Non-Parametric (Distribution-Free) Tests

The Simple Sign Test

Here are data on six patients with cystic fibrosis. Each measurement shows the reduction (ml) in forced vital capacity over a 25 week period. For each patient there are two measurements, one for a period when he was being given a drug intended to slow the process of FVC reduction, and one for a period when he was being given a placebo.

	FEV Reduc (ml)			
Subj	**Placebo**	**Drug**	**Difference**	**Sign**
1	224	213	11	+
2	80	95	-15	-
3	75	33	42	+
4	541	440	101	+
5	74	-32	106	+
6	85	-28	113	+

(For the purpose of illustration, we are using only some of the data from P&G Table 13.1, but pretending that this is the entire sample.)

How can we test the (null) hypothesis that the drug is equivalent to the placebo?

One possibility is the t-test on the six paired differences, *D* = Placebo minus Drug. The test statistic is $\sqrt{n}\,\overline{D}\ /\ s_D$, which has a Student's t distribution with 5 df if the D's are iid normal with mean zero. The value of this test statistic is 2.672, giving a p-value of 0.022 (one-sided).

Worried about non-normality? Then we must find another test statistic. Recall that two properties are required:

(1) The more extreme the value of the test statistic, the stronger the evidence is against H_0.

(2) The distribution of the test statistic, when H_0 is true is <u>known</u>.

Here is a candidate: *Let X be the number of positive differences, and use X itself as the test statistic.*

It satisfies the two conditions:

(1) A positive difference means that FEV loss is less under the drug, so the more positive differences, the stronger the evidence that the drug works as intended.

(2) Under H_0, X has a Binomial(6, 0.5) distribution.

With this test statistic, the p-value is the probability of getting as many positive differences as we observed (5) or more.

It tests the null hypothesis (no difference between drug and placebo) against the alternative that the drug works as expected.

(The two-sided p-value is the probability of 5 or more positive differences or 5 or more negative ones, and is just twice the one-sided p-value. It tests the null hypothesis against the alternative that the drug has an effect, either in the expected direction or the other.)

This test procedure is called the "sign test." It does not require that the differences have a normal distribution.

It is valid so long as

(a) the differences are independent, and

(b) when the hypothesis of no difference between drug and placebo is true, each of the differences is just as likely to be positive as negative.

This would be true if, for example, an independent coin toss determined which treatment each patient got first, the drug or the placebo, and the study was "double blind," so that neither the patient nor the researcher knew which treatment the patient was on at any given time until after all the measurements had been made.

⇨ *Tests like the sign test, which do not require that the distribution have any particular form (like the normal, Poisson, etc.) are called "distribution-free" or "non-parametric" tests.*

For such tests, the correctness of the p-value can often be guaranteed by the researcher's act of randomly assigning treatments, as it could in this example if the coin-tossing assignment scheme were actually used.

If some of the observed differences are zero, Pagano and Gauvreau drop them from the analysis when using the sign test, reducing the sample size accordingly.

(STATA (‘signtest’) does not do this—it counts each zero as half a positive difference and half a negative one, and makes an appropriate adjustment in calculating the p-value if the total number of positives is not an integer.)

⇨ For sample size n, the sign test simply counts the number of positive observations (which has a Binomial(n,θ) probability distribution) and tests the hypothesis that θ=1/2.

The Wilcoxon Signed Rank Test

The sign test is neat, but inefficient. It looks only at the signs of the differences, ignoring their magnitudes.

In our example the absolute differences ranged from 11 ml to 113. Not only did we see just one negative difference, it was one of the two smallest. All of the four biggest changes were in the right direction. This important information is overlooked by the sign test.

	FEV Reduc (ml)			
Subj	**Placebo**	**Drug**	**Difference**	**Sign**
1	224	213	11	+
2	80	95	-15	-
3	75	33	42	+
4	541	440	101	+
5	74	-32	106	+
6	85	-28	113	+

A test that pays some attention to how big the differences are, as well as to their signs, is the **Wilcoxon Signed Rank Test**.

