

Biost 517

Applied Biostatistics I

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

Lecture 12: Two Sample Inference About Means

November 13, 2006

1

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

Testing vs Confidence Intervals

3

Lecture Outline

- Testing vs Confidence Intervals
- Review of Common Approach
- Exceptions to Common Approach
- Comparing Means from Independent Samples
 - Two sample t test

2

Reporting Frequentist Inference

- Three measures (four numbers)
 - Consider whether the observed data might reasonably be expected to be obtained under particular hypotheses
 - Point estimate
 - Confidence interval: all hypotheses for which the data might reasonably be observed
 - P value: probability such extreme data would have been obtained under the null hypothesis
 - Binary decision: Reject or do not reject the null according to whether the P value is low

4

Parallels Between Tests, CIs

- If the null hypothesis not in CI, reject null
 - (Using same level of confidence)
- Relative advantages
 - Test only requires sampling distn under null
 - CI requires sampling distn under alternatives
 - CI provides interpretation when null is not rejected

5

Scientific Information

- “Rejection” uses a single level of significance
 - Different settings might demand different criteria
- P value communicates statistical evidence, not scientific importance
- Only confidence interval allows you to interpret failure to reject the null:
 - Distinguish between
 - Inadequate precision (sample size)
 - Strong evidence for null

6

Hypothetical Example

- Clinical trials of new treatments for high blood pressure
 - Consider four possible scenarios
 - Measure of treatment effect is the difference in average SBP at the end of six months treatment
 - Scenarios differ in
 - Sample size
 - Variability of blood pressure
 - Treatment effect
 - (The scenarios are not replications of the same experiment or even the same scientific setting)

7

Reporting P values

Study	P value
A	0.1974
B	0.1974
C	0.0099
D	0.0099

8

Point Estimates

Study	SBP Diff	P value
A	27.16	0.1974
B	0.27	0.1974
C	27.16	0.0099
D	0.27	0.0099

9

Confidence Intervals

Study	SBP Diff	95% CI	P value
A	27.16	-14.14, 68.46	0.1974
B	0.27	-0.14, 0.68	0.1974
C	27.16	6.51, 47.81	0.0099
D	0.27	0.06, 0.47	0.0099

10

Interpreting Nonsignificance

- Studies A and B are both “nonsignificant”
 - Only study B ruled out clinically important differences
 - The results of study A might reasonably have been obtained if the treatment truly lowered SBP by as much as 68 mm Hg

11

Interpreting Significance

- Studies C and D are both statistically significant results
 - Only study C demonstrated clinically important differences
 - The results of study D are only frequently obtained if the treatment truly lowered SBP by 0.47 mm Hg or less

12

Bottom Line

- If ink is not in short supply, there is no reason not to give point estimates, CI, and P value
- If ink is in short supply, the confidence interval provides most information
 - (but sometimes a confidence interval cannot be easily obtained, because the sampling distribution is unknown under the null)

13

But: Impact of “Three over n”

- The sample size is also important
 - The pure statistical fantasy
 - The P value and CI account for the sample size
 - The scientific reality
 - We need to be able to judge what proportion of the population might have been missed in our sample
 - There might be “outliers” in the population
 - If they are not in our sample, we will not have correctly estimated the variability of our estimates
 - The “Three over n” rule provides some guidance

14

Full Report of Analysis

Study	n	SBP Diff	95% CI	P value
A	20	27.16	-14.14, 68.46	0.1974
B	20	0.27	-0.14, 0.68	0.1974
C	80	27.16	6.51, 47.81	0.0099
D	80	0.27	0.06, 0.47	0.0099

15

Interpreting a “Negative Study”

- This then highlights issues related to the interpretation of a study in which no statistically significant difference between groups was found
 - We have to consider the “differential diagnosis” of possible situations in which we might observe nonsignificance

16

General approach

- Refined scientific question
 - We compare the distribution of some response variable differs across groups
 - E.g., looking for an association between smoking and blood pressure by comparing distribution of SBP between smokers and nonsmokers
 - We base our decisions on a scientifically appropriate summary measure θ
 - E.g., difference of means, ratio of medians, ...

