556: MATHEMATICAL STATISTICS I

FAMILIES OF DISTRIBUTIONS: RESULTS AND EXAMPLES

1. **Parametric Family:** A *parametric family*, $\mathcal{P}$, of distributions is a collection of probability distributions indexed by a finite-dimensional parameter, θ:

$$\mathcal{P} \equiv \{P_X(.;\theta) : \theta \in \Theta\}$$

which may be written equivalently in terms of the cdfs $F_X(.;\theta)$ for $\theta \in \Theta$. The family is *identifiable* if, for $\theta_1, \theta_2 \in \Theta$

$$F_X(.;\theta_1) = F_X(.;\theta_2) \quad \text{for all } x \qquad \Longleftrightarrow \qquad \theta_1 = \theta_2.$$

Typically, θ is an $m \times 1$ vector of real-valued quantities.

- Suppose $\theta_0 \in \Theta$, and suppose $X \sim F_X(x;\theta_0)$. Suppose $\theta_1 \in \Theta$ and consider the *likelihood ratio*

$$R(X;\theta_0,\theta_1) = \frac{f_X(X;\theta_1)}{f_X(X;\theta_0)} = \frac{dF_X(X;\theta_1)}{dF_X(X;\theta_0)}$$

say. Then

$$\mathbb{E}_X[R(X;\theta_0,\theta_1)] = \int \frac{f_X(x;\theta_1)}{f_X(x;\theta_0)} dF_X(x;\theta_0) = \int \frac{dF_X(x;\theta_1)}{dF_X(x;\theta_0)} dF_X(x;\theta_0) = \int dF_X(x;\theta_1) = 1.$$

- Suppose that the pmf/pdf $f_X(x;\theta)$ is differentiable with respect to θ. The *score function*, $\mathbf{S}(x;\theta)$, is a $m \times 1$ vector with jth element equal to

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

The quantity $\mathbf{S}(X;\theta) = (S_1(X;\theta), \ldots, S_m(X;\theta))^\top$ is an m-dimensional *random variable*. Under certain regularity conditions

$$\mathbb{E}_X[\mathbf{S}(X;\theta)] = \mathbf{0} \qquad (m \times 1).$$

Consider first $m = 1$; let

$$\dot{f}_X(x;\theta) = \frac{d}{d\theta} f_X(x;\theta)$$

Then

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

provided that the order of the differentiation wrt θ and the integration wrt x can be exchanged. The result for general m follows by noting that

$$\mathbb{E}_X[\mathbf{S}(X;\theta)] = \begin{bmatrix} \mathbb{E}_X[S_1(X;\theta)] \\ \vdots \\ \mathbb{E}_X[S_m(X;\theta)] \end{bmatrix}$$

and applying the calculation for $m = 1$ for each component.

- The *Fisher Information*, $\mathcal{I}(\theta)$, is an $m \times m$ matrix defined as the variance-covariance matrix of the score random variable $\mathbf{S}$, that is

$$\mathcal{I}(\theta) = \text{Var}_X[\mathbf{S}(X;\theta)] = \mathbb{E}_X[\mathbf{S}(X;\theta)\mathbf{S}(X;\theta)^\top]$$

with (j,k)th element equal to

$$\mathbb{E}_X[S_j(X;\theta)S_k(X;\theta)]$$

The Fisher Information is a constant $m \times m$ matrix with elements that are functions of θ. Under certain regularity conditions, if the pmf/pdf is twice partially differentiable with respect to the elements of θ, then

$$\mathcal{I}(\theta) = -\mathbb{E}_X[\mathbf{\Psi}(X;\theta)]$$

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

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k} \log f_X(X;\theta).$$

In the continuous case, with $m = 1$: from above

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

so therefore, differentiating again wrt θ

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

But

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

so therefore

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

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

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

of equivalently

$$\mathbb{E}_X\left[\frac{d^2}{d\theta^2} \log f_X(x;\theta) \right] = -\mathbb{E}_X\left[\left\{ \frac{d}{d\theta} \log f_X(x;\theta) \right\}^2 \right] = \mathbb{E}_X[S(X;\theta)^2]$$

so that, as $\mathbb{E}_X[S(X;\theta)] = 0$,

$$\mathbb{E}_X\left[\frac{d^2}{d\theta^2} \log f_X(x;\theta) \right] = -\text{Var}_X[S(X;\theta)].$$

Example : $Binomial(n, \theta)$

$$f_X(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad x \in \{0, 1, \dots, n\}$$

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

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

as $X \sim Binomial(n, \theta)$ yields $\mathbb{E}_X[X] = n\theta$. 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

$$\mathcal{I}(\theta) = -\mathbb{E}_X\left[\frac{d^2}{d\theta^2}\log f_X(X;\theta)\right] = \frac{\mathbb{E}_X[X]}{\theta^2} + \frac{n-\mathbb{E}_X[X]}{(1-\theta)^2}$$

and as $\mathbb{E}_X[X] = n\theta$, we have

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

Example : $Poisson(\lambda)$

$$f_X(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x \in \{0, 1, \dots\}$$

so that

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

Hence

$$\mathbb{E}_X[S(X;\lambda)] = \mathbb{E}_X\left[\frac{X}{\lambda} - 1\right] = \frac{\mathbb{E}_X[X]}{\lambda} - 1 = 0$$

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

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

so that

$$\mathcal{I}(\lambda) = -\mathbb{E}_X\left[\frac{d^2}{d\lambda^2}\log f_X(X;\lambda)\right] = \frac{\mathbb{E}_X[X]}{\lambda^2} = \frac{1}{\lambda}.$$

2. **Location-Scale Family:** A *location-scale family* is a family of distributions formed by *translation* and *rescaling* of a standard family member. Suppose that $f_0(x)$ is a pdf. If μ and $\sigma > 0$ are constants then

$$f_X(x;\mu,\sigma) = \frac{1}{\sigma} f_0((x-\mu)/\sigma)$$

is also a pdf.

- if $\sigma = 1$ we have a *location* family: $f_X(x;\mu) = f_0(x-\mu)$
- if $\mu = 0$ we have a *scale* family: $f_X(x;\sigma) = f_0(x/\sigma)/\sigma$

Example : Normal distribution family

$$\begin{aligned} f_0(x) &= \left(\frac{1}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}x^2\right\} \\ f_X(x;\mu,\sigma) &= \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \end{aligned}$$

Example : Exponential distribution family

$$\begin{aligned} f_0(x) &= e^{-x} & x > 0 \\ f_X(x;\mu,\sigma) &= \frac{1}{\sigma} e^{-(x-\mu)/\sigma} & x > \mu \end{aligned}$$

Note that X is a random variable with pdf $f_X(x) = f_X(x;\mu,\sigma)$ (the location-scale family member) **if and only if** there exists another random variable Z with $f_Z(z) = f_0(z)$ (the standard member) such that $X = \sigma Z + \mu$ that is, if X is a linear transformation of a standard random variable Z.

3. **Exponential Families:** A family of pdfs/pmfs is an *Exponential Family* if it can be expressed

$$f_X(x;\theta) = h(x) \exp\left\{\sum_{j=1}^{m} c_j(\theta) T_j(x) - A(\theta)\right\} = h(x) \exp\left\{c(\theta)^\top \mathbf{T}(x) - A(\theta)\right\}$$

for all $x \in \mathbb{R}$, where $\theta \in \Theta$ is a l-dimensional parameter vector (initially we take $l = m$).

- $h(x) \geq 0$ is a function that does not depend on θ
- $A(\theta)$ is a function that does not depend on x
- $\mathbf{T}(x) = (T_1(x), \ldots, T_m(x))^\top$ is a vector of real-valued functions that do not depend on θ.
- $c(x) = (c_1(\theta), \ldots, c_m(\theta))^\top$ is a vector of real-valued functions that do not depend on x.
- The support of $f_X(x;\theta)$ **does not** depend on θ.
- The family is termed *natural* if $m = 1$ and $T_1(x) = x$.

Example : $Binomial(n, \theta)$ **for** $0 < \theta < 1$

For $x \in \{0, 1, \ldots, n\} \equiv \mathbb{X}$,

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

- $m = 1$
- $h(x) = \mathbb{1}_{\mathbb{X}}(x)\binom{n}{x}$.
- $A(\theta) = n\log(1-\theta)$
- $T_1(x) = x$
- $c_1(\theta) = \log(\theta/(1-\theta)) = \log\theta - \log(1-\theta)$

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

For $x \in \mathbb{R}$,

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

- $m = 2$, $\theta = (\mu, \sigma^2)^\top$
- $h(x) = 1/\sqrt{2\pi}$
- $A(\theta) = A(\mu, \sigma^2) = (\log\sigma^2 + \mu^2/\sigma^2)/2$
- $T_1(x) = -x^2/2$, $T_2(x) = x$
- $c_1(\theta) = 1/\sigma^2$, $c_2(\theta) = \mu/\sigma^2$

Example : Suppose, for $\theta > 0$

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

and zero otherwise. Then

- $m = 1$, $\theta = \theta$
- $h(x) = e\mathbb{1}_{[\theta,\infty)}(x)$
- $A(\theta) = \log\theta$
- $T_1(x) = x$
- $c_1(\theta) = -1/\theta$

but the support of $f_X(x;\theta)$ depends on θ so this is **not** an Exponential Family distribution.

- **Parameterization** We can *reparameterize* from θ to $\eta = (\eta_1, \ldots, \eta_m)^\top$ by setting $\eta_j = c_j(\theta)$ for each j, and write

$$f_X(x;\eta) = h(x)\exp\left\{\sum_{j=1}^m \eta_j T_j(x) - K(\eta)\right\} = h(x)\exp\left\{\eta^\top \mathbf{T}(x) - K(\eta)\right\}.$$

  η is termed the *natural* or *canonical* parameter and

$$K(\eta) = A(c^{-1}(\eta))$$

- **Parameter space:** Let $\mathcal{H}$ be the region of $\mathbb{R}^m$ defined by

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

  $\mathcal{H}$ is the *natural parameter space*. For $\eta \in \mathcal{H}$, we must have

$$\exp\{K(\eta)\} = \int_{-\infty}^{\infty} h(x) \exp\left\{ \eta^\top \mathbf{T}(x) \right\} \, dx$$

  It can be shown that $\mathcal{H}$ is a *convex* set, that is, for $0 \leq \lambda \leq 1$,

$$\eta_1, \eta_2 \in \mathcal{H} \quad \Longrightarrow \quad \lambda \eta_1 + (1-\lambda)\eta_2 \in \mathcal{H}.$$

  Note that

$$\mathcal{H}_\Theta = \left\{ c(\theta) = (c_1(\theta), \ldots, c_m(\theta))^\top \; : \; \theta \in \Theta \right\} \subseteq \mathcal{H}.$$

  $\mathcal{H}_\Theta$ can be considered the *natural parameter space induced by* Θ

  Example : $Binomial(n, \theta)$

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

  so that

$$f_X(x;\eta) = \left\{ \binom{n}{x} \mathbb{1}_{\{0,1,\ldots,n\}}(x) \right\} \exp\{\eta x - n\log(1+e^\eta)\}.$$

  Natural parameter space:

$$\int_{-\infty}^{\infty} h(x) \exp\left\{ \eta^\top \mathbf{T}(x) \right\} \, dx = \sum_{x=0}^{n} \binom{n}{x} \exp\{\eta x\} < \infty \quad \forall \eta \qquad \therefore \qquad \mathcal{H} \equiv \mathbb{R}.$$

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

$$\eta = (\eta_1, \eta_2)^\top = (1/\sigma^2, \mu/\sigma^2)^\top$$

  so that

$$f_X(x;\eta) = \left(\frac{\eta_1}{2\pi}\right)^{1/2} \exp\left\{ -\frac{\eta_2^2}{2\eta_1} \right\} \exp\left\{ -\frac{\eta_1 x^2}{2} + \eta_2 x \right\}$$

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

- **Regular Exponential Family:** The family is termed *regular* if
  I. $\mathcal{H} \equiv \mathcal{H}_\Theta$.
  II. In the natural parameterization, neither the η_j nor the $T_j(x)$ satisfy linearity constraints.
  III. $\mathcal{H}$ is an *open* set in $\mathbb{R}^m$.

  If only I. and II. hold, the exponential family is termed *full*.

- **Curved Exponential Family:** The family is termed *curved* if

$$\dim(\theta) = l < m$$

- **Results for the Exponential Family:** If

$$f_X(x;\theta) = h(x)\exp\left\{\sum_{j=1}^{m} c_j(\theta)T_j(x) - A(\theta)\right\}$$

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

$$S_l(x;\theta) = \frac{\partial}{\partial\theta_l}\log f_X(x;\theta) = \sum_{j=1}^{m}\frac{\partial c_j(\theta)}{\partial\theta_l}T_j(x) - \frac{\partial A(\theta)}{\partial\theta_l} = \sum_{j=1}^{m}\dot{c}_{jl}(\theta)T_j(x) - \dot{A}_l(\theta)$$

say. But, for each l, $\mathbb{E}_X[S_l(X;\theta)] = 0$, so therefore, for $l = 1,\ldots,m$,

$$\mathbb{E}_X\left[\sum_{j=1}^{m}\dot{c}_{jl}(\theta)T_j(X)\right] = \dot{A}_l(\theta).$$

By a similar calculation

$$\text{Var}_X\left[\sum_{j=1}^{m}\dot{c}_{jl}(\theta)T_j(X)\right] = \ddot{A}_{ll}(\theta) - \mathbb{E}_X\left[\sum_{j=1}^{m}\ddot{c}_{jll}(\theta)T_j(X)\right]$$

where

$$\ddot{A}_{ll}(\theta) = \frac{\partial^2 A(\theta)}{\partial\theta_l^2} \qquad \ddot{w}_{jll}(\theta) = \frac{\partial^2 c_j(\theta)}{\partial\theta_l^2}$$

Example : $Binomial(n,\theta)$

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

so that

$$c_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \qquad A(\theta) = -n\log(1-\theta) \qquad S(x;\theta) = -\frac{n}{1-\theta} + \frac{x}{\theta(1-\theta)}.$$

From the result above

$$\mathbb{E}_X\left[\dot{c}_{11}(\theta)T_1(X)\right] = \dot{A}_l(\theta)$$

that is

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

Note that in the natural (canonical) parameterization

$$\log f_X(x;\eta) = \log h(x) + \sum_{j=1}^{m}\eta_j T_j(x) - K(\eta)$$

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

$$\mathbb{E}_X\left[T_l(X)\right] = \dot{K}_l(\theta) \qquad \text{Var}_X\left[T_l(X)\right] = \ddot{K}_{ll}(\theta)$$

- **Independent random variables from the Exponential Family**
  Suppose that $X_1, \ldots, X_n$ are independent and identically distributed rvs, with pmf or pdf $f_X(x;\theta)$ in the Exponential Family. Then the joint pmf/pdf for $\mathbf{X} = (X_1, \ldots, X_n)^\top$ takes the form

$$f_{\mathbf{X}}(\mathbf{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = \prod_{i=1}^{n} h(x_i) \exp\left\{ \sum_{j=1}^{m} c_j(\theta) T_j(x_i) - A(\theta) \right\}$$

$$= H(\mathbf{x}) \exp\left\{ \sum_{j=1}^{m} c_j(\theta) T_j(\mathbf{x}) - nA(\theta) \right\}$$

  where

$$H(\mathbf{x}) = \prod_{i=1}^{n} h(x_i) \qquad T_j(\mathbf{x}) = \sum_{i=1}^{n} T_j(x_i).$$

- **Alternative construction of the Exponential Family** Suppose that $f_0(x)$ is a pmf/pdf with corresponding mgf $M(t)$ (presumed to exist in a neighbourhood of zero), so that

$$M(t) = \int e^{tx} f_0(x)\, dx = \exp\{K(t)\}$$

  and $K(t) = \log M(t)$ is the *cumulant generating function*. Now suppose that $f_0(x) = \exp\{g_0(x)\}$. Then

$$\exp\{K(t)\} = M(t) = \int e^{tx} e^{g_0(x)}\, dx = \int e^{tx + g_0(x)}\, dx.$$

  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,

$$f_X(x;t) = \exp\{tx + g_0(x) - K(t)\} = f_0(x) \exp\{tx - K(t)\}$$

  is a valid pdf. If we set $t = \eta$, $h(x) = f_0(x) = \exp\{g_0(x)\}$ then

$$f_X(x;\eta) = h(x) \exp\{\eta x - K(\eta)\}$$

  and we see that $f_X(x;\eta)$ is an exponential family member with natural parameter η. The pmf/pdf $f_X(x;t)$ is termed the *exponential tilting* of $f_0(x)$, with expectation and variance

$$\frac{d}{d\eta} K(\eta) = \dot{K}(\eta) \qquad \frac{d^2 K(\eta)}{d\eta^2} = \ddot{K}(\eta).$$

  respectively. Note further that if

$$f_X(x;\eta) = h(x) \exp\left\{\eta x - K(\eta)\right\}.$$

  then, for t small enough,

$$M_X(t) = \int e^{tx} h(x) \exp\left\{\eta x - K(\eta)\right\}\, dx = \exp\{-K(\eta)\} \int h(x) \exp\left\{(\eta + t)x\right\}\, dx$$

$$= \exp\{K(\eta + t) - K(\eta)\}.$$

- **The Exponential Dispersion Model:** Consider the model

$$f(x;\theta,\phi) = \exp\left\{ d(x,\phi) + \frac{1}{r(\phi)} \sum_{j=1}^{m} c_j(\theta)T_j(x) - \frac{A(\theta)}{r(\phi)} \right\}$$

  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 $1/r(\phi)$.

  Example : $Binomial(n,\theta)$

$$f_X(x;\theta) = \binom{n}{x} \mathbb{1}_{\{0,1,\ldots,n\}}(x) \exp\left\{ \log\left(\frac{\theta}{1-\theta}\right) x - n\log(1-\theta) \right\}.$$

  Let $Y = X/n$, so that

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

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

$$\mathbb{E}_Y[Y] = \theta = \mu$$

  say, and

$$\text{Var}_Y[Y] = \phi\theta(1-\theta) = \phi V(\mu)$$

  where $V(\mu) = \mu(1-\mu)$ is the *variance function*. Thus the exponential dispersion model allows separate modelling of mean and variance.

4. **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

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

   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 $Z_{n1}, \dots, Z_{nn}$ 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
$$\varphi_X(t) = \{\varphi_Z(t)\}^n$$
for some characteristic function φ_Z.
- **Decomposability :** A probability distribution for rv X is *decomposable* if
$$\varphi_X(t) = \varphi_{X_1}(t)\varphi_{X_2}(t)$$
for two characteristic functions φ_{X_1} and φ_{X_2} so that
$$X = X_1 + X_2$$
where X_1 and X_2 are **independent** rvs with characteristic functions φ_{X_1} and φ_{X_2}.
- **Self-Decomposability :** A probability distribution for rv X is *self-decomposable* if
$$\varphi_X(t) = \{\varphi_{X_1}(t)\}^2$$
for characteristic function φ_{X_1} so that
$$X = X_1 + X_2$$
where X_1 and X_2 are **independent identically distributed** rvs with characteristic function φ_{X_1}.

5. **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 :	$X \mid N = n, \theta \sim Binomial(n, \theta)$	

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 $X \sim Poisson(\lambda\theta)$
$$f_X(x; \theta, \lambda) = \frac{(\lambda\theta)^x e^{-\lambda\theta}}{x!} \qquad x = 0, 1, \dots.$$

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 :	$X\|Y = y \sim Poisson(y)$	

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:

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

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

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

where

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

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 $l = 1, \ldots, L : X_{l1}, \ldots, X_{ln_l}\|M_l = m_l \sim Normal(m_l, 1)$ 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.

(i) **Finite Mixture Models**

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

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

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

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

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

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

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

(ii) **Random Sums**

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

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

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

Then, by the law of iterated expectation,

$$\begin{aligned} M_S(t) = \mathbb{E}_S\left[e^{tS}\right] &= \mathbb{E}_X\left[\mathbb{E}_{S|X}\left[e^{tS}|X\right]\right] \\ &= \mathbb{E}_X\left[\mathbb{E}_{f_{\mathbf{Y}|X}}\left[\exp\left\{t\sum_{i=1}^{X} Y_i\right\}\middle| X\right]\right] \\ &= \mathbb{E}_X\left[\{M_Y(t)\}^X\right] \\ &= G_X(M_Y(t)) \end{aligned}$$

where G_X is the factorial mgf (or pgf) for X defined in a neighbourhood $(1-h, 1+h)$ of 1 for some $h > 0$ as

$$G_X(t) = M_X(\log t) = \mathbb{E}_X[t^X] \qquad t \in (1-h, 1+h).$$

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.

(iii) **Location-Scale Mixtures**

LEVEL 3 :	θ	Fixed parameters
LEVEL 2 :	$M, V \sim f_{M,V}(m, v; \theta)$	
LEVEL 1 :	$Y \vert M = m, V = v \sim f_{Y\vert M,V}(y\vert m, v)$	

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

LEVEL 3 :	θ
LEVEL 2 :	$V \sim f_V(v; \theta)$
LEVEL 1 :	$Y \vert V = v \sim f_{Y\vert V}(y\vert v) \equiv Normal(0, g(v))$

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.