75th MORSS CD
Cover Slide

UNCLASSIFIED DISCLOSURE FORM CD
Presentation
12-14 June 2007, at US Naval Academy, Annapolis, MD

Please complete this form 712CD as your cover page
to your electronic briefing submission to the MORSS CD. Do not fax to the MORS office.

Author's Request: I request authority to disclose the following presentation in the MORSS Final Report, for inclusion on the MORSS CD and/or posting on the MORS Website.

Name of Principal Author and all other author(s): Gregory T. Hutto, YD-3 DAF

Principal Author's Organization and address:
Phone: __850,882,0607_________
Fax: ____850.882.5644_________
Email:___gregory.hutto@eglin.af.mil____

Original title on 712 A/B:___________Resampling Statistics___________________________

Revised title:___________________________________________________________________

Presented in: Tutorial Session

This presentation is believed to be:
Distribution A: UNCLASSIFIED AND APPROVED FOR PUBLIC RELEASE
**Resampling Statistics Tutorial**

53rd Wing Air Combat Command USAF Eglin AFB

Approved for public release, distribution unlimited

Resampling Statistics Tutorial
75th MORS Symposium USNA

Presented by:

Greg Hutto
Chief Operations Analyst
53rd Wing
Air Combat Command USAF
gregory.hutto@eglin.af.mil
Tutorial Overview

- A couple of motivational problems
- Why we care & Resampling History
- Resampling vs. Classical Approach to Statistical Testing
  - The one sample bootstrap (binomial p to 1 sample t)
  - The two sample bootstrap & shuffle (two sample t)
  - One sample – sample size (noncentral t power)
  - 1 Way ANOVA (F test)
  - Simple linear regression
  - 2 Way ANOVA
  - Medians, correlations, percentiles
- Potential Advantages of Adopting Resampling
  - Teaching
  - Dealing with missing data
  - Estimating sample size
  - Nonnormal data
  - Getting the “right” answer in Analysis
- Caveats & Directions for the Future
Why we care about resampling

- It’s all about test productivity .., how many *good* tests can you run with $48 million and 2200 people?
- As of Spring 2002, the 53rd Wing has changed its method of test to Design of Experiments (DOE)
- We have 350+ project officers and 70 analysts to teach on a rotating basis
- More than 20 geographically separated units
- As of last count, 256 test projects on the ACC Test Priority List (March 2007)
- We currently service less than 75% of these each year
- And .. Post-Sept 11, the pace continues to quicken
Problems to motivate usefulness of resampling

- Probability of 1 boy & 1 girl in a two child family?
- Given that one child is a boy in a two child family, probability the other child is also a boy?
- JASSM Reliability is low. If “fixed” and now 80% can we determine that with 20 shots?
- Joint Programmable Fuze is 90%+ reliable. Or is it? We had 2 or 3 of 6 failures.
- The SWEAT diet is being proposed to get our Airmen “Fit to Fight’ will it work?
Why Resample? Information gathering & data analysis

- Answering a research question
- Estimating an uncertain quantity
- Determine sampling distributions of statistics whose distributions cannot or have not been mathematically approximated
- To avoid doubtful assumptions about the data
- In the case of permutation tests, to arrive at exact p-values
- To use a conceptually simple, widely applicable general method
Caveat – Resampling is a work in progress in our Wing

- Col (Dr) Pete Vandenbosch suggested we consider resampling in January 02
- Used 2002 MORS Symposium as a forcing function to get to work on solving the problems of applying
- Made some progress – Overall, a B+
  - Teaching – A
  - Nonnormal data – B+
  - Missing data – B-
  - Sample Size – A
  - Applying to DOE – B
- More directions at the summary
Teaching Difficult Statistical Concepts via Resampling

“Statistically Significant Difference”

- Hypothesis testing
- \( \alpha \) and \( \beta \) errors
- Confidence intervals
- Sample size via OC curves
- Random va. Causal variation
- Regression coefficients
Resampling crystallizes thinking

- Why must $\alpha$ be set to compute $\beta$ error?
- What is the root problem behind missing data and unbalanced designs?
- What assumptions are we making about the sample when we apply the t test? Are they appropriate?
- What are the real effects of outliers on linear models?
- What exactly are my null and alternate worlds?
Definitions

**Bootstrap:** Statistics done under the assumption that the population distribution is identical to the sample distribution. The data are sampled with replacement, simulating an infinite population from the sample.

