# Introduction to Experimental Design[1]

## STA305 Winter 2014

[1] See last slide for copyright information.

# Background Reading

## Optional

- Pages 21-26 in Chapter 1 of *Data analysis with SAS*: The correlation-causation issue.
- Chapter 11 of *Data analysis with SAS*: It uses R and is at a lower technical level, but gives the general idea of permutation tests.
- *Wikipedia* under *Resampling statistics* for more detail and references on permutation tests.

# Goal of the course

We want to know the effect of some *treatment* (or combination of treatments) on some *response*. The method should be

- Objective (scientific)
- As efficient as possible.

# Examples

| **Treatment** | **Response** |
|---|---|
| Advertising expenditures | Sales |
| Drug | Health |
| Industrial quality control process | Product quality |
| Fertilizer type | Crop yield |
| Sterilization method | Bacterial/viral load |

# We want to be objective

- Outcome will be measured numerically (including categories).
- Admit that the data will be somewhat noisy.
- Try it more than once (replication).
- Use statistical methods to decide if there was any effect, and if so how much.
- Combination of statistics and research design.
- The data collection should be planned with the statistical analysis in mind!

# Observational versus experimental studies

Definitions from Cox and Reid's (2000) *Theory of the design of experiments*.

- **Observational study**: Allocation of individuals to treatments is not under the control of the investigator.
- **Experimental study**: The system under study (including allocation of individuals to treatments) is mostly under the control of the investigator.

# Experimental units

- The smallest subset of experimental material to which a separate treatment might be applied.
- Usually people, rats, stores, test tubes, plots of land, grocery stores, etc.
- But if one kind of contact lens is put in the left eye and another kind in the right eye, the experimental unit would be the eye.

# Omitted variables

- In any study, some things that affect the response will not be part of the data set.
- At best, they contribute background noise that makes the effect of the treatment harder to see.
- At worst, they are also systematically related to the treatments, and can produce misleading results.
- Observational studies are particularly subject to this bias.

# Example

Hair length at age 25 and length of life

# Regression

- The usual conditional regression model assumes that any omitted variables are independent of the explanatory variables in the model.
- What happens when this is violated?

# Example of omitted variables in regression

Independently for $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. The mean and covariance matrix of the independent variables are given by

$$E\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad cov\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$$

But $X_{i,2}$ is not in the data set, so we just use $X_{i,1}$.

# What happens to $\widehat{\beta}_1$ as $n \to \infty$

When $Xi, 2$ is ignored

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_{i,1} - \overline{X}_1)(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_{i,1} - \overline{X}_1)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_{i,1} - \overline{X}_1)(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_{i,1} - \overline{X}_1)^2} \\
&\to \frac{Cov(X_{i,1}, Y_i)}{Var(X_{i,1})} \\
&= \beta_1 + \frac{\beta_2 \phi_{12}}{\phi_{11}}
\end{aligned}
$$

# Effects of omitted variables

## Illustrated by $\widehat{\beta}_1 \to \beta_1 + \frac{\beta_2 \phi_{12}}{\phi_{11}}$

An omitted variable that is associated with *both* the explanatory and the response variable is sometimes called a *confounding variable*. It can

- Produce an association between explanatory and response variable even when, conditionally on the omitted variable, they are independent
- Mask a real relationship between explanatory and response variable
- Reverse a relationship between explanatory and response variable.

# Include everything?

One possible solution is to include *all* relevant variables, and "control" for them somehow, maybe by regression or subdivision. But,

- You can't know what they all are.
- Data set will be huge and expensive to collect.
- Most variables will be measured with error anyway.

# The solution is an experiment

## Excellent for certain problems, impossible for others

Use control over the setting to break up the association between the treatment and potential confounding variables.

- Random assignment: Thank you Mr. Fisher.
- There are other ways, like alternate assigment to experimental and control treatments.
- The details of how it's done are important.

# Random assignment of experimental units to treatments

Say, to an experimental group versus control group

Use pseudo-random numbers from a computer, a table of random numbers or a physical game of chance to assign experimental units to treatments.

- Make sure there are *no* systematic differences in what happens to the different groups, other than the treatment of interest.
- If there is a systematic relationship between treatment and any omitted variable, it's purely due to chance.
- If the treatment has no effect, any difference in the response must also be due to chance.

# Completely randomized design

- There are $p$ treatments.
- Randomly assign experimental units to treatments .
- Assign $n_1$ to Treatment 1, ..., $n_p$ to treatment $p$.
- So that all assignments are equally likely.
- Could be done with a jar of marbles.
- Or a random permutation.

# Random permutation

## An easy way to do it

- Number the experimental units $1, \ldots, n$, where $n = \sum_{j=1}^{p} n_j$.
- Using software, generate $n$ (pseudo) random variables from a uniform distribution on $(0, 1)$.
- Sort the integers $1, \ldots, n$ by the random numbers.
- The randomly scrambled numbers $1, \ldots, n$ are a *random permutation*.
- Assign the first $n_1$ units to treatment one, the next $n_2$ units to treatment two, and so on.
- All such assignments are equally likely.

