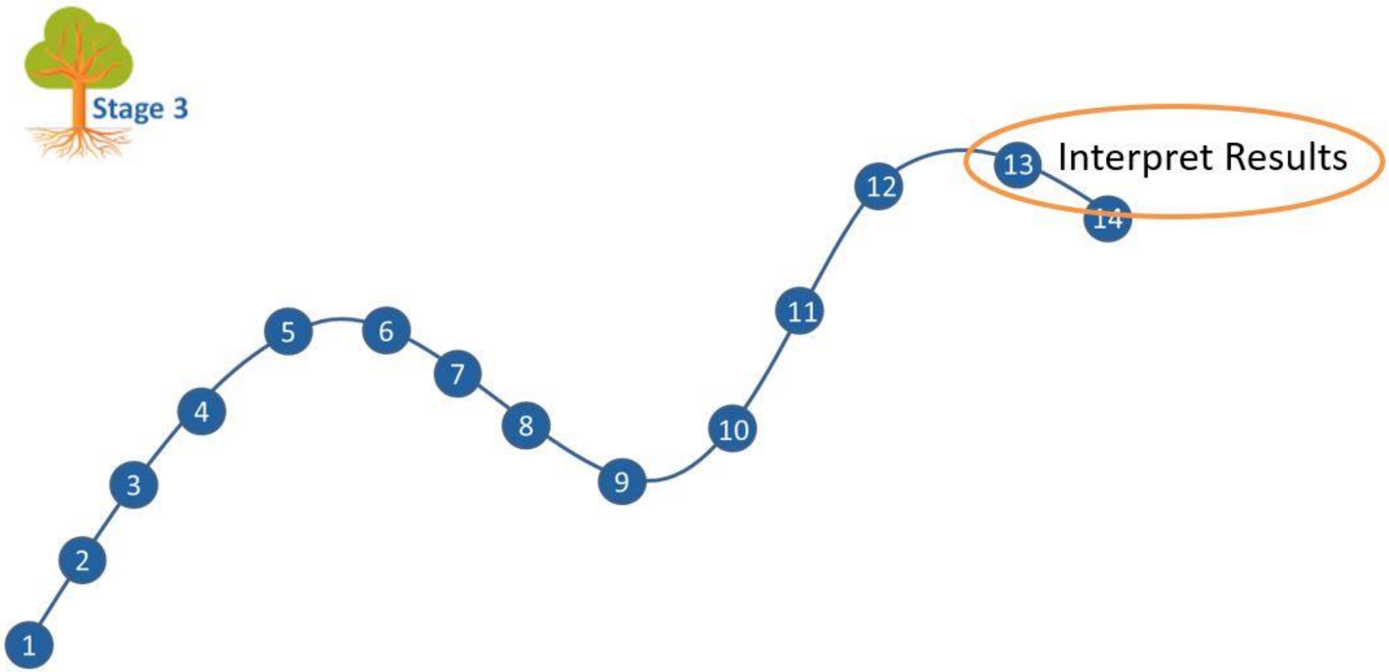
statistically speaking

Exact and Randomization Tests

Steve Simon

Where this fits



Goal

For you to have a good understanding of:

- what randomization and exact tests are
- the steps to implement them
- when it is appropriate to use them

The goal is not to:

- cover every possible application

When should you use exact/randomization tests

You don’t want to rely on

- underlying distributional assumptions
- the Central Limit Theorem

Note: Randomization tests are also commonly called permutation tests. I use the two terms interchangeably.

