**COVID 19 – Random Sample - Preliminary Confidence Interval Calculations**

**David Canning**

**Harvard T.H. Chan School of Public Health**

**April 2020**

Assumptions: Random sample of the population. Will have to be adjusted for a design effect if sampling is stratified/ clustered. Also has to be weighted for non-response. Design effects are frequently around 2 – so may double standard errors.

I provide calculations for sample sizes 1000, 5000, 250,000. We begin by using the approximation that the mean of a large sample is equal to the population average +- 1.96 + sigma/sqrt(n) where sigma is the standard deviation of each observation and n is the sample size.

Graphs show how error bars for a 95% confidence interval – plus or minus percentage points- varies with prevalence percentage.

Figure 1 shows results. For sample size 5000 we are at around +/- 0.5% for prevalence rates under 5%. For prevalence around 50% we are at +/- 1.5%. For sample size 1000 we are at around +/- 1% for prevalence rates under 5%. For prevalence around 50% we are at +/- 3%.

Calculations above are for rates between 2% and 98%. Figure 2 shows the results for low proportions. For extremely low and high rates the normal approximation used to calculate the error bars break down and we have to use exact methods. The exact error bars are conservative to ensure nominal size of 95 % coverage of the confidence interval

Figure 3 shows exact error bars for estimated prevalence percentages below 2%. Note that the error bars are not symmetric. Also even if we find zero cases in the sample the confidence interval can be quite big. Even if there are no positive cases in the sample there may be cases in the community. Table 1 shows the upper confidence interval of the estimate even if we find zero cases in the sample for different sample sizes. Policy actions that require very low prevalence rates may require very big sample to estimate.

Table 1

|  |  |
| --- | --- |
| Sample Size | Upper confidence interval for zero positives in sample |
| 1000 | 0.3682% (4 per 1000) |
| 5000 | 0.0737% (7 per 10,000) |
| 250000 | 0.0014% (14 per million) |
|  |  |

Figure 4 shows exact 95 % confidence intervals at very low prevalence rates. These are from adding the error bar bounds in the previous figures on either side of the sample estimate. We have added here the confidence intervals for a sample of 250,000 for the national population.

We can also look at confidence intervals for the total number infected based on the sample. Current reported cases are at around 32,000. This is an underestimate since many people will be infected without symptoms, and with symptoms but without testing. Conservative estimates give a multiplier of 2 or 3 to get actual cases. However factors as high as 10 and 20 for the multiplier have been suggested. A multiplier of 10 gives about 0.08% infection rate and 20 a 1.7% infection rate for the USA.

Table 2 Cases in the US

|  |  |  |
| --- | --- | --- |
| Tested Positive Cases | Multiplier | Actual Cases |
| 32,000 | 1 | 32000 |
|  | 2 | 64000 |
|  | 3 | 96000 |
|  | 10 | 320000 |
|  | 20 | 640000 |

Figure 5 gives the exact 95% confidence interval for the US number of cases with a sample of 250,000 for multipliers in the range 1-3. We would be able to determine if the true cases were around 100,000 (95 % range around 80,000 to 122,000) rather than just the recorder 32,000

Figure 6 extends the analysis to multiplier factors up to 20 for actual cases. Again, we can accurately determine the multiplier with a sample of 250,000.

At the state level with a sample of 5000 the confidence intervals for the percent infected are the same for every state and are given in Figures 1 -4 . For the absolute numbers infected by state the confidence intervals by state because the same absolute number of infection is at a different percentage of each state’s population. California the largest state has a lot of infections at very low percentage infection rates while Wyoming has very few. We take 3 examples of different states with populations shown in Table 3.

Table 3

|  |  |
| --- | --- |
| State | Population estimate |
| California | 39,512,223 |
| Massachusetts | 6,949,503 |
| Wyoming | 578,759 |

Figure 7 show the confidence intervals for the number infected for California for a sample size of 5000. Figures 8 and 9 show results for Massachesetts and Wyoming. Note that at an estimate of 5000 infections in California the upper bound on the confidence interval is about 40,000 in California. In Massachusetts it would be between about 13,000 and in Wyoming about 6,800.

If we want to know percentages infected the same sample size is fine across states if we want the absolute number bigger states should get bigger samples.

Tobi, Hilde, Paul B. van den Berg, and Lolkje TW de Jong‐van den Berg. "Small proportions: what to report for confidence intervals?." *Pharmacoepidemiology and drug safety* 14.4 (2005): 239-247.