Report on Preliminary Analysis of Smoking Petition

2-17-05

#5

We estimate that there are 33,438 valid signatures on the smoking petition. Using a random sample of a size required by law, the City is 95% confident that the true number of valid signatures on the entire petition is less than 33,795 and is 99% confident that the true number is less than 33,943.

36,907 lines of names were submitted. A random sample of 9,239 of these lines was checked. 297 of the sample lines were disqualified on account of bearing signatures of persons not on the voter list (294) or of not bearing signatures (3). The remaining 8,942 sample lines were validated as bearing signatures of qualified voters. However, 155 qualified voters signed more than once: 142 signed twice, 5 signed three times, and one signed four times. Therefore, the sample contained the signatures of 8,942 – 142 – 2x5 – 3x1 = 8,787 distinct qualified voters, counting each distinct voter's signature only once.

Using these figures, we estimate that there are 33,438 valid signatures on the smoking petition. The method used for calculating this estimate is based on Goodman's method. (The Annals of Mathematical Statistics, 1949, pp. 572-579) supplemented with variance estimate based on Haas and Stokes (Journal of the American Statistical Association, 1998, pp. 1475-1487.) It should be noted that the estimate of 33,438 valid signatures does not simply extrapolate the 8,787 valid signatures by multiplying 8,787 by the petition-to-sample-size ratio 36,907+9,239 (essentially 4). To do so would give a much larger estimate (8,787 x 36,907 + 9,239 = 35,101) of the number of valid signatures. The 35.101 estimate is substantially inflated because it does not deal properly with multiple signatures. An intuitive explanation is that a duplicate signature in the sample will remain a duplicate if the whole petition is checked (or even become a triplicate or more) and will never change to a unique signature. But an apparently unique signature in the sample may become a duplicate (or more) if the entire petition is checked. This can happen in many ways: If an additional quarter of the petition is checked, that quarter will contain duplicates in and of itself. But the second quarter will also contain duplicate matches with signatures in the first quarter. If a third quarter is checked, it will have matches with guarters one and two. And if the fourth guarter is checked, it will have matches with the other three. It is these inter-guarter duplicates that the 35,101 estimate does not adjust for. John Doe can be a duplicate because he appears twice within one of the quarters. But he can also be a duplicate by being a unique signature within any two different quarters. The proportion of multiple signatures in the sample substantially understates the multiple signatures on the petition. Details on proper ways to adjust for multiple signatures are given in the cited references. The 33,438 estimate properly adjusts for all multiple signatures.

Random number generation and all programming was done with SAS[®] (Statistical Analysis System).