Estimating the distribution of mortality risk in a population of newborns

Jonas Schöley

2018-03-16

In this research note I demonstrate various ways to estimate the distribution of mortality risk in a population from individual level data. I'll cover the techniques in order of increasing modelling complexity, from purely descriptive to a full Bayesian probability model. ## Data I use the "NCHS Cohort Linked Birth – Infant Death Data Files" from the *National Center for Health Statistics*[1]. The data set is a complete census of births and infant deaths on the territory of the United States (without its overseas territories) and features most fields present on the birth and death certificates. The size and detail of the data allows to produce mortality estimates for thousands of sub-populations, and thus the quantification of heterogeneity in perinatal mortality.

The following strata contribute information about the distribution of mortality risk in the population of newborns:

- birthweight
- gestation at delivery
- 5 minute apgar score
- presence and severity of congenital anomalies
- plurality

In order to increase the sample size I pool all data from the birth cohorts 2005 to 2010.

```
library(tidyverse)
# set available memory to 100GB (important only on windows)
memory.limit(size = 100000)
```

```
## [1] 1e+05
```

```
# load microdata on births and infant deaths
load('../priv/data/02-harmonized/2017-10-29-ideath.RData')

# basic data pooling, selection and recoding
ideath_sub <-
  ideath %>%
  # subset to study period
  filter(date_of_delivery_y %in% 2005:2010) %>%
  # add variables for survival analysis
  mutate(
    # age at death in (completed) hours
    # we integrate additional information
    # available for the day of birth
    age_at_death_h =
      ifelse(age_at_death_d > 0,  # if death not at first day
             age_at_death_d*24, # convert age in days to hours
             # otherwise check if death happened in first hour
             # or hour 1-23 and code accordingly
             ifelse(age_at_death_c == '1-23 hours', 1, 0)),
    # so now we have interval censored data on the age at death in hours,
```

[1] https://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm

```
    # to deal with it we add the width of the age interval of death [x, x+nx)
    age_at_death_h_width = 24,
    age_at_death_h_width = ifelse(age_at_death_h == 0, 1, age_at_death_h_width),
    age_at_death_h_width = ifelse(age_at_death_h == 1, 23, age_at_death_h_width),
    # perinatal death indicator, i.e. death during first hour of life
    death = age_at_death_h == 0, death = ifelse(is.na(death), FALSE, death),
    # right censoring at hour 1
    # age at death or censoring in (completed) hours
    survtime_h = ifelse(death == TRUE, 0, 1),
    survtime_h_width = ifelse(death == TRUE, 1, 0)
  ) %>%
  select(
    # basic survival information
    death, survtime_h, survtime_h_width,
    # clinical variables
    birthweight_c5, gestation_at_delivery_c4, apgar5,
    congenital_anomalies_c3, plurality_c2
  )
rm(ideath)
```

Life-table function

```
# Calculate Life-tables From Individual Level Survival Times
#
# Individual level survival times may be interval or right censored.
# The life-table can be arbitrarily abridged.
#
# @param df a data frame
# @param x time until death or censoring
# @param nx width of interval [x, x+nx), i.e. precision of measurement
# @param death death (TRUE) or censored (FALSE)
# @param cuts cutpoint for age intervals in abridged life-table [x1, ..., xn)
# @param ... grouping variable
GetLifeTable <- function (df, x, nx, death, cuts, ...) {
  x_i = enquo(x); nx_i = enquo(nx); death_i = enquo(death); strata = quos(...)

  FindIntervalStart <- function (x, breaks) {
    breaks[.bincode(x = x, breaks = breaks, right = FALSE, include.lowest = FALSE)]
  }

  lt <-
    df %>%
    # aggregation into pre-defined age-groups
    mutate(x0 = FindIntervalStart(!!x_i, cuts)) %>%
    group_by(..., x0, add = FALSE) %>%
    summarise(
      nDx = sum(!!death_i),
      nCx = sum(!(!!death_i)),
      # average time spent in interval for
      # those who leave during interval (by death or censoring)
      nax = mean(!!x_i) - first(x0) + 0.5*mean(!!nx_i)
    ) %>%
```

```
    arrange(..., x0) %>%
    group_by(..., add = FALSE) %>%
    # calculation of life-table columns
    mutate(
      nx = c(diff(x0), last(cuts)-last(x0)),
      # assuming no late entry
      Nx = head(cumsum(c(sum(nDx, nCx), -(nCx+nDx))), -1),
      # distribution of censoring or death
      nfx = (nCx + nDx) / sum(nCx + nDx),
      # life-table
      nqx = nDx/Nx,
      lx = Nx/first(Nx),
      ndx = lx*nqx,
      nLx = lead(lx, n = 1, default = NA) * nx + (nfx*nax),
      nLx = ifelse(is.na(nLx), last(nax), nLx),
      nmx = ndx / nLx
    ) %>% ungroup() %>%
    mutate(id = group_indices(., ...)) %>%
    # fill 'gaps' in life-table
    complete(x0 = head(cuts, -1), distinct(., id, !!!strata),
             fill = list(nDx = 0, nCx = 0, ndx = 0, nmx = 0, nqx = 0, nax = 0)) %>%
    arrange(id, x0) %>%
    group_by(id, add = FALSE) %>%
    mutate(
      nx = c(diff(x0), last(cuts)-last(x0)),
      Nx = head(cumsum(c(sum(nDx, nCx), -(nCx+nDx))), -1),
      nEx = (Nx-nDx)*nx + nax*(nDx+nCx),
      nEx = ifelse(is.infinite(nEx), nax*(nDx+nCx), nEx),
      lx = Nx/first(Nx),
      nLx = ifelse(is.na(nLx), lx*nx, nLx),
      Tx = rev(cumsum(rev(nLx))), ex = Tx/lx
    ) %>% ungroup() %>%
    select(id, ..., x = x0, nx, Nx, nEx, nDx, nCx,
           lx, ndx, nqx, nax, nmx, nLx, Tx, ex)

  return(lt)

}
```

