Intro til statistik

V Logistisk regression

Claus Thorn Ekstrøm
Biostatistik, KU
ekstrom@sund.ku.dk

Mandag 25. maj 2020

Slides @ biostatistics.dk/puff/

Plan for i dag

- Logistisk regression
- Fejlgruber

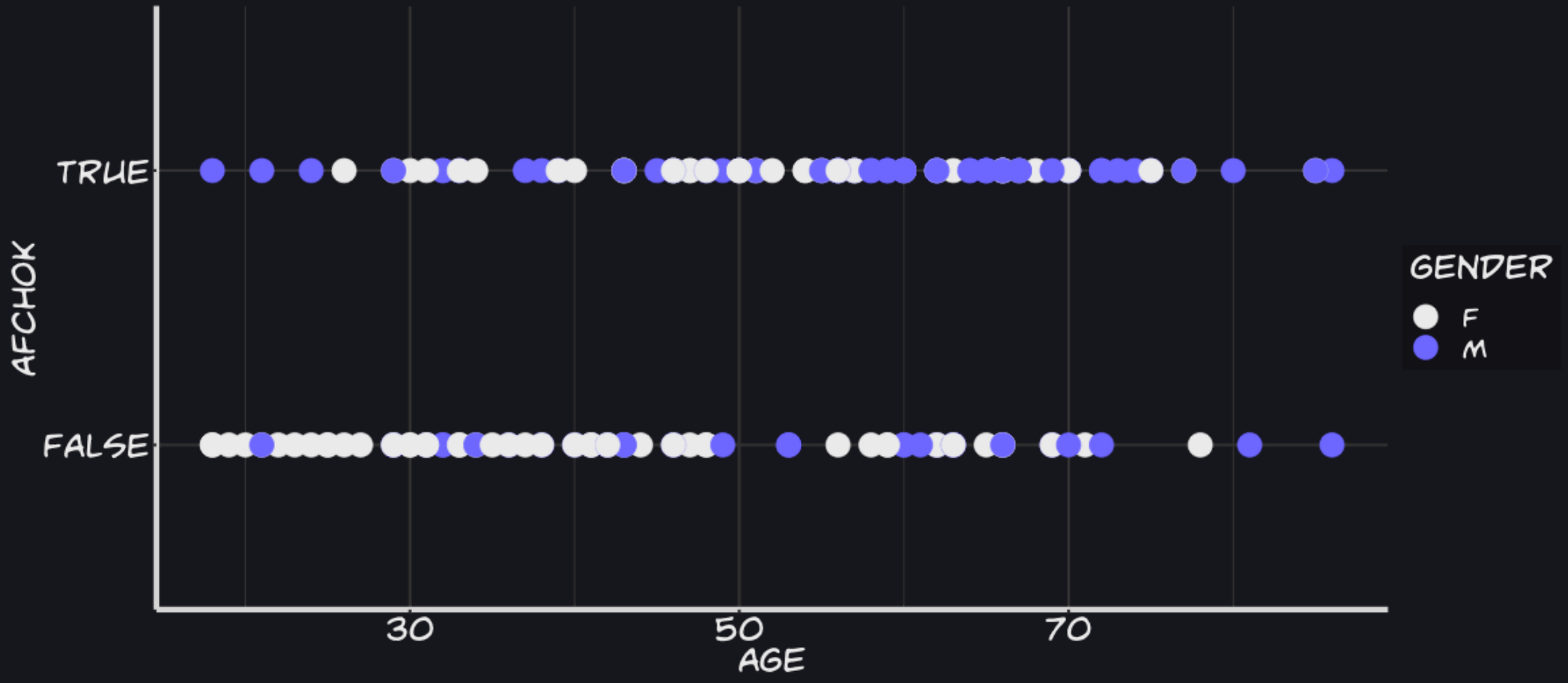
Anæstesi og allergiske reaktioner

```
indata <- read.table("https://publicifsv.sund.ku.dk/~lts/basal14_1/hjemmeopgav
head(indata, 3)
```

```
##   id gender age reacclass tryptase
## 1  1      M  60         3    58.10
## 2  2      F  26         2    22.20
## 3  3      M  39         3     8.86
```

