Sampling Strategies and Costs Control

Dinis Pestana, Maria Luísa Rocha and Fernando Sequeira

To guess is cheap, to guess cheaply can be wrong and expensive.

Abstract Excellent data analysis methodologies fail to produce good results when using bad data. Bad data arise from inadequate strategies at the collecting stage, that are responsible for bias, or insufficient to produce accurate estimates of parameters of interest. Sampling is the statistical subfield that uses randomness as an ally in data gathering, the gold standard in ideal situations being to collect samples without replacement (thus each item bringing in new information, and as a consequence the estimator having reduced variance when compared to the corresponding sampling with replacement estimator). A quick overview of sampling strategies is presented, showing how they deal with cost control in non-ideal circumstances. Caveats on the use of immoderately large samples, on the reuse of samples, and on computational sample augmentation are illustrated, and other critical comments on misuse of statistics, are registered, in the hope that these alerts improve the obtention of statistical findings, so often blurred because sophisticated statistical analysis is useless since it uses bad data.

Dinis Pestana
Universidade de Lisboa, CEAUL and DEIO–FCUL,
Instituto de Investigação Científica Bento da Rocha Cabral, Portugal
e-mail: dinis.pestana@fc.ul.pt

Maria Luísa Rocha
Universidade dos Açores, Departamento de Economa e Gestão and CEEAplA, and
Universidade de Lisboa, CEAUL, Portugal
e-mail: lrocha@uac.pt

Fernando Sequeira
Universidade de Lisboa, CEAUL and DEIO–FCUL
e-mail: fjsequeira@fc.ul.pt

1 Introduction

Happily we can use statistical inference, since observing every population item is is virtually impossible even for modest size populations, and our task is to build knowledge from incomplete information. In fact, there are infinite populations, but performing a census of a whole population is time consuming and can be very expensive (the USA 2010 population census cost $42 per capita, a total of 13 billion!). And, at the end, bitter controversies follow, cf. *PS: Political Science and Politics* **33**, namely [3], [8], [16], *Statistical Science* **9**(4), namely [6], [13], and [49], just to cite a small bunch of papers debating the issue.

It is even arguable that the use of a large number of temporary census officers can endanger the quality of data gathering, and that appropriate sampling, much less expensive, could provide better results, since it uses a much smaller number of high quality trained professional officers. The complementarity of census and samples has also been used for instance in the 2010 U.K. population census, cf. the Office for National Statistics report [42].

No one would dispute that in all fields knowledge is based on partial information. At the end of the XIXth century, Galton was prescient of the importance of Statistics in tackling complex problems, and his well known statement meaning that Statistics is an intellectual swiss jackknife in cutting through the layers of difficult problems

> [Statistics are] *the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of Man.* (quoted by Pearson, in *The Life and Labours of Francis Galton*)

was indeed prophetic.

Yet, although the role of Statistics as the core body of the experimental method theory has been undisputed since Fisher developed Experimental Design, and Neyman and Pearson shaped the theory of statistical testing, *circa* 1930, three decades ago Statistics was seldom used in many branches of Science. Nowadays we have the reverse situation, no serious experimental science journal publishes a paper devoid of statistical analysis supporting the building of knowledge from information.

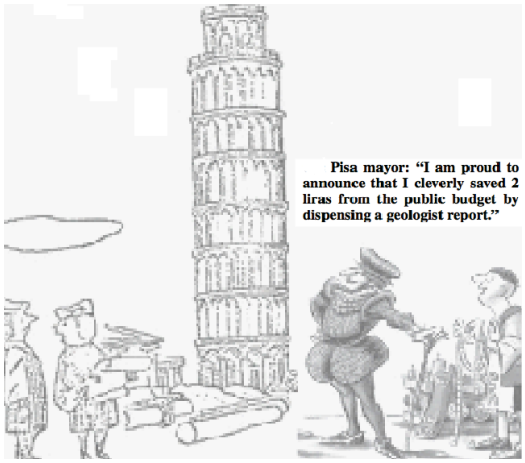
This progress is however controversial, in the sense that in many situations the data analysis is performed using bada data, and hence produces bad science. We can fear that in the near future contributors can mistrust science, since economic interests, incompetence, and the drive to publish hastily, together with a foolish blind trust in bibliometrics evaluation, [2] led to the publication of false conclusions [32]; perusing the documentation [31] on Ig Nobel prizes (that first make laugh, then [sometimes] think) is recommended, to laugh and to have a critical appraisal of the ways of modern academics. It is obvious from the above references, or from editorials of outstanding journals such as *Lancet* or *New England Journal of Medicine*, the ravaging effect of bad data.

Sampling theory has been developed to ascertain how to collect data. Gathering data is a crucial step in knowledge building, and Fisher joke on diagnosis and autopsy in his Presidential Address to the First Indian Statistical Congress, in 1938

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

cleverly points out at one of the principal causes of blunders in science.

Saving costs is a sensible goal, but unfortunately it is often done with the most inadequate strategy: saving the expense of consulting a specialist, and this is not always as rentable as the unexpected success of the Pisa tower.



Data gathering is a well developed statistics subfield, seldom effectively taught to researchers of other field. Observe that Greenfield [26] includes a chapter on principles of sampling in his noteworthy treatise on guidance for postgraduates, but that chapter naturally focuses only in the most general principles.

