Sampling uncertainty in medical research

By Jonas Ranstam

Statistical analysis plays a central role in medical research. With exception for case reports, p-values and significance statements appear in almost all published papers. However, deeper insight into statistical principles seems not widespread (1). P-values and significance statements often appear to be interpreted as obscure properties of examined patients; differences between groups exist only if statistically significant; the interpretation of clinical and statistical significance is confused, etc.

Understanding basic principles is necessary both for statistically analysing data and for interpreting the results of the analyses. Without this understanding severe mistakes are easily made. A few minutes spent on statistical fundamentals may be an investment with an exceptionally high return of the invested time. This is the reason OAC plans a series of short methodological reviews, of which this is the first.

Random sampling

Human subjects have varying characteristics. Repeated sampling of subjects produces samples with slightly different properties. The degree of the heterogeneity depends on the distribution of the characteristics in the population and on sample size. The potential influence of sampling uncertainty on an observation can be calculated using the information in a sample and is usually described by a p-value or a confidence interval.

The calculation of sampling uncertainty is based on the assumption that the observed data have been generated by random sampling from a defined population. Random sampling has a clear definition: All units of the population have had a known (usually equal) probability of being sampled and all units have been sampled independently.

In the world of clinical and epidemiological research study populations are typically not sampled randomly, because there is usually no population with known sampling probabilities to sample from. Data are instead collected where they happen to be available, for example among consecutive patients, neighborhood controls, regional or national registers, etc. Such study populations can be characterized as convenience samples (2).

No sampling uncertainty exists when a sample is assumed to constitute an entire population of its own. P-values and confidence intervals can then of course not be interpreted as measures of sampling uncertainty. However, imperfection in measurement instrument and clerical errors in data registration, may still cause uncertainty in observed findings. This measurement uncertainty may have formidable consequences (3), why an evaluation using statistical methods may be pertinent.

When attempting to understand clinical findings in a sample of carefully examined patients and generalizing these findings to other individuals, which have not been included in the examined sample, sampling uncertainty can be seriously misleading and should always be evaluated. The required statistical methods are based on assumptions of random sampling, also for convenience samples. It is then simply assumed that the selection mechanism behind the specific convenience sample have identified a group of subjects, which in all respects can be considered a random sample.

In clinical and epidemiological research the underlying assumptions of known sampling probabilities and independent sampling play critical roles in the development of a sound study

design.

Types of populations

Problems arise when a convenience sample has different properties than a hypothetical random sample would have had. These problems can easiest be explained vis-à-vis the hypothetical population from which the two samples could be assumed to have be drawn. Two different types of populations can be considered: finite- and super-populations. Let us consider a fictitious convenience sample and discuss what populations this sample could have been sampled from.

Assume that 50 consecutive patients being treated for osteoporosis with a particular treatment have been sampled from Lund University Hospital during March and April 2005. Figure 1 presents the presented phenomena graphically.

The finite population approach

A finite population is real. It could be defined as all patients treated for osteoporosis with the particular treatment at Lund University Hospital during 2003-2007. All these patients exist and can be identified. They have visited the hospital in person and been registered there. If the patients in the sample differ systematically from patients not sampled, for example if male patients have a sampling probability of 0, generalisations from the convenience sample would lack external validity (representativity).

Patients treated during March and April 2005 could also be more alike, with respect to some specific characteristic, than patients treated during other time periods. For example, they may have had a decreased probability of fragility fractures because the winter was unusually mild. If this decreased the variation of bone mineral density among the patients in the sample, this could be seen as a cluster and a violation of the independent sampling assumption. Variance estimates could then be underestimated.

Generalizations beyond the defined finite population would of course be inadequate. Finite population definitions are common in censuses but exceptions in medical research. This can be easily verified by checking if a finite population is explicitly stated or if the statistical calculations include adjustments for the fraction of patients sampled.

The super-population approach

In contrast to finite populations a super-population is always hypothetical. It constitutes an infinite universe of imaginary units. In our example it could be the set of all patients that have been, will be, or could have been, treated for osteoporosis with the particular treatment. The assumptions of representativity and independence are as important for this superpopulation as for the finite population. If the assumptions are fulfilled a convenience sample leads to the same result as a random sample from the super-population.

Figure 1 about here

Two random samples of 50 patients each (the grey rectangles) have been drawn from the finite population. While the finite population has 10% white dots, sampling uncertainty is manifested in the sampled units by the varying proportion of white dots, one sample having8% (4 of 50), the other one 16% (8 of 50).

