Exercise 27.8 of PSLS 3e

Data: 2 samples of corn yields (bushels per acre) for 4 plots without weeds and 4 plots with 9 lamb's-quarter plants per meter of row.

Model: the 2 samples are independent and each a simple random sample (i.i.d. sample) from a distribution with unknown mean, median and standard devation.

Estimation: in each sample, we use the sample statistics (e.g., mean, median, standard deviation); considering the very small sample sizes there is probably little interest in separate scrutiny of those estimates.

(a) When using the Wilcoxon-Mann-Whitney test we may make no further assumptions about the distributions, and test

$$\mathrm{H}_0 : \mathcal{P}_1 = \mathcal{P}_2 \quad \text{versus} \quad \mathrm{H}_a : \mathcal{P}_1 \text{ is systematically larger than } \mathcal{P}_2,$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ are the distributions of corn yields in the two populations (weed-free and weed-filled plots). Alternatively, we may make the "Δ-assumption" that the two distributions are of the same shape (only differ by their location), and test the hypotheses

$$\mathrm{H}_0 : \mathrm{median}_1 = \mathrm{median}_2 \quad \text{versus} \quad \mathrm{H}_a : \mathrm{median}_1 > \mathrm{median}_2,$$

with the alternative expressing higher yields in weed-free plots. The motivation for the one-sided alternatives lies in the wording of the question ("evidence that 9 weeds per meter *reduces* corn yields"), but in practice one could also justify a two-sided alternative. Given the low sample sizes it is perhaps tempting to increase the power of the statistical analysis by using a one-sided alternative.

Minitab commands and output for the test (note that the menu requires the two samples to be in separate columns):

```
Unstack ('yield');
 Subscripts 'weeds';
 After;
 VarNames.
Mann-Whitney 95.0 'yield_0' 'yield_9';
 Alternative 1.
```

Mann-Whitney: yield_0, yield_9

Method

η_1: median of yield_0
η_2: median of yield_9
Difference: $\eta_1 - \eta_2$

Descriptive Statistics

Sample	N	Median
yield_0	4	169.45
yield_9	4	162.55

Estimation for Difference

Difference	Lower Bound for Difference	Achieved Confidence
9.65	2.3	96.97%

Test

Null hypothesis	H_0: $\eta_1 - \eta_2 = 0$
Alternative hypothesis	H_1: $\eta_1 - \eta_2 > 0$

W-Value	P-Value
26.00	0.015

Comments:
The Wilcoxon rank sum test (or Mann-Whitney-Wilcoxon test) gives an approximate P-value of 0.015 for H_0 against a one-sided alternative. This means that there is some evidence that 9 weeds per meter systematically reduce the corn yield. (Note: the exact P-value is 0.014, obtained with another software.)

(b) As the focus here is mostly on the non-parametric analysis, we just give the Minitab commands and output, and add our interpretation.

```
TwoSample 'yield_0' 'yield_9';
 Confidence 95.0;
 Test 0.0;
 Alternative 1.
```

Two-Sample T-Test and CI: yield_0, yield_9

Method

μ_1: mean of yield_0
μ_2: mean of yield_9
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
yield_0	4	170.20	5.42	2.7
yield_9	4	157.6	10.1	5.1

Estimation for Difference

Difference	95% Lower Bound for Difference
12.62	0.39

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
2.20	4	0.046

Comments:
The t-test gives a P-value for the one-sided test of 0.046, just below the 5% significance limit. It is really on the border of 5% significance, which we might express as weak evidence. Note that we do not assume the variances to be equal because of the larger variation in the second sample (due to the outlier).

(c) In order to carry out the analyses without the outlier, we need to replace that value by a missing value ($*$ in Minitab). I recommend to do this in a new column, so that the original value does not get lost, and it becomes visible in the listing that the results are obtained from new (reduced) data.

```
Copy 'yield_9' c7;
 Varnames.
let c7(2)='*'
Mann-Whitney 95.0 'yield_0' 'yield_9_1';
 Alternative 1.
TwoSample 'yield_0' 'yield_9_1';
 Confidence 95.0;
 Test 0.0;
 Alternative 1.
```

Mann-Whitney: yield_0, yield_9_1

Method

η_1: median of yield_0
η_2: median of yield_9_1
Difference: $\eta_1 - \eta_2$

Descriptive Statistics

Sample	N	Median
yield_0	4	169.45
yield_9_1	3	162.70

Estimation for Difference

Difference	Lower Bound for Difference	Achieved Confidence
6.85	2.2	97.41%

Test

Null hypothesis	H_0: $\eta_1 - \eta_2 = 0$
Alternative hypothesis	H_1: $\eta_1 - \eta_2 > 0$

W-Value	P-Value
22.00	0.026

Two-Sample T-Test and CI: yield_0, yield_9_1

Method

μ_1: mean of yield_0
μ_2: mean of yield_9_1
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
yield_0	4	170.20	5.42	2.7
yield_9_1	3	162.633	0.208	0.12

Estimation for Difference

Difference	95% Lower Bound for Difference
7.57	1.18

Test

Null hypothesis	H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis	H_1: $\mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
2.79	3	0.034

Comments:
The outlier reduced the mean yield in the 9 meter group by 5 units (bushels per acre); the value is obtained as $162.6 - 157.6 = 5.0$. It increased the standard deviation by a factor of approximately 50 (10.1/0.208). However, the results in both analyses with and without the outlier are surprisingly similar, although the P-value increases in the non-parametric analyses and decreases with the t-test (why? – answer below).

From this exercise one may get the idea that the Wilcoxon rank sum test is more powerful than the t-test in small samples. That is not true in general, and the P-value obtained by the Wilcoxon rank sum test is in fact the smallest possible with these sample sizes (why? –answer below).

Additional questions
First, we would expect the P-value to increase when removing the outlier because it reduces both the difference between the groups and the power for the test; therefore, the behaviour of the MWW test is as expected. The reason that the t-value gets substantially more extreme (from 2.20 to 2.79) is that the standard deviation is much more strongly affected by the outlier than the mean.

Second, the rank sums for the two groups are the most extreme possible for a dataset with 4 observations in each group. That is because every observation in the weed 0 sample is larger than any of the observations in the weed 9 sample. In this case, the ranks for the weed 9 sample become 1, 2, 3 and 4 (sum = 10), and the ranks for the weed 0 sample become 5, 6, 7 and 8 (sum = 26). No matter the actual values, the ranks in the two groups can never be more different, and therefore the P-value is the smallest possible with two samples of size 4.