The main goal os sampling theory is to provide tools to obtain representative samples, cost being a natural concern, since all rigorous sampling operations are expensive. A representative sample must indeed reflect the heterogeneity of the population units, implying that the sample size must be large enough to reveal variability; but sample size, which is obviously an increasing function of the population dispersion and, in finite populations, of size, depends also on the sampling design, that must take into account questions such as: in sampling from finite populations, do we have a list of items in the target population? Is there the need to infer for separate subgroups of the population? Can we observe the variable of interest, or do we need to infer from observations of another variable, highly correlated with the one that interests us? Is it affordable to get a sample of the size needed with the optimal design?

In the sequel, we describe some of the most common designs, and circumstances advocating their use. But, roughly speaking, the idea is always: control everything you can, and what you cannot control you must randomise, since chance is the ultimate ally in data gathering. Collect as many data as you need, but no much more than that. Beware of bias, and if you are tempted to simulate samples investigate first whether the computational augmentation of samples is truly advisable.

2 Sampling from Finite Populations: the Gold Standard

Most monographs on sampling focus on design-based sampling from finite populations. Design-based, as opposed to model-assisted based, indicates that randomness intervenes through the probability of selection of items from the population to the sample (while in mode-assistedl based sampling, cf. [51], the assumption of some stochastic model for the population has some bearing on the constitution of samples; for a very elementary example assuming Poisson randomness, cf. [53] on quadrats sampling.).

When sampling from a finite population of size N in order to estimate some parameter of interest — often a function of the mean value μ, the mean value itself, or the total $\tau = N\mu$, and observe that a proportion p is the mean value of a Bernoulli model —, an important step is to choose a sampling strategy, and from that to decide the sample size needed to estimate the parameter with the degree of accuracy needed, at a given confidence level.

The simplest situation arises when the selection cost per unit, for planning purposes, is assumed as the same for each possible item, and on the other hand we are not interested in subpopulations.

Under that assumption, the *gold standard* is *srswr — simple random sampling without replacement*. Observe that sampling without replacement implies a mild form of dependency (exchangeability), and hence approximate confidence intervals rely on asymptotic results for sums of mildly dependent random variables.

This sampling strategy is unique in the sense that the probability of selecting any sample of size n is $\dfrac{1}{\binom{N}{n}}$. Observe also that selecting without replacement has an interesting consequence: any observation brings in new information, and that has as ultimate effect reducing the estimator variance.

In fact, when estimating the mean value μ from a simple random sample $X = (X_1, \ldots, X_n)$, with equal selection probabilities $\pi_s = \frac{1}{N}$, either with or without replacement, via the estimator

$$\widetilde{\mu} = \frac{1}{n} \sum_{k=1}^{n} X_k \tag{1}$$

(unbiased minimum variance linear), the variance when using replacement is $\frac{\sigma^2}{n}$, while without replacement there is a finite population correction reflecting the sampling fraction $\frac{n}{N}$, and we obtain smaller variance $\dfrac{\sigma^2}{n} \dfrac{N-n}{N-1}$ (which can be unbiasedly estimated by $\widetilde{V}(\widetilde{\mu}) = \frac{S^2}{n} \frac{N-n}{N}$, where S^2 is the sample variance).

Therefore, in case we wish to estimate μ with an error bound B at a $(1-\alpha) \times 100\%$ confidence level from a population whose standard deviation is σ, denoting $z_{1-\frac{\alpha}{2}}$ the $(1-\frac{\alpha}{2})$-quantile of the standard normal,

- when sampling *with* replacement, the requirement $z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < B$ implies that we should choose a sample size n_w

$$n_w > \frac{z^2_{1-\frac{\alpha}{2}}\,\sigma^2}{B^2}. \tag{2}$$

- when sampling *without* replacement, the requirement $z_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}}\,\sqrt{\frac{N-n}{N-1}} < B$ implies that we should choose a sample size $n_{\overline{w}}$

$$n_{\overline{w}} > \frac{z^2_{1-\frac{\alpha}{2}}\,\sigma^2\,\frac{N}{N-1}}{B^2 + z^2_{1-\frac{\alpha}{2}}\,\frac{\sigma^2}{N-1}} = \frac{N}{1+\frac{B^2}{\frac{z^2_{1-\frac{\alpha}{2}}\,\sigma^2}{N-1}}} \approx \frac{N}{1+\frac{(N-1)\,B^2}{z^2_{1-\frac{\alpha}{2}}\,s^2}}, \tag{3}$$

  an expression that clearly shows that the sampling effort when sampling without replacement should increase with the population size and the population variance.

As $n_w > n_{\overline{w}}$, the cost to achieve a fixed accuracy is reduced when sampling without replacement. For instance, for $N = 1927$, $\sigma = 5.38$, to guarantee an error bound $B = 0.5$,with confidence 95%, $n_{\overline{w}} \geq 362$, while if sampling with replacement we should use $n_w \geq 445$. For smaller values of the standard deviation, less than $\frac{NB^2}{80}$, say, for a population size of 1927, the size reduction is however very tiny, but sampling without replacement still saves costs mainly because it is in general easier and quicker to implement.

