Biost 517
Applied Biostatistics I

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 16:
Two Sample Inference for Correlated Response Data

November 30, 2007

1

Lecture Outline

- Dependent Data Within Clusters
- Matched Continuous Data
  - Paired t Test (means, geometric means)
  - Sign Test (median difference)
  - (Wilcoxon) Signed Rank Test
- Comparing Proportions: Matched Samples

2

Dependent Data Within Clusters

3

Dependent Data

- There are times when data can not be presumed to be totally independent
  - Sampling within families
  - Sampling within schools, hospitals
  - Repeated measurements on individuals taken at a single time
  - Longitudinal data: repeated measurements taken on individuals over time

4

Motivation for Longitudinal Data

- Three settings in which longitudinal studies are performed
  - Convenience of existing study population
  - Efficiency of using subjects as own comparison
  - Scientific questions about effects that occur
    - over time, or
    - within subjects

5

Convenience

- Questions are truly cross-sectional
  - Multiple measurements made on each individual is easier than gathering new subjects
    - Natural variation within individuals provides additional information
  - E.g., Serum osmolality from Na, Glc, BUN
    - Interest is relationships between concurrent measurements

6

Efficiency

- Questions could be answered with cross-sectional study
- Primary comparison within subjects may have less variability
  - Allow detection of smaller effects
  - E.g., Adjusting for baseline measurements
  - E.g., Cross-over study of a new treatment

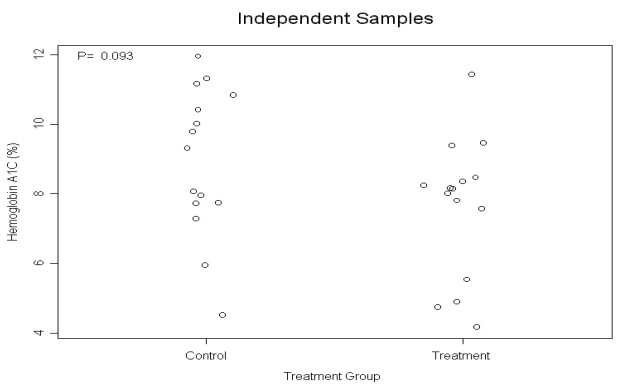
7

Example

- Percent glycosylated hemoglobin is used to monitor long term control in diabetes
  - Hemoglobin A1c
- Consider studies of two insulin delivery strategies
  - Independent groups
  - Cross-over design

8

Graph: Independent Samples



9

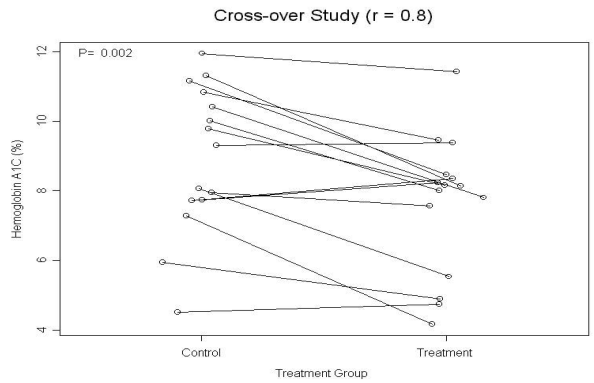
Inference: Independent Groups

- Large between-subject variability hampers our ability to detect differences
  - Between group SE is square root of sum of squared within group SEs
  - Within group SEs are proportional to within group standard deviation divided by the square root of n

$$se\left(\overline{X}-\overline{Y}\right)=\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$$

10

Graph: Cross-over Study



11

Inference: Cross-over Study

- High correlation between measurements taken on the same individual increases precision
  - The “random effect” of patient ID can be thought of as a precision variable

$$se\left(\overline{X}-\overline{Y}\right)=se\left(\overline{D}\right)=\sqrt{\frac{\sigma_X^2+\sigma_Y^2-2\rho\sigma_X\sigma_Y}{n}}$$

12

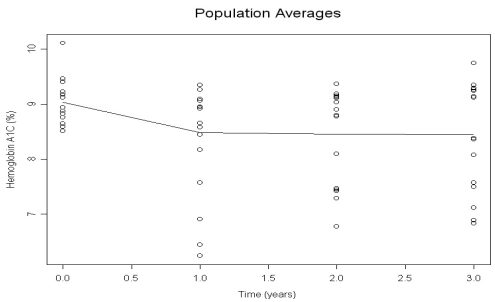
Longitudinal Questions

- Scientific questions about effects that occur over time
  - Studies to detect population time trends in response
    - E.g., rate (slope) of progression of retinopathy in population of diabetics over time
    - E.g., time to development of albuminuria

