Rank-based and nonparametric methods

Patrick Breheny

Rank-based methods

- At the conclusion of the previous lecture, I alluded to using methods that are "robust to the presence of outliers": what are our options?
- A very effective and widely used approach is to use another kind of transformation, and work with the ranks of the data instead of the actual observations themselves
- By ranking the data, the impact of outliers is mitigated: regardless of how extreme an outlier is, it receives the same rank as if it were just slightly larger than the second-largest observation
- Also, any problem of skewness is eliminated, because all ranks are equally far apart from each other

Tailgating ranks

For example, instead of looking at the actual following distances, we could look at the ranks of the following distances:

Distance	Rank
17.89	2
38.96	88
38.31	85
28.58	40
27.70	33
49.76	104
28.91	44
20.38	9

t-test on the ranks

- Now that we've got ranks, we could perform a two-sample t-test on the ranks instead of the actual data
- It turns out that, if you do, you obtain a p-value of 0.02
- This is a much more valid way of achieving statistical significance than throwing away outliers — no arbitrary decisions about which observations to throw away were made

The Mann-Whitney/Wilcoxon test

- The more common approach to testing ranks, however, is to use a permutation test of the ranks (we discussed permutation tests in the "Two sample inference: Continuous data" lecture)
- This approach to hypothesis testing (rank-then-permutation-test) is called either the *Mann-Whitney U test* or the *Wilcoxon rank-sum test*; I'll use the two interchangeably, or use the abbreviation "MWW test"
- It is a very common approach to testing for differences between two groups when one is concerned about normality/skewness/outliers — any of the things that can cause problems with the t-test

Calculating the Mann-Whitney/Wilcoxon test

- As we discussed previously, permutation tests are labor-intensive
- It is possible to carry out an approximate version of the test by hand based on the idea that the sum of the ranks approximately follows a normal distribution
- We won't concern ourselves with the details of this approximation in this class (you can look up those details elsewhere if you are really curious), but will focus on:
  - The concept behind the test (transforming the data by ranking + permutation test)
  - Perform the Mann-Whitney/Wilcoxon test using a computer

Computational considerations

- It is worth pointing out that statisticians have developed clever ways of calculating exact p-values for permutation tests in the special case where the data are consecutive positive integers (i.e., ranks) that are much faster than the brute force permutation test approach
- Thus, many software packages will offer an option to calculate exact p-values for the Mann-Whitney/Wilcoxon test
- This is usually quite fast, although for very large sample sizes it can still be computer-intensive, so software packages may also take shortcuts and calculate an approximate p-value
- Different packages may use different approximations, so p-values for the MWW test sometimes differ slightly depending on the program you are using

The MWW test in R

The syntax for carrying out a Wilcoxon rank sum text in R is very similar to that of a t-test:

```
wilcox.test(Distance ~ Drug, tlg, exact = TRUE)
#
#   Wilcoxon rank sum exact test
#
# data:  Distance by Drug
# W = 2184, p-value = 0.02361
# alternative hypothesis: true location shift is not equal
```

Tailgating study: Mann-Whitney test

- Applying the Mann-Whitney test to the tailgating study, we obtain a p-value of 0.02
  - Exact p-value: 0.0236
  - Approximate p-value: 0.0240 or 0.0238, depending on the approximation
- By ranking the data, we have minimized the impact of the outliers, and conducted a test that doesn't rely on any assumptions about the distribution of the data
- This is a very sound, safe approach to analyzing this data; indeed, it was the approach chosen by the investigators when they published this study
- Conclusion: The study provides substantial evidence that illegal drug users engage in riskier driving habits.

Nonparametric statistics

- Statistical methods like the t-test may be called "parametric", since unknown parameters (i.e., μ) and their effect on the distribution of data are central to the approach
- In contrast, the Mann-Whitney/Wilcoxon test involves no parameters whatsoever; such methods are referred to as *nonparametric* to highlight this fundamental difference
- The advantage of nonparametric methods is that they make fewer assumptions and don't get derailed when those assumptions go wrong — for example, when outliers are present
- The disadvantage of nonparametric methods is that we are often interested in estimating and obtaining confidence intervals for parameters, and nonparametric methods are not particularly helpful in this regard

Wilcoxon confidence intervals?