Outline of topics

1. Historical origins of Fisher’s Exact Test

2. Other exact tests

3. Randomization tests

4. When should you use these tests

1. Historical origins of Fisher's Exact Test

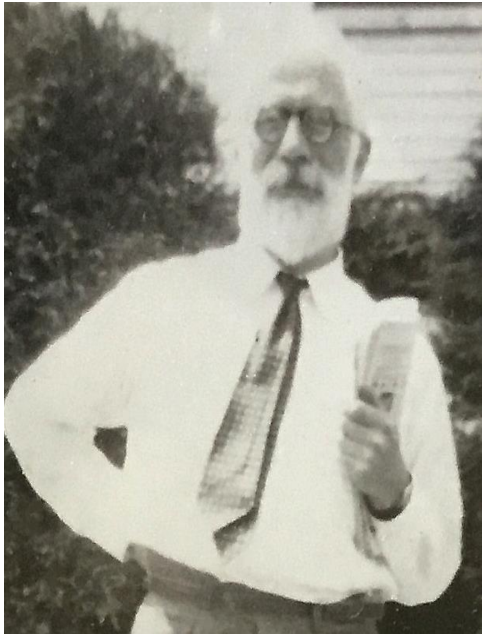


Figure 1. Ronald A. Fisher

The lady tasting tea, tea plus milk

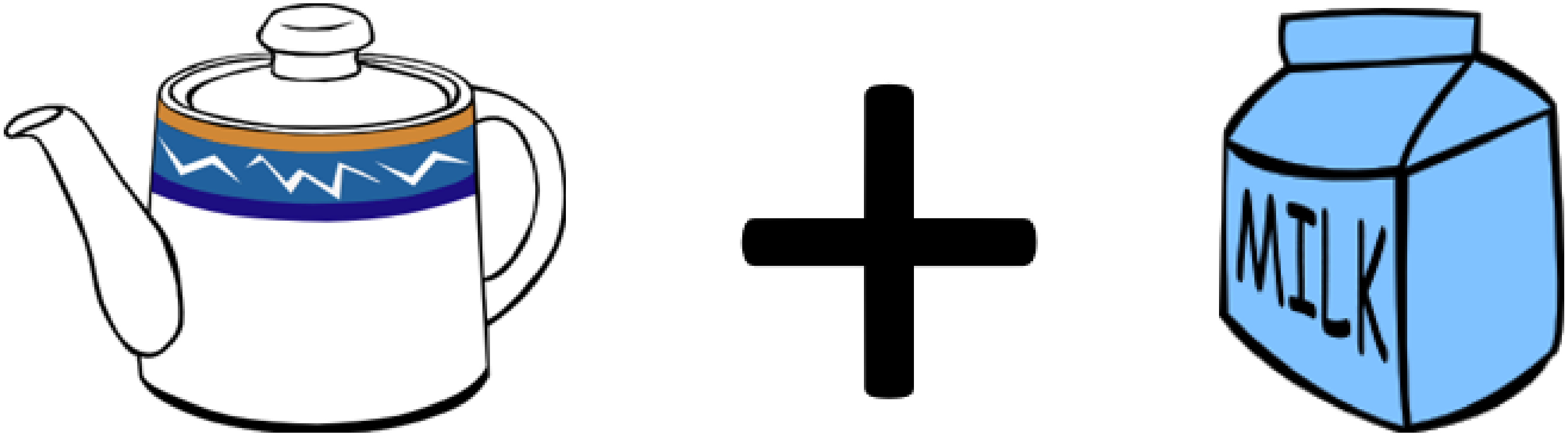


Figure 2. Tea with milk added

Milk plus tea, can you tell the difference?