**Permutation (Shuffling) Methods:** The basis for Fisher’s Exact test and the Tea Experiment. The gold standard of resampling schemes – sampling is without replacement.

**Resampling:** Statistics done under computation-intensive methods involving sampling from some distribution. These include bootstrap, jackknife, data shuffling schemes, and possibly others.

**Monte Carlo:** Simulations done using random (or pseudo-random) numbers. These might include resampling schemes, but for our purposes today, we’ll distinguish between them.
# Historical Resampling Timeline

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1650</td>
<td>Gambling Experiments – Blasé Pascal</td>
</tr>
<tr>
<td>1700</td>
<td>Probability Theory – Bernoulli</td>
</tr>
<tr>
<td>1850</td>
<td>Student t Distribution W.T. Gosset</td>
</tr>
<tr>
<td>1900</td>
<td>The Tea Experiment – R.A. Fisher</td>
</tr>
<tr>
<td>1925</td>
<td>Fisher’s Exact Test (permutations) – R.A. Fisher</td>
</tr>
<tr>
<td>1950</td>
<td>Monte Carlo approaches (RAND et al.)</td>
</tr>
<tr>
<td>1975</td>
<td>Quenouille’s (and later Tukey’s) Jackknife</td>
</tr>
<tr>
<td>1980</td>
<td>Applications to Business and Economics – Julian Simon</td>
</tr>
<tr>
<td>1990</td>
<td>Publication of Efron’s (Stanford) Article on Bootstrap</td>
</tr>
<tr>
<td>2000</td>
<td>Peter Hall’s Publication of Asymptotic Theory of Resampling</td>
</tr>
</tbody>
</table>

---

1. Source: Chapter 1 *Bootstrap Methods and Their Application*, AC Davidson, 1997
“Much to my chagrin, I found myself at the bottom of the lake!” declared the Baron.

“But Baron, how did you save yourself from a watery fate?” exclaimed his lady listener.

“Why, I simply reached down,” the Baron explained “and pulled myself up by my bootstraps.”

Source: www.munchasuen.com
BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

Any experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

- Article originating the Student t distribution by W.T. Gosset, a chemist for Guinness Brewery
- Wrote under a pseudonym for the same reason Romance Writers do ... Statistics not an honorable profession
Gosset’s 1908 Article

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

Any experiment may be regarded as forming an individual of a “population” of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the “error of random sampling” the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabulated, and whose properties are well known.

Translation – we use the normal because we know it…”When all you have is a hammer, everything begins to look like a nail.”
SECTION VI. Practical Test of the foregoing Equations.

Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3000 criminals, from a paper by W. R. Macdonell (Biometrika, Vol. 1. p. 219). The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains the measurements of 3000 criminals in a random order. Finally each consecutive set of 4 was taken as a sample—750 in all—and the mean, standard deviation, and correlation† of each sample determined.

- Simulating a Shuffle test – without replacement
- Could have improved his procedure with bootstrap, I think, and by repeating the sets of 750 -- 3 or 4 times.
In this clever experiment, Fisher justified the \( t \) and \( F \) statistics to examine the null hypothesis of sameness, without appealing to the Central Limit Theorem or the Chi Square Theorem. It makes us more comfortable with violations of the normality and equality of variance assumptions.

\[
\begin{array}{c}
\text{t-statistic:} \\
115 \\
\text{F-statistic:} \\
462
\end{array}
\]

These are two of 462 ways to assign 5 samples to the “Wet” label. Under \( H_0 \), (W=D), the W/D labels make no difference in \( y \) values and thus the samples randomly appear in each column. We plot the distribution of mean differences, and observe how often a difference in 2 of means appears. This is the randomization test.

In this problem, the alphas for both the \( t \) and the randomization test agree:

\[
\begin{align*}
\alpha_t &= .34 \\
\alpha_{\text{rand}} &= .33
\end{align*}
\]
Suppose I am a data collector - only the best! I have historical records of B-1B dry data sets.

