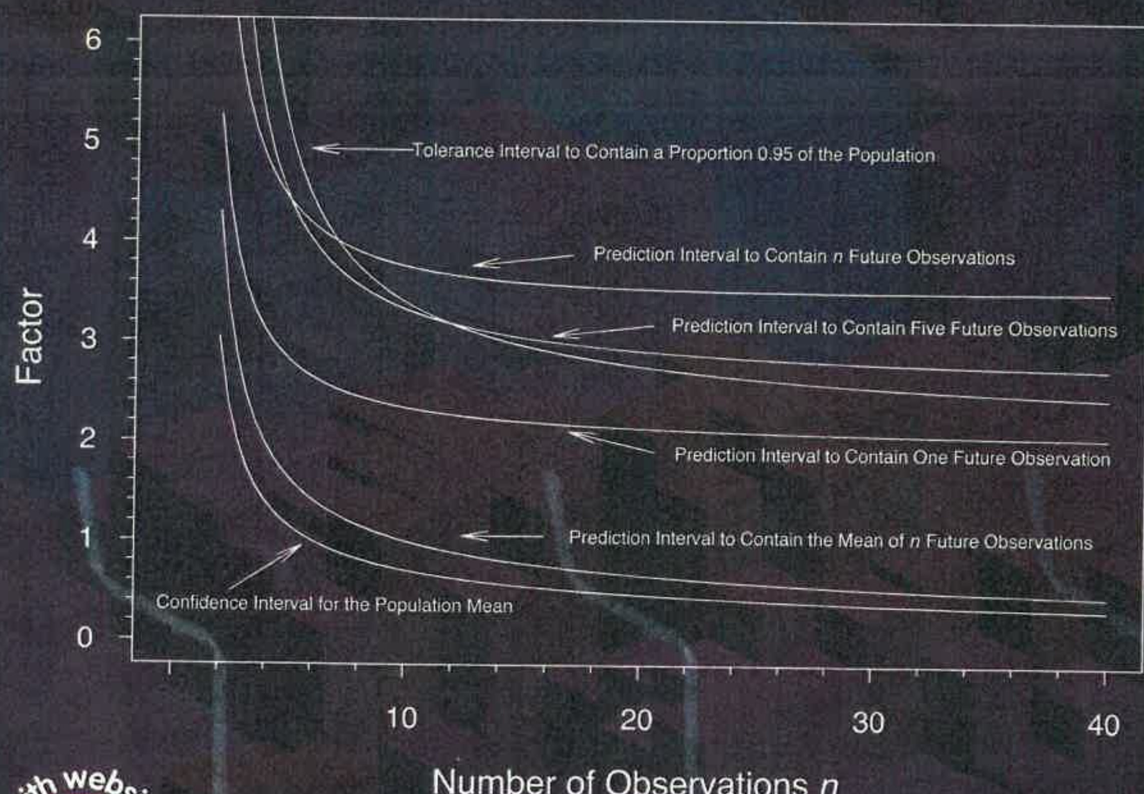
Wiley Series in Probability and Statistics

Second Edition

STATISTICAL INTERVALS

A GUIDE FOR PRACTITIONERS AND RESEARCHERS

William Q. Meeker • Gerald J. Hahn • Luis A. Escobar



with website

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Names: Meeker, William Q. | Hahn, Gerald J. | Escobar, Luis A.
Title: Statistical intervals : a guide for practitioners and researchers.
Description: Second edition / William Q. Meeker, Gerald J. Hahn, Luis A.
Escobar. | Hoboken, New Jersey : John Wiley & Sons, Inc., [2017] |
Includes bibliographical references and index.
Identifiers: LCCN 2016053941 | ISBN 9780471687177 (cloth) | ISBN 9781118595169 (epub)
Subjects: LCSH: Mathematical statistics.
Classification: LCC QA276 .H22 2017 | DDC 519.5/4–dc23 LC record available
at https://lccn.loc.gov/2016053941

