Understanding Statistics and Experimental Design

Exercise about non-parametric tests

1. Ten people who suffer from chronic pain are given a rating scale that measures the intensity of their pain. They are then given a drug (actually a placebo) and encouraging information concerning its effectiveness. After taking the placebo, they again rate the intensity of their pain. Their ratings before and after the placebo is administered are listed below:

Patient	Before	After
1	48	35
2	27	29
3	36	29
4	44	21
5	22	24
6	35	30
7	29	27
8	48	37
9	25	28
10	32	24

a) Do the data meet the assumption of the homogeneity of variances?

Solution. Computing the standard deviations, we find that $s_{before} = 9.407$ while $s_{after} = 4.904$. These differences are quite large in magnitude given the range of data we are working with – if you wish to verify the result statistically, you may conduct e.g. Levene's test for the equality of variances (not covered in class).

b) Does the researcher have an *a priori* reason to expect the drug reduces people's ratings of pain? How does this guide the selection of an appropriate statistical test?

Solution. Either yes, since placebo effects are known to work in a large number of cases – or no, since any effect should only be driven by the placebo and not an intervention of the actual drug. Either answer suffices given a correct justification.

c) Was the drug effective at reducing people's ratings of pain? Choose the appropriate test and use statistical software such as JASP to conduct the test, with $\alpha = 0.05$. Interpret the results appropriately.

Solution. We will use the non-parametric Wilcoxon signed-rank test since the data considered are ordinal scale (can be ranked). However, it is important to note that the homogeneity of variances is not an assumption in selecting this test since the data in question are dependent. Conducting the test in JASP shows that a two-tailed Wilcoxon signed- rank test on a difference in before and after drug treatment has $p = 0.052 > \alpha$, and thus we cannot reject the null hypothesis.

d) BONUS: Compute the statistical test without the use of statistical software.

Solution. First, we state our hypotheses:
H_0: There is no systematic difference in measurements before and after the treatment.
H_1: There is a systematic difference in measurements before and after the treatment.
Set $\alpha = 0.05$.

Second, we compute the differences and the ranks R of the absolute differences. In the case of shared ranking, we apply the median ranking to all.

Patient	Before	After	Difference	Rank of differences
1	48	35	13	9
2	27	29	-2	2
3	36	29	7	6
4	44	21	23	10
5	22	24	-2	2
6	35	30	5	5
7	29	27	2	2
8	48	37	11	8
9	25	28	-3	4
10	32	24	8	7

Second, we sum the ranks of the positive and negative differences, respectively:
$\sum R_+ = 9 + 6 + 10 + 5 + 2 + 8 + 7 = 47$
$\sum R_- = 2 + 2 + 4 = 8$

Finally, we select the smaller of these values, the W-statistic, and look at the critical number in a W-table. We find that the value of 8 for a 2-tailed Wilcoxon test with $n = 10$ and $\alpha = 0.05$ has a critical value of 8. We could reject the null hypothesis on the basis of this result, but we recommend that since the table is approximate, in this edge case we conduct a precise test using statistical software.

Exercises about effect size and statistical power

These questions use statistics derived from the article:

Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an embodiment of importance. Psychological Science, 20(9), 1169-1174.

2. The main result in Study 1 is an F-test comparing participants' judgments of the value of foreign currencies depending on whether a clipboard they were holding was light or heavy. The effect size, partial-eta-squared (η_p^2), is reported as part of the results.

a) What other information, in addition to an effect size, is required for computing a statistical test's *retrospective* (sometimes called *post-hoc*) power?

Solution. In addition to the effect size, we require the sample size, the level of desired statistical significance, and the relevant test.

b) What are the benefits of computing a statistical test's retrospective power? Are there any dangers?

Solution. A benefit is that it helps us understand how well-designed the experiment was, and whether the experiment had sufficient power to find a statistically significant result. Some dangers include underestimation or overestimation of the effect size by chance leading to imprecise measurement of power, and too much focus on post-hoc power instead of other relevant results.

c) Using the statistical software G*Power, compute the retrospective power of Study 1 in Jostmann et al. (2009).

Solution. $\eta_p^2 = 0.12 \Rightarrow power = 0.624$.

3. Study 1 reports an F-test and a partial-eta-squared effect size. We would like to interpret the effect size in a manner more familiar to our Signal Detection Theory framework in the course.

a) The test statistic of the two-sample F-test can be related to the test statistic of a two-sample t-test, using which formula?

Solution. $F = t^2$.

b) Compute the Cohen's d effect size using the t-statistic obtained in step a).

Solution. $n_1 = n_2 = 20, F = 4.86$.
$t = \sqrt{F} = \sqrt{4.86} = 2.204541$
Cohen's d is given by:

$$d = t\sqrt{\frac{n_1 + n_2}{n\ n}} = 2.204541\sqrt{\frac{20 + 20}{20 * 20}} = 0.697137$$

Thus, the Cohen's d effect size is 0.697137, a moderate effect.