Week 7 activity: Nonparametric inference

STAT218

library(tidyverse)  
load('data/brfss.RData')  
load('data/temps.RData')  
ddt <- MASS::DDT

In this activity you’ll learn about nonparametric alternatives to the test for one and two means. The activity is organized much like a lab, but with extra narrative. You should read through the narrative at your own pace and try the exercises provided as you go. At the end there are two practice problems that you’ll be expected to complete before next class.

### Background: parametric and nonparametric inference

Consider the basis for the inferences developed so far: under certain conditions (typically regularity of the underlying population distribution, assessed by checking histograms for unimodality and approximate symmetry) and with sufficient sample sizes, a model is specified for the sampling distribution of a test statistic.

* inference for one mean: model for the sampling distribution of
* inference comparing two means: model for the sampling distribution of
* inference comparing several means: model for the sampling distribution of

These are all what are known as *parametric models*, because they are specified through one or two parameters that determine their exact shape. The parameters in this case are the degrees of freedom terms – for the one-sample test, (usually estimated) for the two-sample test, and and for the test.

As such, these procedures are examples of *parametric inference* – inferences that utilize a parametric model for the data and/or test statistic.

When assumptions for parametric inference aren’t tenable, or when a parametric model is not available, there are so-called **nonparametric** methods of inference: **methods that don’t depend on a parametric model** such as the or models we’ve learned about in class.

We will consider specifically nonparametric procedures based on ranks, *i.e.*, ordering observations from smallest to largest.

### Motivating examples

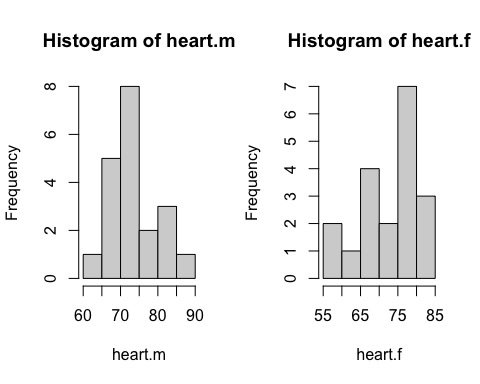
In practice, the situation that most often leads an analyst to consider rank-based nonparametric methods is that the assumptions for the test either don’t hold or are difficult to assess.

Below are two such situations you’ve already encountered in this class.

#### Small sample sizes

When sample sizes are small, it’s awkward to assess assumptions for parametric inference, because with few observations histograms can lack any discernible shape. For example, the most we can say about the following data on heart rates for 19 women and 20 men is that there are no evident outliers.

heart.m <- temps |> filter(sex == 'male') |> pull(heart.rate)  
heart.f <- temps |> filter(sex == 'female') |> pull(heart.rate)  
  
par(mfrow = c(1, 2))  
hist(heart.m)  
hist(heart.f)



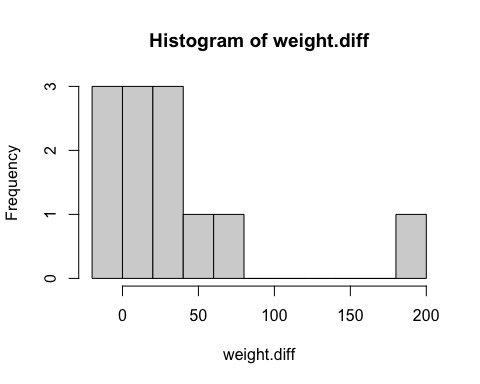
In practical terms, the test is likely still fine under these circumstances; however, some may wish to consider an inference for comparing heart rates between groups that doesn’t depend on distributional assumptions.

#### Assumptions don’t hold

On occasion you may go to check assumptions and find that they’re clearly violated. For example, the pairwise differences between actual and desired weights from BRFSS respondents showed clear skewness and several large outliers. In that case, the sample size was big enough that we overlooked the issue, but it’s not hard to imagine a similar situation cropping up with fewer observations.

Suppose you only had 12 observations that showed the same skew and had one big outlier:

set.seed(51424)  
weight.diff <- sample(brfss$weight - brfss$wtdesire, 12)  
hist(weight.diff, breaks = 10)



Here the test isn’t appropriate, and using it anyway would likely result in a true significance level, coverage, and power quite different from the nominal levels specified in the test, so it would be hard to trust the result. This is a great situation to use a rank-based nonparametric alternative.

