Biostatistika

Loeng 1. Haiguste esinemissageduse mõõtmisest

Märt Möls
martm@ut.ee

Mis on biostatistika?

Statistiliste meetodite kasutamine meditsiini või bioloogia jaoks aktuaalsete probleemide lahendamiseks.

Miks on biostatistika eraldi õppeaine?
Miks lihtsalt statistika loengutest ei piisa?

- Probleemid terminoloogiaga (arstide keel võib olla statistikute jaoks raskesti arusaadav ja ka vastupidi: „leia usaldusintervall levimusele“ vs „leia usaldusintervall tõenäosusele“);
- Andmete eripära ja tausta mõistmine on hädavajalik korrektse statistilise analüüsi läbiviimiseks;
- Uurimisküsimuste tõlkimist statistika/matemaatika keelde tuleb samuti õppida/harjutada. Samuti tuleb harjutada seda, kuidas analüüsi tulemusi arsti/bioloogi jaoks arusaadavaks teha.

Levimus (*prevalence*)

$$P\,(\text{populatsioonist juhuslikult valitu on haige}\,) = \frac{\text{haigete arv populatsioonis}}{\text{populatsiooni suurus}}$$

HIV levimus Eestis:

kõik täiskasvanud: ~ 1%
Tallinna prostituudid (2016): ~ 6%
süstivad narkomaanid, Tallinn (2017): 54%

Süsteemne erütematoosluupus	46 juhtu 100,000 inimese kohta
Epilepsia	5,3/1000
Diabeet (Väike-Maarja, 2008)	8,7%

Levimus (*prevalence*), usaldusintervall

Levimus = P(juhuslikult valitud inimene populatsioonist on haige)

Teame:

Valimi suurus: *n*
haigete osakaal valimis: *PR*

Levimus on tõenäosus (ehk seega keskväärtus):

$$X = \begin{cases} 0, & \text{kui pole haige} \\ 1, & \text{haige} \end{cases}$$

$$\mathrm{E}(X) = 0 \cdot P(X=0) + 1 \cdot P(X=1) = P(X=1).$$

Suure valimi korral $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$

(valim enamasti suur)

Levimus (*prevalence*), usaldusintervall

Levimus = P(juhuslikult valitud inimene populatsioonist on haige)

Teame:

Valimi suurus: *n*
haigete osakaal valimis: *PR*

Levimus on tõenäosus (ehk seega keskväärtus):

$$X = \begin{cases} 0, & \text{kui pole haige} \\ 1, & \text{haige} \end{cases}$$

$$\mathrm{E}(X) = 0 \cdot P(X=0) + 1 \cdot P(X=1) = P(X=1).$$

Suure valimi korral $\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$

(valim enamasti suur)

n=160

$t_{0,025;df=159} \approx 1,97$

$z_{0,025} \approx 1,96$

Levimus (*prevalence*), usaldusintervall

Levimus = P(juhuslikult valitud inimene populatsioonist on haige)

Teame:

Valimi suurus: *n*
haigete osakaal valimis: *PR*

Levimus on tõenäosus (ehk seega keskväärtus):

$$X = \begin{cases} 0, & \text{kui pole haige} \\ 1, & \text{haige} \end{cases}$$

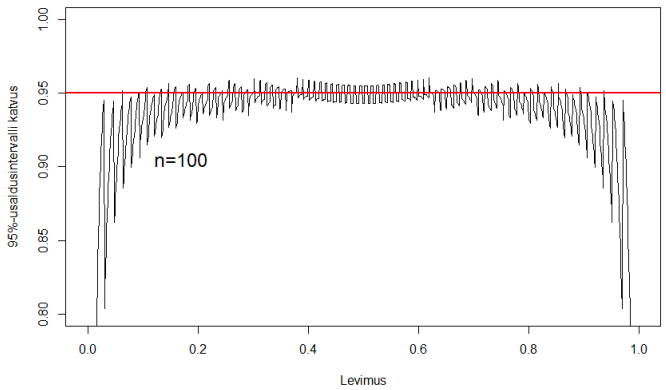
$$\mathrm{E}(X) = 0 \cdot P(X=0) + 1 \cdot P(X=1) = P(X=1).$$

