from a larger population of plants and years, they could be used to form an interval estimate of the mean number of failures per year in that population.

Reactor	Type[6]	Failures during standby	Failures during operation
Dresden 1	BWR	1	2
Yankee	PWR	8	5
Indian Point 1	PWR	7	12
Humboldt Bay 3	BWR	5	7
Big Rock Point	BWR	4	3
San Onofre 1	PWR	3	7
Haddam Neck	PWR	0	3
Nine Mile Point 1	BRW	7	13
Oyster Creek	BWR	10	19
Ginnaa	PWR	1	5
Dresden 2	BWR	8	20
Point Beach 1	PWR	2	4
Millstone 1	BWR	7	22
Robinson 2	PWR	3	17
Monticello	BWR	10	34
Dresden 3	BWR	4	22
Palisades	PWR	6	22

(a) Find the mean, standard deviation, median, and pseudo-standard deviation for the number of failures during power plant standby. Use these statistics to summarize the distribution's shape. Is the normality assumption plausible?

(c) Construct a box plot or a stem-and-leaf display for the number of nuclear power plant failures during operation. Describe the distribution's shape.

(d) Take square roots of the number of failures during operation for each power plant. Construct a stem-and-leaf display of these square roots. Has symmetry been improved?

Radioactive materials. A reconnaissance study of radioactive materials was conducted in Alaska[7] to call attention to anomalous concentrations of uranium in plutonic rocks. The amounts of uranium, in 13 locations under the Darby mountains are

7.92 10.29 19.89 17.73 10.36 13.50 8.81 6.18 7.02 11.71 8.33 9.32 14.61

Find: (i) mean, (ii) standard deviation (iii) Median and Quartiles, and (iv) interquartile range.

1970 Census Population in Millions.

Ala	3.44	Alaska	0.30	Ariz	1.77	Ark	1.92	Calif	19.95
Colo	2.21	Conn	3.03	Del	0.55	Fla	6.79	Ga	4.59
How	0.77	Idaho	0.71	Ill	11.01	Ind	5.19	Iowa	2.83
Kan	2.25	Ky	3.22	La	3.64	Me	0.99	Md	3.92
Mass	5.69	Mich	8.88	Minn	3.81	Miss	2.22	Mo	4.68
Mont	0.69	Neb	1.48	Nev	0.49	Nh	0.74	Nj	7.17
Nm	1.02	Ny	18.24	Nc	5.08	Nd	0.62	Ohio	10.65
Okla	2.56	Ore	2.09	Pa	11.79	Ri	0.95	Sc	2.59
Sd	0.67	Tenn	3.92	Texas	11.2	Utah	1.06	Vt	0.44
Va	4.65	Wash	3.41	W.Va	1.74	Wis	4.42	Wyo	0.33

Frequency Distribution for Size of Catch Among 911 Anglers. Table below presents data from a survey of 911 anglers done during a particular time period on the lower Current River in Canada ("Fisherman's Luck", *Biometrics* (1976): 265-71). Over 50 the fishermen - 56.53

[7]Miller, T., and Bunker, C. *Journal of Research, U.S. Geological Survey (1976), 367-377.*

Number of Fish Caught	Number of (Frequency)	Relative Frequency	Cumulative Relative Frequency
0	515	.5653	.5653
1	65	.0714	.6367
2	60	0659	.7750
3	66	.0724	.8332
4	53	.0582	.8936
5	55	.0604	.8936
6	27	.0296	.9232
7	25	.0274	.9506
8	25	.0274	.9780
9	20	.0220	1.0000
Total	911	1.0000	

Data collected by the Office of Population Censuses and Surveys. The data relate to numbers of live births in England and Wales during 1980. The table below gives information on the distribution of births according to the age of the mother:

Age of mother (years)	Number of births
15 to 19	60754
20 to 24	201541
25 to 29	223438
30 to 34	129908
35 to 39	33893
40 to 44	6075
45 to 49	625
Total	656234

Arrivals at Check-Out. A supermarket chain employed a consultancy firm to investigate service times at their newest store. One aspect that was examined was the time between successive customers arriving at a check-out. Times between arrivals (minutes) were obtained for various periods of a week and those for part of Monday morning are given below:

6. 0	7. 5	2.7	2.0	2.8	2.6	5.0	11.7	1.3	1.7
3.9	1.2	0.6	0.7	1.7	6.7	2.1	9.6	3.7	4.5
0.5	5.9	1.2	3.1	3.3	0.2	1.9	0.2	2.4	2.0
4.4	0.2	1.2	8.6	0.3	1.9	5.1	0.4	10.4	10.3

Use a histogram or a dot diagram to see whether or not it is reasonable to suggest that these data could have come from a symmetric distribution.

