Alternatives to t-Tools

RANK SUM TEST

WELCH'S TEST

Epigenetics & Smoking

This legacy occurs in the form of chemical changes to the surface of DNA that, in turn, affect how particular genes function, known as epigenetic changes. The modification identified in this study was DNA methylation, in which a molecule called a methyl group sits on the surface of DNA and influences whether genes are active or silent.

Related Article: Former smokers share emotional stories for CDC campaign

Studies have showed that smoking can cause these surface changes to DNA and that these changes could be used to measure the risk of particular diseases, such as cancer. But the new study, published in the journal Circulation: Cardiovascular Genetics, identified the diversity of the affected genes, the strength of the association with smoking and what genes are involved in someone's risk of disease.

"We had a very large sample, which gave us a lot of power ... and found sites in the genome where smoking leads to a difference in methylation," said Dr. Stephanie London, deputy chief of the epidemiology branch of the National Institute of Environmental Health Sciences, who led the study.

Let's Start With an Example

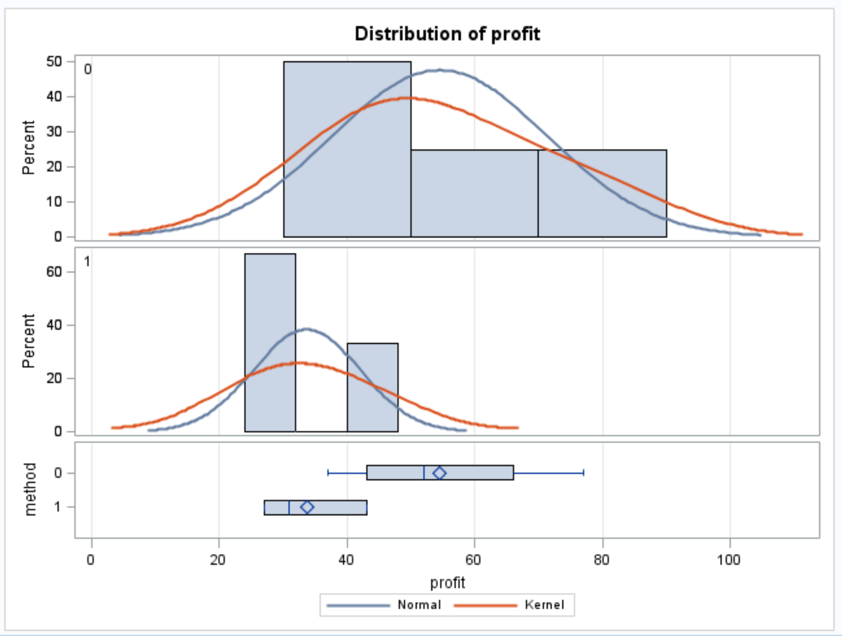
- IBM gives each employee in the marketing department technical training
- Based on further testing, it appears the traditional training method isn't effective
- Hence, a new training method is developed
- Below are the test scores of 4 individuals that just finished the "New Method" and the last 3 test scores from employees trained via the "Traditional Method" course
- Is there evidence to suggest that the "New Method" increases test scores?

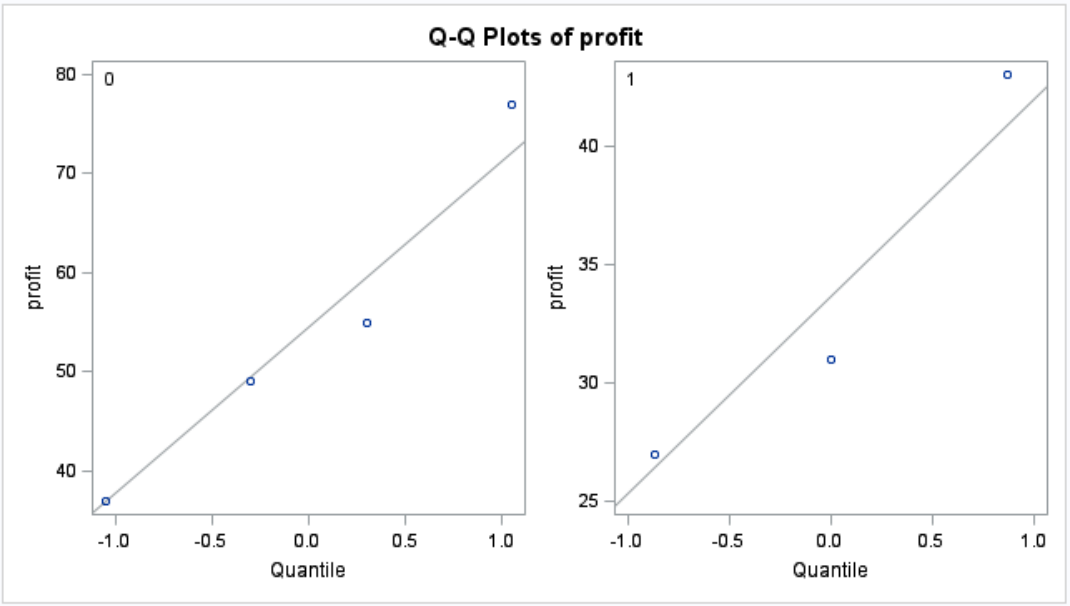
New Method	Traditional Method
37	23
49	31
55	46
77	

```
data example;
input Score Method $;
datalines;
37 New
49 New
55 New
77 New
23 Trad
31 Trad
46 Trad
;
```

