POISSON REGRESSION

Håkon K. Gjessing

Professor/Principal Investigator

Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo

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

Makerere

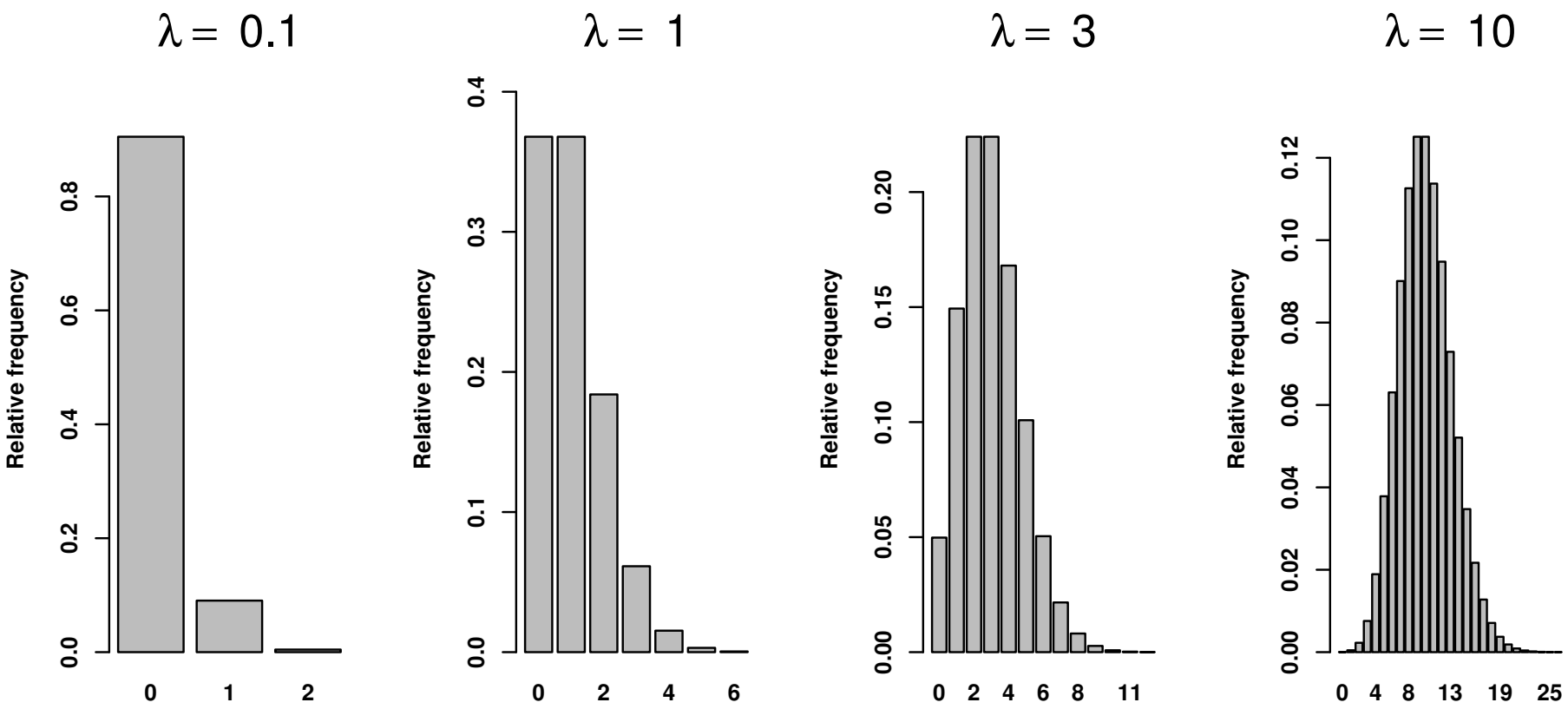
Monday, 12 June 2023

CENTRE FOR FERTILITY AND HEALTH

SM Topher Kisuule, 1993

CENTRE FOR FERTILITY AND HEALTH

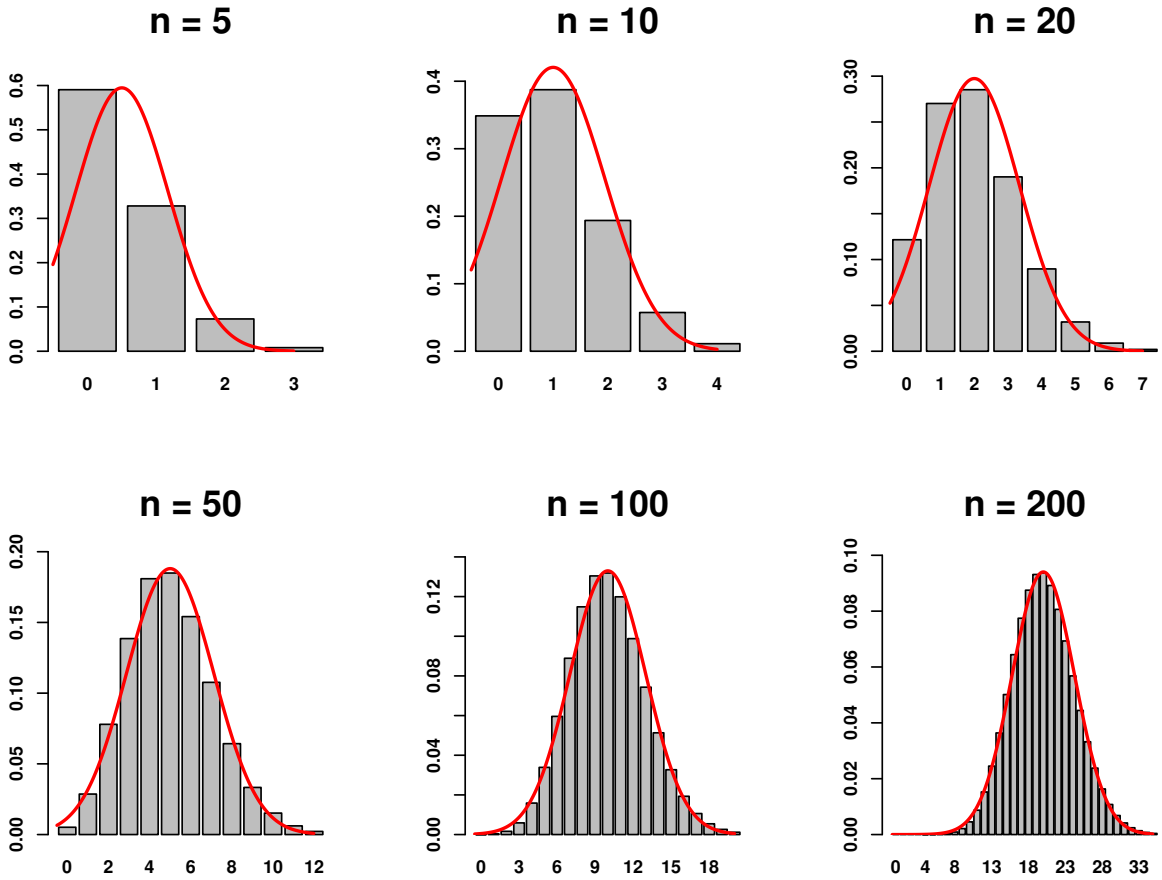
POISSON DISTRIBUTION



- Count data: 0, 1, 2, 3, ... no upper limit
- λ: Both mean and variance of the Poisson distribution

