Statistics for Computer Science

Lecture 01

doc. PaedDr RNDr. Stanislav Katina, PhD.
Institute of Mathematics and Statistics, Masaryk University
Honorary Research Fellow, The University of Glasgow

Syllabus

1. Why computer scientists should study statistics?
2. Computer science related problems with analysed data
3. Why the thought study based on data is useful?
4. Data types
5. Sampling
6. Parametric probabilistic and statistical models
7. Likelihood principle and parameter estimation using numerical methods
8. Descriptive statistics (tables, listings, figures)
9. From description to statistical inference
10. Hypothesis testing and parameters of a model
11. Goodness-of-fit tests
12. Testing hypotheses about one-sample
13. Testing hypotheses about two-samples
14. Testing hypotheses about more than to sample problems
15. Interpretation of statistical findings

Why computer scientist should study statistics?

There is a story about two doctors who are floating above the countryside in a hot-air balloon. They are drifting with the wind and enjoying the scenery, but after couple of hours, they realise that they are totally lost. They see someone down on the ground, and shout down "*Hello! Can you tell us where we are?*" The person on the ground replies, "*You are fifty feet up in the air, in a hot-air balloon.* One doctor turn to the other and says, "*That person on the ground must be a* ***statistician****.*" "*How did you know?*", came astonished reply from the ground. "*Only a statistician would provide an answer that was totally accurate and totally useless at the same time.*" But in this story, the statistician has the last word. "*Very good. But I can also tell that you two are* ***doctors****.*

Why computer scientist should study statistics?

Reality "BBC report"

A UK social atlas suggests that British society is becoming more segregated by class, researchers have said [. . .] It found that:

- An average child in the wealthiest 10% of neighbourhoods can expect to inherit at least 40 times as much wealth as a typical child in the poorest 10%
- In some areas, 16-to-24-year-olds are 50 times more likely to attend an elite university than in others
- In the most impoverished parts of the country young adults in this age group are almost 20 times more likely not to be in education, employment or training than those in the wealthiest neighbourhoods
- There are no large neighbourhoods where under five year olds from the highest social class spend time with any other class of children other the one just beneath them.

(http://news.bbc.co.uk/2/hi/uk_news/6984707.stm)

Why computer scientist should study statistics?

Reason "all is about the data"

In any scientific investigation, there are bound to be some sources of ***bias****. Perhaps the sample wasn't quite* ***random****, or maybe the respondents tended to* ***overestimate*** *their income. We do everything we can to* ***minimize those biases****, but we can never eliminate them entirely. Consequently, small, artifactual relationships are likely to creep into the data. A large sample [or dataset] is like a very* ***sensitive*** *measuring instrument. It's so sensitive that it detects those* ***artifactual relationships*** *along with the* ***true relationships****.*

(Allison 1999, pp.58–59)

Why computer scientist should study statistics?

Don't believe everything you read

It is good to question what a statistic is telling you. Something can appear ***statistically important*** *yet still have* ***no social or scientific meaning****. Be guided by theory and by prior research findings, as well as by* ***what the data are saying****. Try not to 'fit the facts' to the data, just because it is convenient to do so.* ***Statistics are an aid for your intelligence and judgement, not a replacement for them!***

Three types of statistics

Descriptive statistics

Descriptive statistics is providing a summary of a set of measurements. ***Arithmetic averages*** *may be calculated,* ***the most common values*** *identified and* ***the spread of values*** *obtained (for example, by finding the minimum and maximum).*

Descriptive statistics are useful in their own right and as a precursor to more sophisticated analysis. They are ***cognitively useful****: they help to frame understanding of the data and of the real-world events and processes they measure. They are also* ***useful diagnostically****, especially for detecting error (often caused by missing data due to incomplete record keeping)*

EXAMPLE: The (mean) average annual rainfall at Darwin airport in Northern Australia was 1847.1 mm for the period 1971–2007. For Melbourne, on the south coast, it is lower: 654.4 mm. (Source: Australian Government Bureau of Meteorology www.bom.gov.au)

Why computer scientist should study statistics?

