1

What Is an Exponential Family?

We start with a couple of simple, well-known distributions to introduce the common features of the distribution families called *exponential families* or distributions of *exponential type*. First, consider a sample from an *exponential* distribution with intensity parameter λ (expected value $1/\lambda$). A single observation y has the density $\lambda \exp(-\lambda y)$ for $y > 0$, and a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ has the n-dimensional density

$$f(\boldsymbol{y}; \lambda) = \lambda^n e^{-\lambda \sum y_i} = \lambda^n \exp(-\lambda \sum y_i), \tag{1.1}$$

for $\boldsymbol{y}$ with all $y_i > 0$. Another basic example is the density for a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ from a two-parameter *normal* (or *Gaussian*) distribution, $\mathrm{N}(\mu, \sigma^2)$. It can be written

$$\begin{aligned} f(\boldsymbol{y}; \mu, \sigma^2) &= (\sigma\sqrt{2\pi})^{-n} e^{-\frac{\sum (y_i - \mu)^2}{2\sigma^2}} \\ &= (\sigma\sqrt{2\pi})^{-n} \exp\left(\tfrac{-1}{2\sigma^2} \sum y_i^2 + \tfrac{\mu}{\sigma^2} \sum y_i - \tfrac{n\mu^2}{2\sigma^2} \right), \end{aligned} \tag{1.2}$$

where all dependence on data is found in the two sum type functions of data in the exponent, with a parameter in front of each of them.

As discrete distribution examples, we consider the *binomial* and *Poisson* distribution families. Here is first the binomial probability for y successes in n Bernoulli trials, with success probability π_0 as parameter (Greek letter π is preferred to Roman p for a parameter; π_0 is here used to distinguish from the mathematical constant π)

$$f(y; \pi_0) = \binom{n}{y} \pi_0^y (1 - \pi_0)^{n-y} = \binom{n}{y} (1 - \pi_0)^n e^{y \log \frac{\pi_0}{1-\pi_0}}, \tag{1.3}$$

for $y = 0, 1, \ldots, n$. For a single observation (count) y from a Poisson distribution we analogously write the probability

$$f(y; \lambda) = \frac{\lambda^y}{y!} e^{-\lambda} = \frac{1}{y!} e^{-\lambda} e^{y \log \lambda}, \tag{1.4}$$

for $y = 0, 1, \ldots$.

In all four cases we could express the density or probability function as a product of two or three factors, namely one factor depending only on the parameter, another depending only on the data (in some cases not needed), and one factor of exponential form connecting the data with the parameter. The exponent is a product of a (function of the) parameter and a function of data, or more generally a sum of such products.

The second factor was not needed in the normal distribution but was present in the exponential distribution, albeit implicitly, in the form of the characteristic function for the positive real axis (or the product of such functions for a sample).

Many statistical models have these features in common, and are then characterized as being *exponential families*. Generally, let data $\boldsymbol{y}$ be modelled by a continuous or discrete distribution family, with probability density on $\mathbb{R}^n$ or probability (mass) function on $\mathbb{Z}^n$, or on subsets of $\mathbb{R}^n$ or $\mathbb{Z}^n$. The density or probability function (density with respect to the counting measure, or similarly) will be written $f(\boldsymbol{y};\, \boldsymbol{\theta})$ in a parameterization called canonical, where $\boldsymbol{\theta}$ belongs to a k-dimensional parameter space $\boldsymbol{\Theta}$. A *statistic* $\boldsymbol{t}$ is any (measurable) scalar or vector-valued function $\boldsymbol{t}(\boldsymbol{y})$ of data, for example $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{y}) = (\sum y_i,\ \sum y_i^2)$ in the Gaussian example (1.2).

Definition 1.1 ***Exponential family***
A parametric statistical model for a data set $\boldsymbol{y}$ is an *exponential family* (or is of exponential type), with *canonical parameter* vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ and *canonical statistic* $\boldsymbol{t}(\boldsymbol{y}) = (t_1(\boldsymbol{y}), \ldots, t_k(\boldsymbol{y}))$, if f has the structure

$$f(\boldsymbol{y};\, \boldsymbol{\theta}) = a(\boldsymbol{\theta})\, h(\boldsymbol{y})\, e^{\boldsymbol{\theta}^T \boldsymbol{t}(\boldsymbol{y})}, \tag{1.5}$$

where $\boldsymbol{\theta}^T \boldsymbol{t}$ is the scalar product of the k-dimensional parameter vector and a k-dimensional statistic $\boldsymbol{t}$, that is,

$$\boldsymbol{\theta}^T \boldsymbol{t} = \sum_{j=1}^{k} \theta_j\, t_j(\boldsymbol{y}),$$

and a and h are two functions, of which h should (of course) be measurable.

