Nonparametric Statistics

Tom Hettmansperger

Based on notes by S. M. Bendre
University of Hyderabad
India

Contents

1 Parametric and Nonparametric models

A *parametric statistical model* is a model where the joint distribution of the observations involves several unknown constants called *parameters.* The functional form of the joint distribution is assumed to be known and the only unknowns in the model are the parameters. Two parametric models commonly encountered in astronomical experiments are

1. The Poisson model in which we assume that the observations are independent Poisson random variables with unknown common mean θ.
2. The normal model in which the observations are independently distributed with unknown mean μ and unknown variance σ^2.

In the first model θ is the parameter and in the second μ and σ^2 are the parameters.

Anything we can compute from the observations is called a *statistic.* In parametric statistics the goal is to use observations to draw inference about the unobserved parameters and hence about the underlined model.

A *nonparametric model* is the one in which no assumption is made about the functional form of the joint distribution except that the observations are independent identically distributed (i.i.d.) from an arbitrary continuous distribution. As a result, the nonparametric statistics is also called *distribution free* statistics. There are no parameters in a nonparametric model.

A *semiparametric model* is the one which has parameters but very weak assumptions are made about the actual form of the distribution of the observations.

Both nonparametric and semiparametric models are often lumped together and called nonparametric models.

2 Why Nonparametric?

While in many situations parametric assumptions are reasonable (e.g. assumption of Normal distribution for the background noise, Poisson distribution for a photon counting signal of a nonvariable source), we often have no prior knowledge of the underlying distributions. In such situations, the use of parametric statistics can give misleading or even wrong results.

We need statistical procedures which are insensitive to the model assumptions in the sense that the procedures retain their properties in the neighborhood of the model assumptions.

Insensitivity to model assumptions : **Robustness**

In particular, for

- Estimation

  The estimators such that

  - the variance (precision) of an estimator is not sensitive to model assumptions (Variance Robustness).

- Hypothesis Testing

  We need test procedures where

  - the level of significance is not sensitive to model assumptions (Level Robustness).
  - the statistical power of a test to detect important alternative hypotheses is not sensitive to model assumptions (Power Robustness).

Apart from this, we also need procedures which are robust against the presence of outliers in the data.

Examples:

1. The sample mean is not robust against the presence of even one outlier in the data and is not variance robust as well. The sample median is robust against outliers and is variance robust.

2. The t-test does not have t-distribution if the underlined distribution is not normal and the sample size is small. For large sample size, it is asymptotically level robust but is not power robust. Also, it is not robust against the presence of outliers.

Procedures derived for nonparametric and semiparametric models are often called *robust* procedures since they depend on very weak assumptions.

3 Nonparametric Density Estimation

Let $X_1, X_2, \cdots, X_n$ be a random sample from an unknown probability density function f. The interest is to estimate the density function f itself.

Suppose the random sample is drawn from a distribution with known probability density function, say normal with mean μ and variance σ^2. The density f can then be estimated by estimating the values of the unknown parameters μ and σ^2 from the data and substituting these estimates in the expression for normal density. Thus the *parametric density estimator* is

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\{-\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu})^2\}$$

where $\hat{\mu} = \frac{1}{n}\sum_i x_i$ and $\hat{\sigma}^2 = \frac{1}{n-1}\sum_i (x_i - \hat{\mu})^2$.

In case of the *nonparametric* estimation of the density function, the functional form of the density function is assumed to be unknown. We, however, assume that the underlined distribution has a probability density f and determine its form based on the data at hand.

The oldest and widely used *nonparametric density estimator* is the histogram. Given an origin x_0 and a *bandwidth* h, we consider the intervals of length h, also called *bins*, given by $B_i = [x_0 + mh, x_0 + (m+1)h)$ where $m = 0, \pm 1, \pm 2, \cdots$ and define the histogram by

$$\begin{aligned} \hat{f}_n(x) &= \tfrac{1}{nh}[\text{ number of observations in the same bin as} x] \\ &= \tfrac{1}{nh}\textstyle\sum_{i=1}^{n} n_j I[x \in B_j] \end{aligned}$$

where n_j = number of observations lying in bin B_j.

Though it is a very simple estimate, the histogram has many drawbacks, the main one is that we are estimating a continuous function by a non-smooth discrete function. It is not robust against the choice of origin x_0 and bandwidth h. Also, it is not sensitive enough to local properties of f. Various density estimation techniques are proposed to overcome these drawbacks, one of which is the *kernel density estimation.*

Kernel Density Estimation

We consider a specified *kernel function* $K(.)$ satisfying the conditions

- $\int_{-\infty}^{\infty} K(x)dx = 1$
- K(.) is symmetric around 0, giving $\int_{-\infty}^{\infty} xK(x)dx = 0$
- $\int_{-\infty}^{\infty} x^2K(x)dx = \sigma^2(K) > 0$

and define the *kernel density estimator* by

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

The estimate of f at point x is obtained using a weighted function of observations in the h-neighborhood of x where the weight given to each of the observations in the neighborhood depends on the choice of kernel function $K(.)$. Some kernel functions are

- Uniform kernel: $K(u) = \frac{1}{2}I[|u| \leq 1]$
- Triangle kernel: $K(u) = (1 - |u|)I[|u| \leq 1]$
- Epanechnikov kernel: $K(u) = \frac{3}{4}(1 - u^2)I[|u| \leq 1]$
- Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$

The kernel density estimator satisfies the property

$$\int_{-\infty}^{\infty} \hat{f}_n(x)dx = 1$$

and on the whole gives a better estimate of the underlined density. Some of the properties are

- The kernel estimates do not depend on the choice of the origin, unlike histogram.
- The kernel density estimators are 'smoother' than the histogram estimators since they inherit the property of the smoothness of the kernel chosen.
- The kernel density estimator has a faster rate of convergence.

