MSc Research Skills

Topic: Statistical inference for research

D G Rossiter

University of Twente.
Faculty of Geo-information Science & Earth Observation (ITC)

June 26, 2012

Topics

1. Quantification and the inferential paradigm
2. Foundations of statistical inference
3. Bayesian concepts
4. Frequentist concepts and hypothesis testing
5. Statistical modelling

Topic: Quantification and the inferential paradigm

"When you can **measure** what you are speaking about, and express it in **numbers**, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be."

– William Thompson (Lord Kelvin), 1883, *Popular Lectures and Addresses*

"Statistics . . . may best be considered as the determination of the **probable** from the **possible**."

– JC Davis, 2002, *Statistics and data analysis in geology*

Example of qualitative statements

- "The projective transformation can successfully georeference a small-format air photo (SFAP) from ten ground control points measured with a single-receiver GPS".
- "In villages where a participatory GIS was developed there was less conflict between government plans and local goals."
- "Shifting cultivation systems have expanded in the past ten years, mainly at the expense of primary forest."

Example of quantitative statements

- "Primary forest covers 62% of the study area."
- "On 10–September-2000 Lake Naivasha contained $8.36 \cdot 10^9 m^3$ of water."
- "Twice as many boys as girls attend secondary school in District X."

Problem: In almost all cases, it is impossible to observe (measure) the entire population of interest.

Inferential paradigm

- We have a **sample** which **represents** some **population**;
- We want to make a quantitative statement about the population;
- This requires us to **infer** from sample to population.

"Statistics is the branch of applied mathematics that studies ways of **drawing inferences** from **limited and imperfect data**.

We have some data . . . , but we know that our data are **incomplete**, and experience tells us that repeating our experiments or observations, even taking great care to replicate the conditions, gives **more or less different answers** every time. It is foolish to treat any inference from only the data in hand as certain." – Shalizi, *American Scientist* 98:186

Fact vs. inference

"Facts": (with some **measurement uncertainty**):

- "The median sigma of georeferencing of 14 photos was 5.16 m".
- "Participants in the workshop had from two to ten years of formal education."
- "Twelve of the 40 crop fields surveyed in 2004, with an area of 6.3 ha out of the 18 ha total crop land surveyed, were covered by primary forest in 1990".

These are examples of **descriptive** statistics: summarizing many numbers for interpretation.

(continued . . .)

Fact vs. inference (continued)

Inferences:

- "The median sigma of georeferencing with the projective transform is no greater than X m"
- "Small farmers in the district have from X to Y years of formal education."
- "X% of the crop fields active in 2004 and Y% of their area were covered by primary forest in 1990".

These are examples of **inferential** statistics: making a probabilistic statement about the **population**, based on the observation of some **sample**.

(see next).

Populations and samples

- **Population** ("universe of discourse")
  - the set of objects about which we want to make a statement
  - most have not been observed
  - difficult to precisely specify (limit) the population
- **Sample**
  - the portion of the population that has been observed (measured)
  - it must somehow **represent** the population

Sampling design

- The relation between population and sample is called the **sampling design**
- It is specified by the researcher
- The relation must be **explicit** and **operational**
  - What is the **sampling frame?** (list of possible sampling units)
  - How are the **sampling units** selected from the frame?
  - How are the sampling units **identified** in the field?
  - How are the **target variables** measured?

Representing the population with the sample

This is the **basis of statistical inference** – the researcher must be able to:

1. explicitly **identify** the population of interest – the inferential statement is made about **this population** and no other;
2. argue that the **sampling frame** represents the **population**;
3. describe the relation between the actual **sample** and the **sampling frame** – in particular, the *a priori* **probability** of selecting each potential sampling unit.

Note that the third step comes from the nature of the sampling design, but the others require **meta-statistical** arguments.

Topic: Foundations of statistical inference

Deep questions:

- What is the meaning of **probability**?
- What does an **inferential** statement really mean?

Two principal interpretations:

- **Frequentist**, also called **classical** or **British-American**;
- **Bayesian**.

