RCT of Routine Cervical Examinations in Pregnancy

1i π_E =0.04, π_C= 0.05, Δ =0.01. Several options for n/group, ranging from the most exact (Colton p 168)

$$\frac{\{z_{\alpha/2}\ SE[p_E - p_C \mid H_0] + z_\beta\ SE[p_E - p_C \mid H_{alt}]\}^2}{\Delta^2} \qquad [1]$$

$$= \frac{\{z_{\alpha/2}\sqrt{\pi_C[1-\pi_C] + \pi_C[1-\pi_C]} + z_\beta\sqrt{\pi_E[1-\pi_E] + \pi_C[1-\pi_C]}\ \}^2}{\Delta^2}$$

to the rougher

$$\frac{\{z_{\alpha/2}\sqrt{\bar\pi[1-\bar\pi] + \bar\pi[1-\bar\pi]} + z_\beta\sqrt{\bar\pi[1-\bar\pi] + \bar\pi[1-\bar\pi]}\ \}^2}{\Delta^2}$$

$= 2\{z_{\alpha/2} + z_\beta\}^2 \frac{\bar\pi[1-\bar\pi]}{\Delta^2}$ [2]. Here $\bar\pi$ = 0.045.

If α= **0.05** & ß= **0.2**, $z_{\alpha/2}$ = 1.96 & z_β =0.84, so $2\{z_{\alpha/2} + z_\beta\}^2$ = 15.68,

and if we round 15.68 to 16, we get the rougher $16\ \frac{\bar\pi[1-\bar\pi]}{\Delta^2}$

where $\bar\pi[1-\bar\pi]$ is the variance of individual observations on a 0/1 scale (see article by _______ reproduced in Material on Chapter 8).

If π_E =**0.035**, π_C= 0.05, Δ =0.015, then $\bar\pi$ = 0.0425.

ii If α= **0.05** & ß= **0.1**, $z_{\alpha/2}$ = 1.96 & z_β =1.28, so $2\{z_{\alpha/2} + z_\beta\}^2$ = 20.9952,

we can round it to 21, for a fairly exact $21\ \frac{\bar\pi[1-\bar\pi]}{\Delta^2}$

iii If only change ß, and use [2] above, then change $\{z_{\alpha/2} + z_\beta\}^2$ to 20.9952 from 15.68 or an increase of 34%.

2i Data p_E = 169/2521 and p_C =161/2520, p = 330/5041. Or if you are an $\begin{matrix} a & b \\ c & d \end{matrix}$ person who uses columns for groups being contrasted and rows for outcomes, then

```
                G R O U P
                E        C      All
OUTCOME
Bad     a =  169  b=  161 | r1=  330
Good    c = 2352  d= 2359 | r2= 4711
All     c1= 2521 c2= 2520 |  N= 5041
```

Several options for uncorrected X^2,

definitional formula $\sum\frac{\{o-e\}^2}{e}$ or $\frac{N\{ad-bc\}^2}{r_1\ r_2\ c_1\ c_2}$ or $\frac{\{a-E[a \mid H_0]\}^2}{var[a \mid H_0]}$

with $var[\ a \mid H_0\ ]=r_1\ r_2\ c_1\ c_2 / N^3$

Using the 2nd form gives X^2 = **0.20414221..** or $\sqrt{X^2}$ = **0.4518..**

For z test for 2 proportions,

$$Z = \frac{p_E - p_C}{\sqrt{p\{1-p\}\ \{\frac{1}{2521} + \frac{1}{2520}\}}} = \mathbf{0.4518..}$$

For z **test** for 2 π's, must use **common p** [H_0 stipulates $\pi_E = \pi_C$]; for a **CI** for Δ of 2 π's, the SE of the difference uses separate p's, since there is no H_0 to say anything about what the two underlying π's should be. X^2 and Z^2 will not match exactly if use separate p's

ii With continuity correction, X^2 = **0.15592817.., Z = 0.3948**.. = $\sqrt{X^2}$ [the numerator of Z changes from 0.003148.. to 0.002751.. but nothing else is affected]

