Dagens program:

Mandag den 2. maj
Hierarkiske generaliserede lineære modeller

- Eksempel: hierarkisk Poissonfordelingsmodel
  Eksempel: hierarkisk binomialfordelingsmodel
- Hierarkisk model for exponentiel dispersionsfamilie
- Flere eksempler på hierarkiske modeller
- Aposteriorifordeling for middelværdiparameter μ
- Eksempel Aposteriorifordeling for Poisson-Gamma model
  prædiktiv fordeling

Fordeling af dagligt antal tordenvejr ved Cape Kennedy

Antal episoder	Antal dage	Poisson forventet
y_i	$\#i$	
0	803	791.85
1	100	118.78
2	14	8.91
3+	3	0.46

Tælletal, naturlig fordeling er Poissonfordeling med frekvensfunktion

$$\mathsf{P}[\mathsf{Y}=\mathsf{y}] = \mathsf{p}_{\mathsf{Y}}(\mathsf{y};\mu) = \frac{\mu^{\mathsf{y}}}{\mathsf{y}!}\exp(-\mu)$$

for $y = 0, 1, 2, \ldots$ og middelantallet μ episoder/dag.

MEN observeret fordeling har **tykkere haler** end Poissonfordelingen.
Atså **overdispersion**

Hierarkisk model for Poisson fordeling med overdispersion

$$Y|\mu \sim \mathsf{P}(\mu) \text{ Poisson-fordeling}$$

Hvad skal vi vælge som fordeling af μ ?
Det er usmart med en normalfordeling af μ, da $0 < \mu$.

Likelihood'en for μ fra Poissonfordelingen er proportional med

$$p_Y(y;\mu) \propto \mu^y \exp(-\mu)$$

Vi vælger en fordeling af μ, hvis tæthed har samme form som likelihood'en fra Poissonfordelingen, nemlig

$$g_\mu(\mu) \propto (\mu)^{\alpha-1}\exp(-\mu/\beta)$$

Det er jo tætheden for en **Gammafordeling**, dvs

$$g_\mu(\mu;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta}\left(\frac{\mu}{\beta}\right)^{\alpha-1}\exp(-\mu/\beta)$$

Marginal fordeling

Frekvensfunktionen for den marginale fordeling af μ fås ved at integrere $p_Y(y;\mu)$ med hensyn til $g_\mu(\mu;\alpha,\beta)$

$$\begin{aligned} f_Y(y) &= \int_{\mu=0}^{\infty} p_Y(y;\mu)g_\mu(\mu;\alpha,\beta)d\mu \\ &= \int_{\mu=0}^{\infty} \frac{\mu^y}{y!}\exp(-\mu)\frac{1}{\Gamma(\alpha)\beta}\left(\frac{\mu}{\beta}\right)^{\alpha-1}\exp(-\mu/\beta)d\mu \\ &= \int_{\mu=0}^{\infty} \frac{\mu^{y+\alpha-1}}{y!\Gamma(\alpha)\beta^\alpha}\exp\Big\{-\mu\Big/\Big(\beta/(1+\beta)\Big)\Big\}d\mu \end{aligned}$$

Vi bemærker nu, at de faktorer i integranden, der afhænger af μ er, $\mu^{y+\alpha-1}$ og $\exp\Big\{-\mu\Big/\Big(\beta/(1+\beta)\Big)\Big\}$, som netop er kernen i en Gammafordeling med formparameteren $y+\alpha$ og skalaparameter $\beta/(1+\beta)$

Marginal fordeling

Udtrykket kan reorganiseres til

$$\begin{aligned} f_Y(y) &= \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \frac{\beta^y}{(1+\beta)^{y+\alpha}} \\ &\quad \int_{\mu=0}^{\infty} \frac{1}{\Gamma(y+\alpha)\beta/(1+\beta)} \left(\frac{\mu}{\beta/(1+\beta)}\right)^{y+\alpha-1} \exp\Big\{-\mu\Big/\Big(\beta/(1+\beta)\Big)\Big\} d\mu \\ &= \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \frac{\beta^y}{(1+\beta)^{y+\alpha}} \ , \end{aligned}$$

hvor integranden netop er tætheden for en gammafordeling, (med integralet 1).

