Course EPIB-634: Survival Analysis & Related Topics [Winter 2007]

Assignment 1

Answers and some 'bigger picture' comments
(Please let JH know of any errors or unclear items in these answers)

1 **Which is the *single biggest* flaw in the analysis of the scouting injuries [page 3]. List two others that might on their own might be major -- but not nearly as large as the distortion produced by *the big one* !**

Biggest: For the sake of illustration, ignore the small details and think of the 11 per 1000 as a 'person-time' rate i.e. 11 hospitalizations per 1000 child-years (c-y). A child-year is 52 x 7 x 24 = approx 9,000 child-hours , or 9 million hours for 1000 children , so rate = 11 hospitalizations/9million-c-hours. Now for the scouts, at the most, I would estimate that in a year they do 450 hours of scouting (2 hours a week plus (if they even counted it) 2 weeks at camp in summer, i.e. 100 + 2 x 7 x 24 = 450 or so hours, so their 1 accident per 1000 children is really 1 accident in 450,000 scout-hours... i.e., approx. 2 times as high per hour as the general population!
If you assume a child is only at risk when awake, and awake 16 hours a day (2/3 of 24), then the rates are 11/6million c-hours and 1/[1000 x (100 + 2 x 7 x 16)] = 1/0.3million hours , so rate ratio = 1.7

BIG PICTURE: need to use appropriate children-time denominators!!

Other issues.. (1) 11/1000 c-y is (I presume) for all reasons, not just accidents. So the rate of hospitalizations for accidents would be lower that the 11/1000c-y. (2) Are scouts healthier than the average?

BIG PICTURE: denominators are fundamental in epi: they are what create the numerators.

2 **Refer to the first row of Table 1 in the Ayas et al. article "Extended Work Duration and the Risk of Self-reported Percutaneous Injuries in Interns" in JAMA on Sept 6 of 2006.**

(i) Manually calculate the rate per Intern-Month and the 95% CI

Because the Poisson distributions that could yield a count of 498 are virtually Gaussian in shape: { 498 ± 1.96 sqrt[498] } / 17003. More refined approximations are overkill in this instance.

(ii) Re-express these using Intern-Year as the unit of experience. *12 × CI from (i). You don't need to start all over again, since the relative uncertainty doesn't change with a simple change of scale.*

(iii) Repeat (i) and (ii) using software such as the Epitools package for R [see http://www.epitools.net] , or Stata [help epitab, and iri 498 1000000 17003 1000000], or the SAS GENMOD regression program [in c634 Resources].

Most of you used one of the epi-calculators in R or Stata.

The point of showing SAS here is to show that you can have a regression with just 1 data point and one parameter! (will do same in R below)

```
options nocenter; run;
DATA a;
input cases  i_months;
lines;
     498      17003
;
```

Generalized Linear model, in SAS...

```
PROC genmod data=a ;
```

```
  model cases = i_months / dist=poisson link=identity noint* waldci;
RUN;

Data Set                        WORK.A
Distribution                    POISSON
Link Function                   IDENTITY
Dependent Variable              CASES
Observations Used               1

Parameter     Effect

PRM1          INTERCEPT
PRM2          I_MONTHS

Parameter Estimates

Parameter  DF  Estimate  Std Err   ChiSquare  Pr>Chi

INTERCEPT   0    0.0000*  0.0000 (*line forced through 0 (0 cases if 0 P-T !)
I_MONTHS    1    0.0293   0.0013    498.0000  0.0001
SCALE       0    1.0000   0.0000           .       .

NOTE:  The scale parameter was held fixed.

Normal Confidence Intervals For Parameters

Two-Sided Confidence Coefficient: 0.9500

PRM2          Lower        0.0267
PRM2          Upper        0.0319
```

(use i_years = i_months/12 as "X" variable to get point and interval estimate of rate per intern-year)

Generalized Linear model, in R...

```
cases=498;i.months=17003; summary(glm(cases~-1+i.months))
```

***Comments**: Default Identity Link, Gaussian variation;*
-1 is a way to specify an intercept of zero

```
Deviance Residuals:
[1]  0

Coefficients:
         Estimate Std. Error z value Pr(>|z|)
i.months  0.02929        Inf       0        1

Residual deviance:      0  on 0  degrees of freedom  !!!
```

```
summary(glm(cases~ 1+i.months,family=poisson(link="identity")))
```

```
Deviance Residuals:
[1]  0

Coefficients:
         Estimate Std. Error z value Pr(>|z|)
i.months 0.029289   0.001312   22.32   <2e-16

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance:        Inf  on 1  degrees of freedom
Residual deviance: 3.1086e-15  on 0  degrees of freedom
AIC: 10.049
```

(iv) Repeat steps (i) to (iii) for the data from the Psychiatry residency, using a method appropriate to the situation [Rothman, 2002, page 127 says 'such situations are the exception rather than the rule.']

You might find "Exact confidence limits on a Poisson parameter: Excel worksheet" (in Resources) helpful for visualizing Poisson distribution, and for exact CI's and p value calculations when the count is too small to rely on the Gaussian approximation (for p-values, you can also interpolate using the table on page 17; or use the exact Poisson function in Excel

NAIVE... equivalent to using { 1 ± 1.96 sqrt[1] } / 658

```
data a;
input cases  i_months;
LINES;
         1       658
;
PROC genmod data=a ;
  model cases = i_months / dist=poisson link=identity  noint waldci;
RUN;
```

```
Parameter    DF    Estimate     Std Err   ChiSquare  Pr>Chi

INTERCEPT     0      0.0000      0.0000           .       .
I_MONTHS      1      0.0015      0.0015      1.0000  0.3173
SCALE         0      1.0000      0.0000           .       .

Normal Confidence Intervals: Two-Sided Confidence Coefficient: 0.95

Parameter      Confidence Limits

PRM2          Lower     -0.001459 !!!
PRM2          Upper      0.004498
```

Naive (In R...)

```
cases=1;i.months=658; summary(glm(cases~ 1+i.months,family=poisson(link="identity")))

         Estimate Std. Error z value Pr(>|z|)
i.months  0.00152    0.00152       1    0.317
```

Need exact methods that reflect the highly skewed shape of the Poisson distribution at the lower (and even upper) limit(s) compatible with the observed count of 1.

Use CI(based on count of 1) / 658.

Several ways to obtain CI;

if don't want to carry around the CI table, can make your own in R

```
poisson.ci=function(o,conf)(return( c(0.5*qchisq((1-conf)/2, 2*o),0.5*qchisq(conf+(1-conf)/2, 2*o+2)) ))
```

3 (i) Calculate a 95% CI for the SIR and test (at alpha = 0.05 2-sided) H_0: SIR=1 for the Alberta Sour Gas Study [p. 4]. Restrict attention to the 33 vs. 36.3 [Index Area 1970 Cohort Females vs. (1) Southern Alberta excl. Calgary, Lethbridge, & Medicine Hat (RP1)]. Describe your procedures/ steps.

Can think of the 'expected number' E as that for all of Alberta, but scaled down to the size of the index area. Because the *number of cases for all of Alberta is quite large, it remains stable when we scale it down to E; Thus, we say that (at least relative to the observed number O in the index area) the scaled down number E has no statistical variation i.e it is treated as a 'constant' in the SIR -- only the numerator O is a random variable.*

Estimated SIR = 33/36.3 so CI = CI[for numerator] / 36.3

Some of you used slightly more refined approximations, and some of you even used an exact (i.e. Poisson-based) CI. Here I will for sake of illustrating the big picture, use the CI based on Gaussian-approximation to Poisson (since 33 fairly large)

$$\{33 \pm 1.96 \times \text{sqrt}[33]\ \} / 36.3 = \{33 \pm 11.26\ \} / 36.3 = 0.60 \text{ to } 1.22$$

Or use exact CI (spreadsheet or chi-sq link or Stata) of 22.7 to 46.3, divide by 36.3 to get 0.63 to 1.28

Test: CI does include 1, so $p < 0.05$; more specifically $z = (33 - 36.3)/\text{sqrt}[36.3] = –0.55$ The question did not specify whether this was a one-sided test (of most interest to those in the index area), but either way, the chance of a count this or more extreme is quite high, if the null were true.

BIG PICTURE:

(a) uncertainty is in numerator; cannot scale the numerator

(b) One cannot state the null H as observed count = expected count (or observed SIR = 1), since it is impossible to observe 36.3 cancers in a single area. Instead, thinking of this as a generic study of the effects of sour gas on ANY community. Then, can say, the average number of cancers in towns of this size, even if exposed to sour gas, is no higher than the average no. in unexposed Alberta areas of this size.

(c) One challenge for this part of the Alberta study was that any imperfect tracing of those who moved out of the area would create an underestimate (i.e. maybe there were others in addition to the 33 observed). On Feb 12 our department will honour the work of the principal investigator (Spitzer WO) of this project. The publications from this study include

Tousignant P, Groome PA, Spitzer WO, Schechter MT, Montano L, Hutcheon ME. Outmigrant ascertainment for bias assessment in environmental epidemiology. Int J Epidemiol. 1994 Oct;23(5):1091-8.

Schechter MT, Spitzer WO, Hutcheon ME, Dales RE, Eastridge LM, Hobbs C, Suissa S, Tousignant P, Steinmetz N. A study of mortality near sour gas refineries in southwest Alberta: an epidemic unrevealed. Can J Public Health. 1990 Mar-Apr;81(2):107-13.

Spitzer WO, Dales RE, Schechter MT, Suissa S, Tousignant P, Steinmetz N, Hutcheon ME.. Chronic exposure to sour gas emissions: meeting a community concern with epidemiologic evidence. CMAJ. 1989 Oct 1;141(7):685-91.

Dales RE, Spitzer WO, Schechter MT, Suissa S. The influence of psychological status on respiratory symptom reporting. Am Rev Respir Dis. 1989 Jun;139(6):1459-63.

Dales RE, Spitzer WO, Suissa S, Schechter MT, Tousignant P, Steinmetz N. Respiratory health of a population living downwind from natural gas refineries. Am Rev Respir Dis. 1989 Mar;139(3):595-600.

Schechter MT, Spitzer WO, Hutcheon ME, Dales RE, Eastridge LM, Steinmetz N, Tousignant P, Hobbs C. Cancer downwind from sour gas refineries: the perception and the reality of an epidemic. Environ Health Perspect. 1989 Feb;79:283-90.

Dales RE, Spitzer WO, Tousignant P, Schechter M, Suissa S. Clinical interpretation of airway response to a bronchodilator. Epidemiologic considerations. Am Rev Respir Dis. 1988 Aug;138(2):317-20.

Spitzer WO, Dales R, Schechter MT, Tousignant P, Hutcheon M. Subjective fears and objective data: an epidemiologic study of environmental health concerns. Trans Assoc Am Physicians. 1987;100:40-4.

(ii) Carry out the same tasks, but imagine the concerned area or cohort was much smaller, and that 3 cases were observed where 0.45 were expected. Again, describe your procedures/ steps.

CI (exact now, for sure] based on a count of 3, is 0.62 to 8.77,
so SIR = 0.62/.45 to 8.77/.45 = 1.38 to 19.49

For test, for 1 side need to calculate Prob[≥3 | E = 0.45) ; table in JH notes has E=0.4 and E=0.5

Prob[≥3 | E = 0.40) = 0.0072 + 0.0007 + 0.0001 = 0.008
Prob[≥3 | E = 0.50) = 0.0126 + 0.0016 + 0.0002 = 0.014

so Prob[≥3 | E = 0.45) = somewhere between 0008 and 0.014, around 0.011 say.

For overall p-value (2-sided) can EITHER double the 0.011 = 0.022 (Armitage)
OR
find values on other side of 0.45 that are less likely that Prob[3], and add this to the 0.011
But there aren't any, since prob[0] is very high (about 0.64) with E = 0.45.

Poisson tail area in R...

```
ppois(3, lambda=0.45, lower.tail = FALSE) => [1] 0.001195352
```

BIG PICTURE: when Poisson expectation is low, use exact methods (Gaussian not accurate)
some ambiguity re what values are in other tail, when have count outcome

4 Refer to rows 2 and 3 of Table 3 to the Ayas et al article.

(i) Manually calculate ORs and 95% CIs, and repeat by computer software.

These are not OR's, they are Rate Ratios. ie the 26667 and 60763 are 'real' P-T denominators.

Use large-sample CI methods for log[rate ratio] & convert to (asymmetric) limits for Rate Ratio

(fortunately, since denominators much large in magnitude than numerators, using CI for OR or Risk Ratio not all that different; BUT, if had expressed the denominators in say "Intern-years", or "Inter-centuries" the inappropriateness of using a Binomial or Odds-based model would be more serious.. see similar issue with John Snow data below)

BIG PICTURE: just because the estimate (17/26667) / (21/60763) "looks like an or" and "walks like an or" doesn't necessarily mean it is an or for the purposes of statistical modeling. Do not immediately (as Rothman does to save calculator steps) turn it into ad/bc. Go the extra few calculator steps and leave it with its true form i.e., a ratio of two genuine rates.

Some of you rightly questioned whether large-sample (Gaussian) methods are accurate when dealing with counts as low as 4 and 4... see q 8 below)

(ii) Explain why your answers do not match those reported (hint: see the paragraph beginning "To assess the relationships..." in the last column of page 1057 of the article.

The authors used a finely stratified (matched) analysis.

(iii) exactly what (and how many) numbers would you need to carry out *their* analysis for row 3 (injuries in ICU). Answer in the form of a 1-paragraph request to the authors asking for these specific numbers (but do not e-mail the authors! JH has in fact obtained these numbers from Dr Ayas, and they will form the basis for some of next week's homework).

the numerators and denominators for just the 8 md's who had an injury (the others do not contribute to the Mantel-Haenszel summary rate ratio estimate).

(iv) Is OR the correct term for the ratio being estimated here?

No (just because it looks like a duck and walks like a duck doesn't ...)

5 **Refer to the data from John Snow's study, given on bottom of column 3 of page 1 of attached handout for Sept. 05 lecture for Med2 [taken from med2 website, reachable from link at top of 634 website: username med2, password: same as for the cxxx epidemiology courses].**

(i) Calculate a 95% CI to accompany the rate ratio of 13.3.

Again, the denominators are person-time denominators. Some of you subtracted cases from houses and called them 'non-cases'. In reality, the person-time denominators are numbers of houses x average no. persons per house x 4 weeks, but (as long as the average no. persons per house is the same on both sides), these two factors cancel out in the rate ratio. CI for (person-time) rate ratio is appropriate

Do the same for the ratio estimates based on the denominator series of 100 and 1000 (first column, page 2... [in practice, you would not observe the quasi-denominators shown there, but rather these expected numbers ± some sampling variation].

here we have quasi(partial)-denominators, and so the variance of the log rate ratio should reflect the extra uncertainty.. e.g., 1/286 + 1/28 + 1/65 + 1/35

(ii) Why are the CI's based on the 100 or 1000 *wider* than the one based on the actual "return which was made to Parliament"?

because we estimated the denominators by sampling.

It turns out that the 'weakest link' in the variance formula is the 1/14, and so there is no point in becoming more precise about the denominators (with a larger denominator series) since the 1/14 cannot be reduced. This is the reason for denominator ('control') series say 2-4 times the size of the case series. Here the case series is of size 300, so the variance with the denominator series of 1000 is already close to the smallest it can be (i.e. to the variance with the entire denominators). Doubling the denominator series is a lot of work for a 'diminishing return' in terms of variance of the log of the rate ratio. For more on statistical 'efficiency', see at the end of these answers some excerpts from JH's 607 notes on the topic.

6 **Refer to pages 2 & 3 of the Med2 handout of Nov. 11 [attached]. In dealing with CI's for ratios, it used the fact that for *log-based CI's* (instead of the usual ± a margin of error for 'regular' statistics) for *ratios*, one can calculate a "multiplied-by/divided-by" factor in order to arrive at the upper/lower limits. (i) Hand-calculate the CI's for the ratio of 13.1 on page 2, and the 1.44 ratio on page 3*, by your usual manual way, and compare them with the answers from the "multiplied-by/divided-by" method shown**
(ii) Which method do you prefer? (if you have software that does it for you, this is merely a conceptual issue!)
{ * the full article "A population-based study of measles, mumps, and rubella vaccination and autism" can be found under Nov. 11 lecture in med2, reachable from link at top of 634 }

Since some of you were not able to see the 'logic' behind the 'multiply/divide' factor, you were not comfortable with it, even if it saved a few steps on a calculator.

here is the logic...

Traditional: exp[log RR ± z SE]

So..

upper = exp[log RR + z SE] = exp[log RR] x exp[+ z SE] = RR x exp[z SE]

lower = exp[log RR - z SE] = exp[log RR] x exp[- z SE] = RR x exp[- z SE] = RR / exp[z SE]

8 **The large-sample methods for obtaining a CI for a rate ratio are accurate when there are**

enough events in each of the compared categories. But in Q7 above, and in the "Women are Safer Pilots" example on page 3, the small number of events in one of the categories renders large-sample methods inaccurate or even impossible. In such situations, the *conditional* approach, in which one bases the inference on the distribution of the number of events in one category, *conditional* on the *sum* of the numbers of events in the two categories, is a way around this problem (we use a similar *conditioning* strategy when dealing with Fisher's exact test).

Compare the rate of accidents in women relative to men pilots (i.e. the rate *ratio*) (i) Assume that on average, the women pilots fly just as many hours as the men pilots, and that all other relevant factors are equal [although they probably are not!]. Based on the information given, use software to calculate an exact CI for the rate ratio

We do not know the total number of hours (H) involved, but since we are interested in relative rates, it is not critical that we do (the H would in any case cancel out even if we use 0.06H and 0.94H as the two denominators)

point estimate of rate ratio = (2/(0.06H)) / (136/(0.94H)) = (2/(0.06)) / ((136/(0.94))

Again, 'don't be a Rothman' about this; instead leave it as a rate ratio even if it can be made to look like the same "cross-product ratio" form we are familiar with for odds ratios.

For CI, use Binomial-based CI based on conditioning of the total number (138) of cases and treating the proportions 2/138 and 136/138 as Binomial estimates of a parameter π

i.e. lower = (π_{lower}*/0.06)) / (1-*π_{lower}*/0.94))*

Can obtain π_{lower} *by exact method for a Binomial CI (see notes and resources for Chapter 8.1 in course 607)*

Rothman uses (and explains) the conditional approach on page 166, Ch 11, of his 1986 book (see Resources for Rates) but seems to have been unaware at that point that -- just like the exact Poisson tail areas can be obtained from the link with the chi-square distribution -- there is also an exact link between the binomial tail-areas and the F distribution. The Excel sheet provided in the 607 Resources lets you do it by trial and error (as Rothman does), or using the link.

By the way, Rothman says on page 154 of his 1986 book that "the statistical model used for hypothesis testing of person-time data is the binomial distribution". In fact he making a (slight) approximation and simplification here, but minimises it by always using small time units, so he has a binomial with a numerator much smaller than the denominator. he does however then have to evaluate very large factorials. In fact, the correct statistical model used for hypothesis testing of person-time data is the Poisson distribution. It also avoids large factorials. The Poisson distribution can be derived as the limiting case of the Binomial. He would be in trouble if we expressed all of his denominators in person-centuries or even larger units and tried to force it to be Binomial (note that our 0.06 and 0.94 are perfectly good Poisson denominators) .

(ii) Repeat , but now assume that on average the women pilots fly half as many hours as the men.

Now the denominators are in the ratio 0.03 to 0.94

point estimate of rate ratio = (2/(0.03)) / (136/(0.94))

For CI, use Binomial-based CI based on conditioning of the total number (138) of cases and treating the proportions 2/138 and 136/138 as Binomial estimates of a parameter π

i.e. lower = (π_{lower}*/0.03)) / (1-*π_{lower}*/0.94))*

Use same π_{lower} as above.. its just the denominators that have moved, but we still have the same amount of information re the numerators.

(iii) In your own words, and using pages p 29 of Poisson notes, try to describe the basis for the exact method. [JH will use your answers to judge how clear or muddled his description is!]

See Rothman 1986. The point is that if we have very large numerators, we could use the exact parametric relation b/w the expected proportion of exposed cases on the one side and the rate ratio and the 2 denominators on the other side. We solve this equation to get a point estimate of the rate ratio. But in practice that point estimate uses a (possibly quite imprecise) binomial-based estimate, and so we should also get upper and lower limits for the RR by substituting upper and lower binomial limits. In the example above, the 2/138 and the 136/138 can be seen as realizations of a binomial random variable with parameter π. Our 'best' estimate of π is 2/138, but clearly it is subject to sampling variability (time/place). However, we take the 0.06 and 0.94 as measured with no sampling variability.

One way to 'see' what proportion of the cases are exposed cases is to via at the diagram, where the areas of the white rectangles are proportional to the exposed and non-exposed P-T denominators, and the event rate in the exposed is 3 times that in the non-exposed. Then the numbers of exposed cases will be proportional to 0.4 x 3 = and the numbers of non-exposed cases to 0.6 x 1.

So of every (1.2 + 0.6 = 1.8) cases, on average a proportion (1.2/1.8) = 1/3rd would be exposed cases, and the remaining 2/3rds would be non-exposed cases.

This is a specific example of the more general rule

Of all cases, on average a proportion $P = (PT_1 \times RR / (PT_1 \times RR + PT_0 \times 1)$ *would be exposed cases, and the remaining proportion would be non-exposed cases.*

Reversing this, we get $RR = (P/PT_1\ ) / ((1-P)/PT_0)$

This algebraic solution is exactly as one would expect as the definition of a rate ratio:

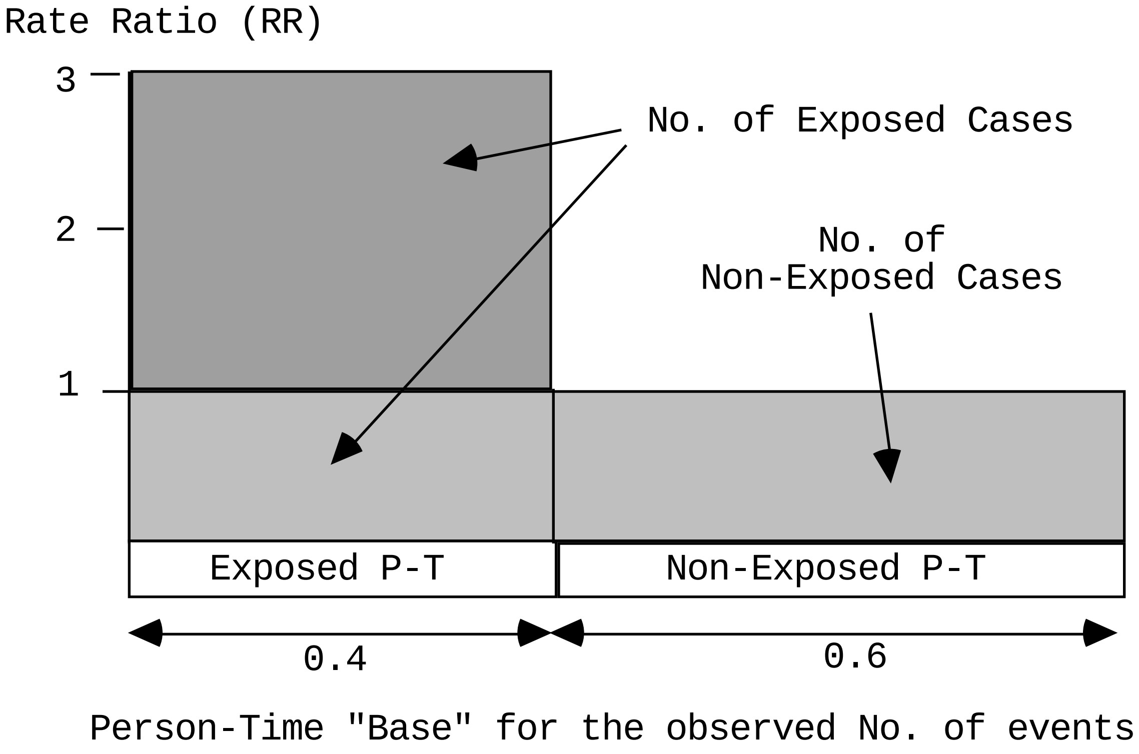
RR = (exposed cases/exposed PT) / (non-exposed cases / non exposed PT)

CONCEPTUAL CORRECTNESS

In "better families", i.e. in modern epidemiology, we always compare event rates in exposed versus event rates in the non-exposed.. i.e., we are students of event rates and event rate ratios

We never compare 'the rate of exposure in the cases' versus 'rate of exposure in controls' , even if the resultant arithmetic looks like a rate (or odds) ratio and computes like a rate (or odds) ratio!

In our example P is estimated by the Binomial 2/138 and 1-P by 136/138


Rate Ratio (RR)
3
2
1
No. of Exposed Cases
No. of
Non-Exposed Cases
Exposed P-T
Non-Exposed P-T
0.4
0.6
Person-Time "Base" for the observed No. of events

Effect of Unequal Sample Sizes ($n_1 \neq n_2$) on precision of estimated differences

If we write the SE of an estimated difference in mean responses as $\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where σ is the (average) per unit variability of the response, then we can establish the following principles:

1 **If costs and other factors (including unit variability) are equal, and if both types of units are equally scarce or equally plentiful**, then for a given total sample size of n = $n_1 + n_2$, an equal division of n i.e. $n_1 = n_2$ is preferable since it yields a smaller SE(estimated difference in means) than any non-symmetric division. However, the SE is relatively unaffected until the ratio exceeds 70:30. This is seen in the following table which gives the value of $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ = SE(estimated difference in means) for various combinations of n_1 and n_2 adding to 100 (the 100 itself is arbitrary) and assuming σ = 1 (also arbitrary).

n_1	n_2	SE(estimated difference in means)	%Increase in SE over $SE_{(50:50)}$ *
50	50	0.200	-----
60	40	0.204	2.1%
65	35	0.210	4.8%
70	30	0.218	9.1%
75	25	0.231	15.5%
80	20	0.250	25.0%
85	15	0.280	40.0%

* if sample sizes are π:(1–π), the % increase is 50 / $\sqrt{\pi(1\text{-}\pi)}$.

2 **If one type of unit is much scarcer, and thus the limiting factor**, then it makes sense to choose all (say n_1) of the available scarcer units, and some $n_2 \geq n_1$ of the other type. The greater is n_2 , the smaller the SE of the estimated difference. However, there is a 'law of diminishing returns' once n_2 is more than a few multiples of n_1. This is seen in the following table which gives the value of $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for n_1 fixed (arbitrarily) at 100 and n_2 ranging from 1 x n_1 to 100 x n_1; again, we assume σ=1.

n_1	n_2	Ratio (K)	$SE(\hat{\mu}_1 - \hat{\mu}_2)$	$SE_{K:1}$ as % of of $SE_{(1:1)}$	$SE_{K:1}$ as % of $SE_{(\infty:1)}$ *
50	50	1.0	0.2000	-	1.414
50	75	1.5	0.1825	91.3%	1.290
50	100	2.0	0.1732	86.6%	1.225
50	150	3.0	0.1633	81.6%	1.155
50	200	4.0	0.1581	79.1%	1.118
50	250	5.0	0.1549	77.5%	1.095
50	300	6.0	0.1527	76.4%	1.080
50	400	8.0	0.1500	75.0%	1.061
50	500	10.0	0.1483	74.2%	1.049
50	1000	20.0	0.1449	72.4%	1.025
50	5000	100.0	0.1421	71.1%	1.005
50	∞	∞	0.1414	70.7%	1

* calculated as $\sqrt{\frac{K + 1}{K}}$; 'efficiency' = $\sqrt{\frac{K}{K + 1}}$

Note: these principles apply to both measurement and count data

Sample size calculation when using unequal sample sizes to estimate / test difference in 2 means or proportions

For power (sensitivity) $1-\beta$, and specificity $1-\alpha$ (2-sided), the sample sizes n_1 and n_2 have to be such that

$$Z_{\alpha/2}\,SE(\bar{x}_1 - \bar{x}_2) - Z_\beta SE(\bar{x}_1 - \bar{x}_2) = \Delta.$$

(if $\beta < 0.5$, then Z_β will be negative). If we assume equal per unit variability, σ, of the x's in the 2 populations, we can write the requirement as

$$Z_{\alpha/2}\ \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} - Z_\beta\,\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \Delta.$$

If we rewrite $\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ as $\sqrt{\frac{1}{n_1}\left\{1+\frac{n_1}{n_2}\right\}}$

and rearrange the inequality, we get

$$n_1 = \left\{1+\frac{n_1}{n_2}\right\}(Z_{\alpha/2}-Z_\beta)^2\left\{\frac{\sigma}{\Delta}\right\}^2$$

or, denoting $\frac{n_2}{n_1}$ by K,

$$n_1 = \left\{1+\frac{1}{K}\right\}(Z_{\alpha/2}-Z_\beta)^2\left\{\frac{\sigma}{\Delta}\right\}^2$$

i.e.

$$\boxed{n_1 = \left\{\frac{K+1}{K}\right\}(Z_{\alpha/2}-Z_\beta)^2\left\{\frac{\sigma}{\Delta}\right\}^2}$$

Notes:

a. If K=1, so that $n_1=n_2$, then we get the familiar "2" at the front of the sample size formula.

b. The same factor applies for **proportions**:

If we use $\sigma_{0/1} = \sqrt{\bar{\pi}\,[1-\bar{\pi}\,]}$

as an "average" standard deviation for the individual 0's and 1's in each population, i.e.

$$\sigma_{0/1} = \sqrt{\pi\,[1-\pi\,]}$$

then, as we get the approximate formula:

$$\boxed{n_1 \approx \left\{\frac{K+1}{K}\right\}(Z_{\alpha/2}-Z_\beta)^2\left\{\frac{\bar{\pi}\,[1-\bar{\pi}\,]}{\Delta^2}\right\}}$$

Sample Size considerations... Test involving OR

Test H_0: OR = 1 vs. H_a: OR ≠ OR :

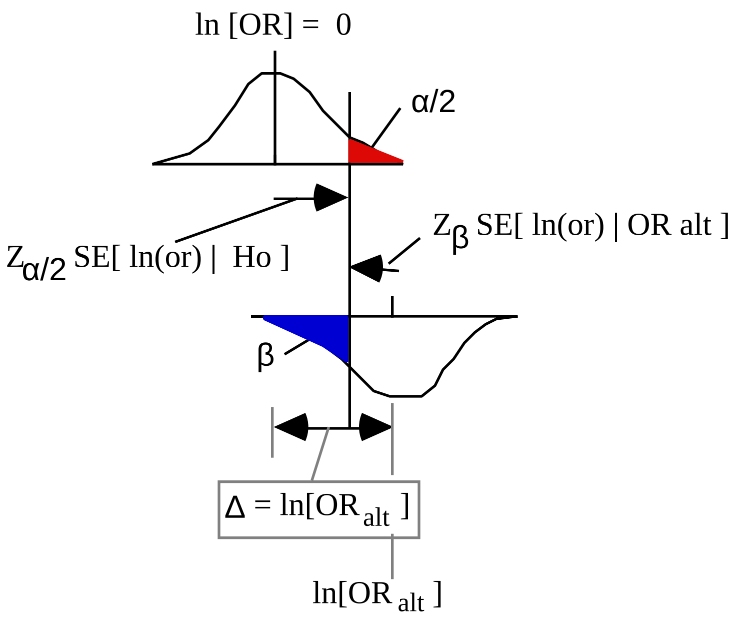
n's for power 1–β if OR = OR_{alt}; prob[type I error] = α

Here I use *ln* for natural log (elsewhere I have used log; I use them interchangeably)

Work in *ln* (or) scale; $SE[\,ln\,(or)\,] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$

Need $Z_{\alpha/2}\ SE[\,ln\,(or)\,]_0 + Z_\beta SE[\,ln\,(or)\,]_{alt} < \text{"}\Delta\text{"}$

where "Δ" = *ln* (**OR_{alt}**)


ln [OR] = 0
α/2
Z_β SE[ln(or) | OR alt]
$Z_{\alpha/2}$ SE[ln(or) | Ho]
β
Δ = ln[OR_{alt}]
ln[OR_{alt}]

Substitute expected a, b, c, d values under null and alt. into SE's and solve for numbers of cases and controls.

<u>References</u>: Schlesselman, Breslow and Day, Volume II, ...

Key points

ln [or] most precise when all 4 cells are of equal size; so...

1 increasing the control:case ratio leads to diminishing marginal gains in precision.

To see this... examine the function

$$\frac{1}{\#\text{ of cases}} + \frac{1}{\text{multiple of this }\#\text{ of controls}}$$

for various values of "multiple"

[like we did back in Chapter 8, for "effect of unequal sample sizes"]

2 The more unequal the distribution of the etiologic / preventive factor, the less precise the estimate

Examine the functions

$$\frac{1}{\#\text{ of exposed cases}} + \frac{1}{\#\text{ of unexposed cases}}$$

and

$$\frac{1}{\#\text{ of exposed controls}} + \frac{1}{\#\text{ of unexposed controls}}$$

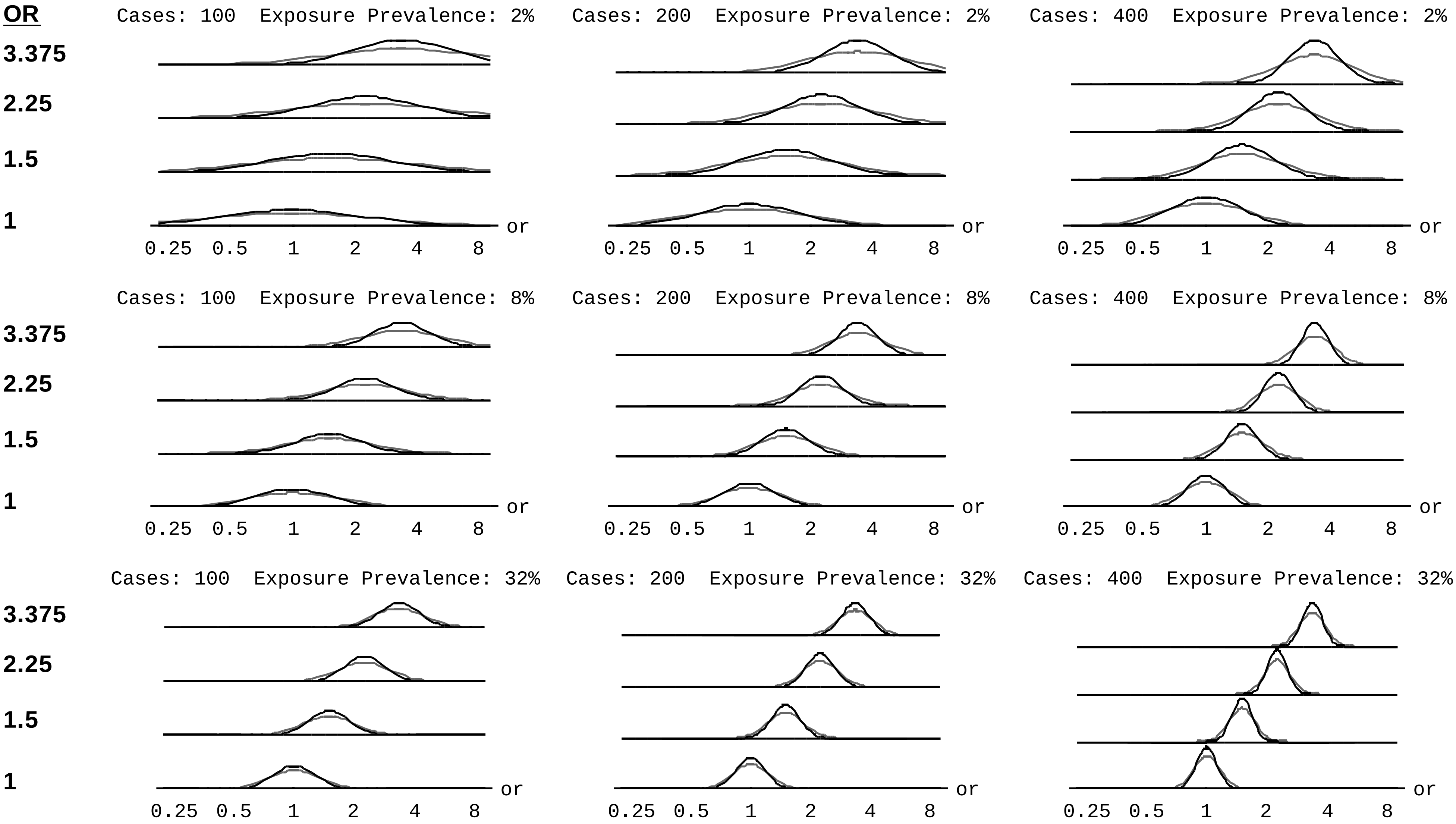
Reading graphs on next page (Note <u>log scale</u> for observed or)

Take as an example the study in the middle panel, with 200 cases, and an exposure prevalence of 8%. Say that the Type I error rate is set at α=0.05 (2sided) so that the upper critical value (the one that cuts off the top 2.5% of the null distribution) is close to or = 2. Draw a vertical line at this critical value, and examine how much of each non-null distribution falls to the right of this critical value. This area to the right of the critical value is the power of the study, i.e., the probability of obtaining a significant or, when in fact the indicated non-null value of OR is correct. Two curves at each OR value are for studies with 1(grey) and 4(black) controls/case. Note that OR values 1, 1.5, 2.25 and 3.375 are also on a log scale.

Power larger if...

i non-null OR >> 1 (cf 2.5 vs 2.25 vs 3.375)
ii exposure common (cf 2% vs 8% vs 32%) and not near universal)
iii use more cases (cf 100 vs 200 vs 400), and controls/case (1 vs 4)

Factors affecting variability of estimates from, and statistical power of, case-control studies



jh 1995-2003