Reason "all is about the data"

The benefit of studying statistics is in ***gaining a skill set that is transferable to other research methods, disciplines and walks of life****.*

Studying statistics encourages an ***approach to research that is reflective, thoughtful and mindful of the limitations of data and their analysis****. Encouraging the researcher to form* ***a clear and manageable research question, a means to answer the question, and awareness of the assumptions, methodological limitations and the researcher's own prejudices is a discipline conducive to all empirical work, quantitative or qualitative****.*

For statistical research, it is good practice to ask ***whether the results have both statistical and substantive meaning****. Focusing on the statistical* ***helps to avoid sensationalising events*** *that are either* ***potentially random or entirely predictable*** *–* ***to avoid claiming they are unusual when they are not.***

Three types of statistics

Inferential statistics

*Inferential statistics go beyond describing a set of data in its own right and **use the information to say something about the population from which the data were sampled.***

EXAMPLE: The UK Conservative Party (as part of the governing coalition) is predicted to have the support of 37% of the electorate against 41% for the Labour Party. **How does a polling company know this when it asked only 1500 people from an electorate of approximately 45 million? It does not know, not for certain, but it can make an inference if it believes the *sample is representative* of the electorate as a whole.** Being representative means those who have been polled are not disproportionately located in, say, Labour-voting industrial areas or Conservative-voting rural areas. The poll is not bia

Three types of statistics

Relational statistics and potentially explanatory

*At a minimum they identify **whether particular events or circumstances coincide**.*

EXAMPLE: Does shaving less than once a day really increase a man's risk of having a stroke (BBC News online, 7 February 2003)? Unlikely, but they could well be related, perhaps by lifestyle.

*Going further, relational statistics become explanatory when they offer firm insight on "**what causes what**".*

EXAMPLE: In an analysis of rural to urban migration in China, Wang and Fan (2006) identify some of the personal, social and economic reasons that can lead a migrant not to continue in a city but instead return home to the countryside. Chief amongst them is age (the older the migrant, the more likely they are to return) but important also is family responsibility – those who are married and/or have children are increasingly likely to return.

*Going further, if statistics are explanatory then they may also be **predictive**. A statistical model might be built to answer '**what if?**' questions.*

Three types of statistics

There are three types of statistic: those that **describe** and summarise a set of data in its own right; those that 'go beyond' the data to **infer** something more general about the population from which the data were sampled; and those which examine **relationships**.

Analysis and errors

Analysis

*The Oxford English Dictionary gives one definition of it: '**a detailed examination of the elements or structure of something**'.*

*For statistical analysis we **examine data** – information that can be stored, retrieved, queried, summarised, classified and used for calculations by a statistical package running on a computer. The data will often be **numeric but not always**.*

EXAMPLE: Use a word processor to count how many times the word statistics appears in this lecture. The result is numeric but the objects of the analysis are words.

*When we examine data we do so because we are interested in **what the data represent**. If they are **measurements**, then our interest is in **what has been measured**.*

EXAMPLE: If I measure the heights of 6, 7, 8, 9, 10-years-old children, I am not merely interested in the numbers. I am interested in **what those numbers tell me about my children's heights** (for example, how much they have changed since the last time I made a measurement).

Analysis and errors

Errors

*The problem is that **any measurement is subject to error**.*

EXAMPLE: The children not standing straight, the tape measure warping, myself not aligning it vertically, and so forth.

An analogy is often used from physics and engineering.

EXAMPLE: Consider a radio wave carrying an analogue transmission of a news broadcast. A number of factors affect the quality of what you hear, including the quality of the radio, the size of its aerial, the distance from a transmitter, atmospheric conditions, and so forth. What you want to hear is the news – **the signal**. Any background hiss is *unwanted **noise***.

*A challenge of statistical analysis is **to separate 'the noise from the signal'**.*

Analysis and errors

Errors

EXAMPLE: Imagine we used a **questionnaire-based survey** to gather data about business and manufacturing for a study about regional economics and competitiveness. What we have to accept is that the **data obtained are imperfect**. **People may misunderstand the questions or make an error when completing a tick box. We may miscode the data or ask questions that are not fully relevant.** Nevertheless, we hope to take the data and use them **to understand** why some regions prosper whilst others decline.

