STAT 532: Foundations of Statistical Inference I **Fall 2020**

Lecture 2: Probability Measures on Euclidean Spaces

Lecturer: Dennis Cox *Scribes: Shu-Hsien Cho*

2.1 Moments and Moment Inequalities

Let X be a random variable. If $E[X^k]$ is finite, where k is a positive integer, then $E[X^k]$ is called the k-th moment of X or P_X (the distribution of X). If $E|X|^a < \infty$ for some real number a, then $E|X|^a$ is called the ath absolute moment of X or P_X. If $\mu = E[X]$ and $E[X-\mu]^k$ are finite for a positive integer k, then $E[X-\mu]^k$ is called the k-th central moment of X or P_X. If $E|X|^a < \infty$ for an $a > 0$, then $E|X|^t < \infty$ for any positive $t < a$ and $E[X]^k$ is finite for any positive integer $k \leq a$.

Now assume further that $\underline{X}$ has finite second moments, i.e. each component of $\underline{X}$ has finite second moment, which is the same as $E[||\underline{X}||^2] < \infty$ of the second moment implies $\underline{X}$ has finite first moments, i.e. $E[||\underline{X}||] < \infty$ follows from the following calculation.

$$\begin{aligned} E[||X||] &= \int_{\mathbb{R}^n} ||\underline{x}|| dP_{\underline{X}}(\underline{x}) = \int_{\{x:||x||<1\}} ||\underline{x}|| dP_{\underline{X}}(\underline{x}) + \int_{\{x:||x||\geq 1\}} ||\underline{x}|| dP_{\underline{X}}(\underline{x}) \\ &\leq \int_{\{x:||x||<1\}} 1 dP_{\underline{X}}(\underline{x}) + \int_{\{x:||x||\geq 1\}} ||\underline{x}||^2 dP_{\underline{X}}(\underline{x}) \leq P[||\underline{X}|| < 1] + E[||\underline{X}||^2] < \infty \end{aligned}$$

In general, we say $\underline{X}$ has finite p'th moment, $0 < p < \infty$, if $E[||\underline{X}||^p] < \infty$. If $\underline{X}$ has finite p'th moment, all the smaller moments are also finite.

For sample mean, we could have empirical distribution $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$. $E[X] = \int_\Omega X d\mathbb{P} = \int_{\mathbb{R}} x dP_X(x)$. $\int x d\hat{P}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$. For sample variance, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \int (x_i - \bar{x})^2 d\hat{P}(x)$

2.1.1 Elementary Moment Bound

Proposition 2.1.1. ***Markov's Inequality****: Suppose $X \geq 0$ a.s., then $\forall \epsilon > 0$, $P[X > \epsilon] \leq \frac{E[X]}{\epsilon}$*

Note*. $P[X > \epsilon]$ is a shorthand way to write $P(\{\omega \in \Omega : X(\omega) > \epsilon\})$, which is equal to $(\mathbb{P} \circ X^{-1})((\epsilon, \infty))$. There are many other inequalities that are trivial corollaries of this one, e.g. $P[X > \epsilon] \leq \frac{E[X]}{\epsilon}$. We will give the proof in some detail, although it is really quite elementary. The student should be able to reproduce the proof and completely justify each detail.*

Proof.

$$\begin{aligned}
E[X] &= \int_0^\infty x dP_X(x) \qquad \text{(by Change of Variables)} \\
&= \int_{[0,\epsilon)} x dP_X(x) + \int_{[\epsilon,\infty)} x dP_X(x) \qquad \text{(by additivity of the integral as applied to } x = I_{[0,\epsilon)}(x)x + I_{[\epsilon,\infty)}(x)x) \\
&\geq \int_{(\epsilon,\infty)} x dP_X(x) \qquad \text{(by integral monotonicity)} \\
&\geq \int_{(\epsilon,\infty)} \epsilon dP_X(x) \qquad \text{(by integral monotonicity because } I_{[\epsilon,\infty)}(x)x \geq I_{[\epsilon,\infty)}(x)\epsilon) \\
&= \epsilon P[X > \epsilon] \qquad \text{(by linearity of integral and the fact that } \int I_A dP = P(A))
\end{aligned}$$

□

Proposition 2.1.2. ***Chebyshev's Inequality:*** *Suppose X is a r.v. with $E[X^2] < \infty$. Let $\mu = E[X]$ and $\sigma^2 = Var[X]$. Then then $\forall k > 0$, $P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$*

Proof. Apply Markov's inequality to the r.v. $(X - \mu_X)^2$ with $\epsilon = (k\sigma_X)^2$ □