17

Interpreting a “Negative Study”

- Possible explanations for no statistically significant difference in θ
 - There is no true difference in the distribution of response across groups
 - There is a difference in the distribution of response across groups, but the value of θ is the same for both groups
 - (i.e., the distributions differ in some other way)
 - There is a difference in the value of θ between the groups, but our study was not precise enough
 - A “type II error” from low “statistical power”

18

Interpreting a “Positive Study”

- Analogous interpretations when we do find a statistically significant difference in θ
 - There is a true difference in the value of θ
 - There is no true difference in θ , but we were unlucky and observed spuriously high or low results
 - Random chance leading to a “type I error”
 - » The p value tells us how unlucky we would have had to have been
 - (Used a statistic that allows other differences in the distribution to be misinterpreted as a difference in θ)
 - » E.g., different variances causing significant t test)

19

Review of Most Common Approach

20

Population Parameters

- Scientific questions are typically answered by making inference about some population parameter θ
 - Quantification of population parameter
 - Mean, geometric mean, median (other quantiles), proportion/odds above threshold, hazard
 - Comparing distributions across groups
 - Difference or ratio of univariate parameters
 - Bivariate parameters
 - Mean ratio, median difference, $\Pr(Y > X)$

21

Approximate Sampling Distrn

- Most often we choose estimators that are asymptotically normally distributed

For large n : $\hat{\theta} \sim N\left(\text{mean } \theta, \text{var } \frac{V}{n}\right)$

V is related to average "statistical information" from each observation

Often : V depends on the value of θ

22

Typical Method for 100(1- α)% CI

- Set of all true means that reasonably result in the observed sample mean
 - "reasonably" = central 100(1- α)% of sampling distn
- When estimate is approximately normal 100(1- α)% confidence interval is $(\text{estimate}) \pm (\text{crit val}) \times (\text{std error})$ where the critical value is the upper $1-\alpha/2$ quantile of the standard normal distribution (or t distribution when the estimate is a sample mean)²³

(or t distribution when the estimate is a sample mean)²³

Typical Method for 100(1- α)% CI

- Set of all true means that reasonably result in the observed sample mean
 - "reasonably" = central 100(1- α)% of sampling distn

100(1- α)% confidence interval is (θ_L, θ_U)

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

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

24

t Distribution Quantiles

- Selected upper quantiles of the t distribution: $t_{df, 1-\alpha}$ (Note $t_{\infty, 1-\alpha} = z_{1-\alpha}$)

df	.005	.01	.025	.05
1	63.657	31.821	12.706	6.314
3	5.841	4.541	3.182	2.353
9	3.250	2.821	2.262	1.833
20	2.845	2.528	2.086	1.725
50	2.678	2.403	2.009	1.676
∞	2.576	2.326	1.960	1.645

25

Comparing Estimates

- Comparisons across strata or studies
 - This is easy, if estimates are independent and approximately normally distributed

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

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

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

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

27

Computing P values using Z

Standardized statistic

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

Stata commands

Lower one - sided P value

$$\text{norm}\left(\frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}\right)$$

Upper one - sided P value

$$1 - \text{norm}\left(\frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}\right)$$

Two - sided P value

$$2 \times \text{norm}\left(-\text{abs}\left(\frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}\right)\right)$$

26

Correlated Estimates

- If estimates are correlated and approximately normally distributed

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

$$\omega = \text{corr}(\hat{\theta}_1, \hat{\theta}_2)$$

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

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

28

Exceptions to Typical Approach

29

There Are Exceptions

- Not all estimators are normally distributed
 - In small samples we tend to use the t distribution when estimate is sample mean
 - If sample sizes too small, still may not be valid
 - The sample minimum and maximum are exponentially distributed in large samples
 - If data are analyzed repeatedly using a stopping rule, estimates are not normal
 - Stopping rules are used for ethical reasons in many clinical trials

30

Stopping Rules in Clinical Trials

- Early stopping of a clinical trial is often considered for ethics or efficiency.
 - Early stopping might be based on
 - Individual ethics
 - the observed statistic suggests efficacy
 - the observed statistic suggests harm
 - Group ethics
 - the observed statistic suggests equivalence
 - Exact choice will vary according to scientific / clinical setting

31

Example

- A study of premature birth is planned with 100 subjects
 - Outcome: Measure difference between treatment groups in estimated gestational age in weeks at birth
 - Monitoring plan
 - Data are analyzed after every group of 25 subjects are accrued
 - Stopping rule based on observed estimate of treatment effect

32

“O’Brien-Fleming” Stopping Rule

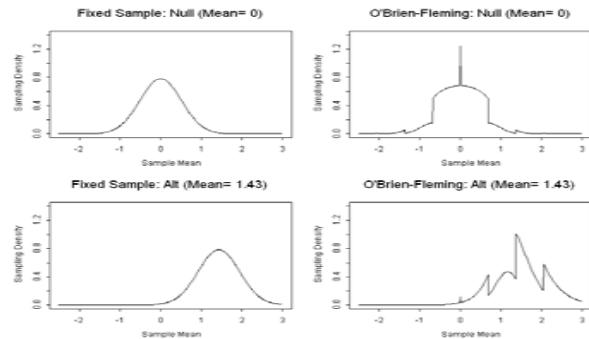
- At each analysis, stop early if estimated treatment effect is in indicated range

<u>N</u>	<u>Harm</u>	<u>Equiv</u>	<u>Efficacy</u>
25	< -4.09	----	> 4.09
50	< -2.05	(-0.006, 0.006)	> 2.05
75	< -1.36	(-0.684, 0.684)	> 1.36

33

Sequential Sampling Density

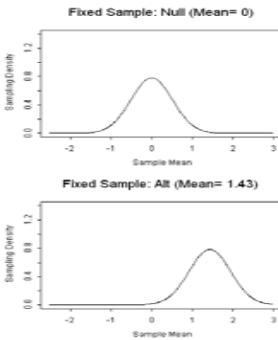
- Sampling density under stopping rule



15

Fixed Sample Sampling Distrn

- If no interim analyses



34

Special Cases

- When estimators are not normally distributed, finding confidence intervals often involves a trial and error search
 - Luckily computers can do this for us

36

Comparisons of Means From Two Independent Samples

37

Difference of Means

- Sampling distribution for sample means computed from independent samples

$$\text{For independent } \bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_Y}\right) \quad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_Y}\right)$$
$$\Downarrow$$

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_Y} + \frac{\sigma_Y^2}{n_Y}\right)$$
$$se(\bar{X} - \bar{Y}) = \sqrt{se^2(\bar{X}) + se^2(\bar{Y})}$$

38

CI Using Small Sample Correction

- We again generally use a t distribution for added conservatism in small samples

Approximate $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$ is

$$(\bar{X} - \bar{Y}) \pm t_{k,1-\alpha/2} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

where degrees of freedom k is a weighted function of the two sample sizes (many methods exist)

39

Test Statistic for Equal Means

- Hypothesis test for equality of means from two populations based on t distribution

A test of $H_0: \mu_X = \mu_Y$ can be based on

$$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \stackrel{H_0}{\sim} t_k$$

40

Test for Difference in Means

- Hypothesis test for difference of means for two populations based on t distribution

A test of $H_0 : \mu_X - \mu_Y = \Delta_0$ can be based on

$$T = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \stackrel{H_0}{\sim} t_k$$

41

T test for Unequal Variances

- The above CI and hypothesis test are referred to as “t test for unequal variances”
 - The distributional theory is only approximate in small samples (even if data are normal)
 - Many ways of handling the degrees of freedom exist
 - E.g., “Satterthwaite” or “Welch”
 - (I can’t think of anything that is less important than trying to decide the best way to choose degrees of freedom in this problem)

42

Stata: Two Sample t Test

- “ttest var, by(groupvar) unequal”
 - Tests that the mean of *var* is the same in the two groups identified by *groupvar*
 - Allows that variances might be unequal
 - Uses Satterthwaite method unless “welch” is specified
 - Provides 95% confidence intervals for each group and for the difference
 - Provides two-sided P value and both upper and lower one-sided P values

43

Ex: Change in Spd 0.075 v 0.4

```
. ttest diffspd if (dose > 0 & dose < 0.1) | dose > 0.3, by(dose) unequal
Two-sample t test with unequal variances
Group | Obs    Mean   StdErr  StdDev    [95% CI]
.075 |  26   -0.41   .269   1.37   -0.97   0.14
.400 |  20   -1.76   .486   2.17   -2.77  -0.74
comb |  46   -1.00   .275   1.87   -1.55  -0.44
diff |        1.34   .555          0.21   2.48
diff = mean(.075) - mean(.400)          t =  2.4204
Ho: diff = 0  Satterthwaite's deg free =  30.2891
Ha: diff < 0      Ha: diff != 0        Ha: diff > 0
Pr(T<t) = 0.9891  Pr(|T|>|t|) = 0.0217  Pr(T>t) = 0.0109
```

44

Ex: Analysis by Dose Group

- Spermidine change over treatment period by dose group
 - Dose 0.075
 - Average decrease of 0.41
 - 95% CI: 0.97 decrease to 0.14 increase
 - Dose 0.4
 - Average decrease of 1.76
 - 95% CI: 0.74 decrease to 2.77 increase
 - (Note overlapping CI)

45

Ex: Comparison of Dose Groups

- Difference between dose groups in Spermidine change over treatment period
 - Point estimate: 1.34
 - Decrease in 0.4 group was 1.34 more than decrease in 0.075 group
 - 95% CI: (0.21, 2.48)
 - Above observation is not atypical if true difference in average decrease between 0.21 and 2.48
 - Two -sided P value: 0.0217
 - Statistically significant diff between dose groups

46

Assuming Equal Variances

- Sometimes researchers assume that the variances for both populations are equal
 - In Stata, just drop the “unequal” option
 - I am generally against such assumptions, because the assumption is more detailed than what we are trying to find out
 - We don’t know whether the means are different across groups, so are we willing to assume we know that the variances are equal?

47

Assuming Equal Variances

- When we do make this assumption, we use a pooled estimate of the variance

100(1 - α)% CI for $\mu_X - \mu_Y$ is

$$(\bar{X} - \bar{Y}) \pm t_{k,1-\alpha/2} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

where s_p^2 is a pooled estimate of variance

$$s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$$

and degrees of freedom $k = n_X + n_Y - 2$

48

Assuming Equal Variances

- The CI and hypothesis test based on the “t test for equal variances” is exact if the data are normally distributed and the population variances are truly equal
 - However, they are not valid inference about the mean if the variances are not equal
 - Thus a rejection of the null hypothesis can correspond to rejection of equality of distributions with 95% confidence, but not rejection of equality of means with 95% confidence

49

Equal vs Unequal Variances

- Comparisons of CI for equal and unequal variances
 - If $n=m$ or the sample variances are equal, there is no difference in the SE estimates
 - There may be differences in the critical value used
 - If variances are truly equal, you lose precision if you assume they are not
 - If variances are not equal, you lose accuracy in statements about your level of confidence if you assume they are

50

If Incorrectly Assume Equal Var

- For inference about the mean
 - Smaller sample size in the group with larger variance leads to anti-conservative inference
 - CI too narrow, P values too small; you reject the hypothesis of equal means too often (type I error larger than you think)
 - Larger sample size in the group with larger variance leads to conservative inference
 - CI too wide; P values too small; type I error smaller than you think

51