$$f_Y(y) = \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \left(\frac{1}{1+\beta}\right)^{\alpha} \left(1 - \frac{1}{1+\beta}\right)^{y}$$

For heltallige positive værdier af α er $f_Y(y)$ netop frekvensfunktionen for en negativ binomialfordeling (fordelingen af "ventetiden" y til den α'te succes i en følge af Bernoulli forsøg, hvor sandsynligheden for succes er $1/(1+\beta)$)

Marginal fordeling

For heltallige positive α bliver frekvensfunktionen

$$f_Y(y) = \frac{(y+\alpha-1)!}{y!(\alpha-1)!} \left(\frac{1}{1+\beta}\right)^{\alpha} \left(1 - \frac{1}{1+\beta}\right)^{y} = \binom{y+\alpha-1}{y} \left(\frac{1}{1+\beta}\right)^{\alpha} \left(1 - \frac{1}{1+\beta}\right)^{y}$$

Vi kan definere en negativ binomialfordeling også for ikke-heltallige værdier af α.

Momenter

Vi har fra gammafordelingen at $\mathsf{E}[\mu] = \alpha\beta$ og $\mathsf{V}[\mu] = \alpha\beta^2$, og fra Poissonfordelingen har vi $\mathsf{E}[\mathsf{Y}|\mu] = \mathsf{V}[\mathsf{Y}|\mu] = \mu$. Momenterne i den marginale fordeling af Y bliver derfor

$$\begin{aligned} \mathsf{E}[\mathsf{Y}] &= \mathsf{E}_\mu[\mathsf{E}[\mathsf{Y}|\mu]] = \mathsf{E}[\mu] = \alpha\beta \\ \mathsf{V}[\mathsf{Y}] &= \mathsf{E}_\mu[\mathsf{V}[\mathsf{Y}|\mu]] + \mathsf{V}_\mu[\mathsf{E}[\mathsf{Y}|\mu]] = \mathsf{E}_\mu[\mu] + \mathsf{V}_\mu[\mu] \\ &= \alpha\beta + \alpha\beta^2 = m + m^2/\alpha \end{aligned}$$

hvor vi har sat $\mathsf{E}[\mu] = \mathsf{m}$.

Parametriseringer

Undertiden er det smartere at parametrisere gammafordelingen ved middelværdi $m = \alpha\beta$, og reciprok formparameter $k = 1/\alpha$, eller ved middelværdi $m = \alpha\beta$ og "signal/støj" forhold

$$\gamma = \frac{\mathsf{V}_\mu[\mathsf{E}[\mathsf{Y}|\mu]]}{\mathsf{E}_\mu[\mathsf{V}[\mathsf{Y}|\mu]]} = \frac{\mathsf{V}_\mu[\mu]}{\mathsf{E}[\ V_{\mathsf{P}}(\mu)]} = \frac{\alpha\beta^2}{\alpha\beta} = \beta = mk$$

hvor $V_P(\cdot)$ angiver variansfunktionen for Poissonfordelingen, $V_P(\mu) = \mu$. Giver mulighed for grænseovergang til en etpunktsfordeling i m for fast m og $k \to 0$ eller, alternativt fast m og $\gamma \to 0$

Ved denne grænseovergang går NB-fordelingen mod en Poissonfordeling med middelværdi m.

Flere observationer fra samme gruppe (samme μ)

Optræder fx. ved bakterietællinger, målinger i "rene rum"
Betragt summen $Z = X_1 + X_2 + \cdots + X_n$, hvor X_i'erne for fastholdt μ (dag, lokalitet el. lgn.) er uafhængige $P(\mu)$-fordelte variable.
Modellen for Z bliver da

$$Z|\mu \sim P(n\mu), \quad \mu \sim G(1/k, m/k)$$

Den marginale fordeling af Z er en NB-fordeling (på ikke-negative heltal).

