Introduction to Biostatistics BC 203

Final Exam
Thursday, December 15, 2005

(Closed Book)

There are 25 questions, each worth 4 points.

Write out your solutions and circle your final answers. Show all your work on these pages. If you need more space write on the back of the page.

You have three hours to complete this exam.

Name: Solutions

1. Suppose we collect data on blood glucose levels from 100 participants. Assume that these really are independent observations from some (well behaved) probability distribution with unknown mean μ and variance σ^2, but that the distribution is not normal.

(a) What characteristic of the distribution is most important for determining whether a large sample 95% confidence interval on the mean will be approximately correct?

SYMMETRY

(b) What is meant by "approximately correct" in part (a)?

the coverage probability of the CI will be approximately 95%.

(c) True or False: As the sample size gets larger, the sampling distribution becomes more symmetric.

the Sampling dist'n is the distribution of $\bar{x}$ not X.

(d) True or False: As the sample size gets larger, the width of the sampling distribution decreases.

b/c $\text{VAR}(\bar{x}) = \frac{\sigma^2}{n}$

1. (continued, each question below refers to the initial sample of 100)

(e) If we repeat this study with a smaller sample, say n = 25, (from the same probability distribution) how will the variance of the new sample mean, $\overline{X}_{25}$, compare to the variance of the old one?

- [x] it will definitely be greater
- [ ] it will definitely be smaller
- [ ] it will probably be greater, although it might be smaller
- [ ] it will probably be smaller, although it might be greater
- [ ] there is no reason to expect it to be greater, and there is no reason to expect it to be smaller

$\frac{\sigma^2}{25} = VaR(\bar{X}_{25}) > VaR(\bar{X}_{100}) = \frac{\sigma^2}{100}$

(f) If we repeat this study with a smaller sample, say n = 25, (from the same probability distribution) how will the new sample variance compare to the old one?

- [ ] it will definitely be greater
- [ ] it will definitely be smaller
- [ ] it will probably be greater, although it might be smaller
- [ ] it will probably be smaller, although it might be greater
- [x] there is no reason to expect it to be greater, and there is no reason to expect it to be smaller

$E[S^2] = \sigma^2$ → does not depend on "n" (the sample size)

(g) If we repeat this study with a smaller sample, say n = 25, (from the same probability distribution), how does the power of this study compare to the old one?

- [ ] it will definitely be greater
- [x] it will definitely be smaller
- [ ] it will probably be greater, although it might be smaller
- [ ] it will probably be smaller, although it might be greater
- [ ] there is no reason to expect it to be greater, and there is no reason to expect it to be smaller

Smaller sample size less power.

2. A study is being planned to investigate the proportion of teenagers who drive while impaired (drunk!). It is thought that this proportion is 32%.

(a) If I wanted to estimate this proportion to within a 5% margin of error, how many teenagers would I need to interview if I planned to use a 90% confidence interval?

$MoE = .05 \Rightarrow L = .1$

$$N = \frac{4(\theta)(1-\theta)(Z_{\alpha/2})^2}{L^2}$$

$\theta = .32$
$\alpha = .1$ $Z_{\alpha/2} = 1.645$
$L = .1$

$= 235.53$

$= 236$

(b) Suppose I could enroll only 50 teenagers. What is the approximate power of my 2-sided test, with size 10%, to reject the null hypothesis that the true proportion is 50%?

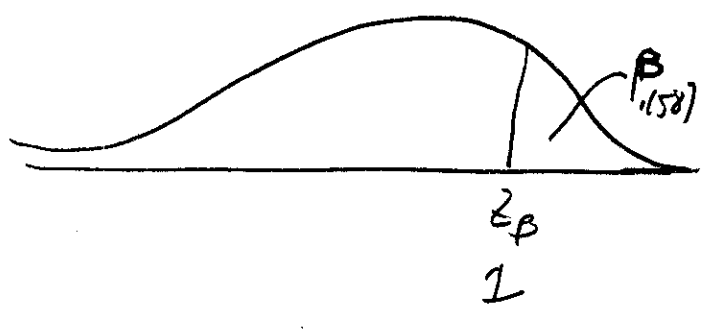
$\theta_0 = .5$
$\theta_a = .32$
"Size" $\alpha = .1$
$n = 50$