Suure valimi korral $PR \pm z_{\alpha/2}\frac{\sqrt{PR(1-PR)}}{\sqrt{n}}$

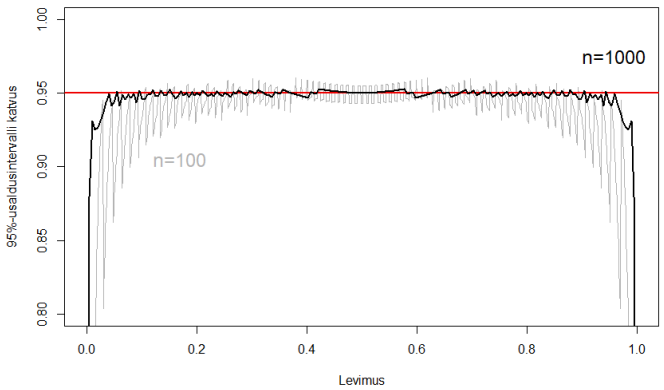
95%-ligikaudne usaldusintervall levimusele

$PR \pm 1,96\sqrt{PR(1-PR)/n}$

Ligikaudne (normaaljaotusel põhinev) 95%-usaldusintervall



Ligikaudne (normaaljaotusel põhinev) 95%-usaldusintervall



Levimus (*prevalence*), Clopper-Pearsoni täpne usaldusintervall

Ligikaudne (asümptootiline)

$$PR \pm z_{\alpha/2}\sqrt{\frac{PR(1-PR)}{n}}$$

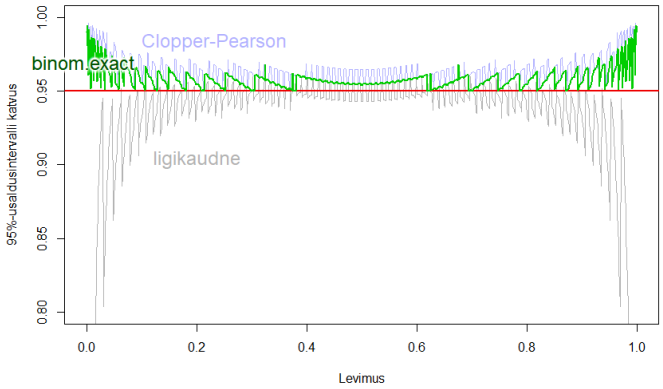
```
0,0009 ... 0,0791
```

Clopper-Pearson'i (täpne) usaldusintervall tõenäosusuele

haigete arv, uuritavaid (*n*)

```
binom.test(4, 100)
0,0110 ... 0,0993
```

Aga eksisteerib ka täpsemaid täpseid usaldusintervalle....



Levimus (*prevalence*), täpsed usalduspiirid

Ligikaudne (asümptootiline)

$$PR \pm z_{\alpha/2}\sqrt{\frac{PR(1-PR)}{n}}$$

```
0,0009 ... 0,0791
```

Clopper-Pearson'i (täpne) usaldusintervall tõenäosusuele

haigete arv, uuritavaid (*n*)

```
binom.test(4, 100)
0,0110 ... 0,0993
```

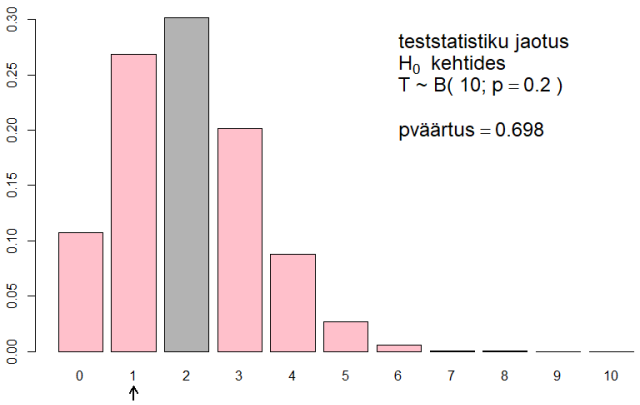
Täpsem täpne usaldusintervall tõenäosusuele (kahepoolse täpse testi pööramine):

```
> library("exactci")
> binom.exact(4, 100, tsmethod="minlike")
0,0138 ... 0,0986
```

