Informatics 1B, 2008

School of Informatics, University of Edinburgh

Data and Analysis

Note 12

Statistical Analysis of Data I

Alex Simpson

Part III — Unstructured Data

Analysis of data

In the absence of structure, often have to *analyse* data.

Typical goals of analysis:

- Discover implicit structure in the data.

  E.g., find patterns in empirical data (such as experimental data).

- Confirm or refute a hypothesis about the data.

  E.g., confirm or refute an experimental hypothesis.

N.B. It is often useful to analyse structured and semistructured data too. The kinds of analysis we consider will be applicable irrespective of whether or not there is pre-existing structure on the data.

Data scales

The type of analysis performed (obviously) depends on:

- The reason for wishing to carry out the analysis.
- The type of data to hand.

  For example, the data may be *quantitative* (i.e., numerical), or it may be *qualitative* (i.e., descriptive).

One important aspect of the kind of data is the form of *data scale* it belongs to:

- *Categorical* (or *nominal*) scale (for qualitative data).
- *Ordinal* scale (between qualitative and quantitative).
- *Interval and ratio* scales (for quantitative data).

This affects the ways in which we can manipulate data.

Categorical scales

Data belongs to a *categorical scale* if each *datum* (i.e., data item) is classified as belonging to one of a fixed number categories.

Example: The British Government (presumably) classifies Visa applications according to the nationality of the applicant. This classification is a categorical scale: the categories are the different possible nationalities.

Example: Insurance companies classify some insurance applications (e.g., home, possessions, car) according to the postcode of the applicant (since different postcodes have different risk assessments).

Categorical scales are sometimes called *nominal scales*, especially in cases in which the value of a datum is a name.

Ordinal scales

Data belongs to an *ordinal scale* if it has an associated ordering but arithmetic transformations on the data are not meaningful.

Example: The *Beaufort wind force scale* classifies wind speeds on a scale from **0** (calm) to **12** (hurricane). This has an obvious associated ordering, but it does not make sense to perform arithmetic operations on this scale. E.g., it does not make much sense to say that scale **6** (strong breeze) is the average of calm and hurricane force.

Example: In many institutions, exam marks are recorded as grades (e.g., A,B,. . . , G) rather than as marks. Again the ordering is clear, but one does not perform arithmetic operations on the scale.

Interval scales

An *interval scale* is a numerical scale (usually with real number values) in which we are interested in *relative value* rather than *absolute value*.

Example: Points in time are given relative to an arbitrarily chosen zero point. We can make sense of comparisons such as: moment x is 2008 years later than moment y. But it does not make sense to say: moment x is twice as large as moment z.

Mathematically, interval scales support the operations of subtraction (returning a real number for this) and weighted average.

Interval scales do not support the operations of addition and multiplication.

Ratio scales

A *ratio scale* is a numerical scale (again usually with real number values) in which there is a notion of *absolute value*.

Example: Most physical quantities such as mass, energy and length are measured on ratio scales. So is temperature if measured in kelvins (i.e. relative to absolute zero).

Like interval scales, ratio scales support the operations of subtraction and weighted average. They also support the operations of addition and of multiplication by a real number.

Question for physics students: Is time a ratio scale if one uses the Big Bang as its zero point?

Visualising data

It is often helpful to *visualise* data by drawing a *chart* or plotting a *graph* of the data.

Visualisations can help us guess properties of the data, whose existence we can then explore mathematically using statistical tools.