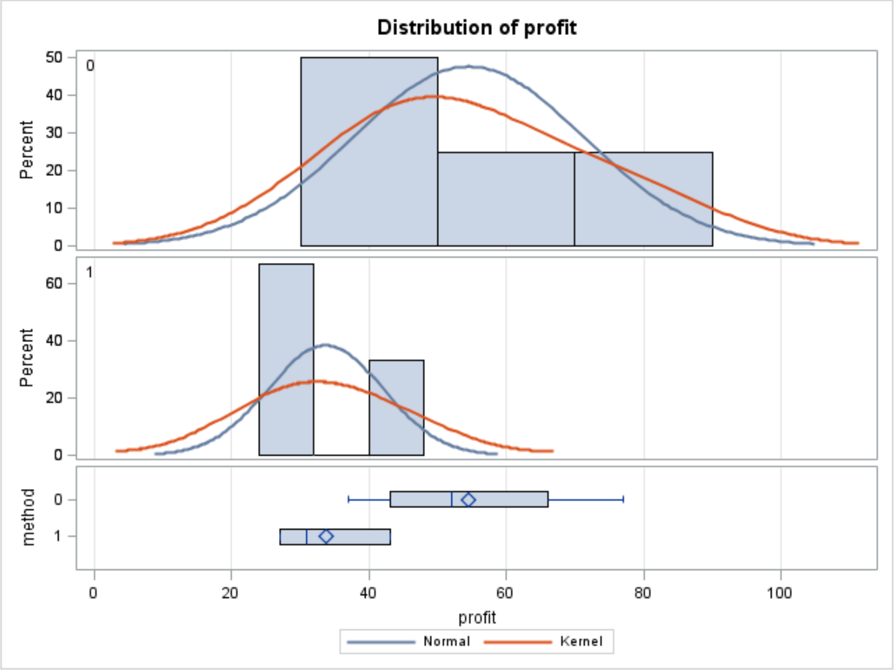
Examining the t-Tools Assumptions

```
data example;
input Score Method $;
datalines;
37 New
49 New
55 New
77 New
23 Trad
31 Trad
46 Trad
;
```





Since the standard deviation appear to be different and the sample sizes are both different and exceptionally small, the pooled t-test was not deemed appropriate and the non parametric rank sum test was performed.



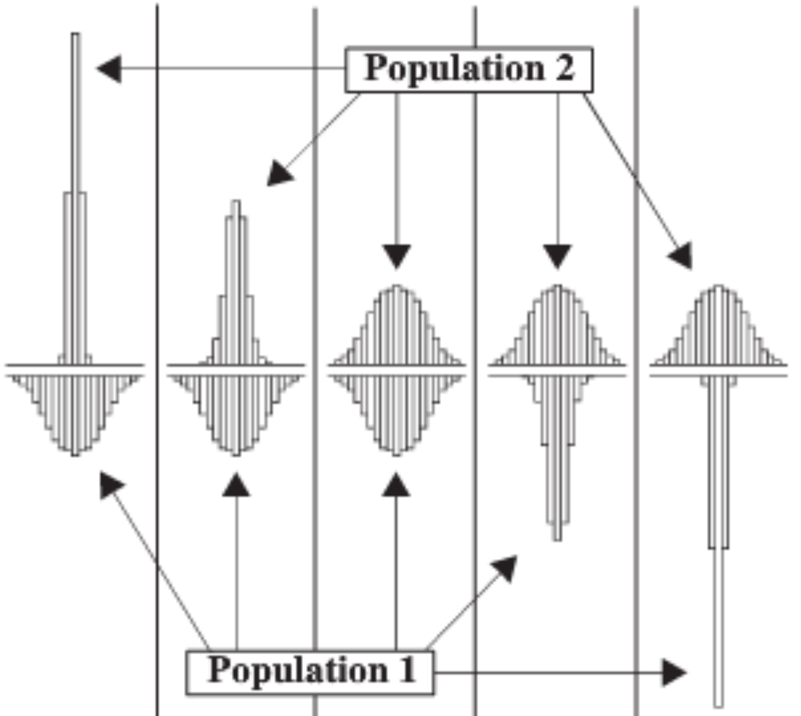
Which situation does it appear we are in?

