1 Nonparametric Statistics

By Steven Arnold
Professor of Statistics-Penn State University

Some good reference for the topics in this course are

1. Higgins, James, (2004) **Introduction to Nonparametric Statistics**
2. Arnold,Steve (1990) **Mathematical Statistics,** *Chapter 17*
3. Hettmansperger, T and McKean, J (1998), **Robust nonparametric Statistical Methodology**

The first book we use for our undergrad course. It is written at about the same level as this course. The second book contains a basic outline of the theory most of which is not presented here. The third book has a detailed description of the theory. It is the source I use when I want to fully understand why a particular procedure is effective.

1.1 Parametric, nonparametric and semiparametric models

A *parametric statistical model* is a model whose joint distribution is dependent on several unknown constants called *parameters.* The only things unknown about 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 observed statistics to draw inference about the unobserved parameters.

All the other classes in this school are concerned with parametric statistics.

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

A *nonparametric model* is one in which the only assumptions made about the distribution of the observations is that they are independently identically distributed (i.i.d.) from an arbitrary continuous distribution. There are no parameters in a nonparametric model.

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

Both nonparametric and semiparametric models used to be (and often still are) lumped together and called nonparametric models. The distinction we have made goes back to Huber in the 1960's and is becoming more and more common.

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

Today's short course will primarily be concerned with nonparametric and semiparametric models.

1.2 Permutation tests

We can often do nonparametric tests with parametric statistics by using permutation tests. We give two examples.

1. Consider a two sample problem with 4 observations X_1, X_2, X_3, X_4 in the first sample from cdf $F\left(x\right)$and 3 observations Y_1, Y_2, Y_3 in the second sample from cdf $G\left(y\right)$ .We want to test the null hypothesis $F(x) = G\left(x\right)$ against the alternative hypothesis $F\left(x\right) \neq G\left(x\right)$ Suppose we observe 37, 49, 55, 57 in the first sample and 23, 31, 46 in the second. Suppose we want a test with size .10.

   (a) 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.

   (b) 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

   $$\left(\begin{array}{c} 7 \\ 4 \end{array} \right) = 35$$

   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 $\alpha = 4/35 = .11$ The p-value for the permutation test is twice the rank of the original data divided by 35.

(c) 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.

(d) For any rearrangement,

$$\sum X_i + \sum Y_i = 298$$

$$\sum X_i^2 + \sum Y_i^2 = 13,650$$

Let $V = \sum X_i$. Then using the facts in the previous equations, it can be shown that for any rearrangement, T is completely determined by V and is an increasing function of V. This means that to find the rearrangements which maximize T, we can find the rearrangements which maximize V, which is somewhat easier. Note that for the best arrangement, $V = 207$ and for the original arrangement $V = 198$

2. Now suppose we have a single sample with 5 observations X_1, X_2, X_3, X_4, X_5. We want to test the null hypothesis that the observations are centered at 0 against the alternative that they are not. We again want a .10 test. We observe -3, 1, 4, 6, 8.

(a) The parametric test for this problem is the one-sample t-test which rejects is

$$|T| = \left| \frac{\sqrt{5}\,(\overline{X})}{S} \right| > t_4^{.05} = 2.132$$

For these data, $T = 1.65$, so we accept the hypothesis that the observations are symmetric about 0. The p-value for these data is .17.

(b) To use a permutation version of this test, we first take the absolute values of all the observations, getting 1,3,4,6,8. Under the null hypothesis that the distribution is symmetric about 0, each of these 5 numbers is equally likely to be positive or negative. Therefore, a rearrangement of the data is to assign each observation to be positive or negative. For example, one such rearrangement is 1,-3,4,-6,8. We look at each of the $2^5 = 32$ rearrangements. For each rearrangement we compute T. If our observed T is one of the 2 largest or 2 smallest, we reject the hypothesis of symmetry about 0. The size α for this test is $4/32 = .125$. The p-value is the twice the rank of the observed T divided by 32. By a similar argument to that for the two-sample problem, we can look at $V = \sum X_i$, instead of T.

(c) Clearly the largest value for V is 22, when we use 1,3,4,6,8. The second largest, 20, occurs for -1,3,4,6,8. and the third largest, 16, occurs for the original observations 1,-3,4,6,8. Therefore we accept the hypothesis that the distribution is symmetric about 0. The p-value is $2 \times 3/32 = .19$.

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. Similarly, for the one-sample problem with n observations, there are 2^n rearrangements. For example

$$2^{20} = 1,048,576$$

so that in a one sample problem with 20 observations, we would need to compute V for 1,048,576 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 fairly robust and use the permutation calculation in small samples where the test is much more sensitive to assumptions.

1.3 Rank tests

1.3.1 Basic discussion

Most of this talk is concerned with so-called rank procedures. The models for these procedures are typically semiparametric models. In these procedures, we jointly rank the observations is some fashion. We take the procedures which we used for the associated parametric model and replace the observations with their ranks. In using these procedures, it is occasionally important that the small ranks go with small observations. Often it does not matter which order we rank in. We always rank with small ranks associated with small observations.

One advantage of using ranks instead of the original observations is that are not changed by monotone transformations. There is no reason to think about transforming the observations before doing a rank procedure.

One other advantage of replacing the observations with the ranks is that the more extreme observations are pulled in closer to the other observations. 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

Power When they were first developed, the rank procedures were called ”quick and dirty” procedures. However, this is completely inaccurate. In order to rank the observations, we have to first order them, something computers do very slowly. The parametric procedures can be computed in one pass through the data, but the ranks cannot. (So the rank procedures are ”slow”.) However as we shall discuss below, they procedures have some strongly positive statistical properties, so they are not ”dirty”.

Later the main motivation was that the size of the rank procedures does not depend on the normal assumption and so the procedures are more robust. (Note that they do depend on all the other assumptions.) However, because of the central limit theorem, the parametric procedures are also robust against the normal assumption.

The main reason we continue to study these rank procedures is power. Suppose 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 is

$$3/\pi = .95$$

and in fact the ARE is always greater than $3/\pi$. However the ARE is ∞ 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 The assumed continuity in the models below implies there are no ties. However, often there are ties in practice. Procedures have been developed for dealing with these ties. Higgins discusses these procedures. They are often rather complicated and not uniquely defined so we do not discuss them here.

1.3.2 The one sample model

In the semiparametric one sample model, we observe a sample from a distribution which is symmetric about an unknown parameter θ. We also assume that the distribution is continuous. We often write this model as one in which we observe X_i

$$X_i = \theta + e_i,\ i = 1, ..., n$$

where the e_i are i.i.d. from a continuous distribution which is symmetric about 0. Note that θ is the median of the distribution and is also the mean if it exists. The assumed continuity implies that there will be no ties and no 0's.

The Wilcoxon signed rank statistic Suppose we want to test the null hypothesis that $\theta = 0$ against the alternative that $\theta \neq 0$.

We first rank the absolute values of the observations getting R_i for the rank of the absolute value of the ith observation. The signed rank of an observation is the rank of the observation times the sign of the observation. We then could compute the one-sample $t - test$ statistic using the signed ranks instead of the observations.

Let

$$S_i = \begin{matrix} 1 & if\ X_i > 0 \\ 0 & otherwise \end{matrix}$$

By similar arguments to one mentioned in the last section, we can base the test on

$$Q = \sum S_i R_i.$$

We reject if Q is too large or too small. This is called the *Wilcoxon signed rank statistic.*

Motivation for shape of test: Suppose the $\theta > 0$. Then the number of positive observations should be greater than it would be if the center were 0. Also the ranks should be higher. Therefore Q should be larger when $\theta > 0$ than when $\theta = 0$. Similarly, when $\theta < 0$, then Q should be smaller than when $\theta = 0$.

To compute the exact critical points and p-values we use permutations test arguments as before.

