Nonparametric/Distribution-Free Tests

Approach with Caution

Andrew J. Spieker, PhD

Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania

Outline

Nonparametric/Distribution-Free Tests

- The Mann-Whitney U test
- Fisher's exact test (for contingency tables)
- The sign test (for paired data)
- Wilcoxon matched-pair signed rank test (for paired data)

Overview and motivation

Parametric tests

- A *parameter* is a value (or a set of values) that summarizes a particular aspect of data for an entire population. Generally, population parameters are unknown. Examples include:
  - Mean: parameter that characterizes central tendency.
  - Variance: parameter that characterizes spread.
  - Risk ratio: parameter that compares risk between two groups.
- We use statistical *inference* to estimate population parameters, quantify uncertainty, and conduct hypothesis tests.

Overview and motivation

Parametric tests

- When one refers to a "parametric test," they *usually* are talking about a test that:
  - Assumes an underlying distributional form for the data (e.g., binomial, Poisson, normal)...
  - For which the *parameters* of that distribution (whose general form is presumed known) are unknown.
- What worries people about a "parametric" test is that sometimes, if its *distributional* assumptions are violated, hypothesis tests are invalid and confidence intervals do not have valid coverage (i.e., they would not cover the truth in 95% of hypothetical study replicates).

Overview and motivation

Parametric tests

- Tests that seek to relax distributional assumptions are sometimes called nonparametric.
- Though a nonparametric test may still seek to provide information about a population parameter, it would not assume that the distribution of the data takes some specific form that can be described by finitely many parameters.
    - Sometimes instead called "distribution-free" to avoid confusion.
- We've seen nonparametric/distribution-free estimation:
    - To estimate population mean, can use sample mean. Can trust that this is valid even if the distribution of the data is not known.

The Mann-Whitney test

Ideas

- Previously noted that t-test does not demand that data are normally distributed if your sample size is large enough.
- In small samples, sometimes worry about heavy skewness or otherwise not approximately "bell-shaped" data.
- The Mann-Whitney test (or Wilcoxon Rank Sum test, or the Mann-Whitney Wilcoxon test), is a nonparametric test that does not make assumptions about the distribution of data.

The Mann-Whitney test

Ideas

- Sometimes presented as an alternative to the t-test for sample sizes that are "small" or data that are not approximately normally distributed.
- This is not an ideal way to view the Mann-Whitney test, specifically because:
  1. The t-test is well behaved in large samples even if the data are not normally distributed.
  2. The Mann-Whitney test answers a *very* different scientific question than the t-test.
  3. The Mann-Whitney test can be a useful test in its own right and should not be viewed exclusively as an *alternative* method.
- That said—what question does the Mann-Whitney test answer?

The Mann-Whitney test

Example: β-carotene

- $n_0 = 8$ individuals randomized to receive 0 mg/day.
- $n_1 = 10$ individuals randomized to receive 60 mg/day (nine of whom were measured at three months to assess plasma β-carotene levels).
- We want to perform an analysis to assess whether those randomized to receive 60 mg/day "tend to have higher plasma β-carotene levels as compared to those receiving 0 mg/day."