CENTRE FOR FERTILITY AND HEALTH

BINOMIAL → NORMAL, P = 0.1, N INCREASES

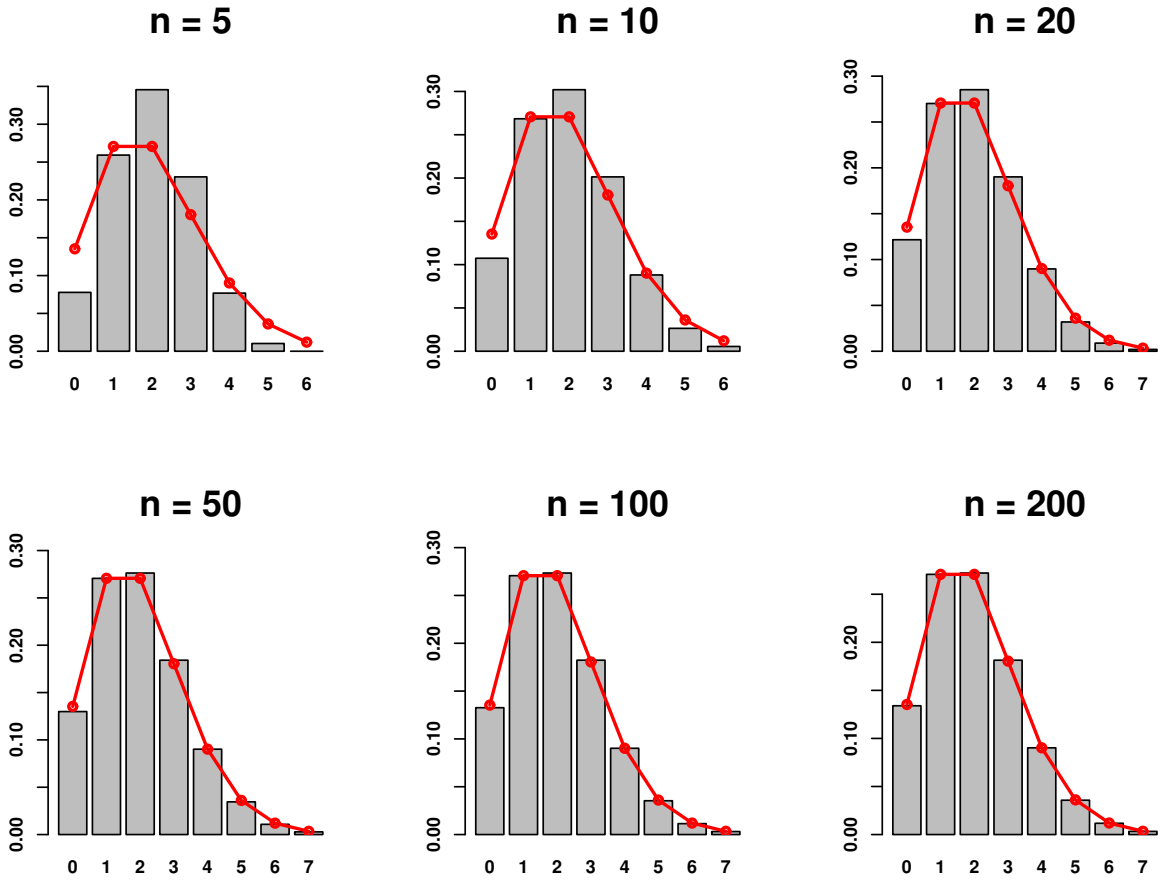


Common requirement:

$$np > 5$$

$$n(1-p) > 5$$

BINOMIAL → POISSON, λ = 2, N INCREASES, P DECREASES



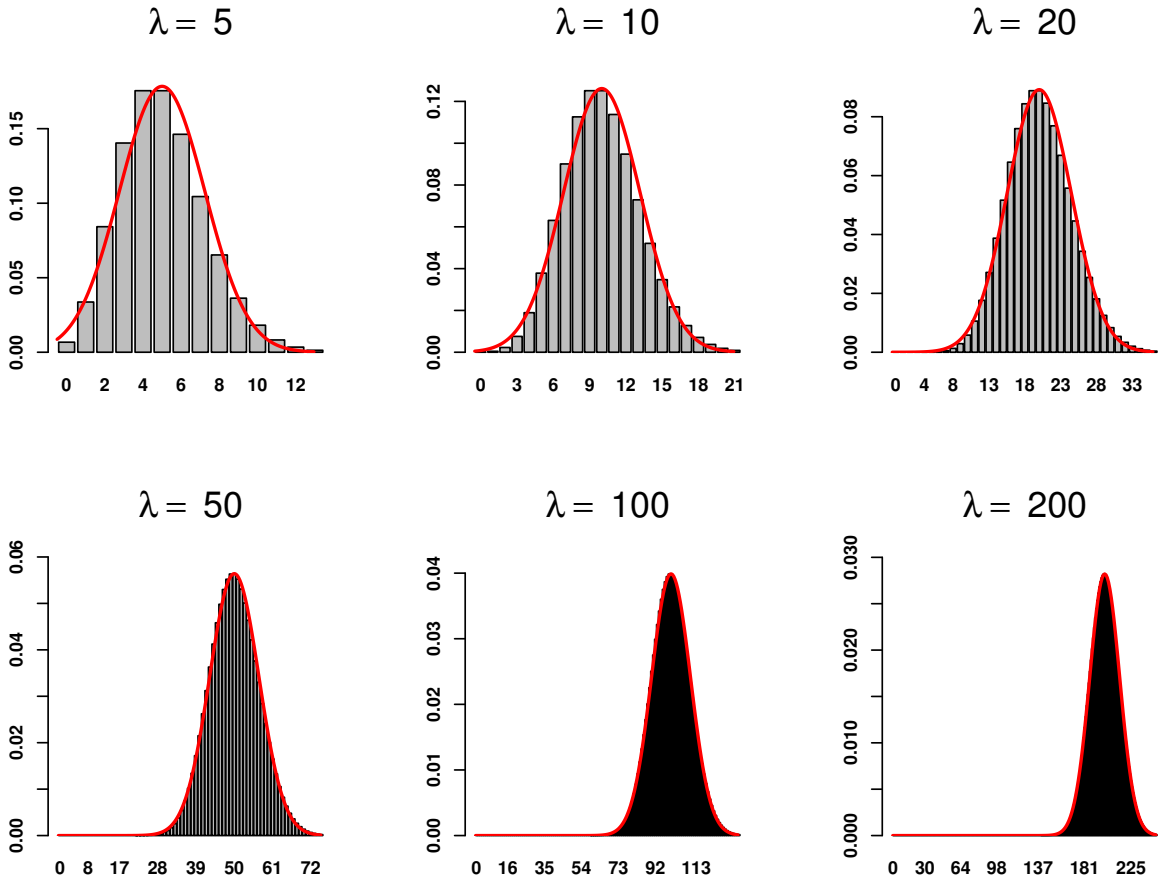
Common requirement:

$$n >= 50$$

$$p <= 0.05$$

CENTRE FOR FERTILITY AND HEALTH

POISSON → NORMAL, λ INCREASES



Common requirement:

$$\lambda >= 5 \text{ (or 10?)}$$

CENTRE FOR FERTILITY AND HEALTH

POISSON REGRESSION

Count traffic accidents in an area in two following weeks

- `count`: number of accidents counted
- `week`: week number 1 and 2
- `week01`: week number, coded 0 and 1
- `at.risk`: amount of traffic, for instance in units of Boda-Boda-kilometers