Motivation for use of permutation test calculations to do computations under the null hypothesis: When $\theta = 0$, the distribution of the observations is symmetric about 0. This means that the probability of any rearrangement of signs should be the same.

Let $Q_{[i]}$ be the order statistics computed from the outcomes of Q from the different rearrangements.

Using the data from the last example -3, 1, 4, 6, 8, we get

data	-3	1	4	6	8
SR	0	1	3	4	5

so that the signed rank statistic is

$$Q = 13.$$

As we said before there are $2^5 = 32$ possible arrangements of the signs. Therefore as before, for a .125 we reject if the observed value for Q is one of the two largest or two smallest. Obviously, the rearrangement which gives the highest value for Q is one in which all 5 ranks are positive giving $Q_{[32]} = 15$, and the second highest is one in which 1 is the only rank which is negative giving $Q_{[31]} = 14$, and the third best is the actual data with $Q_{[30]} = 13$ so that these data are not significant. As before, the p-value for this data set is 2X3/32=.19

Tables of the permutation (exact) distribution of Q are given on p. 346 of Higgins.

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

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

Suppose, as above,we compute $Q = 13$ based on a sample of size 5. In this case $\mu = 7.5$, $\sigma^2 = 13.75$, so the approximate p-value is (using a continuity correction)

$$2P(Q \geq 13) = 2P(Q \geq 12.5) =$$

$$2P\left(\frac{Q - 7.5}{\sqrt{13.75}} \geq \frac{12.5 - 7.5}{\sqrt{13.75}}\right) = 2P(Z \geq 1.35) = .18$$

which is not far from the true p-value derived in the last section even for this small sample size.

Hodges-Lehmann confidence interval and estimator for θ Let

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

be the average of the ith and jth original observations, called a *Walsh average.* For the simple data set above, these 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

Let $W_{[i]}$ be the ith largest W_{ij}. Another representation for the Wilcoxon statistic is

$$Q = \#(W_{ij} \geq 0)$$

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

Now suppose that we do not know θ. Let

$$Q(\theta) = \#(W_{ij} \geq \theta)$$

Then the general distribution of $Q(\theta)$ is the same as null distribution Q.

Suppose that a size $1 - \alpha$ two-sided test Wilcoxon test that $\theta = 0$ accepts if

$$a \leq Q < b.$$

Then a $1 - \alpha$ confidence interval for θ is

$$a \leq Q(\theta) < b \Leftrightarrow U_{[a]} < \theta \leq U_{[b]}$$

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

For our data, we see that the acceptance region for a .125 test is

$$2 \leq Q < 14$$

so that

$$U_{[2]} < \theta \leq U_{[14]} \Leftrightarrow -1 < \theta \leq 7$$

is a .875 confidence interval for θ. Note 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.parametric interval is $-.543 \leq \theta \leq 6.94$

Note that the HL interval is associated with the Wilcoxon test in that the two-sided Wilcoxon test rejects $\theta = 0$ iff 0 is not in the confidence interval.

The Hodges-Lehmann estimator for θ is the median of the Walsh averages. In out data set it is the 8th largest Walsh average, namely

$$\widehat{\theta} = 3.5$$

Note that the parametric estimator is $\overline{X} = 3.2$. The HL estimator is associated with the HL confidence interval, but discussion of this concept is beyond the expectations of this course.

Note that there is no problem with ties in either the HL confidence interval or HL estimator.

1.3.3 The two-sample model

In this model we assume that we observe independent samples $X_1, ..., X_n$ from distribution function $F(x)$, and $Y_1, ..., Y_n$ from distribution $G(y) = F(y + \delta)$. The only additional assumption is the F (and hence G) is a continuous distribution. There is no symmetry assumption in the two sample model. The continuity of the distributions implies there will be no ties. This situation is often called a *shift* family. We could write it as

$$X_i = \delta + e_i, \; Y_j = f_j$$

where all the e_i and f_j are i.i.d. Note that δ is the difference between the means (if they exist) and the difference between the medians, but that the variances must be the same for the two populations.

