PREVIOUSLY

- parametric statistics

  in estimation and hypothesis testing...

  construction of confidence intervals
  computing of p-values
  classical significance testing

  depend on assumptions about the underlying distribution of the data (or on the Central Limit Theorem)

- nonparametric statistics

  no assumptions about underlying distribution of data

  used when...

  assumptions about underlying distributions are not met (data not normally distributed)

  sample sizes are small ($n \leq 30$, cannot rely on Central Limit Theorem)

- advantages

  can be used with non-normally distributed data

  can be used with discrete data (nominal, ordinal)

  simpler computations (questionable since it is rare to have to do any calculations without a program that is designed for statistical methods, e.g. Statcrunch, Statdisk, SAS)

  effect of outlier(s) less than when using a parametric method since actual data values are not used in computation

- disadvantages

  'waste' information (e.g. if evaluating change in some quantity within subjects over time, the sign test would only classify change as -/0/+ while a paired t-test would use the actual measured change)

  not as 'efficient' (most texts use not as 'powerful') - there is more of a chance of not rejecting a false null hypothesis

THREE TESTS

- sign test
- Wilcoxon Signed-Ranks Test for Matched Pairs (equivalent to a paired t-test)
- Wilcoxon Rank-Sum Test or Two Independent Samples (equivalent to two-sample t-test)
- there are MANY more nonparametric tests

SIGN TEST

- used to test

  claims involving matched pairs of data
  claims involving nominal data
  claims about the median of a single population

- requirements

  same issues as to sampling as before (random, representative)

  Note: No assumptions about the underlying distribution of the data.

  n (total of +/- results) $\leq$ 25 table, n > 25 normal approximation

Example from Rosner… 45 people are participants in a study to test whether two treatments differ in their ability to prevent redness resulting from exposure to sunlight… each person has a different treatment applied to their right and left arms...

- results... nominal data (and matched pairs)...

22	treatment A arm not as red as treatment B arm (-)
18	treatment B arm not as red as treatment A arm (+)
5	no difference

- as with McNemar's Test, the only data used are from the discordant results (the 5 with no difference are not used in computation of the test statistic)

- claim… there is no difference between treatments
- null hypothesis…

$$H_0: p = 0.5$$
$$H_1: p \neq 0.5$$

- test statistic…

$$z = \frac{(x + 0.5) - (n / 2)}{\sqrt{n} / 2} = \frac{(18 + 0.5) - (40 / 2)}{\sqrt{40} / 2} = \frac{-1.5}{3.16} = -0.4743$$

where… x=number of times lest frequent sign occurs, and
n=total + and - signs

or…

$$z = \frac{|(n_+) - (n_-)| - 1}{\sqrt{n}} = \frac{|18 - 22| - 1}{\sqrt{40}} = \frac{3}{6.32} = 0.4743$$