### Inference on “location” (not mean)

The usual parametric inferences pertain to the population mean; not so with rank-based nonparametrics. Instead, these inferences pertain simply to “location”.

Often “location” is characterized in terms of the “center” of a distribution so that inferences can be interpreted in a manner similar to parametric tests and intervals.

We will follow this convention and consider the hypotheses to be about the center(s) of the population model(s), denoted . For example, the two-sided test of center would test the hypotheses:

For the two-sided test of difference in centers, we will test the hypotheses:

However, you should keep in mind that “location” is a more general notion that encompasses all measures of location of a distribution.

### One-sample inference: signed rank test

The basic premise of this test is that, if the population model is symmetric, its center should evenly divide the data.

|  |
| --- |
| Warm up |
| Consider the following 15 measurements of DDT in kale in ppm in order from smallest to largest:  sort(ddt)  [1] 2.79 2.93 3.06 3.07 3.08 3.18 3.22 3.22 3.33 3.34 3.34 3.38 3.56 3.78 4.64  Suppose you wish to test whether the center of the population model is ppm and assume a symmetric population model.   1. How many observations would you expect to be smaller than if 3ppm is in fact the center? 2. How many observations are actually smaller than 3 ppm? 3. Based on your answers to 1-2, do you think it is likely that in fact ? |

|  |
| --- |
| Solution |
| If the population is symmetric about , you’d expect roughly half of observations (7.5) to be smaller than 3ppm.  The actual number of observations smaller than 3ppm is:  # number of observations below 3 sum(ddt < 3)  [1] 2  This is much less than half, so it seems unlikely that the center is actually 3ppm. |

The signed rank test is a nonparametric alternative to the one-sample test applicable to any symmetric population model. The particular form of symmetry and presence of outliers do not affect the test.

#### Hypotheses

The test can be directional or two-sided, just like the test. Thus, the possible hypotheses are:

#### Test procedure

While the intuition of the test is that half of observations should be smaller than the true center under population symmetry, the test statistic is not quite as direct as a tally of how many observations are below the hypothetical value. Instead, the procedure is as follows:

1. [center] Calculate deviations
2. [rank] Sort and rank the absolute deviations
   * average ranks in case of ties
   * drop zeros
3. [sum] Add up the positive ‘signed ranks’

This produces the test statistic:

|  |
| --- |
| Check your understanding |
| Try working out the rank sum procedure manually using the DDT data – it’s small enough that you could jot down the steps on some scratch paper. See if you can calculate .  To help, here are the deviations sorted smallest to largest:  # calculate deviations di <- ddt - 3 sort(di)  [1] -0.21 -0.07 0.06 0.07 0.08 0.18 0.22 0.22 0.33 0.34 0.34 0.38 [13] 0.56 0.78 1.64   1. Start by writing the deviations in order of absolute value in a column. 2. Then rank them 1-15 in an adjacent column. If there is a tie – *e.g.*, two absolute deviations of 0.05 – then assign them both the average of the ranks. For example, if 0.05 occurs twice in positions 2 and 3, then give them both rank 2.5. 3. Write down the sign of the deviation in a new column. 4. Then write down the “signed rank”, or product of the sign and the rank, in a fourth column. 5. Add up the positive signed ranks. This is the signed rank statistic. |

|  |
| --- |
| Solution |
| This shows the steps in R, but don’t worry about the codes; the output is meant to illustrate what you would do on paper to find the rank sum statistic.  # this shows the steps, column by column; focus on output ddt.srank <- tibble(di = di) |>  mutate(abs.di = abs(di),   rank = rank(abs.di),  sign = sign(di),  signed.rank = sign\*rank) |>  arrange(abs.di) ddt.srank  # A tibble: 15 × 5  di abs.di rank sign signed.rank  <dbl> <dbl> <dbl> <dbl> <dbl>  1 0.0600 0.0600 1 1 1   2 -0.0700 0.0700 2.5 -1 -2.5  3 0.0700 0.0700 2.5 1 2.5  4 0.0800 0.0800 4 1 4   5 0.180 0.180 5 1 5   6 -0.21 0.21 6 -1 -6   7 0.220 0.220 7.5 1 7.5  8 0.220 0.220 7.5 1 7.5  9 0.33 0.33 9 1 9  10 0.34 0.34 10.5 1 10.5 11 0.34 0.34 10.5 1 10.5 12 0.38 0.38 12 1 12  13 0.56 0.56 13 1 13  14 0.78 0.78 14 1 14  15 1.64 1.64 15 1 15  # signed rank statistic vstat <- ddt.srank |> filter(sign > 0) |> pull(signed.rank) |> sum() vstat  [1] 111.5 |

#### -values for the test

Just like other hypothesis tests, the signed rank test rejects if is sufficiently large in the direction of the alternative.