(Observe that the use of standard normal quantiles of appropriate probability in the case of sampling with replacement is a simple consequence of the classical central limit theorem, since the independence assumption is true. When sampling without replacement, the normal approximation is justified by the Erdös-Rényi central limit theorem extension assuming exchangeability, cf. [18], an information seldom explicated in sampling monographs.)

In many situations, a sample of size greater than $n_{\overline{w}}$ is collected, the goal being to overcome non-response, an important source of error and bias, since there is a tacit belief that statistical inference using large samples will be more accurate. This is trivially true, but irrelevant significance is a possible side effect of immoderately large samples, cf. the example in the Appendix; hence the recommendation is: $n \geq n_{\overline{w}}$ is a sensible guidance, do not exaggerate!

Non-response is in fact one of the most commonly encountered problems in practice, as observed from the very early days of modern sampling theory, cf. [28]. Dealing with non-response quite tricky, cf. [1] on the use of hierarchical models, and [52] on a developed introduction on multiple imputation to counteract non-response. General information on sampling and non-sampling errors and on non-response, issues that no one dealing with sampling can ignore, can be found in [9, 55, 35, 40, 4, 7] and references therein.

3 Deviating from the Gold Standard to Save Costs

Two main reasons to deviate from the *gold standard* are inhomogeneous costs per unit, or unavoidable drawbacks such as the absence of a sampling frame.

On the others, advantages may arise using special strategies, such as pooling units, or adapting the sampling scheme during implementation, according to whether it is providing good or bad results.

Another interesting possibility is to use regression, sampling something that is easy and inexpensive to collect or inspect in order to infer about something else that truly interests us but is difficult and expensive to sample.

3.1 Stratified and group sampling

Stratified sampling is advisable when we can partition the population in a small number of subpopulations. It is a combination of census (of the strata) and sampling, in general *srswr* (within each stratum). For instance, if we want to estimate the amount of claims payed by an insurance company that deals with life insurance, he alt insurance, household insurance, car insurance and travel insurance, it is advisable to partition claims according to the type of insurance originating it, since within each stratum we expect less variability than overall variability, and then to sample within each stratum. Strata estimates of the parameter of interested can then be used, using a simile of the total probability theorem, to build an overall estimate for the whole populations, with the extra advantages of having estimates for subpopulations, when this is needed.

Group (or cluster) sampling combines census and sampling the other way round: as this sampling strategy is fit for the case of many distinct subgroups, we first sample to choose randomly some of the subgroups (and this procedure may be repeated if needed), an at last we select to the sample all the units that can be observed in the chosen groups.

For instance, in a survey on alcohol consumption by university students, a first step may be s *srswr* of universities, to select 5 universities. Then, in each of them we choose 5 courses, again using *srswr*; after that, a dice is used to sample within each course: if 1 or 2 is drawn, the first year is chosen, if 3 or 4 is drawn the second year is selected, if 5 or 6 is drawn the third year is selected. The next step is to select one discipline of the drawn year of the course, to go to the next session of the discipline, and then to select all (census) students that consent to enter the study. This is considered to be the sampling strategy with best return, in the sense that the cost per unit is optimal. In general, it is considered convenient when within groups heterogeneity is big, while the cluster means are approximately equal.

Observe however that there might exist unaccounted dependences spoiling the results, and that on the other hand this type of data collection is often done having in mind the purpose of storing a wealth of information for future use. However, as we comment in Section 4, reuse of samples should be avoided.

So, although we recognise that cluster sampling cuts costs, we shall focus on stratified sampling.

Suppose that a population of size N can be classified in ν non overlapping strata, so that the numbers of items in the strata are $N_k,\ k = 1,...,\nu$. Suppose further that the sampling unit costs are $c_k,\ k = 1,...,\nu$, and that the standard deviation within each stratum is $\sigma_k,\ k = 1,...,\nu$. An unbiased estimate of the population mean is

$$\bar{x}_{st} = \frac{1}{N}\sum_{k=1}^{\nu} N_k \bar{x}_k \tag{4}$$

where the $\bar{x}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} x_{kj}$ are the unbiased mean strata estimators, i.e. $(x_{k1},...,x_{kn_k}$ is the *srsws* collected in the k-th stratum.

We shall consider sampling efforts within strata $w_k,\ k = 1,...,\nu$; in other words, if the sample total size is n, the within strata samples are of size $n_k = n\,w_k$. The approximate sample size n required to estimate the population mean with confidence $(1-a)\times 100\%$ is

$$n = \frac{\sum_{k=1}^{\nu} \frac{N_k^2\sigma_k^2}{w_k}}{\frac{N^2B^2}{z^2_{1-\frac{\alpha}{2}}} + \sum_{k=1}^{\nu} N_k\sigma_k^2} \tag{5}$$

and optimal allocations do exist to minimise costs for fixed variance strata:

$$n_k = n\,\frac{\frac{N_k\sigma_k}{\sqrt{c_k}}}{\sum_{j=1}^{\nu}\frac{N_j.\sigma_j}{\sqrt{c_j}}} \tag{6}$$

Just to give an idea of cost reductions: Assume that $N = 2000$, $N_1 = 800$, $N_2 = 1200$, that we consider an error bound $B = 1.5$ to estimate the mean of the population, at a confidence level 0.95, that the cost of sampling per unit is $c_1 = 1$ is subpopulation 1 and $c_2 = 4$ in the second suppuration. Assume further that the standard deviations are respectively $\sigma_1 = 3.93,\ \sigma_2 = 2.86$. Using *srswr*, $n = 19$ units would be selected at random in the whole population, and so the expected cost would be 53.2, while using stratified random sampling we would select $n_1 = 14$ in the first subpopulation, $n_2 = 8$ in the second population, at a total cost 46.