- Increasing the bandwidth is equivalent to increasing the amount of smoothing in the estimate. Very large $h(\to \infty)$ will give an oversmooth estimate and $h \to 0$ will lead to a needlepoint estimate giving a noisy representation of the data.

- The choice of the kernel function is not very crucial. The choice of the bandwidth, however, is crucial and the optimal bandwidth choice is extensively discussed and derived in the literature. For instance, with Gaussian kernel, the optimal (MISE) bandwidth for a normal distribution is
  $$h_{\text{opt}} = 1.06\sigma n^{-\frac{1}{5}}$$
  where σ is the population standard deviation, which is estimated from the data. This is used for non-normal distributions as well.

- The kernel density estimation can be easily generalized from univariate to multivariate data in theory.

4 Some Nonparametric Goodness-of-Fit Tests

Though the samples are drawn from unknown populations, the investigators wish to confirm whether the data fit some proposed model. The goodness-of-fit tests are useful procedures to confirm whether the proposed model satisfactorily approximates the observed situation. Apart from the usual Chi-Square goodness of fit test, we have Kolmogorov-Smirnov tests which are discussed here.

4.1 One-sample Kolomogorov-Smirnov Test

This is a test of hypothesis that the sampled population follows some specified distribution.

Suppose we observe $X_1, ..., X_n$ i.i.d. from a continuous distribution function $F(x)$. We want to test the null hypothesis that $F(x) = F_0(x)$ for all x, against the alternative that $F(x) \neq F_0(x)$ for some x, where F_0 is a distribution which is completely specified before we collect the data. Let $\widehat{F}_n(x)$ be the empirical distribution function (e.d.f.) defined by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x]$$

The one sample *Kolmogorov-Smirnov* (K-S) statistic is

$$M = \max_x \left| \widehat{F}_n(x) - F_0(x) \right|$$

A large value of M supports $F(x) \neq F_0(x)$ and we reject the null hypothesis if M is too large.

It is not hard to show that the exact null distribution of M is the same for all F_0, but different for different n. Table of critical values are given in many books. For large n

$$P(nM > q) \doteq 2 \sum_{i=1}^{\infty} (-1)^{i-1} \exp\left(-2i^2 q^2\right) \doteq 2\exp\left(-2q^2\right)$$

Use of the last formula is quite accurate and conservative. Hence for a size α test we reject $H_0 : F(x) = F_0(x)$ if

$$nM > \left(-\frac{1}{2}\log\left(\frac{\alpha}{2}\right)\right)^{1/2} = M^{\alpha}$$

The K-S statistic is also called *K-S distance* since it provides a measure of closeness of $\widehat{F}_n$ to F_0. It gives a method of constructing confidence band for the distribution which helps in identifying departures from the assumed distribution F_0, as we now show. First note that the distribution of

$$M(F) = \max_x \left| \widehat{F}_n(x) - F(x) \right|$$

is the same as null distribution for the K-S test statistic. Therefore

$$1 - \alpha = P\left(M(F) \leq M^{\alpha}\right) = P\left(\left|\widehat{F}_n(x) - F(x)\right| \leq \frac{M^{\alpha}}{n} \text{ for all } x\right)$$
$$= P\left(F(x) \in \widehat{F}_n(x) \pm \frac{M^{\alpha}}{n} \text{ for all } x\right).$$

One situation in which K-S is misused is in testing for normality. For K-S to be applied, the distribution F_0 must be completely specified before we collect the data. In testing for normality, we have to choose the mean and the variance based on the data. This means that we have chosen a normal distribution which is a closer to the data than the true F so that M is too small. We must adjust the critical value to adjust for this as we do in χ^2 goodness of fit tests. Lilliefors has investigated the adjustment of p-values necessary to have a correct test for this situation and shown that the test is more powerful than the χ^2 goodness of fit test for normality.

Other tests of this kind for testing $F = F_0$ are the Anderson-Darling test based on the statistic

$$n \int_{-\infty}^{\infty} \left[\widehat{F}_n(x) - F_0(x)\right]^2 \left[F_0(x)(1 - F_0(x))\right]^{-1} dF_0(x)$$

and the Cramer-von Mises test based on the statistic

$$\int_{-\infty}^{\infty} \left(\widehat{F}_n(x) - F_0(x)\right)^2 dF_0(x).$$

In addition, the Shapiro-Wilk test is specifically designed to test for normality.

Kullback-Liebler Distance

Kolmogorov-Smirnov test can be used as a model selection test. However, if there are more than one distributions to choose from, a better measure of closeness between $\widehat{F}_n(x)$ and $F_0(x)$ is given by Kullback-Liebler distance, also called the *relative entropy*.

The Kullback-Liebler (K-L) distance between two distribution functions F and G with corresponding probability density functions $f(.)$ and $g(.)$ respectively is given by

$$KL(f,g) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} d(x).$$

Note that $KL(f,g) \geq 0$ and the equality holds for $f(.) = g(.)$. The K-L distance based on a sample of size n reduces to

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i)}{g(x_i)}.$$

For computing the distance between the empirical distribution function and the specified distribution function F_0, one can use the estimate of the density. Alternatively, for specific F_0 such as Normal, Uniform, Exponential etc, the entropy estimates are available in the literature. In addition, relative entropy based goodness-of-fit tests are also discussed in the literature.

Broadly, to select the distribution which fits best, it is preferable to screen the possible distributions using the Q-Q(P-P) plots and the K-S goodness-of-fit tests and then select one of the screened distributions based on the K-L distance.