A sampling distribution for can be found exactly using combinatorics, or approximated using probability theory. In this caseWe won’t go into details about either approach, except to indicate that there is a sampling distribution for that we can use to obtain -values in the same fashion that the model was used to obtain -values for the test.

In this case, there are 3.2768^{4} possible sign combinations; of these, only about 0.375% give a larger value of . That provides a -value for the test, and since we would reject at the 5% significance level. This result is interpreted as:

The data provide strong evidence that the typical DDT concentration in kale is not 3ppm (signed rank statistic *V* = 111.5, *p* = 0.00375).

#### Implementation with wilcox.test(...)

The implementation in R looks and functions much like t.test:

# signed rank test at 1% level  
wilcox.test(ddt, mu = 3, alternative = 'two.sided',   
 exact = F, conf.int = T, conf.level = 0.99)

Wilcoxon signed rank test with continuity correction  
  
data: ddt  
V = 111.5, p-value = 0.003751  
alternative hypothesis: true location is not equal to 3  
99 percent confidence interval:  
 3.060070 3.780032  
sample estimates:  
(pseudo)median   
 3.26001

Some remarks: - exact = F produces approximate -values and confidence intervals; you may see warning messages if this is excluded - pseudo-median is a measure of center, but not the same as a median or mean (check!)

|  |
| --- |
| Your turn 1 |
| Perform the analogous inference using the test and compare the results. Do the tests agree at the 1% significance level? |

|  |
| --- |
| Solution |
| The tests do not agree: the signed rank test rejects at the 1% level, but the test does not.  # perform t test at 1% level to compare t.test(ddt, mu = 3, alternative = 'two.sided', conf.level = 0.99)  One Sample t-test  data: ddt t = 2.9059, df = 14, p-value = 0.01151 alternative hypothesis: true mean is not equal to 3 99 percent confidence interval:  2.991996 3.664004 sample estimates: mean of x   3.328 |

|  |
| --- |
| Your turn 2 |
| Perform the signed rank test to determine whether actual weight exceeds desired weight by more than 10lbs at the 5% significance level. Report the result in the usual narrative format. Compare your result with the inference obtained from a test. |

|  |
| --- |
| Solution |
| # use signed rank test to determine whether actual exceeds desired by at least 10lbs wilcox.test(weight.diff, mu = 10, alternative = 'greater',   exact = F, conf.int = T, conf.level = 0.95)  Wilcoxon signed rank test with continuity correction  data: weight.diff V = 61, p-value = 0.04559 alternative hypothesis: true location is greater than 10 95 percent confidence interval:  10.00003 Inf sample estimates: (pseudo)median   22.8617  # check t test result to compare t.test(weight.diff, mu = 10, alternative = 'greater', conf.level = 0.95)  One Sample t-test  data: weight.diff t = 1.6127, df = 11, p-value = 0.06755 alternative hypothesis: true mean is greater than 10 95 percent confidence interval:  7.019188 Inf sample estimates: mean of x   36.25  Interpretation of the signed rank test test:  The data provide evidence that actual weight exceeds desired weight by more than 10lbs (signed rank statistic *V* = 61, *p* = 0.0456).  The signed rank test finds evidence that actual weight exceeds desired weight by more than 10lbs, where the test does not. |

### Two-sample inference: rank sum test

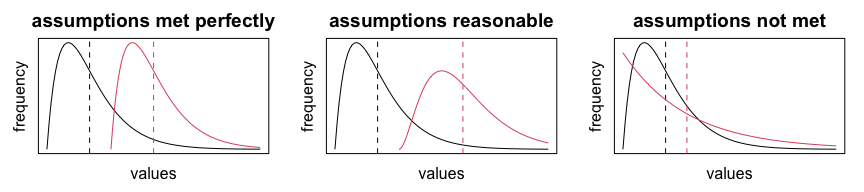
The rank-sum test is a nonparametric alternative to the two-sample test based on ranks. The key idea for the test is that if observations in both groups come from the same population distribution then they should be exchangeable (*i.e.*, groupings don’t matter).