$$N = \frac{\left(Z_{\alpha/2}\sqrt{\theta_0(1-\theta_a)} + Z_\beta\sqrt{\theta_a(1-\theta_a)}\right)^2}{(\theta_0 - \theta_a)^2}$$

$$\Rightarrow \sqrt{50(.5-.32)^2} - 1.645\sqrt{.32(.68)} + Z_\beta\sqrt{\tfrac{1}{2}(\tfrac{1}{2})}$$

$$1.27 = .7674 + Z_\beta \tfrac{1}{2}$$

$$Z_\beta = 1$$



$\beta = .1587$

$1 - \beta = .8413$

3. A (hypothetical) study of the association between stroke and high blood pressure is conducted in 165 participants. The data and *Stata* output are provided below. When enrolled, each participant was classified as having high blood pressure or not and then all participants were followed for five years to see if they had a stroke.

```
. tabi 1 10\ 64 90, exact chi2

           | High blood pressure
    Stroke |         N          Y |     Total
-----------+---------------------+----------
         N |         1         10 |        11
         Y |        64         90 |       154
-----------+---------------------+----------
     Total |        65        100 |       165

          Pearson chi2(1) =   4.5330   Pr = 0.033
           Fisher's exact =                 0.051
  1-sided Fisher's exact =                 0.029
```

(2-sided) (2-sided) (1-sided)

(a) An investigator wants to test the hypothesis that the row and columns are independent at the 5% level, i.e. that there is no association between stroke and high blood pressure. Would he reject the null hypothesis and what p-value should he use?

Use Fisher's Exact Test (it is always correct and in this case the cell counts are low, indicating the usual approximations may not work well).

So, Fail to Reject $p = 0.051$.

(b) Which cell contributes the most to the Chi-square statistic?

- ☒ Top left (Stroke=N, HBP=N, count=1)
- ☐ Top right (Stroke=N, HBP=Y, count=10)
- ☐ Bottom left (Stroke=Y, HBP=N, count=64)
- ☐ Bottom right (Stroke=Y, HBP=Y, count=90)

$\chi^2 = \sum_{i=1}^{4} \frac{(O-E)^2}{E}$ For top cell $O = 1$ $E = \frac{11(65)}{165}$

$\frac{(O-E)^2}{E} = 2.566$ which is the largest.

3. (continued)

(c) Provide a 95% confidence interval for the odds ratio of having a stroke in participants with high blood pressure versus those without. Interpret this interval.

$$\log(\hat{OR}) \pm z_{\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$\log\left(\frac{90(1)}{10(64)}\right) \pm 1.96\sqrt{1+\frac{1}{90}+\frac{1}{10}+\frac{1}{64}}$$

$$-1.96 \pm 2.08$$

CI for log OR $\rightarrow [-4.04, 0.12]$

CI FOR OR

$(e^{-4.04}, e^{0.12})$

$(0.176, 1.1275)$

(d) Provide a 95% confidence interval for difference in stroke risk between the two blood pressure groups. Interpret this interval.

$$\hat{\theta}_1 - \hat{\theta}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}$$

$\hat{\theta}_1 = \frac{90}{100}$

$\hat{\theta}_2 = \frac{64}{65}$

$\Rightarrow .08 \pm 1.96(.0377)$

$(.012, .1479)$

$(.0186^{or}, .15)$] depends on how you Round.

(e) Based on the two confidence intervals above, what would you conclude? (provide scientific and statistical conclusions)

The CI's disagree. How you measure the effect determines if there is a "statistically significant" difference. To me the Risk difference is the most meaningful. Odds ratios are hard to interpret and in this case it may not be a good estimate of the RR).

4. A sample of mercury levels in 18 ACME tuna cans yielded a sample average of 32.4 micrograms with a sample standard deviation of 4. Their competitor, JONES, yielded an average mercury level of 35 micrograms and sample standard deviation of 3 from 20 cans.

(a) Provide a 95% confidence interval for the average mercury level in ACME tuna.

$\bar{x} \pm t^{n-1}_{\alpha/2} \frac{s}{\sqrt{n}}$ $\quad t^{n-1}_{\alpha/2} = t^{17}_{.025} = 2.110$

32.4 ± 1.9808 $\quad (30.4, 34.4)$

(b) Provide a 95% confidence interval for the average mercury level in JONES tuna.

$\bar{x} \pm t^{n-1}_{\alpha/2} \frac{s}{\sqrt{n}}$ $\quad t^{n-1}_{\alpha/2} = t^{19}_{.025} = 2.093$

$(33.6, 36.4)$

(c) Dr. X notices that these two confidence intervals overlap, and he concludes from this that the two sample means are not significantly different at the 5% level. Is he correct? Answer yes or no and justify your answer (no credit without justification).

No! a CI on the difference is

$35 - 32.4 \pm 2.101 \sqrt{\frac{4^2}{18} + \frac{3^2}{20}}$

2.6 ± 2.43

$(.17, 5.03)$

So the means are statistically different at the 5% level ($p < .05$).

5. To test the theory that the intersection of Thayer and Waterman is safe, I decide to stand on the corner – for the whole day - and record the number of accidents at that intersection. An intersection is safe if the rate of accidents is less than or equal to 1 per day (i.e., $\lambda \leq 1$). If I see more than 2 accidents that day, I will reject the Null hypothesis that the corner is safe.

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

(a) What is the statistical null hypothesis?

$H_0: \lambda \leq 1$

↳ this was on the formula sheet

(b) Is this a one-sided or a two-sided test?

one sided

(c) What is the exact Type I error of this test?

$P(K>2 \mid H_0) = P(K \geq 3 \mid H_0) = 1 - P(K=2 \mid \lambda=1) - P(K=1 \mid \lambda=1) - P(K=0 \mid \lambda=1)$

↳ $\lambda = 1$

$= 1 - \frac{e^{-1}1^2}{2!} - e^{-1} - e^{-1}$

$= 1 - 0.1839 - 2(0.3679) = \boxed{0.0803}$

(d) If the rate of accidents is 3 per day, what is the exact power of this test to reject the null?

$P(K>2 \mid H_a) = P(K \geq 3 \mid \lambda = 3) = 1 - P(K=2 \mid \lambda=3) - P(K=1 \mid \lambda=3) - P(K=0 \mid \lambda=3)$

↳ $\lambda = 3$

$= 1 - \frac{e^{-3}3^2}{2!} - \frac{e^{-3}3^1}{1!} - \frac{e^{-3}3^0}{0!}$

$= 1 - 0.224 - 0.1494 - 0.0498$

$= \boxed{0.5768}$

(e) Suppose I observed 2 accidents, what is the p-value?

$P(K \geq 2 \mid H_0) = P(K \geq 2 \mid \lambda = 1) = 1 - P(K=1 \mid \lambda=1) - P(K=0 \mid \lambda=1)$

$= 1 - e^{-1} - e^{-1}$

$= 1 - 2(0.3679)$

$= \boxed{0.2642}$

6. A study is designed to evaluate a new drug that reduces the transmission rate of HIV from the mother to her baby. Data from the study indicate that drug appears to reduce the transmission rate by 30%. That study provided the following three confidence intervals for the reduction in transmission rate:

90% confidence interval for the risk reduction: 5.2% to 55.8%
95% confidence interval for the risk reduction: 0.1% to 60.1%
99% confidence interval for the risk reduction: -4.3% to 65.5%

(a) What is the p-value for testing the null hypothesis that the drug did NOT reduce the transmission rate?

- ☐ .1< p
- ☐ 0.05<p<.1
- ☒ 0.01<p<0.05
- ☐ p<0.01

Zero is in the 99% CI (so p>.01) But not in the 95% CI (so p<.05)

(b) A test of the null hypothesis that the true reduction in transmission rate is 10% would reject at the 10% level?

- ☐ True
- ☒ False

because 10% is in the 90% CI (1-α)% CI, α=.10

(c) Fisher's exact test is named after

- ☒ Sir Ronald A. Fisher, the famous statistician
- ☐ Bobby Fisher
- ☐ Fisher – Price
- ☐ Ronald "Fisher" McDonald
- ☐ Ruth "Fisher" Simmons

HA-HA!