$\sigma_2 < \sigma_1$ and $n_1 < n_2$
(less coverage)

$\sigma_2 > \sigma_1$ and $n_1 < n_2$
(more coverage)

Using a t-test could have low power

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

Rank-Sum Tests

Nonparametric Methods

- A **NONPARAMETRIC** or **DISTRIBUTION-FREE** test doesn't depend on (as many) underlying assumptions
- This makes them ideal for use when the assumptions of non-nonparametric (that is, **PARAMETRIC**) tests aren't met
- The trade-off is that nonparametric methods perform somewhat worse than parametric methods if the assumptions are approximately correct
- We already explored a nonparametric test: the randomization/permutation test from Chapter 1
- Now we will consider the "rank-sum test"

Rank-Sum Test: Discussion and Assumptions

- No distributional assumptions and resistant to outliers
- When t-test assumptions are met, the rank-sum test performs about 95.49% as well
- Performs arbitrarily better if the t-test assumptions are not (approximately) met
- Works well with **ORDINAL** data

(Realistically required for t-tools)

(**NOMINAL:** order is arbitrary. **ORDINAL:** order matters. **INTERVAL:** subtraction is meaningful. **RATIO:** multiplication is meaningful)

- Works with censored values

(Censored means that the actual value was too large/small to be accurately recorded)

- It still requires some assumptions:
  1. All observations are independent
  2. The *Y* values are ordinal

59 patients with arthritis who participated in a clinical trial were assigned to two groups, active and placebo. The response status:
(excellent=5, good=4, moderate=3, fair=2, poor=1)
of each patient was recorded.

Rank-Sum Test: Hypotheses

For the rank-sum test, our null hypothesis is in terms of **DISTRIBUTIONS** instead of means

H_0: The **DISTRIBUTION** of the “new” method scores is the same as the **DISTRIBUTION** of the “traditional” method scores

The Alternative Hypotheses:

H_A: The **DISTRIBUTION** of the “new” method scores is different from the **DISTRIBUTION** of the “traditional” method scores **(TWO SIDED)**

H_A: The **DISTRIBUTION** of the “new” method scores is larger than the **DISTRIBUTION** of the “traditional” method scores **(ONE SIDED)**

Note: “larger than” can be interpreted as “systematically higher than” in the sense that the probability of getting any value from one distribution is larger than for the other distribution

The Rank-Sum test

- We can compute the rank-sum test using the following steps:

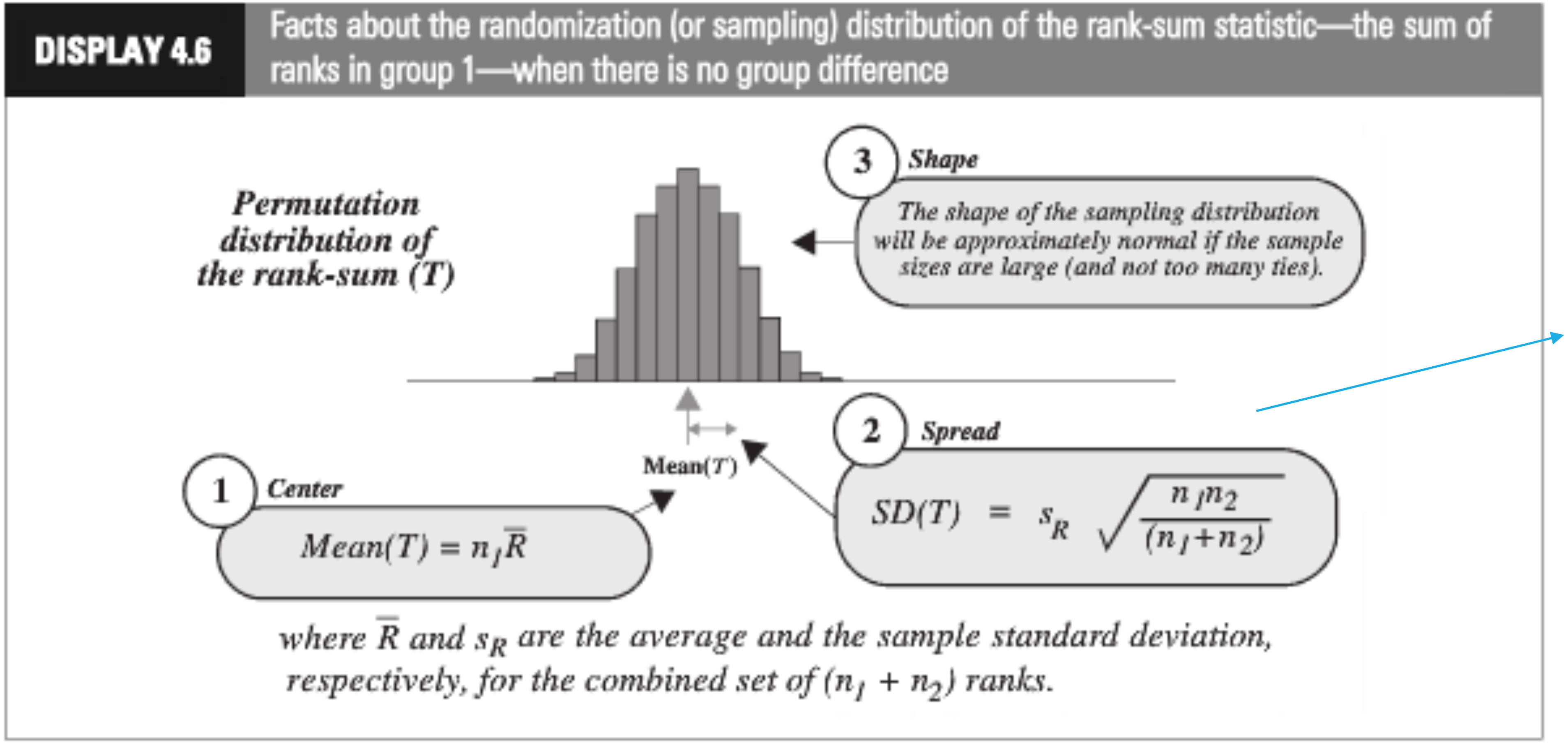
1. List all observations from both groups in increasing order
2. Assign each observation a rank, from 1 to n
3. If there are any ties, assign each such observation's rank to be the average of their ranks.
4. Identify each observation with its group

