Treatment and analysis of data - Applied statistics

What does it mean?

Some definitions from Wikipedia (en.wikipedia.org):

- **Data analysis** is the act of transforming data with the aim of extracting useful information and facilitating conclusions.
- **Statistics** is a mathematical science pertaining to collection, analysis, interpretation and presentation of data.
- **Applied statistics** is the use of statistics and statistical theory in real-life situations.

**Data treatment** (not listed in the Wikipedia) ≈ **data analysis**, perhaps with some emphasis on the practical, computational or automated aspects of the analysis.

**Statistics** includes **data analysis** (but also planning and conducting experiments).

The various terms are thus largely synonymous.

The subject of the present lectures can be summarized: “**how to transform data with the aim of extracting useful information and facilitating conclusions**”.

Relation to probability theory

Although probability theory is very important for data analysis, the two sciences have almost opposite aims, as illustrated by the diagrams below.

(Adapted from J.L. Devore, Probability and Statistics for Engineering and the Sciences, 5th Ed., Duxbury, Pacific Grove 2000)

Schematic representation of (a) deductive logic, or pure mathematics, and (b) plausible reasoning, or inductive logic (adapted from D.S. Sivia, Data Analysis – A Bayesian Tutorial, Oxford Univ. Press 1996)
Lecture plan September-October 2006

1. General concepts
2. Probability theory
3. Sampling and statistics; descriptive statistics and graphical presentation
4. Statistical modelling and estimation
5. Examples of Maximum Likelihood estimation
6. Bayesian estimation
7. Hypothesis testing
8. Numerical techniques (Monte Carlo methods, resampling, ...)

Recommended literature

  - Practical introduction to many methods commonly used in astronomy, lots of examples

- D.S. Sivia, *Data Analysis – A Bayesian Tutorial*, Oxford University Press (1996), or
  - A concise introduction to Bayesian data analysis in general

  - Contains not only programs, but very good introductions to some aspects of data analysis, in particular in Ch. 7, 14 and 15. Freely available via [www.nr.com](http://www.nr.com)

  - A good and authoritative reference book on (mainly classical) probability theory.

  - A magnificent account of probability theory as the rules for plausible reasoning, by a strong advocate of Bayesian thinking. Readable and thought-provoking, but sometimes a bit heavy.
Lecture 1: General concepts

Topics covered:
- Accuracy and precision
- Measures of precision
- Expectation
- Variance and dispersion
- The Law of Large Numbers
- Why averaging (often) works
- Historical note
- Some counter-examples

Accuracy and precision (1): Standards

Standardized terms to express uncertainties and errors in measurement results are defined e.g. in:
- *VIM – International vocabulary of basic and general terms in metrology* (ISO, 1993)
- *Guidelines for evaluating and expressing the uncertainty of NIST measurement results* (NIST TN-1297, 1994) physics.nist.gov/Pubs/guidelines/

Although partly different conventions are used in physical/astronomical literature, it is useful to know what is the standard practice in a wider community.
**Accuracy and precision (2): ISO definitions**

- **Accuracy**: The closeness of agreement between a test result and the accepted reference value
  
  - “a test result” = an observed, calculated or estimated value
  - “the accepted reference value” = the true value

- ** Precision**: The closeness of agreement between independent test results obtained under stipulated conditions
  
  - repeatability: precision under similar conditions
  - reproducibility: precision under different conditions

- **Trueness**: The closeness of agreement between the average value obtained from a large series of test results and the accepted reference value (cf. bias, unbiasedness)

**Accuracy = precision + trueness**

---

**Accuracy and precision (3): Conceptual illustration**

<table>
<thead>
<tr>
<th></th>
<th>high precision</th>
<th>low precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>high trueness</td>
<td><img src="high_accuracy.png" alt="high accuracy" /></td>
<td><img src="low_accuracy.png" alt="low accuracy" /></td>
</tr>
<tr>
<td>(high accuracy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>low trueness</td>
<td><img src="low_accuracy.png" alt="low accuracy" /></td>
<td><img src="low_accuracy.png" alt="low accuracy" /></td>
</tr>
</tbody>
</table>
Accuracy and precision (4): Uses

Accuracy, precision, trueness, repeatability, reproducibility, uncertainty, .... are qualitative concepts. It is recommended that they are not used as synonyms for the corresponding quantitative measures, which need to be specified in statistical terms (standard deviation, bias, etc.; in astronomy often standard error, RMS error, etc.).

