Keener, Theoretical Statistics
Chapter 2: Exponential Families

Junhyeok Kil

July 1, 2025

2. Exponential Families

The big picture:

In statistics, we often don't know the exact distribution of the data.

2. Exponential Families

The big picture:

In statistics, we often don't know the exact distribution of the data.

$\Rightarrow$ Assume the distribution belongs to a known family of distributions.

Many common distributions are members of exponential families. They have a regular, unified structure that makes proving theorems and developing methods much easier.

2.1 Densities and Parameters

Density (Canonical form):

$$p_\eta(x) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x), \quad x \in \mathbb{R}^n$$

This is the general representation for a distribution in an exponential family.

$$p_\eta(x) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x)$$

$$p_\eta(x) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x)$$

η: The **natural/canonical parameters**.

$$p_\eta(x) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x)$$

η: The **natural/canonical parameters**.

$T(x)$: The **sufficient statistics**. These are functions of the data x.

$$p_\eta(x) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x)$$

η: The **natural/canonical parameters**.

$T(x)$: The **sufficient statistics**. These are functions of the data x.

$h(x)$: The **base measure**. It relates to the underlying support of the distribution.

$$p_\eta(x) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x)$$

η: The **natural/canonical parameters**.

$T(x)$: The **sufficient statistics**. These are functions of the data x.

$h(x)$: The **base measure**. It relates to the underlying support of the distribution.

$A(\eta)$: The **log-partition function** or **cumulant-generating function**. It acts as the normalizing constant.

Log-Partition Function & Natural Parameter Space

The log-partition function, $A(\eta)$, is defined by the integral that normalizes the probability distribution:

$$A(\eta) = \log \int \exp \left[\sum_{i=1}^{s} \eta_i T_i(x) \right] h(x) \, d\mu(x)$$

The set of parameters η for which this integral is finite $(A(\eta) < \infty)$ is called the **natural parameter space**, denoted by Ξ.

Example: The Exponential Distribution

The exponential distribution is a member of the exponential family.

The PDF of an exponential distribution is:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

For $x \geq 0$, we can rewrite the density. Reparameterize by setting

$\eta = -\lambda$ (so $\lambda = -\eta$).

Example: The Exponential Distribution

Substituting $\lambda = -\eta$ into the density for $x \geq 0$:

$$
\begin{aligned}
f(x) &= (-\eta)e^{\eta x} \\
&= \exp(\log(-\eta)) \exp(\eta x) \\
&= \exp(\eta x + \log(-\eta))
\end{aligned}
$$

To match the canonical form $p_\eta(x) = \exp\left[\eta T(x) - A(\eta)\right] h(x)$, we can write:

$$f(x) = \exp\left[\eta x - (-\log(-\eta))\right] \cdot 1$$

By comparison, we identify the components:

- Sufficient Statistic: $T(x) = x$
- Log-Partition Function: $A(\eta) = -\log(-\eta) = \log\left(-\frac{1}{\eta}\right)$
- Base Measure: $h(x) = 1$ (for $x \geq 0$)

Example: Verifying $A(\eta)$ and the Parameter Space

We can also find $A(\eta)$ directly from its definition, assuming Lebesgue measure with $h = 1_{(0,\infty)}$, $T_1(x) = x$:

$$A(\eta) = \log \int_0^\infty \exp(\eta \cdot T(x))h(x)\,dx = \log \int_0^\infty \exp(\eta x) \cdot 1\,dx$$

The integral is:

$$\int_0^\infty e^{\eta x}\,dx = \left[\frac{1}{\eta} e^{\eta x}\right]_0^\infty$$

This integral converges to $\log(-\frac{1}{\eta})$ only if $\eta < 0$. If $\eta \geq 0$, the integral is infinite. Therefore, $A(\eta) = \log\left(-\frac{1}{\eta}\right)$ for $\eta < 0$.

The exponential distribution is in the exponential family, with natural parameter space $\Xi = (-\infty, 0)$.

General Parameterizations

To allow for other, more common parameterizations for an exponential family, we can introduce a new parameter θ.

Let η be a function from some space Ω into the natural parameter space Ξ. The density can be written in terms of $\theta \in \Omega$:

$$p_\theta(x) = \exp\left[\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x)$$

- Here, θ is the **usual/common parameter**.
- $\eta(\theta)$ is a function that maps θ to the canonical parameters η.
- The new log-partition function is $B(\theta) = A(\eta(\theta))$.