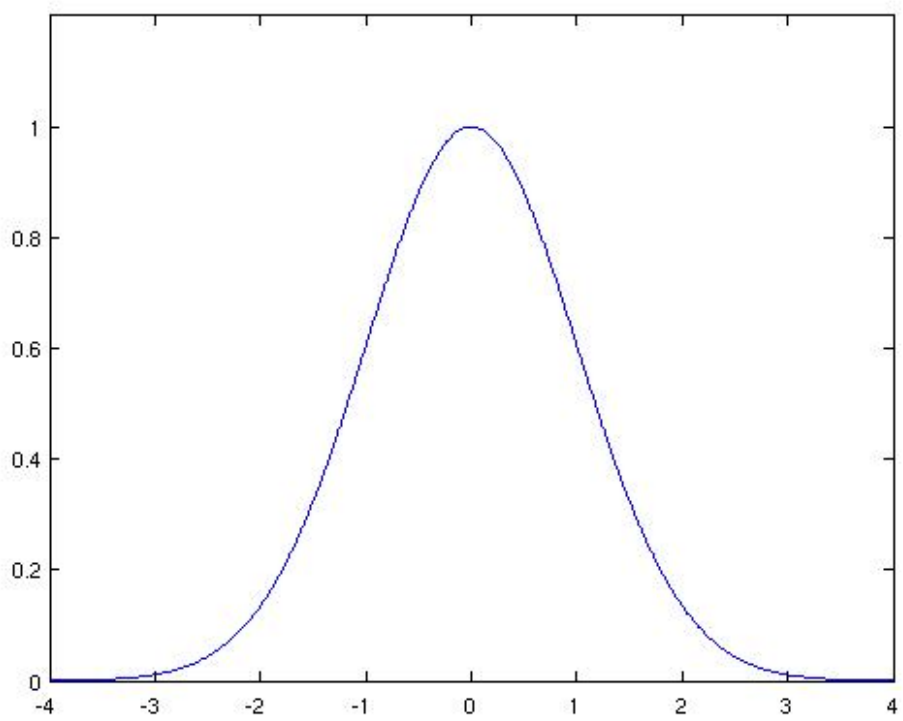
For a collection of data of a categorical or ordinal scale, a natural visual representation is a *histogram* (or *bar chart*), which, for each category, displays the number of occurrences of the category in the data.

For a collection of data from an interval or ratio scale, one plots a *graph* with the data scale as the x-axis and the frequency as the y-axis.

It is very common for such a graph to take a bell-shaped appearence.

Normal distribution

In a *normal distribution*, the data is clustered symmetrically around a central value (zero in the graph below), and takes the bell-shaped appearance below.



Normal distribution (continued)

There are two crucial values associated with the normal distribution.

The *mean*, μ, is the central value around which the data is clustered. In the example, we have $\mu = 0$.

The *standard deviation*, σ, is the distance from the mean to the point at which the curve changes from being *convex* to being *concave*. In the example, we have $\sigma = 1$. The larger the standard deviation, the larger the *spread* of data.

The general equation for a normal distribution is

$$y = c\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(You do not need to remember this formula.)

Statistic(s)

A *statistic* is a numerical value that captures some property of data.

For example, the mean of a normal distribution is a statistic that captures the value around which the data is clustered.

Similarly, the standard deviation of a normal distribution is a statistic that captures the degree of spread of the data around its mean.

The notion of *mean* and *standard deviation* generalise to data that is not normally distributed.

There are also other, *mode* and *median*, which are alternatives to the mean for capturing the "focal point" of data.

Mode

Summary statistics summarise a property of a data set in a single value.

Given data values $x_1, x_2, \ldots, x_N$, the *mode* (or *modes*) is the value (or values) x that occurs most often in $x_1, x_2, \ldots, x_N$.

Example: Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, the mode is 6, which is the only value to occur three times.

The mode makes sense for all types of data scale. However, it is not particularly informative for real-number-valued quantitative data, where it is unlikely for the same data value to occur more than once. (And, anyway, is it meaningful to test two real-number values for equality?)

Median

Given data values $x_1, x_2, \ldots, x_N$, written in non-decreasing order, the *median* is the middle value $x_{(\frac{N+1}{2})}$ assuming N is odd. If N is even, then any data value between $x_{(\frac{N}{2})}$ and $x_{(\frac{N}{2}+1)}$ inclusive is a possible *median*.

