however, and it can also be measured for changes to assumptions in non-Bayesian inference.

There have been some successful attempts to take considerations of sensitivity to assumptions into account explicitly in the formulation of Bayesian models. Model uncertainty (see Draper 1995) and Bayesian model averaging (Hoeting et al. 1999) attempt to develop methodology for expressing directly in the modeling process the belief that more than one out of several competing statistical models might be appropriate for the data, and it may not be clear which one is best. For example, one might not be able to decide whether to model data as normal random variables or as Cauchy random variables. The methods of model uncertainty and model averaging provide ways to pursue both models (as well as others that might seem plausible) without having to pretend as if we really believed firmly in precisely one of them.

Finally, there have been some methods developed to deal with specific common violations of assumptions. The assumption that all observations in a sample have the same distribution has received attention particularly in order to be able to accommodate the occasional outlying observation. One can model data as coming from a mixture of two or more distributions. One of the distributions is the main component on which most interest is focussed. The others represent the types of data that might arise when something goes wrong with the data collection procedure. Each observation can be modeled as coming from one of the component distributions with a probability associated with each component. Indeed, using powerful simulation methods (see *Markov Chain Monte Carlo Methods*) one can even compute, for each observation, the conditional probability that it arose from each of the components given the observed data.

5. Conclusion

This article has discussed the need for carefully acknowledging the probabilistic assumptions made to justify a statistical procedure and some methods for assessing the sensitivity of the procedure to those assumptions. If one believes that there are violations of the assumptions that might reasonably arise in practice, and if the procedure is overly sensitive to violations of those assumptions, one might wish to select an alternative procedure.

There are two popular approaches to choosing alternatives to standard statistical procedures when one fears that assumptions are likely to be violated. One is to use robust procedures, and the other is to use nonparametric procedures. Robust procedures are designed to be less sensitive to violations of specific assumptions without sacrificing too much of the good performance of standard procedures when the standard assumptions hold. Nonparametric procedures are chosen so that their properties can be verified under fewer assumptions. For more details on these types of procedures see *Robustness in Statistics* and *Nonparametric Statistics: The Field*.

See also: Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Robustness in Statistics; Statistics: The Field; Time Series: ARIMA Methods; Time Series: General

Bibliography

Benjamini Y 1983 Is the t test really conservative when the parent distribution is long-tailed? *Journal of the American Statistical Association* **78**: 645–54

Box G E P 1953 Non-normality and tests on variances. *Biometrika* **40**: 318–35

Box G E P 1980 Sampling and Bayes' inference in scientific modeling and robustness (with discussion). *Journal of the Royal Statistical Society*, (*Series A*) **143**: 383–430

Draper D 1995 Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society*, (*Series B*) **57**: 45–97

Efron B 1969 Student's t-test under symmetry conditions. *Journal of the American Statistical Association* **64**: 1278–302

Hoeting J A, Madigan D, Raftery A E, Volinsky C 1999 Bayesian model averaging: A tutorial (with discussion). *Statistical Science* **14**: 382–417

Hotelling H 1961 The behavior of some standard statistical tests under nonstandard conditions. In: Neyman J (ed.) *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA, Vol.1, pp. 319–59

M. J. Schervish

Estimation: Point and Interval

Introduction

When sampling is from a population described by a density or mass function $f(x \mid \theta)$, knowledge of θ yields knowledge of the entire population. Hence, it is natural to seek a method of finding a good estimator of the point θ, that is, a good point estimator. However, a point estimator alone is not enough for a complete inference, as a measure of uncertainty is also needed. For that, we use a set estimator in which the inference is the statement that $\theta \in C$ where $C \subset \Theta$ and $C = C(\mathbf{x})$ is a set determined by the value of the data $\mathbf{X} = \mathbf{x}$ observed. If θ is real-valued, then we usually prefer the set estimate C to be an interval. Our uncertainty is quantified by the size of the interval and its probability of covering the parameter.

