1

Introduction

Why asymptotic statistics? The use of asymptotic approximations is two-fold. First, they enable us to find approximate tests and confidence regions. Second, approximations can be used theoretically to study the quality (efficiency) of statistical procedures.

1.1 Approximate Statistical Procedures

To carry out a statistical test, we need to know the critical value for the test statistic. In most cases this means that we must know the distribution of the test statistic under the null hypothesis. Sometimes this is known exactly, but more often only approximations are available. This may be because the distribution of the statistic is analytically intractable, or perhaps the postulated statistical model is considered only an approximation of the true underlying distributions. In both cases the use of an approximate critical value may be fully satisfactory for practical purposes.

Consider for instance the classical t-test for location. Given a sample of independent observations $X_1, \ldots, X_n$, we wish to test a null hypothesis concerning the mean $\mu = \mathrm{E}X$. The t-test is based on the quotient of the sample mean $\overline{X}_n$ and the sample standard deviation S_n. If the observations arise from a normal distribution with mean μ_0, then the distribution of $\sqrt{n}(\overline{X}_n - \mu_0)/S_n$ is known exactly: It is a t-distribution with $n-1$ degrees of freedom. However, we may have doubts regarding the normality, or we might even believe in a completely different model. If the number of observations is not too small, this does not matter too much. Then we may act as if $\sqrt{n}(\overline{X}_n - \mu_0)/S_n$ possesses a standard normal distribution. The theoretical justification is the limiting result, as $n \to \infty$,

$$\sup_x \left| \mathrm{P}_\mu \left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \le x \right) - \Phi(x) \right| \to 0,$$

provided the variables X_i have a finite second moment. This variation on the central limit theorem is proved in the next chapter. A "large sample" level α test is to reject $H_0 : \mu = \mu_0$ if $\left|\sqrt{n}(\overline{X}_n - \mu_0)/S_n\right|$ exceeds the upper $\alpha/2$ quantile of the standard normal distribution. Table 1.1 gives the significance level of this test if the observations are either normally or exponentially distributed, and $\alpha = 0.05$. For $n \ge 20$ the approximation is quite reasonable in the normal case. If the underlying distribution is exponential, then the approximation is less satisfactory, because of the skewness of the exponential distribution.

Table 1.1. *Level of the test with critical region* $|\sqrt{n}(\overline{X}_n - \mu_0)/S_n| > 1.96$ *if the observations are sampled from the normal or exponential distribution.*

n	Normal	Exponential[a]
5	0.122	0.19
10	0.082	0.14
15	0.070	0.11
20	0.065	0.10
25	0.062	0.09
50	0.056	0.07
100	0.053	0.06

[a] The third column gives approximations based on 10,000 simulations.

In many ways the t-test is an uninteresting example. There are many other reasonable test statistics for the same problem. Often their null distributions are difficult to calculate. An asymptotic result similar to the one for the t-statistic would make them practically applicable at least for large sample sizes. Thus, one aim of asymptotic statistics is to derive the asymptotic distribution of many types of statistics.

There are similar benefits when obtaining confidence intervals. For instance, the given approximation result asserts that $\sqrt{n}(\overline{X}_n - \mu)/S_n$ is approximately standard normally distributed if μ is the true mean, whatever its value. This means that, with probability approximately $1 - 2\alpha$,

$$-z_\alpha \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \leq z_\alpha.$$

This can be rewritten as the confidence statement $\mu = \overline{X}_n \pm z_\alpha S_n/\sqrt{n}$ in the usual manner. For large n its confidence level should be close to $1 - 2\alpha$.

As another example, consider maximum likelihood estimators $\hat{\theta}_n$ based on a sample of size n from a density p_θ. A major result in asymptotic statistics is that in many situations $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normally distributed with zero mean and covariance matrix the inverse of the Fisher information matrix I_θ. If Z is k-variate normally distributed with mean zero and nonsingular covariance matrix Σ, then the quadratic form $Z^T \Sigma^{-1} Z$ possesses a chi-square distribution with k degrees of freedom. Thus, acting as if $\sqrt{n}(\hat{\theta}_n - \theta)$ possesses an $N_k(0, I_\theta^{-1})$ distribution, we find that the ellipsoid

$$\left\{\theta : (\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n} (\theta - \hat{\theta}_n) \leq \frac{\chi^2_{k,\alpha}}{n}\right\}$$

is an approximate $1 - \alpha$ confidence region, if $\chi^2_{k,\alpha}$ is the appropriate critical value from the chi-square distribution. A closely related alternative is the region based on inverting the likelihood ratio test, which is also based on an asymptotic approximation.