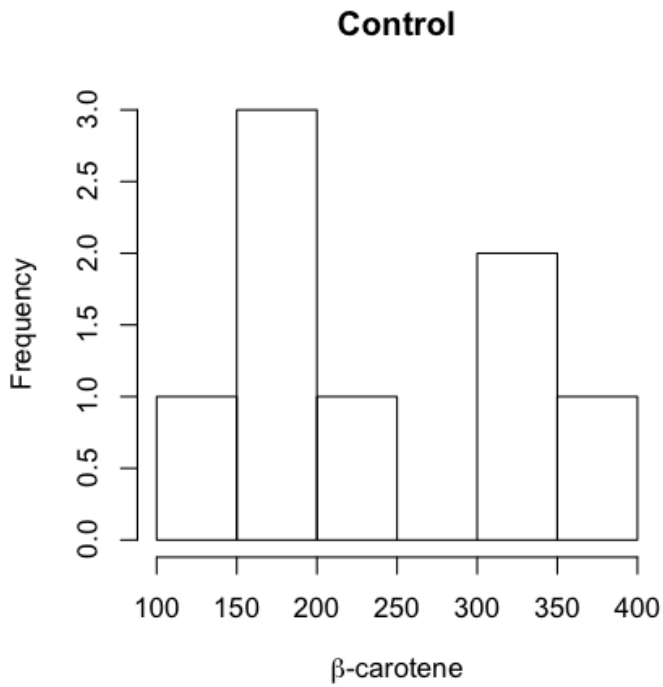
The Mann-Whitney test

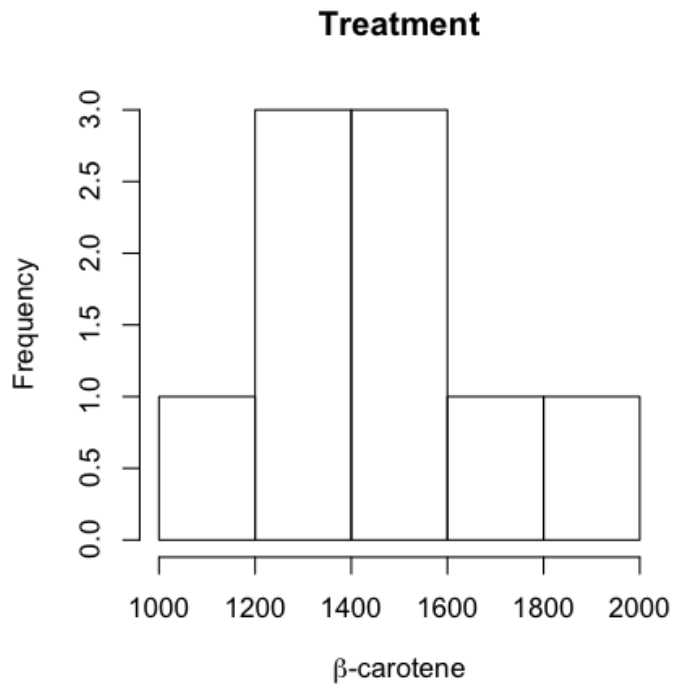
β-carotene: Data

0: 109, 165, 194, 200, 241, 323, 332, 384
60: 1098, 1285, 1292, 1370, 1410, 1586, 1595, 1604, 1960

The Mann-Whitney test

β-carotene: Data





The Mann-Whitney test

Example: β-carotene

- The words "tend to have higher plasma β-carotene" are not nearly specific enough.
- If we wanted to perform an analysis to evaluate if the mean plasma β-carotene" were different between the two groups, we would turn to the two-sample t-test.
- In practice, I wouldn't argue with this since the t-test is well behaved, even in fairly small samples—though some would insist upon employing a nonparametric test like the Mann-Whitney test.
- If you do this. . . *make sure you know what you're actually testing*. . .

The Mann-Whitney test

Example: β-carotene

- X_0: Randomly sampled plasma β-carotene from control group.
- X_1: Randomly sampled plasma β-carotene from tx group.
- $\theta = \mathrm{P}(X_1 > X_0)$ (the probability that a randomly sampled β-carotene value from the treatment group is greater than some randomly sampled plasma β-carotene value from the control group).
- The Mann-Whitney test is a test to evaluate whether θ is different from 0.5:

$$H_0 : \theta = 0.5 \text{ vs. } H_1 : \theta \neq 0.5$$

The Mann-Whitney test

Procedure (Method 1)

- Data for Group 0: $x_{01}, x_{02}, \ldots, x_{0n_0}$
- Data for Group 1: $x_{11}, x_{12}, \ldots, x_{1n_1}$
  1. Directly compute the number of times an observation from Group 1 is greater than an observation from Group 0 in *all* $n_0 \times n_1$ possible pairings:

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} 1(x_{1i} > x_{0j})$$

- The major challenge with this is that the sample sizes grow, this calculation becomes computationally cumbersome.