Levimus (*prevalence*), täpsed usalduspiirid, näide

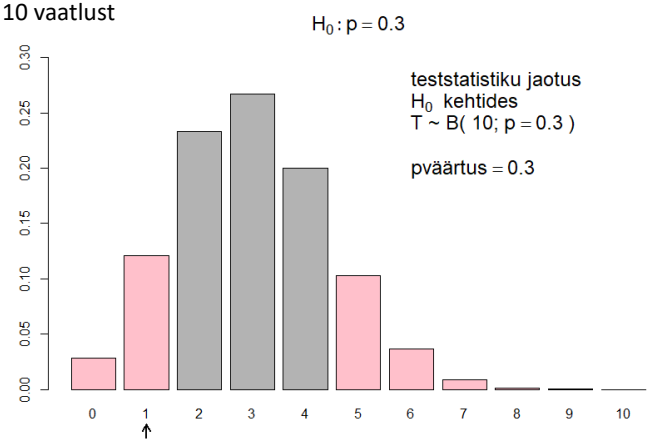
Vaatlused: 1 juhtum; 10 vaatlust

$H_0 : p = 0.2$



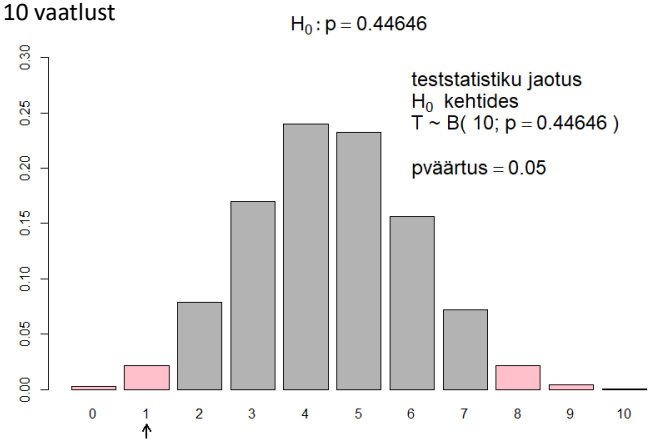
Levimus (*prevalence*), täpsed usalduspiirid, näide

Vaatlused: 1 juhtum; 10 vaatlust



Levimus (*prevalence*), täpsed usalduspiirid, näide

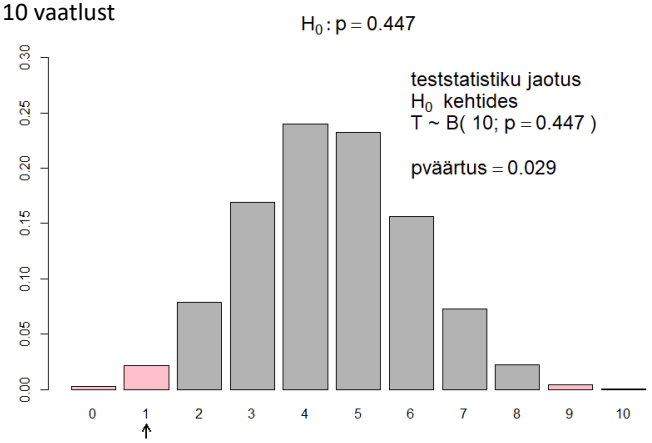
Vaatlused: 1 juhtum; 10 vaatlust



Levimus (*prevalence*), täpsed usalduspiirid, näide

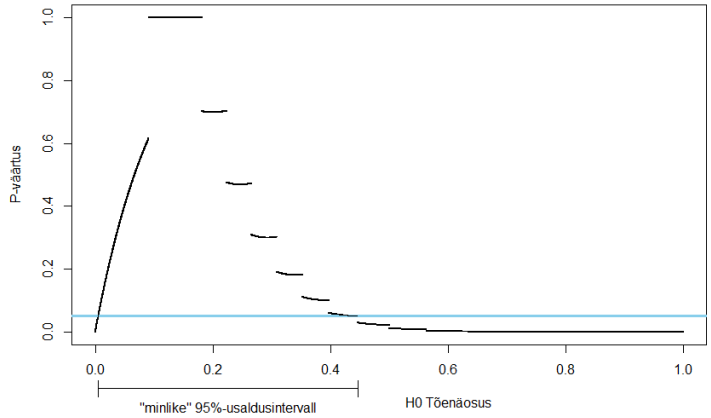
Vaatlused: 1 juhtum; 10 vaatlust

Järelikult 0,447 ei kuulu
enam 95%-usaldusintervalli...



Levimus (*prevalence*), täpsed usalduspiirid, näide

Vaatlused: 1 juhtum; 10 vaatlust



Kuna kasutada usaldusintervalle?

Hüpoteetiline näide:

```
Eesti elanike arv 1. jaanuaril 2020:    1 328 360
       haigeid Eestis (kõik haiged):          100
```

Kas haiguse levimus on 7,52808 juhtu 100 000 inimese kohta või 7,5 (6,1… 9,2) juhtu 100 000 inimese kohta?

```
Eesti elanike arv 2. jaanuaril 2020:    1 328 361
       haigeid Eestis (kõik haiged):          100
```

Haiguse levimus
7,528074 juhtu 100 000 inimese kohta:
MEDITSIINISÜSTEEM/KESKKOND ON EESTIS TÕESTATAVALT PAREMAKS MUUTUNUD!

