

# Biost 517

## Applied Biostatistics I

.....

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

### Lecture 13: Two Sample Inference About Other Measures of Location

November 17, 2006

1

© 2002, 2003, 2005 Scott S. Emerson, M.D., Ph.D.

## Lecture Outline

.....

- General Setting
- Inference About Medians
- Wilcoxon Rank Sum Test

2

## General Setting

.....

3

## Scientific Questions

.....

- The scientific questions most often addressed using statistics
  - Quantifying the distribution of a random variable in a population
  - Association between two random variables
    - Perhaps adjusting for other variables
  - Difference in association between two variables across subgroups
    - Effect modification

4

## Detecting Associations

---

- “No association” = Independence between two variables

- Knowing the value of one random variable confers no knowledge about the other variable

- Mathematical definition of independence

Random variables  $X, Y$  are independent if for every choice of constants  $a, b$

$$\Pr(X \leq a, Y \leq b) = \Pr(X \leq a)\Pr(Y \leq b)$$

5

## Role of Sampling Plan

---

- Most often we detect associations by comparing distributions across groups
  - It is often easiest to choose sample sizes within groups defined by one variable
    - Cohort studies
    - Case-control studies
    - Interventions
  - Even in cross-sectional studies it is often easier to think about comparing groups

6

## Comparing Conditional Distn

---

- Equivalent statements of statistical independence of random variables  $X, Y$

- The distribution of  $X$  does not differ as  $Y$  varies

- The distribution of  $Y$  does not differ as  $X$  varies

Random variables  $X, Y$  are independent if for every choice of constants  $a, b_1, b_2$

$$\Pr(X \leq a | Y = b_1) = \Pr(X \leq a | Y = b_2)$$

$$\Pr(Y \leq a | X = b_1) = \Pr(Y \leq a | X = b_2)$$

(one of these conditions implies the other)

7

## Differences in Distributions

---

- There is only one way two distributions can be the same:
  - The difference between the distribution functions must be zero at every point
- There are an infinite number of ways that two distributions can differ
  - An infinite number of places to differ
  - An infinite number of values for the difference

8

## Comparing Parameters

.....

- To make it more manageable:
  - Inference about associations is usually based on showing differences between population parameters
    - IF some summary measure (e.g., mean, median) is different for two distributions,
  - THEN the distributions must be different

9

## Population Parameters

.....

- Scientific questions are typically answered by making inference about some population parameter, e.g.
  - Mean
  - Geometric mean
  - Median
  - Proportion above threshold
  - Odds above threshold
  - Hazard

10

## Measures of Association

.....

- Most often: difference or ratio of univariate parameters
  - Difference (or ratio) of means
  - Ratio of geometric mean
  - Ratio (or difference) of medians
  - Difference (or ratio) of proportions
  - Odds ratios
  - Hazard ratio (or difference)

11

## Measures of Association

.....

- Sometimes a measure of association is based on summary of bivariate distribution
  - Mean ratio of observations
  - Median difference of observations
  - Probability that a randomly chosen observation from one group might exceed one chosen from the other group

12

## Interpreting Comparisons

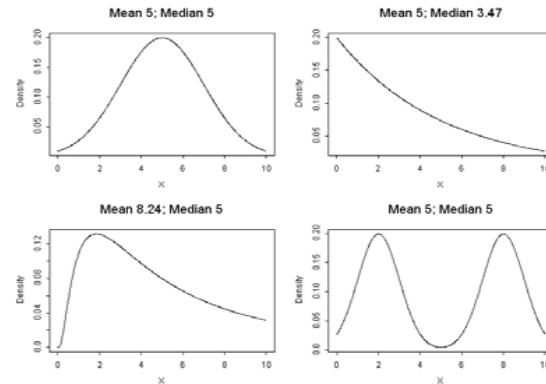
---

- Using summary measures to detect associations does require caution when interpreting results
  - Lack of a difference between population parameters for two distributions does not necessarily imply that the entire distributions are the same

13

## Comparing Distributions

---



14

## Interpreting “Negative” Results

---

- Differential diagnosis for failure to detect a difference between population parameters across two groups
  - Maybe the two distributions are the same
  - Maybe the two distributions are different (an association exists), but the population parameters are the same
  - Maybe the true value of the population parameters are different, but we lacked sufficient precision to detect it with high confidence

15

## Choice of Parameters

---

- It is therefore important to choose parameters that (in order of importance)
  - Capture scientifically important differences
  - Capture differences that will likely exist
  - Can be measured with statistical precision

16

## Inference Using the Sample Median

.....

17

## Median

.....

- Justification for use of the median
  - Scientific relevance
    - When it is most important to show effect across all subjects
      - (The mean would detect large effects that occur only in a very small subset)
  - Statistical issues
    - The sample median tends to be more efficiently estimated than the mean when the data are distributed with heavy tails

18

## Median

.....