1. Point Estimators

In many cases, there will be an obvious or natural candidate for a point estimator of a particular parameter. For example, the sample mean is a natural candidate for a point estimator of the population

mean. However, when we leave a simple case like this, intuition may desert us so it is useful to have some techniques that will at least give us some reasonable candidates for consideration. Those that have stood the test of time include:

1.1 The Method of Moments

The method of moments (MOM) is, perhaps, the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. One of the strengths of MOM estimators is that they are usually simple to use and almost always yield some sort of estimate. In many cases, unfortunately, this method yields estimators that may be improved upon.

Let $X_1,\ldots, X_n$ be a sample from a population with density or mass function $f(x|\theta_1,\ldots, \theta_k)$. MOM estimators are found by equating the first k sample moments to the corresponding k population moments. That is, we define the sample moments by $m_j = \Sigma_{i=1}^n X_i^j/n$ and the population moments by $\mu_j(\theta_1,\ldots, \theta_k) = \mathrm{E}X^j$ for $j = 1,\ldots, k$. We then set $m_j = \mu_j(\theta_1,\ldots, \theta_k)$ and solve for $\theta_1,\ldots, \theta_k$. This solution is the MOM estimator of $\theta_1,\ldots, \theta_k$.

1.2 Maximum Likelihood Estimators

For a sample $X_1,\ldots, X_n$ from $f(x\,|\,\theta_1,\ldots, \theta_k)$, the likelihood function is defined by

$$\begin{aligned} L(\theta\,|\,\mathbf{x}) &= L(\theta_1,\ldots, \theta_k\,|\,x_1,\ldots, x_n) \\ &= \Pi_{i=1}^n f(x_i\,|\,\theta_1,\ldots, \theta_k) \end{aligned} \tag{1}$$

The values of θ_i that maximize this function are those parameter values for which the observed sample is most likely, and are called the maximum likelihood estimators (MLE). If the likelihood function is differentiable (in θ_i), the MLEs can often be found by solving

$$\frac{\partial}{\partial\theta_i}\log L(\theta\,|\,\mathbf{x}) = 0, \quad i = 1,\ldots, k \tag{2}$$

where the vector with coordinates $\frac{\partial}{\partial\theta_i}\log L(\theta\,|\,\mathbf{x})$ is called the score function (see Schervish 1995, Sect. 2.3).

Example If $X_1,\ldots, X_n$ are i.i.d. Bernoulli (p), the likelihood function is

$$L(p\,|\,\mathbf{x}) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \tag{3}$$

and differentiating $\log L(p\,|\,\mathbf{x})$ and setting the result equal to zero gives the MLE $\hat{p} = \Sigma_i x_i/n$. This is also the method of moments estimator.

If we instead have samples $X_1,\ldots, X_n$ from a binomial (k, p) population where p is known and k is unknown, the likelihood function is

$$L(k\,|\,\mathbf{x}, p) = \prod_{i=1}^n \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i} \tag{4}$$

and the MLE must be found by the numerical maximization. The method of moments will give the closed form solution.

$$\hat{k} = \frac{\bar{x}^2}{\bar{x} - (1/n)\Sigma_i(x_i-\bar{x})^2} \tag{5}$$

which can take on negative values. This illustrates a shortcoming of the method of moments, one not shared by the MLE. Another, perhaps more serious shortcoming of the MOM estimator is that it may not be based on a sufficient statistic (see *Statistical Sufficiency*), which means it could be inefficient in not using all of the available information in a sample. In contrast, both MLEs and Bayes estimators (see *Bayesian Statistics*) are based on sufficient statistics.

1.3 Bayes Estimators

In the Bayesian paradigm a random sample $X_1,\ldots, X_n$ is drawn from a population indexed by θ, where θ is considered to be a quantity whose variation can be described by a probability distribution (called the *prior distribution*). A sample is then taken from a population indexed by θ and the prior distribution is updated with this sample information. The updated prior is called the *posterior distribution*.

If we denote the prior distribution, by $\pi(\theta)$, and the sampling distribution by $f(\mathbf{x}\,|\,\theta)$, then the posterior distribution, the conditional distribution of θ given the sample, $\mathbf{x}$, is

$$\pi(\theta\,|\,\mathbf{x}) = f(\mathbf{x}\,|\,\theta)\pi(\theta)/m(\mathbf{x}) \tag{6}$$

where $m(\mathbf{x}) = \int f(\mathbf{x}\,|\,\theta)\pi(\theta)\,\mathrm{d}\theta$ is the marginal distribution of $\mathbf{x}$.

