

# Informatics 1: Data & Analysis

## Lecture 17: Data Scales and Summary Statistics

Ian Stark

School of Informatics  
The University of Edinburgh

Tuesday 17 March 2015  
Semester 2 Week 9

## Data Retrieval

- The **information retrieval** problem
- The **vector space model** for retrieving and ranking

## Statistical Analysis of Data

- Data scales and summary statistics
- Hypothesis testing and correlation
- $\chi^2$  tests and collocations also *chi-squared*, pronounced “kye-squared”

# Analysis of Data

There are many reasons to **analyse** data. For example:

- To **discover** implicit structure in the data;  
e.g., finding patterns in experimental data which might in turn suggest new models or experiments.
- To **confirm** or **refute** a hypothesis about the data.  
e.g., testing a scientific theory against experimental results.

Mathematical **statistics** provide a powerful toolkit for performing such analyses, with wide and effective application.

# Analysis of Data

Mathematical **statistics** provide a powerful toolkit for performing such analyses, with wide and effective application.

This analytic strength is twofold:

- Statistics can sensitively detect information not immediately apparent within a mass of data;
- Statistics can help determine whether or not an apparent feature of data is really there.

Machine assistance is essential for large datasets, and enables otherwise infeasible **resampling** techniques such as *bootstrapping* and *jackknifing*.

There are lots of books for learning about statistics. Here are two, intended to be approachable introductions without requiring especially strong mathematical background.



P. Hinton.

*Statistics Explained: A Guide for Social Science Students.*  
Routledge, third edition, 2014.



D. B. Wright and K. London.

*First (and Second) Steps in Statistics.*  
SAGE Publications Ltd, second edition, 2009.

Here are two more books, for finding out about how statistics are used and abused. Both are easy reading. The second has amusing pictures, too.



M. Blastland and A. Dilnot.

<http://is.gd/tigerisnt>

*The Tiger That Isn't: Seeing Through a World of Numbers.*  
Profile, 2008.

“Makes statistics far, far too interesting”



D. Huff.

*How to Lie with Statistics.*  
W. W. Norton, 1954.

<http://is.gd/huffbook>

“The most widely read statistics book in the history of the world”

# Data Scales

What type of statistical analysis we might apply to some data depends on:

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

Data may be *qualitative* (descriptive) or *quantitative* (numerical).

We can refine this further into different kinds of *data scale*:

- Qualitative data may be drawn from a *categorical* or an *ordinal* scale;
- Quantitative data may lie on an *interval* or a *ratio* scale.

Each of these supports different kinds of analyses.

# Categorical Scales

Data on a **categorical scale** has each item of data being drawn from a fixed number of categories.

## Example: Categorical Scale

A government might classify visa applications from people wishing to visit according to the nationality of the applicant. This classification is a categorical scale: the categories are all the different possible nationalities.

## Example: Categorical Scale

Insurance companies classify some insurance applications (e.g., home, possessions, car) according to the alphanumeric postcode of the applicant, making different risk assessments for different postcodes. Here the categories are all existing postcodes.

Categorical scales are sometimes called *nominal*, particularly where the categories all have names.



# Ordinal Scales

Data on an **ordinal scale** has a recognized ordering between data items, but there is no meaningful arithmetic on the values.

## Example: Ordinal Scale

The *European Credit Transfer and Accumulation System* (ECTS) has a grading scale where course results are recorded as A, B, C, D, E, FX and F. There are no numerical marks. The ordering is clear, but we can't add or subtract grades.

## Example: Ordinal Scale

The *Douglas Sea Scale* classifies the state of the sea on a scale from 0 (glassy calm) through 5 (rough) to 9 (phenomenal). This is ordered, but it makes no sense to perform arithmetic: 4 (moderate) is not the mean of 2 (smooth) and 6 (very rough).

# 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: Interval Scale

Moments in time are given relative to an arbitrarily chosen zero point. We can make sense of comparisons such as “date X is 17 years later than date Y”. But it does not make sense to say “arrival time P is twice as large as departure time Q”.