4.2 Two-sample Kolmogorov-Smirnov Test

Alternatively, one may be interested in verifying whether two independent samples come from identically distributed populations.

Suppose we have two samples $X_1, ..., X_m$ and $Y_1, ..., Y_n$ from continuous distribution functions $F(x)$ and $G(y)$. We want to test the null hypothesis that $F(x) = G(x)$ for all x against the alternative that $F(x) \neq G(x)$ for some x. Let $\widehat{F}_m(x)$ and $\widehat{G}_n(y)$ be the empirical distribution functions for the $x's$ and $y's$. The two sample *Kolmogorov-Smirnov* (K-S) test is based on the statistic

$$M = \max_x \left| \widehat{F}_m(x) - \widehat{G}_n(x) \right|$$

We reject the null hypothesis if M is too large. As in the one sample case, if m and n are large,

$$P(dM > q) \overset{\bullet}{=} 2 \sum_{i=1}^{\infty} (-1)^{i-1} \exp\left(-2i^2 q^2\right) \overset{\bullet}{=} 2 \exp\left(-2q^2\right)$$

(where $d = 1/\left(\frac{1}{m} + \frac{1}{n}\right)$) so that critical values may be determined easily.

5 Nonparametric Tests and Confidence Intervals

The nonparametric tests described here are often called distribution free procedures because their significance levels do not depend on the underlying model assumption i.e., they are level robust. They are also power robust and robust against outliers.

We will mainly discuss the so-called **rank procedures**. In these procedures, the observations are jointly ranked in some fashion. In using these procedures, it is occasionally important that the small ranks go with small observations, though often it does not matter which order we rank in. The models for these procedures are typically semiparametric models.

One advantage of using ranks instead of the original observations is that the ranks are not affected by monotone transformations. Hence there is no need of transforming the observations before doing a rank procedure. Another advantage of replacing the observations with the ranks is that the more extreme observations are pulled in closer to the other observations.

As a consequence, a disadvantage is that nearby observations are spread out.

For example

Obs	1	1.05	1.10	2	3	100	1,000,00
Rank	1	2	3	4	5	6	7

The main reason we continue to study these rank procedures is the power of the procedures. Suppose our statistics depend on the ranks only through the sum of ranks (Wilcoxon signed rank statistic, Mann-Whitney-Wilcoxon rank sum statistic, Kruskal-Wallis statistic discussed below) and the sample size is moderately large. If the observations are really normally distributed, then the rank procedures are nearly as powerful as the parametric ones (which are the best for normal data). In fact it can be shown that Pitman asymptotic relative efficiency (ARE) of the rank procedure to the parametric procedure based on means is

$$3/\pi = .95$$

and in fact the ARE is always greater than 0.864. However the ARE can be artibrarily large for some non-normal distributions. What this means is the rank procedure is never much worse that parametric procedure, but can be much better.

Ties:

We assume that the underlined probability distribution is continuous for the rank procedures and hence, theoretically, there are no ties in the sample. However, the samples often have ties in practice and procedures have been developed for dealing with these ties. Typically average ranks are used for tied observations. There are other approaches to resolving ties; refer to the book by Higgins for details.

5.1 Single Sample Procedures

We introduce the concept of *location parameter* first.

A population is said to be located at μ_0 if the population median is μ_0.

Suppose $X_1, \cdots, X_n$ is a sample from the population. We say that $X_1, \cdots, X_n$ is located at μ if $X_1 - \mu, \cdots, X_n - \mu$ is located at 0.

Thus any statistic

$$S(\mu) = S(X_1 - \mu, \cdots, X_n - \mu)$$

is useful for the location analysis if $E[S(\mu_0)] = 0$ when the population is located at μ_0. This simple fact leads to some test procedures to test the hypothesis of population locations.

Sign Test

This is one of the oldest nonparametric procedures where the data are converted to a series of plus and minus signs. Let $S(\mu)$ be the *sign statistic* defined by

$$\begin{aligned} S(\mu) &= \textstyle\sum_{i=1}^{n} sign(X_i - \mu) \\ &= \#[X_i > \mu] - \#[X_i < \mu] \\ &= S^+(\mu) - S^-(\mu) \\ &= 2S^+(\mu) - n \end{aligned}$$

To find a $\hat{\mu}$ such that $S(\hat{\mu}) = 0$, we get $\hat{\mu} = \text{median}(X_i)$. Thus if μ_0 is the median of the population, we expect $E[S(\mu_0)] = 0$.

Suppose we wish to test the hypothesis that the population median is μ_0 giving

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0.$$

Based on $S(\mu_0)$, the proposed decision rule is:

$$\text{Reject } H_0 \text{ if } |S(\mu_0)| = |2S^+(\mu_0) - n| \geq c$$

where c is chosen such that

$$P_{\mu_0}[|2S^+(\mu_0) - n| \geq c] = \alpha.$$

It is easy to see that under $H_0 : \mu = \mu_0$, the distribution of $S^+(\mu_0)$ is Binomial$\big(n,\ \frac{1}{2}\big)$ irrespective of the underlying distribution of X_i's and hence c can be chosen appropriately. Equivalently, we reject H_0 if

$$S^+(\mu_0) \leq k \quad \text{or} \quad S^+(\mu_0) \geq n - k$$

where

$$P_{\mu_0}[S^+(\mu_0) \leq k] = \frac{\alpha}{2}.$$

This fact can be used to construct a confidence interval for the population median μ. Consider

$$P_d[k < S^+(d) < n - k] = 1 - \alpha$$

and find the smallest d such that [the number of $X_i > d$] $< n - k$. Suppose we get

$$\begin{aligned} d = X_{(k)} &\ :\ \#[X_i > X(k)] = n - k \\ d_{min} = X_{(k+1)} &\ :\ \#[X_i > X(k+1)] = n - k - 1. \end{aligned}$$

On the same lines, we find $d_{max} = X_{(n-k)}$. Then a $(1 - \alpha)100\%$ distribution-free confidence interval for μ is given by $[X_{(k+1)},\ X_{(n-k)}]$. From the normal approximation with continuity correction $k \simeq n/2 - n^{1/2} z_{\alpha/2}/2 - .5$ and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

Note that the median is a robust measure of location and does not get affected by the outliers. The sign test is also robust and insensitive to the outliers and hence the confidence interval is robust too.