Det er naturligere at betragte gennemsnittet $Y = Z/n$

$$\mathsf{V[Y]} = \frac{1}{\mathsf{n}^2}\mathsf{V[Z]} = \mathsf{m}^2 + \frac{\mathsf{m}}{\mathsf{n}} = \mathsf{m}\left(\mathsf{mk} + \frac{1}{\mathsf{n}}\right) = \mathsf{E}[V_{\mathsf{P}}(\mu)]\left(\gamma + \frac{1}{\mathsf{n}}\right)$$

lissom formlen i normalfordelingsituationen

Estimation

Parametrene m og k (eller m og γ) kan estimeres ved momentmetoden fra gennemsnit og empirisk varians for observationerne (hvis designet er balanceret, dvs. samme n i alle grupper).

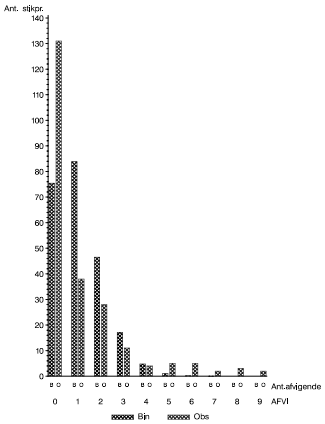
Man kan også maksimum-likelihood estimere.

PROC GENMOD, dist= NEGBIN estimerer m og k og kalder k for "dispersion" (men det er den ikke).

Binomialfordeling, eksempel

Eksempel, Z_i antal dimensionsafvigende låg i stikprøver á 770 låg fra hvert af 229 partier.

Hvis proces i kontrol, naturligt at forvente Z_i er uafhængige ensfordelte med $Z_i \sim B(770, p)$ (binomialfordeling).



Fordeling af antal afvigende låg

Tykkere haler end binomialfordeling med samme middelværdi.

Hierarkisk model:
$Z|p \sim B(n,p)$,

$$p_Z(z) = \binom{n}{z} p^z (1-p)^{n-z}$$

$p \sim$????

Upraktisk med normalfordeling af p, da $0 < p < 1$.

Evt kunne man vælge normalfordeling af den kanoniske parameter, logit'en.

Fordeling af p

Vi vælger $p \sim Be(\alpha,\beta)$ (betafordeling) med tæthed

$$g_p(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \text{for } 0 < p < 1$$

($B(\alpha,\beta)$ er **Betafunktionen**), altså samme form af **kernen** af tætheden, som likelihoodfunktionen for binomialfordelingen.

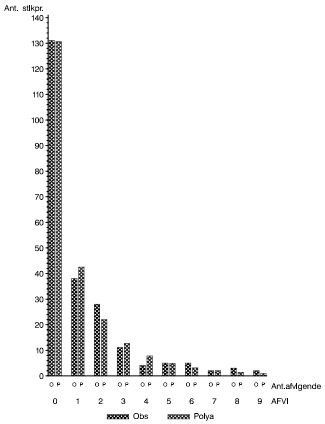
$$\mathsf{E[p]} = \frac{\alpha}{\alpha+\beta} \quad \mathsf{V[p]} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Marginal fordeling

Lad $Z|\mu \sim B(n,\mu)$ og $\mu \sim Be(\alpha,\beta)$. Da kaldes den marginale fordeling af Z for **Polyafordelingen** (efter Polyas urnemodel).