# Basis of the permutation test

For example, compare an experimental group to a control group.

- Calculate a test statistic that expresses *difference* in the response variable values for the different treatment groups. For example, $\overline{Y}_1 - \overline{Y}_2$.
- Under $H_0$, the treatment does nothing.
- So if the test statistic is big, it's just by the luck of random assignment.

# How the permutation test goes

- Consider the response variable values to be fixed constants.
- Under $H_0$, all allocations of these values to treatment groups are equally likely.
- For each allocation, there is a value of the test statistic.
- This induces a probability distribution of the test statistic under $H_0$.
- The $p$-value is the probability of obtaining a test statistic greater than or equal to the one from the actual experiment (perhaps in absolute value).

# A small example

## To illustrate the idea

- Have $n = 6$ automobiles to crash test.
- Randomly assign 3 to the experimental condition (new air bag system), and 3 to the existing system.
- Measure damage to the crash test dummy.
- Damage measurements are 1.3 6.0 3.0 for the new system, and 5.6 6.5 7.1 for the existing system.
- $\overline{Y}_1 = 3.43$, $\overline{Y}_2 = 6.4$, and $\overline{Y}_2 - \overline{Y}_1 = 2.97$.
- Is the new system better?
- We need a one-sided test here.

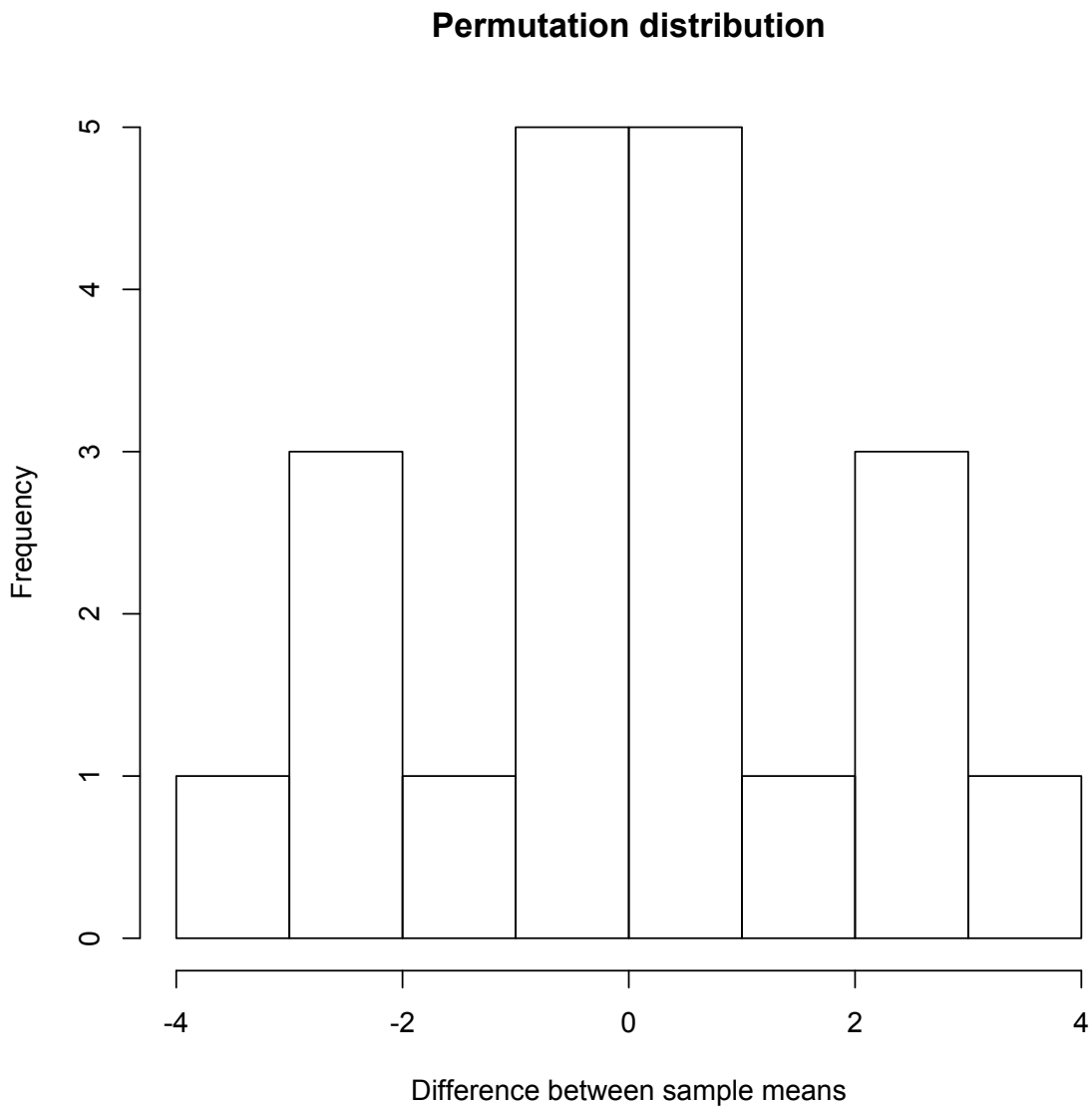
# Get the permutation distribution of the test statistic

- There are $\binom{6}{3} = 20$ ways to divide the observations into 2 groups.
- All equally likely under $H_0$
- List them, and calculate $\overline{Y}_2 - \overline{Y}_1$ for each one.