There will always be noise in data.

EXAMPLE: Take a ruler and measure the length of the following: _________. How long is it? Can you repeatedly get the same exact answer twice? Probably not, because it depends on exactly where you position the ruler, where your eye adjudges to be the end of the line, the ruler being perfectly aligned with the line, a lack of hand shake, no creasing of the paper, and so forth.

Obviously it has length. And you can probably measure it well, just not definitively because you cannot avoid the chance of error. Also, you need to be wary of systematically repeated error: for example, measuring the line **in centimetres** but thinking they are **inches**; or using a plastic ruler that has warped in the Sun.

Analysis and errors

Errors and variability

*Because **noise** – **error** – creeps into any analysis, no single answer, measurement or result will be perfectly definitive (and you would have no way of knowing even if it was).*

*Put another way, **if an experiment is repeated twice or more, there is no guarantee the same exact result will be obtained each time**. Instead, we expect the opposite: **to find variability in the results even if the structure and properties of that being analysed do not change**.*

*However, error is not the only reason to expect data to vary. Social, economic and environmental systems themselves create **outcomes that are not the same everywhere**. That is **science of data**!*

EXAMPLE: Mean temperature in New York is not the same as in Melbourne.

EXAMPLE: Property prices in Manhattan are not the same as in The Bronx.

EXAMPLE: Atmospheric pollution in Los Angeles is different to that in Hawaii.

*Acknowledging this **variation (variability)** does not mean the issue of error goes away. It adds to the challenge of explanatory data analysis: to find and explain **real variations (variability)**, not those due to error.*

Analysis and errors

Errors is unavoidable

Error creeps into any data collection and analysis – **it cannot be avoided because there are so many ways it can arise**:

1. a faulty instrument,
2. misreading,
3. data wrongly coded,
4. 'rounding' errors (for example, treating 1.7 as 1),
5. tiredness,
6. interference (for example, poor atmospheric conditions);
7. misunderstanding the research question; and so forth.

The hope is that **the effects of error will be neutral overall** – that they will not seriously distort what is learned from the data.

The worst errors are those that are

- hard to detect,
- difficult to correct and
- have a large impact on the analysis.

Key points

1. Statistics have three main purposes: to be descriptive, inferential or relational.
2. Knowledge of statistics is important because data collection and analysis are central to the functioning of society.
3. Statistical practice encourages a research rigour and reflectivity that is useful for both quantitative and qualitative research.
4. Users of statistics do so for a variety of reasons and with a range of motivations, beliefs and philosophical perspectives.
5. Data analysis is about detecting, examining, understanding and predicting events and phenomena that are biological, medical, geographical, technical, economical etc. in nature. It tries to separate what is true of those features from error in the data.
6. It is a good idea to ask whether the results of research have both statistical and substantive meaning, probing beneath the surface of what the data appear to be telling you and making links to academic debate and theories.

Statistical vocabulary

- **Population** – large, infinitely large set of statistical units (individuals)
- **Sample** – ***sampling*** statistical units from a population
- **Variable** – ***measuring*** its values (**realisations**, **observations**, **data**, information about variable), ***maximising*** measurement precision
- **Error** (**noise**) – systematic or random (imprecise measurement; randomness in the data), ***minimising*** errors
- **Signal** – searching for a signal, ***maximising* signal-to-noise ratio**, ***separating*** noise from the signal

Statistical vocabulary

- **Variability** – variability (variations) in the data, its **size** and **direction**, ***searching*** for sources of variability
- **Parameter** – a variable characteristics, ***estimating*** it from a sample
- **Estimate** – ***estimating*** parameter (**parameter estimate**), underestimate, overestimate, correct estimate
- **Bias** – sources of bias, ***minimising*** bias, estimation bias, biased estimate
- **Description** and **analysis** of the data
- **Statistics** – **descriptive** (***describing*** data), **inferential** (***inferring*** about parameters, based on data but about situation in a population, generalisation), **relational** (***distinguishing*** artifactual and true relationships)