

**Biost 517**  
**Applied Biostatistics I**  
.....  
Scott S. Emerson, M.D., Ph.D.  
Professor of Biostatistics  
University of Washington

**Lecture 11:**  
**Generalizations of One Sample  
Inference About Means**

November 7, 2007

1

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

**Lecture Outline**  
.....

- Inference for Mean Difference
- Inference for Binomial Proportions
- Inference for Poisson Rates
- Inference for Geometric Means

2

**Inference About Means From  
Matched Samples**  
.....

3

**Inference for Associations**  
.....

- Previously we considered inference about the mean of a distribution within a single group
  - Limited application, because we rarely have some absolute hypothesis about the value of a population parameter
  - Exception: means of differences or ratios
    - Natural comparison of differences to 0 and ratios to 1

4

## Precision of Inference

- Recall standard error of sample mean from independent variables depends on:
  - Variance of measurements within group
  - Sample size

$$se(\bar{Y}) = \sqrt{\frac{\text{Var}(Y_i)}{n}}$$

5

## Increased Precision

- Difference in means across groups can be estimated by mean difference
  - Comparisons within a pair of positively correlated subjects leads greater precision
    - Adjusting for a highly predictive random effect
      - Correlation of matched measurements near 1

Variance of difference with matched samples :

$$\text{Var}(W - X) = \text{Var}(W) + \text{Var}(X) - 2\rho\sqrt{\text{Var}(W)\text{Var}(X)}$$

Variance of difference with independent t samples :

$$\text{Var}(W - X) = \text{Var}(W) + \text{Var}(X)$$

6

## Matched Samples

- Many studies make use of matched samples to study associations
  - E.g., cross-over studies in which each subject receives both treatments *in random order*
  - E.g., “split-plot” designs in which each subject receives both treatments in different locations
    - Eye disease, skin disease
  - E.g., matched subjects in which one of each pair receives a treatment
    - Twin studies, matched communities

7

## Collapsing Data on Subjects

- So far: Inference assuming independent measurements
- When we take several measurements on each subject, we often combine them
  - Take difference between matched data
  - Subjects are independent

8

## Paired Differences

- Measurements  $W_i, X_i$  on  $i$ -th subject made under different conditions to be compared
  - Note difference of means  $E(W) - E(X)$  is the same as the mean difference  $E(W-X)$

For the  $i$ -th subject :

$$W_i \sim (\gamma, \omega^2) \quad X_i \sim (\theta, \tau^2) \quad \text{corr}(W_i, X_i) = \rho$$

$$\text{Difference } D_i = W_i - X_i \sim (\mu, \sigma^2)$$

$$\mu = \gamma - \theta$$

$$\sigma^2 = \omega^2 + \tau^2 - 2\rho\omega\tau$$

9

## Inference on Paired Differences

- Scientific (and statistical) questions relate to distribution of paired differences
  - Estimate / test  $\mu$  = mean of differences using one sample inference about means

10

## Statistics on Differences

- Sample mean, sample variance of differences

For the  $i$ -th subject :  $W_i, X_i$

Compute differences  $D_i = W_i - X_i$

Summary statistics

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i = \frac{(D_1 + \dots + D_n)}{n}$$

$$s_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$$

11

## Inference on Differences

- Inference for  $\mu = E(W - Y) = E(W) - E(Y)$

Point estimate :  $\hat{\mu} = \bar{D}$

100(1 -  $\alpha$ )% CI for  $\mu$  :  $\bar{D} \pm \frac{s_D}{\sqrt{n}} t_{n-1, 1-\alpha/2}$

P values based on :  $\Pr \left( t_{n-1} \leq \frac{\bar{D} - \mu_0}{s_D / \sqrt{n}} \right)$

12

## Stata: Paired t test

- Paired t test is default when you specify two variables
  - "ttest var1 = var2"
    - Tests that the mean of var1 equals the mean of var2 where measurements are made on matched samples
      - Obviously requires data in "wide" format
        - » Rows in your dataset correspond to same subjects
    - Also gives point estimates and 95% CI

13

## Example: SEP data

- Compare n35 peaks on right and left
  - (Why? Should we consider dominant side?)

```
. ttest n35R=n35L
Paired t test
Var | Obs   Mean   StdErr   StdDev   [95% ConfInt]
n35R | 250  35.007   .230     3.639   34.554  35.460
n35L | 250  35.178   .232     3.667   34.722  35.635
diff | 250   -.172    .130     2.054   -.427   .085
mean(diff) = mean(n35R - n35L)      t = -1.3178
Ho: mean(diff) = 0                  deg of fr = 249
Ha: mn(dff) < 0                    Ha: mn(dff) != 0      Ha: mn(diff) > 0
Pr(T<t) = 0.0944                    Pr(|T|>|t|) = 0.1888   Pr(T > t) = 0.9056
```

14

## Example: Interpretation

- Estimate delay of 35.007 msec on R; 35.178 msec on L
  - Difference of 0.172 msec higher on L
  - 95% CI: Such a difference is not unexpected if the true difference were between .427 msec higher on L to .085 higher on R
  - Based on two-sided P value: We would not reject null hypothesis of equal means
    - Two-sided because no reason to presuppose one side higher than other and no different action

15

## Inference for Paired Ratios

- Could look at ratio of paired observations
  - Less stable if denominators near 0
- BUT: Ratio of means is not the mean ratio
  - Consider paired observations (Y,X)
    - (4, 2) (8, 1) (12, 3) (16, 5) (20, 4)
    - $E(Y) = 60 / 5 = 12$ ;  $E(X) = 15 / 5 = 3$
    - $E(Y) / E(X) = 12 / 3 = 4$
  - Consider ratios Y / X
    - 2 8 4 3.2 5
    - $E(Y / X) = 22.2 / 5 = 4.44$

16

## Stata: Inference for Ratios

- In Stata, we would create ratios

```
• g n35ratio= n35R / n35L
• ttest n35ratio=1
```

One-sample t test

Variable	Obs	Mean	StdErr	StdDev	[95% Conf Int]
n35ratio	250	.997	.00374	.0591	.990 1.004

```
mean = mean(n35ratio)          t = -0.8371
Ho: mean = 1                 deg of freedom = 249
```

```
Ha: mean < 1                Ha: mean != 1                Ha: mean > 1
Pr(T<t)= 0.2017  Pr(|T|>|t|)= 0.4033  Pr(T>t)= 0.7983
```

## Inference for Binomial Proportions

Large Samples  
(Uncensored)

18

## Binary Random Variables

- Many variables can take on two values
  - For convenience code as 0 or 1
    - Vital status: “Dead” 0 is (alive) or 1 (dead)
    - Sex: “Female” is 0 (male) or 1 (female)
    - Intervention: “Tx” is 0 (control) or 1 (new therapy)
- Sometimes dichotomize variables
  - For scientific reasons (statistically less precise)
    - Blood pressure less than 160 mm Hg
    - PSA less than 4 ng/ml
    - Serum glucose less than 120 mg/dl

19

## Statistical Hypotheses

- Scientific questions translated into a statistical question about parameter  $p$ 
  - Binary variable has Bernoulli (binomial) distn
    - $p$  is the proportion of the population with the random variable equal to 1
    - $p$  is also the population mean for the random variable

20

## Point Estimate

- Use the sample mean

Data  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} B(1, p)$   $E(X_i) = p$   $Var(X_i) = p(1-p)$

Point estimate:  $\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + \dots + X_n}{n}$

21

## Approximate Distribution

- Use the central limit theorem

Data  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} B(1, p)$   $E(X_i) = p$   $Var(X_i) = p(1-p)$

$$\hat{p} = \bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

– NOTE: A mean – variance relationship

22

## Continuity Correction

- Also, the number of events is discrete
  - In one sample problem we often make a continuity correction

$$\Pr\left(\hat{p} \leq \frac{k}{n}\right) = \Pr\left(\hat{p} \leq \frac{k+0.5}{n}\right)$$

$$\Pr\left(\hat{p} \geq \frac{k}{n}\right) = \Pr\left(\hat{p} \geq \frac{k-0.5}{n}\right)$$

23

## Asymptotic CI: Best Approach

- We do best by considering mean-variance relationship and continuity correction
  - Requires quadratic formula or iterative search

100(1- $\alpha$ )% CI for  $p$ :  $(\hat{p}_L, \hat{p}_U)$

$$\hat{p}_L = \hat{p} - \frac{1}{2n} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}_L(1-\hat{p}_L)}{n}}$$

$$\hat{p}_U = \hat{p} + \frac{1}{2n} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}_U(1-\hat{p}_U)}{n}}$$

24

## Asymptotic CI: Elevator Stats

- Often we can just use best estimate of  $p$  in standard error for confidence intervals and ignore the continuity correction
  - $np$  and  $n(1-p)$  must be large

$$100(1 - \alpha)\% \text{ CI for } p: \quad \hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

25

## Asymptotic P values: Best

- We do best by considering mean-variance relationship and continuity correction

P values for  $H_0: p = p_0$ :

$$\text{Lower one - sided P:} \quad P_{lower} = \Pr \left( Z \leq \frac{\hat{p} + \frac{1}{2n} - p_0}{\sqrt{p_0(1-p_0)/n}} \right)$$

$$\text{Upper one - sided P:} \quad P_{upper} = \Pr \left( Z \geq \frac{\hat{p} - \frac{1}{2n} - p_0}{\sqrt{p_0(1-p_0)/n}} \right)$$

$$\text{Two - sided P:} \quad 2 \times \min(P_{lower}, P_{upper}, 0.5)_{26}$$

## Asymptotic P values: Elevator

- We still consider mean-variance relationship but ignore continuity correction

P values for  $H_0: p = p_0$ :

$$\text{Lower one - sided P:} \quad P_{lower} = \Pr \left( Z \leq \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right)$$

$$\text{Upper one - sided P:} \quad P_{upper} = \Pr \left( Z \geq \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right)$$

$$\text{Two - sided P:} \quad 2 \times \min(P_{lower}, P_{upper}, 0.5)_{27}$$

## Stata: Asymptotic Inference

- Stata explicitly provides exact inference
  - If we want asymptotic inference, we could
    - Compute standard errors, Z statistics
    - Use "norm( )" function to get P values
  - But why not just use exact inference
    - It is better

28

## Inference for Binomial Proportions

.....

### Exact Inference (Uncensored)

29

## Exact Distribution

.....

- Here, we do not have to rely on asymptotic theory
  - A binary variable must be Bernoulli
  - Sums of independent Bernoulli random variables must be binomial
  - We can use the exact binomial distribution to compute our probabilities
    - (Well, computers can)

30

## Binomial Distribution

.....

- Probability theory provides a formula for the distribution of binomial random variables

Data  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} B(1, p)$

↓

$Y = \sum_{i=1}^n X_i = X_1 + \dots + X_n \sim B(n, p)$

For  $k = 0, 1, \dots, n$ :  $\Pr(Y = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$

31

## Exact Point Estimate

.....

- Still use the sample mean

Data  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} B(1, p)$      $E(X_i) = p$      $\text{Var}(X_i) = p(1-p)$

Point estimate:     $\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + \dots + X_n}{n}$

32



## Exact Confidence Intervals

- Use the binomial distribution
  - (But let a computer do it for you)

An exact  $100(1 - \alpha)\%$  confidence interval for  $p$  based on observation  $Y = k$  is  $(\hat{p}_L, \hat{p}_U)$  where an iterative search is used to find

$$\Pr[Y \leq k; \hat{p}_U] = \sum_{i=0}^k \frac{n!}{i!(n-i)!} \hat{p}_U^i (1 - \hat{p}_U)^{n-i} = \alpha/2$$

$$\Pr[Y \geq k; \hat{p}_L] = \sum_{i=k}^n \frac{n!}{i!(n-i)!} \hat{p}_L^i (1 - \hat{p}_L)^{n-i} = \alpha/2$$

33

## Stata: Exact CI for Proportion

- Syntax
  - `"ci varlist, binomial"`
    - Provides exact confidence intervals
    - (Standard errors are based on asymptotics)

34

## Ex: Relapse, Nadir PSA

- PSA dataset: Relapse in 24 months
  - Generating variables of interest

```
. g relapse24=0
. replace relapse24=1 if inrem=="no" & obstime <= 24

. g nadirge2= nadir
. recode nadirge2 min/2=0 2/max=1
```

35

## Ex: CI for Prevalence

- Prevalence of relapse in 24 months

```
. ci relapse24, binomial
```

Variable	Obs	Mean	StdErr	Binomial Exact	
				[95% ConfInt]	
relapse24	50	.44	.070	.300	.587

36

## Ex: CI for 1-Specificity, Sensitivity

- 1-Specificity, Sensitivity of Nadir PSA > 2 for relapse within 24 months

```
. bysort relapse24: ci nadirge2, binomial
-> relapse24 = 0
```

Variable	Obs	Mean	StdErr	[95% Conf Int]	Binomial Exact
nadirge2	28	.143	.066	.040 .327	

```
-> relapse24 = 1
```

Variable	Obs	Mean	StdErr	[95% Conf Int]	Binomial Exact
nadirge2	22	.682	.099	.451 .861	

37

## Ex: Interpretation

- The observed prevalence of relapse within 24 months of 44% was not unusual if the true prevalence were between 30.0% and 58.7%
  - » With 95% confidence reject Prev < 30.0% or >58.7%
- The observed sensitivity of 68.2% was not unusual if the true sensitivity were between 45.1% and 86.1%
- The observed specificity of 85.7% was not unusual if the true specificity were between 67.3% and 96.0%

38

## Compare to Asymptotic CIs

- Compare exact results to asymptotic CI using t statistics
  - Normally we would use Z statistics
    - Std errors differ by square root of (n / n-1)
    - Critical value differs according to df

39

## Compare to Asymptotic CIs

```
. ci relapse24
```

Variable	Obs	Mean	StdErr	[95% ConfInt]
relapse24	50	.44	.071	.297 .583

```
. bysort relapse24: ci nadirge2
-> relapse24 = 0
```

Variable	Obs	Mean	StdErr	[95% ConfInt]
nadirge2	28	.143	.067	.005 .281

```
-> relapse24 = 1
```

Variable	Obs	Mean	StdErr	[95% ConfInt]
nadirge2	22	.682	.102	.470 .893

40

## Elevator Stats: 0 events in n trials

- Two-sided confidence intervals fail in the case where there are either 0 or n events observed in n Bernoulli trials
  - If  $Y=0$ , there is no lower confidence bound
  - If  $Y=n$ , there is no upper confidence bound
- We can, however, derive one-sided confidence bounds in that case

41

## Upper Conf Bnd for 0 Events

- Exact upper confidence bound when all observations are 0

Suppose  $Y \sim B(n, p)$  and  $Y = 0$  is observed

Exact  $100(1 - \alpha)\%$  upper confidence bound for  $p$  is  $\hat{p}_U$

$$\Pr[Y = 0; \hat{p}_U] = (1 - \hat{p}_U)^n = \alpha$$

↓

$$\hat{p}_U = 1 - \alpha^{1/n}$$

42

## Large Sample Approximation

$$(1 - \hat{p}_U)^n = \alpha \Rightarrow n \log(1 - \hat{p}_U) = \log(\alpha)$$

For small  $\hat{p}_U$        $\log(1 - \hat{p}_U) \approx -\hat{p}_U$

so for large  $n$      $\Rightarrow \hat{p}_U \approx -\frac{\log(\alpha)}{n}$

43

## Elevator Stats: 0 Events in n trials

- “Three over n rule”
  - $\log(.05) = -2.9957$
  - In large samples, when 0 events observed, the 95% upper confidence bound for  $p$  is approximately  $3 / n$
- 99% upper confidence bound
  - $\log(.01) = -4.605$
  - Use  $4.6 / n$  as 99% upper confidence bound

44

## Elevator Stats vs Exact

- When  $X=0$  events observed in  $n$  Bernoulli trials

n	95% bound		99% bound	
	Exact	3/n	Exact	4.6/n
2	.7764	1.50	.9000	2.3000
5	.4507	.60	.6019	.9200
10	.2589	.30	.3690	.4600
20	.1391	.15	.2057	.2300
30	.0950	.10	.1423	.1533
50	.0582	.06	.0880	.0920
100	.0295	.03	.0450	.0460

45

## Elevator Stats: n Events in n trials

- We can also use the “Three over n rule” to find the lower confidence bound for  $p$  when every subject has an event
  - Lower 95% confidence bound is  $1 - 3/n$

46

## Exact Tests for a Proportion

- Use binomial distribution under the null
  - (But let a computer do it for you)

For  $Y \sim B(n, p)$  and observation  $Y = k$ :

Test  $H_0: p = p_0$ , calculate P values by

$$\text{Upper one-sided: } P_{upper} = \Pr[Y \geq k; p_0] = \sum_{i=k}^n \frac{n!}{i!(n-i)!} p_0^i (1-p_0)^{n-i}$$

$$\text{Lower one-sided: } P_{lower} = \Pr[Y \leq k; p_0] = \sum_{i=0}^k \frac{n!}{i!(n-i)!} p_0^i (1-p_0)^{n-i}$$

$$\text{Two-sided (easy): } 2 \times \min(P_{lower}, P_{upper}, 0.5)$$

47

## Stata: Tests for Proportion

- Syntax
  - “bitest var = #p”
  - Provides exact test that proportion = #p
  - Gives upper and lower one-sided, two-sided P values
    - Two-sided P value is computed under a slightly more complicated rule, but is valid

48

## Ex: Prevalence of Relapse

- Relapse in 24 months in PSA data
  - Test prevalence of 40% (Why?)

```
. bitest relapse24=0.4
```

Variable	N	Obs k	Exp k	Assumed p	Obs p
relapse24	50	22	20	0.400	0.440

```
Pr(k >= 22) = 0.3299 (one-sided test)
Pr(k <= 22) = 0.7660 (one-sided test)
Pr(k <= 17 or k >= 22) = 0.5668 (two-sided test)
```

49

## Interpretation

- Two-sided inference
  - With 95% confidence, we cannot reject the hypothesis that the true prevalence of relapse within 24 months is 40% (P= 0.57; 95% CI 30.0% to 58.7%)

50

## Exact vs Asymptotic (T test)

- Differences between asymptotic and t test
  - Mean-variance relationship
    - t test would use estimated proportion in standard error instead of hypothesized
  - Computation of standard deviation
    - t test would divide by n-1 to get variance
  - Critical values
    - t test uses t distribution instead of standard normal
- In very large samples none of these make a difference

51

## Exact vs Asymptotic (T test)

```
. ttest relapse24=0.4
One-sample t test
Variable | Obs  Mean  StdErr  StdDev  [95% Conf Int]
relapse24 | 50   .44   .071   .501   .297   .583

mean = mean(relapse24)          t = 0.5641
Ho: mean = 0.4          degrees of freedom = 49

Ha: mean < 0.4   Ha: mean != 0.4   Ha: mean > 0.4
Pr(T<t)=0.7124   Pr(|T|>|t|)=0.5753   Pr(T>t)=0.2876
```

52

## Inference for Binomial Proportions

.....

### Large Samples (Censored)

53

## Dichotomized Continuous Data

.....

- Scientifically it is sometimes of interest to summarize a distribution by the probability of exceeding some threshold
  - E.g., cholesterol greater than 200
  - E.g., survival past 5 years
- Statistically it is sometimes most convenient to do so
  - In right censored data, the mean or median might not be estimable

54

## Inferential Approach

.....

- In the absence of censoring
  - Create dichotomized data
  - Inference as just described
    - Exact versus approximate
- In the presence of right censoring
  - We must use Kaplan-Meier estimates

55

## Right Censored Data

.....

- In the presence of right censored data, we use Kaplan-Meier curves to estimate proportions exceeding a threshold
  - KM estimates asymptotically normally distributed
    - Mean is true proportion
    - Standard error depends on true proportion, sample size, and censoring distribution
      - “Greenwood’s Formula”

56

## Right Censored Data

- Notation:

Unobserved:

True times to event :  $\{T_1^0, T_2^0, \dots, T_n^0\}$

Censoring Times :  $\{C_1, C_2, \dots, C_n\}$

Observed data :

Observation Times :  $T_i = \min(T_i^0, C_i)$

Event indicators :  $D_i = \begin{cases} 1 & \text{if } T_i = T_i^0 \\ 0 & \text{otherwise} \end{cases}$  57

## Kaplan-Meier Notation

- Definition of intervals, number at risk, failures

Ordered distinct observation times :

$$t_1 \leq t_2 \leq \dots \leq t_k$$

Time interval :  $(t_{j-1}, t_j]$

Number at risk at  $t_j$  :  $N_j$

Number of events at  $t_j$  :  $D_j$  58

## Kaplan-Meier Hazard Estimates

- Computation of hazard and conditional probability of survival in interval

Hazard for event in interval :  $\frac{D_j}{N_j}$

Conditional probability of survival in interval :

$$\Pr(T^0 \geq t_j | T^0 \geq t_{j-1}) = 1 - \frac{D_j}{N_j}$$
59

## Kaplan-Meier Survival Estimate

- Estimating survival probability

$$S(t) = \Pr(T^0 > t)$$

Cumulative probability of survival :

$$\Pr(T^0 > t_j) = \Pr(T^0 > t_j | T^0 > t_{j-1}) \Pr(T^0 > t_{j-1})$$

$$\hat{S}(t_j) = \left(1 - \frac{D_j}{N_j}\right) \times \left(1 - \frac{D_{j-1}}{N_{j-1}}\right) \times \dots \times \left(1 - \frac{D_1}{N_1}\right)$$

$$= \prod_{i=1}^j \left(1 - \frac{D_i}{N_i}\right)$$
60

## Std Err: Greenwood's Formula

- Fairly technical, but for statisticians...
  - Hazard estimate is a proportion:  $D_j / N_j$
  - Variance of hazard estimate from theory about binomial proportions
  - Delta method to get variance of  $\log(1 - D_j / N_j)$
  - Then use properties of expectation to get variance of  $\log S(t) = \sum \log(1 - D_j / N_j)$ 
    - Noninformative censoring leads to asymptotically uncorrelated hazard estimates
  - Use delta method to get variance of  $S(t)$
  - Standard error is square root of variance of  $S(t)$

61

## Approximate Distribution

- Suppose interested in  $p = Pr(T^0 \geq c)$  in presence of right censoring

$$\hat{S}(c) \sim N\left( S(c), [se(\hat{S}(c))]^2 \right)$$

62

## Point Estimate

- Suppose interested in  $p = Pr(T^0 \geq c)$  in presence of right censoring

$$\hat{S}(c) \sim N\left( S(c), [se(\hat{S}(c))]^2 \right)$$

Point estimate:  $\hat{p} = \hat{S}(c)$

63

## CI Using Greenwood's Formula

- Suppose interested in  $p = Pr(T^0 \geq c)$  in presence of right censoring

$$\hat{S}(c) \sim N\left( S(c), [se(\hat{S}(c))]^2 \right)$$

100(1 -  $\alpha$ )% Confidence Interval for  $p = S(c)$ :

$$\hat{S}(c) \pm z_{1-\alpha/2} se(\hat{S}(c))$$

64



## Other Methods for CI

- CI constructed with Greenwood's formula sometimes go beyond 0 or 1
  - (This can happen with asymptotic CI with uncensored data, as well)
- If we construct CI based on  $\log(-\log S(t))$  this won't happen
  - Some statistical programs will give you these CI instead

65

## Hypothesis Tests

- Testing null hypothesis  $H_0: p = p_0$  in presence of right censoring

$$\hat{S}(c) \sim N\left( S(c), [se(\hat{S}(c))]^2 \right)$$

$$\text{Lower one-sided P value: } P_{lower} = \Pr\left( Z \leq \frac{\hat{S}(c) - p_0}{se(\hat{S}(c))} \right)$$

$$\text{Lower one-sided P value: } P_{upper} = \Pr\left( Z \geq \frac{\hat{S}(c) - p_0}{se(\hat{S}(c))} \right)$$

$$\text{Two-sided P value: } P_{two} = 2 \times \min(P_{lower}, P_{upper})$$

66

## Example: PSA Data

- Men with prostate cancer
  - Hormonal treatment
  - Followed for signs of progression
- Interested in estimating probability of remaining in remission for three years
  - Testing hypothesis that three year survival probability is 50%
    - (Where did this hypothesis come from?)

67

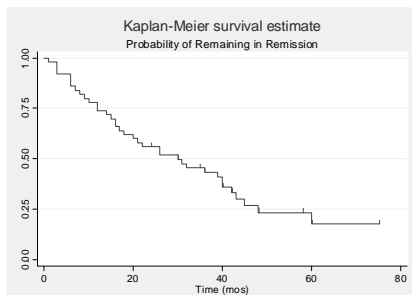
## Example: Stata Commands

- Preparing data
  - infile ... obstime **str8 inrem** using psa.txt
  - g relapse = 0
  - replace relapse = 1 if inrem=="no"
- "Setting" survival variable
  - stset obstime relapse
- Kaplan-Meier estimates
  - sts graph, xtitle("Time from Treatment (mos)")
  - sts list

68

## Stata: KM Graph

- `sts graph, cens(s) xtitle("Time (mos)") t1("Probability of Remaining in Remission")`



69

## Stata: KM Listing

- `sts list`

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]	
1	50	1	0	0.9800	0.0198	0.8664	0.9972
3	49	3	0	0.9200	0.0384	0.8007	0.9692
6	46	3	0	0.8600	0.0491	0.7286	0.9307
7	43	1	0	0.8400	0.0518	0.7054	0.9166
8	42	1	0	0.8200	0.0543	0.6826	0.9020
9	41	1	0	0.8000	0.0566	0.6602	0.8870
10	40	1	0	0.7800	0.0586	0.6381	0.8716
12	39	2	0	0.7400	0.0620	0.5947	0.8399
14	37	1	0	0.7200	0.0635	0.5735	0.8236
15	36	1	0	0.7000	0.0648	0.5525	0.8070
16	35	2	0	0.6600	0.0670	0.5114	0.7730
17	33	1	0	0.6400	0.0679	0.4911	0.7557

--more--

70

## Stata: KM Listing

- `sts list, at(24 27 30 33 36)`

Time	Beg. Total	Fail	Survivor Function	Std. Error	[95% Conf Int]	
24	28	22	0.5600	0.0702	0.4124	0.6842
27	27	2	0.5185	0.0709	0.3725	0.6461
30	25	1	0.4978	0.0710	0.3529	0.6267
33	22	2	0.4545	0.0711	0.3124	0.5860
36	20	1	0.4318	0.0711	0.2913	0.5645

71

## Stata: Two-sided P value

```
disp 2 * norm(- abs( ( 0.4318 - 0.5000) /
0.0711))
.33745177
```

72

## Interpretation

.....

- The Kaplan-Meier estimate of remaining in remission for 3 years after hormonal treatment of prostate cancer is 0.432.
- With 95% confidence, such an observation is not consistent with a true probability less than 0.291 or greater than .565.
- Based on the P value of 0.337, we cannot reject the hypothesis that 50% of hormonally treated men would remain in remission for 3 years.

73

## Inference for Rates

.....

74

## Incidence Rates

.....

- In some studies, we make inference about rates of some event over space and / or time
  - E.g., Estimation of cancer incidence rates
    - Number of new cases of cancer diagnosed per person – year of observation
  - E.g., Number of colon polyps that grow in a person during a 3 year period
  - E.g., Number of respiratory tract infections in cystic fibrosis patients

75

## Incidence Rates

.....

- A mean, normalized to a standard period of time and a standard area of space (population)
  - Most often, inference is based on a probability model involving the Poisson distribution
    - Assumptions that lead to Poisson
      - In a small interval of space and time, only one event can occur
      - The number of events occurring in nonoverlapping intervals are independent
    - Alternatively, Poisson approximation to binomial

76

## Incidence Rates: Data

- Typically, the data for incidence rate data consist of
  - Length of time-space interval a subject is under observation
    - E.g., “Person – years” of observation
  - Number of events observed in that subject
  - Quite often, aggregate data is all that is presented
    - Total person – years of observation
    - Total number of events across subjects

77

## Point Estimate

- Use the “sample mean”

Data  $X_1, \dots, X_n$  independent t with  $X_i \sim P(\lambda t_i)$  ( $t_i$  known)

$$E(X_i) = \lambda t_i \quad \text{Var}(X_i) = \lambda t_i$$

$$Y = \sum_{i=1}^n X_i \sim P(\lambda_0 t) \text{ with } t = \sum_{i=1}^n t_i$$

Point estimate :  $\hat{\lambda} = \frac{Y}{t}$

78

## Approximate Distribution

- From central limit theorem

Data  $X_1, \dots, X_n$  independent t with  $X_i \sim P(\lambda t_i)$  ( $t_i$  known)

$$E(X_i) = \lambda t_i \quad \text{Var}(X_i) = \lambda t_i$$

$$Y = \sum_{i=1}^n X_i \sim P(\lambda_0 t) \text{ with } t = \sum_{i=1}^n t_i$$

$$\hat{\lambda} = \frac{Y}{t} \sim N\left(\lambda, \frac{\lambda}{t}\right)$$

79

## Continuity Correction

- As with the binomial distribution, the number of events is discrete
  - We do not usually bother with the continuity correction, but it would make sense

$$\Pr\left(\hat{\lambda} \leq \frac{k}{t}\right) = \Pr\left(\hat{\lambda} \leq \frac{k+0.5}{t}\right)$$

$$\Pr\left(\hat{\lambda} \geq \frac{k}{t}\right) = \Pr\left(\hat{\lambda} \geq \frac{k-0.5}{t}\right)$$

80

### Asymptotic CI: Best Approach

- We do best by considering mean-variance relationship and continuity correction
  - Requires quadratic formula or iterative search

100(1 -  $\alpha$ )% CI for  $\lambda$ :  $(\hat{\lambda}_L, \hat{\lambda}_U)$

$$\hat{\lambda}_L = \hat{\lambda} - \frac{1}{2t} - z_{1-\alpha/2} \sqrt{\frac{\hat{\lambda}_L}{t}}$$

$$\hat{\lambda}_U = \hat{\lambda} + \frac{1}{2t} + z_{1-\alpha/2} \sqrt{\frac{\hat{\lambda}_U}{t}}$$

81

### Asymptotic CI: Elevator Stats

- Often we can just use best estimate of  $\lambda$  in standard error for confidence intervals and ignore the continuity correction
  - number of events and  $t$  must be large

100(1 -  $\alpha$ )% CI for  $\lambda$ :  $\hat{\lambda} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\lambda}}{t}}$

82

### Asymptotic P values: Best

- We do best by considering mean-variance relationship and continuity correction

P values for  $H_0 : \lambda = \lambda_0$ :

Lower one - sided P:  $P_{lower} = \Pr\left(Z \leq \frac{\hat{\lambda} + \frac{1}{2t} - \lambda_0}{\sqrt{\lambda_0/t}}\right)$

Upper one - sided P:  $P_{upper} = \Pr\left(Z \geq \frac{\hat{\lambda} - \frac{1}{2t} - \lambda_0}{\sqrt{\lambda_0/t}}\right)$

Two - sided P:  $2 \times \min(P_{lower}, P_{upper}, 0.5)$

### Asymptotic P values: Elevator

- We still consider mean-variance relationship but ignore continuity correction

P values for  $H_0 : \lambda = \lambda_0$ :

Lower one - sided P:  $P_{lower} = \Pr\left(Z \leq \frac{\hat{\lambda} - \lambda_0}{\sqrt{\lambda_0/t}}\right)$

Upper one - sided P:  $P_{upper} = \Pr\left(Z \geq \frac{\hat{\lambda} - \lambda_0}{\sqrt{\lambda_0/t}}\right)$

Two - sided P:  $2 \times \min(P_{lower}, P_{upper}, 0.5)$

## Stata Commands

- “ir *countvar timevar*”
  - ir = “incidence rates”
  - timevar = person – years (or area)

85

## Exact Inference

- In the one sample problem, exact inference is possible
  - It is not as common to use exact inference for Poisson rates, however
    - Usually considering Poisson approximation to the binomial
    - Most often we are in a two (or more) sample setting

86

## Incidence Rates: Comments

- The assumption that incidence rate data might follow the Poisson distribution is a very strong one
  - Usually the rate is changing over time, which causes the data to be more variable than the Poisson analysis might allow for
  - But many times, the real reason we are using a Poisson analysis is just as an approximation to the binomial distribution in the presence of a very low probability of event

87

## Inference for Geometric Means

88

## Scientific Indications

- Inference for the geometric mean is sometimes based on scientific issues
  - For some measurements, proportionate change is more important than additive differences
    - E.g., doubling of creatinine is more indicative of loss of kidney function than is the difference in creatinine measurements
    - E.g., the clinical relevance of a change in PSA from 4 to 40 is more similar to a change from 400 to 4000 than from 400 to 436

89

## Statistical Indications

- But, the use of the geometric mean rather than the mean is most often based on statistical issues
  - Relative to the mean, the geometric mean
    - Tends to downweight outlying observations
    - Tends to stabilize variance across groups when the original data has SD proportional to the means
    - Tends to be better behaved when comparisons across groups are to be based on ratios

90

## Inferential Methods

- Analyze means of log transformed data
  - For clarity, usually better to back transform estimates to the original scale
    - E.g., geometric mean of PSA, rather than mean of log PSA
    - E.g., ratio of geometric means, rather than difference of means of log transformed data
  - Exceptions do exist when the scientific community is used to log transformed data
    - pH, Richter scale, decibels, titers

91

## Interpretation

- Note that if the log transformed data is symmetrically distributed, then the geometric mean is the same as the median
  - Hence, IF you are willing to presume symmetry after log transformation, then you can interpret your parameter as the median
  - In this situation, the geometric mean will usually be a more efficient estimator of the median than would be the sample median

92

## Stata Commands

- "means"
  - Provides estimates, CI for geometric means
    - Also arithmetic and harmonic means
- Transforming positive data
  - "gen newvar= log(var)"
    - If zeroes indicate "below limit of detection"
      - Replace 0 by one-half lowest nonzero value?
  - Use "ci" and / or "ttest"
  - Backtransform estimates and CI with
    - "disp exp(#)"

93

## Example: Geometric Mean of FEV

- Scientific / statistical rationale for considering geometric mean of FEV
  - A multiplicative relationship
    - FEV is a volume (cubic dimension)
    - Best predictor is height (linear dimension)
  - Greater statistical precision obtained on log scale

94

## Stata Commands: Estimate, CI

```
. bysort smoker: means fev
-> smoker = 0
```

Var	Type	Obs	Mean	[95% Conf. Interval]
fev	Arithmetic	589	2.566143	2.497314 2.634971
	Geometric	589	2.431225	2.366838 2.497364
	Harmonic	589	2.299331	2.236031 2.36632

```
-> smoker = 1
```

Var	Type	Obs	Mean	[95% Conf. Interval]
fev	Arithmetic	65	3.276862	3.091024 3.462699
	Geometric	65	3.191452	3.011514 3.382142
	Harmonic	65	3.10473	2.927637 3.304627

95

## Stata Commands: Test

```
. gen logfev = log(fev)
. disp log(3)
1.0986123
. bysort smoker: ttest logfev=1.0986123
-> smoker = 0
```

Variab	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
logfev	589	.888	.0136661	.3316671	.861555 .9152357

```
mean = mean(logfev) t = -15.3824
Ho: mean = 1.09861 degrees of freedom = 588
Ha: mean < 1.09861 Ha: mean != 1.09861 Ha: mean > 1.09861
Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000
-> smoker = 1
```

Variab	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
logfev	65	1.160	.0290495	.2342048	1.102443 1.218509

```
mean = mean(logfev) t = 2.1296
Ho: mean = 1.09861 degrees of freedom = 64
Ha: mean < 1.09861 Ha: mean != 1.09861 Ha: mean > 1.09861
Pr(T < t) = 0.9815 Pr(|T| > |t|) = 0.0371 Pr(T > t) = 0.0185
```

96