Data:

count	week	week01	at.risk
50	1	0	2000
36	2	1	1000

- We can use `week`, but `week01` is a "standard dummy", slightly easier in interpretation)
- `at.risk = 2000 means 2000 kilometers by Boda-Boda`
- `NOTE that varying degrees of madness among Boda-Boda drivers may cause` *`overdispersion`* `(although they all seem equally crazy)`

CENTRE FOR FERTILITY AND HEALTH

POISSON REGRESSION, MODEL

To begin with: ignore `at.risk`

Recall GLM (Generalized Linear Model)

- Family **Poisson**
- Link function **log**, i.e. $\boxed{\log(\lambda) = \beta_0 + \beta_1 \cdot x_1 + \cdots}$

Model:

$$\begin{aligned} \textbf{week 1:} \quad \lambda_1 &= \exp(\beta_0) \\ \textbf{week 2:} \quad \lambda_2 &= \exp(\beta_0 + \beta_1) = \exp(\beta_0) \cdot \exp(\beta_1) \\ \textbf{Rate Ratio RR:} \quad \frac{\lambda_2}{\lambda_1} &= \exp(\beta_1) \end{aligned}$$

R model:

```
res <- glm(count ~ week01, family = poisson, data = .data)
```

CENTRE FOR FERTILITY AND HEALTH

POISSON REGRESSION, RESULTS

Estimates and confidence intervals on log scale:

NOTE! In Stata there is the **eform** option that gives you the exponentiated values

```
> summary(res)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.9120     0.1414  27.662   <2e-16 ***
week01       -0.3285     0.2186  -1.503    0.133
```

```
> confint(res)
```

```
                2.5 %     97.5 %
(Intercept)  3.6214057 4.17693452
week01      -0.7635849 0.09655906
```

CENTRE FOR FERTILITY AND HEALTH

POISSON REGRESSION, RESULTS

Estimates and confidence intervals on a multiplicative scale:

```
> exp(res$coefficients)
```

```
(Intercept)      week01
      50.00        0.72
```

```
> exp(confint(res))
```

```
                2.5 %    97.5 %
(Intercept) 37.3900896 65.165782
week01       0.4659929  1.101375
```

Interpretation:

- Estimate for `(Intercept)` is simply the count in week 1 (when we use `week01`)
- Estimate for `week01` is the ratio week 2/week 1 = 36/50 = 0.72, i.e. expected count in week 2 is 72% of week 1
- Count in week 2 can be found as $50 \times 0.72 = 36$
- *Saturated model* number of parameters same as number of observations
- In a saturated model, model expected numbers λ are the same as the observed counts

CENTRE FOR FERTILITY AND HEALTH

POISSON REGRESSION, MODEL

Now, **do it correctly** and take at.risk into account

Model:

$$\begin{aligned} \textbf{week 1:} \quad \lambda_1/2000 &= \exp(\beta_0) \\ \textbf{week 2:} \quad \lambda_2/1000 &= \exp(\beta_0 + \beta_1) \end{aligned}$$

Model:

$$\begin{aligned} \textbf{week 1:} \quad \lambda_1 &= 2000 \times \exp(\beta_0) \\ \textbf{week 2:} \quad \lambda_2 &= 1000 \times \exp(\beta_0 + \beta_1) \end{aligned}$$

Model, on log scale:

$$\begin{aligned} \textbf{week 1:} \quad \log(\lambda_1) &= \log(2000) + \beta_0 \\ \textbf{week 2:} \quad \log(\lambda_2) &= \log(1000) + \beta_0 + \beta_1 \end{aligned}$$

- The values $\log(2000)$ and $\log(1000)$ have no parameter to be estimated... they are called an “offset”
- β_0 is still an intercept to be estimated, but same for all (and “offset” by the value of $\log(\texttt{at.risk})$)

CENTRE FOR FERTILITY AND HEALTH

POISSON REGRESSION, MODEL

The offset has two roles:

- Balance the two groups correctly against one another, according to `at.risk`
- Give the intercept β_0 a more relevant interpretation in terms of expected rates, not expected numbers

Again, a GLM with Poisson family, log link, but with offset!

R model:

```
> res <- glm(count ~ week01 + offset(log(at.risk)),
 family = poisson, data = .data)
```

NOTE that Stata can use either an offset *or* an exposure. The latter is just 2000, 1000.... thus it needs no log, and the output may be easier to read.

CENTRE FOR FERTILITY AND HEALTH

POISSON REGRESSION, RESULTS

Estimates and confidence intervals on log scale:

```
> summary(res)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6889     0.1414 -26.084   <2e-16 ***
week01        0.3646     0.2186   1.668   0.0953 .
```

```
> confint(res)
```

```
                2.5 %     97.5 %
(Intercept) -3.97949677 -3.4239679
week01      -0.07043773  0.7897062
```

CENTRE FOR FERTILITY AND HEALTH

POISSON REGRESSION, RESULTS

Estimates and confidence intervals on multiplicative scale:

```
> exp(res$coefficients)
```

```
(Intercept)      week01
      0.025       1.440
```

```
> exp(confint(res))
```

```
                2.5 %     97.5 %
(Intercept) 0.01869504 0.03258289
week01      0.93198577 2.20274925
```

Interpretation:

- Estimate for `(Intercept)` is the rate in week 1, i.e. count in week 1 per number `at.risk` = 50/2000
- Estimate for `week01` is the rate ratio

$$\frac{\text{\# week 2}/1000}{\text{\# week 1}/2000} = \frac{36/1000}{50/2000} = 1.44,$$

  i.e. expected count in week 2 (per number `at.risk`)
  is 44% higher than in week 1

CENTRE FOR FERTILITY AND HEALTH

CAUSES OF OVERDISPERSION

$-> $ Separate presentation

CENTRE FOR FERTILITY AND HEALTH

Nepal Study: pneumonia hospital visits

Count the number of hospital visits with pneumonia following the first one

- Age range: 2 months to 3 years
- **Dates:** November 2003 to December 2007
- **Main exposure:** Zinc versus placebo

(Data from Tor Strand, Maria Mathisen, and others,
Centre for International Health, UiB)

CENTRE FOR FERTILITY AND HEALTH

Nepal Study: Pneumonia

	id	date	age	sex	weight	length	haemoglobin	bf	born
1	1	2003-11-17	8	1	7.8	73.5	10.0	1	2
2	1	2004-01-25	10	1	NA	NA	10.0	NA	2
3	1	2004-04-02	12	1	NA	NA	10.0	NA	2
4	2	2003-11-20	9	2	8.0	70.0	8.8	1	2
5	2	2004-03-22	13	2	NA	NA	8.8	NA	2

	.zwei	.zlen	.zwfl	block	treat	child.obs	visitnumber	time
1	-0.91	1.30	-2.04	1	1	3	0	0
2	NA	NA	NA	1	1	3	1	69
3	NA	NA	NA	1	1	3	2	68
4	-0.23	-0.06	-0.22	1	1	2	0	0
5	NA	NA	NA	1	1	2	1	123

	tae	pneumonia	severepneumonia	diarrhea	censordate	dof
1	0	0	0	0	2004-05-23	188
2	69	1	0	0	2004-05-23	188
3	137	1	0	0	2004-05-23	188
4	0	0	0	0	2004-05-20	182
5	123	1	0	0	2004-05-20	182