Example Let $X_1,\ldots, X_n$ be i.i.d. Bernoulli (p). Then $Y = \Sigma_i X_i$ is binomial (n, p). If p has a Beta (α, β) prior distribution, that is,

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{7}$$

the posterior distribution of p given y, is

$$\begin{aligned} f(p\,|\,y) &= \frac{f(y, p)}{f(y)} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} \end{aligned} \tag{8}$$

which is a beta distribution with parameters $y+\alpha$ and $n-y+\beta$. The posterior mean, a Bayes estimator of p, is

$$\hat{p}_B = \frac{y+\alpha}{\alpha+\beta+n} \tag{9}$$

2. Evaluating Point Estimators

There are many methods of deriving point estimators (robust methods, least squares, estimating equations, invariance) but the three in Sect. 1 are among the most popular. No matter what method is used to derive a point estimator, it is important to evaluate the estimator using some performance criterion.

One way of evaluating the performance of a point estimator W of a parameter θ is through its mean squared error (MSE), defined by $E_\theta(W-\theta)^2$.

MSE measures the average squared difference between the estimator W and the parameter θ. Although any increasing function of the absolute distance $|W-\theta|$ would serve, there is a nice factorization

$$\begin{aligned} E_\theta(W-\theta)^2 &= \mathrm{Var}_\theta W + (E_\theta W-\theta)^2 \\ &= \mathrm{Var}_\theta W + (\mathrm{Bias}_\theta W)^2 \end{aligned} \tag{10}$$

where we define the bias of a point estimator as $\mathrm{Bias}_\theta W = E_\theta W-\theta$. An estimator whose bias is identically (in θ) equal to zero is called *unbiased*.

For an unbiased estimator we have $E_\theta(W-\theta)^2 = \mathrm{Var}_\theta W$, and so, if an estimator is unbiased, its MSE is equal to its variance. If $X_1,\ldots, X_n$ are i.i.d. from a population with mean μ and variance σ^2, the sample mean is an unbiased estimator since $E\bar{X} = \mu$, and has MSE

$$E(\bar{X}-\mu)^2 = \mathrm{Var}\bar{X} = \frac{\sigma^2}{n} \tag{11}$$

Controlling bias does not guarantee that MSE is minimized. In particular, it is sometimes the case that a trade-off occurs between variance and bias. For example, in sampling from a normal population with variance σ^2, the usual unbiased estimator of the variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ has MSE $= 2\sigma^4/(n-1)$. An alternative estimator for σ^2 is the maximum likelihood estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{n-1}{n}S^2$. This is a biased estimator of σ^2 with MSE

$$\begin{aligned} E(\hat{\sigma}^2-\sigma^2)^2 &= \left(\frac{2n-1}{n^2}\right)\sigma^4 < \left(\frac{2}{n-1}\right)\sigma^4 \\ &= E(S^2-\sigma^2)^2 \end{aligned} \tag{12}$$

showing that $\hat{\sigma}^2$ has smaller MSE than S^2. Thus, by trading off variance for bias, the MSE is improved.

Measuring performance by the squared difference between the estimator and a parameter is a special case of a function called a *loss function*. The study of the performance, and the optimality, of estimators evaluated through loss functions is a branch of decision theory. In addition to MSE, based on squared error loss, another popular loss function is absolute error loss, $L(\theta, W) = |W-\theta|$. Both of these loss functions increase as the distance between θ and W increases, with minimum value $L(\theta, \theta) = 0$. That is, the loss is minimum if the estimator is correct.

In a decision theoretic analysis, the worth of an estimator is quantified in its risk function, that is, for an estimator W of θ, the risk function is $R(\theta, W) = E_\theta L(\theta, W)$, so the risk function is the average loss. If the loss is squared error, the risk function is the MSE.

