AACC

Better health through laboratory medicine.

PEARLS OF LABORATORY MEDICINE

SPECIAL ISSUES:

P-values and Confidence Intervals
Power and Sample Size

Julie E. Buring, ScD

CCTC
Clinical Chemistry Trainee Council

REVIEW OF CHANCE

- Chance is always an explanation for our data, because we are trying to draw a conclusion (make an inference) about all people with an exposure and/or an outcome based on a limited sample of the entire population.
- Chance or sampling variability must be taken into account when we describe our data, as well as when we make comparisons between groups.
- Overriding principle: size of the sample on which we are basing conclusions will play a major role in the likelihood of chance being an explanation for our findings.

- One common way to measure the effect of chance is by conducting a test of statistical significance.
- Set up a null hypothesis (H_0): nothing is going on, no difference, no association.
- Test the alternative hypothesis (H_1): something is happening, there is a difference, there is an association.
- Perform the appropriate test of statistical significance.

Specific tests are for specific situations (ex. t-test, χ^2-test, Fisher's exact test, z-score), but all the tests have the same basic structure, in that each test statistic is a function of the difference between the values that were observed in the study and those that would have been expected if H_0 were true, as well as the amount of variability in the sample (sample size).

- **All tests of significance lead to some measure of the effect of chance on the results of a study.**
- **One measure is the resultant p-value: the probability of obtaining a result as extreme as or more extreme than the actual sample value obtained given that the null hypothesis (H_0) is true.**
- **On basis of p-value and based on an *a priori* chosen cutoff (usually: $p < 0.05$, $p \geq 0.05$), either will reject H_0 (conclude association is statistically significant at $p=0.05$ level) or will not reject H_0(not statistically significant at 0.05 level).**

PROBLEM

- The p-value reflects both the strength of the association and the sample size of the study (i.e., the variability).
- Even a small difference will achieve statistical significance (i.e., be judged unlikely to be due to chance) if the sample size is big enough.
- Even a big difference will not achieve statistical significance (i.e., chance cannot be ruled out as a possible explanation) if the sample size is too small.
- Problem is when you have a small to moderate-sized difference which is not statistically significant - can you conclude that nothing is going on (no effect) or is it that the sample size wasn't large enough to detect an effect that size statistically even if truly there.

- To separate out these two components of the p-value, the confidence interval should always be reported.
- The range of values within which the true magnitude of effect (e.g., RR or absolute difference) lies with a certain degree (e.g., 95%) of confidence.

- The confidence interval can provide the information of the p-value, in assessing whether an association is statistically significant or not at that level.
- But far more importantly, the width of the confidence interval reflects the precision of the estimate, i.e., what the true value is likely to be.
- The interpretation of the confidence interval, then, will depend on the scientific question you are trying to address.

EXAMPLE

- To take postmenopausal hormones (PMH)?
- Clear benefits on postmenopausal symptoms and risk of osteoporosis.
- Increased risk of breast cancer or endometrial cancer?

PMH AND ENDOMETRIAL CANCER

STUDY 1: RR = 7.5 p <0.05

STUDY 2: RR = 7.5 p <0.05

PMH AND ENDOMETRIAL CANCER

STUDY 1: RR = 7.5 $p < 0.05$ 95% CI* = (1.1, 32.1)

STUDY 2: RR = 7.5 $p < 0.05$

CONFIDENCE INTERVALS AND TESTS OF SIGNIFICANCE

If null value (e.g. RR = 1.0) is NOT CONTAINED within 95% confidence interval, then

1. Data not compatible with the null hypothesis
2. Corresponding p-value < 0.05

If null value (e.g. RR = 1.0) IS CONTAINED within 95% confidence interval, then

1. Data compatible with null hypothesis
2. Corresponding p-value ≥ 0.05

PMH AND ENDOMETRIAL CANCER

STUDY 1:	RR = 7.5	$p < 0.05$	95% CI = (1.1 - 32.1)
STUDY 2:	RR = 7.5	$p < 0.05$	95% CI = (7.2 - 8.3)

PMH AND BREAST CANCER

STUDY 1: RR = 1.13 $p \geq 0.05$

STUDY 2: RR = 1.13 $p \geq 0.05$

PMH AND BREAST CANCER

STUDY 1: RR = 1.13 p $\geq$0.05 95% CI* = (0.2 - 13.0)

STUDY 2: RR = 1.13 p $\geq$0.05

* Note: If p $\geq$0.05, then H_0 cannot be rejected. Thus the null value (ex. RR=1) must be in CI.

