ST 411/511 Outline 4

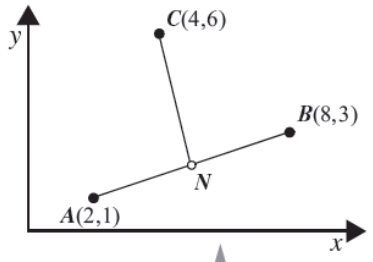
Reading assignment: Chapter 4, omitting Section 4.4.2. This chapter describes some procedures that can be used when the assumptions of the t-test are so badly violated that the results are not reliable. We will skip the Signed-Rank Test of Section 4.4.2.

Chapter 4 Alternatives to the *t*-tools

The Rank-Sum Test (Wilcoxon Rank-Sum Test, Mann-Whitney Test): A nonparametric test for comparing two populations.

Case Study 4.1.2 Cognitive load study. Research question is "do modified materials reduce time to solve problem?"

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:*

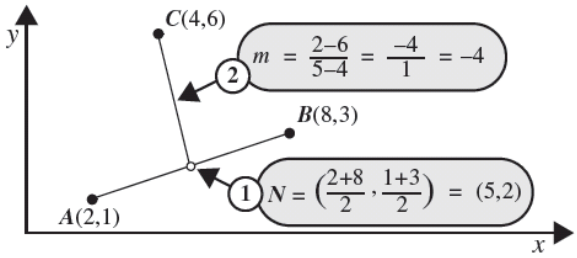
$$N = \left(\frac{2+8}{2}, \frac{1+3}{2}\right) = (5,2)$$

The slope of CN *is:*

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

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

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*)



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

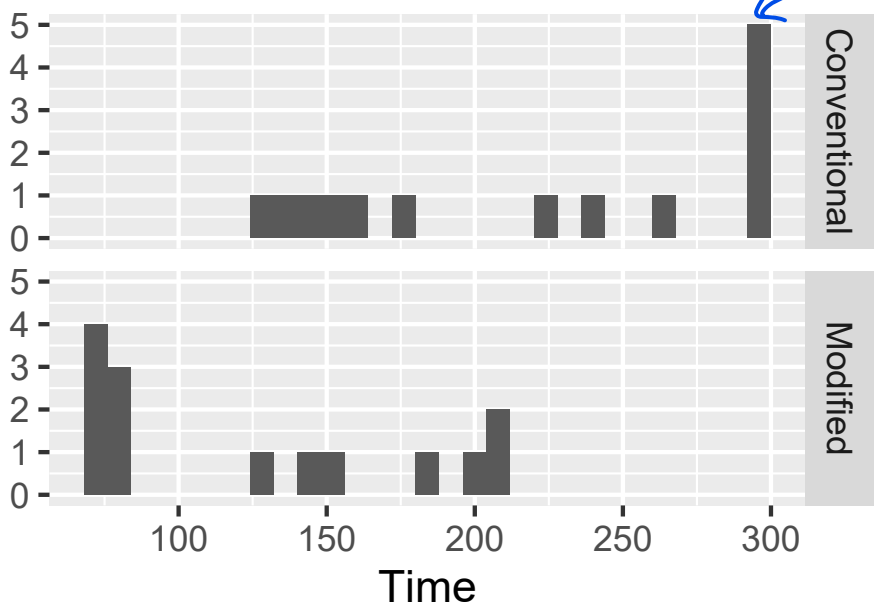
28 students (random sample? Didn't say, so assume not.)
Randomly assigned to groups
Response: Time to solve a math problem.

```
> case0402
   Time    Treatment Censored
1    68     Modified        0
2    70     Modified        0
3    73     Modified        0
4    75     Modified        0
5    77     Modified        0
6    80     Modified        0
```

... (snipped to save space)

```
21  228 Conventional        0
22  242 Conventional        0
23  265 Conventional        0
24  300 Conventional        1
25  300 Conventional        1
26  300 Conventional        1
27  300 Conventional        1
28  300 Conventional        1
```

These students didn't finish within 300 sec. Their times are not known, except they're bigger than 300 (they're "censored.")