Undertiden betegnes Polyafordelingen "binomial-beta" fordelingen fordi den er en **compound** fordeling fremkommet ved at mixe binomialfordelinger ved en betafordeling

Marginal fordeling



Momenter i marginal fordeling

Betragt den relative hyppighed, $Y = Z/n$

Middelværdi, $\pi = \mathsf{E}[\mu] = \alpha/(\alpha+\beta)$ og "signal/støjforhold"

$$\gamma = \frac{\mathsf{V}_\mu[\mathsf{E}[Y|\mu]]}{\mathsf{E}_\mu[\mathsf{V}[Y|\mu]]} = \frac{\mathsf{V}_\mu[\mu]}{\mathsf{E}[\mu(1-\mu)]} = \frac{1}{\alpha+\beta}$$

Vi får derfor

$$\begin{aligned}\mathsf{E}[Y] &= \pi = \frac{\alpha}{\alpha+\beta}\\ \mathsf{V}[Y] &= \frac{\pi(1-\pi)}{1+\gamma}\left(\gamma+\frac{1}{n}\right)\end{aligned}$$

hvor

$$\frac{\pi(1-\pi)}{1+\gamma} = \mathsf{E}[\mu(1-\mu)] = \mathsf{E}[V_{\mathsf{bin}}(\mu)]$$

udtrykker den "gennemsnitlige" binomialvarians

Fortolkning af parametre i konjugeret fordeling

For parametrene m og γ gælder:

$$\begin{aligned}m &= \mathsf{E}[\mu]\\ \gamma &= \frac{\mathsf{V}[\mu]}{\mathsf{E}[V(\mu)]}\,,\end{aligned}$$

hvor $\mu = \mathsf{E}[Y|\theta]$, og hvor $V(\mu)$ angiver variansfunktionen $\kappa''(\tau^{-1}(\mu))$

Marginal fordeling af gruppegennemsnit

$$\begin{aligned}\mathsf{E}[Y_i] &= \mathsf{E}[\mathsf{E}[Y_i|\mu]] = \mathsf{E}[\mu] = m\\ \mathsf{V}[Y_i] &= \mathsf{E}[V(\mu)]\left(\gamma+\frac{1}{n_i}\right) = \mathsf{E}[V(\mu)]\left(\gamma+\frac{1}{n_i}\right)\,,\end{aligned}$$

Konjugeret fordeling giver simpelt analytisk udtryk for den marginale fordeling. Man kunne godt vælge at lave normalfordeling for kanonisk parameter, men det giver komplicerede udtryk for marginalfordelinger.

Eksempel, Monocyttælling

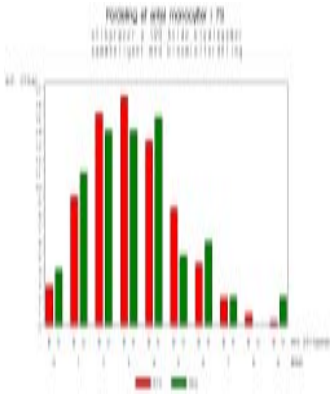
Som led i en undersøgelse af et lægemiddel udtoges en gruppe personer (73 rygere) i en given aldersklasse. For hver person tog man 100 hvide blodlegemer (leukocytter) fra en blodprøve fra personen og optalte antallet af monocytter (encellede). Fordelingen var:

Antal monoc.	0	1	2	3	4	5	6	7	8	9
Antal pers.	4	11	14	14	15	5	6	2	0	2

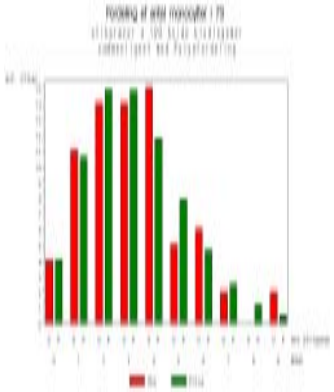
Snit 3.21,

Det kunne være naturligt at sammenligne med en binomialfordeling, $B(n=100, p=0.032)$.