- critical value… with α=0.05, two-tail test, z=1.96

- P-value... from Statcrunch, p=0.3176, P-value=0.6352

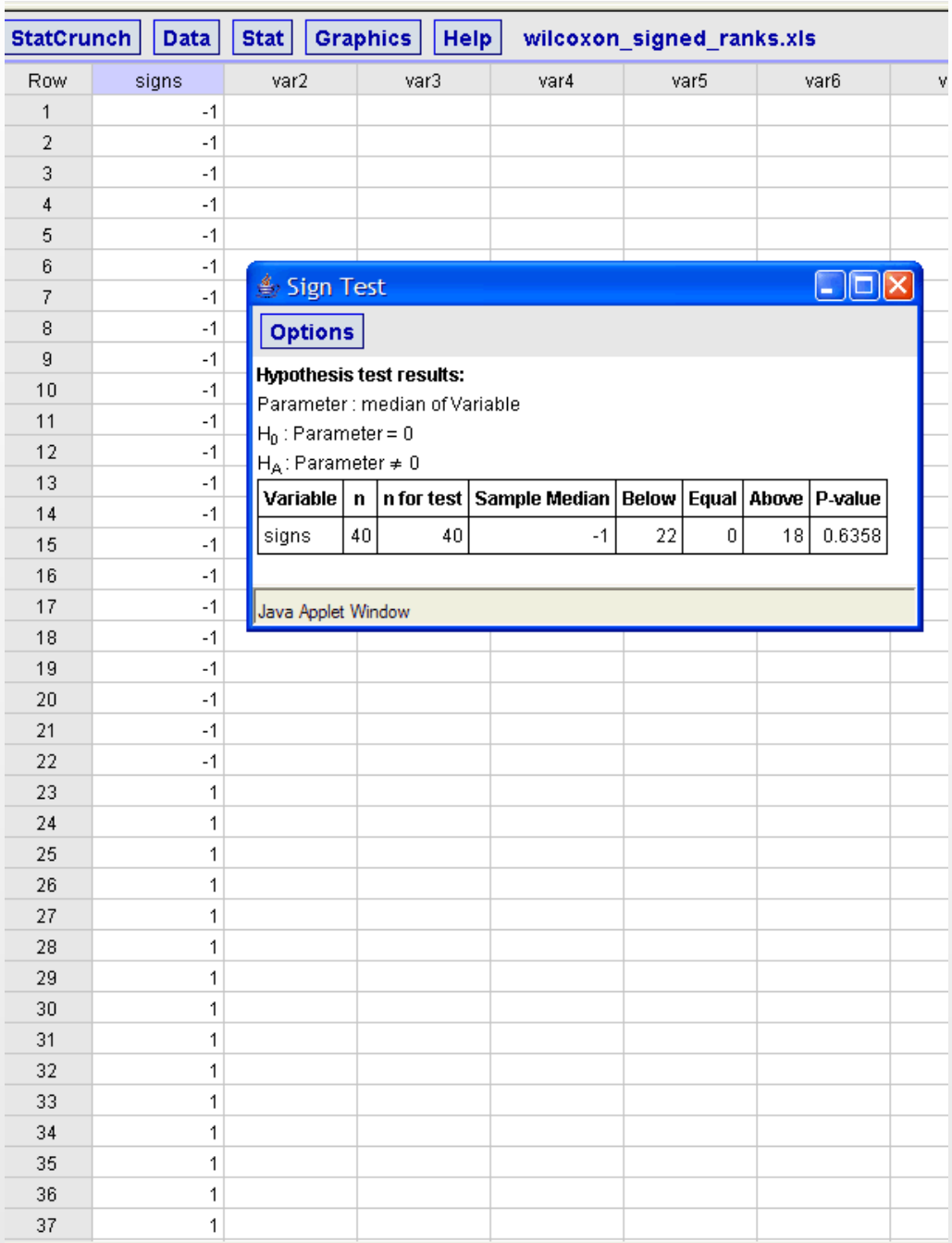
- conclusion... fail to reject the null hypothesis
  no evidence that the two treatments differ

- alternative... fixed number of trials
  trials are independent
  each trial has only two possible outcomes
  probabilities remain constant for each trials

  binomial...
  n=40,k=18,p(k≤18)=0.3179,P-value=0.6358
  n=40,k=22,p(k≥22)=0.3179,P-value=0.6358

- when n≤25... use values from table A-2, values are 'n' and 'k'
  from binomial tables with p=0.50

- using Statcrunch...



Example from Triola... 12 measurements of body temperature... test to see if the body temperatures come from a population with a median body temperature of 98.6...

97.6(-)	97.5(-)	98.6	98.2(-)	98.0(-)	99.0(+)
98.5(-)	98.1(-)	98.4(-)	97.9(-)	97.9(-)	97.7(-)

what is normal body temperature...

http://www.health.harvard.edu/press_releases/normal_body_temperature.htm

- claim… median temperature of population is 98.6
- null hypothesis…

$$H_0: median = 98.6$$
$$H_1: median \neq 98.6$$

- test statistic…

$$z = \frac{|(n_+) - (n_-)| - 1}{\sqrt{n}} = \frac{|1 - 10| - 1}{\sqrt{11}} = \frac{8}{3.32} = 2.41$$

however…
cannot use test statistic with n=11

- P-value… from binomial distribution with n=11, p=0.50, probability of $k \leq 1$ = 0.0107, P-value=0.021
- conclusion… reject the null hypothesis
median not equal to 1.96