## Example: Interval Scale

The Celsius and Fahrenheit temperature scales are interval scales, as the choice of zero is externally imposed.

Mathematically, interval scales support the operations of subtraction and average (all kinds, possibly weighted).

Interval scales do not support either addition or 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: Ratio Scales

Most physical quantities such as mass, energy and length are measured on ratio scales. The Kelvin temperature scale is a ratio scale. So is age (of a person, for example), even though it is a measure of time, because there is a definite zero origin.

Thus one object can have twice the mass of another; or one person can be half as old as someone else.

Like interval scales, ratio scales support subtraction and weighted averages. They also support addition and multiplication by a real number (a *scalar*).

# Summary of Scales

Name	Description	Example
Categorical	Qualitative, fixed set of categories, no order, no possible arithmetic.	Postcodes
Ordinal	Qualitative, fixed set of categories, can be ordered, still no arithmetic.	Exam grades
Interval	Quantitative, values all relative; can take averages, subtract one value from another; no addition or multiplication.	Dates
Ratio	Quantitative, absolute values, can take averages, subtract, add, and take scalar multiples of values.	Mass, energy

# Visualising data

It is often helpful to visualise data by drawing a **chart** or plotting a **graph** of the data. Visualisations may suggest possible properties of the data, whose existence and features we can then explore mathematically with statistics.

What kind of visualisations are possible depends on the kind of data.


For a data on a categorical or ordinal scale, a natural visual representation is a *bar chart*, displaying for each category the number of times it occurs in the data.


Bars in a bar chart are all the same width, and separate.

For data from an interval or ratio scale, we can collect data into bands and draw a *histogram*, giving the frequency with which values occur in the data.

In a histogram the bars are adjacent, and can be of different widths: it is their area, not height, which measures the number of values present.

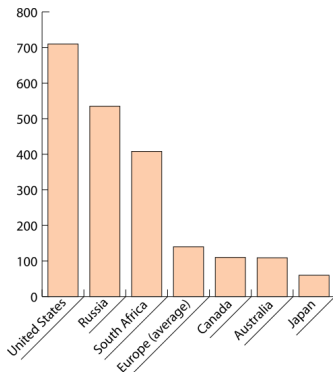


 **Pie I have eaten**

 **Pie I have not yet eaten**

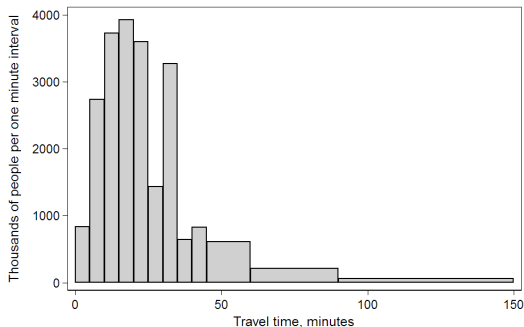
# Bar Chart vs. Histogram

This is a **bar chart**



Credit: Wikipedia, user XcepticZP

This is a **histogram**



Credit: Wikipedia, user Qwfp

Far out in the uncharted backwaters of the unfashionable end of the Western Spiral arm of the Galaxy lies a small unregarded yellow sun.

Orbiting this at a distance of roughly ninety-two million miles is an utterly insignificant little blue green planet. . .

Douglas Adams: The Hitch-Hiker's Guide to the Galaxy





J. Richard Gott III.

*A Grim Reckoning — What has a 16th-century astronomer got to do with the defeat of governments and the possible extinction of the human race?*

New Scientist, 15 November 1997

<http://is.gd/grimreckoning>



Timothy Ferris.

