NST 1B Experimental Psychology
Statistics practical 3

Difference tests (2): nonparametric

Rudolf Cardinal & Mike Aitken
10 / 11 February 2005; Department of Experimental Psychology
University of Cambridge

Slides at
pobox.com / ~rudolf / psychology

These slides are on the web.
No need to scribble frantically.

pobox.com / ~rudolf / psychology

Nonparametric tests

Last time, we looked at the *t* test, a **parametric** test — it made assumptions about parameters of the underlying populations (such as the distribution — e.g. assuming that the data are **normally distributed**).

If these assumptions are violated:

(a) we could *transform* the data to fit the assumptions better (NOT covered at Part 1B level)

or (b) we could use a **nonparametric** ('distribution-free') test that doesn't make the same assumptions.

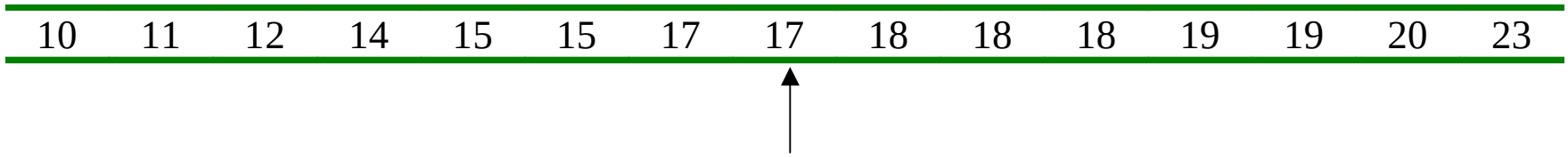
In general, if the assumptions of parametric tests are met, they are the most powerful. If not, we may need to use nonparametric tests. They may, for example, answer questions about medians rather than means. We'll look at some nonparametric tests now that assume only that the data are measured on at least an ordinal scale.

The median

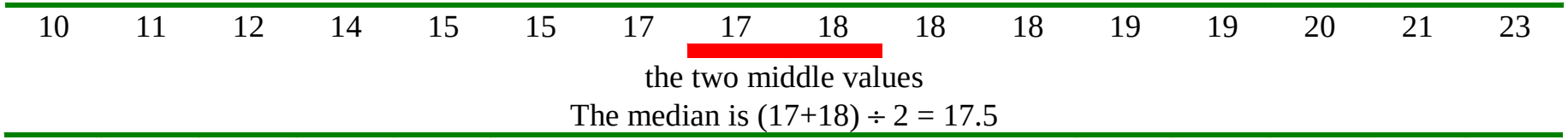
The **median** is the value at or below which 50% of the scores fall when the data are arranged in numerical order.

(Can also be referred to as the 50^{th} centile.)

If *n* is odd, it's the middle value (here, 17):



If *n* is even, it's the mean of the two middle values (here, 17.5):



The median

The **median** is the value at or below which 50% of the scores fall when the data are arranged in numerical order.

(Can also be referred to as the 50th centile.)

Medians are less affected by outliers than means

Ranking removes ‘distribution’ information

Ranking removes information about the distribution.
Whatever the distribution (normal, flat, skewed, bimodal…), the ranks are the same:
1, 2, 3, 4, 5, 6, 7.

How to rank data

Suppose we have ten measurements (e.g. test scores) and want to rank them. First, place them in ascending numerical order:

5	8	9	12	12	15	16	16	16	17

Then start assigning them ranks. When you come to a tie, give each value the mean of the ranks they're tied for — for example, the 12s are tied for ranks 4 and 5, so they get the rank 4.5; the 16s are tied for ranks 7, 8, and 9, so they get the rank 8:

X:	5	8	9	12	12	15	16	16	16	17
rank:	1	2	3	4.5	4.5	6	8	8	8	10

Nonparametric correlation

We've already seen this.

r_s: Spearman's correlation coefficient for **ranked** data

We met this in the first statistics practical. It's a nonparametric version of correlation. You can use it when you obtain ranked data, or when you want to do significance tests on *r* but your data are not normally distributed (violating an assumption of the parametric *t* test that's based on Pearson's *r*).

