1 Introduction

1.1 Random variation

Consider the following experiments:

- Counting: e.g. count the number of mice in a litter to learn about how many offspring mice tend to have;
- Measurement: e.g. measure the heights of ten students in order to learn about how tall students tend to be.

If we were to repeat the experiment, the results would not be exactly the same, for example, because we may choose different students or mice, we may make unpredictable errors in our measurements, or there may be other variation which affects our results in an unpredictable way. We say that such experiments are subject to stochastic (or "random") variation which means the outcome of the experiment cannot be predicted exactly. It is for precisely this reason that we could not expect our observation of the number of mice in one litter, for example, to reveal everything about litter size. On the other hand, if we counted the number of mice in a *sample* of litters, we would start to learn about what litter sizes are more or less likely. Statistics deals with problems of collecting, classifying, summarising and analysing data to enable us to make inferences from the sample data to the population from which the sample has been taken. It does this by providing a language and framework for the study of random variation, allowing us to distinguish it from real trends, differences or signal.

1.2 Fundamental definitions

Any quantity whose value is subject to random variation is called a *random variable*. The weight of individual students in a lecture room, number of policies sold by individual insurance companies in a day, or hormone concentration in the bloodstream over a series of time points are all examples. Each measured value is called an *observation* and we generally refer to any activity which involves making an observation as an *experiment*. For example, "weight of student" is a random variable while "John Smith weighs 65kg" is an observation. Random variables are often denoted by upper case letters such as X. Observations are often denoted by the corresponding lower case letters such as x.

Usually we will not be able to take a complete set of measurements (e.g. weigh every student or measure hormone concentration continuously) but instead we observe a *sample*. The term *population* is used to refer to the collection of individuals or objects from which the sample is drawn, or sometimes the set of all possible measurements they represent. Statistical methods are used to make *inferences* about the population from (the limited information in) the sample and so we want samples to be representative of the population. We therefore usually take *random samples* which are samples with the following properties:

- All members of the population are equally likely to be selected for inclusion, and
- All combinations of members are also equally likely.

1.3 Data

Data are either *qualitative* (e.g. colour, shape) or *quantitative* (e.g. concentration, number of offspring).

A qualitative variable is usually categorical. Sometimes the categories do not have a natural ordering (e.g. gender) and in this case we call the variables *nominal*. In other cases the categories do have a natural ordering (e.g. degree of severity of a disease) and we call the variables *ordinal*.

A quantitative variable takes numerical values that are either *discrete* (e.g. number of online purchases made by a customer in a week, number of cells on a Petri dish) or *continuous* (e.g. dry matter weight of a plant, concentration of a drug). Continuous random variables adopt a smooth range of values, while discrete random variables can only adopt a countable number of values.

1.4 Sample data sets

1.4.1 Survival time data

The data below are the survival times of patients suffering from chronic granulocytic leukaemia, measured in days from the time the patient was diagnosed:

7	47	58	74	177	232	273	285	317	429
440	445	455	468	495	497	532	571	579	581
650	702	715	779	881	900	930	968	1077	1109
1314	1334	1367	1534	1712	1784	1877	1886	2045	2056
2260	2429	2509							

Source: Bryson, M. C. and Siddiqui, M. M. (1969) Survival times: some criteria for aging. *Journal of the American Statistical Association*, **64**, 1472-1483.

1.4.2 Cell viability data

The yeast species *Saccharomyces cerevisiae* has been used in baking and fermenting alcoholic beverages for thousands of years. It is also extremely important as a model organism in modern cell biology research, and is the most thoroughly researched eukaryotic[1] microorganism. Researchers can use it to gather information into the biology of the eukaryotic cell and ultimately human biology.

In investigating lifespan in mutant yeast strains, 100 cell cultures were monitored. The number of cultures which were still viable after 14 days was recorded.

This entire procedure was repeated 40 times, and the following counts were obtained:

88	87	85	91	93	91	94	87	90	91	92	87	91	89
87	90	88	85	90	92	89	86	91	92	91	91	93	93
87	90	91	91	89	90	90	91	91	93	92	85		

[1] *Eukaryotes* are organisms whose cells contain a distinct membrane-bound nucleus.

1.4.3 Cereal purchases data

The data below show the *frequency distribution* for numbers of packets of cereal purchased over 13 weeks by 2000 customers:

No. packets	0	1	2	3	4	5	6	7	8	9	10	11	12	
Frequency	1149	199	129	87	71	43	49	46	44	24	45	22	23	
No. packets	13	14	15	16	17	18	19	20	21	22	23	24	25	
Frequency	33	8	2	7	2	3	1	2	0	0	1	0	1	
No. packets	26	27	28	29	30	31	32	33	34	35	36	37	38	39
Frequency	3	2	0	1	1	0	0	0	0	0	0	0	0	1
No. packets	40	41	42	43	44	45	46	47	48	49	50	51	52	>52
Frequency	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Source: Barnett, V. and Lewis, T. (1984) *Outliers in statistical data.* Chichester: John Wiley & Sons, Table 1.1.

1.4.4 New York air quality data

The full New York air quality data set comprises daily air quality measurements in New York on the 154 consecutive days between May 1st 1973 and September 30th 1973. It can be viewed in R using the commands

```
data(airquality) # Loads the air quality data which is one of R's
                 # built-in data sets.
airquality # Prints the full data set to the screen.
```

More information on the data set can be accessed by typing

```
?airquality # Shows the R documentation for the data set.
```

The variables are the mean ozone in parts per billion, solar radiation in Langleys, average wind speed in miles per hour, maximum daily temperature in degrees Fahrenheit, and the month and day within that month. The data set contains some missing values and these are indicated by NA. Data from the first 40 days are shown below:

Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8.0	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
NA	NA	14.3	56	5	5
28	NA	14.9	66	5	6
23	299	8.6	65	5	7
19	99	13.8	59	5	8
8	19	20.1	61	5	9
NA	194	8.6	69	5	10
7	NA	6.9	74	5	11
16	256	9.7	69	5	12
11	290	9.2	66	5	13
14	274	10.9	68	5	14
18	65	13.2	58	5	15
14	334	11.5	64	5	16
34	307	12.0	66	5	17
6	78	18.4	57	5	18
30	322	11.5	68	5	19
11	44	9.7	62	5	20
1	8	9.7	59	5	21
11	320	16.6	73	5	22
4	25	9.7	61	5	23
32	92	12.0	61	5	24
NA	66	16.6	57	5	25
NA	66	16.6	57	5	25
NA	266	14.9	58	5	26
NA	NA	8.0	57	5	27
23	13	12.0	67	5	28
45	252	14.9	81	5	29
115	223	5.7	79	5	30
37	279	7.4	76	5	31
NA	286	8.6	78	6	1
NA	287	9.7	74	6	2
NA	242	16.1	67	6	3
NA	186	9.2	84	6	4
NA	220	8.6	85	6	5
NA	264	14.3	79	6	6
29	127	9.7	82	6	7
NA	273	6.9	87	6	8
71	291	13.8	90	6	9

Source: Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) *Graphical Methods for Data Analysis.* Belmont, CA: Wadsworth.

Exercise:
For the example data sets, define what type of variables they are: ✎

Data set	Quantitative/Qualitative	Continuous/Discrete
Survival time data		
Cell viability data		
Cereal purchases data		
Air quality **Ozone**		
Air quality **Wind**		
Air quality **Month**		