Observed population mortality rate

I calculate the observed population mortality rate as

$$\bar{\mu} = \frac{\text{Total number of deaths during first hour of life}}{\text{Total population exposure during first hour of life}}.$$

```
obs_pop_nmx <-
  GetLifeTable(ideath_sub, x = survtime_h, nx = survtime_h_width, death = death,
               cuts = c(0, 1, Inf)) %>%
  filter(is.finite(nx)) %>%
  pull(nmx)
obs_pop_nmx
```

```
## [1] 0.0009423652
```

Observed stratified mortality rates

For each sub-population z I calculate the observed mortality rates during the first hour of life as

$$\mu_z = \frac{\text{Number of deaths during first hour of life in stratum } z}{\text{Total exposure during first hour of life in stratum } z}.$$

The relative exposure in stratum z is

$$p_z = \frac{\text{Total exposure during first hour of life in stratum } z}{Total population exposure during first hour of life}.$$

```
lt_strat <-
  GetLifeTable(ideath_sub, x = survtime_h, nx = survtime_h_width, death = death,
               cuts = c(0, 1, Inf),
               birthweight_c5, gestation_at_delivery_c4, apgar5,
               congenital_anomalies_c3, plurality_c2) %>%
  filter(is.finite(nx)) %>%
  select(birthweight_c5, gestation_at_delivery_c4, apgar5,
         congenital_anomalies_c3, plurality_c2,
         x, nx, Nx, nEx, nDx, nCx, nax, nmx) %>%
  mutate(px = nEx/sum(nEx))
```

The overall population mortality rate should be equal to the exposure-weighted average mortality rates for the sub-populations.

```
sum(lt_strat$nmx * lt_strat$px) / obs_pop_nmx
```

```
## [1] 1
```

The distribution of frailties

Frailty is a measure of *relative risk* in a population. For a sub-population z I define frailty as

$$\zeta_z = \frac{\mu_z}{\bar{\mu}},$$

i.e. frailty is the stratum specific mortality rate scaled by the population mortality rate. Therefore modelling the distribution of mortality risk is equal to modelling the distribution of frailties scaled by $\bar{\mu}^{-1}$.

Empirical distribution functions

How is the *observed* risk of death, as defined by the ratio of deaths to exposures, distributed in a population? The empirical *distribution* of the risk is given by the pair (μ_z, p_z), where μ_z is the observed mortality rate in stratum z and p_z the proportion of the total population exposure contributed by the individuals that stratum, i.e. p_z expresses how common the risk μ_z is in the total population. As $\sum_z p_z = 1$ I can define a probability mass function (PMF)

$$f_{mu_z} = p_z.$$

In a case where two or more population strata have identical rates I collapse the equal rates into single rate with a probability mass equal to the sum of their exposures.

Some summary statistics

Most summary statistics have to be calculated with population exposures (p_z) as *weights* as the mortality rates constitute observations on variably sized groups of individuals.

```
range(lt_strat$nmx)
```

```
## [1] 0 2
```

```
with(lt_strat, Hmisc::wtd.mean(nmx, px))
```

```
## [1] 0.0009423652
```

```
with(lt_strat, Hmisc::wtd.var(nmx, px, method = 'ML'))
```

```
## [1] 0.0003489242
```

```
with(lt_strat, Hmisc::wtd.quantile(nmx, px, type = 'i/n'))
```

```
##           0%          25%          50%          75%         100%
## 0.000000e+00 2.835353e-06 3.463493e-06 4.554330e-06 2.000000e+00
```

From the summary statistics it is obvious that the data is *spread over multiple magnitudes*, but with more than 75% of the population experiencing a risk of less than 10 deaths per 1 million person hours of exposure ($\mu_{q.75} = 4.5e-6$). The average mortality rate $\bar{\mu} = 9.4e-4$ is more than 100 times higher than the median mortality ($\mu_{q.75} = 4.5e-6$), pointing towards a distribution of risk that is *extremely* skewed to the right and therefore best analyzed on a *log-scale*.

