What Is Nonparametric Statistics?*

Louis H. Y. Chen
Department of Mathematics
National University of Singapore

Nonparametric statistics can be said to date back to the eighteenth century but its modern development did not begin until the late 1930's. Today nonparametric statistics is one of the major branches of statistics. The objective of this lecture is not to give any survey but to focus on a basic feature of the subject which distinguishes it from parametric statistics. Let us begin with a few examples.

1. Nonparametric tests : Suppose $X_1, \ldots, X_n$ is a random sample from a population with mean μ and suppose we wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. If the population distribution is normal, we use the t statistic

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

where

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Under H_0, $t \sim t_{n-1}$. So at the α level of significance, a critical region is

$$t > t_{n-1,\alpha/2} \quad \text{or} \quad t < -t_{n-1,\alpha/2}.$$

Suppose the population distribution is not normal or we are not sure that it is normal. Do we still use the t statistic ? There are two approaches to answering this question :

(a) Investigate how the distribution of the t statistic is affected by the departure from normality.

---

* Lecture given at the Workshop on Hypothesis Testing and Non-Parametric Statistics for school teachers, 7 September 1987.

(b) Find a test statistic whose distribution remains the same over a larger class of population distributions than the class of normal distributions.

Tests in which the distribution of the test statistic remains the same over a "sufficiently large" class of population distributions are called *nonparametric tests.* We will give a more precise meaning to "sufficiently large" later. A nonparametric test for the above problem is *the sign test.* Assume that the population has a continuous distribution which is symmetric about μ. Then μ is the population mean. Define

$$Z_i = \begin{cases} 1 & \text{if } X_i - \mu_0 > 0 \\ 0 & \text{if } X_i - \mu_0 < 0, \end{cases}$$

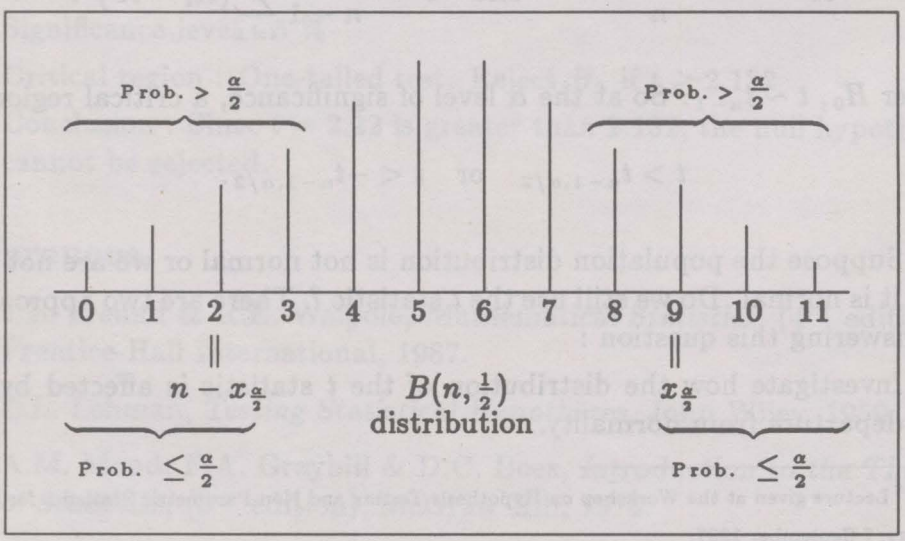
and let $S = \sum_{i=1}^{n} Z_i$, that is S is the number of observations in the random sample which are greater than μ_0. Then under H_0,

$$S \sim B(n, \tfrac{1}{2}).$$

So at the α level of significance, a critical region is

$$S \geq x_{\alpha/2} \quad \text{or} \quad S \leq n - x_{\alpha/2}$$

where $x_{\alpha/2}$ is the smallest integer such that $P(S \geq x_{\alpha/2} \mid H_0) \leq \dfrac{\alpha}{2}$.



The distribution of S remains $B(n, \frac{1}{2})$ over all continuous population distributions which are symmetric about μ_0. This class of distributions includes and is much larger than the class of normal distributions with mean μ_0. It consists of more than one distributional form.