Thus, the only assumption for this test is that data are independent. The test can be directional, and thus it is possible to test the following hypotheses comparing centers:

#### The alternative hypothesis

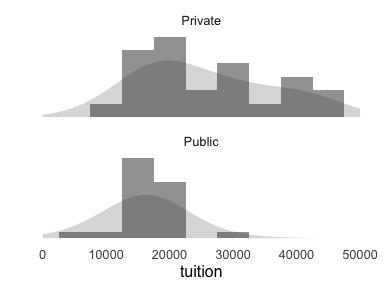
This test is a bit funny in that the null hypothesis is really that *the distributions are exactly the same*. The alternative to this possibility can come about in a number of ways. The alternative is usually interpreted as a *difference in location*, primarily because this is the situation that the test has power to detect.

So, it is often said that the test also assumes that the samples differ only in location. In other words, the test is most appropriate when the histograms look “shifted”, but not fundamentally different, as illustrated below.

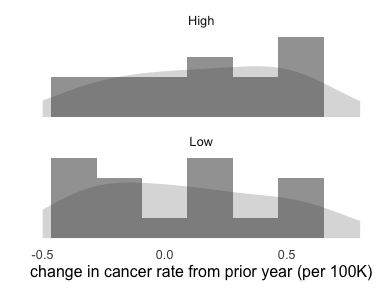


In all of these cases, there is a (true) difference in location, but if shape differs too much, the rank sum test should *not* be used. That said, it can be hard to tell with small samples, so it may not really be practical to to check this assumption.

**Example 1.** Out-of-state tuition costs from 26 public and 26 private universities. These data differ primarily in spread, not location; so the rank-sum test might not work well here.



**Example 2.** Deviations from expected cancer rates in CT in years with high and low sunspot activity. The shape is a bit hard to discern here, but it seems plausible that the distribution for high sunspot years is shifted to the right of that for low sunspot years. So, the rank sum test would be appropriate here.



#### Rank sum test procedure

The rank sum procedure, though a bit opaque, is rather simple:

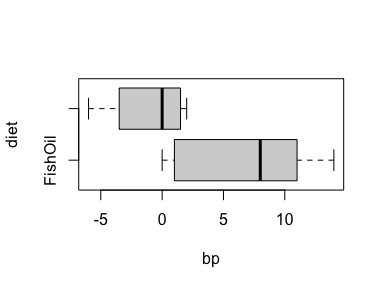
1. [pool] Combine observations from both groups
2. [rank] Sort and rank pooled observations
3. [sum] Add up ranks in the first group
4. [adjust] Subtract , where is the sample size of the first group

The test rejects if the sum is larger than expected in the direction of the alternative. As in the signed rank test, combinatorics are used to determine a -value for the test.

The rationale for this procedure is that if the distributions are the same, then the ranks should be evenly distributed among the two groups; this induces a particular sampling distribution on the sum of the ranks in each group. The adjustment facilitates computation of -values.

Let’s illustrate using data from an experiment in which participants were randomly assigned to receive a fish oil supplement or a regular oil supplement. For each subject, the reduction in blood pressure was measured after a period of time on the treatments.

# load dataset  
fish.oil <- Sleuth3::ex0112 |> rename\_with(tolower)  
  
# make boxplot  
boxplot(bp ~ diet, data = fish.oil, horizontal = T)



There’s a bit of difference in spread, but enough of a shift in location that the rank sum test is reasonable to apply.

|  |
| --- |
| Check your understanding |
| Carry out the rank sum procedure outlined above and compute the rank sum statistic by summing up the ranks in the fish oil group.  To facilitate calculations, here are the data in order of increasing blood pressure:  fish.oil |> arrange(bp)  bp diet 1 -6 RegularOil 2 -4 RegularOil 3 -3 RegularOil 4 0 FishOil 5 0 FishOil 6 0 RegularOil 7 1 RegularOil 8 2 FishOil 9 2 RegularOil 10 2 RegularOil 11 8 FishOil 12 10 FishOil 13 12 FishOil 14 14 FishOil  Add a column of ranks by hand (or in R if you can figure out how!), averaging ranks for any ties. Then add up the ranks in the FishOil group, and subtract to obtain the rank sum statistic. |