Dealing with 0 deaths

There are quite a few cases with *0 observed deaths* (59.8% of all sub-groups or 2.5% of the total population of birth feature zero deaths). The maximum likelihood estimate for the death rate is 0 in such cases, such is the ML-estimate for the probabililty of death. Because zero-risk is implausible and can't be modelled on a log-scale I impute the zero mortality rates by an alternative estimator for the risk of death. Quigley & Revie (2011) propose $\hat{q} \approx \frac{1}{2.5n}$ as a point estimate for the probability of death when 0 events out of n trials have been observed.

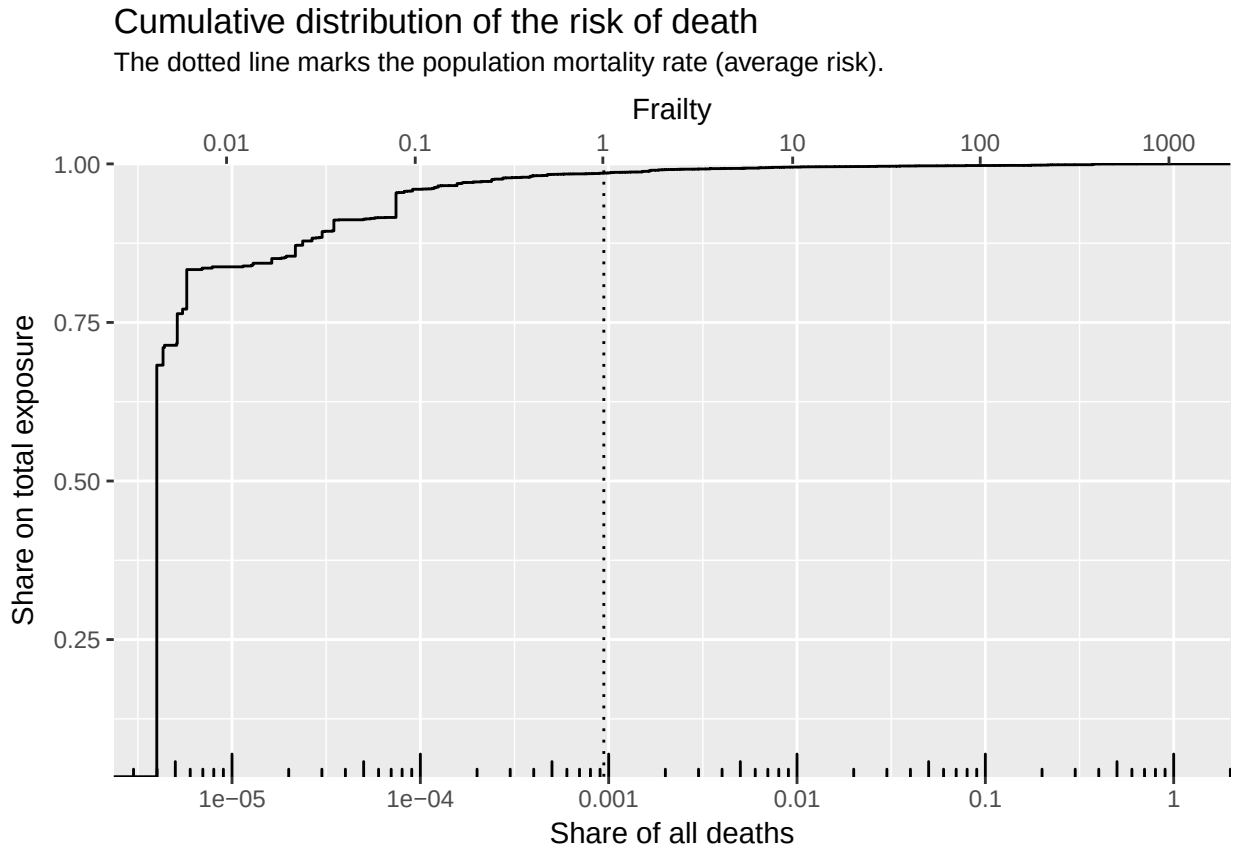
```
# total observed exposure

obs_tot_exp <- sum(lt_strat$nEx)
# empirical distribution of observed rates
empdist_obs_rates <-
  lt_strat %>%
  # impute 0 rates with alternative estimate
  mutate(
    # flag 0s
    is_zero = ifelse(nmx == 0, TRUE, FALSE),
    # use minimax estimator for nqx given 0 deaths
    # and convert back to rate
    nmx = ifelse(is_zero, -log(1-1/(2.5*Nx))/nx, nmx)
  ) %>%
  # collapse equal mortality rates into single rate with
  # probability mass f_i equal to the sum of their exposures
  group_by(nmx) %>%
  summarise(f_i = sum(nEx)/obs_tot_exp) %>%
```

```
  # add other other characterisations of the distribution of rates
  arrange(nmx) %>%
  mutate(
    # rank
    i = 1:n(),
    # cumulative distribution function
    F_i = cumsum(f_i)
  )
empdist_obs_rates
```

```
## # A tibble: 561 x 4
##           nmx     f_i     i    F_i
##         <dbl>   <dbl> <int>  <dbl>
##  1 0.00000236 0.0338      1 0.0338
##  2 0.00000399 0.649       2 0.683
##  3 0.00000430 0.0277      3 0.710
##  4 0.00000438 0.00363     4 0.714
##  5 0.00000509 0.00313     5 0.717
##  6 0.00000512 0.0466      6 0.764
##  7 0.00000546 0.00728     7 0.771
##  8 0.00000575 0.0623      8 0.833
##  9 0.00000693 0.00230     9 0.836
## 10 0.00000784 0.00203    10 0.838
## # ... with 551 more rows
```

