P-values and laws of large numbers: Practical interpretation of significance in very large samples

Michael Brimacombe[1,2]

AFFILIATION
1 Connecticut Children's Research Institute, United States
2 University of Connecticut School of Medicine, United States

CORRESPONDENCE TO
Michael Brimacombe. Connecticut Children's Research Institute, 282 Washington St, Hartford, CT 06106, United States.

E-mail: Mbrimacombe@connecticutchildrens.org ORCID iD https://orcid.org/0000-0002-3276-9071

KEYWORDS
estimation, inference, laws of large numbers, large samples, p-values, sampling distributions, significance

Received: 2 April 2025, **Revised:** 27 April 2025, **Accepted:** 28 April 2025

Popul. Med. 2025;7(April):11

https://doi.org/10.18332/popmed/206958

ABSTRACT

Research in healthcare settings is often conducted using very large amounts of data. This typically leads to a limitation on the use of standard measures of statistical significance, such as p-values, as the related sampling distributions collapse in very large samples. However, while these probability-based calculations are of limited use, application of the laws of large numbers can be very useful in guiding the interpretation of very-large-sample-based results. In specific settings, this may lead to the continued use of the p-value as a tool for inference. In other very-large-sample settings, comparison of observed estimated population characteristics can be reported on a percent difference basis and can be related directly to practical outcomes such as increases in healthcare access or decreases in healthcare cost. These interpretations should be defined pre-experimentally within the context of each study protocol.

INTRODUCTION

Medical research typically involves the careful planning of study designs, the collection of study data and the interpretation of results. Usually, a statistical, probability-based model provides the context, with model parameters interpreted as population characteristics, which can be tested in relation to hypothesized values using p-values or estimated using confidence intervals. Where sample sizes are relatively moderate, the study design will include a sample size justification, along with an assumed clinical effect size. The sample size carefully selected so that statistical significance is closely related to clinical significance[1]. The use of a randomized design to limit bias, the consistent application of the study design protocol, and the drawing of a truly representative sample are often essential components of quality research studies.

In research studies based on very-large-sample sizes, sample sizes that often run into the hundreds of thousands or millions, standard measures of inference become less relevant. Automatic use of p-values, for example, is no longer obvious when clinical significance is no longer closely related to statistical significance. Such challenges have led to consideration of alternative interpretations of the p-value or different approaches such as data-centric machine learning approaches[2,3].

Standard statistical perspectives, however, may remain useful in very-large-samples when viewed from an alternate but related perspective; incorporating large sample convergence as a key element in inference; inference via estimation. This perspective applies to a wide variety of studies that generate very large datasets, for example, genomic studies, healthcare utilization studies based on linked networks of hospital-based databases, internet marketing-based health research, national health surveys, and census data.

Interestingly, in the setting of very-large-sample study designs, while probability is an important component, the sampling distributions of aggregate statistical summaries such as sample means and proportions, and their related statistical likelihood functions, collapse, reflecting the convergence and high accuracy of estimation implied by the laws of large numbers[4] as standard errors go to zero. Probability in relation to sampling distributions in very large samples leads to the comparison of essentially non-random estimated values, without easy reference to standardized units of variation such as standard errors and automatic measures of inference such as p-values.

This setting further leads to the consideration of secondary criteria to assess practical significance in very-large-sample settings. These may include, for example, estimated decreases in healthcare costs, percent increases in access to healthcare, or estimated percent increases in survival rates: practical perspectives for the interpretation of very-large-sample results.

National health surveys and large databases

An example of a setting where such considerations apply is in the examination of very large national databases, a characteristic of health-related research in recent years. These include, among others, US census data[5] and EPIC Cosmos[6], a hospital utilization database with data on 195 million patients. As the size of samples increases to cover a larger proportion of the population of interest, the use of probability-based methods has been called into question[7].

Larger sample sizes restrict the usefulness of the p-value and related hypothesis testing frameworks. An example of an attempt to get around this is an equality or non-inferiority perspective[8]. Assuming that a moderately large sample implies approximate normality due to the central limit theorem, a confidence interval for the standardized difference between, for example, mean values can be obtained. This is then assessed versus an equality range about zero, essentially an effect size for the confidence interval, typically (-2.5, 2.5) or (-3.0, 3.0). If the observed confidence interval lies entirely within the range, the surveys are declared to be equivalent for the relevant measures being examined. Note that in the case of survey data, aggregate statistics such as the sample mean or variance should be weighted appropriately to be representative[9].

