Overview and Descriptive Statistics
Outline

1. Populations, Samples and Processes
2. Pictorial and Tabular Methods in Descriptive Statistics
3. Measures of Location
4. Measures of Variability
Outline

1. Populations, Samples and Processes
2. Pictorial and Tabular Methods in Descriptive Statistics
3. Measures of Location
4. Measures of Variability
1. Populations, Samples and Processes

2. Pictorial and Tabular Methods in Descriptive Statistics

3. Measures of Location

4. Measures of Variability
Outline

1. Populations, Samples and Processes
2. Pictorial and Tabular Methods in Descriptive Statistics
3. Measures of Location
4. Measures of Variability
Statistics is used to

*make intelligent judgements and informed decisions in the presence of uncertainty and variation (text, p. 1)*

In more detail, some of the uses of statistics are:

1. Designing experiments to collect data.
2. Extracting information from data.
3. Making decisions and predictions in the presence of uncertainty and variation.

*Descriptive statistics* summarize or describe important features of data, either graphically or numerically.
1.1 Populations, samples and processes

**Population** - a well-defined collection of objects of interest in the study. It can be a real population (e.g. 18-24 year olds in the U.S.A.) or a virtual population (e.g. all parts that could be produced by a particular machine at particular settings).

**census** - a complete enumeration of a population

**sample** - a subset of the population selected in some prescribed manner.

**variable** - a characteristic of the objects chosen. Often this is a numeric measurement (e.g. number of defects, weight of product produced) but is can also be a category (e.g. male/female).

Data results are sometimes characterized according to the number of variables measured on each object. Names are *univariate* (one variable only), *bivariate* (two variables), *multivariate* (more than two).
Descriptive statistics

- Numerical or graphical representations of data or important features of data.
- For the most part created by computer packages. We will use R, which is similar to S-PLUS. Other statistical packages are Minitab, SAS, SPSS. Many people use spreadsheets such as Excel.

Challenger example  In 1986 the space shuttle Challenger exploded shortly after launch. The cause was failure of several O-rings to seal the fuel tanks properly. This launch was done at colder weather than any previous launch. Example 1.1 (p. 4) gives the temperature (°F) for each test firing or actual launch including the launch that failed.
Challenger example

> str(xmp01.01)

'data.frame': 36 obs. of 1 variable:
$ temp: int 84 49 61 40 83 67 45 66 70 69 ...

> with(xmp01.01, summary(temp))

       Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
63.000   59.500   67.500   65.862   75.000   84.000

> with(xmp01.01, stem(temp))

The decimal point is 1 digit(s) to the right of the |  
3 | 1  
4 | 059  
5 | 23788  
6 | 01136777789  
7 | 000023556689  
8 | 0134  

The temperature at the Challenger launch was 31°F.
Descriptive statistics by themselves provide some information but do not provide conclusions. **Inferential statistical methods** allow us to draw conclusions from data.

Because of the variation in the data, we cannot draw guaranteed conclusions. We must phrase the conclusions as probabilistic statements. (e.g. 95% confidence interval on the median strength).

We begin the course with a short section on descriptive statistics, then discuss probability and probability distributions, then inferential statistics.
Data collection

Simple random sample  If we have a well-defined population from which we are sampling (i.e. an enumerative study), a simple random sample consists of a sample drawn in such a way that all possible samples of the same size have equal probability of being drawn.

Stratified sample  Occasionally we may know that certain subsets of the population are more variable than others and we use different sampling fractions for different groups. Political polls are sometimes conducted like this.

Convenience sample  It is tempting to use a convenience sample (the ones that are most easily available). Doing so affects the population over which inferences can be drawn.
Data collection (cont’d)

Designed experiment  Many engineering studies are conducted as designed experiments, in which the conditions are manipulated. Changes observed in a designed experiment can be attributed to the condition being changed. In an observational study we only know that there is correlation between condition and response, we cannot ascribe causality.

Notation  Generally we use \( n \) for the number of observations in a data set and letters like \( x \) and \( y \) for the data values themselves. The values, in the order that we received them, are indicated by subscripts, as in \( x_1, x_2, \ldots, x_n \). The sorted data values, called the order statistics are written \( x(1) \leq x(2) \leq \cdots \leq x(n) \).
1.2 Pictorial Methods

Stem-and-Leaf displays: A form of hand-drawn histogram invented by John Tukey that preserves some of the information about numeric data values by using a digit for the “leaves” on each “stem”.

Dotplots: Simple pictorial representation of numeric data.

Histograms: A bar plot approximating the density of discrete or continuous numeric data.

Probability plot: Not mentioned until chapter 4. A normal probability plot gives an indication of whether univariate data could reasonably be assumed to follow a normal (Gaussian) distribution, also called “the bell curve”.

Box-and-whisker plot: A pictorial representation of a “five number summary” (minimum, 1st quartile, median, 3rd quartile, maximum). Good for comparing results from different groups.
Golf course yardages

Example 1.6 gives data from *Golf Magazine* on the total yardage of a random sample of golf courses.

```r
> with(xmp01.06, summary(yardage))

          Min. 1st Qu.  Median    Mean 3rd Qu.   Max. 
6433    6674    6872    6845    7042    7280
```

```r
> with(xmp01.06, stem(yardage))

The decimal point is 2 digit(s) to the right of the |
  64 | 3467
  65 | 1338
  66 | 119
  67 | 015779
  68 | 05779
  69 | 0034
  70 | 112455
  71 | 113777
  72 | 18
```
Histograms

Discrete data a histogram of discrete data is a bar plot giving the frequency or relative frequency of the possible values. See example 1.9, pp. 16–17.

