Nonparametric Statistics

Vytaras Brazauskas

Volume 2, pp. 1182–1190

In

Encyclopedia Of Actuarial Science
(ISBN 0-470-84676-3)

Edited by

Jozef L. Teugels and Bjørn Sundt

Nonparametric Statistics

Introduction

Nonparametric statistical methods enjoy many advantages over their parametric counterparts. Some of the advantages are as follows:

- Nonparametric methods require fewer assumptions about the underlying populations from which the data are obtained.
- Nonparametric techniques are relatively insensitive to outlying observations.
- Nonparametric procedures are intuitively appealing and quite easy to understand.

The earliest works in nonparametric statistics date back to 1710, when Arbuthnott [2] developed an antecedent of the *sign test*, and to 1904, when Spearman [43] proposed the *rank correlation* procedure. However, it is generally agreed that the systematic development of the field of *nonparametric statistical inference* started in the late 1930s and 1940s with the seminal articles of Friedman [16], Kendall [24], Mann and Whitney [28], Smirnov [42], and Wilcoxon [46]. Most of the early developments were intuitive by nature and were based on *ranks* of observations (rather than their values) to deemphasize the effect of possible outliers on the conclusions.

Later, an important contribution was made by Quenouille [36]. He invented a clever bias-reduction technique – the *jackknife* – which enabled nonparametric procedures to be used in a variety of situations. Hodges and Lehmann [22] used rank tests to derive point estimators and proved that these estimators have desirable properties. Their approach led to the introduction of nonparametric methods into more general settings such as **regression**.

Among the most important modern advances in the field of nonparametric statistics is that of Efron [11]. He introduced a computer-intensive technique – the *bootstrap* – which enables nonparametric procedures to be used in many complicated situations, including those in which parametric–theoretical approaches are simply intractable.

Finally, owing to the speed of modern computers, the so-called *smoothing techniques*, prime examples of which are *nonparametric density estimation* and *nonparametric regression*, are gaining popularity in practice, including actuarial science applications.

In the section 'Some Standard Problems', some standard nonparametric problems for one, two, or more samples are described. In the section 'Special Topics with Applications in Actuarial Science', more specialized topics, such as *empirical estimation*, *resampling methods*, and *smoothing techniques*, are discussed and, for each topic, a list of articles with actuarial applications of the technique is provided. In the section 'Final Remarks', we conclude with a list of books for further reading and a short note on software packages that have some built-in procedures for nonparametric inference.

Some Standard Problems

Here we present the most common nonparametric inference problems. We formulate them as hypothesis testing problems, and therefore, in describing these problems, we generally adhere to the following sequence: data, assumptions, questions of interest (i.e. specification of the null hypothesis, H_0, and possible alternatives, H_A), test statistic and its distribution, and decision-making rule. Whenever relevant, point and interval estimation is also discussed.

One-sample Location Problem

Let $Z_1, \ldots, Z_n$ be a random sample from a continuous population with cumulative distribution function (cdf) F and median μ. We are interested in the inference about μ. That is, $H_0 : \mu = \mu_0$ versus $H_\mathrm{A} : \mu > \mu_0$, or $H_\mathrm{A} : \mu < \mu_0$, or $H_\mathrm{A} : \mu \neq \mu_0$, where a real number μ_0 and H_A are prespecified by the particular problem.

Also, let us denote the ordered sample values by $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)}$.

Sign Test. In this setting, if besides the *continuity* of F, no additional assumptions are made, then the appropriate test statistic is the *sign test* statistic:

$$B = \sum_{i=1}^{n} \Psi_i, \qquad (1)$$

where $\Psi_i = 1$, if $Z_i > \mu_0$, and $\Psi_i = 0$, if $Z_i < \mu_0$.

Statistic B counts the number of sample observations Z that exceed μ_0. In general, statistics of such a type are referred to as *counting* statistics. Further, when H_0 is true, it is well known (see [37], Section 2.2) that B has a binomial distribution with parameters n and $p = \mathbf{P}\{Z_i > \mu_0\} = 1/2$ (because μ_0 is the median of F). This implies that B is a *distribution-free* statistic, that is, its null distribution does not depend on the underlying distribution F. Finally, for the observed value of B (say, B_{obs}), the rejection region (RR) of the associated level α test is

H_A	RR
$\mu > \mu_0$	$B_{\text{obs}} \geq b_\alpha$
$\mu < \mu_0$	$B_{\text{obs}} \leq n - b_\alpha$
$\mu \neq \mu_0$	$B_{\text{obs}} \leq n - b_{\alpha/2}$ or $B_{\text{obs}} \geq b_{\alpha/2}$

where b_α denotes the upper αth percentile of the binomial (n, $p = 1/2$) distribution.

A typical approach for constructing one- or two-sided *confidence intervals* for μ is to invert the appropriate hypothesis test. Inversion of the level α two-sided sign test leads to the following $(1-\alpha)$ 100% confidence interval for μ: $(Z_{(n+1-b_{\alpha/2})}, Z_{(b_{\alpha/2})})$. The one-sided $(1-\alpha)$ 100% confidence intervals for μ, $(-\infty, Z_{(b_\alpha)})$ and $(Z_{(n+1-b_\alpha)}, \infty)$, are obtained similarly. Finally, the associated point estimator for μ is $\tilde{\mu} = \text{median}\{Z_1, \ldots, Z_n\}$.

