Contents

1. Comparative measures
2. Large-sample CIs for comparative parameters
3. Test-based CI's
4. Sample size considerations
5. Small sample methods: conditional vs. unconditional estimation of OR
6. Examples

1 Comparative measures / parameters:

Measure	Comparative Parameter	Estimate	New Scale
(Risk or Prevalence) **Difference**	$\pi_1 - \pi_2$	$p_1 - p_2$	
(Risk or **NNT**	$1/\{\pi_1 - \pi_2\}$	$1/\{p_1 - p_2\}$	Number Needed to Treat
(Risk or Prevalence) **Ratio**	$\frac{\pi_1}{\pi_2}$	$\frac{p_1}{p_2}$	$\log \frac{p_1}{p_2} = \log p_1 - \log p_2$
Odds Ratio	$\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$	$\frac{odds_1}{odds_2}$	$log[\frac{odds_1}{odds_2}] = logit_1 - logit_2$

Cf. Rothman 2002 p. 135 Eqns 7-2, 7-3 and 7-6.

2 Large-sample CI for Comparative Parameter (if 2 component estimates are uncorrelated)

2.1 In General: (if work in new scale, must back-transform)

$$\begin{array}{lcl} estimate_1 - estimate_2 & \pm & z \times \mathrm{SE}[estimate_1 - estimate_2] \\ estimate_1 - estimate_2 & \pm & z \times (\mathrm{Var}[estimate_1] + \mathrm{Var}[estimate_2])^{1/2}. \end{array}$$

2.2 In Particular

Risk/Prevalence Difference

$$\begin{array}{rcl} p_1 - p_2 \pm z \times SE[p_1 - p_2] & = & p_1 - p_2 \pm z \times (SE^2[p_1] + SE^2[p_2])^{/2} \\ & = & p_1 - p_2 \pm z \times (p_1 q_1/n_1 + p_2 q_2/n_2)^{1/2} \end{array}$$

Risk/Prevalence Ratio

$$\text{antilog } \{\log(p_1/p_2) \pm z \times (SE^2[\log p_1] + SE^2[\log p_2])^{1/2}\},$$

where, for $i = 1, 2$,

$$SE^2[\log p_i] = Var[\log p_i] = 1/\#positive_i - 1/\#total_i.$$

Odds ratio[1]

$$\text{antilog } \{\log[oddsratio] \pm z \times (SE^2[logit_1] + SE^2[logit_2])^{1/2}\}$$

where, for $i = 1, 2$,

$$SE^2[logit_i] = Var[logit_i] = 1/\#positive_i + 1/\#negative_i.$$

$\mathrm{Var}[\log or] = \underline{1/a + 1/b} + \underline{1/c + 1/d}$ for CI_{OR} → "Woolf's Method."

2.3 Large-sample test of $\pi_1 = \pi_2$

Equivalent to test of

$$\begin{array}{rl} \pi_1 - \pi_1 = 0 \rightarrow & \text{Risk or Prevalence Difference} = 0. \\ \pi_1/\pi_2 = 1 \rightarrow & \text{Risk or Prevalence Ratio} = 1. \\ \frac{pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = 1 \rightarrow & \text{Odds Ratio} = 1. \end{array}$$

$$\begin{array}{rcl} z & = & (p_1 - p_2 - \{\Delta = 0\}) \;/\; SE[p_1 - p_2] \\ & = & (p_1 - p_2) \;/\; (p[1-p]/n_1 + p[1-p]/n_2)^{1/2} \end{array}$$

where $p = y/n$, with $y = y_1 + y_2$; $n = n_1 + n_2$.

[1] The Odds Ratio (OR) is close to the Risk Ratio when the 'denominator' odds is low, e.g. under 0.1, and the Risk Ratio is not extreme. For example, if $\pi_1 = 0.16$, and $\pi_2 = 0.08$, so that the Risk Ratio is 2, then OR = (0.16/0.84)/(0.08/0.92) = 2.2; but the approximation worsens with increasing π_2 and increasing Risk Ratio.

Examples:

0 **The generic 2×2 contingency table:**

	+	−	All
sample 1	y_1(%)	$n_1 - y_1$	n_1(100%)
sample 2	y_2(%)	$n_2 - y_2$	n_2(100%)
Total	y(%)	n - y	n(100%)

1 **Bromocriptine for unexplained primary infertility:**[2]

	Became pregnant	Did not	Total no. couples
Bromocriptine	7 (29%)	17	24(100%)
Placebo	5(22%)	18	23(100%)
Total	12(26%)	35	47(100%)

2 **Vitamin C and the common cold:**[3]

	No cold	≥ 1 cold	Total subjects
Vitamin C	105(26%)	302	407(100%)
Placebo	76(18%)	335	411(100%)
Total	181(22%)	637	818(100%)

3 **Stoke Unit vs. Medical Unit for Acute Stroke in elderly?**
Patient status at hospital discharge(BMJ 27 Sept 1980)

	Indep't.	Dep'nt	Total no. pts
Stroke Unit	67(66%)	34	101(100%)
Medical Unit	46(51%)	45	91 (100%)
Total	113(59%)	79	192(100%)

Worked example: Stroke Unit vs. Medical Unit

95% CI for $\Delta\pi$:

$$\begin{aligned} & 0.66 - 0.51 \pm z \times (0.66 \times 0.34/101 + 0.51 \times 0.49/91)^{/2} \\ = \; & 0.15 \pm 1.96 \times 0.07 \\ = \; & 0.15 \pm 0.14. \end{aligned}$$

Test $\Delta\pi = 0$: [carrying several decimal places, for comparison with χ^2]

$$\begin{aligned} z &= (0.6634 - 0.5054) \;/|;\, (0.5885 \times 0.4115 \times \{1/101 + 1/91\})^{1/2} \\ &= 0.1580 \,/\, 0.0711 \\ &= 2.22 \;\rightarrow\; P = 0.026 \text{ (2-sided).} \end{aligned}$$

Worked example: Vitamin C and the common cold

95% CI for $\Delta\pi$:

$$\begin{aligned} & 0.26 - 0.18 \pm z \times (0.26 \times 0.74/407 + 0.18 \times 0.81/411)^{/2} \\ = \; & 0.18 \pm 1.96 \times 0.03 \\ = \; & 0.18 \pm 0.06. \end{aligned}$$

Test $\Delta\pi = 0$:

$$\begin{aligned} z &= (0.258 - 0.185) \;/|;\, (0.221 \times 0.779 \times \{1/407 + 1/411\})^{1/2} \\ &= 0.073 \,/\, 0.029 \\ &= 2.52 \;\rightarrow\; P = 0.006 \text{ (1-sided) or } 0.012 \text{ (2-sided).} \end{aligned}$$

2.4 CI for Risk Ratio (a.k.a. Relative Risk) or Prevalence Ratio cf. Rothman2002 p.135

Example: Vitamin C and the common cold ... Revisited

	No cold	≥ 1 cold	Total no. subjects
Vitamin C	105(26%)	302(74%)	407(100%)
Placebo	76(18%)	335(82%)	411(100%)
Total	181(22%)	637	818(100%)

$$\widehat{RR} = \frac{Prob[\geq 1 \text{ cold} \mid \text{Vitamin C}]}{Prob[\geq 1 \text{ cold} \mid \quad \text{Placebo}]} = \frac{74\%}{82\%} = 0.91$$

CI[RR]:

$$\begin{aligned} & antilog\{\log 0.91 \pm z \times SE[\log p_1 - \log p_2]]\} \\ = \; & antilog\{\log 0.91 \pm z \times (SE^2[\log p_1] + SE^2[\log p_2])^{1/2}\}. \end{aligned}$$

$$SE^2[\log p_1] = Var[\log p_1] = 1/302 - 1/407 = 0.000854;$$
$$SE^2[\log p_2] = Var[\log p_2] = 1/335 - 1/411 = 0.000552.$$

So, CI[RR]:

$$\begin{aligned} & antilog\{\log 0.91 \pm z \times (0.000854 + 0.000552)^{1/2}\} \\ = \; & antilog\{\log 0.91 \pm 0.073\} = 0.85 \text{ to } 0.98. \end{aligned}$$

Shortcut:

Calculate $\exp\{z \times SE[\log \widehat{RR}]\}$ and use it as a multiplier and divider of $\widehat{RR}$.

In our e.g., $\exp\{z \times SE[\log \widehat{RR}]\} = \exp\{0.073\} = 1.076$.

Thus $\{RR_{LOWER}, RR_{UPPER}\} = \{0.91 \div 1.076, 0.91 \times 1.076\} = \{0.85 \text{ to } 0.98\}$.

You can use this shortcut whenever you are working with log-based CI's that you convert back to the original scale, there they become "multiply-divide" symmetric rather than "plus-minus" symmetric.

```
SAS                         Stata

PROC FORMAT;                Immediate:  csi 302 335 105 76
VALUE onefirst 0="z0"       cs stands for 'cohort study'
1="a1";
DATA CI_RR_OR;              input vitc cold npeople
INPUT vitC cold npeople;
LINES;
1 1 302                     1 1 302
1 0 105                     1 0 105
0 1 335                     0 1 335
0 0 76                      0 0 76
;                           end
PROC FREQ data=CI_RR_OR
ORDER=FORMATTED;            cs cold vitc [freq=npeople]
TABLES vitC*cold / CMH;
WEIGHT npeople;
FORMAT vitC cold onefirst;
RUN;
```

2.5 CI for Odds Ratio **cf. Rothman 2002 p. 139**

	Vitamin C	Placebo
had cold(s)	302	335
avoided colds	105	76
# with cold(s) for every 1 who avoided colds	2.88 (:1)	4.41 (:1)
odds of cold(s)	**2.88**	**4.41**
odds **Ratio** = $\frac{2.88}{4.41}$	= 0.65 → $\widehat{OR} = 0.65$	

$$CI[OR] = antilog\ \{\log[oddsRatio] \pm z\ SE[logit_1 - logit_2]\ \}$$

$$SE^2[logit_1] = \frac{1}{\#positive_1} + \frac{1}{\#negative_1}$$
$$SE^2[logit_1] = \frac{1}{\#positive_2} + \frac{1}{\#negative_2}$$

$$SE[logit_1 - logit_2] = \left\{\left(\frac{1}{302} + \frac{1}{105}\right) + \left(\frac{1}{335} + \frac{1}{76}\right)\right\}^{1/2} = 0.17$$

$$z \times SE[logit_1 - logit_2] = 1.96 \times 0.17 = 0.33$$

$$antilog\ \{\log 0.65 \pm 0.33\ \} = \exp\ \{-0.43 \pm 0.33\} = 0.47to0.90$$

```
From SAS
```

See statements for RR (output gives both RR and OR)

Be CAREFUL as to rows / cols. Index exposure category must be 1st row; reference exposure category must be 2nd.

