

Preliminaries; Data-Generating Processes and Statistical Models

Our understanding of phenomena is facilitated by means of a *model*.

A model is a description of the phenomenon of interest. We can formulate a model either as a description of a *data-generating process*, or as a prescription for processing data.

The model is often expressed as a set of equations that relate data elements to each other. It may include probability distributions for the data elements. If any of the data elements are considered to be realizations of random variables, the model is a *stochastic model*.

Models

A class of models may have a common form within which the members of the class are distinguished by values of *parameters*.

In models that are not mathematically tractable computationally intensive methods involving simulations, resamplings, and multiple views may be used to make inferences about the parameters of a model.

Structure in Data

The components of statistical datasets are “observations” and “variables” .

In general, “data structures” are ways of organizing data to take advantage of the relationships among the variables constituting the dataset. Data structures may express hierarchical relationships, crossed relationships (as in “relational” databases), or more complicated aspects of the data (as in “object-oriented” databases).

In data analysis, “structure in the data” is of interest.

Structure in Data

Structure in the data includes such nonparametric features as modes, gaps, or clusters in the data, the symmetry of the data, and other general aspects of the shape of the data.

Because many classical techniques of statistical analysis rely on an assumption of normality of the data, the most interesting structure in the data may be those aspects of the data that deviate most from normality.

Graphical displays may be used to discover qualitative structure in the data.

Model Building

The process of building models involves successive refinements. The evolution of the models proceeds from vague, tentative models to more complete ones, and our understanding of the process being modeled grows in this process.

The usual statements about statistical methods regarding bias, variance, and so on are made in the context of a model.

Model Building

It is not possible to measure bias or variance of a procedure to *select* a model, except in the relatively simple case of selection from some well-defined and simple set of possible models.

Only within the context of rigid assumptions (a “metamodel”) can we do a precise statistical analysis of model selection. Even the simple cases of selection of variables in linear regression analysis under the usual assumptions about the distribution of residuals (and this is a highly idealized situation) present more problems to the analyst than are generally recognized.

Descriptive Statistics, Inferential Statistics, and Model Building

We can distinguish statistical activities that involve:

- data collection;
- descriptions of a given dataset;
- inference within the context of a model or family of models;
and
- model selection.

Once data are available, either from a survey or designed experiment, or just observational data, a statistical analysis begins by considering general descriptions of the dataset.

These descriptions include ensemble characteristics, such as averages and spreads, and identification of extreme points. The descriptions are in the form of various summary statistics and graphical displays.

The descriptive analyses may be computationally intensive for large datasets, especially if there are a large number of variables.

Computational Statistics

The computationally intensive approach also involves multiple views of the data, including consideration of a large number of transformations of the data.

A stochastic model is often expressed as a probability density function or as a cumulative distribution function of a random variable. In a simple linear regression model with normal errors,

$$Y = \beta_0 + \beta_1 x + E,$$

for example, the model may be expressed by use of the probability density function for the random variable E .

The probability density function for Y is

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\beta_0-\beta_1x)^2/(2\sigma^2)}.$$

The elements of a stochastic model include observable random variables, observable covariates, unobservable parameters, and constants.

Statistical Models

The parameters may be considered to be unobservable random variables, and in that sense, a specific data model is defined by a realization of the parameter random variable. In the model, written as

$$Y = f(x; \beta) + E,$$

we identify a “systematic component”, $f(x; \beta)$, and a “random component”, E .

The selection of an appropriate model may be very difficult, and almost always involves not only questions of how well the model corresponds to the observed data, but also the tractability of the model. The methods of computational statistics allow a much wider range of tractability than can be contemplated in mathematical statistics.

Classical Statistical Inference

Formal statistical inference involves use of a sample to make decisions about stochastic models based on probabilities that would result if a given model was indeed the data-generating process.

The heuristic paradigm calls for rejection of a model if the probability is small that data arising from the model would be similar to the observed sample.

Computational Inference

Computationally intensive methods include exploration of a range of models, many of which may be mathematically intractable.

In a different approach employing the same paradigm, the statistical methods may involve direct simulation of the hypothesized data-generating process rather than formal computations of probabilities that would result under a given model of the data-generating process. We refer to this approach as *computational inference*.

In a variation of computational inference, we may not even attempt to develop a model of the data-generating process; rather, we build decision rules directly from the data.

The Empirical Cumulative Distribution Function

Methods of statistical inference are based on an assumption (often implicit) that a discrete uniform distribution with mass points at the observed values of a random sample is asymptotically the same as the distribution governing the data-generating process.

Thus, the distribution function of this discrete uniform distribution is a model of the distribution function of the data-generating process.

For a given set of univariate data, y_1, \dots, y_n , the *empirical cumulative distribution function*, or ECDF, is

$$P_n(y) = \frac{\#\{y_i, \text{ s.t. } y_i \leq y\}}{n}.$$

The ECDF is the basic function used in many methods of computational inference.

The Empirical Cumulative Distribution Function

It is easy to see that the ECDF is pointwise unbiased for the CDF; that is, if the y_i are independent realizations of random variables Y_i , each with CDF $P(\cdot)$, for a given y ,

$$\begin{aligned} \mathbb{E}(P_n(y)) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n \mathbf{I}_{(-\infty, y]}(Y_i)\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left(\mathbf{I}_{(-\infty, y]}(Y_i)\right) \\ &= \Pr(Y \leq y) \\ &= P(y). \end{aligned}$$

Similarly, we find

$$\mathbb{V}(P_n(y)) = P(y)(1 - P(y))/n;$$

indeed, at a fixed point y , $nP_n(y)$ is a binomial random variable with parameters n and $\pi = P(y)$. Because P_n is a function of the order statistics, which form a complete sufficient statistic for P , there is no unbiased estimator of $P(y)$ with smaller variance.