Haiguse levimus
7,5 (95%-UI: 6,1…9,2) juhtu 100 000 inimese kohta:
Tõestatavaid muutuseid pole!

Kuna kasutada usaldusintervalle?

Hüpoteetiline näide:

```
Eesti elanike arv 1. jaanuaril 2020:    1 328 360
       haigeid Eestis (kõik haiged):          100
```

Kas haiguse levimus on 7.52808 juhtu 100 000 inimese kohta või 7,5 (6,1… 9,2) juhtu 100 000 inimese kohta?

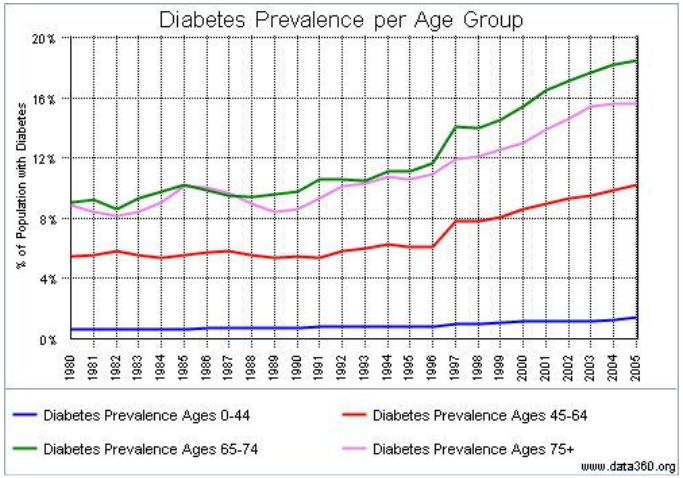
```
Eesti elanike arv 2. jaanuaril 2020:    1 328 361
       haigeid Eestis (kõik haiged):          100
```

Sageli tuleb kõiki riigis elavaid inimesi vaadelda kui ühte suurt juhuslikku valimit (juhuslikult valitud inimesi, kes on teatud keskkonda elama pandud). Järeldusi tahetakse enamasti teha keskkonna kui sellise kohta (aga esineb erandeid).

Levimuse mõõtmise erivormid:

- hetkelevimus (*point prevalence*)
  – kui suur osa elanikkonnast on haige mingil ajahetkel
- perioodlevimus (*period prevalence*)
  – tõenäosus põdeda mingit haigust mingi ajaperioodi jooksul (kui suur osa elanikkonnast kannatab külmetushaiguste all aasta jooksul)
- *Lifetime prevalence*
  – milline on tõenäosus inimesel oma elu jooksul antud haigusega kokku puutuda, haiguse esinemine inimese elu jooksul

Kas diabeeti ikkagi esineb tänapäeval rohkem või püsivad diabeetikud kauem elus?



Esmashaigestumus (*Incidence*)

Vahel ka: avaldumus

Variant 1 - *Incidence Cases; Incidence*
uute haigusjuhtude ehk haiguse esmasjuhtude arv mingil ajavahemikul

Näiteks Eestis (2020) – uued haigusjuhud:

Suhkrutõbi (2020)	4 501
Infarkte (2016)	2 826
Parkinson (2016)	825
COVID-19 (2021)	213 772

Raske (mõttetu) võrrelda:
USA uusi vähijuhte 1975: ~ 650 000
USA uusi vähijuhte 2009: ~ 852 000

elanike arv 1975: 216 mil.
2009: 307 mil.

Esmashaigestumus (*Incidence*)

Variant 2 – Haigestumusrisk (kumulatiivavaldumus; avaldumusrisk)

Incidence Proportion (IP); Cumulative Incidence; Incidence Risk
haigestumuskordaja, mis väljendab uute haigusjuhtude arvu rahvaarvu suhtes mingi ajavahemiku jooksul

$$IP = \frac{\text{uute haigusjuhtude arv ajavahemikus}}{\text{inimeste arv ajavahemiku alguses}}$$

Esitatakse sageli kui juhte 100 000 või 1 000 või … inimese kohta aastas, või kolme aasta jooksul, või viie aasta jooksul,…

Näiteks Eestis (2016) – uued haigusjuhud (100 000 elaniku kohta aastas):

