13. Nonparametric Methods

In Chapters 9-12, we relied on models of probability which could be completely described by knowing the value of a few values we called parameters (few of which are every really known in practice). For this reason, the methods studied in those chapters are sometimes referred to as parametric methods. Because of the CLT, we focused on those methods which made an assumption of normality (at least approximately), among other assumptions. But, your data may suggest departures from these assumptions. If these assumptions are unreasonable in light of your data, then what do we do to analyze our data? Consider these 4 things:

- We have an outlier in our data which may indicate that it did not come from the population of interest. If you can find a reasonable explanation for the outlier (e.g., equipment failure, or you sneezed when typing the value into the computer, or, more generally, something that indicates that the outlying observation did not come from the distribution of the remaining observations, then you might be justified in excluding that value from your analysis.
- We might transform the data in some way so that the transformed data conform to the assumptions behind our existing analysis methods.
- We might use a method centered around an alternative probability distribution (e.g., Poisson vs. Normal)
- We might use nonparametric or distribution free methods (not assumption free)

We will focus on the last of these bullets in this chapter. Generally, nonparametric methods have less restrictive

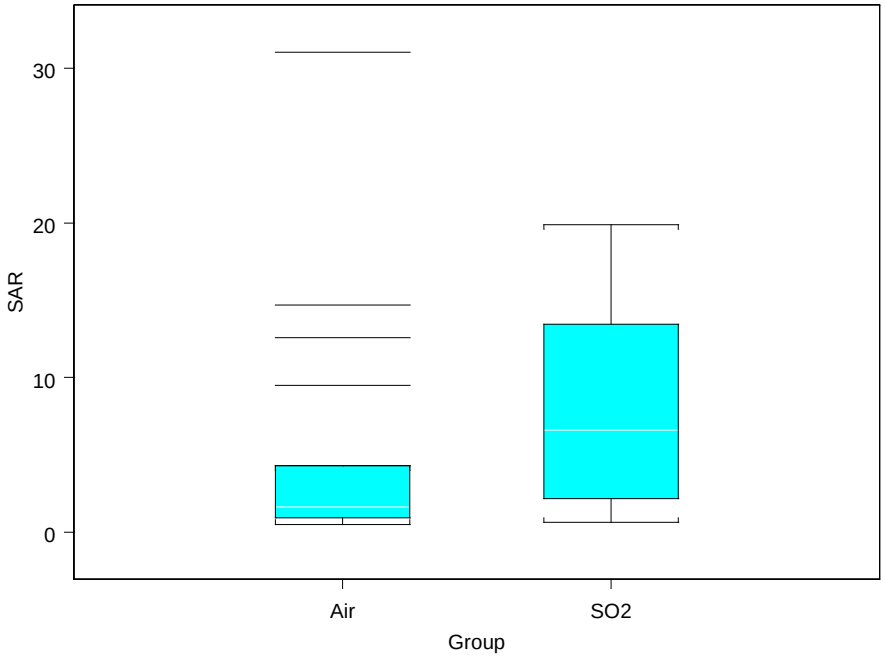
assumptions (e.g., normality is not required), but, on the other hand, these procedures are not as powerful as parametric methods. We will discuss nonparametric analogs to the paired t-test and two-sample t-test. There's even a nonparametric analog to ANOVA, but we won't discuss this.

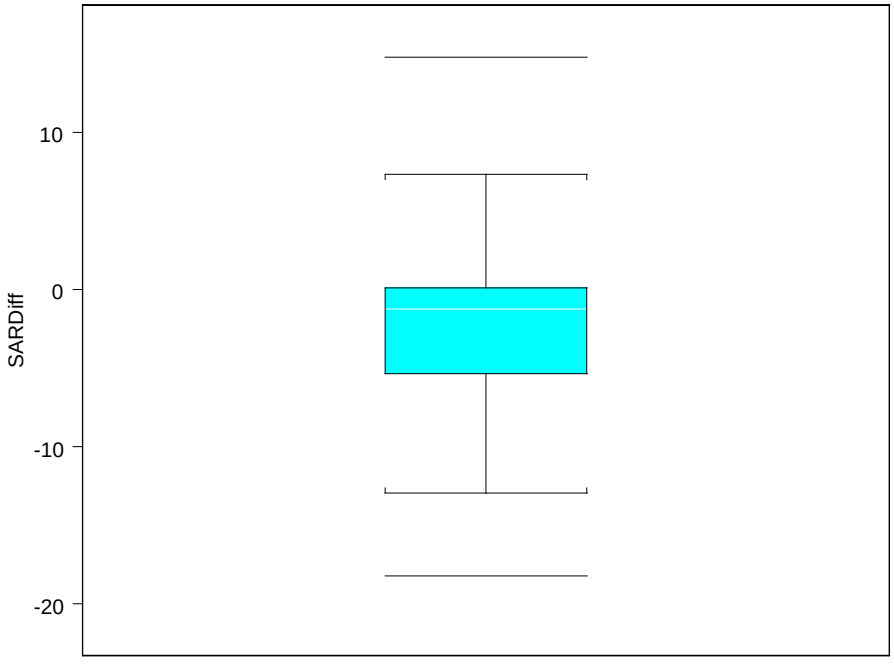
13.1. The Sign Test

The sign test is perhaps the oldest and most simple among the nonparametric testing procedures. It is analogous to the paired t-test, but requires only that your data be ordinal (i.e., for each pair, you must be able to distinguish which is greater/less/equal). Thus, the procedure does not require interval data.

