# Mortalitet og standardisering

Anna-Vera Jørring Pallesen, Laust Hvas Mortensen and Thomas Alexander Gerds

## 1 Mortalitet

SKRIV HER Anna-Vera og Laust

## 2 Standardiserede rater

Vi indfører nu begreberne *aldersspecifikke rater* og *standardiserede rater*. Disse er forskelligt fra de summariske rater fra Kapitel 1. I det følgende betegner vi derfor alle de rater, som Kapitel 1 har omtalt uden prædikat, med prædikatet *summarisk*. *Summarisk* betyder at raterne tæller hændelser og risikotid i hele befolkningen. For at motivere standardiserede rater starter vi med at forklare begrænsingen med de summariske rater når det kommer til sammenligning af forskelige befolkninger.

### 2.1 Sammenligning af summariske rater

Som udgangspunkt har det begrænset interesse at sammenligne forskelige befolkningers summariske rater. Det er især når befolkningerne, som man ønsker at sammenligne, har forskelige aldersfordelinger. Afhængig af formål med undersøgelsen kan det alligevel godt være at man vil sammenligne summariske rater, men det er vigtigt at man ved at de afhænger aldersfordelingen. Problemet som opstår ved sammenligning af summariske rater er ret nemt at indse ved følgende eksempel. En matematisk forklaring (Kitagawas dekomposition) følger i afsnit 5.1.

#### 2.1.1 Eksempel

Vi beregner de summariske mortalitetsrater i året 2011 i den kvindelig danske befolkning og også i den mændlige danske befolkning.

```
library(danstat)
library(tidyverse)
# risikotid i 2011 baseret på middelfolketal metode 1
# middelfolketal fra K3 bliver ganget med 1 år
x <- get_data("FOLK1a",
```

```
              variables=list(list(code="tid",values="2011K3"),
                                 list(code="køn",values=c(2,1))))
# fjern TID fordi den er konstant
x$TID <- NULL
# aendre variable navn fra INDHOLD til RisikoTid
x <- rename(x,"RisikoTid"="INDHOLD")
# number of doedsfald i 2011
d <- get_data("DOD",variables=list(list(code="tid",values="2011"),
                                   list(code="køn",values=c("K","M"))
    ))
# fjern TID fordi den er konstant
d$TID <- NULL
# navngivning af variable
d <- rename(d,"Doed"="INDHOLD")
# join
dat <- left_join(x,d,by="KØN")
# summariske mortalitetsrater per 1000 personaar
dat <- mutate(dat,"Summariske mortalitetsrate"=1000*Doed/RisikoTid)
dat
```

| KØN | RisikoTid | Doed | Summariske mortalitetsrate |
|---|---|---|---|
| Women | 2806716 | 26577 | 9.469073 |
| Men | 2760140 | 25939 | 9.397712 |

Vi ser at den summariske mortalitetsrate i året 2011 var 9,47 døde per 1000 personår for danske kvinder og 9,39 døde per 1000 personår for danske mænd. Hvordan skal disse rater fortolkes? En rate er jo ikke en sandsynlighed og det ville ikke være helt korrekt at konkludere at der døde 9,47 kvinder blandt 1000 kvinder, som man følger igennem 2011, fordi de kvinder som dør i 2011 jo ikke bidrager med et helt personår til risikotiden. En bedre fortolkning opstår når man sammenligner mortalitetsraten med hastigheden af en cykel. Hastigheden er raten cyklen bevæger sig. Dermed kan man fortolke mortalitetsraten som hastigheden befolkningen dør. Denne hastighed betegner vi også med dødelighed. Det vil sige at resultatet kan fortolkes på følgende måde. Danske kvinder har haft en lidt højere dødelighed i 2011 end danske mænd.

På første blik strider dette resultat imod den gængse viden at dansk kvinder lever længere end danske mænd. Det er problemet, som dette eksempel illustrerer: Fordi den kvindelig befolkning i Danmark er ældre end den mændlige, er den summariske mortalitetsrate højere for kvinder end for mænd. Den summariske mortalitetsrate afspejler ikke kun dødeligheden men også aldersfordelingen i befolkningen. Da kvinder lever længere end mænd, er der flere ældre kvinder end ældre mænd og det forøger kvindernes summariske mortalitetsrate, da dødeligheden vokser med alderen.

