The paradox of large samples

S. Kunte and A. P. Gore

For large samples, standard statistical test procedures almost always reject the hypothesis of interest, thus creating a rather paradoxical situation. In this article we show that the basic reason for such a situation is that, the procedures are based upon fixed predetermined level of significance, choice of which does not depend upon the sample size. We propose to resolve the problem by making the choice of the level of significance dependent upon the size of the sample.

WITH progress in instrumentation and electronics many laboratory measurement processes get automated, yielding large and accurate data sets. This is of course a boon. But larger the data set greater is the need for summarizing and modelling. When a plausible model is developed, there is a natural desire to see if the data bear it out. For this the traditional prescription in statistics is the goodness of fit χ^2 test. It gives a measure of discrepancy between observed frequencies and frequencies to be expected if the model is 'good'. If this measure of discrepancy is too large i.e., greater than the so-called 5% value of the χ^2 distribution, the model is prejected. Otherwise the model is accepted as plausible, *at least provisionally. Or such is the traditional* statistical wisdom.

A scientist with huge data sets is likely to get frustrated practising the above approach; for it may turn out to his dismay that the goodness of fit test invariably rejects the model proposed by him. Where is the hitch? Is statistics wrong? Is a precise quantitative model impossible for the data? Is having many observations a 'bad' thing? We propose to argue in this note that the answer lies elsewhere. The difficulty is with the basic logic behind the present practice statistical tests of hypotheses, which was quite adequate for traditional 'small' samples but fails when samples are 'large'. It is not easy to define the terms 'small' and 'large' in this context. But as a thumb rule we will say that sample sizes of order 10^2 are small while those of order 10^4 are large.

We shall illustrate the weakness of the traditional statistical prescriptions with reference to a simple problem namely testing the hypothesis that mean of a normal variable (with unit variance) is zero against the alternative that it is ε where ε is any fixed number greater than 0. The test statistic for this problem is

$$Z = \sqrt{n}\,\bar{X},$$

where $\bar{X}$ is the sample mean and n the sample size. The traditional most powerful test rejects H_0 at 5% level of significance if Z exceeds 1.64. It is well known that if the true mean is even slighly bigger than zero, say ε, then as sample size goes on increasing, the sample mean approaches ε and $\sqrt{n}\,\bar{X}$ goes to infinity guaranteeing rejection of H_0. This is indeed as it should be. Traditional theory lauds this by calling it the consistency of the test procedure.

This brings us to the question of level of significance α the probability of wrongly rejecting H_0 (type I error) and β the probability of wrongly accepting H_0 (type II error). Traditional prescription is to fix α at a pre-assigned level, which is usually taken to be 5% or 1%. Among the tests which satisfy this condition choose that test which maximizes the power $(1-\beta)$. The test described above has this property and so is the most powerful test for the problem.

In this argument since the value of β is not controlled, there is an implicit acceptance of the fact that a somewhat high level of β may have to be lived with even after selecting the most powerful test procedure.

The justification offered for controlling α by fixing it at a suitably low level is that rejecting the null hypothesis when in fact it is true is a more serious error.

The paradox of the situation with large samples now becomes clear. Elementary probability calculations can show that in the above procedure, while α remains fixed at say 0.05, β decreases to zero as sample size increases. Now with large samples, one may more often reject H_0 wrongly but will almost never accept it wrongly. So now a more serious error is committed more often. We shall therefore assert that good tests should necessarily have the property that 'wrong rejection of the null hypothesis (or the model) has a lower probability than wrong rejection of the alternative hypothesis'. We propose to put this as an extra condition on an otherwise reasonable test procedure. The condition, in effect, translates itself into the selection of the level of significance which is not more than a pre-assigned level and should go to zero as n tends to infinity. This

The authors are in the Department of Statistics, University of Poona, Pune 411 007.

suggestion will usually say that the critical value for the standardized test statistic should go to infinity as the sample size goes to infinity. This suggestion that the level of significance should go to zero as n tends to infinity can also be found[1]. However, he had not specified the rate at which this should happen.

It is therefore necessary that the critical value or the cut-off point (and hence the corresponding level of significance) must be made dependent on n. Let us denote such a level of significance by α_n and the cut-off point by c_n. Then for the problem of the mean of a normal distribution we have

$$\alpha_n = P_0(\sqrt{n}\,\bar{X} > c_n)$$
$$= P(Z < -c_n),$$

where Z is a standard normal variate. Similarly,

$$\beta_n = P_1(\sqrt{n}\,\bar{X} < c_n)$$
$$= P(Z < c_n - \sqrt{n}).$$

Our requirement is that $\alpha_n \leqq 0.05$ and $\alpha_n \leqq \beta_n$.