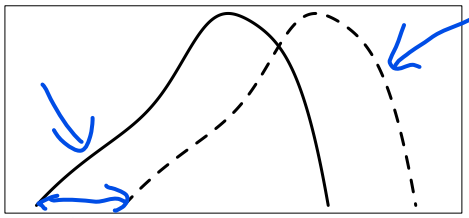
The 5 censored observations

Log transform won't help.

Hypotheses for Wilcoxon Rank-Sum Test

Restrictive formulation of rank-sum test: Suppose Y is the time a random student takes to complete the task using the modified methods. Then $Y + \delta$ is the time for the same student using the conventional methods.

Distn of mod. times



Distn of conv. times

δ

δ is not a parameter of either distn. It describes their relationship.

Null hypothesis for rank-sum test, restrictive formulation:

$H_0: \delta = 0$ No diff between pop'ns.

Alternative hypothesis for rank-sum test, restrictive formulation:

$H_A: \delta > 0$ Longer times with conv. materials

(can have 2-sided test also)

} We will use these

Less restrictive formulation: (You can safely ignore this for ST 411/511.)

Y_C = a randomly-chosen time from population using conventional materials
Y_M = a randomly-chosen time from population using modified materials

Null hypothesis for rank-sum test, less restrictive formulation:

Neither group more likely to have larger times.

$H_0: Pr(Y_C > Y_M) = Pr(Y_M > Y_C)$

"probability"

Alternative hypothesis for rank-sum test, less restrictive formulation:

Conv. time is more likely to be larger

$H_A: Pr(Y_C > Y_M) > Pr(Y_M > Y_C)$

Wilcoxon Rank-Sum Test Procedure:

1. Rank-transform the data.
2. Test statistic is the sum of the ranks in the first group.
3. Do a permutation/randomization test.

Ranks: A *rank transformation* orders a column of data and assigns to each value its order.

```
> dat # Some made-up data
   Y Group
1 18     A
2  6     A
3 14     A
4  2     A
5  7     B
6  8     B
7 20     B
8 47     B
```

smallest

largest

```
> dat$rank <- rank(dat$Y)  # Rank-transform dat$Y
> dat
   Y Group rank
1 18     A    6
2  6     A    2
3 14     A    5
4  2     A    1
5  7     B    3
6  8     B    4
7 20     B    7
8 47     B    8
```

smallest

largest

Midranks: What happens when there are ties?

```
> # 4th and 7th Y's are equal.
   Y Group
1 18     A
2  6     A
3 14     A
4 20     A
5  7     B
6  8     B
7 20     B
8 47     B
```