It ranks the observations according to magnitude (*absolute value*), from smallest (rank=1) to largest (rank=n), and takes as the test statistic the sum of the ranks of the positive observations.

	FEV Reduc (ml)				
Subj	**Placebo**	**Drug**	**Diff.**	**Sign**	**Rank**
1	224	213	11	+	1
2	80	95	-15	-	2
3	75	33	42	+	3
4	541	440	101	+	4
5	74	-32	106	+	5
6	85	-28	113	+	6

It says, in effect, that while all differences that are positive are evidence in favor of the drug, a large positive difference is stronger evidence than a small one.

This test statistic, call it R^+, satisfies our two conditions, since

(1) The bigger it is, the stronger the evidence in favor of the drug.

(2) We can figure out its exact distribution under the null hypothesis.

The reason we can find the exact distribution is that under the null hypothesis, every way of assigning + and - signs to the ranks (1, 2, ..., n) has the same probability, $(1/2)^n$.

To get the p-value, we simply count up how many of these equally probable patterns of +'s and -'s give a rank sum, R^+, as big or bigger than the one we observed.

In our little example, the observed sum of the positive ranks is 1+3+4+5+6 = 19. (Only the second-ranked difference was negative).

Ranks	Signs													
1	+	-	+	+	-	+	-	+	-	+	+	-	+	-
2	+	+	-	+	-	+	+	+	+	-	+	+	-	-
3	+	+	+	-	+	+	-	+	+	-	+	+	+	-
4	+	+	+	+	+	-	+	+	-	+	+	+	-	+
5	+	+	+	+	+	+	+	-	+	+	+	-	+	+
6	+	+	+	+	+	+	+	+	+	+	-	+	+	+
R^+	21	20	19	18	18	17	17	16	16	16	15	15	15	15...
R^-	0	1	2	3	3	4	4	5	5	5	6	6	6	6 ...
$\sum Ranks$	21	21	21	21...										...21...

Since the probability of each column is $(1/2)^6 = 1/64$, the probability of observing $R^+ \geq 19$ is $3/64 = 0.0469$ under the null hypothesis. This is the one-sided p-value for our sample.

The sum of the ranks of positive observations, R^+, and the sum of the ranks of the negative ones, R^-, together must equal the sum of <u>all</u> the ranks, $R^+ + R^- = 1 + 2 + \ldots + n = n(n+1)/2$.

Therefore the test that rejects if R^+ is too large is equivalent to a test that rejects if R^- is too small.

For example in our problem, $R^+ + R^- = 21$, so $R^+ \geq 19$ is equivalent to the condition $R^- \leq 2$. Either way, the p-value is the same:

$$P(R^+ \geq 19) = P(R^- \leq 2) = 3/64 = 0.0469.$$

It is more convenient to make tables for the smaller of the two rank sums, which Pagano and Gauvreau use *T* to represent,

$$T = \min(R^+, \ R^-)$$

To use their Table A.6 for a one-sided test, you first verify that the difference is in the right (hypothesized) direction.

If it is, you find the p-value for your observed T in the table. (If it is in the "wrong" direction, the p-value is 1 minus the value in the table.) For a two-sided test, the p-value is twice the value in the table.

For n=6 we see that the table's values agree with those that we found above:

T_{obs}	$P_0(R^- \leq T_{obs})$
0	1/64 = 0.0156
1	2/64 = 0.0313
2	3/64 = 0.0469
3	5/64 = 0.0781
4	7/64 = 0.1094
5	10/64 = 0.1563
6	14/64 = 0.2188

Note that the line of p-values for the most extreme value, $T_{obs}=0$, is omitted from Table A.6. This is simply an error. There should be another line showing that for $T_{obs}=0$, the p-value is $(½)^n$ (= 0.0156 when n=6).

The tables cover only a small range of sample sizes (Pagano and Gauvreau stop at n=12). When the number of observations gets even moderately large, a normal approximation is used.