The Mann-Whitney test

Procedure (Method 2)

- Data for Group 0: $x_{01}, x_{02}, \ldots, x_{0n_0}$
- Data for Group 1: $x_{11}, x_{12}, \ldots, x_{1n_1}$
  1. Order the values in the two groups combined.
  2. Assign each value its rank r_i.
  3. It turns out that U from the previous slide is given by:

$$U = R - \frac{n_1(n_1+1)}{2},$$

     ... where R is the sum of the ranks in Group 1.
  4. If there are ties, observations are given the "average" rank.
- Exact one-sided/two-sided p-values can be computed. How do we interpret them?

The Mann-Whitney test

Procedure

- Data for Group 0: $x_{01}, x_{02}, \ldots, x_{0n_0}$
- Data for Group 1: $x_{11}, x_{12}, \ldots, x_{1n_1}$
  1. Order the values in the two groups combined.
  2. Assign each value its rank r_i.
  3. Sum the ranks in the first group (let R denote that sum).
  4. If there are ties, observations are given the "average" rank.
- Under $H_0 : \theta = 0.5$, we have that:
  1. $\mathrm{E}[U] = n_0 n_1/2$ (Trust me...)
  2. $\mathrm{Var}[U] = n_0 n_1 (n_0 + n_1 + 1)/12$ (Trust me!)
- In large samples, **under the null hypothesis**:

$$Z = \frac{R - \mathrm{E}[U]}{\sqrt{\mathrm{Var}[U]}} \dot{\sim} \mathcal{N}(0, 1)$$

The Mann-Whitney test

β-carotene: Help from Stata

```
. ranksum carot1, by(dose)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

dose	obs	rank sum	expected
0	8	36	72
60	9	117	81
combined	17	153	153

```
unadjusted variance        108.00
adjustment for ties          0.00
                           ------
adjusted variance          108.00

Ho: carot1(dose==0) = carot1(dose==60)
                z =  -3.464
    Prob > |z| =   0.0005
```

The Mann-Whitney test: Write-up

β-carotene: Writing up results from the test

This study provides evidence that the probability of a randomly sampled β-carotene value from the treatment group being greater than a randomly sampled β-carotene value from the placebo group is differnet from 0.5 ($p = 0.005$).

The Mann-Whitney test

Notes

- The Mann-Whitney test provides you with inference on how well two groups tend to *separate out*. Therefore, it is not an "alternative" to the t-test.
- In fact, $\theta = \mathrm{P}(X_1 > X_0)$ is the area under the ROC curve.
  - AUC can be estimated by $U/(n_0 n_1)$
- "We estimate that a randomly sampled β-carotene from the treatment group has a 100% probability of being larger than a randomly sampled β-carotene from the placebo group."
- However, beware of weirdness!

The Mann-Whitney test

Data (Number of days to recovery? Pain scales?)

- **Treatment Group 1**: (3, 3, 3, 3, 3)
- **Treatment Group 2**: (0, 0, 4, 4, 4)
- **Treatment Group 3**: (2, 2, 2, 5, 5)

Verify on your own!

- $P(X_1 > X_2) = 0.40$, so Tx 1 beats Tx 2.
- $P(X_2 > X_3) = 0.36$, so Tx 2 beats Tx 3.
- $P(X_3 > X_1) = 0.40$, so Tx 3 beats Tx 1.

The Mann-Whitney test

How about that?

The Mann-Whitney test

Beware! Lack of transitivity.

- When comparing multiple treatments, strange things like this can indeed happen.
- Almost all rank tests are weird like this...
  - ...yet informal surveys suggest that fewer than 5% of statisticians are even aware of this problem.

The Mann-Whitney test

Notes

- The Mann-Whitney test is *not* a test on whether or not the medians of the two groups are equal!
- When you hear someone say that it is, you have my permission to say "that is a *lie*!"

The Mann-Whitney test

Raw data

y	grp	freq
-2	1	20
0	1	20
5	1	20
-1	2	20
0	2	20
10	2	20

What can you say about the *medians* in the two groups?

The Mann-Whitney test

Mann-Whitney test in Stata

```
. ranksum y, by(grp)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

      grp |      obs    rank sum    expected
----------+---------------------------------
        1 |       60        3230        3630
        2 |       60        4030        3630
----------+---------------------------------
 combined |      120        7260        7260

unadjusted variance     36300.00
adjustment for ties     -2014.29
                        ---------
adjusted variance       34285.71