E.g. Exercise 8 of Chapter 13

Nineteen patients with asthma were enrolled in a study to investigate the respiratory effects of sulfur dioxide. Measurements specific airway resistance (SAR) were measured from patients in the following way. First, SAR was measured while at rest, then after 5 minutes of exercising. This gives an increase in SAR. This was done while patients breathed "normal" air , and while subjected to 0.25 ppm SO_2. Thus, two measurements of increased SAR were obtained from each patient.


30
20
10
0
SAR
Air
SO2
Group


10
0
-10
-20
SARDiff

You might want to do a paired t-test or a t-based CI, but when assumptions are not met

- p-values from t-tables are not accurate
- nominal coverage rates for CIs are wrong

The idea behind the sign test is simple and is closely tied to the binomial distribution. In order to model a RV (denote the RV as, say, D) using the binomial probability model, we know we need:

- binary outcomes (0/1, -/+, no/yes, failure/success),
- independence between trials,
- constant probability of "success", p, between trials, and
- fixed number of trials, n

For the sign test this may be stated as

- measurement scale is ordinal (i.e., we can produce a binary response (+/-) (throw out pairs with equal values)),
- pairs are mutually independent,
- P(+) is same across trials (perhaps we only need P(+) > P(-) across trials), and
- fixed sample size, n

Thus, for numerical data, we would have the following recipe:

- Difference the data
- Count the number of positive differences, call this D.
- Compare D to a binomial distribution bin(n,p=0.5) (i.e., determine the p-value of the D) (Note: if we have "ties", i.e., differences of 0, then we exclude these pairs from the analysis and adjust the sample size accordingly.)

- If n is “large”, use a normal approximation to determine a p-value

Similar for ordinal data.

We test that the median difference is zero, or, equivalently, p=0.5 in the binomial distribution. (Roughly, we’re asking if the values of the first element of each pair tend to be the same as the values of the second element of each pair.) Why are these the same?

E.g. cont’d. Let’s test the null hypothesis that the median difference in increased SAR between the “air” treatment and “SO_2 is zero versus the alternative that the median difference is not zero.

Data

Increased SAR for Air treatment

```
 0.82  0.86  1.86  1.64 12.57  1.56  1.28  1.08  4.29  1.37
14.68  3.64  3.89  0.58  9.50  0.93  0.49 31.04  1.66
```

Increase SAR for SO_2 treatment

```
 0.72  1.05  1.40  2.30 13.49  0.62  2.41  2.32  8.19  6.33
19.88  8.87  9.25  6.59  2.17  9.93 13.44 16.25 19.89
```

Difference (with signed rank)

```
   1     2    3     4     5    6     7     8    9    10
 0.1 -0.19 0.46 -0.66 -0.92 0.94 -1.13 -1.24 -3.9 -4.96
  11    12    13    14   15 16     17    18     19
-5.2 -5.23 -5.36 -6.01 7.33 -9 -12.95 14.79 -18.23
```

We follow a familiar recipe for hypothesis testing:

Hypotheses:

Test Statistic D (count the plus signs):

p-value (use the binomial(n, 0.5) distribution):

The CLT revisited: Using the standard normal tables to calculate an approximate p-value.

In S-Plus:
Statistics>Compare Samples>Counts and Proportions>Binomial test...

```
     Exact binomial test

data:  5 out of 19
number of successes = 5, n = 19, p-value = 0.0636
alternative hypothesis: true p is not equal to 0.5
```

13.2. Wilcoxon Signed-Rank Test

Notice the sign test did not use much information in the data: just the sign of the "difference" in the data. We may also be able to use the magnitude of the difference, now assuming our data are at least interval data. Thus, we would be using more information about our data, and hence, would expect to conduct a more powerful test. In fact, the sign test is rarely used because the signed-rank test is often just as appropriate and is more powerful, assuming assumptions are met. The signed –rank test can me used to test a variety of hypotheses, most of which can be loosely described as testing whether the values in one element of the pair tend to differ from values in the other element. The signed-rank test has its own set of assumptions for testing a hypothesis about a median difference:

- each difference comes from a symmetric distribution,
- these distributions have the same median,
- the differences are mutually independent (i.e., the pairs of values used to calculate the differences are independent), and
- the difference values are at least interval data

The basic procedure:

- Difference your data
- Rank (i.e., order) your data, ignoring the sign (+/-)
- Put the sign back on your data
- Compute a test statistic. We'll use the sum of the positive ranks OR the sum of the negative ranks, whichever is smaller in magnitude (i.e., we sum the ranks, not the data!); call the test statistic T
- Use an appropriate distribution to determine a(n) (approximate) p-value for T
    - Use the z-table for "large" n
    - Use table A.6 when $n \leq 12$

The idea is that, under the assumptions, the sum of the positive ranks and sum of the negative ranks will be about the same (in absolute value). Thus, we can work with either sum. If the median difference is not zero, then one sum will be smaller, and one larger. How much smaller is determined by the distribution of T and the p-value associate with our calculated T value.