# 3 Aldersfordeling

## 3.1 Alderspyramide

For at sammenligne aldersfordelinger af kvinder og mænd, kan man tegne en alderspyramide. Figur 1 viser alderspyramiden for den danske befolkning baseret på data fra 1 juli 2023. Man kan tydeligt se forskelen mellem mænd og kvinder i toppen af pyramiden. En mere sofistikerede og dynamisk version af den danske alderspyramide findes her `https://extranet.dst.dk/pyramide/pyramide.htm`.

```
library(ggplot2)
library(ggthemes)
## begge køn
dt <- get_data("FOLK1a",variables=list(
                          list(code="alder",values=0:125),
                          list(code="køn",values=1:2),
                          list(code="tid",values="2023K3")))
# formatere ALDER til numerisk
dt <- mutate(dt,ALDER=as.numeric(gsub(" year[s]?","",ALDER)))
# fjern tomme aldre
dt <- subset(dt,ALDER<106)
# separere køn
dt_m <- subset(dt,KØN=="Men") %>% mutate(INDHOLD=-INDHOLD)
dt_k <- subset(dt,KØN=="Women")
# plot
g <- ggplot(dt, aes(x = ALDER, y = INDHOLD, fill = KØN)) +
  geom_bar(data=dt_m, stat = "identity")+
  geom_bar(data=dt_k, stat = "identity")+
  coord_flip() +
  theme_solarized_2()+ylab("Folketal N(t)")+xlab("Alder (år)")+
  theme(legend.title=element_blank())
g <- g+ggtitle("Alderspyramide Danmark 1 juli 2023")
g
```

## 3.2 Folketal i aldersgrupper

Aldersfordelingen af folketal angiver hvor mange personer i en befolkning har en bestemt alder, for alle aldre. Det kan den enten gør i absolutte antal, eller som procent i forhold til antal personer i hele befolkingen. For at beskrive aldersfordelinger, vil man typisk vælge et passende antal aldersintervaller (passende til opgaven man sidder med) og fordele befolkningen på intervallerne. Intervallerne behøver ikke være lige stor. Da alle personers aldre ændrer sig hele tiden, skal man angive det dato, som aldersfordelingen referer til. For eksempel kan vi tale om aldersfordeling af kvinder i Danmark den 8 marts 1910 og om aldersfordeling af Fynens population den 1 juli 1989.

