

Non-parametric tests

A2

Read online at pastopic.com/cambridge/further-mathematics-9231/notes/non-parametric-tests

Last checked 1 October 2026

Read first: [Inference using normal and t-distributions](#)

© 2026 Pastopic. Free for personal study. Please share the link, not the file.

What you need to know

By the end of this topic you should be able to:

1. **Say what a non-parametric test is and when to use one.** It tests a hypothesis about a population **median** without assuming the population follows a particular distribution. Use one when you can't assume the population is normal, for example with a small sample from an unknown or skewed population.
2. **Understand how the three tests work:** the **sign test**, the **Wilcoxon signed-rank test** and the **Wilcoxon rank-sum test**. Know that the Wilcoxon tests are only valid when the distribution is **symmetrical**.
3. **Test a population median from one sample**, with a single-sample sign test or a single-sample Wilcoxon signed-rank test, using a normal approximation when the sample is large.
4. **Test whether two populations are the same**, choosing the right test: a paired-sample sign test or a Wilcoxon matched-pairs signed-rank test for paired data, and a Wilcoxon rank-sum test for two independent samples, again with normal approximations for large samples.

Exam questions never have tied ranks, values equal to the median being tested, or pairs with a difference of zero.

Understand it

Why "non-parametric"?

The t-test and the z-test are **parametric**: they assume the population is normal (or the sample is large enough for the Central Limit Theorem) and test its mean μ . With a small sample from a population whose shape you don't know, that assumption can't be trusted.

A **non-parametric** test makes no assumption about the population's distribution (or only a weak one, such as symmetry). It works with **signs** or **ranks** instead of the actual values, and tests the population **median** m . The price is that it uses less of the information in the data, so it needs stronger evidence before it rejects H_0 .

The sign test

Suppose H_0 : the population median is m_0 . Then each value in a random sample is equally likely to be above or below m_0 . Write + for each value above m_0 and – for each below. If X is the number of + signs in a sample of n ,

$$X \sim B(n, 0.5) \quad \text{when } H_0 \text{ is true.}$$

Too many or too few + signs is evidence against H_0 . Find the probability of a result at least as extreme as the one observed, and compare it with the significance level, exactly as in a binomial hypothesis test.

Paired data (two measurements on each individual) works the same way: take the sign of each difference, such as + when the second score beats the first. H_0 says the two population medians are equal, so each sign is + with probability 0.5.

The sign test uses only the direction of each difference and ignores how big it is. It needs no assumption about the shape of the distribution, so it still works for skewed data or data with outliers.

The Wilcoxon signed-rank test

The Wilcoxon signed-rank test uses the **sizes** of the differences too, through their ranks:

1. Work out each difference $d = x - m_0$ (or the difference within each pair).
2. Rank the differences by size, **ignoring the sign**: the smallest $|d|$ gets rank 1.
3. Put each difference's sign back on its rank.
4. P = sum of the positive ranks, Q = sum of the negative ranks. The test statistic is T , the smaller of P and Q .

For example, testing $m_0 = 50$ with the sample 53, 49, 57, 54, 48, 56: the differences are +3, –1, +7, +4, –2, +6, with signed ranks +3, –1, +6, +4, –2, +5. So $P = 18$ and $Q = 3$, and $T = 3$.

The ranks 1 to n always add up to $\frac{1}{2}n(n+1)$, so $P + Q = \frac{1}{2}n(n+1)$ (here $18 + 3 = 21$). That is a useful check. If H_0 is true, positive and negative differences are equally likely **and equally big**, so P and Q should each be near half of this total. A **small** T means one side has only the small differences: evidence against H_0 . MF19 gives the largest T that is significant; for $n = 6$ at the 5% one-tailed level it is 2, so $T = 3$ here is not quite enough.

Why symmetry matters. "Positive and negative differences are equally big" only follows from H_0 if the population is **symmetrical** about its median. In a skewed population the values on the long side of the median are further from it, so P and Q differ even when m_0 is the true median. So the Wilcoxon signed-rank test assumes the population (for paired data, the population of **differences**) is symmetrical. If the data look skewed or have an outlier, use the sign test instead.