Note: n is the total # of observations

- The test statistic, T, is the sum of the ranks in one of the groups.

- We can find a p-value in two ways:
  - Normal approximation (use this if the sample size is very large)
  - Exact

Rank-Sum Test: Normal Approximation



$$Z = \frac{T - Mean(T)}{SD(T)}$$

Rank-Sum Test: Normal Approximation

H_0: The DISTRIBUTION of the "new" method scores is the same as the DISTRIBUTION of the "traditional" method scores

H_A: The DISTRIBUTION of the "new" method scores is larger than the DISTRIBUTION of the "traditional" method scores

```
proc npar1way data =  example Wilcoxon;
class Method;
var Score;
run;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Score Classified by Variable Method

Method	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
New	4	21.0	16.0	2.828427	5.250000
Trad	3	7.0	12.0	2.828427	2.333333

Wilcoxon Two-Sample Test

Statistic	7.0000
Normal Approximation	
Z	-1.5910
One-Sided Pr < Z	0.0558
Two-Sided Pr > |Z|	0.1116
t Approximation	
One-Sided Pr < Z	0.0814
Two-Sided Pr > |Z|	0.1627

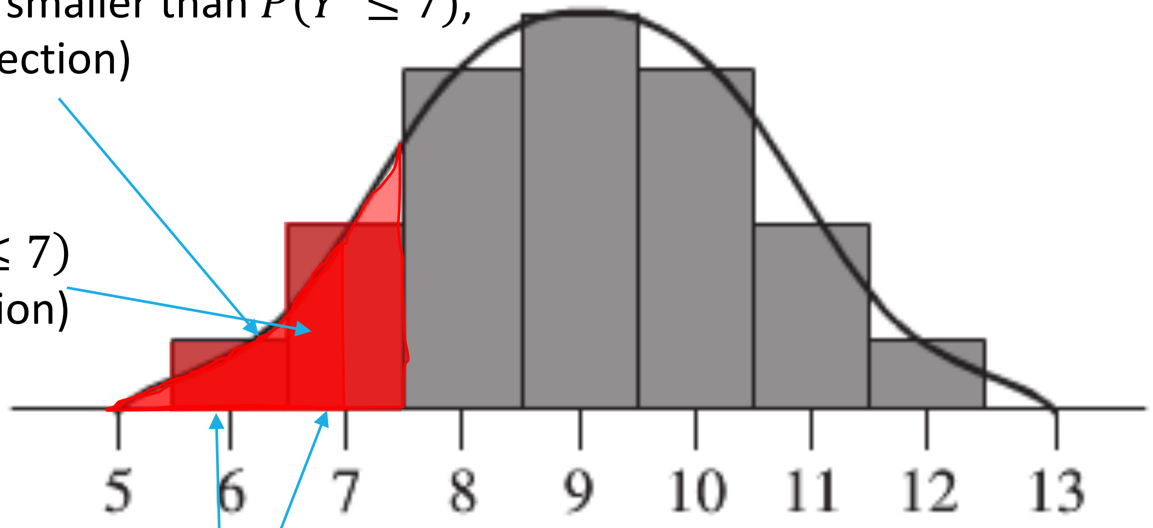
Z includes a continuity correction of 0.5.

There is mild evidence to suggest that the *distribution* of scores from the "New" method is greater than the *distribution* of the "Traditional" method (normal approximation to rank-sum test with continuity correction p-value = 0.0558).

Continuity Correction: Main Idea

$P(Z \le 7)$ is much smaller than $P(Y \le 7)$,
(no continuity correction)

$P(Z \le 7.5) \approx P(Y \le 7)$
(w/ continuity correction)



The exact probability calculation for $P(Y \le 7)$

Rank-Sum Test: Exact

Wilcoxon Scores (Rank Sums) for Variable Score Classified by Variable Method					
Method	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
New	4	21.0	16.0	2.828427	5.250000
Trad	3	7.0	12.0	2.828427	2.333333