Ho: y(grp==1) = y(grp==2)
             z =  -2.160
    Prob > |z| =   0.0308
```

Clearly, the Mann-Whitney test fails as a test of the medians.

The Mann-Whitney test

In short...

- Mann-Whitney test is a nonparametric test (distribution-free) that can help you evaluate whether the distribution of two groups "separate out."
- However, it is not an *alternative* to the t-test.
- Moreover, it is *not* a test on the medians in two groups!

Fisher's exact test

Recall the χ^2 test

- The condition to apply it was to have expected cell counts of at least five.
- This number was deemed the appropriate number such that the "asymptotic properties" of X^2 would be likely to hold.
- If expected cell counts are less than five, sometimes people will turn to *Fisher's exact test*, a nonparametric test to evaluate an association between two binary variables.

Fisher's exact test

The setup

	Disease (D)	No Disease ($\overline{D}$)	TOTAL
Exposed (E)	a	b	a + b
Unexposed ($\overline{E}$)	c	d	c + d
TOTAL	a + c	b + d	n

- What is the *actual* probability of observing these numbers?
- If we assume the table margins are *fixed*:

$$p = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}}$$

Fisher's exact test

Salt intake and CVD death

	CVD death	Non-CVD death	TOTAL
High salt	5	2	7
Low salt	30	23	53
TOTAL	35	25	60

- $\widehat{\text{OR}} = \frac{5 \times 23}{2 \times 30} = 1.92$ (the odds of CVD death is about twice as high among those with a high-salt diet).
- What is the probability of observing this exact table?

$$p = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{\binom{35}{5}\binom{25}{2}}{\binom{60}{7}} \approx 0.252$$

Fisher's exact test

Possible tables (table margins fixed)

0	7	1	6	2	5	3	4
35	18	34	19	33	20	32	21
0.001		0.016		0.082		0.214	
4	3	5	2	6	1	7	0
31	22	30	23	29	24	28	25
0.312		0.252		0.105		0.017	

- To compute two-sided p-value: take the sum of probabilities of all tables whose probabilities did not exceed the probability of the table we observed.
- $p = 0.001 + 0.016 + 0.082 + 0.214 + 0.252 + 0.105 + 0.017 \approx 0.687$.

Fisher's exact test

Possible tables (table margins fixed)

0	7	1	6	2	5	3	4
35	18	34	19	33	20	32	21
0.001		0.016		0.082		0.214	
4	3	5	2	6	1	7	0
31	22	30	23	29	24	28	25
0.312		0.252		0.105		0.017	

- To compute one-sided p-value, compare a to its expectation $\mathrm{E}[a]$: if $a \geq \mathrm{E}[a]$, p-value is the sum of probabilities for tables with $a \geq \mathrm{E}[a]$. Otherwise, p-value is sum of probabilities for tables with $a < \mathrm{E}[a]$. Here, $a = 5$ exceeds $\mathrm{E}[a] = 4.1$. Thus,
- $p = 0.252 + 0.105 + 0.017 \approx 0.374$.

Fisher's exact test

Salt intake and CVD death: Help from Stata!

```
input cvd salt num
1 1 5
1 0 30
0 1 2
0 0 23 5
end
```

Fisher's exact test

Salt intake and CVD death: Help from Stata!

```
tab2 cvd salt [fweight=num], chi2 exp
```

```
. tab2 cvd salt [fweight=num], chi2 exp

-> tabulation of cvd by salt
```

Key
frequency *expected frequency*

cvd	salt 0	1	Total
0	23 22.1	2 2.9	25 25.0
1	30 30.9	5 4.1	35 35.0
Total	53 53.0	7 7.0	60 60.0

Pearson chi2(1) = 0.5591 Pr = 0.455

Fisher's exact test

Salt intake and CVD death: Help from Stata!

