Introducerende Statistik og Dataanalyse med R

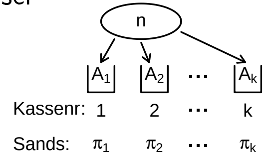
Multinomialmodellen

Jens Ledet Jensen

I DAG

Binomial: falde i en ud af to kasser

Multinomial: falde i en ud af 6 kasser



Multinomial: Inddele data i k kasser

Teste hypotese om sandsynlighederne for at falde i kasserne

Eksempel: Teste at en terning er "fair"

seks kasser, teste sandsynlighed er $\frac{1}{6}$ for hver kasse

Multinomialfordeling

Binomialmodel: n delforsøg hver med to kasser (1 eller 2)

$$X = \text{antal gange hvor kasse 1 rammes } (n - X \text{ rammer kasse 2})$$

Multinomial: n delforsøg hver med k kasser $(1, 2, \ldots, k)$

$$A_j = \text{antal gange kasse } j \text{ rammes:} \quad A_1 + A_2 + \cdots + A_k = n$$

$$\pi_j\text{: sandsynlighed for at ramme kasse } j\text{:} \quad \pi_1 + \pi_2 + \cdots + \pi_k = 1$$

$$(A_1, A_2, \ldots, A_k) \sim \mathrm{multinom}(n, (\pi_1, \pi_2, \ldots, \pi_k))$$

I ord: multinomialfordelt med antalsværdi n og sandsynlighedsparameter $\pi = (\pi_1, \ldots, \pi_k)$

Eksempel: Embryoer af zebrafisk i nanopartikelopløsning

Dødstidspunkt delt op på 3 kasser:

kasse 1: 0-48 timer, kasse 2: 48-96 timer, kasse 3: over 96 timer

$a_1 = 15, \quad a_2 = 19, \quad a_3 = 66, \qquad n = 100$

Hypotese: Samme dødsrate i 48-96t som i 0-48t

Hypotese: $P(\text{kasse 1}) = \theta$, $P(\text{kasse 2}) = (1-\theta)\theta$, $P(\text{kasse 3}) = (1-\theta)^2$

Skøn over θ: $\hat{\theta} = \frac{a_1+a_2}{a_1+a_2+a_2+2a_3} = 0.1838$ (metode: se senere)

Zebrafisk: test

Forventede under hypotesen:

$$e_1 = n\hat{\theta}, \qquad e_2 = n(1-\hat{\theta})\hat{\theta}, \qquad e_3 = n(1-\hat{\theta})^2$$

Kasse	1	2	3	total
Obs a_j	15	19	66	100
Forv e_j	18.38	15.00	66.62	100

Hvordan skal jeg sammenligne 3 observerede tal med 3 forventede ?

Teststørrelse forklares nedenfor

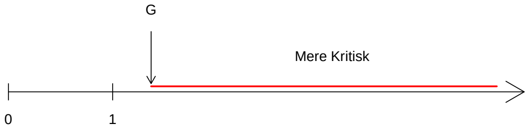
Teststørrelse: $G = 2\Big(a_1 \cdot \log\left(\frac{a_1}{e_1}\right) + a_2 \cdot \log\left(\frac{a_2}{e_2}\right) + a_3 \cdot \log\left(\frac{a_3}{e_3}\right)\Big)$

$$= 2\Big(15 \cdot \log\left(\frac{15}{18.38}\right) + 19 \cdot \log\left(\frac{19}{15.00}\right) + 66 \cdot \log\left(\frac{66}{66.62}\right)\Big) = 1.6522$$

Store værdier er kritiske for hypotesen

Zebrafisk: test

p-værdi
$= P(G(X) \geq 1.6522)$:



hvor ofte vil jeg ved gentagelse af eksperimentet få en værdi større end 1.6522 under forudsætning af samme dødelighed (hypotesen)

P-værdi kan ikke beregnes eksakt. Man kan vise at fordelingen af den stokastiske variabel G ligner en χ^2-fordeling (ki-i-anden)

