Chapter 4: Alternatives to the t tools

Class 7 2/23/09 M

Slide 1 Chapter 4: Alternatives to the t tools

NOTES:

HW 5 due Weds 2/25/09 9:50

Submit as Myname-HW5.doc (or *.rtf)

- Finish Chapter 4 **Wilcoxon rank sum, signed rank tests, Fisher's sign test, Welch's unequal variance t test**
- Comment on Chapter 4 conceptual problems in Blackboard Vista4
- Computation Problem 5
  - Problem 4.31 Effect of group therapy on breast cancer patients.

Slide 2 HW 5 due Weds 2/25/09 9:50

NOTES:

HW 6 due Monday 3/1/09 9:50

Submit as Myname-HW6.doc (or *.rtf)

- Read Chapter 5 **Comparisons among several samples**
- Comment on Chapter 5 conceptual problems in Blackboard Vista4
- Computation Problem 6
  - Problem 4.30 Sunlight protection factor

Slide 3 HW 6 due Monday 3/1/09 9:50

NOTES:

HW 7 due Thursday 3/4/09 Noon

Submit as Myname-HW7.doc (or *.rtf)

- Read Chapter 6 **Comparisons among several samples**
- Comment on Chapter 6 conceptual problems in Blackboard Vista4
- Computation Problem 7
  - Problem 5.25 Duodenal ulcers

Slide 4 HW 7 due Thursday 3/4/09 Noon

NOTES:

Chapter 4: Alternatives to the *t* tools

Note: Sleuth has MANY errors and omissions!

- Permutation tests [Not a solution to unequal variance]
- Wilcoxon's Rank Sum Test (same probability model as Mann-Whitney U test)
  - Ties corrections not in sleuth & exact p values
- Repeated Measures Tests based on ranks: Wilcoxon Sign Rank & Fisher's Sign Test
- Parametric vs. Nonparametrics
  - Power efficiency not an issue, ties not that much of an issue
  - Dealing with covariates & estimating effect sizes can be an issue
    - Hodges-Lehman estimators
- Unequal variance (Welch's) t test: some theoretical and practical problems
- Supplemental material
  - Two-sample binomial test (covered in Sleuth Ch 19)
  - The Fligner-Policello test, a rank-based test for samples with unequal variance

Slide 5 Chapter 4: Alternatives to the t tools

NOTES:

Case 4.1: Space Shuttle O-Ring Failures

See Case 4.1 Movie, solved in Matlab™ & SPSS

Display 4.1

Numbers of O-ring incidents on 24 space shuttle flights prior to the Challenger disaster

Two problems: unequal variance & ties

Launch Temperature	Number of O-Ring Incidents
Below 65° F	1 1 1 3
Above 65° F	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2

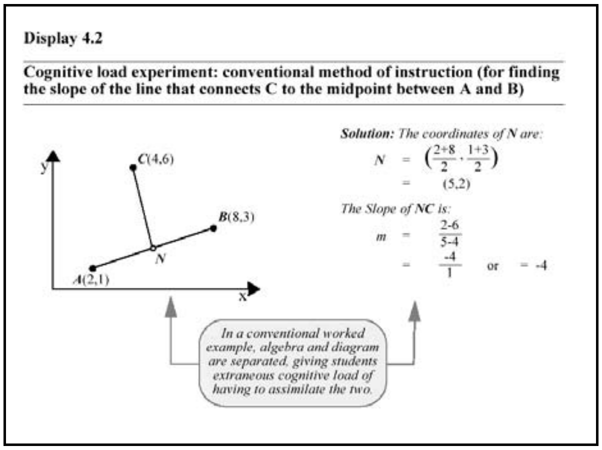
Summary of Statistical Findings

There is strong evidence that the number of O-ring incidents was associated with launch termperature in these 24 launches ...

p- value = 0.009 from a **permutation test** on the *t* statistic

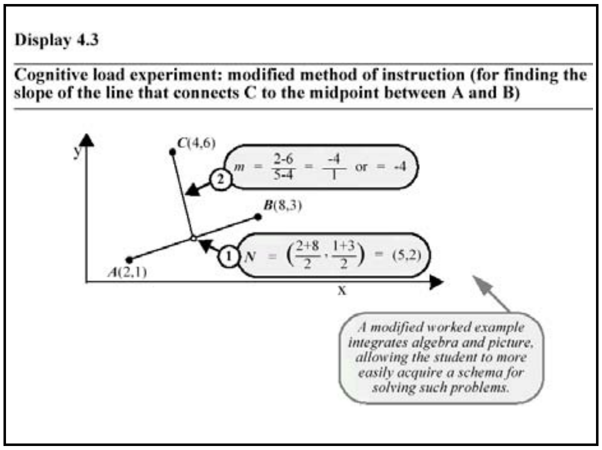
Slide 6 Case 4.1: Space Shuttle O-Ring Failures

NOTES:



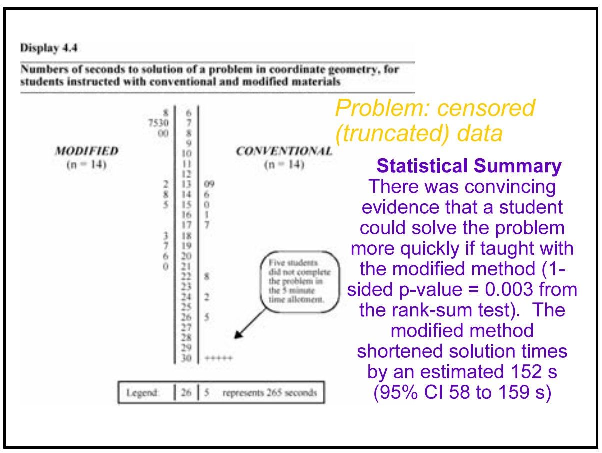
Slide 7 Case 4.2: Cognitive Load

NOTES:



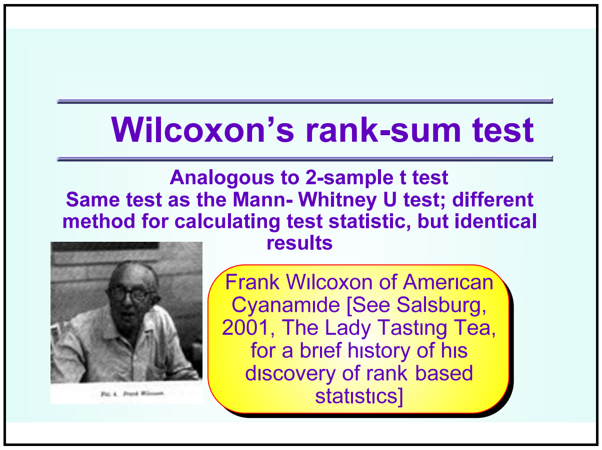
Slide 8

NOTES:



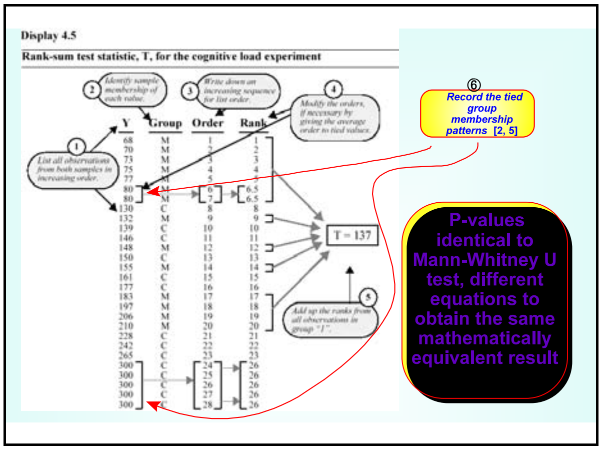
Slide 9

NOTES:



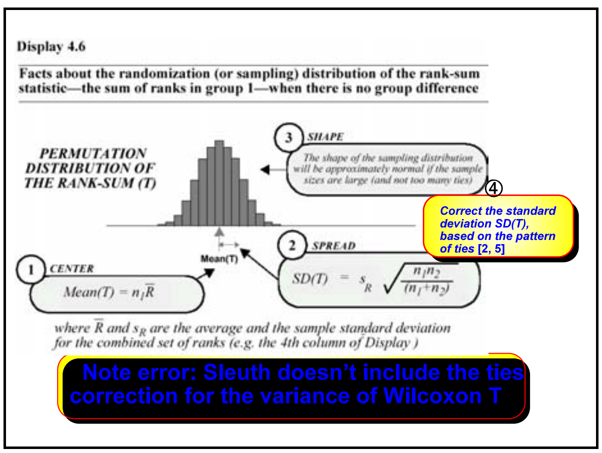
Slide 10 Wilcoxon's rank-sum test

NOTES:



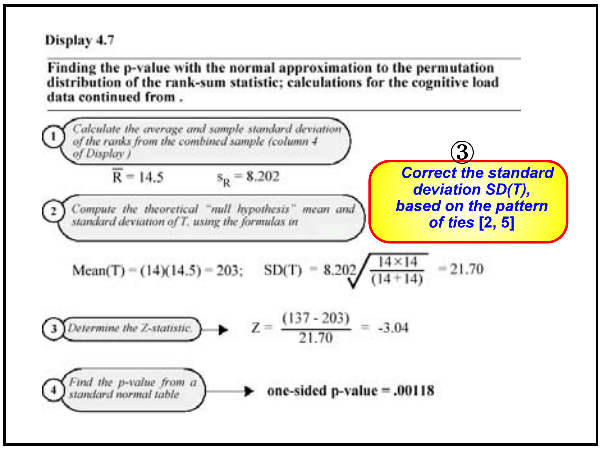
Slide 11 Wilcoxon Rank Sum Statistic

NOTES:



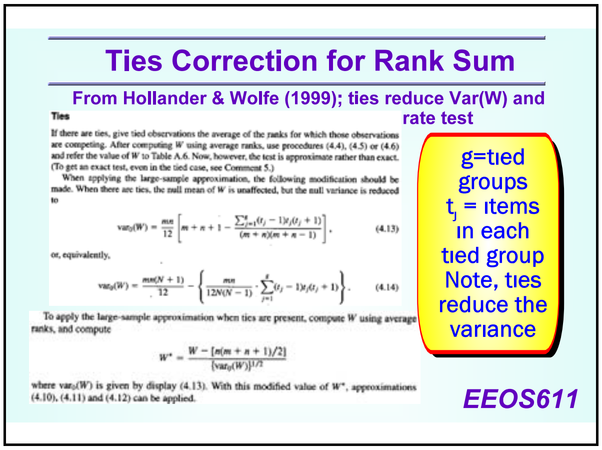
Slide 12

NOTES:



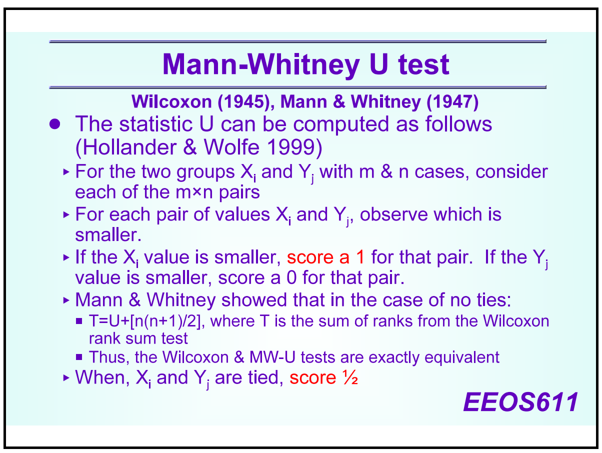
Slide 13

NOTES:



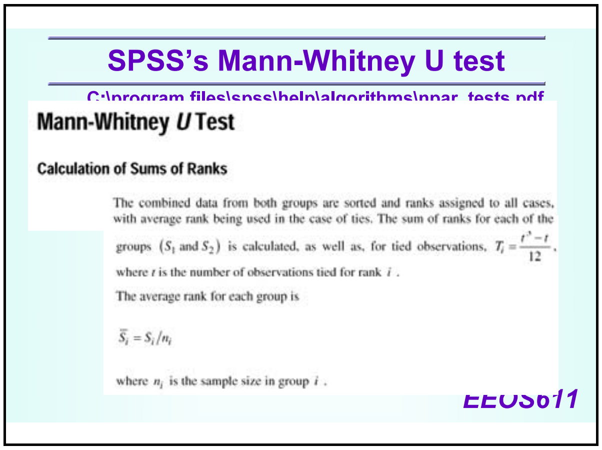
Slide 14 Ties Correction for Rank Sum

NOTES:



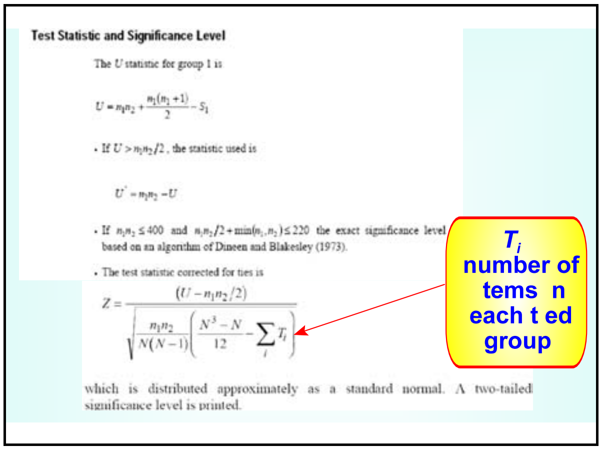
Slide 15 Mann-Whitney U test

NOTES:



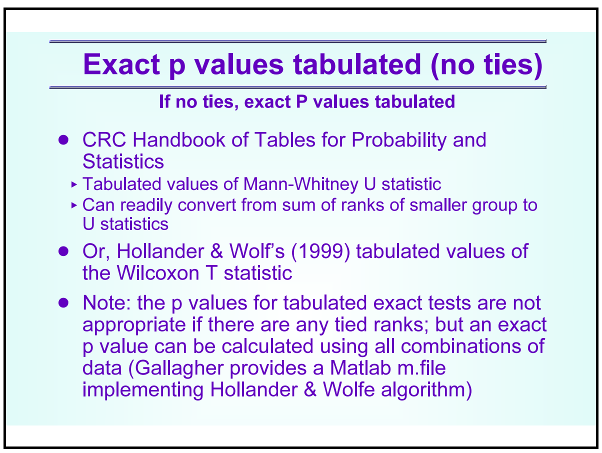
Slide 16 SPSS's Mann-Whitney U test

NOTES:



Slide 17

NOTES:



Slide 18 Exact p values tabulated (no ties)

NOTES:

Other alternatives for two independent samples

4.3.1 Permutation tests

Slide 19 Other alternatives for two independent samples

NOTES:

Ties and Case 4.1

Sleuth argues that rank tests inappropriate because of ties. The real problem is unequal variance (Behrens Fisher)!

Display 4.1

Numbers of O-ring incidents on 24 space shuttle flights prior to the Challenger disaster

Launch Temperature	Number of O-Ring Incidents
Below 65° F	1 1 1 3
Above 65° F	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2

Sleuth argues that tied values pose problems for the Wilcoxon rank sum test, **but** Siegel (1956) states that the Wilcoxon rank sum test is robust in the presence of ties. But the Wilcoxon test p values are only approximate with ties, and the normal approximation is conservative. There is an exact test with ties (computer intensive but Gallagher has programmed).

Slide 20 Ties and Case 4.1

NOTES:

Display 4.1

Numbers of O-ring incidents on 24 space shuttle flights prior to the Challenger disaster

Launch Temperature	Number of O-Ring Incidents
Below 65° F	1 1 1 3
Above 65° F	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2

Independent Samples Test

	t	df	Sig. (2-tailed)	95% Confidence Interval of the Difference Lower	Upper
Equal variances assumed	3.9	22	.00079	.61	2.0
Equal variances not assumed	2.5	3.34	.07690	-.25	2.8

Note the 100-fold difference in p values

Test Statistics[b]

	Incident
Mann-Whitney U	6.000
Wilcoxon W	216.000
Z	-3.301
Asymp. Sig. (2-tailed)	.000963
Exact Sig. [2*(1-tailed Sig.)]	.005082[a]

a. Not corrected for ties.

b. Grouping Variable: Temperature

Slide 21

NOTES:

Wilcoxon rank sum test assumptions

Nonparametric tests distribution-free, NOT assumption free

- The observations of $X_1, \ldots, X_m$ are a random sample from population 1, independent and identically distributed. The observations of $Y_1, \ldots, Y_m$ are a random sample from population 2, independent and identically distributed
- The X's and Y's are mutually independent
- Populations 1 and 2 are continuous [i.e., no ties]
- **"Robustness of level: The significance level of the rank sum test is not preserved if the two populations differ in dispersion or shape. This is also the case for the normal theory 2-sample t test."** Hollander & Wolfe, p. 120

Slide 22 Wilcoxon rank sum test assumptions

NOTES:

Display 4.10

A summary of the t-statistics calculated from all 10,626 rearrangements of the O-ring data into a "Low" group of size 4 and a "High" group of size 20

Number of re-arrangements with identical t-statistics	t-statistic
2,380	-1.188
3,400	-0.463
2,040	0.231
1530	0.939
855	1.716
316	2.643
95	**3.888**
10	**5.952**

Total number of rearrangements into two groups of size 4 and 20: **10,626**

Number of rearrangements with t-statistics greater than or equal to 3.888: **105**

1-sided p-value from a permutation test of the t-statistic: **105/10626 = .00988**

This permutation test is invalid! The underlying t test assumes equal variances, and that problem is not corrected by using permutations. See Manly

Slide 23

NOTES:

nCr=24 Choose 4=24!/((24-4)!*4!)