The Wilcoxon rank-sum test

For **two independent samples** (different individuals in each, so nothing can be paired), rank all the values together from 1 to $m + n$, where m is the size of the **smaller** sample. Add up the ranks of the smaller sample to get R_m .

If the two populations are the same, the smaller sample's ranks are a random choice from 1 to $m + n$, so R_m should be near its mean $\frac{1}{2}m(m + n + 1)$. A very small or very large R_m is evidence of a difference. The table uses

$$W = \text{the smaller of } R_m \text{ and } m(m + n + 1) - R_m,$$

where $m(m + n + 1) - R_m$ is what R_m would have been with the ranks reversed. For example, samples A: 12, 19, 23 and B: 15, 16, 21, 25 rank together as 12 (1), 15 (2), 16 (3), 19 (4), 21 (5), 23 (6), 25 (7), so $R_3 = 1 + 4 + 6 = 11$, $3 \times 8 - 11 = 13$, and $W = 11$. As with T , a **small** W is significant. The test assumes the two populations have the same shape, so that any difference is a shift in the median.

Large samples: normal approximations

The MF19 tables stop at $n = 20$ for the signed-rank test and at $m, n = 10$ for the rank-sum test. Beyond them, the statistics are close to normal.

- **Signed-rank:** under H_0 each rank r is equally likely to be positive or negative, so P gets half of each rank on average: $E(P) = \frac{1}{2} \times \frac{1}{2}n(n + 1) = \frac{1}{4}n(n + 1)$. Adding up the variances $\frac{1}{4}r^2$ of the ranks gives $\text{Var}(P) = \frac{1}{24}n(n + 1)(2n + 1)$. Both are in MF19, and Q has the same distribution.
- **Rank-sum:** R_m is approximately $N\left(\frac{1}{2}m(m + n + 1), \frac{1}{12}mn(m + n + 1)\right)$, also in MF19.
- **Sign test:** $X \sim B(n, 0.5)$ is approximately $N(0.5n, 0.25n)$ when n is large.

T , R_m and X only take whole-number values, so each approximation needs a **continuity correction** of 0.5 towards the mean.

Choosing the test

Data	Test for the median	Needs symmetry?
One sample, compared with a stated median m_0	Single-sample sign test, or single-sample Wilcoxon signed-rank test	Sign: no. Wilcoxon: yes
Two measurements on each individual (paired)	Paired-sample sign test, or Wilcoxon matched-pairs signed-rank test on the differences	Sign: no. Wilcoxon: yes (differences)
Two separate groups of different individuals	Wilcoxon rank-sum test	Same shape for both populations

If the question names the test, use that one. When both a sign test and a Wilcoxon test are possible, the Wilcoxon test is usually more powerful because it uses the sizes, so the two can reach **different** conclusions on the same data.

Key facts and formulas

"Given" means it is in MF19 (the list of formulae and statistical tables) in the exam. "Learn it" means you must remember it.

Formula	Status
Sign test: number of + signs $X \sim B(n, 0.5)$ under H_0	Learn it
Binomial probabilities $p_r = \binom{n}{r} p^r (1-p)^{n-r}$, mean np , variance $np(1-p)$	Given
Sign test for large n : $X \approx N(0.5n, 0.25n)$, with a continuity correction	Learn it
Signed-rank: P = sum of positive ranks, Q = sum of negative ranks, T = smaller of P and Q	Given
$P + Q = \frac{1}{2}n(n+1)$	Learn it
Critical values of T for $n = 6$ to 20 ; reject H_0 if $T \leq$ the table value	Given
Signed-rank for large n : P (or Q) $\approx N(\frac{1}{4}n(n+1), \frac{1}{24}n(n+1)(2n+1))$	Given
Rank-sum: R_m = rank sum of the smaller sample, W = smaller of R_m and $m(m+n+1) - R_m$	Given
Critical values of W for $3 \leq m \leq n \leq 10$; reject H_0 if $W \leq$ the table value	Given
Rank-sum for large samples: $R_m \approx N(\frac{1}{2}m(m+n+1), \frac{1}{12}mn(m+n+1))$	Given
Continuity correction: move the statistic 0.5 towards the mean before standardising	Learn it
Normal critical values: 1.282 (10% one-tailed), 1.645 (5% one-tailed or 10% two-tailed), 1.96 (2.5% one-tailed or 5% two-tailed), 2.326 (1% one-tailed or 2% two-tailed), 2.576 (0.5% one-tailed or 1% two-tailed)	Given

