THE UNIVERSITY of EDINBURGH

Data Science in Medicine

Lecture 2: Introduction to Statistics

Dr Areti Manataki

Usher Institute
The University of Edinburgh

Usher Institute
Population Health Sciences & Informatics

Introduction to Statistical Analysis

- Why analyse data?
  - 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.

Introduction to Statistical Analysis

- Mathematical statistics provide a powerful toolkit for performing such analyses, with wide and effective application.
- 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.
- When carrying out scientific studies, statistics can help describe a class of scientific events and explain these events.

In the next lectures

- Summary statistics
- Visualisation
- Correlation & Hypothesis testing
- χ^2 testing for categorical data
- Statistical analysis with R

Data scales

Data may be:

- qualitative (descriptive)
  - categorical scale
  - ordinal scale
- quantitative (numerical)
  - interval scale
  - ratio scale

Each of these supports different kinds of analyses.

Categorical scale

- Categorical scale: each data item is drawn from a fixed number of categories, where the names of the categories may occur in any sequence and are not orderable.
  - e.g. nationality: French, Japanese, Mexican, etc.
  - e.g. type of transportation: train, bus, car, etc.
- Categorical scales are sometimes called nominal.

Ordinal scale

- Data on an ordinal scale has a recognized ordering between data items, but there is no meaningful arithmetic on the values.
  - e.g. European Credit Transfer and Accumulation System (ECTS) grading scale: A, B, C, D, E, FX and F.
  - e.g. finishing position in a race: 1st, 2nd, 3rd, etc.

Interval scale

- Interval scale: a numerical scale (usually with real number values) in which we are interested in relative rather than absolute value.
  - e.g. Celsius temperature scale
- The differences between the numbers are interpretable, but the variable doesn’t have a “natural” zero value.
- Subtraction and average are meaningful, but addition or multiplication are not.

Ratio scale

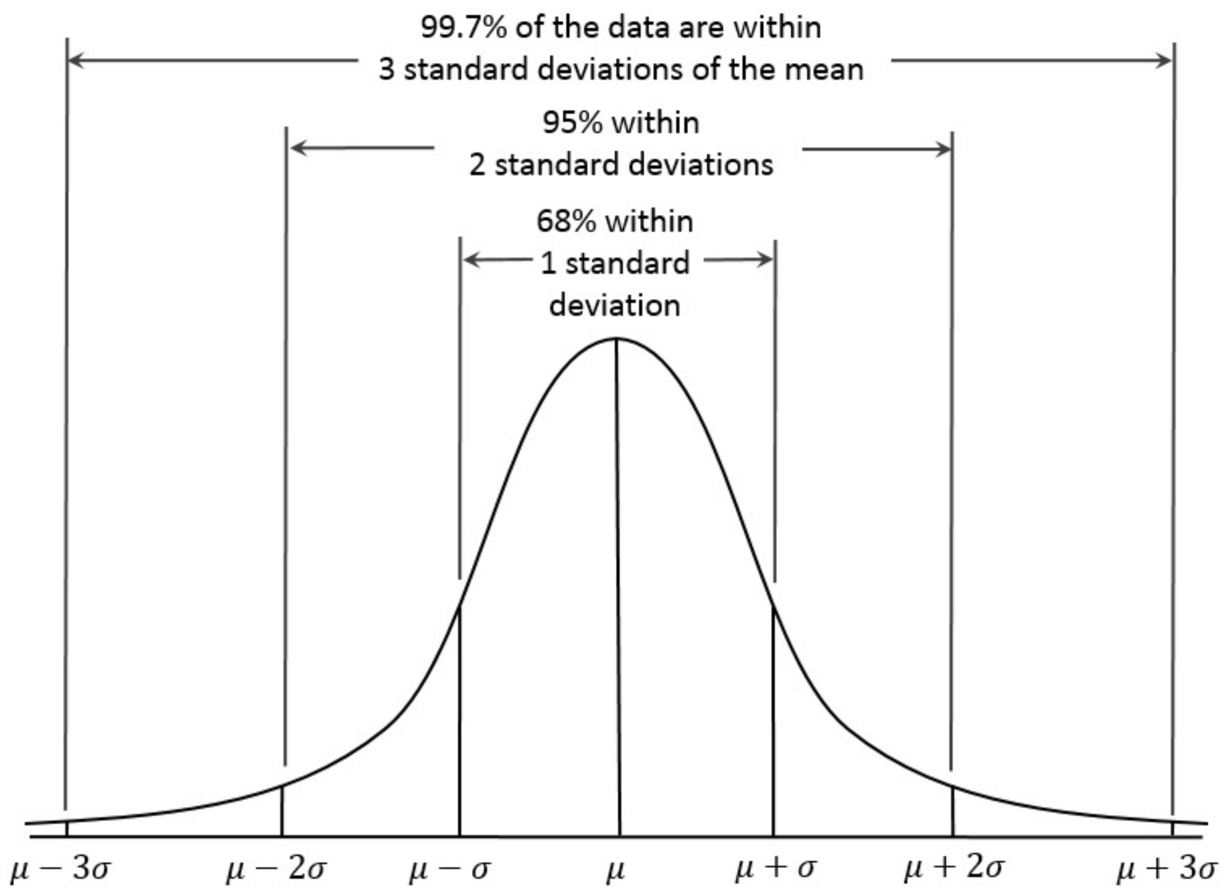
- Ratio scale: a numerical scale (again usually with real number values) in which there is a notion of absolute value.
  - e.g. age in years
  - e.g. response time
- Zero really means zero.
- Subtraction, average, addition and multiplication are meaningful.