13

Example: "Marginal Effects"

- Time trends in group mean HbA1C
  - Note trends in mean and variability



14

Within Subject Effects

- Trends in specific individuals might not look like trends in population means
  - Response over time may be restricted to subgroups of subjects
  - Response over time may be transient

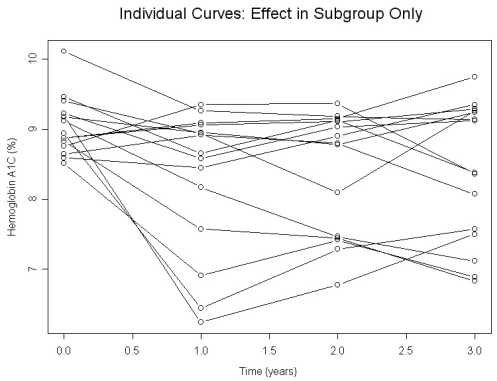
15

Longitudinal Scientific Questions

- Scientific questions about effects that occur within subjects
  - Studies to detect time trends or covariate effects in individual response
    - E.g., distribution of rates (slopes) of progression of retinopathy in population over time
    - E.g., effect of varying risk factors within individuals

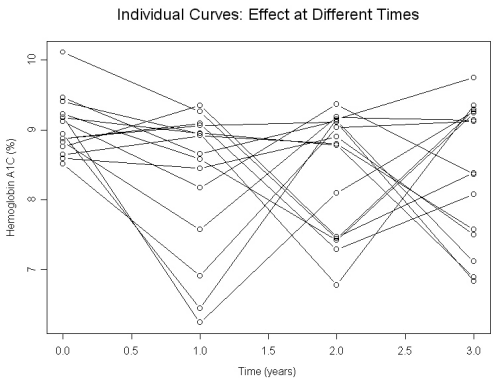
16

Effect in Subgroup



17

Transient Effects



18

Choice of Measures of Outcome

- In order of importance
  - Scientific relevance
    - Including state of current knowledge
  - Plausibility of difference across groups
  - Statistical precision for analysis

19

Longitudinal Outcome Measures

- In longitudinal studies, each individual may have multiple measurements over time
  - Definition of individual response thus can be based on multiple measurements
    - Response at a fixed time
    - Responses at multiple fixed times
    - Average response over time (area under curve)
    - Rate of change in response (slope)
    - Time to attaining some level of response

20

Measures of Outcome

- "Marginal" or population effects
  - Difference or ratio of group means, geometric means, medians, proportion or odds above threshold, hazards
  - Pr (Y > X)
- "Within subject" effects
  - Mean, median difference
  - Mean, geometric mean, median ratio
  - Within subject odds ratio
  - Pr (Y > X)

21

Choice of Longitudinal Outcome

- Should reflect scientific relevance, plausibility of effect, precision
  - Final level of response may be more important than earlier effects
    - (But in the long run, we are all dead)
  - Summarizing response at multiple time points reflects population rather than individuals
  - Average response over time sensitive to transient effects
  - Differences in time to event may be clinically meaningless

22

Statistical Issues

- Repeated measurements on subjects require special analysis techniques
  - May have erroneous conclusion if fail to account for correlated observations
    - Point estimates may be biased for population parameters
      - Too much emphasis placed on some subjects
    - Confidence intervals will not be accurate representation of our true confidence
    - P values will be wrong

23

Statistical Approaches

- Three basic approaches to analyzing correlated data
  - Reduce measurements on each cluster to a single observation; analyze across clusters
  - Estimate correlation within clusters and adjust standard errors for population based models
    - GEE, marginal models
    - "Robust" variance estimates
  - Adjust estimates for "random effects"
    - "Mixed effects models": both fixed and random

24

Easiest Approach

- Reduce data for each individual to a single measurement
  - E.g., response at end of study, average response, rate of change
  - Analyses can then be based on standard methods for independent data
  - But:
    - Does not allow time-varying covariates
    - May not be most efficient statistically

25

Example: Beta-carotene Data

- Randomized clinical trial of beta-carotene supplementation on plasma levels of beta-carotene and vitamin E
  - Subjects randomized to 5 dose groups
  - Measurements at baseline, after 3 and 9 months of treatment, and 3 months after stopping treatment
  - Scientific question: How do plasma beta-carotene levels change over time within dose groups?
    - (effect modification between dose and time)

26

Example: Beta-carotene Data

- Reduce data to a single measurement on each subject
  - Difference between follow-up and baseline
    - Consider average of differences
    - No change corresponds to a difference of 0
  - Ratio between follow-up and baseline
    - Consider average of ratios
    - No change corresponds to a ratio of 1

27

Example: SEP data

- Somatosensory evoked potential measurements on healthy adults
  - Measurements of nerve conduction time
    - Four separate peaks for each leg of each subject
  - Reduce data to a single measurement
    - Consider only one peak on one leg
      - Which one?
    - Average measurements across peaks, legs
      - But will only generalize to similar averages
    - (Differences between peaks?)