Someone offers me a data set claiming it’s B1-B dry data. x - 25m, n = 10

Examining my reference collection, I observe that only 6.75% of my 487 samples equal or exceed this value.

Accordingly, I reject this set with a 6.75% change of being mistaken.

I find likelihood that I see a mean of 25m given this is B-1B dry is low.
Developing a Random Variable Distribution

Step 1. Collect some data

- A random variable converts experimental outcomes into a numeric variable that can be discussed with the tools we are learning.
- What proportion of RWR age outs are greater than 12 seconds?
- With a reference distribution you can compute “unlikelihood”
- **Probability distribution**: Features $P(x)$ values for each outcome. $P(x)$ are between zero and one. The p’s sum to 1.0
- We have several classical theoretical ones: binomial, t, z, chi-square, F ...

Step 2. Assess a discrete or continuous curve that describes how samples are observed

- Samples
  - Age out time
  - 1
  - 2
  - 3-8

- Curve describing how age outs are experienced

June 2007  Resampling Statistics Tutorial  R-17
Concepts of Resampling

- As with classical statistical theory – what we know of the population is contained in the sample
- With classical stats, we use mathematical theorems to estimate the behavior of the population
- With resampling – we draw repeated samples from our original sample and use that evidence to describe the population
  - Conceptually easier to swallow (and teach)
  - Fewer distributional assumptions
  - More flexible in practice
- Let’s see where this idea takes us...
As we re-sample, we obtain different values for a Random Variable

- Original sample of 13 weapon scores
- Resample in groups of 13 (with replacement)
Distribution of the Mean as we re-sample (sampling distribution)

- Resampling statistics (from Simon ‘68 and Efron ‘79) supply us with that empirical reference distribution we had been hoping for.
- We can make probability statements regarding the behavior of any function computed from the sample.
Central Limit Theorem -- Recap

**Given:**

1. Random variable \( x \) distributed arbitrarily with mean \( \mu \) and standard deviation \( \sigma \).
2. Draw samples of size \( n \) are randomly (independently!) from this population.

**Conclusions:**

1. The distribution of sample means \( X \) will, as the sample size increases, approach normal.
2. The mean of the sample means will be the population mean \( \mu \).
3. The standard deviation of the sample means is \( \frac{\sigma}{\sqrt{n}} \) (the standard error of the mean).

**Practical Rules of Thumb**

1. For samples of size \( n \) larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. The approximation gets better as the sample size \( n \) becomes larger.
2. If the original population is itself normally distributed, then the sample means will be normally distributed for any sample size \( n \).
There are combinatoric ways to solve for the sampling distribution of the median, but not with normal-theory methods.
Sampling distributions enable us to judge Significance.

What is the likelihood that the true population average is 8? 19? 35?
With sample of proportions

- Reliability can be viewed as binomial – in any mission, it fails or does not
- JASSM reliability was one of the motivating problems
- We worked a sample size (power) problem with a 20 missile launch
One sample of means, medians, percentiles

- We used the JPF as a motivating example
- Does the test evidence support the claimed reliability?
Two sample bootstrap & shuffle

- We used the “SWEAT Diet” as one of our motivating examples
- Question 1: do the diets differ in weight lost?
- Question 2: if so, by how much?
Resampling Poses Three Sample Size Problems

- How many samples to draw from original data ($b$)
  - Best guidance – $O(n)$ the original sample but NLT 10-15

- How many times to resample original data ($r$)
  - Best guidance -- >500 times. 1000 converges most problems thus far

- How many original sample data values ($n$)
  - Next slide
OC curves for different values of $n$ for the two-sided $t$ test for a level of significance $\alpha = 0.05$

### 1 Sample

- $\text{RWR} \leq 12 \text{ sec}$
- $\alpha_1 \beta = 0.05$
- $\mu_1 - 12 = 1 \text{ sec}$
- $\sigma = 2 \text{ sec}$
- $\Delta = 0.5 \Rightarrow n \approx 60$