PMH AND BREAST CANCER

STUDY 1:	RR = 1.13	$p \geq 0.05$	95% CI = (0.2 - 13.0)
STUDY 2:	RR = 1.13	$p \geq 0.05$	95% CI = (0.96 - 1.2)

When is a confidence interval narrow enough? - this depends on the question being asked.

PMH AND ENDOMETRIAL CANCER

STUDY 1:	RR = 7.5	p <0.05	95% CI = (1.1 - 32.1)
STUDY 2:	RR = 7.5	p <0.05	95% CI = (7.2 - 8.3)

1. QUESTION: Is there something going on? Is the observed association unlikely to be due to chance?

   ANSWER: Both study 1 and 2 would tell you "yes". Both p <0.05, both informative about this question.

2. QUESTION: How sure are we about the precision of the observed magnitude of this association?

   ANSWER: From Study 1, not very sure (uninformative). From Study 2, precise estimate (informative).

PMH AND BREAST CANCER

STUDY 1:	RR = 1.13	$p \geq 0.05$	95% CI = (0.2 - 13.0)
STUDY 2:	RR = 1.13	$p \geq 0.05$	95% CI = (0.96 - 1.2)

1. **QUESTION:** Is there something going on? Is the observed association unlikely to be due to chance?

   ANSWER: Both studies would tell you "no" - both are "null studies" - not statistically significant - chance cannot be ruled out as an explanation for the findings.

2. **QUESTION:** But does this mean there is truly no association between PMH and breast cancer or was the sample size too small to detect this effect even if present.

   ANSWER: Study 1 is an uninformative null result - you cannot distinguish between these two alternatives; Study 2 is an informative null (you can tell the magnitude of effect).

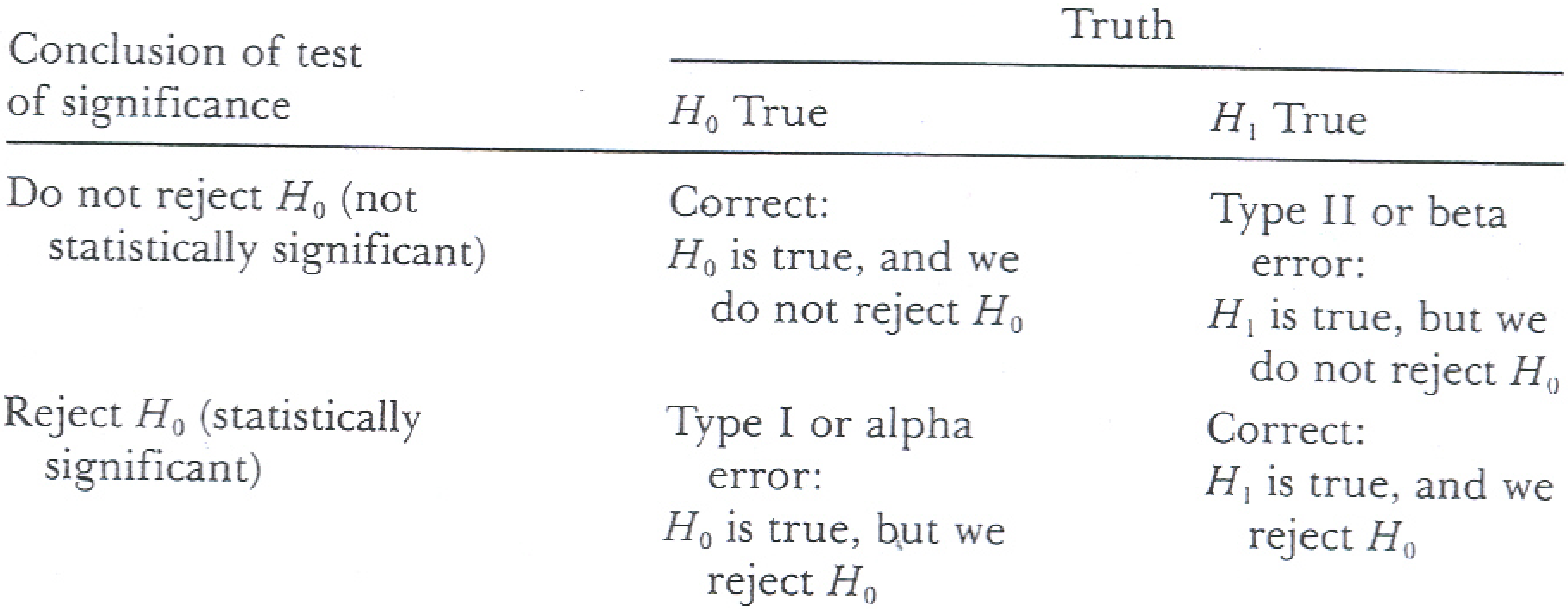
EVALUATION OF THE ROLE OF CHANCE INVOLVES 3 STEPS