Using squared error loss, the risk function (MSE) of the binomial Bayes estimator of p is

$$\begin{aligned} E_p(\hat{p}_B-p)^2 &= \mathrm{Var}_p\hat{p}_B + (\mathrm{Bias}_p\hat{p}_B)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n}-p\right)^2 \end{aligned} \tag{13}$$

In the absence of good prior information about p, we might try to choose α and β to make the risk function of $\hat{p}_B$ constant (called an *equalizer* rule). The solution is to choose $\alpha = \beta = \sqrt{n/4}$, yielding

$$E(\hat{p}_B-p)^2 = \frac{n}{4(n+\sqrt{n})^2}$$

We can also use a Bayesian approach to the problem of loss function optimality, where we would use the prior distribution to compute an average risk $\int_\Theta R(\theta, W)\pi(\theta)d\theta$, known as the *Bayes risk*. We then find the estimator that yields the smallest value of the Bayes risk. Such an estimator is called the Bayes rule with respect to a prior π.

To find the Bayes decision rule for a given prior π we write the Bayes risk as

$$\begin{aligned} &\int_\Theta R(\theta, W)\,\pi(\theta)\,d\theta \\ &\quad = \int_\chi \left[\int_\Theta L(\theta, W(\mathbf{x}))\,\pi(\theta\,|\,\mathbf{x})\,d\theta\right] m(\mathbf{x})\,d\mathbf{x} \end{aligned} \tag{14}$$

where the quantity in square brackets is the expected value of the loss function with respect to the posterior distribution, called the *posterior expected loss*. It is a function only of $\mathbf{x}$, and not a function of θ. Thus, for each $\mathbf{x}$, if we choose the estimate $W(\mathbf{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk.

For squared error loss, the posterior expected loss is minimized by the mean of the posterior distribution. For absolute error loss, the posterior expected loss is minimized by the median of the posterior distribution. If we have a sample $X_1,\ldots, X_n$ from a normal distribution with mean θ and variance σ^2, and the prior is $n(\mu, \tau^2)$, the posterior mean is

$$E(\theta \mid \bar{x}) = \frac{\tau^2}{\tau^2 + (\sigma^2/n)}\,\bar{x} + \frac{\sigma^2/n}{\tau^2 + (\sigma^2/n)}\,\mu \qquad (15)$$

Since the posterior distribution is normal, it is symmetric and the posterior mean is the Bayes rule for both squared error and absolute error loss. The posterior mean in our binomial/beta example, $\hat{p}_B = \frac{y+\alpha}{\alpha+\beta+n}$, is the Bayes estimator against squared error loss.

We can loosely group evaluation criteria into large sample or asymptotic methods, and small sample methods. Our calculations using MSE and risk functions illustrate small sample methods. In large samples, MLEs typically perform very well, being asymptotically normal and efficient, that is, attaining the smallest possible variance. Other types of estimators that are derived in a similar manner (for example, M-estimators—see *Robustness in Statistics*) also share good asymptotic properties. For a detailed discussion see Casella and Berger (2001), Lehmann (1998), Stuart et al. (1999) or Lehmann and Casella (1998, Chap. 6).

3 Interval Estimation

Reporting a point estimator of a parameter θ only provides part of the story. The story becomes more complete if an assessment of the error of estimation is also reported. Informally, this can be accomplished by giving an estimated standard error of the estimator and, more formally, this becomes the reporting of an interval estimate. If $\mathbf{X} = \mathbf{x}$ is observed, an *interval estimate* of a parameter θ is a pair of functions, $L(\mathbf{x})$ and $U(\mathbf{x})$ for which the inference $\theta \in [L(\mathbf{x}), U(\mathbf{x})]$ is made. The *coverage probability* of the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is the probability that $[L(\mathbf{X}), U(\mathbf{X})]$ covers the true parameter, θ, and is denoted by $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$.

By definition, the coverage probability depends on the unknown θ, so it cannot be reported. What is typically reported is the confidence coefficient, the infimum of the coverage probabilities, $\inf_\theta P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$.