The average time for established stores for a corresponding period is 1.2 min. Is there evidence to suggest that the median time for the new store is greater than this average?

Babe Ruth. Here are the number of home runs that Babe Ruth hit in each of his 15 years with the New York Yankes (1920 to 1934) [8]

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

Make a stemplot for these data.

About us. The exam scores for the students in an introductory statistics class are as follows.

88	82	89	70	85
63	100	86	67	39
90	96	76	34	81
64	75	84	89	96

Find the "5-number summary" for the above data set.

Moses. Let us consider an early statistical investigation - namely, Moses's census of 1490 B.C. (in Numbers 1 and 2) given in table below.

Tribe	Number of militarily fit males
Judah	74,600
Issachar	54,400
Zebulun	57,400
Reuben	46,500
Simeon	59,300
Gad	45,650
Ephraim	40,500
Manasseh	32,200
Benjamin	35,400
Dan	62,700
Asher	41,500
Naphtali	53,400
Levi	-
Total	603,550

The tribe of Levi was not included in this or any subsequent census, owing to special priestly exemptions from the misfortunes which inevitably accrue to anyone who gets his name on a governmental roll. A moment's reflection on the table quickly dispels any notions

[8] Data from *The Baseball Encyclopedia*, 3d ed., Macmillan, New York, 1976.

as to a consistent David versus Goliath pattern in Israeli military history. No power in the ancient world could field an army in a Palestinian campaign which could overwhelm a united Israeli home guard by sheer weight of numbers.

Haberdasher's histogram In 1538 A.D., owing to a concern with decreases in the English population caused by the plague, Henry VIII had ordered parish priests of the Church of England to keep a record of christenings, marriages, and deaths. By 1629, monthly and even weekly birth and death lists for the city of London were printed and widely disseminated.

The need for summarizing these mountains of data led John Graunt, a London haberdasher, to present a paper to the Royal Society, in 1661, or the "bills of mortality." This work, *Natural and Political Observations of the Bills of Mortality*, was published as a book in 1662. In that year, on the basis of Graunt's study and on the recommendation of Charles II, the members of the Royal Society enrolled Graunt as a member. (Graund was dropped from the rostrum five years later, perhaps because he exhibited the anti- intellectual attributes of being (a) a small businessman, (2) a Roman Catholic, and (3) a statistician.) Graunt's entire study is too lengthy to dwell on here. We simply show in table below his near histogram , which attempts to give probabilities that an individual in London will die in particular age interval.

Age interval	Probability of death in interval
0-6	.36
6-16	.24
16-26	.15
26-36	.09
36-46	.06
46-56	.04
56-66	.03
66-76	.02
76-86	.01

Yet another statistics classes. At a small New England college, there are only 10 classes of statistics - with the following distribution of class size:

Class Size	Relative Frequency
10	.50
20	.30
90	.20
Total	1.20

The student newspaper reported that the average statistics student faced a class size "over 50." Alarmed, the Dean asked the statistics professors to calculate their average class size, and they reported, "under 30." Who's telling the truth? Or are there two truths? Specifically, calculate:

(a) What class size does the average professor have?

(b) What class size does the average student have?

Annual Alcohol Sales per Person 1975-1977[9]

	State	Alcohol Sales[10]
1	Utah	1.69
2	Idaho	2.55
3	Oregon	2.79
4	New Mexico	2.90
5	Washington	2.97
6	Montana	3.09
7	Arizona	3.15
8	Colorado	3.23
9	Hawaii	3.26
10	Wyoming	3.37
11	California	3.38
12	Alaska	3.96
13	Nevada	6.88

1980 Marriages. An outlier is a value that lies far from the central part of a distribution – more than 1.5(IQR) beyond the first of third quartile. A single outlier can greatly influence the mean and even more so the standard deviation. A good example of outlier distortion appears when we examine marriage rates in the 50 U.S. states by geographical region (Table below).[11]

[9] Linsky, A., J. Colby, and M. Straus (1985). Social stress, normative constraints, and alcohol problems in American states. Paper presented at the annual meeting of the Society for the Study of Social Problems, Washington D.C., August 24.