Imprecision and inaccuracy are also qualitative, merely adding confusion to the previous vocabulary: therefore not recommended.

Examples:

Wrong: “the precision of a single observation is 0.012 mag” (precision used as a quantitative measure, unclear if the standard deviation is meant)

Inexact: “the standard error of a single observation is 0.012 mag” (unclear if precision or accuracy is meant)

Better: “the precision of a single observation is 0.012 mag (standard error)”

Accuracy and precision (5): Other terms

Some other common terms and their definition:

- error = (observed value) – (true value) [could be 0 even if accuracy is low!]
- bias = systematic error = E(observed value) – (true value) [E = expectation]
- random error = error – bias

Thus:

- error = bias + random error
- E(error) = bias
- E(random error) = 0

Note: E(X) = expected value of X ("mean value in the long run")

Also:

- residual = (observed value) – (value from fitted model) [not same as error!]
- correction = something added to an observed value [to improve trueness]
Accuracy and precision (6): Quantitative measures

In astronomy, the most common measure of precision/accuracy is:
- standard deviation
- standard error
- RMS error
- dispersion
- mean error (not recommended)

practically equivalent

Often designated $\sigma$ (sigma), or with $\pm$ as in:

“For the companion ... we obtained a mass of $0.4 \pm 0.1 \, M_\odot$”

[Other measures will be discussed in Lecture 2.]

Sometimes an attempt is made to separate precision and accuracy, as in:

“The observed integral flux above 200 GeV is $(4.8 \pm 1.3_{\text{stat}} \pm 1.0_{\text{syst}}) \times 10^{-12} \, \text{cm}^{-2} \, \text{s}^{-1}$”
(Aharonian et al., A&A 448, L19, 2006)

RMS random error (measuring precision)

An estimate of the likely size of any bias

Total RMS error $= 1.6$ (measuring accuracy)

Error modelling

A basic assumption in data analysis is that observational errors can be modelled as random variables, so that the laws of probability and sampling can be applied.

This means that observations are to some extent expected to behave like random experiments, similar to the flipping of a coin or throwing a die.

Thus we will start by looking at some general properties of such random experiments, taking the coin and die as model examples.
Expectation (1)

We are concerned with the results of an experiment (observation, measurement) where we can assume that

1. the experiment can be repeated as many times as we like (at least in principle);
2. the exact result of any single experiment cannot be predicted with certainty (there is a random element);
3. the result of any single experiment does not affect the results of other experiments (they are independent);
4. for all we know, all the experiments are made under identical conditions (they are thus interchangeable).

This is then a random experiment (from a statistical viewpoint). Examples:

- tossing a coin (possible outcomes: $x = 0$ [head] or $x = 1$ [tail]);
- throwing a die (possible outcomes: $x = 1, 2, 3, 4, 5, 6$);
- repeated measurement of a fixed physical quantity with the same instrument (possible outcomes: values of the real variable $x$).

Expectation (2)

Let $x_1, x_2, x_3, \ldots$ be successive results of such an experiment.

Formally, the values $x_i$ are regarded as different realizations of the random variable $X$.

Let $\xi_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ be the average of the first $n$ values. If the sequence $\xi_1, \xi_2, \xi_3, \ldots$ converges to some value $\xi$, then this is the expectation of $X$:

$$E[X] = \lim_{n \to \infty} \xi_n = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_i$$

(also called the expected value or mean of $X$).
Expectation (3)

Two important remarks:

1. There are cases where the sequence $\xi_1, \xi_2, \xi_3, \ldots$ does not converge and consequently $E[X]$ is undefined (examples later)

2. The limit should be understood in a probabilistic sense, e.g. as convergence in probability:

   \[
   \text{The sequence } \xi_1, \xi_2, \xi_3, \ldots \text{ converges in probability to } \xi \text{ if, for every } \epsilon > 0,
   \lim_{n \to \infty} P(|\xi_n - \xi| \geq \epsilon) = 0
   \]

Expectation (4)

Examples:

- In a fair coin-tossing experiment, with possible outcome $X = 0$ or 1 (for head and tail, respectively), the expected value is $E[X] = 0.5$

- When throwing an unbiased die, with possible outcome $X = 1, 2, 3, 4, 5$ or 6, the expected value is $E[X] = 3.5$

- For an unbiased observation of the fixed physical quantity $\theta$, the expected value (if it exists) is $E[X] = \theta$

The first two examples show that the expected value need not be an expected outcome of the experiment!
Expectation (5)

Three series of “tossing a coin” 100,000 times

Expectation (6)

Three series of “throwing a die” 100,000 times
Variance and dispersion; the Law of Large Numbers

Questions:
1. Under what conditions do the successive averages \(\xi_1, \xi_2, \xi_3, \ldots\) converge?
2. How quickly do they converge?

Introduce the variance and dispersion of \(X\):

\[
\text{Var}[X] = E[(X - \xi)^2], \quad D[X] = \sqrt{\text{Var}[X]}
\]

Then, assuming that \(E[X] = \xi\) and \(\text{Var}[X] < \infty\), we have:

1. The Law of Large Numbers: the mean value of \(X\) converges to \(\xi\) as \(n \to \infty\) (this is almost self-evident in view of our definition of \(E[X]\), but the important thing is that the limit exists if \(\text{Var}[X] < \infty\))
2. \(\text{Var}[\xi_n] = \text{Var}[X] n^{-1}, \quad D[\xi_n] = D[X] n^{-1/2}\)
   (this is exact for any \(n\), and easy to prove – see below)

Calculations with \(E, \text{Var}\) and \(D\) (1)

For any constants \(a\) and \(b\), and random variables \(X\) and \(Y\), we have

\[
E[aX + bY] = aE[X] + bE[Y]
\]

Special cases are:

\[
E[aX] = aE[X] \quad \text{ (for } b = 0) \\
E[X + Y] = E[X] + E[Y] \quad \text{ (for } a = b = 1)
\]

It follows that

\[
\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2
\]

Note how a linear transformation changes \(E, \text{Var}\) and \(D\):

\[
E[aX + b] = aE[X] + b \\
\text{Var}[aX + b] = a^2 \text{Var}[X] \\
D[aX + b] = |a|D[X]
\]
Calculations with E, Var and D (2)

If $X$ and $Y$ are independent, then

\[ E[XY] = E[X]E[Y] \]
\[ \text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] \]
\[ D[X + Y] = \sqrt{D[X]^2 + D[Y]^2} \]

These formulae can be generalized to any number of independent random variables $X, Y, Z, \ldots$;

\[ E[XYZ \ldots] = E[X]E[Y]E[Z] \ldots \]
\[ \text{Var}[X + Y + Z + \ldots] = \text{Var}[X] + \text{Var}[Y] + \text{Var}[Z] + \ldots \]
\[ D[X + Y + Z + \ldots] = \sqrt{D[X]^2 + D[Y]^2 + D[Z]^2 + \ldots} \]

Application to the mean of $n$ data points

If $x_i \ (i = 1, 2, \ldots)$ are independent with constant $\text{Var}[x_i] = \sigma^2$, then:

\[ \text{Var}[x_1 + x_2 + \cdots + x_n] = n\sigma^2 \]
\[ \text{Var}\left[\frac{1}{n}(x_1 + x_2 + \cdots + x_n)\right] = \frac{1}{n^2} \times (n\sigma^2) = \frac{\sigma^2}{n} \]

or

\[ D[x_1 + x_2 + \cdots + x_n] = \sigma\sqrt{n} \]
\[ D\left[\frac{1}{n}(x_1 + x_2 + \cdots + x_n)\right] = \frac{1}{n} \times (\sigma\sqrt{n}) = \frac{\sigma}{\sqrt{n}} \]

This is why averaging works:

When adding many similar data, the dispersion increases as $\sqrt{n}$; but after division by $n$ it decreases as $1/\sqrt{n}$.
Coin-tossing again: Accumulated “errors”

Three series of “tossing a coin” 100,000 times

\[(\xi_n - \xi) \times n \]

Coin-tossing again: Accumulated “errors” / \(\sqrt{n}\)

Three series of “tossing a coin” 100,000 times

\[(\xi_n - \xi) \times \sqrt{n} \]
Historical note

The technique of improving results by repeating and combining observations of the same quantity was first (?) used around 1600:

- Tycho Brahe (1546–1601) **repeated** certain fundamental observations (e.g. of the right ascension of $\alpha$ Ari) many times over several years; **averaged** morning and evening observations to **eliminate systematic errors** from refraction and parallax; and adopted as the final result a value close to the arithmetic mean (next slide).

- Galileo Galilei (1564–1642) used the principle of **least absolute deviation** for hypothesis testing (see below).

The calculation of the mean as a more precise value than a single measurement was well established by 1700–1740 (Flamsteed, Maupertuis). Then:

- Tobias Mayer (1723–1762) studied the libration of the moon, which involved an overdetermined system of equations: 27 equations for 3 unknowns. This he solved after averaging similar equations.

- Ruggero Boscovich (1711–1787) used the idea of **minimising the sum of the absolute deviations** to study the shape of the Earth.

- Carl Friedrich Gauss (1777–1855) and Adrien-Marie Legendre (1752–1833) independently discovered the **method of least squares** around 1800.

### Tycho Brahe’s observations of the R.A. of $\alpha$ Arietis

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26° 00’ 44”</td>
</tr>
<tr>
<td>2</td>
<td>26° 00’ 32”</td>
</tr>
<tr>
<td>3</td>
<td>26° 00’ 30”</td>
</tr>
<tr>
<td>4</td>
<td>26° 00’ 20”</td>
</tr>
<tr>
<td>5</td>
<td>26° 00’ 38”</td>
</tr>
<tr>
<td>6</td>
<td>26° 00’ 18”</td>
</tr>
<tr>
<td>7</td>
<td>26° 00’ 32”</td>
</tr>
<tr>
<td>8</td>
<td>26° 00’ 42”</td>
</tr>
<tr>
<td>9</td>
<td>26° 00’ 37”</td>
</tr>
<tr>
<td>10</td>
<td>26° 00’ 27”</td>
</tr>
<tr>
<td>11</td>
<td>26° 00’ 29”</td>
</tr>
<tr>
<td>12</td>
<td>26° 00’ 14”</td>
</tr>
<tr>
<td>13</td>
<td>26° 00’ 04”</td>
</tr>
<tr>
<td>14</td>
<td>26° 00’ 28”</td>
</tr>
<tr>
<td>15</td>
<td>26° 00’ 39”</td>
</tr>
</tbody>
</table>

**Mean value** 26° 00’ 28.9”

**Adopted by Tycho** 26° 00’ 30”

**Individual R.A. measurements from 1582–1588.**

Each point is normally the average of a morning and evening observation.

It is not clear how Tycho arrived at the value 26° 00’ 30” but it is very close to the mean value.

It also coincides with the median.

Tycho tabulated the R.A. of his standard stars to the nearest 5”.

(R.L. Plackett, *Biometrika* 45, 130, 1958)
Galileo Galilei’s discussion of the new star of 1572

In 1628 Scipio Chiaramonti published a pamphlet in which he “demonstrates” that the famous nova of 1572 was below the moon (sublunar). He combined observations made by Tycho and many others, at different latitudes, to calculate its parallax.

In his book *Dialogue Concerning the Two Chief World Systems* (1632), Galileo refutes Chiaramonti’s work with the following words:

SALVIATI: Now this author takes the observations made by thirteen astronomers at different polar elevations, and comparing a part of these (which he selects) he calculates, by using twelve pairings, that the height of the new star was always below the moon. But he achieves this by expecting such gross ignorance on the part of everyone into whose hands his book might fall that it quite turns my stomach.