METHODOLOGICAL APPROACH

Laws of large numbers

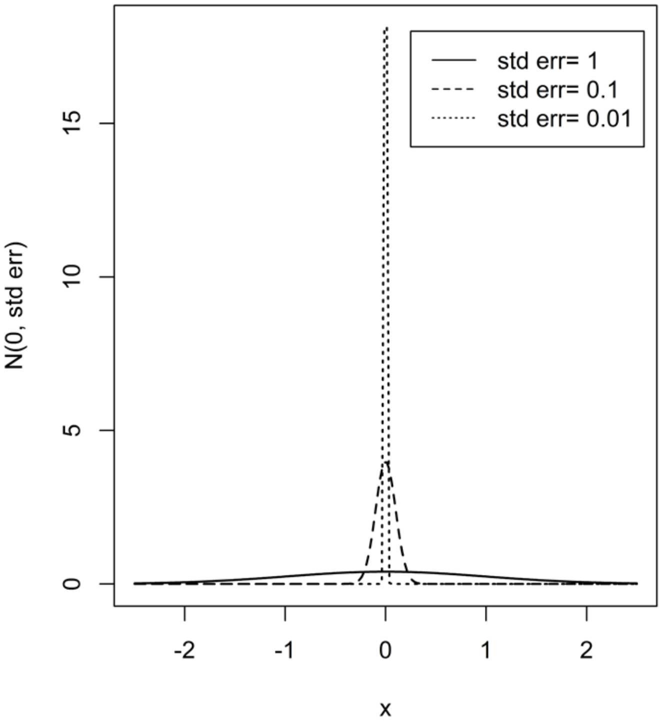
The presence of very large samples and convergent statistics creates challenges for the testing of hypotheses. Under the weak law of large numbers, a sample statistic in a very large sample (often having sample sizes of >100000 responses) will be an essentially exact estimate of, for example, the population mean, or difference of population means, or population proportion of interest. Indeed, in most settings the accuracy of the estimates will be essentially exact after 10000–20000 data values in the case of estimating a proportion, typically much fewer in regard to estimating mean values.

Application of laws of large numbers are most easily displayed using the well-known Chebyshev inequality[10] for the sample mean T(x) with expected population mean value μ, a chosen constant c>0, and variance σ^2:

$$P[|T(x)-\mu|\leq c]\leq 1-\sigma^2/nc^2$$

This implies that the sample mean T(x) converges to the population mean μ and the probability P goes to 1 as $n\to\infty$. In very large samples this implies that the sample mean is practically equivalent to the population mean value.

Figure 1. Setting μ = 0, the distribution of $T(x) - \mu_0$ is shown as $\sigma_{T(x)}\to 0$



While this is a probability-based statement, defined before observing the data, computer-based simulations with data show that these arguments and results hold in very large samples in data-based settings. The most important assumptions supporting such results are the assumptions of independence in the selection of subjects and related data values, sample and population homogeneity, and stable variation. Note that the related strong law of large numbers ensures that the chance of observing a large deviation of T(x) from μ in very large samples is minimal.

Pivotal quantities and p-values

In the context of a very-large-sample national survey, the same context that leads to the weak and strong laws of large numbers giving high accuracy of estimation, also leads to the p-value becoming clinically irrelevant.

Consider the development of a p-value to assess the departure of the observed value of the statistic T(x) from an expected population parameter value θ. A typical first step is to standardize the statistic or pivotal quantity[11] to the general form:

$$Q(x\,;\theta)=[T(x)-\theta]/\sigma_{T(x)}$$

where θ is a population characteristic of interest and $\sigma_{T(x)}$ a measure of variation for T(x) that is a function of the sample size n, allowing for calculation of tail areas based on the N(0,1) distribution. But for large n, $\sigma_{T(x)}\to 0$, implying that when testing H_0: $\theta = \theta_0$, any small deviation $T(x) - \theta_0$ will give a large value for $Q(x\,;\theta_0)$.

As the p-value is typically based on the area to the right of the observed value of $Q(x\,;\theta)$, this implies the p-value will be small and thus significant for any small deviation $T(x) - \theta_0$. Thus, statistical significance is no longer reflective of clinical or scientific significance.

In very large samples, the high accuracy of estimation limits the use of significance testing for hypotheses of the form H_0: $\theta = \theta_0$. The degree of support for the hypothesis provided by the observed data can more reasonably be approached from a different perspective, namely incorporating the high accuracy of estimation provided by T(x) for the true value of θ. This is considered below for several different but related situations in the context of large sample health survey-based research.

Small sample local surveys and very-large-sample national surveys

It is sometimes the case that applied researchers will have a local survey of limited sample size and will be able to reference, for comparison, a similar, weighted national survey of much larger sample size. The goal is often to assess whether the local sample differs from the national sample on various characteristic measures.

When comparing variables drawn from a small-sample local survey versus a comparable very-large-sample national survey, it remains possible to apply p-value based hypothesis testing by applying the laws of large numbers to the national survey. In this setting a one-sample testing approach within the context of the local survey remains useful where the null hypothesis is the essentially known population mean for the very-large-sample population. The difference of the local survey mean from this reference value can be tested, using a one-sample p-value calculated within the context of the local survey with small sample size.