1.2 Asymptotic Optimality Theory

For a relatively small number of statistical problems there exists an exact, optimal solution. For instance, the Neyman-Pearson theory leads to optimal (uniformly most powerful) tests

in certain exponential family models; the Rao-Blackwell theory allows us to conclude that certain estimators are of minimum variance among the unbiased estimators. An important and fairly general result is the Cramér-Rao bound for the variance of unbiased estimators, but it is often not sharp.

If exact optimality theory does not give results, be it because the problem is untractable or because there exist no "optimal" procedures, then asymptotic optimality theory may help. For instance, to compare two tests we might compare approximations to their power functions. To compare estimators, we might compare asymptotic variances rather than exact variances. A major result in this area is that for smooth parametric models maximum likelihood estimators are asymptotically optimal. This roughly means the following. First, maximum likelihood estimators are asymptotically consistent: The sequence of estimators converges in probability to the true value of the parameter. Second, the rate at which maximum likelihood estimators converge to the true value is the fastest possible, typically $1/\sqrt{n}$. Third, their asymptotic variance, the variance of the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$, is minimal; in fact, maximum likelihood estimators "asymptotically attain" the Cramér-Rao bound. Thus asymptotics justify the use of the maximum likelihood method in certain situations. It is of interest here that, even though the method of maximum likelihood often leads to reasonable estimators and has great intuitive appeal, in general it does not lead to best estimators for finite samples. Thus the use of an asymptotic criterion simplifies optimality theory considerably.

By taking limits we can gain much insight in the structure of statistical experiments. It turns out that not only estimators and test statistics are asymptotically normally distributed, but often also the whole sequence of statistical models converges to a model with a normal observation. Our good understanding of the latter "canonical experiment" translates directly into understanding other experiments asymptotically. The mathematical beauty of this theory is an added benefit of asymptotic statistics. Though we shall be mostly concerned with normal limiting theory, this theory applies equally well to other situations.

1.3 Limitations

Although asymptotics is both practically useful and of theoretical importance, it should not be taken for more than what it is: approximations. Clearly, a theorem that can be interpreted as saying that a statistical procedure works fine for $n \to \infty$ is of no use if the number of available observations is $n = 5$.

In fact, strictly speaking, most asymptotic results that are currently available are logically useless. This is because most asymptotic results are limit results, rather than approximations consisting of an approximating formula plus an accurate error bound. For instance, to estimate a value a, we consider it to be the 25th element $a = a_{25}$ in a sequence $a_1, a_2, \ldots$, and next take $\lim_{n\to\infty} a_n$ as an approximation. The accuracy of this procedure depends crucially on the choice of the sequence in which a_{25} is embedded, and it seems impossible to defend the procedure from a logical point of view. This is why there is good asymptotics and bad asymptotics and why two types of asymptotics sometimes lead to conflicting claims.

Fortunately, many limit results of statistics do give reasonable answers. Because it may be theoretically very hard to ascertain that approximation errors are small, one often takes recourse to simulation studies to judge the accuracy of a certain approximation.

Just as care is needed if using asymptotic results for approximations, results on asymptotic optimality must be judged in the right manner. One pitfall is that even though a certain procedure, such as maximum likelihood, is asymptotically optimal, there may be many other procedures that are asymptotically optimal as well. For finite samples these may behave differently and possibly better. Then so-called higher-order asymptotics, which yield better approximations, may be fruitful. See e.g., [7], [52] and [114]. Although we occasionally touch on this subject, we shall mostly be concerned with what is known as "first-order asymptotics."

1.4 The Index n

In all of the following n is an index that tends to infinity, and *asymptotics* means taking limits as $n \to \infty$. In most situations n is the number of observations, so that usually asymptotics is equivalent to "large-sample theory." However, certain abstract results are pure limit theorems that have nothing to do with individual observations. In that case n just plays the role of the index that goes to infinity.

1.5 Notation

A symbol index is given on page xv.

For brevity we often use operator notation for evaluation of expectations and have special symbols for the empirical measure and process.

For P a measure on a measurable space $(\mathcal{X}, \mathcal{B})$ and $f : \mathcal{X} \mapsto \mathbb{R}^k$ a measurable function, Pf denotes the integral $\int f \, dP$; equivalently, the expectation $\mathrm{E}_P f(X_1)$ for X_1 a random variable distributed according to P. When applied to the empirical measure $\mathbb{P}_n$ of a sample $X_1, \ldots, X_n$, the discrete uniform measure on the sample values, this yields

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

This formula can also be viewed as simply an abbreviation for the average on the right. The empirical process $\mathbb{G}_n f$ is the centered and scaled version of the empirical measure, defined by

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(X_i) - \mathrm{E}_P f(X_i) \big).$$

This is studied in detail in Chapter 19, but is used as an abbreviation throughout the book.