Wilcoxon Signed Rank test

The sign test above utilizes only the signs of the differences between the observed values and the hypothesized median. We can use the signs as well as the ranks of the differences, which leads to an alternative procedure.

Suppose $X_1, \cdots, X_n$ is a random sample from an unknown population with median μ. We assume that the population is symmetric around μ. The hypothesis to be tested is $\mu = \mu_0$ against the alternative that $\mu \neq \mu_0$.

We define $Y_i = X_i - \mu_0$ and first rank the absolute values of $|Y_i|$. Let R_i be the rank of the absolute value of Y_i corresponding to the i^{th} observation, $i = 1, \cdots, n$. The signed rank of an observation is the rank of the observation times the sign of the corresponding Y_i.

Let

$$S_i = \begin{cases} 1 & \text{if } (X_i - \mu_0) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

By arguments similar to the one mentioned for earlier test, we can construct a test using the statistic

$$WS = \sum_{i=1}^{n} S_i R_i.$$

WS is called the *Wilcoxon signed rank statistic.*

Note that WS is the sum of ranks with positive signs of Y_i, i.e. the positive signed ranks. If H_0 is true, the probability of observing a positive difference $Y_i = X_i - \mu_0$ of a given magnitude is equal to the probability of observing a negative difference of the same magnitude. Hence, under the null hypothesis the sum of the positive signed ranks is expected to have the same value as that of the negative signed ranks. Thus a large or a small value of WS indicates a departure from the null hypothesis and we reject the null hypothesis if WS is too large or too small.

The critical values of the Wilcoxon Signed Rank test statistic are tabulated for various sample sizes. The tables of exact distribution of WS based on permutations is given in Higgins(2004).

Normal approximation

It can be shown that for large sample, the null distribution of WS is approximately normal with mean μ and variance σ^2 where

$$\mu = \frac{n\,(n+1)}{4}, \ \sigma^2 = \frac{n\,(n+1)\,(2n+1)}{24}$$

and the Normal cut-off points can be used for large values of n.

Hodges-Lehmann confidence Interval for μ

We can construct a $(1-\alpha)100\%$ confidence interval for population median μ using Wilcoxon Signed rank statistic, under the assumption that the underlined population is symmetric around μ.

Let

$$W_{ij} = \frac{X_i + X_j}{2}, \quad n \geq i \geq j \geq 1.$$

be the average of the i^{th} and j^{th} original observations, called a *Walsh average.*

For example, consider a single sample with 5 observations $X_1, \cdots, X_5$ given by $-3, 1, 4, 6, 8$. Then the Walsh averages are

	−3	1	4	6	8
−3	−3	−1	.5	1.5	2.5
1		1	2.5	3.5	4.5
4			4	5	6
6				6	7
8					8

We order the W_{ij} according to their magnitude and let $U_{[i]}$ be the i^{th} largest W_{ij}.

The median of W_{ij}'s provides a point estimation of the population median μ. This median of Walsh averages is known as the *Hodges-Lehmann* estimator of the population median μ.

For instance, in the data set above, the Hodges-Lehman estimator $\hat{\mu}$ is the 8^{th} largest Walsh average, namely $\hat{\mu} = 3.5$ whereas the parametric estimate of μ is $\bar{X} = 3.2$.

Using the Walsh averages, it is easy to see that another representation for the Wilcoxon Signed Rank statistic is

$$WS = \#\,[W_{ij} > 0]$$

(Note that this definition gives $WS = 13$ for the example.)

Now suppose that we do not know μ. Define

$$WS\,(\mu) = \#\,[W_{ij} > \mu]$$

Then the general distribution of $WS\,(\mu)$ is the same as null distribution WS statistic.

Suppose that a size $1-\alpha$ two-sided Wilcoxon Signed Rank test for $\mu = 0$ accepts the null hypothesis if

$$a \le WS < b,$$

where a and b depend on α. Then a $(1-\alpha)100\%$ confidence interval for μ is

$$a \le WS\,(\mu) < b \quad \Leftrightarrow \quad U_{[a]} < \mu \le U_{[b]}$$

This confidence interval is called the *Hodges-Lehmann confidence interval* for θ

For the data above, it can be seen from the table values that the acceptance region for a $\alpha = .125$ test is

$$2 \le WS < 14$$

so that

$$U_{[2]} < \mu \le U_{[14]} \quad \Leftrightarrow \quad -1 < \mu \le 7$$

is a 87.5% confidence interval for μ. Note that the assumed continuity implies that the inequality can be replaced by an equality in the last formula (but not the one before it) or vice versa. Note the similarity with the sign statistic. The sign statistic counts positive observations and the Wilcoxon statistic counts positive pairwise averages.

In general, if $P(WS \le k) = \alpha/2$ then $[U_{[k+1]}, U_{[n(n+1)/2-k]})$ is a $(1-\alpha)100\%$ confidence interval for μ. Using the normal approximation with continuity correction $k \simeq n(n+1)/2 - z_{\alpha/2}[n(n+1)(2n+1)/24]^{1/2} - .5$ and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

Note that the H-L interval is associated with the Wilcoxon signed rank test in that the two-sided Wilcoxon test rejects $\mu = 0$ iff 0 is not in the confidence interval. Also note that there is no problem with ties in either the H-L confidence interval or H-L estimator.