To get a specific value of c_n for large n, we will solve

$$\alpha_n = \beta_n.$$

Then

$$-c_n = c_n - \sqrt{n}\,\varepsilon$$

or

$$c_n = (\sqrt{n}/2)\,\varepsilon.$$

Thus if the sample size is 10,000 and $\varepsilon = 1$, the cut-off point to be used is 50 and not 1.64.

Here we have been able to compute c_n precisely because we could compute β which is a function of n, ε and we have assumed that ε is a fixed known. In statistics this corresponds to saying that we have a simple alternative hypothesis. If the alternative hypothesis is a composite hypothesis like $\theta > 0$, then the probability of type two error β will become a function of θ and our condition of $\alpha_n \leqq \beta_n$ will have to be suitably modified, which involves mathematics of greater complexity.

For the problem of goodness of fit of a model, the alternative hypothesis is much more vague. It is not quite clear how the cut-off point of the χ^2 test of goodness of fit may be adjusted to ensure a similar property, because here the alternative hypothesis is not well-specified. The following intuitive attack on a related problem seems reasonable.

Suppose we wish to test the null hypothesis that the k possible outcomes of a random experiment have probabilities p_{0i}, $i = 1, 2, \ldots, k$ (such that these k values add up to 1). The alternative is that the probabilities are p_{1i}, $i = 1, 2, \ldots, k$. The goodness of fit χ^2 test statistic is

$$T = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}}.$$

T follows, asymptotically, a central χ^2 distribution with $k-1$ degrees of freedom if H_0 is true and a noncentral χ^2 distribution with $k-1$ degrees of freedom and with noncentrality parameter λ when H_1 is true[2], where

$$\lambda = n \sum_{i=1}^{k} \frac{(p_{1i} - p_{0i})^2}{p_{0i}} = n\delta \text{ (say)}$$

The test rejects H_0 if T is larger than the tabulated χ^2 value of assigned level of significance α. As before, notice that given large samples, anything except perfect agreement with the null hypothesis means rejection of H_0. For example consider the null hypothesis that in human births, both sexes are equally likely. We will consider two sample sizes, 20 and 2000 (with the same observed sample proportions). (See Table 1).

The χ^2 values given in the table illustrate the point that the χ^2 statistic T becomes significant even when the relative proportions in observed data remain unchanged, if the sample size increases. The tabulated value of χ^2 with one degree of freedom at 5% level of significance is 3.84. Thus even though the sample proportions remain the same, the conclusion changes as we move from the sample size of 20 to 2000.

The above test procedure does not take into account the change in the probability of type two error, which, for the case of the sample of size 2000 may in fact be smaller than the fixed level 0.05 of type one error. Thus instead of fixing the cut-off point from the χ^2 table, independent of n, we should choose it such that

$$P_0(T > c_n) \leqq P_1(T < c_n).$$

To compute the term on the right hand side, in the case of the multinomial problem with k classes, we can use the noncentral χ^2 distribution with the noncentrality parameter $\lambda = n\delta$, which would be large for large n. For a noncentral χ^2 variate T with degrees of freedom ν—

Table 1. Sensitivity of χ^2 test for the hypothesis of equal proportions in sexes of offsprings (fictitious data)

	Small sample ($n = 20$)		*Large sample* ($n = 2000$)	
	Observed	Expected	Observed	Expected
Male	11	10	1100	1000
Female	9	10	900	1000
χ^2 statistic with 1 d f		0 202		20 2
Conclusion:	Accept the hypothesis of equal proportion		Reject the hypothesis of equal proportion	

and noncentrality parameter λ, it is well known[3] that

$$[T-(\nu+\lambda)]/[2(\nu+2\lambda)]^{\frac{1}{2}}$$

tends to a standard normal variate as $\lambda\to\infty$ for fixed ν (or as $\nu\to\infty$ for fixed λ).

