



Håkon K. Gjessing

Professor/Principal Investigator

⇒ Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo
– a Norwegian Research Council Centre of Excellence

⇒ Department of Global Public Health and Primary Care, University of Bergen



Kathmandu
Tuesday 6 October 2026



Comparing mortality rates in two populations

pop	pop.size	deaths
C1	2e+06	26666
C2	3e+06	33333

(Remember: 2e+06 means 2 000 000, etc.)

Population-specific rates:

$$R_1 = \frac{26666}{2000000} = \frac{1}{75} = 0.01333$$

$$R_2 = \frac{33333}{3000000} = \frac{1}{90} = 0.01111$$

(... often multiplied by, say, 100 000, but let's wait with that.
That's just a matter of presentation)

Comparing mortality rates in two populations

Breaking down on age

Data, broken down by age:

pop	age	pop.size	deaths
C1	A1	1e+06	6666
C1	A2	1e+06	20000
C2	A1	2e+06	13333
C2	A2	1e+06	20000

Population and age-specific observed rates:

$$R_{C1,A1} = \frac{6666}{1000000} = \frac{1}{150} = 0.00666$$

$$R_{C1,A2} = \frac{20000}{1000000} = \frac{1}{50} = 0.02$$

$$R_{C2,A1} = \frac{13333}{2000000} = \frac{1}{150} = 0.00666$$

$$R_{C2,A2} = \frac{20000}{1000000} = \frac{1}{50} = 0.02$$

Age-specific rates are actually identical in the two populations!

Comparing mortality rates in two populations

Breaking down on age

Data, broken down by age, with observed rates

pop	age	pop.size	deaths	rates.obs
C1	A1	1e+06	6666	0.00666
C1	A2	1e+06	20000	0.02
C2	A1	2e+06	13333	0.00666
C2	A2	1e+06	20000	0.02

Comparing mortality rates in two populations

Age standardization!

- The age-specific rates tell the story
- In a sense, this is “enough”
- HOWEVER, age-specific rates are typically different between the two populations
- AND it would be nice to see a single summary rate for each population
- We would like to *standardize* the two rates, correcting for the difference in age distributions

Direct standardization:

- We pick a common “standard” age distribution for the two groups
- We then “apply” the population-specific age-related mortality rates to the common age distribution
- And then average over age groups

Comparing mortality rates in two populations

Age standardization!



The “standard” age distribution:

- From external population distribution
- ... or, maybe better, the total (marginal) distribution you already have

pop	age	pop.size	deaths	rates.obs
C1	A1	1e+06	6666	0.00666
C1	A2	1e+06	20000	0.02
C2	A1	2e+06	13333	0.00666
C2	A2	1e+06	20000	0.02

Marginal age distribution:

age	pop.size.marg
A1	3e+06
A2	2e+06

(Called “marginal” since they are the column sums in a two-by-two CxA table)

Comparing mortality rates in two populations

Age standardization!



The “standard” age distribution:

- From external population distribution
- ... or, maybe better, the total (marginal) distribution you already have

pop	age	pop.size	deaths	rates.obs	pop.size.marg
C1	A1	1e+06	6666	0.00666	3e+06
C1	A2	1e+06	20000	0.02	2e+06
C2	A1	2e+06	13333	0.00666	3e+06
C2	A2	1e+06	20000	0.02	2e+06

Age-standardized rates:

$$R_{std1} = \frac{0.00666 \times 3000000 + 0.02 \times 2000000}{3000000 + 2000000} = 0.012$$
$$R_{std2} = \frac{0.00666 \times 3000000 + 0.02 \times 2000000}{3000000 + 2000000} = 0.012$$

Comparing mortality rates in two populations

Age standardization!

Age-standardized rates:

$$R_{std1} = \frac{0.00666 \times 3000000 + 0.02 \times 2000000}{3000000 + 2000000} = 0.012$$

$$R_{std2} = \frac{0.00666 \times 3000000 + 0.02 \times 2000000}{3000000 + 2000000} = 0.012$$

NOTE:

- It is only in this example the age-specific rates are the same in the two populations (0.00666 and 0.02 in both)
- Usually, those will differ across populations
- But *the same standard population* should always be applied to the two (here: 3000000 and 2000000)

Can we do all this (and more) with Poisson regression??

Comparing mortality rates in two populations

Computing from Poisson

Population-specific rates:

$$R_1 = \frac{26666}{2000000} = \frac{1}{75} = 0.01333$$

$$R_2 = \frac{33333}{3000000} = \frac{1}{90} = 0.01111$$

Add log.pop.size to data file:

```
.data.c
  pop pop.size deaths log.pop.size
1  C1    2e+06  26666    14.50866
2  C2    3e+06  33333    14.91412
```

Comparing mortality rates in two populations

Computing from Poisson

Population-specific rates:

$$R_1 = \frac{26666}{2000000} = \frac{1}{75} = 0.01333$$

$$R_2 = \frac{33333}{3000000} = \frac{1}{90} = 0.01111$$

Using software:

```
glm(deaths ~ offset(log.pop.size) + pop, family = poisson,  
    data = .data.c)
```

Exponentiated coefficients:

(Intercept)	popC2
0.0133330	0.8333458

And the product is $0.0133330 \times 0.8333458 = 0.01111$

... but we can do better...