5.2 Two Sample Procedures

Suppose we observe two independent random samples $X_1, ..., X_n$ from distribution function $F(x)$, and $Y_1, ..., Y_m$ from distribution $G(y)$ where both F and G are continuous distributions.

We discuss the nonparametric procedures for making inference about the difference between the two location parameters of F and G here. In particular, we make the assumption that the distribution functions of the two populations differ only with respect to the location parameter, if they differ at all. This can alternatively be stated by expressing $G(y) = F(y + \delta)$ where δ is the difference between the medians.

There is no assumption of symmetry in the two sample model. The continuity of the distributions implies there will be no ties. In practice when ties occur in the data the average rank is typically used.

Wilcoxon rank sum statistic

Consider testing $\delta = 0$ against $\delta \neq 0$. We first combine and jointly rank all the observations. Let R_i and S_j be the ranks associated with X_i and Y_j. Then we could compute a two-sample t based on these ranks. However, an equivalent test is based on

$$H = \sum_{i=1}^{n} R_i$$

Note that if $\delta > 0$, then the $X_i's$ should be greater than the $Y_j's$, hence the $R_i's$ should be large and hence H should be large. A similar motivation works when $\delta < 0$. Thus we reject the null hypothesis $H_0 : \delta = 0$ if H is too large or too small. This test is called the *Wilcoxon rank-sum test.*

Tables of exact distribution of H are available in Higgins (p 340).

For example, suppose we have two independent random samples of size 4 and 3. Suppose further that we observe 37, 49, 55, 57 in the first sample and 23, 31, 46 in the second. We get

obs	37	49	55	57	23	31	46
rank	3	5	6	7	1	2	4

Therefore, for the observed data

$$H = 21$$

Again we reject if the observed H is one of the two largest or two smallest values. Based on the exact permutation distribution, we reject the null hypothesis as the p-value is $2 \times 2/35 = .101$.

Normal approximation

It can be shown that for large sample the null distribution of H is approximately normal with mean μ and variance σ^2 where

$$\mu = \frac{m\,(m+n+1)}{2},\ \sigma^2 = \frac{mn\,(m+n+1)}{12}$$

Suppose, as above,we compute $H = 21$ based on a samples of size 4 and 3. In this case $\mu = 16$, $\sigma^2 = 8$, so the approximate p-value is (using a continuity correction)

$$2P\,(H \geq 21) = 2P\,(H \geq 20.5) =$$

$$2P\left(\frac{H-16}{\sqrt{8}} \geq \frac{20.5-16}{\sqrt{8}}\right) = 2P\,(Z \geq 1.59) = .11$$

which is close to the true p-value derived above even for this small sample size.

Mann-Whitney test

Let

$$V_{ij} = X_i - Y_j,$$

We define

$$U = \#\,(V_{ij} > 0)$$

which is the *Mann-Whitney* statistic. The Mann-Whitney test rejects the null hypothesis $H_0 : \delta = 0$ if U is too large or too small.

For our example we see that

	23	31	46
37	14	6	−9
49	26	18	3
55	32	24	9
57	34	26	11

Therefore, for this data set $U = 11$.

It can be shown that there is a relationship between the Wilcoxon rank sum H and the Mann-Whitney U :

$$H = U + \frac{n\,(n+1)}{2}.$$

Hence the critical values and p-values for U can be determined from those for H.

The Hodges-Lehmann confidence interval for δ

Analogous to the single sample procedure, we can construct a $(1-\alpha)100\%$ confidence interval for δ using the Mann-Whitney procedure.

We order V_{ij} according to their magnitude and let $V_{[i]}$ be the i^{th} largest V_{ij}. Then the *Hodges Lehmann estimator* for δ is the median of the V_{ij}.

Let

$$U(\delta) = \#\left[V_{ij} > \delta\right].$$

Then the general distribution of $U(\delta)$ is the same as the null distribution of U. Suppose that two-sided size α test the $\delta = 0$ against $\delta \neq 0$ accepts the null hypothesis if

$$a \leq U < b$$

Then a $(1-\alpha)100\%$ confidence region for δ is given by

$$a \leq U(\delta) < b \quad \Leftrightarrow \quad V_{[a]} < \delta \leq V_{[b]}$$

which is the Hodges-Lehmann confidence interval for δ. In our example the estimator is the average of the 6th and 7th largest of the V_{ij}, giving

$$\widehat{\delta} = 16$$

The parametric estimator is $\overline{X} - \overline{Y} = 16.2$.

To find the confidence interval, note that $H = U + 10$

$$.89 = P(12 \leq H < 21) = P(2 \leq U < 11)$$

Therefore the 89% Hodges-Lehmann confidence interval for δ is

$$V_{[2]} \leq \delta < V_{[11]} \Leftrightarrow 3 \leq \delta < 32$$

The classical (t) confidence interval for the data based on t-statistics is $1.12 < \delta \leq 31.22$.

In general, if $P(U \leq k) = \alpha/2$, then $[V_{[k]}, V_{[mn-k]})$ is a $(1-\alpha)100\%$ confidence interval for δ. Using the normal approximation with continuity correction $k \simeq mn/2 - z_{\alpha/2}[mn(m+n+1)/12]^{1/2} - .5$ where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

Paired data

Analogous to the paired t-test in parametric inference, we can propose a nonparametric test of hypothesis that the median of the population of differences between pairs of observations is zero.