=10.626 combinations

Display 4.10

A summary of the t-statistics calculated from all 10,626 rearrangements of the O-ring data into a "Low" group of size 4 and a "High" group of size 20

Number of re-arrangements with identical t-statistics	t-statistic
2,380	-1.188
3,400	-0.463
2,040	0.231
1530	0.939
855	1.716
316	2.643
95	**3.888**
10	**5.952**

Total number of rearrangements into two groups of size 4 and 20: **10,626**

Number of rearrangements with t-statistics greater than or equal to 3.888: **105**

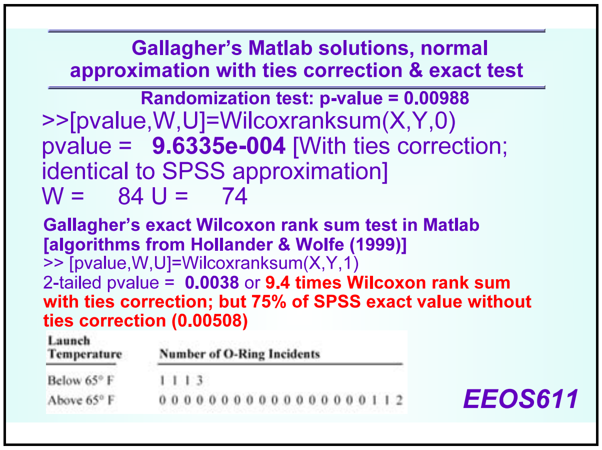
1-sided p-value from a permutation test of the t-statistic: **105/10626 = .00988**

Which test is appropriate?

Independent samples *t* test: exceptionally strong evidence against null ((p=0.00038); Wilcoxon's rank sum test [with ties correction] (0.000963, very strong evidence); Permutation test (strong evidence 0.0098); Unequal variance t test (some evidence, 0.038)

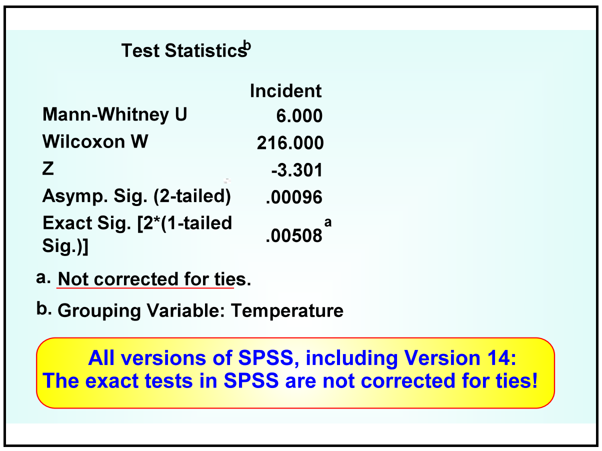
Slide 24 nCr=24 Choose 4=24!/((24-4)!*4!)

NOTES:



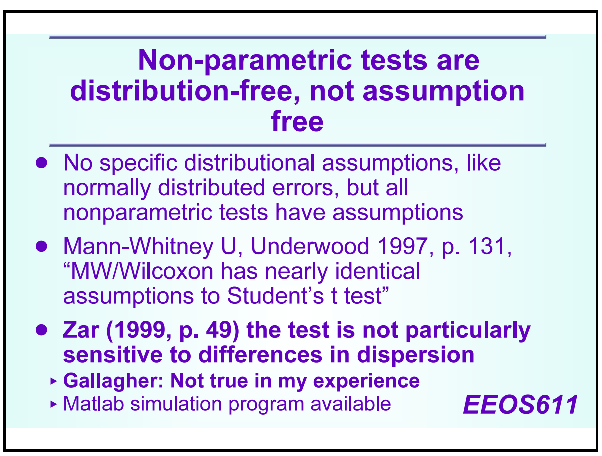
Slide 25 Gallagher's Matlab solutions, normal approximation with ties correction & exact test

NOTES:



Slide 26 SPSS solution with Wilcoxon Rank sum test, not conservative

NOTES:

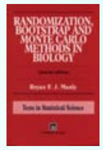


Slide 27 Non-parametric tests are distribution-free, not assumption free

NOTES:

Randomization doesn't solve problems with unequal variance

- Randomization is often superior to the *t*-distribution for 2-sample problems. It does not remedy the common problems with the t distributions though.
- The most common problem with Student's *t* test is the so-called Fisher-Behrens problem: testing for differences in the average if the distributions have different variances
  - This is an open question
  - **Neither** Wilcoxon Rank sum tests **nor** randomization provide a clear solution



EEOS611

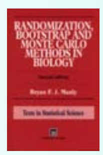
Slide 28 Randomization doesn't solve problems with unequal variance

NOTES:

Neither randomization nor permutation tests solve the unequal variance problem

Manly (1997, p. 141)

- "The randomization test for the difference in two means can be upset if the samples come from sources that have the same mean, but different variances. This is apparent because the null hypothesis for the randomization test is that the samples come from exactly the same source, which is not true if the variances are not constant"
  - A variety of modifications have been proposed, but all require further study.
- O-ring data may not be a test between mean failure rates!



EEOS611

Slide 29 Neither randomization nor permutation tests solve the unequal variance problem

NOTES:

Gallagher's Matlab Case0401b.m

Exact tests based on Student's t test
Why not just use a 2-sample binomial test?