History

- **Frequentist**
  - R A Fisher at Rothamstead Experimental Station (England), 1920's and 1930's
  - developed by well-known workers (Yates, Snedecor, Cochran . . .)
  - Common statistical computing "packages" follow this
- **Bayesian**
  - named for Thomas Bayes (1701–1761) but developed since the 1960's (Jeffreys, de Finetti, Wald, Savage, Lindley . . .)
  - simple applications of Bayes' theorem are not controversial
  - requires sophisticated computing

Principal differences

- Interpretation of the meaning of **probability**
- Hypothesis testing
- Presentation of probabilistic results (e.g. confidence intervals)
- Computational methods

Topic: Bayesian concepts

For a Bayesian, **probability** is:

- the **degree of rational belief** that something is true;
- "rational", so certain rules of consistency must be followed;
- All probability is **conditional** on evidence;
- Any statement has a **probability distribution**: any value of a parameter has a defined probability;
- Probability is continuously **updated** in view of new evidence.

So, there is a degree of **subjectivity**; but this is reduced as more **evidence** is accumulated.

Types of probability

- **Prior** probability: before observations are made, with previous knowledge;
- **Posterior** probability: after observations are made, using this new information;
- **Unconditional** probability: not taking into account other events, other than general knowledge and agreed-on facts;
- **Conditional** probability: in light of other information, specifically some other event(s) that may affect it.

Bayes' rule (simple form)

Bayes' Rule is used to update a **prior** probability $P(A)$, based on new information that an event B with prior probability $P(B)$ has occurred, and knowing that the conditional probability $P(B|A)$ of B given A, to a **posterior** conditional probability $P(A|B)$

$$P(A|B) = P(A) \cdot \frac{P(B|A)}{P(B)} \quad (1)$$

The last factor is the proportion by which the prior is updated, sometimes called the **likelihood function**.

1. Estimate $P(B|A)$: how likely it is to observe B when A is true)
2. Estimate the priors $P(A)$ and $P(B)$
3. The probability of A is updated by the information that B was in fact observed.

   This only works if $P(B|A)$ is a reliable estimate.

Example – medical diagnosis

- Patient has a fever (event B)
- suspect that patient has malaria (event A)

To calculate the probability that the patient in fact has malaria, need to know:

1. The **conditional** probability of a person with malaria having a fever, $P(B|A)$, which we estimate as, say, 0.9 (some people who are infected with malaria don't have a fever);
   - Estimate this from a large number of previous confirmed malaria cases; which proportion of them presented a fever?

   (continued . . .)

2. The **unconditional prior** probability $P(A)$ of having malaria, i.e. the proportion of the population that has it, say 0.2; this is the **prior** probability of having malaria before looking at our symptoms;
   - Estimate this from surveys of malaria prevalence.
3. The **unconditional** probability of having a fever from whatever cause, say $P(B) = 0.25$.
   - Estimate this from a large number of previous cases with any diagnosis; what proportion were presented with fever?

Medical diagnosis – continued

By **Bayes' rule**, compute the **posterior** probability that, given that an individual has a fever, that they have malaria:

$$P(A|B) = 0.2 * (0.9/0.25) = 0.72.$$

The probability of malaria has been greatly increased from the prior (0.2) because the presence of fever is so closely liked to the disease.

The **likelihood function** was thus $0.9/0.25 = 3.6$; the odds increased by 3.6 times in the presence of the information about the symptom.

General form of Bayes' rule

Sample space A of outcomes can be divided into a set of **mutually-exclusive** outcomes $A_1, A_2, \ldots$.

Then the conditional (**posterior**) probability of any of these outcomes A_i, given that event B has occurred, is:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)} \quad (2)$$

Example of general Bayes' rule

Example: land cover class at a particular location; one of several legend categories.

1. The prior probability $P(A_i)$ of a location belonging to class i is estimated from prior knowledge of the area to be mapped, perhaps a previous map or even expert opinion.
2. The conditional probability $P(B|A_i)$ of some event (such as an aspect of a spectral signature) in for all possible land must also be given either from theory or statistical estimation.
3. Compute the **posterior probability** that a given location is in fact in the given class. This is precisely what "Bayesian" image classification algorithms do.