Suppose we observe a sequence of i.i.d. paired observations $(X_1, Y_1), ..., (X_n, Y_n)$. Let μ_D be the median of the population of differences between the pairs. The goal is to draw inference about μ_D. Let

$$D_i = X_i - Y_i$$

The distribution of D_i is symmetric about μ_D. Therefore, we may used the procedures discussed earlier for the one-sample model, based on the observations D_i.

5.3 k-Sample Procedure

Suppose we wish to test the hypothesis that the k samples are drawn from the populations with equal location parameters. The Mann-Witney-Wilcoxon procedure discussed above can be generalized to k independent samples. The test procedure we consider is the *Kruskal-Wallis Test* which is the nonparametric analogue of the parametric one-way analysis of variance procedure.

Suppose we have k independent random samples of sizes $n_i, i = 1, \cdots, k$ each, represented by $X_{ij}, j = 1, \cdots, n_i;\ i = 1, \cdots, k$. Let the underlined location parameters be denoted by $\mu_i, i = 1, \cdots, k$. The null hypothesis to test is that the μ_i are all equal against the alternative that at least one pair $\mu_i,\ \mu_{i*}$ is different.

For the Kruskal Wallis test procedure, we combine the k samples and rank the observations. Let R_{ij} be the rank associated with X_{ij} and let $\overline{R}_{i.}$ be the average of the ranks in the i^{th} sample. If the null hypothesis is true, the distribution of ranks over different samples will be random and no sample will get a concentration of large or small ranks. Thus under the null hypothesis, the average of ranks in each sample will be close to the average of ranks for under the null hypothesis.

The Kruskal-Wallis test statistic is given by

$$KW = \frac{12}{N(N+1)} \sum n_i \left(\overline{R}_{i.} - \frac{N+1}{2} \right)^2$$

If the null hypothesis is not true, the test statistic KW is expected to be large and hence we reject the null hypothesis of equal locations for large values of KW.

The tables of exact critical values are available in the literature. We generally use a χ^2 distribution with $k-1$ degrees of freedom as an approximate sampling distribution for the statistic.

6 Permutation tests

The parametric test statistics can also be used to carry out the nonparametric test procedures. The parametric assumptions determine the distribution of the test statistic and hence the cut-off values under the null hypothesis. Instead, we use permutation tests to determine the cutoff points.

We give an example below.

Consider a two sample problem with 4 observations X_1, X_2, X_3, X_4 in the first sample from cdf $F(x)$and 3 observations Y_1, Y_2, Y_3 in the second sample from cdf $G(y)$. We want to test the null hypothesis $F(x) = G(x)$ against the alternative hypothesis $F(x) \neq G(x)$

Suppose we observe 37, 49, 55, 57 in the first sample and 23, 31, 46 in the second (Section 5.2). Suppose we want a test with size .10.

1. The parametric test for this situation is the two-sample t-test which rejects if
$$|T| = \left| \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{4} + \frac{1}{3}}} \right| > t_{5,.05} = 2.015$$
For this data set, $T = 2.08$ so we reject (barely). The p-value for these data is .092. Note that this analysis depends on the assumptions that the data are normally distributed with equal variances.

2. We now look at rearrangements of the data observed. One possible rearrangement is 31, 37 46, 55 in the first sample and 23, 49, 57 in the second. For each rearrangement, we compute the value of the T. Note that there are
$$\binom{7}{4} = 35$$
such rearrangements. Under the null hypothesis (that all 7 observations come from the same distribution) all 35 rearrangements are equally likely, each with probability 1/35. With the permutation test, we reject if the value of T for the original data is one of the 2 largest or 2 smallest. This test has $\alpha = 4/35 = .11$.

3. If we do this to the data above, we see that the original data gives the second largest value for T. (Only the rearrangement 46, 49, 55, 57 and 23, 31, 37 gives a higher T.) Therefore we reject the null hypothesis. The p-value is $2 \times 2/35 = .11$. Note that the only assumption necessary for these calculations to be valid is that under the null hypothesis the two distributions be the same (so that each rearrangement is equally likely). That is, the assumptions are much lower for this nonparametric computation.

These permutation computations are only practical for small data sets. For the two sample model with m and n observations in the samples, there are

$$\left(\begin{array}{c} m+n \\ m \end{array} \right) = \left(\begin{array}{c} m+n \\ n \end{array} \right)$$

possible rearrangements. For example

$$\left(\begin{array}{c} 20 \\ 10 \end{array} \right) = 184,756$$

so that if we had two samples of size 10, we would need to compute V for a total of 184,756 rearrangements. A recent suggestion is that we don't look at all rearrangements, but rather look a randomly chosen subset of them and estimate critical values and p-values from the sample.

What most people who use these tests would do in practice is use the t-test for large samples, where the t-test is asymptoticall nonparametric and use the permutation calculation in small samples where the test is much more sensitive to assumptions.

Note that the distributions of the rank statistics are permutation distributions and their p-values can be determined via their permutaton distributions also.

7 Correlation coefficients

Pearson's r

The parametric analysis assumes that we have a set of i.i.d. two-dimensional vectors, $(X_1, Y_1), ..., (X_n, Y_n)$ which are normally distributed with correlation coefficient

$$\rho = \frac{cov\,(X_i, Y_i)}{\sqrt{var\,(X_i)\,var\,(Y_i)}}.$$

ρ is estimated by the sample correlation coefficient (Pearson's r)

$$r = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum \left(X_i - \overline{X}\right)^2 \sum \left(Y_i - \overline{Y}\right)^2}}$$

The null hypothesis $\rho = 0$ is tested with the test statistic

$$t = \sqrt{\frac{n-2}{1-r^2}}r \sim t_{n-2}$$

under the null hypothesis.