- `reacclass` Sværhedsgraden af den allergiske reaktion, 1: "Mild", 2: "More serious", 3: "Anafylactic shock"
- `tryptase` Den registrerede værdi af serum tryptase

Anæstesi og allergiske reaktioner



Analyse af data

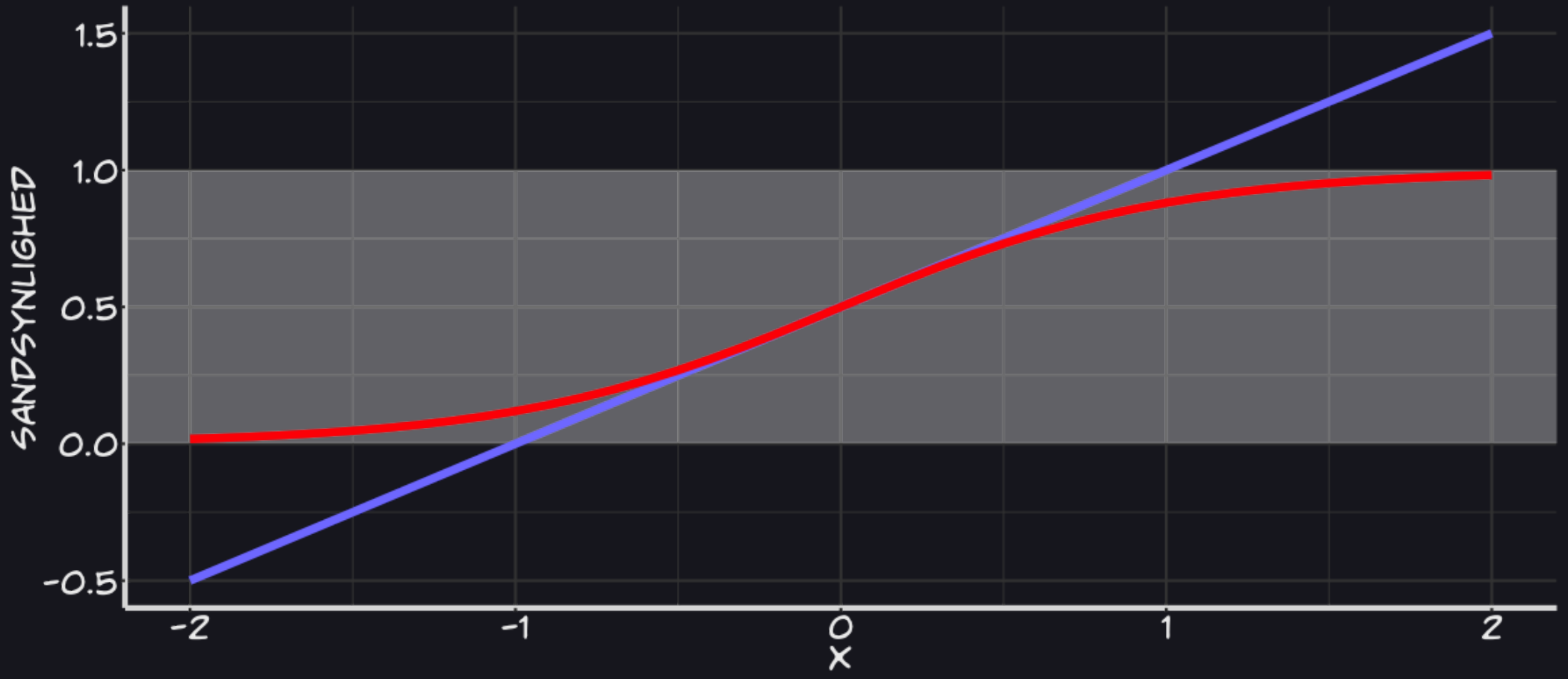
Den forklarende variabel er kontinuert, så χ^2-tabel virker ikke.

Logistisk regression er en regressionsmodel, hvor den afhængige variabel er binær. Konceptuelt lig lineær regression.

I *lineær regression* beskrives, hvordan den gennemsnitlige værdi af den afhængige variabel afhænger af uafhængige variable.

