

# Permutation and Randomization Tests<sup>1</sup>

Jerry Brunner, Department of Statistical Sciences

---

<sup>1</sup>This document is free and open source. See last slide for copyright information.

# Overview

- 1 Permutation Tests
- 2 Randomization Tests

# The lady and the tea

From Fisher's *The design of experiments*, first published in 1935

Once upon a time, there was a British lady who claimed that she could tell from the taste which had been poured into the cup first, the tea or the milk. So Fisher designed an experiment to test it.

- Eight cups of tea were prepared.
- In four, the tea was poured first.
- In the other four, the milk was poured first.
- Other features of the cups of tea (size, temperature, etc.) were held constant.
- Cups were presented in a random order (critical).
- The lady tasted them, and judged.
- She knew there were four of each type.

# The null hypothesis

- The null hypothesis is that the lady has no ability to taste the difference.
- If so, all possible ways of lining up the lady's judgements and the truth about the tea cups should be equally likely.
- Equally likely *because of the random order of presentation*.
- The test statistic is the number of correct judgements.
- What is the distribution of the test statistic under the null hypothesis?

# Data file

	Truth	Judgement
1	tea	milk
2	milk	tea
3	milk	milk
4	milk	milk
5	tea	tea
6	tea	tea
7	tea	milk
8	milk	tea

- Under  $H_0$ , the reasons for the lady's judgements are unknown, except that they have nothing to do with the truth.
- The judgements are what they are; they are fixed.
- Because of randomization, all  $8! = 40,320$  permutations of the cups are equally likely, and each one has its own number of correct judgements.
- But there are lots of repeats.

# Counting argument

- How many ways are there to choose 4 cups to put the tea in first?  
 $\frac{8!}{4!4!} = 70$
- All are equally likely.
- Only one lines up perfectly with the truth.
- The chances of the lady guessing right under  $H_0$  are  
 $\frac{1}{70} = 0.0143 < 0.05$ .
- So  $H_0$  would be rejected at  $\alpha = 0.05$  if she guessed perfectly.

# The permutation distribution

In general

- Decide on a test statistic  $T$ .
- List the possible values of  $T$ .
- Under  $H_0$ , all ways of re-arranging the data are equally likely.
- $P(T = t)$  is proportional to the number of ways of getting the value  $t$ .

The permutation  $p$ -value is the probability of getting a value of  $T$  as extreme or more extreme as the value we observed, “extreme” meaning in a direction inconsistent with  $H_0$ . Alternatively, locate critical value(s). The probability of exceeding these values is  $\leq 0.05$ .

# Fisher's exact test

$2 \times 2$  table, binary response

- As in tea testing example, testing association of two binary variables.
- This time, no requirement of 50-50 split.
- Fisher figured out the probabilities in advance using the hypergeometric distribution.
- Large samples are not required.
- In fact, no sampling at all is required, if there is random assignment.
- Random sampling but no random assignment is okay too.



# Permutation tests are not just for categorical data

Another example from Fisher's *The design of experiments*

Darwin's experiment on self-fertilized versus cross-fertilized corn plants:

- Plants are grown in 15 pairs, one cross-fertilized and one self-fertilized.
- Dependent variable is height.

## Darwin's Data

Crossed.	Self-fert.
Inches.	Inches.
$23\frac{1}{8}$	$17\frac{1}{8}$
12	$20\frac{1}{8}$
21	20
22	20
$19\frac{1}{8}$	$18\frac{1}{8}$
$21\frac{1}{8}$	$18\frac{1}{8}$
$22\frac{1}{8}$	$18\frac{1}{8}$
$20\frac{1}{8}$	$15\frac{1}{8}$
$18\frac{1}{8}$	$16\frac{1}{8}$
$21\frac{1}{8}$	18
$23\frac{1}{8}$	$16\frac{1}{8}$
21	18
$22\frac{1}{8}$	$12\frac{6}{8}$
23	$15\frac{4}{8}$
12	18

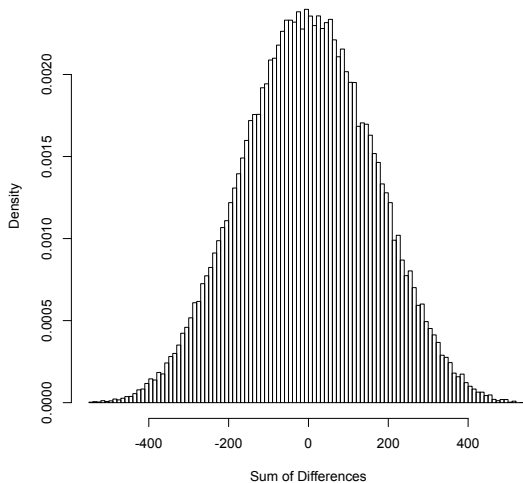
# How to analyze the data?

- Calculate differences.
- Do a matched  $t$ -test, or ...