Families of this form $\{p_\theta : \theta \in \Omega\}$ are called **s-parameter exponential families**.

Example: Normal Distribution

Write the PDF for the Normal distribution $N(\mu, \sigma^2)$ in the exponential family form. The common parameter is $\theta = (\mu, \sigma^2)$.

The PDF is:

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

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

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

Example: Normal Distribution

Continuing by regrouping the terms from the previous slide:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp \left(\underbrace{\left(\frac{\mu}{\sigma^2}\right)}_{\eta_1(\theta)} \underbrace{x}_{T_1(x)} + \underbrace{\left(-\frac{1}{2\sigma^2}\right)}_{\eta_2(\theta)} \underbrace{x^2}_{T_2(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)}_{B(\theta)} \right)$$

This matches the general form

$$p_\theta(x) = h(x) \exp \left[\sum \eta_i(\theta) T_i(x) - B(\theta)\right].$$

The components are:

- $h(x) = \frac{1}{\sqrt{2\pi}}$
- $T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$
- $\eta(\theta) = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix}$
- $B(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$

Example: i.i.d. Normal Sample

For a random sample $\mathbf{x} = (x_1, \ldots, x_n)$, the joint density is the product of the individual marginal densities, since the observations are independent.

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{n} p_\theta(x_i)$$

Using our result for a single observation:

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2} x_i - \frac{1}{2\sigma^2} x_i^2 - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma \right) \right) \right]$$

Since $\prod \exp(a_i) = \exp(\sum a_i)$, we can combine the terms:

$$p_\theta(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(\sum_{i=1}^{n} \left[\frac{\mu}{\sigma^2} x_i - \frac{1}{2\sigma^2} x_i^2 - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma \right) \right] \right)$$

$$= (2\pi)^{-\frac{n}{2}} \exp\left(\frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 - n \left(\frac{\mu^2}{2\sigma^2} + \log\sigma \right) \right)$$

i.i.d. Sample: Identification & General Rule

The final joint density is:

$$p_\theta(\mathbf{x}) = \underbrace{(2\pi)^{-\frac{n}{2}}}_{h(\mathbf{x})} \exp\left(\underbrace{\left(\frac{\mu}{\sigma^2}\right)}_{\eta_1(\theta)} \underbrace{\sum x_i}_{T_1(\mathbf{x})} - \underbrace{\left(\frac{1}{2\sigma^2}\right)}_{-\eta_2(\theta)} \underbrace{\sum x_i^2}_{T_2(\mathbf{x})} - \underbrace{n\left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)}_{B(\theta)} \right)$$

The new components for the i.i.d sample are:

- Sufficient Statistics: $T(\mathbf{x}) = \begin{pmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{pmatrix}$
- Log-Partition Function: $B(\theta) = n\left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)$
- Base Measure: $h(\mathbf{x}) = (2\pi)^{-n/2}$

General Rule: If you have an i.i.d. sample $x_1, \ldots, x_n$ from *any* exponential family, the joint density is also an exponential family. The new sufficient statistics are the sums of the original ones, and the new $B(\theta)$ is n times the original.

2.2 Differential Identities

Goal: We can find the moments of the sufficient statistics by differentiating the log-partition function $A(\eta)$.

Theorem

Let Ξ_f be the set of values for $\eta \in \mathbb{R}^s$ where

$$\int |f(x)| \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x) < \infty$$

Then the function

$$g(\eta) = \int f(x) \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x)$$

is continuous and has continuous partial derivatives of all orders for $\eta \in \Xi_f^o$ (the interior of Ξ_f). Furthermore, these derivatives can be computed by **differentiation under the integral sign**.

Deriving the First Moment

Let $f(x) = 1$. Then,:

$$g(\eta) = \int \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x)$$

By the definition of $A(\eta)$, this is simply $g(\eta) = e^{A(\eta)}$.
Now, we differentiate both sides w.r.t. a single parameter η_j.

Left-Hand Side (LHS): Using the chain rule,

Deriving the First Moment

Let $f(x) = 1$. Then,:

$$g(\eta) = \int \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x)$$

By the definition of $A(\eta)$, this is simply $g(\eta) = e^{A(\eta)}$.
Now, we differentiate both sides w.r.t. a single parameter η_j.