Infarkte:	208
Suhkrutõbi	561
Parkinson	63
Kopsupõletik	975

Esmashaigestumus (*Incidence*)

Variant 3. Esmashaigestumuskordaja – *Incidence Rate (IR)*

haigestumuskordaja, mis väljendab haigusjuhtude arvu riski all oldud aja kohta

$$IR = \frac{\text{uute haigusjuhtude arv}}{\text{riskiaeg}} = \frac{I}{riskiaeg}$$

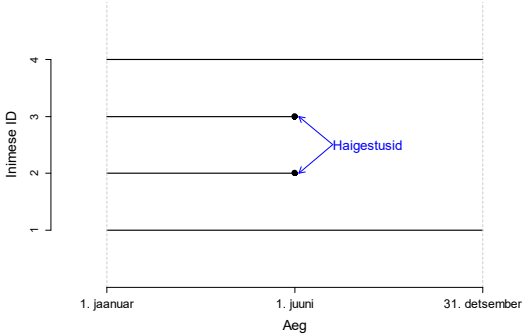
Riskiaeg (*risk time, time at risk*) – haigestuda võivate (nn *vastuvõtlike*) inimeste poolt kokku vaadeldud ajavahemikul elatud aeg. Kasutatakse ka mõisteid *person-years* või *person-time*.

Incidence Proportion 1000 inimese kohta ei saa kunagi olla suurem kui 1000, Incidence Rate võib aga põhimõtteliselt olla kuitahes suur...

Esmahaigestumus (*Incidence*)

Näide – IR ja IP võrdlus

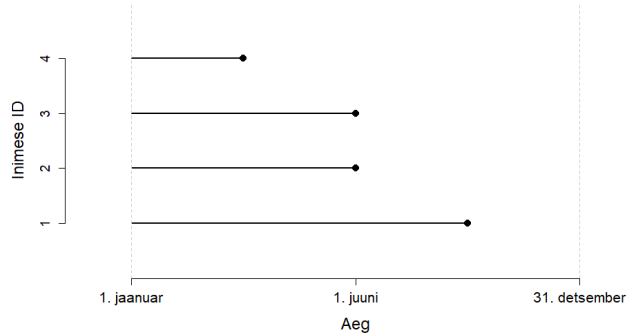
IP = 2 / 4 = 0,5
IR = 2 / 3 = 0,666…



Esmahaigestumus (*Incidence*)

Näide – IR ja IP võrdlus

IP = 4 / 4 = 1
IR = 4 / 2 = 2



Esmahaigestumus (*Incidence*)

CI: cumulative incidence e haigestumusrisk
IR: Incidence rate e esmahaigestumuskordaja

Relation between CI and IR

$\boldsymbol{CI(t) = 1 - e(-IR \times t)}$, where e is the base to the natural logarithm (2.718) and t is the time unit of concern. When the CI < 0.10 (10%), the formula approximate to:

$\exp(\varepsilon) \approx 1 + \varepsilon$

$CI(t) = IR \times t$

X - aeg haigestumiseni

$X \sim \mathrm{Exp}(\lambda)$

Incidence rate

$$F_X(x) = 1 - \exp(-\lambda \cdot x)$$

Esmahaigestumus (*Incidence*)

X - aeg haigestumiseni

$X \sim \mathrm{Exp}(\lambda)$

Incidence rate

$$f_X(x) = \lambda \exp(-\lambda \cdot x)$$
$$F_X(x) = 1 - \exp(-\lambda \cdot x)$$

Suurima tõepära hinnang

$$L = \lambda \exp(-\lambda \cdot x_1) \cdot \ldots \cdot \lambda \exp(-\lambda \cdot x_n) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} \cdot x_i\right)$$

$$l = n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i \qquad \frac{\delta l}{\delta \lambda} = n/\lambda - \sum_{i=1}^{n} x_i$$

$$\frac{\delta l}{\delta \lambda} = 0 \quad \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

Esmahaigestumus (*Incidence*)

X - aeg haigestumiseni

$X \sim \mathrm{Exp}(\lambda)$

Incidence rate

$$f_X(x) = \lambda \exp(-\lambda \cdot x)$$
$$F_X(x) = 1 - \exp(-\lambda \cdot x)$$

Suurima tõepära hinnang, kui osad vaatlused on tsenseeritud

toimus haigestumine (teame aega haigestumiseni)

haigestumine vaadeldud aja jooksul aset ei leidnud

$$L = \lambda \exp(-\lambda \cdot x_1) \cdot \ldots \cdot \lambda \exp(-\lambda x_{n_h}) \cdot P(X > x_{nh+1}) \cdot \ldots P(X > x_n)$$

