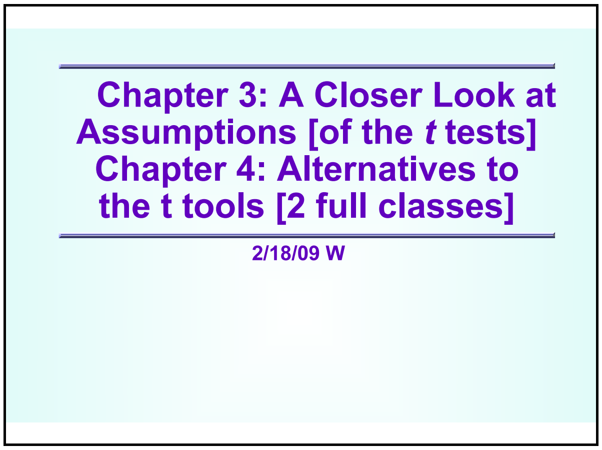


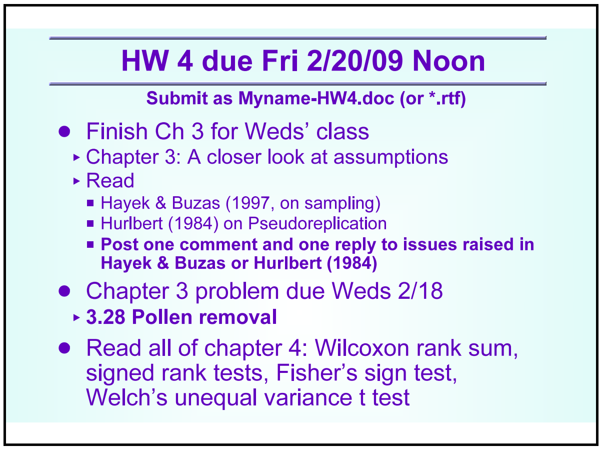
Chapter 3: A Closer Look at Assumptions [of the *t* tests] Chapter 4: Alternatives to the t tools [2 full classes]

2/18/09 W

Slide 1 Chapter 3: A Closer Look at Assumptions [of the t tests]

Chapter 4: Alternatives to the t tools [2 full classes]

NOTES:



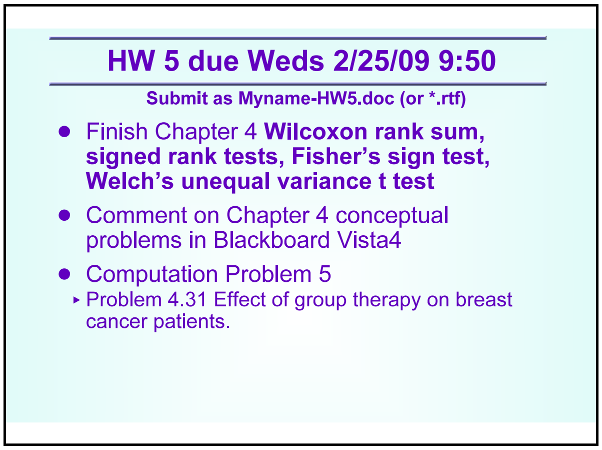
HW 4 due Fri 2/20/09 Noon

Submit as Myname-HW4.doc (or *.rtf)

- Finish Ch 3 for Weds' class
  - Chapter 3: A closer look at assumptions
  - Read
    - Hayek & Buzas (1997, on sampling)
    - Hurlbert (1984) on Pseudoreplication
    - **Post one comment and one reply to issues raised in Hayek & Buzas or Hurlbert (1984)**
- Chapter 3 problem due Weds 2/18
  - **3.28 Pollen removal**
- Read all of chapter 4: Wilcoxon rank sum, signed rank tests, Fisher's sign test, Welch's unequal variance t test

Slide 2 HW 4 due Fri 2/20/09 Noon

NOTES:



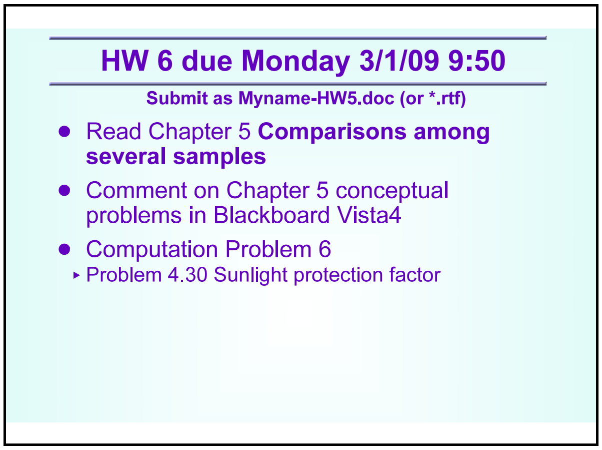
HW 5 due Weds 2/25/09 9:50

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

NOTES:



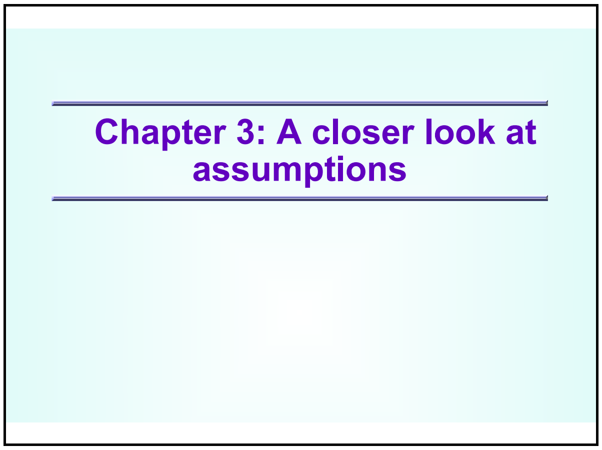
HW 6 due Monday 3/1/09 9:50

Submit as Myname-HW5.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 4 HW 6 due Monday 3/1/09 9:50

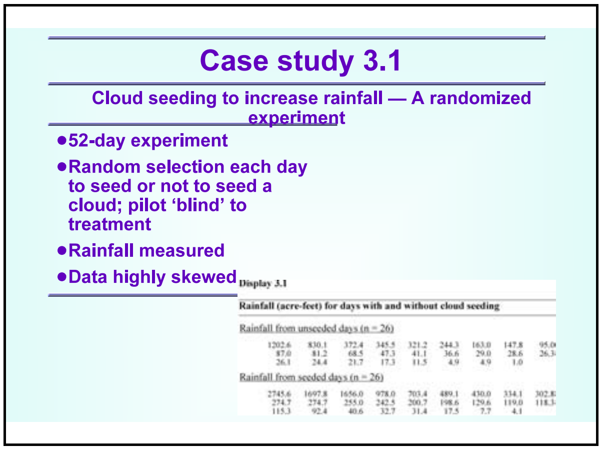
NOTES:



Chapter 3: A closer look at assumptions

Slide 5 Chapter 3: A closer look at assumptions

NOTES:



Case study 3.1

Cloud seeding to increase rainfall — A randomized experiment

- 52-day experiment
- Random selection each day to seed or not to seed a cloud; pilot 'blind' to treatment
- Rainfall measured
- Data highly skewed

Display 3.1

Rainfall (acre-feet) for days with and without cloud seeding

Rainfall from unseeded days (n = 26)

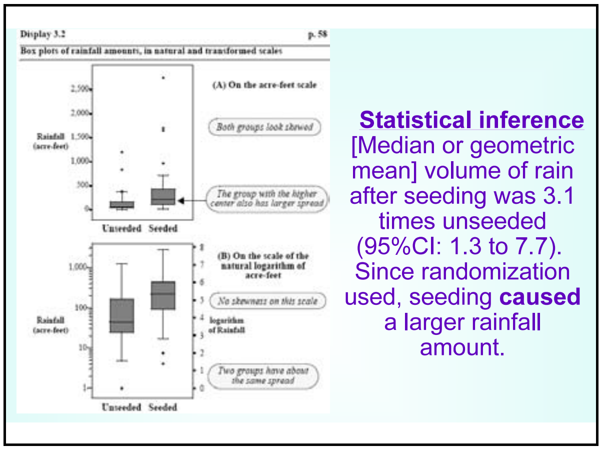
1202.6	830.1	372.4	345.5	321.2	244.3	163.0	147.8	95.0
87.0	81.2	68.5	47.3	41.1	36.6	29.0	28.6	26.3
26.1	24.4	21.7	17.3	11.5	4.9	4.9	1.0	

Rainfall from seeded days (n = 26)

2745.6	1697.8	1656.0	978.0	703.4	489.1	430.0	334.1	302.8
274.7	274.7	255.0	242.5	200.7	198.6	129.6	119.0	118.3
115.3	92.4	40.6	32.7	31.4	17.5	7.7	4.1	

Slide 6 Case study 3.1

NOTES:



Display 3.2 p. 58

Box plots of rainfall amounts, in natural and transformed scales

(A) On the acre-feet scale

Both groups look skewed

The group with the higher center also has larger spread

(B) On the scale of the natural logarithm of acre-feet

No skewness on this scale

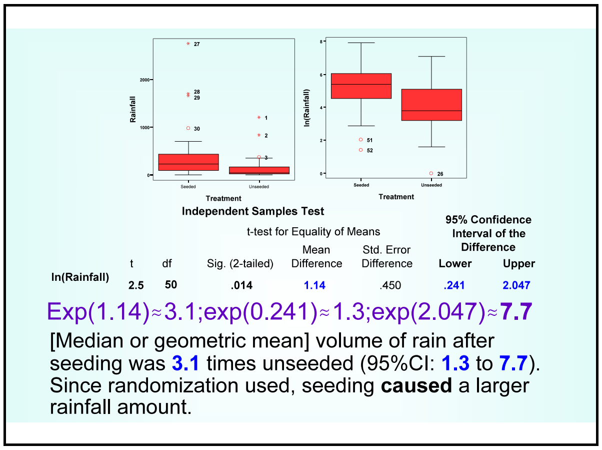
Two groups have about the same spread

Statistical inference

[Median or geometric mean] volume of rain after seeding was 3.1 times unseeded (95%CI: 1.3 to 7.7). Since randomization used, seeding **caused** a larger rainfall amount.

Slide 7

NOTES:



Independent Samples Test

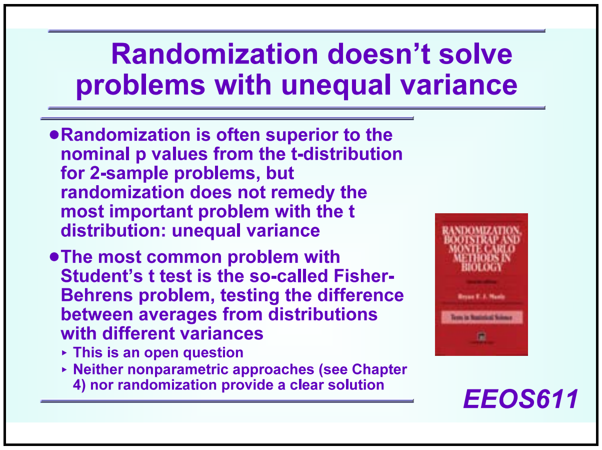
	t-test for Equality of Means					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
ln(Rainfall)	2.5	50	.014	1.14	.450	.241	2.047

Exp(1.14)≈3.1;exp(0.241)≈1.3;exp(2.047)≈**7.7**

[Median or geometric mean] volume of rain after seeding was **3.1** times unseeded (95%CI: **1.3** to **7.7**). Since randomization used, seeding **caused** a larger rainfall amount.

Slide 8

NOTES:


RANDOMIZATION, BOOTSTRAP AND MONTE CARLO METHODS IN BIOLOGY
Bryan F. J. Manly
Texts in Statistical Science

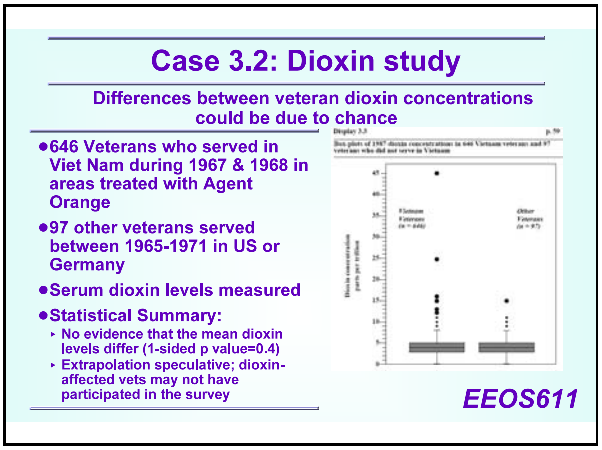
Randomization doesn't solve problems with unequal variance

- **Randomization is often superior to the nominal p values from the t-distribution for 2-sample problems, but randomization does not remedy the most important problem with the t distribution: unequal variance**
- **The most common problem with Student's t test is the so-called Fisher-Behrens problem, testing the difference between averages from distributions with different variances**
    - **This is an open question**
    - **Neither nonparametric approaches (see Chapter 4) nor randomization provide a clear solution**

EEOS611

Slide 9 Randomization doesn't solve problems with unequal variance

NOTES:



Case 3.2: Dioxin study

Differences between veteran dioxin concentrations could be due to chance

- **646 Veterans who served in Viet Nam during 1967 & 1968 in areas treated with Agent Orange**
- **97 other veterans served between 1965-1971 in US or Germany**
- **Serum dioxin levels measured**
- **Statistical Summary:**
  - **No evidence that the mean dioxin levels differ (1-sided p value=0.4)**
  - **Extrapolation speculative; dioxin-affected vets may not have participated in the survey**