>> X = 1 1 1 3; Y = 0 0 0 0 0 0 0 0 0 0 0 0 00 0 0 0 1 1 2

The observed mean difference in O-ring failures was 1.3 incidents per launch with 95% CI of [0.6 2.0]; The t statistic was 3.9 with 22 df and 2-tailed p of **0.00079**;

Exact tests: The total number of ways of selecting 4 items from 24 items is 10626; The 1-tailed probability of observing a t statistic greater than observed (3.9) is **0.009881**; The exact 1-tailed p value is: 5/506; The 1-tailed probability of observing a difference greater than observed (1.3) is **0.009881;** The 2-tailed probability of observing a t statistic with abs value greater than observed (|3.9|) is **0.009881**...The exact two-tailed probability for the Wilcoxon rank sum test is **0.003764 The normal approximation for the 2-tailed probability for the Wilcoxon rank sum test is 0.000963**;

The probability of an 0 ring incident if cold (<65F) was 4/4=1.00; The probability of an 0 ring incident in warm (>=65F) was 3/20=0.15; **The two sample binomial test for equal proportions (0.29) of failure has a 2-sided p value of 0.000640**

Slide 30 Gallagher's Matlab Case0401b.m

NOTES:

Matlab Case0401b.m

Summary of conclusions

The exact two-tailed probability for the Wilcoxon rank sum test is 0.003764 (Can't use: not robust to unequal variances)

The normal approximation for the 2-tailed probability for the Wilcoxon rank sum test is 0.000963 (but this result should NOT be used — it is a large sample approximation)

Binomial test:

The probability of an 0 ring incident if cold (<65F) was 4/4=1.00

The probability of an 0 ring incident in warm (>=65F) was 3/20=0.15

The two-sample binomial test for equal proportions (p=0.29) of failure has a 2-sided p value of 0.000640

Slide 31 Matlab Case0401b.m

NOTES:

Fligner-Policello test

Wilcoxon-rank sum test for unequal variances
Matlab m.file available, $p<3x10^{-14}$ for O-ring data!

% Trujillo-Ortiz, A., F. A. Trujillo-Rodriguez, R. Hernandez-Walls, M. A. Fligner and S. Perez-Osuna (2003). FPtest: Non-parametric Fligner-Policello test of two combined random variables with continuous cumulative distribution. A MATLAB file.

% [WWW document]. URL http://www.mathworks.com/matlabcentral/fileexchange/

% loadFile.do?objectId=4226&objectType=FILE

% References:

Fligner, M. A. and Policello, G. E. (1981), Robust rank procedure for the Behrens-Fisher Problem. Journal of the American Statistical Association, 76(373): 162-168.
Hollander, M. and Wolfe, D. (1999), Nonparametric Statistical Methods (2nd ed.). New York: John Wiley & Sons, Inc. p. 135-139.

Slide 32 Fligner-Policello test

NOTES:

Asymptotic Power Efficiency

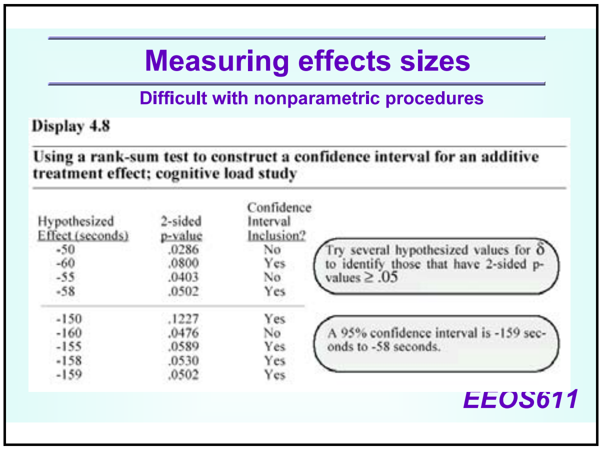
Ratio of sample sizes required to obtain the same p values

- For normally distributed data, the power efficiency of the Wilcoxon rank sum test is 95.5% of the Student's *t* test.
  - For other distributions (e.g., exponential distributions), the power efficiency can be >> 100% (300% for exponential)
    - Hollander & Wolfe p. 140
- Strengths of Wilcoxon's Rank-sum test
  - Resistant to outliers
  - Can handle censored data
- Weakness: generality, determining effect sizes & confidence limits

EEOS611

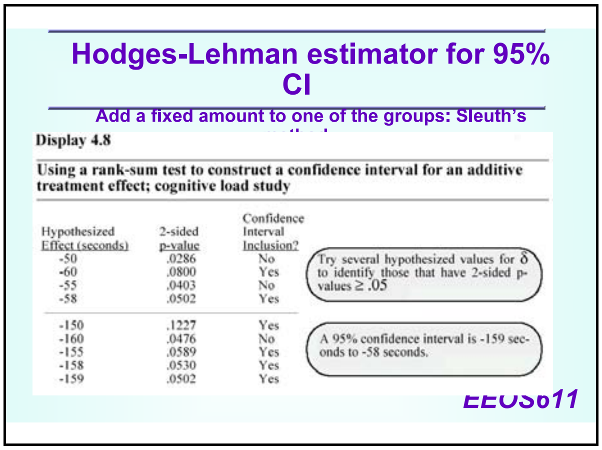
Slide 33 Asymptotic Power Efficiency

NOTES:



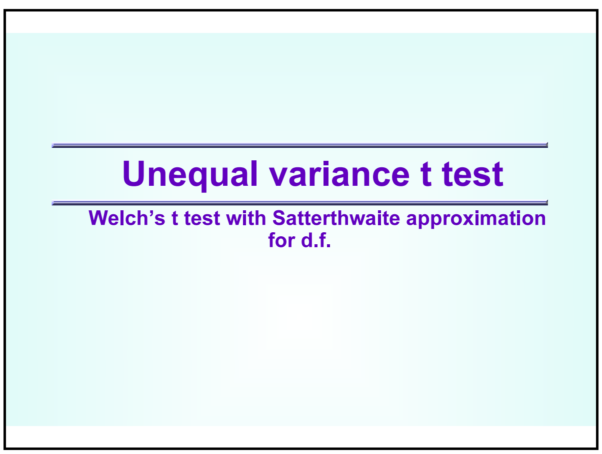
Slide 34 Measuring effects sizes

NOTES:



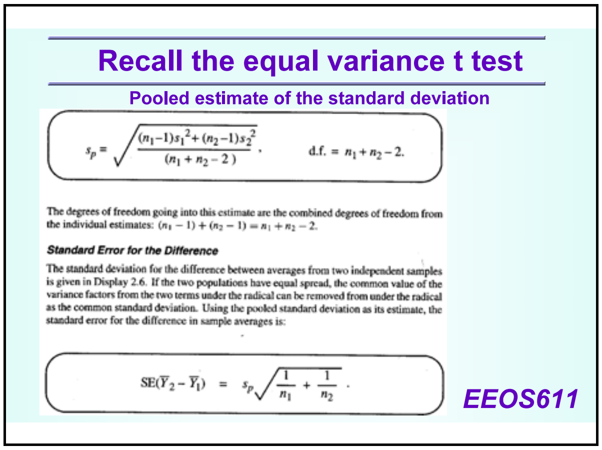
Slide 35 Hodges-Lehman estimator for 95% CI

