Advanced dataanalysis and statistical modelling, Week 11

Hierarchical models

Jan Kloppenborg Møller

April 29, 2018

Outline

Oversigt

Introduction

- A characteristic property of the generalized linear models is that **the variance,** $\mathrm{Var}[Y]$ **is a known function**, $V(\mu)$, that only depends on μ

$$\mathrm{Var}[Y_i] = \lambda_i V(\mu) = \frac{\sigma^2}{w_i}\, V(\mu)$$

  where w_i denotes a known **weight**, associated with the i'th observation, and where σ^2 denotes a common **dispersion parameter**

- The dispersion parameter σ^2 serve to **express overdispersion** in situations where the residual deviance is too large.

- An alternative method for modeling overdispersion, is by **hierarchical models**, analogous to the mixed effects models for the normally distributed observations.

Introduction

- A starting point in a hierarchical modeling is an assumption that the distribution of the random "noise" may be modeled by an **exponential dispersion family** (Binomial, Poisson, etc.), and then it is a matter of choosing a suitable (prior) distribution of the mean-value parameter μ.
- It seems natural to choose a distribution with a **support that coincides with the mean value space $\mathcal{M}$**
- In some applications an approach with a normal distribution of the canonical parameter is used. Such an approach is sometimes called **generalized linear mixed models** (GLMMs)

Introduction

- Although consistent with a formal requirement of equivalence between mean values space the resulting **marginal distribution of the observation is seldom tractable**, and the likelihood of such a model will involve an integral which cannot in general be computed explicitly.
- Instead, we shall describe an approach based on the so-called **standard conjugated distribution** for the mean parameter of the within group distribution for exponential families.
- These distributions combine with the exponential families in a simple way, and lead to **marginal distributions that may be expressed in a closed form** suited for likelihood calculations.

Oversigt

Hierarchical Poisson Gamma model - example

The table shows the distribution of **the number of daily episodes** of thunderstorms at Cape Kennedy, Florida, during the months of June, July and August for the 10-year period 1957–1966, total 920 days.

Number of episodes, z_i	Number of days, $\# \ i$	Poisson expected
0	803	792.71
1	100	118.05
2	14	8.79
3+	3	0.45

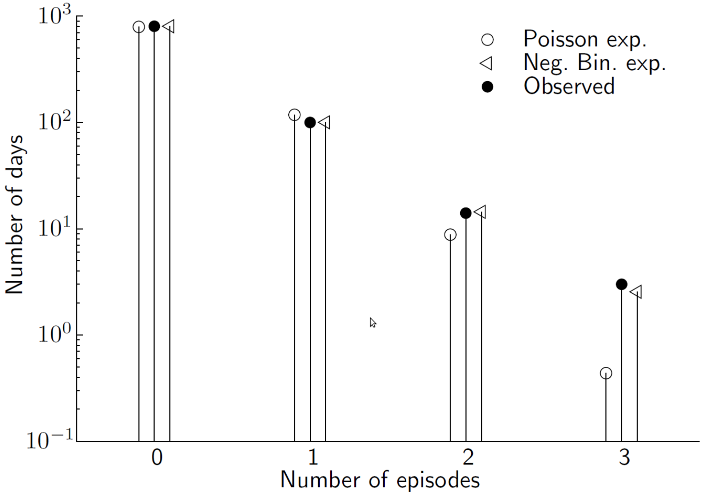
Table: The distribution of days with 0, 1, 2 or more episodes of thunderstorm at Cape Kennedy.

All observational periods are $n_i = 1$ day.

Hierarchical Poisson Gamma model - example

- The data represents **counts** of events (episodes of thunderstorms) distributed in time.
- A completely random distribution of the events would result in a **Poisson distribution** of the number of daily events.
- The variance function for the Poisson distribution is $V(\mu) = \mu$; therefore, a Poisson distribution of the daily number of events would result in the **variance in the distribution of the daily number of events being equal to the mean**, $\widehat{\mu} = \overline{y}_+ = 0.15$ thunderstorms per day.
- The empirical variance is $s^2 = 0.1769$, which is somewhat larger than the average. We further note that the observed distribution has **heavier tails** than the Poisson distribution. Thus, one might be suspicious of overdispersion.

Hierarchical Poisson Gamma model - example



Formulation of hierarchical model

Theorem (Compound Poisson Gamma model)

Consider a hierarchical model for Y specified by

$$\begin{aligned} Y|\mu &\sim Pois(\mu), \\ \mu &\sim G(\alpha, \beta), \end{aligned}$$

i.e. a two stage model.