Topic: Frequentist concepts

The frequentist interpretation of **probability** of an event is is the **proportion** of time it would occur, should the experiment that gives rise to the event be repeated a large number of times.

- Throwing dice: no problem, the experiment can be repeated indefinitely
- Agricultural yield trial: could have selected other locations, can repeat in other (similar?) years
- Probability of a large meteor hitting the Earth within the next ten years?
- Probability that the human species will make itself extinct within the next ten years?

Frequentist hypothesis testing

The **null** and **alternate** hypotheses:

- The **null** hypothesis H_0: Not rejected until proved otherwise ("innocent until proven guilty"); if the evidence is not "strongly" against this, we can't reject it.
- The **alternate** hypothesis H_1: Something we'd like to prove, but we want to be "fairly sure".

The "null" hypothesis does not have to be "no difference" or "no effect".

Classification of inferential errors

Two types of inferential errors:

Type I : **rejecting** the null hypothesis when it is in fact **true**; a **false positive**

Type II : **not rejecting** the null hypothesis when it is in fact **false**; a **false negative**

Significance levels

Quantify the risk of making an incorrect inference. These are of two types:

- α is the risk of a **Type I** error, i.e. a **false positive**;
  - "The probability of convicting an innocent person" (null hypothesis: innocent until proven guilty)
- β is the risk of a **Type II** error, i.e. a **false negative**
  - "The probability of freeing a guilty person"
- The quantity $(1 - \beta)$ is called the **power** of a test to detect a true positive.

Types of errors

	Null hypothesis H_0 is really ...	
Action taken	True	False
Reject	Type I error committed	success
Don't reject	success	Type II error committed

Conventional significance levels

Conventional levels of α, with common-language equivalents of how sure we are that the null hypothesis can not be accepted:

- "Marginally Significant" : $\alpha = 0.1$ ("maybe")
- "Significant" : $\alpha = 0.05$ ("fairly sure")
- "Highly Significant" : $\alpha = 0.01$ ("very sure")
- "Very Highly Significant" : $\alpha = 0.001$ ("extremely sure")

These must be balanced depending on the **consequences** of making each kind of error – it is a **subjective** decision.

Convoluted metastatistics

In frequentist thinking we can never "accept" a hypothesis; all we can say is that we don't have sufficient evidence to reject it.

We can never say that it's probably true, only that it's probably not false.

Deciding on a significance level

- α is set by analyst – it is a **subjective** decision;
- β depends on the form of the test, the true difference, and the variance of the data
  - inherent in the phenomenon (uncontrollable)
  - due to imprecise measurements (controllable)

Some journals or funding agencies have defined levels for certain kinds of studies.

Example

Null hypothesis: a new crop variety is no better than the current one.

- The cost of introducing a new crop variety if it's not really better, and the lost income in case the new crop is in fact worse (Type I error), vs.
- The lost income by not using the truly better variety (Type II error)

Topic: Building a statistical model

A **statistical model** is an **empirical relation** between:

- one or more **responses** ("dependent" variables), and
- one or more **predictors** ("independent" variables).

It **summarizes** relations and also allows **predictions**.

- e.g. correlation, regression, classification, ordination . . .

A provocative example

Reference: Tatem *et al.* Nature 431:525

"Momentous sprint at the 2156 Olympics? Women sprinters are closing the gap on men and may one day overtake them"