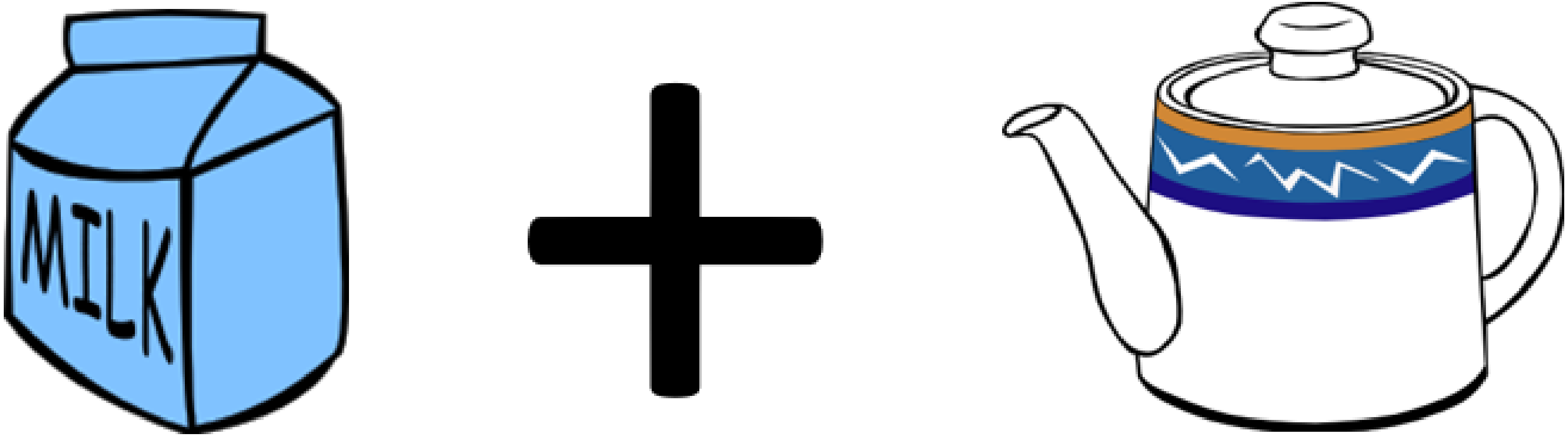


Figure 3. Milk with tea added

The experiment to test the claim

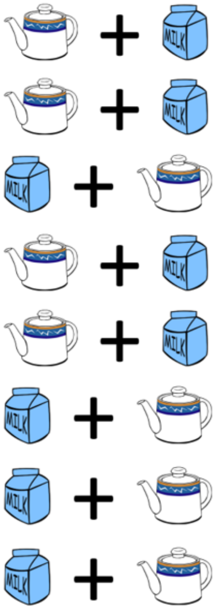


Figure 4. A randomized experiment

The result of the experiment

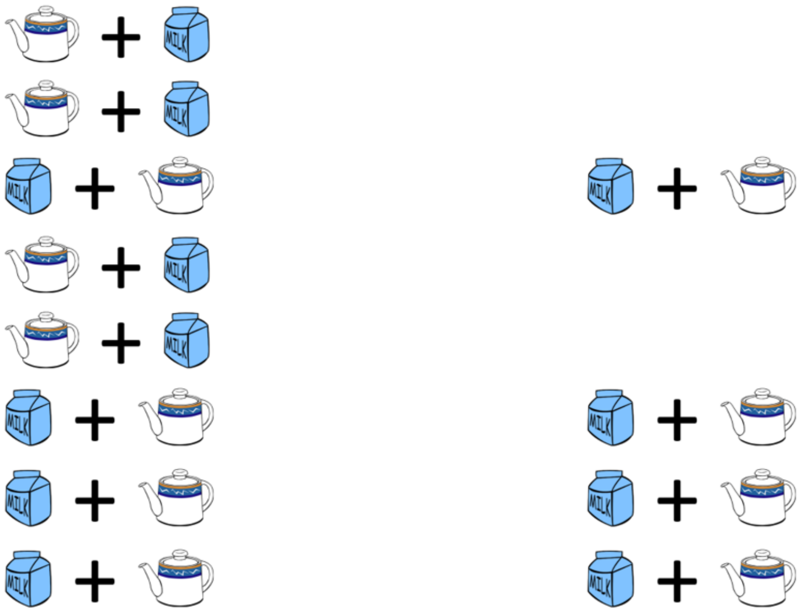


Figure 5. Result of the randomized experiment

How likely is this result?

$$\frac{4}{8} \times \frac{3}{7} \times \frac{2}{6} \times \frac{1}{5} = \frac{1}{70}$$

Note: the probability is NOT $\left(\frac{1}{2}\right)^4$

Break #1

What have you learned?

- Simple application of Fisher's Exact Test

What is coming next?

- The hypergeometric distribution

Any questions?

An alternate result

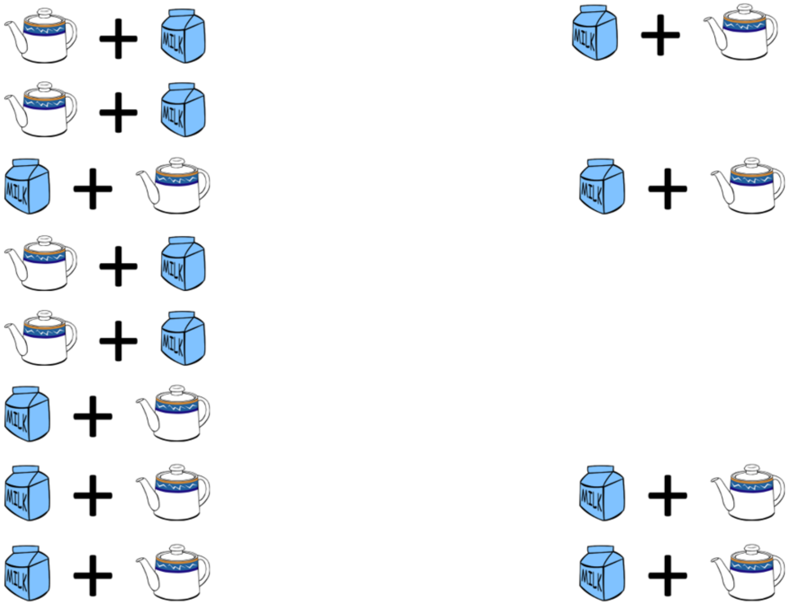


Figure 6. An alternate result with one miss

How likely is three correct results?

$$\frac{4}{8}\times\frac{4}{7}\times\frac{3}{6}\times\frac{2}{5} \;+\; \frac{4}{8}\times\frac{4}{7}\times\frac{3}{6}\times\frac{2}{5} \;+$$

$$\frac{4}{8}\times\frac{3}{7}\times\frac{4}{6}\times\frac{2}{5} \;+\; \frac{4}{8}\times\frac{3}{7}\times\frac{2}{6}\times\frac{4}{5}$$

Too messy! Use the hypergeometric distribution. Note: this is NOT a binomial distribution.

