

Behavioral Risk Factor Survey Module

CONFIDENCE INTERVALS AROUND SAMPLE ESTIMATES

The BRFS, like all surveys, selects and obtains information from a sample of a larger population and calculates **estimated** percentages or subpopulation sizes. Two different samples taken from the same population at the same time will not yield exactly the same estimates. This means that estimates derived from a sample always have some degree of uncertainty around them.

The mathematical principles of statistical theory can be applied to survey design and sampling, observation weighting, and calculating estimates. This allows us to know, with some specific level of confidence, how large that degree of uncertainty around an estimate is. This is a **confidence interval (CI)**. The confidence interval or amount of uncertainty is then measured as, for example, some range of percentage points around an estimated percentage; some plus-or-minus number of points.

A common **level of confidence** used in survey research is 95%. This means that typically 95 of 100 different samples from the same population will produce an estimate within a calculated confidence interval around one sample's estimate. (This assumes the sampling followed procedures that did not seriously violate the mathematical assumptions behind the statistical theory, especially the assumption of random selection of each observation or interview.)

The tables in the BRFS module show both estimated percentages and confidence intervals. For example, in 2005 the statewide estimated percentage of adults currently smoking was 20.7%. The 95% confidence interval around that estimate is +/- 1.1%. We are 95% confident that the actual percentage of smokers in the whole adult Wisconsin population in 2005 was between 19.6% and 21.8% ($20.7\% \pm 1.1\%$).

Researchers would like confidence intervals to be as small as possible. First, less uncertainty around an estimate makes the estimate more useful. Second, when comparing two estimates for differences between, say, men and women, we can make a more precise test for **statistically significant differences** the smaller the confidence intervals are around the estimates. When confidence intervals overlap, we cannot be sure that differences in the estimates are not due just to sampling uncertainty. When CIs do not overlap, then we can say there is at least a statistical difference between the estimates. (This may or may not be a meaningful difference in real-world terms.)

In 2005, the Wisconsin BRFS estimates that 21.9% of men +/- 1.8% smoke, as compared to 19.4% of women +/- 1.5%. Is the difference between men and women statistically significant? (*Answer shown at end of this section.*)

Three main factors affect the size of a confidence interval. One is how confident you wish to be that the true percentage in the population is within the interval. The higher your desired confidence, the wider the interval will need to be: a 99% confidence interval will be wider than a 95% interval. This module calculates confidence intervals around the percentage estimates using a 95% level of confidence.

The second factor is the size of the sample used for the estimate. The larger the sample, the smaller the confidence interval will be. (The advantages of larger samples diminish above a certain point, however. There is little increase in advantage as sample size increases above about 400.)

Note: You can reduce a confidence interval by increasing the sample size. The easiest way to do this is by including more years of sample observations in your query. *We recommend using a sample size of at least 100 in any query and a larger sample than that would be well advised.*

The third influence on the size of a confidence interval is the estimated percentage itself. Percentages close to 50% have the largest CI for a sample size and confidence level. The CI decreases as the estimated percentage becomes smaller or larger than 50%. For a sample of 300 and a confidence level of 95%, the CI around 50% is +/- 5.7%; the CI around 5% or 95% is +/- 2.5%. This factor is automatically built into the calculations in the BRFS module of WISH.

[Answer: Not a statistically significant difference in smoking rates since the CI's overlap.]

HOW TO CALCULATE CONFIDENCE INTERVALS FOR POPULATION ESTIMATES

The output tables show the 95% CI for percentages but do not show the CI around the estimates of the numbers of persons in a category. This population estimate 95% CI may be calculated using the CI from the corresponding percentage.

Population confidence interval range = $R = (A/B)*D$,

where A = Percentage confidence interval range,

B = Estimated percentage, and

D = Estimated population.

Confidence interval around estimated population = $CI = D (+/-)R$

Example

In 2005 the estimated percentage of current smokers among Wisconsin adults was 20.7%,

with a confidence interval of +/- 1.1%. The estimated population of current smokers was 850,900.

First, calculate the range of the population confidence interval:

$$R = (1.1/20.7)*850,900 = 45,217 \text{ (or 45,200 rounded to the nearest 100)}$$

Then calculate the confidence interval by applying the range to the estimated population:

$$CI = 850,900 \pm 45,200$$

In other words, there is a 95% probability that in 2005 the true number of adult smokers in Wisconsin fell within the range of 805,700 - 896,100 people.

Prepared for Wisconsin Statistics on Health (WISH) Data Query System
<http://dhfs.wisconsin.gov/wish/> (January 2007)
Bureau of Health Information and Policy
Division of Public Health
Wisconsin Department of Health and Family Services