I *logistisk regression* beskrives, hvordan odds (indirekte sandsynligheden) for den afhængige variabel afhænger af uafhængige variable.

Lav ikke lineær regression på binære data



Odds

Odds for en hændelse A er

$$\text{Odds}(A) = \frac{\#\text{personer der oplever A}}{\#\text{personer der ikke oplever A}}$$

Sammenlign med risikobegrebet

$$\text{Risiko}(A) = \frac{\#\text{personer der oplever A}}{\text{total antal personer}}$$

Vigtigt at skelne mellem odds og risiko - det er ikke det samme!

Logistisk regression

En logistisk regressionsmodel er defineret som en lineær sammenhæng på *log odds*-skala.

$$\log \text{odds}(y) = a + b \cdot x$$

- log *Odds* er ubegrænsede positive og negative tal, så vi kan bruge en ret linje på denne skala
- effekten af den uafhængige variabel kan beskrives som en relativ forskel i odds: en *odds-ratio*
- kan håndtere flere forklarende variable

Logistisk regression i R

```
model <- glm(afchok ~ age + gender, data=indata,
             family=binomial())
library("broom")
model %>% tidy()
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -1.85      0.533      -3.48 0.000504
## 2 age           0.0369    0.0104      3.56 0.000374
## 3 genderM       0.576     0.326       1.77 0.0774
```

Fortolkning af estimater for log. regression

Logistisk regression er blot en regressionsmodel på log odds skalaen.

$$\log \text{odds}(y) = a + b \cdot \text{age} + c \cdot I(\text{male})$$

- a er skæringspunktet: log odds når $x = 0$ (og for ref. grp.)
- b er hældningskoefficienten: ændring i log odds for en stigning i x på 1
- c er *forskellen* mellem mænd/kvinder

Fortolkning af estimater for log. regression

Svært at fortolke ændringer i log odds. Tag eksponentialfunktionen på begge sider

$$\text{odds}(y) = \exp(a + b \cdot \text{age} + c \cdot I(\text{male}))$$

En forskel (på log odds skala) på 1 år giver

$$\frac{\text{odds}(y \mid x+1)}{\text{odds}(y \mid x)} = \frac{e^{a+b\cdot(x+1)+c\cdot I(\text{male})}}{e^{a+b\cdot x+c\cdot I(\text{male})}} = e^b$$

En odds-ratio!

```
model %>% tidy()
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -1.85      0.533      -3.48 0.000504
## 2 age           0.0369    0.0104      3.56 0.000374
## 3 genderM       0.576     0.326       1.77 0.0774
```

Odds for anafylaktisk chok blandt mænd er $\exp(0.576) = 1.78$ gange odds blandt kvinder.

Logistisk regression på sandsynlighedsskala

Odds-ratioen er en relativ sammenligning af odds, men fortæller os intet om de absolutte risikoer.

Vi kan prædiktere absolutte risikoer ved transformere den logistiske regressionsmodel over på sandsynlighedsskalaen.

Vi har $\log \text{odds}(y) = a + b \cdot x$. Med andre ord:

$$\frac{P(y=1)}{1-P(y=1)} = e^{a+b \cdot x} \Leftrightarrow P(y=1) = \frac{e^{a+b \cdot x}}{1+e^{a+b \cdot x}}$$

Risikoprædiktioner i R

```
DF <- data.frame(gender="M", age=60)
predict(model, newdata=DF, type="response")
```

```
##         1
## 0.7179943
```

Til sammenligning

```
predict(model, newdata=DF)
```

```
##         1
## 0.9345346
```

Typiske situationer

(som vi ikke har dækket, men som optræder tit)

Sammenligning af målemetoder

Sammenligning af målemetoder (maskiner) eller "dommere" (læger).

- Brug ikke korrelationer
- Brug *ikke* parret t test

For numerisk variable: Bland-Altman plots.

For kategoriske variable: Cohens κ eller Fleiss' κ (men der er lidt problemer for disse to).