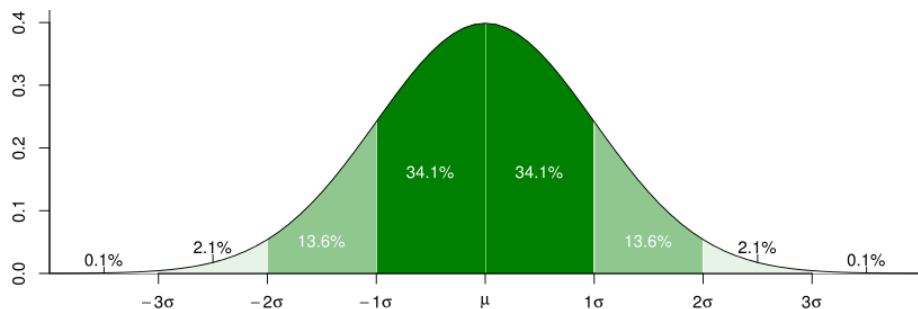
*How to Predict Everything*

The New Yorker, 12 July 1999, pp. 35–39

# Normal Distribution

In the *normal distribution*, data is clustered symmetrically around a central value with a bell-shaped frequency curve.

For sound mathematical reasons, many real-world examples of numerical data do follow a normal distribution. However, not all do so, and the name “normal” can sometimes be misleading.

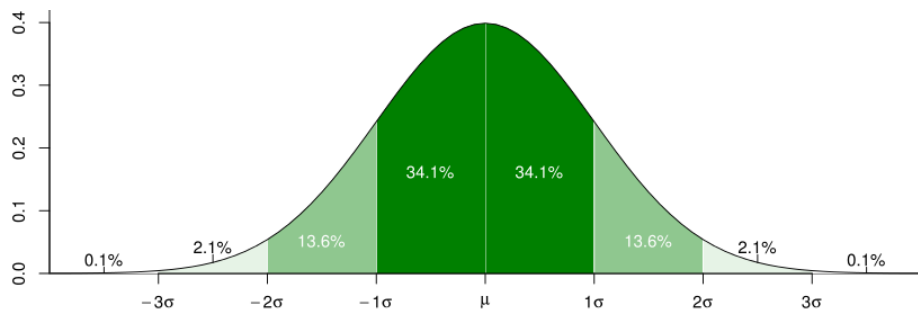


# Normal Distribution

Any normal distribution is described by two parameters.

The *mean*  $\mu$  (mu, said “mew”) is the centre around which the data clusters.

The *standard deviation*  $\sigma$  (sigma) is a measure of the spread of the curve. For a normal distribution, it coincides with the *inflection point* where the curve changes from being convex to concave.



# Statistics

A *statistic* is a single value computed from data that captures some overall property of the 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 any quantitative data, even if it is not normally distributed.

There are also other statistics, the **mode** and **median**, that are alternatives to the mean for summarising the “average value” of some data.

# Mode

For any set of data the *mode* is the value which occurs most often.

## Example: Mode

For the categorical data {red, blue, orange, red, yellow} the mode is red, which is the only value to occur twice.

Data may be *bimodal* (two modes) or even *multimodal* (more than two).

## Example: Bimodal data

For the integer data set {6, 2, 3, 6, 2, 5, 1, 7, 2, 5, 6} both 2 and 6 are modes, each occurring three times.

The mode makes sense for all types of data scale. However, it is not particularly informative for quantitative data with real-number values, where it is uncommon for the same data value to occur more than once.

This is an instance of a more general phenomenon: in general it is neither useful nor meaningful to compare real-number values for equality

# Median

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

## Example: Median

Given the integer data set  $\{6, 2, 3, 6, 2, 5, 0, 7, 2, 5, 6\}$  we can write it in non-decreasing order  $\{0, 2, 2, 2, 3, 5, 5, 6, 6, 6, 7\}$  and identify the middle value as 5.

The median exists for qualitative ordinal data and quantitative interval and ratio data. It does not make sense for categorical data, as that has no appropriate ordering.

Median is a good summary statistic for data where there is a forced cutoff at one end, or possible distortion by extreme outliers. For example, in reporting incomes, hospital waiting times, cancer survival times.

# Mean

Given data values  $\{x_1, x_2, \dots, x_N\}$ , the *mean* is their total divided by the number of values:  $(\sum_{i=1}^N x_i)/N$ .