The sign test can also be applied to matched pair experiments where a t test would again be used if the population distribution is known to be normal. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample where X_i and Y_i represent the responses of a control and a treatment subject respectively. For example, X_i and Y_i may be the hæmoglobin levels of the i^{th} anemic patient before and after treatment with vitamin B_{12}, or X_i and Y_i may be the sales performance before and after a course in salesmanship. Suppose we wish to test

H_0 : the treatment is ineffective

against

H_1 : the treatment is effective.

Under H_0, the distribution of $Y_i - X_i$ is symmetric about 0. Assume that this distribution is continuous. Define

$$Z_i = \begin{cases} 1 & \text{if } Y_i - X_i > 0 \\ 0 & \text{if } Y_i - X_i < 0, \end{cases}$$

and let $S = \sum_{i=1}^{n} Z_i$. Then under H_0,

$$S \sim B(n, \tfrac{1}{2}).$$

At the α level of significance, a critical region is $S \geq x_\alpha$ where as before x_α is the smallest integer such that $P(S \geq x_\alpha \mid H_0) \leq \alpha$. The distribution of S remains $B(n, \frac{1}{2})$ over all continuous distributions of $Y_i - X_i$ which is symmetric about 0. As has been mentioned above, this class of distributions includes and is much larger than the class of normal distributions with mean 0, and consists of more than one distributional form.

2. Nonparametric confidence intervals : Under the normality assumption, the t statistic is used to obtain a confidence interval for the population mean μ. From the fact that

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

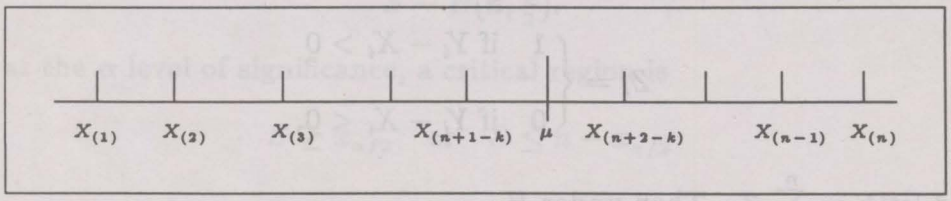
for whatever value of μ, a $100(1-\alpha)\%$ confidence interval for μ can be obtained as follows :

$$\overline{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}.$$

The sign statistic can also be used to obtain a confidence interval for the mean μ of a continuous distribution symmetric about μ. Let $X_1, \ldots, X_n$ be a random sample and let $S(\mu)$ be the number of X_i's which are greater than μ. Then $S(\mu) \sim B(n, \frac{1}{2})$ for whatever value of μ. Let $b(n, \frac{1}{2}, \frac{\alpha}{2})$ be the $100(1-\frac{\alpha}{2})^{\text{th}}$ percentage point of $B(n, \frac{1}{2})$, that is $b(n, \frac{1}{2}, \frac{\alpha}{2})$ is a nonnegative integer such that $P(S(\mu) \geq b(n, \frac{1}{2}, \frac{\alpha}{2})) = \frac{\alpha}{2}$. Since $B(n, \frac{1}{2})$ is symmetric about its centre, $P(S(\mu) \leq n - b(n, \frac{1}{2}, \frac{\alpha}{2})) = \frac{\alpha}{2}$. So

$$P(n - b(n, \tfrac{1}{2}, \tfrac{\alpha}{2}) < S(\mu) < b(n, \tfrac{1}{2}, \tfrac{\alpha}{2})) = 1 - \alpha.$$

Now arrange $X_1, \ldots, X_n$ in an increasing order $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$.



Then $S(\mu) < k$ if and only if $\mu \geq X_{(n+1-k)}$. Similarly $S(\mu) > l$ if and only if $\mu < X_{(n-l)}$. Hence $n - b(n, \frac{1}{2}, \frac{\alpha}{2}) < S(\mu) < b(n, \frac{1}{2}, \frac{\alpha}{2})$ if and only if

$$X_{\left(n+1-b(n,\frac{1}{2},\frac{\alpha}{2})\right)} \leq \mu < X_{\left(b(n,\frac{1}{2},\frac{\alpha}{2})\right)}$$

which is a $100(1-\alpha)\%$ confidence interval for μ. Here the confidence interval applies to the class of all continuous distributions which are symmetric about μ. It is called a nonparametric confidence interval.