Under the null hypothesis, it is easy to show that R^+ has

$$E[R^+]=n(n+1)/4$$
$$Var[R^+]=n(n+1)(2n+1)/24$$

(and R^- has the same mean and variance). Additionally it is less easy, but still possible, to show that

$$\frac{R^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \text{ ~approx } N(0,1)$$

Approximate p-values are then found in a Z-table.

This approximation works well, even for moderate n, because under the null hypothesis, the distribution of R^+ is symmetric about its expected value, and not skewed at all.

Even computer packages, which could easily calculate exact p-values, use this approximation. In fact, if their manual can be believed, (which is doubtful) STATA uses the normal approximation (the Z-test) for all sample sizes. Even a sophisticated package like S-plus uses the normal approximation (with a continuity correction) for n>25.

Pagano and Gauvreau suggest again that any zero differences be dropped from the analysis, as they were in the sign test, with the sample size reduced accordingly. STATA (‘signrank’) does not do this, (according to the manual). It adds one-half of the rank of a zero observation to R^+, and one-half to R^-, etc.

If some of the differences have equal magnitudes (tied observations), they are all assigned a rank equal to the average of the ranks that they would have if they weren't tied.

Pagano and Gauvreau (p. 313, Table 13.4) give an example with n=15. Because four of the observed differences are zero, these are dropped from the analysis, and the rank sum is calculated for the eleven non-zero differences.

The four smallest of the 11 non-zero differences all had the same absolute value, 0.1, so each one gets their average rank, (1+2+3+4)/4 = 2.5. The fifth and sixth were also tied (at 0.2), so they each get their average rank, (5+6)/2 = 5.5, etc.

(The authors also present STATA output for this problem (Table 13.5). Since STATA also drops the zeroes from this analysis, it means that either the manual is wrong or the authors were not using the current version of STATA.)

The Wilcoxon (Mann-Whitney) Rank Sum Test

The sign test and the Wilcoxon signed-rank test are for single samples, like the one-sample t-test.

The distribution-free alternative to the two-sample t-test is the Wilcoxon Rank Sum Test.

(There is another candidate, the Mann-Whitney U-Test, which looks entirely different, but which turns out to be equivalent to Wilcoxon's test. For this reason, various combinations of the three names, Wilcoxon, Mann, and Whitney, are sometimes used to identify this test.)

The model asserts that the two samples are independent, and the null hypothesis states that they come from the same probability distribution. Just as with the one-sample Wilcoxon test, no assumption about the precise form of this common probability distribution is required.

If the alternative hypothesis specifies that the two distributions are not the same, and that observations from a specified one (say the first) will tend to be larger, then a one-sided test is appropriate.

If it simply states that one sample will tend to be larger, without specifying which one, a two-sided test is called for.

The test is again based on the ranks of the observations. We begin by ordering all of the observations (both samples combined, with their signs intact) from smallest to largest, and assigning ranks, 1, 2, ..., $n_1 + n_2$. Let R_1 be the sum of the ranks of the n_1 observations in sample 1 and R_2 be the sum of the ranks of the n_2 observations in sample 2.

The test statistic (one-sided test) is the sum of the ranks of the sample whose members, according to the alternative hypothesis, will tend to be larger. Suppose this is sample 1, so the test statistic is R_1.

(1) The larger R_1 is, the stronger the evidence that sample 1 observations tend to be larger, just as the alternative hypothesis predicted.

(2) The probability distribution of R_1, under the null hypothesis, is known.

The reason (2) is true is that the complete set of ranks (i.e., the integers, 1, 2, ..., n_1+n_2) is fixed. Observations in sample 1 will get some (n_1) of these ranks, and the rest will go to sample 2.

Under the null hypothesis, the particular set of n_1 ranks that go to sample 1 is determined by a process equivalent to choosing a simple random sample of n_1 from the complete set of ranks.

It is as if we put n_1+n_2 pieces of paper, numbered 1, 2, ..., n_1+n_2, into a hat, and drew out n_1 of them (without replacement).

From this we can calculate the exact probability distribution of R_1 under the null hypothesis.