Wilcoxon Signed-rank Test. Consider the same problem as before with an additional assumption that F is *symmetric* about μ. In this setting, the standard approach for inference about μ is based on the *Wilcoxon signed-rank test* statistic (see [46]):

$$T^+ = \sum_{i=1}^{n} R_i \Psi_i, \quad (2)$$

where R_i denotes the rank of $|Z_i - \mu_0|$ among $|Z_1 - \mu_0|, \ldots, |Z_n - \mu_0|$ and Ψ_i is defined as before. Statistic T^+ represents the sum of the $|Z - \mu_0|$ ranks for those sample observations Z that exceed μ_0.

When H_0 is true, two facts about Ψ_i and $|Z_i - \mu_0|$ hold: (i) the $|Z_i - \mu_0|$'s and the Ψ_i's are independent, and (ii) the ranks of $|Z_i - \mu_0|$'s are uniformly distributed over the set of $n!$ permutations of integers $(1, \ldots, n)$ (see [37], Section 2.3). These imply that T^+ is a distribution-free statistic. Tables for the critical values of this distribution, t_α, are available in [23], for example. For the observed value T^+_{obs}, the rejection region of the associated level α test is

H_A	RR
$\mu > \mu_0$	$T^+_{\text{obs}} \geq t_\alpha$
$\mu < \mu_0$	$T^+_{\text{obs}} \leq \dfrac{n(n+1)}{2} - t_\alpha$
$\mu \neq \mu_0$	$T^+_{\text{obs}} \leq \dfrac{n(n+1)}{2} - t_{\alpha/2}$ or $T^+_{\text{obs}} \geq t_{\alpha/2}$

(The factor $n(n+1)/2$ appears because T^+ is symmetric about its mean $n(n+1)/4$.)

As shown previously, the one- and two-sided confidence intervals for μ are derived by inverting the appropriate hypothesis tests. These intervals are based on the order statistics of the $N = n(n+1)/2$ Walsh averages of the form $W_{ij} = (Z_i + Z_j)/2$, for $1 \leq i \leq j \leq n$. That is, the $(1-\alpha)$ 100% confidence intervals for μ are $(W_{(N+1-t_{\alpha/2})}, W_{(t_{\alpha/2})})$, $(-\infty, W_{(t_\alpha)})$, and $(W_{(N+1-t_\alpha)}, \infty)$, where $W_{(1)} \leq \cdots \leq W_{(N)}$ denote the ordered Walsh averages. Finally, the associated point estimator for μ is $\hat{\mu} = \text{median}\{W_{ij}, 1 \leq i \leq j \leq n\}$. (This is also known as the Hodges–Lehmann [22] estimator.)

Remark 1 *Discreteness of the true distribution functions*. Although theoretically the probability that $Z_i = \mu_0$ or that there are ties among the $|Z_i - \mu_0|$s is zero (because F is continuous), in practice, this event may occur due to discreteness of the true distribution function. In the former case, it is common to discard such Z_is, thus reducing the sample size n, and, in the latter case, the ties are broken by assigning average ranks to each of the $|Z_i - \mu_0|$s within a tied group. Also, due to discreteness of the distribution functions of statistics B and T^+, the respective levels of the associated tests (α) and confidence intervals $(1-\alpha)$ are not attained exactly.

Two-sample Location Problem

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent random samples from populations with continuous cdfs F and G respectively. We assume that these populations differ only by a shift of location, that is, we assume that $G(x) = F(x - \Delta)$, and we want to know if there is indeed the nonzero shift Δ. That is, $H_0: \Delta = 0$ versus $H_A: \Delta > 0$, or $H_A: \Delta < 0$, or $H_A: \Delta \neq 0$. (Another possible (though less common) approach is to test $H_0: \Delta = \Delta_0$. In such a case, all procedures described here remain valid if applied to the transformed observations $Y_i^* = Y_i - \Delta_0$.)

In this setting, the most commonly used procedures are based on the *rank sum* statistic of the *Wilcoxon–Mann–Whitney* (see [28, 46]):

$$W = \sum_{i=1}^{n_2} R_i, \tag{3}$$

where R_i is the rank of Y_i among the combined sample of $N = n_1 + n_2$ of observations $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$. (See Remark 1 for how to handle the ties among observations.)

When H_0 is true, it follows from fact (ii) of Section 'Wilcoxon Signed-rank Test' that W is a distribution-free statistic. Tables for the critical values of this distribution, w_α, are available in [23], for example. For the observed value W_{obs}, the rejection region of the associated level α test is

H_A	RR
$\Delta > 0$	$W_{\text{obs}} \geq w_\alpha$
$\Delta < 0$	$W_{\text{obs}} \leq n_2(N+1) - w_\alpha$
$\Delta \neq 0$	$W_{\text{obs}} \leq n_2(N+1) - w_{\alpha/2}$ or $W_{\text{obs}} \geq w_{\alpha/2}$

(The factor $n_2(N+1)$ is due to the symmetricity of W about its mean $n_2(N+1)/2$.)