- Rank the *X* values.
- Rank the *Y* values.
- Correlate the *X* **ranks** with the *Y* **ranks.** (You do this in the normal way for calculating *r*, but you call the result r_s.)
- To ask whether the correlation is 'significant', use the table of critical values of Spearman's r_s in the *Tables and Formulae* booklet.

Nonparametric difference tests

Two unrelated samples: the Mann–Whitney *U* test

Logic

• Suppose we have two samples with n_1 and n_2 observations in each (for a total of $n_1 + n_2 = N$ observations).

• We can rank all observations together, from 1 to N.

• If the two samples come from identical populations, the **sum of the ranks of 'sample 1' scores** is likely to be about the same as the **sum of the ranks of 'sample 2' scores.**

• But if sample 1 comes from a population with much lower values than sample 2, the sum of the ranks of 'sample 1' scores will generally be lower than the sum of the ranks of 'sample 2' scores.

Null hypothesis: the two samples were drawn from identical populations. [Unlike the unpaired *t* test, whose null hypothesis was that the two samples came from populations with the same *mean*.] If we assume the distributions are similar, a significant Mann–Whitney test suggests that the **medians** of the two populations are different.

Calculating the Mann–Whitney *U* statistic

From the Formula Sheet:

Calculating the Mann–Whitney U statistic

1. Call the smaller group ‘group 1’, and the larger group ‘group 2’, so $n_1 < n_2$. (If $n_1 = n_2$, ignore this step.)
2. Calculate the sum of the ranks of group 1 (= R_1) and group 2 (= R_2).
3. $U_1 = R_1 - \dfrac{n_1(n_1+1)}{2}$
4. $U_2 = R_2 - \dfrac{n_2(n_2+1)}{2}$
5. The Mann–Whitney statistic U is the smaller of U_1 and U_2.

Check your sums: verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \dfrac{(n_1+n_2)(n_1+n_2+1)}{2}$.

Mann–Whitney *U* test: EXAMPLE 1

Mothers either received care from **first trimester** onwards — birth weights (in kg):

{1.68, 3.83, 3.11, 2.76, 1.70, 2.79, 3.05, 2.66, 1.40, 2.775}

or from the **third trimester** onwards — babies' birth weights:

{2.94, 3.38, 4.90, 2.81, 2.80, 3.21, 3.08, 2.95}.

Is there a significant difference between the birthweights of the two groups?

- Third trimester group smaller, so is **group 1 (n_1 = 8).** Other is **group 2 (n_2 = 10).**
- **Ranks for group 1:** {10, 16, 18, 9, 8, 15, 13, 11}; **rank sum R_1 = 100.**
- **Ranks for group 2:** {2, 17, 14, 5, 3, 7, 12, 4, 1, 6}; **rank sum R_2 = 71.**

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 100 - \frac{8\times 9}{2} = 64$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} = 71 - \frac{10\times 11}{2} = 16$$

$$U = 16 \quad \text{(the smaller of the two)}$$

Check sums:

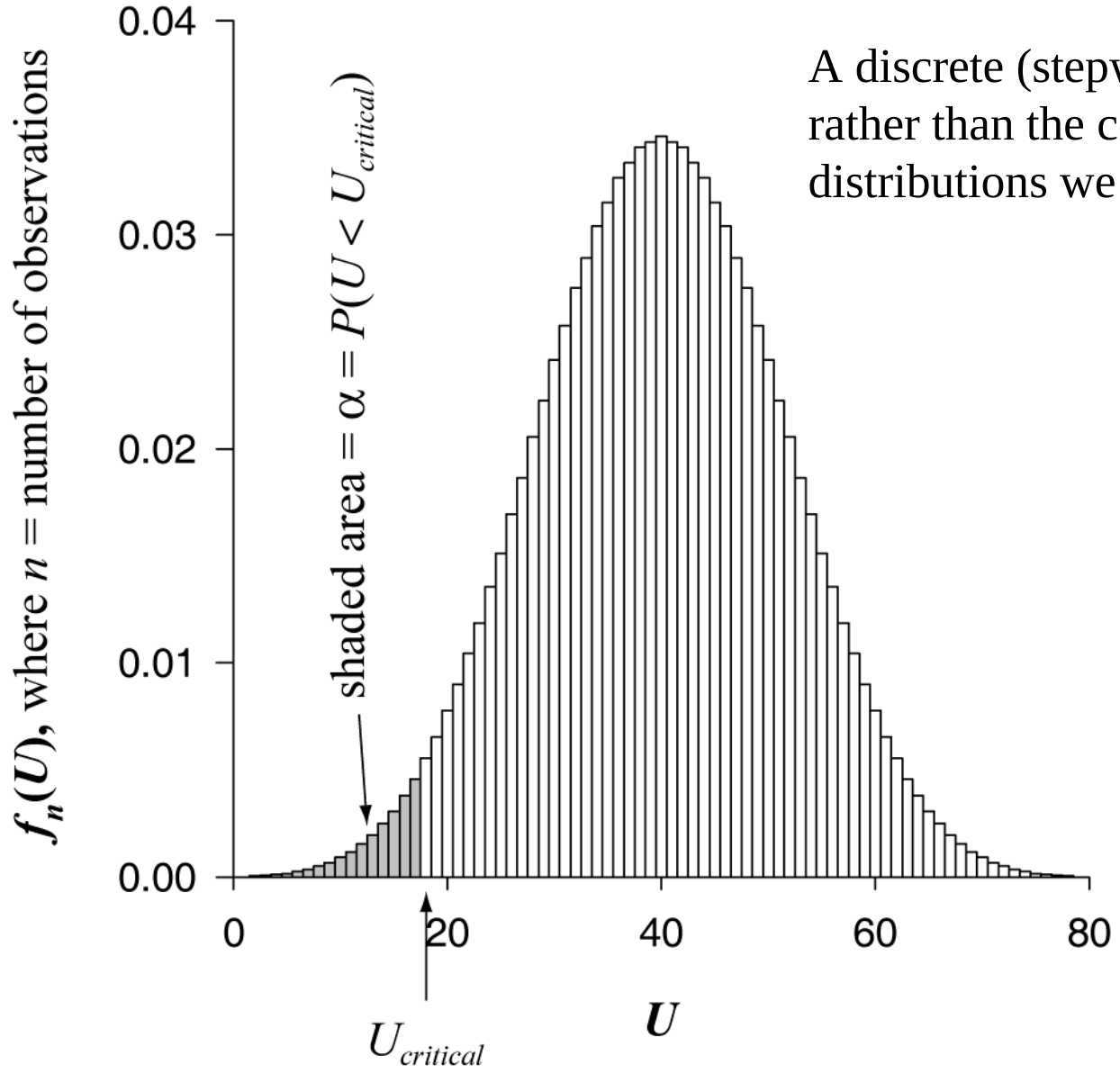
$$U_1 + U_2 = 64 + 16 = 80$$

$$n_1 n_2 = 8\times 10 = 80 \quad \ldots \text{ good, they match}$$

$$R_1 + R_2 = 100 + 71 = 171$$

$$\frac{(n_1+n_2)(n_1+n_2+1)}{2} = \frac{(8+10)(8+10+11)}{2} = 171 \quad \ldots \text{ good, they do too}$$

Distribution of the Mann–Whitney *U* statistic (if H_0 is true)



A discrete (stepwise) distribution, rather than the continuous distributions we've looked at before.

Determining a significance level (*p* value) from *U*

If $n_2 \leq 20$, look up the **critical value** for *U* in your tables. (The critical value depends on n_1 and n_2.)

If your *U* is **smaller** than the critical value, it's significant (you reject the null hypothesis).