Display 3.3 p.59

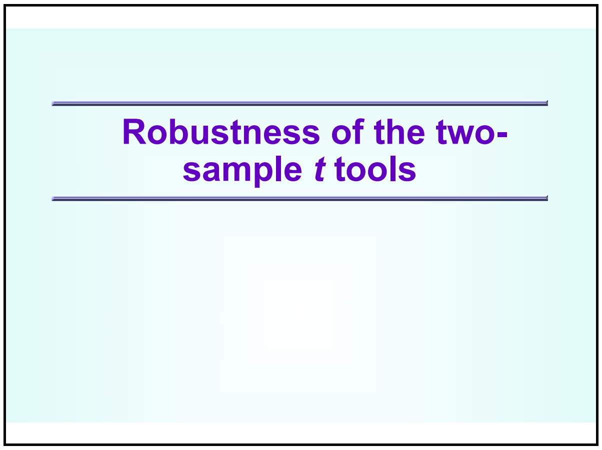
Box plots of 1987 dioxin concentrations in 646 Vietnam veterans and 97 veterans who did not serve in Vietnam

Dioxin concentration parts per trillion
45
40
35
30
25
20
15
10
5
0
Vietnam Veterans (n = 646)
Other Veterans (n = 97)

EEOS611

Slide 10 Case 3.2: Dioxin study

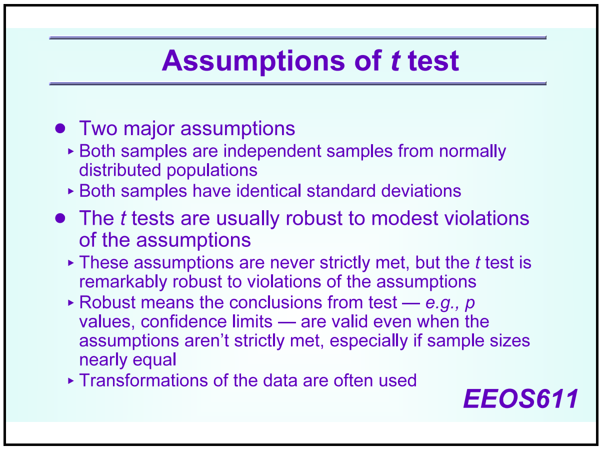
NOTES:



Robustness of the two-sample *t* tools

Slide 11 Robustness of the two-sample t tools

NOTES:



Assumptions of *t* test

- Two major assumptions
  - Both samples are independent samples from normally distributed populations
  - Both samples have identical standard deviations
- The *t* tests are usually robust to modest violations of the assumptions
  - These assumptions are never strictly met, but the *t* test is remarkably robust to violations of the assumptions
  - Robust means the conclusions from test — *e.g., p* values, confidence limits — are valid even when the assumptions aren't strictly met, especially if sample sizes nearly equal
  - Transformations of the data are often used

EEOS611

Slide 12 Assumptions of t test

NOTES:

Violations of assumptions that matter

- With equal sample sizes, the *t*-test is affected moderately by long-tailedness (leptokurtic or peaked distribution) and very little by skewness (the symmetry of the distribution)
  - Kurtosis: peakedness, platykurtic (flat distribution), leptokurtic (peaked)
  - **Skewness: symmetry**
- If the two populations have the same standard deviations and approximately the same shape, with unequal sample size, the *t* tests are affected moderately by long tailedness (leptokurtic) and **substantially by skewness**
- If the **skewness** differs considerably, the **tools can be misleading with small and moderate sample sizes**

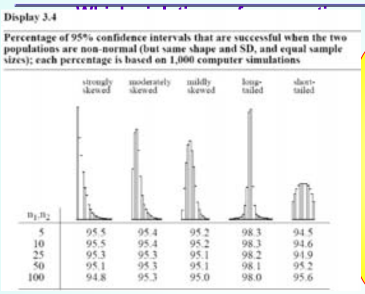
EEOS611

Slide 13 Violations of assumptions that matter

NOTES:

Monte Carlo simulations of violations on p values

really matter?



Display 3.4

Percentage of 95% confidence intervals that are successful when the two populations are non-normal (but same shape and SD, and equal sample sizes); each percentage is based on 1,000 computer simulations

n_1, n_2	strongly skewed	moderately skewed	mildly skewed	long-tailed	short-tailed
5	95.5	95.4	95.2	98.3	94.5
10	95.5	95.4	95.2	98.3	94.6
25	95.3	95.3	95.1	98.2	94.9
50	95.1	95.3	95.1	98.1	95.2
100	94.8	95.3	95.0	98.0	95.6

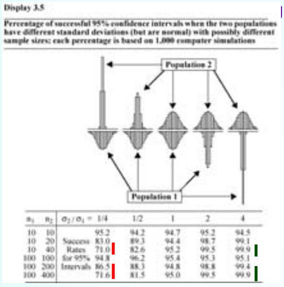
When 2 populations are the same shape with equal n, the results of the *t* test affected moderately (conservative for leptokurtic [peaked] distributions)

Slide 14 Monte Carlo simulations of violations on p values

NOTES:

Different standard deviations & sample sizes

Robust if sample sizes the same, nonconservative if
ger s.d.



Display 3.5

Percentage of successful 95% confidence intervals when the two populations have different standard deviations (but are normal) with possibly different sample sizes; each percentage is based on 1,000 computer simulations

n_1	n_2		$\sigma_2 / \sigma_1 =$ 1/4	1/2	1	2	4
10	10		95.2	94.2	94.7	95.2	94.5
10	20	Success	83.0	89.3	94.4	98.7	99.1
10	40	Rates	71.0	82.6	95.2	99.5	99.9
100	100	for 95%	94.8	96.2	95.4	95.3	95.1
100	200	Intervals	86.5	88.3	94.8	98.8	99.4
100	400		71.6	81.5	95.0	99.5	99.9

P (Type I error) >> stated value (e.g., 0.05) if sd of smaller group larger than larger group

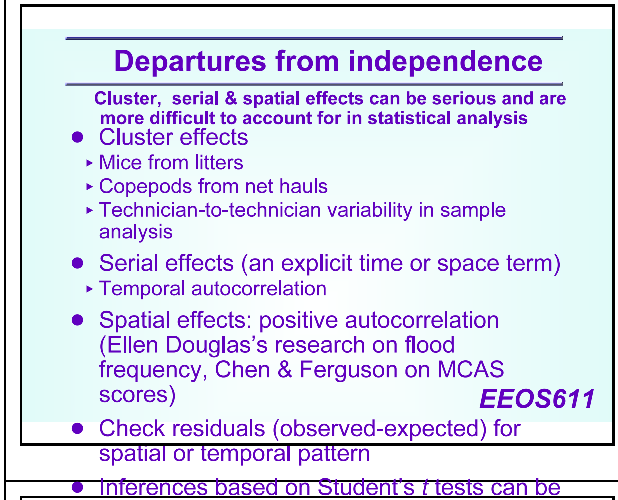
P (Type I error) << stated value (e.g., 0.05) if sd of smaller group smaller than larger group

EEOS611

Slide 15 Different standard deviations & sample sizes

NOTES:

Slide 16 Departures from independence



Departures from independence

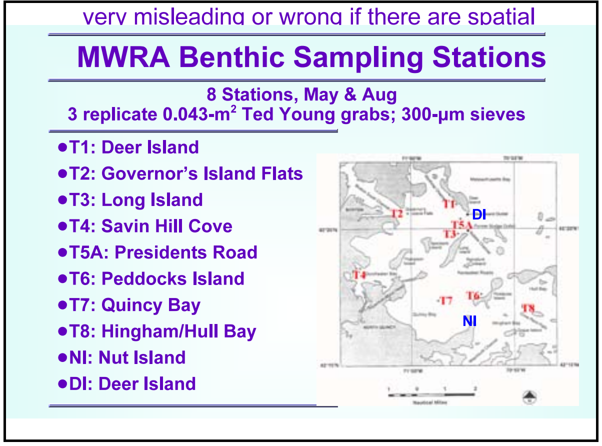
Cluster, serial & spatial effects can be serious and are more difficult to account for in statistical analysis

- Cluster effects
  - Mice from litters
  - Copepods from net hauls
  - Technician-to-technician variability in sample analysis
- Serial effects (an explicit time or space term)
  - Temporal autocorrelation
- Spatial effects: positive autocorrelation (Ellen Douglas's research on flood frequency, Chen & Ferguson on MCAS scores)
- Check residuals (observed-expected) for spatial or temporal pattern
- Inferences based on Student's *t* tests can be very misleading or wrong if there are spatial

EEOS611

NOTES:

Slide 17 MWRA Benthic Sampling Stations



MWRA Benthic Sampling Stations

8 Stations, May & Aug
3 replicate 0.043-m^2 Ted Young grabs; 300-μm sieves

- **T1: Deer Island**
- **T2: Governor's Island Flats**
- **T3: Long Island**
- **T4: Savin Hill Cove**
- **T5A: Presidents Road**
- **T6: Peddocks Island**
- **T7: Quincy Bay**
- **T8: Hingham/Hull Bay**
- **NI: Nut Island**
- **DI: Deer Island**

NOTES:

Slide 18

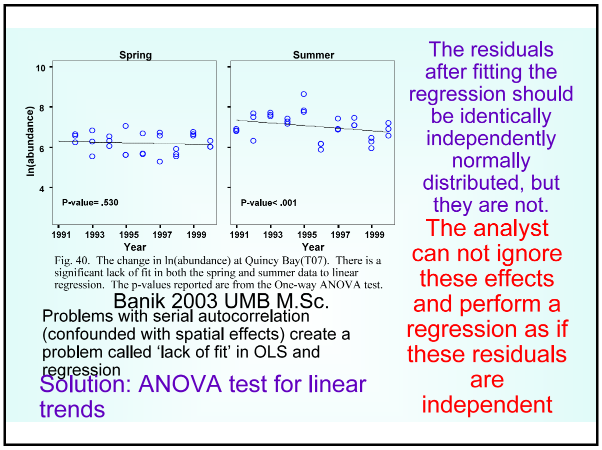


Fig. 40. The change in ln(abundance) at Quincy Bay(T07). There is a significant lack of fit in both the spring and summer data to linear regression. The p-values reported are from the One-way ANOVA test.

Banik 2003 UMB M.Sc.

Problems with serial autocorrelation (confounded with spatial effects) create a problem called 'lack of fit' in OLS and regression

Solution: ANOVA test for linear trends

The residuals after fitting the regression should be identically independently normally distributed, but they are not.

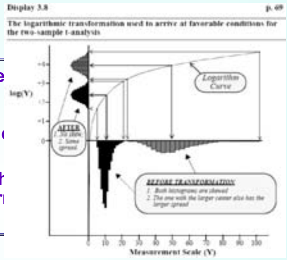
The analyst can not ignore these effects and perform a regression as if these residuals are independent

NOTES:

Log transform of rainfall

See Case 3.1 movie

- The antilogarithm of the me of the log values, the geometric mean, is the median on the original scal of measurement
- Calculate the 95% CI's on th log scale and back transfor they will be asymmetric



Display 3.8 p. 69

The logarithmic transformation used to arrive at favorable conditions for the two-sample t-analysis

EEOS611

Slide 19 Log transform of rainfall

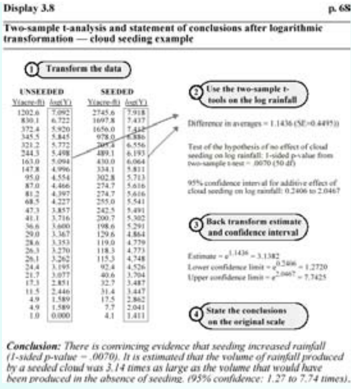
NOTES:

3.5.3 Transformations

- Log (x+1) transform
  - Most biological data, but not usually diversity
  - Needed when there is a multiplicative process in action: growth, bank account interest
  - Marine pollutants: polynuclear aromatic hydrocarbons, fecal coliform bacteria, but not usually metals
  - Calculate the mean and 95% CI and then back-transform. For symmetric data, the mean of the log-transformed data ≈median. Label as the geometric mean
- Many other transforms
  - Arcsin (√Y) for frequency data ranging between 0 and 1 (but the logit transform may be better)
  - % silt clay, but the data must be on the interval 0 to1
  - Logit transform: log [Y/(1-Y)]
  - Square roots for counts, reciprocal for waiting times, logit transforms for proportions between 0 and 1 (log (P/(1-P))
  - "... it is recommended here that a trial-and error approach, with graphical analysis, be used instead."