tied values. These are the 6th & 7th largest values

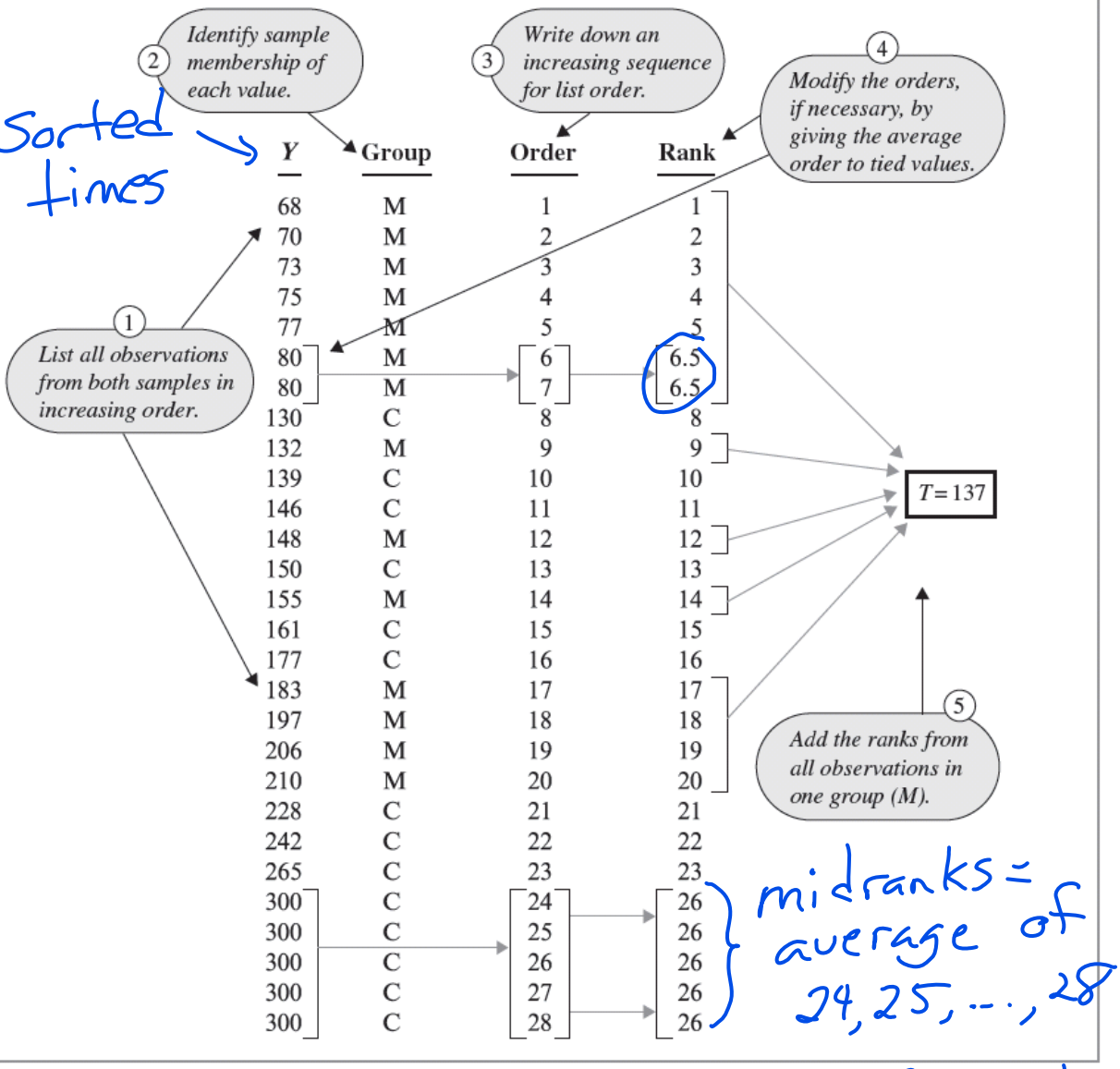
```
> dat$rank <- rank(dat$Y)
> dat
   Y Group rank
1 18     A  5.0
2  6     A  1.0
3 14     A  4.0
4 20     A  6.5
5  7     B  2.0
6  8     B  3.0
7 20     B  6.5
8 47     B  8.0
```

← use rank() to define column.

midrank = average of 6 & 7

Test Statistic for Wilcoxon Rank-Sum Test:

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

Sorted times

midranks = average of 24, 25, ..., 28

Test stat is $T = 137$ = sum of ranks in modified group. This is not our usual t-stat. It does not have a t-dist'n under H_0.