Approksimation: hvis alle $e_j \geq 5$:

$$p\text{-værdi} \approx 1 - \chi^2_{\text{cdf}}(1.6522, 3-1-1) = 0.199$$

Data strider ikke mod samme dødsrate da p-værdi > 0.05

χ^2-fordeling

CDF: cumulative distribution function, $P(X \leq x)$

sandsynligheden for at ligge til venstre for et punkt

Ki-i-anden fordeling med f frihedsgrader: $\chi^2(f)$

Webbog: $\chi^2_{\text{cdf}}(G, f)$

R: pchisq(G,f)

Hvis $U_1, \dots, U_f$ er uafhængige standard normalfordelte, så er

$$U_1^2 + \cdots + U_f^2 \sim \chi^2(f)$$

R-leg

Prøv at finde:

```
pchisq(1.6522,1)

1-pchisq(3.84,1)

1-pchisq(5.99,2)
```

Prøv følgende:

```
obs=c(15, 19, 66)
forv=c(18.38, 15.00, 66.62)

2*sum(obs*log(obs/forv))

sum((obs-forv)^2/forv)
```

Sceneskift

Multinomialmodellen er indført, G-test er vist i eksempel

Næste: simulere fordeling af G-teststørrelsen

Simulering

```
nsim=1000000
th=0.1838
n=100

x=t(rmultinom(nsim,n,c(th,(1-th)*th,(1-th)^2)))

thhat=(x[,1]+x[,2])/(x[,1]+2*x[,2]+2*x[,3])
e1=n*thhat; e2=n*thhat*(1-thhat); e3=n*(1-thhat)^2
x1=ifelse(x==0,1,x)
G=2*(x[,1]*log(x1[,1]/e1)+x[,2]*log(x1[,2]/e2)+
x[,3]*log(x1[,3]/e3))


G0=ifelse(G<exp(-4.99),exp(-4.99),G)
G0=ifelse(G0>exp(2.99),exp(2.99),G0)
hist(log(G0),breaks=-5+c(0:40)*0.2,probability=TRUE)
curve(dchisq(exp(x),df=1)*exp(x),from=-5,to=3,add=TRUE)
abline(v=log(qchisq(0.95,1)),col=2)
```

Sceneskift

Multinomialmodellen er indført, G-test er vist i eksempel

Næste: detaljer i G-testet

Multinomialmodel: detaljer

$$P\big((A_1, A_2, \ldots, A_k) = (a_1, a_2, \ldots, a_k)\big) = \binom{n}{a_1 \cdots a_k} \pi_1^{a_1} \pi_2^{a_2} \cdots \pi_k^{a_k}$$

Model M_0: π kan variere frit:

$$\pi_j \geq 0, \quad \pi_1 + \pi_2 + \cdots + \pi_k = 1$$

skøn over π_j: observeret frekvens: $\hat{\pi}_j(M_0) = \frac{a_j}{n}, \quad j = 1, \ldots, k$

Dette er også maksimumspunktet for likelihoodfunktionen:

$$L(\pi) = \binom{n}{a_1 \cdots a_k} \pi_1^{a_1} \cdot \pi_2^{a_2} \cdots \pi_k^{a_k}$$

= sandsynligheden for det observerede som funktion af parameteren

Bevis

Hvis $(A_1, A_2, \ldots, A_k) \sim \mathrm{multinom}(n, (\pi_1, \pi_2, \ldots, \pi_k))$ så er

$A_1 \sim \mathrm{binom}(n, \pi_1)$

skøn over π_1: maximere $L_{A_1}(\pi_1) = \pi_1^{A_1}(1-\pi_1)^{n-A_1}$, $\hat{\pi}_1 = \frac{A_1}{n}$

Likelihoodfunktion $L(\pi_1, \ldots, \pi_k) = L_{A_1}(\pi_1) L_{A|A_1}(\pi_1, \ldots, \pi_k)$

betingede: $\frac{\pi_1^{A_1}\pi_2^{A_2}\cdots\pi_k^{A_k}}{\pi_1^{A_1}(1-\pi_1)^{n-A_1}} = \left(\frac{\pi_2}{1-\pi_1}\right)^{A_2} \cdots \left(\frac{\pi_k}{1-\pi_1}\right)^{A_k}$

indfør $v_j = \frac{\pi_j}{1-\pi_1}$, $v_j \geq 0$, $v_2 + \cdots v_k = 1$

π_1 og $(v_2 \ldots v_k)$ varierer uafhængigt af hinanden

Multinomialkoefficient

$\binom{n}{a_1 \cdots a_k} = \frac{n!}{a_1! \cdot a_2! \cdots a_k!}$

Eksempel: $\binom{4}{2,1,1} = \frac{24}{2 \cdot 1 \cdot 1} = 12$

1 1 2 3
1 1 3 2
1 2 1 3
1 3 1 2
1 2 3 1
1 3 2 1
2 1 1 3
3 1 1 2
2 1 3 1
3 1 2 1
2 3 1 1
3 2 1 1