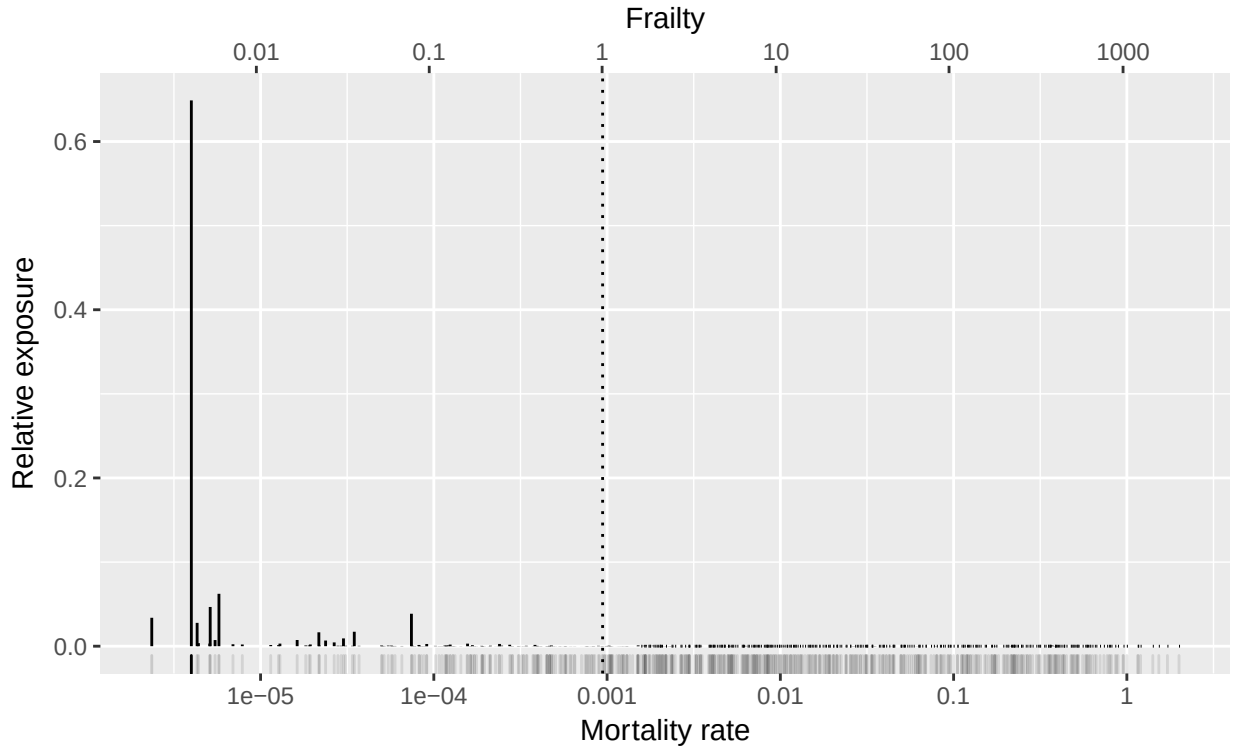
The cumulative distribution function.

```
ggplot(empdist_obs_rates, aes(x = log10(nmx), y = F_i)) +
  geom_step() +
  scale_x_continuous(breaks = log10(10^(-6:1)), labels = 10^(-6:1),
                     sec.axis = sec_axis(~.-log10(obs_pop_nmx),
                                         breaks = log10(10^(-2:10)),
                                         labels = 10^(-2:10),
                                         name = 'Frailty')) +
  geom_vline(xintercept = log10(obs_pop_nmx), lty = 'dotted') +
  annotation_logticks(sides = 'b') +
  coord_cartesian(expand = FALSE) +
  labs(title = 'Cumulative distribution of the risk of death',
       subtitle = 'The dotted line marks the population mortality rate (average risk).',
       x = 'Share of all deaths', y = 'Share on total exposure')
```



