Applying Permutation Methods

Lake Bled 9/22/2010

Phillip Good
http://statcourse.com
http://zanybooks.com
drgood@statcourse.com

1934; 1937

Changes in Theory/Technology Dictate Changes in Practice

Please ask questions

during the lectures

+ written questions during the breaks

No cell phones

No No No

Use Permutation Tests

- Teach Hypothesis Testing
- Multivariate comparison of population means.
- Compare variances.
- Analyze the one-way layout.
- Analyze cross-over designs.
- Analyze contingency tables.

Session I
Two-Sample Comparison

- Univariate Comparison
- Choosing a Test Statistic
- Multivariate Comparison
- Experimental Designs

Two-Sample Comparison

Pitman EJG. Significance tests which may be applied to samples from any population. Roy Statist Soc Suppl. 1937; 4: 119–30.

1 2 3 4 5.3 6

1N 2N 3N 4F 5.3F 6F

null hypothesis:

labels are exchangeable

1N 2N 3F 4N 5.3F 6F

Compare Sums
Some Rearrangements

NEW	FULL
1 2 3	4 5 6
1 2 4	3 5 6
1 2 5	3 4 6
. . .	. . .
4 5 6	1 2 3

Four Steps to a Permutation Test

1. Choose a Test Statistic
2. Compute Statistic
3. Iterate
   - Rearrange labels on data
   - Compute statistic
   - Update permutation distribution
4. Determine p-value

Choose a Test Statistic

$$\sum X_i$$

$$\bar{X}$$

$$\bar{X} - \bar{Y}$$

$$\frac{(\bar{X} - \bar{Y}) / \sqrt{1/m + 1/n}}{\sqrt{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2 / (m + n - 2)}}$$

t-test or Permutation Test?

t-test/perm test

- asymptotically equivalent

Non-normal distribution

- t-test robust
- permutation robust and more powerful

Variances Must be the Same

Choose a Test Statistic II

- Best discriminates between H and K
- Neyman-Pearson Lemma

  Likelihood Ratio
- H: All outcomes equally likely
- Select as R, outcomes most likely under K

  Extreme values of test statistic

Choose a Multivariate Statistic

Hotelling's T^2

Exchange labels on vectors

Significance Level

- Exact
- Distribution-Free

Software

- R—many routines

http://cran.r-project.org/web/packages/permtest/index.html

- NCSS
- Stata

Experimental Designs

- Random Blocks
- Latin Squares
- Fractional Factorials

One-Way Analysis

$EX_{ik} = \mu + ♋_i$

H: $♋_i = 0$ for all $i = 1, \ldots I$

K_1: $♋_I \neq ♋_j$ for some i, j

Choose a Test Statistic

One-Way Analysis

K_1: ♋$_I \neq$ ♋$_j$ for some i, j

Eliminate Invariants from F-Ratio

$$\frac{\text{Between Sum of Squares}}{\text{Within Sum of Squares}}$$

$$F_2 = \sum_{i=1}^{I} \left(\sum_{k=1}^{K_i} X_{\mathrm{ik}} \right)^2$$

Simulation Results

Mixture of Normals

```
s1=rnorm(3,rbinom(3,1,0.3))
s2=ifelse(rbinom(4,1,0.3),rnorm(4,0.5),
        rnorm(4,1.5,1.5))
s3=ifelse(rbinom(5,1,0.3),rnorm(5,1),
         rnorm(5,2,2))
```

Unbalanced Design (3,4,5); alpha = 0.1

Power perm = 30%

Power anov = 18%

Choice of Test Statistics

$$F_2 = \sum_{i=1}^{I} \left(\sum_{k=1}^{K_i} X_{\mathrm{ik}} \right)^2$$

$$F_1 = \sum_{i=1}^{I} \left| \sum_{k=1}^{K_i} X_{\mathrm{ik}} \right|$$

Randomized Blocks

$X_{mik} = \mu_m + ♋_i + \varepsilon_{mik}$ $\qquad$ $F[_{mik}] = F_m$

H: $♋_i = 0$ for all $i=1,\ldots I$

K_1: $♋_I \neq ♋_j$ for some i, j

$$F_2 = \sum_{m=1}^{M} \sum_{i=1}^{I_m} \left(\sum_{k=1}^{K_{mi}} X_{\mathbf{mik}} \right)^2$$

Randomize independently within each block

Latin Squares

Hint:

Analyze an experiment
as it was Designed

Comparison

Permutation Test

1. Choose a Test Statistic
2. Compute Statistic
3. Permutation Distribution
   - Rearrange labels
   - Compute statistic
   - Order Values
4. Significance level

Parametric Test

1. Find Well-Tabulated Statistic
2. Compute Statistic

4. Look up Significance level

Computing the Permutation Distribution

- Exhaustive Enumeration
- Monte Carlo

Monte Carlo

```
So = statistic(data)

cnt=0

for (k in 1:M){

  rdata=sample(data)
  S=statistic(rdata)

  if(S>= So) cnt++

}
```

$\hat{p}$ = cnt/M

$\hat{p}$ is B(p,M)

Transform the Data?