- Approximate inference for the sample median can use asymptotic theory
  - The sample median is asymptotically normally distributed
  - The formula for the standard error is difficult to use

$$X_m \sim N\left(mdn(X), \frac{1}{4n[f(mdn(X))]^2}\right)$$

- Bootstrapping is easiest

19

## Bootstrapped Standard Errors

.....

- Bootstrapping can be used to find sampling distributions when the formulas are too difficult
  - Based on the presumption that the sample adequately represents the true distribution of data
    - Sample size is adequate

20

## Basic Strategy

- We pretend that the sample is the population
  - Sample randomly (and with replacement) from the sample to generate pseudosamples
    - Each pseudosample uses same sample size
    - Each observation equally likely to be sampled at each “draw” from the “pseudopopulation”

21

## Bootstrapped Standard Errors

- From a large number of pseudosamples, we can estimate the sampling distribution of a wide variety of statistics
  - The statistic is calculated on each pseudosample
  - Then analyze the statistics obtained across all replications of the pseudosamples
    - Recall the standard error is just the SD of a statistic computed from replicated experiments

22

## Inference with Bootstrapped SE

- Providing that we know the statistic is approximately normally distributed

100(1- $\alpha$ )% confidence interval is  $(\theta_L, \theta_U)$

$$\theta_L = \hat{\theta} - z_{1-\alpha/2} s\hat{e}(\hat{\theta})$$

$$\theta_U = \hat{\theta} + z_{1-\alpha/2} s\hat{e}(\hat{\theta})$$

Hypothesis tests based on

$$Z = \frac{\hat{\theta} - \theta_0}{s\hat{e}(\hat{\theta})} \sim N(0,1)$$

23

## Ex: SE of Sample Median

- Bootstrapped estimates of the standard error for sample median

	Data	Median
Original sample:	{1, 5, 8, 3, 7}	5
Bootstrap 1	: {1, 7, 1, 3, 7}	3
Bootstrap 2	: {7, 3, 8, 8, 3}	7
Bootstrap 3	: {7, 3, 8, 8, 3}	7
Bootstrap 4	: {3, 5, 5, 1, 5}	5
Bootstrap 5	: {1, 1, 5, 1, 8}	1

etc.

24

## 1000 Bootstrapped Samples

.....

- Descriptive statistics for the sample medians from 1000 bootstrapped samples

n	1000
Mean	4.964
Standard Deviation	1.914
Median	5
Minimum, Maximum	1, 8
25th, 75th %ile	3, 7

25

## Inference for Sample Median

.....

- From the above bootstrapped samples:
  - Estimated SE sample median is 1.914
    - The standard deviation of the sample medians across the 1000 pseudosamples
  - A 95% asymptotic (with  $n=5$ ?) confidence interval (using the 0.975 quantile of the standard normal distribution) is thus
$$5 \pm 1.96 * 1.914 = 1.25, 8.75$$

26

## Bootstrapped Standard Errors

.....

- There are some instances when bootstrapping does not work
  - For instance, no sample of continuous data is ever adequate to bootstrap the sampling distribution of the minimum or maximum
    - We can never mimic the chance to have observed more extreme values than were in our sample
  - But as a general rule, bootstrapping behaves remarkably well for measures of location and variability

27

## Median: Stata Commands

.....

- Stata has some capability to perform inference using the sample median
  - “centile” provides confidence intervals based on binomial distributions
    - Adequate for quantifying population median
    - Such CI could be used with standard normal critical values to get standard errors for use in two sample problems

28

## Stata: SE fo Sample Median

- Stata can find bootstrapped standard errors and confidence intervals
  - E.g, bootstrapped inference for median bilirubin in Primary Biliary Cirrhosis data set
    - bs “summ bili, detail” “\_result(10)”, reps(1000)
      - summ bili, detail returns the median as its 10th result
      - I want 1000 bootstrapped samples (this is overkill)
  - More complicated analyses could be done with Stata command bstrap and a Stata program

29

## Ex: Median Bili in PBC Data

- E.g, bootstrapped inference for median bilirubin in liver data set

```
Bootstrap statistics   Number of obs   =       418
                      Replications   =       1000
```

Vrble	Reps	Obs	Bias	StdErr	[95% Conf Int]
_bs_1	1000	1.4	-.0230	.0966	1.21 1.59 (N)
					1.2 1.6 (P)
					1.3 1.8 (BC)

N = normal; P = percentile; BC = bias-corrected

30

## Two Group Comparisons

- To compare medians, we would compute SE for each group individually, then use methods for combining estimates

For independent  $\hat{\theta}_1 \sim N(\theta_1, se_1^2)$ ,  $\hat{\theta}_2 \sim N(\theta_2, se_2^2)$

$$\hat{\theta}_1 + \hat{\theta}_2 \sim N(\theta_1 + \theta_2, se_1^2 + se_2^2)$$

$$\hat{\theta}_1 - \hat{\theta}_2 \sim N(\theta_1 - \theta_2, se_1^2 + se_2^2)$$

$$\hat{\theta}_1 / \hat{\theta}_2 \sim N\left(\frac{\theta_1}{\theta_2}, \frac{1}{\theta_2^2} \left( se_1^2 + \frac{\theta_1^2}{\theta_2^2} se_2^2 \right)\right)$$

31

## Wilcoxon Rank Sum Test

.....