1. Estimation of magnitude of effect or association (ex. RR).
2. Hypothesis testing: association due to chance? Is this a reasonable alternative explanation?

   p-value: probability that the observed association or one more extreme is due to chance alone, given that there is truly no association between the exposure and disease (i.e., H_0 is true).
3. Estimation of the precision of the effect measure, i.e., calculation of the confidence interval, or the range of values within which the true RR lies with a specified degree of confidence.

POWER AND SAMPLE SIZE

FOUR POSSIBLE OUTCOMES OF HYPOTHESIS TESTING

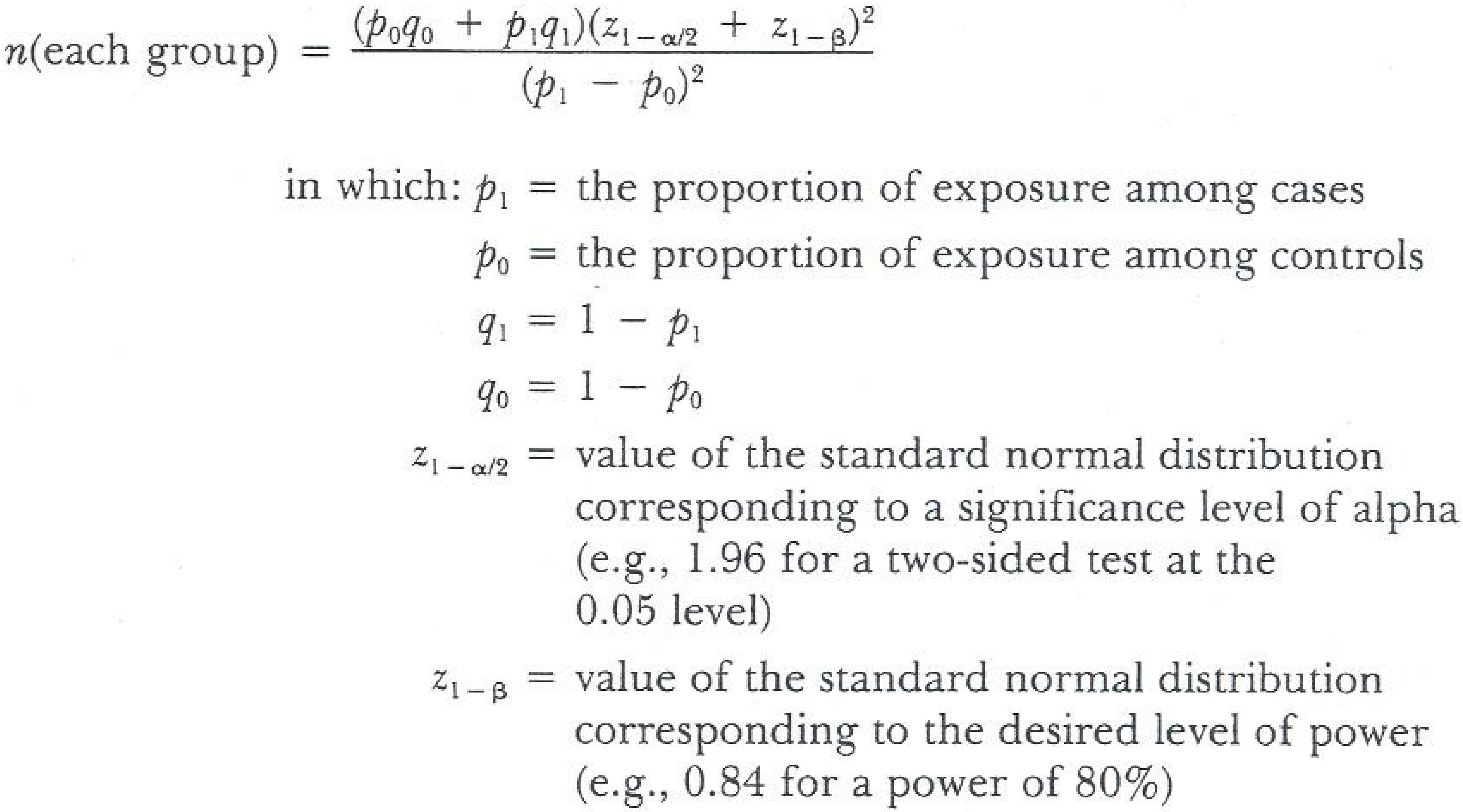


Conclusion of test of significance	Truth	
	H_0 True	H_1 True
Do not reject H_0 (not statistically significant)	Correct: H_0 is true, and we do not reject H_0	Type II or beta error: H_1 is true, but we do not reject H_0
Reject H_0 (statistically significant)	Type I or alpha error: H_0 is true, but we reject H_0	Correct: H_1 is true, and we reject H_0

POWER OF A STUDY

- Power = 1- Type II error.
- Power is the statistically ability to detect a difference or association when one is truly there.
- Probability of rejecting the null hypothesis when the alternative hypothesis is true.
- Just as conventionally test at 0.05 level for Type I error; minimum acceptable power is conventionally 80% (Type II error = 20%).
- Can calculate the sample size that will achieve that power to detect a postulated effect; or calculate the power that can be achieved to detect that effect given a fixed sample size.

Formula For The Calculation of Sample Size in A Case-control Study Evaluating the Difference between Exposure Proportions



$$n(\text{each group}) = \frac{(p_0q_0 + p_1q_1)(z_{1-\alpha/2} + z_{1-\beta})^2}{(p_1 - p_0)^2}$$

in which: p_1 = the proportion of exposure among cases

p_0 = the proportion of exposure among controls

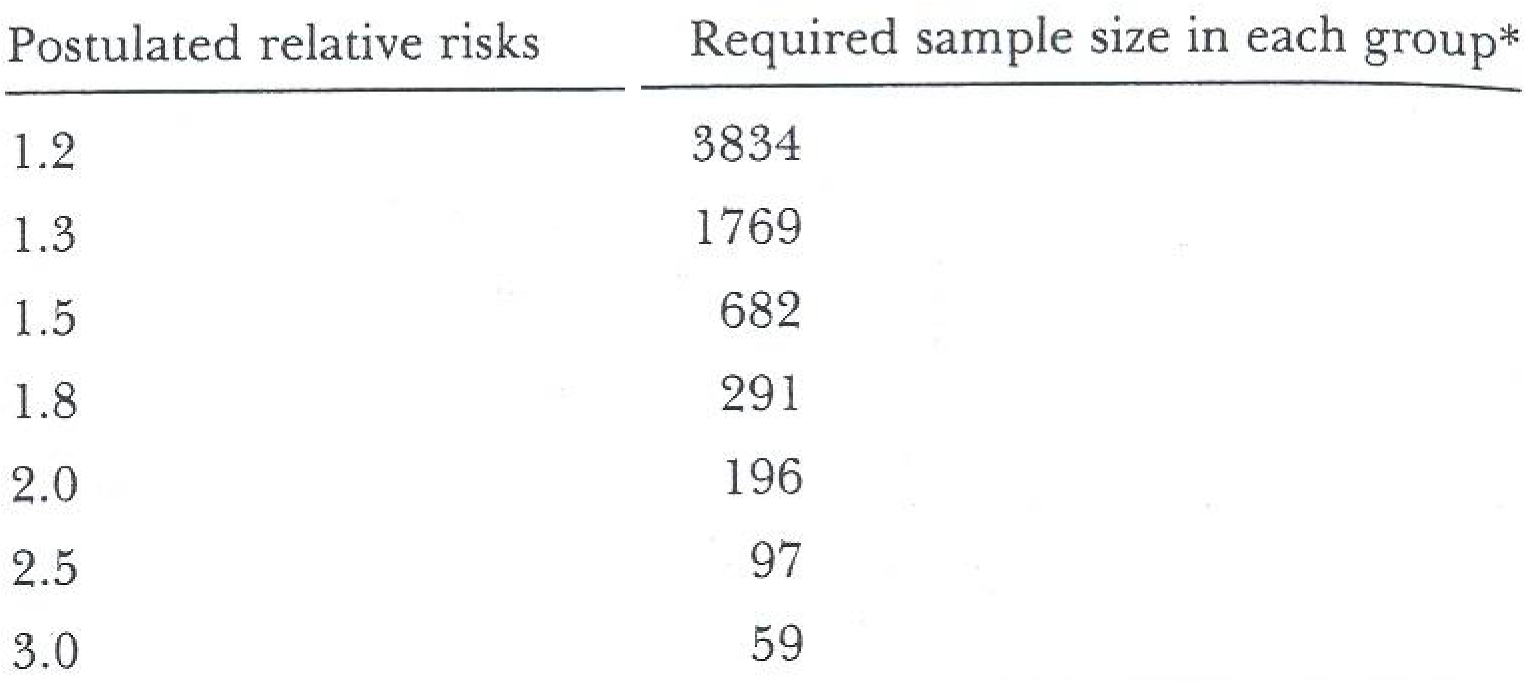
$q_1 = 1 - p_1$

$q_0 = 1 - p_0$

$z_{1-\alpha/2}$ = value of the standard normal distribution corresponding to a significance level of alpha (e.g., 1.96 for a two-sided test at the 0.05 level)

$z_{1-\beta}$ = value of the standard normal distribution corresponding to the desired level of power (e.g., 0.84 for a power of 80%)