Left-Hand Side (LHS): Using the chain rule,

$$\frac{\partial}{\partial \eta_j} e^{A(\eta)} = e^{A(\eta)} \frac{\partial A(\eta)}{\partial \eta_j}$$

Right-Hand Side (RHS): Differentiating under the integral sign,

Deriving the First Moment

Let $f(x) = 1$. Then,:

$$g(\eta) = \int \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x)$$

By the definition of $A(\eta)$, this is simply $g(\eta) = e^{A(\eta)}$.
Now, we differentiate both sides w.r.t. a single parameter η_j.

Left-Hand Side (LHS): Using the chain rule,

$$\frac{\partial}{\partial \eta_j} e^{A(\eta)} = e^{A(\eta)} \frac{\partial A(\eta)}{\partial \eta_j}$$

Right-Hand Side (RHS): Differentiating under the integral sign,

$$\begin{aligned} \frac{\partial}{\partial \eta_j} RHS &= \int \frac{\partial}{\partial \eta_j} \exp\left[\sum \eta_i T_i(x)\right] h(x)\, d\mu(x) \\ &= \int T_j(x) \exp\left[\sum \eta_i T_i(x)\right] h(x)\, d\mu(x) \end{aligned}$$

$$e^{A(\eta)} \frac{\partial A(\eta)}{\partial \eta_j} = \int T_j(x) \exp\left[\sum \eta_i T_i(x)\right] h(x)\, d\mu(x)$$

$$e^{A(\eta)} \frac{\partial A(\eta)}{\partial \eta_j} = \int T_j(x) \exp\left[\sum \eta_i T_i(x)\right] h(x)\, d\mu(x)$$

Dividing both sides by $e^{A(\eta)}$ is equivalent to inserting $e^{-A(\eta)}$:

$$\begin{aligned}\frac{\partial A(\eta)}{\partial \eta_j} &= \int T_j(x) \exp\left[\sum \eta_i T_i(x) - A(\eta)\right] h(x)\, d\mu(x) \\ &= \int T_j(x) \cdot p_\eta(x)\, d\mu(x)\end{aligned}$$

$$e^{A(\eta)} \frac{\partial A(\eta)}{\partial \eta_j} = \int T_j(x) \exp\left[\sum \eta_i T_i(x)\right] h(x)\, d\mu(x)$$

Dividing both sides by $e^{A(\eta)}$ is equivalent to inserting $e^{-A(\eta)}$:

$$\begin{aligned} \frac{\partial A(\eta)}{\partial \eta_j} &= \int T_j(x) \exp\left[\sum \eta_i T_i(x) - A(\eta)\right] h(x)\, d\mu(x) \\ &= \int T_j(x) \cdot p_\eta(x)\, d\mu(x) \end{aligned}$$

By the definition of expectation, the RHS is $\mathbb{E}_\eta[T_j(x)]$.

The first partial derivative of the log-partition function $A(\eta)$ with respect to η_j is the expected value of the corresponding sufficient statistic $T_j(x)$.

$$\frac{\partial A(\eta)}{\partial \eta_j} = \mathbb{E}_\eta[T_j(x)]$$

2.3 Dominated Convergence

We now know that we can differentiate under the integral sign. The formal justification for this is the **Dominated Convergence Theorem**.

Consider the case with a single parameter ($s = 1$) and examine the derivative at $\eta = 0$. The definition of a derivative is a limit. For a function $G(\eta)$, the derivative at $\eta = 0$ is:

$$G'(0) = \lim_{h \to 0} \frac{G(h) - G(0)}{h}$$

2.3 Dominated Convergence

We now know that we can differentiate under the integral sign. The formal justification for this is the **Dominated Convergence Theorem**.

Consider the case with a single parameter ($s = 1$) and examine the derivative at $\eta = 0$. The definition of a derivative is a limit. For a function $G(\eta)$, the derivative at $\eta = 0$ is:

$$G'(0) = \lim_{h \to 0} \frac{G(h) - G(0)}{h}$$

In our case, we used $G(\eta) = e^{A(\eta)} = \int e^{\eta T(x)} h(x)\, d\mu(x)$.

Substituting this in:

$$G'(0) = \lim_{h \to 0} \frac{\int e^{hT(x)} h(x)\, d\mu(x) - \int e^{0 \cdot T(x)} h(x)\, d\mu(x)}{h}$$

$$G'(0) = \lim_{h \to 0} \int \left[\frac{e^{hT(x)} - 1}{h} \right] h(x)\, d\mu(x)$$

Define the function inside the integral: $f_h(x) = \left[\frac{e^{hT(x)}-1}{h} \right] h(x)$.

$$G'(0) = \lim_{h \to 0} \int \left[\frac{e^{hT(x)} - 1}{h} \right] h(x)\, d\mu(x)$$

Define the function inside the integral: $f_h(x) = \left[\frac{e^{hT(x)}-1}{h} \right] h(x)$.

First, we find the pointwise limit of this function as $h \to 0$. Using $\lim_{y \to 0} \frac{e^{ay}-1}{y} = a$, we get:

$$\lim_{h \to 0} f_h(x) = T(x)h(x)$$

$$G'(0) = \lim_{h \to 0} \int \left[\frac{e^{hT(x)} - 1}{h} \right] h(x)\, d\mu(x)$$

Define the function inside the integral: $f_h(x) = \left[\frac{e^{hT(x)}-1}{h} \right] h(x)$.

First, we find the pointwise limit of this function as $h \to 0$. Using

$\lim_{y \to 0} \frac{e^{ay}-1}{y} = a$, we get:

$$\lim_{h \to 0} f_h(x) = T(x)h(x)$$

The problem of finding $G'(0)$ is now about whether we can swap

the limit and the integral:

$$\lim_{h \to 0} \int f_h(x)\, d\mu(x) \stackrel{?}{=} \int \left(\lim_{h \to 0} f_h(x) \right) d\mu(x)$$

The Dominated Convergence Theorem

The theorem that formally permits swapping the limit and integral is the Dominated Convergence Theorem.

Theorem (Dominated Convergence)

Let $\{f_n\}$ be a sequence of functions. If:

1. *Pointwise convergence:* $\lim_{n\to\infty} f_n(x) = f(x)$ *for almost every* x.
2. *Domination: There exists an integrable function g (i.e., $\int |g(x)|\, d\mu(x) < \infty$) such that $|f_n(x)| \leq g(x)$ for all n.*

Then, the limit and integral can be swapped:

$$\lim_{n\to\infty} \int f_n(x)\, d\mu(x) = \int \lim_{n\to\infty} f(x)\, d\mu(x) = \int f d\mu$$

Why Domination Matters

Consider the function: $f_n(x) = \mathbf{1}_{[n,n+1]}(x)$.

Why Domination Matters

Consider the function: $f_n(x) = \mathbf{1}_{[n,n+1]}(x)$.

- **Pointwise Limit:** For any fixed x, $\lim_{n\to\infty} f_n(x) = 0$.

Why Domination Matters

Consider the function: $f_n(x) = \mathbf{1}_{[n,n+1]}(x)$.

- **Pointwise Limit:** For any fixed x, $\lim_{n\to\infty} f_n(x) = 0$.
- **Integral of the Limit:** $\int \left(\lim_{n\to\infty} f_n(x)\right) dx = \int 0 \, dx = 0$.

Why Domination Matters

Consider the function: $f_n(x) = \mathbf{1}_{[n,n+1]}(x)$.

- **Pointwise Limit:** For any fixed x, $\lim_{n\to\infty} f_n(x) = 0$.
- **Integral of the Limit:** $\int (\lim_{n\to\infty} f_n(x))\, dx = \int 0\, dx = 0$.
- **Limit of the Integrals:** The area under the curve is always 1. So, $\lim_{n\to\infty} \int f_n(x)\, dx = \lim_{n\to\infty} 1 = 1$.

Why Domination Matters

Consider the function: $f_n(x) = \mathbf{1}_{[n,n+1]}(x)$.

- **Pointwise Limit:** For any fixed x, $\lim_{n\to\infty} f_n(x) = 0$.
- **Integral of the Limit:** $\int (\lim_{n\to\infty} f_n(x))\ dx = \int 0\, dx = 0$.
- **Limit of the Integrals:** The area under the curve is always 1. So, $\lim_{n\to\infty} \int f_n(x)\, dx = \lim_{n\to\infty} 1 = 1$.

Since $0 \neq 1$, the swap fails.

Why Domination Matters

Consider the function: $f_n(x) = \mathbf{1}_{[n,n+1]}(x)$.

- **Pointwise Limit:** For any fixed x, $\lim_{n\to\infty} f_n(x) = 0$.
- **Integral of the Limit:** $\int (\lim_{n\to\infty} f_n(x))\; dx = \int 0\, dx = 0$.
- **Limit of the Integrals:** The area under the curve is always 1. So, $\lim_{n\to\infty} \int f_n(x)\, dx = \lim_{n\to\infty} 1 = 1$.