Balls in an urn analogy

```
|    W           |
|       B        |
|  B        B    |
|          W     |
|     W       B  |
|        W       |
__________________
```

Combinatorics

Combinatorics = mathematics of defining how many ways you can combine things.

$$\binom{a}{b} = \frac{a!}{b!\,(a-b)!}$$

Example:

$$\binom{4}{3} = \frac{4!}{3!\,(4-3)!} = \frac{24}{6 \times 1} = 4$$

Formula for hypergeometric probabilities

$$\frac{\binom{w_1}{w_0}\binom{b_1}{b_0}}{\binom{n_1}{n_0}}$$

w_1 = # of white balls in the urn

b_1 = # of black balls in the urn

$n_1 = w_1 + b_1$ = total # of balls in the urn

w_0 = # white balls drawn from the urn

b_0 = # black balls drawn from the urn

$n_0 = w_0 + b_0$ = total # of balls drawn

Calculation for 3 correct guesses

$$\frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{\frac{4!}{3!\,1!} \times \frac{4!}{1!\,3!}}{\frac{8!}{4!\,4!}} =$$

$$\frac{\frac{24}{6 \times 1} \times \frac{24}{1 \times 6}}{\frac{40320}{24 \times 24}} = \frac{16}{70}$$

Functions for computing hypergeometric probabilities

SAS: PDF('HYPER', w0, n1, w1, n0)

R: dhyper(w0, w1, b1, n0)

Stata: dis hypergeometricp(n1, w1, n0, w0)

SPSS: PDF.HYPER(w0, n1, w1, n0)

Break #2

What have you learned?

- The hypergeometric distribution

What is coming next?

- Using SPSS and Stata

Any questions?

SPSS data for Fisher's Exact Test

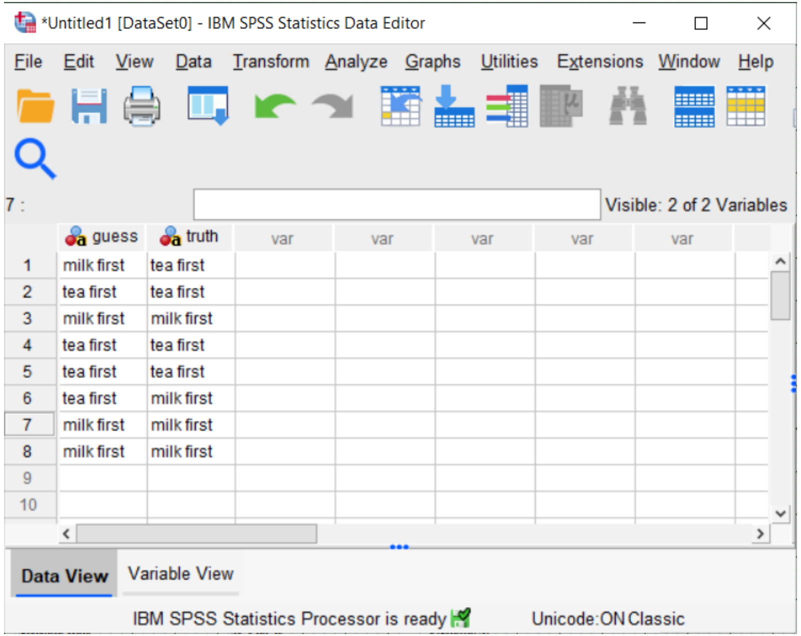


Figure 7. SPSS Dialog box

SPSS dialog boxes for Fisher's Exact Test (1/2)

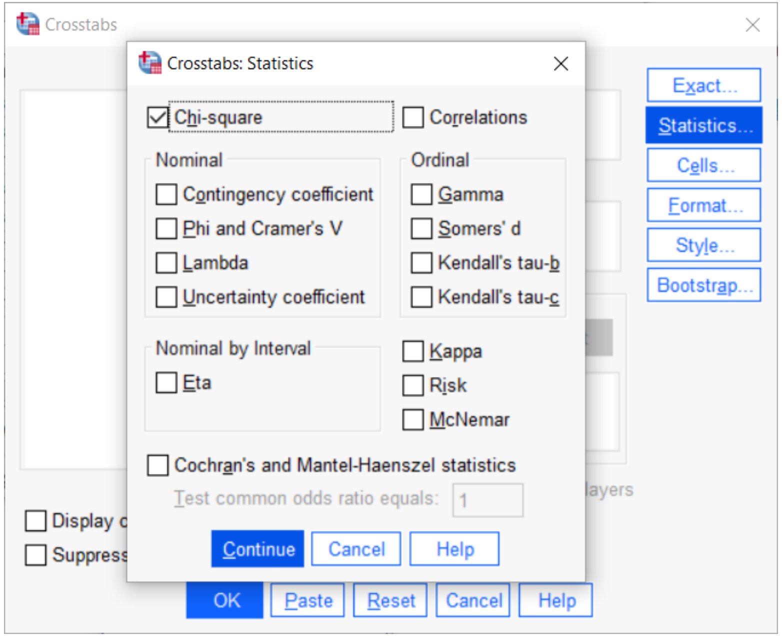


Figure 8. SPSS Dialog box

SPSS dialog boxes for Fisher's Exact Test (2/2)