### 2 Sample ($\text{Ind!}$)

- $\text{RWR} v23 \text{ v40}$
- Range diff $= 3$
- $\beta = 0.10$  $\sigma = 2 \text{ miles}$
- $\Delta = 0.75 \Rightarrow n \approx 20$
- $n_1 = n_2 = n^* = \frac{n + 1}{2} = \frac{21}{2} = 10.5 \approx 11$
- $n_1, n_2 = 11$
How many original samples (n)

- As in other areas, resampling clarifies the problem and illuminates the solution.
- We more clearly understand the sample size big four: \(\alpha, \beta, \delta, \sigma\).
- We can replicate the t and F test operating characteristic curves experimentally.
Classical Approaches to Multivariable Testing are Difficult to Master

- Consider the JSOW delivery:
  - Three terminal delivery parameters
  - Several threat conditions
  - Two profiles
  - Two target types (fixed, mobile)
  - Two sets of launch conditions (in range-in zone)
  - Number of combinations?

- Our problems are multi-variable

- So must be our solutions
  - ANOVA
  - Regression
  - Resampling
Compare to simply checking ANOVA assumptions

<table>
<thead>
<tr>
<th>Classical Assumptions</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 samples independent of t,x,y ...</td>
<td>NPP</td>
</tr>
<tr>
<td>2 equal variances in all cells</td>
<td>e vs. t ✓</td>
</tr>
<tr>
<td>3 residuals distributed normally</td>
<td>e vs. x, other ✓</td>
</tr>
<tr>
<td>4 model is adequate for process</td>
<td>e vs. y ✓</td>
</tr>
</tbody>
</table>

### Tests

<table>
<thead>
<tr>
<th>Tests</th>
<th>NPP</th>
<th>e vs. t</th>
<th>e vs. x, other</th>
<th>e vs. y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

- On Review/Save Residuals tab, choose and save all factors you want to plot against residuals
- On resulting scroll sheet, do scatter plots with residuals as Y and X or time, or other variable as X.
We’re teaching our testers to create multivariable tests

- Factorial experiment in Mobcap Apex threat exploitation examines
  - 4 ECM techniques,
  - two maneuvers, and
  - two directions
- All in a single one hour mission
ANOVA partitions variability around the group means

ANOVA decomposes total variation into explained and unexplained portions
Regression Partitions Variance around line (not grand mean as ANOVA)

Assumptions

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

\[ \varepsilon_{ij} \sim N(0, \sigma^2) \]

\[ \varepsilon_{ij} \text{ independent of } y, x, t \]

Regression Coefficients

\[ \beta_1 = \frac{\text{Rise}}{\text{Run}} = \text{Slope} = \frac{\Delta y}{\Delta x} \]

\[ \beta_0 = \bar{y} - b, \bar{x} \]

Recall

\[ S_y^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1} \]

Summing the Squares of the Components of Variance

\[ (y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \]

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \]

\[ \text{SST} = \text{SSE} + \text{SSR} \]
Applying Resampling to DOE

- Sample case had a two variable design with an overall mean of 40 and an A effect of 60 in the presence of 15 units noise
- Resampling Model estimated as follows (90% confidence interval)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Y Bar</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>34.7</td>
<td>11.9</td>
<td>16.6</td>
<td>4.9</td>
<td>17.0</td>
</tr>
<tr>
<td>b</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>22.5</td>
<td>2.9</td>
<td>13.2</td>
<td>1.6</td>
<td>10.1</td>
</tr>
<tr>
<td>a</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>97.8</td>
<td>69.4</td>
<td>79.5</td>
<td>72.3</td>
<td>79.8</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>101.9</td>
<td>47.2</td>
<td>95.7</td>
<td>84.8</td>
<td>82.4</td>
</tr>
<tr>
<td>Effects</td>
<td>67.5</td>
<td>-2.2</td>
<td>4.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47.3</td>
</tr>
</tbody>
</table>

- This case shows that resampling gives similar answers to classical methods Only A is active in changing Y; B and AB are inactive (effect CI’s includes 0)
Resampling vs. Classical ANOVA

We get substantially identical results with resampling and ANOVA.

The algorithm is simple and resistant to "misuse".