The associated confidence intervals are based on the order statistics of the $n_1 n_2$ differences $U_{ij} = Y_j - X_i$, for $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$. That is, the $(1-\alpha)$ 100% confidence intervals for Δ are $(U_{((n_2(n_1+2n_2+1)/2+1-w_{\alpha/2})}, U_{(w_{\alpha/2}-n_2(n_2+1)/2)})$, $(-\infty,\ U_{(w_\alpha - n_2(n_2+1)/2)})$, and $(U_{(n_2(n_1+2n_2+1)/2+1-w_\alpha)}$, $\infty)$, where $U_{(1)} \leq \cdots \leq U_{(n_1 n_2)}$ denote the ordered differences U_{ij}. Finally, the Hodges–Lehmann [22] point estimator for Δ is $\hat{\Delta} = \text{median}\{U_{ij},\ i = 1, \ldots, n_1,\ j = 1, \ldots, n_2\}$.

Remark 2 *An application*. In the insurance context, Ludwig and McAuley [26] applied the Wilcoxon–Mann–Whitney test to evaluate reinsurers' relative financial strength. In particular, they used the statistic W to determine which financial ratios discriminated most successfully between 'strong' and 'weak' companies.

Related Problems and Extensions

There are two types of problems that are closely related to or directly generalize the location model. One arises from the consideration of other kinds of differences between the two distributions–differences in scale, location and scale, or *any* kind of differences. Another venue is to compare locations of $k > 2$ distributions simultaneously. Inference procedures for these types of problems are motivated by similar ideas as those for the location problem. However, in some cases, technical aspects become more complicated. Therefore, in this section, the technical level will be kept at a minimum; instead, references will be provided.

Problems of Scale, Location–Scale, and General Alternatives. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent random samples from populations with continuous cdfs F_1 and F_2 respectively. The null hypothesis for comparing the populations is $H_0 : F_1(x) = F_2(x)$ (for every x). Here, several problems can be formulated.

Scale Problem. If we assume that, for every x, $F_1(x) = G(x - \mu/\sigma_1)$ and $F_2(x) = G(x - \mu/\sigma_2)$, where G is a continuous distribution with a possibly *unknown* median μ, then the null hypothesis can be replaced by $H_0 : \sigma_1 = \sigma_2$, and possible alternatives are $H_A : \sigma_1 > \sigma_2$, $H_A : \sigma_1 < \sigma_2$, $H_A : \sigma_1 \neq \sigma_2$. In this setting, the most commonly used procedures are based on the *Ansari–Bradley* statistic C. This statistic is computed as follows. First, order the combined sample of $N = n_1 + n_2$ X- and Y-values from least to greatest. Second, assign the score '1' to both the smallest and the largest observations in this combined sample, assign the score '2' to the second smallest and second largest, and continue in this manner. Thus, the arrays of assigned scores are $1, 2, \ldots, N/2, N/2, \ldots, 2, 1$ (for N-even) and $1, 2, \ldots, (N+1)/2, \ldots, 2, 1$ (for N-odd). Finally, the *Ansari–Bradley* statistic is the sum of the scores, $S_1, \ldots, S_{n_2}$, assigned to $Y_1, \ldots, Y_{n_2}$ via this scheme:

$$C = \sum_{i=1}^{n_2} S_i \tag{4}$$

(see [23], Section 5.1, for the distribution of C, rejection regions, and related questions).

Location–Scale Problem. If we do not assume that medians are equal, that is, $F_1(x) = G(x - \mu_1/\sigma_1)$ and $F_2(x) = G(x - \mu_2/\sigma_2)$ (for every x), then the hypothesis testing problem becomes $H_0 : \mu_1 = \mu_2$, $\sigma_1 = \sigma_2$ versus $H_A : \mu_1 \neq \mu_2$ and/or $\sigma_1 \neq \sigma_2$. In this setting, the most commonly used procedures are based on the *Lepage* statistic L, which combines standardized versions of the Wilcoxon–Mann–Whitney W and Ansari–Bradley C:

$$L = \frac{(W - \mathrm{E}_0(W))^2}{\mathrm{Var}_0(W)} + \frac{(C - \mathrm{E}_0(C))^2}{\mathrm{Var}_0(C)}, \qquad (5)$$

where the expected values (E_0) and variances (Var_0) of the statistics are computed *under* H_0 (for further details and critical values for L, see [23, Section 5.3]). Note that an additional assumption of $\mu_1 = \mu_2$ (or $\sigma_1 = \sigma_2$) yields the scale (or the location) problem setup.

General Alternatives. The most general problem in this context is to consider *all* kinds of differences between the distributions F_1 and F_2. That is, we formulate the alternative $H_\mathrm{A} : F_1(x) \neq F_2(x)$ (for at least one x). This leads to the well-known *Kolmogorov–Smirnov* statistic:

$$D = \frac{n_1 n_2}{d} \max_{1 \le k \le N} |\hat{F}_{1,n_1}(Z_{(k)}) - \hat{F}_{2,n_2}(Z_{(k)})|. \qquad (6)$$

Here d is the 'greatest common divisor' of n_1 and n_2, $Z_{(1)} \le \cdots \le Z_{(N)}$ denotes the $N = n_1 + n_2$ ordered values for the combined sample of X's and Y's, and $\hat{F}_{1,n_1}$ and $\hat{F}_{2,n_2}$ are the **empirical distribution functions** for the X and Y samples:

$$\hat{F}_{1,n_1}(z) = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{1}\{X_i \le z\}$$

$$\text{and } \hat{F}_{2,n_2}(z) = \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{1}\{Y_j \le z\}, \qquad (7)$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. For further details, see [23], Section 5.4.