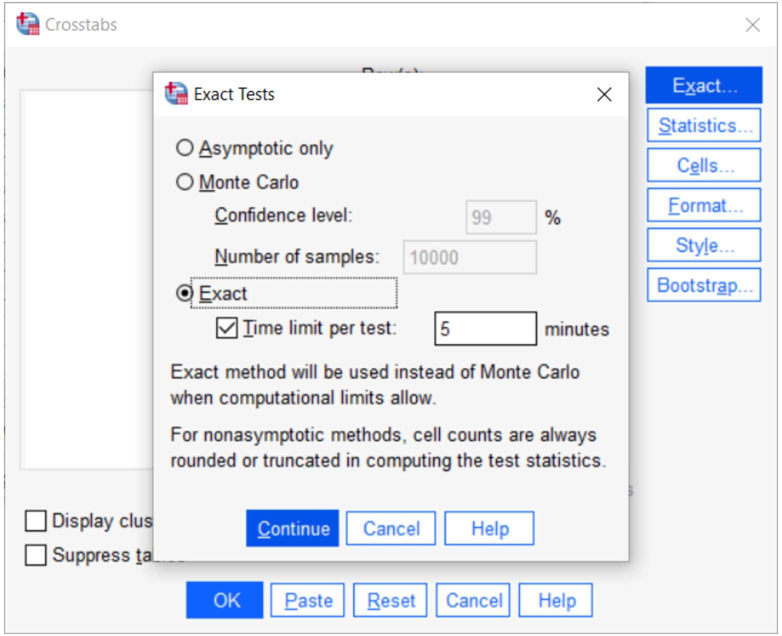
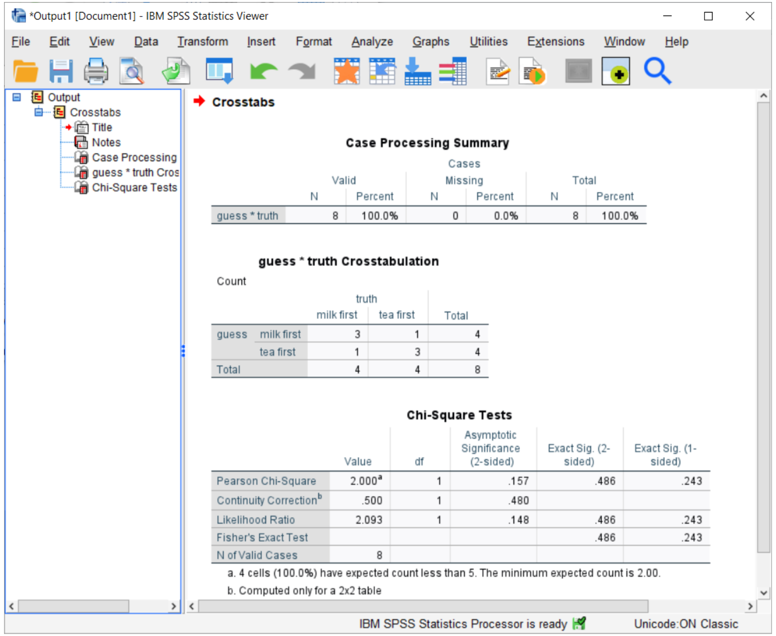


Figure 9. SPSS Dialog box

SPSS output for Fisher's Exact Test



Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
guess * truth	8	100.0%	0	0.0%	8	100.0%

guess * truth Crosstabulation

Count

		truth		
		milk first	tea first	Total
guess	milk first	3	1	4
	tea first	1	3	4
Total		4	4	8

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.000[a]	1	.157	.486	.243
Continuity Correction[b]	.500	1	.480		
Likelihood Ratio	2.093	1	.148	.486	.243
Fisher's Exact Test				.486	.243
N of Valid Cases	8				

a. 4 cells (100.0%) have expected count less than 5. The minimum expected count is 2.00.

b. Computed only for a 2x2 table

Figure 10. SPSS output box

Stata data for Fisher's Exact Test

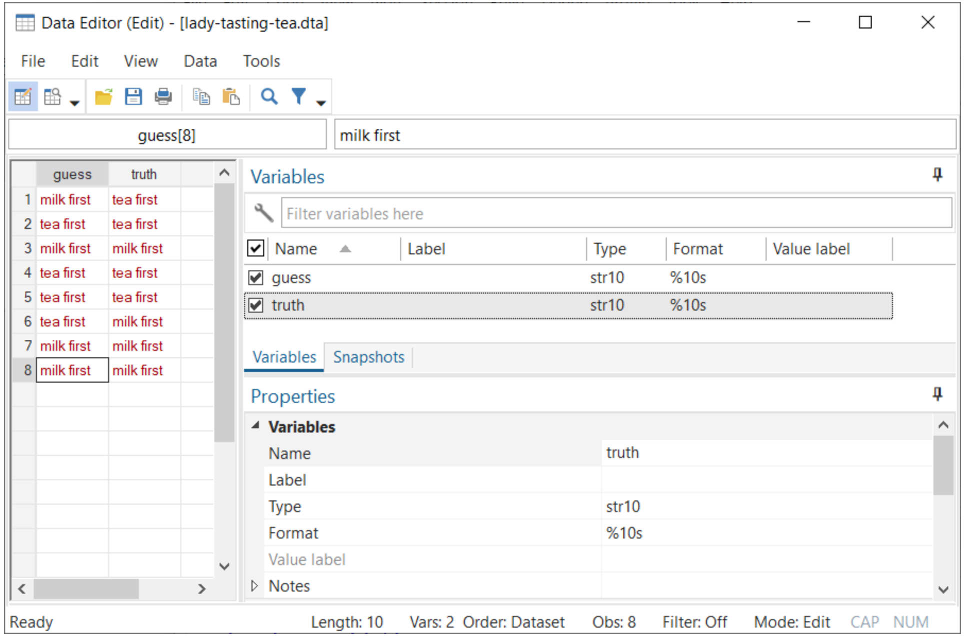


Figure 11. Stata data

Stata code and output for Fisher's Exact Test

```
. tabulate guess truth, exact

           Fisher's exact = 0.486
   1-sided Fisher's exact = 0.243
```

SAS and R code for Fisher's Exact Test

In SAS,

```
proc freq;
  tables guess*truth / fisher;
run;
```

In R,

```
fisher.test(guess, truth)
```

Break #3

What have you learned?

- Using SPSS and Stata

What is coming next?

- Details on the p-value computation

Any questions?

Recall the definition of a p-value

p-value = P[sample results or more extreme| H0]

What does “more extreme” mean?

List all possible 2 by 2 tables

Restricted to common marginal totals (fixed row and column totals)

```
? ? | 4
? ? | 4
----+--
4 4 | 8
```

There are five tables with the same marginal totals

4	0	3	1	2	2	1	3	0	4
0	4	1	3	2	2	3	1	4	0

1/70	16/70	36/70	16/70	1/70
0.014	0.229	0.514	0.229	0.014

Consider only tables that are more extreme

4	0		3	1
0	4		1	3

1/70 + 16/70

0.014 + 0.229

p-value = 17/70 = 0.243 (for a one sided test)

More extreme tables for a two-sided test

4	0	3	1	1	3	0	4
0	4	1	3	3	1	4	0

1/70 + 16/70 + 16/70 + 1/70

0.014 + 0.229 + 0.229 + 0.014

p-value = 34/70 = 0.486 (for a two-sided test)

Computing a two-sided p-value for the asymmetric case

4	0	3	1	2	2	1	3
0	3	1	2	2	1	3	0

1/35	12/35	18/35	4/35
0.029	0.343	0.514	0.114

p-value for 3 correct is 0.343 + 0.029 + 0.114 = 0.486.

Break #4

What have you learned?

- Details on the p-value computation

What is coming next?

- More exact tests

Any questions?

2. Other exact tests

Fisher-Freeman-Halton test

- Generalization of Fisher's Exact Test
- Tabulate all possible R by C tables
  - Fixed row and column totals

R code for Fisher-Freeman-Halton test

```
> v <- c(4, 0, 0, 0, 4, 0, 0, 0, 4)
> m <- matrix(v, nrow=3)
> m
     [,1] [,2] [,3]
[1,]    4    0    0
[2,]    0    4    0
[3,]    0    0    4
```

R output for Fisher-Freeman-Halton test in R

```
> fisher.test(m)

	Fisher's Exact Test for Count Data

data:  m
p-value = 0.0001732
alternative hypothesis: two.sided
```

Code for SAS, Stata, SPSS

SAS: Same as for a 2 by 2 table.

Stata: Same as for a 2 by 2 table.

SPSS: Same as for a 2 by 2 table.