[11] 1980 Marriages per 1,000 population in 50 U.S. States, Soted by region

		Marriage			Marriage
Delaware	South	7.46	Alaska	West	13.34
N. Carolina	South	7.94	Idaho	West	14.23
W. Virginia	South	8.92	Wyoming	West	14.63
Kentucky	South	8.94	Nevada	West	142.83
Louisiana	South	10.33	New Jersey	Northeast	7.58
Maryland	South	10.97	Pennsylvania	Northeast	7.90
Mississippi	South	11.07	Rhode Island	Northeast	7.91
Florida	South	11.12	Massachusetts	Northeast	8.07
Virginia	South	11.26	New York	Northeast	8.23
Arkansas	South	11.60	Connecticut	Northeast	8.38
Alabama	South	12.59	New Hampshire	Northeast	10.05
Texas	South	12.77	Vermont	Northeast	10.22
Tennessee	South	12.89	Maine	Northeast	10.71
Georgia	South	12.93	Wisconsin	Midwest	8.74
Oklahoma	South	15.37	Nebraska	Midwest	9.07
S. Carolina	South	17.27	Minnesota	Midwest	9.23
Oregon	West	8.74	Ohio	Midwest	9.25
California	West	8.91	N. Dakota	Midwest	9.34
Montana	West	10.60	Michigan	Midwest	9.38
Arizona	West	11.12	Iowa	Midwest	9.43
Washington	West	11.55	Illinois	Midwest	9.61
Utah	West	11.61	Kansas	Midwest	10.51
Colorado	West	12.08	Indiana	Midwest	10.54
Hawaii	West	12.29	Missouri	Midwest	11.11
New Mexico	West	12.77	S. Dakota	Midwest	12.74

The above table gives the summary statistics. Judging from the means, marriage rates are lowest in the Northeast ($\bar{X} = 8.78$) and somewhat higher in the Midwest ($\bar{X} = 9.91$) and South ($\bar{X} = 11.47$). The mean marriage rate for the 13 Western states ($\bar{X} = 21.9$) far exceeds the rest. Also, the West's standard deviation is over 10 times higher than elsewhere. Something unusual seems to be going on out West.

Nevada's high marriage rate, like its high alcohol sales does not directly reflect the behavior of Nevadans themselves. Famous for its roadside wedding chapels and loose laws, Nevada draws thousands of visitors for quick marriages every year. Counting these marriages as part of the state's overall marriage rate greatly inflates it. The reason nonresidents have such an impact on alcohol and wedding rates is partly due to Nevada's relatively low population. In California the same number of nonresident weddings would scarcely alter the overall rate.

2.5 Visualizing Multivariate Data: Chernoff Faces

The use of face representation is an interesting approach for a first look at multivariate data which is effective in revealing rather complex relations not always visible from simple correlations based on twodimensional linear theories. It can be used to aid in cluster analysis, discrimonation analysis and to detect subsrtantial changes in time series. People grew up studying faces all the time. Small and barely maesurable differences are easily detected and evoke emotional reactions from a long catalogue buried in memory. The human mind subconsciously operates as a high speed computer, filtering out insignificant phenomena and focusing on the potentially important.

The feature parameters are: Variable 1-area of face; Variable 2-shape of face; Variable 3-length of nose; Variable 4-location of mouth; Variable 5-curve of smile; Variable 6-width of mouth; Variables 7,8,9,10,11-location, separation, angle, shape and width of eyes; Variable 12-location of pupil; and Variables 13,14,15-location, angle and width of eyebrow.