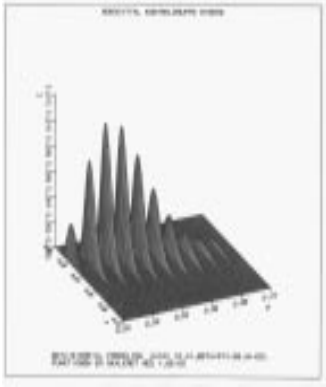
Monocyttælling, sammenligning med binomial fordeling



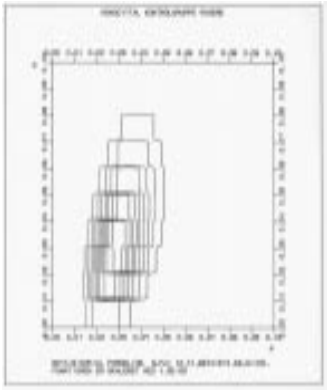
Monocyttælling, sammenligning med Polyafordeling



Monocyttælling, simultan fordeling



Monocyttælling, niveaukurver



Generelt princip for valg af konjugeret fordeling

Betragt en exponentiel dispersionsfamilie (af observationer)

$$f_Y(y;\theta,\lambda) = c(y,\lambda)\exp[\lambda\{\theta y - \kappa(\theta)\}] \quad \text{for } \theta \in \Omega$$

Vælg fordeling af kanonisk parameter θ med tæthed

$$g_\theta(\theta; m, \gamma) = \frac{1}{C(m,\lambda)}\exp[\{\theta m - \kappa(\theta)\}/\gamma]$$

hvor kumulantfrembringeren $\kappa(\cdot)$ er den samme i de to udtryk.
Fordelingen af θ kaldes den **konjugerede fordeling**
I Bayes-sammenhænge kaldes den konjugeret apriorifordeling.
I hierarkiske modeller kaldes den undertiden konjugeret strukturfordeling.

Andre eksempler på hierarkiske modeller

Fordeling af empiriske varianser, afsnit 7.3

Normalfordeling med tilfældigt varierende varians, afsnit 7.4:

$$\begin{aligned} Y_i|\sigma_i^2 &\sim \mathsf{N}(\mu,\sigma_i^2) \\ 1/\sigma_i^2 &\sim \mathsf{G}(\alpha, 1/\beta) \end{aligned}$$

hvor σ_i^2 er indbyrdes uafængige.

Da gælder for den marginale fordeling af Y_i, at

$$\frac{Y_i - \mu}{\sqrt{\beta/\alpha}} \sim \mathsf{t}(2\alpha)$$

hvor $\mathsf{t}(2\alpha)$ er en t-fordeling med 2α frihedsgrader, dvs en fordeling med **tykkere haler** end normalfordelingen.

Aposteriorifordelinger under hierarkisk model

Lad $X_1, X_2, \ldots, X_n$ være uafhængige (for givet θ), frembragt med samme θ og med tæthed

$$g(x|\theta) = d(x)\exp\{[\theta x - \kappa(\theta)]/\sigma^2\} \ ,$$

og lad fordelingen af θ være den konjugerede fordeling med tæthed

$$w(\theta; m, \gamma) = \frac{1}{C(m,\gamma)}\exp\{[\theta m - \kappa(\theta)]/\gamma\}$$

Da er aposteriorifordelingen af θ efter observation af $X_1 = x_1, \ldots, X_n = x_n$ e

$$w\Big(\theta \mid \sum X_i/n = \bar{x}\Big) = \frac{1}{C(m_1,\gamma_1)}\exp\{[\theta m_1 - \kappa(\theta)]/\gamma_1\}$$

med

$$\begin{aligned} m_1 &= m_{\text{apost}} = \frac{m/\gamma + n\bar{x}/\sigma^2}{1/\gamma + n/\sigma^2} \\ \gamma_1 &= \gamma_{\text{apost}} = \frac{1}{1/\gamma + n/\sigma^2} \end{aligned}$$

