Parametric and Non-Parametric Tests

Every hypothesis test makes assumptions. A z-test for a mean assumes the population is normal, or that n is large enough for the Central Limit Theorem to rescue us. A test on a correlation coefficient ρ assumes the underlying data is bivariate normal. Tests like these, which assume the population has a particular distributional form and then test a *parameter* of it, are called **parametric tests**.

What if the assumptions fail? With a sample of size 8 from a skewed, unknown distribution, we cannot assume normality and the CLT is no help. We need tests that work regardless.

Definition. A **non-parametric test** is a hypothesis test that makes few or no assumptions about the distribution of the underlying population.

Non-parametric tests are useful when:

- the sample is small and the population cannot be assumed normal;
- the data is clearly skewed or contains outliers (medians behave better than means);
- the data is only *ordinal* — ranks or preferences rather than true measurements.

The price: when parametric assumptions *do* hold, non-parametric tests are less powerful (they throw away information, as we shall see).

> **Remark.** You have already met one non-parametric test. Testing $H_0 : \rho = 0$ with the product-moment correlation coefficient requires bivariate normality; testing with Spearman's rank correlation coefficient r_s does not, because ranking the data discards its distribution. See the Correlation notes.

Because we drop all distributional assumptions, our hypotheses change character: non-parametric tests are about the **median** m rather than the mean (the median always exists and is meaningful for any shape of distribution).

Textbook Exercises: [CUP.S] Ch 4 §1; [S3&4] S4 Ch 2

The Single-Sample Sign Test

The simplest possible test. To test $H_0: m = m_0$: if the population median really is m_0, then each observation is equally likely to fall above or below m_0. So the number of observations above m_0 is binomial with $p = \frac{1}{2}$.

Tip (The procedure)

1. $H_0: m = m_0$; $H_1: m \neq m_0$ (or $>$, $<$), where m is the population median of [context].
2. **Discard** any observations exactly equal to m_0. Let n be the number remaining.
3. Record the sign (+ or −) of each remaining observation minus m_0, and count the positive signs: $S \sim B(n, \frac{1}{2})$ under H_0.
4. Compute the tail probability of a result at least as extreme as the one observed; for a two-tailed test, double it. Compare with the significance level.

Example (Sign test)

A manufacturer claims the median lifetime of its batteries is more than 30 hours. Twelve batteries are tested, with lifetimes (hours):

$$31,\ 34,\ 28,\ 36,\ 30,\ 32,\ 41,\ 29,\ 33,\ 38,\ 35,\ 31.$$

Test the claim at the 5% significance level using a sign test.

Example (OCR S4, June 2012)

A one-tail sign test of a population median is to be carried out at the 5% significance level using a sample of size n.

(i) Show by calculation that the test can never result in rejection of the null hypothesis when $n = 4$.

(ii) The coach of a college swimming team expects Elena, the best 50 m freestyle swimmer, to have a median time less than 30 seconds. Elena found from records of her previous 72 swims that 44 were less than 30 seconds and 28 were greater than 30 seconds. Stating a necessary assumption, test at the 5% significance level whether Elena's median time for the 50 m freestyle is less than 30 seconds.

Remark (The weakness of the sign test)**.** The sign test uses almost no assumptions — but also almost no information. An observation 0.1 above the median and one 40 above the median count identically. Data like $-1,-2,-1,+30,+45,+38,+29$ would be treated as "3 minus, 4 plus", ignoring the obvious message in the magnitudes. The next test repairs this.

Textbook Exercises: [CUP.S] Ch 4 §1; [S3&4] S4 Ch 2

The Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test uses the *sizes* of the deviations from m_0 as well as their signs — at the cost of one extra assumption.

Fact (Assumption) — The Wilcoxon signed-rank test assumes the population distribution is **symmetric**. Under H_0, deviations of any given size are then equally likely to be positive or negative.

Tip (The procedure)

1. $H_0 : m = m_0$; $H_1 : m \neq m_0$ (or one-tailed).
2. Compute the differences $d_i = x_i - m_0$, discarding any zeros.
3. Rank the $|d_i|$ from 1 (smallest) to n (largest). (We do not consider tied ranks.)
4. Let T^+ = sum of the ranks of the positive differences, T^- = sum of ranks of the negative differences. (Also written W^+ and W^-.) Check: $T^+ + T^- = \frac{n(n+1)}{2}$.
5. Test statistic $T = \min(T^+, T^-)$. Reject H_0 if $T \leq$ the critical value from the formula-booklet table (small T means one side's ranks are suspiciously light).