The Empirical Probability Density Function

We also define the *empirical probability density function* (EPDF) as the derivative of the ECDF:

$$p_n(y) = \frac{1}{n} \sum_{i=1}^n \delta(y - y_i),$$

where δ is the Dirac delta function. The EPDF is just a series of spikes at points corresponding to the observed values. It is not as useful as the ECDF. It is, however, unbiased at any point for the probability density function at that point.

The ECDF and the EPDF can be used as estimators of the corresponding population functions, but there are better estimators.

Statistical Functions of the CDF and the ECDF

Statistical Functions of the CDF and the ECDF

In many models of interest, a parameter can be expressed as a functional of the probability density function or of the cumulative distribution function of a random variable in the model. The mean of a distribution, for example, can be expressed as a functional θ of the CDF P :

$$\theta(P) = \int_{\mathbb{R}^d} y \, dP(y).$$

A functional that defines a parameter is called a *statistical function*.

Estimation of Statistical Functions

A common task in statistics is to use a random sample to estimate the parameters of a probability distribution. If the statistic T from a random sample is used to estimate the parameter θ , we measure the performance of T by the magnitude of the bias,

$$|\mathbb{E}(T) - \theta|,$$

by the variance,

$$\mathbb{V}(T) = \mathbb{E}((T - \mathbb{E}(T))^2),$$

by the mean squared error,

$$\mathbb{E}((T - \theta)^2),$$

and by other expected values of measures of the distance from T to θ .

Properties of Estimators

The order of the mean squared error is an important characteristic of an estimator.

For good estimators of location, the order of the mean squared error is typically $O(n^{-1})$.

Good estimators of probability densities, however, typically have mean squared errors of at least order $O(n^{-4/5})$.

Estimation Using the ECDF

There are many ways to construct an estimator and to make inferences about the population. In the univariate case especially, we often use data to make inferences about a parameter by applying the statistical function to the ECDF. An estimator of a parameter that is defined in this way is called a *plug-in estimator*.

A plug-in estimator for a given parameter is the same functional of the ECDF as the parameter is of the CDF.

Plug-In Estimators

For the mean of the model, for example, we use the estimate that is the same functional of the ECDF as the population mean,

$$\begin{aligned}\Theta(P_n) &= \int_{-\infty}^{\infty} y dP_n(y) \\ &= \int_{-\infty}^{\infty} y d\frac{1}{n} \sum_{i=1}^n I_{(-\infty, y]}(y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} y dI_{(-\infty, y]}(y_i) \\ &= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \bar{y}.\end{aligned}$$

The sample mean is thus a plug-in estimator of the population mean. Such an estimator is called a *method of moments estimator*. This is an important type of plug-in estimator. The method of moments results in estimates of the parameters $E(Y^r)$ that are the corresponding sample moments.

Plug-In Estimators

Statistical properties of plug-in estimators are generally relatively easy to determine. In some cases, the statistical properties, such as expectation and variance, are optimal in some sense.

Estimation Using the ECDF

In addition to estimation based on the ECDF, other methods of computational statistics make use of the ECDF. In some cases, such as in bootstrap methods, the ECDF is a surrogate for the CDF. In other cases, such as Monte Carlo methods, an ECDF for an estimator is constructed by repeated sampling, and that ECDF is used to make inferences using the observed value of the estimator from the given sample.

Viewed as a statistical function, Θ denotes a specific functional form. Any functional of the ECDF is a function of the data, so we may also use the notation $\Theta(Y_1, \dots, Y_n)$. Often, however, the notation is cleaner if we use another letter to denote the function of the data; for example, $T(Y_1, \dots, Y_n)$, even if it might be the case that

$$T(Y_1, \dots, Y_n) = \Theta(P_n).$$

Estimation Using the ECDF

Use of the ECDF in statistical inference does not require many assumptions about the distribution. Other methods discussed below are based on information or assumptions about the data-generating process.

Empirical Quantiles

For $\alpha \in (0, 1)$, the α *quantile* of the distribution with CDF P is the value y_α such that $P(y_\alpha) = \alpha$. (For a univariate random variable, this is a single point. For a d -variate random variable, it is a $(d - 1)$ -dimensional object that is generally nonunique.) For a discrete distribution the quantile may not exist for a given value of α .

We also use the term in a slightly different way: if $P(y) = \alpha$, we say the *quantile of y is α* .

This definition of a quantile applied to the ECDF leads to a quantile of 0 for the smallest sample value, $y_{(1)}$, and a quantile of 1 for the largest sample value, $y_{(n)}$. These values for quantiles are not so useful if the distribution is continuous, because it is likely that the range of the distribution extends beyond the smallest and largest observed values.

Empirical Quantiles

We define the *empirical quantile*, or *sample quantile*, corresponding to the i^{th} order statistic, $y_{(i)}$, in a sample of size n as

$$\frac{i - \iota}{n + \nu}$$

for $\iota \in [0, \frac{1}{2}]$ and $\nu \in [0, 1]$. Values of ι and ν that make the empirical quantiles of a random sample correspond closely to those of the population depend on the distribution of the population, which, of course, is generally unknown. A certain symmetry may be imposed by requiring $\nu = 1 - 2\iota$. Common choices are $\iota = \frac{1}{2}$ and $\nu = 0$.

We use empirical quantiles in Monte Carlo inference, in nonparametric inference, and in graphical displays for comparing a sample with a standard distribution or with another sample. Empirical quantiles can be used as estimators of the population quantiles, but there are other estimators.