Doesn't matter which group you choose. See Lab 4 activity

Sampling Distribution of the Test Statistic Under the Null Hypothesis

Randomization test: randomly re-allocate observed times to groups, and calculate test statistic T. Repeat many times to get distribution of T under H_0.

original data

Time	Treatment	Censored	Rank
68	Modified	0	1.00
70	Modified	0	2.00
73	Modified	0	3.00
75	Modified	0	4.00
77	Modified	0	5.00
80	Modified	0	6.50
80	Modified	0	6.50
132	Modified	0	9.00
148	Modified	0	12.00
155	Modified	0	14.00
183	Modified	0	17.00
197	Modified	0	18.00
206	Modified	0	19.00
210	Modified	0	20.00
130	Conventional	0	8.00
139	Conventional	0	10.00
146	Conventional	0	11.00
150	Conventional	0	13.00
161	Conventional	0	15.00
177	Conventional	0	16.00
228	Conventional	0	21.00
242	Conventional	0	22.00
265	Conventional	0	23.00
300	Conventional	1	26.00
300	Conventional	1	26.00
300	Conventional	1	26.00
300	Conventional	1	26.00
300	Conventional	1	26.00

$T = 1.0 + 2.0 + 3.0 + 4.0 + 5.0 + 6.5 + 6.5 + 9.0 + 12.0 + 14.0 + 17.0 + 18.0 + 19.0 + 20.0 = 137$

obs. value of test stat.

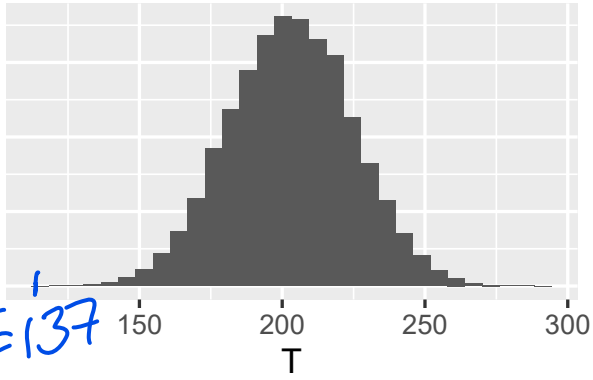
one random reallocation

shuffled

Time	Treatment	Censored	Rank
68	Conventional	0	1.00
70	Conventional	0	2.00
73	Modified	0	3.00
75	Conventional	0	4.00
77	Modified	0	5.00
80	Conventional	0	6.50
80	Modified	0	6.50
132	Conventional	0	9.00
148	Modified	0	12.00
155	Modified	0	14.00
183	Conventional	0	17.00
197	Conventional	0	18.00
206	Modified	0	19.00
210	Modified	0	20.00
130	Modified	0	8.00
139	Conventional	0	10.00
146	Modified	0	11.00
150	Modified	0	13.00
161	Conventional	0	15.00
177	Conventional	0	16.00
228	Conventional	0	21.00
242	Conventional	0	22.00
265	Modified	0	23.00
300	Conventional	1	26.00
300	Modified	1	26.00
300	Modified	1	26.00
300	Modified	1	26.00
300	Conventional	1	26.00

$T = 3.0 + 5.0 + 6.5 + 12.0 + 14.0 + 19.0 + 20.0 + 8.0 + 11.0 + 13.0 + 23.0 + 26.0 + 26.0 + 26.0 = 212.5$

value of T for this randomized arrangement



Histogram of 500,000 Ts

Looks normal to use normal approx. to easily approx. p-value.

T=137

Notation:

$n_1, n_2 =$ sample sizes

$\overline{R} =$ sample mean of ranks $= \frac{n_1+n_2+1}{2}$

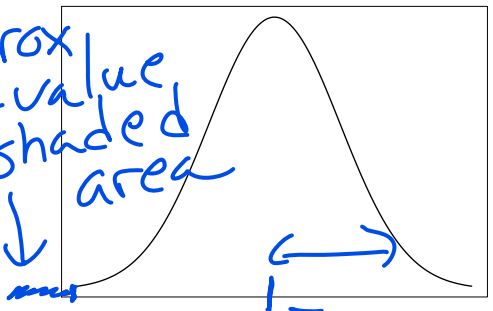
$s_R =$ sample std. dev. of ranks

Mean of sampling distribution of T under H_0: $n_1\overline{R}$ n_1 = sample size in modified group

Standard deviation of sampling distribution of T under H_0: $s_R\sqrt{\frac{n_1 n_2}{n_1+n_2}}$

Not obvious but can be derived mathematically

Normal approximation: of sampling distn of T under Ho



approx p-value = shaded area

T $n_1\overline{R}$ spread determined by $s_R\sqrt{\frac{n_1 n_2}{n_1+n_2}}$

Caution: This does not say we are assuming the population is normal.