Example (Single-sample Wilcoxon signed-rank test)

Bags of flour are sold as having median weight 50 g. A sample of 10 bags weighs (g):

$$52,\ 43,\ 38,\ 45,\ 41,\ 35,\ 51,\ 40,\ 46,\ 42.$$

Assuming the distribution of weights is symmetric, test at the 5% significance level whether the population median weight differs from 50 g.

Fact — Under H_0,

$$\mathbb{E}[T^+] = \mathbb{E}[T^-] = \frac{n(n+1)}{4}.$$

Values of T far below $\frac{n(n+1)}{4}$ are evidence against H_0.

The proof is one line.

Where do the tables come from?

Under H_0 each of the n ranks is positive or negative independently with probability $\frac{1}{2}$, giving 2^n equally likely sign patterns. The critical value at level α is the largest t with $\mathbb{P}(T^+ \le t) \le \alpha$.

Exercise. For $n = 4$, list all $2^4 = 16$ sign patterns and find the distribution of T^+. Show that *no* significant result is possible at the 5% level, so the booklet table has no row for $n = 4$. Then verify the first row of the table ($n = 5$).

Example (OCR Further Statistics, December 2018)

The reaction times, in milliseconds, of all adult males in a standard experiment have a symmetrical distribution with mean and median both equal to 700 and standard deviation 125. The reaction times of a random sample of 6 international athletes are measured and the results are as follows:

$$702, \quad 631, \quad 540, \quad 714, \quad 575, \quad 480.$$

It is required to test whether international athletes have a mean reaction time which is less than 700.

(a) Assume first that the reaction times of international athletes have the distribution $N(\mu, 125^2)$. Test at the 5% significance level whether $\mu < 700$.

(b) Now assume only that the distribution of the data is symmetrical, but not necessarily normal.

(i) State with a reason why a Wilcoxon test is preferable to a sign test.

(ii) Use an appropriate Wilcoxon test at the 5% significance level to test whether the median reaction time of international athletes is less than 700.

(c) Explain why the significance tests in part (a) and part (b)(ii) could produce different results.

Textbook Exercises: [CUP.S] Ch 4 §2; [S3&4] S4 Ch 2

Paired Samples

Often two measurements are made on the *same* subjects — before and after training, two treatments on matched patients, two judges scoring the same performances. Such data is **paired**. The right move is to work with the *differences* within each pair, reducing the problem to a single sample of differences.

Tip (Paired or two-sample?)

If each value in one sample is naturally linked to one particular value in the other (same person, same plot of land, same day), use a **paired** test on the differences. If the two samples are independent groups (possibly of different sizes), a paired test is impossible — use the two-sample rank-sum test of the next section.

Fact (The correct null hypothesis) — For a paired test the null hypothesis is

$$H_0 : m_d = 0,$$

where m_d is the **population median of the differences** in the given context. It is **not** "the two medians are equal": the median of the differences is not the difference of the medians.

Example

Find paired data sets X and Y with equal medians but $m_d \neq 0$.

Example (Paired sign test)

Ten athletes run a time trial before and after a training programme. The differences (before − after, seconds) are

$$+1.2,\ +0.8,\ -0.3,\ +2.1,\ +0.5,\ 0,\ +1.7,\ +0.9,\ +0.4,\ +1.1.$$

Use a sign test at the 5% significance level to decide whether the programme reduces times.

Example (Paired Wilcoxon signed-rank test)

Eight students sit a paper before and after a revision course. The score differences (after – before) are

$$+3,\ +7,\ -2,\ +9,\ +5,\ +11,\ -1,\ +6.$$

Assuming the differences are symmetrically distributed, test at the 5% significance level whether the course improves scores.

Example (OCR S4, June 2014)

A teacher believes that the calculator paper in a GCSE Mathematics examination was easier than the non-calculator paper. The marks of a random sample of ten students are shown in the table.

Student	A	B	C	D	E	F	G	H	I	J
Mark on paper 1 (non-calculator)	66	79	58	87	67	55	75	62	50	84
Mark on paper 2 (calculator)	57	84	70	90	75	42	82	72	65	82

(i) Use a Wilcoxon signed-rank test, at the 5% significance level, to test the teacher's belief.

(ii) State the assumption necessary for this test to be applied.