32



## Motivation

---

- Once upon a time...
  - Computing was not as readily available
    - Finding exact distributions was more difficult
    - Taking logarithms of data was more difficult
      - (I still have a slide rule)
  - People worried about assumptions of normality

33

## Wilcoxon Rank Sum Test

---

- One idea was to downweight influence of outliers by analyzing the ranks of the data instead of the measurements themselves
  - Analogous to comparing the mean ranks

34

## Notation for Ranks

---

Data  $\{X_1, X_2, \dots, X_n\}$   
Order stats  $\{X_{(1)}, X_{(2)}, \dots, X_{(n)}\}$

Ranks  $\{R_1, R_2, \dots, R_n\}$

If  $X_i = X_{(k)} = X_{(k+1)} = \dots = X_{(k+t)}$

$$\text{then } R_i = \frac{k + (k+t)}{2}$$

(untied  $X_i$  is the  $R_i$ -th smallest :  $X_{(R_i)} = X_i$ )

35

## Example of Ranks

---

Data  $\{4, 7, 4, 2, 37, 5, 7, 4\}$

Order stats  $\{2, 4, 4, 4, 5, 7, 7, 37\}$

Ranks  $\{3, 6.5, 3, 1, 8, 5, 6.5, 3\}$

36

## Wilcoxon Rank Sum Test

.....

- Then compare average ranks for two groups
  - Exact distribution from permutation tests
    - What is the probability of obtaining a particular average rank for a group if we just mix up all the observations?
      - Draw  $n$  numbers from the integers from 1 to  $m+n$
    - A test of the null hypothesis that the two distributions are equal
  - A central limit theorem can be used in large samples

37

## Mann-Whitney Formulation

.....

- Rank sum test considers the probability that a randomly chosen subject from one group might be larger than a randomly chosen subject from the other group
  - “Pr ( $Y > X$ )”
    - Intuitive null hypothesis:  $\text{Pr} (Y > X) = 0.5$
  - Not consistent in large samples for just ANY difference in distributions, only for distributions that differ so that  $\text{Pr} (Y > X)$  is not 0.5

38

## Sampling Distribution

.....

- We only know sampling dist under complete equality of distributions
  - We can get point estimates of  $\text{Pr} (Y > X)$ , but this is not often supplied
  - We can not get general confidence intervals for the  $\text{Pr} (Y > X)$ 
    - We do not know the distribution of our test statistic under the alternative
  - We do not have power formulas for sample size computation

39

## Rank Sum Based CI

.....

- Some authors describe CI for differences in the median (or mean) based on the Wilcoxon statistic
  - However, these CI are based on the assumption that the shape of the distribution is the same for each group
    - I think this is an inappropriate assumption: It is assuming you know something that is more detailed (shape of distribution) than what you are trying to detect (general location)

40

## Stata Commands

- "ranksum varname, by(groupvar)"
  - Compares the groups indicated by binary variable groupvar
  - Provides two-sided P values
  - No relevant estimates, CI

41

## Ex: Compare Bilirubin by Sex

```
. ranksum bili, by(sex)
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
      sex |      obs   rank sum   expected
      0 |      36   6623.5     5634
-----+-----
      1 |     276  42204.5    43194
combined |     312   48828     48828
unadjusted variance  259164.00
adjustment for ties    -401.30
-----
adjusted variance    258762.70
Ho: bili(sex==0) = bili(sex==1)
      z =    1.945
      Prob > |z| =    0.0518
```

42

## Interpretation

- We do not have enough evidence to state with 95% confidence that the distribution of bilirubin is different between men and women who have Primary Biliary Cirrhosis
- We have no idea what precision we have to detect a meaningful difference
  - We have no point estimates or CI

43

## Additional Comments

- The Wilcoxon rank sum test can be shown to be "intransitive"
  - It is possible to simultaneously decide that
    - Group A tends to be higher than Group B
    - Group B tends to be higher than Group C
    - Group C tends to be higher than Group A
  - Arises because  $\Pr(Y > X)$  is intransitive
- I think that this is not very useful scientifically

44