Each variable of interest can be assessed in this manner. Note that the standard deviation of each variable in the very-large-sample national survey, essentially known due to the laws of large numbers, need not be a formal element of the pivotal quantity used for the p-value. While it remains a measure of overall accuracy related to the very-large-sample survey, the more relevant standard error of each variable's mean in the very large sample collapses to zero.

For example, a standard pivotal quantity to detect the mean difference of a variable defined in both local and national surveys, using the very-large-sample national survey mean μ as an essentially known population mean value, is given by a one-sample t-test:

$$(\bar{y}-\mu)/(s/\sqrt{n}) \sim t_{n-1}$$

where n is the local survey sample size, $\bar{y}$ is the local survey mean, and s is the local survey standard deviation. This can be used to obtain a standard p-value. Non-parametric methods can also be used here as appropriate. Note that the use of other statistics such as proportions or odds ratios can be approached in a similar manner. Examples are discussed below.

Demographic sub-group analysis

If comparisons are to be carried out within demographic sub-groups similarly defined in both local and national surveys, the relevant population values from the national survey typically remain essentially known, but the corresponding local survey sample size often may be fairly small. As the relevant statistical theory becomes somewhat difficult to interpret in small samples, application of the empirical bootstrap method to obtain the relevant p-value is useful, again assuming the very-large-survey population mean μ can be taken as a reference null hypothesis value.

Here, the pivotal quantity is again $(\bar{y}-\mu)/(s/\sqrt{n})$, where μ is the known population mean for the sub-group of interest within the national survey and a bootstrap distribution can be generated to provide the p-value. The median based Mann-Whitney non-parametric tests or the Fisher exact test for proportions can also be used as appropriate[12].

Very large samples in both local and national surveys: Estimation and comparison

It is a fundamental aspect of probability models that aggregate statistics such as sample means and standard deviations in large samples converge to fixed values. Indeed, this convergence happens fairly quickly[4,13]. This often seems to be forgotten in the interpretation of information drawn from very-large-sample survey-based results when the focus is on the p-value.

If both the local and national surveys are very large samples, the comparative assessment of population values can be based on a comparison of essentially known values, as the sample means will have individually converged to their expected population values. In terms of the sampling distributions of the means, the standard errors have each converged to zero. As such, there is no obvious measure of sampling related dispersion in such settings.

Units of comparison

It may also be possible to use a specific unit of measurement defined as part of the study protocol by the researcher. For example, when comparing two estimated population mean values, for example, $\bar{x}_1=\mu_1$, $\bar{x}_2=\mu_2$ for a response of interest, look at:

$$D=(\mu_1 - \mu_2)/a$$

where a *a* is a value chosen by the researcher and defined in the study protocol so that a value for D is seen as practically meaningful in relation to secondary outcomes such as an increase in a quality-of-life index or a decreased cancer risk score.

It is also possible to simply interpret the value of μ_2 in terms of its percent difference from μ_1 without reference to a unit of measurement. Observed differences between proportions, odds ratios, medians and mean differences may be interpreted using simple percent differences between the estimates, and a determination of the practical effect of such differences on healthcare cost and access.

In very-large-sample settings even a small 1% or 2% increase can have large practical importance in terms of healthcare outcomes and related costs. For example, in a population of 1 million, if the decrease in yearly cost per person is $500 due to prevention-related behavior and a 1% increase in prevention-related behavior is estimated, this may result in savings of $5 million.

Note that in smaller samples, with stable sampling distributions, these considerations can be formulated in terms of risk, the expected value of a defined loss function. The approach here is more limited as the relevant sampling distributions have collapsed.

In some settings, given that individual standard errors are available as exactly estimated values, σ_1 and σ_2, it may be useful to consider the standardized difference:

$$D=\frac{\mu_1}{\sigma_1}-\frac{\mu_2}{\sigma_2}$$

and interpret the estimated difference as a percent of μ_1/σ_1.

Similar arguments apply to inferences based on sample medians and for specific coefficient estimates in linear and generalized linear models. Note that the method of moments becomes very useful as a method of estimation in very-large-sample settings as it reflects the application of laws of large numbers.

Information equivalent sample sizes

Laws of large numbers and the related convergence of sampling distributions have interesting implications. For example, once we are past a certain number of collected data values, say $n \geq 20000$, which is assumed to be attained by very large samples here, differences in the sample sizes become irrelevant to the assessment of information, as the additional amount of data in the larger of the two surveys yields essentially no further information or accuracy for the purposes of statistical inference, as the sampling distributions have collapsed. Thus, in terms of information, two very large samples having, for example, samples sizes of 10000 and 20000, respectively, can be viewed as both having the same effective sample size in terms of estimation and related accuracy.

Additional guidelines on sample sizes might be useful in very large samples when the sample is further stratified for sub-group comparisons. But this is mostly irrelevant when the sample sizes for the sub-groups usually remain large enough themselves for the weak law of large numbers to hold in relation to aggregate statistics.