In the first stage a random mean value μ is selected according to a Gamma distribution, and that Y is generated according to a Poisson distribution with that value as mean value. Then the marginal distribution of Y is a negative binomial distribution, $Y \sim \mathrm{NB}(\alpha, 1/(1+\beta))$

Formulation of hierarchical model

Theorem (Compound Poisson Gamma model, continued)

The probability function for Y is

$$\begin{aligned} P[Y=y] &= g_Y(y;\alpha,\beta) \\ &= \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \; \frac{\beta^y}{(\beta+1)^{y+\alpha}} \\ &= \binom{y+\alpha-1}{y} \frac{1}{(\beta+1)^\alpha} \left(\frac{\beta}{\beta+1}\right)^y \quad \text{for } y=0,1,2,\ldots \end{aligned}$$

where we have used the convention

$$\binom{z}{y} = \frac{\Gamma(z+1)}{\Gamma(z+1-y)\; y!}$$

for z real and y integer values.

Proof.

We have the two densities:

$$f_{Y|\mu}(y) = \frac{\mu^y}{y!}e^{-\mu} \quad \text{and} \quad f_\mu(\mu,\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)}\mu^{\alpha-1}e^{-\frac{\mu}{\beta}}$$

$$g_Y(y) = \int_0^\infty \frac{\mu^y}{y!}e^{-\mu}\frac{1}{\beta^\alpha\Gamma(\alpha)}\mu^{\alpha-1}e^{-\frac{\mu}{\beta}}d\mu \quad \text{[collect, and constants outside]}$$

$$= \frac{1}{y!\beta^\alpha\Gamma(\alpha)}\int_0^\infty \mu^{\overbrace{y+\alpha}^{\widetilde{\alpha}}-1}e^{-\mu\overbrace{\left(\frac{(\beta+1)}{\beta}\right)}^{1/\widetilde{\beta}}}d\mu \quad \text{[recognize as } \Gamma \text{ integral]}$$

$$= \frac{1}{y!\beta^\alpha\Gamma(\alpha)}\frac{\Gamma(y+\alpha)\left(\frac{\beta}{(\beta+1)}\right)^{y+\alpha}}{1} \quad \text{[reduce]}$$

$$= \frac{\Gamma(y+\alpha)\beta^y}{y!\Gamma(\alpha)(\beta+1)^{y+\alpha}} \quad \text{[done!]}$$

□

Formulation of hierarchical model

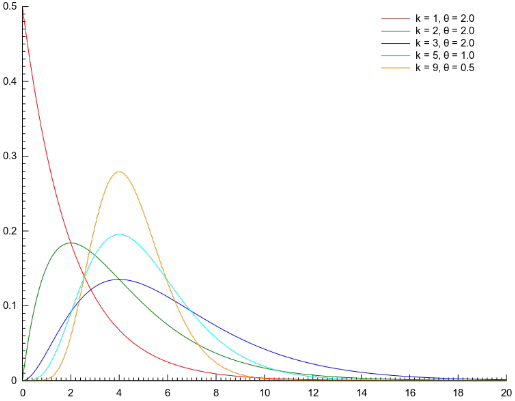
- For integer values of α the negative binomial distribution is known as the distribution of the number of "failures" until the α'th success in a sequence of independent Bernoulli trials where **the probability of success in each trial is** $p = 1/(1+\beta)$

$$\begin{aligned} P[Y=y] &= \binom{y+\alpha-1}{y} \frac{1}{(\beta+1)^\alpha} \left(\frac{\beta}{\beta+1}\right)^y \\ &= \binom{y+\alpha-1}{y} p^\alpha (1-p)^y \quad \text{for } y = 0, 1, 2, \ldots \end{aligned}$$

- For $\alpha = 1$ the distribution is known as the *geometric distribution*.

Why use a Gamma to describe variation between days?

- It has the desired support
- It is a very flexible distribution



- Last but not least the integral can be directly calculated.

Inference on mean μ

Theorem (Conditional distribution of μ)

Consider the hierarchical Poisson-Gamma model and assume that a value $Y = y$ has been observed.