Aposteriorifordelinger

For den konjugerede apriorifordeling gælder altså at aposteriorifordelingen for θ er af samme form som apriorifordelingen, blot med opdaterede parametre m og γ ved brug af den sædvanlige opdateringsformel.

$$\frac{1}{\gamma_{\text{apost}}} = \frac{1}{\gamma_{\text{apriori}}} + n/\sigma^2$$

$$\frac{m_{\text{apost}}}{\gamma_{\text{apost}}} = \frac{m_{\text{apriori}}}{\gamma_{\text{apriori}}} + n\,\bar{x}/\sigma^2$$

$$\gamma_{\text{apost}} = \frac{\mathsf{V}[\mu|\bar{\mathsf{x}}]}{\mathsf{E}[\mathsf{V}(\mu)|\bar{\mathsf{x}}]} \ , \tag{0.1}$$

Eksempel på opdatering af aprioriinformation

Poisson-gamma model:
Vævefejl i stof.

For en given produktion er fejlene tilfældigt spredt over stoffet i overensstemmelse med en Poisson-fordeling.

Gennemsnitligt antal fejl pr m^2, μ varierer fra produktion til produktion i overensstemmelse med en gammafordeling med middelværdi 2.5 og varians 2.0825

Prædiktiv fordeling
Der udtages en stikprøve, og antallet af fejl bestemmes. Dette bruges til bestemmelse af aposterioridfordeling for μ, som derefter bruges til bestemmelse af **prædiktiv fordeling** af antal fejl i hhv een, 10 og 20 m^2 stof fra denne produktion.
Prædiktiv fordeling, mixtur af stikprøvefordeling og apost.

Poisson-gamma

Parameter	Før	Efter observation af		
		$n = 10$	$n = 10$	$n = 20$
		$z = 1$	$z = 60$	$z = 120$
		$\bar{x} = 0.1$	$\bar{x} = 6$	$\bar{x} = 6$
$m = \mathsf{E}[\mu]$	2.5	0.36	5.63	5.80
γ	0.833	0.0893	0.0893	0.0472
$\mathsf{V}[\mu] = \mathsf{m} \times \gamma$	2.0825	0.03	0.50	0.27
α	3	4	63	123
β	1.2	11.2	11.2	21.2
Fordeling af fejl i én ny m $X' \in \mathsf{NB}(\alpha, \beta/(1+\beta))$				
$\mathsf{E}[X'] = \mathsf{m}$	2.5	0.36	5.63	5.80
$\mathsf{V}[X'] = \mathsf{m}(1+\gamma)$	4.58	0.39	6.13	6.07
Fordeling af fejl i 10 nye m $Z' \in \mathsf{NB}(\alpha, \beta/(10+\beta))$				
$\mathsf{E}[Z'] = 10\mathsf{m}$	25	3.6	56.3	58.0
$\mathsf{V}[Z'] = 10\mathsf{m}(1+10\gamma)$	233.32	6.81	106.57	85.38
Fordeling af fejl i 20 nye m $Z' \in \mathsf{NB}(\alpha, \beta/(20+\beta))$				
$\mathsf{E}[Z'] = 20\mathsf{m}$	50	7.2	112.6	116.0
$\mathsf{V}[Z'] = 20\mathsf{m}(1+20\gamma)$	883	20.06	313.7	225.50

Vi så, at

- Variansen i aposteriorifordelingen af μ afhang ikke kun af stikprøvestørrelsen, men også af fejlniveauet (fordi variansen i Poissonfordelingen stiger med middelværdien)
- Det væsentligste bidrag til **spredningen** i den prædiktive fordeling for det totale antal fejl hidrører fra aposteriorivariansen for μ. Bidraget er proportionalt med "prædiktionslængden" og er væsentligt større end bidraget fra Poissonusikkerheden.
  For det totale antal fejl er jo ikke summen af **uafhængige** Poissonstørrelser; de kommer jo fra samme produktion