Slide 20 3.5.3 Transformations

NOTES:



Display 3.8 p. 68

Two-sample t-analysis and statement of conclusions after logarithmic transformation — cloud seeding example

1 Transform the data

UNSEEDED		SEEDED	
Y(acre-ft)	log(Y)	Y(acre-ft)	log(Y)
1202.6	7.092	2745.6	7.918
830.1	6.722	1697.8	7.437
372.4	5.920	1656.0	7.412
345.5	5.845	978.0	6.886
321.2	5.772	703.4	6.556
244.3	5.498	489.1	6.193
163.0	5.094	430.0	6.064
147.8	4.996	334.1	5.811
95.0	4.554	302.8	5.713
87.0	4.466	274.7	5.616
81.2	4.397	274.7	5.616
68.5	4.227	255.0	5.541
47.3	3.857	242.5	5.491
41.1	3.716	200.7	5.302
36.6	3.600	198.6	5.291
29.0	3.367	129.6	4.864
28.6	3.353	119.0	4.779
26.3	3.270	118.3	4.773
26.1	3.262	115.3	4.748
24.4	3.195	92.4	4.526
21.7	3.077	40.6	3.704
17.3	2.851	32.7	3.487
11.5	2.446	31.4	3.447
4.9	1.589	17.5	2.862
4.9	1.589	7.7	2.041
1.0	0.000	4.1	1.411

2 Use the two-sample t-tools on the log rainfall

Difference in averages = 1.1436 (SE=0.4495))

Test of the hypothesis of no effect of cloud seeding on log rainfall: 1-sided p-value from two-sample t-test = .0070 (50 df)

95% confidence interval for additive effect of cloud seeding on log rainfall: 0.2406 to 2.0467

3 Back transform estimate and confidence interval

Estimate = $e^{1.1436} = 3.1382$

Lower confidence limit = $e^{0.2406} = 1.2720$

Upper confidence limit = $e^{2.0467} = 7.7425$

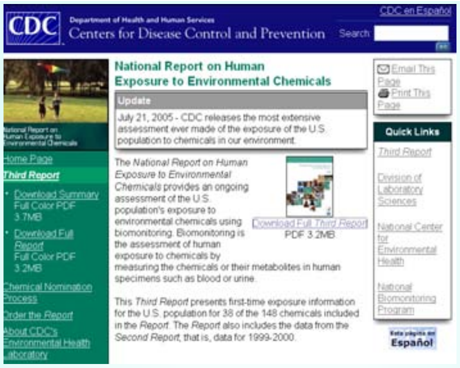
4 State the conclusions on the original scale

Conclusion: *There is convincing evidence that seeding increased rainfall (1-sided p-value = .0070). It is estimated that the volume of rainfall produced by a seeded cloud was 3.14 times as large as the volume that would have been produced in the absence of seeding. (95% confidence: 1.27 to 7.74 times).*

Do the test and calculate the 95% CI on transformed data and then back transform the effect size and confidence limits. Report as ratio of geometric means (Sleuth: ratio of medians).

Slide 21

NOTES:



CDC
Department of Health and Human Services
Centers for Disease Control and Prevention
CDC en Español
Search
National Report on Human Exposure to Environmental Chemicals
Home Page
Third Report
Download Summary Full Color PDF 3.7MB
Download Full Report Full Color PDF 3.2MB
Chemical Nomination Process
Order the Report
About CDC's Environmental Health Laboratory
Email This Page
Print This Page
Quick Links
Third Report
Division of Laboratory Sciences
National Center for Environmental Health
National Biomonitoring Program
Esta página en Español
Download Full Third Report
PDF 3.2MB

National Report on Human Exposure to Environmental Chemicals

Update

July 21, 2005 - CDC releases the most extensive assessment ever made of the exposure of the U.S. population to chemicals in our environment.

The *National Report on Human Exposure to Environmental Chemicals* provides an ongoing assessment of the U.S. population's exposure to environmental chemicals using biomonitoring. Biomonitoring is the assessment of human exposure to chemicals by measuring the chemicals or their metabolites in human specimens such as blood or urine.

This *Third Report* presents first-time exposure information for the U.S. population for 38 of the 148 chemicals included in the *Report*. The *Report* also includes the data from the *Second Report*, that is, data for 1999-2000.

http://www.cdc.gov/exposurereport/

Slide 22

NOTES:

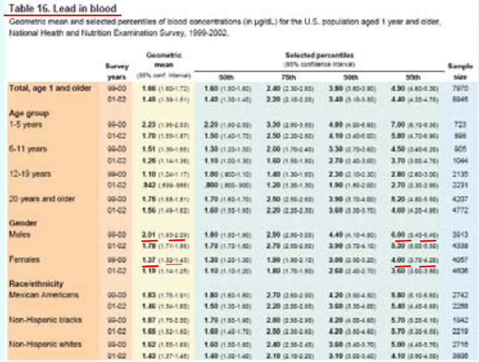


Table 16. Lead in blood

Geometric mean and selected percentiles of blood concentrations (in µg/dL) for the U.S. population aged 1 year and older, National Health and Nutrition Examination Survey, 1999-2002.

	Survey years	Geometric mean (95% conf. interval)	Selected percentiles (95% confidence interval) 50th	75th	90th	95th	Sample size
Total, age 1 and older	99-00	**1.66** (1.60-1.72)	**1.60** (1.50-1.60)	**2.40** (2.30-2.60)	**3.90** (3.60-3.90)	**4.90** (4.60-5.30)	7970
	01-02	**1.45** (1.39-1.51)	**1.40** (1.30-1.40)	**2.20** (2.10-2.20)	**3.40** (3.10-3.60)	**4.40** (4.20-4.70)	8945
Age group							
1-5 years	99-00	**2.23** (1.96-2.53)	**2.20** (1.90-2.50)	**3.30** (2.80-3.80)	**4.80** (4.00-5.60)	**7.00** (6.10-8.30)	723
	01-02	**1.70** (1.55-1.87)	**1.50** (1.40-1.70)	**2.50** (2.20-2.80)	**4.10** (3.40-5.00)	**5.80** (4.70-6.90)	898
6-11 years	99-00	**1.51** (1.36-1.66)	**1.30** (1.20-1.50)	**2.00** (1.70-2.40)	**3.30** (2.70-3.80)	**4.50** (3.40-6.20)	905
	01-02	**1.25** (1.14-1.36)	**1.10** (1.00-1.30)	**1.60** (1.50-1.80)	**2.70** (2.40-3.00)	**3.70** (3.00-4.70)	1044
12-19 years	99-00	**1.10** (1.04-1.17)	**1.00** (.900-1.10)	**1.40** (1.30-1.50)	**2.30** (2.10-2.30)	**2.80** (2.60-3.00)	2135
	01-02	**.942** (.899-.986)	**.800** (.800-.900)	**1.20** (1.20-1.30)	**1.90** (1.80-2.00)	**2.70** (2.30-2.90)	2231
20 years and older	99-00	**1.75** (1.68-1.81)	**1.70** (1.60-1.70)	**2.50** (2.50-2.60)	**3.90** (3.70-4.00)	**5.20** (4.80-5.50)	4207
	01-02	**1.56** (1.49-1.62)	**1.60** (1.50-1.60)	**2.20** (2.20-2.30)	**3.60** (3.30-3.70)	**4.60** (4.20-4.90)	4772
Gender							
Males	99-00	**2.01** (1.93-2.09)	**1.80** (1.80-1.90)	**2.90** (2.80-3.00)	**4.40** (4.10-4.80)	**6.00** (5.40-6.40)	3913
	01-02	**1.78** (1.71-1.86)	**1.70** (1.70-1.80)	**2.70** (2.50-2.80)	**3.90** (3.70-4.10)	**5.30** (5.00-5.50)	4339
Females	99-00	**1.37** (1.32-1.43)	**1.20** (1.20-1.30)	**1.90** (1.80-2.10)	**3.00** (2.90-3.20)	**4.00** (3.70-4.20)	4057
	01-02	**1.19** (1.14-1.25)	**1.10** (1.10-1.20)	**1.80** (1.70-1.80)	**2.60** (2.40-2.70)	**3.60** (3.00-3.90)	4606
Race/ethnicity							
Mexican Americans	99-00	**1.83** (1.75-1.91)	**1.80** (1.60-1.80)	**2.70** (2.60-2.90)	**4.20** (3.90-4.50)	**5.80** (5.10-6.60)	2742
	01-02	**1.46** (1.34-1.59)	**1.50** (1.30-1.60)	**2.20** (2.00-2.50)	**3.60** (3.30-4.00)	**5.40** (4.40-6.60)	2268
Non-Hispanic blacks	99-00	**1.87** (1.75-2.00)	**1.70** (1.60-1.90)	**2.80** (2.50-2.90)	**4.20** (4.00-4.60)	**5.70** (5.20-6.10)	1842
	01-02	**1.65** (1.52-1.80)	**1.60** (1.40-1.70)	**2.50** (2.30-2.80)	**4.20** (3.80-4.60)	**5.70** (5.30-6.00)	2219
Non-Hispanic whites	99-00	**1.62** (1.55-1.69)	**1.60** (1.50-1.60)	**2.40** (2.30-2.40)	**3.60** (3.40-3.70)	**5.00** (4.40-5.70)	2716
	01-02	**1.43** (1.37-1.48)	**1.40** (1.30-1.40)	**2.10** (2.10-2.20)	**3.10** (3.00-3.40)	**4.10** (3.90-4.30)	3806

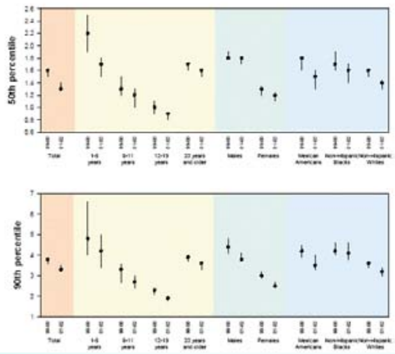
http://www.cdc.gov/exposurereport/3rd/pdf/thirdreport.pdf

Slide 23

NOTES:

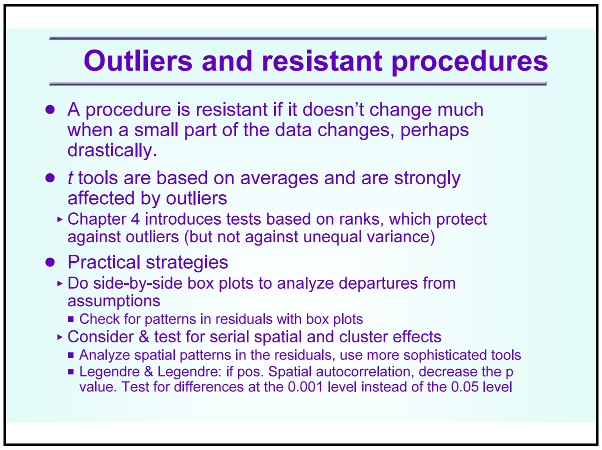
Figure 6. Lead in blood

Selected percentiles with 95% confidence intervals of blood concentrations (in µg/dL) for the U.S. population aged 1 year and older, National Health and Nutrition Examination Survey, 1999-2002.



Slide 24

NOTES:

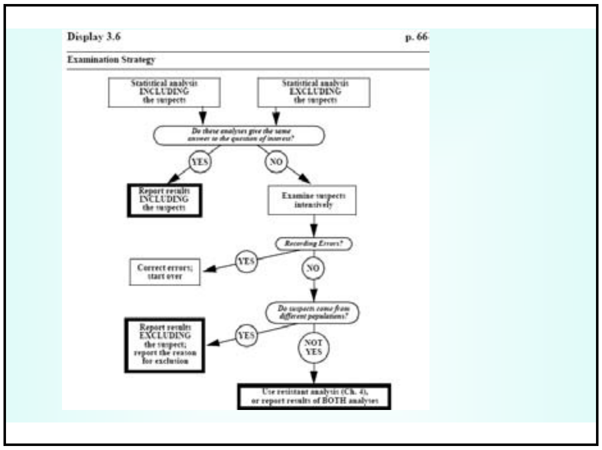


Outliers and resistant procedures

- A procedure is resistant if it doesn't change much when a small part of the data changes, perhaps drastically.
- *t* tools are based on averages and are strongly affected by outliers
  - Chapter 4 introduces tests based on ranks, which protect against outliers (but not against unequal variance)
- Practical strategies
  - Do side-by-side box plots to analyze departures from assumptions
    - Check for patterns in residuals with box plots
  - Consider & test for serial spatial and cluster effects
    - Analyze spatial patterns in the residuals, use more sophisticated tools
    - Legendre & Legendre: if pos. Spatial autocorrelation, decrease the p value. Test for differences at the 0.001 level instead of the 0.05 level

Slide 25 Outliers and resistant procedures

NOTES:



Display 3.6 p. 66

Examination Strategy

Statistical analysis INCLUDING the suspects

Statistical analysis EXCLUDING the suspects

Do these analyses give the same answer to the question of interest?

YES

Report results INCLUDING the suspects

NO

Examine suspects intensively

Recording Errors?

YES

Correct errors; start over

NO

Do suspects come from different populations?

YES

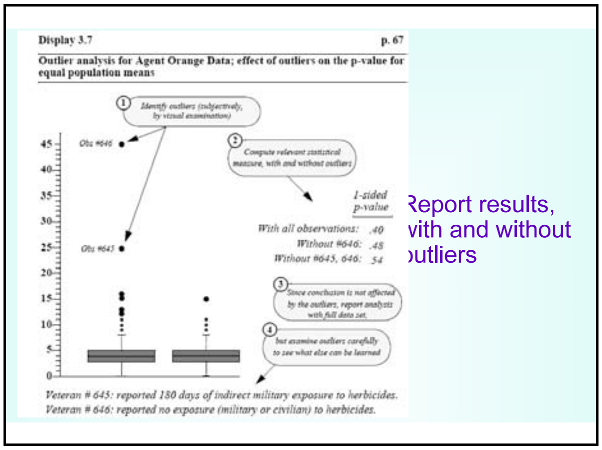
Report results EXCLUDING the suspect; report the reason for exclusion

NOT YES

Use resistant analysis (Ch. 4), or report results of BOTH analyses

Slide 26

NOTES:



Display 3.7 p. 67

Outlier analysis for Agent Orange Data; effect of outliers on the p-value for equal population means

① *Identify outliers (subjectively, by visual examination)*

② *Compute relevant statistical measure, with and without outliers*

	1-sided p-value
With all observations:	*.40*
Without #646:	*.48*
Without #645, 646:	*.54*

③ *Since conclusion is not affected by the outliers, report analysis with full data set,*

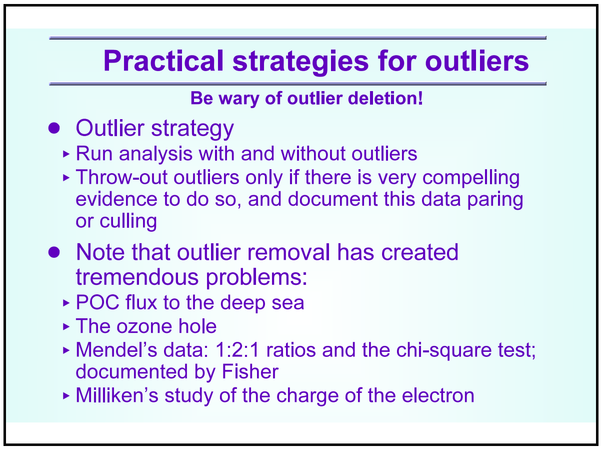
④ *but examine outliers carefully to see what else can be learned*

Veteran # 645: reported 180 days of indirect military exposure to herbicides.
Veteran # 646: reported no exposure (military or civilian) to herbicides.

