Otolaryngology–Head and Neck Surgery (2009) 140, 794-799

INVITED ARTICLE

A practical guide for understanding confidence intervals and *P* values

Eric W. Wang, MD, Nsangou Ghogomu, BS, Courtney C. J. Voelker, MD, D. Phil (Oxon), Jason T. Rich, MD, Randal C. Paniello, MD, Brian Nussenbaum, MD, Ron J. Karni, MD, and **J. Gail Neely, MD,**
St Louis, MO

No sponsorships or competing interests have been disclosed for this article.

ABSTRACT

The 95 percent confidence interval about the mean demarcates the range of values in which the mean would fall if many samples from the universal parent population were taken. In other words, if the same observation, experiment, or trial were done over and over with a different sample of subjects, but with the same characteristics as the original sample, 95 percent of the means from those repeated measures would fall within this range. This gives a measure of how confident we are in the original mean. It tells us not only whether the results are statistically significant because the CI falls totally on one side or the other of the no difference marker (0 if continuous variables; 1 if proportions), but also the actual values so that we might determine if the data seem clinically important. In contrast, the *P* value tells us only whether the results are statistically significant, without translating that information into values relative to the variable that was measured. Consequently, the CI is a better choice to describe the results of observations, experiments, or trials.

Studies use a sample of patients who have a disease or have undergone a treatment to draw conclusions about the larger population of similar individuals. No matter how carefully the study sample is selected to minimize bias and baseline group differences, information gathered from a sample leads to some level of uncertainty and chance. Traditionally, *P* value, or probability, has been used to determine whether the results are due to chance. *P* value is limited in that it provides no information regarding the magnitude and precision of the results. In addition, *P* value does not address how much the results would vary if the study were performed numerous times. Conversely, a confidence interval (CI) is a range of plausible results that attempts to estimate the precision of the results and quantify the uncertainty inherent to studying a sample of the population.[1-6] In other words, the CI is the range of values about the sample mean that we can be relatively certain the true mean of the universal population falls.

There is a difference between actual data and our certainty, or inference, about the data. When we speak of a *P* value, SEM (often shortened to just SE[1]) or CI about the mean, we are discussing inferential statistics in contradistinction to descriptive statistics[7] (see Appendix).

This distinction between actual reported data and inference is important, primarily because we want to be able to generalize the results of a clinical article to our own practice. To do this, it is crucial to remember that the article generally reports *one sample* taken from the universal, or parent, population of subjects with the same characteristic (Fig 1). For example, suppose authors in New York report on the comparison of treatment A versus treatment B. We immediately think that a sample of subjects so treated is just like our patients here in Missouri, California, Texas, or Toronto, and that if we did the same thing, we would get the same results. In other words, we tend to generalize their sample to our own practice. However, to be a bit more discerning, we need to know just how stable their results are or how confident we might be about their reported data. That is where inferential statistics applies. Certainly, many factors of how they obtained the data may be more important, but this article is concerned only with the stability of the obtained data.

Descriptive statistics describes the actual data from a *sample* group, experiment, or trial. Actual data generates individual data points (X_i), the central tendency of the data, such as the mean ($\bar{X}$) or median, and the spread of the data, described as standard deviation (SD or s), interquartile range, or inner percentile ranges (*ipr*)[7] (Fig 2).

Inferential statistics describes what we might expect if the same sampling of a similar group, experiment, or trial was repeated many times to characterize the universe, or *parent population*. This helps us know how confident we

Received January 19, 2009; accepted February 3, 2009.

0194-5998/$36.00
doi:10.1016/j.otohns.2009.02.003

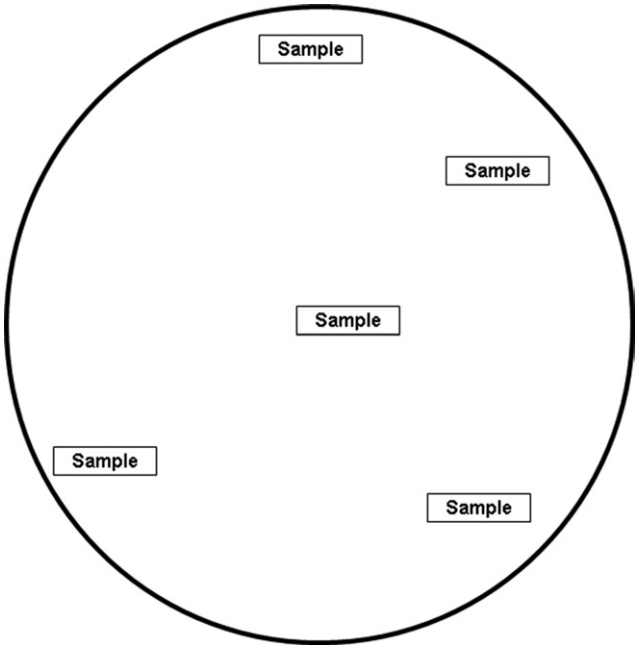

Figure 1 The *large circle* represents the universal, parent population of similar subjects and the *boxes* represent samples of subjects that may be taken from this large population. All studies report only data from samples.

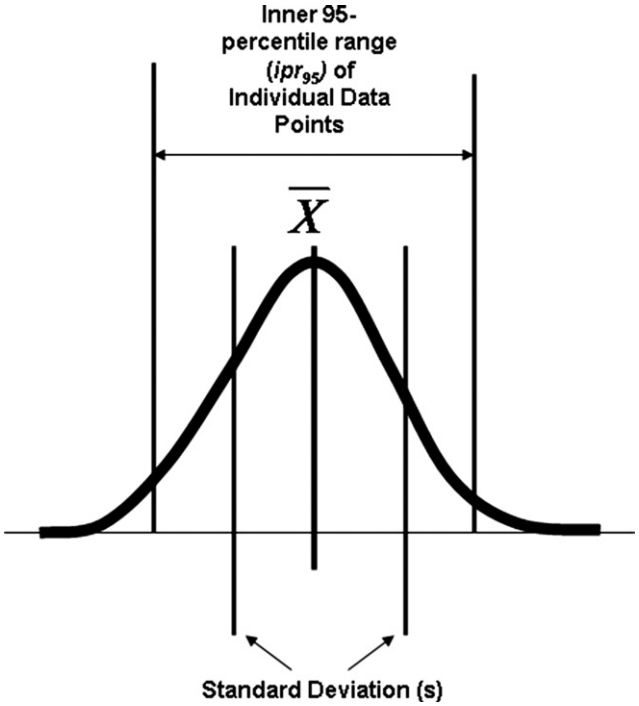

Figure 2 This illustration represents actual data relative to a single variable. The *central vertical line with bar X* is the mean, the *two lines* on either side of the mean represent the SD (s) about the mean, and the *outer lines* inscribe 95 percent of the data, defined as plus and minus 1.96 times the SD, and known as the inner 95th percentile range (ipr_{95}). The data can be further divided into quartiles in which the lower 25th percentile and the upper 75th percentile can be defined and inscribe the central 50th percentile of the data known as the interquartile range.

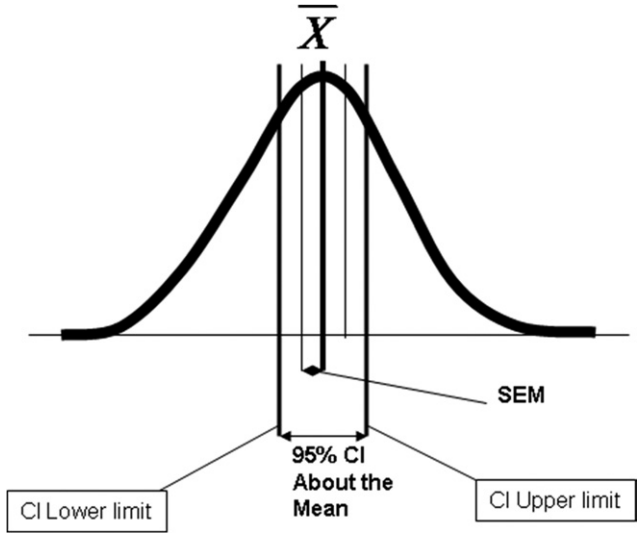

Figure 3 This illustration represents how stable the reported mean of data from a single sample variable is (SEM) and how confident we might be that the reported mean falls within 95 percent of the mean values (95% CI) if the same experiment were done many times over. The CI is demarcated by an upper limit and a lower limit of the mean values from hypothetical repeated measurements.

would be that the mean reported for a variable in an article is within 95 percent of the means obtained by repeated sampling (Fig 3)—or how confident we would be that the comparison between group outcomes is due to the intervention, rather than to chance.

COMPARISON BETWEEN GROUPS: *P* VALUES

When two groups are compared, each group has actual data defining the mean of each group and the spread of the data within that group. In comparing the groups, both the means and the spread are important. What catches our eye immediately is the difference between the means (Fig 4). However, the spread of the data within the groups may be more important (Figs 5A, B and 6).

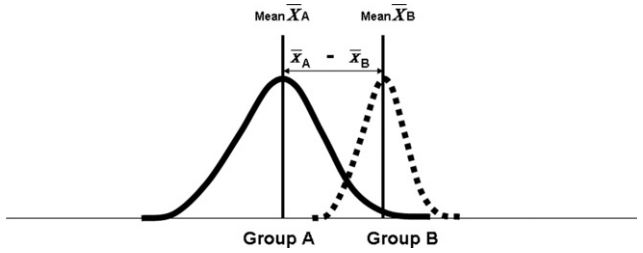

Figure 4 In this illustration of the two groups being compared, each group has its own mean and unique spread of data. The difference between the means immediately catches our eye.

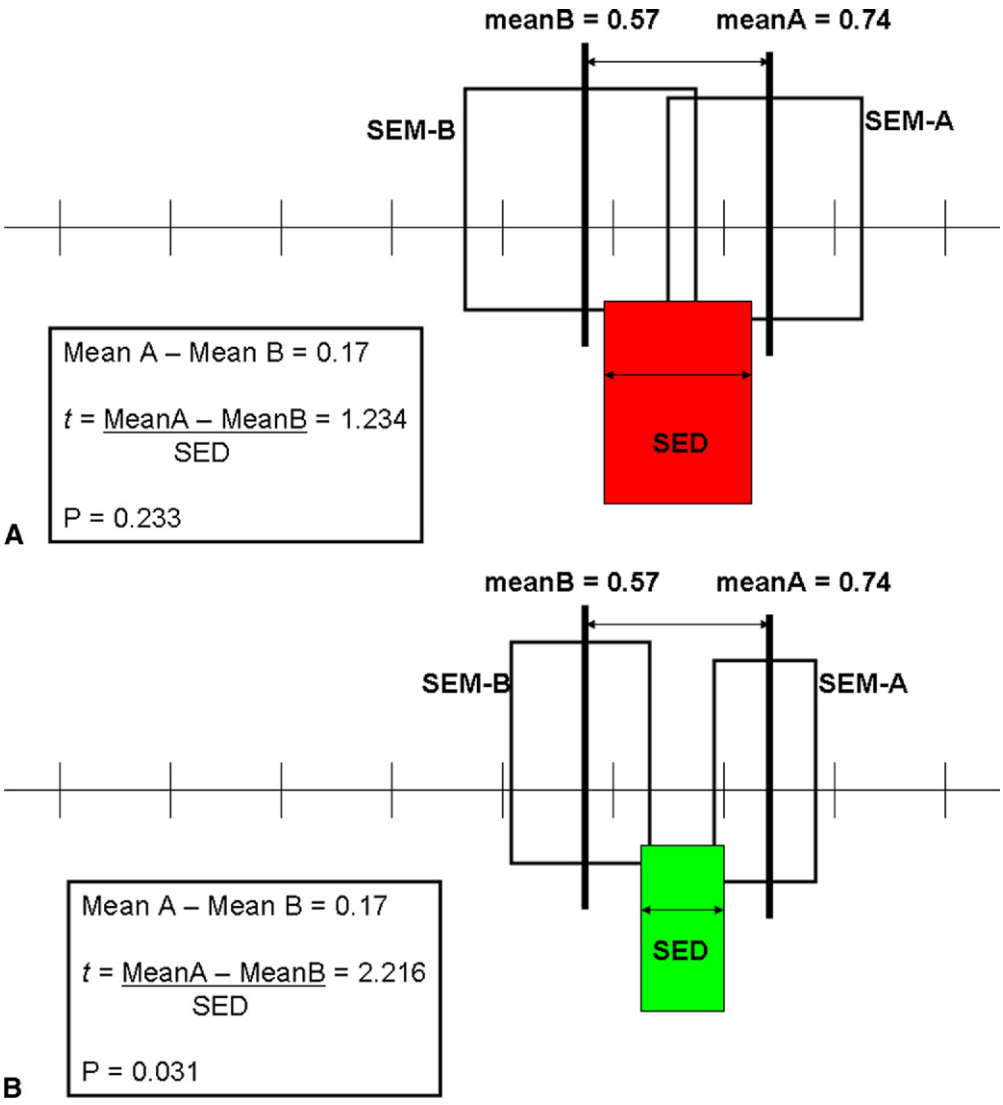

Figure 5 These illustrations show that the difference between means can be the same, but the spread of the data within each group can influence the probability that the groups are significantly different. Here, there is a mixture of actual data, the means of groups A and B, and inferential calculations such as the SEM for each group (SEM-A and SEM-B) and the SED.

In Figure 5, the difference between means are the same in Figure 5A and B; however, the SEMs (SEM-A and SEM-B) for each group are different, causing the standard error of the difference between means (SED) to be different. Explanations of these terms follow.

Understanding *P* Values

The inferential SEM is calculated from the actual standard deviation (s) divided by the square root of the number of subjects in the group. The SD is an index of dispersion of actual sample data, and the SEM is an index of dispersion of a hypothetical series of means repetitively taken from the parent population from which the sample was taken.[7] The SED is inferential as well and is calculated by taking the square root of the sum of the squared SD of group A divided by the number of subjects in group A plus the squared SD of group B divided by the number of subjects in group B.[7] It is not so important to know how these are calculated, but it is useful to know from where these values come (Appendix).

In performing a *t* test, a "critical ratio" is calculated by dividing the difference between means by the SED.[8] This

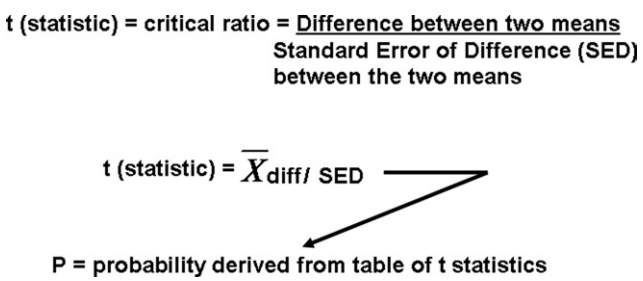

Figure 6 This figure illustrates the generation of probability scores using a *t* test.

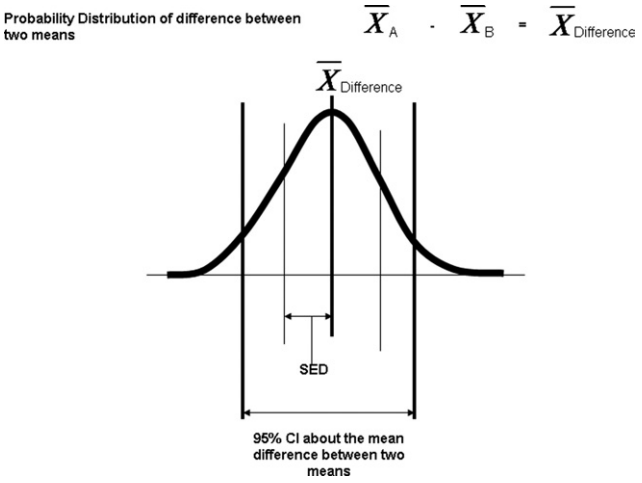

Figure 7 This illustration depicts a frequency distribution curve of the comparison between means of two groups and the generation of the 95 percent CI about the reported difference between the group means. *SED,* standard error of the difference between the two means (see text for calculation).

critical ratio defines the statistic "*t,*" which is then looked up in a probability table for *t* statistics to get the *P* value[9] (Fig 6). The table shows us that the number of subjects plays a major role in the resulting *P* value.

The *P* value assures us of the probability of the results being really from the intervention ($P < 0.05$) versus simply by chance ($P > 0.05$). However, because numbers of subjects are so important, it is possible to have statistically significant results, but the difference is so small as to be clinically meaningless. Thus, the *P* value is qualitative in the sense that it tells us yes or no if chance played a major role; however, it does not tell us quantitatively about the generalizable values associated with the sample means. More discussion of this issue will be presented later.

COMPARISON BETWEEN GROUPS: CONFIDENCE INTERVALS

The CI about the mean is also inferential but serves as an index of the *values* associated with a sample mean as well as an index of statistical significance if the experiment or trial were done over and over to create a hypothetical series of means. The values give us a sense of just how meaningful the sample data is to generalize to our practice. CIs may be set at any percent desired; however, in most cases 95 percent is usually chosen (95% CI). This means that if samples of the parent population were taken over and over, 95 percent of the values of the resulting means would fall within this CI, demarcated by the upper limit and lower limit of the CI (Figs 3 and 7).

The statistical interpretation is easy. If the whole CI is on one side or the other of the no difference marker, 0 in continuous data and 1 in ratios, then the results are statistically significant; however, if the interval crosses the marker, the results are not statistically significant. Note that this is true for data with normal distributions; however, if the data is markedly skewed, other measures such as the use of the median are more appropriate. Discussion of skewed data is outside the scope of this article.

The interpretation of clinical significance, or importance, is an additional value of the CI. The CI values may show us that even statistically significant results really do not seem to mean much clinically. For example, in a study, a large number of subjects are treated with canal wall up mastoidectomy and a similarly large number of subjects are treated with a canal wall down mastoidectomy, and a mean difference in hearing of 3 dB is calculated. The mean difference may be statistically significant in favor of canal wall up procedures because $P < 0.05$ and the CI about the mean difference of 3 dB ranges from 1 to 5 dB. However, we might not think that this small difference is clinically meaningful. On the other hand, an article that finds, in a smaller number of subjects, that the mean difference is 15 dB, but because the *P* value is >0.05 and the CI ranges from –1 to 31 dB, the results would not be statistically significant. However, we might feel that this may be a clinically very important preliminary finding. Thus, it might be worthwhile to test a larger number of subjects to determine if this difference holds up and is both statistically and clinically meaningful. Explanations of terms in calculating CI follow.

Understanding the Calculation of Confidence Intervals

Again, the SEM is an index of dispersion of a hypothetical series of means repetitively taken from the parent population from which the sample was taken.[7] As seen in Figure 3, a CI also may be obtained for univariate data for a single variable. The CI is calculated as the mean $\pm$ Z_α(SEM); in larger sample sizes, z_α for a two-tailed description = 1.96 for 95 percent CI[7] (Appendix).

When comparing two groups, a new mean value, the difference between means, is generated. Likewise, the new SEM in this new group is the SED. If repetitively the same experiment were done many times, a new frequency distribution of a series of values representing the *differences between means* would be constructed, but the SE this time would be the SED. Thus, the CI is calculated, as above, as

Table 1
2 × 2 contingency table

	Dependent (outcome) variable		
Independent (predictor) variable	+	−	
Group A	**a**	**b**	a + b
Group B	**c**	**d**	c + d
	a + c	b + d	

the *difference between means* ± Z_{α}*(SED)* (Fig 7). Note that in this new distribution of differences between means, the SED is much like the SE and the 95 percent CI looks like the inner 95th percentile range in descriptive statistics of actual data (Appendix).

Most CIs are equidistant about the mean, having an upper limit and a lower limit often reported as such, as with the SPSS statistical program (SPSS Inc, Chicago, IL), or as a range separated by a hyphen or two numbers separated by a comma. When only one number is given for a CI, as is the case with the SigmaStat program (Systat Software Inc, Richmond, CA), it is assumed that the symbol ± precedes that CI single number to determine the upper and lower limit values about the mean.

However, occasionally, the upper and lower limits are asymmetrical about the mean. This is the case with odds ratios (OR), also called the cross-product ratio (ad/bc), because the CI is first calculated as the natural log (Log_e; ln) of the OR (ln(OR)) and the results (lnLower Limit and lnUpper Limit) are then converted back to a range of ORs by taking the antilogarithm of each using them as exponents of e[7,9] (Appendix, Table 1).

CONCLUSION

The 95 percent CI about the mean demarcates the range of values in which the mean would fall if many samples from the universal parent population were taken. In other words, if the same observation, experiment, or trial were done over and over with a different sample of subjects, but with the same characteristics as the original sample, 95 percent of the means from those repeated measures would fall within this range. This gives a measure of how confident we are in the original mean. It not only tells us whether the results are statistically significant because the CI falls totally on one side or the other of the no difference marker (0 if continuous variables; 1 if proportions), but it gives us the actual values so that we might determine whether the data seem clinically important. In contrast, the *P* value tells us only whether the results are statistically significant, without translating that information into values relative to the variable that was measured. Consequently, the CI is a better choice to describe the results of observations, experiments, or trials.

AUTHOR INFORMATION

From the Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine.

Corresponding author: J. Gail Neely, MD, Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine, 660 S. Euclid Ave, Box 8115, St Louis, MO 63110.

E-mail address: jgneely@aol.com;

neelyg@ent.wustl.com (alternative).

AUTHOR CONTRIBUTIONS

Eric W. Wang, primary contributor; **Nsangou Ghogomu,** primary contributor; **Courtney C. J. Voelker,** reader, editor; **Jason T. Rich,** reader, editor; **Randal C. Paniello,** reader, editor; **Brian Nussenbaum,** reader, editor; **Ron J. Karni,** reader, editor; **J. Gail Neely,** primary author.

DISCLOSURE

Competing interests: None.

Sponsorships: None.

REFERENCES

1. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994;121:200–6.
2. Shakespeare TP, Gebski VJ, Veness MJ, et al. Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. Lancet 2001;357:1349–53.
3. Sim J, Reid N. Statistical inference by confidence intervals; issues of interpretation and utilization. Phys Ther 1999;79:186–95.
4. Smith SD. Statistical tools in the quest for truth: hypothesis testing, confidence intervals, and the power of clinical studies. Ophthalmology 2008;115:423–4.
5. Visintainer PF, Tejani N. Understanding and using confidence intervals in clinical research. J Matern Fetal Med 1998;7:201–6.
6. Zou GY, Donner A. Construction of confidence limits about effect measures: a general approach. Stat Med 2007;27:1693–702.
7. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders; 1985. p. 102–3, 113, 145, 161, 124–5, 432.
8. Jekel JF, Katz DL, Elmore JG. Epidemiology, biostatistics, and preventive medicine. 2 ed. Philadelphia: WB Saunders; 2001. p. 158.
9. Motulsky H. Intuitive biostatistics. New York: Oxford University Press; 1995. p. 78 and Table A5.4.

APPENDIX

Formulas used in descriptive and inferential assessments

	Descriptive (of actual data sample)	Inferential (use of actual data to infer values if sample is repeated many times)
Continuous variables		
Mean ($\bar{X}$)	$\bar{X} = (\sum X_i)/n$	
SD (s)	$s = \sqrt{\sum(X_i - \bar{X})^2/N - 1}$	
Inner 95th percentile range (ipr_{95})	$ipr_{95} = \bar{X} \pm Z_\alpha s = \bar{X} \pm 1.96s$	
SEM ($S_{\bar{X}}$)		$s/\sqrt{N}$
95% CI about a single variable mean, large sample size, two-tailed		$CI = \bar{X} \pm Z_\alpha s_{\bar{X}} = \bar{X} \pm 1.96 s_{\bar{X}}$
95% CI about a single variable mean, small sample size (<30), two-tailed		$CI = \bar{X} \pm t_{\alpha v} s_{\bar{X}}$ Example: d t 29 1.699 20 1.725 10 1.812 df = n − 1
95% CI about the difference between means, two-tailed		$CI_{95} = \bar{X}_{difference} \pm Z_\alpha (SED) = \bar{X}_{difference} \pm 1.96 (SED)$
SED		$SED = \sqrt{\frac{S_A^2}{n_A}} + \sqrt{\frac{S_B^2}{n_B}}$
Proportions		
SD of a proportion	$s = \sqrt{pq}$	
SE of a proportion		$s_{\bar{X}} = \sqrt{pq/N}$
95% CI of a proportion		$95\%\ CI = p \pm Z_\alpha (\sqrt{pq/n})$
95% CI of OR		$95\%\ CI \ln(OR) = \ln (OR) \pm 1.96 \sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}$ $\rightarrow Lower \lim 95\%\ CI (OR) = OR - e^{\ln LL}$ $\rightarrow Upper \lim 95\%\ CI (OR) = OR + e^{\ln UL}$

Certainly the universal parent population has a mean μ, a standard deviation σ, and a proportion π, if a dichotomous. However, the importance of these two columns, described in terms of the sample nomenclature, is to emphasize those items that predominantly describe the sample and those that require significant inference from the parent population to allow a degree of confidence about the sample data.

X_i, individual value; $\sum(X_i)$, sum of all the individual values in a group; n or N, number of all individual values in the group under consideration; Z_α, Z frequency (probability) distribution of specific alpha level, which is usually 0.05, meaning the level set to demarcate statistical significance in which 5 percent error is acceptable; $\sqrt{}$, square root; $t_{\alpha v}$, t distribution at α alpha level (usually 0.05) and υ, degrees of freedom, which is n minus the number of times a mean is calculated (usually n − 1 per group); p, proportion of interest; q, 1-p.

$\sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}$ where A,B,C,D are the values from the cells (a,b,c,d) in the original data 2 × 2 table generating the OR, also known as the cross-product ratio (ad/bc). ln = natural log = Log_e. e = 2.718281828. lnLL = natural log of the CI lower limit and lnUL = natural log of CI upper limit.