2.1.2 Convexity and Jensen's Inequality

A set $K \subset \mathbb{R}^n$ is called convex iff for any finite subset $\{\underline{x_1}, \underline{x_2}, \cdots, \underline{x_m}\} \subset K$, and ant real number $p_1, \cdots, p_m$ with $p_i \geq 0\ \forall i$, and $\sum_{i=1}^m p_i = 1$. We have $\sum_{i=1}^m p_i \underline{x}_i \in K$. A linear combination with the coefficient settings above is called a convex combination. Thus, a set K is convex iff it is closed under taking convex combinations. Assuming $m = 2$, one can see geometrically that as p_1 and p_2 vary over values, the set of vectors $p_1\underline{x}_1 + p_2\underline{x}_2$ obtained is the line segment between $\underline{x}_1$ and $\underline{x}_2$. Thus, a set K is convex iff for every two points in K, the line segment between the two points is contained in K. Let $f : K \to \mathbb{R}$ where K is a convex subset of $\mathbb{R}^n$. Then f is called a convex function iff $\{\underline{x}_1, \cdots, \underline{x}_m\} \subset K$ and $p_i \geq 0\ \forall i$ and $\sum_{i=1}^m p_i = 1$ implies $f\left(\sum_{i=1}^m p_i \underline{x}_i\right) \leq \sum_{i=1}^m p_i f(\underline{x}_i)$. We could replace the $\geq$ and $\leq$, and we can get a *strictly convex*.

We wish to give an easily checked sufficient condition for convexity of a function, but some definitions are needed first. Suppose $g : \mathbb{R}^n \to \mathbb{R}$ has continuous second order partial derivatives. The Hessian matrix $D^2 f(x) = H(x)$ is given by $H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$ where $f(x + h) = f(x) + Df(x)h + o(||h||)$ as $h \to 0$. $o(||h||)$ stands for a function $c(h)$ s.t. $\lim_{h \to 0} \frac{1}{||h||} c(h) = 0$.

Note that H is actually a mapping of n-vectors to $n \times n$ matrices. If B is an $n \times m$ matrix with (i, j) entry B_{ij}, the the transpose of B, denotes B^T, is an $m \times n$ matrix obtained by interchanging rows and columns, i.e. the (i, j) entry of B^T is B_{ji}. A $n \times n$ matrix A is symmetric iff $A^T = A$, which is the same as $A_{ij} = A_{ji}$ for all i and j. Observe that our assumption of continuity of the second order partial derivatives of f implies equality of mixed partials (i.e. $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$), and hence that the Hessian is symmetric. A symmetric matrix A is called **nonnegative definite (positive semi-definite)** iff $y^T A y \geq 0$ for all n-vectors y. Note

that y^T is an $1 \times n$ matrix, and $y^T Ay = \sum_{i=1}^{n} \sum_{j=1}^{n} j_i A_{ij} y_j$. A symmetric matrix A is called **strictly positive definite** iff $y^T Ay > 0$ for all nonzero n-vectors y.

Partial order on symmetric matrices: $A \succeq B$ iff $A - B$ is nonnegative definite, $A \succ B$ iff $A - B$ is strictly positive definite.

Theorem 2.1.3. *Suppose $f : K \to \mathbb{R}$ where $K \subset \mathbb{R}^n$ is a convex open set and f has twice continuously differentiable on K.*

1. *If the Hessian matrix $H(\underline{x}) \succeq 0$, nonnegative definite for all $x \in K$, then f is convex.*
2. *If the Hessian matrix $H(\underline{x}) \succ 0$, strictly positive definite for all $x \in K$, then f is strictly convex.*

*Fix arbitrary $\underline{x}_0$ and $\underline{x}_1$ in K, and consider $g(p) = (1-p)f(\underline{x}_0) + pf(\underline{x}_1)$ for $p \in (0,1)$. It suffices to check that g is convex or strictly convex, which is a one dimensional problem. This illustrates a common theme in convex analysis: general problems involving convex functions can often be reduced to problems involving functions of a single real variable. A real valued function f defined on a convex set K is called **concave** if $-f$ is convex, and similarly f is strictly concave if $-f$ is strictly convex.*

Example 2.1.1. Some example of convex function

1. $f(\underline{x}) = ||x||^p$, $\underline{x} \in \mathbb{R}^n$, where $p \geq 1$
2. $f(\underline{x}) = x^{-p}$, $x \in (0, \infty$, where $p \geq 0$
3. $f(\underline{x}) = \exp(ax)$, $x \in \mathbb{R}$
4. $f(\underline{x}) = \underline{x}^T Q \underline{x}$, $x \in \mathbb{R}^n$,where Q is nonnegative definite

Some example of strictly concave function

1. $f(\underline{x}) = \log x$, $x \in (0, \infty)$
2. $f(\underline{x}) = ||x||^p$, $\underline{x} \in \mathbb{R}^n$, where $0 \leq p \leq 1$