It follows immediately that $1/a(\boldsymbol{\theta})$, to be denoted $C(\boldsymbol{\theta})$, can be interpreted as a normalizing constant, that makes the density integrate to 1,

$$C(\boldsymbol{\theta}) = \int h(\boldsymbol{y})\, e^{\boldsymbol{\theta}^T \boldsymbol{t}(\boldsymbol{y})}\, \mathrm{d}\boldsymbol{y}, \tag{1.6}$$

or the analogous sum over all possible outcomes in the discrete case. Of course $C(\boldsymbol{\theta})$ or $a(\boldsymbol{\theta})$ are well-defined only up to a constant factor, which can be borrowed from or lent to $h(\boldsymbol{y})$.

In some literature, mostly older, the canonical parameterization is called the *natural* parameterization. This is not a good term, however, because the canonical parameters are not necessarily the intuitively natural ones, see for example the Gaussian distribution above.

We think of the vector $\boldsymbol{t}$ and parameter space $\boldsymbol{\Theta}$ as in effect k-dimensional (not $< k$). This demand will later be shown to imply that $\boldsymbol{t}$ is minimal sufficient for $\boldsymbol{\theta}$. That $\boldsymbol{t}$ is really k-dimensional means that none of its components t_j can be written as a linear expression in the others. Unless otherwise explicitly told, $\boldsymbol{\Theta}$ is taken to be maximal, that is, comprising all $\boldsymbol{\theta}$ for which the integral (1.6) or the corresponding sum is finite. This maximal parameter space $\boldsymbol{\Theta}$ is called the *canonical parameter space*. In Section 3.1 we will be more precise about regularity conditions.

Before we go to many more examples in Chapter 2, we look at some simple consequences of the definition. Consider first a *sample*, that is, a set of independent and identically distributed (iid) observations from a distribution of exponential type.

Proposition 1.2 ***Preservation under repeated sampling***
If $\boldsymbol{y} = (y_1, \ldots, y_n)$ *is a sample from an exponential family, with distribution*

$$f(y_i;\, \boldsymbol{\theta}) = a(\boldsymbol{\theta})\, h(y_i)\, e^{\boldsymbol{\theta}^T \boldsymbol{t}(y_i)},$$

then the sample $\boldsymbol{y}$ *follows an exponential family with the same canonical parameter space* $\boldsymbol{\Theta}$ *and with the sum* $\sum \boldsymbol{t}(y_i)$ *as canonical statistic,*

$$f(\boldsymbol{y};\, \boldsymbol{\theta}) = a(\boldsymbol{\theta})^n\, e^{\boldsymbol{\theta}^T \sum \boldsymbol{t}(y_i)} \prod_i h(y_i)\,. \tag{1.7}$$

Proof Formula (1.7) follows immediately from $f(\boldsymbol{y};\, \boldsymbol{\theta}) = \prod_i f(y_i;\, \boldsymbol{\theta})$. □

Exponential families are preserved not only under repeated sampling from one and the same distribution. It is sufficient in Proposition 1.2 that the observations are independent and have the same canonical parameter. Two important examples are log-linear models and Gaussian linear models, discussed in Chapter 2 (Examples 2.5 and 2.9).

The canonical statistic $\boldsymbol{t}$ itself necessarily also has a distribution of exponential type:

Proposition 1.3 ***Distribution for t***
If $\boldsymbol{y}$ *has a distribution of exponential type, as given by (1.5), then the (marginal) distribution of* $\boldsymbol{t}$ *is also of exponential type. Under certain regularity conditions on the function* $\boldsymbol{t}(\boldsymbol{y})$*, the distribution of* $\boldsymbol{t}$ *also has a density or probability function, which can then be written*

$$f(\boldsymbol{t};\, \boldsymbol{\theta}) = a(\boldsymbol{\theta})\, g(\boldsymbol{t})\, e^{\boldsymbol{\theta}^T \boldsymbol{t}}\,, \tag{1.8}$$

where the structure function g *in the discrete case can be expressed as*

$$g(\boldsymbol{t}) = \sum_{\boldsymbol{t}(\boldsymbol{y})=\boldsymbol{t}} h(\boldsymbol{y})\,, \tag{1.9}$$

or an analogous integral in the continuous case, written

$$g(\boldsymbol{t}) = \int_{\boldsymbol{t}(\boldsymbol{y})=\boldsymbol{t}} h(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\,. \tag{1.10}$$

Proof In the discrete case, the probability function for $\boldsymbol{t}$ follows immediately by summation over the possible outcomes of $\boldsymbol{y}$ for given $\boldsymbol{t}(\boldsymbol{y}) = \boldsymbol{t}$. (Reader, check this!) In the continuous case, the integral representation of $g(\boldsymbol{t})$ is obtained by a limiting procedure starting from a narrow interval of width $d\boldsymbol{t}$ in $\boldsymbol{t}$, corresponding to a thin shell in $\boldsymbol{y}$-space. See also Section 6.1 for a general formula. □

