STATISTICAL INFERENCE [1]

Richard A. Johnson
Professor Emeritus
Department of Statistics
University of Wisconsin

Key words : Bayesian approach, classical approach, confidence interval, estimation, randomization, test of hypotheses.

At the heart of statistics lie the ideas of statistical inference. Methods of statistical inference enable the investigator to argue from the particular observations in a sample to the general case. In contrast to logical deductions made from the general case to the specific case, a statistical inference can sometimes be incorrect. Nevertheless, one of the great intellectual advances of the twentieth century is the realization that strong scientific evidence can be developed on the basis of many, highly variable, observations.

The subject of statistical inference extends well beyond statistics' historical purposes of describing and displaying data. It deals with collecting informative data, interpreting these data, and drawing conclusions. Statistical inference includes all processes of acquiring knowledge that involve fact finding through the collection and examination of data. These processes are as diverse as opinion polls, agricultural field trials, clinical trials of new medicines, and the studying of properties of exotic new materials. As a consequence, statistical inference has permeated all fields of human endeavor in which the evaluation of information must be grounded in data-based evidence.

A few characteristics are common to all studies involving fact finding through the collection and interpretation of data. First, in order to acquire new knowledge, relevant data must be collected. Second, some variability is unavoidable even when observations are made under the same or very similar conditions. The third, which sets the stage for statistical inference, is that access to a complete set of data is either not feasible from a practical standpoint or is physically impossible to obtain.

To more fully describe statistical inference, it is necessary to introduce several key terminologies and concepts. The first step in making a statistical inference is to model the population(s) by a *probability distribution* which has a numerical feature of interest called a *parameter*. The problem of statistical inference arises once we want to make generalizations about the *population* when only a *sample* is available.

A *statistic*, based on a sample, must serve as the source of information about a parameter. Three salient points guide the development of procedures for statistical inference

1. Because a sample is only part of the population, the numerical value of the statistic will not be the exact value of the parameter.

[1] Based on an article from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science +Business Media, LLC

2. The observed value of the statistic depends on the particular sample selected.

3. Some variability in the values of a statistic, over different samples, is unavoidable.

The two main classes of inference problems are *estimation* of parameter(s) and *testing hypotheses* about the value of the parameter(s). The first class consists of point estimators, a single number estimate of the value of the parameter, and interval estimates. Typically, the interval estimate specifies an interval of plausible values for the parameter but the subclass also includes prediction intervals for future observations. A test of hypotheses provides a yes/no answer as to whether the parameter lies in a specified region of values.

Because statistical inferences are based on a sample, they will sometimes be in error. Because the actual value of the parameter is unknown, a test of hypotheses may yield the wrong yes/no answer and the interval of plausible values may not contain the true value of the parameter.

Statistical inferences, or generalizations from the sample to the population, are founded on an understanding of the manner in which variation in the population is transmitted, via sampling, to variation in a statistic. Most introductory texts (see Johnson and Bhattacharyya [11], Johnson, Miller, and Freund [12]) give expanded discussions of these topics.

There are two primary approaches, *frequentist* and *Bayesian*, for making statistical inferences. Both are based on the *likelihood* but their frameworks are entirely different.

The frequentist treats parameters as fixed but unknown quantities in the distribution which governs variation in the sample. Then, the frequentist tries to protect against errors in inference by controlling the probabilities of these errors. The long-run relative frequency interpretation of probability then guarantees that if the experiment is repeated many times only a small proportion of times will produce incorrect inferences. Most importantly, using this approach in many different problems keeps the overall proportion of errors small.

To illustrate a frequentist approach to confidence intervals and tests of hypotheses, we consider the case were the observations are a random sample of size n from a normal distribution having mean μ and standard deviation σ. Let $X_1, ..., X_n$ be independent observations from that distribution, $\overline{X} = \sum_{i-1}^{n} X_i/n$, and $S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 / (n-1)$. Then, using the fact that the sampling distribution of $\sqrt{n}\,(\overline{X} - \mu)/S = T$ is the $t-$distribution with $n-1$ degrees of freedom

$$1 - \alpha = P[\, -t_{n-1}(\alpha/2) < \frac{\sqrt{n}\,(\overline{X} - \mu)}{S} < t_{n-1}(\alpha/2)\,]$$

where $t_{n-1}(\alpha/2)$ is the upper $100\alpha/2$ percentile of that $t-$distribution.

Rearranging the terms, we obtain the probability statement

$$1 - \alpha = P[\, \overline{X} - t_{n-1}(\alpha/2)\,\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{n-1}(\alpha/2)\,\frac{S}{\sqrt{n}}\,]$$

which states that, prior to collecting the sample, the random interval with end points $\overline{X} \pm t_{n-1}(\alpha/2)S/\sqrt{n}$ will cover the unknown, but fixed, μ with the specified probability $1-\alpha$. After the sample is collected, $\overline{x}$, s and the endpoints of the interval are calculated. The interval is now fixed and μ is fixed but unknown. Instead of probability we say that the resulting interval is a $100(1-\alpha)$ percent confidence interval for μ.

To test the null hypothesis that the mean has a specified value μ_0, we consider the test statistic $\sqrt{n}\,(\overline{X} - \mu_0)/S$ which has the $t-$ distribution with $n-1$ when the null hypothesis prevails. When the alternative hypothesis asserts that μ is different from μ_0, the null hypothesis should be rejected when $|\sqrt{n}\,(\overline{X} - \mu_0)/S| \geq t_{n-1}(\alpha/2)$. Before the sample is collected, with specified probability α, the test will falsely fail to reject the null hypothesis.

Frequentists are divided on the problem of testing hypotheses. Some statisticians (see Cox [4]) follow R. A. Fisher and perform *significance tests* where the decision to reject a *null hypothesis* is based on values of the statistic that are extreme in directions considered important by subject matter interest. R. A. Fisher [7] also suggests using *fudicial probabilities* to interpret significance tests but this is no longer a popular approach.

It is more common to take a *Neyman-Pearson* approach where an *alternative hypothesis* is clearly specified together with the corresponding distributions for the statistic. *Power*, the probability of rejecting the null hypothesis when it is false, can then be optimized. A definitive account of the Neyman-Pearson theory of testing hypotheses is given by Lehmann and Ramono [14] and that for the theory of estimation by Lehmann and Casella [13].

In contrast, Bayesians consider unknown parameters to be random variables and, prior to sampling, assign a *prior distribution* for the parameters. After the data are obtained, the Bayesian multiplies the likelihood by the prior distribution to obtain the *posterior distribution* of the parameter, after a suitable normalization. Depending on the goal of the investigation, a pertinent feature or features of the posterior distribution are used to make inferences. The mean is often a suitable point estimator and a suitable region of highest posterior density gives an interval of plausible values.

More generally, under a Bayesian approach, a distribution is given for anything that is unknown or uncertain. Once the data become known, the prior distribution is updated using the appropriate laws of conditional probability. See Box and Tiao[1] and Gelman, Carlin and Rubin [8] for discussions of Bayesian approaches to statistical inference.

A second phase of statistical inference, *model checking*, is required for both frequentist and Bayesian approaches. Are the data consonant with the model or must the model be modified in some way? Checks on the model are often subjective and rely on graphical diagnostics.

D. R. Cox [4] gives an excellent introduction to statistical inference where he also compares Bayesian and frequentist approaches and highlights many of the important issues underlying their differences.

The advent of designed experiments has greatly enhanced the opportunities for making statistical inferences about differences between methods, drugs,

or procedures. R. A. Fisher pioneered the development of both the design of experiments and also their analysis which he called the Analysis of Variance (ANOVA). Box, Hunter, and Hunter [2] and Seber and Lee [17], together with the material in the references therein, provide comprehensive coverage.

When one or more variables may influence the expected value of the response, and these variables can be controlled by the experimenter, the selection of values used in the experiment can often be chosen in clever ways. We use the term *factor* for a variable and *levels* for its values. In addition to the individual factors, the response may depend on terms such as the product of two factors or other combination of the factors. The expected value of the response is expressed as a function of these terms and parameters. In the classical linear models setting, the function is linear in the parameters and the error is additive. These errors are assumed to be independent and normally distributed with mean zero and the same variance for all runs. This setting, which encompasses all linear regression analysis, gives rise to the normal theory sampling distributions; the chi square, F, normal and t distributions.

The two simplest designs are the matched pairs design and the two samples design. Suppose n experimental units are available. When the two treatments can be assigned by the experimenter, the experimenter should randomly select n_1 of them to receive treatment 1 and then treatment 2 is applied to the other $n - n_1 = n_2$ units. After making a list, or physically arranging the units in order, n_1 different random numbers between 1 and n, inclusive, can be generated. The corresponding experimental units receive treatment 1.

In the matched pairs design, the experimental units are paired according to some observable characteristic that is expected to influence the response. In each pair, treatment 1 is applied to one unit and treatment 2 to the other unit. The assignment of treatments within a pair should be done randomly. A coin could be flipped, for each pair, with heads corresponding to the assignment of treatment 1 to the first unit in that pair.

Both the two samples design and the matched pairs design are examples of randomized designs. R. A. Fisher [7] introduced the idea of *randomized tests* in his famous example of the tea tasting lady who claimed she could tell if milk or the tea infusion were added first to her cup. A small example, employing the two samples design, illustrates the concepts. In order to compare two cake recipes, or treatments, cakes are made using the two different recipes. Three of the seven participants are randomly assigned to receive treatment 1, and the other four receive treatment 2. Suppose the responses, ratings of the taste, are

Treatment 1 :	11	13	9			$\overline{x} =$ 11
Treatment 2 :	8	7	12	5		$\overline{y} =$ 8

Randomization tests compare the two treatments by calculating a test statistic. Here we use the difference of means $11 - 8 = 3$. Equivalently we could use the mean of the first sample or a slightly modified version of the t statistic.

The observed value of the test statistic is compared, not with a tabled distribution, but with the values of test statistic evaluated over all permutations

of the data. As a consequence, randomization tests are also called *permutation tests*.

Here the first person in the treatment 1 group can be selected in 7 ways, the second in 6 ways, and the third in 5 ways. The succession of choices can be done in $7 \times 6 \times 5 = 210$ ways but there are $3 \times 2 \times 1 = 6$ different orders that lead to the same set of three people. Dividing the number of permutations 210 by 6, we obtain 35 different assignments of participants to treatment 1. If there is no difference in treatments, these 35 re-assignments should all be comparable.

One such case results in

Treatment 1 :	11	13	12		$\overline{x} =$	12
Treatment 2 :	8	7	9	5	$\overline{y} =$	7.25

and the corresponding difference in means is $12 - 7.25 = 4.75$. After calculating all 35 values we find that this last case and the observed one give the largest differences. The one-sided randomization test, for showing that the first treatment has higher average response, would then have P-value $2/35 = 0.057$.

In applications where there are far to many cases to evaluate to obtain the complete randomization distribution, it is often necessary to take a Monte Carlo approach. By randomly selecting, say, 10,000 of the possible permutations and evaluating the statistic for each case, we usually obtain a very good approximation to the randomization distribution.

We emphasize that randomization tests (i) do not require random samples and (ii) make no assumptions about normality. Instead, they rely on the randomization of treatments to deduce whether or not there is a difference in treatments. Also, randomized designs allow for some inferences to be based on firmer grounds than observational studies where the two groups are already formed and the assignment of treatments is not possible.

Edgington and Onghena [5] treat many additional randomization tests. *Rank tests* are special cases of permutation tests where the responses are replaced by their ranks. (see Hajek and Sidak [9]).

The same ideas leading to randomized designs also provide the underpinning for the traditional approach to inference in sample surveys. It is the random selection of individuals, or random selection within subgroups or strata, that permits inferences to made about the population. There is some difference here because the population consists of a finite number of units. When all subsets of size n have the same probability of being selected, the sample is called a random sample and sampling distributions of means and proportions can be determined from the selection procedure. The random selection is still paramount when sampling from strata or using multiple stage sampling. Lohr [15] gives a very readable introduction and the classical text by Cochran [3] presents the statistical theory of sample surveys.

Bootstrap sampling (see Efron and Tibshirani [6]) provides another alternative for obtaining a reference distribution with which to compare the observed value of statistic or even a distribution on which to base interval estimates. Yet another approach is using the *empirical likelihood* as discussed in Owen [16] .

Much of the modern research on methods of inference concerns infinite dimensional parameters. Examples include function estimation where the function may be a nonparametric regression function, cumulative distribution function, or hazard rate. Additionally, considerable research activity has been motivated by genomic applications where the number of variables far exceeds the sample size.

Major advances are also being made in developing computer intensive methods for *statistical learning*. These include techniques with applications to the cross-disciplinary areas of *data mining* and *artificial intelligence*. See Hastie, Tibshirani and Friedman [10] for a good summary of statistical learning techniques.

REFERENCES

1 Box, G. E. P and G.C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading.

2 Box, G. E. P., J. S. Hunter, and W. G. Hunter (2005), *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd Edition, John Wiley, New York

3 Cochran, W. G. (1977), *Sampling Techniques*, 3rd Edition, John Wiley, New York.

4 Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge University Press.

5 Edgington, E. and P. Onghena(2007), *Randomization tests*, 4th edition, Chapman and Hall, Boca Rotan.

6 Efron, B. and R. Tibshirani (2007), *An Introduction to the Bootstrap*, Chapman and Hall/CRC, Boca Rotan.

7 Fisher, R. A. (1935), *Design of Experiments*, Hafner, New York.

8 Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004), *Bayesian Data Analysis* 2nd ed., Chapman and Hall/CRC, Boca Rotan.

9 Hajek, J. and Sidak, Z. (1967), *Theory of Rank Tests*, Academic press, New York.

10 Hastie, T., R. Tibshirani and J. Friedman (2009), *The Elements of Statistical Learning* 2nd ed., Springer, New York.

11 Johnson, Richard and G. K. Bhattacharyya (2010), *Statistics–Principles and Methods* 6th ed., John Wiley, New York.

12 Johnson, Richard (2010), *Miller and Freund's Probability and Statistics for Engineers* 8th ed., Prentice Hall, Boston.

13 Lehmann E. L. and G. C. Casella (2003), *Theory of Point Estimation* 2nd ed., Springer, New York.

14 Lehmann E. L. and J. P. Romano (2005), *Testing Statistical Hypotheses* 3rd ed., Springer, New York.

15 Lohr, S. (2010), *Sampling: Design and Analysis*, 2nd edition, Brooks/Cole, Boston.

16 Owen, A. (2001), *Empirical Likelihood*, Chapman and Hall/CRC, Boca Rotan.

17 Seber, G. and A. Lee (2003), *Linear Regression Analysis*, 2nd edition, John Wiley, New York.