If $n_2 > 20$, the tables don't give you critical values, but by this point the *U* statistic is approximately **normally distributed,** so we can calculate a ***Z* score** from *U* and test that in the usual way, using tables of *Z*.

The formula for calculating *Z* from *U* is on the Formula Sheet.

Our example: $U = 16$, $n_1 = 8$, $n_2 = 10$; critical value of *U* is 18 (from tables). Our *U* less than this, so **birthweight difference was significant** ($p < 0.05$ two-tailed).

Mann–Whitney *U* test: EXAMPLE 2 (a)

Transfer along a continuum practical (**a previous year's** data, I'm afraid). Different groups of subjects were trained with {1 or 3} blocks of {easy or hard} discriminations before being tested on similar discriminations. Here are the test scores for the 3-block groups (high = good). Is there an effect of training difficulty? You could use either an unpaired *t* test or a Mann–Whitney *U* test; try the latter.

3 Blocks Easy	15	20	30	40	42.5	45	50	50	55	55	65	65	75	80	82.5
3 Blocks Hard	–10	–5	7.5	10	15	25	27.5	30	35	35	45	50	50	57.5	60

1. Call the smaller group 'group 1', and the larger group 'group 2', so $n_1 < n_2$. (If $n_1 = n_2$, ignore this step.)
2. Calculate the sum of the ranks of group 1 (= R_1) and group 2 (= R_2).
3. $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$
4. $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$
5. The Mann–Whitney statistic U is the smaller of U_1 and U_2.

Check your sums: verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$.

Mann–Whitney *U* test: EXAMPLE 2 (b)

3 Blocks Easy	15	20	30	40	42.5	45	50	50	55	55	65	65	75	80	82.5
Ranks	**5.5**	**7**	**10.5**	**14**	**15**	**16.5**	**19.5**	**19.5**	**22.5**	**22.5**	**26.5**	**26.5**	**28**	**29**	**30**
3 Blocks Hard	–10	–5	7.5	10	15	25	27.5	30	35	35	45	50	50	57.5	60
Ranks	**1**	**2**	**3**	**4**	**5.5**	**8**	**9**	**10.5**	**12.5**	**12.5**	**16.5**	**19.5**	**19.5**	**24**	**25**

- Both groups same size. Arbitrarily, call the Easy group 'group 1' **(n_1 = 15).** Hard group is 'group 2' **(n_2 = 15).**
- **Rank sum R_1** = 5.5 + 7 + … + 30 = **292.5**
- **Rank sum R_2** = 1 + 2 + … + 25 = **172.5**

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 292.5 - \frac{15\times16}{2} = 172.5$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} = 172.5 - \frac{15\times16}{2} = 52.5$$

$U = 52.5$ (the smaller of the two)

Critical U ($n_1 = n_2 = 15$) for $\alpha = 0.05$ two-tailed is 65. So **significant.**

Check sums:

$$U_1 + U_2 = 172.5 + 52.5 = 225$$

$$n_1 n_2 = 15\times15 = 225 \quad \text{… good, they match}$$

$$R_1 + R_2 = 292.5 + 172.5 = 465$$

$$\frac{(n_1+n_2)(n_1+n_2+1)}{2} = \frac{(15+15)(15+15+1)}{2} = 465 \quad \text{… good, they do too}$$

Last year's (2003) data

Time-saving tip…

If the ranks **do not overlap at all,** $U = 0$.

Example:

Group A	55	65	75	80	82.5
Ranks	**6**	**7**	**8**	**9**	**10**
Group B	10	15	39	40	48
Ranks	**1**	**2**	**3**	**4**	**5**

$$U = 0$$

If you find a significant difference…

If you conduct a Mann–Whitney test and find a significant difference, **which group had the larger median and which group had the smaller median?**

group 1	40	41	43							median = 41
ranks 1	**9**	**10**	**12**							rank sum $R_1 = 31$
group 2	5	7	9	10	11	15	16	17	42	median = 11
ranks 1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**11**	rank sum $R_2 = 47$

Here, $U = 2$. Significant (critical $U = 3$, $\alpha = 0.05$ two-tailed).