To make this test more robust, we can use a permutation test to get nonparametric critical values and p-values. To do the rearrangements for this test, we fix the $X's$ and permute the $Y's$.

Some Semiparametric correlation coefficients

A semiparametric model alternative for the normal correlation model above is to assume that the $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d. from a continuous bivariate distribution, implying no ties.

Spearman's rank correlation

We rank the X's and Y's separately getting ranks R_i and S_i. The sample correlation coefficient between the R_i and S_i is called *Spearman's rank correlation.* Suppose, for example the we observe

x	1	3	6	9	15
r	1	2	3	4	5
y	1	9	36	81	225
s	1	2	3	4	5

Then the rank correlation r_S is obviously one. Note that this happens because $Y = X^2$. Since Y is not a linear function of X, the correlation coefficient is less than 1. In fact the correlation coefficient is .967.

We often want to test that X and Y are independent. We reject if r_S is too large or too small. We determine the critical values and p-values from the permutation test as described above. For reasonably large sample sizes, it can be shown that under the null hypothesis

$$r_S \overset{\bullet}{\sim} N\left(0, \frac{1}{n-1}\right)$$

Kendall's coefficient of concordance

We say two of the vectors (X_i, Y_i) and (X_{i*}, Y_{i*}) are concordant if

$$(X_i - Y_i)(X_{i*} - Y_{i*}) > 0$$

Kendall's τ is defined by

$$\tau = 2P\left[(X_i - Y_i)(X_{i*} - Y_{i*}) > 0\right) - 1$$

We estimate Kendall's τ by

$$r_K = 2\frac{\#\left(concordant\ pairs\right)}{\left(\begin{array}{c} n \\ 2 \end{array}\right)} - 1$$

To test $\tau = 0$, we would use r_K. One and two sided (exact) critical values can be determined from permutation arguments. Approximate critical value and p-values can be determined from the fact that for reasonably large n, the null distribution is

$$r_K \overset{\bullet}{\sim} N\left(0, \frac{4n+10}{9\left(n^2 - n\right)}\right).$$

8 Nonparametric Regression

Suppose we have n observations $(Y_1, X_1), \cdots, (Y_n, X_n)$ on (Y, X) where Y is the response variable and X is the predictor variable and the aim is to model Y as a function of X.

The most widely used statistical procedure for such a problem is *linear regression model* where we assume that $E[Y|X = x]$ is a linear function of X, specified by

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \cdots, n,$$

and the errors ϵ_i are taken to be uncorrelated with zero mean and variance σ^2. When not appropriate, fitting a linear regression model to a nonlinear relationship may result in a misleading and unreliable inference.

Nonparametric regression is a more general alternative to this set up when the functional form of $E[Y|X = x]$ can not be assumed. In particular, the model considered is

$$Y_i = m(X_i) + \epsilon_i$$

where the regression curve $m(x)$ is the conditional expectation $m(x) = E[Y|X = x]$ with $E[\epsilon|X = x] = 0$ and $\mathrm{Var}[\epsilon|X = x] = \sigma^2(x)$. The model removes the parametric restrictions on $m(x)$ and allows the data to dictate the alternative structure of $m(x)$ by using the data based estimate of $m(x)$. Note that $\sigma^2(x)$ now depends on x as well.

The statistical procedures which estimate the regression curve using the information available in the neighborhood are called *smoothing techniques*. Different smoothing techniques lead to different nonparametric regression estimators.

8.1 Kernel Estimator

We have

$$\begin{array}{rcl} m(x) & = & E[Y|X = x] \\ & = & \int y \frac{f(x,y)}{f(x)} dy \end{array}$$

where $f(x)$ and $f(x, y)$ are the marginal density of X and the joint density of X and Y respectively. On substituting the univariate and bivariate kernel density estimates of the two densities and noting the properties of kernel function $K(.)$ specified in Section 3, we get

$$\begin{array}{rcl} \hat{m}_{NW}(x) & = & \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} \\ \\ & \equiv & \sum_{i=1}^n W_{hi}(x) Y_i \end{array}$$

which is a weighted average of the response variables in a fixed neighborhood around x with weights

$$W_{hi}(x) = (nh)^{-1} \frac{K\left(\frac{x-X_i}{h}\right)}{\hat{f}(x)}.$$

$\hat{m}_{NW}(x)$ is called the *Nadaraya-Watson kernel estimator*. Note that

- The weights depend on the kernel function $K(.)$, the bandwidth h and the whole sample $\{X_i, i = 1, \cdots, n\}$ through the kernel density estimate $\hat{f}(x)$.
- For the uniform kernel, the estimate of $m(x) = E[Y|X = x]$ is the average of Y_j's corresponding to the X_j's in the h-neighborhood of x.

- Observations Y_i obtain more weight in those areas where the corresponding X_i are sparse.
- When the denominator is zero, the numerator is also equal to zero and the estimate is set to be zero.
- Analogous to kernel density estimation, the bandwidth h determines the level of smoothness of the estimate and is called the smoothing parameter. Decreasing bandwidth leads to a less smooth estimate. In particular, for $h \to 0$ the estimate $\hat{m}(X_i)$ converges to Y_i and for $h \to \infty$ the estimate converges to $\bar{Y}$. The criteria of bandwidth selection and guidelines for selecting the optimal bandwidth are available in the literature.

In case the predictors $X_i,\ i = 1, \cdots, n$ are not random, alternative estimators such as *Gasser-Müller kernel estimator* are more appropriate.