```
tab2 cvd salt [fweight=num], exact
```

```
. tab2 cvd salt [fweight=num], exact

-> tabulation of cvd by salt

           |       salt
       cvd |         0          1 |     Total
-----------+----------------------+----------
         0 |        23          2 |        25
         1 |        30          5 |        35
-----------+----------------------+----------
     Total |        53          7 |        60

          Fisher's exact =                 0.688
  1-sided Fisher's exact =                 0.375
```

Fisher's exact test

Challenges

- Almost *never* are all margins for a study fixed.
  - In a case-control study, we often fix the column margins, but not the row margins.
  - In a cohort study, we can sometimes fix the row margins in advance, but generally not the column margins.
- Strictly speaking, not actually an *exact* test. I argue that a better name is Fisher's *conditional* exact test.
- Some argue that this is okay because margin totals contain almost no information about the tested property (in statistical terms, it is almost an *ancillary* statistic).

Fisher's exact test

Challenges

- Fisher's exact test has been criticized for being conservative (actual rejection rate is *below* nominal significance level, α).
- In short (and without too much detail) happens because we are combining a discrete statistic with fixed significance levels.
- On the other hand, the χ^2 test may not be valid if the expected table counts fall below 5.
    - Solution: Always argue for larger samples? ☺

The sign test

Matched pairs

- Recall the matched/paired data setup.
- Observations in two groups not independent (matched; pre-post).
- The sign test is a nonparametric test to evaluate whether one group of measurements tend to be less than/greater than their paired counterparts.

The sign test

Example: β-carotene

- We've seen this data set multiple times! Now, we will focus on the $n = 10$ individuals receiving 15 mg/day of β-carotene supplementation.
- These ten individuals had measurements at baseline (prior to supplementation) and at three months.
- We want to perform an analysis to evaluate if those randomized to receive 15 mg/day "tend to have higher plasma β-carotene levels by the end of three months of treatment."

The sign test

Example: β-carotene