If $X_1,\ldots, X_n$ are i.i.d. with mean μ and variance σ^2, a common interval estimator for μ is

$$\mu \in \bar{x} \pm 2\frac{s}{\sqrt{n}} \qquad (16)$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation. The validity of this interval can be justified from the Central Limit Theorem, because

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \to n(0, 1) \qquad (17)$$

the standard normal distribution. We then see that the coverage probability (and confidence coefficient) of Eqn. 16 is approximately 95 percent.

The above interval is a large sample interval since its justification is based on an asymptotic argument. There are many methods for constructing interval estimators that are valid in small samples, including these.

3.1 Inverting a Test Statistic

There is a correspondence between acceptance regions of tests (see *Hypothesis Testing in Statistics*) and confidence sets, summarized in the following theorem.

Theorem *For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of H_0: $\theta = \theta_0$. For each $\mathbf{x} \in \chi$, define a set $C(\mathbf{x})$ in the parameter space by*

$$C(\mathbf{x}) = \{\theta_0 \colon \mathbf{x} \in A(\theta_0)\} \qquad (18)$$

Then the random set $C(\mathbf{X})$ is a $1-\alpha$ confidence set. Conversely, let $C(\mathbf{X})$ be a $1-\alpha$ confidence set. For any $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{\mathrm{x} \colon \theta_0 \in C(\mathbf{x})\} \qquad (19)$$

Then $A(\theta_0)$ is the acceptance region of a level α test of H_0: $\theta = \theta_0$.

Example If $X_1,\ldots, X_n$ are i.i.d. $n(\mu, \sigma^2)$, with σ^2 known, the test of $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$ will accept the null hypothesis at level α if

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu_0 \leqslant \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (20)$$

The interval of μ values, $[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{\alpha/2}\sigma\sqrt{n}]$, for which the null hypothesis will be accepted at level α, is a $1-\alpha$ confidence interval for μ.

3.2 Pivotal Inference

Perhaps one of the most elegant methods of constructing set estimators is the use of pivotal quantities (Barnard 1949). A random variable $Q(\mathbf{X}, \theta) = Q(X_1,\ldots, X_n, \theta)$, is a *pivotal quantity* (or pivot) if the distribution of $Q(\mathbf{X}, \theta)$ is independent of all parameters. If we find a set C such that $P(Q(\mathbf{X}, \theta) \in C) =$

$1-\alpha$, then the set $\{\theta: Q(\mathbf{X}, \theta) \in C\}$ has coverage probability $1-\alpha$.

In location and scale cases, once we calculate the sample mean $\bar{X}$ and the sample standard deviation S, we can construct the following pivots:

Form of pdf	Type of pdf	Pivotal quantity
$f(x-\mu)$	location	$\bar{X}-\mu$
$\frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$	scale	$\frac{\bar{X}}{\sigma}$
$\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$	location-scale	$\frac{\bar{X}-\mu}{S}$

(21)

In general, differences are pivotal for location problems, while ratios (or products) are pivotal for scale problems. See also *Fiducial and Structural Statistical Inference*.

Example Suppose that $X_1, \ldots, X_n$ are i.i.d. exponential (λ). Then $T = \sum_i X_i$ is a sufficient statistic for λ and $T \sim$ gamma (n, λ). In the gamma pdf, t and λ appear together as t/λ and, in fact the gamma (n, λ) pdf $(\Gamma(n)\lambda^n)^{-1} t^{n-1} e^{-t/\lambda}$ is a scale family. Thus, if $Q(T, \lambda) = 2T/\lambda$, then

$$Q(T, \lambda) \sim \text{gamma}(n, \lambda(2/\lambda)) = \text{gamma}(n, 2)$$

which does not depend on λ. The quantity $Q(T, \lambda) = 2T/\lambda$ is a pivot with a gamma $(n, 2)$, or χ^2_{2n}, distribution, and a $1-\alpha$ pivotal interval is

$$\frac{2T}{\chi^2_{2n,\,\alpha/2}} \leqslant \lambda \leqslant \frac{2T}{\chi^2_{2n,\,1-\alpha/2}}$$

where $P(\chi^2_{2n} > \chi^2_{2n,\,a}) = a$.