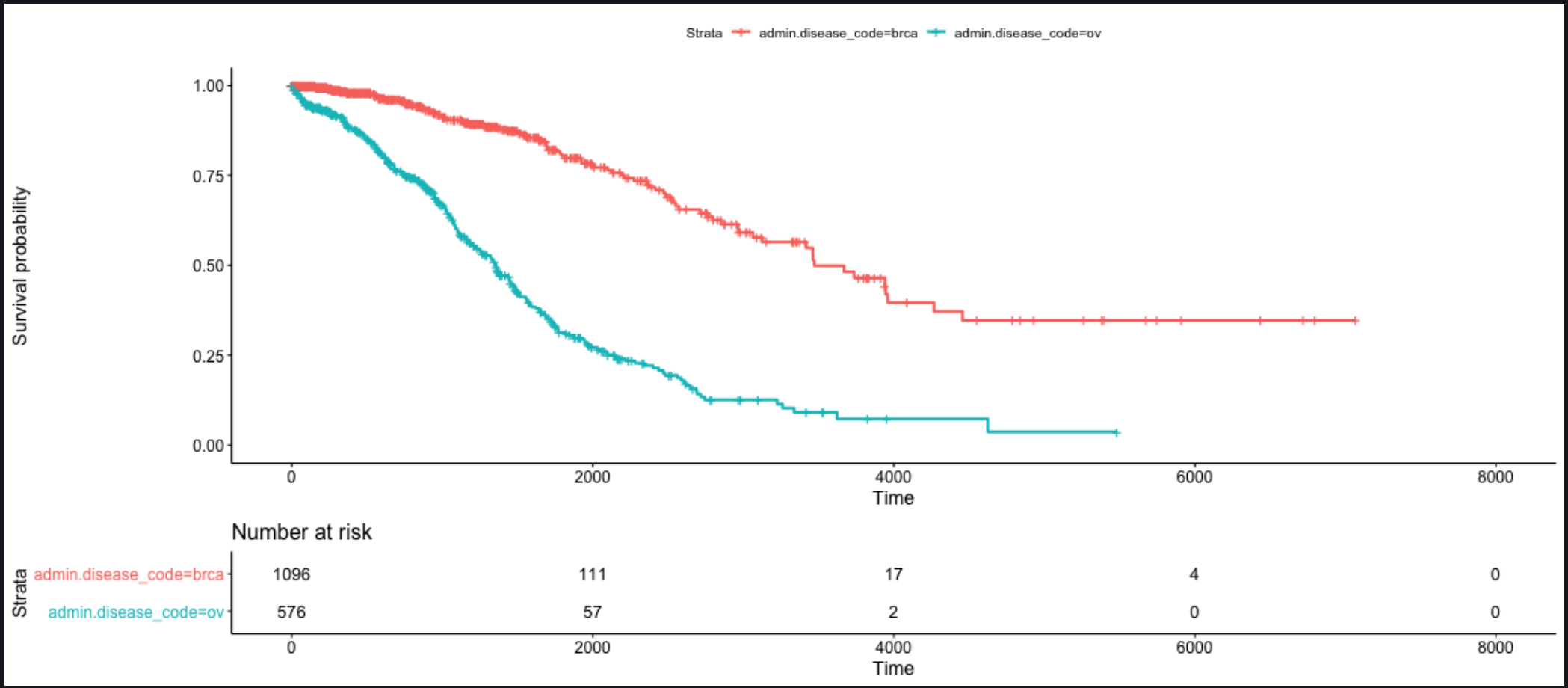
Overlevelsesanalyse

Time-to-event data. Modellerer *tiden* til en hændelse indtræffer.

Ofte problem med **censurering** (specielt med overlevelse som hændelse). Ved, at en person var i live til tid t, og ikke mere.

Competing risks (konkurrerende dødsårsager) komplicerer tingene yderligere (men kan håndteres).

Kaplan-Meier overlevelsesfunktion



Cox regression

```
library("survival")
data("lung")
res.cox <- coxph(Surv(time, status) ~ age + sex +
                 ph.ecog, data =  lung)
res.cox %>% tidy()
```

```
## # A tibble: 3 x 7
##   term    estimate std.error statistic   p.value conf.low conf.high
##   <chr>      <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 age       0.0111   0.00927      1.19 0.232     -0.00710    0.0292
## 2 sex      -0.553    0.168       -3.29 0.000986  -0.881     -0.224
## 3 ph.ecog   0.464    0.114        4.08 0.0000445  0.241      0.686
```

Gentagne målinger

Har indtil nu antaget uafhængige data.

