Biostatistics II

Dr Vedaste Ndahindwa

University of Rwanda
School of Public Health

Contents

Tests of means covered so far

Compare a mean from one group to a constant:

- One sample z-test (variance known)
  - Assumes that the sample mean is normally distributed, with mean μ, and variance σ^2/n
- One sample t-test (variance unknown)
  - Assumes that the sample mean is normally distributed, with mean μ, and variance σ^2/n
  - Assumes that the test statistic, standardized using the sample variance, s^2, follows a t-distribution with $n-1$ degrees of freedom

Nonparametric Analysis | Why nonparametric analysis?

Tests of means covered so far

Compare a mean from one group to mean of another group (paired):

- Paired t-test (variance unknown)
  - Assumes that the sample mean of the differences is normally distributed, with mean μ_d, and variance σ^2/n
  - Assumes that the test statistic, standardized using the sample variance, s^2, follows a t-distribution with $n-1$ degrees of freedom (with n being number of observations per group)

Tests of means covered so far

Compare a mean from one group to mean of another group (unpaired):

- Two-sample t-test
  - Assumes that the sample means of each group is normally distributed, with mean μ_1 and μ_2 and variance σ_1^2/n_1 and σ_2^2/n_2
  - Assumes that the test statistic, standardized using the sample variances, s_1^2 and $s^2 - 2$, follows a t-distribution with ν degrees of freedom (ν depends on whether assume variance are equal)

Distributional assumptions

All of these test make an assumption about the sample means – that they are approximately normally distributed.

The **t-test** also relies on the assumption that your underlying population is normal.

What about non-normal population, or small samples?

If you have small samples and/or an underlying distribution that is not normal, you may not want to make the necessary assumptions.

In this case, we can rely on a set of **nonparametric tests** - tests that do not have assumptions about the distributions of parameters

One sample median test

After the change in HIV treatment guidelines, the goal is for patients to come before their CD4 is below 200.

We have a sample of 15 patients starting treatment at our ART clinics.

Using this data we want to test the hypothesis that the median CD4 count at initiation is different from 200.

- $H_0 : median = 200$
- $H_a : median \neq 200$

One sample median test

Individual	CD4 baseline	Hypothesized median	Difference	Sign
1	147	200	-53	-
2	115	200	-85	-
3	144	200	-56	-
4	255	200	55	+
5	231	200	31	+
6	128	200	-72	-
7	253	200	53	+
8	366	200	166	+
9	221	200	21	+
10	115	200	-85	-
11	75	200	-125	-
12	53	200	-147	-
13	204	200	4	+
14	282	200	82	+
15	291	200	91	+

Sign Rank Test

Step 1: Calculate the difference (d_i) between observations and hypothesized medians and rank according to the magnitude of the differences

Step 2: Sum the ranks of the positive differences (T)

Step 3: Calculate the test statistic

$$z = \frac{T - [n(n+1)/2]/2}{\sqrt{Var(ranks)}}$$

where: $Var(ranks) = \frac{\sum_{i=1}^{n} r_i^2}{4}$ $where\ (d_i \neq 0)$

Step 4: Compare to standard normal.

No	CD4 baseline	Hypoth. median	Difference	Absolute value	Rank	Sign	Rank squared
1	147	200	-53	53	4.5	-	20.25
2	115	200	-85	85	10.5	-	110.25
3	144	200	-56	56	7	-	49
4	255	200	55	55	6	+	36
5	231	200	31	31	3	+	9
6	128	200	-72	72	8	-	64
7	253	200	53	53	4.5	+	20.25
8	366	200	166	166	15	+	225
9	221	200	21	21	2	+	4
10	115	200	-85	85	10.5	-	110.25
11	75	200	-125	125	13	-	169
12	53	200	-147	147	14	-	196
13	204	200	4	4	1	+	1
14	282	200	82	82	9	+	81
15	291	200	91	91	12	+	144
				Sum of Pos Ranks	52.5		
				Sum of Neg Ranks	67.5		
				Var of Ranks	309.75		

Sum of Pos Ranks	52.5
Sum of Neg Ranks	67.5
Var of Ranks	309.75

$$z = \frac{52.5 - (15 * 16/2)/2}{\sqrt{309.75}} = -0.426$$

$$P(|z| > 0.426) = 0.670$$

Stata Output

```
. signrank  cd4=200

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |        8        52.5          60
    negative |        7        67.5          60
        zero |        0           0           0
-------------+---------------------------------
         all |       15         120         120

unadjusted variance        310.00
adjustment for ties         -0.25
adjustment for zeros         0.00
                       ----------
adjusted variance          309.75

Ho: cd4 = 200
             z =   -0.426
    Prob > |z| =    0.6700
```


Two sample paired non-parametric tests

We use the same tests – sign test and signed-rank test – for paired data, but now applied to differences between pairs.