```
. summarize carot0 if dose == 15
```

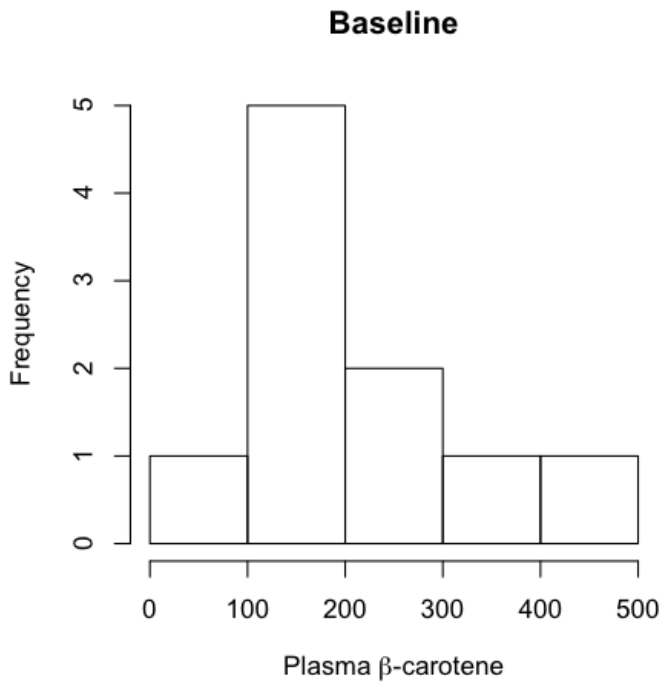
Variable	Obs	Mean	Std. Dev.	Min	Max
carot0	10	220.06	127.944	64.75	496

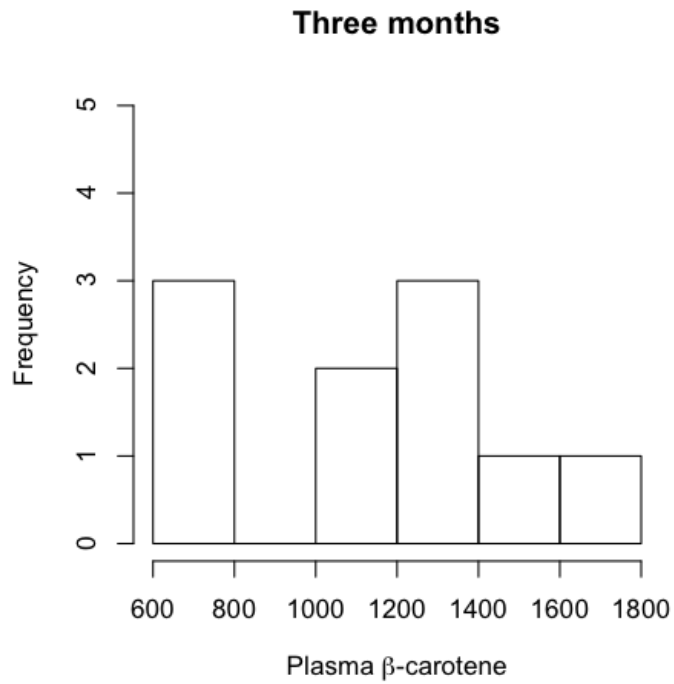
```
. summarize carot1 if dose == 15
```

Variable	Obs	Mean	Std. Dev.	Min	Max
carot1	10	1116.367	317.3562	699	1602.667

The sign test

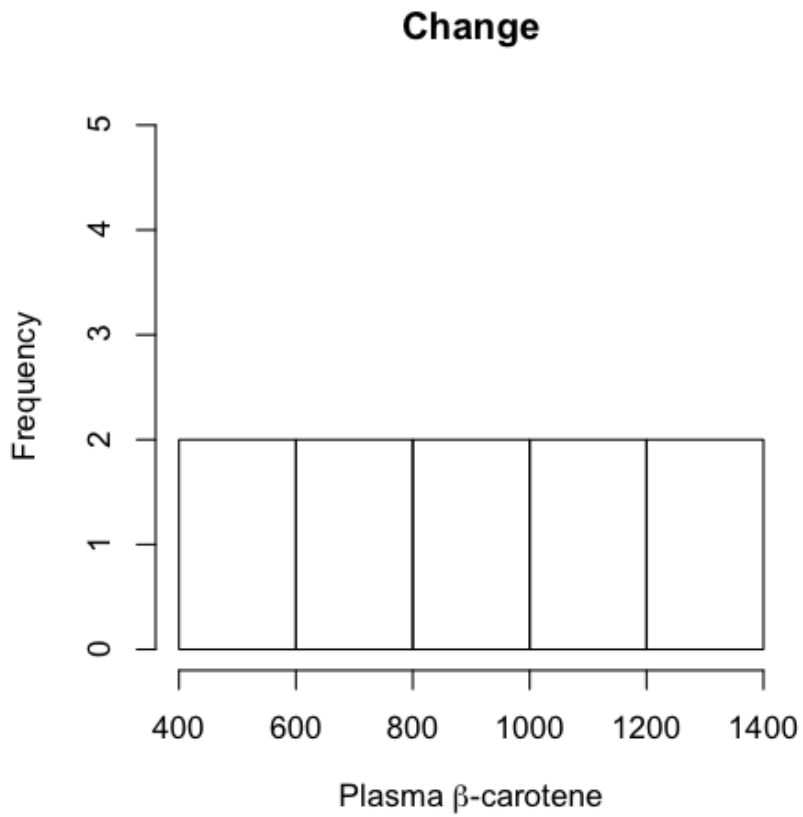
Example: β-carotene





The sign test

Example: β-carotene



The sign test

Example: β-carotene

- The words "tend to have higher plasma β-carotene" are again not nearly specific enough.
- If we wanted to perform an analysis to evaluate if the mean plasma β-carotene were different after three months of treatment, we would turn to the paired t-test.
- In practice, I again wouldn't argue with this in most circumstances. Though, some insist upon employing a nonparametric test like the sign test or the Wilcoxon matched-pairs signed rank test.
- If you do this, *make sure you know what you're actually testing...*

The sign test

Example: β-carotene

- Baseline β-carotene: $x_{01}, x_{02}, \ldots, x_{0n}$
- Three-month β-carotene: $x_{11}, x_{12}, \ldots, x_{1n}$
- Change in β-carotene: $d_1, d_2, \ldots, d_n$
  - Define new variable: $D_i = \text{sign}(d_i)$ (any zeros excluded and taken away from total sample size).
- Null hypothesis: "The median of the differences is zero."
- In which case, $D \sim \text{Binomial}(n, p = 0.5)$, with $\text{E}[D] = n/2$, and $\text{Var}[D] = n/4$.
- Either binomial probabilities or normal approximation.

The sign test

Example: β-carotene (Help from Stata)