Stratified sampling is highly recommended when within strata heterogeneity is small, and on the other hand heterogeneity of strata means is considerably higher.

3.2 Ratio and other regression estimators

It would be very expensive to monitor daily the glucose level in my blood if this sould be done by traditional chemical analysis. Fortunately I own a small device

that "reads" the glucose blood level — in fact, as I learned from the accompanying booklet, what it reads is the refraction angle of light in my blood. This way and time expensive operation is replaced by a very simple indirect reading, since scientists have been able to relate glucose blood level and refraction angle of the light in the blood.

When sampling is difficult, expensive, or even impossible in large scale, for instance forest wealth, in general there is a possibility of quite moderately sampling *pairs* (x_k, y_k) where y_k is a measurement that in fact interests us but is difficult or expensive to execute, x_k the measurement of a related variable that is quite easy and inexpensive to obtain (for instance, actual number of grown up trees in a randomly chosen circle of radius 10m, and the correspondent count in aerial photography. In many situations, the second measurements is proportional to the first one, $\frac{Y}{X} = R$, and this ratio can be estimated by $r = \dfrac{\sum_{k=1}^{n} y_k}{\sum_{k=1}^{n} x_k}$.

As sampling X is easy and relatively inexpensive, we can estimate the mean $\overline{x}$ of a much larger sample, and then estimate the mean $\overline{y}$ by $r\overline{x}$).

This *ratio estimation* is just a special case of regression estimators, the general idea being to use correlation to perform indirect measurements, i.e., to measure something that is simple and cheap to measure, and then using regression to transform it in the measurement we need. If instead of proportionality we suspect deviation (such as it often happens in accountancy), we investigate the special case of linear regression of the form $Y = X + b$, a technique known as *difference estimation*. For further details, cf. any of the recommended books in the bibliographic comments.

3.3 Systematic sampling

The unavailability of a sampling frame hinders *srsw*, but we can use alternative designs to randomly select units from the population. For instance, we may be interested the average caloric content of lunches of university students using a campus restaurant, but there is no listing of the population, so *srswr* is out of question.

One of the most advisable ways of dealing with the problem is the following: suppose we wish to take some measurement on 5% elements from the population, and that we can have sequential access to the members of the population. We can then select a random integer in $\{1, 2, \ldots, 20\}$, say 17, and therefore to select to the sample the 17th, the 37th, ..., the $(20k + 17)$th element from the sequence, $k = 1, 2, ..., n$, with n determined by $N \in [17 + 20n, 17 + 20(n + 1))$.

Observe that with this systematic sampling strategy, the probability of selection is the same for all individual units, but that the probabilities of selecting different samples of the same size are radically diverse, in fact 0 for most subsets of the population.

Systematic sampling is an interesting alternative in the absence of a sample frame, since we can fix *a priori* the sampling effort, and it emulates quite well

simple sampling in two extreme circumstances: when there is none structure in the sequence of population items as we observe it, or, on the other hand, when the units roughly appear in monotone order. In the hospital balance of debts, for instance, where chronological ordering of files may be correlated to costs, this sampling strategy provides rich information, since is balances the representativity of debts from several periods.

Obviously systematic sampling can be very misleading when sampling from periodic sequences, namely when $\frac{1}{f}$, where f stands for the sample fraction, is approximately the period. On the other hand, if $\frac{1}{f}$ is much smaller than the period, systematic sample can provide some interesting insights, since it artificially creates something similar to post-stracta or to clusters. For instance, many phenomena have a 12 months period, and a $f = \frac{1}{4}$ sampling effort will provide interesting quarterly data.

Saving the cost of building up a rigorous and manageable sampling frame is justification enough to use systematic sampling instead of *srswr*.

3.4 Combined, sequential, and adaptive sampling

Combined and sequential sampling have been developments contributing to the US II World War effort, and sequential analysis developments, because of their impact in quality control, have been classified information until the end of the war, so that the original Wald paper [59] publication has been delayed until the hostilities ceased. The idea is to dispense with fixing in advance the sample size; instead, data are evaluated as they are collected, and further sampling is stopped in accordance with a pre-defined stopping rule as soon as significant results are observed, and this of course is bound to save costs in many instances. Observe however that this is not properly a random sampling strategy, contrarily to the other that we consider. However, as it can be considered a intercessor of more modern adaptive sampling techniques, and it produces in fact important results, this brief mention seems worth keeping.

In the early forties, the need to detect recruits with venereal diseases led Dorfman [17] to investigate the idea of pooling blood samples of several soldiers, say 10, and to check whether the analysis would return a positive result. In that case, separate 10 analyses would be done to detect the infected ones, otherwise this single analysis would 'clean" the 10 members of the group.