Then the conditional distribution of μ for given $Y = y$ is a Gamma distribution,

$$\mu |\ Y = y\ \sim G(\alpha + y, \beta/(\beta + 1))$$

with mean

$$\mathrm{E}[\mu |\ Y = y] = \frac{\alpha + y}{(1/\beta + 1)}$$

Proof is: 1. Bayes' theorem, 2. Collect terms, 3. Recognize Gamma

Back to the thunder storm example

The data was:

Number of episodes, z_i	Number of days, $\#\ i$	Poisson expected
0	803	791.85
1	100	118.78
2	14	8.91
3+	3	0.46

- Notice that the observations have been summarized for us
- The real data would be something like:

Day	Number of storms
1	0
2	0
3	1
.	.
.	.
.	.
920	0

- The model we want to setup is fairly simple:

$$Y_i \sim NB(\alpha, 1/(1+\beta)), \quad \text{where } i = 1 \ldots 920.$$

- As the observations are collected, so can we collect the likelihood calculations

$$803 \cdot \ell(0) + 100 \cdot \ell(1) + 14 \cdot \ell(2) + 3 \cdot \ell(\geq 3)$$

- Remember that:

$$P(Y \geq 3) = 1 - P(Y = 0) - P(Y = 1) - P(Y = 2)$$

Detour: Bayesian inference

- Purely likelihood based inference (a.k.a. Frequentist inference) is based on drawing information from data Y about the model parameters θ via the **likelihood function**:

$$L(Y|\theta)$$

- In Bayesian inference **prior beliefs** about the model parameters are expressed as a probability density, so we have:

$$L(Y|\theta) \qquad \text{and} \qquad q(\theta|\psi)$$

- Inference about the model parameters are drawn from the **posterior density**:

$$p(\theta|Y=y) = \frac{L(Y=y|\theta)q(\theta|\psi)}{\int L(Y=y|\theta)q(\theta|\psi)d\theta}$$

which is computed via Bayes' rule.

Detour: Bayesian inference

- What is done here is to **update the prior beliefs** with data
- **If the data part is dominating** results close to likelihood inference can be expected
- Notice that the prior parameters ψ **are not influenced by data**. In hierarchical/mixed/random effects models we would **estimate those**.
- Notice that the **prior assumption is entirely subjective** and not subject to model validation. In hierarchical/mixed/random effects models we can - to some extend - validate our assumed distribution.

Detour: Bayesian inference

- Notice that the **integral in the posterior denominator** in general cannot be calculate analytically.
- Before the widespread use of MCMC* it was very **important to specify priors** such that the denominator integral could be calculated.
- A prior density is said to be **conjugated to a certain likelihood** if the posterior density has the **same parametric form** as the prior density.
- Using **conjugate priors simplifies the modeling**. To derive the posterior distribution, it is not necessary to perform the integration, as the posterior distribution is simply obtained by updating the parameters of the prior one.

---

*Markov Chain Monte Carlo methods are simulations techniques that allow you to sample a Markov chain with a desired equilibrium density, when that density is only know unnormalized

Reparameterization of the Gamma distribution

Instead of the usual parameterization of the gamma distribution of μ by its shape parameter α and scale parameter β, we may choose a parameterization by the **mean value**, $m = \alpha\beta$, and the **signal/noise ratio** $\gamma = \beta$

$$\gamma = \beta$$
$$m = \alpha\beta$$

The parameterization by m and γ implies that the degenerate **one-point distribution** of μ in a value m_0 may be obtained as **limiting distribution** for Gamma distributions with mean m_0 and signal/noise ratios $\gamma \to 0$. Moreover, under that limiting process the corresponding marginal distribution of Y (negative binomial) will **converge towards a Poisson distribution** with mean m_0.

Oversigt

Conjugate prior distributions

Definition (Standard conjugate distribution for an exponential dispersion family)

Consider an exponential dispersion family $\mathrm{ED}(\mu, V(\mu)/\lambda)$ for $\theta \in \Omega$. Let $\mathcal{M} = \tau(\Omega)$ denote the mean value space for this family. Let $m \in \mathcal{M}$ and consider

$$g_\theta(\theta; m, \gamma) = \frac{1}{C(m,\gamma)} \exp\left(\frac{\theta m - \kappa(\theta)}{\gamma}\right)$$

with

$$C(m,\gamma) = \int_\Omega \exp\left(\frac{\theta m - \kappa(\theta)}{\gamma}\right)\, d\theta$$

for all (positive) values of γ for which the integral converges.
This distribution is called the **standard conjugate distribution** for θ. The concept has its roots in the context of Bayesian parametric inference to describe a family of distributions whose densities have the structure of the likelihood kernel.