WILCOXON SIGNED-RANKS TEST

- used to test

  claims involving matched pairs of data (do differences come from a population with a median value of zero)

  claims about the median of a single population

- requirements

  same issues as to sampling as before (random, representative)

  Note: No assumptions about the underlying distribution of the data.

  n (total of +/- results) $\leq$ 30 table, n > 30 normal approximation

Example from Rosner... 45 people are participants in a study to test whether two treatments differ in their ability to prevent redness resulting from exposure to sunlight... each person has a different treatment applied to their right and left arms...

- previously... results... nominal data (and matched pairs)...

22	treatment A arm not as red as treatment B arm (-)
18	treatment B arm not as red as treatment A arm (+)
5	no difference

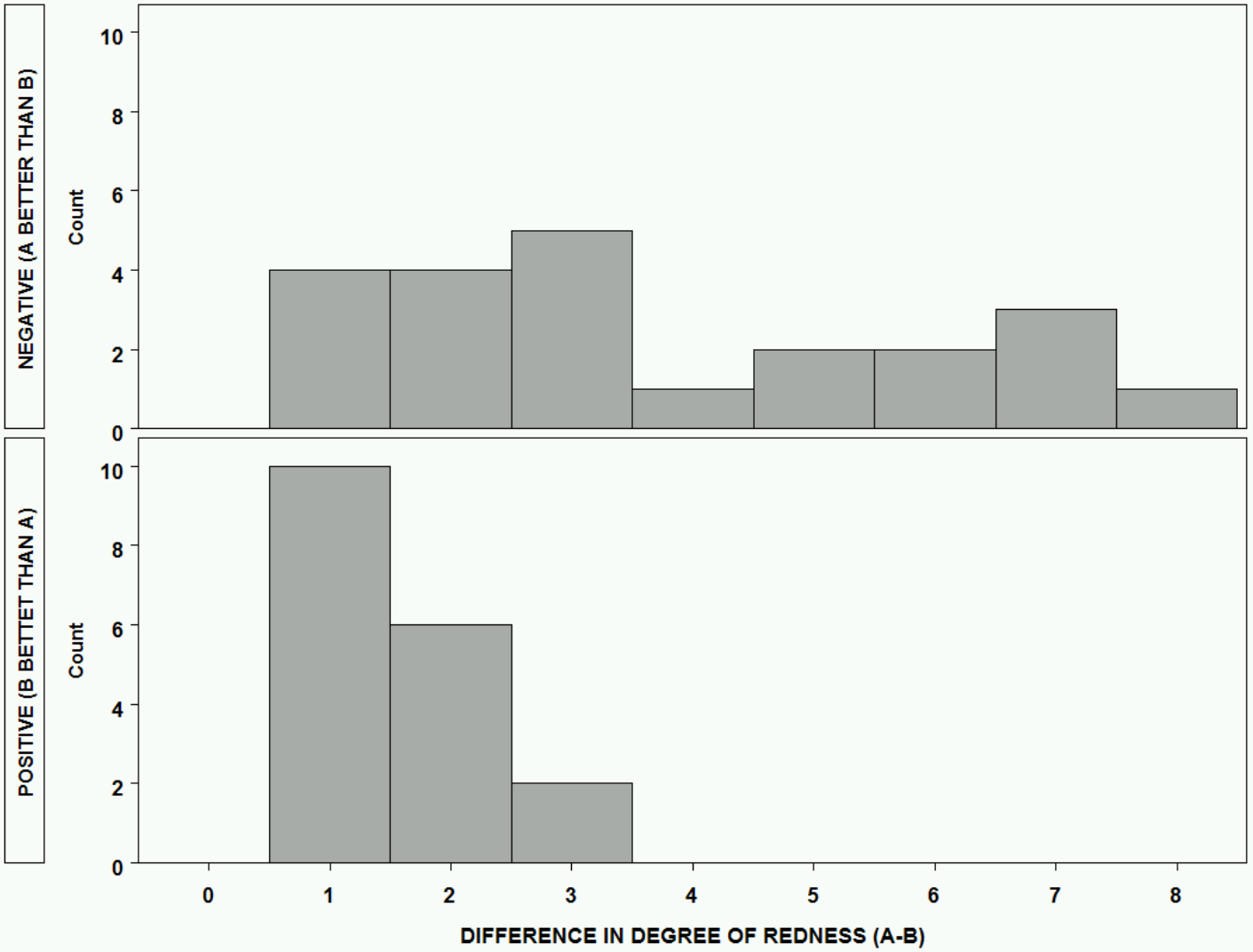
- new results measured in difference in degree of redness…