The density function.

```
ggplot(empdist_obs_rates) +
  geom_segment(aes(x = log10(nmx), y = 0, xend = log10(nmx), yend = f_i)) +
  scale_x_continuous(breaks = log10(10^(-6:1)), labels = 10^(-6:1),
                     sec.axis = sec_axis(~.-log10(obs_pop_nmx),
                                         breaks = log10(10^(-2:10)),
                                         labels = 10^(-2:10),
                                         name = 'Frailty')) +
  geom_vline(xintercept = log10(obs_pop_nmx), lty = 'dotted') +
  geom_rug(aes(x = log10(nmx), alpha = f_i), show.legend = FALSE) +
  coord_cartesian(expand = TRUE) +
  labs(title = 'Density distribution of the risk of death',
       subtitle = 'The dotted line marks the population mortality rate (average risk).',
       x = 'Mortality rate', y = 'Relative exposure')
```



Instead of estimating the distribution of mortality rates one can also look at the *concentration* of deaths over quantiles of a birth cohort. The *Lorenz curve* – commonly used to show the distribution of income in a population – can also be used to show how (un-)equal deaths are distributed in a cohort of newborns.

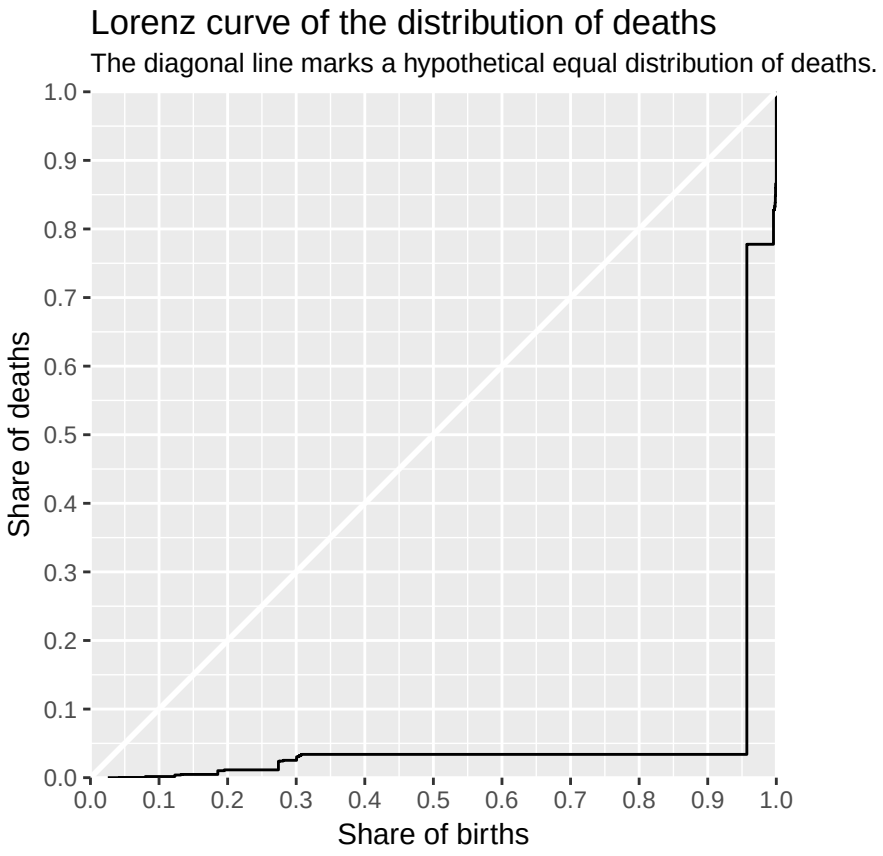
```
# empirical distribution of observed deaths
total_population_size <- sum(lt_strat$Nx)
empdist_obs_deaths <-
  lt_strat %>%
  # collapse equal death counts into single
  # probability mass f_i equal to the sum of their population shares
  group_by(nDx) %>%
  summarise(f_i = sum(Nx)/total_population_size) %>%
  # add other other characterisations of the distribution of deaths
  arrange(nDx) %>%
  mutate(
    # cumulative distribution function
    F_i = cumsum(f_i),
    # Lorenz function
    S_i = cumsum(f_i*nDx),
    L_i = S_i/last(S_i)
  )
empdist_obs_deaths
```

```
## # A tibble: 78 x 5
##       nDx      f_i    F_i    S_i      L_i
##     <dbl>    <dbl>  <dbl>  <dbl>    <dbl>
##  1    0. 0.0253   0.0253 0.     0.
```

```
##  2    1. 0.0173   0.0425 0.0173 0.000304
##  3    2. 0.0374   0.0800 0.0921 0.00162
##  4    3. 0.0430   0.123  0.221  0.00390
##  5    4. 0.00890  0.132  0.257  0.00453
##  6    5. 0.00207  0.134  0.267  0.00471
##  7    6. 0.0517   0.186  0.577  0.0102
##  8    7. 0.00953  0.195  0.644  0.0113
##  9    8. 0.000121 0.195  0.645  0.0114
## 10    9. 0.0788   0.274  1.35   0.0239
## # ... with 68 more rows
```