Report results, with and without outliers

Slide 27 Identifying outliers with boxplots

NOTES:



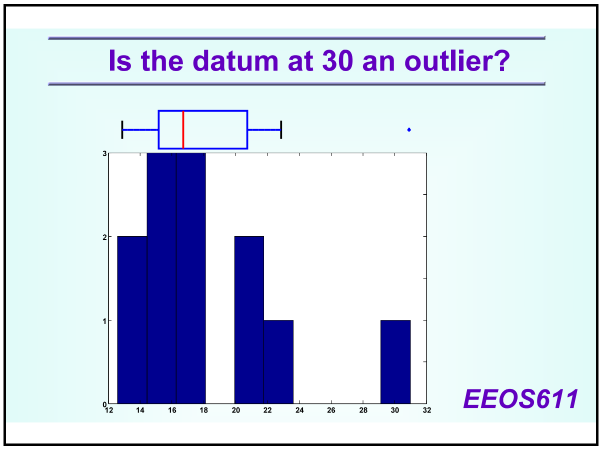
Practical strategies for outliers

Be wary of outlier deletion!

- Outlier strategy
  - Run analysis with and without outliers
  - Throw-out outliers only if there is very compelling evidence to do so, and document this data paring or culling
- Note that outlier removal has created tremendous problems:
  - POC flux to the deep sea
  - The ozone hole
  - Mendel's data: 1:2:1 ratios and the chi-square test; documented by Fisher
  - Milliken's study of the charge of the electron

Slide 28 Practical strategies for outliers

NOTES:

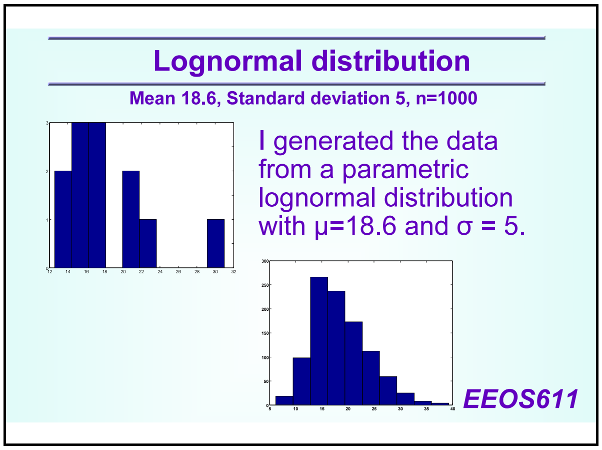


Is the datum at 30 an outlier?

EEOS611

Slide 29 Is the datum at 30 an outlier?

NOTES:



Lognormal distribution

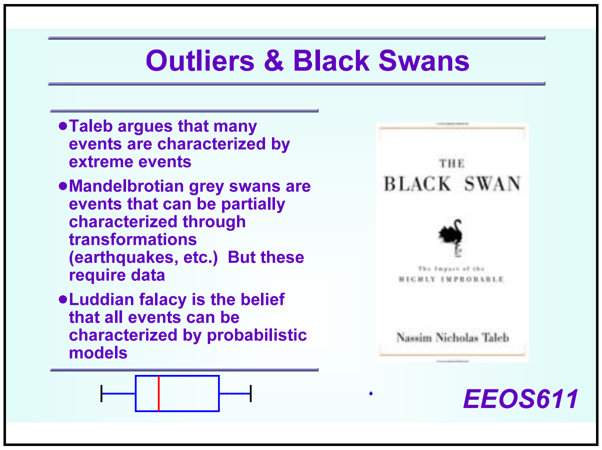
Mean 18.6, Standard deviation 5, n=1000

I generated the data from a parametric lognormal distribution with μ=18.6 and σ = 5.

EEOS611

Slide 30 Lognormal distribution

NOTES:



Outliers & Black Swans

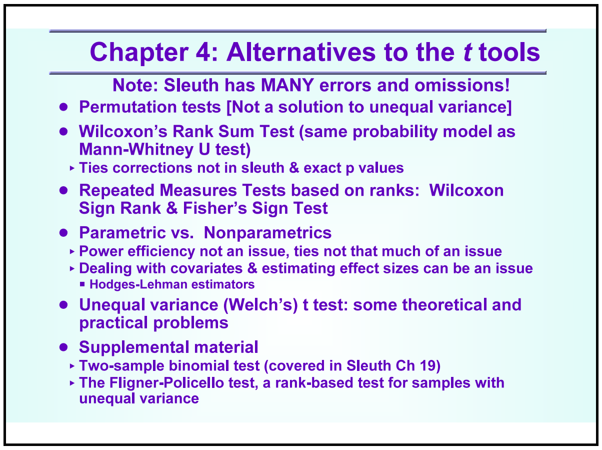
- Taleb argues that many events are characterized by extreme events
- Mandelbrotian grey swans are events that can be partially characterized through transformations (earthquakes, etc.) But these require data
- Luddian falacy is the belief that all events can be characterized by probabilistic models

THE
BLACK SWAN
The Impact of the
HIGHLY IMPROBABLE
Nassim Nicholas Taleb

EEOS611

Slide 31 Outliers & Black Swans

NOTES:



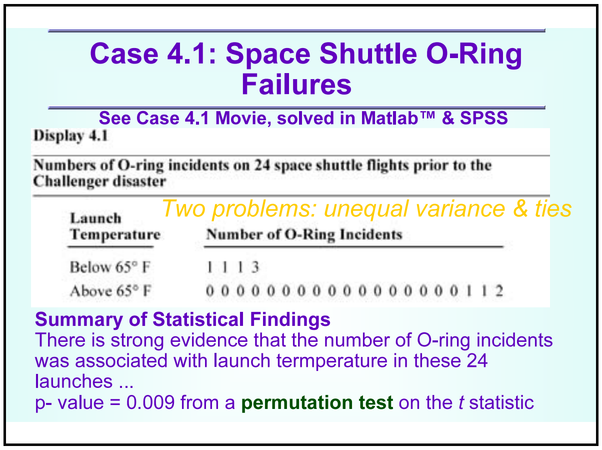
Chapter 4: Alternatives to the *t* tools

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 32 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

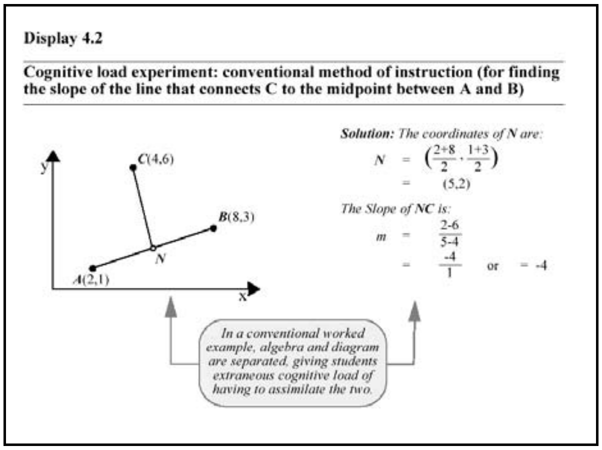
Summary of Statistical Findings

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 33 Case 4.1: Space Shuttle O-Ring Failures

NOTES:



Display 4.2

Cognitive load experiment: conventional method of instruction (for finding the slope of the line that connects C to the midpoint between A and B)

Solution: The coordinates of N are:

$$N = \left(\frac{2+8}{2}, \frac{1+3}{2}\right) = (5,2)$$

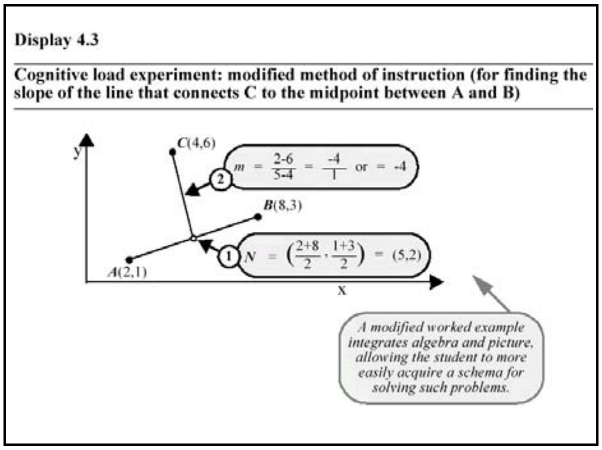
The Slope of NC is:

$$m = \frac{2-6}{5-4} = \frac{-4}{1} \text{ or } = -4$$

In a conventional worked example, algebra and diagram are separated, giving students extraneous cognitive load of having to assimilate the two.

Slide 34 Case 4.2: Cognitive Load

NOTES:



Display 4.3

Cognitive load experiment: modified method of instruction (for finding the slope of the line that connects C to the midpoint between A and B)

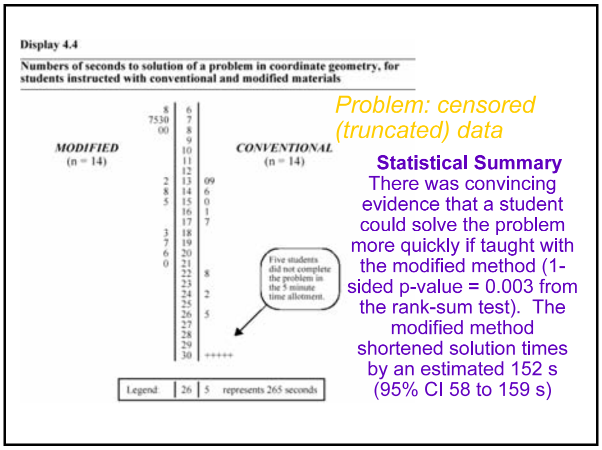
$m = \frac{2-6}{5-4} = \frac{-4}{1}$ or $= -4$

$N = \left(\frac{2+8}{2}, \frac{1+3}{2}\right) = (5,2)$

A modified worked example integrates algebra and picture, allowing the student to more easily acquire a schema for solving such problems.

Slide 35

NOTES:



Display 4.4

Numbers of seconds to solution of a problem in coordinate geometry, for students instructed with conventional and modified materials

MODIFIED (n = 14)		CONVENTIONAL (n = 14)
8	6	
7530	7	
00	8	
	9	
	10	
	11	
	12	
2	13	09
8	14	6
5	15	0
	16	1
	17	7
3	18	
7	19	
6	20	
0	21	
	22	8
	23	
	24	2
	25	
	26	5
	27	
	28	
	29	
	30	+++++

Five students did not complete the problem in the 5 minute time allotment.

Legend | 26 | 5 represents 265 seconds

Problem: censored (truncated) data

Statistical Summary
There was convincing evidence that a student could solve the problem more quickly if taught with the modified method (1-sided p-value = 0.003 from the rank-sum test). The modified method shortened solution times by an estimated 152 s (95% CI 58 to 159 s)

Slide 36

NOTES:

Slide 37 Wilcoxon's rank-sum test

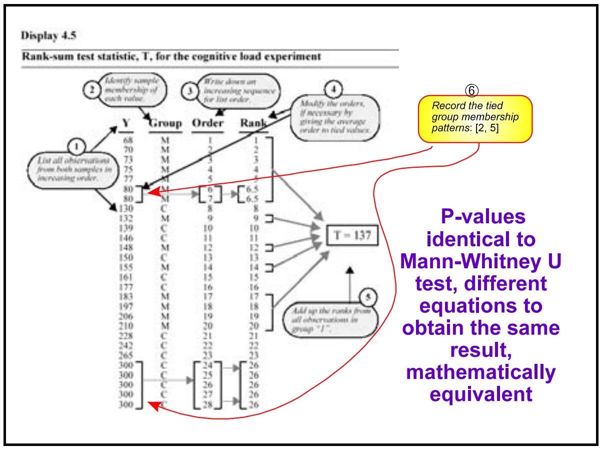
Wilcoxon's rank-sum test

Analogous to 2-sample t test
Same test as the Mann- Whitney U test

Frank Wilcoxon of American Cyanamide [See Salsburg, 2001, The Lady Tasting Tea]

NOTES:

Slide 38 Wilcoxon Rank Sum Statistic



Display 4.5

Rank-sum test statistic, T, for the cognitive load experiment

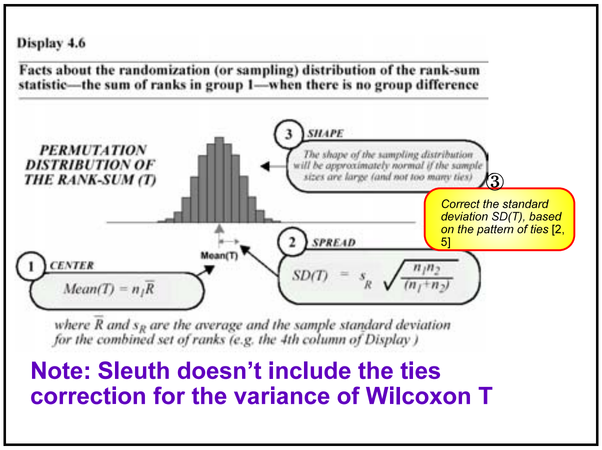
Y	Group	Order	Rank
68	M	1	1
70	M	2	2
73	M	3	3
75	M	4	4
77	M	5	5
80	M	6	6.5
80	M	7	6.5
130	C	8	8
132	M	9	9
139	C	10	10
146	C	11	11
148	M	12	12
150	C	13	13
155	M	14	14
161	C	15	15
177	C	16	16
183	M	17	17
197	M	18	18
206	M	19	19
210	M	20	20
228	C	21	21
242	C	22	22
265	C	23	23
300	C	24	26
300	C	25	26
300	C	26	26
300	C	27	26
300	C	28	26