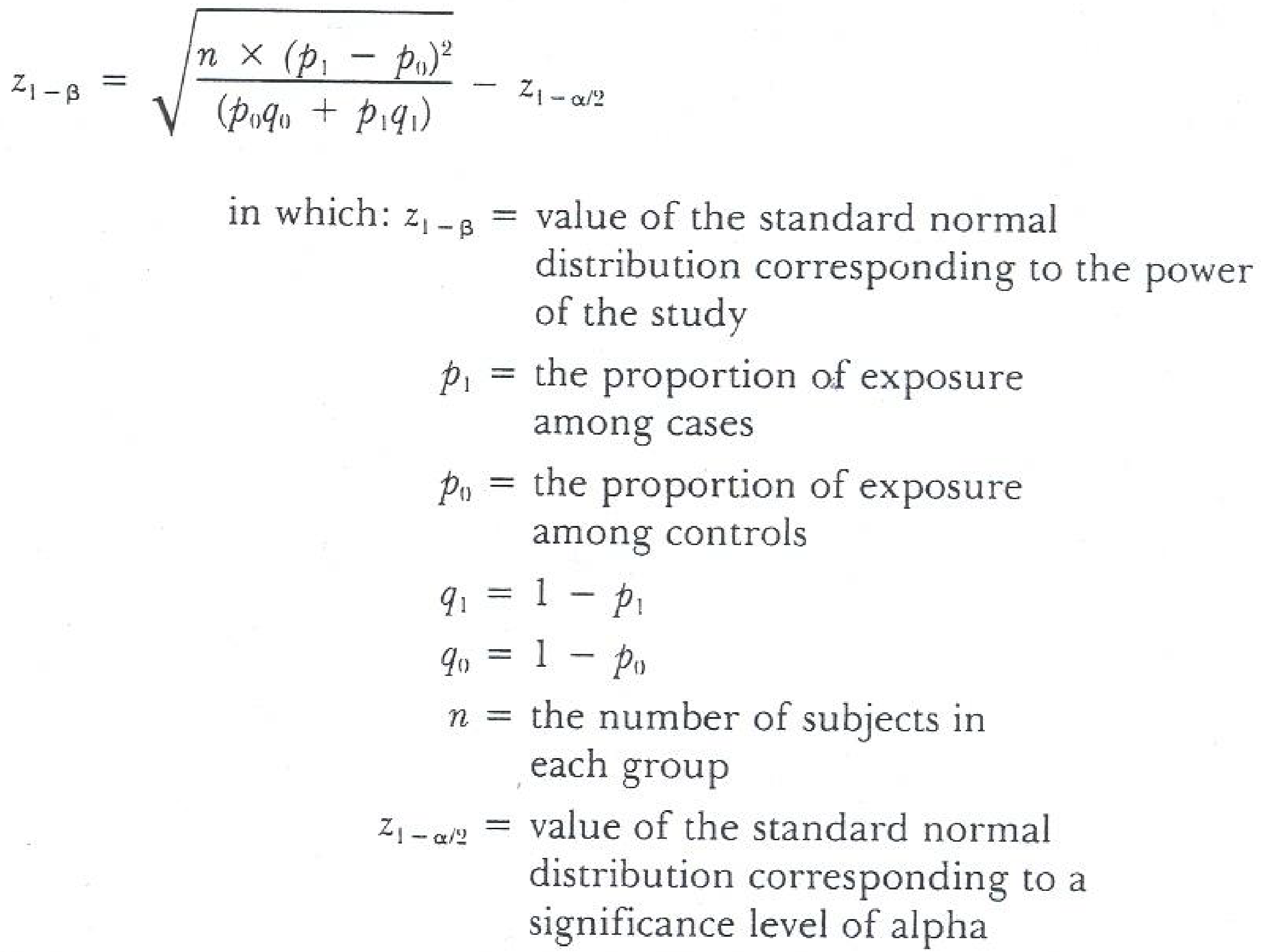
Sample Size Estimates for a Case-control Study of OC Use and Heart Attack Among Women



Postulated relative risks	Required sample size in each group*
1.2	3834
1.3	1769
1.5	682
1.8	291
2.0	196
2.5	97
3.0	59

*** Assuming proportion of current OC use in general population of women of childbearing age = 10%, power = 80%, type I error (two-sided) = 5%.**

Power Calculation Formula for Case-control Studies Using Discrete Data



$$z_{1-\beta} = \sqrt{\frac{n \times (p_1 - p_0)^2}{(p_0 q_0 + p_1 q_1)}} - z_{1-\alpha/2}$$

in which: $z_{1-\beta}$ = value of the standard normal distribution corresponding to the power of the study

p_1 = the proportion of exposure among cases

p_0 = the proportion of exposure among controls

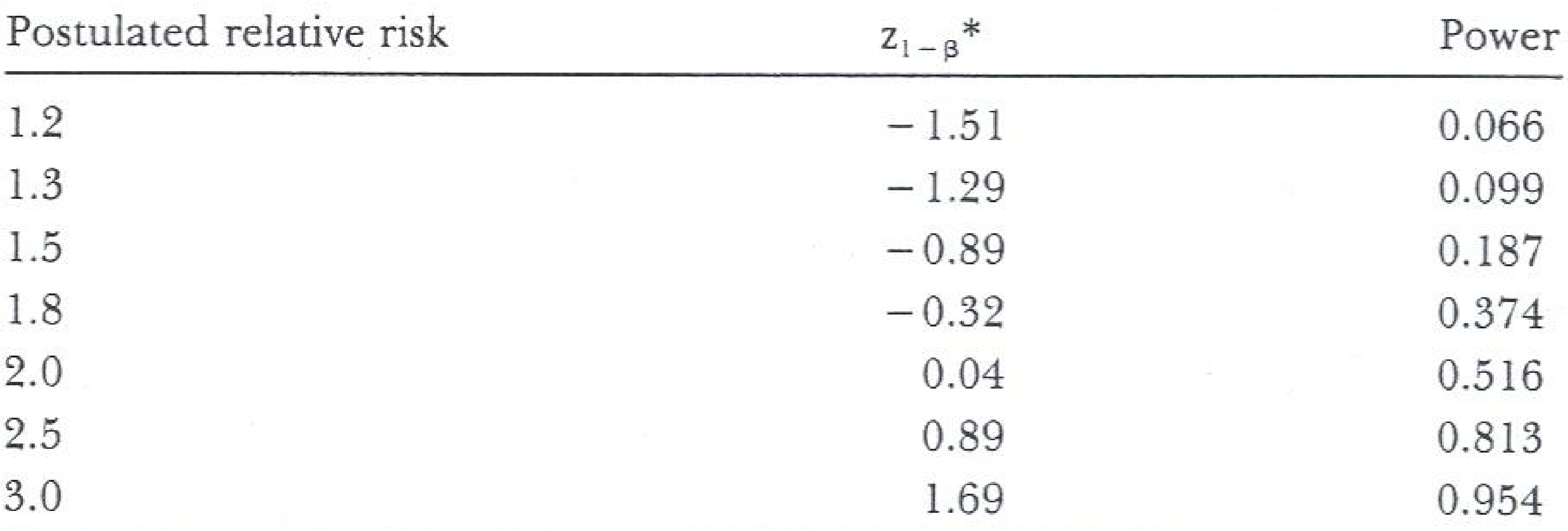
$q_1 = 1 - p_1$

$q_0 = 1 - p_0$

n = the number of subjects in each group

$z_{1-\alpha/2}$ = value of the standard normal distribution corresponding to a significance level of alpha

Power Associated with a Study of OC Use and Risk of MI with a Sample Size of 100 Cases and Controls, with Various Postulated Relative Risks



Postulated relative risk	$z_{1-\beta}$*	Power
1.2	−1.51	0.066
1.3	−1.29	0.099
1.5	−0.89	0.187
1.8	−0.32	0.374
2.0	0.04	0.516
2.5	0.89	0.813
3.0	1.69	0.954

When this number is ≤0.0, the power equals the area in one tail of the standard normal distribution corresponding to that value. When >0.0, the power equals (1.0 – that area).

POINTS TO CONSIDER

- Determining sample size and power is a back-and-forth discussion between the epidemiologist and the statistician: can be a reality check as to whether a study can be realistically achieved as proposed.
- When designing a study and writing the grant, calculate power; when interpreting a study, evaluate confidence intervals (theoretical power no longer relevant, only observed results of the study).
- Remember: sample size is based not on number of people, but number of endpoints. So may be OK for main endpoint, but not rarer events or subgroups. If these matter, need to power study for them too.

Thank you for participating in this *Clinical Chemistry* Trainee Council Webcast

Find our upcoming Webcasts and other Trainee Council information at www.traineecouncil.org

Follow us