- Original observations
- More Powerful
  - F-Scores
- More Robust
  - M-Tests
  - Ranks

Use Rank Tests Only To

- Reduce Influence of Outliers
- Combine Observations taken on different scales
- Combine Observations of differing precision

Session II
Applications

- Comparing Variances
- Crossover Designs
- Microarrays
- Contingency Tables

Steps to a Permutation Test

- Specify Hypotheses
- Choose a Test Statistic
- Specify Design
- Choose Set of Permutations

Comparing Variances

- Two Samples
- Multiple Samples
- Non-Responders

Comparing Variances

$$x'_i = x_i - medianX$$

$$MAD = \sum_{i=1}^{n-1} |x'_i|$$

Comparing Variances II

^‡o‡o‡t‡ ‡E‡n‡t‡r‡y‡‡‡‡‡‡‡‡‡‡‡

Comparing Variances I>2 samples

$$x'_{ij} = x_{ij} - medianX_i$$

$$MAD = \sum_i \left(\sum_{j=1}^{n-1} |x'_{ij}|\right)^2$$

Choose a Test Statistic VI
Nonresponders

H: $F_2[x]=F_1[x]$

K: $F_2[y]=(1\text{-}p)F_1[y] + p\,F_1[(y\text{-}d)/s]$ d>0 s ¤ 1 p>0

$$T = u(\bar{X}_\pi - \bar{X}_O) + (1-u)(MAD_\pi - MAD_O)$$

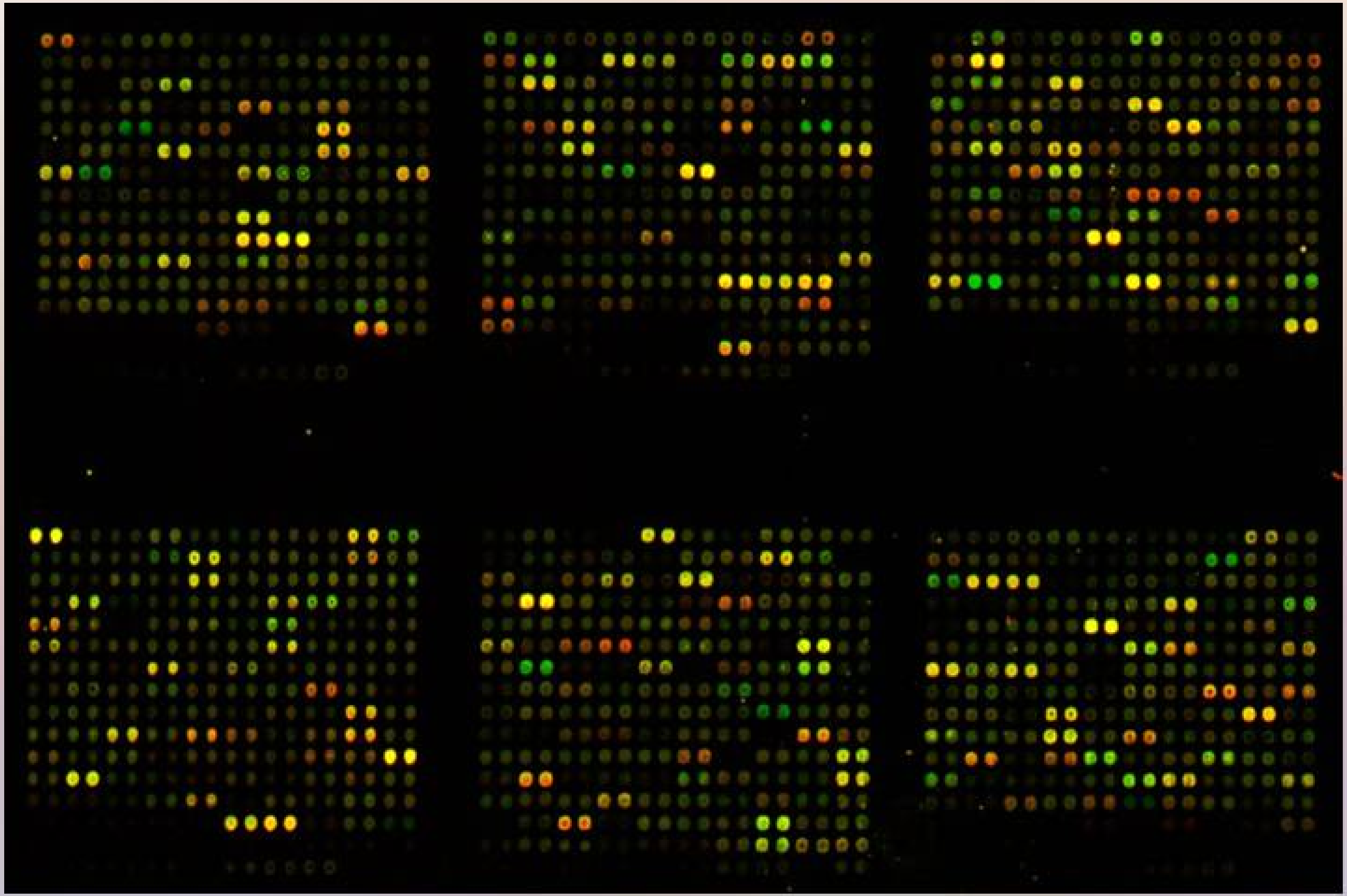
$$x'_i = x_i - medianX;\ \text{eliminate } 0$$