3. Nonparametric estimators : Here we make no assumption whatsoever on the population distribution. Suppose $X_1, \ldots, X_n$ is a random sample from a population whose distribution function is F. Define

$$\hat{F}_n(x) = \frac{\text{No. of } X_i \leq x}{n}.$$

The function $\hat{F}_n$ is called the *empirical distribution function.* It is known that $\hat{F}_n(x)$ converges to $F(x)$ in some very strong sense as $n \longrightarrow \infty$. If a characteristic of the population distribution can be expressed as

$$\gamma = \int_{-\infty}^{\infty} \psi(x) dF(x),$$

then

$$\hat{\gamma} = \int_{-\infty}^{\infty} \psi(x) d\hat{F}_n(x)$$

is a reasonable estimator of γ. Such estimators are nonparametric estimators because no assumption is made on the population distribution. Some special cases of $\hat{\gamma}$ are :

(i) $\psi(x) = \begin{cases} 1 & \text{if } x \in I \\ 0 & \text{if } x \notin I \end{cases}$ where I is an interval.

$$\hat{\gamma} = \frac{\text{No. of } X_i \in I}{n}, \quad \gamma = P(X_i \in I).$$

(ii) $\psi(x) = x$.

$\hat{\gamma}$ = sample mean $\overline{X}$, γ = population mean.

(iii) $\psi(x) = \begin{cases} x & \text{if } -a \leq x \leq a \\ 0 & \text{otherwise.} \end{cases}$

$$\hat{\gamma} = \text{truncated sample mean} = \frac{1}{n} \sum_{|X_i| \leq a} X_i,$$

$$\gamma = \text{truncated mean} = \int_{|x| \leq a} x dF(x).$$

In a statistical problem, we often do not know or are uncertain about certain aspects of the population distribution. A random sample is then drawn from the population and on the basis of the sample we infer about these unknown aspects. A population distribution some of whose aspects are unknown may be thought of as one of a family of possible distributions $\{P_\theta : \theta \in \Theta\}$. Such a family of distributions is called a *statistical model* where θ is called a *parameter* and Θ the *parameter space.* For example, if we know that the number of customers at a certain shop in a day has a Poisson distribution but do not know the average number of customers per day, then the statistical model is the family of Poisson distributions with mean θ, where θ is the parameter and the parameter space Θ is $(0, \infty)$. Similarly, the statistical model for the measurement of an unknown physcial quantity may be the family of normal distributions with

mean μ and variance σ^2 where the parameter $\theta = (\mu, \sigma^2)$ and the parameter space $\Theta = \mathbf{R} \times (0, \infty)$ or $(0, \infty) \times (0, \infty)$. In these two examples, the distributions are "naturally" parametrized with the parameter space Θ being a subset of an euclidean space and therefore finite dimensional. Any statistical procedure for such a statistical model is classified as a *parametric procedure*. The t test and the construction of confidence intervals using the t statistic mentioned in the above examples are parametric procedures.

The family of continuous distributions which are symmetric about some point cannot be "naturally" expressed as $\{P_\theta : \theta \in \Theta\}$ with Θ being a finite dimensional set. More so is the class of all distributions considered in Example 3. Such statistical models are called *nonparametric models*. The sign test, the construction of confidence intervals using the sign statistic, and the method of estimation using the empirical distribution function, which are statistical procedures for such nonparametric models, are therefore called *nonparametric procedures*. The field of nonparametric statistics can be said to consist of the theory and applications of nonparametric procedures.

Nonparametric procedures have gained much popularity because they require less restrictive assumptions on the population distribution than their parametric counterparts. Since not all the information in the random sample is used in the procedures, one might think that there would be a great loss of efficiency. Surprisingly, at times nonparametric procedures perform almost just as well as parametric procedures when the population distribution is assumed to be normal.

References

[1] Bickel, P J & Doksum, K A, *Mathematical Statistics : Basic Ideas and Selected Topics*. Holden-Day, Inc., 1977.

[2] Breiman, L, *Statistics : With a View Toward Applications*. Houghton Mifflin Company, 1973.

[3] Larsen, R J & Marx, M L, *An Introduction to Mathematical Statistics and Its Applications*, Second Edition. Prentice Hall, 1986.

[4] Lehmann, E L, *Nonparametrics : Statistical Methods Based on Ranks*. Holden-Day, Inc., 1975.

[5] Randles, R H & Wolfe, D A, *Introduction to the Theory of Nonparametric Statistics*. John Wiley & Sons, 1979.