Permutation and Randomization Tests[1]

[1] This slide show is an open-source document. See last slide for copyright information.

Overview

1. Introduction
2. Permutation Tests
3. Randomization Tests
4. Multiple Comparisons
5. Bootstrap

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? $\binom{8}{4} = 70$
- All are equally likely.
- Only one lines up perfectly with the lady's judgements.
- The chances of this 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.

Fisher's exact test

- Testing association of two binary variables.
- Unlike the tea-teasing example, no requirement of 50-50 split.
- Numbers of A = Yes, No and B = Yes, No are fixed.
- Subject to those restrictions, the count in one cell is free to vary ($df = 1$).
- Number in the (Yes, Yes) cell is one-to-one with the odds ratio.
- If counts in the (Yes, Yes) cell were completely random subject to the restriction of row and column totals (that's H_0), what's the probability of getting such a large (or small) oddes ratio?
- p-values are exact probabilities based on the hypergeometric distribution.
- Large samples are not required.

The idea

Thank you Mr. Fisher

- Experimental study, with random assignment of units to conditions.
- Under H_0, the treatment has no effect at all.
- The process producing values of y is unspecified.
- Except that it has nothing to do with experimental condition.
- The particular values of the response variable are what they are.
- The only reason for differences among conditions is the random assignment.

The permutation distribution

- Pick a test statistic (more on that later).
- Under H_0, all the ways of distributing y values into experimental conditions are equally likely.
- Each re-arrangement (permutation) of the y values produces a value of the test statistic.
- Compute the test statistic for each re-arrangement.
- Relative frequencies are the *permutation distribution* of the test statistic.
- Make a histogram.

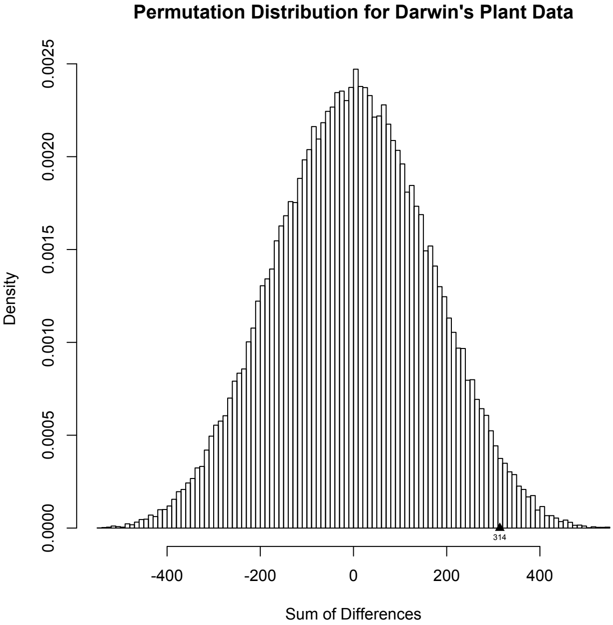
Another example from Fisher's *Design of experiments*

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

- Plants are grown in 15 pairs, one cross and one self-fertilized.
- Response variable is height.
- Calculate differences.
- Do a matched t-test, or …

A randomization 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.
- So the absolute value of the difference is what it is, but the plus or minus sign is by chance alone (under H_0).
- Test statistic is sum of the differences.
- There are $2^{15} = 32,768$ ways to swap the plus and minus signs, 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
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
314
-400
-200
0
200
400
Sum of Differences

Permutation p-value

The permutation test p-value is the proportion of values in the permutation distribution that equal or exceed the observed value of the test statistic from the un-scrambled data — in the direction(s) of the alternative hypothesis.

Advantages of the permutation test idea

- Simplicity. The distribution theory is elementary.
- 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.
- Random assignment actually happens. Random sampling often does not.

More comments

- Applies to 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 $\alpha = 0.05$.
- One good choice 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 unlikely is this?
- The answer is given by the permutation p-value.

Choice of test statistic

- 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 usually a good choice.
- When the assumptions happen to be approximately satisfied, they often are nearly optimal.

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.

Using the p-value from a traditional test

As the test statistic

- The p-value from a traditional test is sometimes a more convenient test statistic than the original test statistic.
- p-value is $1-1$ function of the test statistic, so the permutation p-value is the same.
- The permutation p-value is the proportion of p-values from the scrambled data that are less than or equal to the observed p-value.
- That's exactly the cdf of the permutation distribution of p-values.
- One-sided, two-sided does not matter.
- Handy for multiple comparisons (More later).