The curve shows that less than 5% of the births in a cohort of newborns result in more that 95% of the deaths. The Lorenz curve is a characterization of a probability distribution that shows how unequal a trait (in our case, the risk of death) is distributed in a population.

```
ggplot(empdist_obs_deaths, aes(x = F_i, y = L_i)) +
  geom_step() +
  coord_fixed(xlim = 0:1, ylim = 0:1, expand = FALSE) +
  scale_x_continuous(breaks = seq(0, 1, 0.1)) +
  scale_y_continuous(breaks = seq(0, 1, 0.1)) +
  geom_abline(color = 'white', lwd = 1) +
  labs(title = 'Lorenz curve of the distribution of deaths',
       subtitle = 'The diagonal line marks a hypothetical equal distribution of deaths.',
       x = 'Share of births', y = 'Share of deaths')
```



Spline-based density estimate

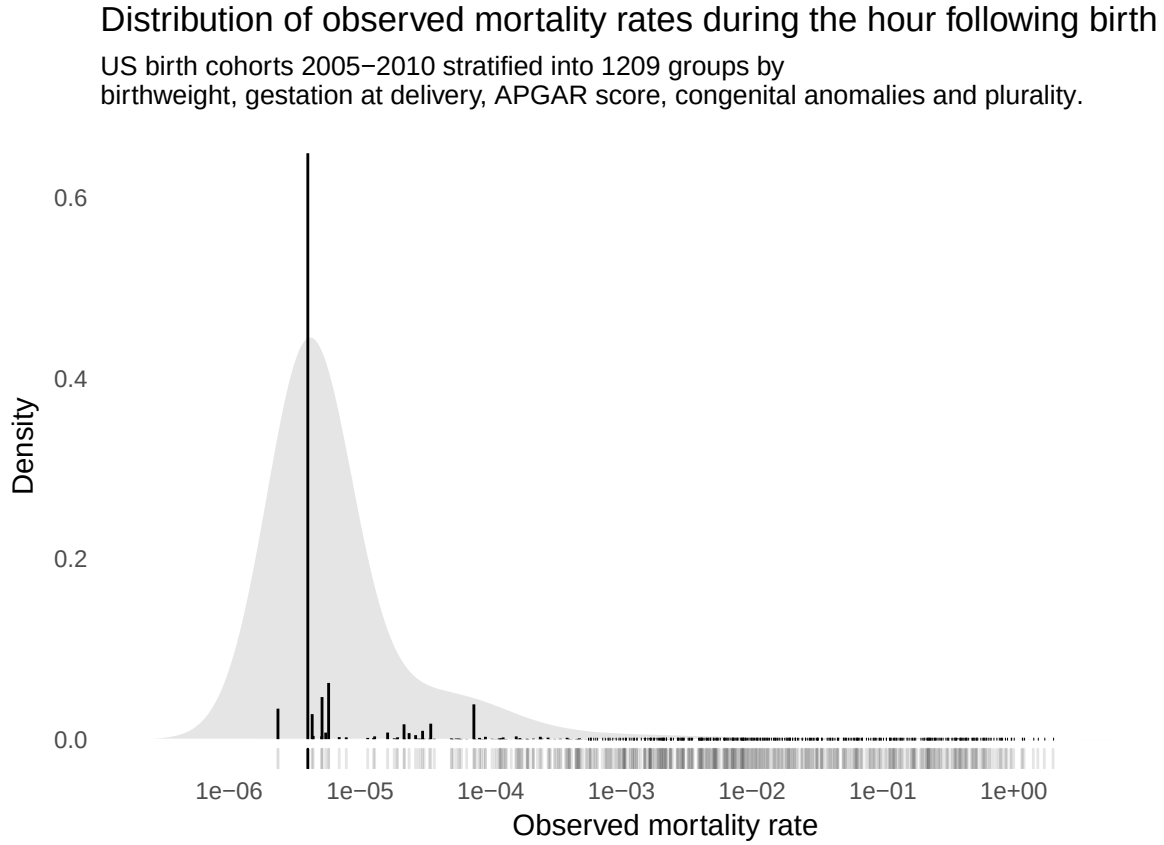
Kernel-density estimate

The Kernel Density Estimator (KDE) is a non-parametric estimator of a random variables probability density. Given observed mortality rates μ_z and observed exposures n_z I use the KDE to estimate a smooth density distribution of mortality $\hat{f}_{\text{KDE}}(\mu_z)$.

In order to facilitate the KDE fit I model the distribution of the log transformed mortality instead. The distribution of the μ_z is then recovered by evaluating $\hat{f}_{\text{KDE}}(x = log(\mu_z))$ at $\exp(x)$.

I fit the KDE to an empirical probability mass function (PMF) of observed, 0-adjusted mortality rates. To attain the PMF I aggregate multiple identical observed mortality rates into a single point-mass at that rate with the mass being the sum of the exposure times associated with each of the identcal rates. I then divide the masses by the total exposure time of the population to attain a distribution of probabilities. Thus the data used for fitting the KDE will consist of a vector of unique mortality rates and their corresponding weights given by the relative exposure.