Example: Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, we write this in non-decreasing order:

$$1, 1, 2, 2, 3, 5, 5, 6, 6, 6, 7$$

The middle value is the sixth value 5.

The median makes sense for ordinal data and for interval and ratio data. It does not make sense for categorical data, because categorical data has no associated order.

Mean

Given data values $x_1, x_2, \ldots, x_N$, the *mean* μ is the value:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Example: Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, the mean is

$$\frac{6+2+3+6+1+5+1+7+2+5+6}{11} = 4.$$

Although the formula for the mean involves a sum, the mean makes sense for both interval and ratio scales. The reason it makes sense for data on an interval scale is that interval scales support *weighted averages*, and a mean is simply an equally-weighted average (all weights are set as $\frac{1}{N}$).

The mean does *not* make sense for categorical and ordinal data.

Variance and standard deviation

Given data values $x_1, x_2, \ldots, x_N$, with mean μ, the *variance*, written $\mathbf{Var}$ or σ^2, is the value:

$$\mathbf{Var} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

The *standard deviation*, written σ, is defined by:

$$\sigma = \sqrt{\mathbf{Var}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}}$$

Like the mean, the standard deviation makes sense for both interval and ratio data. (The values that are squared are real numbers, so, even with interval data, there is no issue about performing the multiplication.)

Variance and standard deviation (example)

Given data: $\mathbf{6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6}$, we have $\boldsymbol{\mu = 4}$.

$$\begin{aligned}
\mathbf{Var} &= \frac{\mathbf{2^2 + 2^2 + 1^2 + 2^2 + 3^2 + 1^2 + 3^2 + 3^2 + 2^2 + 1^2 + 2^2}}{\mathbf{11}} \\
&= \frac{\mathbf{4 + 4 + 1 + 4 + 9 + 1 + 9 + 9 + 4 + 1 + 4}}{\mathbf{11}} \\
&= \frac{\mathbf{50}}{\mathbf{11}} \\
&= \mathbf{4.55} \quad \text{(to 2 decimal places)}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\sigma} &= \sqrt{\frac{\mathbf{50}}{\mathbf{11}}} \\
&= \mathbf{2.13} \quad \text{(to 2 decimal places)}
\end{aligned}$$

Several variables

Often, one wants to relate data in several variables (i.e., multi-dimensional data).

For example, the table below tabulates, for eight students (A–H), their weekly time (in hours) spent: studying for Data & Analysis, drinking and eating. This is juxtaposed with their Data & Analysis exam results.

	A	B	C	D	E	F	G	H
Study	0.5	1	1.4	1.2	2.2	2.4	3	3.5
Drinking	25	20	22	10	14	5	2	4
Eating	4	7	4.5	5	8	3.5	6	5
Exam	16	35	42	45	60	72	85	95

Thus, we have four variables: study, drinking, eating and exam. (This is four-dimensional data.)

Correlation

We can ask if there is any *relationship* between the values taken by any two variables.