Reading the Wilcoxon tables: each column is headed with a one-tailed and a two-tailed level. A 5% two-tailed test uses the same column as a 2.5% one-tailed test. For example, with $n = 11$ the table gives 13 (5% one-tailed) and 10 (5% two-tailed).

How to do it

In the Paper 4 questions we have tagged for this topic (34 questions, 2021–2025, of which 11 are Paper 41 copies of another paper), the types came up this often. 20 of the 34 questions mix more than one type, so the counts add up to more than 34.

Question type	Questions
Wilcoxon matched-pairs signed-rank test (tables)	12
Justify the choice of test	9

Question type	Questions
Sign test (single sample or paired)	6
Single-sample Wilcoxon signed-rank test with a normal approximation	5
Single-sample Wilcoxon signed-rank test (tables)	4
Wilcoxon rank-sum test (tables)	4
Wilcoxon rank-sum test with a normal approximation	4
State the assumption needed for a Wilcoxon test	4
Effect of a corrected data value on the conclusion	4
Find the least or greatest test statistic for a given outcome	2
Comment on the validity of the experiment	2

Every test starts and ends the same way:

- **Hypotheses** about the **population median**, in words or with m defined as the population median (never μ , never "mean"). One sample: " H_0 : the population median is 800; H_1 : the population median is less than 800." Two populations: " H_0 : the population median scores on the two keyboards are equal; H_1 : they are not equal." Read the claim to decide whether H_1 is one-tailed ("more", "less", "increased", "quicker") or two-tailed ("different", "changed", "any evidence against").
- **Conclusion**: compare, say "reject H_0 " or "do not reject H_0 ", then give one sentence in context that sounds uncertain: "there is enough evidence at the 5% level to suggest ..." or "there is not enough evidence to suggest ...". Never "this proves", never "there is no evidence".

1. Wilcoxon matched-pairs signed-rank test, with tables (12 questions)

1. Hypotheses on the two population medians.
2. **Differences** for each pair, always the same way round (for example second minus first). A table with columns for the pairs, d , $|d|$ rank and signed rank keeps you accurate.
3. **Rank** the differences by size, ignoring signs, then attach the signs.
4. Find P and Q and check that $P + Q = \frac{1}{2}n(n + 1)$. T is the smaller.
5. **Critical value** from the table for n pairs at the right one- or two-tailed level.
6. $T \leq$ critical value: reject H_0 . $T >$ critical value: do not reject. Conclude in context.

Variations you will meet:

- **One-tailed tests.** H_1 predicts which sum will be small. If the claim is "the second score is higher", the differences (second minus first) should be mostly positive, so Q should be small. If instead the sum that H_1 predicts to be small is the **larger** one, the data point the other way and you cannot reject H_0 .
- **Two-tailed at 10%** uses the 5% one-tailed column (for $n = 9$ that is 8, not 5).
- The question may ask for the test and then for a second test, an assumption or a corrected value (types 3, 8 and 9).

2. Justify the choice of test (9 questions)

One-mark reasons; always talk about the **population** (or underlying distribution), not the sample.