If necessary, use FORMAT to have table come out this way ... (note trick to reverse rows / cols)

SAS doesn't know if it data come from a 'case-control' or 'cohort' study.

```
From Stata
Immediate: cci 302 335 105 76, woolf

cc stands for 'case control study'

input vit_c cold n_people
        1     1     302
        1     0     105
        0     1     335
        0     0     76

end
cc cold vit_c [freq=n_people], woolf
```

3 "Test-based CI's"

3.1 Preamble

In 1959, when Mantel and Haenszel developed their summary Odds Ratio measure over 2 or more strata, they did not supply a CI to accompany this point estimate. From 1955 onwards, the main competitor was the weighted average (in the log OR scale) and accompanying CI obtained by Woolf. But this latter method has problems with strata where one or more cell frequencies are zero. In 1976, Miettinen developed the "test-based" method for epidemiologic situations where the summary point estimate is easily calculated, the standard error estimate is unknown or hard to compute, but where a statistical test of the null value of the parameter of interest (derived by aggregating a "sub-statistic" from each stratum) is already available. Although the 1886 development, by Robins, Breslow and Greenland, of a direct standard error for the log of the Mantel-Haenszel OR estimator, the "test-based" CI is still used (see A&B KKM).

Even though its main usefulness is for summaries over strata, the idea can be explained using a simpler and familiar (single starum) example, the comparison of two independent means using a z-test with large *df* (the principle does not depend on t *vs.* z). Suppose all that was reported was the difference in sample means, and the 2-sided p-value associated with a test of the null hypothesis that the mean difference was zero. From the sample means, and the p-value, how could we obtain a 95%CI for the difference in the 'population' means? The trick is to

1. work back (using a table of the normal distribution) from the p-value to the corresponding value of the z-statistic (the number of standard errors that the difference in sample means is from zero);
2. divide this observed difference by the observed z value, to get the standard error of the difference in sample means, and
3. use the observed difference, and the desired multiple (1.645 for 90% CI, 1.96 for 95% etc.) to create the CI.

The same procedure is directly applicable for the difference of two independently estimated proportions. If one tests the (null) difference using a z-test, one can obtain the SE of the difference by dividing the observed difference in proportions by the z statistic; if the difference was tested by a chi-square statistic, one can obtain the z-statistic by taking the square root of the observed chi-square value (authors call this square root an observed 'chi' value). Either way, the observed z-value leads directly to the SE, and from there to the CI. This is worked out in the next example, where it is assumed that the null hypothesis is tested via a chi-squared (χ^2) test.

3.2 "Test-based" CI's ... specific applications

- **Difference of 2 proportions** $\pi_1 - \pi_2$ (Risk or Prevalence Difference)

  Observe: p_1 and p_2 and (maybe via p-value) the calculated value of X^2 This implies that

  $$(observed\ X^2\ value)^{1/2} = observed\ X\ value = observed\ z\ value;$$

  But... observed z statistic $= (p_1 - p_2)\ /\ SE[p_1 - p_2]$.

  So... $SE[p_1 - p_2] = (p_1 - p_2)\ /\ observed\ z\ statistic \qquad \{use\ +\text{ve}\ sign\}$

  95% CI for $p_1 - p_2$:

  $$(p_1 - p_2) \mp \{z\ value\ for\ 95\%\} \times SE[p_1 - p_2]$$

  i.e. ...

  $$(p_1 - p_2) \mp \{z\ value\ for\ 95\%\} \times \frac{p_1 - p_2}{observed\ z\ statistic}$$

  i.e., after re-arranging terms ...

  $$(p_1 - p_2)\left\{1 \mp \frac{z\ value\ for\ 95\%\ CI}{observed\ z\ statistic}\right\} \qquad (1a)$$

or, in terms of a reported chi-squared statistic

$$(p_1 - p_2)\left\{1 \mp \frac{z\ value\ for\ 95\%\ CI}{Sqrt[observed\ chi - squared\ statistic]}\right\}. \quad (1b)$$

See Section 12.3 of Miettinen's "Theoretical Epidemiology".

Technically, when the variance is a function of the parameter (as is the case with binary response data), the test-based CI is most accurate close to the Null. However, as you can verify by comparing test-based CIs with CI's derived in other ways, the inaccuracies are not as extreme as textbooks and manuals (e.g. Stata) suggest.

- **Ratio** of 2 proportions π_1 / π_2
  (Risk Ratio; Prevalence Ratio; Relative Risk; "RR")

  Observe:

  1. $rr = p_1 \,/\, p_2$ and
  2. (maybe via p-value) the value of X^2 statistic (H_0: RR = 1) $\rightarrow (observed\, X^2 value)^{1/2} = observed\, X\ value = observed\, z\ value.$

  In log scale, in relation to $log[RR_{null}] = 0$, observed z value would be:

$$observed\ z\ value = \frac{\log rr - 0}{SE[\log rr]}$$

  This implies that

$$SE[\log rr] = \frac{log[rr]}{observed\ z\ value} \quad \{use\ \ +ve\ sign\}$$

  95% CI for $\log RR$:

$$\log rr \mp \{z\ value\ for\ 95\%\ CI\} \times SE[\log rr]$$

  i.e. ...

$$\log rr \mp \{z\ value\ for\ 95\%\ CI\} \times \frac{log[rr]}{observed\ z\ value}$$

  i.e., after re-arranging terms ...

$$log[rr] \times \left\{1 \pm \frac{z\ value\ for\ 95\%\ CI}{observed\ z\ statistic}\right\} \quad (2a)$$

  Going back to RR scale, by taking antilogs[4]...

  95% CI for RR:

$$rr\ to\ power\ of\ \left\{1 \pm \frac{z\ value\ for\ 95\%}{observed\ z\ statistic}\right\} \quad (2b)$$

- **Ratio** of 2 odds $\pi_1/(1-\pi_1)$ and $\pi_2/(1-\pi_2)$ (Odds Ratio; "OR")

  Observe:

  1. $or = \frac{p_1/(1-p_2)}{p_2/(1-p_2)} = \frac{ad}{bc}$ and
  2. (maybe via p-value) the value of X^2 statistic (H_0: OR = 1) $\rightarrow (observed\ X^2 value)^{1/2} = observed\ X\ value = observed\ z\ value$

  In log scale, in relation to $\log[OR_{null}] = 0$, observed z value would be:

$$observed\ z\ value \ = \frac{\log or - 0}{SE[\log or]}$$

  This implies that

$$SE[\log or] = \frac{\log or}{observed\ z\ value} \quad use\ \ +ve\ sign$$

  95% CI for $\log OR$:

$$\log or \mp \{z\ value\ for\ 95\%\ CI\} \times SE[\log or] \quad (3a)$$

  i.e. ...

$$\log or \pm \{z\ value\ for\ 95\%\ CI\} \times \frac{\log or}{observed\ z\ value} \quad (3b)$$

  i.e., after re-arranging terms ...

$$\log or \pm \times \left\{1 \pm \frac{z\ value\ for\ 95\%\ CI}{observed\ z\ statistic}\right\}$$

  Going back to OR scale, by taking antilog$s$[5]...

  95% CI for OR:

$$or\ \ to\ power\ of \left\{1 \pm \frac{z\ value\ for\ 95\%}{observed\ z\ statistic}\right\}$$

  See Section 13.3 of Miettinen's "Theoretical Epidemiology"

[4] $antilog[log[a] \infty b] = \exp[log[a] \infty b] = \{\exp[log[a]]\}$ to power of $b = a$ to power of b

4 Sample Size considerations...

4.0.1 CI for $\pi_1 - \pi_2$

n's to produce CI for difference in π's of pre specified margin of error (ME) at stated confidence level

- large-sample CI: $p_1 - p_2 \pm Z\, SE[p_1 - p_2] = p_1 - p_2 \pm ME$
- $SE[p_1 - p_2] = \{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2\}^{1/2}$.

  Simplify (involves some approximation) by using an average p.

  If use equal n's, then

  $$n\ per\ group = \frac{2 \times p(1-p) \times Z^2_{\alpha/2}}{ME^2}$$

  M&M use the fact that if $p = 1/2$ then $p(1-p) = 1/4$, and so $2p(1-p) = 1/2$, so the above equation becomes

  $$[max]\ n\ per\ group = \frac{\frac{1}{2}\, Z^2_{\alpha/2}}{ME^2}$$

4.0.2 Test involving π_T and π_C

Test H_0: $\pi_T = \pi_C$ vs. H_a: $\pi_T \neq \pi_C$:

n's for power $1-\beta$ if $\pi_T = \pi_C + \Delta$; $prob[Type\ I\ error] = \alpha$

n per group

$$\begin{aligned} &= \frac{\{Z_{\alpha/2}\sqrt{2\pi_C\{1-\pi_C\}} - Z_\beta\sqrt{\pi_C\{1-\pi_C\} + \pi_T\{1-\pi_T\}}\}^2}{\Delta^2} \\ &\approx 2(Z_{\alpha/2} - Z_\beta)^2\left\{\frac{\bar{\pi}(1-\bar{\pi})}{\Delta^2}\right\} \\ &= 2\{Z_{\alpha/2} - Z_\beta\}^2\left\{\frac{\sigma_{0,1}}{\Delta}\right\}^2 \qquad (4) \end{aligned}$$

If $\alpha = 0.05(2-sided)$ & $\beta = 0.2 ... Z_\alpha = 1.96$; $Z_\beta = -0.84$, then $2(Z_{\alpha/2} - Z_\beta)^2 = 2\{1.96 - (-0.84)\}^2 \approx 16$, i.e. $n\,per\,group \approx 16 \times \frac{\bar{\pi}\{1-\bar{\pi}\}}{\Delta^2}$.

→ $n_T \approx 100$ & $n_C \approx 100$ if $\pi_T = 0.6$ & $\pi_C = 0.4$.

See Sample Size Requirements for Comparison of 2 Proportions (from text by Smith and Morrow) under Resources for Chapter 8.