We have a sample of 15 patients starting treatment at our ART clinics.

Using this data we want to test the hypothesis that the median CD4 count at initiation is different from after one-year of treatment.

- $H_0 : median_1 = median_2$
- $H_a : median_1 \neq median_2$

No	CD4 baseline	CD4 year1	Difference
1	147	765	618
2	115	298	183
3	144	157	13
4	255	731	476
5	231	453	222
6	128	177	49
7	253	716	463
8	366	866	500
9	221	353	132
10	115	282	167
11	75	242	167
12	53	167	114
13	204	993	789
14	282	375	93
15	291	614	323

Stata Output

```
. signrank cd4baseline= cd4year1

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |        0           0          60
    negative |       15         120          60
        zero |        0           0           0
-------------+---------------------------------
         all |       15         120         120

unadjusted variance       310.00
adjustment for ties        -0.13
adjustment for zeros        0.00
                      ----------
adjusted variance         309.88

Ho: cd4baseline = cd4year1
             z =  -3.408
    Prob > |z| =   0.0007
```

What about unmatched data?

Suppose we want to know if the median baseline CD4 differs between men and women.

We have a sample of 20 randomly sampled individuals – 11 women, 9 men

- H_0 : $median_1 = median_0$
- H_a : $median_1 \neq median_0$

Patient	Female	CD4y0
1	0	147
2	1	115
3	0	144
4	1	194
5	0	255
6	1	231
7	0	128
8	0	253
9	0	112
10	1	176
11	0	366
12	1	221
13	1	115
14	1	75
15	1	53
16	1	204
17	1	85
18	0	329
19	1	282
20	0	33

Wilcoxon Rank Sum Test

Step 1: Order outcomes and assign ranks by order

Step 2: Sum ranks by group

- Group 0 = T0
- Group 1= T1

Step 3: Calculate the test statistic

$$z = \frac{T_1 - n_1(n_1 + n_0 + 1)/2}{\sqrt{n_1(n_0)s^2/(n_1 + n_0)}}$$

where $s^2 = \frac{sum_{i=1}^{n_1+n_0}(r_i - \bar{r})^2}{n_1+n_0-1}$

Step 4: Compare to standard normal

Patient	Female	CD4y0	Ranks	$(r_i - \bar{r})^2$
20	0	33	1	100
15	1	53	2	81
14	1	75	3	64
17	1	85	4	49
9	0	112	5	36
2	1	115	6.5	20.25
13	1	115	6.5	20.25
7	0	128	8	9
3	0	144	9	4
1	0	147	10	1
10	1	176	11	0
4	1	194	12	1
16	1	204	13	4
12	1	221	14	9
6	1	231	15	16
8	0	253	16	25
5	0	255	17	36
19	1	282	18	49
18	0	329	19	64
11	0	366	20	81
Sum of G0 Ranks		105		
Sum of G1 Ranks		105		
Average rank		11		
S^2		35.23684		

$$z = \frac{105 - 11(11+9+1)/2}{\sqrt{11(9)35.2/(11+9)}}$$

$$z = -0.795$$

$$P(|z| > 0.795) = 0.424$$

Stata Output

```
. ranksum cd4y0, by(female)

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

      female |      obs    rank sum    expected
-------------+---------------------------------
           0 |        9         105        94.5
           1 |       11         105       115.5
-------------+---------------------------------
    combined |       20         210         210

unadjusted variance      173.25
adjustment for ties       -0.13
                     ----------
adjusted variance        173.12

Ho: cd4y0(female==0) = cd4y0(female==1)
             z =   0.798
    Prob > |z| =   0.4249
```

Parametric versus Nonparametric Tests

Parametric Tests

- Used if the underlying distributions of the data are **known** or can be assumed to be **normally** distributed
- μ and σ are known as **parameters** of a population

Nonparametric Tests

- Used if the underlying distributions of the data are **unknown** or are **not normally** distributed
- Fewer assumptions about the underlying distribution
    - Wilcoxon Rank-sum test assumes same shape
- Compares medians rather than means