- **Why a Wilcoxon (or sign) test rather than a t-test?** The t-test needs the population (for paired data, the population of differences) to be normal. Here that isn't known, and the sample is too small for the Central Limit Theorem to help.
- **Why not a Wilcoxon signed-rank test?** Its population must be symmetrical. A value far from the rest (an outlier) or data bunched on one side suggest the population is skewed, so use a sign test instead.
- **Why not a paired test?** Paired tests need each value in one sample to belong with one value in the other (the same person, item or day). Two groups of **different** individuals can't be paired, and their sizes may even differ. Use the rank-sum test.
- **Why might a t-test and a rank-sum test disagree?** The t-test compares means and is only valid if both populations are normal with a common variance; the rank-sum test compares medians and needs neither. One very large or small value can drag a sample mean a long way but moves only one rank. Different questions, different conditions: the two tests can easily disagree.

3. Sign test, single sample or paired (6 questions)

1. Hypotheses on the population median (or on the two population medians).
2. Record the sign of each value minus m_0 (or of each difference). Count the + and the – signs.
3. $X \sim B(n, 0.5)$. Find the probability, **in the tail H_1 points to**, of a result at least as extreme as observed: $P(X \leq \text{count})$ if the count is small, $P(X \geq \text{count})$ if it is large (by symmetry these are equal for the other sign's count).
4. **Compare** with the significance level. For a **two-tailed** test compare the tail probability with **half** the level (or double it and compare with the full level). Conclude.

For example, 12 pairs with only 2 minus signs, tested two-tailed at 5%: $P(X \leq 2) = \frac{1+12+66}{4096} = 0.0193 < 0.025$, so reject H_0 .

Variations you will meet:

- **Large n :** use $N(0.5n, 0.25n)$ with a continuity correction. With 32 plus signs out of 50: $z = \frac{31.5-25}{\sqrt{12.5}} = 1.838$. That is significant against 1.645 (5% one-tailed) but not against 1.96 (5% two-tailed).
- **Sign test beside a Wilcoxon test on the same data**, "show that the conclusions differ (or agree)": carry out both fully, then say which rejects H_0 . The sign test throws away the sizes, so it can miss evidence that the Wilcoxon test finds (see Example 1).

4. Single-sample Wilcoxon signed-rank test with a normal approximation (5 questions)

Used when $n > 20$ (typically 22 to 25 values).

1. Hypotheses, then differences from m_0 (**not** from 0), ranks, P and Q exactly as in type 5. Check $P + Q = \frac{1}{2}n(n+1)$.

- Mean $\frac{1}{4}n(n+1)$ and variance $\frac{1}{24}n(n+1)(2n+1)$ from MF19.
- $z = \frac{T + 0.5 - \text{mean}}{\sqrt{\text{variance}}}$: T is the smaller sum, so it is below the mean and the correction is $+0.5$.
- Compare with the negative critical value (-1.645 for 5% one-tailed or 10% two-tailed, -1.96 for 5% two-tailed), or compare the tail probability $\Phi(z)$ with the level. Keep the signs consistent.

For example, with $n = 30$ the mean is 232.5 and the variance is 2363.75. If $T = 160$, then $z = \frac{160.5 - 232.5}{\sqrt{2363.75}} = -1.481$, which is not significant at 5% one-tailed ($-1.481 > -1.645$).

5. Single-sample Wilcoxon signed-rank test, with tables (4 questions)

The same steps as type 1, with $d = x - m_0$ for each value: rank the $|d|$, attach signs, T = smaller of P and Q , compare with the table for n values. Example 1 is this type. A question may ask you to show that a sign test and a Wilcoxon test on the same sample give the **same** conclusion: do both in full and state the shared conclusion once, in context.

6. Wilcoxon rank-sum test, with tables (4 questions)

- Hypotheses: the two population medians are equal; H_1 one- or two-tailed from the claim.
- Rank **all** $m + n$ values together from 1 (smallest). Label each rank with its sample.
- R_m = sum of the ranks of the **smaller** sample (size m). Work out $m(m+n+1) - R_m$; W is the smaller of the two.
- Critical value from the table for m and n (for example $m = 6, n = 9$ at 5% one-tailed: 33). $W \leq$ critical value: reject H_0 .

One-tailed direction: if H_1 says the smaller sample's population has the **larger** median, its ranks should be high, so the relevant value is $m(m+n+1) - R_m$; if it says **smaller**, it is R_m . If the relevant value is the larger of the two, the data go the wrong way and you cannot reject H_0 .