Effect of Unequal Sample Sizes ($n_1 \neq n_2$) on precision of estimated differences: See Notes on Sample Size Calculations for Inferences Concerning Means.

4.0.3 Test involving OR

Test H_0: OR = 1 vs. H_a: OR $\neq$ OR:

n's for power $1-\beta$ if $OR = OR_{alt}$; Prob[Type I error] = α.

Work in $\log or$ scale; $SE[\log or] = (1/a + 1/b + 1/c + 1/d)^{1/2}$.

Need

$$Z_{\alpha/2}\, SE_0[\log or] + Z_\beta SE_{alt}[\log or] < \Delta.$$

where

$$\Delta = \log[OR_{alt}]$$

Substitute expected $a, |; b,\ c,\ d$ values under null and alt. into SE's and solve for number of cases and controls.

References: Schlesselman, Breslow and day, Volume II, ...

Key Points: $\log or$ 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

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

   for various values of "multiple" [cf earlier notes "effect of unequal sample sizes"]

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

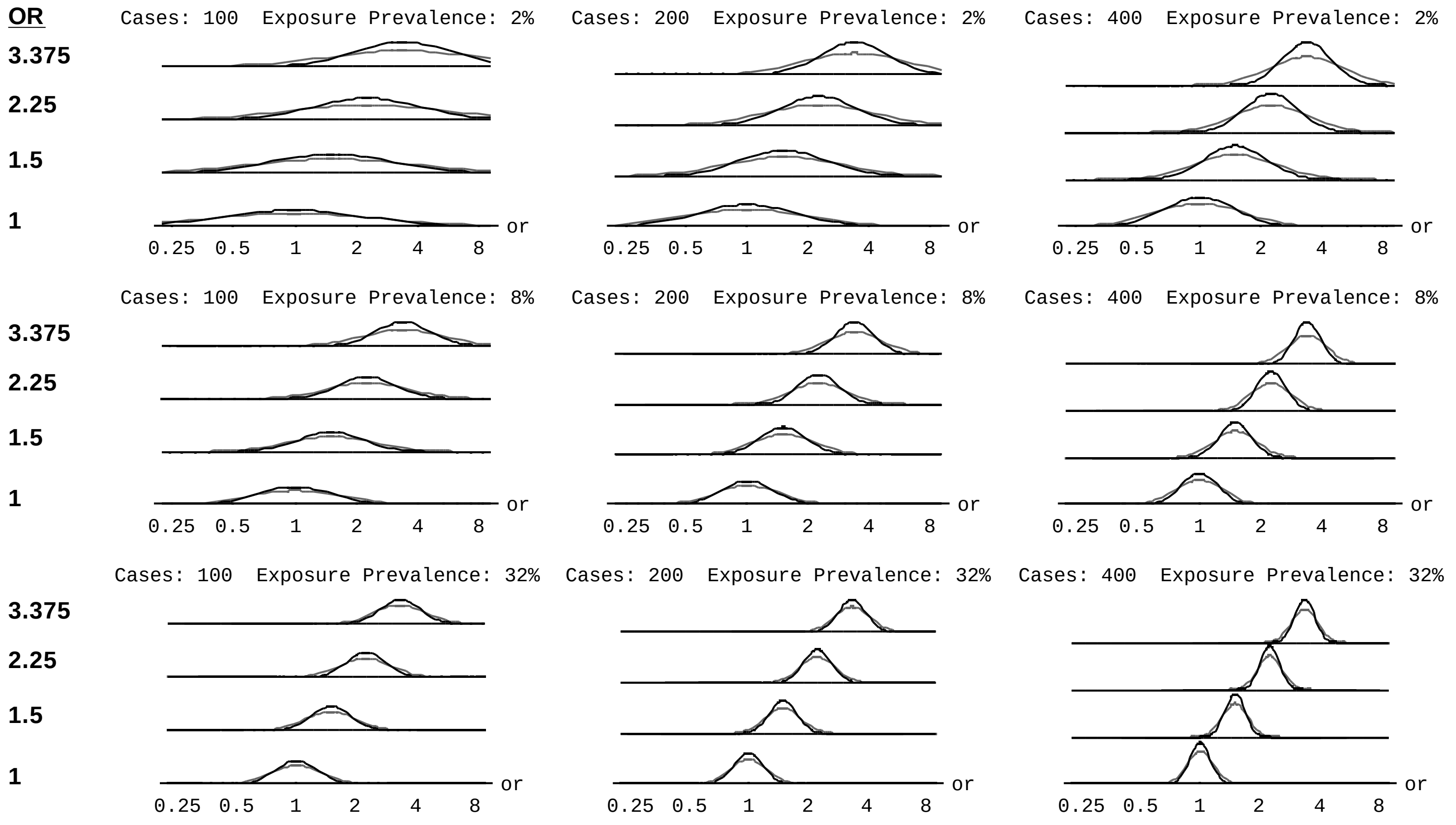
   Examine the functions

   $$1/(\text{no. of exposed cases}) + 1/(\text{no. of unexposed cases})$$

   and

   $$1/(\text{no. of exposed controls}) + 1/(\text{no. of unexposed controls}).$$

Factors affecting variability of estimates from, and statistical[5] power of, case-control studies



jh 1995-2003

Reading graphs: (Note log scale 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 $\alpha = 0.05$ (2-sided) 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.

[5] **Power larger if ...** 1. non-null OR >> 1 (cf. 2.5 vs 2.25 vs 3.375); 2 exposure common (cf. 2% vs 8% vs 32%) and not near universal; 3 use more cases (cf. 100 vs. 200 vs. 400), and controls/case (1 vs 4).

5 Small sample methods:

Test:

Since a risk difference of zero implies a risk ratio, or odds ratio, of 1, all three can be tested in the same way.

U (unconditional)
Suissa S; Shuster JJ. Exact Unconditional Sample Sizes for the 2×2 Binomial Trial; Journal of the Royal Statistical Society. Series A (General) Vol. 148, No. 4 (1985), pp. 317-327.

C (conditional)
Fisher 1935, JRSS Vol 98, p 48. (central) Hypergeometric distribution, obtained by conditioning on (treating as fixed) all marginal frequencies.

Confidence Interval:

5.1 Risk Difference

See section 3.1.2 of Sahai and Khurshid (1996).

5.2 Risk Ratio

See section 3.1.2 of Sahai and Khurshid (1996).

5.3 Odds Ratio: Point- and Interval-estimation

See section 4.1.2 of Sahai and Khurshid (1996), and Chapter of Volume I of Breslow and Day. See also example 1, pp 48-51, in Fisher 1935.

Elaboration on equation 4.11 in Sahai and Khurshid , and on the (what we now call the *non-central* hypergeometric random variable whose distribution is given in the middle of p 50 of Fisher's article.

Let $Y_i \sim$ Binomial(n_i, π_i), $i = 1, 2$, be 2 independent binomial random variables.

We wish to make inference regarding the parameter

$$\psi = \{\pi_1/(1-\pi_1\}/\{\pi_2/(1-\pi_2\}.$$

We can do so by considering only those data configurations which have the same total number of 'positives', $y_1 + y_2 = y$, say, as were observed in the actual study, and then considering the distribution of $Y_1 \mid y$.

$$Prob[Y_1 = y_1 \,;\, Y_2 = y_2] = \;{}^{n_1}\mathrm{C}_{y_1}\, \pi_1^{y_1}(1-\pi_1)^{n_1-y_1} \times \;{}^{n_2}\mathrm{C}_{y_2}\, \pi_2^{y_2}(1-\pi_2)^{n_2-y_2}.$$

If we condition on $Y_1 + Y_2 = y$, then

$$Prob[Y_1 = y_1 \mid Y_1 + Y_2 = y] = Prob[Y_1 = y_1 \,;\, Y_2 = y - y_1]/Prob[Y_1 + Y_1 = y].$$

If we rewrite the quantity

$$\pi_1^{y_1}(1-\pi_1)^{n_1-y_1} \times \pi_2^{y_2}(1-\pi_2)^{n_2-y_2}$$

as

$$\pi_1^{y_1}(1-\pi_1)^{-y_1}\pi_2^{-y_2}(1-\pi_2)^{y_1} \times (1-\pi_1)^{n_1}\pi_2^{y}(1-\pi_2)^{n-y}$$

we see that it simplifies to

$$\psi^{y_1} \times (1-\pi_1)^{n_1}\, \pi_2^{y}\, (1-\pi_2)^{n-y}$$

and that the last three terms do not involve ψ and do not involve the random variable y_1. Since they appear in both the numerator and the denominator of the conditional probability, they cancel out.

This we can write the conditional probability $Prob[Y_1 = y_1 \mid Y_1 + Y_2 = y]$ as

$$Prob[\, y_1 \mid y\,] = \;{}^{n_1}\mathrm{C}_{y_1}\; {}^{n_2}\mathrm{C}_{y-y_1}\; \psi^{\,y_1} \,/\, \Sigma\; {}^{n_1}\mathrm{C}_{y_1'}\; {}^{n_2}\mathrm{C}_{n-y_1'}\; \psi^{\,y_1'},$$

where the summation is over those y_1' values that are compatible with the 4 marginal frequencies.

Aside: you will note that if we set $\psi = 1$, the probabilities are the same as those in the central hypergeometric distribution, used for Fisher's exact test of two binomial proportions. Indeed, Fisher, in page 48-49 of his 1935 paper, first computes the null probabilities for the 2×2 table.

Conviction of Like-sex Twins of Criminals

	Convicted.	Not Convicted.	Total.
Monozygotic	10(a)	3(b)	13
Dizygotic	2(c)	15(d)	17
Total	12	18	30

[We use y_1 and y_2 where epidemiologists typically use a and c.]

He calculated that the probability that 1, 2, 3, ... monozygotic twins would escape conviction[6] was $(1/6\ 652\ 325) \times \{1, 102, 2992, ...\}$. Thus, "a discrepancy from proportionality as great or greater than that observed, will arise, subject to the conditions specified by the ancillary information, in exactly 3,095 trials out of 6,652,325 or approximately once in 2,150 trials."

He then went on to work out the lower limit of the 90% 2-sided CI (or a 95% 1-sided CI), for the odds ratio: i.e. for the odds, $\pi_{mono-z}/(1-\pi_{mono-z})$, of criminals to non-criminals in twins of monozygotic criminals divided by the corresponding odds $\pi_{di-z}/(1-\pi_{di-z})$, in twins of dizygotic criminals.