28

Matched Continuous Data

29

Comparing Means

- Paired t test
  - Compute differences for each pair
  - One sample t test that mean difference is 0
- Note that mean difference is difference of means
  - Same answer for population ("marginal") and within subject questions (providing they both make sense)
    - May be inherent confounding, effect modification
    - E.g., age vs time vs birth year cohort effects

30

Comparing Geometric Means

- Paired t test on log transformed data
  - Compute differences for each pair
  - One sample t test that mean difference is 0
  - Back transform to consider geometric mean of ratios
    - Also geometric mean of ratios

31

Sign Test

- A very simple alternative test to the paired t test (which compares means) is to test whether the median of the differences is zero
  - If the median of the differences is zero, we would expect as many differences to be above zero as below zero
    - The differences that are exactly zero do not contribute much information about which measurement tends to be higher

32

Median Difference

- Compute differences of observations
  - Consider whether differences tend to be negative or positive

33

Median Difference Properties

- Median difference is not difference in medians
  - Ex: X = (1, 3, 10); Y = (2, 5, 10)
    - mdn(Y) – mdn(X) = 5 – 3 = 2
    - Difference: D = X – Y = (1, 2, 0); mdn(D) = 1
- The median difference is not transitive
  - Ex: X = (1, 2, 3); Y= (2, 3, 1); Z = (3, 0, 2)
    - mdn (Y – X) = 1 > 0 (so "Y larger than X")
    - mdn (Z – Y) = 1 > 0 (so "Z larger than Y")
    - mdn (X – Z) = 1 > 0 (so "X larger than Z")

34

Sign Test (Elevator Statistics)

- Proportion positive among nonzero differences

$$X_i \overset{iid}{\sim} (\mu, \sigma^2) \quad Y_i \overset{iid}{\sim} (\nu, \tau^2) \quad D_i = X_i - Y_i \overset{iid}{\sim} (\mu - \nu, \omega^2)$$

$$P = \text{number of } D_i\text{'s} > 0$$

$$N = \text{number of } D_i\text{'s} < 0$$

If the median difference is 0, the number of positive differences is binomially distributed:

$$H_0 : P \sim B(P + N, 0.5)$$

35

Sign Test: Stata Commands

- Stata has a command to perform the sign test
  - "`signtest var1 = var2`"
    - Provides one-sided and two-sided P values
    - Does not provide any meaningful estimates or confidence intervals
  - (The sign test can also be performed by creating the differences, changing the zeroes to missing, and then using "`bitest`")

36

Sign Test: Stata Example

- Example: Change in plasma beta-carotene in placebo group

```
. signtest carot3=carot0 if dose==0
Sign test
    sign |    observed    expected
---------+------------------------
positive |           1         3.5
negative |           6         3.5
    zero |           0           0
---------+------------------------
     all |           7           7
```

37

Sign Test: Stata Example

```
One-sided tests:
Ho: mdnn of carot3 - carot0  = 0 vs.
Ha: median of carot3 - carot0  > 0
Pr(#pos >= 1) = Binomial(n=7, x>=1, p=0.5)=  0.9922

Ho: median of carot3 - carot0  = 0 vs.
Ha: median of carot3 - carot0  < 0
Pr(#neg >= 6) = Binomial(n=7, x>=6, p=0.5)=  0.0625

Two-sided test:
Ho: median of carot3 - carot0  = 0 vs.
Ha: median of carot3 - carot0  ~= 0
Pr(#pos >= 6 or #neg >= 6)                =  0.1250
```

38

Interpretation

- We can not with 95% confidence reject the null hypothesis that the median change in plasma beta-carotene levels after 9 months of treatment with placebo was 0

39

(Wilcoxon) Signed Rank Test

- The sign test is simple to perform, but it ignores a lot of information
  - Intuitively, you would expect that there is some information in the magnitude of the differences as well as the sign
    - For instance, there may be nearly as many negative differences as positive differences, but the positive differences tend to be far larger (in absolute value) than the negative differences