## Example: Mean

For the integer data set  $\{6, 2, 3, 6, 2, 5, 0, 7, 2, 5, 6\}$ , the mean is  $(6 + 2 + 3 + 6 + 2 + 5 + 0 + 7 + 2 + 5 + 6)/11 = 4$ .

Although the formula for the mean involves a sum, the mean makes sense for both interval and ratio scales; it does not depend on an absolute zero in the scale. Mean does not make sense for categorical or ordinal data.

A mean incorporates all the data and is a genuine summary; however, it is not always the right choice of summary statistic, and can be distorted if there are extremely high or low values.

“The mean is like a loaded gun, which in the inexperienced hand can lead to serious accidents, as means can give hopelessly distorted results”

Karl Pearson, 1857–1936

# Variance and Standard Deviation

Given data values  $\{x_1, x_2, \dots, x_N\}$  with mean  $\mu$ , their *variance*  $\sigma^2$  is the **mean square deviation** from  $\mu$ :

$$\text{Variance} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Variance measures the spread of data, but it changes as the square of the data. A more common measure of spread is its square root, known as the *standard deviation*  $\sigma$ :

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

As for the mean, the standard deviation makes sense for both interval and ratio data; but has no meaning for qualitative data scales.



# Variance and Standard Deviation

Given data values  $\{x_1, x_2, \dots, x_N\}$  with mean  $\mu$ , their *variance*  $\sigma^2$  is the **mean square deviation** from  $\mu$ :

$$\text{Variance} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \left( \frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \mu^2$$

Not obvious, but the algebra works.

Variance measures the spread of data, but it changes as the square of the data. A more common measure of spread is its square root, known as the *standard deviation*  $\sigma$ :

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} = \sqrt{\left( \frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \mu^2}$$

As for the mean, the standard deviation makes sense for both interval and ratio data; but has no meaning for qualitative data scales.

## Example

For the integer data set  $\{6, 2, 3, 6, 2, 5, 0, 7, 2, 5, 6\}$  we compute:

$$\text{Variance} = \frac{(6-4)^2 + (2-4)^2 + (3-4)^2 + (6-4)^2 + (2-4)^2 + (5-4)^2 + (0-4)^2 + (7-4)^2 + (2-4)^2 + (5-4)^2 + (6-4)^2}{11}$$

$$= \frac{4 + 4 + 1 + 4 + 4 + 1 + 16 + 9 + 4 + 1 + 4}{11}$$

$$= \frac{52}{11}$$

$$= 4.73 \quad \text{to 3 significant figures}$$

$$\sigma = \sqrt{\frac{52}{11}}$$

$$= 2.17 \quad \text{to 3 significant figures}$$

## Example

For the integer data set  $\{6, 2, 3, 6, 2, 5, 0, 7, 2, 5, 6\}$  we compute:

$$\text{Variance} = \frac{6^2 + 2^2 + 3^2 + 6^2 + 2^2 + 5^2 + 0^2 + 7^2 + 2^2 + 5^2 + 6^2}{11} - 4^2$$

$$= \frac{4 + 4 + 1 + 4 + 4 + 1 + 16 + 9 + 4 + 1 + 4}{11}$$

$$= \frac{52}{11}$$

$$= 4.73 \quad \text{to 3 significant figures}$$

$$\sigma = \sqrt{\frac{52}{11}}$$

$$= 2.17 \quad \text{to 3 significant figures}$$

# Populations and Samples

So far we have seen different statistics for a given set of data, and how to compute them exactly.

Very often, however, data is only a **sample** drawn from a larger **population**, and we really want to know — or find out some information about — the statistic on the whole population. For example:

- Experiments in social sciences where one wants to discover information about some section of society — say, university students.
- Surveys and polls — for marketing, opinion gathering, etc.
- In software design when questioning a number of potential users in order to understand general user requirements.

In such cases it is impractical to obtain exhaustive data about the population as a whole; instead, we must work with a sample.

# Sampling

Sampling from a population needs to be done carefully to ensure analysis of the sample is a reliable basis for estimating properties of the whole population.