1 List all observations from both samples in increasing order.
2 Identify sample membership of each value.
3 Write down an increasing sequence for list order.
4 Modify the orders, if necessary by giving the average order to tied values.
5 Add up the ranks from all observations in group "1".
T = 137
⑥ Record the tied group membership patterns: [2, 5]

P-values identical to Mann-Whitney U test, different equations to obtain the same result, mathematically equivalent

NOTES:

Slide 39



Display 4.6

Facts about the randomization (or sampling) distribution of the rank-sum statistic—the sum of ranks in group 1—when there is no group difference

3 SHAPE: The shape of the sampling distribution will be approximately normal if the sample sizes are large (and not too many ties)

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

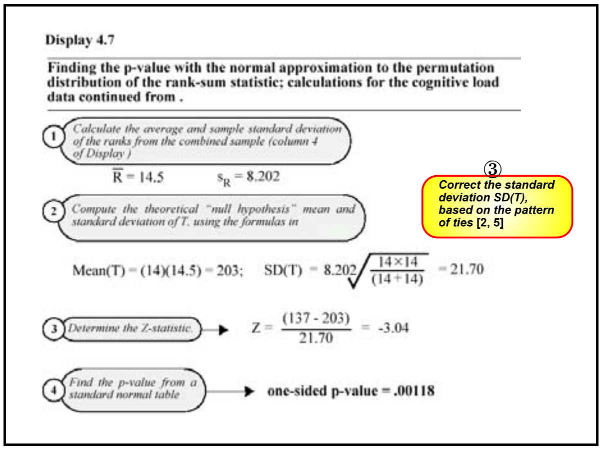
2 SPREAD: $SD(T) = s_R \sqrt{\frac{n_1 n_2}{(n_1+n_2)}}$

1 CENTER: $Mean(T) = n_1\bar{R}$

where $\bar{R}$ and s_R are the average and the sample standard deviation for the combined set of ranks (e.g. the 4th column of Display)

Note: Sleuth doesn't include the ties correction for the variance of Wilcoxon T

NOTES:



Display 4.7

Finding the p-value with the normal approximation to the permutation distribution of the rank-sum statistic; calculations for the cognitive load data continued from .

1 *Calculate the average and sample standard deviation of the ranks from the combined sample (column 4 of Display)*

$\bar{R} = 14.5$ $s_R = 8.202$

2 *Compute the theoretical "null hypothesis" mean and standard deviation of T, using the formulas in*

$$\text{Mean(T)} = (14)(14.5) = 203; \quad \text{SD(T)} = 8.202\sqrt{\frac{14\times14}{(14+14)}} = 21.70$$

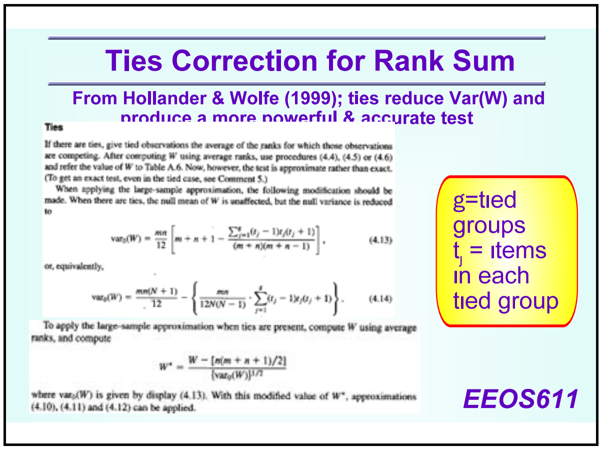
3 *Determine the Z-statistic.* → $Z = \frac{(137 - 203)}{21.70} = -3.04$

4 *Find the p-value from a standard normal table* → **one-sided p-value = .00118**

③ ***Correct the standard deviation SD(T), based on the pattern of ties* [2, 5]**

Slide 40

NOTES:



Ties Correction for Rank Sum

From Hollander & Wolfe (1999); ties reduce Var(W) and produce a more powerful & accurate test

Ties

If there are ties, give tied observations the average of the ranks for which those observations are competing. After computing W using average ranks, use procedures (4.4), (4.5) or (4.6) and refer the value of W to Table A.6. Now, however, the test is approximate rather than exact. (To get an exact test, even in the tied case, see Comment 5.)

When applying the large-sample approximation, the following modification should be made. When there are ties, the null mean of W is unaffected, but the null variance is reduced to

$$\text{var}_0(W) = \frac{mn}{12}\left[m + n + 1 - \frac{\sum_{j=1}^{g}(t_j - 1)t_j(t_j + 1)}{(m + n)(m + n - 1)}\right], \qquad (4.13)$$

or, equivalently,

$$\text{var}_0(W) = \frac{mn(N + 1)}{12} - \left\{\frac{mn}{12N(N - 1)} \cdot \sum_{j=1}^{g}(t_j - 1)t_j(t_j + 1)\right\}. \qquad (4.14)$$

To apply the large-sample approximation when ties are present, compute W using average ranks, and compute

$$W^* = \frac{W - [n(m + n + 1)/2]}{\{\text{var}_0(W)\}^{1/2}}$$

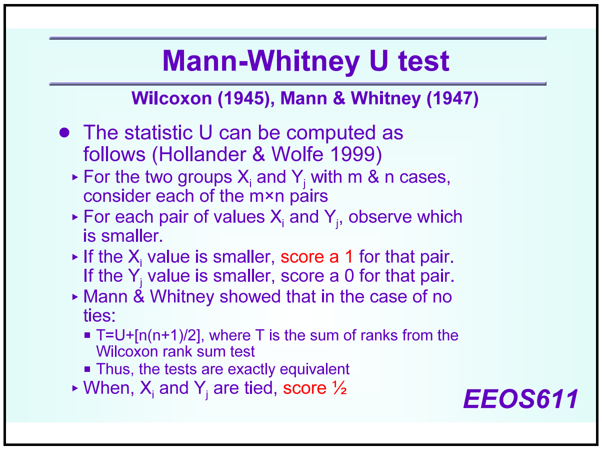
where $\text{var}_0(W)$ is given by display (4.13). With this modified value of W^*, approximations (4.10), (4.11) and (4.12) can be applied.

g=tied groups t_j = items in each tied group

EEOS611

Slide 41 Ties Correction for Rank Sum

NOTES:



Mann-Whitney U test

Wilcoxon (1945), Mann & Whitney (1947)

- The statistic U can be computed as follows (Hollander & Wolfe 1999)
  - For the two groups X_i and Y_j with m & n cases, consider each of the m×n pairs
  - For each pair of values X_i and Y_j, observe which is smaller.
  - If the X_i value is smaller, score a 1 for that pair. If the Y_j value is smaller, score a 0 for that pair.
  - Mann & Whitney showed that in the case of no ties:
    - T=U+[n(n+1)/2], where T is the sum of ranks from the Wilcoxon rank sum test
    - Thus, the tests are exactly equivalent
  - When, X_i and Y_j are tied, score ½

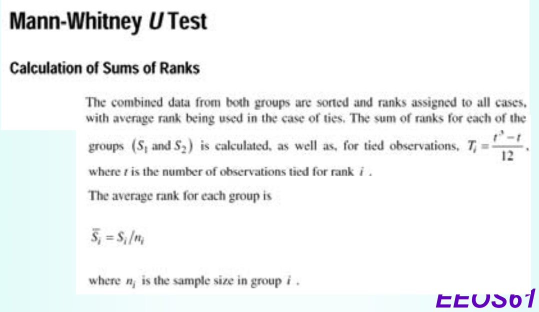
EEOS611

Slide 42 Mann-Whitney U test

NOTES:

SPSS's Mann-Whitney U test

C:\program files\spss\help\algorithms\npar_tests.pdf



Mann-Whitney *U* Test

Calculation of Sums of Ranks

The combined data from both groups are sorted and ranks assigned to all cases, with average rank being used in the case of ties. The sum of ranks for each of the groups $(S_1 \text{ and } S_2)$ is calculated, as well as, for tied observations, $T_i = \frac{t^3 - t}{12}$, where t is the number of observations tied for rank i .

The average rank for each group is

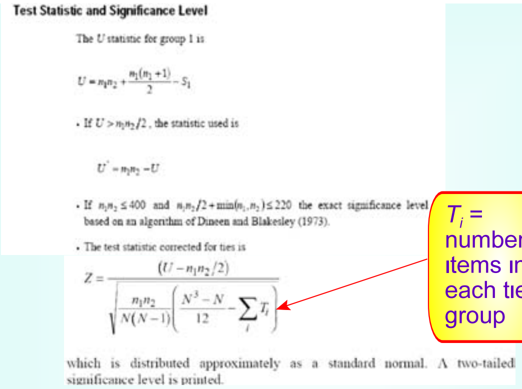
$$\bar{S}_i = S_i / n_i$$

where n_i is the sample size in group i .

EEOS611

Slide 43 SPSS's Mann-Whitney U test

NOTES:



Test Statistic and Significance Level

The U statistic for group 1 is

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - S_1$$

- If $U > n_1 n_2 / 2$, the statistic used is

$$U^{'} = n_1 n_2 - U$$

- If $n_1 n_2 \leq 400$ and $n_1 n_2 / 2 + \min(n_1, n_2) \leq 220$ the exact significance level based on an algorithm of Dineen and Blakesley (1973).
- The test statistic corrected for ties is

$$Z = \frac{(U - n_1 n_2 / 2)}{\sqrt{\frac{n_1 n_2}{N(N-1)} \left(\frac{N^3 - N}{12} - \sum_i T_i \right)}}$$

which is distributed approximately as a standard normal. A two-tailed significance level is printed.

T_i = number of items in each tied group

Slide 44

NOTES:

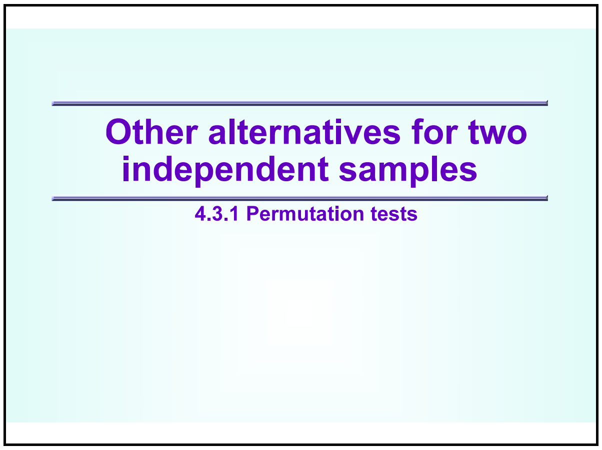
Exact p values tabulated (no ties)

If no ties, exact P values tabulated

- CRC Handbook of Tables for Probability and Statistics
  - Tabulated values of Mann-Whitney U statistic
  - Can readily convert from sum of ranks of smaller group to U statistics
- Or, Hollander & Wolf's (1999) tabulated values of the Wilcoxon T statistic
- Note: the p values for tabulated exact tests are not appropriate if there are any tied ranks; but an exact p value can be calculated using all combinations of data (Gallagher provides a Matlab m.file implementing Hollander & Wolfe algorithm)

Slide 45 Exact p values tabulated (no ties)

NOTES:

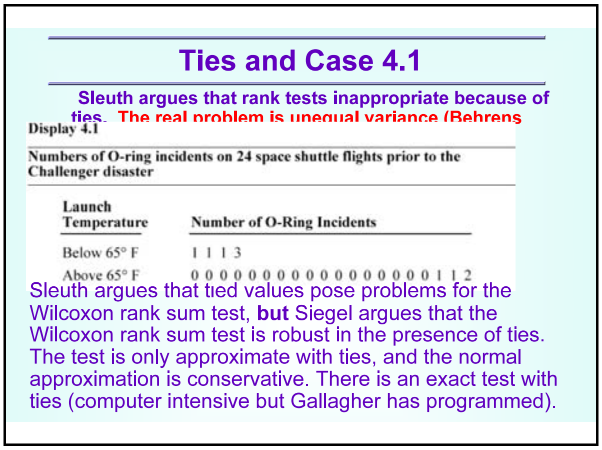


Other alternatives for two independent samples

4.3.1 Permutation tests

Slide 46 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

Display 4.1

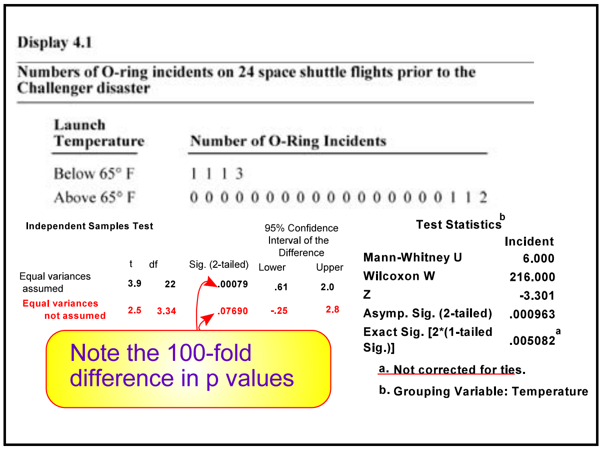
Numbers of O-ring incidents on 24 space shuttle flights prior to the Challenger disaster

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 argues that the Wilcoxon rank sum test is robust in the presence of ties. The test is only approximate with ties, and the normal approximation is conservative. There is an exact test with ties (computer intensive but Gallagher has programmed).