Distn of T under Ho is approx. normal. Normal approx. is valid if $n_1, n_2 > 5$ and not too many ties.

Wilcoxon Rank-Sum Test in R

```
> summary(case0402$Treatment)
Conventional     Modified
          14           14
```

R's group 1 is conv.

Conv. times are larger

```
> wilcox.test(Time~Treatment, data=case0402, alternative="greater",
+             exact=FALSE, correct=FALSE)
Wilcoxon rank sum test

data:  Time by Treatment
W = 164, p-value = 0.001178
alternative hypothesis: true location shift is greater than 0
```

use normal approx.

Don't use a continuity correction.

$$W = T - \underbrace{\frac{n_1(n_1+1)}{2}}_{\text{smallest possible value of } T}$$

Statistical Conclusion: Strong evidence that solution time using conv. materials is longer than using mod. materials ($p \approx 0.0012$, one-sided Wilcoxon rank-sum test)

Note: not reporting on pop'n means.

A Rank-Sum Confidence Interval For Shift Parameter δ

If Y is a random student's time from modified group, then $Y+\delta$ would be their time using conv. materials.

```
> wilcox.test(Time~Treatment, data=case0402, exact=FALSE,
+             correct=FALSE, conf.int=TRUE)

Wilcoxon rank sum test

data:  Time by Treatment
W = 164, p-value = 0.002356
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 58.99997 158.00005
sample estimates:
difference in location
                    94
```

← 2-sided test

← 95% CI for δ

← pt. est. of δ

Statistical Conclusion: We estimate that modified materials cause* a decrease in solution time** by 59 to 158 sec. (95% Wilcoxon CI).

* because students randomized to groups

** Not mean

How `wilcox.test()` calculates the confidence interval

Recall the relationship between a hypothesis test and a confidence interval: If μ_0 is in the two-sided 95% confidence interval for μ, then the p-value of a two-sided test of $H_0 : \delta = \delta_0$ is greater than 0.05.

```
> with(case0402, wilcox.test(x=Time[Treatment=="Modified"]+58,
+                            y=Time[Treatment=="Conventional"],
+                            exact=FALSE, correct=FALSE))

Wilcoxon rank sum test

data:  Time[Treatment == "Modified"] + 58
            and Time[Treatment == "Conventional"]
W = 55, p-value = 0.04754
alternative hypothesis: true location shift is not equal to 0
```

p-value < 0.05, so $\delta_0 = 58$ is not a plausible value for δ, so 58 is not in the 95% CI

```
> with(case0402, wilcox.test(x=Time[Treatment=="Modified"]+59,
+                            y=Time[Treatment=="Conventional"],
+                            exact=FALSE, correct=FALSE))

Wilcoxon rank sum test

data:  Time[Treatment == "Modified"] + 59
            and Time[Treatment == "Conventional"]
W = 57, p-value = 0.05868
alternative hypothesis: true location shift is not equal to 0
```

p-value > 0.05, so 59 is in the 95% CI.

```
> with(case0402, wilcox.test(x=Time[Treatment=="Modified"]+60,
+                            y=Time[Treatment=="Conventional"],
+                            exact=FALSE, correct=FALSE))

Wilcoxon rank sum test

data:  Time[Treatment == "Modified"] + 60
       and Time[Treatment == "Conventional"]
W = 59.5, p-value = 0.07601
alternative hypothesis: true location shift is not equal to 0
```