Set up a prediction data set:

pop	pop.size	log.pop.size
C1	1	0
C2	1	0

Then, use the glm model to predict on the *new* data:

pop	pop.size	log.pop.size	pred
C1	1	0	0.013333
C2	1	0	0.011111

Comparing mortality rates in two populations

Predicting with Poisson

... and if we want rates per 100 000, say...

Set up a prediction data set and add prediction from glm:

pop	pop.size	log.pop.size	pred
C1	1e+05	11.51293	1333.3
C2	1e+05	11.51293	1111.1

Comparing mortality rates in two populations

Predicting with Poisson

... and check with the original...

Set up a prediction data set and add prediction from glm:

pop	pop.size	log.pop.size	pred
C1	2e+06	14.50866	26666
C2	3e+06	14.91412	33333

NOTE: With a *saturated model* the prediction will always recover the observed numbers here

Comparing mortality rates in two populations

Predicting with Poisson

Data file, broken down by age:

```
.data
  pop age pop.size deaths log.pop.size
1  C1  A1      1e+06   6666      13.81551
2  C1  A2      1e+06  20000      13.81551
3  C2  A1      2e+06  13333      14.50866
4  C2  A2      1e+06  20000      13.81551
```

Using software, *wait with age*:

```
glm(deaths ~ offset(log.pop.size) + pop, family = poisson, data = .data)
```

Exponentiated coefficients:

```
(Intercept)      popC2
  0.0133330    0.8333458
```

And the product is $0.0133330 \times 0.8333458 = 0.01111$

Just as before!

Comparing mortality rates in two populations

Predicting with Poisson



Data file, broken down by age:

```
.data
  pop age pop.size deaths log.pop.size
1  C1  A1      1e+06   6666      13.81551
2  C1  A2      1e+06  20000      13.81551
3  C2  A1      2e+06  13333      14.50866
4  C2  A2      1e+06  20000      13.81551
```

Using software, *adjusting for age*:

```
glm(deaths ~ offset(log.pop.size) + pop + age, family = poisson, data = .data)
```

Exponentiated coefficients:

(Intercept)	popC2	ageA2
0.006666231	1.00	3.00

Comparing mortality rates in two populations

Predicting with Poisson



Note that the exponentiated coefficients

(Intercept)	popC2	ageA2
0.006666231	1.00	3.00

show that there is no difference between the two groups.
(But a threefold difference by age group)

Set up prediction data, and predict using the glm model:

pop	age	pop.size	log.pop.size	pred
C1	A1	1	0	0.006666
C1	A2	1	0	0.020000
C2	A1	1	0	0.006666
C2	A2	1	0	0.020000

Comparing mortality rates in two populations

Adjusting for age

Population and age-specific rates:

pop	age	pop.size	log.pop.size	pred
C1	A1	1	0	0.006666
C1	A2	1	0	0.020000
C2	A1	1	0	0.006666
C2	A2	1	0	0.020000

And checking the products:

(Intercept)	popC2	ageA2
0.006666231	1.00	3.00

$$\begin{aligned}0.006666231 &= 0.006666231 \\0.006666231 \times 3.00 &= 0.02 \\0.006666231 \times 1.00 &= 0.006666231 \\0.006666231 \times 1.00 \times 3.00 &= 0.02\end{aligned}$$

Note:

- The rate multiplications are of exactly the same type as we did in the interaction session
- Here, we see *no rate difference* between the populations after adjusting for age (RR for popC2 is 1)

Warning:

- The model predicted rates fit perfectly to the observed rates.
- We have fitted a log-additive model with three parameters (to four cells), and there is no sign of an interaction (on log scale)
- This is because this particular example was constructed *without interactions*
- In reality, there can be different age-specific mortality rates in the two populations

But let's check with an interaction in the model!

Comparing mortality rates in two populations

Adjusting for age

Using software, *adjusting for age*, checking for interaction:

```
glm(deaths ~ offset(log.pop.size) + pop + age + pop:age,
    family = poisson, data = .data)
```

Result (log scale):

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.010735	0.012248	-409.104	<2e-16	***
popC2	0.000075	0.015001	0.005	0.996	
ageA2	1.098712	0.014143	77.688	<2e-16	***
popC2:ageA2	-0.000075	0.018028	-0.004	0.997	

- No difference between populations
- No significant interaction

Comparing mortality rates in two populations

Adjusting for age

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.010735	0.012248	-409.104	<2e-16	***
popC2	0.000075	0.015001	0.005	0.996	
ageA2	1.098712	0.014143	77.688	<2e-16	***
popC2:ageA2	-0.000075	0.018028	-0.004	0.997	

- No difference between populations
- No significant interaction

What we have learned:

- The population rates are equal after adjusting for age
- The age-specific rates are the same in the two populations (here by construction)

Comparing mortality rates in two populations

Age standardization!