For example, the smallest possible value of R_1 is the sum of the n_1 smallest ranks, 1 + 2 + ... + n_1 = n_1(n_1+ 1)/2, and the probability of this value is

$$\frac{1}{\binom{n_1+n_2}{n_1}}$$

When the alternative hypothesis specifies that observations in sample 1 will tend to be larger, the p-value is P($R_1 \geq R_{1obs}$).

<u>Example</u>: Sample 1: 10, 12, 19, 20 (n_1 = 4)
Sample 2: -15, -2, 1, 9, 15 (n_2 = 5)

Sample 1					⇓	⇓		⇓	⇓	
Combined sample:	-15,	-2,	1,	9,	10,	12,	15,	19,	20	
Ranks	1	2	3	4	5	6	7	8	9	
										$R_{1\ obs}$
Observed Sample 1 Ranks					5	6		8	9	28
				4			7	8	9	28
					5		7	8	9	29
						6	7	8	9	30

The one-sided p-value is

$$P(\ R_1 \geq\ R_{1\ obs}\ ) = \frac{4}{\binom{9}{4}} = \frac{4}{126} = 0.032$$

Because there are 4 such outcomes ‘more extreme’ than R_1 =28.

As with the one-sample rank test, it is more convenient to make tables in terms of the smaller of the two rank sums, R_1 and R_2. Pagano and Gauvreau (Table A.7) let *W* denote the smaller of the two rank sums. And when the sample sizes get moderately large, a normal approximation is used instead of the exact p-value. If n_W is the size of the sample whose rank sum is W, the approximately normal test statistic is

$$\frac{W - \frac{n_W(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}}$$

When there are ties (more than one observation with the same value), all observations with the same value are given the average rank that they would have if there were no ties. For example, if the 7th, 8th, 9th, and 10th largest observations are all equal, they are all given rank (7+8+9+10)/4 = 8.5.

When there are only a few ties, they are sometimes ignored, but there is a correction for ties that is often used when the Wilcoxon rank sum test is applied to "highly tied" data.

So for a single sample (which often consists of differences between two paired samples), for testing the hypothesis that the mean is zero, we can use the test statistic

$$\frac{\sqrt{n}\,\overline{X}_n}{S_n},$$

looking up the p-value in a t-table. This test is exact only when the distribution is normal. For a non-normal distribution the test is only approximate, but when n is large (and the p-value comes from a Z-table) the approximation is very good.

As alternatives, we have the sign test and the Wilcoxon signed rank test. If the distribution is normal, these tests are less powerful than the t-test. But their Type I error probabilities (and p-values) are exact under much more general conditions.

Similarly, for testing whether two normal distributions are the same, vs the alternative that they have different means, we have, as an alternative to the two-sample t-test (which is exact only when the distributions are normal), the Wilcoxon rank sum test. It has less power when the distributions are normal, but exact size under much more general conditions.

The best place to learn more about nonparametric tests is in books that specialize in that topic. General statistical methods books are sometimes not very accurate in their treatment of nonparametric methods.

For example, the awkward question about whether the two distributions have equal variances, which is so annoying when two normal distributions are being compared, cannot be avoided by using the Wilcoxon rank sum test instead of the t-test. The correctness of the rank test's p-value can be upset by inequality of the variances, just like the t-test's p-value can.

Many general statistical methods books are inaccurate on this point.

Distribution-free ("non-parametric") tests have some advantages.

(1) They enable you to calculate exact significance levels and p-values for small samples when the precise form of the distribution is not known.

(2) They can be used with "ordinal" data, when the observations are ordered or consist of ordered categories, like "poor", "fair", and "good," but do not have meaningful numerical values.

(3) They are insensitive to gross numerical errors.

Distribution-free tests also have disadvantages. One is their loss of power, compared to parametric tests like Student's t, when the parametric tests are valid.

A more important disadvantage, and the one that probably explains why non-parametric tests have not achieved greater popularity, is that they have no natural, easily-understood, link to estimation, like the link between the t-test and the sample mean.