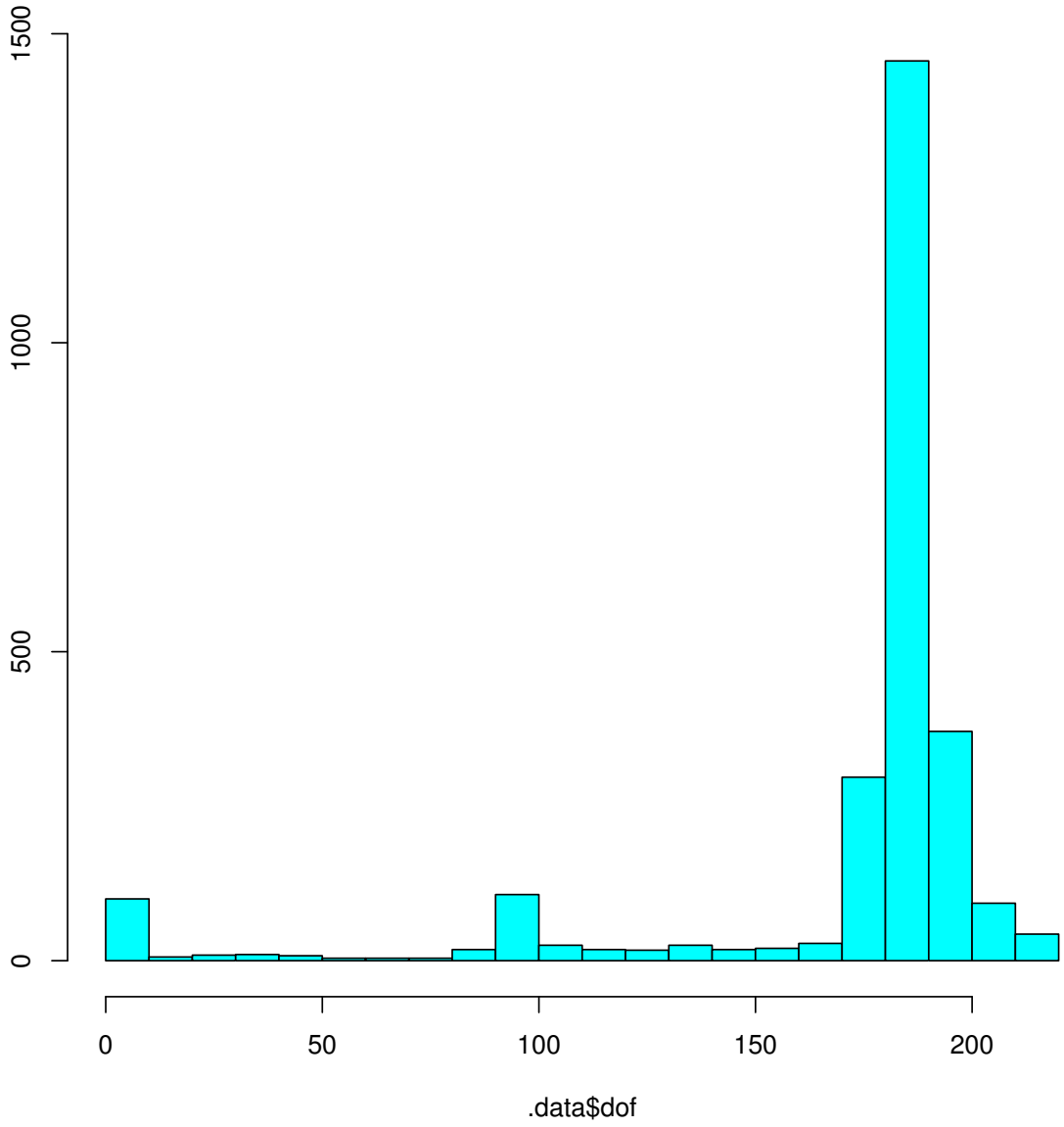
CENTRE FOR FERTILITY AND HEALTH

VARIABLES

- Original file contains
  - `age`: Age of child (in months)
  - `child.obs`: Number of observations per child
  - `dof`: Total number of days of follow-up for each child
  - ... and much more
- Compute: `count` = `child.obs` - 1