- We said at the outset of the course that tests and confidence intervals form pairs that agree with one another
- This begs the question: if we flip the MWW test around, do we get a confidence interval for something? If so, what?
- It turns out that inverting the MWW test does, in fact, produce a confidence interval for something: the median difference between two randomly chosen observations (one from each group)
- By this measure, we estimate that illegal drug users tailgate 4.3 m closer than other drivers (95% CI: 0.6 to 7.5 m)
- This is interesting, albeit a bit abstract, and not widely reported in practice

One-sample studies

- The idea of ranks can be used to analyze one-sample continuous data as well
- For example, let's consider our crossover cystic fibrosis study:

Placebo-Drug	Sign	Rank_{abs}	Placebo-Drug	Sign	Rank_{abs}
11	+	1	155	+	8
−15	−	2	158	+	9
42	+	3	−178	−	10
101	+	4	185	+	11
106	+	5	245	+	12
113	+	6	460	+	13
−152	−	7	680	+	14

Where Rank_{abs} denotes the rank of the absolute value of the Placebo - Drug difference

The Wilcoxon signed-rank test

- After setting up the data in this way, we can test the null hypothesis (no difference in response between drug and placebo) in the following way:
  - The sum of the ranks in the "+" group is 86
  - If there were no difference between drug and placebo, we should be equally likely to get a "+" and a "-" response for any given patient
  - If we randomly flip all the $+/-$ signs around, we only get a sum of the ranks for the "+" group of 86 or larger 2% of the time
  - The two sided p-value is therefore $p = 0.04$
- This approach to one-sample hypothesis testing is known as the *Wilcoxon signed-rank test*