δ_0

60 is in 95% CI

...etc. Tried all values $61, \ldots, 157$. All p-values were > 0.05.

```
> with(case0402, wilcox.test(x=Time[Treatment=="Modified"]+158,
+                            y=Time[Treatment=="Conventional"],
+                            exact=FALSE, correct=FALSE))

Wilcoxon rank sum test

data:  Time[Treatment == "Modified"] + 158
       and Time[Treatment == "Conventional"]
W = 140.5, p-value = 0.05015
alternative hypothesis: true location shift is not equal to 0
```

δ_0

158 is in 95% CI

```
> with(case0402, wilcox.test(x=Time[Treatment=="Modified"]+159,
+                            y=Time[Treatment=="Conventional"],
+                            exact=FALSE, correct=FALSE))

Wilcoxon rank sum test

data:  Time[Treatment == "Modified"] + 159
       and Time[Treatment == "Conventional"]
W = 141, p-value = 0.04754
alternative hypothesis: true location shift is not equal to 0
```

δ_0

159 is not in 95% CI

One δ_0 in the middle:

```
> with(case0402, wilcox.test(x=Time[Treatment=="Modified"]+110,
+                            y=Time[Treatment=="Conventional"],
+                            exact=FALSE, correct=FALSE))

Wilcoxon rank sum test

data:  Time[Treatment == "Modified"] + 110
       and Time[Treatment == "Conventional"]
W = 117, p-value = 0.3812
alternative hypothesis: true location shift is not equal to 0
```

near middle

large p-value

δ_0's in middle of 95% CI have large p-values ("no evidence" range)

δ_0's near endpoints of CI have smaller p-values ("suggestive" range).

(See Display 2.12)

Case Study 4.1.1 Space Shuttle O-ring Data. Research question is, "are cooler temperatures associated with more O-ring incidents?"

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

Definitely not normal - all those zeros! Too many ties for Wilcoxon rank-sum test.

Permutation/Randomization Tests

Use repeated random re-allocation of observed data to the different groups, keeping sample sizes the same as observed data, to get the sampling distribution of the test statistic under H_0. It's a permutation test when the sample is simple enough (small sample size, few different values) that we can enumerate all possible values of the test statistic.

test stat: $\dfrac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ = t-stat

This is a standardized version of diff. in sample means.

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

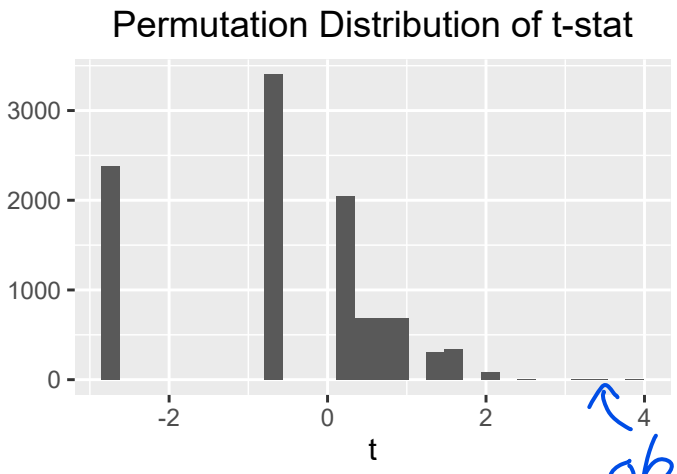
only 8 distinct t-stats with these data

Number of rearrangements with identical *t*-statistics	*t*-statistic
2,380	−1.188
3,400	−0.463
2,040	0.231
1,530	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

one-sided p*-value from a permutation test of the* t*-statistic:*
105/10,626 = 0.00988



t-stat doesn't have a t distn here because assumptions not met.

obs. t-stat = 3.888

exact p-value = proportion of permutations of data where t-stat $\geq 3.888 = \frac{95+10}{10,626} \approx 0.0099$

```
> library(perm)
> permTS(Incidents~Launch, data=case0401, alternative="greater",
         exact=TRUE)

Exact Permutation Test (network algorithm)

data:  Incidents by Launch
p-value = 0.009881
alternative hypothesis:
       true mean Launch=Cool - mean Launch=Warm is greater than 0
sample estimates:
mean Launch=Cool - mean Launch=Warm
                                 1.3
```

Statistical conclusion

strong evidence that cooler temps are associated* with more incidents ($p \approx 0.0099$, one sided 2-sample perm. test).