The Wilcoxon rank sum statistic Consider testing that $\delta = 0$ against $\delta \neq 0$. We first 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

$$H = \sum R_i$$

We reject if H is too large or too small. This test is called the *Wilcoxon rank-sum test.* We compute critical values and p-values using permutation test calculations.

Motivation for shape: If $\delta > 0$, then the $X's$ should be greater than the $Y's$, hence the $R_i's$ should be large and hence H should be large. A similar motivation works when $\delta < 0$.

Motivation for using permutation calculations to determine critical values and p-values: Under the null hypothesis that $\delta = 0$, the $X's$ and $Y's$ are a big sample from F. Therefore every rearrangement of the ranks should have the same probability.

For the data set used in the first section we see that

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

Therefore, for the data

$$H = 21$$

Again we reject if the observed H is one of the two largest or two smallest values. We recall that there are a total of 35, so this has size 4/35=.11. Note that rearrangement with the largest value for H has ranks 4,5,6,7, so that $H_{[35]} = 22$.The second largest rearrangement is the observed data so that $H_{[34]} = 21$, and we reject the null hypothesis. The p-value is $2 \times 2/35 = .101$. Similarly

$$H_{[1]} = 10,\ H_{[2]} = 11,\ H_{[3]} = 12$$

Tables of the permutation (exact) distribution of H are given on p.340 of Higgins.

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)

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

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

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

The Mann-Whitney test Let

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

The Mann-Whitney test rejects if U is too large.

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 :

$$H = U + \frac{m\left(m+1\right)}{2}.$$

Therefore, critical values and p-values for U can be determined from those for H.

The Hodges-Lehmann confidence interval and estimator for δ The *Hodges Lehmann estimator* for δ is the median of the V_{ij}

Let

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

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

$$a \leq U < b$$

Then a $1 - \alpha$ confidence region for δ is

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

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 $H = U + 10$

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

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 is $1.12 < \delta \leq 31.22$.

1.3.4 Paired data

In this model we observe a sequence $(X_1, Y_1), ..., (X_n, Y_n)$ of i.i.d. random $2-$dimensional random vectors such that

$$(X_i, Y_i - \theta) \sim (Y_i - \theta, X_i)$$

The goal is to draw inference about θ. Let

$$D_i = X_i - Y_i$$

By the equation above, the distribution of D_i is symmetric about θ. Therefore, we may used the procedures discussed earlier for the one-sample model, which we do. (Note that is the same thing we do for paired data in parametric case. Take differences and use one-sample procedures.)

I suppose most practical applications of the signed rank test are to paired data. In fact, Higgins does not discuss the signed rank test in the one-sample chapter, but only in the chapter on paired models.

1.3.5 K-sample model

The F-test In the parametric version of the k-sample model. we observe X_{ij}, independent, where

$$X_{ij} \sim N\left(\mu_i, \sigma^2\right),\ i = 1, ...k;\ j = 1, ..., n_i,\ N = \sum n_i$$

We want to test whether the μ_i are equal. The test we often use is to reject when

$$F > F^{\alpha}_{k-1,N-k},\ F = \frac{MSTR}{MSE}$$

$$MSTR = \frac{\sum n_i \left(\overline{X}_{i.} - \overline{X}_{..}\right)^2}{k-1},\ MSE = \frac{\sum\sum \left(X_{ij} - \overline{X}_i\right)^2}{N-k}$$

If we want to make this procedure into a more robust one, we can find a new critical value by doing a permutation test on this model. The calculations are an obvious extension of those for the two-sample model.

Parametric multiple comparisons After rejecting with the F-test, we want to find out which of the cells have different means, leading to multiple comparisons procedures. We say that procedure controls the per comparison error rate if the probability of a particular false rejection is α and the procedure controls the experiment-wide error rate if the probability of at least 1 false rejection is α (across the whole experiment.) Early work on multiple comparisons focused on controlling the per comparison error rate, but in the last 30-40 years the emphasis is on controlling the experiment-wide error rate.

