Data collection
- observational studies - confounding
- sample surveys - generalizable to population
- experiments
  - treatment imposed
  - cause-effect relation

outcome measurement
- predictor variables are called “factors”
- the values of a factor are its “levels”
- each combination of levels of the factors is a “treatment”

experiment
- a design is “designed” “balanced” if each treatment is applied the same number of times
Key step - randomization
eliminates effects of confounding
variables; differences among
treatment groups are random differences
- eliminates bias

Principles of designs of experiments:
- randomization - randomize
treatment assignment, time,
location
- control - a group for comparison
- replication - need multiple
observations per treatment
- ethically can't always carry out an
experiment

If subjects are human or animals,
its nice:
- placebo - to control for "placebo
effect"
- double-blind - subject nor experimenter
Final comments:

"Effective sample size"
- t-tests, analyses of variance
  assume observations are independent

Fishing expeditions
- if doing 100 tests at $\alpha=.05$ significance level, expect 5 of the 100 show
  significant difference from $H_0$ even when $H_0$ is always true.

One useful and widely applicable fix
"Controlling for Type I error rate"

Bonferroni inequality

\[
P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_k) \\
\leq P(A_1) + P(A_2) + \ldots + P(A_k)
\]
If $A_i$ is event $i$th test has Type I error

$A_i \cup A_2 \cup \ldots \cup A_n$ at least 1 Type I error

$P(A_i \cup A_2 \cup \ldots \cup A_n) = P(\text{at least Type I error})$

$\leq k \alpha \quad \Rightarrow \quad \text{"overall significance level"}$

"experimentwise error rate."

If use significance level of $\alpha/k$ for each individual test, the overall significance level will be at most $\alpha$

Note this inequality is very conservative.
Analysis of Variance (Ch 13, 14)

Example: The Spock Conspiracy Trial

Benjamin Spock

In 1968 Spock was tried in Boston by conspiring to violate the Selective Service Act by encouraging young men to avoid the draft.

The defense challenged the method by which jurors were selected claiming that women were underrepresented.
- there were no women on Spock's jury.

How juries were selected in Boston:
- 300 names selected at random from city directory
- before a trial, 30 jurors selected at random from 300
  the "venire"
- each side can then exclude jurors from this 30 (non-random)
Spock's trial the "venire contained only
  1 woman who was released by the prosecution.
Defence - Judge had history of underrepresenting women
- Compared judge's recent verdicts with 6 other Boston judges

Data: 70 women

Two key questions:
(1) Is there evidence that women are underrepresented on Spock's judges' venires compared to other judges? (t-test p<10^-6)
(2) Is there any evidence that there are differences in women's representation among other 6 judges?

Analysis of Variance (ANOVA)
- Generalization of 2-sample pooled variance t-test
- ANOVA - are there differences in the means among more than 2 groups?
Outcome variable: $y$

Measure $Y_{ij}$ for $j^{th}$ subject/in $i^{th}$ group (e.g., judges, etc.).

Model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

- Random overall mean value
- Random error effect of being in $i^{th}$ group

Assumptions:
- $\alpha_i$ defined so that $\sum_{i=1}^{a} \alpha_i = 0$
- $\epsilon_{ij}$ i.i.d. $N(0, \sigma^2)$

Note: same variance for each group
Want a test for:

\[ H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_a \]

versus \( H_a : \) at least 2 of the \( \alpha_i \)'s are not equal

Why "analysis of variance"?

Decompose sum of squares

\[
Y_{ij} = \overline{Y}_{..} + Y_{ij} - \overline{Y}_{..} \]

\[
\overline{Y}_{..} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{n_i} Y_{ij}}{N}
\]

where \( \overline{Y}_{..} \) is average of all observations

\( N = N_1 + N_2 + \ldots + N_a \)

\[
Y_{ij} - \overline{Y}_{..} = (\overline{Y}_{ii} - \overline{Y}_{..}) + (Y_{ij} - \overline{Y}_{ii})
\]
Square:

\[(Y_{ij} - \overline{Y}_\cdot)^2 = (\overline{Y}_i - \overline{Y}_\cdot)^2 + (Y_{ij} - \overline{Y}_i)^2 + 2(\overline{Y}_i - \overline{Y}_\cdot)(Y_{ij} - \overline{Y}_i)\]

Sum over \(i, j\):

\[\frac{1}{2} \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_\cdot)^2 = \sum_{i=1}^{a} \left( n_i (\overline{Y}_i - \overline{Y}_\cdot)^2 \right) + \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 + 2 \sum_{i=1}^{a} \frac{n_i}{n_c} (\overline{Y}_i - \overline{Y}_\cdot) \sum_{j=1}^{n_c} (Y_{ij} - \overline{Y}_i)\]

Total Sum of Squares

\[SS_{\text{Total}}\]

Error Sum of Squares

\[SS_{\text{E}}\]

Within group \[SS\]

\[\sum_{i=1}^{n_i} Y_{ij} j_i = n_i \overline{Y}_i\]
Summarize on ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>a-1</td>
<td>SSTr</td>
<td>SSTr</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>N-a</td>
<td>SSE</td>
<td>SSE</td>
<td>SSE/N-a</td>
</tr>
<tr>
<td>Total</td>
<td>N-1</td>
<td>SSTT</td>
<td>SSTT</td>
<td>MS_T</td>
</tr>
</tbody>
</table>

MS_T = SSTr/(a-1) = MS_Tr

MS_E = SSE/(N-a) = MS_E