3.3 Bayesian Intervals

If $\pi(\theta \mid \mathbf{x})$ is the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$, then for any set $A \subset \Theta$ the posterior probability of A is

$$P(\theta \in A \mid \mathbf{x}) = \int_A \pi(\theta \mid \mathbf{x})\, d\theta$$

If $A = A(\mathbf{x})$ is chosen so that this posterior probability is $1-\alpha$, then A is called a $1-\alpha$ *credible set* for θ. If $\pi(\theta \mid \mathbf{x})$ corresponds to a discrete distribution, we replace integrals with sums in the above expressions.

The interpretation of the Bayes interval estimator is different from the classical intervals. In the classical approach, to assert 95 percent coverage is to assert that in 95 percent of repeated experiments, the realized intervals will cover the true parameter. In the Bayesian approach, a 95 percent coverage means that the probability is 95 percent that the parameter is in the realized interval. In the classical approach the randomness comes from the repetition of experiments, while in the Bayesian approach the randomness comes from uncertainty about the value of the parameter (summarized in the prior distribution).

Example Let $X_1, \ldots, X_n$ be i.i.d. Poisson (λ) and assume that λ has a gamma prior, $\lambda \sim$ gamma (a, b), where a is an integer. The posterior distribution of λ is

$$\pi(\lambda \mid \textstyle\sum_i X_i = \sum_i x_i) = \text{gamma}(a + \sum_i x_i, [n + (1/b)]^{-1}) \tag{22}$$

Thus the posterior distribution of $2[n+(1/b)]\lambda$ is $\chi^2_{2(a+\Sigma x)}$, and a $1-\alpha$ Bayes credible interval for λ is

$$\left\{ \lambda: \frac{\chi^2_{2(a+\Sigma x),\,1-\alpha/2}}{2[n+(1/b)]} \leqslant \lambda \leqslant \frac{\chi^2_{2(a+\Sigma x),\,\alpha/2}}{2[n+(1/b)]} \right\} \tag{23}$$

We can also form a Bayes credible set by taking the highest posterior density (HPD) region of the parameter space, by choosing c so that

$$1-\alpha = \int_{\{\lambda: \pi(\lambda \mid \Sigma x) \geqslant c\}} \pi(\lambda \mid \Sigma_i x_i)\, d\lambda \tag{24}$$

Such a construction is optimal in the sense of giving the shortest interval for a given $1-\alpha$ (although if the posterior is multimodal the set may not be an interval).

4. Other Intervals

We have presented two-sided parametric intervals that are constructed to cover a parameter. Other types of intervals include (*a*) one-sided intervals, (*b*) distribution-free intervals, (*c*) prediction intervals, (*d*) tolerance intervals.

One-sided intervals are those in which only one endpoint is estimated, such as $\theta \in [L(\mathbf{X}), \infty)$. Distribution-free intervals are intervals whose probability guarantee holds with little (or no) assumption on the underlying distribution. The other two interval definitions, together with the usual confidence interval, provide us with a hierarchy of inferences, each more stringent than the previous.

If $X_1, \ldots, X_n$ are i.i.d. from a population with cdf $F(x \mid \theta)$, and $C(\mathbf{x}) = [L(\mathbf{x}), U(\mathbf{x})]$ is an interval, for a specified value $1-\alpha$ it is a:

(a) *confidence interval* if, for all θ, $P_\theta[L(\mathbf{X}) \leqslant \theta \leqslant U(\mathbf{X})] \geqslant 1-\alpha$;

(b) *prediction interval* if, for all θ, $P_\theta[L(\mathbf{X}) \leqslant X_{n+1} \leqslant U(\mathbf{X})] \geqslant 1-\alpha$;

(c) *tolerance interval* if, for all θ and for a specified value p, $P_\theta([F(U(\mathbf{X})\,|\,\theta)-F(L(\mathbf{X})\,|\,\theta) \geqslant p] \geqslant 1-\alpha$.

So a confidence interval covers a mean, a prediction interval covers a new random variable, and a tolerance interval covers a proportion of the population. Thus, each gives a different inference, with the appropriate one being dictated by the problem at hand. Vardeman (1992) argues that these 'other intervals' provide an inference that is different from that of a confidence interval and are as important as confidence intervals.

