2007

Nonparametric Statistics with Applications to Science and Engineering

Paul Kvam
University of Richmond, pkvam@richmond.edu

Brani Vidakovic

1

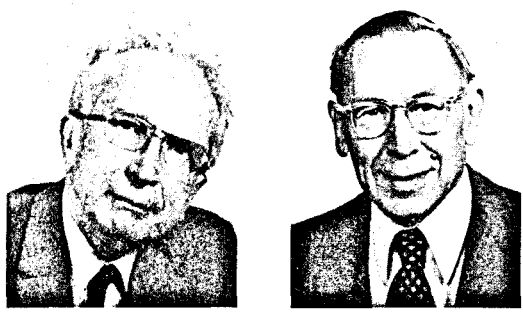
Introduction

For every complex question, there is a simple answer.... and it is wrong.

H. L. Mencken

Jacob Wolfowitz (Figure 1.1a) first coined the term *nonparametric*, saying "We shall refer to this situation [*where a distribution is completely determined by the knowledge of its finite parameter set*] as the parametric case, and denote the opposite case, where the functional forms of the distributions are unknown, as the non-parametric case" (Wolfowitz, 1942). From that point on, nonparametric statistics was defined by what it is not: traditional statistics based on known distributions with unknown parameters. Randles, Hettmansperger, and Casella (2004) extended this notion by stating "nonparametric statistics can and should be broadly defined to include all methodology that does not use a model based on a single parametric family."

Traditional statistical methods are based on parametric assumptions; that is, that the data can be assumed to be generated by some well-known family of distributions, such as normal, exponential, Poisson, and so on. Each of these distributions has one or more parameters (e.g., the normal distribution has μ and σ^2), at least one of which is presumed unknown and must be inferred. The emphasis on the normal distribution in linear model theory is often justified by the central limit theorem, which guarantees *approximate normality* of sample means provided the sample sizes are large enough. Other distributions also play an important role in science and engineering. Physical failure mechanisms often characterize the lifetime distribution of industrial compo-



(a) (b)

Fig. 1.1 (a) Jacob Wolfowitz (1910–1981) and (b) Wassily Hoeffding (1914–1991), pioneers in nonparametric statistics.

nents (e.g., Weibull or lognormal), so parametric methods are important in reliability engineering.

However, with complex experiments and messy sampling plans, the generated data might not be attributed to any well-known distribution. Analysts limited to basic statistical methods can be trapped into making parametric assumptions about the data that are not apparent in the experiment or the data. In the case where the experimenter is not sure about the underlying distribution of the data, statistical techniques are needed which can be applied regardless of the true distribution of the data. These techniques are called *nonparametric methods*, or *distribution-free methods.*

> The terms nonparametric and distribution-free are not synonymous... Popular usage, however, has equated the terms ... Roughly speaking, a nonparametric test is one which makes no hypothesis about the value of a parameter in a statistical density function, whereas a distribution-free test is one which makes no assumptions about the precise form of the sampled population.
>
> J. V. Bradley (1968)

It can be confusing to understand what is implied by the word "nonparametric". What is termed *modern nonparametrics* includes statistical models that are quite refined, except the distribution for error is left unspecified. Wasserman's recent book *All Things Nonparametric* (Wasserman, 2005) emphasizes only modern topics in nonparametric statistics, such as curve fitting, density estimation, and wavelets. Conover's *Practical Nonparametric Statistics* (Conover, 1999), on the other hand, is a classic nonparametrics textbook, but mostly limited to traditional binomial and rank tests, contingency tables, and tests for goodness of fit. Topics that are not really under the distribution-free umbrella, such as robust analysis, Bayesian analysis, and statistical learning also have important connections to nonparametric statistics, and are all

featured in this book. Perhaps this text could have been titled *A Bit Less of Parametric Statistics with Applications in Science and Engineering*, but it surely would have sold fewer copies. On the other hand, if sales were the primary objective, we would have titled this *Nonparametric Statistics for Dummies* or maybe *Nonparametric Statistics with Pictures of Naked People.*

1.1 EFFICIENCY OF NONPARAMETRIC METHODS

It would be a mistake to think that nonparametric procedures are simpler than their parametric counterparts. On the contrary, a primary criticism of using parametric methods in statistical analysis is that they oversimplify the population or process we are observing. Indeed, parametric families are not more useful because they are perfectly appropriate, rather because they are perfectly convenient.