As a simple guideline, in relation to proportions, a sample size of n=10000 conservatively gives an accuracy of measurement, in terms of the length of a confidence interval, at approximately a 1% level. Sample sizes above the n=10000–20000 range give levels of accuracy that are essentially exact if the sample and reference population are homogeneous. Once past this level, sample sizes, when comparing two groups, in terms of information content, are essentially information equivalent.

Testing a null hypothesis within each survey

In some settings, a null hypothesis value specific to each survey, local or national, is also of interest. In a small-sample local survey using standard test statistics, small-sample testing for mean $\mu = \mu_0$, the odds ratio, OR=1 or a regression coefficient value, $\beta=0$, can be carried out.

In a national, very-large-sample survey, for a given variable, a hypothesized parameter value μ_0 can be compared directly to the estimated, essentially known, population value μ in terms of a chosen measure of accuracy, for example, standard deviation σ or interquartile range (IQR) based distance:

$$D=(\mu-\mu_0)/a$$

and interpreted using a practically meaningful cutoff rule. A chosen value for *a* might depend on the particular medical, social science or science context. It may again also be more appropriate to simply report the percent difference between

μ and μ_0 without reference to a chosen *a* value. This would then be interpreted in terms of its practical import in relation to, for example, healthcare outcomes, cost and access.

The importance of context

The laws of large numbers can most usefully be applied in settings where the underlying sampling process and the data values generated are independent and fairly stable. The population from which the sample is drawn should be homogeneous for the most part, or a collection of homogeneous sub-groups of respondents with appropriate weighting to reflect sampling design and overall demographic[14]. This assumption can be assessed to some extent by the random re-sampling of the database, but should be a focus of the initial design of the study.

The representative aspect of the sampled survey data remains important. Proper weighting of the data, using probability-based sampling weights, is often necessary in all large survey settings[14]. National surveys often have a wide variety of regions, cities and rural areas to obtain a demographic reflective of the entire country. In some settings the means of interest may actually be weighted averages of means from smaller regions or sub-groups. These issues are relevant to any analysis, but in very-large-sample population-based surveys, even the regional means and their weighted averages will converge at a rate that reflects application of the laws of large numbers.

In general, sample surveys should be representative of the population, carried out within given time periods, having limited outliers and stable measures of variation. The quality of the study design is always the most important aspect of a study and the criteria to be used to assess survey information should be defined in the study protocol for each given study.

PRACTICAL EXAMPLES

Example 1: Small local survey and very-large-sample national survey comparison

Consider a national survey based on an overall sample size of 500000 individuals, chosen at random from the general US population. The subset of responses representing individuals aged 18–25 years who suffer from a moderately rare disease A are of interest. This sub-group has a total number of 50000.

A local health alliance of local hospitals and health professionals in a fairly rural district conducts a survey yielding responses from n=300 individuals and 49 of these are from a similar subgroup. When planning the survey, the local alliance followed the national survey fairly closely in item development and many of the questions are comparable.

Mean values

Clinical health data for each participant are available and have been merged with the survey results. Body mass index (BMI) values are available in both local and national surveys. The national survey gives 50000 values and an estimated population mean BMI of 20 kg/m^2. Applying the weak law of large numbers to the national survey estimate and seeing this as practically equal to its population counterpart allows us to apply a one sample t-test for continuous measures in the context of the local study, using the national sample mean as a null hypothesis of interest. The test statistic of interest for the BMI measurement is given by:

$$(\bar{y}\text{-}20)/(s/\sqrt{49}) \sim t_{48}$$

If the local BMI (kg/m^2) sample mean value is 22.2, standard deviation 1.6, with a sample size of 49, this gives a test statistic value of 9.63, with an associated $p<0.001$. Therefore, the local survey mean differs significantly from the national survey mean value.

Proportions

In the national survey sub-group, 21500/50000 participants self-report as having moderate to liberal political beliefs. The local survey found 39/49. It is of interest to test for a difference, testing the null hypothesis of whether the local proportion is equal to the national proportion.

Using a one-sample test statistic for an estimated proportion (p) and taking the national survey percentage as known due to convergence of the weak law of large numbers, the usual pivotal quantity having an N(0,1) sampling distribution is given by:

$$\frac{p\text{-}f}{\sqrt{(p(1\text{-}p)/n)}} \sim N(0,1)$$

where for n=49, p = 39/49 = 0.7959 and fraction of participants f = 21500/50000 = 0.43, the pivotal quantity here yields a value of 6.35 and associated $p<0.001$. The local survey percentage differs significantly from the national survey. Note the small sample Fisher's exact or binomial test can also be used here to derive the p-value.

Odds ratios

Given, for example, comparable cases and controls in both the local survey and national survey, a similar approach to testing the equality of odds ratios can be obtained. Most standard assessments using the sample odds ratio require a large sample interpretation.