Definition of the convex function can be interpreted probabilistically. Let $\underline{X}$ be a discrete random n-vector with distribution given by $\mathbb{P}[\underline{X} = \underline{x}_i] = p_i$ i.e. $\text{Law}[\underline{X}] = \sum_{i=1}^{m} p_i \delta_{\underline{x}_i}$. This summation is a probability measure. Then $E[f(\underline{X})] = \sum_{i=1}^{m} p_i f(\underline{x}_i) \geq f(E[\underline{X}]) = f\left(\sum_{i=1}^{m} p_i \underline{x}_i\right)$, and we can get Jensen's Inequality

Theorem 2.1.4. ***Jensen's Inequality:*** *Let f be a convex function on a convex set $K \subset \mathbb{R}^n$ and suppose X is a random n-vector with $E||X|| < \infty$ and $X \in K$ a.s. Then $E[X] \in K$ and $f(E[\underline{X}]) \leq E[f(\underline{X})]$. Furthermore, if f is strictly convex and Law$[\underline{X}]$ is nondegenerate (i.e. $\underline{X}$ is not a.s. equal to a constant, or equivalently Law$[\underline{X}]$ is not a unit point mass), then strict inequality holds in the above.*

2.1.3 Covariance Matrix

Remark. Inner product spaces: a linear space L (over $\mathbb{R}$) such that there is a scalar valued binary operation, denoted $\langle x, y \rangle$, and satisfying:

1. symmetry: $\forall x, y \in L, \langle x, y\rangle = \langle y, x\rangle$
2. bilinearity: $\forall x_1, x_2, y \in L \& \forall a_1, a_2 \in \mathbb{R}, \langle a_1 x_1 + a_2 x_2, y\rangle = a_1\langle x_1, y\rangle + a_2\langle x_2, y\rangle$
3. positivity: $\forall x \in L, x \neq 0$ implies $\langle x, x\rangle = 0$

There is a norm associated with an inner product: $||x|| = \sqrt{\langle x, x\rangle}$. An inner prodcut space which is complete in the norm (all Cauchy sequences converge) is called a Hilbert space

Example 2.1.2. Let $(\Omega, \mathcal{F}, \mu)$ be any measure space. Define

$$L_2(\Omega, \mathcal{F}, \mu) = L_2(\mu) = \left\{ f : \Omega \to \mathbb{R} : \int f^2 d\mu < \infty \right\}$$

Let L be the linear space constructed from L_2 by identifying functions that are equal μ-a.e. The inner product on L is $\langle f, g\rangle = \int fg d\mu$. L_2 is a Hilbert Space.

Theorem 2.1.5. ***Cauchy-Schwarz Inequality:*** *For any r.v. X and Y, $(E[XY])^2 \leq E[X^2]E[Y^2]$. Assume the l.h.s. is finite. Then equality holds iff either $X = 0$ a.s. or $Y = cX$ a.s. for some constant c.*

Theorem 2.1.6. ***Holder's Inequality:*** $E|XY| \leq (E|X|^p)^{1/p} (E|Y|^q)^{1/q}$. *Cauchy-Schwarz is a special case of Holder's Inequality.*

Let $\underline{X} = (X_1, \cdot, X_n)$ be a random n-vector with finite second moments. We have by Holder's Inequality that

$$E|(X_i - \mu_i)(X_j - \mu_j)| \leq E[(X_i - \mu_i)^2]E[(X_j - \mu_j)^2]^{(1/2)}$$

and

$$E[(X_i - \mu_i)2] = Var[X_i] = E[X_i^2] - \mu_i^2 \in [0, \infty)$$

, so $(X_i - \mu_i)(X_j - \mu_j)$ is integrable. Assuming $E[||\underline{X}||^2] < \infty$, we define the covariance matrix $V = Cov[\underline{X}]$ by $V_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)], 1 \leq i, j \leq n$, or, in a more compact matrix notation,

$$V = E[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T].$$

Note that V is an $n \times n$ matrix with real (i.e. finite) entries. In fact, by the above $|V_{ij}| \leq \sqrt{Var[X_i]Var[X_j]}$. Also, one can further check that $Cov[\underline{X}]$ is symmetric and nonnegative definite. If A is any $m \times n$ matrix and $b \in \mathbb{R}^m$, then $Cov[A\underline{X} + \underline{b}] = ACov[\underline{X}]A^T$. If $\underline{X}$ is a random n-vector and $\underline{Y}$ is a random m-vector, then the covariance between $\underline{X}$ and $\underline{Y}$ is

$$Cov[\underline{X}, \underline{Y}] = E[(\underline{X} - E[\underline{X}])(\underline{Y} - E[\underline{Y}])^T]$$

which is an $n \times m$ matrix. $Cov[\underline{X}] = Cov[\underline{X}, \underline{X}]$, and $Cov[\underline{Y}, \underline{X}] = Cov[\underline{X}, \underline{Y}]^T$. If $\underline{Z}$ is the random $n + m$-vector $(\underline{X}, \underline{Y})$, then

$$\begin{bmatrix} Cov[\underline{X}] & Cov[\underline{X}, \underline{Y}] \\ Cov[\underline{Y}, \underline{X}] & Cov[\underline{Y}] \end{bmatrix}$$

If $Cov[\underline{X}, \underline{Y}] = 0$ (where the latter is a matrix of zeroes), then we say $\underline{X}$ and $\underline{Y}$ are uncorrelated. One can show that if $\underline{X}$ and $\underline{Y}$ are independent then they are uncorrelated, provided both have finite second moments, but the converse is false.