Fisher said

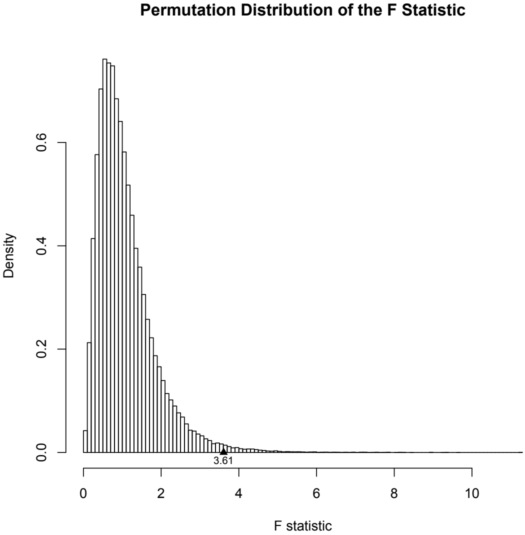
Statistical methods for research workers, 1936

> *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.*

See Cox and Reid (2000) *The Theory of the Design of Experiments* for the research literature.

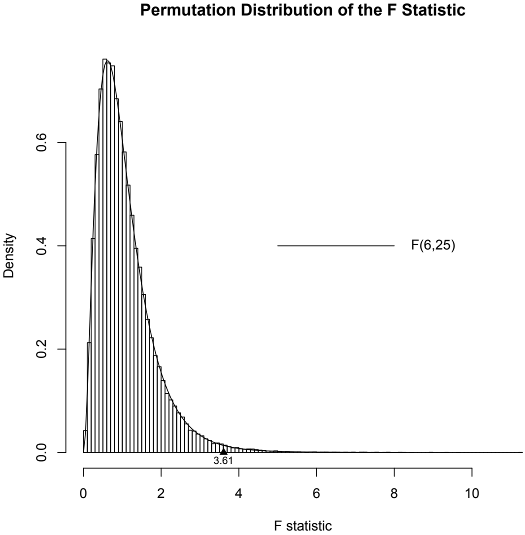
Scab disease data

Illustrating Fisher's claim



Scab disease data

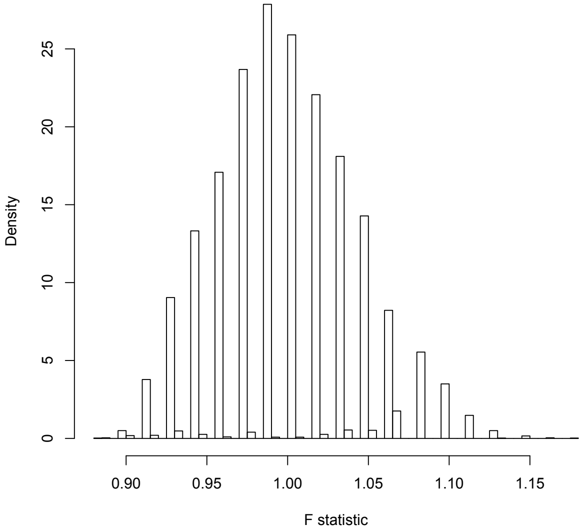
Illustrating Fisher's claim

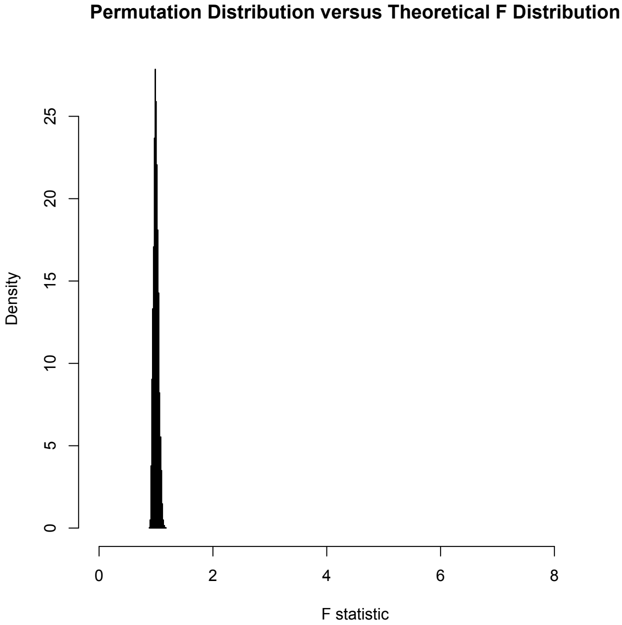


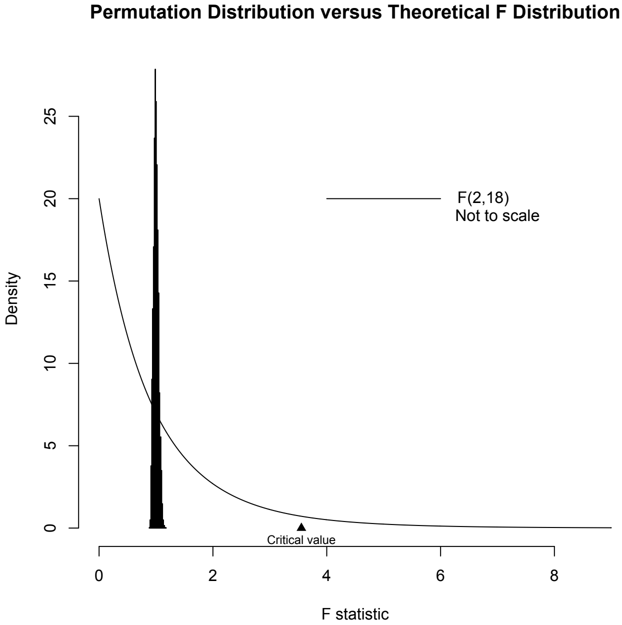
The approximation is not always so good

Group1	Group2	Group3
220	1	4
0	0	0
1	2	0
0	4	3
1	2	1
1	0	4
0	1	1

Permutation Distribution of the F Statistic

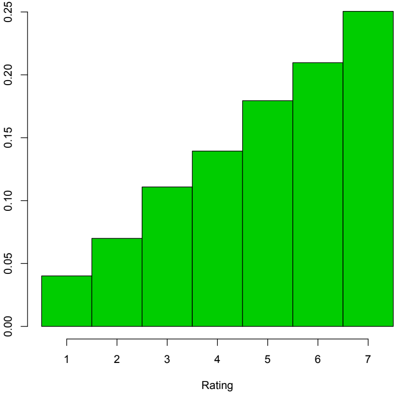



Permutation Distribution versus Theoretical F Distribution
Density
0
5
10
15
20
25
0
2
4
6
8
F statistic


Permutation Distribution versus Theoretical F Distribution
Density
0
5
10
15
20
25
F(2,18)
Not to scale
Critical value
0
2
4
6
8