40

(Wilcoxon) Signed Rank Test

- The Wilcoxon signed rank test attempts to use the information about the magnitude of the differences
  - The null hypothesis of the Wilcoxon signed rank test is that
    - the number of positive and negative differences should tend to be equal, and
    - there should be no tendency for the positive differences to be further from (or closer to) zero than the negative differences

41

(Wilcoxon) Signed Rank Test

- Basic approach of the signed rank test
  - Compute the differences and rank the absolute value of the differences
  - Sum up the ranks of the positive differences
  - Under the null hypothesis of equality of distributions, the sampling distribution for that sum should be the same as randomly choosing n/2 numbers from the integers 1 to n
    - Adjustment for ties and zeroes
    - (Computers can figure this out for us)

42

Example of Signed Ranks

X	$\{9,\ 7,\ 4,\ 2,\ 37,\ 9,\ 7,\ 4\}$
Y	$\{3,\ 8,\ 4,\ 5,\ 7,\ 5,\ 9,\ 5\}$
Diff	$\{6, -1,\ 0,\ -3,\ 30,\ 4,\ -2,\ -1\}$
Ranks	$\{7, 2.5,\ 1,\ 5,\ 8,\ 6,\ 4,\ 2.5\}$

Sum of Positive Ranks : 21

43

Summary Measure

- It is not immediately clear (or easily explained) what aspect of the distributions the signed rank test is comparing
  - Can be significant because
    - Number of positive differences is unusually high
    - Mean positive difference is high
  - It provides some sort of a balance between the two

44

Interpretation

- In any case, it is clear that a significant signed rank test can only be interpreted as a difference in distributions
  - The standard error of the test statistic is based on a permutation distribution, and thus
    - is only testing equality of distributions with the appropriate type I error,
    - but because it is not a consistent test of arbitrary differences between distributions
      - the differences must be something that the signed rank test can detect

45

Stata Commands

- Stata has a command to perform the signed rank test
  - "`signrank var1 = var2`"
    - Provides one-sided and two-sided P values
    - Does not provide any meaningful estimates or confidence intervals

46

Stata Example

- Example: Change in plasma beta-carotene in placebo group

```
. signrank carot3=carot0 if dose==0
Wilcoxon signed-rank test
    sign |      obs   sum ranks    expected
---------+---------------------------------
positive |        1           1          14
negative |        6          27          14
    zero |        0           0           0
---------+---------------------------------
     all |        7          28          28
(some purely technical output omitted)
Ho: carot3 = carot0    z =  -2.197 Prob > |z| =   0.0280
```

47

Interpretation

- We can with 95% confidence reject the null hypothesis that there was no systematic trend toward increasing or decreasing plasma beta-carotene levels after 9 months of treatment with placebo
  - (Note that we were able to reject the null with the signed rank, but not the sign test.)

48

Comparing Proportions: Matched Samples

49

Matched Binary Data

- In some studies, we make comparisons of proportions across samples which are not independent
  - E.g., Cross-over studies
    - Relief of headaches from aspirin vs Tylenol
    - Each subject receives each treatment (in random order)
  - E.g., Ophthalmology studies
    - Cure of conjunctivitis: new treatment vs placebo
    - Each subject receives each treatment (randomize which eye receives new treatment)

50

Presentation of Data

- We tend to alter the format of contingency table to reflect the matched data
  - Instead of response by group, we display concordance of response in each group

Response

Treatment	+	−	
New	r	s	n
Plc	t	u	n
	m_0	m_1	n

Resp on Plc

Resp on New	+	−	
+	a	b	r
−	c	d	s
	t	u	n

Estimate

- Usual estimate of difference of proportions

Resp on Plc

Resp on New	+	−	
+	a	b	r
−	c	d	s
	t	u	n

Estimated difference in proportions

$$\frac{r}{n} - \frac{t}{n} = \frac{b-c}{n}$$

52

Analysis of Data

- The analysis of the matched data can proceed along two lines
  - Least frequently used
    - Compare proportion with response in each group taking matching into account
    - Analogous to paired t test (which would be a valid test in large samples)
  - Most often used: McNemar's test
    - Focus on the "discordant pairs" only
    - Evaluate whether discordant pairs are evenly distributed between (+, -) and (-, +)