Fisher's least significant difference (LSD) only controls the per comparison error rate and so most statisticians don't use it. Tukey's honest significant difference (HSD) controls the experiment-wide error rate and is typically recommended. This procedure says the ith and i*th mean are significantly different if

$$\left|\overline{X}_i - \overline{X}_{i*}\right| > \frac{q^{\alpha}_{k,N-k}}{\sqrt{2}}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_{i*}}\right)}$$

where $q^{\alpha}_{k,N-k}$ is the upper α critical point for a studentized range distribution. Tables of these critical point are in most analysis of variance books. One is also in Higgins, p. 345.

The Kruskal-Wallis test The semiparamteric model we use for the k-sample problem is that we observe

$$X_{ij} = \theta_i + e_{ij},\ i = 1, ..., k;\ j = 1, ..., n,\ N = \sum n_i$$

where the θ_i are unobserved parameters and the e_{ij} are i.i.d. from a continuous distribution. Note that this model allows the means (if they exist) to be different for the different samples, but assumes that the variances (if they exist) are the same. Note also that the continuity implies no ties.

We want to test the null hypothesis that the θ_i are equal against the alternative that at least one pair θ_i, θ_{i*} are different.

To do the Kruskal-Wallis test, we first jointly rank the k-samples as we did the two-samples earlier. Let R_{ij} be the rank associated with X_{ij} and let $\overline{R}_{i.}$ be the average of the ranks in the ith sample. We could then replace the X_{ij} in the F-statistic with ranks X_{ij}. It can be shown that an equivalent test can be based on

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

Note that is just a constant multiple time MSTR with ranks replacing observations. (Not that $\overline{R}_{..} = (N+1)/2$). Clearly we reject if KW is too large.

To find the critical values for this test, can use permutation tests in the obvious way (kind of a mess). Table of these exact critical values are given in Higgins, p.343.

The large sample approximation for null distribution of KW is

$$KW \overset{\bullet}{\sim} \chi^2_{k-1}$$

Rank-based HSD If we have done a Kruskal-Wallis test and rejected, we know that there is at least one pair θ_i and θ_{i*} which is significantly different and perhaps many such pairs. We can use the following method to test which pairs are significantly different. We say that θ_i and θ_{i*} are significantly different if

$$\left|\overline{R}_i - \overline{R}_{i*}\right| > q^{\alpha}_{k-1,\infty}\sqrt{\frac{N(N+1)}{24}\left(\frac{1}{n_i}+\frac{1}{n_{i*}}\right)}$$

This procedure controls the experiment-wide error rate for reasonable sample sizes.

1.3.6 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

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

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 non-parametric 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. This means 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 *Kendall'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 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

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

We estimate Kendall's τ by

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

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

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

1.4 Robust regression

As we have mentioned earlier, robust seems to be one of those terms whose meaning depends on context. Robust regression means procedures for regression problem which are less sensitive to extreme values of the response. They are unfortunately quite sensitive to extreme values in the predictors.

In the robust regression model we have response observations (Y_i) and predictor row vectors $(\mathbf{x}_i)$. (Although we shall assume that $\mathbf{x}'s$ are known constants and the $Y's$ are observed random variables, we can usually handle the case of random predictors by conditioning. The model we assume is that

$$Y_i = \mathbf{x}_i\beta + e_i,\ \ i = 1, \cdots, n$$

the β is an unobserved p-dimensional parameter and the e_i are i.i.d. from a symmetric continuous distribution with mean 0 and variance σ^2. If we assume that the e_i are normally distributed, we have the usual multiple regression model.

Let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},\ \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix},\ \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

so that

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

We assume as usual that the $\mathbf{X}$ matrix has rank p.