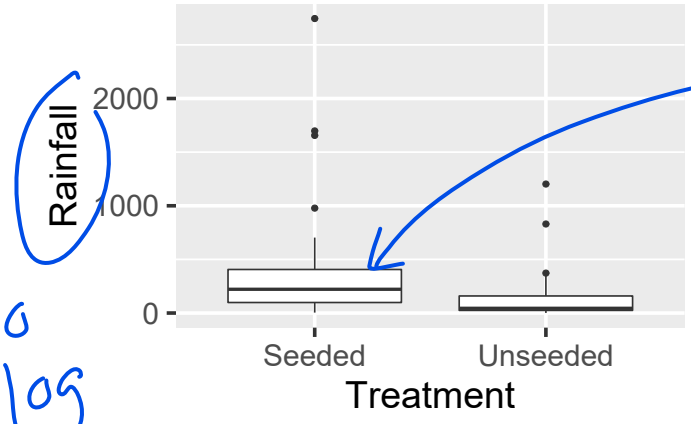
* not caused because the launched weren't randomized to temps.

```
> wilcox.test(Incidents~Launch, data=case0401, alternative="greater")

Wilcoxon rank sum test with continuity correction

data:  Incidents by Launch
W = 74, p-value = 0.000572
alternative hypothesis: true location shift is greater than 0

Warning message:
In wilcox.test.default(x = c(1L, 1L, 1L, 3L), y = c(0L, 0L, 0L,  :
  cannot compute exact p-value with ties
```

similar p-value "strong evidence"

Too many ties.

Welch's t-test An approximation to the two-sample t-test that doesn't assume populations have equal standard deviations.



bigger spread (box is taller), so don't assume pop'n variances equal.

no log

no var.equal= TRUE

```
> t.test(Rainfall~Treatment, data=case0301, alternative="greater")

        Welch Two Sample t-test

data:  Rainfall by Treatment
t = 1.9982, df = 33.855, p-value = 0.02689
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 42.63408      Inf
sample estimates:
  mean in group Seeded mean in group Unseeded
              441.9846                164.5885
```

approx. df not n-2

See p. 98 of textbook for formula if you're curious.

```
> t.test(Rainfall~Treatment, data=case0301)

Welch Two Sample t-test

data:  Rainfall by Treatment
t = 1.9982, df = 33.855, p-value = 0.05377
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -4.764295 559.556603
sample estimates:
  mean in group Seeded mean in group Unseeded
              441.9846                164.5885
```

Statistical Conclusion: (for both hypothesis test & CI)

There is moderate evidence that seeding caused higher mean rainfall ($p \approx 0.027$, one-sided Welch's t-test). We estimate that mean rainfall on seeded days is between about 4.8 acre-feet less to 559.6 acre-feet more than of unseeded days (95% CI).

Or report 1-sided CI:

We estimate the increase in mean rainfall due to seeding is at least 42.6 acre-feet (95% one-sided CI).

Summary of alternatives to the two-sample t-test

- Welch's t-test–useful when normality assumption is reasonable but standard deviations are not equal.

Welch's t-test OK for moderately skewed distns.

- Rank-sum test (aka Wilcoxon rank-sum test)–useful when normality assumption fails badly. Does not test population means or medians. Confidence interval for "shift parameter" δ. Doesn't like too many ties.

restrictive formulation

- Permutation/randomization test–general tool only requiring the independence assumption, but does not test population parameters, only that the two groups come from the same population.