Suppose that the prevalence rate of the disease is p, and that we analyse in the first step the amalgamated blood of n individuals. The expected number of analyses needed to screen each group of n individuals with this pooling technique is then $n* = (1-p)^n + (n+1)\,[1-(1-p)^n]$, and in general $n* \ll n$, considerably lowering costs.

The optimal group size can the be easily be computed, and it naturally increases with the inverse of the prevalence rate. It is worth mentioning that this technique is

worth considering for $p < 0.30663$, and that there is an interesting discontinuity, in the sense that there is an abrupt chafe of the optimal size from $n = 1$ to $n = 3$.

The idea of pooling also occurred to Turing and the team working at Bletchley Park with the Banburismus technique: to test hypotheses about whether different messages coded by German Enigma machines should be connected and analysed together, but this and Turing's work on sequential analysis, that it seems he devised at the same time and independently of Wald, remained secrete until 1975 ([48]).

Combined sampling has important applications in quality control, cf. [10] for an extensive bibliography on the subject, that unfortunately doesn't go beyond 1992. Santos and co-workers have been investigating composite sampling when qualitative analyses have imperfect sensibility and/or specificity, and to quantitative analyses, cf. [50, 38] and references therein.

Adaptive sampling is a nice evolution of sequential sampling, in the sense that it shares the same type of common sense: random samples on average perform nicely, but there is in the essence of sampling the possibility of getting "outlier" samples performing very bad. Sequential analysis is optimistic, in the sense that it is expected that a stopping rule will act so that our purposes can be achieved with smaller size samples than recommended using for instance *srswr*.

An example of adaptive sampling helps to understand why we claim that is is more realistic: suppose that we want to sample zones in the Newfoundland sea to evaluate cod stocks, and that in a first step n randomly chosen square regions with a 3 marine miles diagonal are chosen in the ocean chart.

When in one of those spots the research ship sails 0.25 marine miles without cod catching, this spot is abandoned. On the other hand, if the catch of cod occurs in some zone, the 8 neighbouring spots in the chart are added to the sampling plan. Hence, there is a tacit recognition that some spots are useless, while others are useful and hint that the neighbouring ones are also useful.

It is obvious that this adaptive scheme is much more rewarding than a fixed scheme, that at the end of the day could eventually provide none information whatsoever for our purposes. The many facets of adaptive sampling are detailed in [57, 54].

3.5 Distance sampling

As we said before, in a large majority of cases sampling is done with the purpose of cleverly estimate some parameter, namely a function of the mean value. This is done taking for granted that the population size is a known constant N.

However, for instance in wildlife studies, N is unknown, and the purpose of sampling is to estimate the population size. Crude but clever methods began to develop at the end of the XIXth Century, for instance Petersen's capture-recapture based estimator, [44], but in fact similar methods had already been used *circa* 1650 by Bacon to estimate game abundance, and by Laplace in 1780 to estimate the population of France departments.

The crucial clue given by Dr. H. Lectner to Clarice Starling

> We begin by coveting what we see every day. Don't you feel eyes moving over your body, Clarice? And don't your eyes seek out the things you want?

in the successful *The Silence of the Lambs*, could be a hint to explain to newcomers to the field the interest of modern distance sampling methods, where it is assumed that using an appropriate adjustment function what is observed from transects can provide at lower cost good estimates of the population size. For further details, cf. [39, 36] and references therein

4 Further Comments

The first steps of statistical inference were severely constrained by the incapacity of dealing with complex models. Fortunately most of the usual statistics are functions of sums, and the central limit theorem asserted that a "normal" approximation could be used for sums of random variables, under very broad circumstances. This was very fortunate, since the cumulants of of the normal are 0 but for the first and the second (this is the ultimate reason why the central limit theorem holds for sums of independent identically distributed random variables with finite variance), and hence it is a pure location/scale parametrized family, linearly amenable to the very well tabulated standard normal distribution. And assuming normality, excellent exact results did follow to deal with means (Student's t), variances (chi-square) and quotient of variances (F), appropriate to deal with samples of all sizes, *including small size or moderately sized samples*.

Also in the first half of the XXth Century, important results of nonparametric statistics served to deal with important location/scale problems, distribution fitting, asymmetry, randomness, association, and many other important questions, also with *small size or moderately sized samples*.

In the mid century, the availability of computing devices capable of easily dealing with large datasets and tricky distributions brought in great developments in statistics. Namely, the team led by Tukey [58], [41] advocated so convincingly EDA (Exploratory Data Analysis) that a later return to Confirmatory Data Analysis has been necessary. Other side effects of the development of computers have been the ease of use of Monte Carlo, Jacknife, Bootstrap, and many other techniques, an interesting turn point: for instance, question of robustness were suddenly as relevant as sufficiency, or even more relevant.

Resampling techniques, and real time data collecting data, changed abruptly the state of the art: instead of dealing with small samples, very large samples needed new *data mining* techniques, in a sense devices to separate gold from ore in haphazard collected data. This has been particularly important in fields such as geophysics, astronomy, economics, where the daily automated collection of thousands of data is easily feasible.

But even in other areas, such as life sciences, in which many experiences deal with small samples, the situation changed abruptly due to two developments: on one hand, formal computational augmentation of samples and simulation; on the other hand, the ease of communications strengthened collaboration bonds between groups investigating similar questions, and retrieval of experimental data became a recommendation, if not a standard (namely using the Cochrane Collaboration), so that systematic reviews and meta analysis, an expression first appearing in [24] and nowadays ubiquitous in research, superseded the restrictive situation of analysing small data sets.