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

1.2 DIFFERENT TYPES OF STATISTICAL INTERVALS: AN OVERVIEW

Various types of statistical intervals may be calculated from sample data. The appropriate interval depends upon the specific application. Frequently used intervals are:

- A *confidence interval* to contain an unknown characteristic of the sampled population or process. The quantity of interest might be a population property or "parameter," such as the mean or standard deviation of the population or process. Alternatively, interest might center on some other property of the sampled population, such as a quantile or a probability. Thus, depending upon the question of interest, one might compute a confidence interval that one can claim, with a specified high degree of confidence, contains (1) the mean tensile strength, (2) the standard deviation of the distribution of tensile strengths, (3) the 0.10 quantile of the tensile strength distribution, or (4) the proportion of specimens that exceed a stated threshold tensile strength value.
- A *statistical tolerance interval* to contain a specified proportion of the units from the sampled population or process. For example, based upon a random sample of tensile strength measurements, we might wish to compute an interval to contain, with a specified degree of confidence, the tensile strengths of at least a proportion 0.90 of the units from the sampled population or process. Hereafter we will generally simply refer to such an interval as a "tolerance interval."
- A *prediction interval* to contain one or more future observations, or some function of such future observations, from a previously sampled population or process. For example, based upon a random sample of tensile strength measurements, we might wish to construct an interval to contain, with a specified degree of confidence, (1) the tensile strength of a randomly selected single future unit from the sampled process (this was of interest in the turbine efficiency example), (2) the tensile strengths for *all* of five future units, or (3) the mean tensile strength of five future units.

Most users of statistical methods are familiar with (the common) confidence intervals for the population mean and for the population standard deviation, but often not for population quantiles or the probability of exceeding a specified threshold value. Some, especially in industry, are also aware of tolerance intervals. Despite their practical importance, however, most practitioners, and even many professional statisticians, know very little about prediction intervals except, perhaps, for their application to regression problems. A frequent mistake is to calculate a confidence interval to contain the population mean when the problem requires a tolerance interval or a prediction interval. At other times, a tolerance interval is used when a prediction interval is needed. Such confusion is understandable, because statistics textbooks typically focus on the common confidence intervals, occasionally make reference to tolerance intervals, and consider prediction intervals only in the context of regression analysis. This is unfortunate because in applications, tolerance intervals, prediction intervals, and confidence intervals on distribution quantiles and on exceedance probabilities are needed almost as frequently as the better-known confidence intervals. Moreover, the calculations for such intervals are generally no more difficult than those for confidence intervals.

1.3 THE ASSUMPTION OF SAMPLE DATA

In this book we are concerned only with situations in which uncertainty is present because the available data are from a random sample (often small) from a population or process. There are, of course, some situations for which there is little or no such statistical uncertainty. This is

Equivalently, one would, in the long run, be correct $100(1-\alpha)\%$ of the time in claiming that the actual value of θ is contained within the confidence interval." More commonly, but less precisely, a two-sided confidence interval is described by a statement such as "we are 95% confident that the interval $[\underset{\sim}{\theta}, \widetilde{\theta}]$ contains the unknown actual parameter value θ." In fact, the observed interval either contains θ or does not. Thus the 95% refers to the *procedure* for constructing a confidence interval, and not to the observed interval itself. One-sided confidence bounds can be similarly interpreted.

2.3 PREDICTION INTERVALS

2.3.1 Prediction Interval to Contain a Single Future Observation

A prediction interval for a single future observation is an interval that will, with a specified degree of confidence, contain a future randomly selected observation from a distribution. Such an interval would interest the purchaser of a single unit of a particular product and is generally more relevant to such an individual than, say, a confidence interval to contain average performance. For example, the purchaser of a new automobile might wish to obtain, from the data on a previous sample of five similar automobiles, an interval that contains, with a high degree of confidence, the gasoline mileage that the new automobile will obtain under specified driving conditions. This interval is calculated from the sample data under the important assumption that the previously sampled automobiles and the future one(s) can be regarded as random samples from the same distribution; this assumes identical production processes and similar driving conditions. In many applications, the population may be conceptual, as per our discussion of analytic studies in Chapter 1.

2.3.2 Prediction Interval to Contain All of m Future Observations

A prediction interval to contain the values of all of m future observations generalizes the concept of a prediction interval to contain a single future observation. For example, a traveler who must plan a specific number of trips may not be interested in the amount of fuel that will be needed on the average for all future trips. Instead, this person would want to determine the amount of fuel that will be needed to complete each of, say, one, or three, or five future trips.

Prediction intervals to contain all of m future observations are often of interest to manufacturers of large equipment who produce only a small number of units of a particular type product. For example, a manufacturer of gas turbines might wish to establish an interval that, with a high degree of confidence, will contain the performance values for all three units in a future shipment of such turbines, based upon the observed performance of similar past units. In this example, the past units and those in the future shipment would conceptually be thought of as random samples from the population of all turbines that the manufacturer might build (see the discussion in Section 1.2).

Prediction intervals are especially pertinent to users of one or a small number of units of a product. Such individuals are generally more concerned with the performance of a specific sample of one or more units, rather than with that of the entire process from which the sample was selected. For example, based upon the data from a life test of 10 systems, one might wish to construct an interval that would have a high probability of including the lives of all of three additional systems that are to be purchased. Prediction intervals to contain all of m future observations are often referred to as *simultaneous* prediction intervals, because one is concerned with simultaneously containing *all* of the m observations within the calculated interval (with the associated level of confidence).

2.3.3 Prediction Interval to Contain at Least k out of m Future Observations

A generalization of a prediction interval to contain all m future observations is one to contain at least k out of m of the future observations. We will refer to this type of interval again in Section 2.4.1.

2.3.4 Prediction Interval to Contain the Sample Mean or Sample Standard Deviation of a Future Sample

Sometimes one desires an interval to contain the *sample mean* (or sample standard deviation or other estimated quantity) of a future sample of m observations, rather than one to contain *all* (or at least k) of the future sample values. Such an interval would be pertinent, for example, if acceptance or rejection of a particular design were to be based upon the sample mean of a future sample from a previously sampled distribution.

Consider the following example: A manufacturer of a high voltage insulating material must provide a potential customer "performance limits" to contain average breakdown strength of the material, estimated from a destructive test on a sample of 10 units. Here "average" is understood to be the sample mean of the readings on the units. The tighter these limits, the better are the chances that the manufacturer will be awarded a forthcoming contract. The manufacturer, however, has to provide the customer five randomly selected units for a test. If the sample mean for these five units does not fall within the performance limits stated by the manufacturer, the product is automatically disqualified. The manufacturer has available a random sample of 15 units from production. Ten of these units will be randomly selected and tested by the manufacturer to establish the desired limits. The remaining five units will be shipped to and tested by the customer. Based on the data from the sample of 10 units, the manufacturer will establish prediction limits for the sample mean of the five future readings so as to be able to assert with 95% confidence that the sample mean of the five units to be tested by the customer will lie in the interval. A 95% prediction interval to contain the future sample mean provides the desired limits.

Alternatively, suppose that in the preceding example the concern is uniformity of performance, as measured by the sample standard deviation, rather than sample mean performance. In this case, one might wish to compute a prediction interval, based upon measurements on a random sample of 10 units, to contain the standard deviation of a future sample of five units from the same process or population.

2.3.5 One-Sided Prediction Bounds

Some applications call for a one-sided prediction bound, instead of a two-sided prediction interval, to contain future sample results. An example is provided by a manufacturer who needs to warranty the efficiency of all units (or of their average) for a future shipment of three motors, based upon the results of a sample of five previously tested motors from the same process. This problem calls for a one-sided lower prediction bound, rather than a two-sided prediction interval.