Conjugate prior distributions

- When the variance function, $V(\mu)$ is at most quadratic, the parameters m and γ have a simple interpretation in terms of the mean value parameter, $\mu = \tau(\theta)$, viz.

$$m = \mathrm{E}[\mu]$$

$$\gamma = \frac{\mathrm{Var}[\mu]}{\mathrm{E}[\mathrm{Var}(\mu)]}$$

  with $\mu = \mathrm{E}[Y|\theta]$, and with $\mathrm{Var}(\mu)$ denoting the variance function
- The use of the symbol γ is in agreement with our introduction of γ as **signal to noise ratio** for normally distributed observations and for the Poisson-Gamma hierarchical model.

Conjugate prior distributions

- When the variance function for the exponential dispersion family is at most quadratic, the standard **conjugate distribution** for μ **coincides** with the standard conjugate distribution for θ.
- However, for the Inverse Gaussian distribution, the standard conjugate distribution for μ is improper.
- The parameterization of the natural conjugate distribution for μ by the parameters m and γ has the advantage that **location and spread are described by separate parameters**. Thus, letting $\gamma \to 0$, the distribution of μ will converge towards a degenerate distribution with all its mass in m.

Oversigt

Density for Y_i	Sufficient statistic $T(Y_1,\ldots,Y_n)$	Density for T	$\mathsf{E}[T\|\theta]$	$\mathsf{V}[T\|\theta]$
$\mathsf{Bern}(\theta)$	$\sum Y_i$	$\mathsf{B}(n,\theta)$	$n\theta$	$n\theta(1-\theta)$
$\mathsf{B}(r,\theta)$	$\sum Y_i$	$\mathsf{B}(rn,\theta)$	$rn\theta$	$rn\theta(1-\theta)$
$\mathsf{Geo}(\theta)$	$\sum Y_i$	$\mathsf{NB}(n,\theta)$	$n\frac{1-\theta}{\theta}$	$n\frac{1-\theta}{\theta}^2$
$\mathsf{NB}(r,\theta)$	$\sum Y_i$	$\mathsf{NB}(rn,\theta)$	$rn\frac{1-\theta}{\theta}$	$rn\frac{1-\theta}{\theta}^2$
$\mathsf{P}(\theta)$	$\sum Y_i$	$\mathsf{P}(n\theta)$	$n\theta$	$n\theta$
$\mathsf{P}(r\theta)$	$\sum Y_i$	$\mathsf{P}(rn\theta)$	$rn\theta$	$rn\theta$
$\mathsf{Ex}(\theta)$	$\sum Y_i$	$\mathsf{G}(n,\theta)$	$n\theta$	$n\theta^2$
$\mathsf{G}(\alpha,\theta)$	$\sum Y_i$	$\mathsf{G}(n\alpha,\theta)$	$\alpha n\theta$	$\alpha n\theta^2$
$\mathsf{U}(0,\theta)$	$\max Y_i$	$\mathsf{Inv\text{-}Par}(\theta,n)$	$\frac{n\theta}{n+1}$	$\frac{n\theta^2}{(n+1)^2(n+2)}$
$\mathsf{N}(\theta,\sigma^2)$	$\sum Y_i$	$\mathsf{N}(n\theta,n\sigma^2)$	$n\theta$	$n\sigma^2$
$\mathsf{N}(\mu,\theta)$	$\sum (Y_i-\mu)^2$	$\mathsf{G}(n/2,2\theta)$	$n\theta$	$2n\sigma^2$
$\mathsf{N}_k(\boldsymbol{\theta},\boldsymbol{\Sigma})$	$\sum \boldsymbol{Y}_i$	$\mathsf{N}_k(n\boldsymbol{\theta},n\boldsymbol{\Sigma})$	$n\boldsymbol{\theta}$	$n\boldsymbol{\Sigma}$
$\mathsf{N}_k(\boldsymbol{\mu},\theta\boldsymbol{\Sigma})$	$\sum (\boldsymbol{Y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i-\boldsymbol{\mu})$	$\mathsf{G}(n/2,2\theta)$	$n\theta$	$2n\sigma^2$
$\mathsf{N}_k(\boldsymbol{\mu},\boldsymbol{\theta})$	$\sum (\boldsymbol{Y}_i-\boldsymbol{\mu})(\boldsymbol{Y}_i-\boldsymbol{\mu})^T$	$\mathsf{Wis}(k,n,\boldsymbol{\theta})$	$n\boldsymbol{\theta}$	

Table: Sufficient statistic $T(Y_1,\ldots,Y_n)$ (see p. 16 in the book) given a sample of n iid random variables $Y_1,Y_2,\ldots,Y_n$. Notice that in some cases the observation is a k dimensional random vector, and here a bold notation $\boldsymbol{Y}_i$ is used.