```
# kernel density estimate of distribution of observed rates
kde_rates <- as.data.frame(with(empdist_obs_rates,
                                density(log(nmx), weights = f_i),
                                kernel = 'gaussian')[c('x', 'y')])

# plot KDE versus observed frequency distribution of rates
ggplot(empdist_obs_rates) +
  geom_polygon(data = kde_rates, aes(x = exp(x), y = y),
               color = NA, fill = 'grey90') +
  geom_segment(aes(x = nmx, xend = nmx, y = 0, yend = f_i)) +
  geom_rug(aes(x = nmx, alpha = f_i), sides = 'b', show.legend = FALSE) +
  scale_x_continuous(trans = 'log10', breaks = 10^(-8:0)) +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(y = 'Density', x = 'Observed mortality rate',
       title = 'Distribution of observed mortality rates during the hour following birth',
       subtitle = 'US birth cohorts 2005-2010 stratified into 1209 groups by\nbirthweight, gestation at
```



- Quigley, J., & Revie, M. (2011). Estimating the Probability of Rare Events: Addressing Zero Failure Data. Risk Analysis, 31(7), 1120-1132. https://doi.org/10.1111/j.1539-6924.2010.01568.x

Summary statistics

If X is random variable "log-mortality" with distribution $\hat{f}_{\text{KDE}}(x = log(\mu_z))$, then $\text{E}[\exp(X)]$ is the estimated average population mortality. Via the *Law of the unconscious statistician* we have $\text{E}[exp(X)] = \int_x exp(X = x)\hat{f}_{\text{KDE}}(x)\ \text{d}x$. Numeric integration yields the result.

```
kde_mean <-
  integrate(
    approxfun(x = kde_rates$x, y = exp(kde_rates$x)*kde_rates$y),
    lower = min(kde_rates$x), upper = max(kde_rates$x)
  )$value
kde_mean
```

```
## [1] 0.001257112
```

