[CCLHDME thread](https://groups.io/g/CCLHDME/topic/age_adjusted_rates_with/104779159?p=,,,20,0,0,0::recentpostdate/sticky,,,20,2,0,104779159,previd%3D1709851783152166626,nextid%3D1705443097060839253&previd=1709851783152166626&nextid=1705443097060839253)

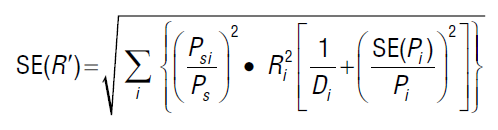
**[Matt Beyers]**

When calculating denominators from PUMS, these population data will come with error which can be calculated with survey weights. This is in addition to the normal error calculated in standard age-adjusted or other rates. You would have to account for error in both the numerator and denominator to calculate the confidence intervals or standard error for the rate.

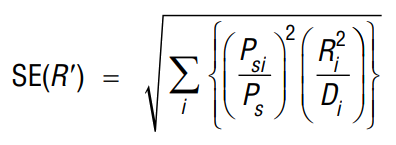
Does anyone have experience in calculating age-adjusted rates and their error, given error in the denominator?

**[Joshua Endow-Monteiro]**

There is some relevant information in the technical notes for the [Deaths: Final Data for 2020 report](https://stacks.cdc.gov/view/cdc/131355). Well not directly in those technical notes. To save you the search, those technical notes mention that supplemental tables I-5 through I-7 “are calculated using population figures that are subject to sampling error.” And if you follow the hyperlink to the [Deaths: Final Data for 2020 Supplemental Tables](https://www.cdc.gov/nchs/data/nvsr/nvsr72/nvsr72-10-tables.pdf), the Technical Notes Addendum for Supplemental Tables of “Deaths: Final Data for 2020” include the information Matt is looking for in the random variation section. Specifically, those tables include rates based on population estimates derived from 1-year ACS PUMS for 2020, and the following Formula 2 for calculating standard error of the age-adjusted death rate is provided:



For reference, the standard method for calculating standard error of the age-adjusted death rate when the population is not subject to sampling error is given in Formula 4 of the main report’s technical notes:



Even for those not planning on using PUMS, it might be a good idea to keep in mind that nearly all 2020 census population data have associated margins of error due to the infusion of noise from differential privacy.

The technical notes also mention the problem of misclassification of race and ethnicity, so [The Validity of Race and Hispanic-origin Reporting on Death Certificates in the United States: An Update](https://www.cdc.gov/nchs/data/series/sr_02/sr02_172.pdf) might be of interest as well. While that study only investigated the OMB race and ethnicity categories, I would expect misclassification to be more of a challenge with increasing disaggregation. Moreover, one needs to account for missingness due to lack of specificity in the responses.

**[Adam Readhead]**

I agree that it would be great to incorporate the errors in both the numerator and the denominator.  I have chewed on the problem a bit for earlier projects.  Here are a few of my thoughts on the matter:

1. I did some work to incorporate denominator errors for TB incidence rates by country of birth using Bayesian methods (see [Readhead et al. 2018 Challenges and solutions to estimating tuberculosis disease incidence by country of birth in Los Angeles County](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209051)).  I thought it was going make a noticeable difference but the upshot was that incorporating the errors of the denominator did not change things much.  I think that it because the denominator errors (as modeled) were normally distributed and not huge in any case.  The actual denominator error is probably different.
2. The misclassifications errors in the numerator and denominator will not be accounted for in the systematic errors one can calculate. And the misclassification errors are likely to be unruly that is not normally distributed and grossly different by category.  The misclassification of race and ethnicity on death certificates is a known issue (as Joshua pointed out) and misclassification on race and ethnicity of race and ethnicity in the ACS data is likely a problem too.  So even if you get the systematic errors combined correctly, you still have the ghost of misclassification hanging over the calculation.  Plus, unlikely the well-behaved errors from standard error of a rate and standard error of population survey, the misclassification is probably   I don’t have any fixes for that, except that we have to report out this issue when we report the stat.  And just to be clear, I think it’s worth calculating the stat anyway.
3. More broadly, I wonder whether going straight to an incidence rate ratio wouldn’t be more helpful.  I believe that presenting the age-adjusted incidence rates is a little misleading, because the whole point of doing the age-adjustment is to make an apples-to-apples comparison; it’s a construction we use to better compare between groups.  So if we are trying to compare groups, why not just skip to the ratio?  I would be concerned that someone might cherry pick the rate (“the age-adjusted mortality for group A is X”) which is kind of meaningless in my opinion.  I’d rather them cherry pick the rate ratio.
4. The regular error calculations for incidence rates are based on the assumption that the data follow a Poisson distribution which, in my experience, is usually not true.  It’s worth trying to model the data using both Poisson and negative binomial distributions to see how “off” the Poisson distribution is.  I have also found that the distribution is quite sensitive to the categorization of the covariates which was disconcerting.  And these models take some getting used the input data is aggregate (one row per numerator, denominator, and covariates) rather than individual (one row per person).