Mann-Whitney U

Hypothetical data

```
T: 14, 23, 37
C: 12, 13, 15, 25
```

Rank the data

```
T:   3,   5,   7
C:   1,   2,   4,   6
```

Sum of the ranks

```
T = 15
C = 13
```

How likely is this result under the null hypothesis?

List all possible ranking for T

1,2,3	1,2,4	1,2,5	1,2,6	1,2,7
1,3,4	1,3,5	1,3,6	1,3,7	1,4,5
1,4,6	1,4,7	1,5,6	1,5,7	1,6,7
2,3,4	2,3,5	2,3,6	2,3,7	2,4,5
2,4,6	2,4,7	2,5,6	2,5,7	2,6,7
3,4,5	3,4,6	3,4,7	3,5,6	3,5,7
3,6,7	4,5,6	4,5,7	4,6,7	5,6,7

Select as extreme or more extreme rankings

				2,6,7
				3,5,7
3,6,7	4,5,6	4,5,7	4,6,7	5,6,7

p-value = 7/35 = 0.20

Rankings for a two-sided test

1,2,3	1,2,4	1,2,5	1,2,6	
1,3,4	1,3,5			
2,3,4				
				2,6,7
				3,5,7
3,6,7	4,5,6	4,5,7	4,6,7	5,6,7

p-value = 14/35 = 0.40

SAS code for Mann-Whitney test in SAS

```
proc npar1way wilcoxon;
  class grp;
  var x;
  exact wilcoxon;
run;
```

SAS output for Mann-Whitney test (1/2)

The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable x Classified by Variable grp					
grp	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
T	3	15.0	12.0	2.828427	5.000
C	4	13.0	16.0	2.828427	3.250

Figure 12. SAS output

SAS output for Mann-Whitney test (2/2)

Wilcoxon Two-Sample Test	
Statistic (S)	15.0000
Normal Approximation	
Z	0.8839
One-Sided Pr > Z	0.1884
Two-Sided Pr > |Z|	0.3768
t Approximation	
One-Sided Pr > Z	0.2054
Two-Sided Pr > |Z|	0.4108
Exact Test	
One-Sided Pr >= S	0.2000
Two-Sided Pr >= |S - Mean|	0.4000
Z includes a continuity correction of 0.5.	

Figure 13. SAS output

R, Stata, and SPSS

R: `wilcox.test`

Stata: `ranksum`

SPSS: `Analyze, Nonparametric tests, Independent Samples`

Mechanics for additional exact tests

- General algorithm
  - Assume a null hypothesis
  - List all possible outcomes
  - Find probabilities for each
  - Add up as extreme or more extreme probabilities

- Exact tests have very few assumptions
  - Usually only independence

- StatXact software

Break #5

What have you learned

- More exact tests

What is coming next

- Randomization tests

Any questions?

3. Randomization tests

Randomization tests

- Impractical to list all possible outcomes
- Randomly sample instead

Titanic data

	Alive	Dead	Total
Female	308 (67%)	154 (33%)	462
Male	142 (17%)	709 (83%)	851
Total	450 (34%)	863 (66%)	1,313

Average age

Alive	29.4
Dead	31.1
Overall	30.4

R code for the Titanic data

```
prop_male_survivors <- rep(NA, 10000)
avg_age_survivors <- rep(NA, 10000)
for (i in 1:10000) {
  prop_male_survivors[i] <-
    sum(sample(t$Sex, 450)=="male")/851
  avg_age_survivors[i] <-
    mean(sample(t$Age, 450), na.rm=TRUE)
}
```