```
data example;
input Score Method $;
datalines;
37 New
49 New
55 New
77 New
23 Trad
31 Trad
46 Trad
;

proc npar1way data =  example Wilcoxon;
class Method;
var Score;
exact;
run;
```

Wilcoxon Two-Sample Test	
Statistic (S)	7.0000
Normal Approximation	
Z	-1.5910
One-Sided Pr < Z	0.0558
Two-Sided Pr > |Z|	0.1116
t Approximation	
One-Sided Pr < Z	0.0814
Two-Sided Pr > |Z|	0.1627
Exact Test	
One-Sided Pr <= S	0.0571
Two-Sided Pr >= |S - Mean|	0.1143
Z includes a continuity correction of 0.5.	

Normal approximation p-values

Exact p-values

Which p-value should we use in problem?

Rank-Sum Test: Hypotheses about medians

For the rank-sum test, our null hypothesis is in terms of **DISTRIBUTIONS** instead of means

This can be stated in terms of medians if we make an additional assumption that the **SHAPE** of the distributions are the same

H_0: The **MEDIAN** of the "new" method score = the **MEDIAN** of the "traditional" method score

The Alternative Hypotheses:

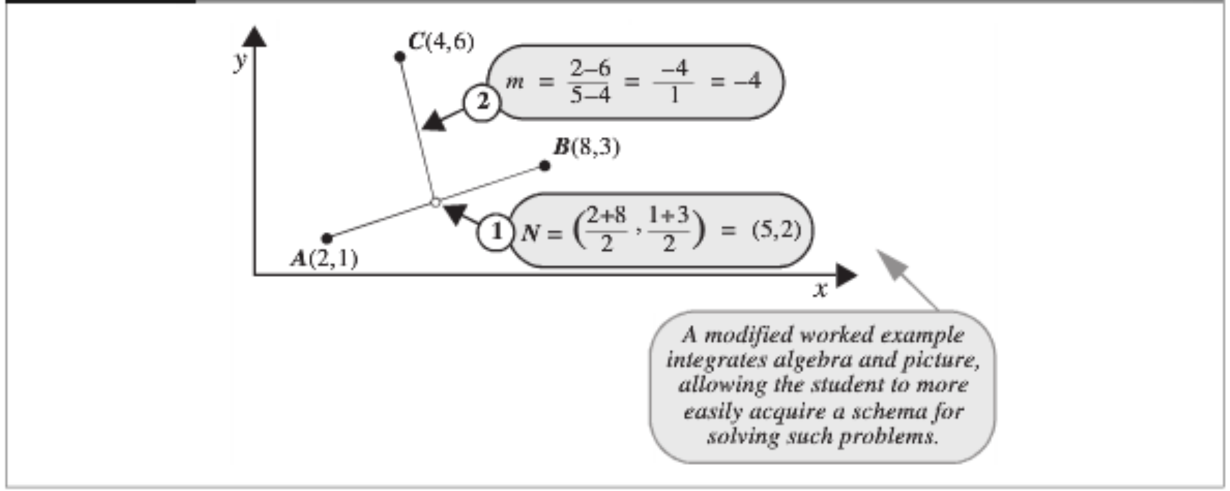
H_A: The **MEDIAN** of the "new" method score $\neq$ the **MEDIAN** of the "traditional" method score (and the shape of distributions are the same) **(TWO SIDED)**

H_A: The **MEDIAN** of the "new" method score $>$ the **MEDIAN** of the "traditional" method score (and the shape of distributions are the same) **(ONE SIDED)**

Cognitive Load Experiment

- Researchers compared the effectiveness of conventional textbook examples to modified ones
- They selected 28 ninth-year students, who had no previous exposure to coordinate geometry
- The students were randomly assigned to one of two self study instructional groups, using conventional and modified instructional materials.
- After instruction they were given a test and the time to complete one of the problems was recorded.

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*)



Is there sufficient evidence to suggest that the cognitive load theory (modified instruction) shortened response times?

Cognitive Load Experiment

DISPLAY 4.4 Number of seconds to solution of a problem in coordinate geometry, for students instructed with conventional and modified materials

Modified ($n = 14$)	Stem	*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	+++++

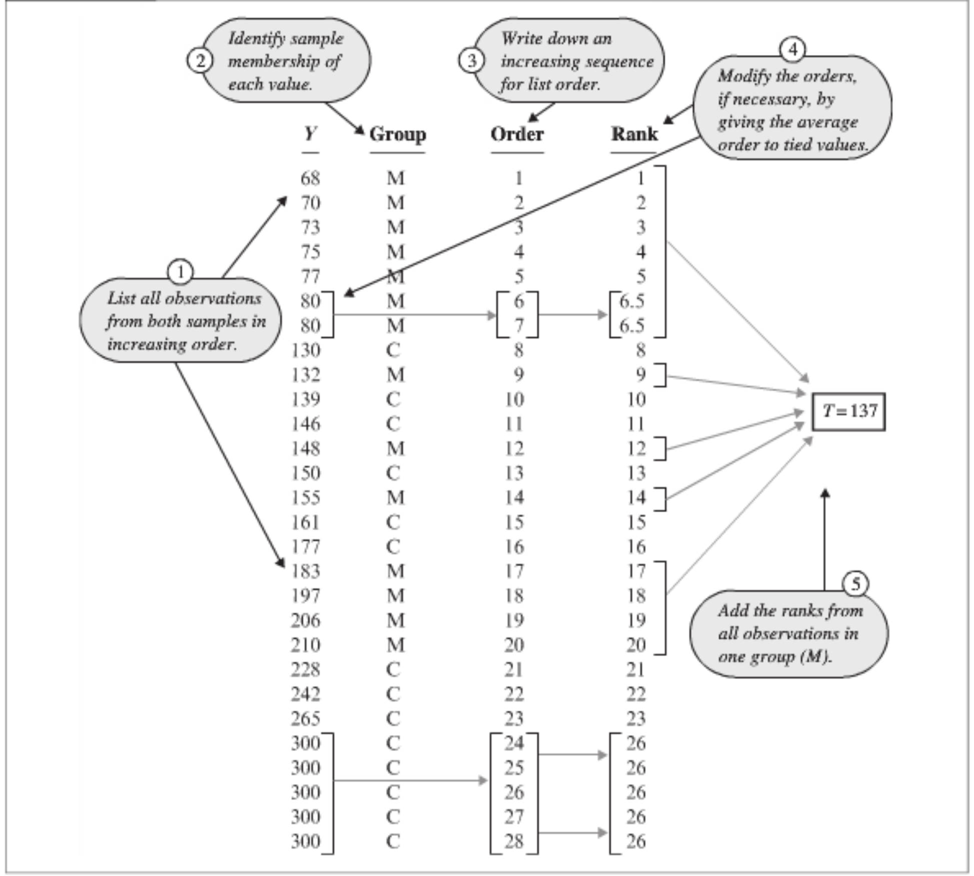
Five students did not complete the problem in the five-minute time allotment.

