Introduction to Probability and Statistics

Sarah Filippi
Departmental Lecturer, University of Oxford
filippi@stats.ox.ac.uk

22/10/2014
Objectives of Course

▶ To teach underlying theory and applications of statistical data analysis
▶ To emphasis the importance of exploratory data analysis, prior to any formal statistical testing or modelling
▶ To define some useful summary statistics (numbers) that help characterise features of data sets, such as location and spread
▶ To introduce probability as a mathematical framework for reasoning with uncertainty
▶ To demonstrate the use of MatLab to implement these techniques
What is Statistics?

- Statistics is the *science* and *art* of data analysis
  - from observational studies
  - from planned experiments
- Statistics is concerned with the collection, analysis and interpretation of data
- It is the science of the scientific method
Branches of statistics

- Statistics covers a range of areas from data collection (optimal design of experiments) to the construction of stochastic (empirical) models
- It is a broad church
- Some areas of note include
  - graphical displays of quantitative data
  - stochastic modelling of systems
  - forecasting algorithms
Uncertainty

At the heart of statistics is the rigorous treatment of uncertainty or random variation as characterised via probability

- probability provides a formal system for inference and inductive logic
- allows for coherent accumulation of evidence supporting or refuting a hypothesis of interest
My use of statistics

- I work on a range of topics in computational statistics focused on understanding biological processes and their relation to disease.
- Using Bayesian statistics and stochastic models.
- I did a PhD in reinforcement learning, a field at the intersection of statistics, machine learning and artificial intelligence.
First steps

- The starting point of ALL good statistical data analysis begins with graphical plots and summary statistics of the data
- ALWAYS, ALWAYS, ALWAYS, PLOT YOUR DATA!!!
- Why?
First steps

- The starting point of ALL good statistical data analysis begins with graphical plots and summary statistics of the data
- ALWAYS, ALWAYS, ALWAYS, PLOT YOUR DATA!!!
- Why?

  Graphics reveal data, communicate complex ideas and dependencies with clarity, precision and efficiency
Excellent graphics:

- show the data
- induce the viewer to think about the substance
- avoid bias
- make large complex data sets coherent
- encourage data exploration and debate
Types of data

**Categorical/Qualitative**
- Nominal
  - Binary: Gender, Smoker (yes/no)
  - Non-Binary: Eye colour, ethnicity
- Ordinal
  - Likert scale: 1=Strongly disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly agree.
  - Socioeconomic class
  - Highest level of educational attainment

**Quantitative**
- Discrete (Counts)
  - Number of children
- Continuous
  - Height, weight
  - Age
  - IQ
  - Femur length
Bar chart for categorical data

Sex of babies born in one day on an Australian hospital

Frequency

Girl       Boy
0 4 8 12 16 20 24
Histograms for quantitative data

Birth Weight (g) vs. Frequency
How to make a histogram?

1. Find range of the data
2. Choose bin size
3. Count up observation in each bin
4. Plot boxes
Effect of bin size

Too few categories

Birth Weight (g)

Frequency

0 5 10 15 20 25 30 35

1500 2500 3500 4500

Too many categories

Birth Weight (g)

Frequency

0 1 2 3 4 5 6 7

1500 2500 3500 4500
Histograms: frequency vs density

Birth Weight (g) vs Frequency

- Birth Weight (g): 1500, 2000, 2500, 3000, 3500, 4000, 4500
- Frequency: 0, 5, 10, 15, 20

Birth Weight (g) vs Density (babies/g)

- Birth Weight (g): 1500, 2000, 2500, 3000, 3500, 4000, 4500
- Density (babies/g): 0, 0.01, 0.02, 0.03, 0.04
Typical shapes of histograms

- bell shapes
- left skewed shapes
- right skewed shapes
- bimodal
Bell shape
Right skewed
Bimodal

Percentage acetylation of isoniazid

Frequency

0 10 20 30 40 50

25 30 35 40 45 50 55 60 65 70 75 80 85 90 95

Percentage acetylation of isoniazid
Summary statistics are numbers (univariate statistics) that characterise certain (important) features of your data.

Definition: A "statistic" is a number derived from a population; formally, it is a function on a set of random variable(s).

You can think about summary statistics as data reduction; condensing your $n$ data points to a small number of pertinent features.

In data analysis two common attributes of interest are the location and spread of the data.
It is vital at this point to introduce the notion of an unknown underlying population, characteristics of which you are trying to estimate using a sample taken from the population.

For example, you would like to know the average weight of a Welshman. There IS a true value for this (but you and I will never know it). Because we don't know it we estimate it using a data set of observations taken from the population.