Another caveat, for those dealing with retrospective studies: observe that in many situations the data that are being analysed are an haphazard sample, and inference in that case is hazardous.

And a final caveat: gathering data is expensive, tiring, eventually boring, and this often causes a very serious blunder: the same data are reused and reused to investigate different questions. Using already collected data, eventually to do some confirmatory data analysis on hypothesis suggested by those data at the exploratory stage is indeed a source of questionable results.

Finally, a brief reference to some sources of more detailed information for those who want to lean more on sampling. The Sage *Sampling Toolkit* [11, 12, 19, 20, 21, 22, 23, 43] provides useful practical information, and can be complemented by [37]; [53] is an excellent primer, as well as [5, 56], and [52] a thorough treatment of more advanced topics; [29], and the fact that Hansen and Hurwitz are co-authors is in itself a recommendation. The pioneer papers by Hansen, and Hurwitz [27] and by Horvitz and Thompson[30] are well worth reading. Sampling rare populations [33] requires specific strategies, be it in the are of small domains estimation, or in important issues such as estimating the *VaR* (Value at Risk), high quantiles of great importance in the applications of extreme value theory to risk.

Appendix

Herein we present some examples supporting our claim that immoderately increasing the sample size can bring in problems, and that computationally augmenting samples can also be counter-effective.

Large samples and irrelevant significance

Suppose that we collect a moderate sample size and classify the observations using a cross dichotomous criterion, leading to the 2×2 contingency table with free margins,

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

appropriate to test independence of factors. Suppose further that from the observed value of the test statistic, $X^2_{2,2} = \frac{n(ad-bc)^2}{(a+b)(a+c),(b+d)(c+d)}$ we decide that the corresponding p-value is large enough to maintain the null hypothesis of independence.

It is usual to say that to maintain the null hypothesis is a weak decision, to reject the null hypothesis is a strong decision. This indeed builds up from the experience that small samples do not provide enough information to challenge the null hypothesis.

Suppose now that a larger sample of size Kn leads to

Ka	Kb	$K(a+b)$
Kc	Kd	$K(c+d)$
$K(a+c)$	$K(b+d)$	Kn

This time $X^{2*}_{2,2} = KX^2_{2,2}$ will be larger, and for large enough K it will lead to rejection of the null hypothesis at as low levels as we wish, although the odds ratio $\frac{Ka}{Kb} = \frac{a}{b}$.

To a certain extent this makes sense: the odds ratio is the same, but there exists more accumulated evidence against the null hypothesis.

But the point is: any null hypothesis will ultimately be rejected, if we increase the sample size beyond reasonable values.

Many submitted papers deserve to be rejected with the argument "the sample size is insufficient to draw conclusions". But when ill informed users of statistics react collecting large data samples in future work, they can be abusing statistics another way.

Observe that the philosophy of data mining and of cumulative meta-analysis should spend some time assessing the possible effects of excess of information.

Combining p-values and loss of power with computational sample augmentation

Combining p-values is an elementary and popular technique in meta-analysis, developed much earlier than meta analysis has been named. In fact, under validity of the null hypothesis, a sample of p-values $p = (p_1, ..., p_n)$ obtained on independent tests is a sample from the standard uniform. Thus, to investigate the compsite hypothesis $H^*_0 :\ H_{0,k}$ is true, $\forall k = 1, ..., n$ *vs.* an alternative $H^*_A :\ \exists k \in \{1, ...n\} :\ H_{A,k}$ is true, we can use test statistics [45] such as Tippett's

$$\min\{P_1, ..., P_n\} \underset{|_{H^*_0}}{\frown} Beta(n, 1), \tag{7}$$

or Fisher's

$$-2\sum_{k=1}^{n}\ln(P_k) \underset{|_{H_0^*}}{\frown} \chi^2_{2n}. \tag{8}$$

On the other hand, it is well known that the absence of significance isn't properly an asset to have submitted papers accepted, and apart from publication bias [46] we can fear that it is a general practice in research to make a second experiment when the result of the first one is not satisfactory, and to select the "best" of the two, be it the minimum or the maximum, according to the setup. The minimum of of two independent uniforms is a $Beta(2,1)$, and the maximum is a $Beta(1,2)$. Hence, in composite testing, we may expect to have a set o p-values that, instead of standard uniform, are from a population X_m with density function that is a convex mixture $f_{X_m}(x) = [1+m(x-0.5)]\ \mathbb{I}_{0,1}(x)$ of uniform, with coefficient $1-\frac{|m|}{2}$, and, with mixing coefficient $\frac{|m(}{2}$, of $Beta(2,1)$ if $m \in [-2,0)$ or of $Beta(1,2)$ if $m \in (0,2]$. This is, for instance, the "resolution" of Mendel-Fisher controversy given in [47].