$1 - F_X(x_n)$

$$l = n_h \log(\lambda) - \lambda \sum_{i=1}^{n} x_i \qquad \hat{\lambda} = \frac{n_h}{\sum_{i=1}^{n} x_i}$$

Esmahaigestumus (*Incidence*)

X - aeg haigestumiseni

$X \sim \text{Exp}(\lambda)$

Incidence rate

$$f_X(x) = \lambda \exp(-\lambda \cdot x)$$
$$F_X(x) = 1 - \exp(-\lambda \cdot x)$$

hinnang esmahaigestumuskordajale IR

nähtud haigusjuhtude arv

$$\hat{\lambda} = \frac{n_h}{\sum_{i=1}^{n} x_i}$$

jälgimisaeg

Esmahaigestumus (*Incidence*)

X - aeg haigestumiseni

$X \sim \text{Exp}(\lambda)$

Incidence rate

$$f_X(x) = \lambda \exp(-\lambda \cdot x)$$
$$F_X(x) = 1 - \exp(-\lambda \cdot x)$$

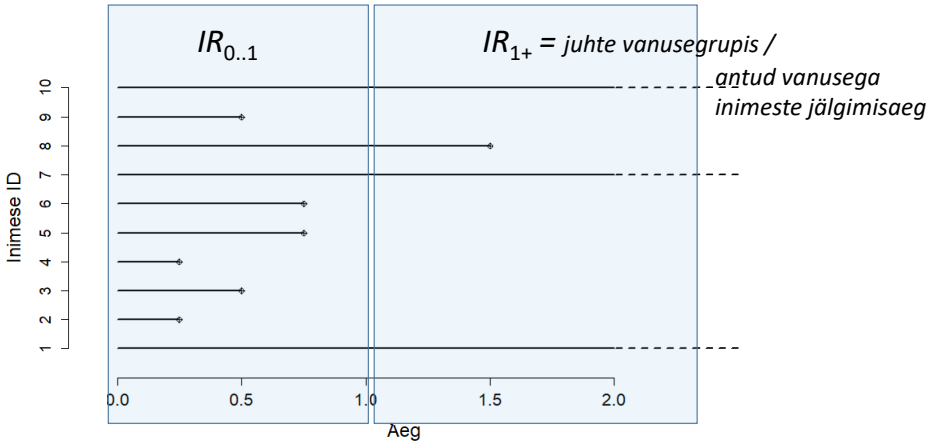
Eksponentjaotus on mäluta ehk mittevananev jaotus – tõenäosus haigestuda järgmisel aastal ei sõltu sellest, kui kaua keegi on haigust suutnud vältida:

$$P(X < t+1 | X > t) = P(X < 1)$$

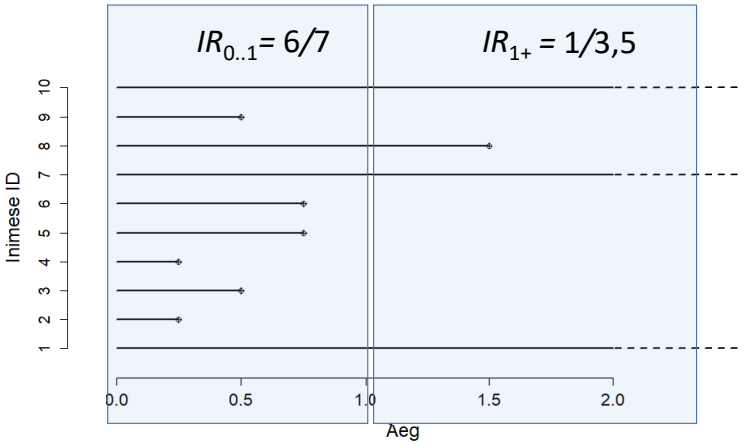
Selline tingimus pole aga sageli rahuldatud – näiteks aasta jooksul peale operatsiooni on komplikatsiooni esinemise tõenäosus märksa suurem kui järgmisel aastal peale operatsiooni...

Lahendus: vanusest (ajast) sõltuvad haigestumuskordajad...