Group 1 has a significantly larger median (even though the rank sums convey the opposite impression).

You have to calculate the medians. But this is quick.

Two related samples: Wilcoxon matched-pairs signed-rank test

Logic

• Suppose we have a set of *n* **paired** scores — for each subject, say, we have one score from condition 1 and one score from condition 2.

• We can calculate the difference score $condition_1 - condition_2$ for each pair. Then we **rank** the non-zero differences.

• If, on average, there is no difference between performance in condition 1 and condition 2, then the **sum of the ranks of the positive differences** should be about the same as the **sum of the ranks of the negative differences.**

• But if there *is* a difference between condition 1 and condition 2, the + and – rank sums should differ.

Null hypothesis: the distribution of differences between the pairs of scores is symmetric about zero. Since the median and mean of a symmetric population are the same, this can be restated as 'the differences between the pairs of scores are symmetric with a mean and median of zero'.

Calculating the Wilcoxon *T* statistic

Easy. From the Formula Sheet:

Calculating the Wilcoxon matched-pairs signed-rank statistic, T

1. Calculate the difference scores.
2. Ignore any differences that are zero.
3. Rank the difference scores, *ignoring their sign* (+ or –).
4. Add up all the ranks for difference scores that were positive; call this T^+.
5. Add up all the ranks for difference scores that were negative; call this T^-.
6. The Wilcoxon matched-pairs statistic T is the smaller of T^+ and T^-.

Check your sums: verify that $T^+ + T^- = \frac{n(n+1)}{2}$.

Wilcoxon matched-pairs signed-rank test: EXAMPLE 1

Measure blood pressure (BP_1). Make subjects run a lot. Measure blood pressure again (BP_2) in the same subjects. Has their blood pressure changed?

Before (BP_1):	130	148	170	125	170	130	130	145	119	160
After (BP_2):	120	148	163	120	135	143	136	144	119	120
Difference ($BP_1 - BP_2$):	10	0	7	5	35	–13	–6	1	0	40
Rank of difference (ignoring zero differences and sign):	5		4	2	7	6	3	1		8
'Signed rank'	*5*		*4*	*2*	*7*	*–6*	*–3*	*1*		*8*
Ranks of positive differences:	5		4	2	7			1		8
Ranks of negative differences:						6	3			

Sum of positive ranks $\boldsymbol{T^+} = 5 + 4 + 2 + 7 + 1 + 8 = \mathbf{27}$

Sum of negative ranks $\boldsymbol{T^-} = 6 + 3 = \mathbf{9}$

Wilcoxon T = the smaller of T^+ and T^- = **9.** And $n = 8$.