Thus it seems interesting to use nice results such as

Theorem 1. *If X with support $(0,1)$ and $U \frown Uniform(0,1)$ are independent, than $\min\left\{\frac{U}{X}, \frac{1-U}{1-X}\right\} = Y \frown Uniform(0,1)$, with X and Y independent*

to augment computationally the set of p-values using uniform pseudo-random numbers $u_k,\ k)1,...,n$ and computing

$$p_{n+k} = \min\left\{\frac{u_k}{p_k}, \frac{1-u_k}{1-p_k}\right\} \tag{9}$$

expecting to obtain increased power when testing uniformity using the size computationally augmented sample $(p_1,...,p_n,p_{n+1},...,p_{2n})$.

In fact this is not so, cf. [25], [15], the reverse situation holds, the explanation being that if X_{m_1}, X_{m_2} have convex mixture distributions as described above,

$$\min\left\{\frac{X_{m_1}}{X_{m_2}}, \frac{1-X_{m_1}}{1-X_{m_2}}\right\} = X_{\frac{m_1 m_2}{6}}. \tag{10}$$

Hence, the result is always nearer to a uniform, and in particular if one of the random variables is uniform (i.e., $m=0$), the result is uniform. In fact, the maximal entropy of the uniform among random variables with $(0,1)$ support confers a strong attraction character to this model, cf. Bril2.

So, computationally augmenting the sample is not always beneficial!

Acknowledgements This research has been supported by National Funds through FCT — Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2011, and PTDC/FEDER.

References

1. Aleixo, S., Brilhante, M.F., Diamantino, F, Mendona, S., Pestana, D.: (2007). Non-response and sample size. Bulletin of the International Statistical Institute **LXII**, 4804–4807 (2007)
2. Arnold, D.N.: Integrity Under Attack: The State of Scholarly Publishing. http://www.ima.umn.edu/ arnold//siam-columns/ integrity-under-attack.pdf
3. Anderson, M., Fienberg, S.E.: History, Myth-Making and Statistics: A Short Story about the Reapportionment of Congress and the 1990 Census. PS: Political Science and Politics **33** 783–794 (2000)
4. Assael,M., and Keon, J.: Non-sampling vs sampling errors in survey research. J. Marketing 46 114–123 (1982)
5. Barnett, V.: Sample Surveys: Principles and Methods. Arnold, London (2002)
6. Belin, T.R., Rolph, J.E.: Can We Reach Consensus on Census Adjustment? Statistical Science **9** 4586–508 (1994)
7. Bethlehem, J.: Nonresponse in surveys, in Lovric, M. (ed.) *International Encyclopedia of Statistical Science*, Springer, New York, 982–983 (2011)
8. Billard, L., The Census Count: Who Counts? How Do We Count? When Do We Count? PS: Political Science and Politics **33** 767–774 (2000)
9. Biemer, P.P.: Overview of design issues: total survey error. In Marsden, P., and Wright, J. (eds.) *Handbook of Survey Research*. Emerald, United Kingdom, pp. 27–58 (2010)
10. Boswell, M.T., Gore, S.D., Lovison, G., Patil, G.P.: Annotated bibliography of composite sampling Part: 1936–92. Envirinmental and Ecological Statistics **3** 1–50 (1996)
11. Bourque, L.B., and Fielder, E.P. (2003). How to Conduct Self-Administered and Mail Surveys. Sage Publ., Thousand Oaks (2003)
12. Bourque, L.B., and Fielder, E.P.: How to Conduct Telephone Surveys. Sage Publ., Thousand Oaks (2003)
13. Breiman, L.: The 1991 Census Adjustment: Undercount or Bad Data? Statistical Science **9** 458–475 (1994)
14. Brilhante, M.F., Mendona, S., Pestana, D., Sequeira, F.: Using Products and Powers of Products to Test Uniformity, In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, IEEE CFP10498-PRT, pp. 509-514.
15. Brilhante, M.F., Pestana, D., Sequeira, F.: Combining p-values and random p-values, In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, IEEE CFP10498-PRT, pp. 515–520. (2010)
16. Brunell, T.L.: Statistical Sampling to Estimate the U.S. population: the methodological and political debate over census 2000. PS: Political Science and Politics **33** 775–782 (2000)
17. Dorfman, R.: The detection of defective members in large populations. Ann. Math. Statistics **14** 436–440 (1943)
18. Erdös, P., and Rényi, A.: On a central limit theorem for samples from a finite population. Publ. Math. Instit. Hungar. Acad. Sci. **4**, 49–61(1959)
19. Fink, A.: The Survey Handbook. Sage Publ., Thousand Oaks (2003)
20. Fink, A.: How to Ask Survey Questions,. Fink, A.: How to Design Survey Studies. Sage Publ., Thousand Oaks (2003)
21. Fink, A.: How to Sample in Surveys. Sage Publ., Thousand Oaks (2003)
22. Fink, A.: How to Manage, Analyze, and Interpret Survey Studies. Sage Publ., Thousand Oaks (2003)
23. Fink, A.: How to Report on Surveys. Sage Publ., Thousand Oaks (2003)
24. Glass, G.V.: Primary, secondary, and meta-analysis of resealch. Educational Researcher. **5** 3–8 (1976)
25. Gomes, M.I., Pestana, D.D., Sequeira, F., Mendonca, S., Velosa, S.: Uniformity of offsprings from uniform and non-uniform parents. In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), Proceedings of the ITI 2009, 31st International Conference on Information Technology Interfaces, pp. 243–248 (2009)