Slide 47 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]

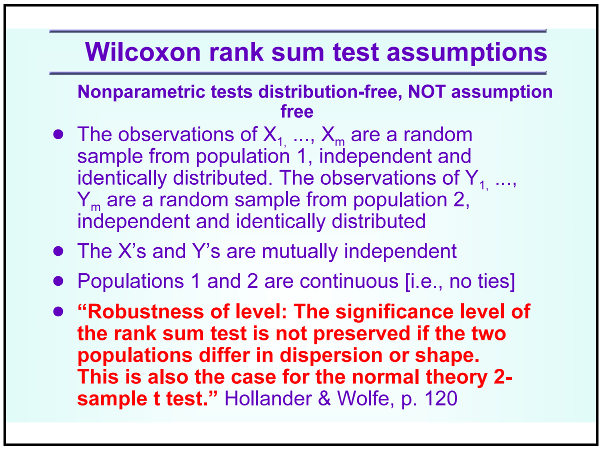
a. Not corrected for ties.

b. Grouping Variable: Temperature

Slide 48

NOTES:

Slide 49 Wilcoxon rank sum test assumptions



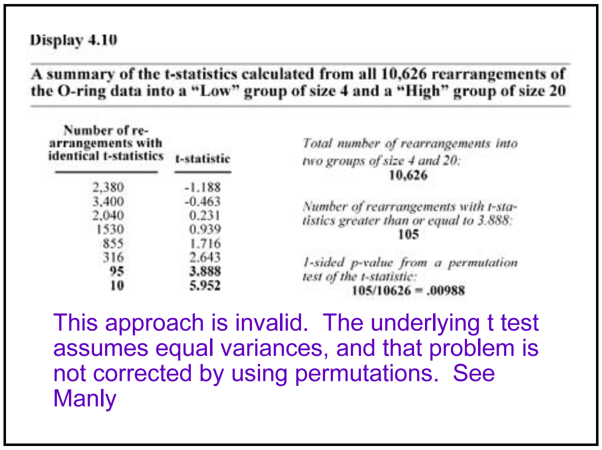
Wilcoxon rank sum test assumptions

Nonparametric tests distribution-free, NOT assumption free

- The observations of $X_{1,}$ ..., X_m are a random sample from population 1, independent and identically distributed. The observations of $Y_{1,}$ ..., 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

NOTES:

Slide 50



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

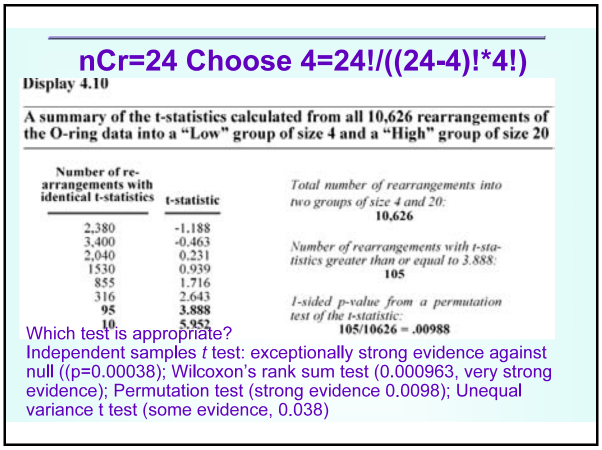
Number of rearrangements with t-statistics greater than or equal to 3.888:
105

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

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

NOTES:

Slide 51 nCr=24 Choose 4=24!/((24-4)!*4!)



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

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**

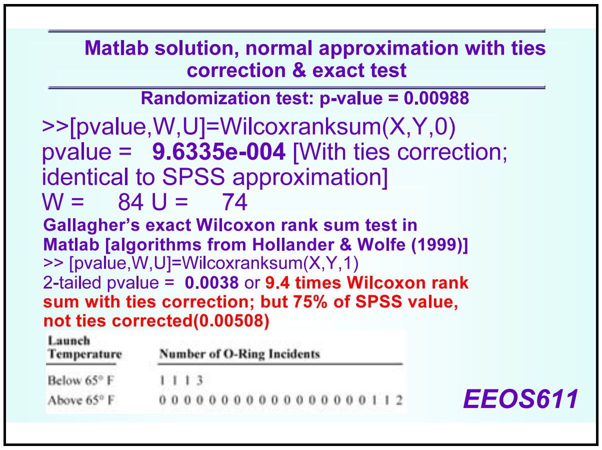
Total number of rearrangements into two groups of size 4 and 20:
10,626

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 (0.000963, very strong evidence); Permutation test (strong evidence 0.0098); Unequal variance t test (some evidence, 0.038)

NOTES:



Matlab solution, normal approximation with ties correction & exact test

Randomization test: p-value = 0.00988

>>[pvalue,W,U]=Wilcoxranksum(X,Y,0)
pvalue = **9.6335e-004** [With ties correction; identical to SPSS approximation]
W = 84 U = 74

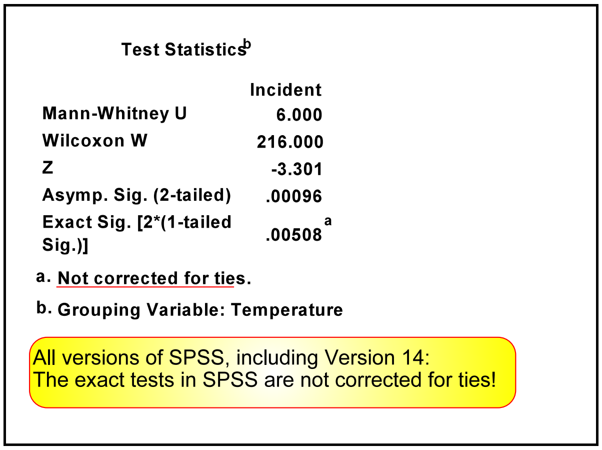
Gallagher's exact Wilcoxon rank sum test in Matlab [algorithms from Hollander & Wolfe (1999)]
>> [pvalue,W,U]=Wilcoxranksum(X,Y,1)
2-tailed pvalue = **0.0038** or **9.4 times Wilcoxon rank sum with ties correction; but 75% of SPSS value, not ties corrected(0.00508)**

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

EEOS611

Slide 52 Matlab solution, normal approximation with ties correction & exact test

NOTES:



Test Statistics[b]

	Incident
Mann-Whitney U	6.000
Wilcoxon W	216.000
Z	-3.301
Asymp. Sig. (2-tailed)	.00096
Exact Sig. [2*(1-tailed Sig.)]	.00508[a]

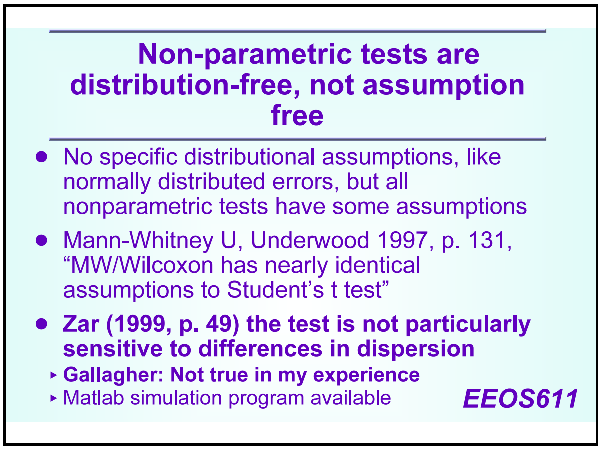
a. Not corrected for ties.

b. Grouping Variable: Temperature

All versions of SPSS, including Version 14:
The exact tests in SPSS are not corrected for ties!

Slide 53 SPSS solution with Wilcoxon Rank sum test, not conservative

NOTES:



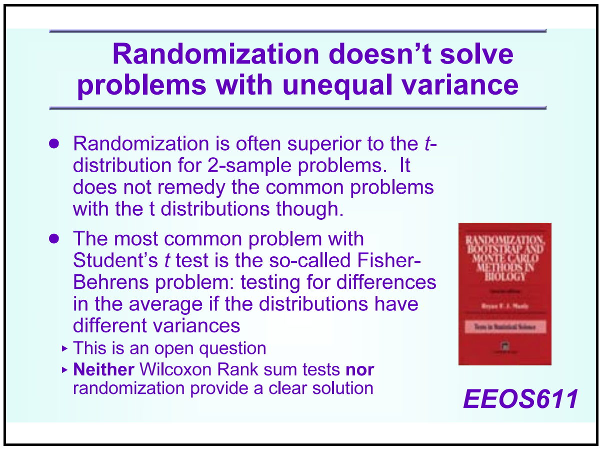
Non-parametric tests are distribution-free, not assumption free

- No specific distributional assumptions, like normally distributed errors, but all nonparametric tests have some assumptions
- Mann-Whitney U, Underwood 1997, p. 131, "MW/Wilcoxon has nearly identical assumptions to Student's t test"
- **Zar (1999, p. 49) the test is not particularly sensitive to differences in dispersion**
  - **Gallagher: Not true in my experience**
  - Matlab simulation program available

EEOS611

Slide 54 Non-parametric tests are distribution-free, not assumption free

NOTES:



Randomization doesn't solve problems with unequal variance

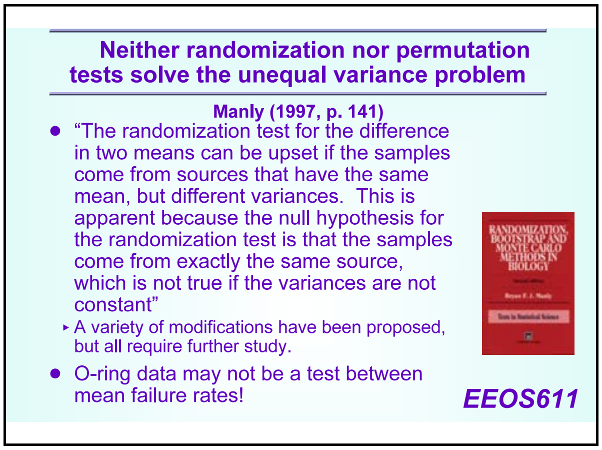
- Randomization is often superior to the *t*-distribution for 2-sample problems. It does not remedy the common problems with the t distributions though.
- The most common problem with Student's *t* test is the so-called Fisher-Behrens problem: testing for differences in the average if the distributions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

RANDOMIZATION, BOOTSTRAP AND MONTE CARLO METHODS IN BIOLOGY
Bryan F. J. Manly
Texts in Statistical Science

EEOS611

Slide 55 Randomization doesn't solve problems with unequal variance

NOTES:



Neither randomization nor permutation tests solve the unequal variance problem

Manly (1997, p. 141)

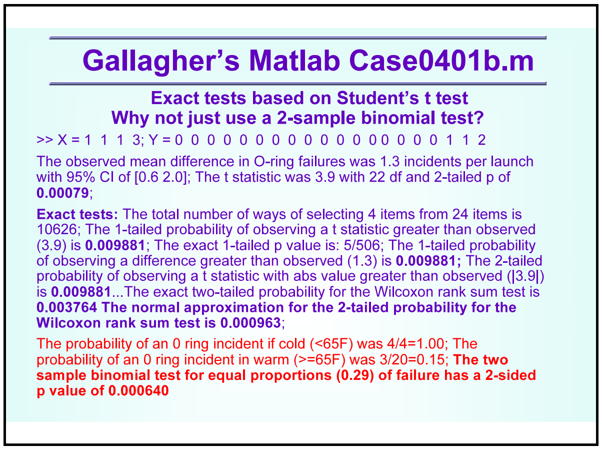
- "The randomization test for the difference in two means can be upset if the samples come from sources that have the same mean, but different variances. This is apparent because the null hypothesis for the randomization test is that the samples come from exactly the same source, which is not true if 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!