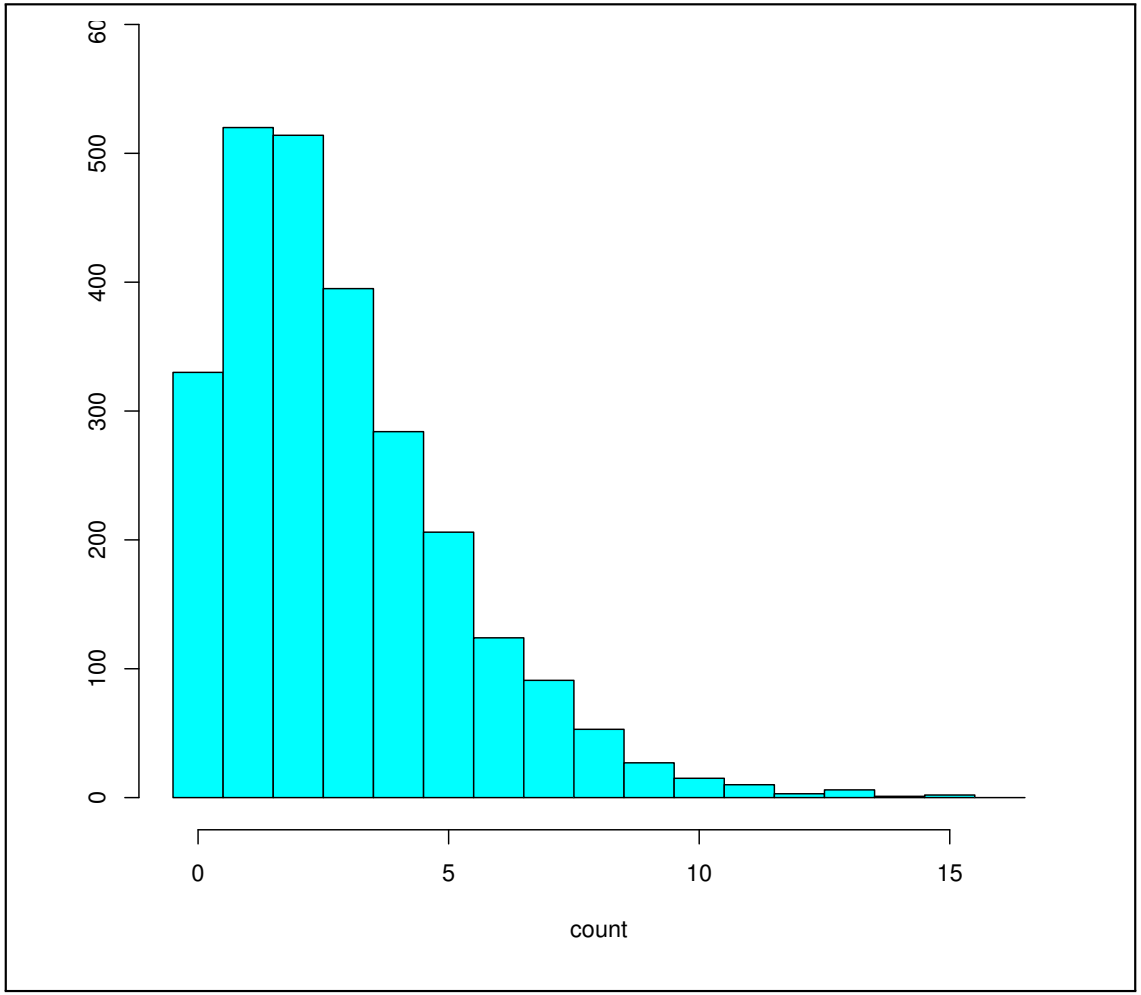
CENTRE FOR FERTILITY AND HEALTH

Days of follow-up



CENTRE FOR FERTILITY AND HEALTH

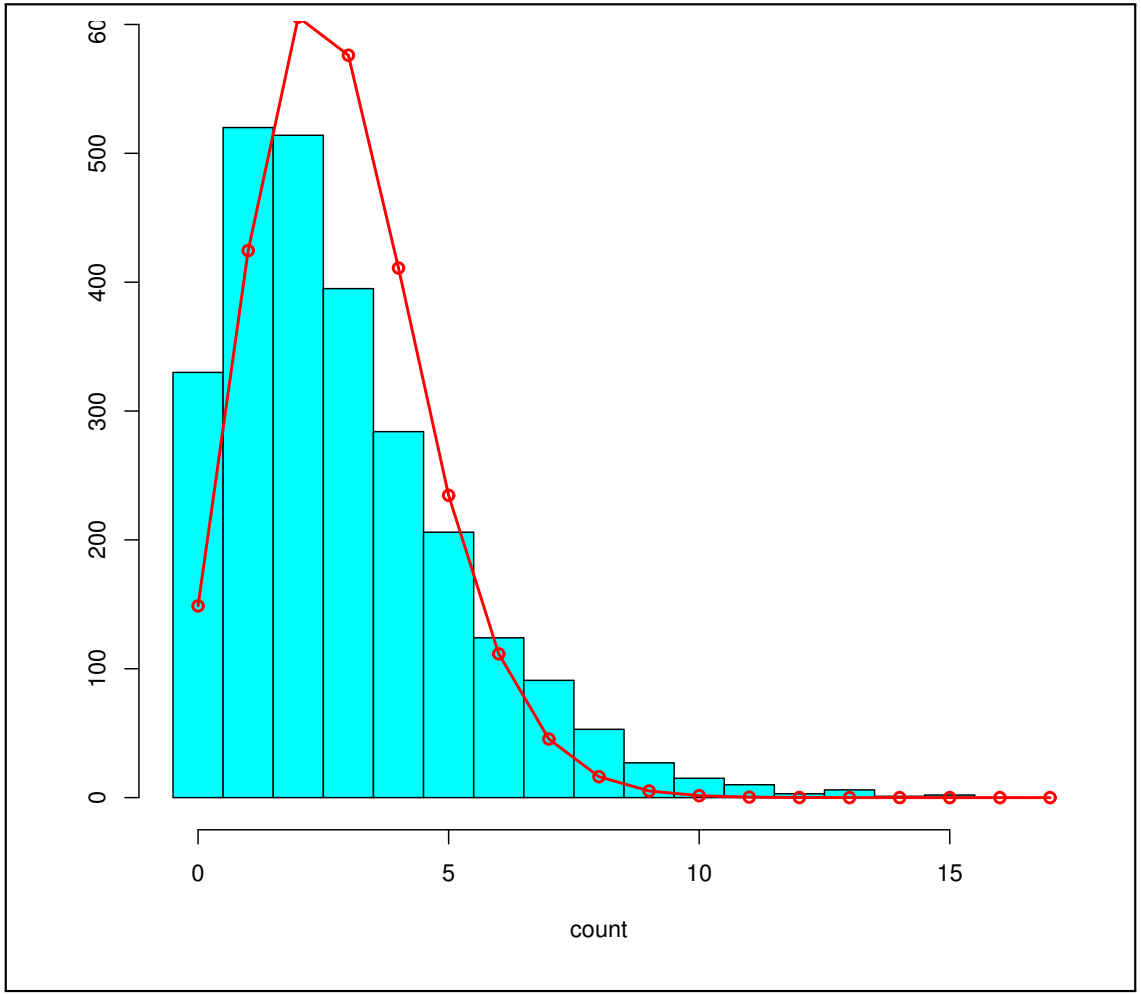
OVERALL DISTRIBUTION COUNT



λ = Mean = 2.85

CENTRE FOR FERTILITY AND HEALTH

Overall distribution count, Poisson added (red)



λ = Mean = 2.85, BUT λ = Variance = 5.47 ???

CENTRE FOR FERTILITY AND HEALTH

OVERDISPERSION

Variable follow-up for children creates *overdispersion*

$-> $ Separate presentation

CENTRE FOR FERTILITY AND HEALTH

Effect of age on count, Poisson regression

```
glm(count ~ offset(log(dof)) + age, family = poisson, data = .data)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.774897   0.021046 -179.37   <2e-16 ***
age         -0.027972   0.001522  -18.38   <2e-16 ***
```

Transform to exp(coefficient):

```
(Intercept)         age
 0.02293944  0.97241583
```

Transform to exp(12*coefficient) (year scale):

```
      age
0.7148653
```

TASK: SIDE EFFECTS OF ASTRAZENECA VACCINE

- **In Norway:** 4 deaths had been confirmed as side effects among 134,000 vaccinated with the AstraZeneca vaccine

Risk: $4/134,000 \approx 3$ out of 100,000

- If the vaccine is removed from the program, this will cause delays in vaccination
- Delays will cause further deaths from infections
- Continuing with the vaccine may cause more deaths as side effects

---

Question:

To an individual, what is most/least dangerous:

1. Take the vaccine
2. Wait until a safer vaccine is available, but then risk dying from Covid-19 in the meantime

---

- Should be considered in light of the current situation of infections in Norway at that time
- Should take into consideration differences in mortality rates between women and men, and across different age groups

CENTRE FOR FERTILITY AND HEALTH

Background

Data source: BeredtC19

- Several Norwegian registries linked together
- Exclusively to answer questions related to the pandemic
- Infections (positive Covid-19 PCR-test)
- What vaccines received, and when
  - Pfizer (mRNA vaccine)
  - Moderna (mRNA vaccine)
  - AstraZeneca (adeno-vector)

Outcomes

1. Hospital admissions related to Covid-19
2. ICU admissions related to Covid-19
3. Deaths related to Covid-19

CENTRE FOR FERTILITY AND HEALTH

Survival analysis

- Outcomes: Hospitalization, ICU admission, Death
- Analyses at individual levels?
- Data in aggregated form is much easier to obtain and work on

CENTRE FOR FERTILITY AND HEALTH

DATA

	sex	age_gr	isoyearweek	ICU	hospital	ASZ03_partly_vacc	ASZ03_fully_vacc
1	Female	00-04	2020-15	x	x	x	x
2	Female	00-04	2020-16	x	x	x	x
3	Female	00-04	2020-17	x	x	x	x
4	Female	00-04	2020-18	x	x	x	x
5	Female	00-04	2020-19	x	x	x	x
6	Female	00-04	2020-20	x	x	x	x

	BNT03_partly_vacc	BNT03_fully_vacc	JAN03_partly_vacc	JAN03_fully_vacc
1	x	x	x	x
2	x	x	x	x
3	x	x	x	x
4	x	x	x	x
5	x	x	x	x
6	x	x	x	x

	MOD03_partly_vacc	MOD03_fully_vacc	unkno_partly_vacc	unkno_fully_vacc
1	x	x	x	x
2	x	x	x	x
3	x	x	x	x
4	x	x	x	x
5	x	x	x	x
6	x	x	x	x

	infect	dead	tested	pop	period	year	week	age_num	pop.red
1	x	x	x	140650	1	2020	15	1	140650
2	x	x	x	140650	2	2020	16	1	140650
3	x	x	x	140650	3	2020	17	1	140650
4	x	x	x	140650	4	2020	18	1	140650
5	x	x	x	140650	5	2020	19	1	140650
6	x	x	x	140650	6	2020	20	1	140650