Nonparametric methods are inherently less powerful than parametric methods. This must be true because the parametric methods are assuming more information to construct inferences about the data. In these cases the estimators are inefficient, where the efficiencies of two estimators are assessed by comparing their variances for the same sample size. This inefficiency of one method relative to another is measured in power in hypothesis testing, for example.

However, even when the parametric assumptions hold perfectly true, we will see that nonparametric methods are only slightly less powerful than the more presumptuous statistical methods. Furthermore, if the parametric assumptions about the data fail to hold, only the nonparametric method is valid. A t-test between the means of two normal populations can be dangerously misleading if the underlying data are not actually normally distributed. Some examples of the relative efficiency of nonparametric tests are listed in Table 1.1, where asymptotic relative efficiency (A.R.E.) is used to compare parametric procedures (2^{nd} column) with their nonparametric counterparts (3^{rd} column). Asymptotic relative efficiency describes the relative efficiency of two estimators of a parameter as the sample size approaches infinity. The A.R.E. is listed for the normal distribution, where parametric assumptions are justified, and the double-exponential distribution. For example, if the underlying data are normally distributed, the t-test requires 955 observations in order to have the same power of the Wilcoxon signed-rank test based on 1000 observations.

Parametric assumptions allow us to extrapolate away from the data. For example, it is hardly uncommon for an experimenter to make inferences about a population's extreme upper percentile (say 99^{th} percentile) with a sample so small that none of the observations would be expected to exceed that percentile. If the assumptions are not justified, this is grossly unscientific.

Nonparametric methods are seldom used to extrapolate outside the range

Table 1.1 Asymptotic relative efficiency (A.R.E.) of some nonparametric tests

	Parametric Test	Nonparametric Test	A.R.E. (normal)	A.R.E. (double exp.)
2-Sample Test	t-test	Mann-Whitney	0.955	1.50
3-Sample Test	one-way layout	Kruskal-Wallis	0.864	1.50
Variances Test	F-test	Conover	0.760	1.08

of observed data. In a typical nonparametric analysis, little or nothing can be said about the probability of obtaining future data beyond the largest sampled observation or less than the smallest one. For this reason, the actual measurements of a sample item means less compared to its rank within the sample. In fact, nonparametric methods are typically based on *ranks* of the data, and properties of the population are deduced using *order statistics* (Chapter 5). The measurement scales for typical data are

> *Nominal Scale:* Numbers used only to categorize outcomes (e.g., we might define a random variable to equal one in the event a coin flips heads, and zero if it flips tails).
>
> *Ordinal Scale:* Numbers can be used to order outcomes (e.g., the event X is greater than the event Y if X = *medium* and Y = *small*).
>
> *Interval Scale:* Order between numbers as well as distances between numbers are used to compare outcomes.

Only interval scale measurements can be used by parametric methods. Nonparametric methods based on ranks can use ordinal scale measurements, and simpler nonparametric techniques can be used with nominal scale measurements.

The binomial distribution is characterized by counting the number of independent observations that are classified into a particular category. Binomial data can be formed from measurements based on a *nominal scale* of measurements, thus binomial models are most encountered models in nonparametric analysis. For this reason, Chapter 3 includes a special emphasis on statistical estimation and testing associated with binomial samples.

1.2 OVERCONFIDENCE BIAS

> Be slow to believe what you worst want to be true
>
> Samual Pepys

Confirmation Bias or *Overconfidence Bias* describes our tendency to search for or interpret information in a way that confirms our preconceptions. Business and finance has shown interest in this psychological phenomenon (Tversky and Kahneman, 1974) because it has proven to have a significant effect on personal and corporate financial decisions where the decision maker will actively seek out and give extra weight to evidence that confirms a hypothesis they already favor. At the same time, the decision maker tends to ignore evidence that contradicts or disconfirms their hypothesis.

Overconfidence bias has a natural tendency to effect an experimenter's data analysis for the same reasons. While the dictates of the experiment and the data sampling should reduce the possibility of this problem, one of the clear pathways open to such bias is the infusion of parametric assumptions into the data analysis. After all, if the assumptions seem plausible, the researcher has much to gain from the extra certainty that comes from the assumptions in terms of narrower confidence intervals and more powerful statistical tests.