RANDOMIZATION, BOOTSTRAP AND MONTE CARLO METHODS IN BIOLOGY
Bryan F. J. Manly
Texts in Statistical Science

EEOS611

Slide 56 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
```

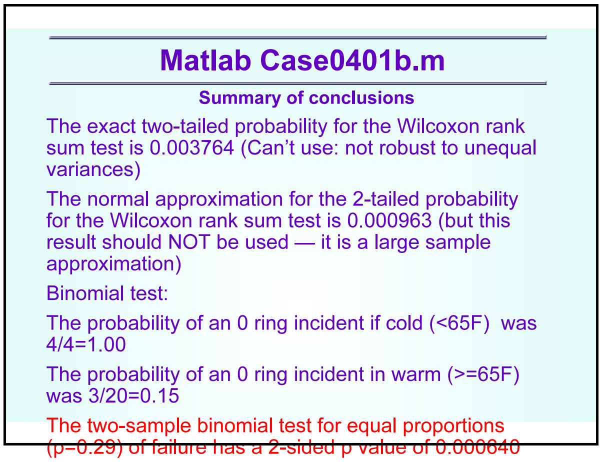
The observed mean difference in O-ring failures was 1.3 incidents per launch with 95% CI of [0.6 2.0]; The t statistic was 3.9 with 22 df and 2-tailed p of **0.00079**;

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 57 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:

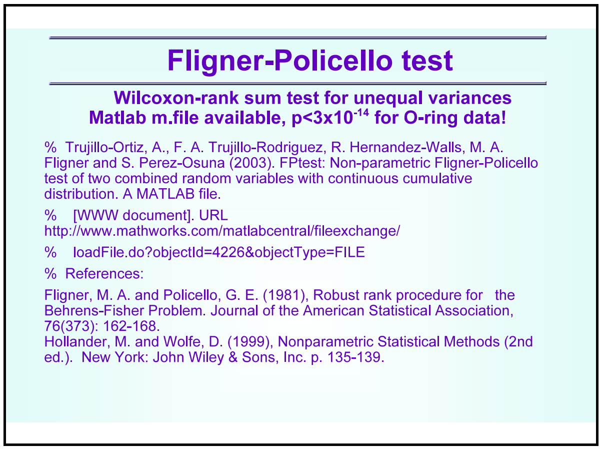
The probability of an 0 ring incident if cold (<65F) was 4/4=1.00

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 58 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/

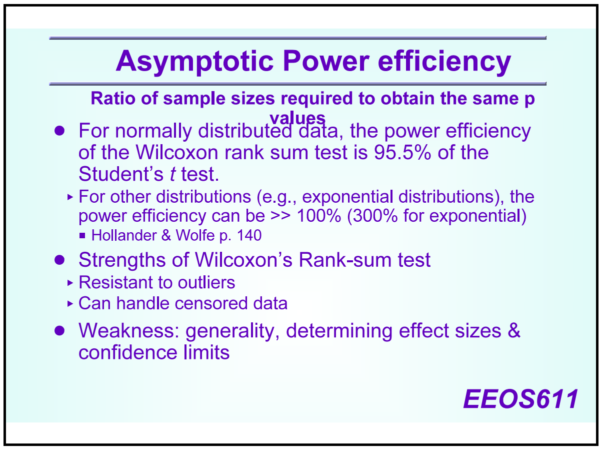
% loadFile.do?objectId=4226&objectType=FILE

% 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 59 Fligner-Policello test

NOTES:



Asymptotic Power efficiency

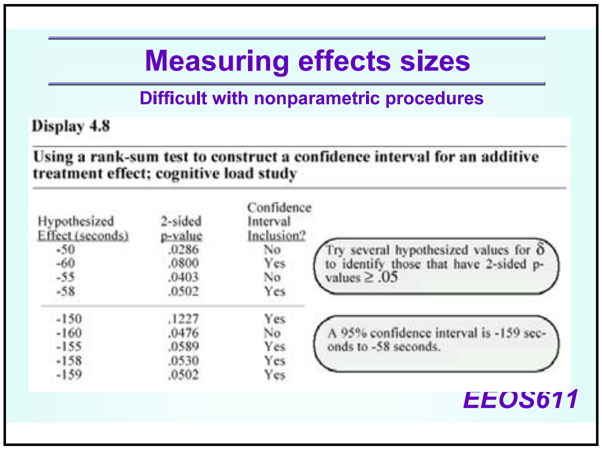
Ratio of sample sizes required to obtain the same p values

- 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 60 Asymptotic Power efficiency

NOTES:



Measuring effects sizes

Difficult with nonparametric procedures

Display 4.8

Using a rank-sum test to construct a confidence interval for an additive treatment effect; cognitive load study

Hypothesized Effect (seconds)	2-sided p-value	Confidence Interval Inclusion?
-50	.0286	No
-60	.0800	Yes
-55	.0403	No
-58	.0502	Yes
-150	.1227	Yes
-160	.0476	No
-155	.0589	Yes
-158	.0530	Yes
-159	.0502	Yes

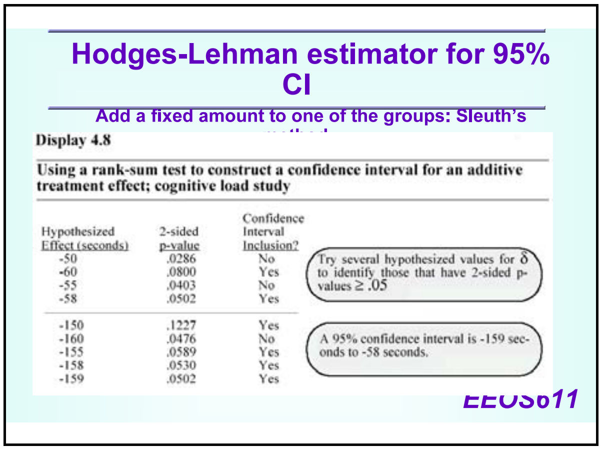
Try several hypothesized values for δ to identify those that have 2-sided p-values ≥ .05

A 95% confidence interval is -159 seconds to -58 seconds.

EEOS611

Slide 61 Measuring effects sizes

NOTES:



Hodges-Lehman estimator for 95% CI

Add a fixed amount to one of the groups: Sleuth’s

Display 4.8

Using a rank-sum test to construct a confidence interval for an additive treatment effect; cognitive load study

Hypothesized Effect (seconds)	2-sided p-value	Confidence Interval Inclusion?
-50	.0286	No
-60	.0800	Yes
-55	.0403	No
-58	.0502	Yes
-150	.1227	Yes
-160	.0476	No
-155	.0589	Yes
-158	.0530	Yes
-159	.0502	Yes

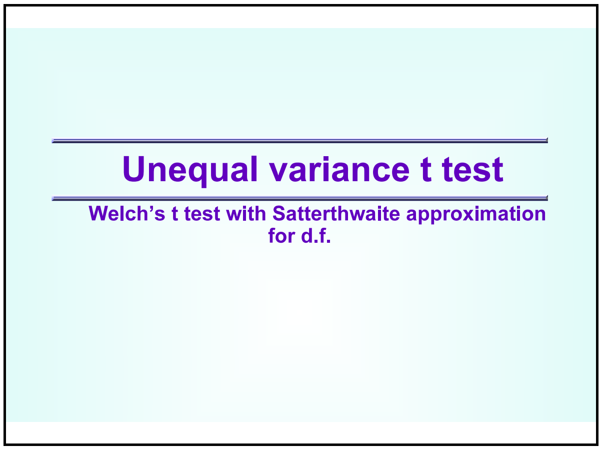
Try several hypothesized values for δ to identify those that have 2-sided p-values ≥ .05

A 95% confidence interval is -159 seconds to -58 seconds.

EEOS611

Slide 62 Hodges-Lehman estimator for 95% CI

NOTES:

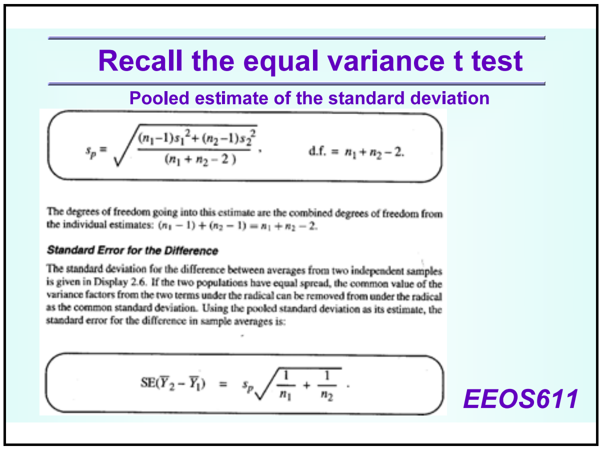


Unequal variance t test

Welch’s t test with Satterthwaite approximation for d.f.

Slide 63 Unequal variance t test

NOTES:



Recall the equal variance t test

Pooled estimate of the standard deviation

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}. \qquad \text{d.f.} = n_1+n_2-2.$$

The degrees of freedom going into this estimate are the combined degrees of freedom from the individual estimates: $(n_1-1)+(n_2-1)=n_1+n_2-2$.

Standard Error for the Difference

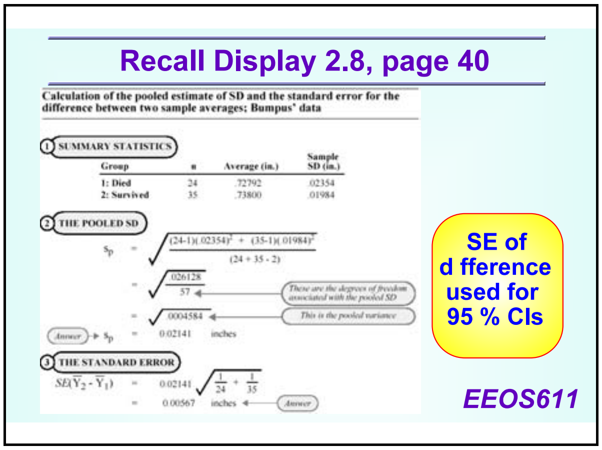
The standard deviation for the difference between averages from two independent samples is given in Display 2.6. If the two populations have equal spread, the common value of the variance factors from the two terms under the radical can be removed from under the radical as the common standard deviation. Using the pooled standard deviation as its estimate, the standard error for the difference in sample averages is:

$$\mathrm{SE}(\overline{Y}_2-\overline{Y}_1) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

EEOS611

Slide 64 Recall the equal variance t test

NOTES:



Recall Display 2.8, page 40

Calculation of the pooled estimate of SD and the standard error for the difference between two sample averages; Bumpus' data

1 SUMMARY STATISTICS

Group	n	Average (in.)	Sample SD (in.)
1: Died	24	.72792	.02354
2: Survived	35	.73800	.01984

2 THE POOLED SD

$$s_p = \sqrt{\frac{(24-1)(.02354)^2 + (35-1)(.01984)^2}{(24+35-2)}}$$

$$= \sqrt{\frac{.026128}{57}}$$

These are the degrees of freedom associated with the pooled SD

$$= \sqrt{.0004584}$$

This is the pooled variance

Answer → s_p = 0.02141 inches

3 THE STANDARD ERROR

$$SE(\overline{Y}_2-\overline{Y}_1) = 0.02141\sqrt{\frac{1}{24}+\frac{1}{35}}$$

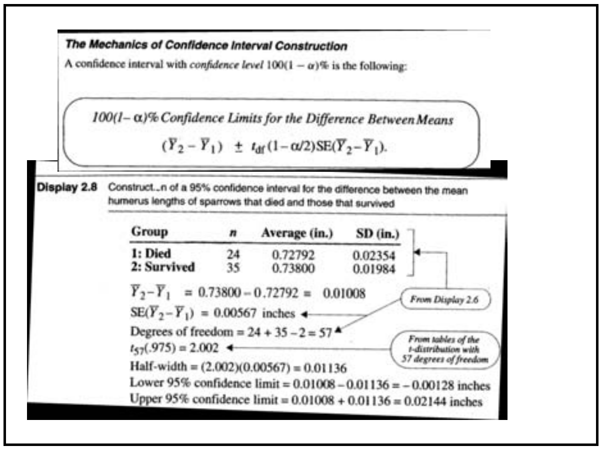
= 0.00567 inches ← Answer

SE of d fference used for 95 % CIs

EEOS611

Slide 65 Recall Display 2.8, page 40

NOTES:



The Mechanics of Confidence Interval Construction

A confidence interval with *confidence level* $100(1-\alpha)\%$ is the following:

100(1– α)% Confidence Limits for the Difference Between Means

$$(\overline{Y}_2-\overline{Y}_1) \pm t_{df}(1-\alpha/2)\mathrm{SE}(\overline{Y}_2-\overline{Y}_1).$$

Display 2.8 Construct..n of a 95% confidence interval for the difference between the mean humerus lengths of sparrows that died and those that survived