CENTRE FOR FERTILITY AND HEALTH

Data

Table consists of frequency counts (x for anonymity)

One row for each combination of `sex`, `age_gr`,

and week of pandemic (`isoyearweek`)

```
> dim(data)
2160 24
```

```
> length(unique(data$isoyearweek))
54
```

`sex`: 2

`age_gr`: 20

`isoyearweek`: 54

$2 \cdot 20 \cdot 54 = 2160$

CENTRE FOR FERTILITY AND HEALTH

Poisson regression in R

Plain age group analysis:

```
> glm.result <- glm(dead ~ age_gr + offset(log.pop),
    + data = data, family = poisson)
```

- `dead` is the number of deaths. A *count* variable: $0, 1, 2, 3, \ldots$
- `age_gr` is a categorical variable of 5-year categories
- `log.pop` is the log of the population background size in each group

CENTRE FOR FERTILITY AND HEALTH

Poisson regression

$$\lambda_i = \eta_i \cdot \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots\right)$$

- λ_i is the expected count in cell i
- η_i is the "background" population size in cell i
- x_{1i} etc. are dummies for each age category

Or, on log scale:

$$\log(\lambda_i) = \log(\eta_i) + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots$$

Note:

- η_i serves as a "scaling." No parameters are estimated for the offset.

CENTRE FOR FERTILITY AND HEALTH

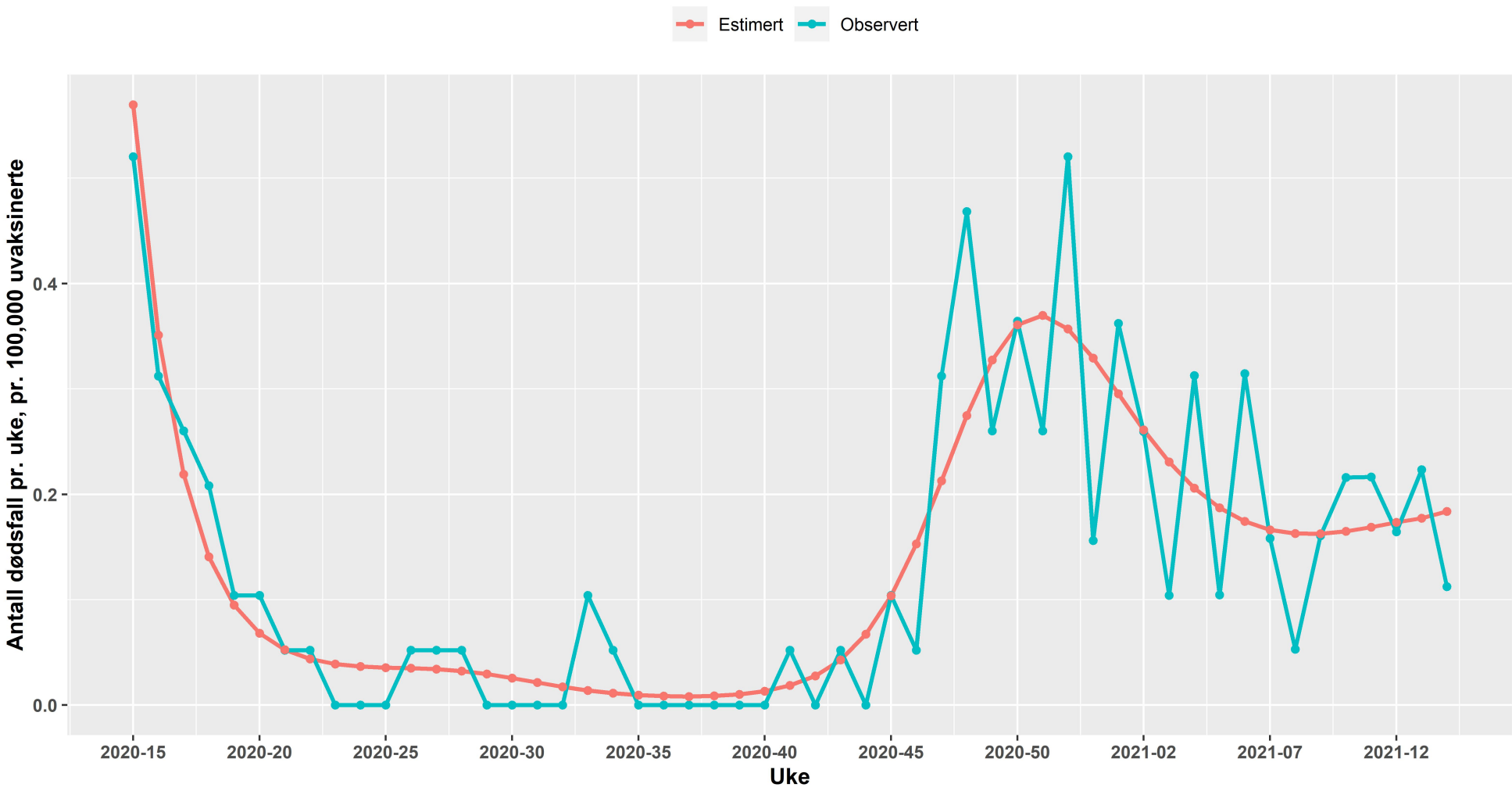
MODEL

```
> library(mgcv)

> gam.result <- gam(dead ~ s(period, fx = T, k = 13)
    + s(age_num, fx = T, k = 4, by = sex) + sex
    + offset(log.pop), data = data, family = poisson)

> gam.pred <- predict.gam(gam.result, newdata = xxxxx,
    type = "response")

> gam.pred.logscale <- predict.gam(gam.result, newdata = xxxxx,
    type = "link", se.fit = T)
```

- `newdata` can contain a log.pop variable equal to, say, $\log(100000)$
- `s(period....)` is a spline function of week
- BEWARE that infectious disease data will often have overdispersion