2.3.6 Interpretation of Prediction Intervals and Bounds

If all the parameters of a probability distribution are known, one can compute a probability interval to contain the values of a future sample. For example, for a normal distribution with mean μ and standard deviation σ, the probability is 0.95 that a single future observation will be contained in the interval $\mu \pm 1.96\sigma$. In the more usual situation where one has only sample data, one can construct a $100(1-\alpha)\%$ prediction interval to contain the future observation with a specified degree of confidence. Such an interval may be formally characterized as follows: "If from many independent pairs of random samples, a $100(1-\alpha)\%$ prediction interval is computed from the data of the first sample to contain the value(s) of the second sample, $100(1-\alpha)\%$ of the intervals would, in the long run, correctly bracket the future value(s)." Equivalently, one would, in the long run, be correct $100(1-\alpha)\%$ of the time in claiming that the future value(s) will be contained within the prediction interval. The requirement of independence holds both with regard to the different pairs of samples and the observations within each sample.

The probability that a particular prediction interval will contain the future value that it is supposed to contain is unknown because the probability depends on the unknown parameters. As with confidence intervals, the $100(1-\alpha)\%$ refers to the *procedure* used to construct the prediction interval and not any particular interval that is computed.

2.4 STATISTICAL TOLERANCE INTERVALS

2.4.1 Tolerance Interval to Contain a Proportion of a Distribution

Prediction intervals are in general useful to predict the performance of one, or a small number, of future units. Consider now the case where one wishes to draw conclusions about the performance of a relatively large number of future units, based upon the data from a random sample from the distribution of interest. Conceptually, one can also construct prediction intervals for such situations (e.g., a prediction interval to contain all 100, 1,000, or any number m, future units). Such intervals would, however, often be very wide. Also, the exact number of future units of interest is sometimes not known or may be conceptually infinite. Moreover, rather than requiring that the calculated interval contain *all* of a specified number of units, it is generally sufficient to construct an interval to contain a *large proportion* of such units.

As indicated in Section 2.3.3, there are procedures for calculating prediction intervals to contain at least k out of m future observations, where $k \leq m$. More frequently, however, applications call for the construction of intervals to contain a specified proportion, β, of the entire sampled distribution. This leads to the concept of a tolerance interval.

Specifically, a tolerance interval is an interval that one can claim to contain at least a specified proportion, β, of the distribution with a specified degree of confidence, $100(1-\alpha)\%$. Such an interval would be of particular interest in setting limits on the process capability for a product manufactured in large quantities. This is in contrast to a prediction interval which, as noted, is of greatest interest in making predictions about a small number of future units.

Suppose, for example, that measurements of the diameter of a machined part have been obtained on a random sample of 25 units from a production process. A tolerance interval calculated from such data provides limits that one can claim, with a specified degree of confidence (e.g., 95%), contains the (measured) diameters of at least a specified proportion (e.g., 0.99) of units from the sampled process.

The two numbers in the preceding statement should not create any confusion when one recognizes that the 0.99 refers to the proportion of the distribution to be contained, and the 95% deals with the degree of confidence associated with the claim.

2.4.2 One-Sided Tolerance Bounds

Practical applications often require the construction of one-sided tolerance bounds. For example, in response to a request by a regulatory agency, a manufacturer has to make a statement concerning the maximum noise that, under specified operating conditions, is met (i.e., is not exceeded) by a high proportion of units, such as 0.99 of a particular model of a jet engine. The statement is to be based upon measurements from a random sample of 10 engines and is to

The standard model also assumes that the observed y values are statistically independent with a variance σ^2 that does not depend on the x_j values. In some applications, these x_j values might be known functions of some explanatory variables (e.g., 1/temperature, log(voltage), $(\text{time})^2$).

Users of regression analysis should watch for departures from the model assumptions. Methods for doing this can involve graphical analysis of the residuals (i.e., the differences between the observed and the predicted response values) from the fitted model, as described in textbooks on applied regression analysis. Some ways of handling possible departures from the assumed model are:

1. If the mean of y cannot be expressed as a linear function of the parameters, special (usually iterative) nonlinear least squares methods for estimating the parameters may be required; see, for example, Bates and Watts (1988), Seber and Wild (1989), or Ritz and Streibig (2008).
2. If σ^2 is not the same for all observations, a transformation—for example, a Box and Cox (1964) transformation, given in Section 4.12.4—of the response variable might be appropriate. Sometimes the method of weighted least squares is used. Both approaches are described in Carroll and Ruppert (1988). In other cases, it might be desirable to model both the mean and the standard deviation as separate functions of the explanatory variable (see Nelson, 1984, for an example).
3. If the observed response values (y) are not statistically independent, either generalized least squares or time series analysis methods—see, for example, Wei (2005), Cryer and Chan (2008), Bisgaard and Kulahci (2011), or Box et al. (2015)—might be appropriate.
4. If some of the values of the response variable are censored (i.e., the actual response is unknown other than being less than a known left-censoring value, greater than a known right-censoring value, or to lie between known lower and upper censoring values), or if they do not follow a normal distribution, the method of maximum likelihood, rather than the method of least squares, should be used for estimating the parameters. See Nelson (1990), Meeker and Escobar (1998), or Lawless (2003) for details.
5. If the observed values of the explanatory variables contain significant measurement error, the methods given by Fuller (1987), Carroll et al. (2006), or Buonaccorsi (2010) might be appropriate.

The rest of this section provides references to methods for computing statistical intervals for the standard linear regression model, when all of the assumptions hold. For other situations, some of the references given above provide similar methods. Some of the procedures require factors that are too numerous to tabulate and thus require specialized computer software (e.g., Eberhardt et al., 1989). In Chapter 12 we will describe some other more general methods that can be applied to regression analysis, but these also generally require special computer software.

4.13.1 Confidence Intervals for Linear Regression Analysis

Many textbooks on statistical methods and specialized textbooks on regression analysis give details on the construction of confidence intervals for

- The parameters $(\beta_0, \beta_1, \ldots, \beta_p)$ of the regression model.
- The expected value or mean $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ of the response variable for a specified set of conditions for the explanatory variable(s).

- Quantiles of the distribution of the response variable for a specified set of conditions for the explanatory variable(s).
- The variance σ^2 (or standard deviation σ) of the observations (which may or may not depend on a given set of x_j values).

4.13.2 Tolerance Intervals for Linear Regression Analysis

Tolerance intervals for the response variable for one or more conditions of the explanatory variable are *not* provided in most of the standard textbook chapters and textbooks on regression analysis (one exception is Graybill, 1976). This may be because special factors are required to compute these intervals.

4.13.3 Prediction Intervals for Regression Analysis

Many textbooks on statistical methods and specialized textbooks on regression analysis show how to compute a prediction interval for a single future observation on a response variable for a specified set of conditions for the explanatory variable(s). In fact, in introductory textbooks, regression analysis is the only situation in which prediction intervals are generally discussed. The methods for a single future observation easily extend to a prediction interval for the mean of m future response variable observations. It is also possible to construct prediction intervals to contain at least k of m future observations, but special methods would be required.

4.14 STATISTICAL INTERVALS FOR COMPARING POPULATIONS AND PROCESSES

Designed experiments are often used to compare two or more competing products, designs, treatments, packaging methods, etc. Statistical intervals are useful for presenting the results of such experiments. Most textbooks on elementary statistical methods show how to compute confidence intervals for the difference between the means of two randomly sampled populations or processes, assuming normal distributions and using various assumptions concerning the variances of the two distributions, based upon either paired or unpaired observations. References dealing with the use of statistical intervals to compare more than two normal distributions are also provided. The statistical methods for (fixed effects) analysis of variance and analysis of covariance models frequently used in the analysis of such data are special cases of the more general regression analysis methods that were outlined in the previous section. See the Bibliographic Notes section at the end of this chapter for references to some books that cover these and related topics.

BIBLIOGRAPHIC NOTES

General theory

Odeh and Owen (1980) describe the distribution theory methods behind the computation of many of the intervals in this chapter.

Tolerance intervals

Guttman (1970) describes the general theory behind tolerance intervals and tolerance regions (i.e., multivariate tolerance intervals), including Bayesian tolerance intervals (covered in our