Since $0 \neq 1$, the swap fails.

Reason

The domination condition failed. Any function $g(x)$ that dominates the sequence would have to be at least 1 on $[0, \infty)$, making its integral infinite.

Domination in Exponential Families

On the other hand, the sequence of functions from our derivative calculation *does* satisfy the domination condition.

Recall our function: $f_h(x) = \left[\frac{e^{hT(x)}-1}{h}\right] h(x)$. We need to find an integrable function $g(x)$ that dominates $|f_h(x)|$ for all h in a neighborhood of 0.

Domination in Exponential Families

On the other hand, the sequence of functions from our derivative calculation *does* satisfy the domination condition.

Recall our function: $f_h(x) = \left[\frac{e^{hT(x)}-1}{h}\right] h(x)$. We need to find an integrable function $g(x)$ that dominates $|f_h(x)|$ for all h in a neighborhood of 0.

We use the key inequality $|e^t - 1| \leq |t|e^{|t|}$, which holds for all $t \in \mathbb{R}$. Letting $t = hT(x)$, we can find an upper bound for $|f_h(x)|$:

$$\begin{aligned} |f_h(x)| &= \left|\frac{e^{hT(x)} - 1}{h}\right| |h(x)| \\ &\leq \frac{|hT(x)|e^{|hT(x)|}}{|h|} |h(x)| \\ &= |T(x)|e^{|h||T(x)|} |h(x)| \end{aligned}$$

Constructing the Dominating Function

$$|f_h(x)| \leq |T(x)|e^{|h||T(x)|}|h(x)|$$

The goal is to find a single function $g(x)$, which does not depend on h, that is greater than $|f_h(x)|$ for all h in a neighborhood of 0 (e.g., for $|h| \leq \epsilon$ for some small $\epsilon > 0$).

For a small, fixed $\epsilon > 0$, we can define the following dominating function:

$$g(x) := \frac{1}{\epsilon}\left(e^{2\epsilon T(x)} + e^{-2\epsilon T(x)}\right) h(x)$$

This function $g(x)$ works as an upper bound for $|f_h(x)|$.

Constructing the Dominating Function

$$|f_h(x)| \le |T(x)|e^{|h||T(x)|}|h(x)|$$

The goal is to find a single function $g(x)$, which does not depend on h, that is greater than $|f_h(x)|$ for all h in a neighborhood of 0 (e.g., for $|h| \le \epsilon$ for some small $\epsilon > 0$).

For a small, fixed $\epsilon > 0$, we can define the following dominating function:

$$g(x) := \frac{1}{\epsilon}\left(e^{2\epsilon T(x)} + e^{-2\epsilon T(x)}\right) h(x)$$

This function $g(x)$ works as an upper bound for $|f_h(x)|$. Crucially,

$g(x)$ has a finite integral, provided $\pm 2\epsilon$ are in the natural parameter space:

$$\int g(x)\, d\mu(x) = \frac{1}{\epsilon}\left[\int e^{2\epsilon T(x)} h(x)\, d\mu(x) + \int e^{-2\epsilon T(x)} h(x)\, d\mu(x)\right]$$
$$= \frac{1}{\epsilon}\left(e^{A(2\epsilon)} + e^{A(-2\epsilon)}\right) < \infty$$

Constructing the Dominating Function

$$\begin{aligned}\int g(x)\,d\mu(x) &= \frac{1}{\epsilon}\left[\int e^{2\epsilon T(x)} h(x)\,d\mu(x) + \int e^{-2\epsilon T(x)} h(x)\,d\mu(x)\right] \\ &= \frac{1}{\epsilon}\left(e^{A(2\epsilon)} + e^{A(-2\epsilon)}\right) < \infty\end{aligned}$$

Constructing the Dominating Function

$$\int g(x)\,d\mu(x) = \frac{1}{\epsilon}\left[\int e^{2\epsilon T(x)} h(x)\,d\mu(x) + \int e^{-2\epsilon T(x)} h(x)\,d\mu(x)\right]$$
$$= \frac{1}{\epsilon}\left(e^{A(2\epsilon)} + e^{A(-2\epsilon)}\right) < \infty$$

Conclusion

Since we have found an integrable function $g(x)$ that dominates $|f_h(x)|$, the Dominated Convergence Theorem applies.

$$\lim_{h\to 0}\int f_h(x)\,d\mu(x) = \int \lim_{h\to 0} f(x)\,d\mu(x)$$