$$MAD = \sum_{i=1}^{n-1} |x'_i|$$

Simulation Results

```
data1=rnorm(size[1],1)    N(0,1)
data2=ifelse(rbinom(size[2],1,0.5),
        rnorm(size[2],1), rnorm(size[2],2,2))
                      .5N(0,1) + .5N(2,2)
alpha =0.1, size=c(11,11)
```

β: t-test 17% perm 18% permT 24%



Microarrays: Selecting Genes

Null Hypothesis:

1. Genes in S have same associations with *phenotype* as rest of genes.

2. S does not contain genes whose expression levels are associated with the *phenotype* of interest.

Analyzing Microarray Data I

1. Put data in matrix. Rows are genes. Columns are subjects.

2. Calculate measure of association t_i between each gene i and the phenotype of interest.

3. Form the statistic T by summing over all the t_i for the genes in the set of interest.

Analyzing Microarray Data II

Hypothesis I. Genes in S have same associations with *phenotype* as rest of genes.

Form the permutation distribution by rearranging the *row* labels of the data matrix and computing T for each rearrangement.

Analyzing Microarray Data II

Hypothesis II. S does not contain genes whose expression levels are associated with the *phenotype* of interest.

Form the permutation distribution by rearranging the *column labels* of the data matrix and computing T for each rearrangement.

Analyzing Microarray Data III

1. Rearrange rows (or columns) of t_i data matrix.

2. Form the statistic T by summing over all the genes in the rearranged set.

3. Repeat steps 1 and 2 to form Permutation distribution of T. Determine p-value.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, and Park PJ. Discovering statistically significant pathways in expression profiling studies. *PNAS*. USA. 2005; 102:13544–13549.

Cross-Over Design (Intent)

1. TTC TCT CTT
2. TCT CTT TTC
3. CTT TTC TCT

Patients randomized to treatment sequence

Cross-Over Design Analysis

Analyze as two-sample comparison

Test statistic ΣC

Randomize treatment sequence labels among patients

Has no effect under null hypothesis

Cross-Over Design (Reality)

1. TTC TCT CTT
2. TCT CTT TTC 61 patients
3. CTT TTC TCT

TTC TCT
TCT CTT 9 patients
CTT TTC

4 patients

Good P. and Xie F. Analysis of a crossover clinical trial by permutation methods. *Contemporary Clinical Trials* 2008; 29:565-568.