# $\binom{6}{3} = 20$. We don't need all $6! = 720$ permutations.

| | | diff |
|---|---|---|
| (1.3 6.0 7.1) | (3.0 5.6 6.5) | 0.2333333 |
| (1.3 3.0 7.1) | (5.6 6.0 6.5) | 2.2333333 |
| (1.3 5.6 7.1) | (3.0 6.0 6.5) | 0.5000000 |
| (1.3 6.5 7.1) | (3.0 5.6 6.0) | -0.1000000 |
| (3.0 6.0 7.1) | (1.3 5.6 6.5) | -0.9000000 |
| (5.6 6.0 7.1) | (1.3 3.0 6.5) | -2.6333333 |
| (6.0 6.5 7.1) | (1.3 3.0 5.6) | -3.2333333 |
| (3.0 5.6 7.1) | (1.3 6.0 6.5) | -0.6333333 |
| (3.0 6.5 7.1) | (1.3 5.6 6.0) | -1.2333333 |
| (5.6 6.5 7.1) | (1.3 3.0 6.0) | -2.9666667 |
| (3.0 5.6 6.5) | (1.3 6.0 7.1) | -0.2333333 |
| (5.6 6.0 6.5) | (1.3 3.0 7.1) | -2.2333333 |
| (3.0 6.0 6.5) | (1.3 5.6 7.1) | -0.5000000 |
| (3.0 5.6 6.0) | (1.3 6.5 7.1) | 0.1000000 |
| (1.3 5.6 6.5) | (3.0 6.0 7.1) | 0.9000000 |
| (1.3 3.0 6.5) | (5.6 6.0 7.1) | 2.6333333 |
| (1.3 3.0 5.6) | (6.0 6.5 7.1) | 3.2333333 |
| (1.3 6.0 6.5) | (3.0 5.6 7.1) | 0.6333333 |
| (1.3 5.6 6.0) | (3.0 6.5 7.1) | 1.2333333 |
| (1.3 3.0 6.0) | (5.6 6.5 7.1) | 2.9666667 |

# Permutation distribution of $\overline{Y}_2 - \overline{Y}_1$



# Observed $\overline{Y}_2 - \overline{Y}_1 = 2.97$

```
> sort(diff)
 [1] -3.2333333 -2.9666667 -2.6333333 -2.2333333 -1.2333333 -0.9000000
 [7] -0.6333333 -0.5000000 -0.2333333 -0.1000000  0.1000000  0.2333333
[13]  0.5000000  0.6333333  0.9000000  1.2333333  2.2333333  2.6333333
[19]  2.9666667  3.2333333
```

So one-sided $p = 0.10$

# This is beautiful

## Thank you, Mr. Fisher

- The probability theory is elementary.
- There is no pretence of random sampling from some population.
- It still works if there is random sampling.
- All the randomness in the model comes from random assignment.
- Distribution-free.
- Small samples are okay.
- Any test statistic you want is okay under $H_0$; it's up to you.
- Some test statistics are better than others under $H_1$; it depends on *how* $H_0$ is wrong.

# The only problem

- For larger samples, listing the permutations or combinations of the data and calculating the test statistic for each one can be a huge task.
- It was impossible before electronic computers, in most cases.
- Now, it's possible to *estimate* the $p$-value of a permutation test even when it can't be obtained exactly.

# Monte Carlo estimation of the permutation $p$-value

See the *Wikipedia* under *Resampling statistics*.

- Place the values of the response variable in a random order.
- 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 this procedure a large number of times.
- By the Law of Large Numbers, the $p$-value is approximated by the proportion of randomly generated values that exceed or equal the observed value of the test statistic.

# Another kind of approximation

Fisher himself considered permutation tests to be entirely theoretical. In his classic *Statistical Methods for Research Workers* (1936) he wrote, after describing the procedure,

> *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.*

# Analysis of variance is an approximation

- Standard ANOVA methods can be justified as approximate permutation tests.
- Use the theory of Sample Surveys.
- Random assignment to a treatment group is like sampling without replacement from a relatively small population.
- There is a Central Limit Theorem.
- Convergence is fast, but it's still a large-sample argument.
- See Cox and Reid's (2000) *Theory of the design of experiments* and the references therein.

# Vocabulary

You need to know these terms.

- Observational study
- Experimental study
- Experimental unit
- Confounding variable
- Completely randomized design
- Random permutation
- Permutation distribution
- Permutation test

# Copyright Information