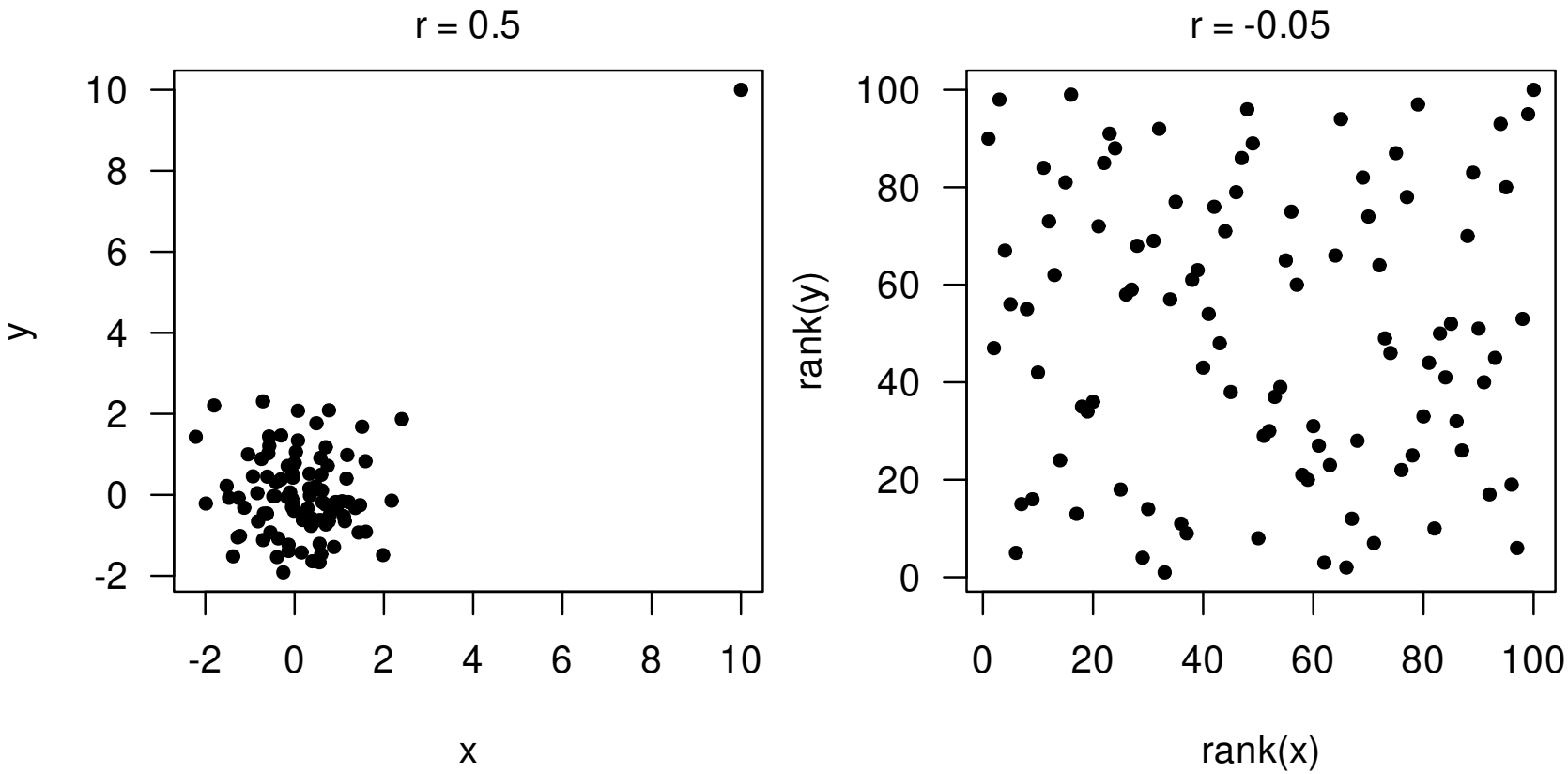
Wilcoxon signed-rank test: Power

- We've now analyzed the cystic fibrosis crossover study three ways:
  - Binomial test: $p = 0.06$
  - One-sample t-test: $p = 0.04$
  - Signed-rank test: $p = 0.04$
- All three tests more or less agree, although we do tend to get more powerful results from methods which take into account the magnitude of the difference (how much better the patient did on one treatment than the other) instead of just the direction/sign

Spearman correlation

- Finally, it is also sometimes useful to use ranks when calculating correlations
- To do this, we simply calculate the usual Pearson correlation coefficient between the ranks of x and the ranks of y; this is known as the *Spearman correlation*
- For normally distributed data such as height, the two correlation coefficients are usually quite similar
- For example, with the father-son height data, we calculated a (Pearson) correlation of 0.50 between fathers' heights and sons' heights; the Spearman correlation for the same data is 0.51

Spearman correlation & outliers



Nonparametric confidence intervals

- Nonparametric confidence intervals can be constructed by inverting rank-based tests, but tend to provide confidence intervals for esoteric quantities (e.g., "pseudomedians")
- A different approach to making nonparametric confidence intervals is the *bootstrap*
- The idea behind the bootstrap is fairly simple; we will illustrate with the tailgating data to obtain a nonparametric confidence interval for the difference in median following times

Bootstrap procedure: Difference in medians

- To "bootstrap" a sample, we simply place all 55 observed following distance values for the illegal drug user group in an urn and randomly draw 55 observations back out again (with replacement)
- Calculate the median for this "bootstrapped" sample
- Do the same for the non-illegal drug user group, and calculate the difference in medians
- Repeat the above a large number of times (say, 10,000), obtaining a long list of differences in medians
- The bootstrap confidence interval is the interval that contains the middle 95% of this list of values

Bootstrap results: Tailgating study

- For the tailgating study, this interval is (1.1, 7.7); very similar to the Wilcoxon interval from earlier, although not exactly the same, since they are intervals for subtly different quantities
- The bootstrap has no special connection to medians; the technique is extremely versatile and can be used to obtain nonparametric confidence intervals for other quantities using the same idea
- *Why* the bootstrap works is an excellent question, but beyond the scope of this course
- However, it is an important idea to be aware of, as it is certainly the most useful method for constructing nonparametric confidence intervals

Permutation tests have low power when n is small

- The Mann-Whitney/Wilcoxon test is an essential alternative to the t-test, and requires no assumptions about the population distribution
- However, it is a permutation test, and like any permutation test, it has little to no power for very small sample sizes
- For example, consider the following made-up data: the response in one group is 1, 2, 3, while the response in the other group is 101, 102, 103
- The t-test has no difficulty rejecting the null hypothesis: $p = 3 \times 10^{-8}$
- However, the Mann-Whitney/Wilcoxon test only comes up with a p-value of 0.1

Power and nonparametric tests

- Don't read too much into this, however
- The difference in power is far less dramatic when the sample size is larger (for large sample sizes, the Mann-Whitney/Wilcoxon test is about 95% as powerful as the t-test, even when the outcome is normally distributed)
- Furthermore, as we saw in the driving study, and as you will see in lab, when outliers/skewness are present, nonparametric methods can be much more powerful than t-tests

Summary

- Rank-based methods are a powerful way to analyze data when distributional assumptions are questionable, and particularly effective in the presence of outliers
- This idea can be applied to several kinds of studies:
  - Two-sample studies: Mann-Whitney/Wilcoxon rank sum test
  - One-sample studies: Wilcoxon signed-rank test
  - Both variables continuous: Spearman correlation
- Parametric vs. nonparametric:
  - Parametric advantages: More powerful when parametric assumptions hold, straightforward confidence intervals
  - Nonparametric advantages: Minimal assumptions, more powerful when parametric assumptions are wrong