F statistic

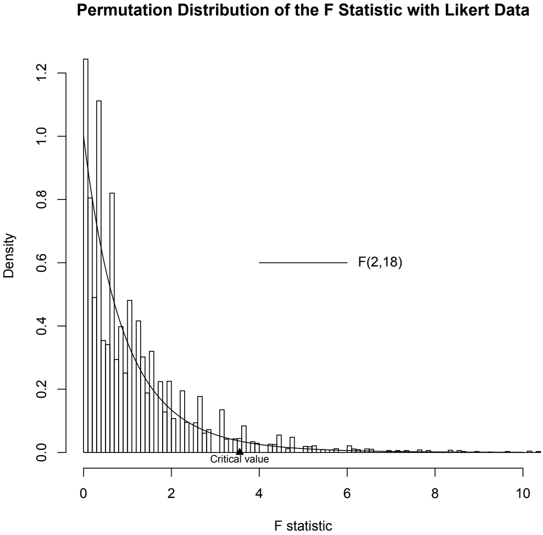
It was the outlier

- Approximation was excellent for exponential data.
- Awful for absolute Cauchy not rounded.
- Likert scale: 7-point scale, strongly disagree to strongly agree.



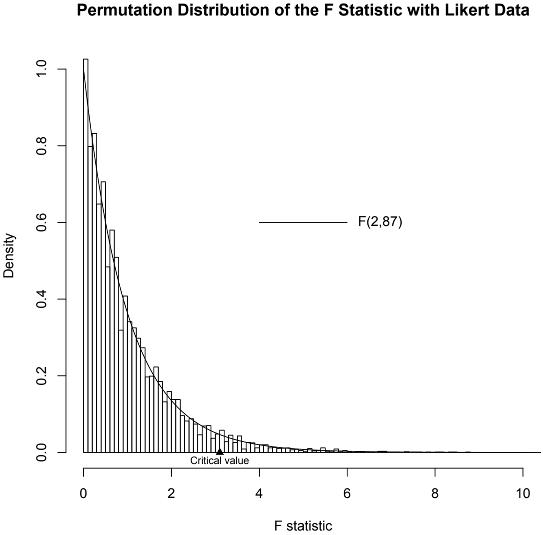
$n = 7$ for each of three treatments

0.0542 of the permutation distribution is above the F critical vlue



$n = 30$ for each of three treatments

0.0472 of the permutation distribution is above the F critical vlue



Main drawback of permutation tests are that they're 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.

Scab disease

- Eight plots of land in the control condition.
- Four plots in each of 6 experimental conditions.
- Total $n = 32$.
- This is a small sample.
- There are $\frac{32!}{8!\,4!\,4!\,4!\,4!\,4!\,4!}$ ways to place the observed data into treatment conditions.
- Calculate the F statistic for each one.
- SAS will do it. Or anyway, it will try.