Now we introduce some matrix theory which is extremely useful in many areas of statistics. Recall that a square matrix U is called orthogonal iff $U^{-1} = U^T$. Assuming U is $n \times n$, then U is an orthogonal matrix iff the columns of U form an orthonormal basis for $\mathbb{R}^n$. A square matrix D is *diagonal* if the off diagonal entries are zero, i.e. $D_{ij} = 0$ if $i \neq j$. It will be convenient to write $D = diag[d]$, where d is the vector of diagonal entries.

Theorem 2.1.7. ***Spectral Decomposition of a Symmetric Matrix:*** *Let A be a symmetric matrix. Then there is an orthogonal matrix U and a diagonal matrix Λ s.t. $A = U\Lambda U^T$, where $\Lambda = diag(\lambda_1, \cdots, \lambda_n), \lambda_1 \geq \cdots \geq \lambda_n$ being the eigenvalues of A, and $U = [u_1, \cdots, u_n]$, with $u_1, \cdots, u_n$ being corresponding unit eigenvectors.*

Proposition 2.1.8. *Suppose X is a random n-vector with finite second moments. Then there is an orthogonal matrix U such that $Cov[U^T X]$ is a diagonal matrix.*

Proof. Since $V = Cov[X]$ is symmetric there is an orthogonal matrix U and a diagonal matrix Λ such that $V = U\Lambda U^T$. Then multiplying on the left by U^T and on the right by U and using the defining property of an orthogonal matrix, $\Lambda = U^T V U$. The result now follows with $A = U$. □

Assume $\underline{X}$ is a random n-vector with finite second moments and put $\mu = E[\underline{X}]$, $V = Cov[\underline{X}]$. Write $V = U\Lambda U^T$, where U is orthogonal and Λ is diagonal, as in the proof of the last result. Since V is nonnegative definite, the eigenvalues (which are the diagonal entries of Λ) are nonnegative. Assume that the number of positive eigenvalues is r, so there are $n - r$ zero eigenvalues. The **null space** of V (which is defined to be the set of vectors $\underline{x}$ such that $V\underline{x} = \underline{0}$) is given by

$$\mathbf{N}(V) = span[\underline{u}_{r+1}, \cdots, \underline{u}_n]$$

and the **column space or range** (which is defined to be $\{V\underline{x} : \underline{x} \in \mathbb{R}^n\}$) is given by

$$\mathbf{R}(V) = span[\underline{u}_1, \cdots, \underline{u}_r] = \left\{ \sum_{i=1}^{r} a_i \underline{u}_i : a_i \in \mathbb{R}, \forall i \right\}$$

Here, the *span* of a collection of vectors is the set of all linear combinations of the given collection, i.e. the smallest linear subspace which includes the given collection. Also, $r = rank(V)$, the dimension of the range of V , is known as the *rank* of the linear transformation V, and also the number of positive eigenvalues. Null and vector spaces follow since any $\underline{x} \in \mathbb{R}^n$ may be expanded as $\underline{x} = \sum_{i=1}^{n} (\underline{x}^T \underline{u}_i)\underline{u}_i$ because $\{\underline{u}_i : 1 \leq i \leq n\}$ form an orthonormal basis for $\mathbb{R}^n$. Thus,

$$V\underline{x} = \sum_{i=1}^{n} (\underline{x}^T \underline{u}_i \underline{x}^T \underline{u}_i) V\underline{u}_i = \sum_{i=1}^{n} \lambda_i (\underline{x}^T \underline{u}_i)\underline{u}_i = \sum_{i=1}^{r} \lambda_i (\underline{x}^T \underline{u}_i)\underline{u}_i \tag{2.1}$$

Thus, $V\underline{x} = 0$ iff $\underline{x}^T \underline{u}_i = 0$ for $1 \leq i \leq r$, which is true iff $\underline{x} \in span[\underline{u}_{r+1}, \cdots, \underline{u}_n]$. Also, $\underline{y} = V\underline{x}$ for some $\underline{x} \in \mathbb{R}^n$ iff $\underline{y}$ has the form of the last expression in (2.1), which us true that iff $\underline{y} \in span[\underline{u}_1, \cdots, \underline{u}_r]$. Note that in this latter case we may take $\underline{y} = V\underline{x}$ where $\underline{x} = \sum_{i=1}^{r} \lambda_i^{-1} (\underline{y}^T \underline{u}_i)\underline{u}_i = V^- \underline{y}$

Here, the last equation defines the linear transformation V^-, which is known as the **Moore-Penrose generalized inverse** of V . Note that $V^- \underline{y}$ is just one of infinitely many $\underline{x}$'s satisfying $\underline{y} = V\underline{x}$ when $rank(V) < n$. If $rank(V) = n$, i.e. V is nonsingular, then $V^- = V^{-1}$.