(CENSORED DATA)

Legend: 26 | 5 represents 265 seconds.

Cognitive Load Experiment

DISPLAY 4.5 Rank-sum test statistic T for the cognitive load experiment



Cognitive Load Experiment: Normal Approximation

DISPLAY 4.7 Finding the p-value with the normal approximation to the permutation distribution of the rank-sum statistic, using a continuity correction. Calculations for the cognitive load data are continued from Display 4.5

1. *Calculate the average and sample standard deviation of the ranks from the combined sample (column 4 of Display 4.5).*

$\overline{R} = 14.5 \qquad s_R = 8.2023$

2. *Compute the theoretical "null hypothesis" mean and standard deviation of T using the formulae in Display 4.6.*

$$\text{Mean}(T) = 14 \times 14.5 = 203 \qquad \text{SD}(T) = 8.2023\sqrt{\frac{14 \times 14}{(14+14)}} = 21.7013$$

3. *Calculate the Z-statistic using a continuity correction.*

$$Z = \frac{(137.5 - 203)}{21.7013} = -3.0183$$

4. *Find the p-value from a standard normal table.* **One-sided p-value = 0.0013**

(CONTINUITY CORRECTION)

Statistical Conclusion: The data provide convincing evidence that the distribution of times to solve the problem for students in "modified" group is larger for the "conventional" group (one-sided rank-sum test, normal approximation w/ C.C. p-value = 0.0013).

Cognitive Load Experiment: Using SAS

```
DATA pvalue_nocc;
    pval = CDF('NORMAL',(137-203)/21.7013);
RUN;
PROC PRINT DATA = pvalue_nocc;
```

Obs	pval
1	.001177825

```
DATA pvalue_yescc;
    pval = CDF('NORMAL',(137.5-203)/21.7013);
RUN;
PROC PRINT DATA = pvalue_yescc;
```

Obs	pval
1	.001271185

```
PROC NPAR1WAY DATA = cognitiveLoad WILCOXON;
    CLASS treatment;
    VAR time;
    EXACT;
RUN;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable time Classified by Variable treatment

treatment	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Modified	14	137.0	203.0	21.701254	9.785714
Conventi	14	269.0	203.0	21.701254	19.214286

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic (S)	137.0000
Normal Approximation	
Z	-3.0183
One-Sided Pr < Z	0.0013
Two-Sided Pr > |Z|	0.0025
t Approximation	
One-Sided Pr < Z	0.0027
Two-Sided Pr > |Z|	0.0055
Exact Test	
One-Sided Pr <= S	0.0008
Two-Sided Pr >= |S - Mean|	0.0016

Z includes a continuity correction of 0.5.