53

McNemar's Test: Rationale

- If response were equal in the two groups, discordant pairs should be equally likely to be in either order
  - Condition on the number of discordant pairs
    - Intuitively, the number of discordant pairs does not contribute much information as to which group does better
  - Under the null hypothesis, the discordant pairs should be equally likely to be in either the "b" or the "c" cell of the contingency table
    - Use the one sample test of a binomial proportion

54

McNemar's Test

- One sample binomial test

$$\begin{array}{cc|cc|c} & & \multicolumn{2}{c}{\text{Resp on Plc}} & \\ & & + & - & \\ \hline \text{Resp on New} & + & a & b & r \\ & - & c & d & s \\ \hline & & t & u & n \end{array}$$

If response rates are equal for both treatments, under the null we would have binomial distribution

$$b \sim B(b+c, 0.5)$$

55

Stata: Exact McNemar's

- Example: Prevalence of edema vs ascites in liver data
  - Are ascites and edema equally prevalent?
    - Stata does not perform McNemar's using exact distributions, but we can get it to perform the test quite easily

56

Stata: Exact McNemar's

```
table edema ascites
----------+-----------
          |  ascites
    edema |    0     1
----------+-----------
        0 |  268     7
        1 |   20    17
----------+-----------
```

57

Stata: Exact McNemar's

```
. bitesti 27 7 0.5
N   Obs k   Exp k   Assumed p   Observed p
-----------------------------------------
27      7    13.5    0.50000      0.25926

Pr(k>= 7)           = 0.9970     (one-sided test)
Pr(k<= 7)           = 0.0096     (one-sided test)
Pr(k<= 7 or k>= 20) = 0.0192     (two-sided)
```

58

McNemar's Test

- Test statistic can be based on asymptotic distribution
  - Standardized Z statistic or (more commonly) a chi squared statistic

Under the null we would have binomial distribution

$$b \sim B(b+c, 0.5)$$

$$Z = \frac{\frac{b}{b+c} - 0.5}{\sqrt{0.25/(b+c)}} \overset{H_0}{\sim} N(0,1) \qquad \chi^2 = Z^2 = \frac{(b\text{-}c)^2}{(b+c)}$$

Stata: Large Sample

- Stata uses asymptotic theory
  - "`mcc casevar ctrlvar`"
    - mcc = matched case-control
    - Labels are by "Cases" and "Controls"
    - Provides two-sided P-values
    - Provides confidence interval for difference in proportions

60

Stata Commands: Example

- Prevalence of edema vs ascites in liver data

```
mcc edema ascites
                 Controls
Cases     |Exposed  Unexposed  |   Total
----------+--------------------+---------
  Exposed |     17         20  |      37
Unexposed |      7        268  |     275
----------+--------------------+---------
    Total |     24        288  |     312
McNemar's chi2(1)=  6.26  Pr>chi2= 0.0124
```

61

Stata Commands: Example

- Prevalence of edema vs ascites in liver data

```
Proportion with factor
Cases       .1186
Controls    .0770     [95% CI]
---------     --------------------
difference  .0417   .0061   .0772
ratio      1.5467  1.0954  2.1698
rel. diff.  .0451   .0106   .0797

odds ratio 2.8571  1.1605  7.9971 (exact)
```

62

Compare Paired t Test

```
ttest edema=ascites
Paired t test           Number of obs =      312
------------------------------------------------
Variable | Mean  St Err  t     P>|t|   [95% CI]
---------+--------------------------------------
   edema | .1186  .0183 6.469  0.0000 .0825  .1547
 ascites | .0769  .0151 5.091  0.0000 .0472  .1067
---------+--------------------------------------
    diff | .0417  .0165 2.523  0.0121 .0092  .0742
```

63

Compare Paired t Test

```
Degrees of freedom: 311

             Ho: mean diff = 0

Ha: diff < 0    Ha: diff ~= 0      Ha: diff > 0
t =  2.523      t =  2.523         t =  2.523
P < t = 0.9939  P > |t| = 0.0121   P > t = 0.0061
```

64

Comments

- It is useful to highlight the difference between the questions answered by the chi square test and McNemar's test
  - Consider test of edema and ascites
    - McNemar's test
      - Are ascites and edema equally prevalent?
    - Chi square test
      - Does the prevalence of ascites differ between subjects with and without edema?

65

Sign Test vs McNemar's Test

- McNemar's test is just the sign test performed on binary data
  - The sign test is a more general description of the procedure, and thus I prefer using that name even when using binary data
  - Hence, I introduced the word "McNemar" only because you will sometimes see it referred to in the literature
    - I wish the word "McNemar" would disappear from the literature (my brain is full)

66