If both samples have 10 or fewer values, you **must** use the table: a normal approximation here loses the final marks.

7. Wilcoxon rank-sum test with a normal approximation (4 questions)

When either sample has more than 10 values:

- Rank and find R_m as in type 6. Sometimes the question gives only m, n and R_m ; then all the work is in steps 2 and 3.
- Mean $\frac{1}{2}m(m+n+1)$, variance $\frac{1}{12}mn(m+n+1)$ (the denominator is 12).
- Standardise with a continuity correction **towards the mean**: $R_m + 0.5$ if R_m is below the mean, $R_m - 0.5$ if it is above.
- Compare z with $\pm 1.96, \pm 1.645$ and so on, and conclude.

For example, $m = 15$ and $n = 20$ give mean 270 and variance 900. If $R_m = 318$ (above the mean), $z = \frac{317.5 - 270}{30} = 1.583$.

8. State the assumption needed for a Wilcoxon test (4 questions)

One mark, and the words matter:

- **Single-sample signed-rank:** the population (underlying distribution) is symmetrical about its median.
- **Matched pairs:** the population of **differences** is symmetrical.
- **Rank-sum:** the two population distributions have the same shape (only their medians might differ), for example both symmetrical.

"The data are symmetrical" or anything about the mean or normality earns nothing.

9. Effect of a corrected data value on the conclusion (4 questions)

1. Recalculate the difference for the corrected pair and find its new rank and sign. Check whether it changes the order of the other ranks.
2. Find the new P , Q and T . Often one rank just changes sign, so T drops (or rises) by that rank.
3. Compare the new T with the **same** critical value and say whether the conclusion changes. Give the numbers: "Explain, with numerical justification" needs the new T and the critical value, not just words. Example 2(b) is this type.

10. Find the least or greatest test statistic for a given outcome (2 questions)

The test is run backwards: you know the outcome and want T .

1. Write the mean and variance for the normal approximation.
2. Write the boundary condition with the critical z for the stated level. "Significant" means $\frac{T + 0.5 - \text{mean}}{\sqrt{\text{var}}} \leq -z_{\text{crit}}$; "not significant" means the same fraction is $> -z_{\text{crit}}$ (for example -1.96 for a two-tailed 5% test).
3. Solve for T and round to the right **whole number**: up for "least T not significant", down for "greatest T significant". Example 4(b) is this type.

11. Comment on the validity of the experiment (2 questions)

Look for anything other than the factor being tested that could explain the result, and say it in context:

- the **order** of tasks: if everyone did task A before task B, B might go better because they have warmed up, or worse because they are weary, whatever the tasks themselves are like; a fair design would give half the people A first;
- samples that are not random or not representative;
- conditions that differ between the two measurements (time of day, equipment).

Worked examples

The bold codes show where marks are earned, as in Cambridge mark schemes. The number is how many marks.

- **M1** method mark: for a correct method, even if the numbers slip.
- **A1** accuracy mark: for a correct answer; it needs the M mark before it.
- **B1** mark for a correct result on its own, with no method needed.

Example 1

A bakery says that the median mass of its large loaves is 800 g. A customer thinks the loaves are lighter than this. The masses, in grams, of a random sample of 10 loaves are

786 803 791 779 806 783 798 789 788 805

- (a) Carry out a Wilcoxon signed-rank test at the 5% significance level to test the customer's belief. **[7]**
- (b) Carry out a sign test on the same data at the same level, and explain how its result compares with part (a). **[3]**
- (c) State an assumption needed for the test in part (a). **[1]**

(a) H_0 : population median mass = 800 g; H_1 : population median mass < 800 g. **B1**

Mass	786	803	791	779	806	783	798	789	788	805
$d = x - 800$	-14	+3	-9	-21	+6	-17	-2	-11	-12	+5
Signed rank	-8	+2	-5	-10	+4	-9	-1	-6	-7	+3

M1 for the differences, **M1** for ranking them by size.