Resampling Results (Within Rows)

<table>
<thead>
<tr>
<th>Effect</th>
<th>L</th>
<th>U</th>
<th>Estimate</th>
<th>&quot;Real&quot;</th>
<th>CI Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>56.2</td>
<td>79.1</td>
<td>67.5</td>
<td>60.0</td>
<td>22.9</td>
</tr>
<tr>
<td>B</td>
<td>-13.7</td>
<td>8.3</td>
<td>-2.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>AB</td>
<td>-7.4</td>
<td>15.6</td>
<td>4.8</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Warning – the Akerson\textsuperscript{1} Smear Effect

- Suppose I have two strong effects – A at 60 and AB at 20
- Estimate these effects via shuffling

<table>
<thead>
<tr>
<th>Case</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Y Bar</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>24.1</td>
<td>19.6</td>
<td>29.7</td>
<td>49.4</td>
<td>30.7</td>
</tr>
<tr>
<td>b</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-24.3</td>
<td>-29.3</td>
<td>-1.6</td>
<td>-8.0</td>
<td>-15.8</td>
</tr>
<tr>
<td>a</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>32.4</td>
<td>57.8</td>
<td>34.4</td>
<td>80.0</td>
<td>51.2</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>85.6</td>
<td>61.0</td>
<td>69.2</td>
<td>70.5</td>
<td>71.6</td>
</tr>
<tr>
<td>Effects</td>
<td>53.9</td>
<td>-13.0</td>
<td>33.5</td>
<td>34.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- When looking up the values in the resampling distribution
  - A of 53.9 is at the 99.4 percentile – a strong effect
  - AB=33.5 is at the 87.5 percentile – an insignificant value
- What happened? The strong A effect is smeared over the rest of the data set, masking the AB effect. Permutation fails here.
- When resampling \textit{within rows} (to keep A where it belongs)
  - AB=33.5 is at the 99.9\textsuperscript{th} percentile where it belongs

\textsuperscript{1} Discovered by LtCol Jerome Akerson, May 2002
30 years experience teaches:

“When resampling and classical methods are applied to classical problems they agree. When applied to no-classical problems, resampling usually gets the right answer.”

Procedure:

- Create and examine data that violates normality & equal variance assumptions for ANOVA

Findings:

- For Boot Strap and transformed-ANOVA confidence intervals will not at all agree
- Boot strap can other useful distribution statistics--including median and percentile
Some cautionary notes on resampling non-normal data

1. There is a bias in bootstrap estimates of variance --, in Confidence intervals -- therefore in significance tests.
2. Bias can be fixed using a second bootstrap procedure.
3. Normal assumptions turn out to be pretty good until the data get way off. Strongly bivariate data are the case we’ll show.
4. So far, large skewness doesn't appear to be a major factor.
5. Regardless of this bias, there are numerous important roles for bootstrap.

Warning for beginning users -- dangerous curves ahead.
Mechanics of analysis via Resampling are simple

**Algorithm:**
- Set up a (possibly constructive) sample
- Compute a function of the sample observations (Random Variable)
- Choose a resampling method
  - Permutation (shuffle)
  - With replacement (Bootstrap)
  - With single item deletion (Jackknife)
- Resample the original sample
- Review the output for the distribution of the Random Variable
- Draw conclusions
Summary and Questions
Summary of Resampling

- 53rd Wing power and confidence across broad battlespace as test method
- Real world poses challenges to classical assumptions
- Resampling provides a method (potentially) that is:
  - Simple to teach and use
  - Resistant to misuse
  - Robust against violations of assumptions
  - Helps with missing data
  - Dealing directly with the problem
  - Illustrative of difficult statistical concepts
- Progress report after 5 years of investigation – very promising – but Akerson Smear a troublesome issue

“To call in the statistician after the experiment is ... asking him to perform a postmortem examination: he may be able to say what the experiment died of.”

Address to Indian Statistical Congress, 1938.
Resampling Resources

- Peter Bruce, Resampling Stats, Inc. pbruce@resample.com, www.resample.com, the single best resource to learn these methods
- *Resampling: the New Statistics*, Simon and Bruce
- *Bootstrap Methods*, Michael Chernick
- *Permutation Tests*, Phillip Good
- R: A statistics language, freeware for download – numerous sites
- Gregory.hutto@eglin.af.mil
Questions ? ? ?