For the characteristic of interest (for example, the onset of hypertension or high blood pressure) let the relevant very-large-sample national survey-based value be given by OR = 1.8. Assuming this is a known population characteristic via application of the law of large numbers, a standardized test statistic with an approximately N(0,1) distribution can typically be defined as:

$$[\log(OR)\text{-}\log(1.8)] \,/\, \sqrt{(1/n_1+1/n_2+1/n_3+1/n_4)} \sim N(0,1)$$

Assume the local survey gives a cross-tabulation as given in Table 1, then this gives a sample relative risk or odds ratio of 12.0 and thus a test statistic value 1.62 and a related one-sided p=0.053, showing a non-significant difference between local and national OR using a Type I error = 0.05.

Example 2: Very-large-sample local survey and national survey comparison

Assume that the local (sample 1) and national surveys (sample 2) are conducted with sample sizes of 500000 and 1500000, respectively. The sample statistics for both can be assumed to have converged to their population mean values. The sampling distributions have collapsed about their expected population mean values.

Mean values

The resulting sample means and standard deviations for a continuous variable, a cancer-related risk score, are given by:
Sample mean 1 = 1.47, SD 1 = 5.6
Sample mean 2 = 22.1, SD 2 = 4.7
The observed mean difference of 7.4 on its own can be interpreted as a 33.48% decrease from the 22.1 sample 2 baseline, an impressive percent change and may have important practical importance in relation to the onset of cancer-related illness and the related cost of healthcare.

Also relevant here is:

$$D=\frac{22.1}{4.7}-\frac{14.7}{6.5}=4.7-2.6=2.1$$

which reflects the view that with very-large-sample sizes and effectively equal sample sizes, there is no need to pool the standard errors to improve the estimation of variation. Here, the standardized mean difference is >2.

Proportions

A discrete variable, for example whether a respondent requires mental health treatment (yes/no) after participating in a specified treatment program, is measured in each sample and gives the following observed values: Sample proportion p_1=0.85, and sample proportion p_2=0.79. The difference:

$$D=(p_1-p_2)=0.06$$

Table 1. Cross-tabulation example of cases with and without hypertension

	Hypertension	No Hypertension	Total
Cases	n_1 = 40	n_3 = 10	50
Non-cases	n_2 = 10	n_4 = 30	40
Total	50	40	90

represents an absolute change of 6%, or a decrease of 7.1% of 0.85. This can then be interpreted in terms of the number of subjects and the practical importance of this decrease. For example, if the yearly cost per patient is $50000 and a 7.1% decrease in those requiring treatment corresponds to approximately 200 patients, then the yearly cost saving is $10 million.

Note again that at sample sizes of approximately 10000, differences of 1–2% are statistically significant using standard large sample sampling distributions. At levels above 20000, statistical significance loses its meaning as a point of reference for interpretation of observed percent differences and sample estimates are essentially exact.

Odds ratios

For the purpose of interpretation, in settings where there are two proportions to compare and both the local and national samples are very large, the odds ratio can also be computed:

$$OR=\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

Interpretation can be based on the distance of the observed OR value from 1. For the above example with p_1=0.85 and p_2=0.79, the OR=1.51. Inference can then be based on whether the value is sufficiently greater than 1.0. However, the relevant sampling distribution has collapsed.

It seems more reasonable here to interpret the original 6% difference in proportions, translating the percentage to actual counts and related cost per patient. The approach chosen to interpreting the result, even when not primarily probability based, should be set pre-experimentally as part of the study design.

Multiple comparison issues

In settings where the local survey has a small sample size and the national survey a very-large-sample size, standard multiple comparisons can be applied to a set of specific comparative hypotheses being considered, as appropriate. Typically, adjustment to the Type I error suffices using, for example, the Bonferroni correction.

In a setting where both the local and national surveys have very-large-sample sizes, assessment of possible significant or important differences is no longer based on probability calculations. The laws of large numbers have given almost exact estimates of, for example, both mean and variance.

Comparisons for a set of null hypotheses, for example across a set of clinical variables, might generally be of the form:

$$D_j=\frac{\mu_{1j}}{\sigma_{1j}}-\frac{\mu_{2j}}{\sigma_{2j}}$$

with j = 1, ... ,10 standardized differences, or a simple comparison of the percent differences between each μ_{1j} and μ_{2j}. In each case, the difference can be interpreted in terms of related practical consequences: for example, a decrease in healthcare cost, an increase in access or a decrease in risk score. These interpretations can be listed out clearly in the study protocol.

Here, there is no α = P(Type I error) that is useful to such an assessment, as the laws of large numbers imply convergence of the sample statistic to its population value and the collapse of the related sampling distribution. Thus, the rejection regions for such significance testing have lost relevance as the standard error of the test statistics has converged to zero. Direct interpretation for each of the estimated population values is required. Standard multiple comparison corrections, based on family-wide assessment of the set of hypotheses to be considered and their related rejection regions, are no longer relevant in comparisons of very large samples.