```
. gen carotdiff = carot1 - carot0

. signtest carotdiff = 0

Sign test

        sign |     observed    expected
-------------+------------------------
    positive |           10           5
    negative |            0           5
        zero |            0           0
-------------+------------------------
         all |           10          10

One-sided tests:
  Ho: median of carotdiff = 0 vs.
  Ha: median of carotdiff > 0
      Pr(#positive >= 10) =
         Binomial(n = 10, x >= 10, p = 0.5) =  0.0010

  Ho: median of carotdiff = 0 vs.
  Ha: median of carotdiff < 0
      Pr(#negative >= 0) =
         Binomial(n = 10, x >= 0, p = 0.5) =  1.0000

Two-sided test:
  Ho: median of carotdiff = 0 vs.
  Ha: median of carotdiff != 0
      Pr(#positive >= 10 or #negative >= 10) =
         min(1, 2*Binomial(n = 10, x >= 10, p = 0.5)) =  0.0020
```

The sign test

The good

- Simple.
- Distribution-free.

The bad

- Loss of data.
- Throw away information on magnitude of difference.

The ugly

- The median of the differences is simply *not* the difference in the medians!

The Wilcoxon matched-pairs signed rank test

Example: β-carotene

- Same setup as previously:
  - Baseline β-carotene: $x_{01}, x_{02}, \ldots, x_{0n}$
  - Three-month β-carotene: $x_{11}, x_{12}, \ldots, x_{1n}$
  - Change in β-carotene: $d_1, d_2, \ldots, d_n$
- Procedure:
  1. Order the differences d_i.
  2. Ignore the sign; assign each its *rank*, r_i (zeros are given the lowest rank, and ties are given the average rank).
  3. Then, re-assign the sign to the rank.
  4. Let R be the sum of the *positive* ranks.
- Null hypothesis: "pre- and post-tx plasma β-carotene values are being sampled from same distribution."
- Under H_0, $\mathrm{E}[R] = n(n+1)/4$; $\mathrm{Var}[R] = n(n+1)(2n+1)/24$. Yup, you're gonna have to believe me on that.

The Wilcoxon matched-pairs signed rank test

Example: β-carotene

- Define test statistic:

$$Z = \frac{R - \mathrm{E}[R|H_0]}{\sqrt{\mathrm{Var}[R|H_0]}}$$

- Under H_0, $Z \dot{\sim} \mathcal{N}(0,1)$. Generally need $n \geq 20$ in order for this approximation to hold up well.

The Wilcoxon matched-pairs signed rank test

Example: β-carotene (Help from Stata)

```
. signrank carotdiff = 0

Wilcoxon signed-rank test

        sign |      obs    sum ranks    expected
-------------+---------------------------------
    positive |       10           55        27.5
    negative |        0            0        27.5
        zero |        0            0           0
-------------+---------------------------------
         all |       10           55          55

unadjusted variance        96.25
adjustment for ties         0.00
adjustment for zeros        0.00
                           ------
adjusted variance          96.25

Ho: carotdiff = 0
                z =   2.803
    Prob > |z| =   0.0051
```

The Wilcoxon matched-pairs signed rank test

The good

- Distribution-free.
- Doesn't throw away information on magnitudes like the sign test does.

The bad

- A little complicated.
- Exact p-values slightly cumbersome (normal approximation okay for $n \geq 20$.

The ugly

- What on earth are we testing? It is a test of the strong null hypothesis that the two distributions are the same.

Summary

Nonparametric/distribution-free tests

- The Mann-Whitney test (and its challenges)
- Fisher's exact test (and *its* challenges)
- Tests for matched pairs (and yes, *their* challenges)

Not to discourage one from using nonparametric tests—but I do discourage one from using them without knowing exactly what he or she is doing and why!

I leave you with these thoughts

"*Non-parametric tests reflect a desire to get a free lunch.*"
(Ruben Roa-Ureta)

"*An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.*"
(John Tukey)