26. Greenfield, T.: Research Methods for Postgraduates. Arnold, London (2002)
27. Hansen, M.M., and Hurwitz, W.N.: On the theory of sampling from finite populations. Annals of Mathematical Statistics **14**, 333–362. (1943)
28. Hansen, M.M., and Hurwitz, W.N.: The problem of nonresponse in sample survys. J. Am. Statist. Assoc. **41**, 517–529 (1946)
29. Hansen, M.M., Hurwitz, W.N., and Madow, W.G.: Sample Survey Methods and Theory. Wiley, New York (1962)
30. Horvitz, D.G. and Thompson, D.J.: A Generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. **47**, 663–685 (1952)
31. Improbable Research. http://www.improbable.com/ig/
32. Ioannidis, J.P.A.: Why Most Published Research Findings Are False. http://www.plosmedicine.org/article/info:doi/10.1371/ journal.pmed.0020124
33. Kalton, G., and Anderson, D.: Sampling rare populations. J. Royal Statist. Soc. **A 149**, 65–82 (1986)
34. Lynn, P.: Principles of sampling. In: Greenfield (ed.), T: Research Methods for Postgraduates, pp. 185-194, Arnold, London (2002)
35. Manfreda, K. L., Berzelak, N., Vehovar, V.: Nonresponse in web surveys, in Lovric, M. (ed.) International Encyclopedia of Statistical Science, Springer, New York, pp. 984987 (2011)
36. Marques, T.A., Buckland, S.T., Bispo, R., Howland, B.: Accounting for animal density gradients using independent information in distance sampling surveys. Stat. Methods Appl. DOI: 10.1007/s10260-012-0223-2 (2013)
37. Marsden, P., and Wright, J. (eds.) Handbook of Survey Research. Emerald, United Kingdom (2010)
38. Martins, J.P., Santos, R., Felgueiras, M.: A Maximum Likelihood Estimator for the Prevalence Rate Using Pooled Sample Tests Notas e Comunicações do CEAUL 27 (2013)
39. Morrison, M.l., Block, W.M., Strickland, M.D., Collier, B.A.: Wildlife Study Design. Springer, New York (2008) Sample survey strategies 137–198 Sampling strategies: applications 199–228
40. Mosteller, F. (1978). Errors I: nonsampling errors. In: Kruskal, W.H., and Tanur, J.M. (eds.) *International Encyclopedia of Statistics*. Free Press, New York, 208–229.
41. Mosteller, F., Tukey, J.W.: Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley (1977)
42. Office of National Statistics: Census Coverage Survey Sample Balance Adjustment. 2011 Census: Methods and Quality Report, www.ons.gov.uk/.../census/.../census...census.../ccs-sample-balance-adjus...? (2011)
43. Oishi, S.M.: How to Conduct In-Person Interviews for Surveys. Sage Publ., Thousand Oaks (2003)
44. Petersen, C.G.J.: The yearly immigration of young plaice into the Limfjord from the German Sea. Dan. Biol. St. **6**, 5–84 (1895)
45. Pestana, D.: Combining p-values, in M. Lovric (Ed.), International Encyclopaedia of Statistical Science, pp. 1145–1147, Springer Verlag, New York (2011)
46. Pestana, D., Rocha, M.L., Vasconcelos, R., Velosa, S.: Publication Bias and Meta-Analytic Syntheses. In Lita da Silva, J.; Caeiro, F.; Natrio, I.; Braumann, C.A. (Eds.): Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and other Statistical Applications, pp. 347–354, Springer, Berlin (2013)
47. Pires, A.M., Branco, J.A.: A statistical model to explain the Mendel-Fisher controversy. Statistical Science **25** 545–565 (2010)
48. Randell, B.: The Colossus, in N. Metropolis, J. Howlett, G.C. Rota, Eds.), A History of Computing in the Twentieth Century. Academic Press, New York, pp.47-92 (1980)
49. Ronzio, C.R.: Ambiguity and discord in U.S. Census data on the undercount, race/ethnicity and SES: Responding to the challenge of complexity. International Journal of Critical Statistics **1** 11–18 (2007)
50. Santos, R., Martins, J.P., Felgueiras, M.: 24/2013 Discrete Compound Tests and Dorfmans Methodology in the Presence of Misclassification. Notas e Comunicações do CEAUL 26 (2013)

51. Särndal,C.-E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (2003)
52. Singh, S.: Advanced Sampling Theory with Applications, How Michael 'Selected' Amy. Kluwer, Dordrecht (2003)
53. Scheaffer, R.L., Mendenhall III, W., Ott, R.L., GEROW, K.: Elementary Survey Sampling, Duxbury, Belmont (2012)
54. Seber, G.A.F., Salehi, M.M.: Adaptive Sampling Designs: Inference for Sparse and Clustered Populations, Springer, New Yor (2012)
55. Tanur, J. M.: Nonsampling errors in surveys, in Lovric, M. (ed.) *International Encyclopedia of Statistical Science*, Springer, New York, pp. 988–991 (2011)
56. Thompson, S.K.: Sampling. Wiley, New York (2012)
57. Thompson, S.K., Seber, G.A.F.: Adaptive Sampling. Wiley, New York (1996)
58. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley (1977)
59. Wald, A.: Sequential tests of statistical hypotheses. Ann. Math. Statist. **16** 117–186 (1945)