Let Y_{mono} be the number of MZ twins convicted. Fisher finds the value ψ_L such that

$$Prob[\,Y_{mono} \ge 10 \mid \psi_L\ ,\ y = 12\,] = 0.05.$$

He reports that this value is $1/0.28496 \approx 3.509$. In the Excel spreadsheet for Fisher's exact test and exact CI for OR (on website), you can verify that indeed, with $\psi_L = 3.509$, $Prob[\,Y_{mono} \ge 10 \mid \psi = 3.509\ ,\ y = 12\,] = 0.05$.

One has to admire Fisher's ability, in 1935, to solve a polynomial equation of order 12, namely

$$\frac{1 + 102\psi + 2992\psi^2}{1 + 102\psi + 2992\psi^2 + \cdots + 476\psi^{12}} = 0.05.$$

5.3.1 Point estimation of ψ under Hypergeometric Model

See section x.x of Breslow and Day, Volume I.

It will come as a surprise to many that *there are 2 point estimators of* ψ:

one, the familiar – *unconditional* – based on the "2 independent Binomials" model, with two random variables y_1 and y_2, and

the other – *conditional* – based on the *single* random variable $y_1 \mid y$ with a Non-Central Hypergeometric distribution.

While the two estimators yield similar estimates when sample sizes are large, the estimates can be quite different from each other in small sample situations.

Estimator, based on Unconditional Approach:

The estimator derives from the principle that if there are two parameters θ_1 and θ_2, with Maximum Likelihood Estimators $\hat{\theta_1}$ and $\hat{\theta_2}$, then the Maximum Likelihood Estimator of θ_1/θ_2 is $\hat{\theta_1}/\hat{\theta_2}$.

Thus, since $\hat{\pi}_1 = 10/13$, and $\hat{\pi}_2 = 2/17$, we have

$$\hat{\psi}_{UMLE} = \frac{(10/13)/(2/13)}{(2/17)/(15/17)} = \frac{10 \times 15}{3 \times 2} = 25 = \frac{a \times d}{b \times c}.$$

Estimator, based on Conditional Approach:

The Maximum Likelihood Estimate $\hat{\psi}_{CMLE}$ is the solution of $d \log L/d\psi = 0$.

If we use Σ as shorthand for the denominator of $prob[\,y_1 \mid y\,]$, then $\hat{\psi}_{CMLE}$ is the solution of

$$\frac{y_1}{\psi} = \frac{d \log \Sigma}{d\psi} = \frac{d\Sigma}{d\psi} \times \frac{1}{\Sigma}.$$

Re-arranging, we find that $\hat{\psi}_{CMLE}$ is the solution of

$$y_1 = \mathrm{E}[\,Y_1 \mid \psi\,].$$

In this case the CMLE of ψ is the same as the estimate obtained by equating the observed and expected moment (the "Method of Moments").

Using the same spreadsheet used above, we find that the value of ψ that satisfies this estimating equation is

$$\hat{\psi}_{CMLE} = 21.3.$$

It can be shown that, in any given dataset, $\hat{\psi}_{CMLE}$ is *closer to the null* (i.e., to $\psi = 1$) than the $\hat{\psi}_{MLE}$ is. Indeed, it the CMLE can be can be seen as a UMLE that has been shrunk towards the null.[7]

[8]

[6] the range is 1 to 13; 0 cannot escape, since then there would be 13 convicted in the first row, but there are only 12 convicted in all.

[7] See Hanley JA, Miettinen OS. An Unconditional-like Structure for the Conditional Estimator of Odds Ratio from 2 x 2 Tables. Biometrical Journal 48 (2006) 1, 2334 DOI: 10.1002/bimj.200510167

[8] [Notes from JH]:

- The 5 tables from the tea-tasting experiment with the 2x2 tables with all marginal totals = 4 are another example of this hypergeometric distribution
- Excel has the Hypergeometric probability function. It is like the Binomial , except that instead of specifying p, one specifies the size of the POPULATION and the NUMBER OF POSITIVES IN THE POPULATION .. example, to get P_1 above, one would ask for HYPERGEODIST(a;r1;c1;N)
  The spreadsheet "Fisher's Exact test" uses this function; to use the spreadsheet, simply type in the 4 cell frequencies, a, b, c, and d. the spreadsheet will calculate the probability for each possible table. then you can find the tail areas yourself. You can also use it for the non-null (non-central) hypergeometric distribution.

5.4 The “Exact” Test for 2 x 2 tables

5.4.1 Material taken from Armitage & Berry §4.9. Material on hand-calculation of null probabilities is omitted

Even with the continuity correction there will be some doubt about the adequacy of the χ^2 approximation when the frequencies are particularly small. An exact test was suggested almost simultaneously in the mid-1930s by R. A. Fisher, J. O. Irwin and F. Yates. It consists in calculating the exact probabilities of the possible tables described in the previous subsection. The probability of a table with frequencies

$$\begin{array}{cc|c} a & b & r_1 \\ c & d & r_2 \\ \hline c_1 & c_2 & N \end{array}$$

is given by the formula

$$P[a|r_1, r_2, c_1, c_2] = \frac{r_1!r_2!r_3!r_4!}{N!a!b!c!d!} \tag{5}$$

This is, in fact, the probability of the observed cell frequencies conditional on the observed marginal totals, under the null hypothesis of no association between the row and column classifications. Given any observed table, the probabilities of all tables with the same marginal totals can be calculated, and the P value for the significance test calculated by summation. Example 4.14 illustrates the calculations and some of he difficulties of interpretation which may arise. The data in Table 4.6, due to M. Hellman, are discussed by Yates (1934).

Table 1: Data on malocclusion of teeth in infants (Yates, 1934)

	Infants with		
	Normal teeth	Malocclusion	Total
Breast-fed	4	16	20
Bottle-fed	1	21	22
Total	5	37	42

There are six possible tables with the same marginal totals as those observed. since neither a nor c (in the notation given above) can fall below 0 or exceed 5, the smallest marginal total in the table. The cell frequencies in each of these tables are shown in Table 2 Below them are shown the probabilities of these tables, calculated under the null hypothesis.

Table 2: Cell frequencies in tables with the same marginal totals as those in Table 1

	0	20	20	1	19	20	2	18	20	3	17	20	4	16	20
	5	17	22	4	18	22	3	19	22	2	20	22	1	21	22
	5	37	42	5	37	42	5	37	42	5	37	42	5	37	42
a	0			1			2			3			4		
P_a	0.1720			0.3440			0.3096			0.1253			0.0182		

Table 2 continued ...

	5	15	20
	0	22	22
	5	37	42
a	5		
P_a	0.0182		

This is the complete conditional distribution for the observed marginal totals, and the probabilities sum to unity as would be expected. Note the importance of carrying enough significant digits in the first probability to be calculated; the above calculations were carried out with more decimal places than recorded by retaining each probability in the calculator for the next stage. The observed table has a probability of 0.1253. To assess its significance we could measure the extent to which it falls into the tail of the distribution by calculating the probability of that table or of one more extreme. For a one-sided test the procedure clearly gives $P = 0.1253 + 0.0182 = 0.1435$. The result is not significant at even the 10% level.

For a two-sided test the other tail of the distribution must be taken into account, and here some ambiguity arises. Many authors advocate that the one-tailed P value should be doubled. In the present example, the one-tailed test gave $P = 0.1435$ and the two-tailed test would give P = 0.2870. An alternative approach is to calculate P as the total probability of tables, in either tail, which are at least as extreme as that observed in the sense of having a probability at least as small. In the present example we should have

$$P = 0.1253 + 0.0182 + 0.0310 = 0.1745$$

The first procedure is probably to be preferred on the grounds that a significant result is interpreted as strong evidence for a difference in the observed direction, and there is some merit in controlling the chance probability of such a result to no more than half the two-sided significance level.

The results of applying the exact test in this example may be compared with those obtained by the χ^2 test with Yates's correction. We find $X^2 = 2.39$, $(P = 0.12)$ without correction and $X^2_C = 1.14$, $(P = 0.29)$ with correction. The probability level of 0.29 for X^2_C agrees well with the two-sided value 0 29 from the exact test, and the probability level of 0.12 for X^2 is a fair approximation to the exact mid-P value of 0.16.

Cochran (1954) recommends the use of the exact test, in preference to the X^2 test with continuity correction, (i) if $N < 20$, or (ii) $20 < N < 40$ and the smallest expected value is less than 5. With modern scientific calculators and statistical software the exact test is much easier to calculate than previously and should be used for any table with an expected value less than 5.

The exact test and therefore the χ^2 test with Yates's correction for continuity have been criticized over the last 50 years on the grounds that they are conservative in the sense that a result significant at, say, the 5% level will be found in less than 5% of hypothetical repeated random samples from a population in which the null hypothesis is true. This feature was discussed in §4.7 and it was remarked that the problem was a consequence of the discrete nature of the data and causes no difficulty if the precise level of P is stated. Another source of criticism has been that the tests are conditional on the observed margins, which frequently would not all be fixed. For example, in Example 4.14 one could imagine repetitions of sampling in which 20 breast-fed infants were compared with 22 bottle-fed infants but in many of these samples the number of infants with normal teeth would differ from 5. The conditional argument is that, whatever inference can be made about the association between breast-feeding and tooth decay, it has to be made within the context that exactly five children had normal teeth. If this number had been different then the inference would have been made in this different context, but that is irrelevant to inferences that can be made when there are five children with normal teeth. Therefore, we do not accept the various arguments that have been put forward for rejecting the exact test based on consideration of possible samples with different totals in one of the margins. The issues were discussed by Yates 1984) and in the ensuing discussion, and by Barnard (1989) and Upton (1992), and we will not pursue this point further. Nevertheless, the exact test and the corrected χ^2 test have the undesirable feature that the average value of the significance level, when the null hypothesis is true, exceeds 0.5. The mid-P value avoids this problem, and so is more appropriate when combining results from several studies (see §4.7).

As for a single proportion, the mid-P value corresponds to an uncorrected χ^2 test, whilst the exact P value corresponds to the corrected χ^2 test. The confidence limits for the difference, ratio or odds ratio of two proportions based on the standard errors given by (4.14), (4.17) or (4.19) respectively are all approximate and the approximate values will be suspect if one or more of the frequencies in the 2 x 2 table are small. Various methods have been put forward to give improved limits but all of these involve iterations and are tedious to carry out on a calculator. The odds ratio is the easiest case. Apart from exact limits, which involve an excessive amount of calculation, the most satisfactory limits are those of Cornfield (1956); see Example 16.1 and Breslow and Day (1980, §4.3) or Fleiss (1981, §5.6). For the ratio of two proportions a method was given by Koopman (1984) and Miettinen and Nurminen (1985) which can be programmed fairly readily. The confidence interval produced gives a good approximation to the required confidence coefficient, but the two tail probabilities are unequal due to skewness. Gart and Nam (1988) gave a correction for skewness but this is tedious to calculate. For the difference of two proportions a method was given by Mee (1984) and Miettinen and Nurminen (1985). This involves more calculation than for the ratio limits, and again there could be a problem due to skewness (Gart and Nam, 1990).

Notes by JH

- The word "exact" means that the p-values are calculated using a finite discrete reference distribution – the hypergeometric distribution (cousin of the binomial) rather than using large-sample approximations. It doesn't mean that it is the correct test. [see comment by A&B in their section dealing with Mid-P values].

  While greater accuracy is always desirable, this particular test uses a 'conditional' approach that not all statisticians agree with. Moreover, compared with some unconditional competitors, the test is somewhat conservative, and thus less powerful, particularly if sample sizes are very small.

- Fisher's exact test is usually used just as a test*; if one is interested in the difference $\Delta = \pi_1 \pi_2$, the conditional approach does not yield a corresponding confidence interval for Δ. [it does provide one for the comparative odds ratio parameter $\psi = \frac{\pi_1/(1-\pi_1}{\pi_2/(1-\pi_2}$.

- Thus, one can find anomalous situations where the (conditional) test provides $P > 0.05$ making the difference 'not statistically significant', whereas the large-sample (unconditional) CI for Δ, computed as $p_1 - p_2 \pm z\, SE(p_1 - p_2)$, does not overlap 0, and so would indicate that the difference is 'statistically significant'. [* see the Breslow and Day text Vol I , §4.2, for CI's for ψ derived from the conditional distribution]

- See letter from Begin & Hanley re 1/20 mortality with pentamidine vs 5/20 with Trimethoprim-Sulfamethoxazole in patients with Pneumocystis carinii Preumonia-Annals Int Med 106 474 1987.

- Miettinen's test-based method of forming CI's, while it can have some drawbacks, keeps the correspondence between test and CI and avoids such anomalies.

- This illustrates one important point about parameters related to binary data – with means of interval data, we typically deal just with differences*; however, with binary data, we often switch between differences and ratios, either because the design of the study forces us to use odds ratios (case-control studies), or because the most readily available regression software uses a ratio (i.e. logistic regression for odds ratios) or because one is easier to explain that the other, or because one has a more natural interpretation (e.g. in assessing the cost per life saved of a more expensive and more efficacious management modality, it is the difference in, rather than the ratio of, mortality rates that comes into the calculation). [* the sampling variability of the estimated ratios of means of interval data is also more difficult to calculate accurately].

6 (Mis-)Application; Costly Application

6.1 Fisher's Exact Test in a Double-Blind study of Symptom Provocation to Determine Food Sensitivity (N Engl J Med 1990; 323: 429-33)

Abstract

Background Some claim that food sensitivities can best be identified by intradermal injection of extracts of the suspected allergens to reproduce the associated symptoms. A different dose of an offending allergen is thought to "neutralize" the reaction.

Methods To assess the validity of symptom provocation, we performed a double-blind study that was carried out in the offices of seven physicians who were proponents of this technique and experienced in its use. Eighteen patients were tested in 20 sessions (two patients were tested twice) by the same technician, using the same extracts (at the same dilutions with the same saline diluent) as those previously thought to provoke symptoms during unblinded testing. At each session three injections of extract and nine of diluent were given in random sequence. The symptoms evaluated included nasal stuffiness, dry mouth, nausea, fatigue, headache, and feelings of disorientation or depression. No patient had a history of asthma or anaphylaxis.

Results The responses of the patients to the active and control injections were indistinguishable, as was the incidence of positive responses: 27 percent of the active injections (16 of 60) were judged by the patients to be the active substance, as were 24 percent of the control injections (44 of 180). Neutralizing doses given by some of the physicians to treat the symptoms after a response were equally efficacious whether the injection was of the suspected allergen or saline. The rate of judging injections as active remained relatively constant within the experimental sessions, with no major change in the response rate due to neutralization or habituation.

Conclusions When the provocation of symptoms to identify food sensitivities is evaluated under double-blind conditions, this type of testing, as well as the treatments based on "neutralizing" such reactions, appears to lack scientific validity. The frequency of positive responses to the injected extracts appears to be the result of suggestion and chance

Calculated according to Fisher's exact test, which assumes that the hypothesized direction of effect is the same as the direction of effect in the data. Therefore, when the effect is opposite to the hypothesis, as it is for the data below those of Patient 9, the P value computed is testing the null hypothesis that the results obtained were due to change as compared with the possibility that the patients were more likely to judge a placebo injection as active than an active injection.

Responses of 18 Patients Forced to Decide Whether Injections Contained an Active Ingredient or Placebo

Pt. No*	Active Injection		Placebo Injection		P Value
	resp	no resp	resp	no resp	
3	2	1	1	8	0.13
1	2	1	2	7	0.24
14a	2	1	2	7	0.24
12	1	2	0	9	0.25
16	2	1	3	6	0.36
18	2	1	4	5	0.50
14b	1	2	2	7	0.87
4	1	2	2	7	0.87
5	1	2	2	7	0.87
9	0	3	0	9	—
2a	0	3	1	8	0.75
13	0	3	1	8	0.75
15	1	2	3	6	0.76
6	0	3	2	7	0.55
8	0	3	2	7	0.55
17	1	2	5	4	0.50
2b	0	3	3	6	0.38
7	0	3	3	6	0.38
10	0	3	3	6	0.38
11	0	3	3	6	0.38

*Patients were numbered in the order they were studied

The order in the table is related to the degree that the results agree with the hypothesis that patients could distinguish active injections from placebo injections. The results listed below those of Patient 9 do not support this hypothesis, placebo injections were identified as active at a higher rate than were true active injections. The letters a and b denote the first and second testing sessions, respectively, in Patients 2 and 14. true active injections. ID denotes intradermal, and SC subcutaneous.

The value is the P value associated with the test of whether the common odds ratio (the odds ratio for all patients) is equal to 1.0. The common odds ratio was equal to 1.13 (computed according to the Mantel-Haenszel test).

Notes on P-Values from Fisher's Exact Test in above article

Patient number 3:

	Response		Total
	+	-	
Active Injection	2	1	3
Placebo Injection	1	8	9
	3	9	

All possible tables with a total of 3 +ve responses

	0 3 3 6	1 2 2 7	2 1 1 8	3 0 0 9
Prob	$\frac{9\times8\times7}{12\times11\times10}$ $= 0.382$	$0.382 \times \frac{3\times3}{1\times7}$ $= 0.491$	$0.491 \times \frac{2\times2}{2\times8}$ $= 0.123$	$0.123 \times \frac{1\times1}{3\times9}$ $= 0.005$
(pt #)	(2b, 7, 10, 11)	(14b, 4, 5)	(3)	
P-Value*	1.0	0.618	0.128	0.005

Patient number 1:

	Response		Total
	+	-	
Active Injection	2	1	3
Placebo Injection	2	7	9
	4	8	

All possible tables with a total of 4 +ve responses

	0 3 4 5	1 2 3 6	2 1 2 7	3 0 1 8
Prob	$\frac{8\times7\times6}{12\times11\times10}$ $= 0.255$	$0.255 \times \frac{3\times4}{1\times6}$ $= 0.510$	$0.510 \times \frac{2\times3}{2\times7}$ $= 0.218$	$0.218 \times \frac{1\times2}{3\times8}$ $= 0.018$
(pt #)		(15)	(1, 14a)	
P-Value	1.0	0.745	0.236	0.018

*1-sided, guided by H_{alt}:

π of +ve responses with Active > π of +ve responses with Placebo.

Patient number 18:

	Response		Total
	+	-	
Active Injection	2	1	3
Placebo Injection	4	5	9
	6	6	

All possible tables with a total of 6 +ve responses

	0 3 6 3	1 2 5 4	2 1 4 5	3 0 3 6
Prob	$\frac{6\times5\times4}{12\times11\times10}$ $= 0.091$	$0.091 \times \frac{3\times6}{1\times4}$ $= 0.409$	$0.409 \times \frac{2\times5}{2\times5}$ $= 0.409$	$0.409 \times \frac{1\times4}{3\times6}$ $= 0.091$
(pt #)		(17)	(18)	
P-Value	1.0	0.909	0.500	0.091
(1-sided, as above)				

In the Table, the P-values for patients below patient 9 are calculated as 1-sided, but guided by the opposite H_{alt} from that used for the patients in the upper half of the table, i.e. by

H_{alt}:

π of +ve responses with Active $<$ π of +ve responses with Placebo.

It appears that the authors decided the "sided-ness" of the H_{alt} after observing the data!!!

And they used different H_{alt} for different patients!!!

Message: Tail areas for this test are tricky: it is best to lay out all the tables, so that one is clear which tables are being included in which tail!

6.2 Fisher's Exact Test and Rhinoceroses

Note: The Namibian government expelled the authors from Namibia following the publication of the following article; the reason given was that their "data and conclusions were premature."

Since 1900 the world's population has increased from about 1.6 to over 5 billion) the U.S. population has kept pace, growing from nearly 75 to 260 million. While the expansion of humans and environmental alterations go hand in hand, it remains uncertain whether conservation programs will slow our biotic losses. Current strategies focus on solutions to problems associated with diminishing and less continuous habitats, but in the past, when habitat loss was not the issue, active intervention prevented extirpation. Here we briefly summarize intervention measures and focus on tactics for species with economically valuable body parts, particularly on the merits and pitfalls of biological strategies tried for Africa's most endangered pachyderms, rhinoceroses.

[...]

Given the inadequacies of protective. legislation and enforcement, Namibia. Zimbabwe, and Swaziland are using a controversial preemptive measure, dehorning (Fig. D) with the hope that complete devaluation will buy time for implementing other protective measures (7) In Namibia and Zimbabwe, two species, black and white rhinos (Ceratotherium simum), are dehorned, a tactic resulting in sociological and biological uncertainty: Is poaching deterred? Can hornless mothers defend calves from dangerous predators?

On the basis of our work in Namibia during the last 3 years (8) and comparative information from Zimbabwe, some data are available. Horns regenerate rapidly, about 8.7 cm per animal per year, so that 1 year after dehorning the regrown mass exceeds 0.5 kg. Because poachers apparently do not prefer animals with more massive horns (8), frequent and costly horn removal may be required (9). In Zimbabwe, a population of 100 white rhinos, with at least 80 dehorned, was reduced to less than 5 animals in 18 months (10). These discouraging results suggest that intervention by itself is unlikely to eliminate the incentive for poaching. Nevertheless, some benefits accrue when governments, rather than poachers, practice horn harvesting, since less horn enters the black market Whether horn stockpiles may be used to enhance conservation remains controversial, but mortality risks associated with anesthesia during dehorning are low (5).

Biologically, there have also been problems. Despite media attention and a bevy of allegations about the soundness of dehorning (11), serious attempts to determine whether dehorning is harmful have been remiss. A lack of negative effects has been suggested because (i) horned and dehorned individuals have interacted without subsequent injury; (ii) dehorned animals have thwarted the advance of dangerous predators; (iii) feeding is normal; and (iv) dehorned mothers have given birth (12) However, most claims are anecdotal and mean little without attendant data on demographic effects. For instance, while some dehorned females give birth, it may be that these females were pregnant when first immobilized. Perhaps others have not conceived or have lost calves after birth. Without knowing more about the frequency of mortality, it seems premature to argue that dehorning is effective. We gathered data on more than 40 known horned and hornless black rhinos in the presence and absence of dangerous carnivores in a 7,000 km^2 area of the northern Namib Desert and on 60 horned animals in the 22,000 km^2 Etosha National Park. On the basis of over 200 witnessed interactions between horned rhinos and spotted hyenas (Crocura crocura) and lions (Panthera leo) we saw no cases of predation, although mothers charged predators in about 45% of the cases. Serious interspecific aggression is not uncommon elsewhere in Africa, and calves missing ears and tails have been observed from South Africa, Kenya, Tanzania, and Namibia (13).

To evaluate the vulnerability of dehorned rhinos to potential predators, we developed an experimental design using three regions:

- Area A had horned animals with spotted hyenas and occasional lions
- Area B had dehorned animals lacking dangerous predators,
- Area C consisted of dehorned animals that were sympatric with hyenas only.

Populations were discrete and inhabited similar xeric landscapes that averaged less than 125 mm of precipitation annually. Area A occurred north of a country long veterinary cordon fence, whereas animals from areas B and C occurred to the south or east, and no individuals moved between regions.

The differences in calf survivorship were remarkable. All three calves in area C died within 1 year of birth, whereas all calves survived for both dehorned females living without dangerous predators (**area B**; $n = 3$) and for horned mothers in **area A** ($n = 4$). Despite admittedly restricted samples, the differences are striking [Fisher's (3 x 2) exact test, $P = 0.017$; area B versus C, $P = 0.05$; area A versus C, $P = 0.0291$ ††. The data offer a first assessment of an empirically derived relation between horns and recruitment.

Our results imply that hyena predation was responsible for calf deaths, but other explanations are possible. If drought affected one area to a larger extent than the others, then calves might be more susceptible to early mortality. This possibility appears unlikely because all of western Namibia has been experiencing drought and, on average, the desert rhinos in one area were in no poorer bodily condition than those in another. Also, the mothers who lost calves were between 15 to 25 years old, suggesting that they were not first time, inexperienced mothers (14). What seems more likely is that the drought induced migration of more l than 85% of the large, herbivore biomass (kudu, springbok, zebra, gemsbok, giraffe, and ostrich) resulted in hyenas preying on an alternative food, rhino neonates, when mothers with regenerating horns could not protect them.

Clearly, unpredictable events, including drought, may not be anticipated on a short-term basis. Similarly, it may not be possible to predict when governments can no longer fund antipoaching measures, an event that may have led to the collapse of Zimbabwe's dehorned white rhinos. Nevertheless, any effective conservation actions must account for uncertainty. In the case of dehorning, additional precautions must be taken. [...]

	A	B	C
survived	4	3	0
died	0	0	3
	4	3	3

††

B vs C

	B	C	B	C	B	C	B	C	*total**
survived	3	0	2	1	1	2	0	3	*3*
died	0	3	1	2	2	1	3	3	*3*
	3	3	3	3	3	3	3	3	

A vs C

	A	C	A	C	A	C	A	C	*total**
survived	4	0	3	1	2	2	1	3	*4*
died	0	3	1	2	2	1	3	3	*3*
	4	3	4	3	4	3	4	3	
Prob	$\frac{1}{35}$		$\frac{12}{35}$		$\frac{18}{35}$		$\frac{4}{35}$		

"Data and conclusions were premature."

Agree?

0 Exercises

0.1 Questions re van Belle et al.

1. From Problems 6.1 to 6.30, find 1 that involves (i) a test of a single proportion (ii) a CI for a single proportion (iii) a test of the equality of two proportions, a CI for (iv) a RD (v) a Risk Ratio and (vi) an Odds Ratio.

2. Problem 6.1(b) raises a subtle and important point, but does not say why the requested probability calculation helps in evaluating the complaint. Explain what is the appropriate probability to calculate in order to judge if the clinic's complaint is valid.

0.2 Sample size to assess risk of abortion after chorionic villus sampling

The following letter is by Holzgreve et al. to The Lancet (p. 223, January 26, 1985). They use symbols P_1 and P_2 in the same way we use the Greek (for "population" or "parameter") symbols "π_1" and "π_1". Also, they use the term 'rate' where we might use 'proportion' and they use it as a percentage i.e., their $P_2 = 4.4\%$ is our $P_2 = 0.044$. Note also that in the 1st sentence at the top of the page, they reverse the 2 subscripts. The correct subscripts are those used later on i.e., 1= ultrasonically normal pregnancies and 2=chorionic villous biopsy (cvb). Below, lower case p is used for a proportion observed in a sample, i.e., the 'statistic.'

> We agree with Dr Wilson and colleagues (Oct 20, p 920) that background rates of spontaneous abortion in ultrasonically normal pregnancies are an important requirement for evaluating the of chorionic villus sampling in the first trimester. For an unbiased assessment of the risk of spontaneous abortion with this new method of prenatal diagnosis, however, the rate of fetal losses should be compared with matched pregnancies without invasive procedures in a prospective, randomised trial.
>
> To be able to state with confidence that the fetal loss rate in a group of patients (P) after chorionic villus biopsy differs from that in a control group of ultrasonically normal pregnancies (P_2) we have calculated the required sample size for the two populations, based on a probability of a type I error (α) of 1% and of a type II error (b) of 10%. The most recent international survey^{ref} revealed a spontaneous abortion rate of about 4.4% after chorionic villus sampling, and this was the figure we used for the rate in P_2 when calculating sample sizes by the Fleiss formula, the arc-sine formula, and the formula of Casagrande, Pike, and Smith[9] for different assumed risk figures for P_1:
>
> | P_1 | P_2 | Fleiss | Arcsine | Casagrande |
> |---|---|---|---|---|
> | 4.0 | 4.4 | 654,33 | 65,965 | 75,831 |
> | 3.0 | 4.4 | 4,691 | 4872 | 5,690 |
> | 4.1 | 4.4 | 117,677 | 118,376 | 135 884 |
> | 2.5 | 4.4 | 2,357 | 2,504 | 2,950 |
>
> These calculations show that if chorionic villus biopsy increases the spontaneous abortion rate by 0.4%, which would be equivalent to the risk for second-trimester amniocentesis, about 69,000 pregnancies would be required in each group. The background rate of spontaneous abortion in the first trimester strongly influences the required numbers of patients – e.g. a drop to about 2,600 patients in the two groups if the difference in abortion rates is about 2%. Even though the numbers required to achieve statistical significance are large, a study with matched controls allows a more meaningful statement about the added risk of spontaneous abortion after chorionic villus biopsy than the mere comparison with fetal loss rates in ultrasonically normal pregnancies now available. Only a well-designed, statistically sound, multicentre (preferably international) study can answer the very important questions about the safety of chorionic villus sampling.
> W. Holzgreve. Women's Clinic, Dept Biomed. Statistics & Inst of Human Genetics, Westphalian Wilhelma Uni., Munster, Germany.

Questions on above letter:

1. Why do the authors propose a 2-sample study? i.e., why not compare the proportion, p_2, of fetal losses observed following cvb in a single sample of n_2 pregnancies, against a "background rate" of $P_1 = 3.7$? Assume that this 3.7 is the figure they would have obtained by combining data from the literature, consulting experts, etc.

2. What form would the data-analysis of such a "one-arm" study take? Use a numerical example with $n_2 = 500$ to illustrate.

[9] Fleiss JL Statistical Methods for Rates an Proportions, 1973.

3. Calculate the required sample size for such a "one-arm" study, using the same α and β as they did (cf. Notes, or vanBelle, or Colton p161).

4. What form will the data-analysis of the "two-arm" study proposed by the authors take? Use a numerical example with $n_1 = n_2 = 500$ to illustrate.

5. Calculate the required sizes n_1 and n_2 for this study that the authors propose (cf Notes, or vanBelle, or Colton p168). Use $P_1 = 3.0$ (3rd row of table) and the same α and β. Note that the sample sizes may differ somewhat depending on the method of analysis, and on the formula used.

6. Assume that a study of this size has been done and that the observed losses were $p_1 = 3.8\%$ and $p_2 = 4.3\%$. What do you conclude? Use language that is understandable to those who will need to understand it.

7. In the now-completed Canadian collaborative trial of cvb, the investigators plan to analyze the difference in all fetal losses and so are using $P_1 = 6.6\%$ and $P_2 = 9.5\%$ in their calculations. They used $\alpha = 0.05$ and $\beta = 0.20$. What impact do these design differences have on sample size? Full calculations are not required.

0.3 Analysis of un-matched case-control studies

A 1982 Swedish study (Arch. Env. Health, March/April 1982, p.81-) examined the association between exposure of female physiotherapists to non-ionizing radiations (shortwaves, microwaves,.) and the risk in subsequently delivered infants of a serious malformation or perinatal death. Two *series* of working physiotherapists were compared: ($Y = 1$) the 33 mothers of the (33) infants who were born with serious malformations or who died perinatally; and ($Y = 0$) the (66) mothers of 66 randomly chosen "normal" infants. The resulting data, presented in a somewhat simplified form for this exercise, are:

	Y			Y	
Shortwave Use	1	0	*Microwave Use**	1	0
never/seldom	24	54	never	29	63
often/daily	9	9	sometimes	4	0

* data missing on 3 mothers for whom $Y = 0$.

1. What comparative parameter can one estimate from these data? Think of the $Y = 1$ data as coming from the *numerator series*; think of the $Y = 0$ data as coming from the *denominator series* that supplies *estimates* of the fractions of the source population that are in the higher- and lower-use categories.

2. For each the two exposures, what is the point-estimate of this parameter?

3. Derive a 95% CI for the parameter, by "Woolf's" method for shortwave, the exact conditional method (Fisher) for microwave (see spreadsheet).

4. Perform a 2-sided test of significance to test the null hypothesis of no association between each of the two exposures and the subsequent delivery outcome.

0.4 A simple way to improve the chances for acceptance of your scientific paper

To the Editor: During the past few years we have witnessed a revolution in the way manuscripts, abstracts, and grant proposals are being typed. With improved typewriters and computer programs it is possible to produce manuscripts of typeset quality. It is generally assumed that data should be judged by its scientific quality and that this judgment should not be influenced by typing style.

I challenged this premise by analyzing the rate of acceptance of abstracts by a large national meeting. All abstracts submitted to the 1986 annual meeting of the American Pediatric Society and the Society of Pediatric Research (APS/SPR) appeared in Volume 20, No. 4 (Part 2) (April 1986) of Pediatric Research. Contrary to the practice of many other meetings, this volume also includes all the abstracts that were not accepted for presentation, and accepted papers are identified by symbols.

Abstracts were defined as "regularly typed" or "typeset printed." Each abstract was categorized as accepted if chosen for presentation or rejected.

A total of 1965 abstracts were evaluated. Excluded were 47 abstracts assigned for joint internal medicine-pediatric presentation, because the majority of them were submitted to the meeting of the American Federation for Clinical Research, and there was no indication of their rejection rate; only those that had been accepted appeared in the APS/SPR book of abstracts.

Of the 1918 evaluable abstracts, 1706 were regularly typed and 212 were "typeset." The acceptance rate was significantly higher for the "typeset" abstracts: 107 of 212 (51.4 percent) vs. 747 of 1706 (44 percent) (P<0.05).

Eighty-eight investigators submitted five or more abstracts to the meeting. Here, too, there was a higher rate of acceptance for the "typeset" abstracts (62 of 107:57.9 percent) as compared with the regularly typed abstracts (184 of 451:40.8 percent) (P = 0.002).

One may argue that investigators who can afford the new equipment for printing abstracts have more money and can afford better research, and therefore that their abstracts are accepted at higher rates. To explore this possibility. I analyzed data on the 15 investigators who submitted five or more abstracts each and who used both typing methods. In this subgroup, 19 or 55 regularly typed abstracts were accepted (34.5 percent), whereas 31 of 53 of the "typeset" abstracts were accepted (58.5 percent) (P = 0.015).

These results demonstrate that the new "typeset" appearance of data increases the chance of acceptance. It may mean that "typeset" printing may cause the data to look more impressive. Alternatively, it may mean that the new printing makes it easier for reviewers to read the data and to appreciate its meaning.

Most important, it means that this technological innovation reduces the chance of success of those not currently using it.

Questions

i Display the data in the 5th paragraph in a 2 × 2 table.

ii What test (and what hypotheses) are appropriate to compare the "107 of 212 vs. 747/1706"? Notice that P < 0.05. (Paragraph 5)

iii-v See after rebuttal below

...ACCEPTANCE OF ABSTRACTS - A REBUTTAL

To the Editor: Dr. Koren claims that the use of a new "typeset" method for preparing an abstract may improve the chances for its acceptance at a national meeting, specifically, at the 1986 annual meeting of the American Pediatric Society and the Society for Pediatric Research (Nov 13 issue). This assertion, if correct, should raise alarm among investigators submitting their work for peer review and seeking a fair and objective critique. Although Dr. Koren lists several possibilities to explain why typeset printing may enhance the rate of acceptance of an abstract, including the possibility that printing may make the data appear more impressive or may make the reading of an abstract easier, his data can be interpreted differently.

Koren reports that 107 of 212 "typeset-printed" abstracts were accepted, as compared with 747 of 1706 "regularly typed" abstracts, the relative acceptance rates being 51.4 versus 44 percent (P < 0.05). Because of the disparity in the sizes of the groups, we are uncertain what form of statistical analysis he employed. If one uses the technique of hypothesis testing of the differences between two proportions, the proportions 107 of 212 versus 747 of 1706 have a z value of 1849 with P<0.06. Thus, when an appropriate statistical method is used, a significant difference between the two proportions is not found at the 0.05 level.

These data can be examined in another way: 107 of a total of 854 accepted abstracts (12.5 percent) were "typeset," whereas 212 of 1918 abstracts submitted (11.1 percent) were "typeset." The difference between these proportions is obviously not significant. The difference in the sizes of the groups also makes it difficult to compare them. Furthermore, some abstracts were judged independently of this process in order to be placed in a poster symposium dealing with a specific topic (ie, "AIDS in Pediatric Patients"). Of the 30 abstracts chosen for these poster symposia, 15 were (we think) 'typeset printed" and may appropriately be removed from the pool of accepted "typeset" abstracts.

Most important, a reviewer is judging the merit of a given abstract from a photocopy of the actual abstract, not its appearance in the April 1986 issue of Pediatric Research. "Typeset" abstracts that appear impressive in the abstract book do not necessarily stand out on the actual abstract form.

For these reasons, Koren's conclusion that a "technological innovation reduces the chance of success of those not currently using it" may not be entirely correct. Other reasons can be advanced to account for the apparent success of "typeset" abstracts.

Finally, in order to ensure that objective criteria are being used, all reviewers of abstracts for the 1987 meeting will receive a copy of Dr. Koren's letter so that they are aware of this potential problem.

R W. Chesney, M.D. Society for Pediatric Research, University of California.

Questions (continued)

iii The rebuttal claims that the difference between these two proportions is associated with a P-value of p=0.06 (2nd paragraph). Why do you think the "rebutting" authors arrive at a different p-value? [The typographical error (1819 for 1.849) is not the problem] (Paragraph 2, last two sentences)

iv In the 3rd paragraph of the reply, the authors look at the data regarding the same 1918 abstracts "in another way" i.e. in a type of case-control analysis. This is a legitimate way to look at the data; however, the "obviously nonsignificant" pvalue associated with the comparison of 107/854 vs 212/1918 is not legitimate. Why? (Paragraph 3, fourth line)

v The rebuttal mentions "the disparity in the sizes of the groups" in two places. The second time, in paragraph 3, it is stated that "the difference in the sizes of the two groups also makes it difficult to compare them."

(Third paragraph, fifth line) Do you agree? Why / Why not?

0.5 Test of a proposed mosquito repellent

An entomologist carried out the following experiment as a test of a proposed mosquito repellent. Thirty-five volunteers had one forearm treated with a small amount of repellent and the other with a control solution. The subjects did not know on which forearm the repellent had been used. At dusk the volunteers exposed themselves to mosquitoes and reported which forearm was bitten first. In 10/35, the arm with the repellent was bitten first.

1. Make a statistical report on the findings.
2. How would you analyze the results if:
   - some arms were not bitten at all?
   - some people were not bitten at all?

0.6 Perioperative Normothermia

Refer to the report of this study (scanned version of text as images [.gif files] under Resources for Chapter 5; full version, using optical character recognition, and reformatting in a word processor, as a pdf file in Resources for Chapter 7)

1. Using the same 'inputs' as the authors did (2nd paragraph of Methods), calculate the sample size requirements.

   Some formulae do not use different null and non-null variances, instead, for simplicity, they use the same null and non-null variance –calculated at the average of the null and non-null p's; and some authors use a formula based not on the difference of the proportions, but of the arcsine transformations of these proportions. Thus, you should not be surprised if you don't get exactly the same numbers.

   See also footnote concerning the choice of 'delta.' The difference that would be important (the clinically important difference) is a matter of judgment; it should not be left to be 'dictated' empirically by Nature (the authors used as their 'delta' the empirical difference 9/38 - 4/42 = 14.2% found in their pilot study!). Imagine what the authors' 'delta' could gave been if they had done a pilot study of say 2 patients vs. 3 patients, or just 1 vs. 2! And , even with increasing sample sizes, Nature is just going to show you more precise estimates of what the difference is, not of "the difference that would make a difference." After all, Nature doesn't know how much these normothermia blankets cost, or how acceptable and practical they would be! Indeed, it is ironic that the observed difference in the study proper is only 19% - 6% = 13%; it is "statistically significant" but less than the 'clinically important delta' used by the authors in their sample size formula.

2. State the null and alternative hypotheses, and re-calculate the P-value in the first row of Table 2.
3. Calculate a 95%CI for the difference in infection rates.
4. You can convert the point estimate of the difference into the "number required to treat." The formula for this is 1/(Infection Rate if Do Not Treat - Infection Rate if Treat). The logic is that if 19/100 would develop an infection without the intervention, and 6/100 despite it, then intervening on 100 would prevent 19 - 6 = 13 infections, i.e.. one would need to intervene on approximately 8 (i.e. 100/13) to prevent 1 infection. Convert the upper and lower 95% limits for the difference (from part iii) into the corresponding limits on the number required to treat.

0.7 Women are Safer Pilots: Study

London- Initial results of a study by Britain's Civil Aviation Authority shows that women behind the controls of a plane might be safer than men. The study shows that male pilots in general aviation are more likely to have accidents than female pilots. Only 6 per cent of Britain's general aviation pilots are women. According to the aviation magazine Flight International, there have been 138 fatal accidents in general aviation in the last 10 years, and only two involved women - less than 1.5 per cent of the total.

[Montreal Gazette, WomanNews, page F1]

1. What is the comparative parameter at issue here?
2. Comment on the epidemiologic soundness of the comparison reported.
3. Assuming that the comparison reported is a sound one, or that it can be made so using additional information, translate the data into point and interval estimates of the comparative parameter. Also, carry out a test of the null value of the comparative parameter.

0.8 Equivalent Forms of the X^2 statistic from a 2×2 table

Consider a 2×2 table with frequencies $y_1 = a,\ b,\ y_2 = c,\ d$, row totals $n_1 = r_1,\ y_2 = r_2$, column totals $c_1,\ c_2$, overall total n, observed proportions $p_1 = y_1/n_1$ and $p_2 = y_2/n_2$, overall proportion $p = (y_1 + y_2)/n$, and $Var[a|H_0]$ based on the '2-independent-binomials' model. Show that

$$
\begin{aligned}
X^2 &= \sum \frac{(Observed\ Frequency - Expected\ Frequency)^2}{Expected\ Frequency} \\
&= n \times \frac{(a \times d\ -\ b \times c)^2}{r_1 \times r_2 \times c_1 \times c_2} \\
&= \frac{\{p_1 - p_2\}^2}{p(1-p) \times (1/n_1 + 1/n_2)} \\
&= \frac{\{a - E\widehat{[a \mid H_0]}\}^2}{Var[a \mid H_0]}.[seenote]
\end{aligned}
$$

0.9 Bone mineral density and body composition in boys with distal forearm fractures

J Pediatr 2001 Oct;139(4):509-15.

Goulding et al (New Zealand)

Abstract

Objective: To determine whether boys with distal forearm fractures differ from fracture-free control subjects in bone mineral density (BMD) or body composition. **Study design**: A case-control study of 100 patients with fractures (aged 3 to 19 years) and l00 age-matched fracture-free control subjects was conducted. Weight, height, and body mass index were measured anthropometrically. BMD values and body composition were determined by dual-energy x-ray absorptiometry. **Results**: More patients than control subjects (36 vs l4) were overweight (body mass index > 85th percentile for age, P < .001). Patients had lower areal (aBMD) and volumetric (BMAD) bone mineral density values and lower bone mineral content but more fat and less lean tissue than fracture-free control subjects. The ratios (95% CIs) for all case patients/control subjects in age and weight-adjusted data were ultradistal radius aBMD 0.94 (0.91-0.97); 33% radius aBMD 0.96 (0.93-0.98) and BMAD 0.95 (0.91-0.99); spinal L2-4 BMD 0.92 (0.89-0.95) and BMAD 0.92 (0.89-0.94); femoral neck aBMD 0.95 (0.92-0.98) and BMAD 0.95 (0.91-0.98); total body aBMD 0.97 (0.96-0.99), fat mass 1.14 (1.04-1.24), lean mass 0.96 (0.93-0.99), and total body bone mineral content 0.94 (0.91-0.97). **Conclusions**: Our results support the view that low BMC, aBMD, and BMAD values and high adiposity are associated with increased risk of distal forearm fracture in boys. This is a concern, given the increasing levels of obesity in children today. (J Pediatr 2001;139:509-15)

		Fracture?	
		Yes	No
Overweight?	Yes:	36	14
	No:	64	86
	Total	100	100

1. Rewrite the sentence "A case-control study of 100 patients with fractures (aged 3 to 19 years) and l00 age-matched fracture-free control subjects was conducted" using terminology that better reflects the purpose of the 100 fracture-free subjects.
2. All of the fractures occurred over a 1-year period, ten of them in persons aged 11. Suppose one could choose a random sample, of size ten, from **all** 600 11-year old boys living in the city of Dunedin, what is the probability that this denominator series would have an overlap of 0, 1, 2, .. with the case series of ten? [10]

   What if age-matching were to the nearest month of age, and that there were two cases in boys aged 11 years and 3 months, so we took a sample of two from all of the 600?
3. Estimate the ratio of the fracture rate in the overweight to the fracture rate in the not-overweight, and use Woolf's method to calculate a 95% CI for it (ignore the age-matching).
4. We can repeat the point- and interval estimation using logistic regression: e.g., in `R`,

   ```
   y=c(rep(1,100),rep(0,100)); over=c(rep(1,36),rep(0,64),rep(1,14),rep(0,86))
   summary(glm(y~over,family=binomial))
   ```

   yielding...

   ```
               Estimate Std. Error z value Pr(>|z|)
   (Intercept)  -0.2955     0.1651  -1.790 0.073490 .
   over          1.2399     0.3556   3.487 0.000489
   ```

   Verify that 1.2399 represents the $\log or$ and 0.3556 its SE.

0.10 Theoretical basis for "odds ratio" as estimator of Rate Ratio, together with statistical model for the estimator

The old-fashioned and very loose justification for using the empirical odds ratio, *or*, as an estimator of the theoretical rate ratio goes back to Cornfield in the 1950s. Unfortunately it still is the one given in many 'modern' texts, despite the much more general justification provided by Miettinen in 1976.

The old justification rested on algebraic arguments using *persons*, not *population time.* The outcome *proportions* involved refer to *cumulative* incidence.

The truly modern way is to think of the cases as arising in population-time, and to think of the population time involved as an infinite number of person-moments - think of a person-moment as a person at a particular moment. Say that a proportion π_E of these are

[10] In fact, the "age-matched denominator series" was assembled as follows: All patients with fractures were asked to supply the names of 3 friends of their own age: the first friend who had never fractured a bone at any time of his life and who agreed to take part as a fracture-free control subject was then enrolled."

"exposed" person moments, and the remaining proportion π_0 are "non-exposed" person-moments. Suppose further that the (theoretical) event rates in the exposed and unexposed amounts of population-time are

$$\lambda_E = \frac{E[no.events]}{PT_E} \ ; \ \lambda_0 = \frac{E[no.events]}{PT_0},$$

with (theoretical) Rate Ratio $\theta = \lambda_E/\lambda_0$.

Denominator Series

Suppose we take a finite random sample, of size d, of the infinite number of person moments in the base that generated the cases, and classify them into d_E "exposed" person moments and $d_0 = d - d_E$ "non-exposed" person-moments. We will refer to this sample of d as the *denominator* series. What is the statistical model for $d_E \mid d$? Clearly, it is

$$d_E \sim Binomial(d, \pi_E).$$

Numerator (Case) Series

Denote by c the observed number of events; we classify them into c_E events in "exposed" population-time and $c_0 = c - c_E$ in the "non-exposed" population-time. We will refer to this sample of c as the *case* series.

What is the statistical model for $c_E \mid c$? We can think of c_E as the realization of a Poisson r.v. with mean (expectation) $\mu_E = (PT_E \times \pi_E) \times \lambda_E$. Likewise, think of for c_0 as the realization of a Poisson r.v. with mean (expectation) $\mu_0 = (PT_0 \times \pi_0) \times \lambda_0$.

Now, it is a statistical theorem (Casella and Berger, p194, exercise 4.15) that

$$c_E \mid c \sim Binomial(c, \mu_E/[\mu_E + \mu_0]).$$

Thus we can identify the distribution of the 4 random variables involved in the OR estimator

$$\hat{OR} = or = c_E/d_E \ \div \ c_0/d_0 \ = \ c_E/c_0 \ \div \ d_E/d_0 = (c_E \times d_0) \ \div \ (c_0 \times d_E).$$

The $c_E : c_0$ split is governed by one binomial, involving θ and other parameters, while the $d_E : d_0$ split is governed by a separate binomial, involving the same other parameters, but not involving θ.

If one replaces μ_E and μ_0 by their constituents, one can show that the odds that an unexposed person-moment in the series of $c + d$ represents a "case" is $c : d$, whereas the corresponding odds for an exposed person moment is $(\theta \times c) \ : \ d$.

In other words, in the dataset of $c + d$,

$$logit[Prob[case|0] = \log(c/d) = \beta_0 \ ; \ logit[Prob[case|E] = \log(c/d) + \log\theta = \beta_0 + \beta_E E,$$

where E is an indicator variable.

So, one can estimate $\log\theta = \log OR$ by a logistic regression of the $c + d$ Y's i.e. $Y = 1$ if in case series; $= 0$ if in denominator series, on the corresponding set of $c + d$ indicators of exposure (1 if exposed, 0 if not).

==================================

Note, added 2009.11.11

==================================

-8- Equivalent Forms of the X^2 statistic from a 2×2 table

Consider a 2×2 table with frequencies $y_1 = a, \ b, \ y_2 = c, \ d$, row totals $n_1 = r_1, \ y_2 = r_2$, column totals $c_1, \ c_2$, overall total n, observed proportions $p_1 = y_1/n_1$ and $p_2 = y_2/n_2$, overall proportion $p = (y_1 + y_2)/n$, and $Var[a \mid H_0]$ based on the '2-independent-binomials' model. Show that

$$\begin{aligned} X^2 &= \sum \frac{(Observed\ Frequency - Expected\ Frequency)^2}{Expected\ Frequency} \\ &= n \times \frac{(a \times d \ - \ b \times c)^2}{r_1 \times r_2 \times c_1 \times c_2} \\ &= \frac{\{p_1 - p_2\}^2}{p(1-p) \times (1/n_1 + 1/n_2)} \\ &= \frac{\{a - E[a \mid H_0]\}^2}{Var[a \mid H_0]}. \end{aligned}$$

I'm embarrassed I didn't see the 4th form right away. The key is that under a 2-independent-binomials model, the numerator is $a - E\widehat{[a \mid H_0}$, i.e., the second component of it is also a random variable. Indeed with $y_i \sim Binomial(r_i, \pi)$,

$$E\widehat{[a \mid H_0}] = (y_1 + y_2)/(r_1 + r_2)$$

so (with y_1 as a duplicate name for a, and a bit of algebra),

$$y_1 - E\widehat{[a \mid H_0}] = y_1(1 - r_1/n_1) - y_2/(r_1/n)$$

so its variance is

$$Var\{y_1 - E\widehat{[a \mid H_0}]\} = r_1\pi(1-\pi)(1 - r_1/n_1)^2 + r_2\pi(1-\pi)/(r_1/n)^2$$

Now, substitute c_1/n for π, and c_2/n for $(1 - \pi)$ and you get, again after some algebra,

$$\widehat{Var} = \hat{\pi}(1 - \hat{\pi})r_1 r_2/n = (c_1/n)(c_2/n)r_1 r_2/n = c_1 c_2 r_1 r_2/n^3.$$

The numerator of the statistic can be simplified to $(ad - bc)^2/n^2$. Putting the numerator over the denominator leads to the result.

And the **correct form** should be:

$$X^2 = \frac{\{a - E[a \mid H_0]\}^2}{Var\Big[a - E\widehat{[a \mid H_0]}\Big]}.$$