As the super-population is infinitely large it can not be described graphically, but if the two

finite samples can be assumed to be randomly sampled from the super-population, its proportion of white dots can be estimated to 8% (2.2% - 19.2%) and to 16% (7.2% - 29.1%) respectively. The range within brackets are 95% confidence intervals describing the sampling uncertainty.

The interpretation of an observation in a sample

While sampling uncertainty may be easily recognized as an important concept in itself, its consequences for the interpretation of observed findings is, however, often neglected. For example, in a randomized clinical trial with patients who had a single chronic symptomatic cartilage defect on the femoral condyle (4), forty patients were treated with autologous chondrocyte implantation, and forty were treated with microfracture. At the five-year follow-up, there were nine failures in each of the two groups. The authors observed and concluded that "there was no significant difference".

However, even if no difference can be observed in the sample, there may well be one in the population. Sampling of patients with individually varying failure rates may prevent a true average treatment difference from being observed in the sample.

The sampling uncertainty can be described by a 95% confidence interval for the estimated true failure rate ratio. Using the data presented in the paper the interval can be calculated as (0.3 - 3.2), which implies that observed data actually only speaks against an average failure rate difference in the population greater than about 300%.

A claim that the two treatments have the same failure rates would thus not have much empirical basis.

The population is important for the statistical analysis

Super-populations are seldom explicitly defined, but doing so may be a good exercise for avoiding unit of analysis errors. Many orthopedic papers present statistical analyses of joints or cells and of phantom measurements, and errors in analyses of hips, knees, hands, feet, shoulders and elbows appears, according to a recently published systematic review (5), to be a surprisingly common problem; of the 142 reviewed papers 42% involved unit of analysis errors.

What populations are involved in these studies? What relevance have the reasoning on superpopulations on the statistical analysis? Three examples of common study situations will be discussed.

First, an experimental unit, and the unit of analysis in the statistical analysis, is usually defined as the smallest amount of experimental material that can be independently assigned to a treatment. It can be patients in clinical studies, animals in in vivo studies and cell cultures in in vitro studies (6).

If independent subjects (or animals) contribute multiple observations to a sample, and within-subject variance is less than between-subject variance, observations are not independent. This violates the independence assumption of random sampling.

Example 1.

What superpopulation is relevant when studying two groups of 5 patients each, treated for osteoporosis with drugs A and B, who have had the increase in bone mineral density in percent (BMD) measured in their left tibia (L). Observations from 5 right tibiae (R) of the same patients

have also been included.

Drug	A	A	A	A	A	B	B	B	B	B	A	A	B	B	B
Patient	1	2	3	4	5	6	7	8	9	10	1	4	8	9	10
Tibia	L	L	L	L	L	L	L	L	L	L	R	R	R	R	R
BMD	4.2	2.1	3.3	4.5	8.7	6.1	9.9	5.4	6.1	8.1	2.9	5.1	3.7	6.3	7.7

Ignoring that some patients contribute two observations each, the number of observations is 15 and the between-patient variance in BMD 4.0.

When tested with Student's t-test, the 2.3%-units increase in BMD among patients treated with drug B, as compared to those treated with drug A is shown to be statistically significant, $p < 0.05$.

However, the dataset includes only 10 patients, and assessing variation among both patients and tibiae within patients using a mixed effects model (7) shows that Student's t-test underestimates the between-patient variance, because of the contamination with within-patient variance. Separating within- and between patient variance shows that the latter is 4.8, not 4.0. The intraclasscorrelation coefficient can be calculated to 0.9.

Furthermore, the correct analysis shows that the increase in BMD is slightly greater (2.4%- units instead of 2.3%-units) but fails to reach statistical significance. The correct p-value is 0.13 instead of <0.05.

The super-population of independent analysis units in this example is of course patients, not measurements.

Example 2.

The definition of experimental unit often causes controversies in experimental research. What super-population should be considered when statistically analysing cells?

Again, the principle is that the analysis unit should be the smallest amount of experimental material that can be independently assigned to a treatment. The principle is, unfortunately, not always appreciated in laboratory research.

It is, for example, reported in a systematic review of subfertility trials (8) that 82% of the 39 reviewed papers had unit of analysis errors. Live birth rates are often reported per started cycle, per oocyte collection and per embryo transfer, but not as appropriate per person rates. For a cell to be regarded as an independent experimental unit, it would have to be possible to assign independent cells to different treatments. Cells sampled and treated together with other cells in a cell culture on a dish are not independent, and would therefore not be meaningful as elements of the discussed super-population.