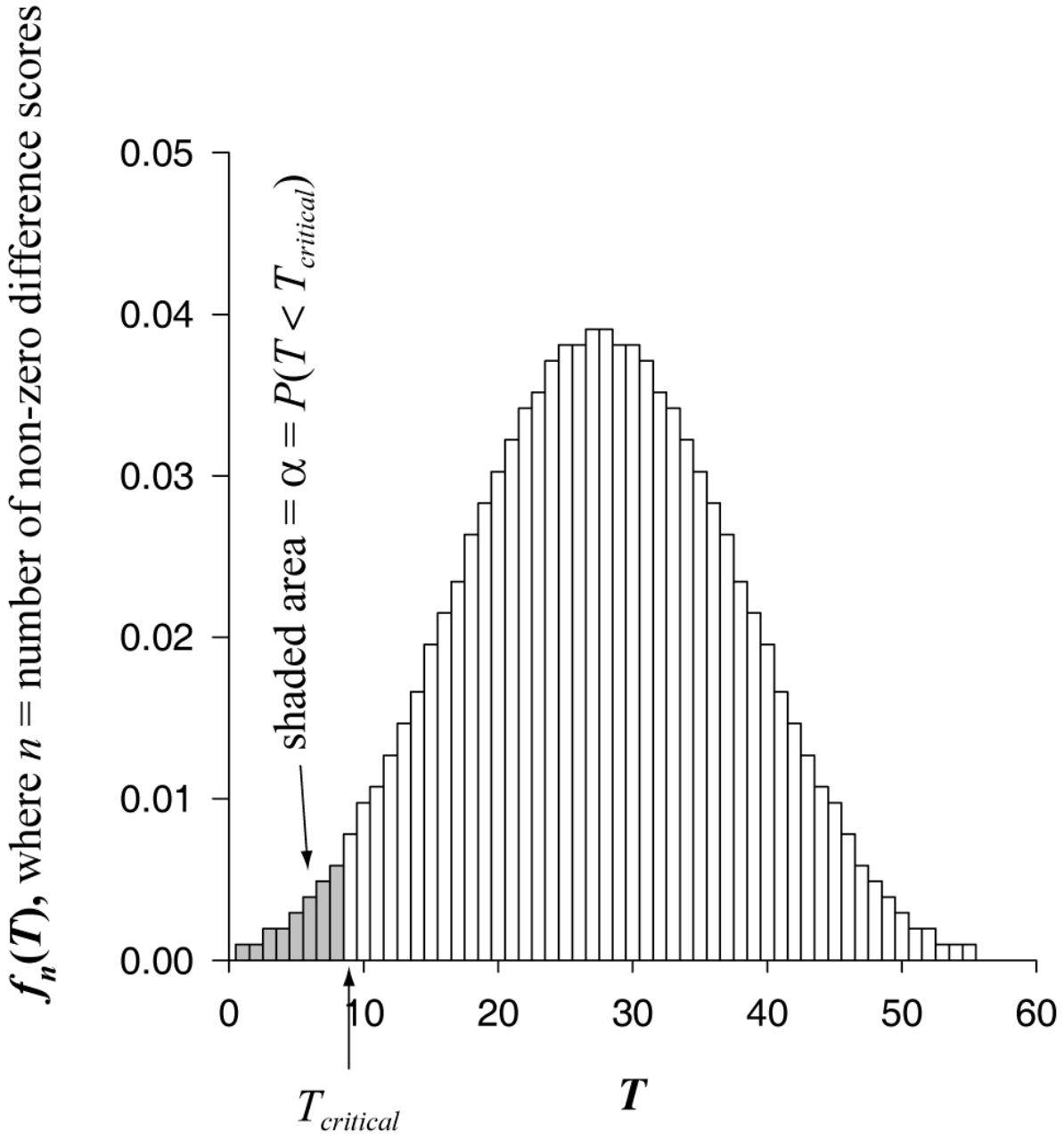
Check sums:

$$T^+ + T^- = 27 + 9 = 36$$

$$\frac{n(n+1)}{2} = \frac{8 \times 9}{2} = 36$$

… good, they match

Distribution of the Wilcoxon *T* statistic (if H_0 is true)



Determining a significance level (*p* value) from *T*

If $n \leq 25$, look up the **critical value** for *T* in your tables. (The critical value depends on *n*.)

If your *T* is **smaller** than the critical value, it's significant (you reject the null hypothesis).

If $n > 25$, the tables don't give you critical values, but by this point the *T* statistic is approximately **normally distributed,** so we can calculate a ***Z*** **score** from *T* and test that in the usual way, using tables of *Z*.

The formula for calculating *Z* from *T* is on the Formula Sheet.

In our example, $T = 9$ and $n = 8$. Critical value of *T* is 4 (for $\alpha = 0.05$ two-tailed); since our *T* is **not** smaller than this, the BP difference was **not** significant.

Wilcoxon matched-pairs signed-rank test: EXAMPLE 2 (a)

Proactive interference practical (subset of **a previous year's** data, I'm afraid). Subjects hear and repeat trigram (e.g. CXJ), perform distractor task, recall trigram. Compare trials 9 & 10 (after many similar trigrams) with trials 11 & 12 (after shift to new *type* of trigram, e.g. 925). Is there 'release' from proactive interference? Note very non-normal difference scores; parametric test unsuitable.