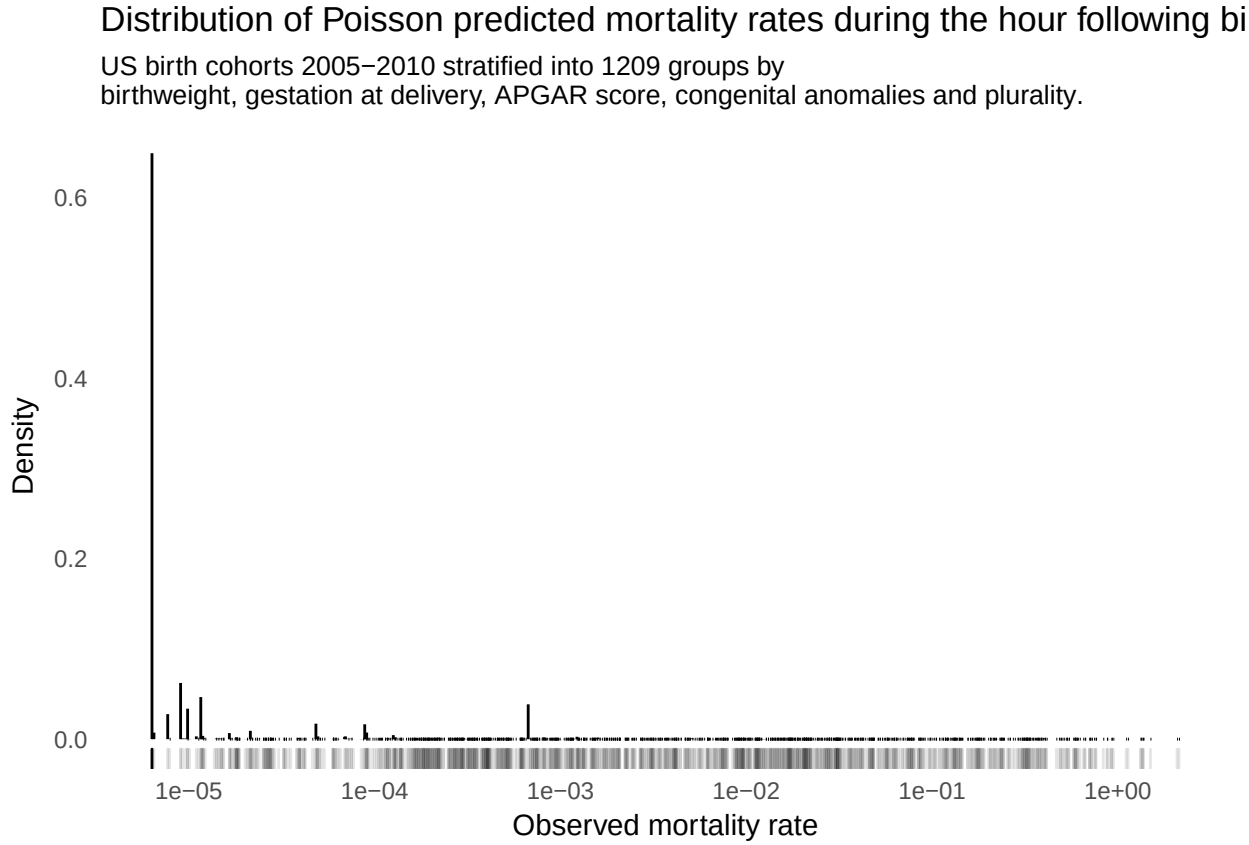
It is convenient to calculate the variance of the estimated risk distribution via the identity $\text{Var}(Y) = \text{E}(Y^2) - \text{E}(Y)^2$, with $Y = \exp(X)$ and $X\ \hat{f}_{\text{KDE}}(x = log(\mu_z))$. Therefore we have $\text{Var}[\exp(X)] = \text{E}[\exp(X)^2] - \text{E}[exp(X)]^2$.

```
kde_variance <-
  integrate(
    approxfun(x = kde_rates$x, y = exp(kde_rates$x)^2*kde_rates$y),
    lower = min(kde_rates$x), upper = max(kde_rates$x)
  )$value - kde_mean^2
kde_variance
```

```
## [1] 0.001050066
```

Regression estimation

```
rate_fit <-
  glm(nDx~
        birthweight_c5+
        gestation_at_delivery_c4+
        apgar5+
        congenital_anomalies_c3+
        plurality_c2+
        offset(log(nEx)),
      data = mutate_at(lt_strat, vars(birthweight_c5, gestation_at_delivery_c4,
                                      apgar5, congenital_anomalies_c3,
                                      plurality_c2),
                       funs(ifelse(is.na(.), 'unknown', .))),
      family = poisson(link = 'log'))

lt_strat$pois_rate <- predict(rate_fit, type = 'response')/lt_strat$nEx

# Poisson predicted weighted unique rates
pois_weighted_unique_rates <-
  lt_strat %>%
  # sum up the share for equal mortality rates
  group_by(pois_rate) %>%
  summarise(px = sum(nEx)/obs_tot_exp)

# poisson predicted population mortality rate
pois_pop_nmx <- sum(with(pois_weighted_unique_rates, pois_rate*px))

# plot Poisson predicted frequency distribution of rates
ggplot(pois_weighted_unique_rates) +
  geom_segment(aes(x = pois_rate, xend = pois_rate, y = 0, yend = px)) +
  geom_rug(aes(x = pois_rate, alpha = px), sides = 'b', show.legend = FALSE) +
  scale_x_continuous(trans = 'log10', breaks = 10^(-8:0)) +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(y = 'Density', x = 'Observed mortality rate',
       title = 'Distribution of Poisson predicted mortality rates during the hour following birth',
       subtitle = 'US birth cohorts 2005-2010 stratified into 1209 groups by\nbirthweight, gestation at
```



Bayesian estimation

My basic assumptions concerning the distribution of mortality risk in a population of newborns are:

- there is no zero risk, the risk must be positive
- the distribution of risk is smooth
- the distribution of risk has a long right tail of improbable, but high risk newborns