Group	*n*	Average (in.)	SD (in.)
1: Died	24	0.72792	0.02354
2: Survived	35	0.73800	0.01984

$\overline{Y}_2-\overline{Y}_1$ = 0.73800 – 0.72792 = 0.01008

$\mathrm{SE}(\overline{Y}_2-\overline{Y}_1)$ = 0.00567 inches

From Display 2.6

Degrees of freedom = 24 + 35 – 2 = 57

$t_{57}(.975) = 2.002$

From tables of the t-distribution with 57 degrees of freedom

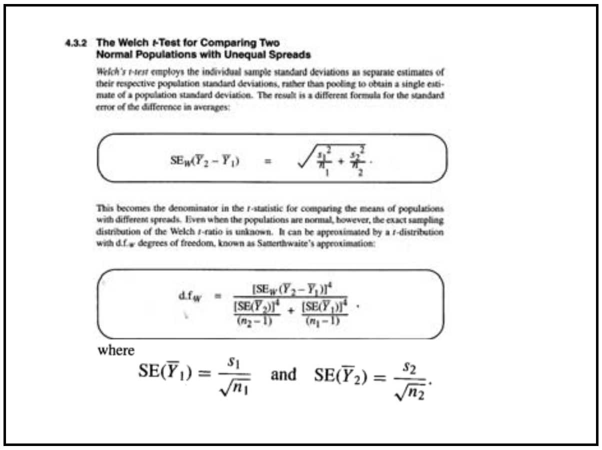
Half-width = (2.002)(0.00567) = 0.01136

Lower 95% confidence limit = 0.01008 – 0.01136 = –0.00128 inches

Upper 95% confidence limit = 0.01008 + 0.01136 = 0.02144 inches

Slide 66

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:

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

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:

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

where

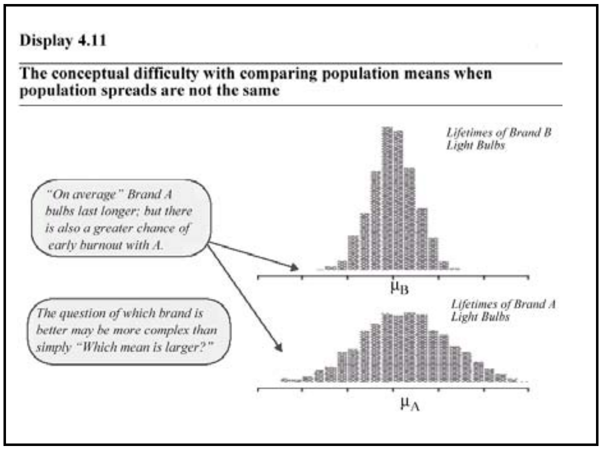
$$\mathrm{SE}(\overline{Y}_1) = \frac{s_1}{\sqrt{n_1}} \quad \text{and} \quad \mathrm{SE}(\overline{Y}_2) = \frac{s_2}{\sqrt{n_2}}.$$

Slide 67 Welch's t test, p. 97 in Sleuth

NOTES:

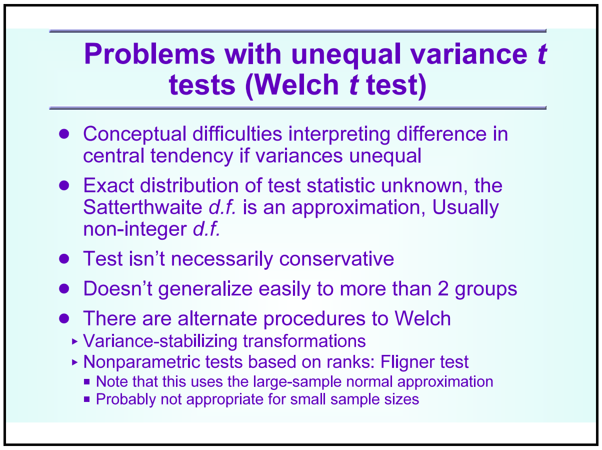
Display 4.11

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



Slide 68

NOTES:

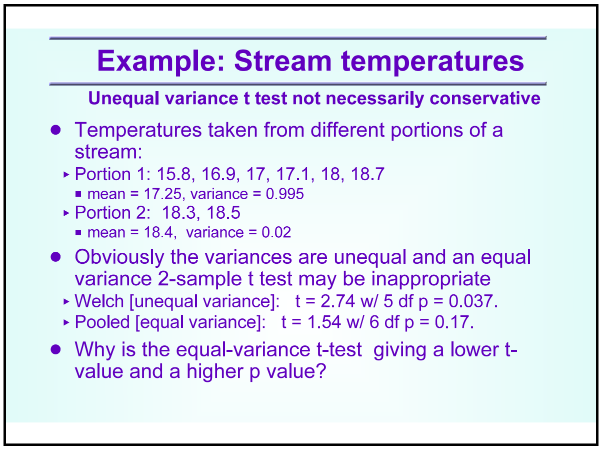


Problems with unequal variance *t* tests (Welch *t* test)

- 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 69 Problems with unequal variance t tests (Welch t test)

NOTES:



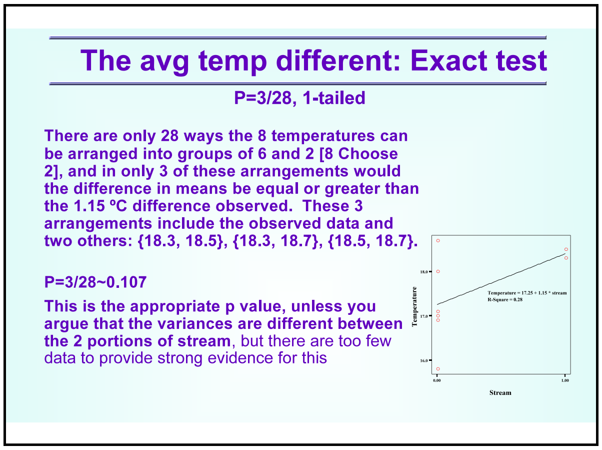
Example: Stream temperatures

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 70 Example: Stream temperatures

NOTES:



The avg temp different: Exact test

P=3/28, 1-tailed

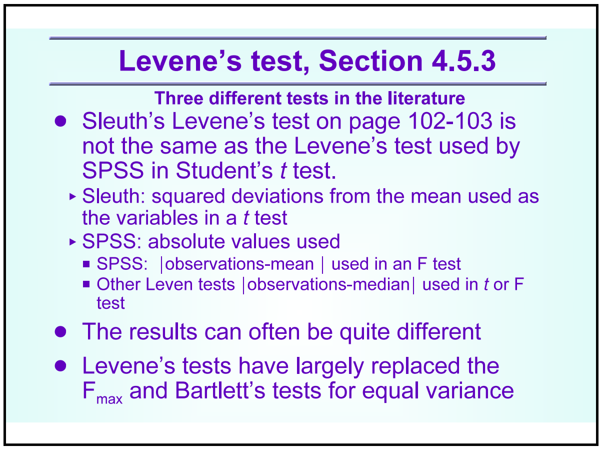
There are only 28 ways the 8 temperatures can be arranged into groups of 6 and 2 [8 Choose 2], and in only 3 of these arrangements would the difference in means be equal or greater than the 1.15 ºC difference observed. These 3 arrangements include the observed data and two others: {18.3, 18.5}, {18.3, 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 71 The avg temp different: Exact test

NOTES:



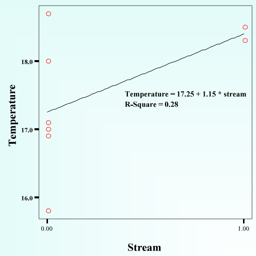
Levene’s test, Section 4.5.3

Three different 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 72 Levene’s test, Section 4.5.3

NOTES:



		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
Temperature	Equal variances assumed	-1.54	6	.174	-1.15	0.75	-3.0	0.7
	Equal variances not assumed	-2.74	5.5	0.037	-1.15	0.42	-2.2	-0.1

EEOS611

Slide 73

NOTES:

4.4 Alternatives to the paired t test

Wilcoxon sign-rank and Fisher Sign tests

Slide 74 4.4 Alternatives to the paired t test

NOTES:

Schizophrenia data

Sleuth p. 99

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: 1⁴

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]

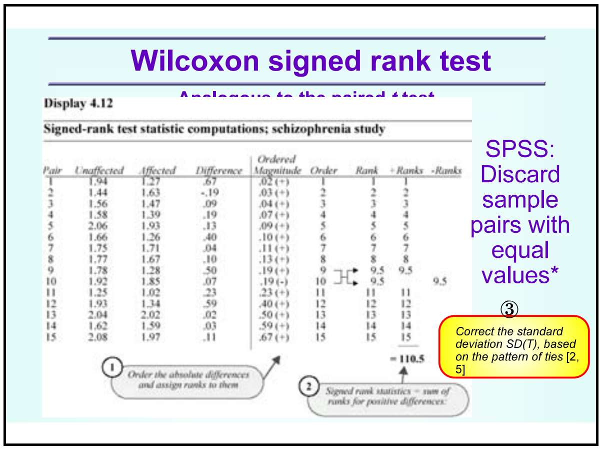
	AFFECTED - UNAFFECT
Exact Sig. (2-tailed)	.0010[a]

a. Binomial distribution used.
b. Sign Test

Paired *t* test, p=0.006

Slide 75 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

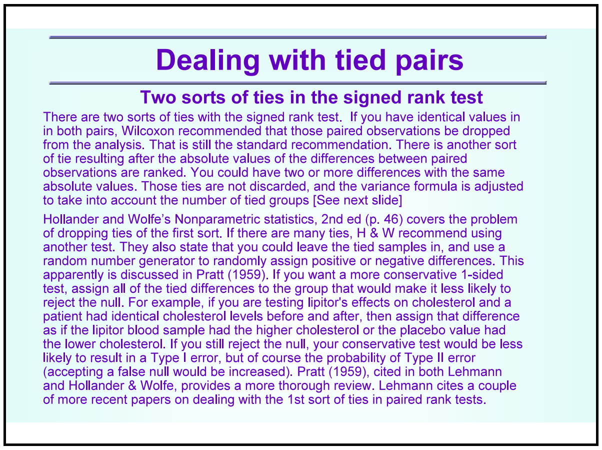
② Signed rank statistics = sum of ranks for positive differences.

SPSS: Discard sample pairs with equal values*

③ Correct the standard deviation SD(T), based on the pattern of ties [2, 5]

Slide 76 Wilcoxon signed rank test

NOTES:



Dealing with tied pairs

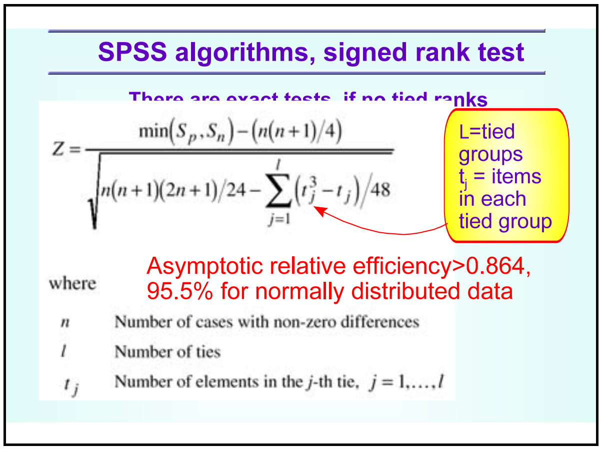
Two sorts of ties in the signed rank test

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 77 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

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

where

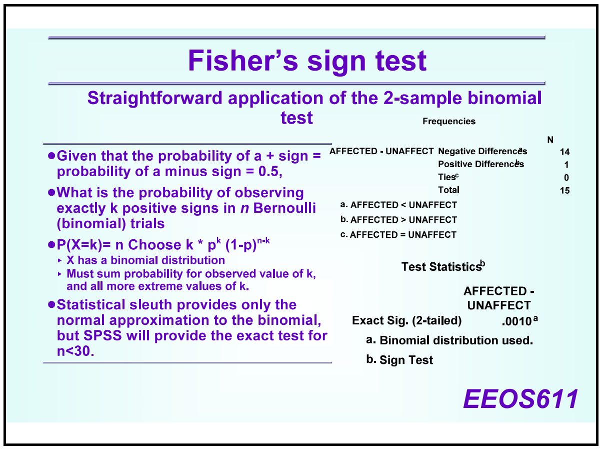
n Number of cases with non-zero differences

l Number of ties

t_j Number of elements in the j-th tie, $j = 1, \ldots, l$

Slide 78 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 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]

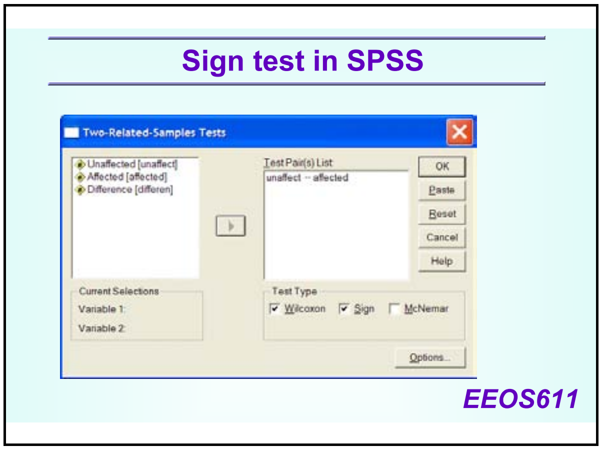
a. Binomial distribution used.

b. Sign Test

EEOS611

Slide 79 Fisher's sign test

NOTES:



Sign test in SPSS

Two-Related-Samples Tests

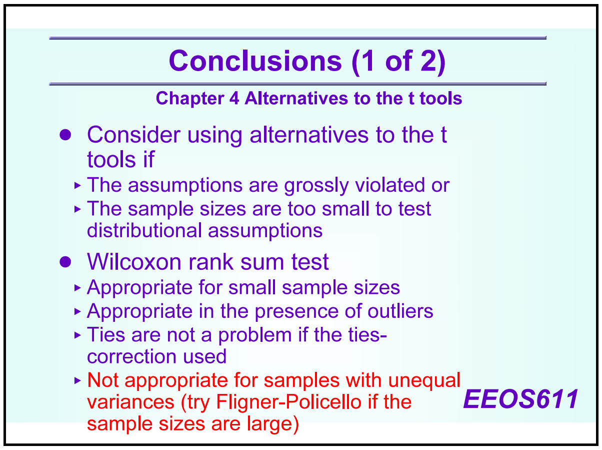
Test Pair(s) List: unaffect -- affected

Test Type: ☑ Wilcoxon ☑ Sign ☐ McNemar

EEOS611

Slide 80 Sign test in SPSS

NOTES:



Conclusions (1 of 2)

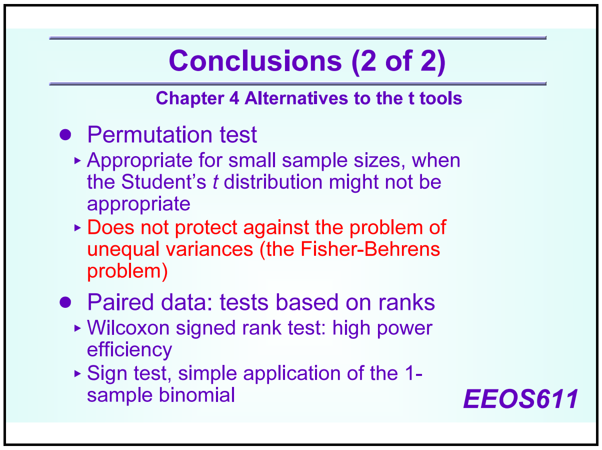
Chapter 4 Alternatives to the t tools

- 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 81 Conclusions (1 of 2)

NOTES:



Conclusions (2 of 2)

Chapter 4 Alternatives to the t tools

- Permutation test
  - Appropriate for small sample sizes, when the Student's *t* distribution might not be appropriate
  - Does not protect against the problem of unequal variances (the Fisher-Behrens problem)
- Paired data: tests based on ranks
  - Wilcoxon signed rank test: high power efficiency
  - Sign test, simple application of the 1-sample binomial

EEOS611

Slide 82 Conclusions (2 of 2)

NOTES: