Tools of the Trade:

SAMPLE ESTIMATES COMPARED USING CONFIDENCE INTERVALS

In public health, much of the knowledge that we rely upon is obtained from scientific samples because it is the only feasible way to obtain the information. Results from samples of human populations are always estimates. The estimates vary by the error or bias associated with the sample survey process. Sample bias may occur as the result of a large array of things, such as poorly designed questions, sample coverage, or the recall, honesty and understanding of the respondent. Most biases can't be quantified, but we endeavor through careful sample design and care in crafting and testing survey questions to minimize these biases. The aim is to keep non-sampling error (biases) small, expecting some of them to cancel out one another. Hopefully the overall effect of the biases, which we can never be sure we have eliminated, will be smaller than the statistical sample error which we can accurately quantify.

All estimates obtained from probabilistic (scientific) samples have some error attributable to the sampling process. This sampling error is accurately quantifiable and is usually represented as the confidence interval (CI) or confidence bounds. Any level of confidence could be determined but the 95% level is almost universally used. So if a sample of Pennsylvania residents indicates that 54% (95% CI 51%-57%) believe that the sky is falling, then we can be confident that, if we conducted 100 different samples using the same sample design at least 95 of them would result in estimates between 51% and 57% of Pennsylvanians believing the sky is falling.
To accurately compare two sample estimates to determine if they are actually different, after accounting for the sample error, a statistical test (significance test) is required. A significance test will yield the probability (P-value) that the two sample estimates being compared are actually no different from one another. In order to calculate the P-value, one must have the actual data collected in the sample. Additionally, nearly all of the samples we encounter use a complex sample design in order to make the sample more efficient or affordable. If the sample has a complex design you will also need knowledge of the design, the weighting, and specialized software to account for the sample design when you calculate the variances.

A common level of significance is $P<.05$, or a chance of less than 1 in 20 , of the compared values actually being the same when the statistic says they are different. Although, you should keep in mind that if you are observing a large number of relationships with P-values near .05 then approximately 1 out of 20 of those significantly different estimates will not really be different (false positive). You may want to rely on a higher level of statistical significance, such as $P<.01$ or 1 in 100. Unless a particular hypothesis is being tested, it is not common to have sample data reported with significance tests calculated. However, competent analysis of sample data will provide some quantification of the sample error in the sample. This is most often reported as a confidence interval (CI).

Even though you can't determine the actual level of statistical significance for the difference between two sample estimates by comparing the CIs, they can give you a good idea of how important the difference might be. There are two conditions that can be employed to help evaluate how different two estimates really are.

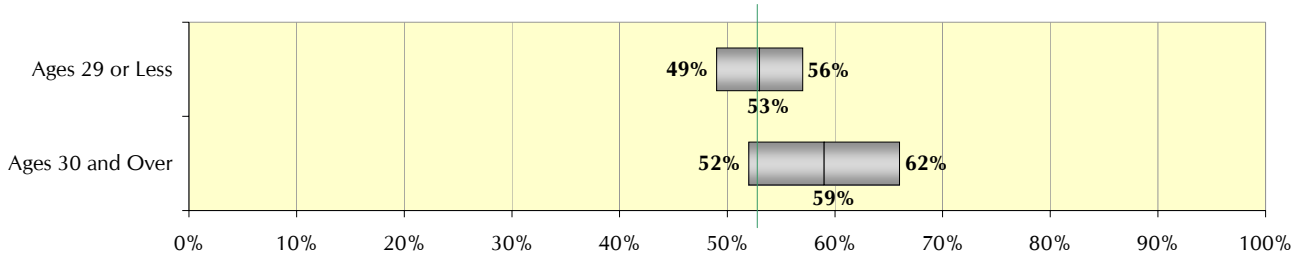
1) If either point estimate is contained within the 95% confidence interval for either of the estimates being compared, the difference IS NOT statistically significant at the <.05 level.

2) If the confidence bounds do not overlap, then the difference between the estimates being compared is most likely statistically significant. When the bounds do not overlap, the significance is almost certainly statistically significant at the <.05 level and it is very likely statistically significant at a <.01 level if the bounds for the two estimates are not very close together.

For example, let us compare the estimates of persons who think the sky is falling for different groups of Pennsylvanians.

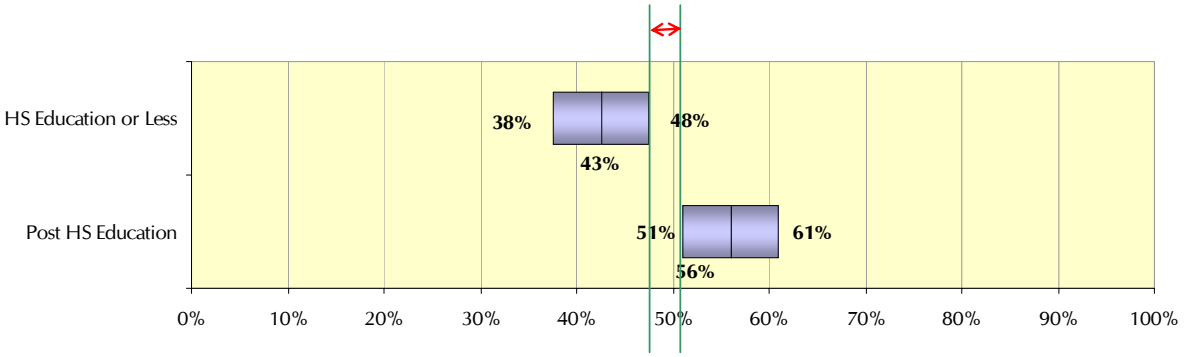
Condition 1 (see Chart 1) – Difference NOT statistically significant (P > .05): An estimate is contained within the bounds of the other estimate.

Chart 1 – Condition 1
Not Significantly Different Sample Estimates
One Sample Estimate Contained Within the Confidence Interval of the Other Estimate
(Hypothetical Example)



Condition 2 (see Chart 2) – Difference probably statistically significant (at least P < .05): When the confidence intervals for the estimates do not overlap.

Chart 2 – Condition 2
Significantly Different Sample Estimates
Confidence Intervals for the Sample Estimates Do Not Overlap
(Hypothetical Example)



When the difference in the estimates fall between the two above conditions, a determination of the statistical importance of the difference would require the calculation of a statistical test (see

Chart 3). You may want to think about these as “borderline different” or as differences you may want to consider as "unclear" since you would not be very confident in the differences.

Chart 3
Significant Difference Unclear
Confidence Intervals Overlap But Sample Estimate Not Contained in Other Estimate's Confidence Interval
(Hypothetical Example)