$P = 2 + 4 + 3 = 9$ and $Q = 55 - 9 = 46$ (check: $P + Q = \frac{1}{2} \times 10 \times 11 = 55$). H_1 predicts few positive differences, so the statistic is $T = P = 9$. **A1**

The critical value for $n = 10$ at 5% one-tailed is 10. **B1**

$9 \leq 10$, so reject H_0 . **M1** There is enough evidence at the 5% level to suggest that the median mass of the loaves is less than 800 g. **A1**

(b) There are 3 plus signs out of 10. Under H_0 , $X \sim B(10, 0.5)$ and

$$P(X \leq 3) = \frac{1 + 10 + 45 + 120}{1024} = \frac{176}{1024} = 0.172.$$

B1

$0.172 > 0.05$, so do not reject H_0 : the sign test does not give enough evidence that the median is below 800 g. **M1** So the two tests disagree: only the signed-rank test finds the loaves lighter than claimed. **A1**

(The positive differences are all small; the sign test ignores that, the Wilcoxon test uses it.)

(c) The population of loaf masses is symmetrical about its median. **B1**

Example 2

Eleven people each type a passage on keyboard A and on keyboard B. Their speeds, in words per minute, are shown.

Person	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
Keyboard A	42	51	38	47	55	60	44	36	49	52	40
Keyboard B	50	47	49	45	58	54	57	37	64	57	49

(a) Carry out a Wilcoxon matched-pairs signed-rank test at the 5% significance level to test whether there is a difference in typing speed between the two keyboards. **[7]**

(b) It is found that person P4's speed on keyboard A was recorded wrongly and should have been 43. Determine whether this changes the conclusion of the test. **[2]**

(a) H_0 : the population median speeds on the two keyboards are equal; H_1 : they are not equal. **B1**

Person	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
$d = B - A$	+8	-4	+11	-2	+3	-6	+13	+1	+15	+5	+9
Signed rank	+7	-4	+9	-2	+3	-6	+10	+1	+11	+5	+8

M1 differences, **M1** ranks.

$Q = 4 + 2 + 6 = 12$, $P = 66 - 12 = 54$, so $T = 12$. **A1**

The critical value for $n = 11$ at 5% two-tailed is 10. **B1**

$12 > 10$, so do not reject H_0 . **M1** There is not enough evidence at the 5% level to suggest that typing speeds differ between the two keyboards. **A1**

(b) P4's difference becomes $45 - 43 = +2$. Its size is unchanged, so it keeps rank 2 but is now positive: $Q = 4 + 6 = 10$. **B1**

$10 \leq 10$, so H_0 is now rejected: the conclusion changes, and there is now enough evidence of a difference in speeds. **B1**

Example 3

A gardener grows 5 tomato plants with a new fertiliser and 7 without it. The heights, in cm, after six weeks are

- With fertiliser: 17.3, 18.8, 19.5, 20.3, 21.7
- Without fertiliser: 14.2, 15.0, 15.8, 16.1, 16.9, 17.6, 18.4

(a) Carry out a Wilcoxon rank-sum test at the 5% significance level to test whether the fertiliser increases the median height. **[7]**

(b) Give a reason why a paired-sample test would not be appropriate. **[1]**

(a) H_0 : the population median heights with and without fertiliser are equal; H_1 : the population median height with fertiliser is greater. **B1**

Ranking all 12 heights: without fertiliser 14.2 (1), 15.0 (2), 15.8 (3), 16.1 (4), 16.9 (5), 17.6 (7), 18.4 (8); with fertiliser 17.3 (6), 18.8 (9), 19.5 (10), 20.3 (11), 21.7 (12). **M1**

The smaller sample has $m = 5$: $R_5 = 6 + 9 + 10 + 11 + 12 = 48$. **A1**

$m(m + n + 1) - R_5 = 5 \times 13 - 48 = 17$. **M1** H_1 predicts high ranks for the fertiliser group, so the relevant value is $W = 17$ (the smaller). **A1**

The critical value for $m = 5$, $n = 7$ at 5% one-tailed is 21. **B1**

$17 \leq 21$: reject H_0 . There is enough evidence at the 5% level to suggest that the fertiliser increases the median height. **A1**

(b) The two groups are different plants, so no plant in one group belongs with a particular plant in the other (and the groups are different sizes). **B1**

Example 4

A random sample of 32 commuters record their journey times after a new timetable starts, and each is compared with the old median journey time. A Wilcoxon signed-rank test is used to test whether the median journey time has decreased. The sum of the ranks of the positive differences is 172.