Conditional density of T given θ	Conjugate prior for θ	Posterior density for θ after the obs. $T = t(y_1, \ldots, y_n)$	Marginal density of $T = t(Y_1, \ldots, Y_n)$
$\mathsf{B}(n, \theta)$	$\mathsf{Beta}(\alpha, \beta)$	$\mathsf{Beta}(t + \alpha, n + \beta - t)$	$\mathsf{PI}(n, \alpha, \alpha + \beta)$
$\mathsf{NB}(n, \theta)$	$\mathsf{Beta}(\alpha, \beta)$	$\mathsf{Beta}(n + \alpha, \beta + t)$	$\mathsf{NPI}(n, \beta, \alpha + \beta)$
$\mathsf{P}(n\theta)$	$\mathsf{G}(\alpha, 1/\beta)$	$\mathsf{G}(t + \alpha, 1/(\beta + n)$	$\mathsf{NB}(\alpha, \beta/(\beta + n))$
$\mathsf{G}(n, \theta)$	$\mathsf{Inv\text{-}G}(\alpha, \beta)$	$\mathsf{Inv\text{-}G}(n + \alpha, \beta + t)$	$\mathsf{Inv\text{-}Beta}(\alpha, n, \beta)$
$\mathsf{Inv\text{-}Par}(\theta, n)$	$\mathsf{Par}(\beta, \mu)$	$\mathsf{Par}(\max(t, \beta), n + \mu)$	$\mathsf{BPar}(\beta, \mu, n)$
$\mathsf{N}(n\theta, n\sigma^2)$	$\mathsf{N}(\mu, \sigma_0^2)$	$\mathsf{N}(\mu_1, \sigma_1^2)$ $\mu_1 = (\mu/\sigma_0^2 + t/\sigma^2)$ $1/\sigma_1^2 = 1/\sigma_0^2 + n/\sigma^2$	$\mathsf{N}(n\mu, n\sigma^2 + n^2\sigma_0^2)$
$\mathsf{N}_k(n\boldsymbol{\theta}, n\boldsymbol{\Sigma})$	$\mathsf{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma_0})$	$\mathsf{N}_k(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$ $\boldsymbol{\mu_1} = \boldsymbol{\Sigma_1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}^{-1}\boldsymbol{t})$ $\boldsymbol{\Sigma_1}^{-1} = \boldsymbol{\Sigma_0}^{-1} + n\boldsymbol{\Sigma}^{-1}$	$\mathsf{N}_k(n\boldsymbol{\mu}, n\boldsymbol{\Sigma} + \boldsymbol{\Sigma_0})$

Table: Conditional densities of the statistic T given the parameter θ, conjugate prior densities for θ, posterior densities for θ after having observed the statistic $T = t(y_1, \ldots, y_n)$, and the marginal densities for $T = t(Y_1, \ldots, Y_n)$ – cf. also the discussion on page 16 and 17 in the book.(Notice that in some cases the observation is a random vector)

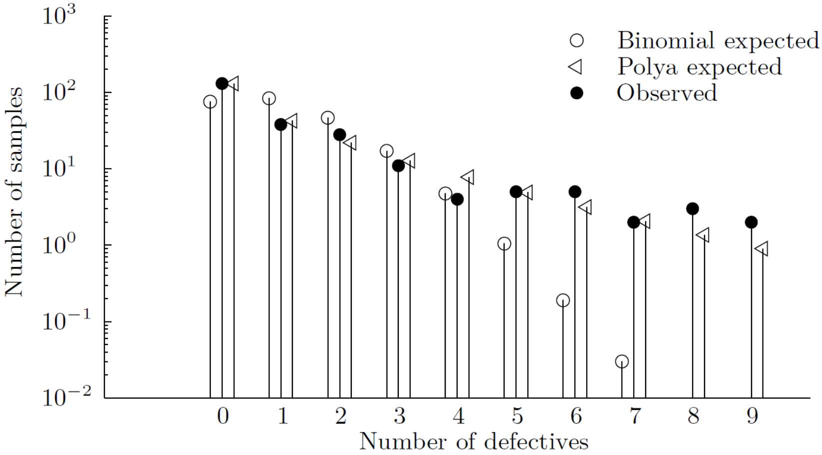
Oversigt

Hierarchical Beta-Binomial model

- Data describing the number of defective lids in samples of 770 lids from each of 229 samples.