NOTES:



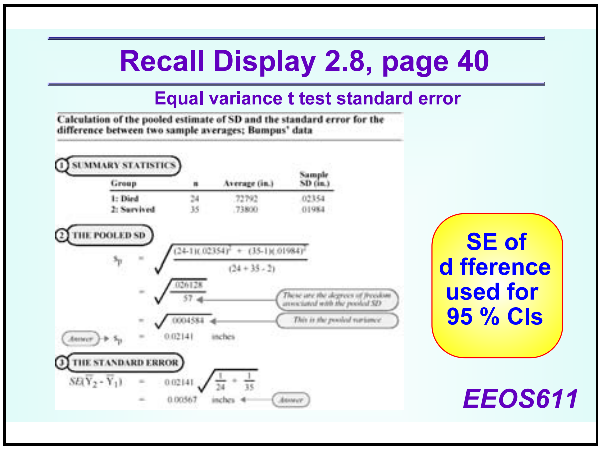
Slide 36 Unequal variance t test

NOTES:



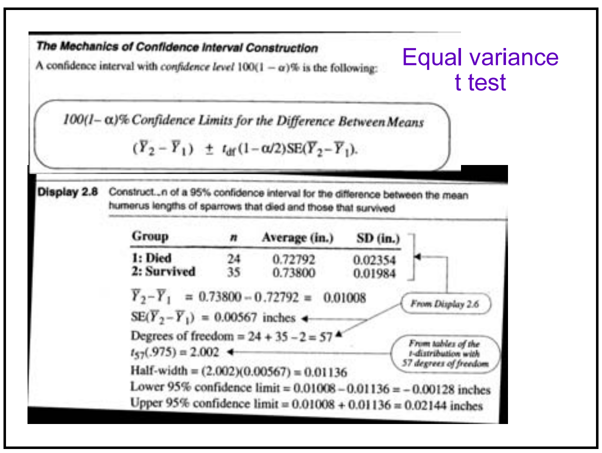
Slide 37 Recall the equal variance t test

NOTES:



Slide 38 Recall Display 2.8, page 40

NOTES:



Slide 39

NOTES:

4.3.2 The Welch *t*-Test for Comparing Two Normal Populations with Unequal Spreads

Welch's t-test employs the individual sample standard deviations as separate estimates of their respective population standard deviations, rather than pooling to obtain a single estimate of a population standard deviation. The result is a different formula for the standard error of the difference in averages:

$$SE_W(\overline{Y}_2 - \overline{Y}_1) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

This becomes the denominator in the *t*-statistic for comparing the means of populations with different spreads. Even when the populations are normal, however, the exact sampling distribution of the Welch *t*-ratio is unknown. It can be approximated by a *t*-distribution with $d.f._W$ degrees of freedom, known as Satterthwaite's approximation:

$$d.f._W = \frac{[SE_W(\overline{Y}_2 - \overline{Y}_1)]^4}{\frac{[SE(\overline{Y}_2)]^4}{(n_2 - 1)} + \frac{[SE(\overline{Y}_1)]^4}{(n_1 - 1)}}.$$

where

$$SE(\overline{Y}_1) = \frac{s_1}{\sqrt{n_1}} \quad \text{and} \quad SE(\overline{Y}_2) = \frac{s_2}{\sqrt{n_2}}.$$

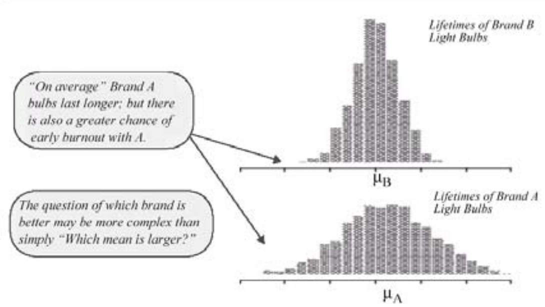
Note the reduction in degrees of freedom. This is the Satterthwaite approximation

Slide 40 Welch's t test, p. 97 in Sleuth

NOTES:

Display 4.11

The conceptual difficulty with comparing population means when population spreads are not the same



Slide 41

NOTES:

Problems with unequal variance *t* tests (Welch *t* test)

- Conceptual difficulties interpreting difference in central tendency if variances unequal
- Exact distribution of test statistic unknown, the Satterthwaite *d.f.* is an approximation, Usually non-integer *d.f.*
- Test isn't necessarily conservative
- Doesn't generalize easily to more than 2 groups
- There are alternate procedures to Welch
  - Variance-stabilizing transformations
  - Nonparametric tests based on ranks: Fligner test
    - Note that this uses the large-sample normal approximation
    - Probably not appropriate for small sample sizes

Slide 42 Problems with unequal variance t tests (Welch t test)

NOTES:

Example: Stream temperatures

Unequal variance t test not necessarily conservative

- Temperatures taken from different portions of a stream:
  - Portion 1: 15.8, 16.9, 17, 17.1, 18, 18.7
    - mean = 17.25, variance = 0.995
  - Portion 2: 18.3, 18.5
    - mean = 18.4, variance = 0.02
- Obviously the variances are unequal and an equal variance 2-sample t test may be inappropriate
  - Welch [unequal variance]: t = 2.74 w/ 5 df p = 0.037.
  - Pooled [equal variance]: t = 1.54 w/ 6 df p = 0.17.
- Why is the equal-variance t-test giving a lower t-value and a higher p value?

Slide 43 Example: Stream temperatures

NOTES:

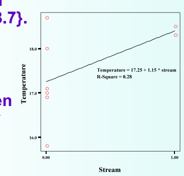
The avg temp different: Exact test

P=3/28, 1-tailed

There are only 28 ways the 8 temperatures can be arranged into groups of 6 and 2 [8 Choose 2], and in only 3 of these arrangements would the difference in means be equal or greater than the 1.15 ºC difference observed. These 3 arrangements include the observed data and two others: {18.3, 18.5}, {18.3, 18.7}, {18.5, 18.7}.

P=3/28~0.107

This is the appropriate p value, unless you argue that the variances are different between the 2 portions of stream, but there are too few data to provide strong evidence for this



Slide 44 The avg temp different: Exact test

NOTES:

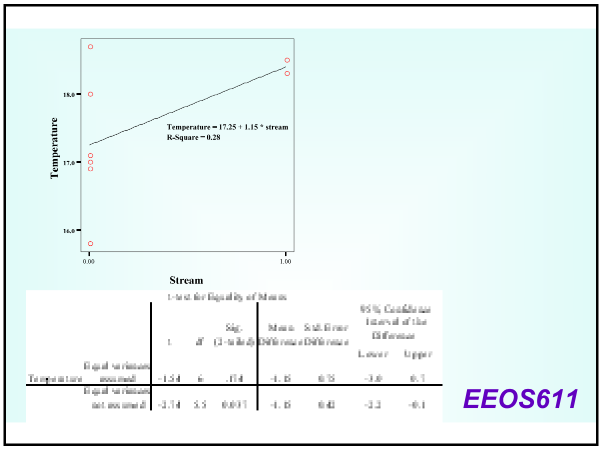
Levene's test, Section 4.5.3

Three different Levene's tests in the literature

- Sleuth's Levene's test on page 102-103 is not the same as the Levene's test used by SPSS in Student's *t* test.
  - Sleuth: squared deviations from the mean used as the variables in a *t* test
  - SPSS: absolute values used
    - SPSS: |observations-mean | used in an F test
    - Other Leven tests |observations-median| used in *t* or F test
- The results can often be quite different
- Levene's tests have largely replaced the F_{max} and Bartlett's tests for equal variance

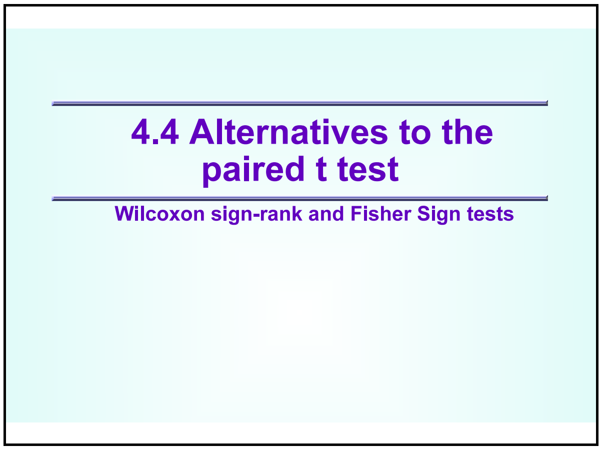
Slide 45 Levene's test, Section 4.5.3

NOTES:



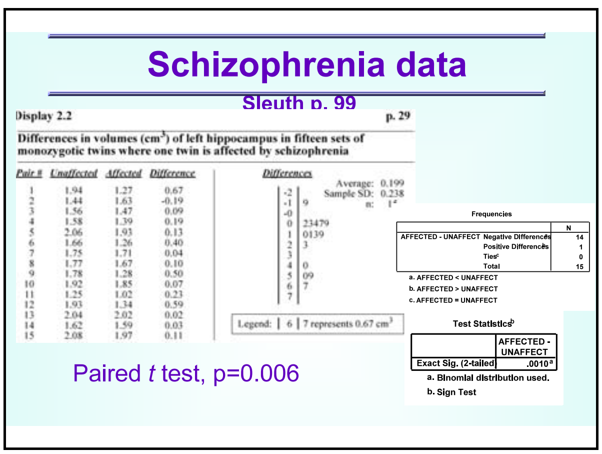
Slide 46

NOTES:



Slide 47 4.4 Alternatives to the paired t test

NOTES:



Display 2.2 p. 29

Differences in volumes (cm^3) of left hippocampus in fifteen sets of monozygotic twins where one twin is affected by schizophrenia

Pair #	Unaffected	Affected	Difference
1	1.94	1.27	0.67
2	1.44	1.63	-0.19
3	1.56	1.47	0.09
4	1.58	1.39	0.19
5	2.06	1.93	0.13
6	1.66	1.26	0.40
7	1.75	1.71	0.04
8	1.77	1.67	0.10
9	1.78	1.28	0.50
10	1.92	1.85	0.07
11	1.25	1.02	0.23
12	1.93	1.34	0.59
13	2.04	2.02	0.02
14	1.62	1.59	0.03
15	2.08	1.97	0.11

Differences

-2	
-1	9
-0	
0	23479
1	0139
2	3
3	
4	0
5	09
6	7
7	

Average: 0.199
Sample SD: 0.238
n: 1[illegible]

Legend: | 6 | 7 represents 0.67 cm^3

Frequencies

		N
AFFECTED - UNAFFECT	Negative Differences[a]	14
	Positive Differences[b]	1
	Ties[c]	0
	Total	15

a. AFFECTED < UNAFFECT
b. AFFECTED > UNAFFECT
c. AFFECTED = UNAFFECT

Test Statistics[b]

	AFFECTED - UNAFFECT
Exact Sig. (2-tailed)	.0010[a]

a. Binomial distribution used.
b. Sign Test

Slide 48 Schizophrenia data

NOTES:

Wilcoxon signed rank test

Analogous to the paired t test

Display 4.12

Signed-rank test statistic computations; schizophrenia study

Pair	Unaffected	Affected	Difference	Ordered Magnitude	Order	Rank	+Ranks	-Ranks
1	1.94	1.27	.67	.02 (+)	1	1	1	
2	1.44	1.63	-.19	.03 (+)	2	2	2	
3	1.56	1.47	.09	.04 (+)	3	3	3	
4	1.58	1.39	.19	.07 (+)	4	4	4	
5	2.06	1.93	.13	.09 (+)	5	5	5	
6	1.66	1.26	.40	.10 (+)	6	6	6	
7	1.75	1.71	.04	.11 (+)	7	7	7	
8	1.77	1.67	.10	.13 (+)	8	8	8	
9	1.78	1.28	.50	.19 (+)	9	9.5	9.5	
10	1.92	1.85	.07	.19 (-)	10	9.5		9.5
11	1.25	1.02	.23	.23 (+)	11	11	11	
12	1.93	1.34	.59	.40 (+)	12	12	12	
13	2.04	2.02	.02	.50 (+)	13	13	13	
14	1.62	1.59	.03	.59 (+)	14	14	14	
15	2.08	1.97	.11	.67 (+)	15	15	15	
							= 110.5	

1 Order the absolute differences and assign ranks to them

2 Signed rank statistics = sum of ranks for positive differences.

SPSS: Discard sample pairs with equal values*

3

Correct the standard deviation SD(T), based on the pattern of ties [2, 5]

Slide 49 Wilcoxon signed rank test

NOTES:

Dealing with tied pairs

Two sorts of ties in the signed rank test

There are two sorts of ties with the signed rank test. If you have identical values in in both pairs, Wilcoxon recommended that those paired observations be dropped from the analysis. That is still the standard recommendation. There is another sort of tie resulting after the absolute values of the differences between paired observations are ranked. You could have two or more differences with the same absolute values. Those ties are not discarded, and the variance formula is adjusted to take into account the number of tied groups [See next slide]

Hollander and Wolfe's Nonparametric statistics, 2nd ed (p. 46) covers the problem of dropping ties of the first sort. If there are many ties, H & W recommend using another test. They also state that you could leave the tied samples in, and use a random number generator to randomly assign positive or negative differences. This apparently is discussed in Pratt (1959). If you want a more conservative 1-sided test, assign all of the tied differences to the group that would make it less likely to reject the null. For example, if you are testing lipitor's effects on cholesterol and a patient had identical cholesterol levels before and after, then assign that difference as if the lipitor blood sample had the higher cholesterol or the placebo value had the lower cholesterol. If you still reject the null, your conservative test would be less likely to result in a Type I error, but of course the probability of Type II error (accepting a false null would be increased). Pratt (1959), cited in both Lehmann and Hollander & Wolfe, provides a more thorough review. Lehmann cites a couple of more recent papers on dealing with the 1st sort of ties in paired rank tests.

Slide 50 Dealing with tied pairs

NOTES:

SPSS algorithms, signed rank test

There are exact tests, if no tied ranks

$$Z = \frac{\min(S_p, S_n) - (n(n+1)/4)}{\sqrt{n(n+1)(2n+1)/24 - \sum_{j=1}^{l}(t_j^3 - t_j)/48}}$$

L=tied groups
t_j = items in each tied group

where

Asymptotic relative efficiency>0.864, 95.5% for normally distributed data

n	Number of cases with non-zero differences
l	Number of ties
t_j	Number of elements in the j-th tie, $j = 1,\ldots,l$

Slide 51 SPSS algorithms, signed rank test

NOTES:

Fisher's sign test

Straightforward application of the 2-sample binomial test

- Given that the probability of a + sign = probability of a minus sign = 0.5,
- What is the probability of observing exactly k positive signs in *n* Bernoulli (binomial) trials
- $P(X=k) = n \text{ Choose } k * p^k (1-p)^{n-k}$
  - X has a binomial distribution
  - Must sum probability for observed value of k, and all more extreme values of k.
- Statistical sleuth provides only the normal approximation to the binomial, but SPSS will provide the exact test for n<30.

Frequencies

		N
AFFECTED - UNAFFECT	Negative Differences[a]	14
	Positive Differences[b]	1
	Ties[c]	0
	Total	15

a. AFFECTED < UNAFFECT
b. AFFECTED > UNAFFECT
c. AFFECTED = UNAFFECT

Test Statistics[b]

	AFFECTED - UNAFFECT
Exact Sig. (2-tailed)	.0010[a]

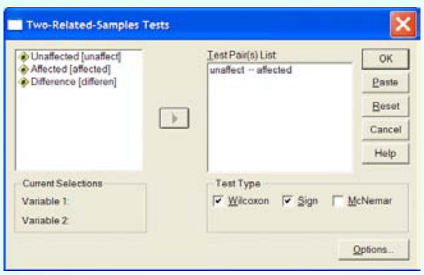
a. Binomial distribution used.
b. Sign Test

EEOS611

Slide 52 Fisher's sign test

NOTES:

Sign test in SPSS



EEOS611

Slide 53 Sign test in SPSS

NOTES:

Conclusions (1 of 2)

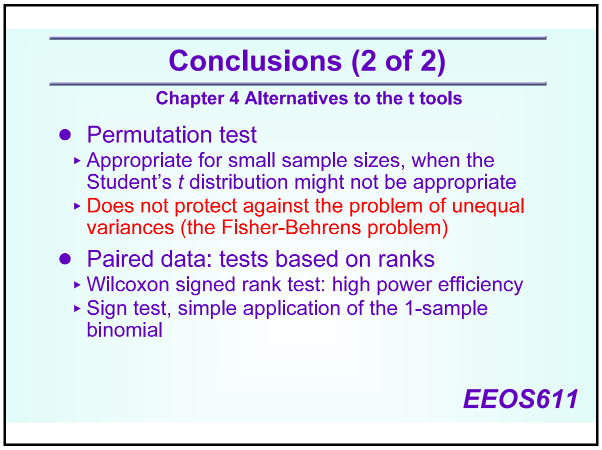
Chapter 4 Alternatives to the t tools

- Consider using alternatives to the t tools if
  - The assumptions are grossly violated or
  - The sample sizes are too small to test distributional assumptions
- Wilcoxon rank sum test
  - Appropriate for small sample sizes
  - Appropriate in the presence of outliers
  - Ties are not a problem if the ties-correction used
  - Not appropriate for samples with unequal variances (try Fligner-Policello if the sample sizes are large)

EEOS611

Slide 54 Conclusions (1 of 2)

NOTES:



Slide 55 Conclusions (2 of 2)

NOTES: