PHSD Rate Calculation Guidance

This document provides formulas and the equivalent computations in Excel, R, and SAS to calculate rates, standard errors, and confidence intervals for crude and age-adjusted rates. Parts of formulas in brackets and **in green text** should be replaced with their respective variables or cell references. See the references section at the end of the document for additional information regarding the derivation of formulas. Dataset-agnostic rate calculators in Excel, R, and SAS utilizing these methods are available on Gitlab.

Calculating crude rates

Because the proportion of many conditions or public health-related indicators among the population is very low, both crude and age-adjusted rates involve a multiplier to produce a whole number that is more readily understood by the general population (15 per 100k population, for example). In most public health applications, the multiplier will be 100,000, but can be any multiple of 10 depending upon the context of the outcome measure. For example, prescribing rates calculated with prescription drug registry data use a multiplier of 100. The result should be a whole number that makes sense for the type of outcome measure.

Crude rate

$$rate_{crude} = multiplier * \frac{count}{population}$$

Standard error of crude rate

$$SE_{crude} = multiplier * \frac{\sqrt{count}}{population}$$

In Excel: `=[multiplier]*SQRT([count])/[population]`

In R: `[multiplier]*sqrt([count])/[population]`

In SAS: `[multiplier]*sqrt([count])/[population]`

Variance of crude rate

$$v_{crude} = SE_{crude}^2$$

Confidence interval for crude rates based on counts less than 100

Rates should be suppressed for counts under 16 per PHSD Guidelines for the Release of Public Health Data. Confidence intervals for rates based on counts between 16 and 99 should be calculated using the chi-square distribution as shown below. These equations will produce the same results as using a Poisson critical values table but are easier to program. Below, $chiinv$ is the inverse of the chi-squared distribution and α refers to the desired p-value, which is typically set to be 0.05 to obtain a 95% confidence interval.

Last updated: 10/2/25 by Sandra.biller@mt.gov & Hannah.yang@mt.gov

$$CI_{lower} = multiplier * \frac{\frac{1}{2}\left(chiinv\left(\frac{\alpha}{2}, 2 * count\right)\right)}{population}$$

In Excel: `=[multiplier] *(0.5*CHISQ.INV([alpha]/2, 2*[count])/[population])`

In R: `[multiplier]*(0.5*qchisq([alpha]/2, 2*[count])/[population])`

In SAS:
```
[multiplier]*(0.5*quantile("chisq", [alpha]/2,
2*[count])/[population])
```

$$CI_{upper} = multiplier * \frac{\frac{1}{2}\left(chiinv\left(1 - \frac{\alpha}{2}, 2 * (count + 1)\right)\right)}{population}$$

In Excel:
```
=[multiplier] *(0.5*CHISQ.INV(1-([alpha]/2),
2*([count]+1))/[population])
```

In R:
```
[multiplier]*(0.5*qchisq(1-[alpha]/2,
2*([count]+1))/[population])
```

In SAS:
```
[multiplier]*(0.5*quantile("chisq", 1-[alpha]/2,
2*([count]+1))/[population])
```

Confidence interval for crude rates based on counts 100 and above

For counts 100 and above, the normal approximation using the Wald interval can be used to obtain confidence intervals.

$$CI = rate_{crude} \pm z_{\alpha/2}\frac{\sqrt{count}}{population} = rate_{crude} \pm z_{\alpha/2}SE$$

In Excel:
```
=[cruderate]+NORM.S.INV([alpha]/2)*[SE]
```
```
=[cruderate]+NORM.S.INV(1-[alpha]/2)*[SE]
```

In R:
```
[cruderate]+ qnorm([alpha]/2)*[SE]
```
```
[cruderate]+ qnorm(1-[alpha]/2)*[SE]
```

In SAS:
```
[cruderate]+ quantile("normal", [alpha]/2)*[SE]
```
```
[cruderate]+ quantile("normal", 1-[alpha]/2)*[SE]
```

Calculating age-adjusted rates

PHSD uses the 2000 US standard population as the standard population in age-adjustment.

Age-adjusted rate

$$rate_{adj} = \sum_{i=x}^{y}\left[multiplier * \frac{count_i}{population_i} * \frac{stdpop_i}{\sum_{j=x}^{y} stdpop_j}\right] = \sum_{i=x}^{y}[multiplier * agerate_i * w_i]$$

To break this into pieces:

- $multiplier$ is typically 100,000 (can vary based on the context of the outcome measure)
- $\frac{count_i}{population_i}$ is the age-specific rate for age group i, where $count_i$ is the number of events in age group i and $population_i$ is the relevant population in age group i.
- $\frac{stdpop_i}{\sum_{j=x}^{y} stdpop_j}$ is equivalent to w_i, the weight for age group i and is calculated by taking the standard population of age group i and dividing it by the total of the standard population across all age groups. It can be easier to compute weights separately. The Centers for Disease Control and Prevention (CDC) and the National Center for Health Statistics (NCHS) provide weights for commonly used age groupings.

These components are multiplied together for each age group and then added together.