# A permutation test for matched pairs

- Fisher wishes the self-fertilized plants had been randomly assigned to be on either the left or the right. Otherwise he loves the experiment.
- Under null hypothesis that self-fertilized versus cross-fertilized does not matter at all, only chance determined whether  $A$  was subtracted from  $B$  or  $B$  was subtracted from  $A$ .
- Each pair of numbers is what it is, but only chance determined which one was called “self-fertilized” (under the null hypothesis).
- Test statistic is sum of the differences, in eights of an inch.
- There are  $2^{15} = 32,768$  ways to swap the pairs of heights, all equally likely under  $H_0$ .
- Calculate the sum of differences for each one, yielding a permutation distribution for the test statistic under  $H_0$ .
- The  $p$ -value is the proportion of these that equal or exceed in absolute value the sum of differences Darwin observed:  $D = 314$ .
- Fisher's answer is  $p = 0.05267$ , compared to  $p = 0.0497$  from a  $t$ -test.
- He used his brain as well as doing a lot of tedious calculation.

## Permutation Distribution for Darwin's Plant Data



Observed sum of differences = 314 in eights of an inch:  $p = 0.05267$

## Some big advantages of the permutation test idea

- Test is distribution-free under the null hypothesis. There is no assumption of the normal or any other distribution.
- Some non-parametric methods depend on large sample sizes for their validity. Permutation tests do not. Even for tiny samples, the chance of false significance cannot exceed 0.05.
- $p$ -values are exact and not asymptotic.
- There is no pretense of random sampling from some imaginary population.
- All the probability comes from random assignment.
- Can easily be extended to tests comparing several treatments.
- For repeated measures, re-arrange within subjects, as with Darwin's plants.

## More comments

- For observational studies too, the null hypothesis is that the explanatory variable(s) and response variable(s) are independent.
- It's even better than that. Bell and Doksum (1967) proved that *any* valid distribution-free test of independence *must* be a permutation test (maybe a permutation test in disguise).
- It doesn't matter if data are categorical or quantitative. By scrambling the data, any possible relationship between explanatory and response variables is destroyed.
- If either explanatory or response variable is multivariate, scramble *vectors* of data.

# What is “the” test statistic?

- It's up to you.
- No matter what you choose, the chance of wrongly rejecting the null hypothesis cannot exceed  $\alpha = 0.05$ .
- The best test statistic is a descriptive statistic that accurately reflects the phenomenon as you understand it.
- Could that number (or greater) have been produced by random assignment or random sampling? No doubt.
- The question is, how likely is this?
- The answer is given by the permutation  $p$ -value.



# Choice of test statistic from a statistician's point of view

- We are testing a null hypothesis based on the value of the test statistic.
- The probability of wrongly rejecting  $H_0$  (and making a false discovery) is limited to  $\alpha = 0.05$ . Good.
- Some test statistics are better than others, depending on *how*  $H_0$  is false: Statistical power.
- See Good (1994) *Permutation tests*.
- Traditional test statistics are a popular choice, and often a good choice.
- When the assumptions happen to be approximately satisfied, they often are nearly optimal.
- But scientific meaningfulness is more important than power, in my opinion.

## To summarize

A permutation test is conducted by following these three steps.

- ① Compute some test statistic using the set of original observations.
- ② Re-arrange the observations in all possible orders, computing the test statistic each time. Re-arrangement corresponds exactly to the details of random assignment.
- ③ Calculate the permutation test  $p$ -value, which is the proportion of test statistic values from the re-arranged data that equal or exceed the value of the test statistic from the original data. Or, locate the critical value(s) in the permutation distribution.

Fisher said

*Statistical methods for research workers*, 1936

*Actually, the statistician does not carry out this very tedious process but his conclusions have no justification beyond the fact they could have been arrived at by this very elementary method.*

## Main drawback is that it's hard to compute

- Fisher considered permutation tests to be mostly hypothetical, but that was before computers.
- Even with computers, listing all the permutations can be out of the question, and combinatoric simplification may be challenging.
- One way around the computational problem is to convert the data to ranks, and then do it.
- Then, permutation distributions can be figured out in advance.
- All the common non-parametric rank tests are permutation tests carried out on ranks.

# Randomization tests: A modern solution

- Scramble the values of the response variable in a random order.
- Compute the test statistic for the randomly shuffled data.
- In this way, we have randomly sampled a value of the test statistic from its permutation distribution.
- Carry out the procedure a large number of times.
- By the Law of Large Numbers, the the permutation  $p$ -value is approximated by the proportion of randomly generated values that exceed or equal the observed value of the test statistic.
- This proportion is the  $p$ -value of the randomization test.

# There is some programming

An alternative for simple problems is the `coin` package

- 
- First calculate the test statistic on the original, unscrambled data.
- Set up an array to hold the randomized test statistic values.
- Loop through the array with the `for` function.
  - Use the `sample` function to scramble the data in a way that mimics the random assignment.
  - Calculate the test statistic on the scrambled data.
  - Store the computed test statistic in the array.
- Calculate the randomization  $p$ -value or locate the critical value.
- Output.

# The `sample` function

Random permutations the easy way

```
> letters = c("A","B","C","D")
```

```
> sample(letters)
```

```
[1] "A" "D" "B" "C"
```

```
> sample(letters)
```

```
[1] "B" "A" "C" "D"
```

```
> indices = sample(1:4); indices
```

```
[1] 4 3 2 1
```

```
> letters[indices]
```

```
[1] "D" "C" "B" "A"
```

# Copyright Information

This slide show was prepared by **Jerry Brunner**, Department of Statistical Sciences, University of Toronto. It is licensed under a **Creative Commons Attribution - ShareAlike 3.0 Unported License**. Use any part of it as you like and share the result freely. The  $\text{\LaTeX}$  source code is available from

<http://www.utstat.toronto.edu/~brunner/workshops/randomization>