Prøv i R

Firsidet terning kastes 3 gange. Hvad har størst sandsynlighed:

3 ens, 2 ens eller 3 forskellige

gang hver af følgende sandsynlighed med passende tal

```
c(dmultinom(c(3,0,0,0),3,rep(0.25,4)),
dmultinom(c(2,1,0,0),3,rep(0.25,4)),
dmultinom(c(1,1,1,0),3,rep(0.25,4)))
```

Hypotese og test

Hypotese (model M_1): $(\pi_1, \ldots, \pi_k)$ har en bestemt form

$\pi_j = p_j(\theta)$, p_j kendt funktion, θ ukendt parameter

θ har d koordinater, $\theta \in \mathbf{R}^d$

Skøn over θ: den værdi der maksimerer $L(p_1(\theta), \ldots, p_k(\theta))$

$\hat{\pi}_j(M_1) = p_j(\hat{\theta})$, forventede $e_j = n \cdot p_j(\hat{\theta})$

Eksempel:

$$L(\theta, \theta(1-\theta), (1-\theta)^2) = \theta^{A_1}(\theta(1-\theta))^{A_2}((1-\theta)^2)^{A_3}$$

$$= \theta^{A_1+A_2}(1-\theta)^{A_2+2A_3}$$

$$\hat{\theta} = \frac{A_1+A_2}{A_1+A_2+A_2+2A_3}$$

Hypotese og test

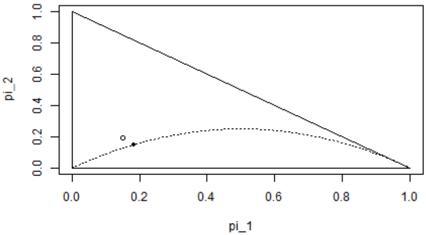
Teststørrelse: $G = 2\sum_{i=1}^{k} A_j \log\left(\frac{A_j}{e_j}\right)$

p-værdi $\approx 1 - \chi^2_{\text{cdf}}(G, \text{k-1-d})$ (hvis alle $e_j \geq 5$)

R: $1 - \text{pchisq}(G, k-1-d)$

$$G = -2\log(Q) = -2\log\left(\frac{\max_{M_1} L(\pi)}{\max_{M_0} L(\pi)}\right) = -2\log\left(\frac{L(\hat{\pi}_1(M_1),\ldots,\hat{\pi}_k(M_1))}{L(\hat{\pi}_1(M_0),\ldots,\hat{\pi}_k(M_0))}\right)$$

Baggrund



(π_1, π_2, π_3), $\pi_1 + \pi_2 + \pi_3 = 1$, $L(\pi_1, \pi_2, \pi_3)$"="$L(\pi_1, \pi_2)$

cirkelpunkt: $(\hat{\pi}_1, \hat{\pi}_2)$, sorte punkt: $(\hat{\theta}, \hat{\theta}(1 - \hat{\theta}))$

Baggrund

$$G = -2\log\left(\frac{L(\hat{\pi}_1(M_1),\ldots,\hat{\pi}_k(M_1))}{L(\hat{\pi}_1(M_0),\ldots,\hat{\pi}_k(M_0))}\right) = -2\log\left(\frac{\max_{M_1} L(\pi)}{\max_{M_0} L(\pi)}\right)$$