Bayesian perspective

In some settings a Bayesian perspective may be of interest. In the Bayesian context, researchers will not only have information from the current surveys in question, but also from expert opinion or beliefs regarding model parameters based on past surveys or studies regarding possible population values in the form of a pre-existing probability distribution for the population characteristic of interest θ. Here the effects of the laws of large numbers on the current study of interest remain relevant and impactful.

Standard Bayesian calculations require both prior density and likelihood function, but the likelihood function, based on the sampling distribution of the key statistic(s) of interest, will typically collapse in very large samples. Note that the prior, not affected by the information in the current sample, does not collapse, as it serves as a pre-existing baseline assessment of existing knowledge regarding possible θ values. This is a possible point of contention as some theoretical Bayesian arguments will assume the variance of the prior distribution is somehow related to the sample size in the current study[15]. This is not assumed here.

In the case of the local survey having small sample size and the national survey having very-large-sample size, a standard Bayesian analysis based on the local sample and related probability models can be carried out, comparing the known, very-large-sample national survey mean, to the local mean. The posterior odds ratio or the Bayes factor can be applied.

In the case of both surveys having very-large-sample sizes and the difference in population means being the parameter of interest, an initial prior probability distribution for the difference in population means for a given variable, could be taken as normally distributed, say:

$$D=(\mu_1-\mu_2) \sim N(D_p,\sigma_p^2)$$

with an assumed known difference of means D_p and assumed known variance. Ideally, this is based on a systematic review of past relevant studies. A central 95% prior region for is then given by $D_p \pm 1.96\sigma_p$.

This can be then updated by the observed information in the surveys, here using the laws of large numbers which provide an essentially known population value for the mean difference. The collapse of the likelihood function reflects the collapse of the sampling distribution in large samples, implying the related standard error has converged to zero and is irrelevant to updating the prior and the accuracy of the resulting estimate. The new study reports an essentially known population value D_{obs} for the difference in survey means.

If the prior information and new information are given the same weight, the updated or posterior result might be given by:

$$D=(\mu_1-\mu_2) \sim N(D_{po},\sigma_p^2)$$

where $D_{po} = (D_p + D_{obs})/2$. This simply shifts the prior distribution without altering its shape, as the shape of the likelihood has collapsed and adds no new information. For the purposes of inference, where the information from prior studies and the current study are given an equivalent weight, a central 95% highest posterior density (HPD) interval can be obtained for the difference of interest: $D_{po} \pm 1.96\sigma_p$.

If the scale of interest is taken to be the number of studies considered, with, for example, the prior distribution based on 29 previous studies, this weighting might be expressed conservatively, as: $[(29/30)\ D_p + (1/30)\ D_{obs}] \pm 1.96\sigma_p$, with the prior density again remaining the only source of variation.

Again, the practical setting should guide the interpretation given to the differences observed and assumptions should be clearly outlined in the study protocol. The practical implication of the estimated difference in terms of related cost or access to care implications can be reported.

As an example, consider studying differences in the average income between defined socio-economic groups A and B in a national survey of 250000 individuals. The sample sizes in the two groups are 50000 and 35000, respectively. The observed mean difference is 7500 and the respective standard deviation is 2000. The standard errors for the mean estimates themselves are 0.001 and 0.0003, respectively. Historically, from comparable survey data, the observed difference in mean incomes has followed an approximate normal distribution with mean 10000 and standard deviation of 1000. Taking this as a prior density for the mean and taking the current observed mean difference as the known value of the population mean difference, the resulting, updated prior information or posterior density in light of the survey data, can be expressed as:

$$N([10000 + 7500]/2)$$

if the prior information and current study are given equal importance.

Random resamples of very large samples

A suggestion is sometimes made that the use of standard p-values and confidence intervals can be made relevant in very-large-sample settings by drawing a random set of small resamples (n=50, for example) from the very large sample and using these as the basis for comparing local and national surveys, generating a set of p-values and confidence intervals. This can be done, but in very-large-sample settings, the number of such random resamples required to cover even 10% of the total data sample will yield a very large number of p-values, which will then be subject to multiple comparison issues in their interpretation, limiting the usefulness of this strategy.

In the context of the large sample approaches presented here, such second-stage resampling primarily might better be viewed as representing a simple approach to assessing the assumption of homogeneity in the dataset, useful as a diagnostic for application of laws of large numbers and assessing the underlying stability of very-large-sample results.