Nonparametric procedures serve as a buffer against this human tendency of looking for the evidence that best supports the researcher's underlying hypothesis. Given the subjective interests behind many corporate research findings, nonparametric methods can help alleviate doubt to their validity in cases when these procedures give statistical significance to the corporations's claims.

1.3 COMPUTING WITH MATLAB

Because a typical nonparametric analysis can be computationally intensive, computer support is essential to understand both theory and applications. Numerous software products can be used to complete exercises and run nonparametric analysis in this textbook, including SAS, **R**, S-Plus, MINITAB, StatXact and JMP (to name a few). A student familiar with one of these platforms can incorporate it with the lessons provided here, and without too much extra work.

It must be stressed, however, that demonstrations in this book rely entirely on a single software tool called MATLAB© (by MathWorks Inc.) that is used widely in engineering and the physical sciences. MATLAB (short for *MATrix LABoratory*) is a flexible programming tool that is widely popular in engineering practice and research. The program environment features user-friendly front-end and includes menus for easy implementation of program commands. MATLAB is available on Unix systems, Microsoft Windows and

Apple Macintosh. If you are unfamiliar with MATLAB, in the first appendix we present a brief tutorial along with a short description of some MATLAB procedures that are used to solve analytical problems and demonstrate nonparametric methods in this book. For a more comprehensive guide, we recommend the handy little book *MATLAB Primer* (Sigmon and Davis, 2002).

We hope that many students of statistics will find this book useful, but it was written primarily with the scientist and engineer in mind. With nothing against statisticians (some of our best friends know statisticians) our approach emphasizes the application of the method over its mathematical theory. We have intentionally made the text less heavy with theory and instead emphasized applications and examples. If you come into this course thinking the history of nonparametric statistics is dry and unexciting, you are probably right, at least compared to the history of ancient Rome, the British monarchy or maybe even Wayne Newton[1]. Nonetheless, we made efforts to convince you otherwise by noting the interesting historical context of the research and the personalities behind its development. For example, we will learn more about Karl Pearson (1857–1936) and R. A. Fisher (1890–1962), legendary scientists and competitive arch-rivals, who both contributed greatly to the foundation of nonparametric statistics through their separate research directions.

Fig. 1.2 "Doubt is not a pleasant condition, but certainty is absurd" – Francois Marie Voltaire (1694–1778).

In short, this book features techniques of data analysis that rely less on the assumptions of the data's good behavior – the very assumptions that can get researchers in trouble. Science's gravitation toward distribution-free techniques is due to both a deeper awareness of experimental uncertainty and the availability of ever-increasing computational abilities to deal with the implied ambiguities in the experimental outcome. The quote from Voltaire

[1]Strangely popular Las Vegas entertainer.

(Figure 1.2) exemplifies the attitude toward uncertainty; as science progresses, we are able to see some truths more clearly, but at the same time, we uncover more uncertainties and more things become less "black and white".

1.4 EXERCISES

1.1. Describe a potential data analysis in engineering where parametric methods are appropriate. How would you defend this assumption?

1.2. Describe another potential data analysis in engineering where parametric methods may not be appropriate. What might prevent you from using parametric assumptions in this case?

1.3. Describe three ways in which overconfidence bias can affect the statistical analysis of experimental data. How can this problem be overcome?

REFERENCES

Bradley, J. V. (1968), *Distribution Free Statistical Tests*, Englewood Cliffs, NJ: Prentice Hall.

Conover, W. J. (1999), *Practical Nonparametric Statistics*, New York: Wiley.

Randles, R. H., Hettmansperger, T.P., and Casella, G. (2004), Introduction to the Special Issue "Nonparametric Statistics," *Statistical Science*, 19, 561-562.

Sigmon, K., and Davis, T.A. (2002), *MATLAB Primer*, 6th Edition, MathWorks, Inc.; Boca Raton, FL: CRC Press.

Tversky, A., and Kahneman, D. (1974), "Judgment Under Uncertainty: Heuristics and Biases," *Science*, 185, 1124-1131.

Wasserman, L. (2006), *All Things Nonparametric*, New York: Springer Verlag.

Wolfowitz, J. (1942), "Additive Partition Functions and a Class of Statistical Hypotheses," *Annals of Statistics*, 13, 247-279.