One-way Analysis of Variance. Here, the data are $k > 2$ mutually independent random samples from continuous populations with medians $\mu_1, \ldots, \mu_k$ and cdfs $F_1(x) = F(x - \tau_1), \ldots, F_k(x) = F(x - \tau_k)$, where F is the cdf of a continuous distribution with median μ, and $\tau_1 = \mu_1 - \mu, \ldots, \tau_k = \mu_k - \mu$ represent the location effects for populations $1, \ldots, k$. We are interested in the inference about medians $\mu_1, \ldots, \mu_k$, which is equivalent to the investigation of differences in the population effects $\tau_1, \ldots, \tau_k$. This setting is called the *one-way layout* or *one-way analysis of variance*. Formulation of the problem in terms of the effects instead of the medians has interpretive advantages and is easier to generalize (to two-way analysis of variance, for example).

For testing $H_0 : \tau_1 = \cdots = \tau_k$ versus general alternatives H_A: *at least one* $\tau_i \neq \tau_j$, the *Kruskal–Wallis* test is most commonly used. For testing H_0 versus ordered alternatives $H_\mathrm{A} : \tau_1 \le \cdots \le \tau_k$ (*with at least one strict inequality*), the *Jonckheere–Terpstra* test is appropriate, and versus umbrella alternatives $H_\mathrm{A} : \tau_1 \le \cdots \le \tau_{q-1} \le \tau_q \ge \tau_{q+1} \ge \cdots \ge \tau_k$ (*with at least one strict inequality*), the most popular techniques are those of Mack and Wolfe [27].

Also, if H_0 is rejected, then we are interested in finding which of the populations are different and then in estimating these differences. Such questions lead to the *multiple comparisons* and *contrast estimation* procedures. For further details and extensions of the one-way analysis of variance, the reader is referred to [23], Chapter 6.

Independence

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a continuous bivariate population. We are interested in the null hypothesis H_0 : *X and Y are independent*. Depending on how the strength of association between X and Y is measured, different approaches for testing H_0 are possible. Here we briefly review two most popular procedures based on *Spearman*'s ρ and *Kendall*'s τ. For further details and discussion, see [23], Chapter 8.

Spearman's ρ. The Spearman rank correlation coefficient ρ was introduced in 1904, and is probably the oldest nonparametric measure in current use. It is defined by

$$\rho = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \left(R_i^X - \frac{n+1}{2}\right)\left(R_i^Y - \frac{n+1}{2}\right)$$

$$= 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i^X - R_i^Y)^2, \qquad (8)$$

where R_i^X (R_i^Y) denotes the rank of X_i (Y_i) in the sample of Xs (Ys). Using the first expression, it is a straightforward exercise to show that ρ is simply the classical Pearson's correlation coefficient applied to the rank vectors R^X and R^Y instead of the actual X and Y observations respectively. The second expression is more convenient for computations.

For testing the independence hypothesis H_0 versus alternatives H_A : *X and Y are positively associated*, H_A : *X and Y are negatively associated*, H_A: *X and Y are not independent*, the corresponding rejection regions are $\rho \geq \rho_\alpha$, $\rho \leq -\rho_\alpha$, and $|\rho| \geq \rho_{\alpha/2}$. Critical values ρ_α are available in [23], Table A.31.

Kendall's τ. The Kendall population correlation coefficient τ was introduced in 1938 by Kendall [24] and is given by

$$\tau = 2\mathbf{P}\{(X_2 - X_1)(Y_2 - Y_1) > 0\} - 1. \qquad (9)$$

If X and Y are independent, then $\tau = 0$ (but the opposite statement is not necessarily true). Thus, the hypotheses of interest can be written as H_0 : $\tau = 0$ and $H_A : \tau > 0$, $H_A : \tau < 0$, $H_A : \tau \neq 0$. The appropriate test statistic is

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q((X_i, Y_i), (X_j, Y_j)), \qquad (10)$$

where $Q((a, b), (c, d)) = 1$, if $(d - b)(c - a) > 0$, and $= -1$, if $(d - b)(c - a) < 0$. The corresponding rejection regions are $K \geq k_\alpha$, $K \leq -k_\alpha$, and $|K| \geq k_{\alpha/2}$. Critical values k_α are available in [23], Table A.30. If ties are present, that is, $(d - b)(c - a) = 0$, the function Q is modified by allowing it to take a third value, $Q = 0$. Such a modification, however, makes the test only approximately of significance level α. Finally, the associated point estimator of τ is given by $\hat{\tau} = 2K/(n(n - 1))$.

Besides their simplicity, the main advantage of Spearman's ρ and Kendall's τ is that they provide estimates of dependence between two variables. On the other hand, some limitations on the usefulness of these measures exist. In particular, it is well known that zero correlation does not imply independence; thus, the H_0 statement is not 'if and only if'. In the actuarial literature, this problem is addressed in [6]. A solution provided there is based on the *empirical approach*, which we discuss next.

Special Topics with Applications in Actuarial Science

Here we review some special topics in nonparametric statistics that find frequent application in actuarial research and practice. For each topic, we provide a reasonably short list of actuarial articles where these techniques (or their variants) are implemented.

Empirical Approach

Many important features of an underlying distribution F, such as mean, variance, skewness, and kurtosis can be represented as functionals of F, denoted $T(F)$. Similarly, various actuarial parameters, such as *mean excess function*, *loss elimination ratio*, and various types of reinsurance **premiums** can be represented as functionals of the underlying *claims* distribution F. For example, the net premium of the **excess-of-loss reinsurance**, with *priority* β and expected number of claims λ, is given by $T(F) = \lambda \int \max\{0, x - \beta\}\, dF(x)$; the loss elimination ratio for a **deductible** of d is defined as $T(F) = \int \min\{x, d\}\, dF(x) / \int x\, dF(x)$.

In estimation problems using *empirical nonparametric approach*, a parameter of interest $T(F)$ is estimated by $T(\hat{F}_n)$, where $\hat{F}_n$ is the empirical distribution function based on a sample of n observations from a population with cdf F, as defined in the section 'Problems of Scale, Location–Scale, and General Alternatives'. (Sometimes this approach of replacing F by $\hat{F}_n$ in the expression of $T(\cdot)$ is called the *plug-in* principle.) Then, one uses the *delta method* to derive the asymptotic distribution of such estimators. Under certain regularity conditions, these estimators are *asymptotically normal* with mean $T(F)$ and variance $(1/n) \int [T'(F)]^2\, dF(x)$, where $T'(F)$ is a directional derivative (Hadamard-type) of $T(F)$. (In the context of **robust statistics** (*see* **Robustness**), $T'(F)$ is also known as the *influence function*.)

The empirical approach and variants are extensively used in solving different actuarial problems. Formal treatment of the delta-method for actuarial statistics is provided by Hipp [21] and Præstgaard [34]. Robustness properties of empirical nonparametric estimators of the excess-of-loss and **stop-loss reinsurance** (*see* **Stop-loss Reinsurance**) premiums, and the probability of ruin are investigated by Marceau and Rioux [29]. More extensive robustness study of reinsurance premiums, which additionally includes **quota-share** (*see* **Quota-share Reinsurance**), **largest claims and ECOMOR** (*see* **Largest Claims and ECOMOR Reinsurance**) treaties, is carried out by Brazauskas [3]. Carriere [4, 6] applies empirical approach to construct nonparametric tests for **mixed Poisson distributions** and a test for independence of claim frequencies and severities. In a similar fashion, Pitts [32] developed a nonparametric estimator for the **aggregate claims distribution** (*see* **Aggregate Loss Modeling**) and for the probability of ruin in the **Poisson risk model**. The latter problem is also treated by Croux and Veraverbeke [8] using empirical estimation and U-statistics (for a comprehensive account on U-statistics, see Serfling [39], Chapter 5). Finally, Nakamura and Pérez-Abreu [30] proposed an empirical probability **generating function** for estimation of claim frequency distributions.

Resampling Methods

Suppose we have an estimator $\hat{\theta} = T(\hat{F}_n)$ for the parameter of interest $\theta = T(F)$, where $\hat{F}_n$ is defined as in the preceding section. In all practical situations, we are interested in $\hat{\theta}$ as well as in the evaluation of its performance according to some measure of error. The bias and variance of $\hat{\theta}$ are a primary choice but more general measures of statistical accuracy can also be considered. The nonparametric estimation of these measures is based on the so-called **resampling methods** (*see* **Resampling**) also known as the *bootstrap*.

The concept of bootstrap was introduced by Efron [11] as an extension of the Quenouille–Tukey *jackknife*. The latter technique, invented by Quenouille [36] for nonparametric estimation of bias and later extended by Tukey [45] for estimation of variance, works as follows. It is based on the idea of sequentially deleting points X_i and then recomputing $\hat{\theta}$. This produces n new estimates $\hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(n)}$, where $\hat{\theta}_{(i)}$ is based on a sample of size $n-1$, that is, the corresponding empirical cdf $\hat{F}_n^{(i)}(x) = 1/(n-1)\sum_{j\neq i} \mathbf{1}\{X_j \leq x\}$, $i = 1, \ldots, n$. Then, the jackknife bias and variance for $\hat{\theta}$ are

$$\widehat{\text{BIAS}}(\hat{\theta}) = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$$

$$\text{and}\quad \widehat{\text{Var}}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2, \quad (11)$$

where $\hat{\theta}_{(\cdot)} = 1/n \sum_{i=1}^{n} \hat{\theta}_{(i)}$.

More general approaches of this 'delete-a-point' procedure exist. For example, one can consider deleting $d \geq 2$ points at a time and recomputing $\hat{\theta}$. This is known as the delete-d jackknife. The bootstrap generalizes/extends the jackknife in a slightly different fashion. Suppose we have a random sample $X_1, \ldots, X_n$ with its empirical cdf $\hat{F}_n$ putting probability mass $1/n$ on each X_i. Assuming that $X_1, \ldots, X_n$ are distributed according to $\hat{F}_n$, we can draw a sample (of size n) *with replacement* from $\hat{F}_n$, denoted $X_1^*, \ldots, X_n^*$, and then recompute $\hat{\theta}$ based on these observations. (This new sample is called a *bootstrap sample*.) Repeating this process b number of times yields $\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(b)}$. Now, based on b realizations of $\hat{\theta}$, we can evaluate the bias, variance, standard deviation, confidence intervals, or some other feature of the distribution of $\hat{\theta}$ by using empirical formulas for these measures. For example, the bootstrap estimate of standard deviation for $\hat{\theta}$ is given by

$$\widehat{\text{SD}}_{\text{boot}}(\hat{\theta}) = \sqrt{\frac{1}{b-1}\sum_{i=1}^{b}\left(\hat{\theta}^*_{(i)} - \hat{\theta}^*_{(\cdot)}\right)^2}, \quad (12)$$

where $\hat{\theta}^*_{(\cdot)} = 1/b \sum_{i=1}^{b} \hat{\theta}^*_{(i)}$. For further discussion of bootstrap methods, see [9, 12].

Finally, this so-called 'sample reuse' principle is so flexible that it can be applied to virtually any statistical problem (e.g. hypothesis testing, parametric inference, **regression**, **time series**, **multivariate statistics**, and others) and the only thing it requires is a high-speed computer (which is not a problem at this day and age). Consequently, these techniques received a considerable share of attention in the actuarial literature also. For example, Frees [15] and Hipp [20] use the bootstrap to estimate the ruin probability; Embrechts and Mikosch [13] apply it for estimating the **adjustment coefficient**; Aebi, Embrechts, Mikosch [1] provide a theoretical justification for bootstrap estimation of perpetuities and **aggregate claim amounts**; England and Verrall [14] compare the bootstrap prediction errors in **claims reserving** with their analytic equivalents. A nice account on the applications of bootstrap and its variants in actuarial practice is given by Derrig, Ostaszewski, and Rempala [10]. Both the jackknife and the bootstrap are used by Pitts, Grübel, and Embrechts [33] in constructing confidence bounds for the adjustment coefficient. Zehnwirth [49] applied the jackknife for estimating the variance part of the **credibility premium**. See also a subsequent article by Sundt [44], which provides a critique of Zehnwirth's approach.

Smoothing Techniques

Here we provide a brief discussion of two leading areas of application of smoothing techniques – nonparametric density estimation and nonparametric regression.

Nonparametric Density Estimation. Let $X_1, \ldots, X_n$ be a random sample from a population with the density function f, which is to be estimated nonparametrically, that is, without assuming that f is from a known parametric family. The naïve histogram-like estimator, $\tilde{f}(x) = 1/n \sum_{i=1}^{n}(\mathbf{1}\{x - h < X_i < x + h\})/2h$, where $h > 0$ is a small (subjectively selected) number, lacks 'smoothness', which results in some technical difficulties (see [40]). A straightforward correction (and generalization) of the naïve estimator is to replace the term $\mathbf{1}\{\cdot\}/2$ in the expression of $\tilde{f}$ by a smooth function. This leads to the *kernel estimator*

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{13}$$

where $h > 0$ is the *window width* or *bandwidth*, and $K(\cdot)$ is a *kernel* function that satisfies the condition $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$. (Note that the estimator $\hat{f}$ is simply a sum of 'bumps' placed at the observations. Here function K determines the shape of the bumps while the bandwidth h determines their width.)

In this setting, two problems are of crucial importance – the choice of the kernel function K and the choice of h (i.e. how much to smooth). There are many criteria available for choosing K and h. Unfortunately, no unique, universally best answer exists. Two commonly used *symmetric* kernels are the Gaussian kernel $K_{\mathrm{G}}(x) = (1/\sqrt{2\pi})\mathrm{e}^{-x^2/2}$, for $-\infty < x < \infty$, and the Epanechnikov kernel $K_{\mathrm{E}}(x) = (3/4\sqrt{5})(1 - x^2/5)$, for $-\sqrt{5} < x < \sqrt{5}$. Among all symmetric kernels, the Epanechnikov kernel is the most efficient in the sense of its smallest mean integrated square error (MISE). The MISE of the Gaussian kernel, however, is just about 5% larger, so one does not lose much efficiency by using K_{G} instead of K_{E}. The MISE criterion can also be used in determining optimal h. The solution, however, is a disappointment since it depends on the unknown density being estimated (see [40], p. 40). Therefore, additional (or different) criteria have to be employed. These considerations lead to a variety of methods for choosing h, including subjective choice, reference to a standard distribution, cross-validation techniques, and others (see [40], Chapter 3).

There are many other nonparametric density estimators available, for example, estimators based on the general weight function class, the orthogonal series estimator, the nearest neighbor estimator, the variable kernel method estimator, the maximum penalized likelihood approach, and others (see [40], Chapter 2). The kernel method, however, continues to have leading importance mainly because of its well-understood theoretical properties and its wide applicability. In the actuarial literature, kernel density estimation was used by Young [47, 48] and Nielsen and Sandqvist [31] to derive the *credibility*-based estimators.

For multivariate density estimation techniques, see Chapter 4 in [40, 41].

Nonparametric Regression. The most general **regression model** is defined by

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{14}$$

where $m(\cdot)$ is an unknown function and $\varepsilon_1, \ldots, \varepsilon_n$ represent a random sample from a continuous population that has median 0.

Since the form of $m(\cdot)$ is not specified, the variability in the response variable Y makes it difficult to describe the relationship between x and Y. Therefore, one employs certain smoothing techniques (called *smoothers*) to dampen the fluctuations of Y as x changes. Some commonly used smoothers are *running line smoothers*, *kernel regression smoothers*, *local regression smoothers*, and *spline regression smoothers*. Here we provide a discussion of the spline regression smoothers. For a detailed description of this and other approaches, see [38] and Chapter 5 in [41].

A **spline** is a curve pieced together from a number of individually constructed *polynomial* curve segments. The juncture points where these curves are connected are called *knots*. The underlying principle is to estimate the unknown (but smooth) function $m(\cdot)$ by finding an optimal trade-off between smoothness of the estimated curve and goodness-of-fit of the curve to the original data. For regression data, the residual sum of squares is a natural measure of goodness-of-fit, thus the so-called *roughness penalty* estimator $\hat{m}(\cdot)$ is found by solving the minimization problem

$$L = \frac{1}{n}\sum_{i=1}^{n}(Y_i - m(x_i))^2 + \phi(m), \qquad (15)$$

where $\phi(\cdot)$ is a roughness penalty function that decreases as $m(\cdot)$ gets smoother. To guarantee that the minimization problem has a solution, some restrictions have to be imposed on function $\phi(\cdot)$. The most common version of L takes the form

$$L = \frac{1}{n}\sum_{i=1}^{n}(Y_i - m(x_i))^2 + h\int (m''(u))^2\,\mathrm{d}u, \quad (16)$$

where h acts as a smoothing parameter (analogous to the bandwidth for kernel estimators) and functions $m(\cdot)$ and $m'(\cdot)$ are absolutely continuous and $m''(\cdot)$ is square integrable. Then, the estimator $\hat{m}(\cdot)$ is called a *cubic smoothing spline* with knots at predictor values $x_1, \ldots, x_n$.

Carriere [5, 7] used smoothing splines (in conjunction with other above-mentioned nonparametric approaches) for valuation of American options and **instantaneous interest rates**. Qian [35] applied nonparametric regression methods for calculation of **credibility premiums**.

Remark 3 *An alternative formulation*. A natural extension of classical nonparametric methods to regression are the procedures based on ranks. These techniques start with a specific regression model (with associated parameters) and then develop distribution-free methods for making inferences about the unknown parameters. The reader interested in the rank-based regression procedures is referred to [19, 23].

Final Remarks

Textbooks and Further Reading

To get a more complete picture on the theory and applications of nonparametric statistics, the reader should consult some of the following texts. Recent applications-oriented books include [17, 23]. For theoretical intermediate-level reading, check [18, 25, 37]. A specialized theoretical text by Hettmansperger and McKean [19] focuses on **robustness** aspects of nonparametric regression. Silverman [40] presents density estimation techniques with special emphasis on topics of methodological interest. An applications-oriented book by Simonoff [41] surveys the uses of smoothing techniques in statistics. Books on bootstrapping include Efron and Tibshirani's [12] introductory book and Davison and Hinkley's [9] intermediate-level book. Both texts contain many practical illustrations.

Software

There is a rich variety of statistical packages that have some capabilities to perform nonparametric inference. These include `BMDP`, `Minitab`, `SAS`, `S-Plus`, `SPSS`, `Stata`, `STATISTICA`, `StatXact`, and others. Of particular interest in this regard is the `StatXact` package because of its unique ability to produce exact p-values and exact confidence intervals. Finally, we should mention that most of the packages are available for various platforms including Windows, Macintosh, and UNIX.

References

[1] Aebi, M., Embrechts, P. & Mikosch, T. (1994). Stochastic discounting, aggregate claims, and the bootstrap, *Advances in Applied Probability* **26**, 183–206.

[2] Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes, *Philosophical Transaction of the Royal Society of London* **27**, 186–190.

[3] Brazauskas, V. (2003). Influence functions of empirical nonparametric estimators of net reinsurance premiums, *Insurance: Mathematics and Economics* **32**, 115–133.

[4] Carriere, J.F. (1993). Nonparametric tests for mixed Poisson distributions, *Insurance: Mathematics and Economics* **12**, 3–8.

[5] Carriere, J.F. (1996). Valuation of the early-exercise price for options using simulations and nonparametric regression, *Insurance: Mathematics and Economics* **19**, 19–30.

[6] Carriere, J.F. (1997). Testing independence in bivariate distributions of claim frequencies and severities, *Insurance: Mathematics and Economics* **21**, 81–89.

[7] Carriere, J.F. (2000). Non-parametric confidence intervals of instantaneous forward rates, *Insurance: Mathematics and Economics* **26**, 193–202.

[8] Croux, K. & Veraverbeke, N. (1990). Nonparametric estimators for the probability of ruin, *Insurance: Mathematics and Economics* **9**, 127–130.

[9] Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*, Cambridge University Press, New York.

[10] Derrig, R.A., Ostaszewski, K.M. & Rempala, G.A. (2000). Applications of resampling methods in actuarial practice, *Proceedings of the Casualty Actuarial Society* **LXXXVII**, 322–364.

[11] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7**, 1–26.

[12] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

[13] Embrechts, P. & Mikosch, T. (1991). A bootstrap procedure for estimating the adjustment coefficient, *Insurance: Mathematics and Economics* **10**, 181–190.

[14] England, P. & Verrall, R. (1999). Analytic and bootstrap estimates of prediction errors in claims reserving, *Insurance: Mathematics and Economics* **25**, 281–293.

[15] Frees, E.W. (1986). Nonparametric estimation of the probability of ruin, *ASTIN Bulletin* **16**, S81–S90.

[16] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.

[17] Gibbons, J.D. (1997). *Nonparametric Methods for Quantitative Analysis*, 3rd edition, American Sciences Press, Syracuse, NY.

[18] Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks*, Wiley, New York.

[19] Hettmansperger, T.P. & McKean, J.W. (1998). *Robust Nonparametric Statistical Methods*, Arnold, London.

[20] Hipp, C. (1989). Estimators and bootstrap confidence intervals for ruin probabilities, *ASTIN Bulletin* **19**, 57–70.

[21] Hipp, C. (1996). The delta-method for actuarial statistics, *Scandinavian Actuarial Journal*, 79–94.

[22] Hodges, J.L. & Lehmann, E.L. (1963). Estimates of location based on rank tests, *Annals of Mathematical Statistics* **34**, 598–611.

[23] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd edition, Wiley, New York.

[24] Kendall, M.G. (1938). A new measure of rank correlation, *Biometrika* **30**, 81–93.

[25] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, Holden Day, San Francisco.

[26] Ludwig, S.J. & McAuley, R.F. (1988). A nonparametric approach to evaluating reinsurers' relative financial strength, *Proceedings of the Casualty Actuarial Society* **LXXV**, 219–240.

[27] Mack, G.A. & Wolfe, D.A. (1981). K-sample rank tests for umbrella alternatives, *Journal of the American Statistical Association* **76**, 175–181.

[28] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.

[29] Marceau, E. & Rioux, J. (2001). On robustness in risk theory, *Insurance: Mathematics and Economics* **29**, 167–185.

[30] Nakamura, M. & Pérez-Abreu, V. (1993). Empirical probability generating function: an overview, *Insurance: Mathematics and Economics* **12**, 287–295.

[31] Nielsen, J.P. & Sandqvist, B.L. (2000). Credibility weighted hazard estimation, *ASTIN Bulletin* **30**, 405–417.

[32] Pitts, S.M. (1994). Nonparametric estimation of compound distributions with applications in insurance, *Annals of the Institute of Statistical Mathematics* **46**, 537–555.

[33] Pitts, M.S., Grübel, R. & Embrechts, P. (1996). Confidence bounds for the adjustment coefficient, *Advances in Applied Probability* **28**, 802–827.

[34] Præstgaard, J. (1991). Nonparametric estimation of actuarial values, *Scandinavian Actuarial Journal* **2**, 129–143.

[35] Qian, W. (2000). An application of nonparametric regression estimation in credibility theory, *Insurance: Mathematics and Economics* **27**, 169–176.

[36] Quenouille, M.H. (1949). Approximate tests of correlation in time-series, *Journal of the Royal Statistical Society, Series B* **11**, 68–84.

[37] Randles, R.H. & Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics*, Wiley, New York.

[38] Ryan, T.P. (1997). *Modern Regression Methods*, Wiley, New York.

[39] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

[40] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[41] Simonoff, J.S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

[42] Smirnov, N.V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples, *Bulletin of Moscow University* **2**, 3–16 (in Russian).

[43] Spearman, C. (1904). The proof and measurement of association between two things, *American Journal of Psychology* **15**, 72–101.

[44] Sundt, B. (1982). The variance part of the credibility premium without the jackknife, *Scandinavian Actuarial Journal*, 179–187.

[45] Tukey, J.W. (1958). Bias and confidence in not-quite large samples, *Annals of Mathematical Statistics* **29**, 614.

[46] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

[47] Young, V.R. (1997). Credibility using semiparametric models, *ASTIN Bulletin* **27**, 273–285.

[48] Young, V.R. (1998). Robust Bayesian credibility using semiparametric models, *ASTIN Bulletin* **28**, 187–203.

[49] Zehnwirth, B. (1981). The jackknife and the variance part of the credibility premium, *Scandinavian Actuarial Journal*, 245–249.

(*See also* **Adverse Selection**; **Competing Risks**; **Dependent Risks**; **Diffusion Processes**; **Estimation**; **Frailty**; **Graduation**; **Survival Analysis**; **Value-at-risk**)

VYTARAS BRAZAUSKAS