556: MATHEMATICAL STATISTICS I

FAMILIES OF DISTRIBUTIONS

4.1 Location-Scale Families

Definition: Location Scale Family
A **location-scale family** is a family of distributions formed by *translation* and *rescaling* of a standard family member.

Suppose that $f(x)$ is a pdf. Then if μ and $\sigma > 0$ are constants then

$$f(x|\mu,\sigma) = \frac{1}{\sigma} f((x-\mu)/\sigma)$$

is also a pdf; $f(x|\mu,\sigma) \geq 0$, and

$$\int_{-\infty}^{\infty} f(x|\mu,\sigma)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sigma} f((x-\mu)/\sigma)\,dx = \int_{-\infty}^{\infty} f(y)\,dy = 1$$

setting $y = (x-\mu)/\sigma$ in the penultimate integral.

- $f(x|\mu,\sigma)$ is termed a **location-scale** family
- if $\sigma = 1$ we have a **location** family: $f(x|\mu) = f(x-\mu)$
- if $\mu = 0$ we have a **scale** family: $f(x|\sigma) = f(x/\sigma)/\sigma$

Example : Normal distribution family

$$\begin{aligned} f(x) &= \left(\frac{1}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}x^2\right\} \\ f(x|\mu,\sigma) &= \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \end{aligned}$$

Example : Exponential distribution family

$$\begin{aligned} f(x) &= e^{-x} & x > 0 \\ f(x|\mu,\sigma) &= \frac{1}{\sigma} e^{-(x-\mu)/\sigma} & x > \mu \\ f(x|\mu) &= e^{-(x-\mu)} & x > \mu \end{aligned}$$

Note that X is a random variable with pdf $f_X(x) = f(x|\mu,\sigma)$ (the location-scale family member) **if and only if** there exists another random variable Z with $f_Z(z) = f(z)$ (the standard member) such that

$$X = \sigma Z + \mu$$

that is, if X is a linear (location-scale) transformation of a standard random variable Z.

4.2 Exponential Families

Definition: Exponential Family
A family of pdfs/pmfs is called an **exponential family** if it can be expressed

$$f(x|\underset{\sim}{\theta}) = h(x)c(\underset{\sim}{\theta}) \exp\left\{\sum_{j=1}^{k} w_j(\underset{\sim}{\theta}) t_j(x)\right\} = h(x)c(\underset{\sim}{\theta}) \exp\left\{\underset{\sim}{w}(\underset{\sim}{\theta})^\mathsf{T} \underset{\sim}{t}(x)\right\}$$

for all $x \in \mathbb{R}$, where $\underset{\sim}{\theta} \in \Theta$ is a d-dimensional parameter vector, and

- $h(x) \geq 0$ is a function that does not depend on $\underset{\sim}{\theta}$
- $c(\underset{\sim}{\theta}) \geq 0$ is a function that does not depend on x
- $\underset{\sim}{t}(x) = (t_1(x), \ldots, t_k(x))^\mathsf{T}$ is a vector of real-valued functions that do not depend on $\underset{\sim}{\theta}$
- $\underset{\sim}{w}(x) = (w_1(\underset{\sim}{\theta}), \ldots, w_k(\underset{\sim}{\theta}))^\mathsf{T}$ is a vector of real-valued functions that do not depend on x

An exponential family distribution is termed **natural** if $k = 1$ and $t_1(x) = x$.

Example : $Binomial(n, \theta)$ **for** $0 < \theta < 1$
For $x \in \{0, 1, \ldots, n\} \equiv \mathbb{X}$,

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{n}{x}(1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^x = \binom{n}{x}(1-\theta)^n \exp\left\{\log\left(\frac{\theta}{1-\theta}\right) x\right\}$$

- $k = 1$
- $h(x) = I_{\mathbb{X}}(x)\binom{n}{x}$, where $I_A(x)$ is the **indicator function** for set A

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

- $c(\underset{\sim}{\theta}) = (1-\theta)^n$
- $t_1(x) = x$
- $w_1(\underset{\sim}{\theta}) = \log\left(\frac{\theta}{1-\theta}\right)$

Example : $Normal(\mu, \sigma^2)$
For $x \in \mathbb{R}$,

$$f(x|\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right\}$$

- $k = 2$, $\underset{\sim}{\theta} = (\mu, \sigma^2)^\mathsf{T}$
- $h(x) = 1$
- $c(\underset{\sim}{\theta}) = c(\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}$
- $t_1(x) = -\frac{x^2}{2}$, $t_2(x) = x$
- $w_1(\underset{\sim}{\theta}) = \frac{1}{\sigma^2}$, $w_2(\underset{\sim}{\theta}) = \frac{\mu}{\sigma^2}$

Note that the support of an exponential family distribution $f(x|\underset{\sim}{\theta})$ cannot depend on $\underset{\sim}{\theta}$.

Example :

Suppose, for $\theta > 0$,

$$f(x|\theta) = \frac{1}{\theta} \exp\left\{1 - \frac{x}{\theta}\right\} \qquad x > \theta$$

and zero otherwise. Then

- $k = 1$, $\underset{\sim}{\theta} = \theta$
- $h(x) = eI_{[\theta,\infty)}(x)$
- $c(\underset{\sim}{\theta}) = 1/\theta$
- $t_1(x) = x$
- $w_1(\underset{\sim}{\theta}) = 1/\theta$

but the support of $f(x|\theta)$ depends on θ so this is not an exponential family distribution.

4.2.1 Parameterization

We can **reparameterize** an exponential family distribution from $\underset{\sim}{\theta}$ to $\underset{\sim}{\eta} = (\eta_1, \ldots, \eta_k)^\mathsf{T}$ by setting $\eta_j = w_j(\underset{\sim}{\theta})$ for each j, and write

$$f(x|\underset{\sim}{\eta}) = h(x)c^\star(\underset{\sim}{\eta}) \exp\left\{\sum_{j=1}^{k} \eta_j t_j(x)\right\} = h(x)c^\star(\underset{\sim}{\eta}) \exp\left\{\underset{\sim}{\eta}^\mathsf{T} \underset{\sim}{t}(x)\right\}.$$

$\underset{\sim}{\eta}$ is termed the **natural** or **canonical** parameter

Let $\mathcal{H}$ be the region of $\mathbb{R}^k$ defined by

$$\mathcal{H} \equiv \left\{\underset{\sim}{\eta} : \int_{-\infty}^{\infty} h(x) \exp\left\{\underset{\sim}{\eta}^\mathsf{T} \underset{\sim}{t}(x)\right\}\, dx < \infty\right\}$$

Then, for $\underset{\sim}{\eta} \in \mathcal{H}$, we must have

$$c^\star(\underset{\sim}{\eta}) = \left[\int_{-\infty}^{\infty} h(x) \exp\left\{\underset{\sim}{\eta}^\mathsf{T} \underset{\sim}{t}(x)\right\}\, dx\right]^{-1}$$

Note that

$$\left\{\underset{\sim}{w}(\underset{\sim}{\theta}) = (w_1(\underset{\sim}{\theta}), \ldots, w_k(\underset{\sim}{\theta}))^\mathsf{T} \ : \ \theta \in \Theta\right\}$$

is a subset of $\mathcal{H}$.

Example : $Binomial(n, \theta)$

Natural parameter:

$$\eta = \log\left(\frac{\theta}{1-\theta}\right) \qquad \Longleftrightarrow \qquad \theta = \frac{e^\eta}{1+e^\eta}$$

so that

$$f(x|\eta) = \left\{\binom{n}{x} I_{\{0,1,\ldots,n\}}(x)\right\} \frac{e^{\eta x}}{(1+e^\eta)^n}$$

Natural parameter space: here (interpreting the integral as a Lebesgue integral)

$$\int_{-\infty}^{\infty} h(x) \exp\left\{\underset{\sim}{\eta}^{\mathsf{T}} \underset{\sim}{t}(x)\right\} \, dx = \sum_{x=0}^{n} \binom{n}{x} \exp\{\eta x\} < \infty$$

for all finite values of η, so $\mathcal{H} \equiv \mathbb{R}$.

Example : $Normal(\mu, \sigma^2)$
Natural parameters:

$$\underset{\sim}{\eta} = (\eta_1, \eta_2)^{\mathsf{T}} = (1/\sigma^2, \mu/\sigma^2)^{\mathsf{T}}$$

so that

$$f(x|\underset{\sim}{\eta}) = \left(\frac{\eta_1}{2\pi}\right)^{1/2} \exp\left\{-\frac{\eta_2^2}{2\eta_1}\right\} \exp\left\{-\frac{\eta_1 x^2}{2} + \eta_2 x\right\}$$

Natural parameter space: this density will be integrable with respect to x if and only if $\eta_1 > 0$, so $\mathcal{H} \equiv \mathbb{R}^+ \times \mathbb{R}$.

Definition: Curved Exponential Family
An exponential family indexed by parameter θ is termed **curved** if

$$\dim(\underset{\sim}{\theta}) = d < k$$

4.2.2 Expectation and Variance for Exponential Families

Definition: Score Function
For pmf/pdf f_X with d-dimensional parameter $\underset{\sim}{\theta}$, the **score function**, $\underset{\sim}{S}(x; \underset{\sim}{\theta})$, is a $d \times 1$ vector with jth element equal to

$$S_j(x; \underset{\sim}{\theta}) = \frac{\partial}{\partial \theta_j} \log f_X(x|\underset{\sim}{\theta}).$$

The quantity $\underset{\sim}{S}(X; \underset{\sim}{\theta})$ is a d-dimensional **random variable**.

Lemma Under certain regularity conditions

$$E_{f_X}[\underset{\sim}{S}(X; \underset{\sim}{\theta})] = \underset{\sim}{0}$$

Proof In the case $d = 1$; let

$$\dot{f}_X(x|\theta) = \frac{d}{d\theta} f_X(x|\underset{\sim}{\theta})$$

Then

$$\begin{aligned} E_{f_X}[S(X;\theta)] = \int S(x;\theta) f_X(x|\theta) \, dx &= \int \left\{\frac{d}{d\theta} \log f_X(x|\underset{\sim}{\theta})\right\} f_X(x|\theta) \, dx \\ &= \int \left\{\frac{\dot{f}_X(x|\theta)}{f_X(x|\theta)}\right\} f_X(x|\theta) \, dx \\ &= \int \frac{d}{d\theta} f_X(x|\underset{\sim}{\theta}) \, dx \\ &= \frac{d}{d\theta} \left\{\int f_X(x|\underset{\sim}{\theta}) \, dx\right\} = 0 \end{aligned}$$

provided that the order of the differentiation wrt θ and the integration wrt x can be exchanged.

Definition: Fisher Information
For pmf/pdf f_X with d-dimensional parameter $\underset{\sim}{\theta}$, the **Fisher Information**, $\mathcal{I}(\underset{\sim}{\theta})$, is a $d \times d$ matrix defined as the variance-covariance matrix of the score random variable $\underset{\sim}{S}$, that is

$$\mathcal{I}(\underset{\sim}{\theta}) = Var_{f_X}[\underset{\sim}{S}(X;\underset{\sim}{\theta})] = E_{f_X}[\underset{\sim}{S}(X;\underset{\sim}{\theta})\underset{\sim}{S}(X;\underset{\sim}{\theta})^\mathsf{T}]$$

with (i,j)th element equal to

$$E_{f_X}[S_i(X;\underset{\sim}{\theta})S_j(X;\underset{\sim}{\theta})]$$

The Fisher Information is a constant $d \times d$ matrix in which each of the elements is a function of $\underset{\sim}{\theta}$.

Lemma Under certain regularity conditions, if the pmf/pdf is twice partially differentiable with respect to the elements of $\underset{\sim}{\theta}$, then

$$\mathcal{I}(\underset{\sim}{\theta}) = -E_{f_X}[\mathbf{\Psi}(X;\underset{\sim}{\theta})]$$

where $\mathbf{\Psi}(X;\underset{\sim}{\theta})$ is the $d \times d$ matrix of second partial derivatives with (i,j)th element equal to

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_X(x|\underset{\sim}{\theta}).$$

Proof In the case $d = 1$; from above

$$\int \left\{\frac{d}{d\theta}\log f_X(x|\underset{\sim}{\theta})\right\} f_X(x|\theta)\,dx = 0$$

so therefore, differentiating again wrt θ

$$\int \left[\left\{\frac{d^2}{d\theta^2}\log f_X(x|\underset{\sim}{\theta})\; f_X(x|\theta)\right\} + \left\{\frac{d}{d\theta}\log f_X(x|\underset{\sim}{\theta})\;\frac{d}{d\theta}f_X(x|\theta)\right\}\right]\,dx = 0 \tag{1}$$

But

$$\frac{d}{d\theta}\log f_X(x|\underset{\sim}{\theta}) = \frac{\dot{f}_X(x|\theta)}{f_X(x|\theta)} \qquad \therefore \qquad \dot{f}_X(x|\theta) = \frac{d}{d\theta}f_X(x|\theta) = f_X(x|\theta)\frac{d}{d\theta}\log f_X(x|\underset{\sim}{\theta})$$

so therefore

$$\int \frac{d}{d\theta}\log f_X(x|\underset{\sim}{\theta})\;\frac{d}{d\theta}f_X(x|\theta)\,dx = \int\left\{\frac{d}{d\theta}\log f_X(x|\underset{\sim}{\theta})\right\}^2 f_X(x|\theta)\,dx$$

and so substituting into equation (1) above, we have

$$\int\left\{\frac{d^2}{d\theta^2}\log f_X(x|\underset{\sim}{\theta})\; f_X(x|\theta)\right\}\,dx = -\int\left\{\frac{d}{d\theta}\log f_X(x|\underset{\sim}{\theta})\right\}^2 f_X(x|\theta)\,dx$$

of equivalently

$$E_{f_X}\left[\frac{d^2}{d\theta^2}\log f_X(x|\underset{\sim}{\theta})\right] = -E_{f_X}\left[\left\{\frac{d}{d\theta}\log f_X(x|\underset{\sim}{\theta})\right\}^2\right] = E_{f_X}[S(X;\theta)^2]$$

so that, as $E_{f_X}[S(X;\theta)] = 0$,

$$E_{f_X}\left[\frac{d^2}{d\theta^2}\log f_X(x|\underset{\sim}{\theta})\right] = -Var_{f_X}[S(X;\theta)].$$

Note that if $\underset{\sim}{a} = (a_1,\ldots,a_d)^\mathsf{T}$, then

$$Var_{f_X}[\underset{\sim}{a}^\mathsf{T}\underset{\sim}{S}(X;\underset{\sim}{\theta})] = \underset{\sim}{a}\mathcal{I}(\underset{\sim}{\theta})\underset{\sim}{a}^\mathsf{T}$$

Example : $Binomial(n, \theta)$

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad x \in \{0, 1, \ldots, n\}$$

so that

$$S(x;\theta) = \frac{d}{d\theta} \log f_X(x|\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{x - n\theta}{\theta(1-\theta)}.$$

Hence

$$E_{f_X}[S(X;\theta)] = E_{f_X}\left[\frac{X - n\theta}{\theta(1-\theta)}\right] = \frac{E_{f_X}[X] - n\theta}{\theta(1-\theta)} = 0$$

as $X \sim Binomial(n, \theta)$ yields $E_{f_X}[X] = n\theta$. For the second derivative

$$\frac{d^2}{d\theta^2} \log f_X(x|\theta) = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}$$

so that

$$\mathcal{I}(\theta) = -E_{f_X}\left[\frac{d^2}{d\theta^2} \log f_X(x|\theta)\right] = \frac{E_{f_X}[X]}{\theta^2} + \frac{n - E_{f_X}[X]}{(1-\theta)^2}$$

and as $E_{f_X}[X] = n\theta$, we have

$$\mathcal{I}(\theta) = \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

Example : $Poisson(\lambda)$

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x \in \{0, 1, \ldots\}$$

so that

$$S(x;\lambda) = \frac{d}{d\lambda} \log f_X(x|\lambda) = \frac{x}{\lambda} - 1$$

Hence

$$E_{f_X}[S(X;\lambda)] = E_{f_X}\left[\frac{X}{\lambda} - 1\right] = \frac{E_{f_X}[X]}{\lambda} - 1 = 0$$

as $X \sim Poisson(\lambda)$ yields $E_{f_X}[X] = \lambda$. For the second derivative

$$\frac{d^2}{d\lambda^2} \log f_X(x|\lambda) = -\frac{x}{\lambda^2}$$

so that

$$\mathcal{I}(\lambda) = -E_{f_X}\left[\frac{d^2}{d\lambda^2} \log f_X(x|\lambda)\right] = \frac{E_{f_X}[X]}{\lambda^2}$$

and as $E_{f_X}[X] = \lambda$, we have

$$\mathcal{I}(\lambda) = \frac{1}{\lambda}$$

Results for the Exponential Family
If

$$f_X(x|\underset{\sim}{\theta}) = h(x)c(\underset{\sim}{\theta}) \exp\left\{\sum_{j=1}^{k} w_j(\underset{\sim}{\theta})t_j(x)\right\}$$

then, for $l = 1, \ldots, d$,

$$S_l(x;\underset{\sim}{\theta}) = \frac{\partial}{\partial \theta_l} \log f_X(x|\underset{\sim}{\theta}) = \frac{\partial}{\partial \theta_l} \log c(\underset{\sim}{\theta}) + \sum_{j=1}^{k} \dot{w}_{jl}(\underset{\sim}{\theta})t_j(x) = \frac{\dot{c}_l(\underset{\sim}{\theta})}{c(\underset{\sim}{\theta})} + \sum_{j=1}^{k} \dot{w}_{jl}(\underset{\sim}{\theta})t_j(x)$$

where

$$\dot{c}_l(\underset{\sim}{\theta}) = \frac{\partial c(\underset{\sim}{\theta})}{\partial \theta_l} \qquad \dot{w}_{jl}(\underset{\sim}{\theta}) = \frac{\partial w_j(\underset{\sim}{\theta})}{\partial \theta_l}.$$

But, for each l, $E_{f_X}[S_l(X;\underset{\sim}{\theta})] = 0$, so therefore, for $l = 1, \ldots, d$,

$$E_{f_X}\left[\sum_{j=1}^{k} \dot{w}_{jl}(\underset{\sim}{\theta})t_j(X)\right] = -\frac{\dot{c}_l(\underset{\sim}{\theta})}{c(\underset{\sim}{\theta})} = -\frac{\partial}{\partial \theta_l} \log c(\underset{\sim}{\theta}).$$

By a similar calculation

$$Var_{f_X}\left[\sum_{j=1}^{k} \dot{w}_{jl}(\underset{\sim}{\theta})t_j(X)\right] = -\frac{\partial^2}{\partial \theta_l^2} \log c(\underset{\sim}{\theta}) - E_{f_X}\left[\sum_{j=1}^{k} \ddot{w}_{jll}(\underset{\sim}{\theta})t_j(X)\right]$$

where

$$\ddot{w}_{jll}(\underset{\sim}{\theta}) = \frac{\partial^2 w_j(\underset{\sim}{\theta})}{\partial \theta_l^2}$$

Example : $Binomial(n, \theta)$

$$f(x|\theta) = \binom{n}{x}(1-\theta)^n \exp\left\{\log\left(\frac{\theta}{1-\theta}\right)x\right\}$$

so that

$$w_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \qquad \log c(\theta) = n\log(1-\theta) \qquad S(x;\theta) = -\frac{n}{1-\theta} + \frac{x}{\theta(1-\theta)}.$$

From the result above

$$E_{f_X}[\dot{w}_{11}(\theta)t_1(X)] = -\frac{\partial}{\partial \theta_l} \log c(\underset{\sim}{\theta})$$

that is

$$E_{f_X}\left[\frac{1}{\theta(1-\theta)}X\right] = \frac{n}{1-\theta} \qquad \therefore \qquad E_{f_X}[X] = n\theta.$$

Note that in the **natural** (canonical) parameterization

$$\log f_X(x|\underset{\sim}{\eta}) = \log h(x) + \log c^\star(\underset{\sim}{\eta}) + \sum_{j=1}^{k} \eta_j t_j(X)$$

so that, using the arguments above for $l = 1, \ldots, d$,

$$E_{f_X}\left[t_l(X)\right] = -\frac{\partial}{\partial \eta_l} \log c^\star(\underset{\sim}{\eta}) \qquad Var_{f_X}\left[t_l(X)\right] = -\frac{\partial^2}{\partial \eta_l^2} \log c^\star(\underset{\sim}{\eta})$$

4.2.3 Independent random variables from the Exponential Family

Suppose that $X_1, \ldots, X_n$ are independent and identically distributed random variables, with pmf/pdf $f_X(x|\underset{\sim}{\theta})$ in the Exponential Family. Then the joint pmf/pdf for $\underset{\sim}{X} = (X_1, \ldots, X_n)^\mathsf{T}$ takes the form

$$f_{\underset{\sim}{X}}(\underset{\sim}{x}|\underset{\sim}{\theta}) = \prod_{i=1}^{n} f_X(x_i|\underset{\sim}{\theta}) = \prod_{i=1}^{n} h(x_i) c(\underset{\sim}{\theta}) \exp\left\{\sum_{j=1}^{k} w_j(\underset{\sim}{\theta}) t_j(x_i)\right\} = H(\underset{\sim}{x}) C(\underset{\sim}{\theta}) \exp\left\{\sum_{j=1}^{k} w_j(\underset{\sim}{\theta}) T_j(\underset{\sim}{x})\right\}$$

where

$$H(\underset{\sim}{x}) = \prod_{i=1}^{n} h(x_i) \qquad C(\underset{\sim}{\theta}) = \{c(\underset{\sim}{\theta})\}^n \qquad T_j(\underset{\sim}{x}) = \sum_{i=1}^{n} t_j(x_i).$$

4.2.4 Alternative construction of the Exponential Family

Suppose that $f(x)$ is a pmf/pdf with corresponding mgf $M(t)$ (presumed to exist in a neighbourhood of zero), so that

$$M(t) = \int e^{tx} f(x)\, dx = \exp\{K(t)\}$$

and $K(t) = \log M(t)$ is the cumulant generating function. Now suppose that $f(x) = \exp\{g(x)\}$. Then

$$\exp\{K(t)\} = M(t) = \int e^{tx} f(x)\, dx = \int e^{tx} e^{g(x)}\, dx = \int e^{tx+g(x)}\, dx.$$

Hence, dividing through by $\exp\{K(t)\}$, we have that

$$\int e^{tx+g(x)-K(t)}\, dx = 1$$

and also that the integrand is non-negative. Thus, for all t for which $M(t)$ exists,

$$f(x|t) = \exp\{tx + g(x) - K(t)\} = f(x) \exp\{tx - K(t)\}$$

is a valid pdf. If we set $t = \eta$, $h(x) = f(x) = \exp\{g(x)\}$ and $c^\star(\eta) = \exp\{-K(t)\}$, then

$$f(x|\eta) = h(x) c^\star(\eta) \exp\{\eta x\}$$

and we see that $f(x|\eta)$ is an exponential family member with natural parameter η. The pmf/pdf $f(x|t)$ is termed the **exponential tilting** of $f(x)$, with expectation

$$-\frac{d}{dt} \log c^\star(t) = -\frac{d}{dt}\{-K(t)\} = \dot{K}(t)$$

and variance

$$-\frac{d^2}{dt^2} \log c^\star(t) = -\frac{d^2}{dt^2}\{-K(t)\} = \ddot{K}(t).$$

4.2.5 The Exponential Dispersion Model

Consider the model

$$f(x|\underset{\sim}{\theta}, \phi) = \exp\left\{ d(x,\phi) + \frac{\log c(\underset{\sim}{\theta})}{r(\phi)} + \frac{1}{r(\phi)} \sum_{j=1}^{k} w_j(\underset{\sim}{\theta}) t_j(x) \right\} = h(x) c(\underset{\sim}{\theta}) \exp\left\{ \underset{\sim}{w}(\underset{\sim}{\theta})^{\mathsf{T}} \underset{\sim}{t}(x) \right\}$$

where $r(\phi) > 0$ is a function of **dispersion** parameter $\phi > 0$.

In this model, using the previous results, we see that the expectation is unchanged compared to the Exponential Family model by the presence of the term $r(\phi)$, but the variance is modified by a factor of $r(\phi)$.

Example : $Binomial(n, \theta)$

$$f_X(x|\theta) = \binom{n}{x} I_{\{0,1,\ldots,n\}}(x) \exp\left\{ n\log(1-\theta) + \log\left(\frac{\theta}{1-\theta}\right) x \right\}$$

Let $Y = X/n$, so that

$$f_Y(y|\theta,\phi) = \binom{1/\phi}{y/\phi} I_{\{0,\phi,2\phi,\ldots,1\}}(y/\phi) \exp\left\{ \frac{1}{\phi}\left[\log(1-\theta) + y\log\left(\frac{\theta}{1-\theta}\right)\right]\right\}$$

where $\phi = 1/n$. Note that

$$E_{f_Y}[Y] = \theta = \mu$$

say, and

$$Var_{f_Y}[Y] = \phi\theta(1-\theta) = \phi V(\mu)$$

where $V(\mu) = \mu(1-\mu)$ is the **variance function**.

Thus the exponential dispersion model allows separate modelling of mean and variance.

4.3 Convolution Families

The **convolution** of functions g and h is a function written $g \circ h$, which is defined by

$$g \circ h(y) = \int_{-\infty}^{\infty} g(x) h(y-x)\, dx.$$

Now if X_1 and X_2 are independent random variables with marginal pdfs f_{X_1} and f_{X_2} respectively, then the random variable $Y = X_1 + X_2$ has a pdf that can be determined using the multivariate transformation result. If we use dummy variable $Z = X_1$, then

$$\left.\begin{array}{rcl} Z & = & X_1 \\ Y & = & X_1 + X_2 \end{array}\right\} \quad \Longleftrightarrow \quad \left\{\begin{array}{rcl} X_1 & = & Z \\ X_2 & = & Y - Z \end{array}\right.$$

which is a transformation with Jacobian 1. Thus

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Z,Y}(z,y)\, dz = \int_{-\infty}^{\infty} f_{X_1,X_2}(z, y-z)\, dz = \int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(y-x)\, dx$$

so we can see that the pdf of Y is computed as the convolution of f_{X_1} and f_{X_2}.

A family of distributions, $\mathcal{F}$, is **closed under convolution** if

$$f_1, f_2 \in \mathcal{F} \qquad \Longrightarrow \qquad f_1 \circ f_2 \in \mathcal{F}$$

For independent random variables X_1 and X_2 with pdfs f_1 and f_2 in a family $\mathcal{F}$, closure under convolution implies that the random variable $Y = X_1 + X_2$ also has a pdf in $\mathcal{F}$.

This concept is closely related to the idea of **infinite divisibility**, **decomposibility**, and **self-decomposibility**.

- **Infinite Divisibility :** A probability distribution for rv X is *infinitely divisible* if, for all positive integers n, there exists a **sequence of independent and identically distributed** rvs $Z_{n1}, \ldots, Z_{nn}$ such that X and

$$Z_n = \sum_{j=1}^{n} Z_{nj}$$

have the same distribution, that is, the characteristic function of X can be written

$$C_X(t) = \{C_Z(t)\}^n$$

for some characteristic function C_Z.

- **Decomposability :** A probability distribution for rv X is *decomposable* if

$$C_X(t) = C_{X_1}(t) C_{X_2}(t)$$

for two characteristic functions C_{X_1} and C_{X_2} so that

$$X = X_1 + X_2$$

where X_1 and X_2 are **independent** rvs with characteristic functions C_{X_1} and C_{X_2}.

- **Self-Decomposability :** A probability distribution for rv X is *self-decomposable* if

$$C_X(t) = \{C_{X_1}(t)\}^2$$

for characteristic function C_{X_1} so that

$$X = X_1 + X_2$$

where X_1 and X_2 are **independent identically distributed** rvs with characteristic function C_{X_1}.

4.4 Hierarchical Models

A hierarchical model is a model constructed by considering a series of distributions at different levels of a "hierarchy" that together, after marginalization, combine to yield the distribution of the observable quantities.

Example : A three-level model

Consider the **three-level** hierarchical model:

LEVEL 3 :	$\lambda > 0$	Fixed parameter
LEVEL 2 :	$N \sim Poisson(\lambda)$	
LEVEL 1 :	$X \vert N = n, \theta \sim Binomial(n, \theta)$	

Then the marginal pmf for X is given by

$$f_X(x|\theta, \lambda) = \sum_{n=0}^{\infty} f_{X|N}(x|n, \theta, \lambda) f_N(n|\lambda).$$

By elementary calculation, we see that $X \sim Poisson(\lambda\theta)$

$$f_X(x|\theta, \lambda) = \frac{(\lambda\theta)^x e^{-\lambda\theta}}{x!} \qquad x = 0, 1, \ldots.$$

Example : A three-level model

Consider the **three-level** hierarchical model:

LEVEL 3 :	$\alpha, \beta > 0$	Fixed parameters
LEVEL 2 :	$Y \sim Gamma(\alpha, \beta)$	
LEVEL 1 :	$X \vert Y = y \sim Poisson(y)$	

Then the marginal pdf for X is given by

$$f_X(x|\alpha, \beta) = \int_0^\infty f_{X|Y}(x|y) f_Y(y|\alpha, \beta)\, dy.$$

A general K-level hierarchical model can be specified in terms of K vector random variables:

$$\begin{array}{rcl} \text{LEVEL K} & : & \underset{\sim}{X}_K = (X_{K1}, \ldots, X_{Kn_K})^\mathsf{T} \\ \vdots & : & \vdots \\ \text{LEVEL 2} & : & \underset{\sim}{X}_2 = (X_{21}, \ldots, X_{2n_2})^\mathsf{T} \\ \text{LEVEL 1} & : & \underset{\sim}{X}_1 = (X_{11}, \ldots, X_{1n_1})^\mathsf{T} \end{array}$$

The hierarchical model specifies the joint distribution via a series of **conditional independence** assumptions, so that

$$f_{\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_K}(\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_K) = f_{\underset{\sim}{X}_K}(\underset{\sim}{x}_k) \prod_{k=1}^{K-1} f_{\underset{\sim}{X}_k|\underset{\sim}{X}_{k+1}}(\underset{\sim}{x}_k|\underset{\sim}{x}_{k+1})$$

where

$$f_{\underset{\sim}{X}_k|\underset{\sim}{X}_{k+1}}(\underset{\sim}{x}_k|\underset{\sim}{x}_{k+1}) = \prod_{j=1}^{n_k} f_k(x_{kj}|\underset{\sim}{x}_{k+1})$$

that is, at level k in the hierarchy, the random variables are taken to be **conditionally independent** given the values of variables at level $k+1$.

The uppermost level, Level K, can be taken to be a degenerate model, with mass function equal to 1 at a set of fixed values.

Example : A three-level model

Consider the **three-level** hierarchical model:

LEVEL 3 :	$\theta, \tau^2 > 0$ Fixed parameters	
LEVEL 2 :	$M_1, \ldots, M_L \sim Normal(\theta, \tau^2)$	Independent
LEVEL 1 :	For $l = 1, \ldots, L$: $X_{l1}, \ldots, X_{ln_l} \vert M_l = m_l \sim Normal(m_l, 1)$ where all the X_{lj} are conditionally independent given $M_1, \ldots, M_L$	

For random variables X, Y and Z, we write $X \perp Y \mid Z$ if X and Y are conditionally independent given Z, so that in the above model

$$X_{l_1 j_1} \perp X_{l_2 j_2} \mid M_1, \ldots, M_L$$

for all l_1, j_1, l_2, j_2.

Special Cases of Hierarchical Models

1. **Finite Mixture Models**

$$\text{LEVEL 3}: \quad L \geq 1 \text{ (integer)}, \pi_1, \ldots, \pi_l \text{ with } 0 \leq \pi_l \leq 1 \text{ and } \sum_{l=1}^{L} \pi_l = 1, \text{ and } \theta_1, \ldots, \theta_L$$

$$\text{LEVEL 2}: \quad X \sim f_X(x|\underset{\sim}{\pi}, L) \text{ with } \mathbb{X} \equiv \{1, 2, \ldots, L\} \text{ such that } \Pr[X = l] = \pi_l$$

$$\text{LEVEL 1}: \quad Y|X = l \sim f_l(y|\theta_l)$$

where f_l is some pmf or pdf with parameters θ_l. Then

$$f_Y(y|\underset{\sim}{\pi}, \underset{\sim}{\theta}, L) = \sum_{l=1}^{L} f_{Y|X}(y|x) f_X(x) = \sum_{l=1}^{L} f_l(y|\theta_l)\pi_l$$

This is a **finite mixture distribution**: the observed Y are drawn from L distinct sub-populations characterized by pmf/pdf $f_1, \ldots, f_L$ and parameters $\theta_1, \ldots, \theta_L$, with sub-population proportions $\pi_1, \ldots, \pi_L$.

Note that if $M_1, \ldots, M_L$ are the mgfs corresponding to $f_1, \ldots, f_L$, then

$$M_Y(t) = \sum_{l=1}^{L} \pi_l M_l(t)$$

2. **Random Sums**

$$\text{LEVEL 3}: \quad \underset{\sim}{\theta}, \underset{\sim}{\phi} \qquad \text{(fixed parameters)}$$

$$\text{LEVEL 2}: \quad X \sim f_X(x|\underset{\sim}{\phi}) \text{ with } \mathbb{X} \equiv \{0, 1, 2, \ldots\}$$

$$\text{LEVEL 1}: \quad Y_1, \ldots, Y_n|X = x \sim f_Y(y|\underset{\sim}{\theta}) \text{ (independent), and } S = \sum_{i=1}^{x} Y_i$$

Then, by the law of iterated expectation,

$$\begin{aligned} M_S(t) = E_{f_S}\left[e^{tS}\right] &= E_{f_X}\left[E_{f_{S|X}}\left[e^{tS}|X = x\right]\right] \\ &= E_{f_X}\left[E_{f_{\underset{\sim}{Y}|X}}\left[\exp\left\{t\sum_{i=1}^{x} Y_i\right\}|X = x\right]\right] \\ &= E_{f_X}\left[\{M_Y(t)\}^X\right] \\ &= G_X(M_Y(t)) \end{aligned}$$

where G_X is the factorial mgf (or pgf) for X. By a similar calculation,

$$G_S(t) = G_X(G_Y(t)).$$

For example, if $X \sim Poisson(\phi)$, then

$$G_S(t) = \exp\{\phi(G_Y(t) - 1)\}$$

is the pgf of S. Expanding the pgf as a power series in t yields the pmf of S.

Example : Branching Process

Consider a sequence of generations of an organism; let S_i be the total number of individuals in the ith generation, for $i = 0, 1, 2, \ldots$. Suppose that f_X is a pmf with support $\mathbb{X} \equiv \{0, 1, 2, \ldots\}$.

- **Generation 0** : $S_0 \sim f_X(x|\phi)$
- **Generation 1** : Given $S_0 = s_0$, let

$$S_{11}, \ldots, S_{1s_0}|S_0 = s_0 \quad \text{such that} \quad S_{1j} \sim f_X(x|\phi), \text{ with } S_{1j_1} \perp S_{1j_2} \text{ for all } j_1, j_2$$

  and set

$$S_1 = \sum_{j=1}^{s_0} S_{1j}$$

  is the total number of individuals in the 1st generation. S_{1j} is the number of offspring of the jth individual in the zeroth generation.

- **Generation i** : Given $S_{i-1} = s_{i-1}$, let

$$S_{i1}, \ldots, S_{is_{i-1}}|S_{i-1} = s_{i-1} \quad \text{such that} \quad S_{ij} \sim f_X(x|\phi) \text{ (independent)}$$

  and set

$$S_i = \sum_{j=1}^{s_{i-1}} S_{ij}$$

Let G_i be the pgf of S_i. Then, by recursion, we have

$$G_i(t) = G_{i-1}(G_X(t)) = G_{i-2}(G_X(G_X(t))) = \cdots = G_X(G_X(\cdots G_X(G_X(t))\cdots))$$

that is, an $i+1$-fold iterated calculation.

3. **Location-Scale Mixtures**

$$\begin{array}{lll} \text{LEVEL 3 :} & \underset{\sim}{\theta} & \text{Fixed parameters} \\ \text{LEVEL 2 :} & M, V \sim f_{M,V}(m, v|\underset{\sim}{\theta}) & \\ \text{LEVEL 1 :} & Y|M = m, V = v \sim f_{Y|M,V}(y|m, v) & \end{array}$$

where

$$f_{Y|M,V}(y|m, v) = \frac{1}{v} f\left(\frac{y-m}{v}\right)$$

that is a location-scale family distribution, mixed over different location and scale parameters with *mixing distribution* $f_{M,V}$.

Example : Scale Mixtures of Normal Distributions

$$\begin{array}{ll} \text{LEVEL 3 :} & \underset{\sim}{\theta} \\ \text{LEVEL 2 :} & V \sim f_V(v|\underset{\sim}{\theta}) \\ \text{LEVEL 1 :} & Y|V = v \sim f_{Y|V}(y|v) \equiv Normal(0, g(v)) \end{array}$$

for some positive function g.

For example, if

$$Y|V = v \sim Normal(0, v^{-1}) \qquad V \sim Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$$

then by elementary calculations, we find that

$$f_Y(y) = \frac{1}{\pi}\frac{1}{1+y^2} \qquad y \in \mathbb{R} \qquad \therefore \qquad Y \sim Cauchy.$$

The scale mixture of normal distributions family includes the *Student*, *Double Exponential* and *Logistic* as special cases.

Moments of location-scale mixtures can be computed using the law of iterated expectation. The location-scale mixture construction allows the modelling of

- **skewness** through the mixture over different *locations*
- **kurtosis** through the mixture over different *scales*

Example : Location-Scale Mixtures of Normal Distributions

Suppose M and V are independent, with

$$M \sim Exponential(1/2) \qquad V \sim Gamma(2, 1/2)$$

and

$$Y|M = m, V = v \sim Normal(m, 1/v)$$

Then the marginal distribution of Y is given by

$$f_Y(y) = \int_0^\infty \int_0^\infty f_{Y|M,V}(y|m,v) f_M(m) f_V(v)\, dm\, dv$$

which can most readily be examined by simulation. The figure below depicts a histogram of 10000 values simulated from the model, and demonstrates the skewness of the marginal of Y.