- The sample should be chosen **at random** from the population.
- The sample should be as large as is practically possible (given constraints on gathering data, storing data and calculating with data).

These improve the likelihood that a sample is *representative* of the population, reducing the chance of building *bias* into the sample.

Given a sample, we can calculate its statistical properties, and use that to infer information about similar properties of the whole population.

It is a significant topic in statistics, but beyond this course, to work out how to quantify and maximise the reliability of these techniques.

# Estimating Population Statistics

Suppose we have a **sample**  $\{x_1, \dots, x_n\}$  of size  $n$  from a **population** of size  $N$ , where  $n \ll N$  (i.e.,  $n$  is much smaller than  $N$ ).

We use the sample  $\{x_1, \dots, x_n\}$  to estimate statistics for the whole population. These estimates may not be correct; but knowing the sample and population size, we can often make estimates about the errors, too.

For mean, the best estimate of the **population mean**  $\mu$  is in fact the **sample mean**  $m$ :

$$m = \frac{1}{n} \sum_{i=1}^n x_i$$

This is an *unbiased estimator* — its value, given a random sample, is evenly distributed around the mean of the overall population  $\mu$ .

$$E(m) = \mu$$

# Estimating Population Variance

The variance and standard deviation of a sample are not appropriate estimates for the equivalent statistics on the population from which the sample is drawn; they turn out to be slightly too small, because a sample will be distributed more closely around its own mean than to the population mean.

The best estimate for the **variance** of the whole population is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2.$$

Note the denominator  $(n-1)$  rather than  $n$ . This is known as the *Bessel correction* and gives an unbiased estimator for the variation of the larger population:

$$E(s^2) = \sigma^2.$$

# Estimating Population Standard Deviation

Using this to estimate the variance of a whole population, based on a sample:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2$$

gives us an estimate for the **standard deviation** of the whole population:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2}$$

Again the denominator  $n$  for standard deviation has been replaced by  $(n - 1)$ ; and the mean  $m$  used is that of the sample, not the (unknown) population mean  $\mu$ .



# Beware

The use of samples to estimate statistics of a larger population is so common that the formula on the previous slide is very often the one needed, rather than the standard deviation of the sample itself.

Its usage is so widespread that sometimes it is wrongly given as **the** definition of standard deviation.

The existence of two different formulas for calculating standard deviations in different circumstances can lead to confusion. So take care.

Often calculators make both formulas available: as  $\sigma_n$  for the formula with denominator  $n$ ; and  $\sigma_{n-1}$  for the formula with denominator  $(n - 1)$ .

# Summary

**Data Scales:** Qualitative, Quantitative

**Qualitative:** Categorical, Ordinal

**Quantitative:** Interval, Ratio

**Visualisation:** Bar chart vs. Histogram

**Normal Distribution:** Bell curve, fixed by mean and standard deviation

**Statistic:** Single value computed from a set of data

**Averages:** mean, median, mode

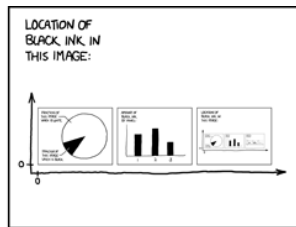
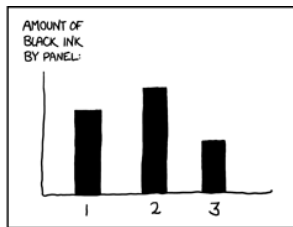
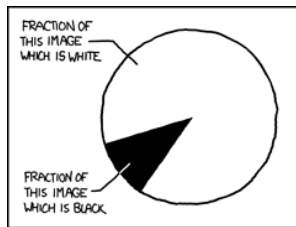
**Spread of Data:** variance, standard deviation

**Sample:** Chosen at random from population

**Estimates:** Population statistics from data about a sample

**Mean:** Use mean of the sample

**Variance, Standard Deviation:** Use the  $(n - 1)$  versions.



XKCD: 688 Self-Description

Q. How is this done?

Which panel do you draw first?

Can you do it without machine assistance?