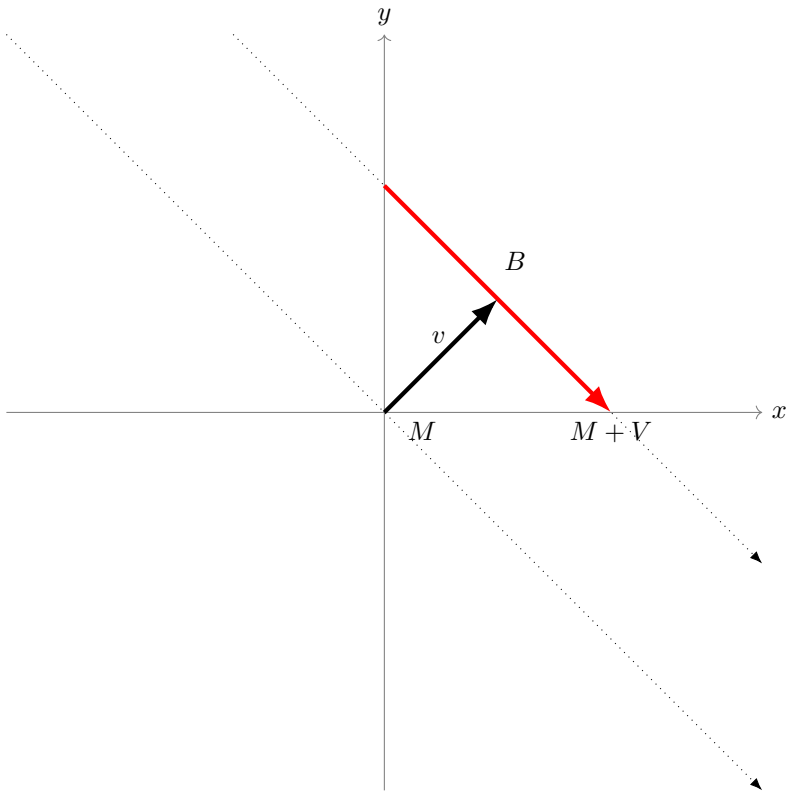
Proposition 2.1.9. *If $\underline{X}$ is a random n-vector with $E[||\underline{X}||^2] < \infty$, $\underline{\mu} = E[\underline{X}]$, and $V = Cov[\underline{X}]$, then*

$$\mathbb{P}[\underline{X} \in \mathbf{R}(V) + \underline{\mu}] = 1$$

where $\mathbf{R}(V) + \underline{\mu} = \{\underline{(y)} + \underline{\mu} : \underline{(y)} \in \mathbf{R}(V)\}$

Proof. Let $\underline{Y} + \underline{X} - \underline{\mu}$, so $\underline{Y} \in \mathbf{R}(V)$ iff $\underline{X} \in \mathbf{R}(V) + \underline{\mu}$. Write $\underline{Y} = \sum_{i=1}^{n} Y_i \underline{u}_i$, $Y_i = \underline{Y}^T \underline{u}_i$. Then $Cov[\underline{X}]$ =Cov$[\underline{Y}]$ and $E[Y_i^2] = \underline{u}_i^T V \underline{u}_i = \lambda_i$. Hence, $E[Y_i^2] = 0$ iff $i > r$, which is equivalent to $Y_i = 0$ a.s. iff $i > r$, hence $\underline{Y} = \sum_{i=1}^{r} Y_i \underline{u}_i$ a.s. , which implies $\underline{Y} \in \mathbf{R}(V)$ □

Proposition 2.1.10. ***Linear constraint in*** $\mathbb{R}^n$***:*** *$a \in \mathbb{R}^n$ is fixed, $b \in \mathbb{R}$ is fixed, and $x \in \mathbb{R}^n$ variable, $a^T x = b$. $B \subseteq \mathbb{R}^n$, B satisfies a linear constraint iff $\exists a \in \mathbb{R}^n, b \in \mathbb{R}$ s.t. $\forall x \in B$, $a^T x = b$. $\exists n-1$ dimensional subspace of M s.t. $B \subseteq V + M = \{v + x : x \in M\}$. e.g.*



Claims if B satisfies a linear constraint $a^T x = b, a \neq 0$, then $m^n(B) = 0$

$$m^n(B) = \int_{\mathbb{R}^n} I_B(\underline{x}) d\underline{x} = \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} I_B(x_1, \underline{y}) dm(x_1) dm(\underline{y}).$$

$a_1 x_1 = b - \sum_{i=2}^{n} a_i x_i$, $x_i = a_i^{-1}(b - \sum_{i=2}^{n} a_i x_i$, $\int_{\mathbb{R}} I_B(x_1, \underline{y}) dm(x_1) = 0$. *That is, if $Cox[\underline{X}] = 0$, then we can take ant eigenvector of $Cov[\underline{X}]$ with eigenvalue 0, say u_n, $u_n^T(X - \mu_X) = 0$ a.s. Thus, $u_n^T X = u_n^T \mu_X)$ is the linear constraint a.s. and $P_X \not\ll m^n$*

Proposition 2.1.11. *Let $\underline{X}$ be as in the previous proposition. If $P_X = Law[X] \ll m^n$ then $rank[Cov[\underline{X}]] = n$*

Proof. If $rank[V] = r < n$, then $R(V)$ is a proper linear subspace of $\mathbb{R}^n$, and $R(V) + \underline{\mu}$ is a proper linear manifold, i.e. a translate of a proper linear subspace. Such a set is closed, hence a Borel set, and we claim its Lebesgue measure is 0. We have that $R(V) + \underline{\mu} \subset \{\underline{x} \in \mathbb{R}^n : (\underline{x} - \underline{\mu})^T \underline{u}_n = 0\} \equiv B$. Applying Fubini's theorem and the fact that $m^n = m^{n-1} \times m$ by definition, we have,

$$m^n(B) = \int_{\mathbb{R}^n} I_B(\underline{x}) d\underline{x} = \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} I_B(\underline{y}, x_n) dx_n d\underline{y}$$

where $\underline{y} = (x_1, \cdots, x_{n-1})$. Now $I_B(\underline{y}, x_n) = 1$ iff $(x_n - \mu_n)u_{nn} = -\sum^{n-1} -i = 1(y_j - \mu_j)u_{nj}$ where $\underline{u}_m = (u_{1n}, \cdots, u_{nn})$ Assuming $u_{nn} \neq 0$, then for fixed $\underline{y} \in \mathbb{R}^{n-1}$, $I_B(\underline{y}, x_n) \neq 0$ iff $x_n = \mu_n - u_{nn}^{-1} \sum_{i=1}^{n-1} (y_i - \mu_i)u_{in}$ which is only a single point. Hence, the inner integral $\int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} I_B(\underline{y}, x_n) dx_n d\underline{y}$ is 0. If $u_{nn} = 0$, n is nonzero, then some component of $\underline{u}_n$ is nonzero (since $\underline{u}_n$ is an element of an orthornormal basis for $\mathbb{R}^n$), say $u_{nj} \neq 0$, and then replace x in the integral with u_{nj} and $\underline{y}$ with the remaining components of $\underline{x}$. □

2.2 Characteristic and Moment Generating Functions.

2.2.1 General Results

Definition 2.2.1. The characteristic function of a random vector $\underline{X}$ is the complex valued function$\phi_{\underline{X}} : \mathbb{R}^n \to \mathbb{C}$ given by

$$\phi_{\underline{X}}(\underline{u}) = E[\exp(i\underline{u}^T \underline{X})] = E[\cos(\underline{u}^T \underline{X})] + iE[\sin(\underline{u}^T \underline{X})] = \psi_{\underline{X}}(i\underline{u})$$

. The moment generating function is given by

$$\psi_{\underline{X}}(\underline{u}) = E[\exp(\underline{u}^T \underline{X})], \underline{u} \in \mathbb{R}^n$$

We say that the m.g.f. exists in a neighborhood of $\underline{0}$ iff there is an $\epsilon > 0$ s.t. $\psi_{\underline{X}}(\underline{u}), \forall \underline{u} \in \mathbb{R}^n$ s.t. $||\underline{u}|| < \epsilon$. Then it uniquely identifies P_X

The ch.f. is defined and finite for all $\underline{u} \in \mathbb{R}^n$, since it is the expectation of a bounded continuous function (or more simply, its real and imaginary components are bounded and continuous). In fact, $|\phi_{\underline{X}}(\underline{u})| \leq 1$ for all $\underline{u} \in \mathbb{R}^n$ since $|\exp(it)| \leq 1$ for all $t \in \mathbb{R}$. The m.g.f. is defined for all $\underline{u}$ but may be ∞ everywhere except $\underline{u} = \underline{0}$. Many of the results for ch.f.'s given in Chung or Ash for r.v.'s carry over to random vectors as well, and also to the m.g.f. Some of the results of most interest to us are in the next proposition. Further discussion and proofs may be found in Billinsgley, pp. 352-356.

Theorem 2.2.1. *Let $\underline{X}$ be a random n-vector with ch.f and m.g.f.*

1. ***(Continuity):*** *ϕ is uniformly continuous on $\mathbb{R}^n$, and ψ is continuous at every point $\underline{u}$ s.t. $\psi(\underline{v}) < \infty$ for all $\underline{v}$ in a neighborhood of $\underline{u}$.*

2. ***(Relation to moments):*** *If $\underline{X}$ is integrable, then the gradient*

$$\nabla\phi = \left(\frac{\partial\phi}{\partial u_1}, \cdots, \frac{\partial\phi}{\partial u_n}\right)$$

is defined at $\underline{u} = \underline{0}$ and equals $iE[\underline{X}]$. Also, $\underline{X}$ has finite second moments iff the Hessian $D^2\phi(\underline{u}) = H(\underline{u})$ of ϕ exists at $\underline{u} = \underline{0}$ and then $H(\underline{0}) = -E[\underline{X}\underline{X}^T]$. If ψ is finite in a neighborhood of $\underline{0}$, then $E[||\underline{X}||^p] < \infty$ for all $p \geq 0$. Further, $\nabla\psi(\underline{0}) = E[\underline{X}]$, and $D^2\psi(\underline{0}) = E[\underline{X}\underline{X}^T]$.

3. ***(Linear Transformation Formulae):*** *Let $\underline{Y} = A\underline{X} + \underline{b}$ for some $m \times n$ matrix A and some m-vector $\underline{b}$. Then for all $\underline{v} \in \mathbb{R}^m$,*

$$\phi_{\underline{Y}}(\underline{v}) = \exp(i\underline{v}^T\underline{b})\phi_{\underline{X}}(A^T\underline{v})$$
$$\psi_{\underline{Y}}(\underline{v}) = \exp(\underline{v}^T\underline{b})\psi_{\underline{X}}(A^T\underline{v})$$

4. ***(Uniqueness):*** *If $\underline{Y}$ is a random n-vector and if $\phi_{\underline{X}}(\underline{u}) = \phi_{\underline{Y}}(\underline{u})$ for all $\underline{u} \in \mathbb{R}^n$, then Law[$\underline{X}$] =Law[$\underline{Y}$]. If both $\psi_{\underline{X}}$ and $\psi_{\underline{Y}}$ are defined and equal in a neighborhood of $\underline{0}$, then Law[$\underline{X}$] =Law[$\underline{Y}$]*

Proof. (a) The result follows from the following inversion formula whose proof can be found, for example, in Billingsley (1986, p. 395): for any $a = (a_1, \cdots, a_k) \in \mathbb{R}^k, b = (b_1, \cdots, b_k) \in \mathbb{R}^k$, and $(a, b] = (a_1, b_1] \times \cdots \times (a_k, b_k]$ satisfying Law[$\underline{X}$] (the boundary of $(a, b]$) $= 0$,

$$P_X((a,b]) = \lim_{c\to\infty} \int_{-c}^{c} \cdots \int_{-c}^{c} \frac{\phi_X(t_1, \cdots, t_k)}{(-1)^{k/2}(2\pi)^k} \prod_{i=1}^{k} \frac{e^{-it_i a_i} - e^{-it_i b_i}}{t_i} dt_i$$

(b) First consider the case of $k = 1$. From $e^{s|x|} \le e^{sx} + e^{-sx}$, we conclude that $|X|$ has an m.g.f. that is finite in the neighborhood $(-c, c)$ for some $c > 0$ and $|X|$ has finite moments of all order. Using the inequality $|e^{itx}[e^{iax} - \sum_{j=0}^{n}(iax)^j/j!]| \le \frac{|ax|^{n+1}}{(n+1)!}$, we obtain that

$$\left| \phi_X(t+a) - \sum_{j=0}^{n} \frac{a^j}{j!} E[(iX)^j e^{itX}] \right| \le \frac{|a|^{n+1} E|X|^{n+1}}{(n+1)!}$$

For any $t \in \mathbb{R}$,

$$\phi_X(t+a) = \sum_{j=0}^{\infty} \frac{\phi_X^{(j)}(t)}{j!} a^j, |a| < c \tag{2.2}$$

Similarly, (2.2) holds with ϕ_X replaced by ϕ_Y. Under the assumption that $\psi_X = \psi_Y < \infty$ in a neighborhood of 0, X and Y have the same moments of all order. $\phi_X^{(j)}(0) = \phi_Y^{(j)}(0)$ for all $j = 1, 2, \cdots$, which and (2.2) with $t = 0$ imply that ϕ_X and ϕ_Y are the same on the interval $(-c, c)$ and hence have identical derivatives there. Considering $t = c - \epsilon$ and $-c + \epsilon$ for an arbitrarily small $\epsilon > 0$ in (2.2) shows that ϕ_X and ϕ_Y also agree on $(-2c + \epsilon, 2c - \epsilon)$ and hence on $(-2c, 2c)$. By the same argument ϕ_X and ϕ_Y are the same on $(-3c, 3c)$ and so on. Hence, $\phi_X(t) = \phi_Y(t)$ for all t and, by part (a) ,$P_X = P_Y$. Consider now the general case of $k \ge 2$. If $P_X \ne P_Y$, then by part (a) there exists $t \in \mathbb{R}^k$ such that$\phi_X(t) \ne \phi_Y(t)$. Then $\phi_{t^T X}(1) \ne \phi_{t^T Y}(1)$, which implies that $P_{t^T X} \ne P_{t^T Y}$. But $\psi_X = \psi_Y < \infty$ in a neighborhood of $0 \in \mathbb{R}^k$ implies that $\psi_{t^T X} = \psi_{t^T Y} < \infty$ in a neighborhood of $0 \in \mathbb{R}$ and, by the proved result for $j = 1$, $P_{t^T X} = P_{t^T Y}$ This contradiction shows that $P_X = P_Y$