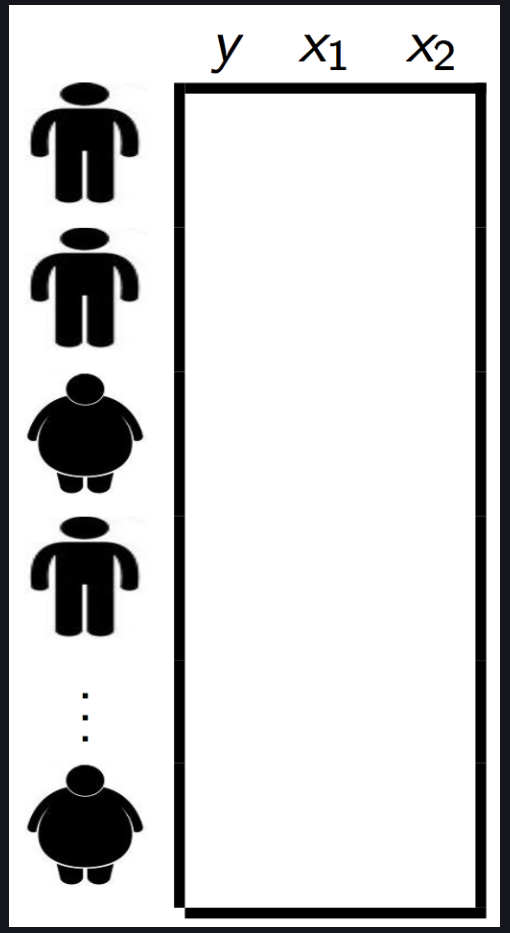
Gentagne målinger er et ret bredt begreb og optræder ofte:
overkrydsningsforsøg, follow-up / center / clusters

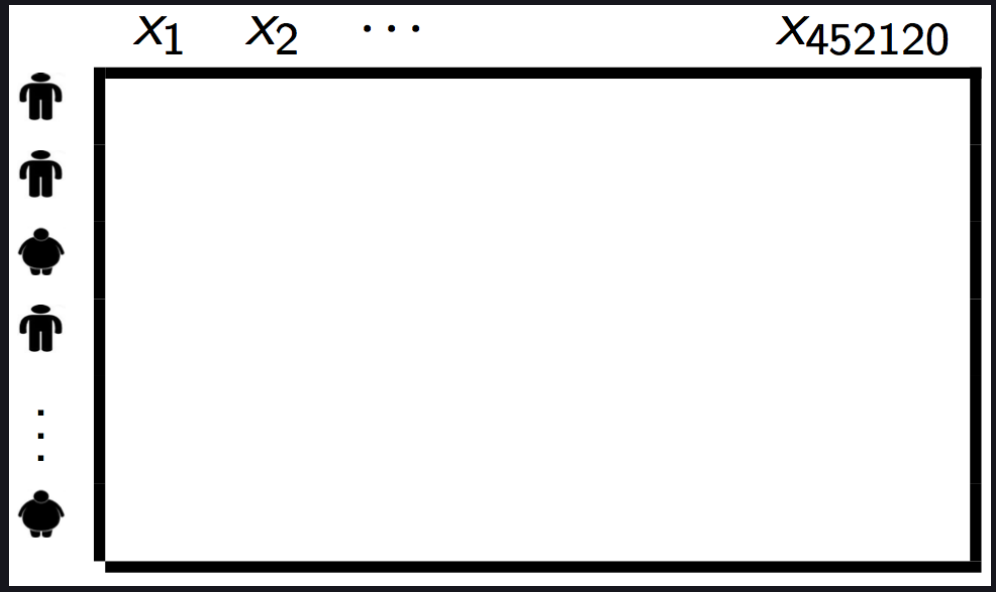
Generelt en dum ide at

- Ignorere problemet
- Opsummere den fulde profil i et enkelt tal

Se fx på `lme4` pakken og `lmer()`.

Højdimensionelle data





Hjælp!

www.biostat.ku.dk

www.sandsynligvis.dk