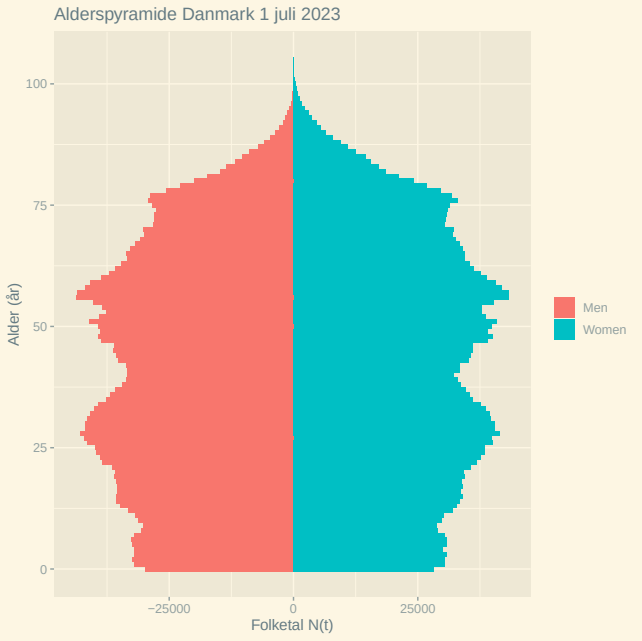


Figure 1: Data fra statistikbankens FOLK1a

### 3.2.1 Eksempel

Vi finder aldersfordeling af folketal for hele den danske befolkning den 1 januar 2023 og inddeler den i 4 intervaller: $[0, 25], (25, 50], (50, 75], (75, 125]$. Bemærk at vores syntaks for intervaller betyder at intervalgrænsen er inkluderet hvis parentesen er rundt og ekskluderet hvis parentesen er firkantet. Det vil sige at personer, som er præcis 25 år gamle falder i intervallet $[0, 25]$ og personer som er 50 falder ikke i intervallet $(50, 75]$. Vi beregner også andelen, som de enkelte aldersgrupper udgør og angiver den i procent.

```
library(danstat)
library(tidyverse)
## meta <- get_table_metadata("FOLK1a")
## meta$variables[3,]$values[[1]][-1,"id"]
dt <- get_data("FOLK1a",variables=list(
                          list(code="alder",values=0:125),
                          list(code="tid",values="2023K3")))
# formatere ALDER til numerisk
dt <- mutate(dt,ALDER=as.numeric(gsub(" year[s]?","",ALDER)))
# Aldersintervaller
dt <- mutate(dt,Aldersinterval=cut(ALDER,
                                   breaks=c(0,25,50,75,125),
                                   include.lowest = TRUE))
# antal person i de 4 aldersintervaller
```

```
af <-dt%>% group_by(Aldersinterval) %>% summarise(Antal=sum(INDHOLD))
# procent
af <- af %>% mutate(Procent=100*Antal/sum(Antal))
af
```

```
 Aldersinterval Antal    Procent
[0,25]          1742979   29.3
(25,50]         1882860   31.7
(50,75]         1778084   29.9
(75,125]         540222    9.09
```

### 3.2.2 Notation og formler

En hver definition af aldersintervaller opdeler en befolkning i aldersgrupper. For $x = 1, \ldots, m$ aldersgrupper betegner vi med $N_x(t)$ folketal i aldersgruppe $x$ til tid $t$. Vi betegner fortsæt med $N(t)$ folketal i hele befolkningen til tid $t$ og kan nu udtrykke den som sum af folketal i alle aldersgrupper:

$$N(t) = N_1(t) + \cdots + N_m(t) = \sum_{x=1}^{m} N_x(t).$$

Vi beregner andelen af befolkningen i aldersgruppe $x$ ved at dividere folketal i aldersgruppen med folketal i hele befolkningen til tid $t$:

$$\frac{N_x(t)}{N(t)} = \{\text{Andel af befolkningen i aldersgruppe } x \text{ til tid } t\}.$$

### 3.2.3 Sammenligning af aldersfordelinger

Vi kan bruge denne inddeling af aldersspektrum i 4 grupper for at sammenligne aldersfordeling i hovedstadsområdet med aldersfordeling i landdistrikter i Danmark i 2023. For at gør det henter vi folketal data fra statistikbankens register BY2 hvor man kan specificere bystørrelse.

```
## meta <- get_table_metadata("BY2")
b2 <- get_data("BY2",variables=list(
                      list(code="alder",values=0:125),
                      list(code="BYST",values=c("HOVEDS","LAND")),
                      list(code="tid",values="2023")))
# formatere ALDER til numerisk
b2 <- mutate(b2,ALDER=as.numeric(gsub(" year[s]?","",ALDER)))
# aldersintervaller
b2 <- mutate(b2,Aldersinterval=cut(ALDER,
                                   breaks=c(0,25,50,75,125),
                                   include.lowest = TRUE))
# antal person i de 4 aldersintervaller
af <-b2 %>% group_by(BYST,Aldersinterval) %>% summarise(Antal=sum(
    INDHOLD))
```

```
# procent
af <- af %>% mutate(Procent=100*Antal/sum(Antal))
af
```

```
 BYST                       Aldersinterval  Antal Procent
Greater Copenhagen Region [0,25]          424524   31.1
Greater Copenhagen Region (25,50]         520217   38.2
Greater Copenhagen Region (50,75]         329994   24.2
Greater Copenhagen Region (75,125]         88561    6.50
Rural areas               [0,25]          184556   26.8
Rural areas               (25,50]         198151   28.8
Rural areas               (50,75]         258161   37.5
Rural areas               (75,125]         46720    6.79
```

En sammenligning af de to aldersfordelinger viser at andelen af mennesker, der er over 75 år gamle, er cirka det samme, men at andelen af unge mennesker er højst i hovedstadsområdet og andelen af mennesker mellem 50 og 75 er højst i landdistrikterne.