As an exapmple we give 18 observations of nummulited specimens from the eocene yellow limestone formation, Jamaica.[12]

ID	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
1	160	51	10	28	70	450
2	155	52	8	27	85	400
3	141	49	11	25	72	380
4	130	50	10	26	75	560
6	135	50	12	27	88	570
41	85	55	13	33	81	355
42	200	34	10	24	98	1210
43	260	31	8	21	110	1220
44	195	30	9	20	105	1130
45	195	32	9	19	110	1010
46	220	33	10	24	95	1205
81	55	50	10	27	128	205
82	70	53	7	28	118	204
83	85	49	11	19	117	206
84	115	50	10	21	112	198
85	110	57	9	26	125	230
86	95	48	8	27	114	228
87	95	49	8	29	118	240

In S, we read data as a matrix where one (multivariate) observation is in a row:

[12]Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically, JASA, 68, 361-366.

```
> jam
      [,1] [,2] [,3] [,4] [,5] [,6]
 [1,]  160   51   10   28   70  450
 [2,]  155   52    8   27   85  400
 [3,]  141   49   11   25   72  380
 [4,]  130   50   10   26   75  560
 [5,]  135   50   12   27   88  570
 [6,]   85   55   13   33   81  355
 [7,]  200   34   10   24   98 1210
 [8,]  260   31    8   21  110 1220
 [9,]  195   30    9   20  105 1130
[10,]  195   32    9   19  110 1010
[11,]  220   33   10   24   95 1205
[12,]   55   50   10   27  128  205
[13,]   70   53    7   28  118  204
[14,]   85   49   11   19  117  206
[15,]  115   50   10   21  112  198
[16,]  110   57    9   26  125  230
[17,]   95   48    8   27  114  228
[18,]   95   49    8   29  118  240
> faces(jam)
```

If we plot the pairs of variables in Jamaica data the overall picture on relations of variables is not very clear

```
> pairs(jam)
```

2.6 Exercises

Survey. Ann Landers once asked her readers, “If you had it to do over again, would you have children?” She received nearly 10,000 responses, of which 70% said “No.”

a. Why is this not a representative sample of all American parents?

b. For the population of American parents, is the true proportion of “No” responses likely to be higher or lower than 70%?