- The exponentiated estimate $RR = 1$ tells you the rate ratio between the two populations
- In a sense, this is “enough”
- BUT AGAIN, it would be nice to see the *actual rates* rather than just the ratio between them
- We would like to *standardize* the two rates, correcting for the difference in age distributions

Direct standardization:

- We pick a common “standard” age distribution for the two groups
- We then “apply” the population-specific age-related mortality rates to the common age distribution
- And then average over age groups

Comparing mortality rates in two populations

Age standardization!

Using software, *adjusting for age*, including interaction:

```
glm(deaths ~ offset(log.pop.size) + pop + age + pop:age,  
    family = poisson, data = .data)
```

Predict on new data:

	pop	age	pop.size	log.pop.size	pred
1	C1	A1	1	0	0.006666
2	C1	A2	1	0	0.020000
3	C2	A1	1	0	0.006666
4	C2	A2	1	0	0.020000

- From here, the standardization can be done as before

Step-by-Step

- Get the variables `count` (here deaths), `age` (preferably in groups), `group` (here C1 and C2), and population size `pop.size` in all subgroups.
- Run the model (with interaction `group x age` included, and offset)
- Predict on new data with `pop.size = 1`
- “Apply” (multiply) the predicted rates with a standard population
- Sum up over age groups!

Or even better: use a ready-made package

- `marginalmeans` seems to be just right for our purpose
- `stdReg2`, `emmeans`, `Epi`, `popEpi`, `epitools`, and `epiR` may also be worth exploring.

Using marginalmeans

```
.data
  pop age pop.size deaths log.pop.size
  C1  A1    1e+06   6666    13.81551
  C1  A2    1e+06  20000    13.81551
  C2  A1    2e+06  13333    14.50866
  C2  A2    1e+06  20000    13.81551
```

```
glm.result <- glm(deaths ~ offset(log.pop.size) + pop + age + pop:age, family = poisson,
  data = .data)
```

```
newdata <- .data
newdata$pop.size <- newdata$deaths <- NULL # remove pop.size and deaths, for clarity
newdata$log.pop.size <- 0 # set population size to 1 (or, for instance, use log(1e05))
```

```
pop age log.pop.size
C1  A1          0
C1  A2          0
C2  A1          0
C2  A2          0
```

Using `marginalmeans`

Compute marginal frequencies of age, add to data frame `newdata`:

```
age.freq.all <- tapply(.data$pop.size, .data$age, sum))
```

```
age.freq.all
```

```
  A1    A2
```

```
3e+06 2e+06
```

```
newdata$w <- age.freq.all[newdata$age]
```

```
pop age log.pop.size      w
```

```
1  C1  A1           0 3e+06
```

```
2  C1  A2           0 2e+06
```

```
3  C2  A1           0 3e+06
```

```
4  C2  A2           0 2e+06
```

Using marginalmeans

```
library(marginalmeans)
```

```
avg_pred <- avg_predictions(glm.result, newdata = newdata, by = "pop", wts = "w")
```

pop	Estimate	Std. Error	z	Pr(> z)	S	2.5 %	97.5 %
C1	0.012	7.48e-05	160	<0.001	Inf	0.0119	0.0121
C2	0.012	6.63e-05	181	<0.001	Inf	0.0119	0.0121

D is outcome, for example death. A is age. The two groups are “C1” and “C2”.

Law of total probability:

Population-specific mortality rates

$$P(D|C1) = \sum_A P(D|A, C1) \times P(A|C1)$$

$$P(D|C2) = \sum_A P(D|A, C2) \times P(A|C2)$$

Recall that

$$\sum_A$$

signifies the sum over all age groups.

Formulas? Direct standardization

Direct standardization replaces the population-specific age distributions with one common distribution. That is,

$$P(A|C1) \text{ and } P(A|C2) \text{ replaced by } P(A|C)$$

where C is a “standard” population.

Directly standardized mortality rates

$$P(D|C1) = \sum_A P(D|A, C1) \times P(A|C)$$

$$P(D|C2) = \sum_A P(D|A, C2) \times P(A|C)$$

Formulas? Standardization

NOTE:

- Indirect standardization is used if target population is too small to calculate population and age-specific rates
- BUT an advantage with model-based standardization is that a model can be used for

$$P(D|A, C1) \text{ and } P(D|A, C2),$$

thus possibly obtaining more stable estimates

- The usual standardizations assume an interaction between age A and population C when computing rates. This may not always be necessary.

Formulas? Extensions

NOTE:

- Model-based standardization makes it easier to add more standardization variables (in addition to age), such as sex.
- The outcome does not have to be dichotomous (D). The formulas also work for continuous outcomes, for instance.
- The formulas can be used for more than two groups to be compared