3i At issue is the sampling variability of the **statistic** $\bar y_E - \bar y_C$. With the large n's involved this should be Gaussian around a mean of zero, no matter what the SD_{ind} of **individual** y's (CLT). The SD or SE of the Gaussian distribution of the statistic will be the square root of the sum of the squares of the SEM's of the two component statistics ie. SE = $\sqrt{\frac{SD_{ind}^2}{2521} + \frac{SD_{ind}^2}{2520}}$

$= SD_{ind}\sqrt{\frac{1}{2521} + \frac{1}{2520}} = 0.028SD_{ind}$. Thus there is a 95% probability that possible $\bar y_E - \bar y_C$'s will be within the range –1.95SE to +1.96SE or between **$-0.055SD_{ind}$ and $+0.055SD_{nd}$.**

The variation of individual ages is probably not Gaussian and so SD_{ind}. cannot be used with the Z table to describe the limits of individual variation;

however it can be used for the limits of the statistic $\bar{y}_E - \bar{y}_C$. If we thought that maternal ages varied from 15 to 45, the SD_{ind} cannot be more than half the range i.e. (45–15)/2 =15. Since the ages have a strong central tendency, SD_{ind} is probably closer to **5**. Substituting this in the limits above gives **±0.275 years**. In other words there is a good chance that the difference between the average age of 2 randomly chosen samples of maternal ages will not be more than 0.3 years.

ii For the difference in the proportion of primipara's, the only difference is that we are dealing with a mean of a 0/1 variable. Since 45% of the y's are 1 and 55% are 0's, the $\mathbf{SD_{ind}}$ or these 0's and 1's is $\sqrt{0.45 \times 0.55} \approx \mathbf{0.5}$. So the difference in the two proportions will again have a Gaussian distribution with mean zero and 95% limits of ±0.055(0.5) i.e. **±0.0275 or ±2.75%.**

A large number of students spoke about CI's. **CI's are for parameters**. Here we are dealing with **distribution of a statistic**. i.e.

$$\bar{y}_E - \bar{y}_C \sim N(\mathbf{0},\ SD_{ind}\sqrt{\frac{1}{2521} + \frac{1}{2520}}\ )$$

4i p=44/5400 = 0.00809 or 8.09 per 1000, based on a sample of n=5440. Large sample 95%CI for π, the frequency of multiple births, based on Binomial variation, given by p±1.96$\sqrt{p[1-p]/5440}$ = 0.00809±0.00238 or 5.71 to 10.47 per 1000. [Some of you got the scales mixed up: better to stay in (0,1) proportion scale until end and then convert to rate per 1000. Also, some of you based CI on n = 1000 rather than 5440].

CI constructed from margin of error, and a strict funcction of n and p; any biases in sampling are not included in the calculation. Some high risk pregnancies were excluded, and this could well mean that multiple births are therefore under-represented (other exclusions too). Also, should be suspicious of 'national' samples based on two to 6 clinics or hospitals per country.

ii Chi-square and z tests of 2 proportions are (equivalent) large sample tests that try to approximate binomial sampling variation by Gaussian variation. For some of the country-specific comparisons, the <u>expected</u> numbers of bad events are below 5, a threshold below which users are often warned about the accuracy of the approximations. Cannot always be sure of direction of inaccuracies: I usually calculate by large sample and 'exact' methods and find close agreement even when E's are well < 5.

A more fundamental issue in this example is whether it is reasonable to count the outcomes in 2 twins as independent observations [the binomial, just like every SE involving √n, is built on independent observations]. The outcomes are likely to be positively correlated and so using n= no. of infants gives a falsely small SE, and a more extreme p-value than we should have. However, the multiple births counted in the n is not large and the 'false inflation' of n is inconsequential.

5i When intervention is left to caregivers, authors must assure us that it was carried out well. If it was only halfheartedly done, that might explain why no effect was seen. It is a little like in a drug trial checking that the patients filled their prescriptions and actually took their medications. It does not make any sense to compare this process variable in the two groups by a formal statistical test. The null hypothesis is not of interest; the issue is whether the intervention was implemented <u>intensively enough</u> that we might expect a result; a small difference in the number of examinations could be statistically significant if the n's were large, but simply telling the reader that the difference was "statistically significant" would be of no use; the reader wants to be sure that the intervention was given in a big enough 'dose' to work.

ii Because the distributions of the variables were skewed, and a median is more descriptive of central tendency, and is robust ('resistant') to extreme observations than a mean.

iii Total numbers of bed days or lengths of stay can be reconstructed/estimated from the mean but not from the median.

iv Even with very skewed distributions of individual observations, with the large sample sizes involved, the sampling variability of the sample mean (a statistic) is quite Gaussian (CLT).

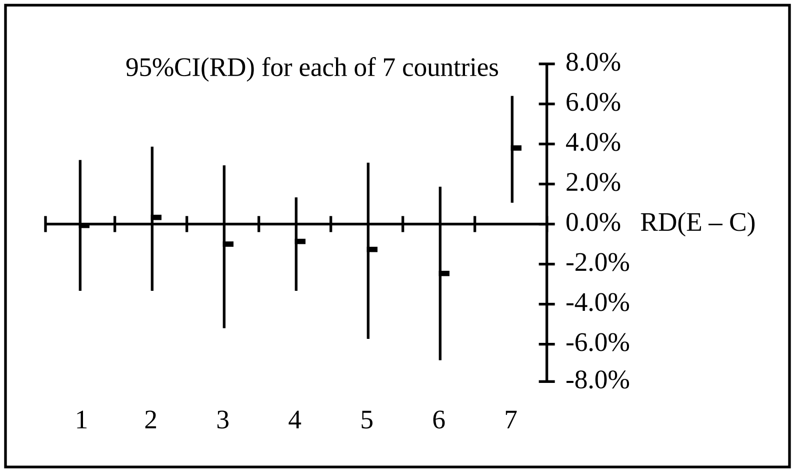
v Both are reasonable in this large-sample example. However, if the natural parameter is the mean, why not test means by parametric tests.

vi Use Gaussian approximation to distribution of Sum of Ranks in smaller sample. The null distribution of this statistic is very Gaussian even for sample sizes as low as 10. The SD of the Sum of Ranks, needed for using the Gaussian table, is simply a function of the two n's involved (thats why these tests are called 'distribution-free') and the formula is given in the material from A&B.

6 A zero value of the RD parameter corresponds to a unity value of the RR parameter (contrary to what many will tell you, 'null' does not necessarily mean zero). Thus if we are consistent in our statistical approach, an RR interval estimate (CI) that does not include 1 should correspond to an RD interval estimate (CI) that does not include 0, and both of these imply a point estimate that if tested against the null would not be statistically significant. Sadly, a consistent approach to RD's and RR's is absent from most textbooks.

7i CI's for RR are calculated by first constructing symmetric CI's on log scale and then converting them to asymmetric RR scale. (ratios of statistics tend to be skewed)

ii RD... using symmetric CI's based on
$ME = 1.96 \sqrt{p_1[1-p_1]/n_1 + p_2[1-p_2]/n_2}$ from difference of two independent sample proportions, each subject to binomial variation.
[Graph is from Excel spreadsheet of the type used to display 4 items 'volume-high-low- closing' for stock market activity. using volumes of zero and deleting the corresponding left vertical axis]



8i If Ho true and if carry out many tests, increase chance of a false positive result. Stricter alpha level for individual tests reduces risk of this.

ii Prob[at least 1 FP] = 1–Prob[all 14 negative] = 1–0.95^{14} = 1–0.49 = 0.51.

iii Prob[at least 1 FP] = 1–$(1-0.05/14)^{14}$ = 1–0.951 = 0.049.

iv The two outcomes Low Birth Weight and Preterm Delivery are biologically linked and so are not independent.

v If results were in same direction in neighbouring countries with same setup, might be inclined to think that "chance does not strike the same region twice". But this isn't the case [RR 0.67 in Portugal but 1.84 in Spain]. So no consistency, maybe 1.84 is chance.

vi I did use geographic contiguity to try to interpret them; had the pattern in the two been similar I might have been more willing to think it was real. So I am using 'outside' information to read something into the p-values: ie even if p-values are same blinded and unblinded, I do not interpret them in exactly same way. [Then again, I know very little about the medical setup and the possibility of such outcomes in these countries-- it may well be that the two neighbouring countries are as different as Ireland and Hungary in so far as what one would expect of the interventions. Other factors, such as integrity of the randomization, compliance, etc etc come into the interpretations too]

vii The same p-values, should, if combined with unequal prior beliefs, lead to unequal posterior beliefs. We must not interpret p-values in a vacuum.

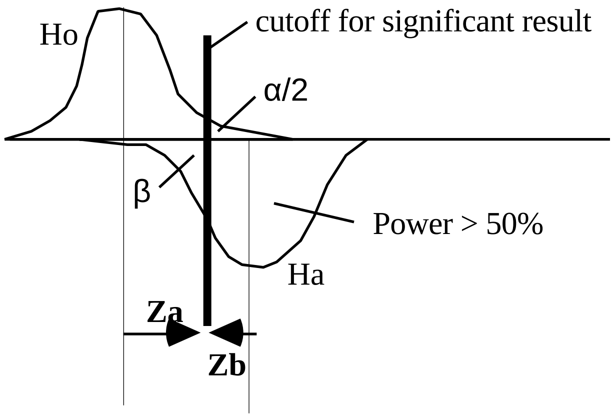
9i Data in form of 7 Binomial proportions, or counts in a 2x7 table. So can use a chi-square test of equality of proportions (or lack of associoation between rows and columns in table).

ii π=proportion of Low Birth Weight.
H_0: $\pi_{\text{Belgium}} = \pi_{\text{Denmark}} = \ldots = \pi_{\text{Spain}}$
Ha: Some variation among the 7 π's.

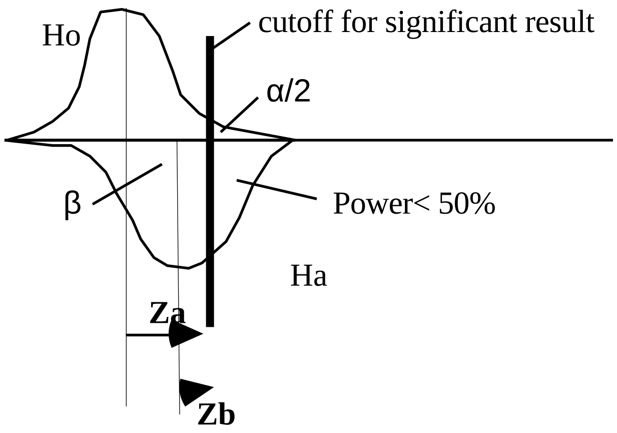
iii That have evidence against H_0. Not clear where source of variation is.

10i $2750 = 2\{z_{\alpha/2} + z_\beta\}^2 \dfrac{\bar{\pi}[1-\bar{\pi}]}{\Delta^2}$. Here $\bar{\pi} = 0.045$ and $\Delta = 0.01$.
Thus $\{z_{\alpha/2} + z_\beta\} = 1.79$.
The correct formula involves $\{z_{\alpha/2} - [\, z_\beta\ ]\}$ When we have a power of > 50%, z_β will be a negative quantity so the difference comes to more than 1.96. Many, as I do, will write the formula as $\{z_{\alpha/2} + z_\beta\}$ with the understanding that $\{z_{\alpha/2} + z_\beta\}$ will be a positive quantity. Here the amalgam of the two z's comes to less than $z_{\alpha/2}$ alone, so in fact we have less than 50% power [Less than half of the Ha distribution is beyond the cutoff for statistical significance]. Strictly speaking the 1.79 is {1.96 – [+0.17]}. It is best to draw a diagram to illustrate.

Usual Situation (Power > 50%)



Low Power Situation (Power < 50%)



The power is the amount of the Ha distribution that is beyond the cutoff for significance. Here it is the proportion of the Z distribution that is to the right of Z=+0.17. This is approximately 43%.

ii The comment was to illustrate that if use CI of .89 to 1.29, we are 'ruling out' RR's below 0.85; and before the study the authors agreed that the RR had to be below 0.8 to be clinically important. So the study provides a 'definitive negative' result rather than an inconclusive one.

iii The probability that if a given difference exists, the study will produce a positive [ie 'statistically significant'] result.

iv See the text. A CI is an interval produced using a procedure that yields a 'correct' answer (within the margin of error specified) with a given confidence.

v We should not interpret the results in isolation but use them to revise what we believed before the study.

11i Cost, acceptability, displacement of resources, gravity of problem, etc... Clinically significant Δ = the value of Δ at which we should switch from old to new or "the Δ that makes a Δ" [à la W. Spitzer]

ii 95%CI(RD) = 0.077–0.066 ±1.95$\sqrt{\frac{0.077x0.923}{2683} + \frac{0.066x0.934}{2688}}$ = 0.011±0.014 or –0.003 to +0.025. These are all on the proportion scale. In the percentage scale the interval becomes –0.3% to +2.5%. [Note also that, unlike for a test, the CI does not have to use a common p in the variance p{1–p}. In actual practice, the use of a common p for both will barely change the margin of error]

The 2.5% means that if one intervenes on 100, one would prevent an average of 2.5 ie NRT=100/2.5 or 40. The reverse sign on the –0.3% means that one might cause 1 for every 100 intervened on, i.e. no benefit. This translates to 100/0 or infinity. So the 95% CI for NRT is (40,infinity).

Differences in Proximal Femur Bone Density over Two Centuries

1 They took 5 measurements for one femur, calculated the mean and SD, and from them the CV = 100x(SD/mean). They did the same thing for the other two femora. Presumably the 1.245 is the mean [or median] of these 3 CV's.

2 Assuming a linear relation, the estimated difference between mean density of femaor from persons aged x years and persons aged x+1 is 0.197(%)

3i Positive gradient (top) negative (bottom; both lines got through (xbar, ybar)

3ii When I made up this Q last year, I thought the answer was no, since we did not have the SE for the slope. However a number of students last year and this year pointed out that in simple linear regression the test of beta=0 is equivalent computationally to a test of rho=0, and the latter can be done directly from knowledge of r and n (Colton has a table for this, or one can use the test statistic $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ as in M&M p 669). Others of you used the nomogram for CI's for rho based on a given r; if you want to use this nomogram for a test, you can either enter the diagram at the given r value and read the CI from the vertical axis, or else enter the diagram at rho=0 on the vertical axis and determine the horizontal 95% range of possible r's and see if the observed r is in that range... this latter way is equivalent to Colton's approach. This year I got from one student one further method that I had overlooked: in a simple linear regression, one can obtain all of the relevant coefficients from just 5 summary numbers, the sums of the x's, the squared x's, the sums of the Y's, the squared y's, and the cross products i.e. x•y 's . From the means and SDs in Table I one use the mean and the Sd to get the sums of the x's and the squared x's, from Table II the same for the y's and squared y's, and from the r in table III one can reconstruct the sum of products. Then using the steps in M&M p659- (Calculations for regression inference) one can get to the SE's for the coefficients, etc.

4 form the test ratio $\frac{-0.658 - 0.197}{\sqrt{SE[-0.658]^2 + SE[0.197]^2}}$ and test against t-table with df = sum of df's for the residual sums of squares [23+60 premenopausal]

5 Ho: rho = 0 6 Most are covered in the discussion.