-8	-7	-7	-7	-6	-6	-5	-5	-4	-3
-3	-3	-3	-3	-2	-2	-2	-2	-1	-1
-1	-1	3	3	2	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1

  - \- treatment A arm not as red as treatment B arm (n=22)
  - \+ treatment B arm not as red as treatment A arm (n=18)
  - 0 treatment A and B not different (n=5, not shown in table)

- as with the Sign Test, the only data used are from the discordant results (the 5 with no difference are not used in computation of the test statistic)

- distribution of actual differences shows much more information than merely comparing the number of positive and negative values...



- arrange results in a table… rank differences from lowest to highest without regard to the sign of the difference…

1(-)	1(-)	1(-)	1(-)	1	1	1	1	1	1
1	1	1	1	2(-)	2(-)	2(-)	2(-)	2	2
2	2	2	2	3(-)	3(-)	3(-)	3(-)	3(-)	3
3	4(-)	5(-)	5(-)	6(-)	6(-)	7(-)	7(-)	7(-)	8(-)

- assign ranks to each value… ties are assigned average rank…

7.5(-)	7.5(-)	7.5(-)	7.5(-)	7.5	7.5	7.5	7.5	7.5	7.5
7.5	7.5	7.5	7.5	19.5(-)	19.5(-)	19.5(-)	19.5(-)	19.5	19.5
19.5	19.5	19.5	19.5	28(-)	28(-)	28(-)	28(-)	28(-)	28
28	32(-)	33.5(-)	33.5(-)	35.5(-)	35.5(-)	38(-)	38(-)	38(-)	40(-)

- sum the ranks assigned to the positive and negative values…

  positive… 7.5(10) + 19.5(6) + 28(2) = 248

  negative… 7.5(4) + 19.5(4) + 28(5) + 32(1) + 33.5(2) + 35.5(2) + 38(3) + 40(1) = 572

  (ranks should sum to n(n+1)/2 = 40(41)/2 = 820, if no difference, the positive and negative ranks should each sum to 410 and the test is trying to find if the sums of 248 and 572 are different enough from 410 to be a rare event)

  T = smaller of +/- ranks = 248

- test statistic… $$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{248 - \frac{40(41)}{4}}{\sqrt{\frac{40(41)(81)}{24}}} = \frac{-162}{74.40} = -2.18$$

- critical value… from normal table, α=0.5, two-tail, 1.96
- P-value… from normal table, p=0.0146,
  P-value=0.29
- conclusion… reject the null hypothesis
  there is a difference in the effect of treatments A and B (a different conclusion than was found with the sign test - sign test is less powerful so one is more likely to accept an incorrect null hypothesis)

- using Statcrunch...

Row	a-b	var2	var3	var4	var5	var6	var7
1	-8						
2	-7						
3	-7						
4	-7						
5	-6						
6	-6						
7	-5						
8	-5						
9	-4						
10	-3						
11	-3						
12	-3						
13	-3						
14	-3						
15	-2						
16	-2						
17	-2						
18	-2						
19	-1						
20	-1						
21	-1						
22	-1						
23	3						
24	3						
25	2						
26	2						
27	2						
28	2						
29	2						
30	2						
31	1						
32	1						
33	1						
34	1						
35	1						
36	1						
37	1						

Wilcoxon Signed Ranks

Options

Hypothesis test results:

Parameter : median of Variable

H_0 : Parameter = 0

H_A : Parameter ≠ 0

Variable	n	n for test	Median Est.	Wilcoxon Stat.	P-value	Method
a-b	40	40	-1	248	0.0287	Norm. Approx.

Java Applet Window

- alternative... imagine n (number of +/-) = 5

  think of five chips numbered 1 through 5 on one side, 0 on the other (0 represents a negative value, 1 through 5 is a positive rank)... toss the five chips in the air... 32 possible ways that the five chips can land (2^5)

  sample space (sum of positive ranks)...

sum=0	0,0,0,0,0		p=1/32
sum=1	1,0,0,0,0		p=1/32
sum=2	0,2,0,0,0		p=1/32
sum=3	0,0,3,0,0	1,2,0,0,0	p=2/32
sum=4	0,0,0,4,0	1,0,3,0,0	p=2/32
sum=5	0,0,0,0,5		p=1/32

  only 'rare event' is sum=1 (p=2/32=.0625, 1-tail)
  same logic applies to any n (entries in table A-8)

Example from Triola... 12 measurements of body temperature... test to see if the body temperatures come from a population with a median body temperature of 98.6...

rather than use just signs (below/above median), use the actual differences from the median value of 98.6 and follow the same procedure as just used with the Rosner data...

97.6(-1.0)	97.5(-1.1)	98.6 (0)	98.2(-0.4)	98.0(-0.6)	99.0(+0.4)
98.5(-0.1)	98.1(-0.5)	98.4(-0.2)	97.9(-0.7)	97.9(-0.7)	97.7(-0.9)

0.1 (-)	0.2 (-)	0.4 (-)	0.4 (+)	0.5 (-)	0.6 (-)
0.7 (-)	0.7 (-)	0.9 (-)	1.0 (-)	1.1 (-)	

1(-)	2(-)	3.5(-)	3.5	5(-)	6(-)
7.5(-)	7.5(-)	9(-)	10(-)	11(-)	

- sum the ranks assigned to the positive and negative values...

  positive... 3.5
  negative... 1(1) + 2(1) + 3.5(1) + 5(1) + 6(1) + 7.5(2) + 9(1) + 10(1) + 11(1) = 62.5

  (ranks should sum to n(n+1)/2 = 11(12)/2 = 66, if no difference, the positive and negative ranks should each sum to 33 and the test is trying to find if the sums of 3.5 and 62.5 are different enough from 33 to be a rare event)

  T = smaller of +/- ranks = 3.5

- test statistic... T since $n \leq 30$

- critical value... from table A-8 with $n=11, \alpha=0.5$, two-tail, 11

- conclusion... no reason to reject the null hypothesis that the median body temperature is 98.6

WILCOXON RANK-SUM TEST

- used to test

  claims involving independent samples (do samples come from two populations with equal medians), sometimes referred to as the Mann-Whitney U Test

- requirements

  same issues as to sampling as before (random, representative)

  Note: No assumptions about the underlying distribution of the data.

  n>10 in both samples, use normal approximation, otherwise must use table values (no table in Triola)

- from… http://www.socr.ucla.edu/Applets.dir/WilcoxonRankSumTable.html

Wilcoxon Rank-Sum Table

You can also find a Calculator for the Wilcoxon-Mann-Whitney Significance here

This table shows the critical values values of the Wilcoxon-Mann-Whitney statistics (U_S) for various sample sizes (N_1 and N_2) and p-values (p).

N_1	N_2	p=0.2	0.1	0.05	0.02	0.01	0.002	0.001
3	2	U_S = 6						
3	3	8	9					
4	2	8						
4	3	11	12					
4	4	13	15	16				
5	2	9	10					
5	3	13	14	15				
5	4	16	18	19	20			
5	5	20	21	23	24	25		
6	2	11	12					
6	3	15	16	17				
6	4	19	21	22	23	24		
6	5	23	25	27	28	29		
6	6	27	29	31	33	34		
7	2	13	14					
7	3	17	19	20	21			
7	4	22	24	25	27	28		
7	5	27	29	30	32	34		
7	6	31	34	36	38	39	42	
7	7	36	38	41	43	45	48	49
8	2	14	15	16				
8	3	19	21	22	24			
8	4	25	27	28	30	31		
8	5	30	32	34	36	38	40	
8	6	35	38	40	42	44	47	48
8	7	40	43	46	49	50	54	55
8	8	45	49	51	55	57	60	62
9	1	9						
9	2	16	17	18				
9	3	22	23	25	26	27		
9	4	27	30	32	33	35		
9	5	33	36	38	40	42	44	45
9	6	39	42	44	47	49	52	53
9	7	45	48	51	54	56	60	61

Example from Rosner… A protocol of meditation therapy is administered once a day to 20 patients with anxiety. The patients are given a psychiatric exam at baseline and at a follow-up exam 2 months later. The degree of improvement is rated on a 10-point scale, with 1 indicating the most improvement and 10 the least improvement. Similarly, 26 comparably affected patients with anxiety are given standard psychotherapy and are asked to come back 2 months later for a follow-up exam.

- study group...

1	1	1	2	2	2	2	3	3	3
3	3	3	3	4	4	4	5	5	6

- control group

2	2	3	3	3	3	3	4	4	4
5	5	5	5	5	5	5	5	6	6
6	6	7	7	8	9				

- combined groups... study group values are shaded...

1	1	1	2	2	2	2	2	2	3
3	3	3	3	3	3	3	3	3	3
3	4	4	4	4	4	4	5	5	5
5	5	5	5	5	5	5	6	6	6
6	6	7	7	8	9				

- assign ranks...

2	2	2	6.5	6.5	6.5	6.5	6.5	6.5	15.5
15.5	15.5	15.5	15.5	15.5	15.5	15.5	15.5	15.5	15.5
15.5	24.5	24.5	24.5	24.5	24.5	24.5	32.5	32.5	32.5
32.5	32.5	32.5	32.5	32.5	32.5	32.5	40	40	40
40	40	43.5	43.5	45	46				

- sum the ranks assigned to study group…

  2(3) + 6.5(4) + 15.5(7) + 24.5(3) + 32.5(2) +40(1) = 319 = R

  control ranks sum = 762

  (ranks should sum to n(n+1)/2 = 46(47)/2 = 1081, if no difference, the positive and negative ranks should each sum to 540.5 and the test is trying to find if the sums of 319 and 762 are different enough from 540.5 to be a rare event)

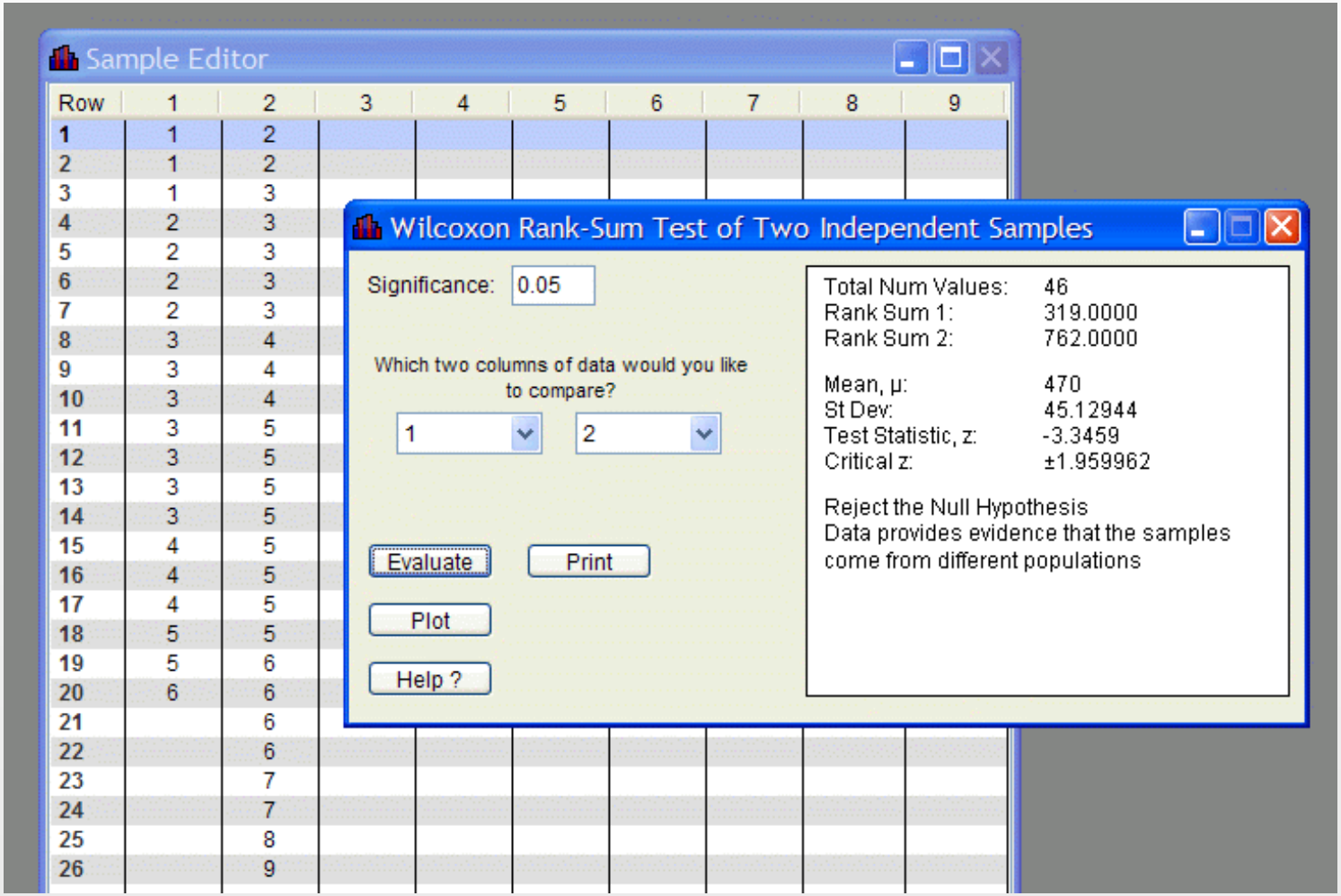
- test statistic… $z = (R - \mu_R) / \sigma_R$

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{20(20 + 26 + 1)}{2} = 10(47) = 470$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{20(26)(20 + 26 + 1)}{12}} = 45.129$$

- test statistic... $z = (R - \mu_R) / \sigma_R = (319 - 470) / 45.129 = -3.35$
- critical value... from normal table, α=0.05, two tail, 1.96
- P-value... from normal table, p=0.0004, P-value=0.0008
- conclusion... reject the null hypothesis
  study group improved more than the control group

- same problem using Statdisk (on Triola CD)...



- same problem using Statcrunch (notice that Statcrunch uses the Mann-Whitney test, see problem 12-4.11 in Triola)...

Row	study	control	var3	var4	var5	var6	var7
1	1	2					
2	1	2					
3	1	3					
4	2	3					
5	2	3					
6	2	3					
7	2	3					
8	3	4					
9	3	4					
10	3	4					
11	3	5					
12	3	5					
13	3	5					
14	3	5					
15	4	5					
16	4	5					
17	4	5					
18	5	5					
19	5	6					
20	6	6					
21		6					
22		6					
23		7					
24		7					
25		8					
26		9					

Mann-Whitney

Options

Hypothesis test results:

m1 = median of study

m2 = median of control

Parameter : m1 - m2

H_0 : Parameter = 0

H_A : Parameter ≠ 0

Difference	n1	n2	Diff. Est.	Test Stat.	P-value	Method
m1 - m2	20	26	-2	319	0.0007	Norm. Approx.

Java Applet Window