(Adding `WILCOXON` here produces the exact statistics and confidence intervals)

Cognitive Load Experiment: Confidence Interval (Chap. 4.2.4)

```
PROC NPAR1WAY DATA = cognitiveLoad WILCOXON ALPHA=0.05;
    CLASS treatment;
    VAR time;
    EXACT HL;
RUN;
```

Looks at the median value of all pairwise differences between the two groups in the data

DISPLAY 4.8 Using a rank-sum test to construct a confidence interval for an additive treatment effect (cognitive load study)

Hypothesized effect (seconds)	Two-sided *p*-value	Confidence interval inclusion?
50	0.0286	no
60	0.0800	yes
55	0.0403	no
58	0.0502	yes
150	0.1227	yes
160	0.0476	no
155	0.0589	yes
158	0.0530	yes
159	0.0502	yes

Try several hypothesized values for δ to identify those that have two-sided *p*-values ≥ 0.05.

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

Hodges-Lehmann Estimation

Location Shift (Modified - Conventi) -94.0000

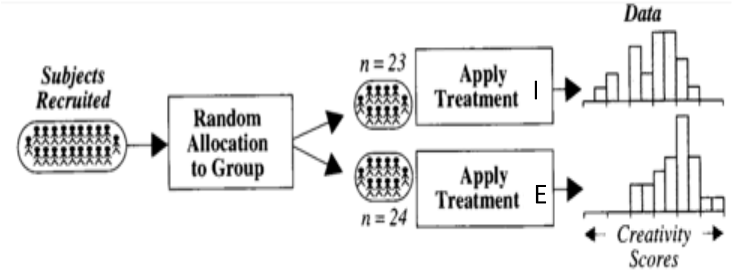
Type	95% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
Asymptotic (Moses)	-160.0000	-57.0000	-108.5000	26.2760
Exact	-158.0000	-59.0000	-108.5000	

Asymptotic:
midpoint - $z_{\alpha/2}$SE = -108.5 – $1.96 * 26.2760$ = [-160,-57]

Statistical Conclusion (continued): A range of plausible values the difference in median for the "modified" distribution vs. the "traditional" is [-158, -59] s. (95% confidence interval based on a rank-sum test) with a point-estimate of 108.5 s.

Welch’s T-Tools

Creativity Study: Reminder



→ Population mean: μ_I
→ Population sd: σ_I

→ Population mean: μ_E
→ Population sd: σ_E

- We additionally need to know/estimate the standard deviation of $\bar{Y}_I - \bar{Y}_E$
- There are two ways mentioned in the book
    1. Pooled SD
    2. Welch's SD
- To create the pooled SD, we need to assume that $\sigma_I = \sigma_E$
- Then, we can form an estimate of this common standard deviation via
- $s_p = \sqrt{\dfrac{(n_I - 1)\, s_I^2 + (n_E - 1)\, s_E^2}{n_I + n_E - 2}}$
- $SE(\bar{Y}_I - \bar{Y}_E) = \sqrt{\dfrac{\sigma_I^2}{n_I} + \dfrac{\sigma_E^2}{n_E}} \leftrightarrow SE(\bar{Y}_I - \bar{Y}_E) = s_p\sqrt{\dfrac{1}{n_I} + \dfrac{1}{n_E}}$

What if this assumption isn't true?

Welch’s t-Test

The only differences between Welch’s t-Test and the “pooled” t-test are:

- The standard error ($SE(\bar{Y}_I - \bar{Y}_E)$)
- The degrees of freedom (df)

(The new degrees of freedom are formed via a Satterthwaite approximation)

Luckily, we already know how to get the output from a Welch’s t-Test: PROC TTEST

Testing Hypothesis: Welch’s t-Tools

```
PROC TTEST DATA=creativity ORDER=DATA;
    CLASS intrinsic;
    VAR SCORE;
RUN;
```

