Experimental Evaluation: Data Analysis

- Data analysis is the systematic study of data in order to understand its meaning, organisation, structure, relationships etc.
  - where "data" is a group of measurements, or facts
- it can be done by using various techniques:
  - statistical
  - logical
  - graphical
  - tabular
  - etc.
- in this course, we'll focus on *Exploratory Data Analysis*

Note: source for this material is the *e-Handbook of statistical methods*, available at http://www.itl.nist.gov/div898/handbook/index.htm

Exploratory Data Analysis (EDA)

- main characteristic: use of (mostly) graphical techniques to understand data properties
- but it's not only a matter of using graphics, it's a different "philosophy" to approach the problem
- in classical data analysis, the experimenter
  1. starts with a general problem to explore
  2. collects some data, with an experiment
  3. makes a hypothesis on the model this data should follow
  4. carries out an analysis of data
  5. and draws some conclusions on the features that the data exhibits
- while in EDA the experimenter:
  1. starts with a general problem to explore
  2. collects some data, with an experiment
  3. carries out an analysis of data
  4. infers the model that is appropriate to represent the data
  5. and draws some conclusions on the features that the data exhibits

EDA vs classical DA

- model:
  - in classical DA it is imposed *a priori*: e.g. the most common probabilistic model assumes that the errors about the deterministic model are "normally" distributed (they have a bell shaped function)
  - in EDA, it's the data that suggests the most appropriate model
- focus:
  - in classical DA is on the model
  - in EDA is on the data
- techniques:
  - in classical DA are mostly quantitative
  - in EDA are mostly graphical

EDA vs classical DA (2)

- rigor:
  - in classical DA techniques are rigorous, formal and "objective"
  - in EDA they are suggestive, and are subjective to the interpretation of the analyst
- data treatment:
  - in classical DA the aim is producing few "numbers" (*estimates*) that summarise the data properties
  - in EDA all data is on focus (e.g. plotted in a graphic)
- assumptions:
  - in classical DA one can easily discover "statistically significant" variations from the assumed model, but this is only true if the initial assumption was correct
  - in EDA there are very few assumptions, the analysis of the data has priority

Why EDA?

- it's oriented towards the future, rather than the past
  - it uses data to understand and improve, rather than summarise what happened
  - it has an important role in research
- a good "feel" for the data is invaluable
  - the main goal is to gain insight into the *process* behind data
  - but also to understand what is NOT in the data
- this can almost only be obtain by graphical techniques
  - graphics can give information that no number can replace
  - they rely on the human's ability to recognise patterns and make comparisons

Assumptions for Measurement Processes

- these are the same for all techniques
- the data from a process should "behave" like:
  1. random drawings
     - so, data should not be *related* to one another, they should not influence one another
  2. from a fixed distribution
     - data are scattered in a fixed way (graphically, according to a "shape")
  3. where the distribution has a fixed location
     - the expected value of data is fixed
  4. and a fixed variation
     - the way in which data differs from the expected value is fixed

Assessing the distribution parameters

- the usual estimate of location is the mean:

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

from N measurements: $Y_1, Y_2, \ldots, Y_n$

- the usual estimate of variation is the standard deviation

$$s_Y = \frac{1}{\sqrt{(N-1)}}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}$$

from N measurements: $Y_1, Y_2, \ldots, Y_n$, where $\overline{Y}$ is the mean
  - it tells you how tightly the values cluster around the mean
  - the smaller s_Y, the less spread the distribution is