Contingency Tables

	Standard Drug	New Drug
Patient Got Better	5	9
No Improvement	5	1

p-value 3% or 7%

Is The Sample Large Enough?

p=3% chi-square approximation

p=7% Fisher's Exact Test

.sampsi 05. 0.9 , n1(10) power = 0.29

.sampsi 05. 0.9 , n1(30) power = 0.90

Table A	Lived	Died	Total
Men	9	1	10
Women	4	10	14
Total	13	11	24

Table B	Lived	Died	Total
Men	10	0	10
Women	3	11	14
Total	13	11	24

Table C	Lived	Died	Total
Men	8	2	10
Women	5	9	14
Total	13	11	24

Contingency Tables

- Common Odds Ratio in 2x2's.
- RxC tables
  - UnOrdered
  - Ordered
- Use StatXact (cytel.com) or R (fisher.test)

TEST FOR HOMOGENEITY OF ODDS RATIOS

Asymptotic p-value: (based on Chi-Square distribution with 17 df)

Pr { BD .GE. 25.78 } = 0.0785

Monte Carlo estimate of p-value:

Pr { ZE .LE. 9.481e-009 } = 0.0126

TEST FOR HOMOGENEITY OF ODDS RATIOS

Asymptotic p-value: (based on Chi-Square distribution with 17 df)

Pr { BD .GE. 25.78 } = 0.0785

Monte Carlo estimate of p-value:

Pr { ZE .LE. 9.481e-009 } = 0.0126

LIKELIHOOD RATIO TEST

Statistic based on **9 x 3** table (rows/cols with 0 ignored):

Likelihood ratio = 23.30

Asymptotic p-value: (based on Chi-Square distribution with 16 df)

Pr { LI(X).GE. 23.30 } = 0.1060

Monte Carlo estimate of p-value :

Pr { LI(X).GE.23.30 } = 0.0349

99% CI = (0.0302, 0.0396)

Session III
Applications and Limitations

- Your Problems
- Validation
- Type I Censoring
- Assumptions and Limitations
- Books and Software

Validation

- Learning Data vs Test Data
- Test Data vs Model Predictions
- Metric

Type I Censoring

- Radio-Immune Assay
cannot detect small values

- Median Time to Failure

- Animal Survival Studies

Type I Censoring

- Reliability; Assays

(C=#Censored, L=Mean Uncensored)

$C\pi >= C0$ & $L\pi >= L0$ π in R

$C\pi < C0$ & $L\pi < L0$ π in A

$C\pi >= C0$ & $L\pi < L0$ π in I

Permutation Test Assumptions

Test exact if

$\{x_{ijk}\}$ Exchangeable

$$F = \prod\prod\prod f_{ijk}[x_{ijk}]$$

F: Invariant under all rearrangements

Exchangeable if

- $\{X_i\}$ Identically distributed

and

independent.

or

- $\rho_{ij} = \rho$ for all i and j.

Use Permutation Tests

- To Obtain Exact p-values
- For More Powerful Tests
- For Combining Independent Tests
- Application-Specific (non-tabulated) statistics

Pesarin's Combination Test

Convert Data to Vector of p-values

Iterate—create rows of matrix

1. Rearrange Labels on Data
2. Convert Data to Vector of p-values

Convert p-values to ranks within columns

Combine ranks within row

Convert Data to Vector of p-values

$$\begin{pmatrix} x_{11} & \\ & x_{34} \end{pmatrix} \to \left(p_1, \ldots p_4, \ldots\right)$$

Convert p-values to ranks within columns

$$\begin{pmatrix} p_1 & . & . & p_4 & \mathbf{K} \\ p_{1\pi} & . & . & p_{4\pi} & \mathbf{K} \\ & & & & \\ & & & & \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 8 & . & 3 & \mathbf{K} \\ 7 & . & . & 9 & \\ 2 & . & . & 1 & \mathbf{K} \\ & 2 & & & \end{pmatrix}$$

To Learn More: Books

- *Introduction to Statistics via Resampling Methods and R/S-Plus*, Wiley, NJ 2005.
- *Resampling Methods*, Birkhauser, Boston, 1999, 2001 (3rd edition, 2005).
- *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, NY. (3rd edition, 2005).
- *Common Errors in Statistics*. Wiley, 2nd Ed, 2006

Analyzing Large Arrays of Data, Wiley, 2011

Managers' Guide to the Design and Conduct of Clinical Trials, Wiley, NY, 2nd ed, 2006.

Applications of Statistics in the Courtroom, Chapman Hall, London, 2001.

Pinkie: Stories of a Homeless Man, Zanybooks, CA, 2008.

2011:Military Takeover of America, Jay Cuze, Zanybooks, CA, 2010.

The Canyon, Luke Jackson, Zanybooks, CA, 2009

Software

- R—many routines
- NCSS—Hotelling's T^2
- NPCtest—Pesarin's omnibus test.
- StatXact—RxC contingency tables