5. Conclusions

Point estimation is one of the cornerstones of statistical analysis, and the basic element on which many inferences are based. Inferences using point estimators gain statistical validity when they are accompanied by an interval estimate, providing an assessment of the uncertainty. We have mainly discussed parametric point and interval estimation, where we assume that the underlying model is correct. Such an assumption can be questioned, and considerations of nonparametric or robust alternatives can address this (see *Robustness in Statistics*). For more on these subjects see, for example, Hettmansberger and McKean (1998) or Staudte and Sheather (1990). Full treatments of parametric point and interval estimation can be found in Casella and Berger (2001), Stuart et al. (1999), or Schervish (1995).

Bibliography

Barnard G A 1949 Statistical inference (with discussion). *Journal of the Royal Statistical Society, Series B* **11**: 115–39

Casella G, Berger R L 2001 *Statistical Inference*, 2nd edn. Wadsworth/Brooks Cole, Pacific Grove, CA

Hettmansperger T P, McKean J W 1998 *Robust Nonparametric Statistical Methods*. Wiley, New York

Lehmann E L 1998 *Introduction to Large-Sample Theory*. Springer-Verlag, New York

Lehmann E L, Casella G 1998 *Theory of Point Estimation*, 2nd edn. Springer-Verlag, New York

Schervish M J 1995 *Theory of Statistics*. Springer-Verlag, New York

Staudte R G, Sheather S J 1990 *Robust Estimation and Testing*. John Wiley, New York

Stuart A, Ord J K, Arnold S 1999 *Advanced Theory of Statistics, Classical Inference and the Linear Model,* 6th edn. Arnold, Oxford University Press, London, Vol. 2A

Vardeman S B 1992 What about other intervals. *The American Statistician* **46**: 193–7

G. Casella and R. L. Berger

Ethical Behavior, Evolution of

The distinction between good or bad and what we ought or ought not do constitutes the subject matter of ethics. Early students of the evolution of ethical behavior (EEB) and some sociobiologists attempted to direct the study of EEB into the domain of prescriptive ethics. Twenty-first century sociobiologists are not concerned with the nature of vice, virtue, or the rules of moral behavior, but with the question of the biological origin of ethical behavior. To disregard the distinction between the questions of ethics and the questions of the evolutionary causes and the bases of human ethical behavior is to misunderstand the discipline of EEB, which is not a branch of philosophy or of ethics, but a subdiscipline of sociobiology. No prescriptive code can be derived from the theory of evolution; therefore, the previously used term 'evolutionary ethics' is a misnomer that must be dropped from sociobiological usage. An excellent philosophical examination (Flew 1967) and a superior historical analysis (Richards 1987) on EEB are available.

The ideas on EEB will be exemplified by a review of arguments, first, of the originators of EEB; second, of ecologists, ethologists, geneticists, linguists, neurophysiologists, and cognitive scientists; and third, of sociobiologists and their supporters and opponents.

1. The Originators of the Idea

Charles Darwin, in *On the Origin of Species* (1859), established the homology between human and nonhuman primate morphology, but not between human and nonhuman behavior. In his revised edition of *The Descent of Man* (1885), he recognized that the moral sense or conscience differentiates humans from the other animals, and that 'The imperious word *ought* seems merely to imply the consciousness of the existence of a rule of conduct, however it may have originated.' He believed that the understanding of the EEB would be enlarged by studies of nonhuman behavior, and that nonhuman animals possessing parental and filial instinct would exhibit a kind of rudimentary intellectual and moral sense. He supported this by arguing that, first, social instincts lead animals to social groups and to perform 'social services'; second, as the mental faculties develop, images and feelings endure and are subject to recall; third, development of language facilitates communication; and last, acquisition of habits helps to guide the conduct of individuals within the community.

Humans, to Darwin, are social animals who have lost some early instincts but retain the protohuman instinctive love and sympathy for others. Some instincts are subject to group selection; the altruistic social behaviors of early humans were not for the good

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7