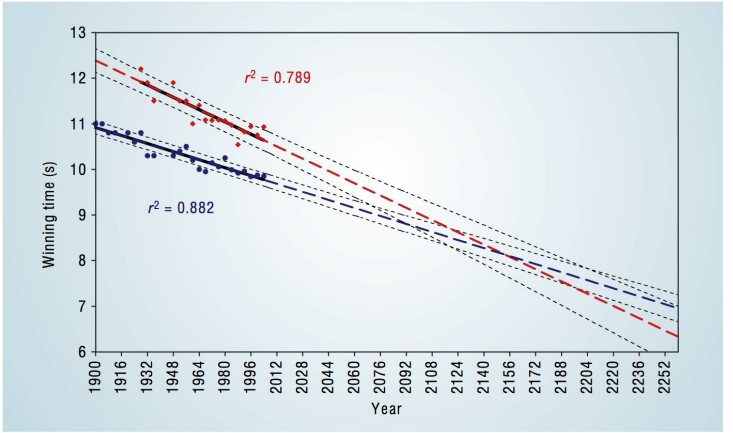


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Questions about this example

1. How to **interpret** this graph? (i.e. read the axes, say in words what the graph shows)

2. What conclusions can we draw from it?
   (a) Within the time period of observations (**interpolation**)
   (b) For the future (**extrapolation**)

3. How can we **explain** the interpretation? I.e. what could be the **causes** of the observations and fitted model?

4. Is there reason to suspect the extrapolation?
   (a) From internal evidence (model fit)
   (b) From external evidence (real-world knowledge)

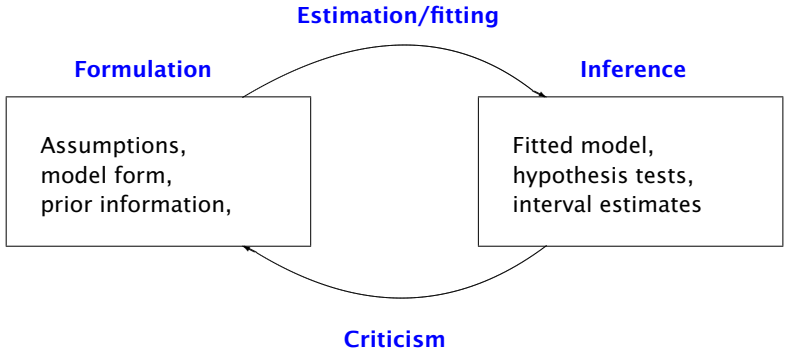
Partial answers

1. The graph shows:
   (a) At every Olympics, men have been faster than women;
   (b) Both men and women are improving;
   (c) Women are improving faster than men;
   (d) The improvement is linear (no other curve shape has a better parsimonious fit);
   (e) The linear fit is better for men than for women (R^2) because (a) the women's times are more variable around the line; (b) women have participated in fewer Olympics
   (f) Some interesting details of individual performances.

Steps in modelling

1. Selecting a **functional form**, i.e. the model to be fitted;
2. Determining the **parameters** of the model; this is called **calibration** or **parameter estimation**;
3. Determining how well the model describes reality; this is called **validation**.
4. **Criticising** (re-examining) the assumptions and possibly re-cycling.

The modelling paradigm



– after Cook & Weisberg (1982) *Residuals and influence in regression*

Note **criticism** of the **assumptions**, especially **model form**.

Structure vs. noise

- Observations = f(Structure, Noise)
- Observations = f(model, unexplained variation)

Observations are a subset of **Reality**, so:

- Reality = f(Structure, Noise)
- Reality = f(deterministic **processes**, **random** variation)

The aim is to match our **model** with the true **deterministic process** . . .

. . . and match our estimate of the **noise** with the actual **random variation**.

It is equally an error to model the noise (**over-fit** the model) as to not model the process (**under-fit** the model).

Evidence that a model is suitable

Two levels of evidence:

1. **external** to the model:
   (a) what is known or suspected about the **process** that gave rise to the data
   (b) this is the connection to the **reality** that the model is trying to explain or summarise;
   (c) how well the model fits further data from the same population: success of **validation** against an independent dataset
2. **internal**: from the model itself:
   (a) how well the model fits the data (success of **calibration**);
   (b) how well the fitted model meets the **assumptions** of that functional form (e.g. examination of regression diagnostics).

Model calibration

(Also called **parameter estimation**)

Goodness-of-fit: how well the model matches the calibration data.