Continuous data a histogram of continuous data is constructed by dividing the responses into bins or class intervals - usually of equal width. The choice of the number of bins and their end points can affect the visual appearance of the histogram and, possibly, conclusions drawn from it. With large data sets a density plot is more effective than a histogram.

Example Power consumption data from Wisconsin Power and Light.
Unequal bin widths

It is possible, though unusual, to use different widths for the bins. In such cases it is important to have the area of each bar proportional to the frequency or to the relative frequency. Failure to do so is deceptive.

Example The corrosion data in Example 1.11 (p. 20) from J. of Structural Engineering is badly skewed to the right. A histogram is more informative if we use unequal intervals or take a transformation of the data.
In a *Categorical data* set we only observe the category of the response. It does not correspond to a numerical measurement.

**Example** In exercise 1.29 (p. 27) 120 observations of the type of health complaint (J = joint swelling, F = fatigue, B = back pain, M = muscle weakness, C = coughing, N = nose running/irritation, O = other) are given. There is no sense in which these are ordered.

**Use in R** In *R* a *factor* is used to represent categorical data. An *ordered factor* is a special type of factor in which the levels have a natural order.

**Plotting** A bar chart or bar plot is a convenient way to plot such data
Bar plot for health complaints

> with(ex01.29, barplot(table(complaint), horiz = TRUE))
Sorted bar plot for health complaints

For data like these where the categories do not have a natural order we can order the categories according to their frequency of occurrence.

```r
> with(ex01.29, barplot(sort(table(complaint)), horiz = TRUE))
```
Multivariate data

The most common plot for bivariate data is a *scatterplot*.

```r
> xyplot(OSA ~ palprebal, xmp12.01, type = c("g", "p", +   "r"))
```

Data from Example 12.1, page 490
Measures of location

Sample mean - the average data value.

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

If each data point was a point mass on a number line, then \( \bar{x} \) would be the balance point. The population mean, \( \mu \), is the average over the whole population.

Median

The Sample Median is the “middle” data value. When the data are skewed, this is more representative than the sample mean. Recall that the sorted data values, called the order statistics are written \( x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \). If \( n \) is odd, the median is \( x_{(\lfloor n+1/2 \rfloor)} \), otherwise \( \left[ x_{(n/2)} + x_{(n/2+1)} \right] / 2 \).
Quartiles, percentiles, and quantiles

**Quartiles** Just as the median divides the sorted data in half, the first and third quartiles divide the data into quarters.

**Percentiles and quantiles**

*Definition* The $p$'th quantile ($0 < p < 1$) is any value such that a proportion $p$ of the data is below it. A percentile is the same but expressed as a percentage rather than a fraction.

*Common example:* Scores on standardized tests such as the SAT are often converted to percentiles. “Class rank” is a similar measure.

*Calculation:* Use interpolation on the order statistics. (See `?quantile` in R).
Trimmed means

A sample mean is strongly influenced by outliers. A median is robust against outliers but does not use the data effectively. (Sampling variation of medians is greater than sampling variation of means.) A compromise that is sometimes used is a trimmed mean where some fraction of the data at the ends is dropped and the remaining data averaged.

See the trim argument in ?mean in R and check

```r
> example(xmp01.15)

x01.15> with(xmp01.15, summary(lifetime))

     Min.  1st Qu.   Median      Mean  3rd Qu.     Max.  
 612.00   894.20  1010.00    965.00 1086.00 1201.00
x01.15> with(xmp01.15, mean(lifetime, trim = 0.1))
[1] 979.125

x01.15> with(xmp01.15, mean(lifetime, trim = 0.2))
[1] 999.9167

x01.15> with(xmp01.15, dotchart(lifetime))
```
Summary of categorical data

Use the sample proportions.

```r
> (freq <- xtabs(~complaint, ex01.29))

complaint
  B  C  F  J  M  N  O
  7  3  9 10  4  6 21

> freq/sum(freq)

complaint
  B  C  F  J  M  N  O
  0.11666667 0.05000000 0.15000000 0.16666667 0.06666667 0.10000000 0.00000000

  0  0.35000000
```
Sample Variance and standard deviation

The *Sample Variance* tells us how far away the data points are from the center of the data. That is, it measures spread or dispersion.

\[ s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n - 1) \]

The *Sample Standard Deviation* is the square root of sample variance. It has the same units as the original data and the sample mean.

\[ s = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n - 1)} \]
The variance and standard deviation are not affected by shifting the data by a constant but they are affected by scaling the data. If the data are $x_1, x_2, \ldots, x_n$ and $c$ is a constant then

1. If $y_1 = x_1 + c, y_2 = x_2 + c, \ldots, y_n = x_n + c$, then $s_y^2 = s_x^2$
2. If $y_1 = cx_1, y_2 = cx_2, \ldots, y_n = cx_n$, then $s_y^2 = c^2 s_x^2$ and $s_y = |c| s_x$.

Notice the absolute value in the last expression. You cannot have a negative variance or a negative standard deviation.
We have seen several examples of boxplots. The box spans the “middle half” of the data - the region from the first quartile to the third quartile. The whiskers extend to the minimum and maximum unless these appear to be outliers based on the *inter-quartile range*.

```r
> bwplot(~C1, xmp01.19, xlab = "Pulse widths")
```

Boxplots show us location, scale, outliers, and symmetry.
Comparative boxplots

When we have two or more groups of observations of a numeric variable, *comparative boxplots* allow us to examine differences between the groups in location, scale, or symmetry.

> bwplot(type ~ strength, xmp10.01)

Data from Example (compare Figure 10.1, p. 405)