Randomization results for proportion of male survivors

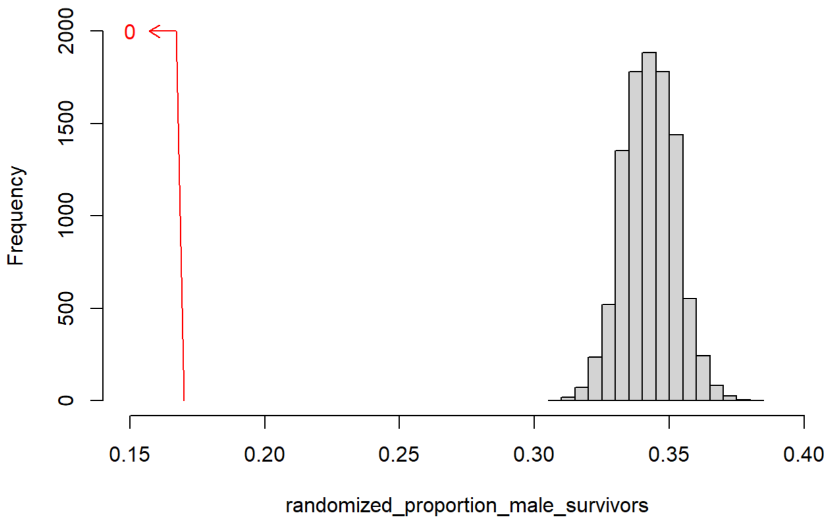


Figure x. Histogram of randomized counts of male survivors

Randomization results for average age of survivors

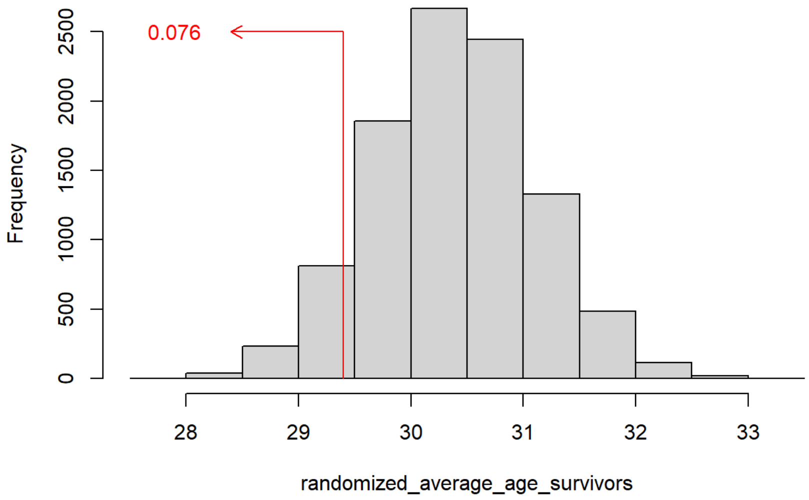


Figure x. Histogram of randomized average ages of survivors

Break #6

What have you learned

- Randomization tests

What is coming next

- A practical example

A practical example

```
Therapy
  Old: -1 -1 -1  0  0  0  0  0
  New:  0  1  1  1  2  2  2  2  2  2  3  3

-1 = slight decline
 0 = no change
 1 = slight improvement
 2 = moderate improvement
 3 = large improvement
```

A practical example

```
Average
   Old therapy: -0.38
   New therapy:  1.75
   All patients: 0.84
   Difference:   2.13
```

Randomize

2	-1	0	2	3	1	-1	1	Old mean =	0.88
1	0	3	2	2	-1	3	0	Old mean =	1.25
0	1	3	2	3	0	0	2	Old mean =	1.38
0	-1	0	3	1	0	2	2	Old mean =	0.88
-1	-1	2	3	0	1	0	1	Old mean =	0.62
1	0	0	-1	0	2	2	2	Old mean =	0.75
2	0	-1	0	3	1	0	2	Old mean =	0.88
1	1	1	3	2	2	0	0	Old mean =	1.25
1	-1	2	3	0	-1	2	2	Old mean =	1

Programming a randomization test

Can you get it directly (without programming)?

If not,

R:	`for` and `sample`.
SAS:	`do` and `ranperm` in IML.
Stata:	`ritest`.
SPSS:	not recommended (maybe with Python add-on?)

Break #7

What have you learned.

- A practical example of the randomization test

What is coming next

- When should you use exact and randomization tests?

Questions?

4. When should you use these tests

When should you use exact or randomization tests?

Fisher's Exact Test

- Lots of guidance

Other exact or randomization tests

- Not so much guidance

When should you use Fisher's Exact Test?

Your alternative is the Pearson Chi-squared test

$$T = \Sigma \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Under H_0, T is approximately $\chi^2(1)$

- Poor approximation if any E_{ij} < 5

Criticisms of Fisher's Exact Test

1. Too conservative

2. Fixed row and column totals are unrealistic.

When should you use other exact/randomization tests?

- Concern about small sample sizes
- Concern about distributional assumptions
- As a safety/sensitivity check

Criticisms of other exact/randomization tests?

- No easy way to get confidence intervals
- No easy extension to more complex settings
  - Risk adjustment
  - Longitudinal/hierarchical models
- Sometimes too computationally difficult
  - Inadequate computer speed and capacity
  - Difficulty in programming

Exact versus randomization tests

Use exact test if

- it is pre-programmed

And

- sample size is small or moderate

Use randomization test if

- you have to program it yourself

Or

- sample size is large

Randomization tests versus bootstrap

Bootstrap

= repeated sample WITH replacement

Advantages

- Very easy confidence intervals
- Applicable to descriptive statistics

Randomization

= repeated samples WITHOUT replacement.

Advantages

- Simplicity

Neither extends easily to complex settings.

Conclusion

What have you learned?

- Fisher's Exact Test (the lady tasting tea)
- Fisher-Freeman-Halton test
- Mann-Whitney test
- Randomization tests
  - Titanic data
  - Practical example
- When to use/not use exact and randomization tests

Questions?