Machine learning and large sample convergence

Machine learning methods such as artificial neural networks (ANNs) or random forest methods do not formally use probability-based methods. They are data-centric, often using one half of the data (training) to predict the response in the remaining half (testing). This is actually more a measure of agreement between the two data components. There is little formal use of theoretical probabilistic modeling[16]. The ANN model for example uses a multi-nodal network template that guides the fitting of such models. From a statistical perspective these methods use a very large number of fitting parameters, called weights and biases, that massively overfit the model. This model is highly nonlinear as well, often using numerical least squares related convergence criteria augmented with stochastic algorithms and backward fitting approaches. The convergence patterns of such models can be complex, sometimes de-converging before converging[17]. Such models are often more stable in very large samples. See for example Golas et al.[18] where over 11000 subjects and 3500 variables are the context of the modeling. ANN and machine learning models are best applied in very-large-samples. Note also that the laws of large numbers discussed here apply to individual parameters defined within a larger, usually linear, model. ANNs most usefully provide a classification, without a focus on providing information on individual parameters.

DISCUSSION

The issues discussed here apply in many modern research settings: healthcare, genomics, ecology, internet data, marketing, economics and others. In all settings where hundreds of thousands or millions of samples are to be collected and organized as databases, laws of large numbers will affect the sampling distributions related to the aggregate statistics of interest. Here, examples were discussed in a healthcare survey context.

The growing use of large databases and related data-centric approaches has led to challenges in the continued application of probability-based approaches, for example the p-value in relation to hypothesis testing. However, probability in the form of large sample convergence allows for the simplification of calculations related to statistical inference in the form of estimation where very-large-sample national surveys are to be interpreted or compared with smaller local surveys.

When comparing very-large-sample surveys, the concept of essentially equivalent sample sizes arises. Sample sizes beyond a given value, for example n=20000, are equivalent in relation to the information and accuracy they provide due to large sample convergence and the laws of large numbers. These threshold values will differ according to the statistic of interest, with sample means often achieving convergence more quickly than proportions or rates.

Different areas of science use different statistics and related pivotal quantities, so it is difficult to generalize criteria. However, in most settings, with very large samples, the central limit theorem and related laws of large numbers apply. Whether odds ratios, proportions, coefficients in a regression model or more generally maximum likelihood estimates, they are assessed, on the appropriate scale, in relation to the normal distribution or related chi-squared distribution. Continuous variables tend to converge rather quickly, for moderate sample sizes, assuming a representative and unbiased sample with stable variation. A general rule for interpretation can be derived by looking at the standard error of the statistic of interest. If it is less than 0.01 or 0.001, convergence of sample statistics to their expected population values can be assumed. This should be defined in the study protocol.

With categorical responses, larger sample sizes are typically required for convergence. Again, examining the standard error can guide the interpretation of convergence. While a sample size >20000 in each sample considered should be sufficient, a standard error value less than 0.001 typically implies convergence.

While the above examples compare univariate aggregate measures subject to laws of large number related effects, multivariate and higher order statistics (correlations, parameters in generalized linear models, vectors of means etc.) are also affected by large numbers and the issues of interpretation discussed here. For example, in settings where the data can be organized into a summary matrix, and the elements of the matrix are essentially random, even the eigenvalues of the data matrices will have laws of large numbers[19].

Inferential cutoff values are always somewhat subjective and reflect the area of application and goal of the study in question. The use of 0.05 and 0.01 as cutoffs for interpreting p-values has become somewhat standard, but in very-large-sample cases where the tail area of the sampling distributions collapse, the cutoff criteria can be modified to: 1) reasonable percentage differences, motivated by chosen practical outcomes such as improved healthcare access or lower healthcare costs; or 2) 1 or 2σ-fold changes for standardized statistics, allowing for reasonable, practical definitions of σ.

The approach suggested here involves:

1. Adjustments to the study protocol to include guidance for the comparison of estimated, essentially known quantities.
2. Defining meaningful differences or percent changes, which can be small and remain practically meaningful.
3. Careful assessment of population homogeneity and the independence of the sampling process. These underlie the application of the laws of large numbers.
4. Careful interpretation and links to practical outcomes, for example increasing access to care, the lowering of healthcare costs or lowering of disease-related risk score averages.

These can provide meaningful guidelines for researchers and should be made part of the relevant study protocol. Practical considerations such as related health cost savings or improved access to healthcare should be used to set thresholds for the interpretation of observed percent differences. These should be set pre-experimentally, as part of the initial study design. Such practical considerations imply the need for clear understanding of the medical, socio-economic or scientific context on the part of data analysts. In a sense this approach is similar to defining decision theoretic loss functions without access to a sampling distribution and the concept of average loss.

The Bayesian statistical perspective is also affected. The laws of large numbers can be viewed as implying collapsed sampling distributions, implying the collapse of the likelihood function. The pieces of information available are then: 1) the prior density for the population characteristic of interest, and 2) the estimated population value via the law of large numbers in the new study. The observed population value for the current study can be used to update the prior to a posterior density through a simple shift of prior density mean value. Note the accuracy of this estimate reflects the variation in the assumed prior density, making thoughtful selection of the prior density a requirement.

Limitations

There are of course other challenges when collecting and analyzing a large sample. Missing data issues are particularly challenging, often causing bias to arise in the dataset. Simpson's paradox and the possibility of latent variable effects and residual confounding may arise, affecting the interpretation of results[20].