It can be shown that the Nadaraya-Watson kernel estimator is the solution of the weighted least squares estimator obtained on minimizing

$$\sum_{i=1}^{n} (Y_i - \beta_0)^2 K\left(\frac{x - X_i}{h}\right)$$

over β_0. This corresponds to locally approximating $m(x)$ with a constant while giving higher weights to the Y_j's corresponding to the X_j's in the h-neighborhood of x.

This concept is further generalized to fitting higher order polynomials 'locally', i.e. in the neighborhood of x. In particular, we consider minimizing

$$\sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1(x - X_i) - \beta_2(x - X_i)^2 - \cdots - \beta_p(x - X_i)^p]^2 K\left(\frac{x - X_i}{h}\right)$$

over $\beta_0, \beta_1, \cdots, \beta_p$. The resulting estimator is called the *local polynomial regression estimator* and the appropriate choice of the degree of polynomial p can be made based on the data.

8.2 k-Nearest Neighbor Estimator

The k-Nearest Neighbor or $k - NN$ estimator of $m(x)$ is also a weighted average of response variables in the neighborhood of x. However, unlike the kernel estimator with a fixed h-neighborhood, we consider a varying neighborhood around x here which is defined through the k X_j's which are closest to x.

In particular, for every x, we define the set of indexes

$$J_x = \{i \ : \ X_i \text{ is one of the } k \text{ nearest observations to } x\}$$

and construct the weight sequence $\{W_{ki}(x), \ i = 1, \cdots, n\}$ given by

$$W_{ki}(x) = \begin{cases} \frac{n}{k} & \text{if } \ i \in J_x \\ 0 & \text{otherwise} \end{cases}$$

Then the $k - NN$ estimator of $m(x)$ is defined as

$$\hat{m}_k(x) = n^{-1} \sum_{i=1}^{n} W_{ki}(x) Y_i.$$

k is the smoothing parameter here as it controls the degree of smoothness of the estimated curve. For $k = n$ the neighborhood covers the entire sample for each x, giving $\hat{m}_{ni}(x) = \bar{Y}$. On the other hand, $k = 1$ gives a step function which is equal to Y_i for $x = X_i$ and jumps in the middle between two adjacent values of X. Variations of the $k - NN$ estimator using different weight sequences are also proposed in the literature.

8.3 LOESS Estimator

LOESS (also called LOWESS) stands for a LOcally Estimated Scatter plot Smoothing technique which combines the two smoothing techniques discussed above and is more flexible and robust. It initially selects varying bandwidth based on the nearest neighbors and iteratively uses the polynomial weighted least squares fit in each neighborhood. The polynomials considered are either linear or quadratic and the weights given to the response variables corresponding to X_i's in the neighborhood of x are determined by the choice of the kernel function. LOESS can not be expressed in a closed form and estimating it is a computer-intensive technique.

9 References

1. Arnold, Steven (1990), *Mathematical Statistics* (Chapter 17). Prentice Hall, Englewood Cliffs, N. J

2. Beers, Flynn, and Gebhardt (1990), Measures of Location and Scale for Velocities in Cluster of Galaxies-A Robust Approach. *Astron Jr*, 100, 32-46.

3. Härdle, W. (1990) *Applied Nonparametric Regression.* Cambridge Univ Press, Cambridge.

4. Hettmansperger, T and McKean, J (1998), *Robust nonparametric Statistical Methods.* Arnold, London.

5. Higgins, James (2004), *Introduction to Modern Nonparametric Statistics.* Duxbury Press.

6. Hollander and Wolfe, (1999), *Nonparametric Statistical Methods* John Wiley, N.Y.

7. Johnson, Morrell, and Schick (1992), Two-Sample Nonparametric Estimation and Confidence Intervals Under Truncation, *Biometrics*, 48, 1043-1056.

8. Summer School in Statistics for Astronomers(2005). Lecture Notes by Steven Arnold Website: http://astrostatistics.psu.edu/

9. Summer School in Statistics for Astronomers(2008). Lecture Notes by Tom Hettmansperger. Nonparametrics Statistics. Website: http://astrostatistics.psu.edu/su08/program.html

10. Website: http://www.stat.wmich.edu/slab/RGLM/

* Part of these notes borrow heavily from the material in 8 and 9

Nonparametric statistics is a method that makes statistical inference without regard to any underlying distribution. The method fits a normal distribution under no assumptions. Habitually, the approach uses data that is often ordinalOrdinal DataIn statistics, ordinal data are the type of data in which the values follow a natural order. One of the most notable features of ordinal data is that because it relies on rankings rather than on numbers. Nonparametric statistics can be contrasted with parametric statistics. Nonparametric statistics includes nonparametric descriptive statistics, statistical models, inference, and statistical tests. The model structure of nonparametric models is not specified a priori but is instead determined from data. The term nonparametric is not meant to imply that such models completely lack parameters, but rather that the number and nature of the parameters are flexible and not fixed in advance. A histogram is an example of a nonparametric estimate of a probability distribution. Key Takeaways. or nonparametric. Parametric statistics are more com-. monly used in the social sciences, and they include. such widely recognizedtests as the Studentâ€™s t test and. the analysis of variance (ANOVA). Researchers using.Â application of nonparametric statistics. We begin with. an explanation of how and under what circumstances. nonparametric statistics can be meaningfully applied, thendescribe speciï¬c nonparametrctechniquesthatcan. be used to answer certain types of research questions. Nonparametric statistics may be defined as statistical methods that contribute valid testing and estimation procedures under less stringent assumptions than the classical parametric statistics. However, there is no general agreement in the literature regarding the exact specification of the term â€nonparametric statistics.â€ In the past, nonparametric statistics and the term â€distribution-free methodsâ€ were commonly used interchangeably in the literature, although they have different implications.