Understanding Exponential Families: Theory, Proofs, and Applications

Wanrun Yang

March 15, 2024

Abstract

Exponential families play an important role in probability theory and statistics, providing a powerful framework for modeling a wide range of probability distributions. In this project, we delve into the theory of exponential families, exploring their properties and the implications they have for calculating the mean and variance of random variables. We begin by presenting the two key statements of the theorem on exponential families, followed by detailed proofs and illustrative examples to aid understanding. Additionally, we demonstrate how these concepts are applied in practical scenarios.

Introduction

Exponential families are a class of probability distributions characterized by elegant mathematical properties that make them particularly useful in statistical modeling. They encompass a broad range of distributions, including the normal, exponential, Poisson, and gamma distributions, among others. In this paper, we aim to provide a comprehensive understanding of exponential families, focusing on their theoretical underpinnings and practical applications.

Definition

An exponential family refers to a probability density function or probability mass function that can be represented in the following expression:

$$f_X(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right)$$

Note:

1. $h(x), t_1(x), ..., t_k(x)$ do not depend on θ
2. $c(\theta)$ does not depend of x

Examples of Exponential Families

Binomial Distribution

The probability mass function (PMF) of the binomial distribution is given by:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where:

- x is the number of successes,
- n is the number of trials,
- p is the probability of success in each trial.

Let's express it in the exponential family form:

$$\begin{aligned} p(x) &= \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \\ &= \left(\begin{array}{c} n \\ x \end{array} \right) \left(\frac{p}{1-p} \right)^x (1-p)^n \\ &= \left(\begin{array}{c} n \\ x \end{array} \right) (1-p)^n e^{\log\left(\frac{p}{1-p}\right)^x} \\ &= \left(\begin{array}{c} n \\ x \end{array} \right) (1-p)^n e^{x \log\left(\frac{p}{1-p}\right)} \end{aligned}$$

We can conclude that this pmf is an exponential family with $h(x) = \left(\begin{array}{c} n \\ x \end{array} \right), c(p) = (1-p)^n, t_1(x) = x, w_1(p) = \log \frac{p}{1-p}$.

Poisson Distribution

The probability mass function (PMF) of the Poisson distribution is given by:

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where:

- x is the number of events,
- λ is the average rate of occurrence.

Let's express it in the exponential family form: $p(x) = \frac{\lambda^x e^{-x}}{x!} = \frac{1}{x!} e^{-\lambda} e^{\log \lambda^x} = \frac{1}{x!} e^{-\lambda} e^{(\log \lambda) x}$

We can conclude that this pmf is an exponential family with

$$h(x) = \frac{1}{x!}, c(\lambda) = e^{-\lambda}, w_1(\lambda) = \log \lambda, \quad t(x) = x$$

Theorem on Exponential Families

The theorem on exponential families states two crucial statements: Suppose a random variable X has a probability density function (pdf) or probability mass function (pmf) that can be expressed in the form of exponential family.

Statement 1

$$\text{(a)} \quad E\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(x) \right) = -\frac{\partial}{\partial \theta_j} \log c(\boldsymbol{\theta}).$$

Statement 2

$$\text{(b)} \quad \operatorname{var}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(x) \right) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) - E\left(\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(x) \right).$$

Proofs

Proof of Statement 1

$$\int_x f(x \mid \boldsymbol{\theta}) dx = 1$$

$$\int_x h(x) c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right) dx = 1$$

Differentiate both sides w.r.t. θ_j :

$$\int_x h(x) \frac{\partial c(\boldsymbol{\theta})}{\partial \theta_j} \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right) dx$$
$$+ \int_x h(x) c(\boldsymbol{\theta}) \sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(x) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right) dx = 0$$

Multiply the first integral by $\frac{c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})}$ and note that $\frac{\partial \log c(\boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial c(\boldsymbol{\theta})}{\partial \theta_j} \frac{1}{c(\boldsymbol{\theta})}$.

$$\int_x h(x) \frac{\partial c(\boldsymbol{\theta})}{\partial \theta_j} \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right) \frac{c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})} dx$$
$$+ \quad \int_x h(x) c(\boldsymbol{\theta}) \sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(x) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right) dx = 0$$

After rearranging we get

$$\int_x \sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(x) h(x) c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right) dx =$$
$$- \frac{\partial \log c(\boldsymbol{\theta})}{\partial \theta_j} \int_x h(x) c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right) dx$$

Or

$$E\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(x)\right) = -\frac{\partial}{\partial \theta_j} \log(\boldsymbol{\theta}).$$

Proof of Statement 2

Given the theorem:

$$\operatorname{var}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(x)\right) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) - E\left(\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(x)\right)$$

We'll start by calculating the variance of the given expression. Let's denote the random variable as $Y = \sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(x)$. Then, the variance of Y can be expressed as:

$$\operatorname{var}(Y) = \operatorname{var}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(x)\right)$$

Using the linearity of variance, this can be expanded as:

$$\text{var}(Y) = \sum_{i=1}^{k} \sum_{l=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial w_l(\boldsymbol{\theta})}{\partial \theta_j} \text{cov}(t_i(x), t_l(x))$$