We now review normal regression. We estimate β by ordinary least squares (OLS) i.e. by minimizing

$$\|\mathbf{Y} - \mathbf{X}\beta\|^2 = \sum \left(Y_i - \mathbf{x}_i\beta\right)^2$$

getting the OLS estimator

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

Actually most software does not use this formula but rather solves the normal equations

$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{Y}$$

We also note that

$$\widehat{\beta} \sim N_p\left(\beta,\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

Most conclusions that that we are interested follow from this result and

$$\widehat{\sigma}^2 = \frac{\left\|\mathbf{Y}-\mathbf{X}\widehat{\beta}\right\|^2}{n-p} = \frac{\sum\left(Y_i - \mathbf{x}_i\widehat{\beta}\right)^2}{n-p},$$

$$(n-p)\,\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}.$$

One important problem with the normal development is that OLS is too sensitive too outliers. Robust regression is an answer to this problem.

Other problems with regression which robust regression does not help: unequal variance, influence, multi-collinearity, association vs. causation.

1.5 M-estimators

Consider once again the regression model without the normal assumption. Let $\rho(y)$ be a symmetric function with a unique minimum at $y = 0$. Then an *M-estimator* of β minimizes

$$\sum \rho\left(\frac{Y_i - \mathbf{x}_i\beta}{\widehat{\sigma}}\right)$$

For example if $\rho(y) = y^2$, we just get the OLS. If $\rho(y) = |y|$ we get the minimum absolute deviation. One recent favorite for people using M-estimators is the Tukey Bisquare function

$$\rho(x) = \min\left(x^6 - 3x^4 + 3x^2, 1\ \right)$$

Another one, for which I cannot find a formula is the optimal weight function of Yohai and Zamar.

One very interesting fact about an M-estimator $\widehat{\beta}_M$ is that

$$\widehat{\beta}_M \overset{\bullet}{\sim} N_p\left(\beta, \tau^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

where τ^2 is dependent on the distribution of the errors and on the function ρ. It is possible to estimate τ^2 from the data. After that M-estimator inference about β essentially follows least squares inference about β.

1.6 Rank estimators

Consider the following measure of dispersion for $\mathbf{u} = (u_1, ..., u_n)'$

$$D(\mathbf{u}) = \sum \sum |u_i - u_j| = 4 \sum \left(R_i - \frac{n+1}{2} \right) u_i$$

Let

$$e_i = Y_i - \mathbf{x}_i \beta, \ \mathbf{e} = (e_1, ..., e_n)'$$

Then the R-estimator minimizes $D(\mathbf{e})$. Note that the intercept drops out of $D(\mathbf{e})$, so that intercept must be estimated separately. Usually the Hodges-Lehmann estimator defined earlier is used. As for M-estimators, we can show the rank estimator $\widehat{\beta}_r$ satisfies

$$\widehat{\beta} \overset{\bullet}{\sim} N_p \left(\beta, \kappa^2 (\mathbf{X}'\mathbf{X})^{-1} \right)$$

where κ^2

1.7 Some nonparamtric procedures

In this last section, we consider two procedures which involve no parameters (and so are actually non-parametric).

1.7.1 One-sample Kolomogorov-Smirnov

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 $\widehat{F}(x)$ be the empirical distribution function (e.d.f.) The one sample *Kolmogorov-Smirnov* (KS) statistic is

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

We want to reject 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. A large sample result is for large n

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

Use of the last formula is quite accurate and conservative. There for a size α test we reject if

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

We can also construct a confidence band for the distribution as we now show. First note that the distribution of

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

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

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

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

On situation in which K-S is misused is in testing for normality. The problem is that 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 gladness of fit test for normality. The Anderson-Darling and Shapiro-Wilk tests are specifically designed to test for normality.

Another test of this kind for testing $F = F_0$ is the Cramer-von Mises test based on

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

1.7.2 Two-sample Kolmogorov-Smirnov

For this problem, 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 $\widehat{F}(x)$ and $\widehat{G}(y)$ be the empirical distribution functions (edf's) for the $x's$ and $y's$. The two sample *Kolmogorov-Smirnov* (K-S) test is based on

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

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

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

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