This is one of the most important concepts in Statistics.

Think carefully about this.
Location: mean

- The **mean** is the most commonly used measure of location.
- The **true unknown population mean** is estimated by the "average" value in your data

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

- The mean is also known as the central moment; which measures the expectation of your data.
- It defines the centre of mass (measure) of your data.
Location: mean

- The mean value is often not representative of an actual value.
Location: mean

- The mean value is often not representative of an actual value
  - "The average Welshman has less than two legs"
  - What is the average (expected) value of the throw of a dice?
- The mean value can be distorted by outliers
  - In 1984 U. Virginia's Dept. Communication announced that starting graduate salary was $55,000 in 1984
  - However, class contained 7ft 4inch, Ralph Sampson, leading N.B.A. center for Houston Rockets
Location: median

- The **median** provides an alternative measure of location.
- The **true unknown population median** is estimated by the central value of the data.
- That is, by first sorting the data (by value) and then reporting the middle value (or the average of the two middle values if the number of data points are even).
- The median provides a useful complement to the mean as it is more robust. That is, it is less sensitive to extreme values,
- Compare the mean and median values of the following data set:
  
  \[(2.3, 5.2, 6.4, 7.8, 987234.3)\]
Location

Symmetric

Positive Skew

Negative Skew
Spread: standard deviation

- The **standard deviation** measures spread of the data around the mean.
- The true unknown population standard deviation is estimated by
  
  $$sd(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- The **variance** is simply the standard deviation squared.
- The standard deviation is more useful than the variance as it has the same units of measurement as the data (rather than on squared units for the variance).
An alternative measure of spread is provided by the quantiles.

The quantiles are estimated by sorting the data by value and then reading off the values at particular points.

The 0% quantile is the minimum value; the 100% quantile is the maximum value; the 50% quantile is the median. In general the A% quantile is the A/100 sorted value.

The inter-quartile range is defined as the range 75% – 25% quantile.

The inter-quartile range is a robust measure of spread as it discards the top and bottom 25% of the data.
Boxplots

- Introduced by John Tukey, a founder of graphical statistics
- Middle bar denotes the median
- Rectangle enclose the **Inter-Quartile Range (IQR)**
  note: position of median bar within IQR reveals skewness in the distribution
- **Outliers** are drawn as points; defined as greater than 1.5 IQR from rectangle
- "**Whiskers**" extend out to furthest points which are not outliers
Principle of good graphical display

Good graphs should.....

- Display information in an unbiased and uncluttered manner
- Have a clear purpose
- Reveal interesting and pertinent features of the data distribution
  - and highlight outliers
- Encourage the viewer to think about the data generating mechanisms (abductive inference)
Pie Chart

Rules for using Pie Charts:

Never use them!
Pie Chart

Rules for using Pie Charts:

Never use them!

Pie charts are rarely justified. If you must use them, do so only when the total is fixed (and meaningful) and never (ever!) use 3d pie charts as they’re worse still!!

What’s the matter with pie charts:

► people are not good at interpreting areas
► zero is often a very meaningful number but gets lost in a pie chart
► very hard to compare two pie charts
► loose overall scale (size) of the underlying figures

Barplots are usually a much better choice
Pie Chart

- 5’ to known gene
- Pseudogene
- Novel
- Within 3’ flanking

Number of TFBS regions
Barplot

- Barplots are useful for comparing single figures across groups.
- Barplots should **not(!)** be used to compare distributions of data across groups.
- Most of the information is lost in this case.
Terrible Graph of Olympic 100m winning times, by gender, since 1928
Barplot

Winning times of Olympic 100m, by gender, since 1928

Winning Times (s)
Barplot

▶ Note how the information is revealed by the boxplot whereas interesting features are lost in the barplot of the same data

*** Aside: you will have noted that the 100m times is a poor example as the data follows a time trend (i.e. not exchangeable) ***

▶ A graph is an opportunity to present, with clear purpose, the data in the most informative way

▶ Don’t waste this opportunity
Scatter plots can be used to reveal associations between variables.

- Also useful to breakdown association into categories using conditional scatterplots (see the examples using scatter and plotmatrix).
Summary

- Type of data
- Representing data with graphs: bar charts, histograms, boxplots, scatter plots
- Summarising data: location, spread
- Graphical displays enhance our understanding of the state of nature
- Good graphs serve a well defined purpose, and present data in an uncluttered and unbiased manner
- Good graphs reveal pertinent features and invite us to think about the causal mechanisms that underly the data
Some references

- Gilks, W et al. eds. ”Markov chain Monte Carlo in Practice”. (1998) - a great collection of papers illustrating the power of modern computational statistics