No. defective	**No. samples**
0	131
1	38
2	28
3	11
4	4
5	5
6	5
7	2
8	3
9	2

- Notice that the data is summarized



Binomial expected
Polya expected
Observed
Number of samples
Number of defectives
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
0
1
2
3
4
5
6
7
8
9

Hierarchical Binomial-Beta distribution model

The natural conjugate distribution to the binomial is a Beta-distribution.

Theorem

Consider the generalized one-way random effects model for $Z_1, Z_2, \ldots, Z_k$ *given by*

$$\begin{aligned} Z_i|p_i &\sim B(n, p_i) \\ p_i &\sim \mathrm{Beta}(\alpha, \beta) \end{aligned}$$

i.e. the conditional distribution of Z_i *given* p_i *is a Binomial distribution, and the distribution of the mean value* p_i *is a Beta distribution. Then the marginal distribution of* Z_i *is a Polya distribution*
with probability function

$$P[Z = z] = g_Z(z) = \binom{n}{z} \frac{\Gamma(\alpha + z)}{\Gamma(\alpha)} \frac{\Gamma(\beta + n - z)}{\Gamma(\beta)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)}$$

for $z = 0, 1, 2, \ldots, n$.

Hierarchical Beta-Binomial distribution model

- The Polya distribution is named after the Hungarian mathematician G. Polya, who first described this distribution – although in another context.
- This distribution has:

$$\mathrm{E}[p] = \frac{\alpha}{\alpha+\beta}$$

$$\mathrm{Var}[p] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Altenative representation of Beta model ($\mu \in (0,1)$, $\phi > 0$)

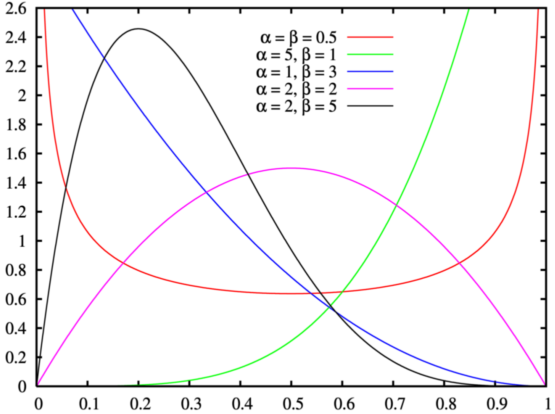
$$\alpha = \mu\phi; \quad \beta = (1-\mu)\phi$$

$$\mathrm{E}[p] = \mu; \quad \mathrm{Var}[p] = \frac{\mu(1-\mu)}{(\phi+1)}$$

ϕ serves as a precision parameter.

Why use a Beta to describe variation between samples?

- It has the desired support
- It is a very flexible distribution



- Last but not least the integral can be directly calculated.

Hierarchical Binomial-Beta distribution model

Theorem

The conditional distribution of $p|Z$ *is*

$$p|Z = z \sim Beta(\alpha + z, \beta + n - z)$$

Beetles exposed to ethylene oxide

Ten groups beetles were exposed to different concentrations of ethylene oxide and it was recorded how many died.

```
> conc <- c(24.8, 24.6, 23, 21, 20.6, 18.2, 16.8, 15.8, 14.7, 10.8)
> n <- c(30, 30, 31, 30, 26, 27, 31, 30, 31, 24)
> y <- c(23, 30, 29, 22, 23, 7, 12, 17, 10, 0)
```

The natural model is a binomial, and we wish to setup a logit-linear model as a function of the logarithm of the concentrations

$$
\begin{aligned}
y_i &\sim \text{Bin}(n_i, p_i)\ , \ \text{where} \\
\text{logit}(p_i) &= \mu + \beta \log(\text{conc}_i)
\end{aligned}
$$

```
> resp <- cbind(y, n - y)
> fit <- glm(resp ~ I(log(conc)), family = binomial())
```

Oversigt

Normal distributions with random variance

As a non-trivial example (and not given in the table) of a hierarchical distribution we consider the hierarchical normal distribution model with random variance:

Theorem

Consider a generalized one-way random effects model specified by

$$Y_i|\sigma_i^2 \sim N(\mu, \sigma_i^2)$$
$$1/\sigma_i^2 \sim G(\alpha, 1/\beta)$$

where σ_i^2 are mutually independent for $i = 1, \ldots, k$.
The marginal distribution of Y_i under this model is

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

where $\mathrm{t}(2\alpha)$ is a t-distribution with 2α degrees of freedom, i.e. a distribution with heavier tails than the normal distribution.

Overview