Continuous vs. discrete data

- A separate type of distinction.
- Continuous variable: it is possible to have another value between any two values
  - e.g. response time
- Discrete variable: a variable that is not continuous
  - e.g. graduation year

	continuous	discrete
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

Normal distribution



Normal distribution

- Any normal distribution is described by two parameters:
  - The mean μ is the centre around which the data clusters.
  - The standard deviation σ is a measure of the spread of the curve.

Summary statistics

- A statistic is a single value computed from data that captures some overall property of the data.
- When describing data, we're typically interested in:
  - measures of central tendency: these give us an idea of what a typical or common value for a given variable is
    - mean, median, mode, etc.
  - measures of dispersion: these give us an idea of how spread out data values are
    - range, variance, standard deviation, etc.

Mean

- Given data values $\{x_1, x_2, \ldots, x_N\}$, the mean is their total divided by the number of values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- Appropriate for both interval and ratio scales; it does not depend on an absolute zero in the scale.
- It does not make sense for categorical or ordinal data.

Mean example

- Given data values $\{x_1, x_2, \ldots, x_N\}$, the mean is their total divided by the number of values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- Suppose that these are the grades that students got last year in the Statistics course: {69, 70, 86, 42, 54, 79, 69}

$$\mu = \frac{69+70+86+42+54+79+69}{7} = \frac{469}{7} = 67$$

Median

- The median of a data set is the middle value when the values are ranked in ascending or descending order.
- Given data values $\{x_1, x_2, . . . , x_N\}$ sorted into in non-decreasing order, the median is:
  - $x_{(N+1)/2}$ for N odd
  - any value between $x_{N/2}$ and $x_{(N/2)+1}$ for N even

Median

- Appropriate 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.

Median example

- Given data values $\{x_1, x_2, \ldots, x_N\}$ sorted into in non-decreasing order, the median is:
  - $x_{(N+1)/2}$ for N odd
  - any value between $x_{N/2}$ and $x_{(N/2)+1}$ for N even
- We can write the Statistics course grades dataset in non-decreasing order: {42, 54, 69, 69, 70, 79, 86}
- The median is 69.

Mode

- The mode of a data set is the most commonly occurring value.
- The mode of {69, 70, 86, 42, 54, 79, 69} is 69.
- It is most typically used for ordinal or categorical data.
- 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.

Range

- The range of a dataset is the difference between the highest and the lowest values.
- Often the minimum and maximum values are also reported.
- The range of {69, 70, 86, 42, 54, 79, 69} is 86-42 = 44.
- The interquartile range is an alternative measure that is less influenced by extreme values. This is used a lot.

Variance

- Given data values $\{x_1, x_2, . . . , x_N\}$ with mean μ, their variance σ^2 is the mean square deviation from μ:

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

- For the Statistics course grades dataset {69, 70, 86, 42, 54, 79, 69} we have:

$$\sigma^2 = \frac{(69-67)^2 + (70-67)^2 + (86-67)^2 + (42-67)^2 + (54-67)^2 + (79-67)^2 + (69-67)^2}{7}$$

$$= 188$$

Standard Deviation

- A more common measure of spread is its square root, known as the standard deviation:

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

- The standard deviation makes sense for both interval and ratio data; but has no meaning for qualitative data scales.
- This is perhaps the most popular measure of dispersion.

Standard Deviation example

- A more common measure of spread is its square root, known as the standard deviation:

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

- For the Statistics course grades dataset {69, 70, 86, 42, 54, 79, 69} we have:

$$\sigma = \sqrt{\frac{(69-67)^2 + (70-67)^2 + (86-67)^2 + (42-67)^2 + (54-67)^2 + (79-67)^2 + (69-67)^2}{7}}$$

$$= \sqrt{188} = 13.71$$

Populations vs. samples

- It is often impractical to obtain exhaustive data about the population as a whole; instead, we must work with a sample.
- So we use the sample to estimate statistics about the whole population.

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.

Estimating Population Statistics

- Suppose we have a sample $\{x_1, x_2, . . . , x_n\}$ of size n from a population of size N, where n << N.
- We use the sample $\{x_1, x_2, . . . , x_n\}$ 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.

Estimating Population Mean

- The best estimate of the population mean μ is the sample mean m:

$$m = \frac{\sum_{i=1}^{N} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Estimating Population Variance

- The estimate for the variance of the whole population is:

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

- Note the denominator (n - 1) rather than n. This is known as the Bessel correction.
- Note that the mean m used is that of the sample, not the (unknown) population mean.

Estimating Population Standard Deviation

- The estimate for the standard deviation of the whole population:

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

- Again, the denominator is (n - 1) rather than n, and the mean m is used.

Conclusions

- Data may be qualitative (categorical or ordinal scale) or quantitative (interval or ratio scale).
- Summary statistics involve measures of central tendency (e.g. mean, median, mode) and measures of dispersion (e.g. range, variance, standard deviation).
- Beware of different formulas for calculating population and sample statistics.