Lab 5

(Exponential Families), (Inequalities: Part I), (Multivariate RV's)

5.1. Exponential Families

A family of pdfs or pmfs is called an *exponential family* if it can be expressed as

$$f(x|\theta) = h(x)c(\theta)\exp\{\boldsymbol{\eta}(\theta)^\intercal \mathbf{t}(x)\}$$

where $\boldsymbol{\eta}(\theta) = (\eta_1(\theta), \cdots, \eta_k(\theta))$ and $\mathbf{t}(x) = (t_1(x), \cdots, t_k(x))$. The terms θ and x could be vectors as well. Another parametrization is to have $A(\theta) = -\log(c(\theta))$, and hence we have

$$f(x|\theta) = h(x)\exp\{\boldsymbol{\eta}(\theta)^\intercal \mathbf{t}(x) - A(\theta)\}$$

In the latter form, each has an interpretation: $\boldsymbol{\eta}$ is called the natural parameter, $\mathbf{t}$ is called the sufficient statistic, and A is called the log-partition function.The term $h(x)$ doesn't really have any interpretation; we'll just call this the base measure.

In the same way that specific families of distributions are important, exponential families are crucial to our understanding of many advanced statistical subjects (which will be covered in future courses and include topics like sufficient statistics, conjugate priors, generalized linear models, etc). Distributions which fall in the exponential family include the normal, exponential, gamma, beta, binomial, poisson, and negative binomial. Note that the same exponential family distribution can be parametrized in a number of ways. Generally it is a good idea to keep $\mathbf{t}(X)$ as 'clean' and interpretable as possible.

Sometimes there is a restriction on the values x can take in a pdf. We express these restrictions through *indicator functions:*

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

For example, the exponential pdf can be written as

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} I_{(0,\infty)}(x)$$

Note that the indicator function can be a function of θ and x or just x alone.

Theorem 5.1.1. If X is a random variable with pdf or pmf of exponential family form, then

$$\begin{aligned}\mathbb{E}\frac{\partial}{\partial\theta_j}\boldsymbol{\eta}(\theta)^\intercal\mathbf{t}(X) &= \frac{\partial}{\partial\theta_j}A(\theta)\\ \mathrm{Var}\left(\frac{\partial}{\partial\theta_j}\boldsymbol{\eta}(\theta)^\intercal\mathbf{t}(X)\right) &= \frac{\partial^2}{\partial\theta_j^2}A(\theta) - \mathbb{E}\frac{\partial^2}{\partial\theta_j^2}\boldsymbol{\eta}(\theta)^\intercal\mathbf{t}(X)\end{aligned}$$

Definition 5.1.1. A *curved exponential family* is a family of densities for which the dimension of the vector θ is equal to $d < k$. If $d = k$ then the family is a *full exponential family.*

5.2. Markov/Chebyshev's Inequality

The Chebyshev inequality is one of the most important and most heavily-used inequalities in statistics. It is also one of the most elegant.

Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,

$$\mathbb{P}(g(X) \geq r) \leq \frac{\mathbb{E}\left[g(X)\right]}{r}$$

The classic use of Chebyshev is to bound the probability of deviating from the mean:

$$\begin{aligned}\mathbb{P}\left(\frac{(X-\mu)^2}{\sigma^2} \geq t^2\right) &\leq \frac{1}{t^2}\mathbb{E}\left[\frac{(X-\mu)^2}{\sigma^2}\right]\\ &= \frac{1}{t^2}\\ &\Rightarrow \mathbb{P}(|X-\mu| \geq t\sigma) \leq \frac{1}{t^2}\end{aligned}$$

5.3. Multiple Random Variables

A surprisingly large amount of statistical research is concerned with univariate models, but it is certainly also necessary to understand what happens when we are interested in probability models which involve more than one random variable. Recall that our univariate random variable was defined to be a function from the sample space S into the real numbers. A random vector of multiple random variables is defined similarly.

Definition 5.3.1. An *n-dimensional random vector* is a function from a sample space S into $\mathbb{R}^n$, n-dimensional Euclidean space.

Definition 5.3.2. Let (X, Y) be a discrete bivariate random vector. Then the function $f_{XY}(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ defined by

$$f_{XY}(x, y) = P(X = x, Y = y)$$

is called the *joint probability mass function* or *joint pmf* of (X, Y). The joint pmf of (X, Y) completely defines the probability distribution of the random vector (X, Y).

We first list some properties of discrete bivariate random vectors, but the principles generalize for dimensions greater than two. In the case where X and Y are discrete, many results from the univariate setting have analogues in the multivariate case.

1. The joint pmf can be used to compute the probability of any event defined in terms of (X, Y). Let A be any subset of $\mathbb{R}^2$. Then

$$\mathbb{P}((X,Y) \in A) = \sum_{(x,y)\in A} f_{XY}(x,y)$$

2. Let $g(x, y)$ be a real-valued function defined for all possible values (x, y) of the discrete random vector (X, Y). Then $g(X, Y)$ is itself a random variable and its expected value $E[g(X, Y)]$ is given by

$$\mathbb{E}[g(X,Y)] = \sum_{(x,y)\in\mathbb{R}^2} g(x,y) f_{XY}(x,y)$$

3. Let (X, Y) be a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$. Then the marginal pmfs of X and Y, $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by

$$\begin{aligned} f_X(x) &= \sum_{y\in\mathbb{R}} f_{X,Y}(x,y) \\ f_Y(y) &= \sum_{x\in\mathbb{R}} f_{X,Y}(x,y) \end{aligned}$$

Definition 5.3.3. A function $f_{XY}(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ is called a *joint probability density function* or *joint pdf* of the continuous bivariate random vector (X, Y) if, for every $A \subset \mathbb{R}^2$

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{XY}(x,y)\,dxdy$$

Any function $f_{XY}(x, y)$ satisfying $f_{XY}(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$ and

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dxdy$$

is the joint pdf of some continuous bivariate random vector (X, Y). Again the properties below generalize for dimensions greater than 2.