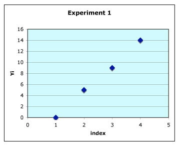
Example

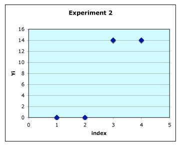
- we want to measure how many wrong menu options the user tries out before opening the right one
- we observe four users (so $N = 4$), asking them to find three different menu options. The mistakes they make are:
  - Experiment 1: $Y_1 = 0, Y_2 = 5, Y_3 = 9, Y_4 = 14$
  - Experiment 2: $Y_1 = 0, Y_2 = 0, Y_3 = 14, Y_4 = 14$
  - Experiment 3: $Y_1 = 5, Y_2 = 6, Y_3 = 8, Y_4 = 9$
- the means are:
  - Experiment 1: $\overline{Y} = \frac{1}{4}\sum_{i=1}^{4} Y_i = \frac{0+5+9+14}{4} = 7$
  - Experiment 2: $\overline{Y} = \frac{1}{4}\sum_{i=1}^{4} Y_i = \frac{0+0+14+14}{4} = 7$
  - Experiment 3: $\overline{Y} = \frac{1}{4}\sum_{i=1}^{4} Y_i = \frac{5+6+8+9}{4} = 7$
- while the standard deviations are:
  - Experiment 1: $s_Y = \frac{1}{\sqrt{3}}\sqrt{(0-7)^2 + (5-7)^2 + (9-7)^2 + (14-7)^2} = 5.94$
  - Experiment 2: $s_Y = \frac{1}{\sqrt{3}}\sqrt{(0-7)^2 + (0-7)^2 + (14-7)^2 + (14-7)^2} = 8.08$
  - Experiment 3: $s_Y = \frac{1}{\sqrt{3}}\sqrt{(5-7)^2 + (6-7)^2 + (8-7)^2 + (9-7)^2} = 1.82$

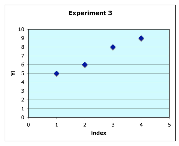
EDA technique: Run Sequence Plot

- the EDA technique for assessing a fixed location and fixed variation distribution is the Run Sequence Plot
  - all values of Y_i are simply plotted on a chart where
    - the vertical axis is the variable Y_i
    - and the horizontal axis is the index $1, 2, \ldots, n$
- this should give a feeling of how fixed is the location
  - whether many of the Y_i are distributed around the same values
- and how fixed is the variation
  - whether there are shifts or outliers

Run Sequence Plot applied to the example





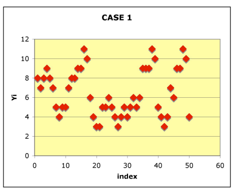


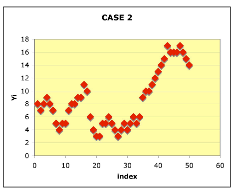
data does not fluctuate around the mean	there is a big jump after 2	data is well close to the mean
there is an increasing level of expertise among the users?	the first two users are expert, the second two are not?	all four users have similar expertise?

- in the first two cases the experimenter may presume that there is a factor (user's expertise) which has not been considered, and biases the result

Other examples

- the application of the Run Sequence Plot is more significant when there are more data to analyse
- for example compare the following two plots:





data fluctuates around the mean	there is a shift in location after about 35 items
$\overline{Y} = 2.4; s_Y = 6.54$	$\overline{Y} = 4.3; s_Y = 8.56$

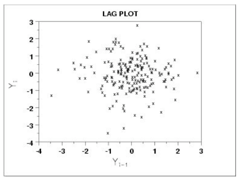
Assessing the randomness of the distribution

- a non random distribution is one in which the values depend on one another
  - e.g., in the example before, if one user could tell another where the menu option is
- if a distribution is not random, then all statistical tests have little sense
- one example of non randomness is autocorrelation
- this is when a value Y_i in the distribution is related to another value Y_{i-k}
  - this means that each value is highly dependent on the value k values before
  - if $k = 1$ then each value is highly dependent on the previous one
- the number k is called the lag of the autocorrelation

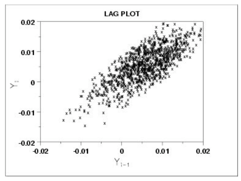
EDA technique: Run Sequence Plot

- a lag plot can be used to check if data is random or not
- first, one needs to choose the lag, that is the value for k
  - usually, $k = 1$
- then one plots the values of Y_i against the values of Y_{i-k}
  - usually, the value of Y_i against the valued of Y_{i-1}
- for truly random data, the plot should show no structure
- a structure (a recognisable shape of the plot) would suggest that one could tell, knowing one value, which the next one is
  - therefore data is not random
  - the plot can also suggest which is the most appropriate model to represent the data

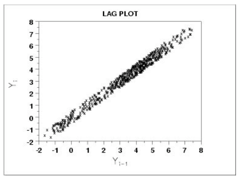
Examples of Lag Plots



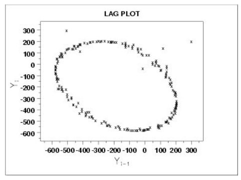
data show no correlation



there is some correlation



very strong correlation



very strong correlation