(a) Carry out the test at the 5% significance level. **[6]**

(b) Find the greatest value of the test statistic that would lead to the conclusion that the median has decreased at the 1% significance level. **[4]**

(a) H_0 : the population median journey time is unchanged; H_1 : the population median journey time has decreased. **B1**

$P = 172$, so $Q = \frac{1}{2} \times 32 \times 33 - 172 = 528 - 172 = 356$, and $T = 172$ (a decrease means few positive differences, so P is the one to test).

Mean = $\frac{1}{4} \times 32 \times 33 = 264$, variance = $\frac{1}{24} \times 32 \times 33 \times 65 = 2860$. **B1**

$$z = \frac{172.5 - 264}{\sqrt{2860}} = -1.711.$$

M1 A1

$-1.711 < -1.645$, so reject H_0 . **M1** There is enough evidence at the 5% level to suggest the median journey time has decreased. **A1**

(b) Significant at 1% one-tailed means **B1** (for -2.326) **M1**:

$$\frac{T + 0.5 - 264}{\sqrt{2860}} \leq -2.326 \Rightarrow T \leq 264 - 0.5 - 2.326\sqrt{2860} = 139.1.$$

A1. T is a whole number, so the greatest value is $T = 139$. **A1**

Common mistakes

- **Hypotheses about the mean.** Writing μ , "mean", or just "median" without "population" loses the mark every year. Write "population median", or define m as the population median. "Keyboard A = keyboard B" isn't a hypothesis either: say what is equal.
- **Using the wrong test.** Two groups of different people are not matched pairs, so the matched-pairs test (and a paired t-test) is wrong for them; use the rank-sum test. And if the question names a test, use that one.
- **Differences from the wrong value.** In a single-sample test subtract the stated median m_0 from each value, not 0.
- **Slips in ranking.** Accuracy decides these questions. Set out a table (value, difference, rank, signed rank) instead of writing on the data, and check $P + Q = \frac{1}{2}n(n + 1)$.
- **The wrong column of the table.** A one-tailed test at 5% with $n = 9$ uses 8; the two-tailed value 5 is a common mistake, and so is the reverse (17 instead of 21 for $n = 13$).
- **Dropping the continuity correction** in a normal approximation, or copying the rank-sum variance with 2 instead of 12 in the denominator.
- **Mixing one- and two-tailed.** A two-tailed test at 10% compares z with ± 1.645 and a tail probability with 0.05, not -1.282 or 0.1. Keep the sign of z and the critical value consistent.
- **Sign test not known.** Many candidates could do the Wilcoxon test but not the sign test. It is just a binomial test with $p = 0.5$ on the number of + signs.
- **Vague assumptions.** For matched pairs, the assumption is about the **population of differences** being symmetrical. Comments about normality, randomness or "the data" earn nothing.
- **Conclusions that are too sure.** A test gives evidence; it never proves. Avoid "there is no evidence" as well: say "not enough evidence".
- **Words only when numbers are asked for.** "With numerical justification" means quote the new T and the critical value.

How to get the marks

- **Hypotheses (B1):** population median, both H_0 and H_1 , in context. For two populations, name both, for example " H_1 : the population median for B is greater than for A".