1. The joint pmf can be used to compute the probability of any event defined in terms of (X, Y). Let A be any subset of $\mathbb{R}^2$. Then

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{XY}(x,y)\,dxdy$$

2. Let $g(x, y)$ be a real-valued function defined for all possible values (x, y) of the random vector (X, Y). Then $g(X, Y)$ is itself a random variable and its expected

value $E[g(X,Y)]$ is given by

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dx dy$$

3. The marginal pmfs of X and Y are given by

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dy, \quad -\infty < x < \infty \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dx, \quad -\infty < y < \infty \end{aligned}$$

Definition 5.3.4. Let (X,Y) be a discrete bivariate random vector with joint pmf $f_{XY}(x,y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $P(X=x) = f_X(x) > 0$, the *conditional pmf of Y given $X = x$* is the function of y denoted by $f_{Y|X}(y|x)$ and defined by

$$f_{Y|X}(y|x) = P(Y=y|X=x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

For any y such that $P(Y=y) = f_Y(y) > 0$, the *conditional pmf of X given that $Y=y$* is the function of x denoted by $f_{X|Y}(x|y)$ and defined by

$$f_{X|Y}(x|y) = P(X=x|Y=y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

The definition is the same for the continuous case.

Definition 5.3.5. Remember that conditional pdfs and pmfs have all the properties of their marginal counterparts. In particular, they can be used to calculated conditional expectations and variances. If $g(Y)$ is a function of Y, then the *conditional expected value of g(Y) given that $X = x$* is given by

$$\begin{aligned} \mathbb{E}[g(Y)|x] &= \sum_y g(y) f_{Y|X}(y|x) \\ \text{or } \mathbb{E}[g(Y)|x] &= \int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x) dy \end{aligned}$$

imilarly the *conditional variance of $g(Y)$ given that $X = x$* is given by

$$\begin{aligned} \text{Var}(Y|x) &= \sum_y g^2(y) f_{Y|X}(y|x) - \left(\sum_y g(y) f_{Y|X}(y|x)\right)^2 \\ \text{or } \text{Var}(Y|x) &= \int_{-\infty}^{\infty} g^2(y) f(y|x) dy - \left(\int_{-\infty}^{\infty} g(y) f(y|x) dy\right)^2 \end{aligned}$$

Definition 5.3.6. Let (X,Y) be a bivariate random vector with joint pdf or pmf $f_{Y|X}(x,y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called *independent random variables* if, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f_{XY}(x,y) = f_X(x) f_Y(y)$$

Furthermore, if X and Y are independent, the conditional pdf of Y given $X = x$ is

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} \\ &= \frac{f_X(x)f_Y(y)}{f_X(x)} \\ &= f_Y(y) \end{aligned}$$

regardless of the value of x. Thus, for any $A \subset \mathbb{R}$ and $x \in \mathbb{R}$,

$$\mathbb{P}(Y \in A|x) = \int_A f_{Y|X}(y|x)dy = \int_A f_Y(y)dy = P(Y \in A)$$

The knowledge that $X = x$ gives us no additional information about Y.

Lemma 5.3.1. Let (X, Y) be a bivariate random vector with joint pdf or pmf $f_{XY}(x, y)$. Then X and Y are indepeendent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f_{XY}(x, y) = g(x)h(y)$$

Theorem 5.3.1. Let X and Y be independent random variables.

- For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$; that is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.
- Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

5.4. Additional Problems

1. To get in the "mood" for inequalities to come, prove the following elementary inequalities:

   - $-\log(1-u) - u \le \frac{u^2}{2(1-u)}$ for $u \in (0,1)$
   - $\log(1-\theta+\theta e^u) - \theta u \le \frac{1}{8}u^2$ for $u, \theta > 0$
   - $$\min_{1\le i\le n} \frac{a_i}{b_i} \le \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \le \max_{1\le i\le n} \frac{a_i}{b_i}$$

     for any positive $a_1, \cdots, a_n, b_1, \cdots, b_n$.

2. Let X have finite second moment. Denote $\mu = \mathbb{E}X, \sigma^2 = \mathrm{Var}(X)$, and $\widetilde{x}$ as the median of X. Prove that

$$|\mu - \widetilde{x}| \le \sigma$$

3. (The "random meeting" problem.) Person A and person B have agreed to meet between 7:00 p.m. and 8:00 p.m. at a particular location. They have both forgotten the exact meeting time and choose their respective arrival times randomly and independently from each other between 7:00 p.m. and 8:00 p.m., according to the uniform distribution on the interval [7:00, 8:00]. They both have patience to wait no longer than 10 min. Prove that the probability that A and B will actually meet equals 11/36.

4. Suppose $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$. Show that $X|(X+Y = n) \sim \text{Bin}(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$.

5. Suppose X, Y satisfy $\mathbb{E}|X|^2, \mathbb{E}|Y|^2 < \infty$ and that $\mathbb{E}(X|Y) = Y, \mathbb{E}(Y|X) = X$. Prove that $X \overset{\text{a.s}}{=} Y$.
Remark: The statement is still true when $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$, although the problem is much harder.