For linear modelling: R^2 (the **coefficient of determination**), the complement of the **residual sum of squares** (RSS) as a proportion of the **total sum of squares** (TSS):

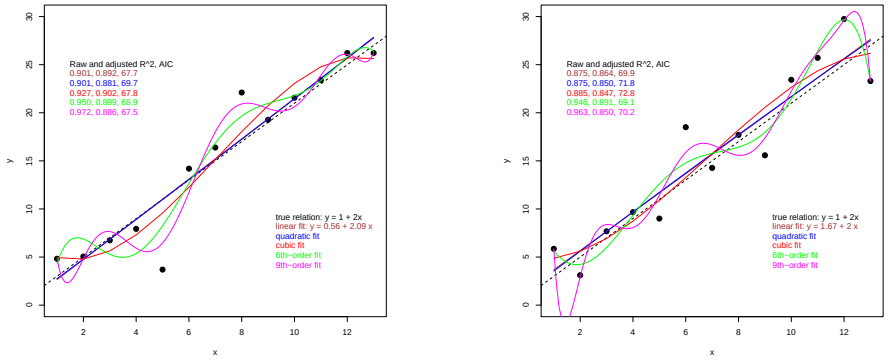
$$\begin{aligned} R^2 &= 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \\ RSS &= \sum_{i=1}^{n} (z_i - \hat{z}_i)^2 \\ TSS &= \sum_{i=1}^{n} (z_i - \bar{z})^2 \end{aligned}$$

where $\hat{z}_i$ is the predicted (modelled) value and $\bar{z}$ is the mean response.

Parsimony

- Principle: the **simplest** relation that explains the data is the best
- especially applicable in **multiple regression** models
  - unadjusted $R^2 \rightarrow 1$ as more predictors are added
  - but some of this is fitting **noise**, not **relation**

Fit vs. parsimony



Same true relation, different noise → different empirical fits

Higher-order fits always have lower R^2; this is compensated by the adjustment.

Correcting for over-fitting

For linear models, use the **adjusted** R^2 in place of the un-adjusted coefficient of determination.

This decreases the apparent R^2, computed from the ANOVA table, to account for the number of predictive factors:

$$R^2_{\mathrm{adj}} \equiv 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

The proportion of variance not explained by the model $(1 - R^2)$ is **increased** with the number of predictors k. As n, the number of observations, increases, the correction decreases.

Can also use **information-theoretic** measures, e.g. the **Akaike Information Criterion** (AIC).

Model validation

Given an **independent sample**, apply the model and compare the predictions ($\hat{y}_i$) with reality (y_i):

- **Root mean squared error** (RMSE) of the **residuals**: the **actual** (observed) less the **estimate** (from the model) in the **validation** dataset; lower is better:

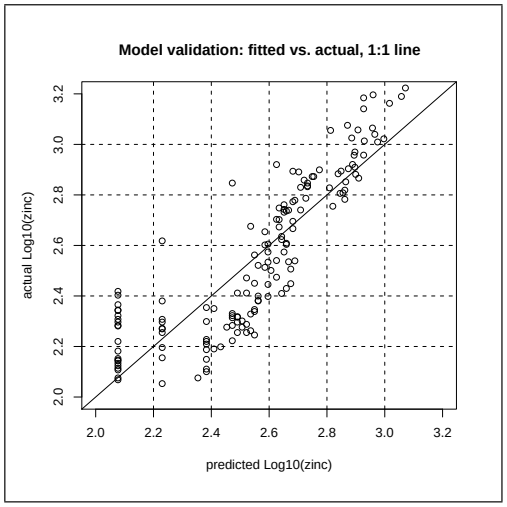
$$\text{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right]^{1/2}$$

- **Bias** or mean error (ME) of estimated vs. actual mean of the **validation** dataset; should be zero (0) if the model was supposed to be unbiased:

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$$

- **Gain** of the least-square fit of estimated vs. actual data; this should be 1, otherwise the estimate does not increase at the same rate as the actual data.

Validation against a 1:1 line



To be continued ...

There is **much more** to this story!!