Alternatives for paired data

Sign Test: Have we observed an unusual number of positive (or negative) differences?

DISPLAY 2.2 Differences in volumes (cm^3) of left hippocampus in 15 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: 15

Legend: 6 | 7 represents 0.67 cm^3

Null hypothesis: H_0: median of popn of diffs = 0
H_A: median of diffs > 0 (one-sided test)

Test statistic: K = # positive diffs

Sampling distribution of test statistic under H_0
is like tossing a fair coin 15 times
50% chance a random diff > 0
" " < 0
K has a binomial distn
with index n = 15 and 50%
chance of success.

```
> # Sign test
> n <- 15
> K <- 14
> binom.test(K, n, alternative="greater")

Exact binomial test

data:  K and n
number of successes = 14, number of trials = 15, p-value = 0.0004883
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.7206038 1.0000000
sample estimates:
probability of success
             0.9333333
```

"index" = "# trials" = sample size

← # pos. diffs

H_A: pos. diff. more likely

CI & pt. est. are for probability of pos. diff. We won't use these here.

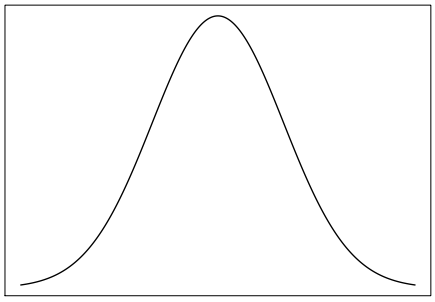
Statistical conclusion:

Strong evidence that median diff is larger than 0 ($p \approx 0.0005$, one-sided sign test).

Normal approximation to binomial

```
> 1 - pnorm(K, mean=n/2, sd=sqrt(n/4))
[1] 0.0003945565
```

Book does a continuity correction.



Normal approx. is a way to get approx. p-value with less computation. This is not so much an issue anymore.

Wilcoxon Signed-Rank Test: Not to be confused with the Wilcoxon rank-sum test or the sign test.

Null hypothesis: Population is symmetric about 0.

Ignore

Summary of alternatives to the paired t-test

- Sign test
- Signed-rank test

Levene's (Median) Test for Equality of Variances

Note: Please don't use this to test the equal standard deviation assumption. Use diagnostic plots (e.g. boxplots) and summary statistics instead.

would you then check assumptions of Levene's test with another hypothesis test?

```
> library(car)
> with(case0102, leveneTest(Salary~Sex))
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.1876 0.6659
      91
```

We will not use this test in ST 411/511

Statistical vs. practical significance

Significant (adjective)

Signifying something; carrying meaning.

a significant word or sound; a significant look

Having a noticeable or major effect; notable.

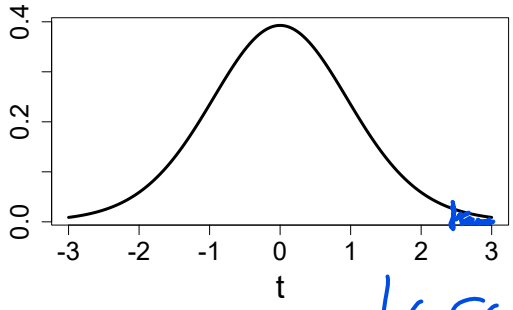
The First World War was a significant event.

(statistics) Having a low probability of occurring by chance (for example, an unusually large t-statistic).

so small p-value

Recall t-statistic to test $H_0 : \mu_1 - \mu_2 = 0$

$$t\text{-stat} = \frac{\overline{Y_1} - \overline{Y_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



Can get a small p-value even if $\overline{Y}_1 - \overline{Y}_2$ is insignificantly small.

large t-stat $\Rightarrow$ small p-value

$\overline{Y}_1 - \overline{Y}_2$ estimates $\mu_1 - \mu_2$ = "effect size"

Effect size may be so small as to not make any practical diff, yet p-value could be small enough to yield "stong evidence."