# Exponential Family Distributions

Stephen J. Mildenhall

2020-10-20

# 1 Exponential Family Distributions

## 1.1 Introduction

Part I introduced an exponential family as a set of non-degenerate distributions having a density or probability mass function that factors as

$$f(y;\theta) = c(y)k(\theta)e^{y\theta}.$$

$\theta$ is called the canonical parameter and $c$ and $k$ are non-negative functions. The factorization has symmetric roles for the observation $y$ and parameter $\theta$, reflecting the dual meaning of the density as the probability of an observation and the likelihood of a parameter. Strictly speaking, this form defines a Natural Exponential Family (NEF). The exponential family is more general, as explained in Section 1.6. See [1] for another actuary-friendly introduction to the exponential family.

NEF Nonet refers to nine equivalent ways of defining a NEF. Each definition highlights a different property of exponential families. They are described in the next section.

Throughout Part II, $Y$ denotes random variable in a NEF, with density $f$. $Y$ is used rather than $X$ in deference to the modeling application, where $y$ is an observation or **unit**, and $x$ are covariates. Dependence on a parameter $\theta$ will be denoted $Y_\theta$ and $f(\cdot;\theta)$. The **support** of a function or random variable is the set of points in the domain where it takes a non-zero value. All functions are real valued and defined on a subset of the real numbers. Integrals with no explicit limits are over the whole real line. Terminology and notation for the components of an exponential family is not standardized, so other books and papers may not track perfectly with our usage.

## 1.2 Nine Ways of Defining a Natural Exponential Family

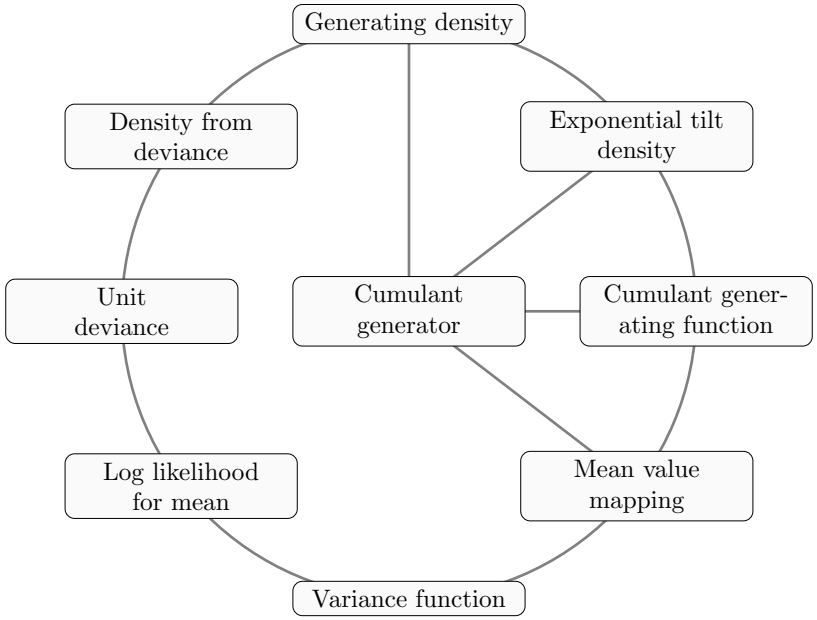


Figure 1: *Figure 1:* Nine different ways of defining a NEF.

### 1.2.1 The Generating Density

A **generator** or **carrier** is a real valued function $c(y) \geq 0$. If $c$ is a probability density, $\int c(y)dy = 1$, then it is called a **generating density**. However, $\int c(y)dy \neq 1$ and even $\int c(y)dy = \infty$ are allowed. The generating density sits at the top of the circle, reflecting its fundamental importance.

How can we create a probability density from $c$? It must be normalized to have integral 1. Normalization is not possible when $\int c(y)dy = \infty$. However, it will be possible to normalize the adjusted generator $c(y)e^{\theta y}$ when its integral is finite. To that end, define

$$\Theta = \{\theta \mid \int c(y)e^{\theta y}dy < \infty\}.$$

The **Natural Exponential Family** generated by $c$ is the **set** of probability densities

$$\text{NEF}(c) = \left\{ \frac{c(y)e^{\theta y}}{\int c(z)e^{\theta z}dz} \mid \theta \in \Theta \right\}$$

proportional to $c(y)e^{\theta y}$. $\theta$ is called the **natural** or **canonical parameter** and $\Theta$ the **natural parameter space**. Naming the normalizing constant $k(\theta) = \left(\int c(y)e^{\theta y}dy\right)^{-1}$ shows all densities in NEF($c$) have a factorization

$$c(y)k(\theta)e^{\theta y},$$

as required by the original definition of an exponential family.

There are two technical requirements for a NEF.

First, a distribution in a NEF cannot be degenerate, i.e., it cannot take only one value. Excluding degenerate distributions ensures that the variance of every member of a NEF is strictly positive, which will be very important.

Second, the natural parameter space $\Theta$ must contain more than one point. If $c(y) = \frac{1}{\pi}\frac{1}{1+y^2}$ is the density of a Cauchy, then $\Theta = \{0\}$ because $c$ has a fat left and right tails. By assumption, the Cauchy density does not generate a NEF.

The same NEF is generated by any element of NEF($c$), although the set $\Theta$ varies with the generator chosen. Therefore, we can assume that $c$ is a probability density. When $c$ is a density $\int c = 1$, $k(0) = \log(1) = 0$ and $0 \in \Theta$.

We will show below that $\Theta$ is an interval. If $c$ is supported on the non-negative reals then $(-\infty, 0) \subset \Theta$. If $\Theta$ is open the the NEF if called **regular**. In general, $\Theta$ might contain an endpoint.

All densities in a NEF have the same support, defined by $\{y \mid c(y) \neq 0\}$ because $e^{\theta y} > 0$ and $k(\theta) > 0$ on $\Theta$.

Many common distributions belong to a NEF, including the normal, Poisson and gamma. The Cauchy distribution does not. The set of uniform distributions on $[0, x]$ as $x$ varies is not a NEF because the elements do not all have the same support.

### 1.2.2 Cumulant Generator

Instead of the generator we can work from the **cumulant generator** or **log partition function**

$$\kappa(\theta) := \log \int e^{\theta y} c(y)\, dy,$$

which is defined for $\theta \in \Theta$. The cumulant generator is the log Laplace transform of $c$ at $-\theta$, and so there is a one-to-one mapping between generators and cumulant generators. The cumulant generator sits in the center of the circle because it is directly linked to several other components. In terms of $\kappa$, a member of NEF($c$) has density

$$c(y)e^{\theta y - \kappa(\theta)}.$$

The cumulant generator is a convex function, and strictly convex if $c$ is not degenerate. Convexity follows from Hölder's inequality. Let $\theta = s\theta_1 + (1-s)\theta_2$. Then

$$\begin{aligned}\int e^{\theta y} c(y) dy &= \int (e^{\theta_1 y})^s (e^{\theta_2 y})^{1-s} c(y) dy \\ &\leq \left(\int e^{\theta_1 y} c(y) dy\right)^s \left(\int e^{\theta_2 y} c(y) dy\right)^{1-s}.\end{aligned}$$

Now take logs. As a result $\Theta$ is an interval. Hölder's inequality is an equality iff $e^{\theta_1 y}$ is proportional to $e^{\theta_2 y}$, which implies $\theta_1 = \theta_2$. Thus provided $\Theta$ is not degenerate $\kappa$ is strictly convex.

### 1.2.3 Exponential Tilting

The **exponential tilt** of $Y$ with parameter $\theta$ is a random variable with density

$$f(y;\theta) = c(y)e^{\theta y - \kappa(\theta)}.$$

It is denoted $Y_\theta$. The exponential tilt is defined for all $\theta \in \Theta$. Tilting, as its name implies, alters the mean and tail thickness of $c$. For example, when $\theta < 0$ multiplying $c(y)$ by $e^{\theta y}$, decreases the probability of positive outcomes, increases that of negative ones, and therefore lowers the mean.

A NEF consists of all valid exponential tilts of a generator density $c$, and all distributions in a NEF family are exponential tilts of one another. They are parameterized by $\theta$. We have seen they all have the same support, since $e^{-\kappa(\theta)} > 0$ for $\theta \in \Theta$. Therefore they are equivalent measures. In finance, equivalent measures are used to model different views of probabilities and to determine no-arbitrage prices.

An exponential tilt is also known as an **Esscher transform**.

**Exercise**: show that all Poisson distributions are exponential tilts of one another as are all normal distributions with standard deviation 1. The tilt directly adjusts the mean. The cumulant generator of a standard normal is $\theta^2/2$ and for a Poisson($\lambda$) it is $\lambda(e^\theta - 1)$.

### 1.2.4 Cumulant Generating Functions

The **moment generating function** (MGF) of a random variable $Y$ is[1]

$$M(t) = \mathsf{E}[e^{tY}] = \int e^{ty} f(y)\, dy.$$

The MGF contains the same information about $Y$ as the distribution function, it is just an alternative representation. Think of distributions and MGFs as the random variable analog of Cartesian and polar coordinates for points in the plane.

The moment generating function owes its name to the fact that

$$\mathsf{E}[Y^n] = \frac{d^n}{dt^n} M_Y(t)\Big|_{t=0},$$

provided $\mathsf{E}[Y^n]$ exists. That is, the derivatives of $M$ evaluated at $t = 0$ give the non-central moments of $Y$. The moment relationship follows by differentiating $\mathsf{E}[e^{tY}] = \sum \mathsf{E}[(tY)^n/n!]$ through the expectation integral.

The MGF of a sum of independent variables is the product of their MGFs

$$\begin{aligned} M_{X+Y}(t) &= \mathsf{E}[e^{t(X+Y)}] \\ &= \mathsf{E}[e^{tX}e^{tY}] \\ &= \mathsf{E}[e^{tX}]\mathsf{E}[e^{tY}] \\ &= M_X(t)M_Y(t). \end{aligned}$$

[1] Strictly, we should use the **characteristic function**, defined by $\phi(s) = \mathsf{E}[e^{isY}]$ where $i = \sqrt{-1}$. The characteristic function exists for all $Y$ and all real $s$ because $|e^{isy}| = 1$, whereas for certain thick tailed $Y$ the MGF does not always exist. However, imaginary numbers can be intimidating and often we can get by with the MGF. The cognoscenti should replace MGF with CF and sprinkle with $i$.

Independence is used to equate the expectation of the product and the product of the expectations. Similarly, the MGF of a sum $X_1 + \cdots + X_n$ of iid variables is $M_X(t)^n$.

The **cumulant generating function** is the log of the MGF, $K(t) = \log M(t)$. The $n$th **cumulant** is defined as

$$\frac{d^n}{dt^n} K(t)\Big|_{t=0}.$$

Cumulants are additive for independent variables because $K_{X+Y} = \log M_{X+Y} = \log(M_X M_Y) = \log(M_X) + \log(M_Y) = K_X + K_Y$. Higher cumulants are translation invariant because $K_{k+X}(t) = kt + K_X(t)$. The first three cumulants are the mean, the variance and the third central moment, but thereafter they differ from both central and non-central moments.

**Exercise**: show that all cumulants of a Poisson distribution equal its mean.

The MGF of the exponential tilt $Y_\theta$ in NEF($c$) is

$$\begin{aligned} M(t;\theta) &= \mathsf{E}[e^{tY_\theta}] \\ &= \int e^{ty}\, c(y) e^{\theta y - \kappa(\theta)}\, dy \\ &= e^{\kappa(\theta+t)-\kappa(\theta)}. \end{aligned}$$

Therefore the cumulant generating function of $Y_\theta$ is simply

$$K(t;\theta) = \kappa(\theta + t) - \kappa(\theta).$$

### 1.2.5 The Mean Value Mapping

The mean of $Y_\theta$ is the first cumulant, computed by differentiating $K(t;\theta)$ with respect to $t$ and setting $t = 0$. The second cumulant, the variance, is the second derivative. Thus

$$\begin{cases} \mathsf{E}[Y_\theta] = K'(0;\theta) = \kappa'(\theta) & \text{and} \\ \mathsf{Var}(Y_\theta) = \kappa''(\theta). \end{cases}$$

The **mean value mapping** (MVM) function is $\tau(\theta) = \kappa'(\theta)$. Since a NEF distribution is non-degenerate, $\tau'(\theta) = \kappa''(\theta) = \mathsf{Var}(Y_\theta) > 0$ showing again that $\kappa$ is convex and that $\tau$ is monotonically increasing and therefore invertible. Thus $\theta = \tau^{-1}(\mu)$ is well defined. The function $\tau^{-1}$ is called the **canonical link** in a GLM. The link function, usually denoted $g$, bridges from the mean domain to the linear modeling domain.

The **mean domain** is $\Omega := \tau(\operatorname{int}\Theta)$, the set of possible means. It is another interval. The NEF is called **regular** if $\Theta$ is open, and then the mean parameterization will return the whole family. But if $\Theta$ contains an endpoint the mean domain may need to be extended to include $\pm\infty$. The family is called **steep** if the mean domain equals the interior of the convex hull of the support. Regular implies steep. A NEF is steep iff $\mathsf{E}[X_\theta] = \infty$ for all $\theta \in \Theta \setminus \operatorname{int}\Theta$.

When we model, the mean is the focus of attention. Using $\tau$ we can parameterize NEF($c$) by the mean, rather than $\theta$, which is usually more convenient.

### 1.2.6 The Variance Function

The variance function determines the relationship between the mean and variance of distributions in a NEF. It sits at the bottom of the circle, befitting its foundational role. In many cases the modeler will have prior knowledge of the form of the variance function. Part I explains how NEFs allow knowledge about $V$ to be incorporated without adding any other assumptions.

Using the MVM we can express the variance of $Y_\theta$ in terms of its mean. Define the **variance function** by

$$\begin{aligned} V(\mu) &= \mathsf{Var}(Y_{\tau^{-1}(\mu)}) \\ &= \tau'(\tau^{-1}(\mu)) \\ &= \frac{1}{(\tau^{-1})'(\mu)}. \end{aligned}$$

The last expression follows from differentiating $\tau(\tau^{-1}(\mu)) = \mu$ with respect to $\mu$ using the chain rule. Integrating, we can recover $\theta$ from $V$

$$\begin{aligned} \theta &= \theta(\mu) \\ &= \tau^{-1}(\mu) \\ &= \int_{\mu_0}^{\mu} (\tau^{-1})'(m) dm \\ &= \int_{\mu_0}^{\mu} \frac{dm}{V(m)}. \end{aligned}$$

We can phrase this relationship as

$$\frac{\partial \theta}{\partial \mu} = \frac{1}{V(\mu)}$$

meaning $\theta$ is a primitive or anti-derivative of $1/V(\mu)$. Similarly,

$$\frac{\partial \kappa}{\partial \mu} = \frac{\kappa'(\theta(\mu))}{V(\mu)} = \frac{\mu}{V(\mu)},$$

meaning $\kappa(\theta(\mu))$ is a primitive of $\mu/V(\mu)$.

$V$ and $\Theta$ uniquely characterize a NEF. It is necessary to specify $\Theta$, for example, to distinguish a gamma family from its negative. $(V, \Theta)$ do not characterize a family within all distributions. For example, the family $kX$ for $X$ with $\mathsf{E}[X] = \mathsf{Var}(X) = 1$ has variance proportional to the square of the mean, for any $X$. But the gamma is the only NEF family of distributions with square variance function.

### 1.2.7 Log Likelihood for the Mean

The NEF density factorization implies the sample mean is a sufficient statistic for $\theta$. The log likelihood for $\theta$ is $l(y; \theta) = \log(c(y)) + y\theta - \kappa(\theta)$. Only the terms of the log likelihood involving $\theta$ are relevant for inference about the mean. The portion $y\theta - \kappa(\theta)$ is often called the **quasi-likelihood**.

Differentiating $l$ respect to $\theta$ and setting equal to zero shows the maximum likelihood estimator (MLE) of $\theta$ given $y$ solves the score equation $y - \kappa'(\theta) = 0$. Given a sample of

independent observations $y_1, \ldots, y_n$, the MLE solves $\bar{y} - \kappa(\theta) = 0$, where $\bar{y}$ is the sample mean.

Using the mean parameterization

$$f(y;\mu) = c(y)e^{y\tau^{-1}(\mu)-\kappa(\tau^{-1}(\mu))}$$

the log likelihood of $\mu$ is

$$l(y;\mu) = \log(c(y)) + y\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu)).$$

Differentiating with respect to $\mu$, shows the maximum likelihood value occurs when

$$\begin{aligned}\frac{\partial l}{\partial \mu} &= \frac{\partial l}{\partial \theta}\frac{\partial \theta}{\partial \mu} \\ &= \{y - \kappa'(\tau^{-1}(\mu))\}\,\frac{1}{V(\mu)} \\ &= \frac{y-\mu}{V(\mu)} \\ &= 0\end{aligned}$$

since $\kappa'(\tau^{-1}(\mu)) = \tau(\tau^{-1}(\mu)) = \mu$. Thus the most likely tilt given $y$ has parameter $\theta$ determined so that $\mathsf{E}[Y_\theta] = y$. Recall, $\partial l/\partial\mu$ is called the **score** function.

In a NEF, the maximum likelihood estimator of the canonical parameter is unbiased. Given a sample from a uniform $[0, x]$ distribution, the maximum likelihood estimator for $x$ is the maximum of a sample, but it is biased low. The uniform family is not a NEF.

### 1.2.8 Unit Deviance

A statistical unit is an observation and a deviance is a measure of fit that generalizes the squared difference. A unit deviance is a measure of fit for a single observation.

Given an observation $y$ and an estimate (fitted value) $\mu$, a **unit deviance** measures of how much we care about the absolute size of the residual $y - \mu$. Deviance is a function $d(y;\mu)$ with the similar properties to $(y-\mu)^2$:

1. $d(y;y) = 0$ and
2. $d(y;\mu) > 0$ if $y \neq \mu$.

If $d$ is twice continuously differentiable in both arguments it is called a **regular** deviance. $d(y;\mu) = |y-\mu|$ is an example of a deviance that is not regular.

We can make a unit deviance from a likelihood function by defining

$$\begin{aligned}d(y;\mu) &= 2(\sup_{\mu\in\Omega} l(y;\mu) - l(y;\mu)) \\ &= 2(l(y;y) - l(y;\mu)),\end{aligned}$$

provided $y \in \Omega$. This is where we want steepness. It is also where we run into problems with Poisson modeling. $y = 0$ is a legitimate outcome but not a legitimate mean value. For a steep family we know the only such values occur on the boundary of the support, generally at 0. The factor 2 is included to match squared differences. $l(y;y)$ is the

likelihood of a saturated model, with one parameter for each observation; it is the best a distribution within the NEF can achieve. $d$ is a relative measure of likelihood, compared to the best achievable. It obviously satisfies the first condition to be a deviance. It satisfies the second because $\tau$ is strictly monotone (again, using the fact NEF distributions are non-degenerate and have positive variance). Finally, the nuisance term $\log(c(y))$ in $l$ disappears in $d$ because it is independent of $\theta$.

We can construct $d$ directly from the variance function. Since

$$\frac{\partial d}{\partial \mu} = -2\frac{\partial l}{\partial \mu} = -2\frac{y-\mu}{V(\mu)}$$

it follows that

$$d(y;\mu) = 2\int_{\mu}^{y} \frac{y-t}{V(t)}\, dt.$$

The limits ensure $d(y;y) = 0$ and that $d$ has the desired partial derivative wrt $\mu$. The deviance is the average of how much we care about the difference between $y$ and the fitted value, between $y$ and $\mu$. The variance function in the denominator allows the degree of care to vary with the fitted value.

**Example.** When $d(y;\mu) = (y-\mu)^2$, $\partial d/\partial \mu = -2(y-\mu)$ and hence $V(\mu) = 1$.

We can make a deviance function from a single-variable function $d^*$ via $d(y;\mu) = d^*(y-\mu)$ provided $d^*(0) = 0$ and $d^*(x) \neq 0$ for $x \neq 0$. $d^*(x) = x^2$ shows square distance has this form. We can then distinguish **scale** vs. **dispersion** or **shape** via

$$d\left(\frac{y-\mu}{\sigma}\right) \quad \text{vs.} \quad \frac{d(y-\mu)}{\sigma^2}.$$

Scale and shape are the same in a normal-square error model. Part III shows they are different for other distributions such as the gamma or inverse Gaussian. Densities with different shape cannot be shifted and scaled to one-another.

### 1.2.9 Density From Deviance

Finally, we can write a NEF density in terms of the deviance rather than as an exponential tilt. This view further draws out connections with the normal. Starting with the tilt density, and parameterizing by the mean, we get

$$\begin{aligned} f(y;\tau^{-1}(\mu)) &= c(y)e^{y\tau^{-1}(\mu)-\kappa(\tau^{-1}(\mu))} \\ &= e^{l(y;\mu)} \\ &= e^{-d(y;\mu)/2+l(y;y)} \\ &= c^*(y)e^{-d(y;\mu)/2} \end{aligned}$$

where $c^*(y) := e^{l(y;y)} = c(y)e^{y\tau^{-1}(y)-\kappa(\tau^{-1}(y))}$.

Although it is easy to compute the deviance from $V$, it is not necessarily easy to compute $c^*$. It can be hard (or impossible) to identify a closed form expression for the density of a NEF member in terms of elementary functions. This will be the case for the Tweedie distribution.

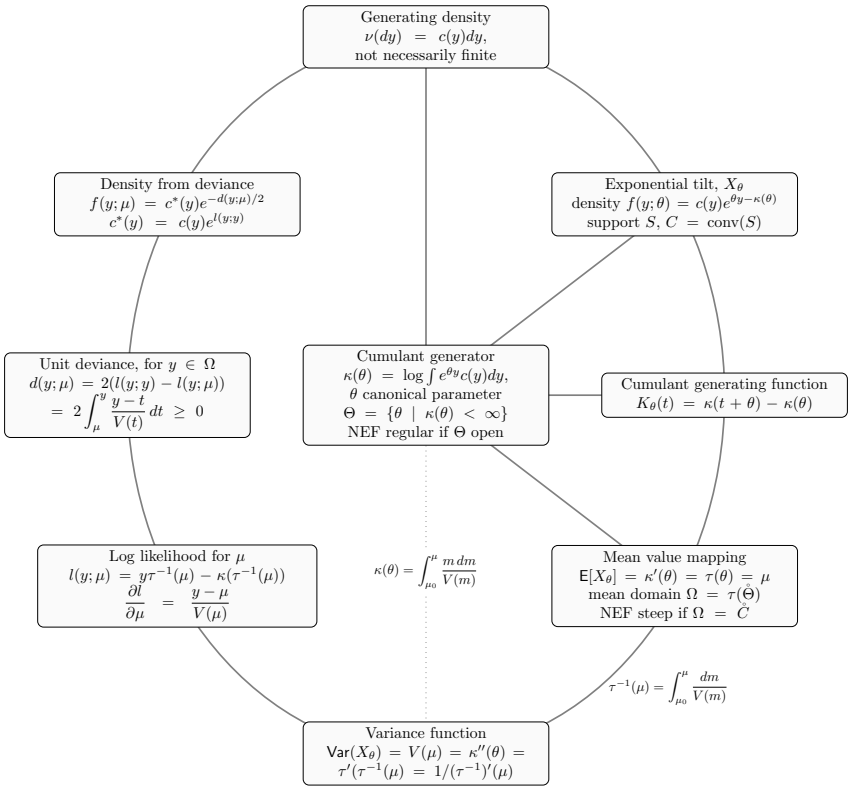


Figure 2: Relationships between actors for a NEF defined by the generating density $c$. If the NEF is steep (e.g., if it is regular) then the condition $y \in \Omega$ in the unit deviance can be dropped. If the NEF is not steep then the $l(y;y)$ may not be defined and must be replaced with $\sup_{\theta\in\Theta} y\theta - \kappa(\theta)$ in the unit deviance. $\text{conv}(S)$ denotes the smallest convex set (i.e., interval) containing $S$. For the Poisson $S = \{0, 1, 2, ...\}$ and $C = [0, \infty)$. $\mathring{\Theta}$ denotes the interior of $\Theta$.

## 1.3 The NEF Circle

We have defined enough terms to make your head spin. Let's use the formulas summarized in Figure 2 to work out the details for four examples: the standard normal, a Poisson, an exponential (gamma), and an inverse Gaussian. The results are shown in the four diagrams below, Figure 3 to Figure 6.

The normal distribution is the archetype, corresponding to squared distance deviance and the constant variance function. Scale and shape coincide. The Poisson, like the normal, has no shape parameter. But because it is defined on the non-negative integers, scaling changes the support and creates in the over-dispersed Poisson model. For the gamma and inverse Gaussian, scale and shape are different operations.

### 1.3.1 NEFs associated with the standard normal, Poisson, exponential (gamma), and standard inverse Gaussian distributions.

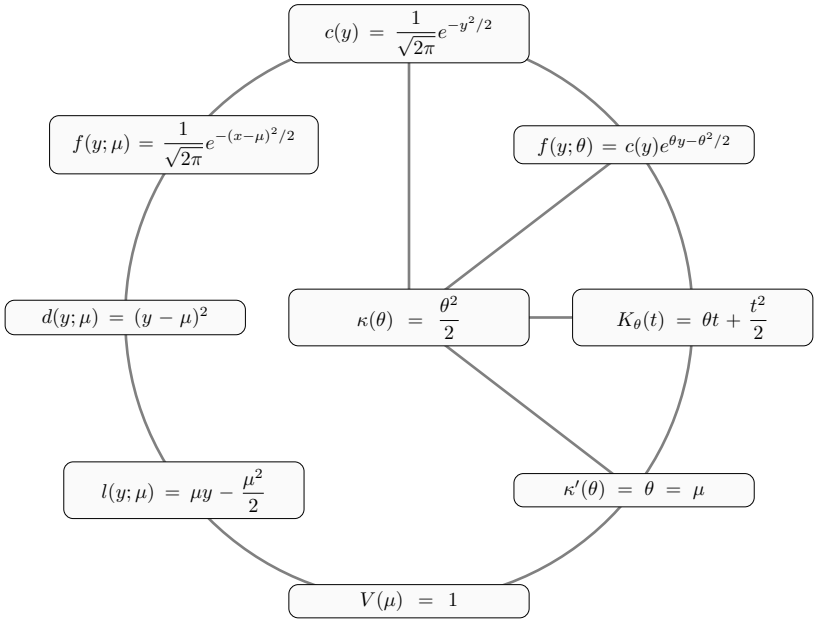


Figure 3: The normal NEF, with $\sigma^2 = 1$.

## 1.4 Completing the NEF Circle

This section presents an algorithm to compute each element of the NEF Circle from a variance function, as well as a less formulaic approach starting with the generator density. The formal approach is described in [2] and [3].

### 1.4.1 Starting From the Variance Function.

The variance function $V$ for a NEF variable $Y$ satisfies

$$\mathsf{Var}(Y) = V(\mathsf{E}[Y]).$$

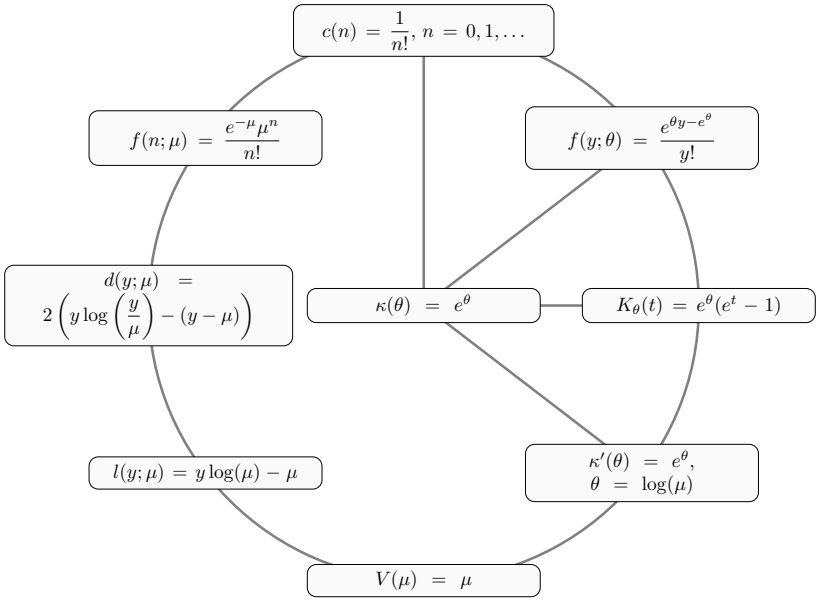


Figure 4: The Poisson NEF. Counting base measure, $\kappa(\theta) = \log \sum_n e^{\theta n}/n!$.

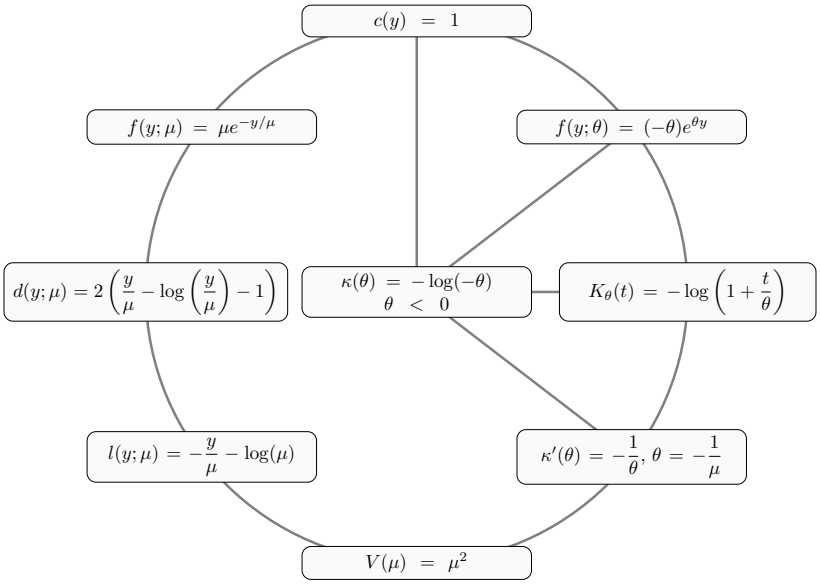


Figure 5: The Gamma NEF with shape parameter 1. $\tau^{-1}(\mu) = -\mu^{-1}$ and $V(\mu) = 1/(\tau^{-1})'(\mu) = \mu^2$. Note $\theta > 0$ for $y < 0$ is another solution.

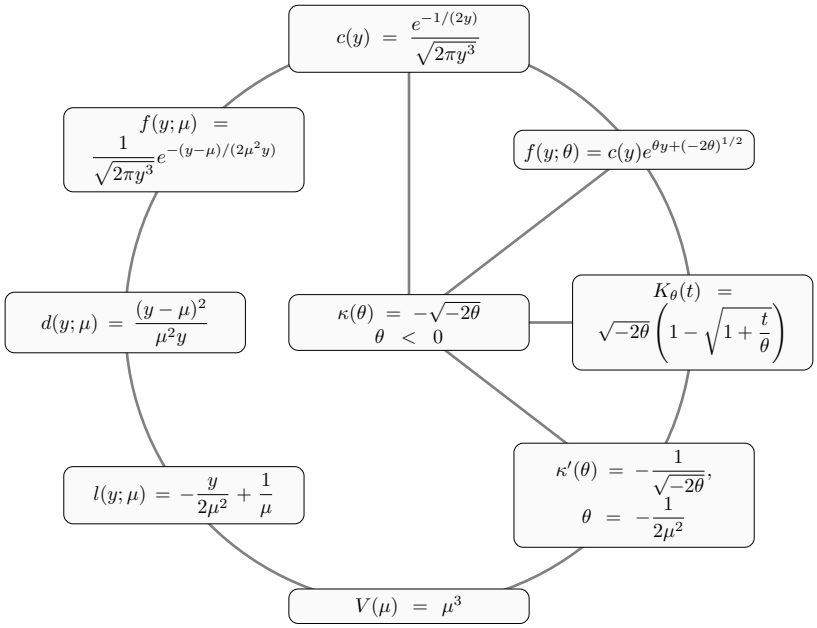


Figure 6: The inverse Gaussian NEF with shape parameter 1. $\tau^{-1}(\mu) = -1/(2\mu^2)$ and $V(\mu) = 1/(\tau^{-1})'(\mu) = \mu^3$. Compute $d$ by integrating $(y-t)/V(t)$.

| **Name** | $V(\mu)$ | $d(y;\mu)$, $\sigma^2 = \lambda = 1$ | $c(y;\lambda)$, $\lambda = 1/\sigma^2$ | $\kappa(\theta)$ | $\tau(\theta)$ | $\Theta$ | $\Omega$ | $S$ |
|---|---|---|---|---|---|---|---|---|
| Gaussian$(\mu,\sigma^2)$ | 1 | $(y-\mu)^2$ | $\frac{1}{\sqrt{2\pi}\sigma}\exp(-y^2/2\sigma^2)$ | $\frac{1}{2}\theta^2$ | $\theta$ | $\mathbf{R}$ | $\mathbf{R}$ | $\mathbf{R}$ |
| Poisson | $\mu$ | $2\,(y\log(y/\mu)-(y-\mu))$ | $\delta_k\lambda^k/k!$ | $e^\theta$ | $e^\theta$ | $\mathbf{R}$ | $\mathbf{R}_{>0}$ | $\{0,1,2,\ldots\}$ |
| Tweedie$(\mu,p)$ | $\mu^p$ | $2\left(\frac{\max(y,0)^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right)$ | | $\frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^\alpha$ | | $\mathbf{R}_{<0}$ | $\mathbf{R}_{>0}$ | $\mathbf{R}_{\geq 0}$ |
| Gamma$(\mu,\sigma^2)$ | $\mu^2/\lambda$ | $2\left(\log\frac{y}{\mu} + \frac{y}{\mu} - 1\right)$ | $\lambda^\lambda y^{\lambda-1}/\Gamma(\lambda)$ | $-\log(-\theta)$ | $-1/\theta$ | $\mathbf{R}_{<0}$ | $\mathbf{R}_{>0}$ | $\mathbf{R}_{>0}$ |
| Positive extreme stable | $\mu^p$, $p > 2$ | as Tweedie | | | | $\mathbf{R}_{\leq 0}$ | $\mathbf{R}_{>0}$ | $\mathbf{R}_{>0}$ |
| Inverse Gaussian | $\mu^3/p^3$ | $\frac{(y-\mu)^2}{\mu^2 y}$ | $\frac{1}{\sqrt{2\pi y^3}}e^{-1/2x}$ | | | $\mathbf{R}_{\leq 0}$ | $\mathbf{R}_{>0}$ | $\mathbf{R}_{>0}$ |
| Extreme stable | $\mu^p$, $p < 0$ | as Tweedie | | | | $\mathbf{R}_{\geq 0}$ | $\mathbf{R}_{>0}$ | $\mathbf{R}$ |
| Extreme Cauchy, $\beta$ | $e^\mu$ | $2\beta^{-2}\left(e^{-\beta y} + e^{-\beta\mu}(\beta y - \beta\mu - 1)\right)$ | | $\theta(1+\log\theta)$ | | $\mathbf{R}_{<0}$ | $\mathbf{R}$ | $\mathbf{R}$ |
| Binomial$(n,p)$, $n$ known | $\mu(1-\frac{\mu}{n})$ | $2\left(y\log\frac{y}{\mu} + (1-y)\log\frac{1-y}{1-\mu}\right)$ | $\lambda k\delta_k$ | $\log(1+e^\theta)$ | $e^\theta/(1+e^\theta)$ | $\mathbf{R}$ | $(0,1)$ | $\{0,1,2,\ldots\}$ |
| Negative binomial$(n,\lambda)$, $n$n known | $\mu(1+\frac{\mu}{p})$ | $2\left(y\log\frac{y}{\mu} + (1+y)\log\frac{1+\mu}{1+y}\right)$ | $p+k-1k\delta_k$ | $-\log(1-e^\theta)$ | $e^\theta/(1-e^\theta)$ | $\mathbf{R}_{<0}$ | $\mathbf{R}_{>0}$ | $\{0,1,2,\ldots\}$ |
| Generalized hyperbolic secant | $p(1+\frac{\mu^2}{p^2})$ | $2y(\operatorname{atan}(y) - \operatorname{atan}(\mu)) + \log\left(\frac{1+\mu^2}{1+y^2}\right)$ | $\frac{\lambda\lvert\Gamma(\lambda(1+iy)/2)\rvert^2}{\pi\Gamma(\lambda)2^{2-\lambda}}$ | $-\log(\cos\theta)$ | $\tan\theta$ | $(-\pi/2,\pi/2)$ | $\mathbf{R}$ | $\mathbf{R}$ |
| Abel, generalized Poisson | $\mu(1+\frac{\mu}{p})^2$ | | $p(p+k)^{k-1}\frac{\delta_k}{k!}$ | | | | $\mathbf{R}_{>0}$ | $\{0,1,2,\ldots\}$ |
| Taskacs, $a>0$ | $\mu(1+\frac{\mu}{p})(1+\frac{a+1}{a}\frac{\mu}{p})$ | | $ap\prod_{j=1}^{k-1}(a(p+k)+j)\,\frac{\delta_k}{k!}$ | | | | $\mathbf{R}_{>0}$ | $\{0,1,2,\ldots\}$ |
| Strict arcsine | $\mu(1+\frac{\mu^2}{p^2})$ | | $\frac{p_n(a)}{n!}$ | | | | $\mathbf{R}_{>0}$ | $\{0,1,2,\ldots\}$ |
| Large arcsine, $a>0$ | $\mu(1+2\frac{\mu}{p}+\frac{1+a^2}{a^2}\frac{\mu^2}{p^2})$ | | $\frac{p}{p+k}p_k(a(p+k))\frac{\delta_k}{k!}$ | | | | $\mathbf{R}_{>0}$ | $\{0,1,2,\ldots\}$ |
| Ressel | $\frac{\mu^2}{p}(1+\frac{\mu}{p})$ | | $\frac{px^{x+p-1}e^{-1}}{\Gamma(x+p+1)}$ | | | | $\mathbf{R}_{>0}$ | $\mathbf{R}_{>0}$ |

Figure 7: Various natural exponential family distributions. In the strict and large arcsine distributions, $p_n(a)$ are defined as the coefficients in $\exp(a\arcsin(z)) = \sum_n \frac{p_n(a)}{n!}z^n$. For the Abel family on down $\kappa$ is defined implicitly.

It is independent of the parameterization used for $Y$ because it only expresses how the variance behaves as a function of the mean. The mean is denoted $\mu$. There is a one-to-one relationship between values of the canonical parameter $\theta$ and values of $\mu$ given by $\tau(\theta) := \kappa'(\theta) = \mu$. Thus we can consider the family parameterized by $\theta$ or $\mu$. To complete the NEF Circle starting from $V$:

1. Integrate $(\tau^{-1})'(\mu) = 1/V(\mu)$ to determine the canonical parameter $\theta = \tau^{-1}(\mu)$ as a primitive of $1/V(\mu)$
$$\begin{aligned}\theta(\mu) &= \tau^{-1}(\mu) \\ &= \int (\tau^{-1})'(\mu)\, d\mu \\ &= \int \frac{d\mu}{V(\mu)}.\end{aligned}$$
2. Rearrange to obtain $\mu = \kappa'(\theta)$ as a function of $\theta$.
3. Integrate $\kappa'(\theta)$ to determine the cumulant generator $\kappa(\theta)$. Change variables $\mu = \kappa'(\theta)$, $d\mu = \kappa''(\theta)d\theta$, to see $\kappa(\theta)$ is a primitive of $\mu/V(\mu)$:
$$\kappa(\theta) = \int \kappa'(\theta)\, d\theta = \int \frac{\mu}{V(\mu)}\, d\mu.$$
4. The cumulant generating function is $K_\theta(t) = \kappa(\theta + t) - \kappa(\theta)$.
5. The deviance can be computed directly as
$$d(y;\mu) = 2\int_\mu^y \frac{y-m}{V(m)} dm = l(y;y) - l(y;\mu).$$
Notice that equally
$$d(y;\mu) = 2\{y\theta(y) - \kappa(\theta(y)) - (y\theta(\mu) - \kappa(\theta(\mu)))\}$$
using the results of Steps 1 and 3. As a result, $l(y;\mu) = y\theta(\mu) - \kappa(\theta(\mu))$ up to irrelevant factors.

This algorithm can always be computed numerically. It can run into problems if the functions in Steps 1 and 3 are not integrable, or in Step 2 if $\theta$ cannot be inverted.

### 1.4.2 Starting from the Density.

A. Starting with the density or probability mass function, find the factorization
$$c(y)e^{\theta y - \kappa(\theta)}.$$
in terms of the original parameterization.

B. Identify $\theta$ as a function of the original parameters.

Working from the density is less algorithmic, but is easier if the form of the density is known. Note that you can then confirm $V(\mu) = \kappa''(\tau^{-1}(\mu))$.

We now present several examples of these techniques.

### 1.4.3 The Binomial Distribution

Let $Y \sim \text{binomial}(n, p)$, with integer $n > 0$ known and $0 < p < 1$ unknown. $\mathsf{E}[Y] = \mu = np$ and $\mathsf{Var}(Y) = npq = \mu(1-p) = \mu(1-\mu/n)$ so the variance function is

$$V(\mu) = \mu\left(1 - \frac{\mu}{n}\right).$$

**Binomial: Starting From the Variance Function.**

1. Integrate $1/V$ (using the partial fraction decomposition) to determine $\theta$

$$\begin{aligned}\theta &= \int \frac{d\mu}{V(\mu)} \\ &= \int \frac{n}{\mu(n-\mu)}\, d\mu \\ &= \int \left(\frac{1}{\mu} + \frac{1}{n-\mu}\right) d\mu \\ &= \log\left(\frac{\mu}{n-\mu}\right) \\ &= \log\left(\frac{p}{1-p}\right).\end{aligned}$$

   Rearranging gives $p = e^\theta/(1+e^\theta)$. This step has identified the canonical parameter.
2. Invert $\theta = \log(\mu/(n-\mu))$ to obtain $\mu = \tau(\theta) = ne^\theta/(1+e^\theta) = np$.
3. Integrate $\tau(\theta)$ to determine the cumulant generator

$$\begin{aligned}\kappa(\theta) &= \int \tau(\theta)\, d\theta \\ &= \int \frac{ne^\theta}{1+e^\theta} d\theta \\ &= n\log(1+e^\theta).\end{aligned}$$

   Alternatively, substituting,

$$\kappa(\theta) = \int_{\mu_0}^{\mu} \frac{\mu}{V(\mu)}\, d\mu$$

   where $\tau(\theta) = \mu$. If $p = 0$ then $\mu = 0$ and the distribution is degenerate with $\kappa = 0$. Therefore, the cumulant generator is given by

$$\begin{aligned}\kappa(\theta) &= \int_0^{\mu} \frac{m}{V(m)} dm \\ &= \int_0^{\mu} \frac{n}{n-m} dm \\ &= -n\log\left(1 - \frac{\mu}{n}\right) \\ &= n\log\left(1+e^\theta\right)\end{aligned}$$

4. The cumulant generating function is

$$\begin{aligned} K_\theta(t) &= \kappa(\theta + t) - \kappa(\theta) \\ &= n \log\left(\frac{1 + e^{\theta+t}}{1 + e^\theta}\right) \\ &= n \log\left((1-p) + pe^t\right). \end{aligned}$$

   Exponentiating yields $(1 - p + pe^t)^n$, the MGF of the binomial.
5. The deviance is

$$\begin{aligned} d(y;\mu) &= 2\int_\mu^y \frac{y-m}{V(m)} dm \\ &= 2\left\{ y \log\left(\frac{y}{\mu}\right) + (1-y)\log\left(\frac{1-y}{1-\mu}\right)\right\} \end{aligned}$$

   and hence $l(y;\mu) = y\log(\mu) + (1-y)\log(1-\mu)$.

**Binomial: Starting from the Density.**

A. Write the probability mass function as

$$\begin{aligned} P(N = y) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= \binom{n}{y} \exp\left\{ y \log\left(\frac{p}{1-p}\right) - n(-\log(1-p))\right\} \end{aligned}$$

B. Identify $c(y) = \binom{n}{y}$, $\theta = \log\left(\frac{p}{1-p}\right)$, which confirms $\kappa(\theta) = n\log(1 + e^\theta)$.

### 1.4.4 The Negative Binomial Distribution

Let $Y \sim$ negative binomial$(n, p)$, with $n$ known and $p$ unknown. $\mathsf{E}[Y] = \mu = np/(1-p)$ and $\mathsf{Var}(Y) = np/(1-p)^2$. Since $\frac{p}{(1-p)^2} = \frac{p}{1-p}(1 + \frac{p}{1-p})$, we can write

$$\begin{aligned} \mathsf{Var}(Y) &= \frac{np}{(1-p)^2} \\ &= n\left(\frac{p}{1-p}\left(1 + \frac{p}{1-p}\right)\right) \\ &= \frac{np}{1-p}\left(1 + \frac{1}{n}\frac{np}{1-p}\right) \\ &= \mu\left(1 + \frac{\mu}{n}\right), \end{aligned}$$

so variance function is

$$V(\mu) = \mu\left(1 + \frac{\mu}{n}\right).$$

**Negative Binomial: Starting From the Variance Function.**

1. Integrate $1/V$ (again, using partial fractions) to determine $\theta$

$$\begin{aligned}\theta &= \int \frac{d\mu}{V(\mu)} \\ &= \int \frac{n}{\mu(n+\mu)}\, d\mu \\ &= \int \frac{1}{\mu} - \frac{1}{n+\mu}\, d\mu \\ &= \log\left(\frac{\mu}{n+\mu}\right) \\ &= \log(p).\end{aligned}$$

   Hence $p = e^\theta$. This step has identified the canonical parameter.

2. Invert $\theta = \log(\mu/(n+\mu))$ to obtain $\mu = \tau(\theta) = \dfrac{ne^\theta}{1-e^\theta}$.

3. Integrate $\tau(\theta)$ to determine the cumulant generator

$$\begin{aligned}\kappa(\theta) &= \int \tau(\theta)\, d\theta \\ &= \int \frac{ne^\theta}{1-e^\theta} d\theta \\ &= -n\log(1-e^\theta).\end{aligned}$$

4. The cumulant generating function is

$$\begin{aligned}K_\theta(t) &= \kappa(\theta+t) - \kappa(\theta) \\ &= -n\log\left(\frac{1+e^{\theta+t}}{1+e^\theta}\right) \\ &= n\log\left(\frac{1-p}{1-pe^t}\right).\end{aligned}$$

   Exponentiating yields the MGF of the negative binomial.

5. The deviance, using the second expression for $\theta$ from part 2 and the cumulant $\kappa(\theta(\mu)) = n\log(n/(n+\mu))$ from part 3, is

$$\begin{aligned}d(y;\mu) &= 2\left\{y\log\left(\frac{y}{n+y}\right) - n\log\left(\frac{n+y}{n}\right) - y\log\left(\frac{\mu}{n+\mu}\right) + n\log\left(\frac{n+\mu}{n}\right)\right\} \\ &= 2\left\{y\log\left(\frac{y}{\mu}\right) - (n+y)\log\left(\frac{n+y}{n+\mu}\right)\right\}\end{aligned}$$

   and

$$l(y;\mu) = y\log\left(\frac{\mu}{n+\mu}\right) + n\log\left(\frac{n}{n+\mu}\right)$$

   up to irrelevant factors.

**Negative Binomial: Starting from the Density.**

A. Factorizing the probability mass function as

$$\begin{aligned}P(N=k) &= \binom{n+k-1}{k} p^k(1-p)^n \\ &= \binom{n+k-1}{k} \exp\left\{k\log(p) - n(-\log(1-p))\right\}\end{aligned}$$

shows $c(k) = \binom{n+k-1}{k}$, $\theta = \log(p) < 0$ and $\kappa(\theta) = -n\log(1-e^{\theta})$.

### 1.4.5 The Gamma Distribution

Let $Y \sim \text{gamma}(\mu, \alpha)$, with known shape parameter $\alpha$. $\mathsf{E}[Y] = \mu$ and $\mathsf{Var}(Y) = \mu^2/\alpha$, so $\alpha$ is $1/CV^2$. The variance function is simply $V(\mu) = \mu^2/\alpha$.

The shape and rate parameters are $\alpha$ and $\beta = \alpha/\mu$. The density is

$$\frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} = \frac{\alpha^{\alpha}}{\mu^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-y\alpha/\mu}$$

**Gamma: Starting From the Variance Function.**

1. Integrate $1/V$ to determine $\theta$
$$\begin{aligned}\theta &= \int \frac{d\mu}{V(\mu)} \\ &= \int \frac{\alpha}{\mu^2}\, d\mu = -\frac{\alpha}{\mu}.\end{aligned}$$
2. Invert, to obtain $\mu = \tau(\theta) = -\dfrac{\alpha}{\theta}$.
3. Integrate $\tau(\theta)$ to determine the cumulant generator
$$\begin{aligned}\kappa(\theta) &= \int \tau(\theta)\, d\theta \\ &= \int -\frac{\alpha}{\theta} d\theta \\ &= -\alpha\log(-\theta).\end{aligned}$$
Beware: if $F(x) = \int f(x)dx$ then $\int f(-x)dx = -\int f(y)dy = -F(y) = -F(-x)$, substituting $y = -x$.
4. The cumulant generating function is
$$\begin{aligned}K_{\theta}(t) &= \kappa(\theta + t) - \kappa(\theta) \\ &= -\alpha\left\{\log(-\theta - t) + \log(-\theta)\right\} \\ &= -\alpha\log\left(1 + \frac{t}{\theta}\right).\end{aligned}$$
Exponentiating yields the MGF of the gamma; note $\theta < 0$.
5. The deviance is
$$d(y;\mu) = \alpha\left\{\frac{y-\mu}{\mu} - \log\frac{y}{\mu}\right\}$$
and $l(y;\mu) = -y/\mu - \log\mu$ up to irrelevant factors.

**Gamma: Starting from the Density.**

A. Factorizing the probability mass function as
$$\begin{aligned}\frac{y^{\alpha-1}}{\Gamma(\alpha)}\left(\frac{\alpha}{\mu}\right)^{\alpha} e^{-y\alpha/\mu} &= \frac{y^{\alpha-1}}{\Gamma(\alpha)} \exp\left(-y\frac{\alpha}{\mu} + \alpha\log\frac{\alpha}{\mu}\right) \\ &= \frac{y^{\alpha-1}}{\Gamma(\alpha)} \exp\left(y\theta - (-\alpha\log(-\theta))\right)\end{aligned}$$
where $\theta = -\alpha/\mu$ is the canonical parameter and $\kappa(\theta) = -\alpha\log(-\theta)$

### 1.4.6 The Generalized Hyperbolic Secant (GHS) Distribution.

Let's start with a possible variance function

$$V(\mu) = 1 + \mu^2.$$

Since $V(0) = 1$, the support must include positive and negative values. $V$ is valid for all $\mu \in \mathbf{R}$.

**GHS: Starting From the Variance Function.**

1. Integrate $1/V$ to determine $\theta$

$$\begin{aligned}\theta &= \int \frac{d\mu}{V(\mu)} \\ &= \int \frac{1}{1+\mu^2}\, d\mu = \arctan(\mu)\end{aligned}$$

   for $\mu \in \mathbf{R}$ and $\theta \in (-\pi/2, \pi/2)$.
2. Invert, to obtain $\mu = \tau(\theta) = \tan\theta$.
3. Integrate $\tau(\theta)$ to determine the cumulant generator

$$\begin{aligned}\kappa(\theta) &= \int \tau(\theta)\, d\theta \\ &= \int \tan\theta\, d\theta \\ &= -\log(\cos\theta).\end{aligned}$$

   The corresponding MGF for the carrier density is $\sec\theta$, which is the MGF for the hyperbolic secant distribution. It has probability density function given by $\frac{1}{2}\text{sech}(\pi y/2)$ (it has characteristic function and density sech, [4], p.503. Note $\cos(iz) = \cosh(z)$.) The general tilted density is therefore

$$\frac{1}{2}\text{sech}\left(\frac{\pi y}{2}\right) e^{\theta y + \log(\cos\theta)} = \frac{1}{2}\text{sech}\left(\frac{\pi y}{2}\right)\cos(\theta) e^{\theta y}.$$

4. The cumulant generating function is

$$\begin{aligned}K_\theta(t) &= \kappa(\theta + t) - \kappa(\theta) \\ &= -\log(\cos t - \tan\theta \sin t).\end{aligned}$$

5. The deviance is

$$\begin{aligned}d(y;\mu) &= 2\left\{y\theta(y) - \kappa(\theta(y)) - (y\theta(\mu) - \kappa(\theta(\mu)))\right\} \\ &= 2y(\arctan(y) - \arctan(\mu)) - \log\left(\frac{1+y^2}{1+\mu^2}\right)\end{aligned}$$

   since $\theta(\mu) = \arctan(\mu)$, $\cos\arctan(\mu) = 1/\sqrt{1+\mu^2}$ using high school trigonometry, and therefore

$$\kappa(\theta(\mu)) = -\log\left(\frac{1}{\sqrt{1+\mu^2}}\right).$$

   Notice how the 2 disappears with the square root.

The density is

$$f(y;\theta) = \frac{\exp(\theta y + \log(\cos\theta))}{2\cosh(\pi y/2)}$$

---

---

## 1.5 Appendix: Likelihood, Scores, and the Cramer-Rao Minimum Variance Bounds

Consider the likelihood $L$ and log likelihood $l$ (a monotone transformation). If $l$ as a function of $\mu$ is very peaked around its MLE then an observation contains a lot of information about the parameter, else not. At the MLE the derivatives $\partial L/\partial\mu$ and $\partial l/\partial\mu$ are both zero. The function is peaked (high curvature) if the second derivative is large. At the maximum the second derivative will be negative.

Let $s(x;\mu) = \partial l/\partial\mu$ be the **score** function and $f' = \partial f/\partial\mu$. If $\mu$ is the true parameter, then the statistic $s(X) = s(X;\mu)$ has mean zero

$$\begin{aligned}
\mathsf{E}[s(X)] &= \int sf = \int \frac{\partial l}{\partial\mu} f \\
&= \int \frac{f'}{f} f \\
&= \int \frac{\partial f}{\partial\mu} \\
&= \frac{\partial}{\partial\mu} \int f = 0,
\end{aligned}$$

pulling the differential through the integral in the last step. Hence $\mathsf{Var}(s) = \mathsf{E}[s^2]$. Differentiating $\mathsf{E}[s] = 0$ wrt $\mu$ gives

$$\begin{aligned}
0 &= \frac{\partial}{\partial\mu}\mathsf{E}[s] \\
&= \int \frac{\partial}{\partial\mu}(sf) \\
&= \int \frac{\partial s}{\partial\mu} f + \int s\frac{\partial f}{\partial\mu} \\
&= \int \frac{\partial^2 l}{\partial\mu^2} f + \int s\frac{\partial f}{\partial\mu}\frac{f}{f} \\
&= \mathsf{E}\left[\frac{\partial^2 l}{\partial\mu^2}\right] + \mathsf{E}[s^2]
\end{aligned}$$

and so we get two expressions for the **Fisher information** $\mathcal{I}(\mu) := \mathsf{Var}(s)$

$$\begin{aligned}
\mathcal{I}(\mu) &= \mathsf{E}\left[\left(\frac{\partial l}{\partial\mu}\right)^2\right] \\
&= -\mathsf{E}\left[\frac{\partial^2 l}{\partial\mu^2}\right]
\end{aligned}$$

If the information is large then a small deviation between $x$ and $\mu$ leads to a big change in the score and so the likelihood has a high slope, meaning the likelihood function is very peaked around $\mu$. As a result $\mu$ is easy to estimate from the data, because it is clustered around $\mu$. Hence the name information.

If $X$ is normal$(\mu, \sigma^2)$, $\sigma^2$ known, then $s(x;\mu) = (x-\mu)/\sigma^2$ and $\mathsf{Var}(s) = 1/\sigma^2$. Therefore large $\sigma$ generates small information. It is harder to infer the mean from a population with a larger variance.

The **Cramer-Rao bound** says that when $r(X)$ is an unbiased estimator for $\mu$ then $\mathsf{Var}(r(X)) \geq 1/\mathcal{J}(\mu)$. Unbiased means $\mathsf{E}[r(X)] = \mu$, so $\int (r-\mu)f = 0$. Differentiating wrt $\mu$, remember $r$ is a function of the data and does not depend on $\mu$, swap differential and integral, use product rule, and the $\frac{\partial f}{\partial \mu} = \frac{\partial f}{\partial \mu}\frac{f}{f} = \frac{\partial l}{\partial \mu} f$ trick to get

$$0 = \int (r-\mu)\frac{\partial l}{\partial \mu} f - \int f.$$

Applying the Cauchy-Schwarz inequality

$$\begin{aligned} 1 &= \int (r-\mu)\frac{\partial l}{\partial \mu} f \\ &= \int (r-\mu)\sqrt{f}\frac{\partial l}{\partial \mu}\sqrt{f} \\ &\leq \int (r-\mu)^2 f \cdot \int \left(\frac{\partial l}{\partial \mu}\right)^2 f \\ &= \mathsf{Var}(r)\mathcal{J}(\mu) \end{aligned}$$

yields the famous **Cramer-Rao minimum variance bound** (MVB) for unbiased estimators

$$\mathsf{Var}(r) \geq \frac{1}{\mathcal{J}(\mu)}.$$

The greater the information, the tighter it is possible to estimate $\mu$.

When is the MVB attained? The Cauchy-Schwarz inequality is an equality iff the two terms are proportional, which translates into

$$(r-\mu) \propto \frac{\partial l}{\partial \mu}.$$

The constant of proportionality varies with $\mu$, so we can write

$$\frac{\partial l}{\partial \mu} = A(\mu)(r-\mu)$$

for $A$ independent of the observations. Multiply by $r-\mu$ and take expectations. Using the first equality in the Cauchy-Schwarz derivation, gives

$$1 = A(\mu)\mathsf{Var}(r)$$

and therefore the MVB is attained iff

$$\frac{\partial l}{\partial \mu} = \frac{r-\mu}{\mathsf{Var}(r)},$$

i.e., precisely when the distribution is exponential family! For a given mean-variance relationship the exponential family has **minimum information.** Hence a one-parameter exponential family makes fewest additional assumptions beyond the mean-variance relationship, [5, end of section 4.].

See [6] Chapter 17 for more on the topics covered in this appendix.

## 1.6 Appendix: General Exponential Family Distributions

An **exponential family** (EF) distribution generalizes a NEF in three ways. Rather than a density factored as $c(y)e^{y\theta-\kappa(\theta)}$, an EF requires the density factors as

$$f(y;\theta) = c(y)e^{\eta(\theta)\cdot T(y)-\kappa(\theta)}$$

where

1. The canonical parameter is replaced by a vector of parameters, still denoted $\theta$,
2. $\eta(\theta)$ is a new vector-valued **natural parameter**, and
3. $y$ is replaced by a vector-valued function $T(y)$ of **sufficient statistics**.

The two vectors $\eta(\theta)$ and $T(y)$ must have the same dimension, to compute the dot product. When $\eta(\theta) = \theta$ and $T(y) = y$ the exponential family is in **canonical form** and when $\theta$ is a scalar it becomes a NEF.

**Example**. EF distributions can be used to model the mean and variance of a normal distribution with $T(y) = (y, y^2)$.

The separation and symmetry between the roles of $y$ and $\theta$ ensures that $T(y)$ is a sufficient statistic for $\eta$[2].

The Pitman-Koopman-Darmois theorem says that if a parametric class of distributions whose domain does not depend on the parameter has a set of minimal sufficient statistic whose number does not depend on sample size then the distribution belongs to an exponential family.

The uniform distribution on $[0, \theta]$ with parameter $\theta > 0$ has a single sufficient statistics, $\max X_i$, but it does not belong the exponential family because the domain depends on the parameter $\theta$. Thus, the Pitman-Koopman-Darmois theorem does not apply to the uniform distribution.

**Example.** If $x_i$ are sampled from a normal with mean $\theta$ and known unit variance, then, up to irrelevant terms and factors, the log likelihood for $\theta$ is

$$-\frac{1}{2}\sum_i (x_i - \theta)^2 = -n\frac{\theta^2}{2} + \theta\sum_i x_i - \frac{1}{2}\sum_i x_i^2.$$

The density of the sample factors as $f(x_i;\theta) = g(\theta, \sum_i x_i)h(x_i)$, showing that $\sum_i x_i$ is a sufficient statistic. Knowing $\sum_i x_i$ is sufficient to know the maximum likelihood estimate of $\theta$, even though the value of the likelihood is unknown, since it depends on $\sum_i x_i^2$. The sum of quadratics is another quadratic, and in particular, it is concave.

[2]The Fisher-Neyman factorization theorem says that $T(x)$ is a sufficient statistic for the parameter $\theta$ if and only if $f$ factors as a function of $x$ alone times a function of $\theta$ and $T(x)$.

**Example.** If $x_i$ are sampled from a NEF with canonical parameter $\theta$ and cumulant generator $\kappa$, then, up to irrelevant terms and factors, the log likelihood for $\theta$ is

$$\sum_i (x_i\theta - \kappa(\theta)) = -n\kappa(\theta) + \theta \sum_i x_i.$$

Again, $\sum_i x_i$ is a sufficient statistic. Since $\kappa$ is convex, the log likelihood is concave and has a unique maximum. The pattern is identical to the normal, where $\kappa(\theta) = \theta^2/2$.

**Example.** The location parameter of a Cauchy distribution has no sufficient statistic of dimension smaller than the data. It is also known not to be in an exponential family because its density

$$f(x;\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$$

cannot be factored to separate $x$ and $\theta$ as required. The likelihood of $\theta$ is

$$-\sum_i \log(1+(x_i-\theta)^2).$$

Plotting this as a function of $\theta$ reveals a nasty, non-concave, multi-modal curve—it contains a lot of information that cannot be summarized in a sufficient statistic. In contrast, all normal log likelihood plots look essentially the same. See Figure 8.

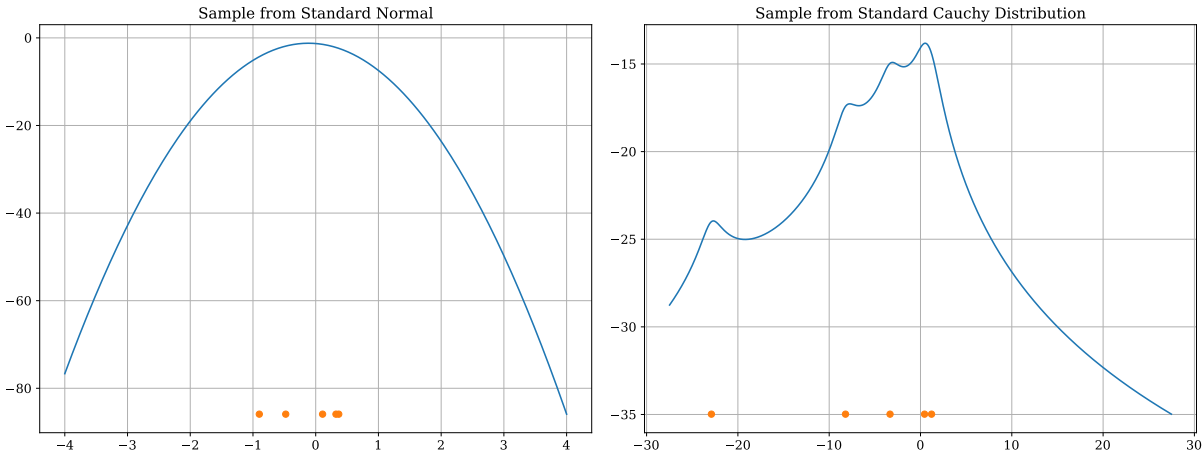


Figure 8: Location likelihood functions (blue) derived from five samples (orange) drawn from a standard normal (left) and standard Cauchy (right). The normal likelihood is determined, up to vertical translation, by $\sum_i x_i$. It is concave and has a unique maximum. For the Cauchy distribution, the likelihood is a complex curve depending on the particular samples drawn. It cannot be summarized. It is not concave and has multiple local maximums. Note different $x$ axis scales.

One last fact that we will not pursue: all exponential family distributions solve a maximum entropy problem. The exponential family distribution has the greatest entropy of any distribution with given values of $T(x)$, when the underlying distribution of $x$ is given by $c$.

See [7] for a formal treatment of the general theory of exponential distributions and [8] for a less formal one.

## References


1. Clark, D.R., Thayer, C.A.: A primer on the exponential family of distributions. Casualty Actuarial Society Spring Forum. 117–148 (2004)
2. Jørgensen, B.: The theory of dispersion models. CRC Press (1997)
3. Letac, G.: Associated natural exponential families and elliptic functions. In: The fascination of probability, statistics and their applications: In honour of ole e. Barndorff-nielsen. pp. 53–83 (2015)
4. Feller, W.: An Introduction to Probability Theory and its Applications, Volume 2. J. Wiley; Sons (1971)
5. Wedderburn, R.W.M.: Quasi likelihood functions, generalized linear models, and the Gauss Newton method. Biometrika. 61, 439–447 (1974). https://doi.org/10.2307/2334725
6. Stuart, A., Ord, J.K., Arnold, S.F.: Kendall's Advanced Theory of Statistics: Classical Inference and the Linear Model. Volume 2A. Arnold; Oxford University Press (1999)
7. Brown, L.D.: Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. Institute of Mathematical Statistics, Hayward CA (1986)
8. Efron, B.: Exponential Families in Theory and Practice. Stanford University (2018)