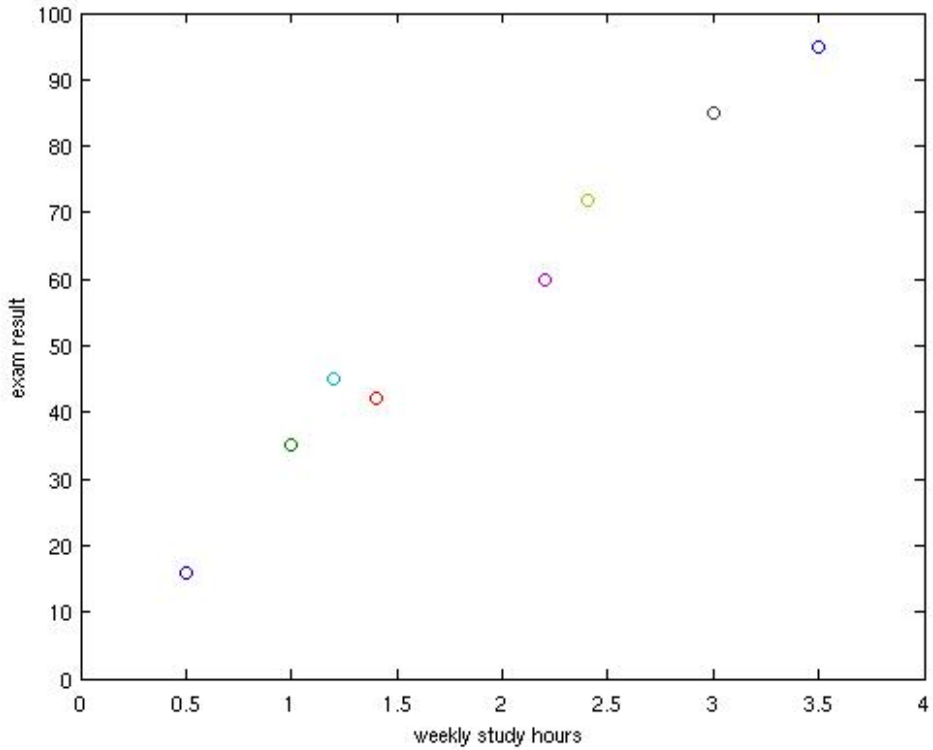
If there is no relationship, then the variables are said to be *independent*.
If there is a relationship, then the variables are said to be *correlated*.

Example: Is there a correlation between study hours and exam results? What about between drinking hours and exam results? What about eating and exam results?

How do we detect whether there is correlation?

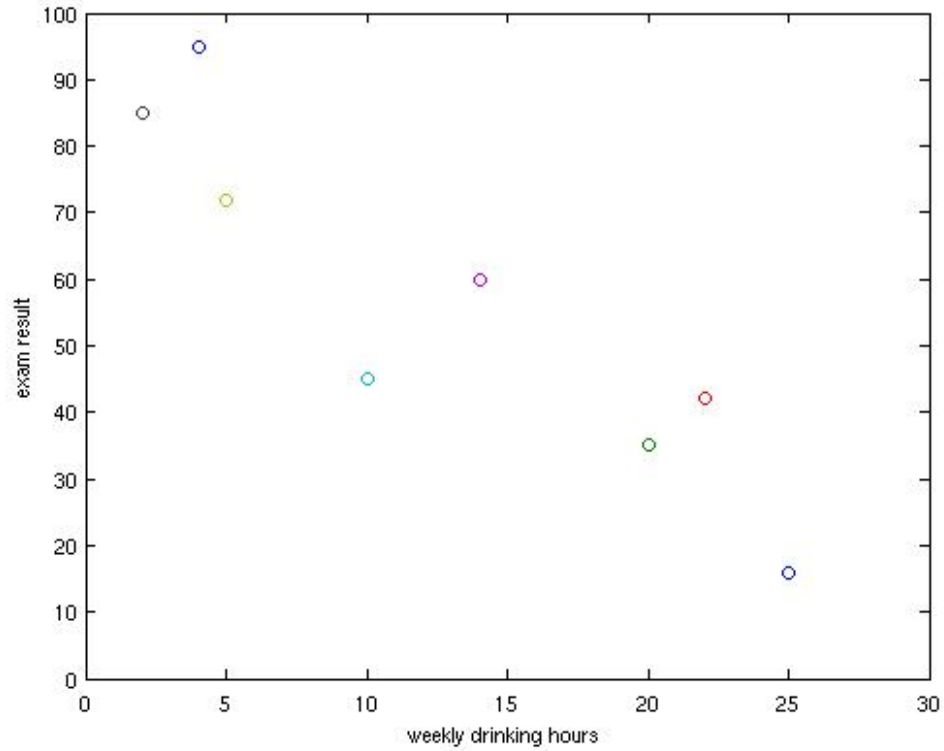
A simple visual guide is to draw a *scatter plot* using one variable for the x-axis and one for the y-axis.