CENTRE FOR FERTILITY AND HEALTH

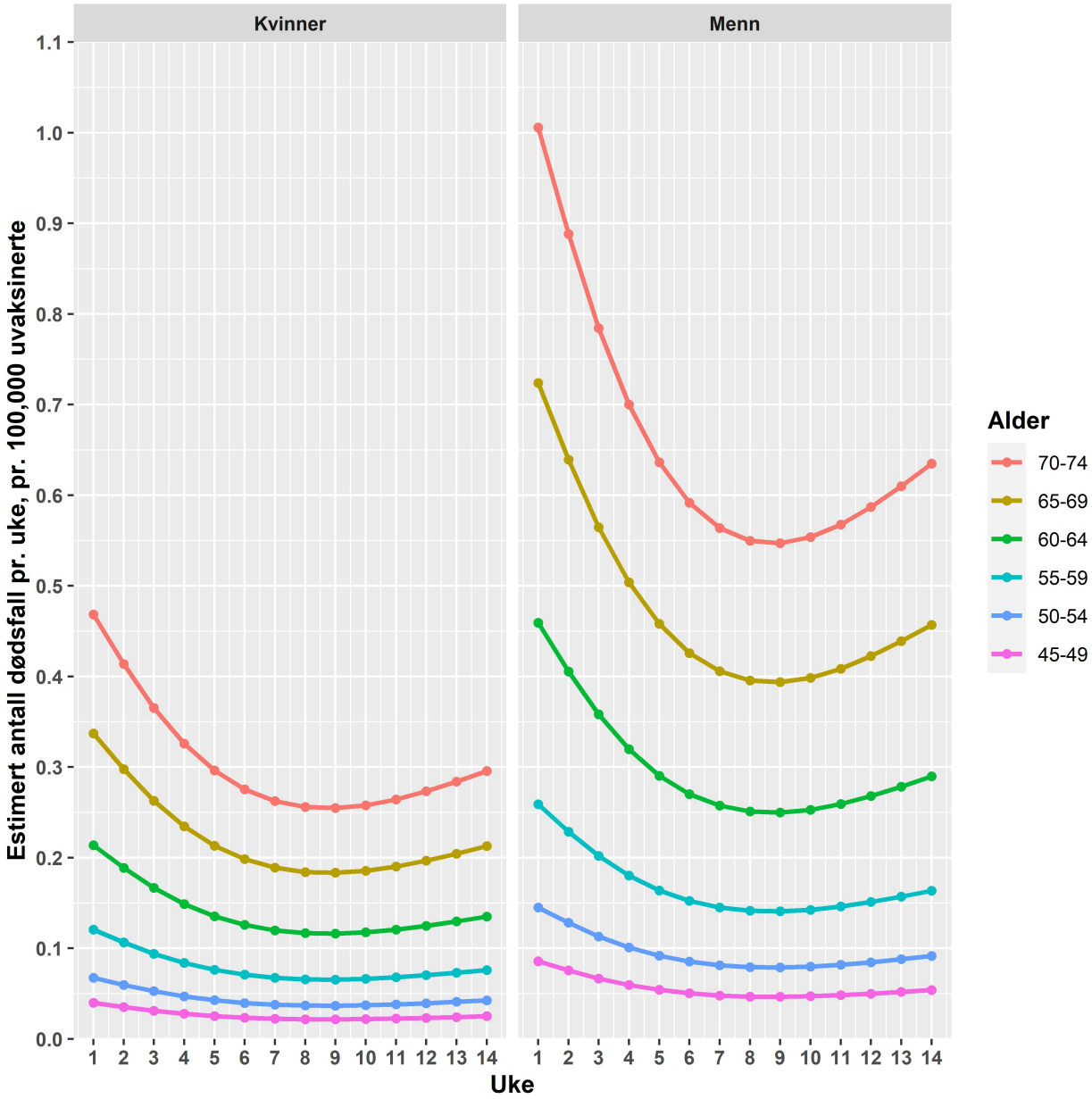
MORTALITY RATES ACROSS PANDEMIC, WEEK BY WEEK



(Aggregated over selected age groups)