## 3.3 Risikotid i aldersgrupper

Med hensyn til mortalitetsrater, har vi brug for aldersfordeling af risikotid i en bestemt kalenderperiode. Vi betegner med $R_x[t_1, t_2]$ den samlede gennemlevede tid i perioden $[t_1, t_2]$ af alle personer i aldersgruppe $x$. Vi bemærker at en person, som har levet i befolkingen i perioden $[t_1, t_2]$ kan bidrage med risikotid til et eller flere aldersintervaller. Det sker for personer som har fødselsdag mellem dato $t_1$ og dato $t_2$, hvis de den dag skifter fra aldersgruppe $x$ til aldersgruppe $x + 1$. Vi betegner fortsæt med $R[t_1, t_2]$ risikotiden for hele befolkningen og kan nu udtrykke den som sum af de aldersspecifikke risikotider:

$$R[t_1, t_2] = R_1[t_1, t_2] + \cdots + R_m[t_1, t_2] = \sum_{x=1}^{m} R_x[t_1, t_2].$$

Vi beregner andelen af risikotid i aldersgruppe $x$ ved at dividere risikotid i aldersgruppen med risikotid i hele befolkningen i perioden $[t_1, t_2]$ og betegner den med $V_x$:

$$V_x[t_1, t_2] = \frac{R_x[t_1, t_2]}{R[t_1, t_2]} = \{\text{Andel af risikotid i aldersgruppe } x \text{ i perioden } [t_1, t_2]\}.$$

Risikotid beregnes ofte ved at gange middelfolketal med periodens længde. I den særlige situation, hvor perioden er 1 år lang, altså når $t_2 - t_1 = 1$ år, har middelfolketal (antal) og risikotid (personår) den samme værdi, men forskelige endheder. Vi skal bruge $V_x$ som vægte i definitionen af aldersstandardiserede rater (Afsnit 5).

### 3.3.1 Eksempel

Vi finder aldersfordeling af risikotid for hele den danske befolkning i perioden mellem den 1 januar 2022 og den 1 januar 2023 og inddeler den i 4 intervaller: $[0, 25], (25, 50], (50, 75], (75, 125]$.

```
library(danstat)
library(tidyverse)
dt <- get_data("FOLK1a",variables=list(
                          list(code="alder",values=0:125),
                          list(code="tid",values=c("2022K3","2023K3")
    )))
# formatere ALDER som numerisk
dt <- mutate(dt,ALDER=as.numeric(gsub(" year[s]?","",ALDER)))
# Risikotid= 1* Middelfolketal metode 2
dt <- dt %>% group_by(ALDER) %>% summarise(Risikotid=1*mean(INDHOLD))
# Aldersintervaller
dt <- mutate(dt,Aldersinterval=cut(ALDER,
                                   breaks=c(0,25,50,75,125),
                                   include.lowest = TRUE))
# antal personår i de 4 aldersintervaller
af <-dt%>% group_by(Aldersinterval) %>% summarise(Personår=sum(
    Risikotid))
# procent
af <- af %>% mutate(Procent=100*Personår/sum(Personår))
af
```

```
Aldersinterval Personår   Procent
[0,25]         1746486    29.5
(25,50]        1876958    31.7
(50,75]        1775912    30.0
(75,125]        528006     8.91
```

## 3.4 Lexis diagram

Lexis diagrammet visualiserer sammenhæng mellem kalendertid (vertikal) og alder (horisontal). Hver person er repræsenteret af sin livslinje i et Lexis diagram. I en såkaldt *lukket befolkning* starter alle livslinjer i fødselsdagen hvor personen er 0 år gamle og ender i dødsdatoen den alder personen har livet til. I en åben befolkning, starter livslinjer for immigranter den dag de immigrerer og slutter for emigranter den dag de emigrerer.

Figur 2 viser 5 personers livslinjer fra en åben befolkning. Den mørkeblå linje representerer en person som bliver født i foråret 2015 og forbliver i befolkingen indtil foråret 2020 – måske også læ

Figur 3 viser hvordan man kan følge en aldersgruppe igennem kalendertid, en fødselskohorte igennem kalendertid og alder, og hvilken del en periode

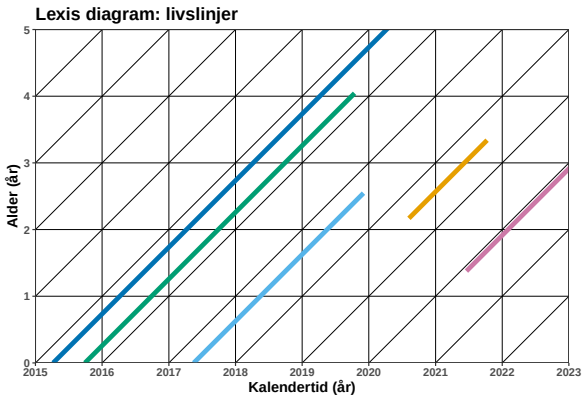


Figure 2: Figuren viser 5 personers livslinjer i (den nederste del af) et Lexis diagram. Livslinjer der ikke starter i alder '0' representerer immigranter og livslinjer som stopper representerer enten dødsfald eller emigranter.