Studying vs. exam results



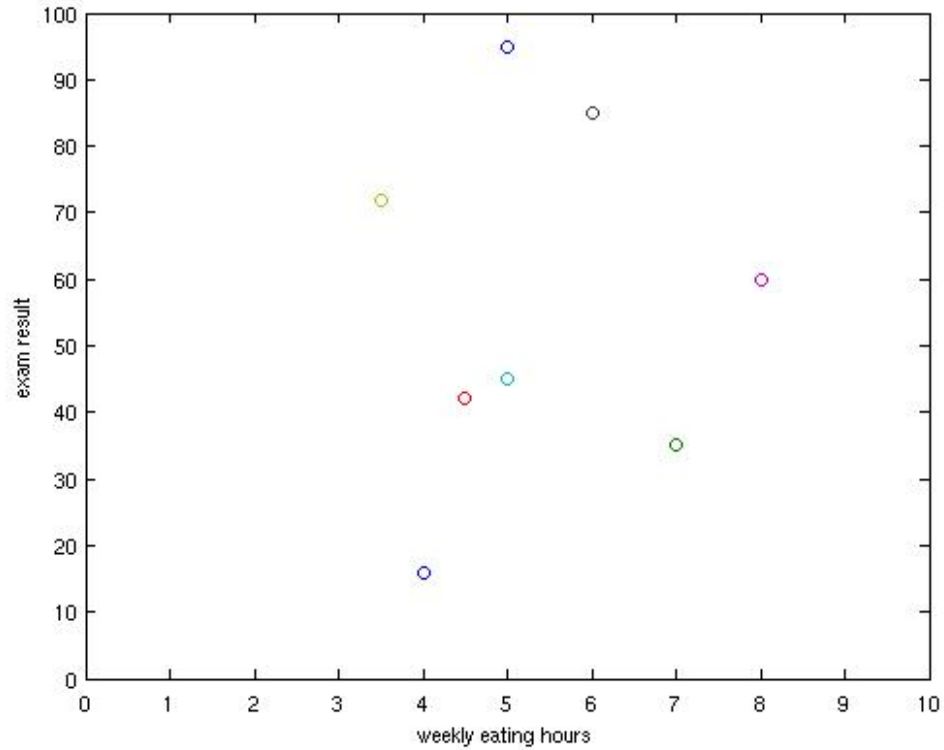
This looks like a *positive* correlation.

Drinking vs. exam results



This looks like a *negative* correlation.

Eating vs. exam results



There is no obvious correlation.

Pearson's correlation coefficient (Not examinable!)

To investigate whether the data values $x_1, \ldots, x_N$ are correlated with $y_1, \ldots, y_N$.

Let μ_x and σ_x be the mean and standard deviation of the x values.
Let μ_y and σ_y be the mean and standard deviation of the y values.

Pearson's corelation coefficient, $r_{x,y}$, is defined by:

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}$$

If $r_{x,y}$ is close to 1 then x, y are *positively correlated*.
If $r_{x,y}$ is close to -1 then x, y are *negatively correlated*.
If $r_{x,y}$ is close to 0 then there is no correlation.

Correlation subtleties

Causality: A correlation does *not* imply a *causal relationship* between one variable and another. For example, there is a positive correlation between incidences of lung cancer and time spent watching television, but neither causes the other.

However, in cases in which there *is* a causal relationship between two variables, then there often will be an associated correlation between the variables.

Linearity: The Pearson correlation coefficient measures how close a scatter plot of $\boldsymbol{x}, \boldsymbol{y}$ values is to a straight line. Nonetheless, a high correlation does not mean that the relationship between $\boldsymbol{x}, \boldsymbol{y}$ is linear. It just means it can be reasonably closely approximated by a linear relationship.