Calculate the F statistic for each re-arrangement of the data

$$\frac{32!}{8!\,4!\,4!\,4!\,4!\,4!\,4!} = 34,149,454,710,484,113,000,000$$

- That's a big number.
- Maybe we can distribute the computation among lots of computers.
- World population is approximately 7.51 billion.
- That's $34149454710484113/7510 \approx 4.547 \times 10^{12}$ calculations per person.
- If they all had computers and could do one test every 0.01 seconds,
- It would take around 1,441.9 years to finish the job.

Some problems can be figured out in advance

- If both explanatory and response variable are binary (an important case), Fisher derived the permutation distribution of the number of observations in the Yes, Yes cell (equivalent to the odds ratio) based on the hypergeometric distribution.
- The result is called Fisher's exact test.
- For non-binary response variables, one can convert the data to ranks.
- 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

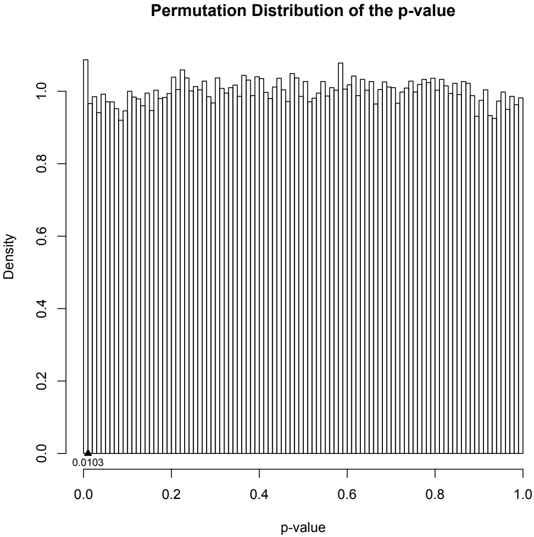
- Scramble the values of the response variable in a random order, leaving the explanatory variable values in place.
- Compute the test statistic for the randomly shuffled data.
- We have randomly sampled a value of the test statistic from its permutation distribution.
- Carry out the procedure a large number of times.
- By the Law of Large Numbers, the 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.
- The p-value of a randomization test is an *estimate* of the p-value of the corresponding permutation test.
- SAS does this (among other options) in `proc npar1way`.
- With a confidence interval for the permutation test p-value.

Multiple Comparisons

Using randomization

- You could Bonferroni protect a collection of randomization tests, but this is better.
- It's *not* conservative.
- You do need to know what all the tests are, in advance.

Use a standard p-value as the test statistic



p-values

- Have a family of tests, and an observed p-value from each one.
- The event that at least one p-value is less than some critical value is the event that the *minimum* is less than the critical value.
- If the distribution of the test statistic is continuous and H_0 is true, p-values are uniformly distributed on $(0, 1)$.
- It's easy to derive the distribution of the minimum of a collection of independent uniforms.
- Except the p-values are not independent.

The randomization test solution

Approximate permutation distribution for a family of tests

- Randomly permute the data, scrambling y against x.
- Calculate p-values for all the tests and take the minimum.
- Repeat.
- The result is a randomization distribution of minimum p-values.
- This is an approximation of the corresponding permutation distribution.
- Compare each observed p-value to the distribution of the minimum.
- The proportion of minimum p-values at or below any given observed p-value is an *adjusted* p-value.
- If all null hypotheses are true, the probability of getting at least one adjusted p-value less than 0.05 equals 0.05.
- Give or take discreteness and Monte Carlo sampling error.
- `proc multtest` does this.

Bootstrap (Efron, 1979)

For comparison

- If the sample size is large enough, the histogram of the sample data is a lot like the histogram of the entire population.
- Thus, sampling from the sample *with replacement* is a lot like sampling from the population.
- Sampling from the sample is called **resampling**.

Bootstrap distribution

One can approximate the sampling distribution of a statistic as follows.

- Select a random sample of size n from the sample data, *with replacement*.
- Compute the statistic from the resampled data.
- Do this over and over again, accumulating the values of the statistic.
- A histogram of the values you have accumulated will resemble the sampling distribution of the statistic.
- Use it to construct tests and confidence intervals.

Bootstrap vs. Randomization tests

Similarities and differences

- Both are computer-intensive Monte Carlo methods based on random number generation.
- Neither requires any assumption about the distribution of the data.
- Both substitute computing power for probability theory.
- Bootstrap assumes random sampling and is justifiable only as $n \to \infty$, though it often seems to work well with moderate sample sizes.
- Randomization tests do *not* assume random sampling and are *exact* for small samples (almost).

Copyright Information