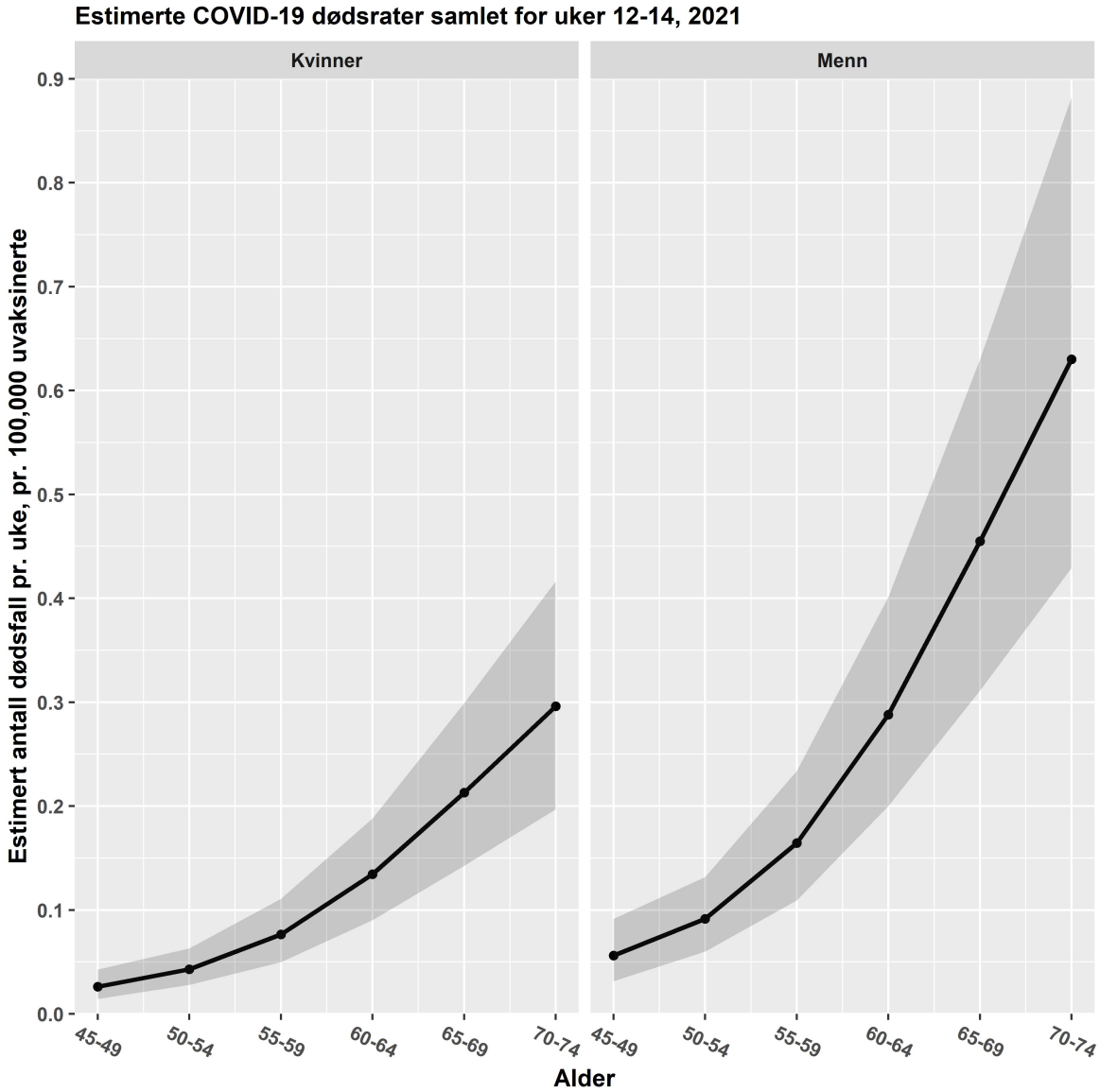
CENTRE FOR FERTILITY AND HEALTH

MORTALITY RATES IN AGE CATEGORIES, WEEK BY WEEK, 2021



CENTRE FOR FERTILITY AND HEALTH

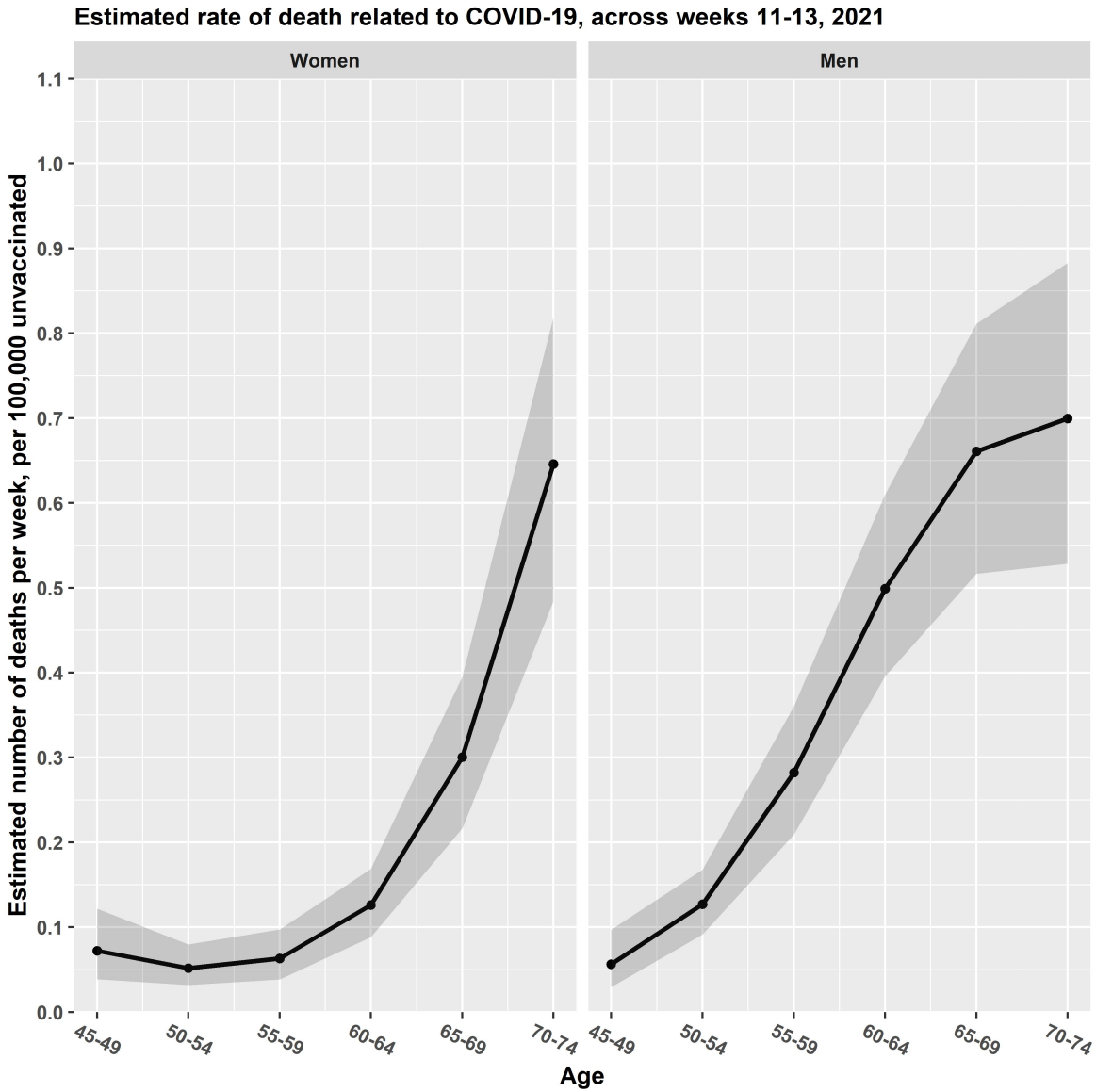
MORTALITY RATES IN AGE CATEGORIES



(Aggregated over weeks 12-14, 2021)

CENTRE FOR FERTILITY AND HEALTH

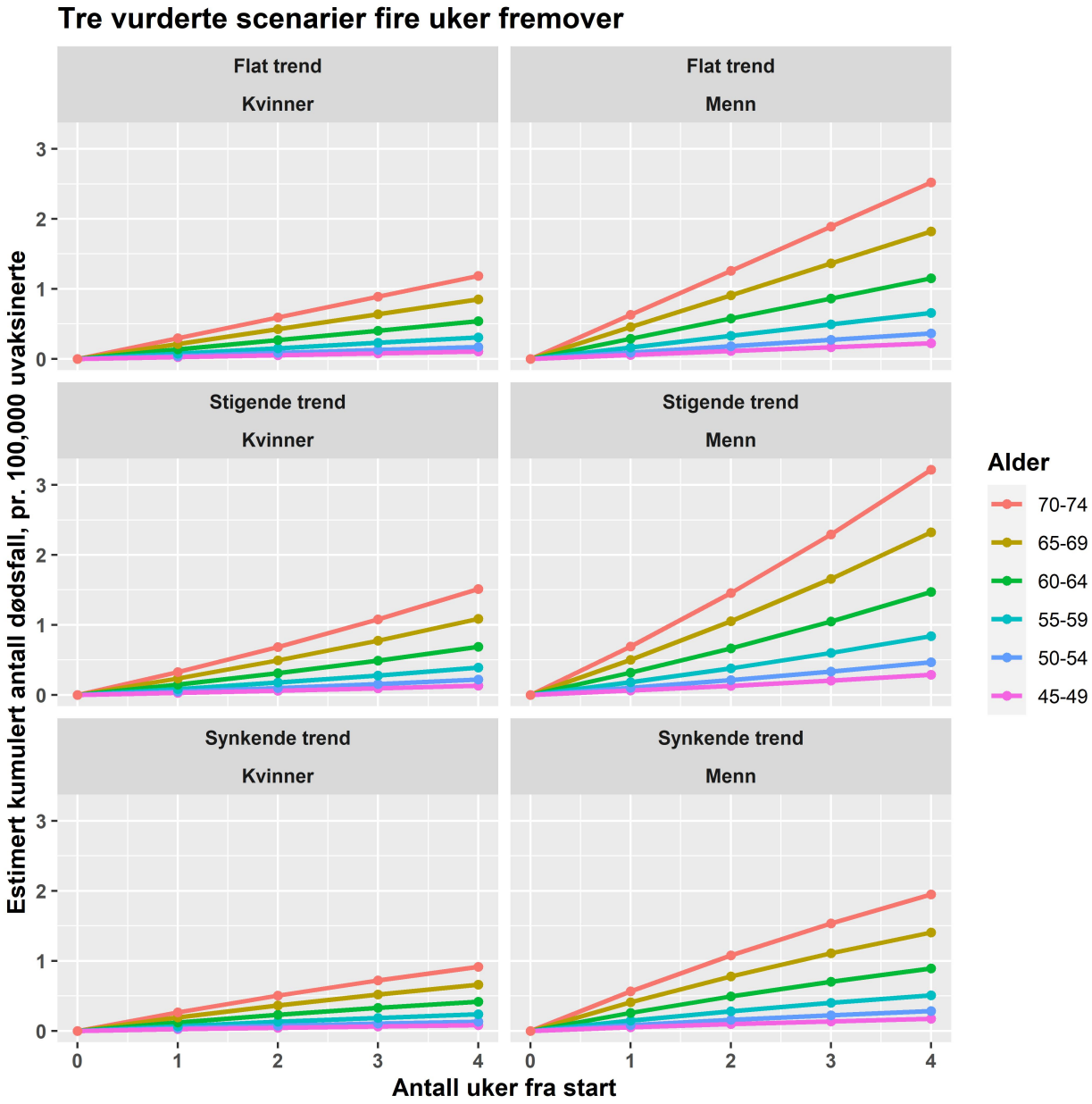
MORTALITY RATES IN AGE CATEGORIES, WITH INTERACTION



(Aggregated over weeks 11-13, 2021)

CENTRE FOR FERTILITY AND HEALTH

LOOKING INTO THE FUTURE(?)



CENTRE FOR FERTILITY AND HEALTH