Textbook Exercises: [CUP.S] Ch 4 §3; [S3&4] S4 Ch 2

The Wilcoxon Rank-Sum Test

For two *independent* (unpaired) samples, of sizes m and n with $m \le n$, we use the **Wilcoxon rank-sum test**, also known as the **Mann–Whitney U test**. The hypotheses are

$$H_0: \text{the two population distributions are identical}$$
$$H_1: \text{the two population distributions are not identical.}$$

More commonly we assume the two distributions have the same *shape* and may differ only in *location*, so the hypotheses become $H_0: m_A = m_B$ against $H_1: m_A \neq m_B$ (or one-tailed), where m_A, m_B are the population medians.

Tip (The procedure)

1. Combine both samples into a single list and rank all $m+n$ values from 1 (smallest) upwards.
2. Let W be the sum of the ranks of the **smaller** sample (size m).
3. Reject H_0 if $W \le W_{\text{crit}}$ (booklet table) or $W \ge m(m+n+1) - W_{\text{crit}}$. For a one-tailed test, use whichever tail H_1 points to.

Remark. The Mann–Whitney statistic is $U = W - \frac{m(m+1)}{2}$, the number of (smaller-sample, larger-sample) pairs in which the smaller-sample value wins; some tables are written in terms of U instead of W. OCR uses W.

Example (Rank-sum test)

Two groups of students learn a routine by different methods, then are timed performing it. The times (minutes) are

$$\text{Method A } (m=4): \quad 11,\ 12,\ 15,\ 19$$
$$\text{Method B } (n=6): \quad 17,\ 18,\ 21,\ 24,\ 25,\ 30$$

Test at the 5% significance level whether the median times under the two methods differ. (You may assume the two distributions have the same shape.)

Fact — Under H_0, the distribution of W is symmetric about its mean

$$\mathbb{E}[W] = \frac{m(m+n+1)}{2}.$$

The proof is a pleasant symmetry argument.

Where do the tables come from?

Example ($m = 2$, $n = 5$)
Under H_0 all $\binom{7}{2} = 21$ choices of the two ranks held by the smaller sample are equally likely. Find the distribution of W and hence the one-tailed critical value at the 5% level.

Textbook Exercises: [CUP.S] Ch 4 §4; [S3&4] S4 Ch 2

Normal Approximations for Large Samples

The booklet tables stop at modest sample sizes. For larger samples, T^+ and W are sums of many small independent contributions, so (by CLT-style reasoning) they are approximately normal, with the means we have already found.

Fact (In the formula booklet) — For large samples, under H_0:

$$\text{Wilcoxon signed-rank:} \quad T \sim N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right) \quad \text{approximately}$$

$$\text{Wilcoxon rank-sum:} \quad W \sim N\left(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12}\right) \quad \text{approximately}$$

Since T and W are discrete (integer-valued) and the normal distribution is continuous, a **continuity correction** must be used. A continuity correction is essential here.

The means are the ones we derived earlier; the variance of T^+ is no mystery either.

Remark. The approximations are for T^+ (equivalently T^-) and W, so we can test at either tail, choosing the tail H_1 points to; for a two-tailed test, double the tail probability.

Example (Signed-rank, large sample)

A website claims the median time to complete its checkout is 35 seconds. The times of 20 customers are recorded; none equals 35, and the Wilcoxon signed-rank calculation gives $T^+ = 60$ (sum of ranks of times above 35). Test at the 5% significance level whether the median time is less than claimed.

Example (OCR S4, June 2018)

A Wilcoxon signed-rank test is carried out at the 5% level of significance on a random sample of size 32. The hypotheses are $H_0: m = m_0$, $H_1: m < m_0$, where m is the population median and m_0 is a specific numerical value. The value obtained for the test statistic T is 162. Find the outcome of the test.

Example (Rank-sum, large sample)

Independent samples of $m = 10$ and $n = 12$ plants are grown with two fertilisers and their heights ranked together. The rank sum of the smaller sample is $W = 85$. Test at the 5% significance level whether the two fertilisers produce different median heights.

Choosing the right test

Situation	Test	Assumptions
One sample, median	Sign test	none
One sample, median	Wilcoxon signed-rank	symmetric distribution
Paired samples	Sign test on differences	none
Paired samples	Wilcoxon signed-rank on differences	symmetric differences
Two independent samples	Wilcoxon rank-sum	same shape distributions

Tip

In conclusions, never write "accept H_0" or "accept H_1". Either "reject H_0: there is evidence at the [level] to suggest that…" or "do not reject H_0: there is insufficient evidence to suggest that…" — always in the context of the question.

Textbook Exercises: [CUP.S] Ch 4 §5; [S3&4] S4 Ch 2