Independence between analysis units is the key. A specific recommendation for dealing with this in practice is to consider whether a proposed analysis unit really can be treated and sampled independently of other analysis units.

Sampling individual cells, or groups of cells, from a dish where all cells have been exposed to a common treatment does not generate independent samples. They are related by the common treatment, and should be regarded as repeated observations from this treatment, like BMD measurements from the right and left tibia of the same patient represent this patient's BMD, not different relationships between treatment and BMD.

The independence assumption is an issue that may need to be addressed thoroughly. A motivation for the definition of experimental unit is often, or should be, requested by reviewers during the peer review process.

Example 3.

What super-population corresponds to a series of measurements using a mechanical device especially developed for simulating a clinical phenomenon?

For evaluating sampling uncertainty, more than one device would have to be constructed. Let us assume that such variation of devices has no clinical relevance. P-values and confidence intervals related to only one experimental unit do not refer to sampling uncertainty but to measurement uncertainty, and it would probably facilitate many readers' interpretation of the presentation if the measurement uncertainty was expressed in terms of accuracy or reliability instead of p-values.

The adequate super-population in this example consists of the measurements that can or could be performed using the developed device.

Recommendations

When designing laboratory experiments, clinical trials and observational studies, consider carefully which sample unit is adequate and what population is at hand.

Then consider what sample size would be sufficient, without being unnecessarily large, to avoid an outcome that is inconclusive because of too much sample uncertainty. This can, and should, be calculated during the planning stages of a study.

In some cases collection of dependent data may be advantageous. Repeated measurements on analysis units can for example be used to provide a better estimate of between-subject variance than that available from a single measurement. Matching controls to cases in a case-control study to increase internal validity is another example.

The statistical analysis in such studies need, however, to account for the dependencies in data. Which statistical method to use for a specific purpose depends to a great extent on factors like study design, data structure, adequacy of distributional assumptions, and the needs to address multiplicity issues.

Selecting the right choice may not be trivial. The reader is recommended to discuss specific applications with a professional statistician.

When presenting data in a study with dependent observations, always describe both number of independent experimental units and the number of observations. This information is necessary for assessing the correct number of degrees of freedom.

When presenting results, use 95% confidence intervals to describe the sampling uncertainty. Confidence intervals are more informative than p-values, because the confidence intervals provide easily interpretable uncertainty limits in terms of the analyzed variable's measurement units.

Finally, the aim of this short methodological review has been to provide an overview on the nature of uncertainty in medical research. Basic statistical principles and their relation to uncertainty are important for both authors and readers. Understanding them is necessary for rational interpretation

and communication of scientific findings.

References

1. Glaser DN. The controversy of significance testing: misconceptions and alternatives. Am J Crit Care. 1999;8:291-296.

2. Lunsford TK, Lunsford BR. The Research Sample, Part I: Sampling. J Prosthet Orthot 1995;7:105-112.

3. Ranstam J, Wagner P, Robertsson O, Lidgren L. Healthcare quality registers: outcomeoriented ranking of hospitals is unreliable. J Bone Joint Surg Br 2008;90-B:1558-1561.

4. Knutsen G, Drogset JO, Engebretsen L, Grøntvedt T, Isaksen V, Ludvigsen TC, et al. A Randomized Trial Comparing Autologous Chondrocyte Implantation with Microfracture. Findings at Five Years. J Bone Joint Surg Am 2007;89-A:2105-2112.

5. Bryant D, Harvey TC, Roberts R, Guyatt G. How many patients? How many limbs? Analysis of patients or limbs in the orthopaedic literature: a systematic review. J Bone Joint Surg Am 2006; 88:41-45.

6. Lovell DP, Omori T. Statistical issues in the use of the comet assay. Mutagenesis 2008;23:171-82.

7. Pinheiro J, Bates DM. Mixed-effects models in S and S-PLUS. Springer, New York, 2000.

8. Vail A, Gardener E. Common statistical errors in the design and analysis of subfertility trials. Hum Reprod 2003;18:1000-1004

Legend

Figure 1. A super- and a finite population of osteoporosis patients (symbolized by black and white dots; white dots representing some clinically meaningful characteristic; the finite population defined by the continuous frame).

Figure 1.