The TTEST Procedure
Variable: score

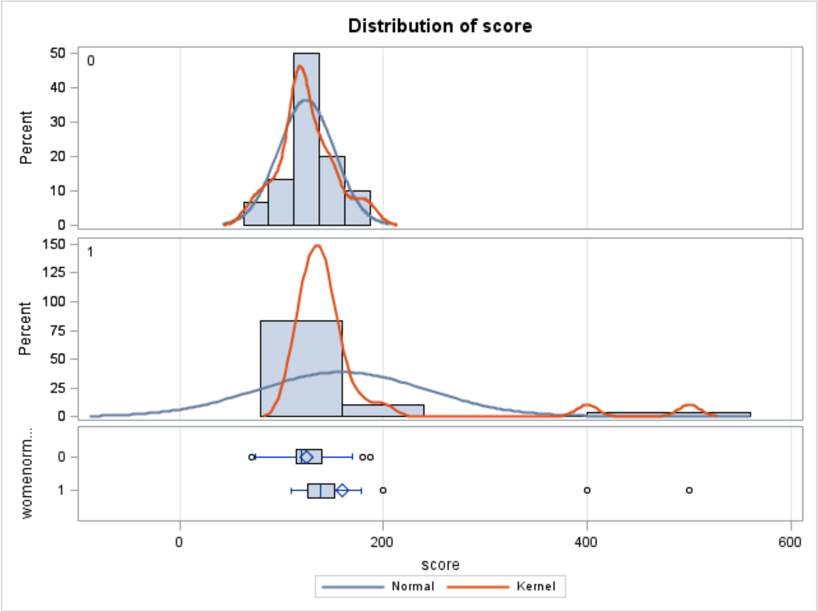
intrinsic	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	24	19.8833	4.4395	0.9062	12.0000	29.7000
0	23	15.7391	5.2526	1.0952	5.0000	24.0000
Diff (1-2)		4.1442	4.8541	1.4164		

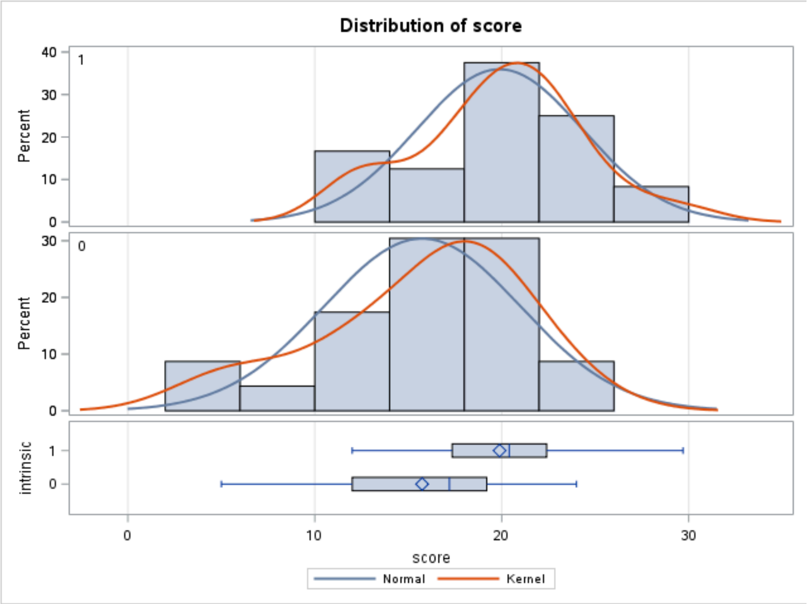
intrinsic	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		19.8833	18.0087	21.7580	4.4395	3.4504	6.2276
0		15.7391	13.4677	18.0105	5.2526	4.0623	7.4343
Diff (1-2)	Pooled	4.1442	1.2914	6.9970	4.8541	4.0261	6.1138
Diff (1-2)	Satterthwaite	4.1442	1.2776	7.0108			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	45	2.93	0.0054
Satterthwaite	Unequal	43.108	2.92	0.0056

“This experiment provides strong evidence that the intrinsic rather than extrinsic is associated with a higher scoring poem (p-value = 0.0056 from a Welch’s two-sample t-test). The estimated treatment effect is 4.14 pts (95% confidence interval $[1.28, 7.01]$ pts) on a 40 pt scale”

How to Decide?





- Use Welch's if the standard deviations are different but all other assumptions of t-Tools are met
- If the t-Tool assumptions are at all questionable, use a nonparametric test

Example: Forest Fires

When wildfires ravage forests, the timber industry argues that logging the burned trees enhances forest recovery; the EPA argues the opposite. The 2002 Biscuit Fire in southwest Oregon provided a test case. Researchers selected 16 fire-affected plots in 2004-before any logging was done-and counted tree seedlings along a randomly located transect pattern in each plot. They returned in 2005, after nine of the plots had been logged, and counted the tree seedlings along the same transects. The percent of seedlings lost from 2004 to 2005 is recorded in the table below for logged (L) and unlogged (U) plots:

Test the EPA's assertion that logging decreased the percentage of seedlings from 2004 to 2005.