Weights

In Excel: `=[standardpop_i]/sum([standardpop_x:standardpop_y])`

In R: `round([standardpop]/sum([standardpop]), 6)`

In SAS: `Round([standardpop]/sum([standardpop]), 0.000001)`

In R and SAS, `[standardpop]` should be a vector of the standard population counts for each age group.

Rates

In Excel:

`=[multiplier]*SUMPRODUCT([count_x:count_y],1/[population_x:population_y],[weight_x:weight_y])`

The R and SAS calculations accommodate calculating an age-adjusted rate for a single group or for any number of groups (ie- by sex, race, county):

- *`[count]` and `[population]` should each be an* n *x* k *matrix, where* n *is the number of groups, and* k *is the number of age groups. For example, when calculating rates for females and males using 12 age groups, the matrices should both be 2 x 12.*
- *`[weights]` should be a 1 x* k *vector, where* k *is the number of age groups. For example, when using 12 age groups, the vector will be 1 x 12.*

In R: `[multiplier]*([count]/[population]) %*% t([weights])`

In SAS: `[multiplier]*(([count]/[population]) # [weights])[,+]`

Standard error of age-adjusted rate

$$SE_{adj} = multiplier * \sqrt{\sum_{i=x}^{y}\left(\left(\frac{count_i}{population_i^2}\right) * \left(\frac{stdpop_i}{\sum_{j=x}^{y} stdpop_j}\right)^2\right)}$$

$$= multiplier * \sqrt{\sum_{i=x}^{y}\left(\left(\frac{count_i}{population_i^2}\right) * w_i^2\right)}$$

A description of the variables used in this equation can be found with the age-adjusted rate formula on page 3.

In Excel:
```
=[multiplier]*SQRT(SUMPRODUCT(([countx:county]),
    (1/[populationx:populationy])^2,
    ([weightx:weighty])^2))
```

In R:
```
[multiplier]*sqrt(([count]/[population]^2) %*% t([weights]^2))
```

In SAS:
```
[multiplier]*sqrt(([count]/[population]##2 # [weights]##2)[,+])
```

Variance of age-adjusted rate

$$v_{adj} = SE_{adj}^2$$

Confidence interval for age-adjusted rates based on counts less than 100

Rates should be suppressed for counts under 16 per PHSD Guidelines for the Release of Public Health Data. Either the chi-squared distribution or gamma distribution can be used to calculate age-adjusted rates when the count is 16-99 and the results will be equivalent. The chi-squared distribution is used in the programs developed and available on Gitlab.

Chi-squared distribution

$$CI_{loweradj} = \left(\frac{v_{adj}}{2 * rate_{adj}}\right) * chiinv\left(\frac{\alpha}{2}, \frac{2 * rate_{adj}^2}{v_{adj}}\right)$$

In Excel: `=([v_adj]/(2*[rate_adj]))*CHISQ.INV([alpha]/2, (2*(([rate_adj])^2))/[v_adj])`

The R and SAS calculations accommodate calculating confidence intervals for an age-adjusted rate for a single group or for any number of groups (ie- by sex, race, county):

- *`[v_adj]`, `[rate_adj]`, and `[w_max]` should each be an* n *x* 1 *vector, where* n *is the number of groups. For example, when calculating CI's for female and male rates, the vectors should all be 2 x 1.*

In R: `([v_adj]/(2*[rate_adj]))*qchisq([alpha]/2, 2*([rate_adj]^2/[v_adj]))`

In SAS:
```
([v_adj]/(2#[rate_adj]))#quantile("CHISQ",[alpha]/2,
2#([rate_adj]##2/[v_adj]))
```

$$CI_{upperadj} = \left(\frac{v_{adj} + w_{max}^2}{2(rate_{adj} + w_{max})}\right) * chiinv\left(1 - \frac{\alpha}{2}, \frac{2(rate_{adj} + w_{max})^2}{v_{adj} + w_{max}^2}\right)$$

Where $w_{max} = max\left(\frac{w_i}{population_i}\right)$.

- *Note that `[w_max]` represents a vector of maximum scaled weights:*
  `[multiplier]*([weights]/[population])`
  Where `[weights]` is the 1 x k vector of weights based on the standard population and `[population]` is an n *x* k *matrix, where* n *is the number of groups, and* k *is the number of age groups will produce the `[w_max]` vector*

In Excel:
```
=(([v_adj]+[w_max]^2)/(2*([rate_adj] + [w_max])))*CHISQ.INV(1-[alpha]/2,
                    (2*([rate_adj] + [w_max])^2)/([v_adj] + [w_max]^2))
```

In R:
```
(([v_adj] + [w_max]^2)/(2*([rate_adj] + [w_max])))*qchisq(1-[alpha]/2,
                    (2*([rate_adj] + [w_max])^2/([v_adj] + [w_max]^2)))
```

In SAS:
```
(([v_adj] + [w_max]##2)/(2#([rate_adj] + [w_max])))#quantile("CHISQ",
        1-[alpha]/2, (2#([rate_adj] + [w_max])##2/([v_adj] + [w_max]##2)))
```

Gamma distribution

$$CI_{loweradj} = gammainv\left(\frac{\alpha}{2}, \frac{rate_{adj}^2}{v_{adj}}, \frac{v_{adj}}{rate_{adj}}\right)$$

Note that the 3rd argument, $\frac{v_{adj}}{rate_{adj}}$ is the scale parameter (used by default in Excel and SAS). In R, the qgamma function defaults to using the rate parameter, which is $\frac{1}{scale}$. To correct for this, enter the inverse of the scale parameter as the 3rd argument or use `scale=[v_adj]/[rate_adj]` *to tell qgamma to use the scale parameter.*

In Excel: `GAMMA.INV([alpha]/2, [rate_adj]^2/[v_adj], [v_adj]/[rate_adj])`

In R: `qgamma([alpha]/2, [rate_adj]^2/[v_adj], [rate_adj]/ [v_adj])`

In SAS: `quantile("GAMMA",[alpha]/2, [rate_adj]##2/[v_adj], [v_adj]/[rate_adj])`

$$CI_{upperadj} = gammainv\left(1-\frac{\alpha}{2}, \frac{(rate_{adj}+w_{max})^2}{v_{adj}+w_{max}^2}, \frac{v_{adj}+w_{max}^2}{rate_{adj}+w_{max}}\right)$$

In Excel:
```
=GAMMA.INV(1-[alpha]/2,
      ([rate_adj] + [w_max])^2/([v_adj] + [w_max]^2),
      ([v_adj] + [w_max]^2)/( [rate_adj] + [w_max]))
```

In R:
```
qgamma(1-[alpha]/2,
      ([rate_adj] + [w_max])^2/([v_adj] + [w_max]^2),
      ([rate_adj] + [w_max])/( [v_adj] + [w_max]^2))
```

In SAS:
```
quantile("GAMMA",1-[alpha]/2,
      ([rate_adj] + [w_max])##2/([v_adj] + [w_max]##2),
      ([v_adj] + [w_max]##2)/([rate_adj] + [w_max]))
```

Confidence interval for age-adjusted rates based on counts 100+

For counts 100 and above, the normal approximation using the Wald interval can be used to obtain confidence intervals. Note that these formulas are the same as for crude rates, but the rates and standard errors used in the present formulas have been age-adjusted.

$$CI = rate_{adj} \pm z_{\alpha/2} SE$$

In Excel:
```
=[rate_adj]+NORM.S.INV([alpha]/2)*[SE_adj]
=[rate_adj]+NORM.S.INV(1-[alpha]/2)*[SE_adj]
```

In R:
```
=[rate_adj]+ qnorm([alpha]/2)*[SE_adj]
=[rate_adj]+ qnorm(1-[alpha]/2)*[SE_adj]
```

In SAS:
```
[rate_adj]+ probit([alpha]/2)#[SE_adj]
```

```
[rate_adj]+ probit(1-[alpha]/2)#[SE_adj]
```

About data interpretation and relative standard error

Confidence intervals express the uncertainty around an estimate: this is the range in which, with a specified level of confidence, we can assume contains the true population value. 95% confidence intervals are most commonly used, but confidence levels can be set to other values such as 90% or 99%. A wide confidence interval suggests that there is a lot of uncertainty about an estimate. If a confidence interval is too wide, an estimate may hold little practical value and should be suppressed.

How do we know if a confidence interval is too wide? Some of this will be largely dependent upon the context of the problem, how the data will be used, and how much uncertainty is tolerable. It may also depend upon the value of the estimate. For example, say we have a confidence interval of length 40. This would be an undesirable level of uncertainty if the rate estimate is 5 per 100,000, but may be okay if the rate estimate is 500 per 100,000.

To obtain an estimate of uncertainty relative to the rate estimate, we can obtain the relative standard error using

$$SE_{relative} = \frac{SE}{rate}$$

You can multiply this value by 100 to express it as a percentage. For example, if the SE is 10 and the rate estimate is 500,

$$SE_{relative} = \frac{10}{500} = 0.02 = 2\%.$$

As a general guide, an RSE of 30% or greater has historically been used as a threshold for data suppression due to the instability of the data. PHSD does not currently require suppression based on RSE.

Note: In the past, the National Center for Health Statistics (NCHS) suppressed or flagged estimates as unreliable if the RSE was greater than 23-30% or sample size criteria were unmet, depending on the data type. They now use a combination of sample size (at least 10) and relative confidence interval width (greater than 160%).

Resources

1. PHSD Guidelines for the Release of Public Health Data
2. Age Adjustment Using the 2000 Projected U.S. Population
3. Confidence Intervals for Directly Standardized Rates: A Method Based on the Gamma Distribution (Fay and Feuer, 1997)
4. Rate Algorithms (National Cancer institute)
5. Guidelines for Using Confidence Intervals for Public Health Assessment (Washington State Department of Health)
6. Relative Standard Error (National Center for Health Statistics)
7. Healthy People 2010 Criteria for Data Suppression (National Center for Health Statistics)
8. Data Presentation Standards for Rates and Counts (National Center for Health Statistics)