The focus here is on the interpretation of estimated individual model parameters. In broader settings, for example logistic regression, where the probability of a success (p) is also estimated (a function of the estimated values in the logistic function), the ROC is generated for various values of p and the set of standard classification diagnostics for p=0.5. Poor fit and resulting false positives can still occur, as the variables included in the model may not be the appropriate set of variables for accurate prediction of the outcome of interest. So false positives can still exist when modeling with large datasets.

CONCLUSIONS

In very large samples, almost exact estimation and the practical interpretation of such observed differences become an essential paradigm for statistical inference via the laws of large numbers, and related criteria should be a component of the study protocol. In specific settings, this leads to the continued use of the p-value as a tool for inference. In other very-large-sample settings, this leads to practical comparisons of estimated population values. In general, it is important when interpreting study results, for any sample size, not to become too dependent on p-values or other specific inferential summaries alone and to carefully consider the context and quality of the study.

REFERENCES

1. Sharma H. Statistical significance or clinical significance? A researcher's dilemma for appropriate interpretation of research results. Saudi J Anaesth. 2021;15(4):431-434. doi:10.4103/sja.sja_158_21
2. Gómez-de-Mariscal E, Guerrero V, Sneider A, et al. Use of the p-values as a size-dependent function to address practical differences when analyzing large datasets. Sci Rep. 2021;11(1):20942. doi:10.1038/s41598-021-00199-5
3. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018;15(4):233-234. doi:10.1038/nmeth.4642
4. Regazzini E. Concentration Comparisons between Probability Measures. Sankhya: The Indian Journal of Statistics, Series B (1960-2002;54(2):129-149.
5. Walker K. Analyzing US Census Data: Methods, Maps, and Models in R. CRC Press, Chapman & Hall; 2023.
6. Tarabichi Y, Frees A, Honeywell S, et al. The Cosmos Collaborative: A Vendor-Facilitated Electronic Health Record Data Aggregation Platform. ACI open. 2021;5(1):e36-e46. doi:10.1055/s-0041-1731004
7. Halsey LG. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum?. Biol Lett. 2019;15(5):20190174. doi:10.1098/rsbl.2019.0174
8. Tatem KS, Romo ML, McVeigh KH, et al. Comparing Prevalence Estimates From Population-Based Surveys to Inform Surveillance Using Electronic Health Records. Prev Chronic Dis 2017;14:160516. doi:10.5888/pcd14.160516
9. Graubard BI, Korn EL. Modelling the sampling design in the analysis of health surveys. Stat Methods Med Res. 1996;5(3):263-81. doi:10.1177/096228029600500304
10. Grimmett G, Welsh DJ. Probability: An Introduction. Oxford

University Press; 2014.
11. Brimacombe M. Likelihood Methods in Biology and Ecology: A Modern Approach to Statistics. CRC Press, Chapman & Hall; 2019.
12. Hollander M, Wolfe DA, Chicken E. Nonparametric Statistical Methods. 3rd ed. Wiley; 2013.
13. Franklin A. Shifting standards: Experiments in Particle Physics in the Twentieth Century. University of Pittsburgh Press; 2013.
14. Little RJ, Lewitzky S, Heeringa S, Lepkowski J, Kessler RC. Assessment of weighting methodology for the National Comorbidity Survey. Am J Epidemiol. 1997;146(5):439-449. doi:10.1093/oxfordjournals.aje.a009297
15. Johnstone IM. High dimensional Bernstein-von Mises: simple examples. Inst Math Stat Collect. 2010;6:87-98. doi:10.1214/10-IMSCOLL607
16. Brimacombe M. Data Flow-Based Strategies to Improve the Interpretation and Understanding of Machine Learning Models. Bioengineering (Basel). 2024;11(12):1189. doi:10.3390/bioengineering11121189
17. Colbrook MJ, Vegard A, Hansen AC. The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smales 18th problem. PNAS. 2022;119(12):e2107151119. doi:10.1073/pnas.2107151119
18. Golas SB, Shibahara T, Agboola S, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak. 2018;18(1):44. doi:10.1186/s12911-018-0620-z
19. Paul D, Aue A. Random matrix theory in statistics: A review. Journal of Statistical Planning and Inference. 2014;150:1-29. doi:10.1016/j.jspi.2013.09.005
20. Kutner MH, Nachtsheim CJ, Neter J, Li W, eds. Applied Linear Statistical Models, 5th ed. McGraw-Hill Irwin; 2005.

CONFLICTS OF INTEREST

The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none was reported.

FUNDING

There was no source of funding for this research.

ETHICAL APPROVAL AND INFORMED CONSENT

Ethical approval and informed consent were not required for this study.

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created.

PROVENANCE AND PEER REVIEW

Not commissioned; externally peer-reviewed.