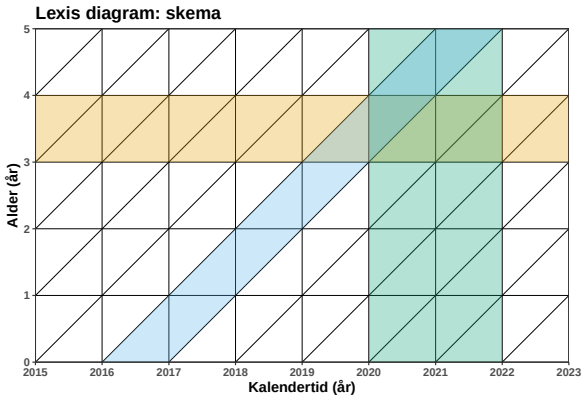


Figure 3: I et Lexis diagram kan man følge en aldersgruppe igennem kalendertid (gul) eller en fødselskohorte igennem både alder og kalendertid (blå). Det grønne område viser en kalenderperiode.

# 4 Aldersspecifikke mortalitetsrater

Vi ser på en befolkning i en kalenderperiode $[t_1, t_2]$ og inddeler den i $\{x = 1, \ldots, m\}$ aldersgrupper. Vi betegner med $D_x[t_1, t_2]$ antal dødsfald i perioden hvor personens alder ved dødsdatoen falder i aldersgruppe $x$. For at lette notationsbyrden dropper vi kalenderperioden og forkorter $D_x[t_1, t_2]$ til $D_x$ og ligeledes skriver vi $R_x$ for den aldersspecifikke risikotid $R_x[t_1, t_2]$ i samme periode. De aldersspecifikke mortalitetsrater er defineret som

$$M_x = \frac{D_x}{R_x}, \qquad x = 1, ..., m.$$

Bemærk at den aldersspecifikke mortalitetsrate $M_x$ afhænger også kalenderperioden og den langform notation er $M_x[t_1, t_2]$.

# 5 Aldersstandardisering

Formålet med alderstandardisering er at befolkninger bliver sammenlignelig selv hvis aldersfordelingerne er afvigende. For eksempel, kan vi spørge hvor meget højere er mortalitetsraten blandt danske mænd sammenlignet med danske kvinder hvis aldersfordeling havde været den samme blandt mænd og kvinder. Det mangler at specificere den aldersfordeling som de standardiserede rater skal have i fælles. Her er der umiddelbart fire forskelige muligheder: aldersfordeling bland mænd, aldersfordeling bland kvinder, aldersfordeling bland alle uanset køn, en helt tredje aldersfordeling. Vi beskriver to standardiseringsformer, *direkte standardisering* og *indirekte standardisering*. Vi motiverer deres formler ved en matematisk forklaring af resultatet fra afsnit 2.1.

## 5.1 Kitagawas dekomposition

Forskellen på de summariske mortalitetsrater mellem to populationer (population A vs population B) skyldes både:

- Forskelle i aldersspecifikke mortalitetsrater
- Forskelle i aldersfordeling

$$\begin{aligned} M^A - M^B &= \sum_x M_x^A V_x^A - \sum_x M_x^B V_x^B \\ &= \underbrace{\sum_x (M_x^A - M_x^B) \frac{V_x^A + V_x^B}{2}}_{\text{Komponent 1}} + \underbrace{\sum_x (V_x^A - V_x^B) \frac{M_x^A + M_x^B}{2}}_{\text{Komponent 2}} \end{aligned}$$

**Komponent 1** = (Forskel i aldersspecifikke mortalitetsrater) vægtet med gennemsnitlig aldersfordeling

**Komponent 2** = (Forskel i aldersfordeling) vægtet med gennemsnitlig mortalitetsrate