- **Method marks for the working (M1 M1):** show the differences and the ranks, ideally in a table. A couple of slips still earn the method marks; a bare T earns much less.
- **The test statistic (A1):** identify it clearly, " $T = 12$ " or " $W = 17$ ", as the smaller sum (or the one H_1 points to).
- **Critical value (B1):** quote it from the correct table and column.
- **Comparison (M1):** write the comparison (" $12 > 10$ ") and a result ("do not reject H_0 "). With a normal approximation, compare z with the critical z with matching signs, or a probability with the level.
- **Conclusion (A1):** in context, about the median or "average", with uncertain language. It is only given after correct working, so check your arithmetic.
- **"Show that ... opposite (or same) conclusion":** carry out both tests completely and finish with a sentence comparing the two results.
- **Reasons and assumptions (B1):** one clear sentence about the **population** or **underlying distribution**, tied to the context (an outlier in this sample, different people in the two groups).
- **Normal approximations:** write the mean and variance separately (each can earn a mark), then z to at least 3 significant figures.
- **Don't use a normal approximation when the table covers the sample sizes;** it caps your marks.

Quick check

1. Fourteen people taste two versions of a drink. Eleven prefer the new version and three prefer the old. Use a sign test at the 5% significance level to test whether people in general prefer the new version.
2. A random sample of 8 values is 25.4, 23.7, 27.1, 25.7, 24.8, 26.8, 27.6, 26.1. Use a Wilcoxon signed-rank test at the 10% significance level to test whether the population median differs from 25.
3. Independent samples of 14 and 18 values are ranked together. The sum of the ranks of the sample of 14 is 180. Test at the 5% significance level whether the two population medians differ.
4. Which test would you use? (a) Blood pressure of 9 patients before and after a treatment, with one very large change. (b) Marks of 8 students in one class and 10 students in another, from populations of unknown shape. (c) Reaction times of 12 people, compared with a stated median, from a population thought to be symmetrical.
5. A Wilcoxon signed-rank test on 40 paired differences gives $T = 300$. Test at the 5% significance level whether the two population medians differ.

Answers

1. H_0 : in the population, a person is equally likely to prefer either version; H_1 : the new version is more likely to be preferred. $X \sim B(14, 0.5)$: $P(X \geq 11) = P(X \leq 3) = \frac{1+14+91+364}{16384} = 0.0287 < 0.05$. Reject H_0 : there is enough evidence that people in general prefer the new version.
2. Differences from 25: +0.4, -1.3, +2.1, +0.7, -0.2, +1.8, +2.6, +1.1. Signed ranks: +2, -5, +7, +3, -1, +6, +8, +4. $Q = 5 + 1 = 6$, $P = 36 - 6 = 30$, $T = 6$. Critical value for $n = 8$, 10% two-tailed: 5. $6 > 5$: do not reject H_0 ; not enough evidence that the population median differs from 25.
3. $m = 14$, $n = 18$: mean = $\frac{1}{2} \times 14 \times 33 = 231$, variance = $\frac{1}{12} \times 14 \times 18 \times 33 = 693$. 180 is below the mean, so add 0.5: $z = \frac{180.5-231}{\sqrt{693}} = -1.918$. Two-tailed 5%: $-1.918 > -1.96$, so do not reject H_0 ; not enough evidence that the population medians differ (it would be significant in a one-tailed test).
4. (a) A paired-sample sign test: the extreme change suggests the differences aren't symmetrical. (b) A Wilcoxon rank-sum test: two separate groups, sizes within the table. (c) A single-sample Wilcoxon signed-rank test: symmetry makes it valid, and it uses more information than the sign test.
5. Mean = $\frac{1}{4} \times 40 \times 41 = 410$, variance = $\frac{1}{24} \times 40 \times 41 \times 81 = 5535$. $z = \frac{300.5-410}{\sqrt{5535}} = -1.472$. Two-tailed 5%: $-1.472 > -1.96$, so do not reject H_0 ; not enough evidence that the population medians differ.

Past questions to try

- 9231/42/M/J/23 Q4: a matched-pairs test, then a corrected value with numerical justification.
- 9231/42/M/J/24 Q2: a paired-sample sign test.
- 9231/42/O/N/24 Q6: a single-sample signed-rank test with a normal approximation.
- 9231/43/O/N/25 Q4: the least test statistic for a given outcome.