Given that $t_i(x)$'s are the sufficient statistics, they are uncorrelated under the assumed model, so $\text{cov}(t_i(x), t_l(x)) = 0$ for $i \neq l$. Therefore, the above expression simplifies to:

$$\text{var}(Y) = \sum_{i=1}^{k} \left(\frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \right)^2 \text{var}(t_i(x))$$

Now, we can rewrite the variance using the second moment about the mean as:

$$\text{var}(Y) = \sum_{i=1}^{k} \left(\frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \right)^2 \left(E\left(t_i(x)^2\right) - E\left(t_i(x)\right)^2 \right)$$

Since $t_i(x)$ are sufficient statistics, $E(t_i(x)^2) = \text{var}(t_i(x)) + E(t_i(x))^2$. Substituting this into the expression:

$$\text{var}(Y) = \sum_{i=1}^{k} \left(\frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \right)^2 \left(\text{var}(t_i(x)) + E(t_i(x))^2 - E(t_i(x))^2 \right)$$

$$\text{var}(Y) = \sum_{i=1}^{k} \left(\frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \right)^2 \text{var}(t_i(x))$$

Now, we utilize the definition of the variance of a sufficient statistic:

$$\text{var}(t_i(x)) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta})$$

Substituting this back into the expression for $\text{var}(Y)$:

$$\text{var}(Y) = \sum_{i=1}^{k} \left(\frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \right)^2 \left(-\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) \right)$$

$$\text{var}(Y) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) \sum_{i=1}^{k} \left(\frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \right)^2$$

Now, we use the fact that the expectation of the second derivative of $w_i(\boldsymbol{\theta})$ with respect to θ_j can be expressed as:

$$E\left(\frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} \right) = -E\left(\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(x) \right)$$

Thus, we have:

$$\text{var}(Y) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) - E\left(\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(x) \right)$$

Examples

Example 1:

Let $X \sim \text{Poisson}(\lambda)$. Use the theorem above to show that $E[X] = \lambda$ and $\text{var}[X] = \lambda$.

Solution:

$$E\left(\frac{\partial w_i \theta}{\partial \theta i} ti(x)\right) = -\frac{\partial}{\partial \theta j} \log c(\theta)$$

$$\left(\frac{\partial \log \lambda}{\partial \lambda} \cdot x\right) \quad E\left(\frac{1}{\lambda} \cdot x\right) = -(-1)$$

$$\frac{1}{\lambda} E(x) = 1$$

$$E(x) = \lambda$$

$$\text{Var}\left(\frac{1}{\lambda} \cdot x\right) = 0 - E\left(-\frac{1}{\lambda^2} \cdot x\right) \frac{\partial^2 c(A)}{\partial^2 \lambda}$$

$$= 0$$

$$\frac{1}{\lambda^2} \text{var}(x) = \frac{1}{\lambda^2} E(x)$$

$$\text{Var}(x) = E(x) = \lambda$$

Example 2:

Let $x \sim N(\mu, \theta)$. Use the theorem above to show $E[X]$ and $\text{var}[X]$.

Solution:

$$x \sim N(\mu, \theta)$$

$$f(x) = \frac{1}{\theta\sqrt{2\pi}} e^{-\frac{1}{2\theta^2}(x-\mu)^2}$$

$$= \left(2\pi\theta^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2\theta^2}x^2 + \frac{\mu^2}{\theta^2}x}$$

$$h(x) = 1$$

$$c(\theta) = \left(2\pi\theta^2\right)^{-\frac{1}{2}} e^{-\frac{\mu^2}{2\theta^2}}$$

$$w_1(\theta) = -\frac{1}{\theta^2}$$

$$\omega_2(\theta) = \frac{\mu}{\theta^2}$$

$$t_1(x) = -\frac{x^2}{2}$$

$$t_2(x) = x$$

$$\log g(c(\theta)) = -\frac{1}{2}\log 2\pi\theta^2 - \frac{\mu^2}{2\theta^2}$$

$$E\left(\Sigma \frac{\partial \omega_1(\mu)}{\partial \mu} t_1(x)\right) = -\frac{\partial^2}{\partial \mu} \log c(\theta)$$

$$E\left(\frac{1}{\theta^2} \cdot x\right) = -\left(-\frac{\mu}{\theta^2}\right)$$

$$\frac{1}{\theta^2} E(x) = \frac{\mu}{\theta^2}$$

$$E(x) = \mu$$

$$\operatorname{Var}\left(\sum \frac{\partial \omega_1(\mu)}{\partial \mu} t_1(x)\right) = \frac{\partial^2}{\partial \mu^2} \log c(\theta) - E\left(\sum \frac{\partial \omega_1^2(\mu)}{\partial^2 \mu} t_1(x)\right)$$

$$\operatorname{Var}\left(\frac{1}{\theta^2} \cdot x\right) = -\left(-\frac{1}{\theta^2}\right)$$

$$\frac{1}{\theta^\mu} \operatorname{Var}(x) = \frac{1}{\theta^2}$$

$$\operatorname{Var}(x) = \theta^2$$

Conclusion

In conclusion, exponential families offer a versatile framework for modeling probability distributions, with implications for computing essential statistical quantities such as the mean and variance. The theorem on exponential families provides a theoretical foundation for understanding these concepts and their applications in real-world scenarios. By mastering the principles outlined in this paper, readers can enhance their proficiency in probability theory and statistical analysis.