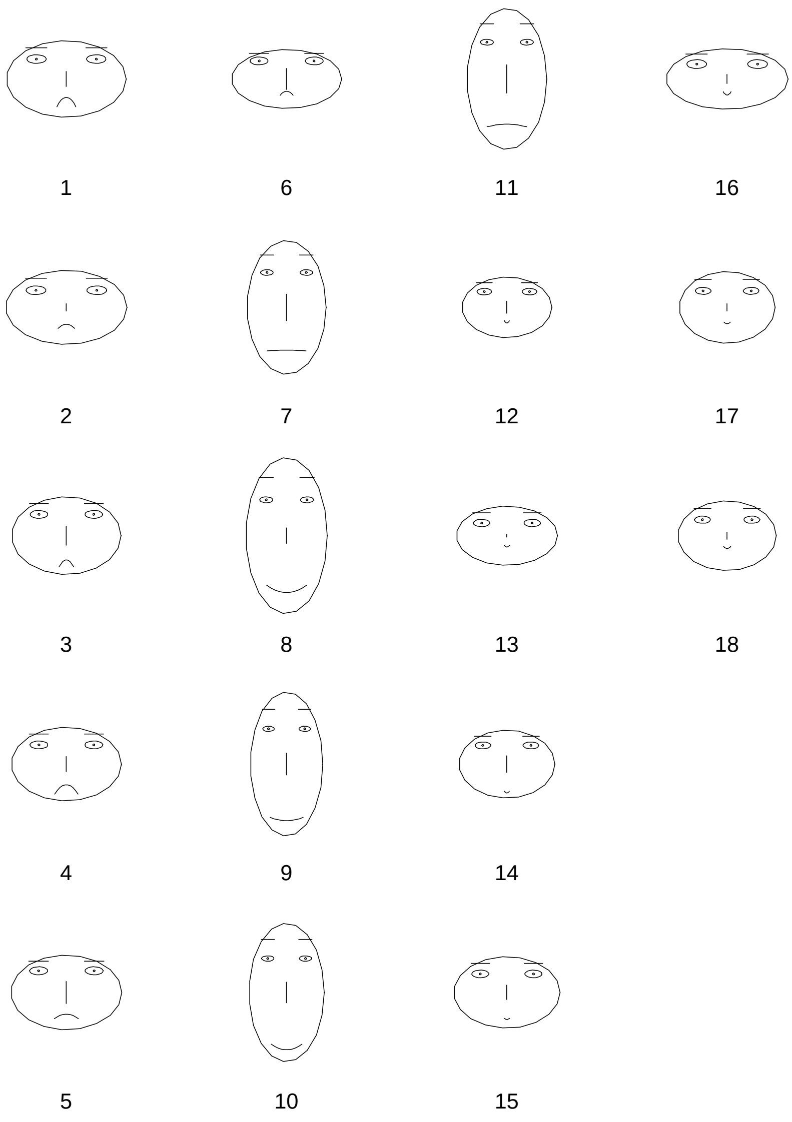

Figure 4: Chernoff faces for Jamaica Data

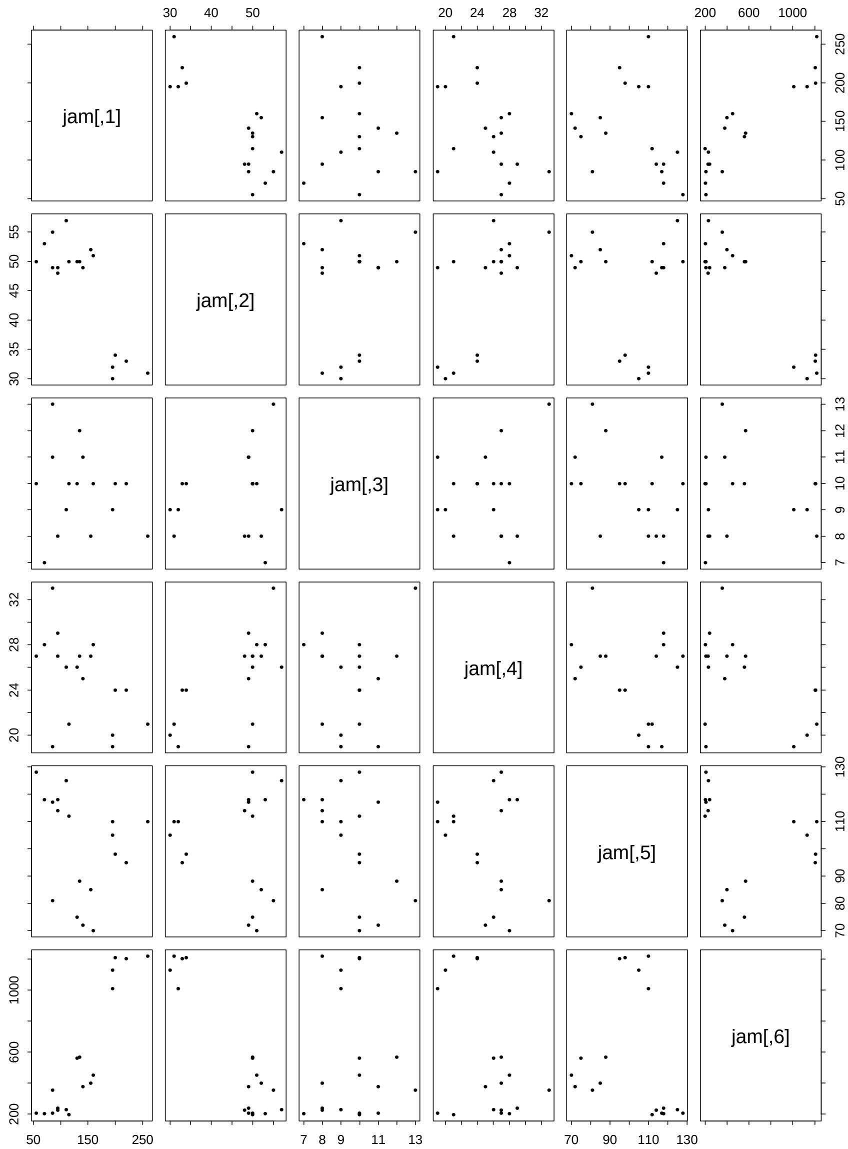

Figure 5: Pairs for Jamaica Data

Wechsler Scale. Scores on the Wechsler Adult Intelligence Scale for Americans in the 60–64 age group are approximately normally distributed with a mean of 90 and standard deviation 25.
a. About what percent of these people have scores between 65 and 115?

b. What is the 90^{th} percentile for this age group?

c. What is the inter-quartile range for this age group?

Mushrooms. The unhappy outcome of uninformed mushroom-picking is poisoning. In many cases such poisoning is due to ignorance or a superficial approach to identification. The most dangerous fungi are Death Cap (*Amanita Phaloides*) and two species akin to it *Amanita Verna* and Destroying Angel (*Amanita Virosa*). These three toadstools cause the majority of fatal poisoning.