Esmahaigestumus (*Incidence*): haigestumuse vanusekordaja



Esmahaigestumus (*Incidence*): haigestumuse vanusekordaja

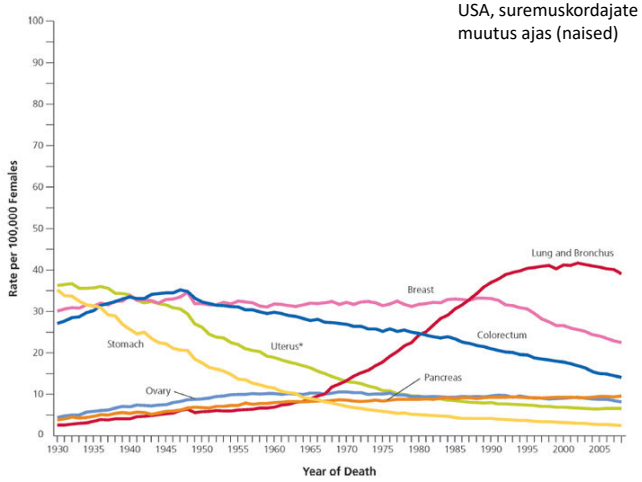


Suremuskordaja (*mortality rate / death rate*):

MR = surmade arv mingil ajavahemikul / riskiaeg

Nimetus muutub, matemaatiline sisu/sobivad meetodid ei muutu.

USA, suremuskordajate muutus ajas (naised)



Haigestumus(suremus)kordaja oma toorel kujul võib kergesti inimesi eksiteele viia. Näiteks oli suremuskordaja Inglismaal 1901. aastal 15,7 (1000 inimese kohta aastas), 1981. aastal aga 15,6 (1000 inimese kohta aastas). Kas see tähendab, et asjalood olid 1981.a. sama halvad kui 1901. aastal?

Vaatame suremust vanusegrupiti:

Vanusegrupp	suremuskordaja (MR)		vanusegrupi osakaal %	
	1901	1981	1901	1981
15-19	3,5	0,8	15,36	11,09
20-24	4,7	0,8	14,07	9,79
25-34	6,2	0,9	23,76	18,81
35-44	10,6	1,8	18,46	15,99
45-54	18,0	6,1	13,34	14,75
55-64	33,5	17,7	8,68	14,04
65-74	67,8	45,6	4,57	10,65
75-84	139,8	105,2	1,58	4,28
84+	276,5	226,2	0,17	0,64

suremuskordaja või esmashaigestumiskordaja leidmine kahe vanusegrupi korral

$$IR = (I_1 + I_2)/(aeg_1 + aeg_2)$$
$$= I_1/(aeg_1 + aeg_2) + I_2/(aeg_1 + aeg_2)$$
$$= I_1/aeg_1 \cdot aeg_1/(aeg_1 + aeg_2) + I_2/aeg_2 \cdot aeg_2/(aeg_1 + aeg_2)$$
$$= IR_1 \cdot osakaal_1 + IR_2 \cdot osakaal_2$$

Suremus(haigestumus)kordaja standardimine I

Leiame, milline oleks olnud 1901. aasta suremuskordaja, kui ühiskonna vanuseline struktuur oleks olnud samasugune kui 1981.a.

Vanusegrupp	vanuserühma suremuskordaja aastal 1901	vanuserühma osakaal 1981.a.	
15-19	3,5	0,1109	3,5*0,1109=0,38815
20-24	4,7	0,0979	0,46013
25-34	6,2	0,1881	1,16622
35-44	10,6	0,1599	1,69494
45-54	18,0	0,1475	2,65500
55-64	33,5	0,1404	4,70340
65-74	67,8	0,1065	7,22070
75-84	139,8	0,0428	5,98344
84+	276,5	0,0064	1,76960
MR			**26,04**

Ehk kui peaksime praeguse elanikkonnaga hakkama saama 1901.a meditsiini ja elutingimustega, sureks aastas 1000 inimese kohta 10 inimest rohkem kui praegu.

Suremus(haigestumus)kordaja standardimine II

Kahjuks tuleb sageli ette olukordi, kus pole selge, millise riigi või rahvastikugrupi suremusega millisel aastal lugejad uuringu tulemust võrdlema soovib hakata. Lahenduseks on nn. standardrahvastiku kasutamine: tulemuste võrreldavaks tegemiseks kasutatakse etteantud teoreetilist vanusestruktuuri.