Altid: $\max_{M_1} L(\pi) \leq \max_{M_0} L(\pi)$, og dermed $G \geq 0$

$\max_{M_1} L(\pi)$ langt under $\max_{M_0} L(\pi)$: G er stor

store værdier af G er kritiske

Centrale grænseværdisætning + taylorudvikling: $G \approx \chi^2(k-1-d)$

$$Q = \frac{\binom{n}{A_1,\ldots,A_k} p_1(\hat{\theta})^{A_1} \cdots p_k(\hat{\theta})^{A_k}}{\binom{n}{A_1,\ldots,A_k}\left(\frac{A_1}{n}\right)^{A_1}\cdots\left(\frac{A_k}{n}\right)^{A_k}} = \frac{1}{\left(\frac{A_1}{np_1(\hat{\theta})}\right)^{A_1}\cdots\left(\frac{A_k}{np_k(\hat{\theta})}\right)^{A_k}},$$

og dermed $G = -2\log(Q) = 2\sum_{j=1}^{k} A_j \log\left(\frac{A_j}{e_j}\right)$, $\quad e_j = np_j(\hat{\theta})$.

Likelihoodratiotest generelt

$$G = -2\log(Q) = -2\log\left(\frac{\max_{M_1} L(\pi)}{\max_{M_0} L(\pi)}\right)$$

Generelt vil G som en approksimation følge en $\chi^2(f)$ med

$$f = d_0 - d_1$$

d er antallet af frie parametre i modellen

Fulde multinomialmodel:

$\pi_j \geq 0$ $\pi_1 + \cdots + \pi_k = 1$ har $k-1$ frie parametre

Model $(\pi_1, \ldots, \pi_k) = (\theta, \theta(1-\theta), (1-\theta)^2)$ har 1 fri parameter

Sceneskift

Fra: generelt test i multinomialmodel

Til: teste at data følger en bestemt fordeling

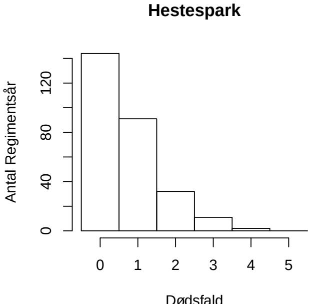
Klassisk goodness of fit test

Bortkiewicz: $n = 280$ observationer af dødsfald i et preussisk regiment ved hestespark inden for et år (20 år, 14 regimenter). Obervationer: $x_1, \ldots, x_n$

Lad X være antal dødsfald i et år i et regiment

> Undersøge om $X \sim \text{pois}(\lambda)$ for et $\lambda \geq 0$

Skabe overblik over data gennem (antals-) histogram



Grafisk sammenligning

Vælge λ således at poissonsandsynligheder passer bedst med histogram

- $\hat{\lambda} = \bar{x} = \frac{196}{280} = 0.7$
- forventede med værdi 0: $e_1 = 280 \cdot \text{dpois}(0, \hat{\lambda})$
- forventede med værdi 1: $e_2 = 280 \cdot \text{dpois}(1, \hat{\lambda})$, og så videre
- forventede med værdi ≥ 4: $e_5 = 280 \cdot (1 - \text{ppois}(3, \hat{\lambda}))$
- R: dpois(x, λ) udregner sandsynligheder i en pois(λ)-fordeling

Multinomialmodel

Når vi laver et antalshistogram bliver vi naturligt ledt over i multinomialmodel

A_1 : antal med værdien 0, A_2 : antal med værdien 1

A_3 : antal med værdien 2, A_4 : antal med værdien 3

A_5 : antal med værdier ≥ 4

Model M_0: $(A_1, \ldots, A_5) \sim \text{multinom}(n, \pi)$

π_j vilkårlig, $\pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 = 1$

fortolkning: $\pi_1 = P(X = 0)$, $\pi_2 = P(X = 1)$, $\pi_3 = P(X = 2)$

$\pi_4 = P(X = 3)$, $\pi_5 = P(X \geq 4)$

Multinomialmodel

Hypotese: $\pi_1, \pi_2, \pi_3, \pi_4, \pi_5$ er givet ved at X er poissonfordelt:

der findes $\lambda \geq 0$ således at

$$\pi_1 = \frac{\lambda^0}{0!} e^{-\lambda}, \quad \pi_2 = \frac{\lambda^1}{1!} e^{-\lambda}, \quad \pi_3 = \frac{\lambda^2}{2!} e^{-\lambda}$$

$$\pi_4 = \frac{\lambda^3}{3!} e^{-\lambda}, \quad \pi_5 = 1 - \pi_1 - \pi_2 - \pi_3 - \pi_4$$

Under hypotesen betegnes modellen med M_1:

teste reduktion fra model M_0 til model M_1

Skøn over λ: benytter $\hat{\lambda} = \bar{x}$ (MSRR Prop 6.1.2)

Observerede og forventede

Index	1	2	3	4	5
Værdi	0	1	2	3	≥ 4
A_j	144	91	32	11	2
e_j	139.04	97.33	34.07	7.95	1.61

$\hat{\lambda} = \frac{196}{280} = 0.7$

$e_1 = 280 \cdot \text{dpois}(0, 0.7) = 139.04, \quad e_5 = 280 \cdot (1 - \text{ppois}(3, 0.7)) = 1.61$

Ønsker $e_j \geq 5$: nødvendigt at slå kasse 4 og 5 sammen

Index	1	2	3	4
Værdi	0	1	2	≥ 3
A_j	144	91	32	13
e_j	139.04	97.33	34.07	9.56

Slå sammen

Index	1	2	3	4
Værdi	0	1	2	≥ 3
A_j	144	91	32	13
e_j	139.04	97.33	34.07	9.56

$k = 4$ kasser, $\quad d = 1$ parameter

teststørrelse: $G = 2\sum_j a_j \log\left(\frac{a_j}{e_j}\right) = 1.84$

p-værdi $= 1 - \text{pchisq}(1.84, 4 - 1 - 1) = 0.399$

Konklusion: data strider ikke mod poissonfordelingen

Prøv i R

```
a=c(144, 91, 32, 13)

ex=c(139.04, 97.33, 34.07,  9.56)

sum((a-ex)^2)/ex
```

Goodness of fit test generelt

Målinger $x_1, x_2, \ldots, x_n$ fra stokastisk variabel X

Del talakse op i intervaller:

endepunkter: $z_0, z_1, \ldots, z_k \quad (z_0 : \; -\infty, \quad z_k : \; +\infty)$

A_j: antal i interval $(z_{j-1}, z_j]$

Model M_0: $(A_1, \ldots, A_k) \sim \text{multinom}(n, \pi)$

π_j vilkårlig, $\pi_1 + \pi_2 + \cdots + \pi_k = 1$; fortolkning: $\pi_j = P(z_{j-1} < X \leq z_j)$

Model M_1: $\pi_j = p_j(\theta) = F(z_j, \theta) - F(zj - 1, \theta)$, $j = 1, \ldots, k$

θ ukendt parameter

$P(X \leq z)$ er givet ved $F(z, \theta)$

Poissoneksemplet ovenfor: $F(z, \lambda) = \sum_{x=0}^{z} \frac{\lambda^x}{x!} e^{-\lambda}$

Goodness of fit test generelt

1) Find skøn $\hat{\theta}$ over θ

2) Udregn $\hat{\pi}_j(M_1) = p_j(\hat{\theta}) = F(z_j, \hat{\theta}) - F(zj - 1, \hat{\theta})$

3) Udregn forventede: $e_j = n \cdot p_j(\hat{\theta})$

4) slå eventuelt kasser sammen

5) Beregn G-teststørrelse og p-værdi

Sceneskift

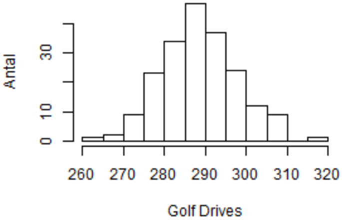
Goodness of fit test er indført:

$$G = 2 \sum \text{observerede} \log \left(\frac{\text{observerede}}{\text{forventede}} \right)$$

Næste: Goodness of fit test for kontinuert stokastisk variable

Kontinuert variabel: alle værdier er mulige

The average drive distance (in yards) for 199 professional golfers during a week on the men's PGA tour in 2015 (DASL)



Antal i interval $(a, b]$: $\approx n \cdot P(a < X \leq b) = n \int_a^b f(x)dx$

hvordan viser jeg tæthed $f(x)$ i histogrammet ?