One of the keys for mushroom identification is the spore deposit. Spores of *Amanita Phaloides* are colorless, almost spherical, and smooth. Measurements in $m\mu$ of 28 spores are given below:

9.2	8.8	9.1	10.1
8.5	8.4	9.3	8.7
9.7	9.9	8.4	8.6
8.0	9.5	8.8	8.1
8.3	9.0	8.2	8.6
9.0	8.7	9.1	9.2
7.9	8.6	9.0	9.1

Find the **five number summary** for the spore measurements data.
Find the mean and the mode.

Favorite. Known: $\bar{y} = 6$ and $s_y = 3.32$, and $n = 11$.
Observation $y_{11} = 7$ removed.
Find $\bar{y}_{new}$ and $s_{y(new)}$.

Solution:
$\bar{y}_{new} = \frac{11\cdot 6-7}{10} = \frac{59}{10} = 5.9.$
$s^2_{y(old)} = 11.0224 = \frac{1}{10}(\sum_{i=1}^{11} y_i^2 - 11\cdot 6^2).$
Thus, $\sum_{i=1}^{11} y_i^2 = 506.224;$
$\sum_{i=1}^{10} y_i^2 = 506.224 - 7^2 = 457.224.$
$s^2_{y(new)} = \frac{1}{9}(457.224 - 10\cdot 5.9^2) = 12.12489.$
$s_{y(new)} = 3.48.$

Sleep deprivation. A psychologist is interested in the effect of sleep deprivation on motor performance. Thirty subjects are randomly assigned to a 12-hour sleep deprivation group or a 36-hour sleep deprivation group. After being “sleep-deprived,” the subjects’ (ordered) reaction times on a task assessing fine-motor skills are as follows:

12-hour sleep deprivation group	3.42	3.55	3.59	3.65	3.77
	3.87	3.94	3.96	3.97	4.00
	4.11	4.18	4.22	4.28	4.31
36-hour sleep deprivation group	7.59	7.74	7.83	7.90	8.00
	8.01	8.03	8.03	8.15	8.16
	8.19	8.20	8.46	8.50	8.50

Find the 5-number summaries for the 12-hour and 36-hour sleep deprivation groups.

Toxic Emissions. The total toxic emissions reported by the EPA for ten counties in the United States are reported below. Data are in millions of pounds.

5200	612	581	512	423	404	349	329	309	284

(a) Find the *five-number summary* for the above data.

(b) Find the sample mean and sample standard deviation for the last 4 measurements: 349, 329, 309, 284.

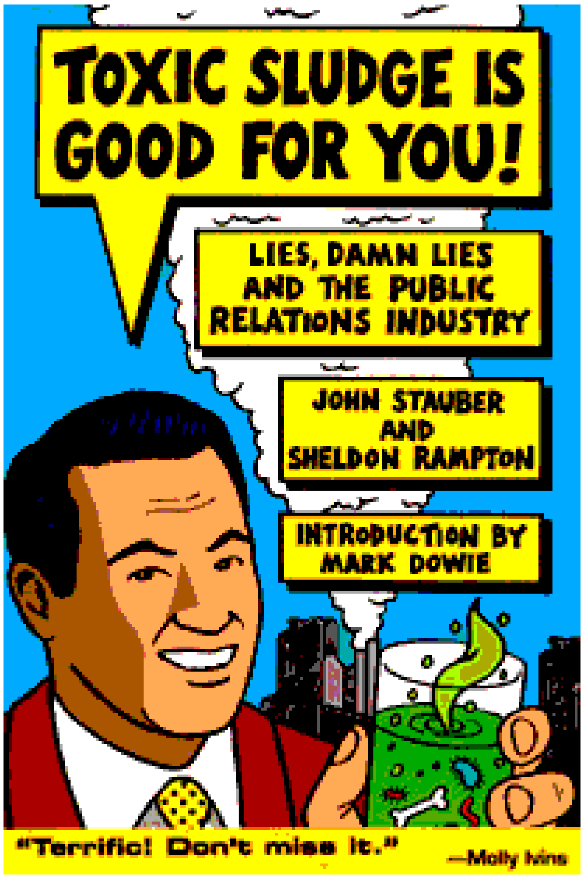

Figure 6: Toxic Emissions