Galileo shows that if *all* the observations are used, they rather support the hypothesis that the nova was infinitely high. His analysis contains some rudiments of statistical theory:

1. There is just one value that gives the true distance to the star
2. All observations are encumbered by errors
3. The errors are symmetric about zero
4. Small errors occur more frequently than large errors

From this Galileo concludes that one should select the hypothesis that requires the smallest corrections to the observations (method of least absolute deviations). (A. Hald, preprint, Copenhagen, 1985)

A counter-example (1): The Cauchy distribution

The Cauchy (or Lorentzian) distribution

\[ p(x \mid \alpha, \beta) = \frac{\beta}{\pi} \frac{1}{\beta^2 + (x - \alpha)^2} \]

\[ p(x \mid 0, 1) = \frac{1}{\pi(1 + x^2)} \quad \text{(standard form)} \]

has undefined E\[x\] and infinite Var\[x\]. Therefore the Central Limit Theorem does not apply. In fact, if \( x_i \sim \text{Cauchy}(\alpha, \beta) \) then

\[ \xi_n = \frac{1}{n} \sum_{i=1}^{n} x_i \sim \text{Cauchy}(\alpha, \beta) \]

With such an error law the mean of \( n \) observations is not more or less precise than any single observation!

(The median works better in this case.)
The Cauchy distribution (from Wikipedia)

Non-convergence of the average of Cauchy variates
A counter-example (2): “Heavy-tail distributions”

A non-negative random variable $X$ is called heavy-tailed if

$$P[X > x] \propto x^{-\alpha}$$

for some $0 < \alpha < 2$. This is also called a power law, scaling distribution, or a self-similar process. The parameter $\alpha$ is called the tail index.

- $1 < \alpha < 2$: $X$ has infinite variance, but finite mean
- $0 < \alpha < 1$: the variance and mean of $X$ are infinite

Heavy-tailed distributions have been used to model a wide range of phenomena, such as:

- Communication network loads
- Data file sizes
- Queueing times
- Insurance risks
- Natural disasters

Disasters follow a heavy-tail distribution with $\alpha = 1$

20th Century’s 100 largest disasters worldwide (W. Willinger, AT&T)
Micro-meteoroid impacts on Gaia (1)

In space, the omni-directional flux of particles with mass $> m$ is (Yamakoshi, 1994)

$$F(m) = (1.1 \times 10^{-5} \text{ m}^{-2} \text{ s}^{-1}) \times \begin{cases} (m/m_0)^{1/2} & \text{for } m < m_0 \\ (m/m_0)^{-7/6} & \text{for } m > m_0 \end{cases}$$

where $m_0 = 2.8 \times 10^{-11}$ kg. The mean impact velocity relative to Gaia is $v = 12,000 \text{ m s}^{-1}$.

An impact changes the angular velocity of the satellite by

$$\Delta \omega \approx |r \times v| \ m I^{-1},$$

where $r$ is the impact vector (rel. c.o.m. of Gaia) and $I \approx 3000 \text{ kg m}^2$ Gaia’s moment of inertia.

For omni-directional influx on the solar screen ($R = 4.75 \text{ m}$) the rate of impacts with $m > m_0$ is

$$0.5 \pi R^2 F(m_0) = 1.4 \text{ hr}^{-1}.$$ 

Also,

$$|r \times v| \approx (\pi/6) R v = 30,000 \text{ m}^2 \text{ s}^{-1},$$

so the mean effect of $m_0$ particle is

$$\Delta \omega \approx 2.8 \times 10^{-10} \text{ rad s}^{-1} \approx 60 \mu\text{as s}^{-1}.$$ 

Micro-meteoroid impacts on Gaia (2)

The effect of the impact must be compared with

- the photon-statistical measurement error, which for a single measurement (one CCD crossing = 3.6 s) of a bright star ($V = 10$) is of order 60 $\mu$as
- the attitude control, which requires a star to be found within $\pm 200$ mas of the predicted position after 8 s

Thus:

- a disturbance $\sim 60 \mu\text{as s}^{-1}$ happens every 40 min: clearly detectable
- a disturbance $\sim 600 \mu\text{as s}^{-1}$ happens every 10 hours: temporary loss of science data
- a disturbance $\sim 6$ mas s$^{-1}$ happens every 150 hours (6 days): upsets attitude control
- a disturbance $\sim 60$ mas s$^{-1}$ happens every 100 days: temporary loss of attitude
- a disturbance $\sim 600$ mas s$^{-1}$ happens every 4 years: major recovery operation needed

This is a heavy-tailed stochastic process with tail index $\alpha = 7/6$. The mean is in principle finite but in practice meaningless; the variance is infinite.