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
% correct trials 9 & 10	0	0	0	0	50	50	0	0	50	0	50	50
% correct trials 11 & 12	100	100	100	100	100	100	50	50	100	50	100	100
Subject	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
% correct trials 9 & 10	50	50	50	100	100	100	50	50	50	100	100	100
% correct trials 11 & 12	100	100	100	100	100	100	50	50	0	50	50	50

The procedure is:

1. Calculate the difference scores.
2. Ignore any differences that are zero.
3. Rank the difference scores, *ignoring their sign* (+ or –).
4. Add up all the ranks for difference scores that were positive; call this T^+.
5. Add up all the ranks for difference scores that were negative; call this T^-.
6. The Wilcoxon matched-pairs statistic T is the smaller of T^+ and T^-.

Check your sums: verify that $T^+ + T^- = \frac{n(n+1)}{2}$.

Wilcoxon matched-pairs signed-rank test: EXAMPLE 2 (b)

subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
% correct trials 9 & 10	0	0	0	0	50	50	0	0	50	0	50	50	50	50	50	100	100	100	50	50	50	100	100	100
% correct trials 11 & 12	100	100	100	100	100	100	50	50	100	50	100	100	100	100	100	100	100	100	50	50	0	50	50	50
Difference (9&10 – 11&12)	–100	–100	–100	–100	–50	–50	–50	–50	–50	–50	–50	–50	–50	–50	–50	0	0	0	0	0	+50	+50	+50	+50
Rank of difference (ignoring zero differences and sign):	17.5	17.5	17.5	17.5	8	8	8	8	8	8	8	8	8	8	8						8	8	8	8
Rank of positive differences:																					8	8	8	8
Rank of negative differences:	17.5	17.5	17.5	17.5	8	8	8	8	8	8	8	8	8	8	8									

Sum of positive ranks $\boldsymbol{T^+} = (4 \times 8) = \mathbf{32}$

Sum of negative ranks $\boldsymbol{T^-} = (4 \times 17.5) + (11 \times 8) = \mathbf{158}$

Wilcoxon T = the smaller of T^+ and T^- = **32.** And n = 19.

Check sums:

$$T^+ + T^- = 32 + 158 = 190$$

$$\frac{n(n+1)}{2} = \frac{19 \times 20}{2} = 190$$

… good, they match

From our tables, for n = 19 and $\alpha = 0.05$ two-tailed, the critical value of T is 47. Our T is smaller, so the difference is significant. (In fact, it's significant at $\alpha = 0.01$ two-tailed.) Subjects did better on trials 11 & 12 than on trials 9 & 10.

Last year's (2003) data

One sample: Wilcoxon signed-rank test with only one sample

Very easy.
Null hypothesis: the median is equal to M.

For each score x, calculate a difference score ($x - M$). Then proceed as for the two-sample Wilcoxon test using these difference scores.

(Logic: if the median is M, then the sum of the ranks of the positive differences — from scores where $x > M$ — should be the same as the sum of the ranks of the negative differences — from scores where $x < M$. If the median isn't M, then the two rank sums should differ.)

Calculating the Wilcoxon matched-pairs signed-rank statistic, T

The procedure is:

1. Calculate the difference scores.
2. Ignore any differences that are zero.
3. Rank the difference scores, *ignoring their sign* (+ or –).
4. Add up all the ranks for difference scores that were positive; call this T^+.
5. Add up all the ranks for difference scores that were negative; call this T^-.
6. The Wilcoxon matched-pairs statistic T is the smaller of T^+ and T^-.

Comparison of parametric and non-parametric tests

Parametric test	Equivalent nonparametric test
Two-sample unpaired t test	Mann–Whitney *U* test
Two-sample paired t test	Wilcoxon signed-rank test with matched pairs
One-sample t test	Wilcoxon signed-rank test, pairing data with a fixed value


© Maki Kawai