What happens if a difference is zero?

What happens if there are tied ranks?

As hinted above, when the sample size, n, is "large", then, under the null hypothesis of zero median difference, T can be

viewed as approximately normal. T has a mean and variance as follow:

Again, for smaller sample sizes ($n \leq 12$), we can use table A.6.

E.g., SAR cont'd

Data

Increased SAR for Air treatment

```
 0.82  0.86  1.86  1.64 12.57  1.56  1.28  1.08  4.29  1.37
14.68  3.64  3.89  0.58  9.50  0.93  0.49 31.04  1.66
```

Increase SAR for SO_2 treatment

```
 0.72  1.05  1.40  2.30 13.49  0.62  2.41  2.32  8.19  6.33
19.88  8.87  9.25  6.59  2.17  9.93 13.44 16.25 19.89
```

Difference (with signed rank)

```
   1     2    3     4     5    6     7     8    9    10
 0.1 -0.19 0.46 -0.66 -0.92 0.94 -1.13 -1.24 -3.9 -4.96
  11    12    13    14   15 16     17    18     19
-5.2 -5.23 -5.36 -6.01 7.33 -9 -12.95 14.79 -18.23
```

The Recipe:

Hypotheses

test statistics T

p-value

In Splus:

Statistics>Compare Samples>Two Samples>Wilcoxon Rank Test...(corrected version)

```
        Exact Wilcoxon signed-rank test

data:  x: Air in sar , and y: SO2 in sar
signed-rank statistic V = 43, n = 19, p-value = 0.0361
alternative hypothesis: true mu is not equal to 0
```

OR

```
        Wilcoxon signed-rank test

data:  x: Air in sar , and y: SO2 in sar
signed-rank normal statistic without correction Z = -2.0926, p-value = 0.0364
alternative hypothesis: true mu is not equal to 0
```

13.3. Wilcoxon Rank Sum Test

The sign test and signed rank test are appropriate for paired data (or a single sample) like in the case of the paired t-test. The Wilcoxon rank sum test can be used in situations analogous to the two-sample t-test, assuming 2 independent samples (possibly of different sizes now). Roughly, assumptions are

- Independence within and between samples
- The two samples come from distributions with the "same shape" (possibly different "locations")
- The data are at least ordinal

Roughly, the idea behind the test is the following:

- Treat the data from each sample as one big sample.
- Rank the data in the "big" sample.
- Sum the big sample rank values associated with each of the 2 samples.
- If the each sample came from distributions with the same median (null hypothesis here) the sum of the ranks will tend to be the same.
- If the samples came from distributions with different medians, then the summed ranks will tend to differ.
- Similar to the case of the signed rank test, we need only consider the smaller of the two summed ranks when computing a p-value. We'll call this test statistic, W (for Wilcoxon)

As is common with many statistics, W can be assumed to be normal if the sample sizes are “large”. Its mean and variance are:

E.g., Carbon monoxide diffusing capacity for samples of individuals with and without emphysema (Table 13.6, page 315 POB). The data are measurements of carbon monoxide diffusion capacity ($D1_{CO}$) for two groups people: those without emphysema (“healthy”) and those with emphysema. We wish to test the null hypothesis of equal medians between these 2 populations.

Rank Sums
Emphysema: 168
No Emphysema: 498
Use the smaller: W=168

n_S=13
n_L=23

The Recipe:

Hypothesis

test statistic W

p-value

if samples sizes are “large” use the normal distribution

if sample sizes are not “large” use table A.7

Splus Says:

Statistics>Compare Samples>Two Samples>Wilcoxon Rank Test…(etc.)

```
        Exact Wilcoxon rank-sum test

data:  x: dlco with emphysema = No , and y: dlco with emphysema = Yes
rank-sum statistic W = 498, n = 23, m = 13, p-value = 0.0162
alternative hypothesis: true mu is not equal to 0
```

OR the normal approximation:

```
        Wilcoxon rank-sum test

data:  x: dlco with emphysema = No , and y: dlco with emphysema = Yes
rank-sum normal statistic without correction Z = 2.3878, p-value =
0.017
alternative hypothesis: true mu is not equal to 0
```

Summary

- Check for departures from normality, etc. Or, perhaps data are ordinal (sign test or rank sum).
- Parametric procedures are more powerful if assumptions are met.
- Paired samples
  - paired t-test
  - sign test
  - signed rank test
- Independent samples
  - two-sample t-test
  - rank sum test
- Multiply p-values by 2 for 2-sided tests