Example 1.1 ***The structure function for repeated Bernoulli trials***
If the sequence $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the realization of n Bernoulli trials, with common success probability π_0, with $y_i = 1$ representing success, the probability function for the sequence $\boldsymbol{y}$ represents an exponential family,

$$f(\boldsymbol{y};\, \pi_0) = \pi_0^t\,(1-\pi_0)^{n-t} = (1-\pi_0)^n\, e^{\,t\,\log\frac{\pi_0}{1-\pi_0}}\,, \tag{1.11}$$

where $t = t(\boldsymbol{y}) = \sum y_i$ is the number of ones. The structure function $g(t)$ is found by summing over all the equally probable outcome sequences having t ones and $n-t$ zeros. The well-known number of such sequences is $g(t) = \binom{n}{t}$, cf. the binomial example (1.3) above. The distribution for the statistic t, induced by the Bernoulli distribution, is the binomial, $\mathrm{Bin}(n;\, \pi_0)$. △

The distribution for $\boldsymbol{t}$ in the Gaussian example requires more difficult calculations, involving n-dimensional geometry and left aside here.

The conditional density for data $\boldsymbol{y}$, given the statistic $\boldsymbol{t} = t(\boldsymbol{y})$, is obtained by dividing $f(\boldsymbol{y};\boldsymbol{\theta})$ by the marginal density $f(\boldsymbol{t};\boldsymbol{\theta})$. We see from (1.5) and (1.8) that the parameter $\boldsymbol{\theta}$ cancels, so $f(\boldsymbol{y}|\,\boldsymbol{t})$ is free from $\boldsymbol{\theta}$. This is the general definition of $\boldsymbol{t}$ being a *sufficient statistic* for $\boldsymbol{\theta}$, with the interpretation that there is no information about $\boldsymbol{\theta}$ in primary data $\boldsymbol{y}$ that is not already in the statistic $\boldsymbol{t}$. This is formalized in the *Sufficiency Principle* of statistical inference: Provided we trust the model for data, all possible outcomes $\boldsymbol{y}$ with the same value of a sufficient statistic $\boldsymbol{t}$ must lead to the same conclusions about $\boldsymbol{\theta}$.

A sufficient statistic should not be of unnecessarily high dimension, so the reduction of data to a sufficient statistic should aim at a *minimal sufficient* statistic. Typically, the canonical statistic is minimal sufficient, see

Proposition 3.3, where the mild additional regularity condition for this is specified.

In statistical modelling we can go a reverse way, stressed in Chapter 6. We reduce the data set $\boldsymbol{y}$ to a small-dimensional statistic $\boldsymbol{t}(\boldsymbol{y})$ that will take the role of canonical statistic in an exponential family, and thus is all we need to know from data for the inference about the parameter $\boldsymbol{\theta}$.

The corresponding parameter-free distribution for $\boldsymbol{y}$ given $\boldsymbol{t}$ is used to check the model. Is the observed $\boldsymbol{y}$ a plausible outcome in this conditional distribution, or at least with respect to some aspect of it? An example is checking a normal linear model by use of studentized residuals (i.e. variance-normalized residuals), e.g. checking for constant variance, for absence of auto-correlation and time trend, for lack of curvature in the dependence of a linear regressor, or for underlying normality.

The statistical inference in this text about the parameter $\boldsymbol{\theta}$ is frequentistic in character, more precisely meaning that the inference is primarily based on the principle of repeated sampling, involving *sampling distributions* of parameter estimators (typically *maximum likelihood*), hypothesis testing via *p-values*, and confidence regions for parameters with prescribed *degree of confidence*. Appendix A contains a summary of inferential concepts and principles, intended to indicate what is a good background knowledge about frequentistic statistical inference for the present text.

Exercise 1.1 ***Scale factor in*** $h(y)$
Suppose $h(y)$ is changed by a constant factor c, to $c\,h(y)$. What effect does this have on the other constituents of the exponential family? △

Exercise 1.2 ***Structure function for Poisson and exponential samples***
Calculate these structure functions by utilizing well-known distributions for t, and characterize the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{t}$:
(a) Sample of size n from the Poisson $\mathrm{Po}(\lambda)$. Use the reproducibility property for the Poisson, that $\sum y_i$ is distributed as $\mathrm{Po}(\sum \lambda_i)$.
(b) Sample of size n from the exponential with intensity λ. Use the fact that $t = \sum y_i$ is gamma distributed, with density

$$f(t;\,\lambda) = \frac{\lambda^n\, t^{n-1}}{\Gamma(n)}\, e^{-\lambda t}\,,$$

and $\Gamma(n) = (n-1)!$ (Section B.2.2). See also Example 2.7 below.
(c) Note that the conditional density for $\boldsymbol{y}$ is constant on some set, Y_t say. Characterize Y_t for the Poisson and the exponential by specifying its form and its volume or cardinality (number of points). △