|  |
| --- |
| Solution |
| The output below illustrates the rankings and selection of which ones to add up. Don’t worry about the codes  # again, ignore the codes; look at output fish.oil.ranksum <- fish.oil |>   mutate(rank = rank(bp),  ranks.fish = rank\*(diet == 'FishOil')) |>  arrange(bp) fish.oil.ranksum  bp diet rank ranks.fish 1 -6 RegularOil 1 0 2 -4 RegularOil 2 0 3 -3 RegularOil 3 0 4 0 FishOil 5 5 5 0 FishOil 5 5 6 0 RegularOil 5 0 7 1 RegularOil 7 0 8 2 FishOil 9 9 9 2 RegularOil 9 0 10 2 RegularOil 9 0 11 8 FishOil 11 11 12 10 FishOil 12 12 13 12 FishOil 13 13 14 14 FishOil 14 14  # rank sum statistic n1 <- count(fish.oil, diet)$n[1] sum(fish.oil.ranksum$ranks.fish) - n1\*(n1 + 1)/2  [1] 41 |

#### Implementation using wilcox.test(...)

The wilcox.test function also implements the rank sum test, using the same syntax as t.test(...). For example, using the fish oil data, we might test at the 1% significance level whether the fish oil supplement caused a greater reduction in blood pressure:

wilcox.test(bp ~ diet, data = fish.oil,   
 mu = 0, alternative = 'greater',   
 exact = F, conf.int = T, conf.level = 0.99)

Wilcoxon rank sum test with continuity correction  
  
data: bp by diet  
W = 41, p-value = 0.01957  
alternative hypothesis: true location shift is greater than 0  
99 percent confidence interval:  
 -0.9999444 Inf  
sample estimates:  
difference in location   
 7.999962

For this test, the -value gives the percentage of possible rank allocations among the groups for which the rank sum is at least as favorable to ; again this is computed using combinatorics or approximation methods.

In this instance, the -value indicates that only about 1.96% of all possible rank allocations among the two groups would produce a rank sum statistic at least as large. Thus, testing at the 1% significance level:

The data provide do not provide sufficient evidence at the 1% significance level that the fish oil supplement caused a greater reduction in blood pressure than the regular oil supplement (rank sum statistic *W* = 41, *p* = 0.01957).

|  |
| --- |
| Your turn 3 |
| Perform the corresponding test for comparison at the 1% significance level. Do the tests agree? |

|  |
| --- |
| Solution |
| The tests do not agree; the test supports evidence of an effect at the 1% level where the rank sum test does not.  # t test for effect of treatment at 1% level t.test(bp ~ diet, data = fish.oil, mu = 0,   alternative = 'greater', conf.level = 0.95)  Welch Two Sample t-test  data: bp by diet t = 3.0621, df = 9.2643, p-value = 0.006542 alternative hypothesis: true difference in means between group FishOil and group RegularOil is greater than 0 95 percent confidence interval:  3.111056 Inf sample estimates:  mean in group FishOil mean in group RegularOil   6.571429 -1.142857 |

|  |
| --- |
| Your turn 4 |
| Use the cancer dataset (specifically the delta variable) to test whether the change in cancer rate is higher in years with high sunspot activity. Carry out the test at the 5% level, and report the result in the usual narrative style. |

|  |
| --- |
| Solution |
| # test whether change in cancer rate is higher in years with high sunspot activity at the 5% level wilcox.test(delta ~ sunspot, data = cancer, mu = 0,  exact = F, alternative = 'greater',   conf.int = T, conf.level = 0.95)  Wilcoxon rank sum test with continuity correction  data: delta by sunspot W = 157.5, p-value = 0.3072 alternative hypothesis: true location shift is greater than 0 95 percent confidence interval:  -0.1000484 Inf sample estimates: difference in location   0.09999013  Interpretation:  The data provide no evidence that the change in cancer rate is higher in years with high sunspot activity (rank sum statistic *W* = 157.5, *p* = 0.3072). |

### Practice problem

1. Is there a difference in cholesterol associated with consuming oat bran compared with consuming corn flakes? Use the cholesterol dataset to test for a difference at the 5% level using a nonparametric test.
   1. Construct boxplots to compare the distributions for location shift. Does the nonparametric test seem appropriate?
   2. Carry out the test.
   3. Report the test result in the usual narrative style.