Vanuserühm	Maailm (Segi, 1960)	WHO World (2000-2025)	Euroopa (ESP, 1976)	EU-27+EFTA (2011-2030)
0-4	12 000	88 569	8 000	5 226
5-9	10 000	86 870	7 000	5 334
10-14	9 000	85 970	7 000	5 343
15-19	9 000	84 670	7 000	5 401
20-24	8 000	82 171	7 000	5 727
25-29	8 000	79 272	7 000	6 210
30-34	6 000	76 073	7 000	6 664
35-39	6 000	71 475	7 000	6 953
40-44	6 000	65 877	7 000	7 030
45-49	6 000	60 379	7 000	7 012
50-54	5 000	53 681	7 000	6 884
55-59	4 000	45 484	6 000	6 636
60-64	4 000	37 187	5 000	6 247
65-69	3 000	29 590	4 000	5 606
70-74	2 000	22 092	3 000	4 772
75-79	1 000	15 159	2 000	3 811
80-84	500	9 097	1 000	2 719
85-	500	6 348	1 000	2 425
Kokku	**100 000**	**1 000 000**	**100 000**	**100 000**

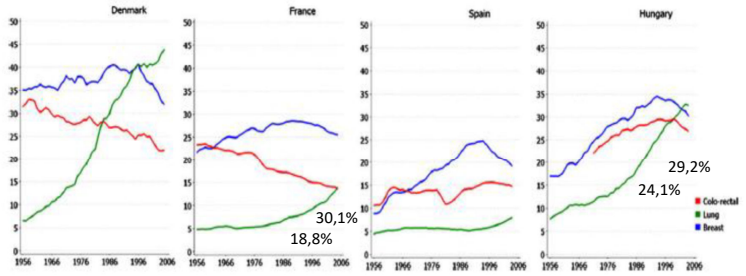


Fig. 4b – Trends in age-standardised (European standard) mortality rates from colorectal, lung and breast cancers in females in Denmark, France, Spain and Hungary (Source: WHO mortality database[8]).

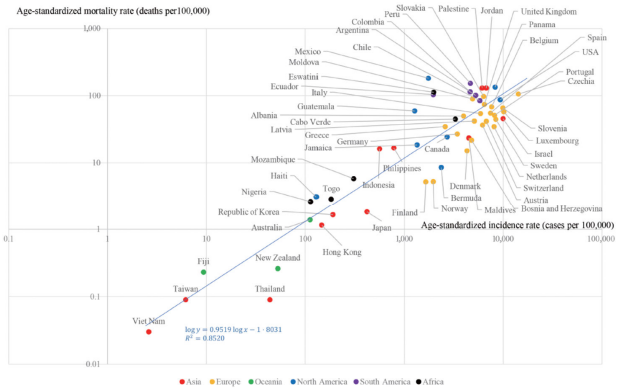


Figure 1. Age-standardized incidence and mortality rates (all ages). R^2, coefficients of determination.

Usaldusintervallide leidmisel alustame tähelepanekust, et uute haigusjuhtude arv (*I*) on enamasti (vähemalt ligikaudu) Poissoni jaotusega juhuslik suurus:

$$I \sim \text{Poi}(IR^* \, riskiaeg);$$

Poissoni jaotuse korral on aga juhusliku suuruse dispersioon võrdne keskväärtusega ehk Poissoni jaotuse parameetriga:

$$\text{D}(I) = IR^* \, riskiaeg.$$

Seega

$$\text{D}(\widehat{IR}) = \text{D}(I/riskiaeg) = \text{D}(I)/riskiaeg^2 = IR/riskiaeg = IR^2/I$$

ja ligikaudne 95%-usaldusintervall tegelikule esmahaigestumusele on leitav valemiga

$$\widehat{IR} \pm 1,96\frac{\widehat{IR}}{\sqrt{I}}$$

Standardiseeritud suremuskordaja täpsus

Eri vanuserühmadele leitud suremuskordaja hinnangud on sõltumatud. Seega:

$$\text{D}(0,25\ \text{MR}_{\text{noored}} + 0,55\ \text{MR}_{\text{keskealised}} + 0,2\ \text{MR}_{\text{vanad}}) =$$
$$0,25^2\ \text{D}(\text{MR}_{\text{noored}}) + 0,55^2\ \text{D}(\text{MR}_{\text{keskealised}}) + 0,2^2\ \text{D}(\text{MR}_{\text{vanad}}),$$

kus $\text{D}(\text{MR}_{\text{noored}})$ on noorte suremuskordaja dispersioon (arvutatud samamoodi kui esmashaigestumuskordaja dispersioon). Kui valimimahud on (enamikes vanuserühmades) piisavalt suured võib standardiseeritud haigestumuskordaja hinnangut vaadelda kui normaaljaotusega juhuslikku suurust.