Tæthedshistogram

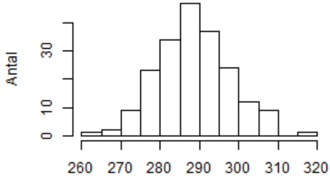
Inddel data-akse i intervaller. Tegn kasser med højde:

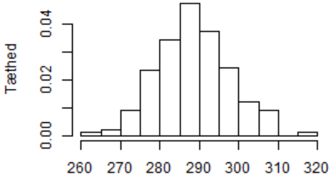
antalshistogram: Antal i interval (R: hist(x))

tæthedshistogram: $\frac{\text{Antal i interval}}{n \cdot \text{intervallængde}}$ = frekvens per længde

$\approx$ sandsynlighed per længde = tæthed

(R: hist(x,probability=TRUE))





Prøv i R

```
par(mfrow=c(1,2))

hist(log(rivers))

hist(log(rivers),probability=TRUE)
```

Normalfordelingen

Generel normalfordeling med middelværdi μ og spredning σ:

tæthed $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$

notation: $X \sim N(\mu, \sigma^2)$

standard normalfordeling: $N(0, 1)$

Er drivelængden i golf normalfordelt ?

R: tæthed: dnorm(x,μ, σ), fordelingsfunktion ($P(X \leq x)$: pnorm(x,μ, σ)

Goodness of fit test: data

Indlæsning: goldfdrives=scan("golf2015.txt")

Intervalinddeling: endePkt=260+c(0:12)*5

antal i intervaller: a=hist(golfdrives,breaks=endePkt)$counts

Model M_0: $(A_1, \ldots, A_{12}) \sim (multinom(199, (\pi_1, \ldots, \pi_{12}))$
$(pi_1, \ldots, \pi_{12})$ kan variere frit

Hypotese (model M_1): der findes μ og σ således at:
$\pi_1 = p_1(\mu, \sigma) = pnorm(265, \mu, \sigma)$,
$\pi_2 = p_2(\mu, \sigma) = pnorm(270, \mu, \sigma) - pnorm(265, \mu, \sigma)$, $\ldots$
$\pi_{11} = p_{11}(\mu, \sigma) = pnorm(315, \mu, \sigma) - pnorm(310, \mu, \sigma)$,
$\pi_{12} = p_{12}(\mu, \sigma) = 1 - pnorm(315, \mu, \sigma)$

Goodness of fit test: Finde parameterskøn

I vil typisk ikke blive bedt om at gøre dette i opgaverne

```
endePkt=260+c(0:12)*5
a=hist(golfdrives,breaks=endePkt,plot=FALSE)$counts

loglik=function(th){
prob=pnorm(endePkt[2:12],th[1],th[2])
probInterval=c(prob,1)-c(0,prob)
return(sum(-a*log(probInterval)))
}

nlm(loglik,c(288,9))
```

Goodness of fit test: forventede

Skøn over μ og σ: $\hat{\mu} = 288.5796$ og $\hat{\sigma} = 9.1315$

Beregning af forventede:

```
endePkt=260+c(0:12)*5

prob=pnorm(endePkt[2:12],288.5796,9.1315)

ex=199*(c(prob,1)-c(0,prob))
```

Kasse	1	2	3	4	5	6	7	8	9	10	11	12
a_j	1	2	9	23	34	47	37	24	12	9	0	1
e_j	1.0	3.2	9.5	20.9	34.6	42.6	39.2	27.0	13.8	5.3	1.5	0.4

Goodness of fit test: forventede

Vi slår kasse 1 og 2 samme og kasserne 10, 11 og 12 (Cochrans regel)

Kasse	1+2	3	4	5	6	7	8	9	10-12
$a1_j$	3	9	23	34	47	37	24	12	10
$e1_j$	4.2	9.5	20.9	34.6	42.6	39.2	27.0	13.8	7.2

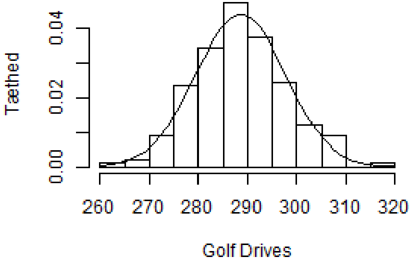
Beregning af G og p-værdi:

```
G=2*sum(a1*log(a1/e1))

c(G,1-pchisq(G,9-1-2))

[1] 2.7293672 0.8419664
```