In our case, the degrees of freedom $k-1$ are fixed while the noncentrality parameter $n\delta$ tends to infinity. Hence using the normal approximation, the cut off point c_n satisfies

$$P_0(T>c_n)=P\left(Z<\frac{c_n-(k-1+n\delta)}{[2(k-1+2n\delta)]^{\frac{1}{2}}}\right).$$

Using the normal approximation for the term on the left hand side, (i.e. assuming that k is large) we get

$$P\left(Z<-\frac{c_n-(k-1)}{[2(k-1)]^{\frac{1}{2}}}\right)=P\left(Z<\frac{c_n-(k-1+n\delta)}{[2(k-1+2n\delta]^{\frac{1}{2}}}\right).$$

Equating the upper limits on Z we get

$$c_n=\frac{(k-1+n\delta)[2(k-1)]^{\frac{1}{2}}+(k-1)[2(k-1+2n\delta)]^{\frac{1}{2}}}{[2(k-1)]^{\frac{1}{2}}+[2(k-1)+2n\delta]^{\frac{1}{2}}}.$$

Ignoring terms of order $1/\sqrt{n}$

$$c_n=k-1+\{\tfrac{1}{2}[n\delta(k-1)]\}^{\frac{1}{2}}.$$

If δ is known, this provides the suitable cut-off point to carry out the test. In general, δ is unknown. A common prescription for choosing 'class limits' in a χ^2 goodness of fit test with k classes is that under H_0, the classes should be equiprobable i.e. $p_{0i}=1/k$ for every i. Suppose our alternative value of p_{1i} is such that values of $p_{0i}-p_{1i}$, $i=1,2,\ldots,k$ are of the order $1/k^2$. Then δ is also of the order $1/k^2$.

In these circumstances the cut-off point c_n, suitable for large n and $k(>30)$ is

$$c_n=k-1+(n/2k)^{\frac{1}{2}}$$

If k is 30, the values of c_n are

n	10,000	50,000	100,000
c_n	40	58	70

When we compare these with the 5% cut-off value for 29 d.f. namely 42.6 we see the implication. Clearly there is no need to work with a value of α larger than the traditional value. Hence we recommend using the larger of the two values, one being the value from the χ^2 table and the other being given by the formula above.

The calculations of c_n so far are based on normal approximation to central and noncentral χ^2 distributions. This requires that not only is n large but k is also bigger than 30. Such is indeed the case in many problems with large data. However, it is also common to have a model with a fairly small value of k. This would require alternative treatment to the extent that the level of significance is calculated using the χ^2 distribution with $k-1$ d.f. while the probability of type II error is obtained using normal approximation. The appropriate value of the cut-off point c_n for which $\alpha_n=\beta_n$ can be obtained with the help of a computer program.

In the example of the human sex ratio, the null hypothesis says that the probability of a male birth is 0.5. Suppose the alternative is that the above probability is 0.6. Then δ is $2\frac{(0.6-0.5)^2}{0.5}=0.04$. Hence the cut-off point c_n satisfies the equation

$$P(\chi_1^2>c_n)=P\left[Z<\frac{c_n-(1+0.04n)}{[2(1+0.08n)]^{\frac{1}{2}}}\right].$$

Using a computer it can be seen that this is approximately satisfied by $c_n=22.1$ for $n=2000$, which gives $\alpha_n=0.259\times10^{-5}$ and $\beta_n=0.268\times10^{-5}$. Such a cut-off point, one notes, will lead to acceptance of H_0 in the present problem.

In conclusion we wish to point out the following. Firstly, the paradox of large samples is caused because of the insistence of conventional *fixed* levels of significance which leads to committing the more serious error more often. It is not a weakness of a test. It happens whenever a 'consistent' test is applied to massive data. The way out is use of unconventionally small significance levels such as two or three per million as seen in our illustrative example. Secondly, no general answer is available in statistical literature as to what the precise cut-off point in a test for large samples should be. We have obtained the answers in two particular cases where the alternative hypothesis is simple. Perhaps the most important message to the experimental scientist is that text book prescriptions of 1 or 5% level of significance for statistical tests may not be suitable for data involving large samples. A more careful analysis would be needed before a model is to be rejected.

Acknowledgements. We thank Prof. V. Sitaramam for useful suggestions. Research of A. P. G. was supported by a grant from the Department of Science and Technology, New Delhi.

1. Zellner, Arnold, *An Introduction to Bayesian Inference in Econometrics*, John Wiley, New York, 1971, p. 304
2. Kotz, S. and Johnson, N. L., (eds) *Encyclopaedia of Statistical Sciences*, Wiley Interscience, New York, 1985, p. 279
3. Johnson, N. L. and Kotz, S., *Continuous Univariate Distributions – 2*, Houghton Mifflin, Boston, 1970, p. 135.