Goodness of fit test: figur



Sceneskift

Goodness of fit test for normalfordelingen er vist

Næste: Sammenligne poissonfordelinger

Teste samme rate i flere poissonfordelinger

8 dør af hestespark i løbet af 20 år i regiment 1

17 dør af hestespark i løbet af 20 år i regiment 2

hypotese: samme rate af dødsfald i de to regimenter

Model: $X_1 \sim \text{poisson}(20\lambda_1), \quad X_2 \sim \text{poisson}(20\lambda_2)$

hypotese: $\lambda_1 = \lambda_2$

Intuitivt: hvis samme rate skal de 25 dødsfald fordeles tilfældigt på de to regimenter

Teste proportionale poissonparametre

Generelt: $X_i \sim \text{poisson}(\lambda_i), \quad i = 1, \ldots, k$

hypotese: $\lambda_i = m_i \lambda, \quad m_i$ kendte tal

Lad $n = X_1 + \cdots + X_k$. For "fastholdt" n vil vi have modellen

$$(X_1, X_2, \ldots, X_k) \sim \text{multinom}(n, (\pi_1, \ldots, \pi_k))$$
$$\pi_j = \lambda_j / \lambda_\bullet,\ j = 1, \ldots, k, \quad \lambda_\bullet = \sum_j \lambda_j$$

I denne model testes hypotesen

$$\pi_j = \frac{m_i}{m_\bullet},\ j = 1, \ldots, k, \quad m_\bullet = m_1 + \cdots + m_k$$

Forventede: $e_j = n \frac{m_j}{m_\bullet},\ j = 1, \ldots, k,$

Beregn G og p-værdi $= 1 - \chi^2_{\text{cdf}}(G, k - 1 - 0)$

Bevis for betingning

$$X_i \sim \text{poisson}(\lambda_i), \quad i = 1, \ldots, k, \quad X_\bullet = X_1 + \cdots X_k \sim \text{poisson}(\lambda_\bullet)$$

$$\lambda_\bullet = \lambda_1 + \cdots \lambda_k$$

$$P(X_1 = x_1 \ldots, X_k = x_k | X_\bullet = n) = \frac{P(X_1 = x_1 \ldots, X_k = x_k)}{P(X_\bullet = n)} = \frac{\prod_j \frac{\lambda_j^{x_j}}{x_j!} e^{-\lambda_j}}{\frac{\lambda_\bullet^n}{n!} e^{-\lambda_\bullet}}$$

$$\binom{n}{x_1 \ldots, x_k} \left(\frac{\lambda_1}{\lambda_\bullet}\right)^{x_1} \cdots \left(\frac{\lambda_k}{\lambda_\bullet}\right)^{x_k}$$

$$\sim \text{multinom}(n, (\pi_1, \ldots \pi_k)), \quad \pi_j = \frac{\lambda_j}{\lambda_\bullet}$$

Sceneskift

Sammenligne poissonfordelinger er vist

Afslutte med: chi-i-anden test

χ^2-test

I webbogen bruges teststørrelsen: $G = 2 \sum \text{observeret} \cdot \log\left(\frac{\text{observeret}}{\text{forventet}}\right)$

Ude i verden (og i MSRR) bruges ofte af historiske grunde:

$$C = \sum \frac{(\text{observeret} - \text{forventet})^2}{\text{forventet}}$$

For store datasæt er der ikke forskel. For små datasæt kan χ^2-approksimationen være lidt bedre for G end for C

Simulering

```
nsim=1000000
p0=0.3
n=100

x1=rbinom(nsim,n,p0)
x2=n-x1
phat=rep(p0,nsim)
e1=n*phat
e2=n*(1-phat)
x11=ifelse(x1==0,1,x1)
x22=ifelse(x2==0,1,x2)

G=2*(x1*log(x11/e1)+x2*log(x22/e2))
Ctest=(x1-e1)^2/e1+(x2-e2)^2/e2

100*c(sum(G>3.8415),sum(Ctest>3.8415))/nsim
```

Sceneskift

Slut for i dag