□

5. ***Ch.f. for sums of independent r.v.'s:*** *Suppose $\underline{X}$ and $\underline{Y}$ are independent random p-vectors and let $\underline{Z} = \underline{X} + \underline{Y}$. Then $\phi_{\underline{Z}}(\underline{u}) = \phi_{\underline{X}}(\underline{u})\phi_{\underline{Y}}(\underline{u})$*

Proof. (b) Fort he second part of (b), assume the m.g.f. is defined in a neighborhood of $\underline{0}$, say $\psi(\underline{u}) < \infty$ for $||\underline{u}|| < \epsilon$. If suffices to prove the result for $p \ge 2$. For $E[||\underline{X}||^p] < \infty$ when $p \ge 2$, it suffices for $E[||\underline{X}_i||^p] < \infty$, $1 \le i \le n$, since by convexity

$$||\underline{X}||^p = n^{\frac{p}{2}} \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right)^{\frac{p}{2}} \le n^{\frac{p}{2}} \frac{1}{n} \sum_{i=1}^{n} (X_i^2)^{\frac{p}{2}} = n^{\frac{p}{2}} \frac{1}{n} \sum_{i=1}^{n} |X_i|^p$$

Now $\psi_{X_i}(\pm\frac{\epsilon}{2}) = E[\exp(\pm\epsilon X_i/2)] < \infty$ by taking $u = (\pm\epsilon/2, 0, 0, \cdots, 0)$. Since exponential functions "grow faster" than power functions, there is some $M > 0$ such that $|x|^p \le \exp(\epsilon|x|/2)$ for all $|x| > M$.

Hence,

$$\begin{aligned} E[|X_i|^p] &= \int_{-\infty}^{\infty} |x|^p dP_{X_i}(x) \\ &\leq \int_{-\infty}^{-M} \exp[-\epsilon|x|/2] dP_{X_i}(x) + \int_{-M}^{M} |x|^p dP_{X_i}(x) + \int_{M}^{\infty} \exp[\epsilon|x|/2] dP_{X_i}(x) \\ &\leq \psi_{X_i}(-\epsilon/2) + M^p + \psi_{X_i}(\epsilon/2) < \infty \end{aligned}$$

We show that $\partial\psi/\partial u_1$ exists and can be computed by differentiation under the integral sign. (An extension of this argument will show that ψ has partial derivatives of all orders on the interior of $\{u : \psi(u) < \infty\}$, and can be computed by differentiation under the integral sign.) For simplicity, assume $n = 2$. Fix u_2 and let $\delta = \sqrt{\epsilon^2 - u_2^2}$. Then for $|u_1| < \delta$, $||\underline{u}||^2 = u_1^2 + u_2^2 < \epsilon^2$. Now take any $\delta_0 < \delta$ and $\delta_1 \in (\delta_0, \delta)$. Put $g(\underline{x}, u_1) = \exp[u_1 x_1 + u_2 x_2]$. Then $\dfrac{\partial}{\partial u_1} g(\underline{x}, u_1) = x_1 \exp[u_1 x_1 + u_2 x_2]$. Since the exponential $\exp[(\delta_1 - \delta_0)|x_1|]$ "grows faster" than $|x_1|$ as $|x_1| \to \infty$, there is a constant $M > 0$ s.t. $|x_1| \leq M \exp[(\delta_1 - \delta_0)|x_1|]$ for all x_1. Also, if $|u_1| < \delta_0$, then $0 < \exp[u_1 x_1] < \exp[\delta_0 |x_1|]$. Combining these last two estimates we have

$$\begin{aligned} \left| \frac{\partial}{\partial u_1} g(\underline{x}, u_1) \right| = |x_1 \exp[u_1 x_1 + u_2 x_2]| &\leq (M \exp[(\delta_1 - \delta_0)|x_1|]) \exp[\delta_0 |x_1|] \exp[u_2 x_2] \\ &\leq G(\underline{x}) = M(\exp(\delta_1 x_1 + u_2 x_2) + \exp(-\delta_1 x_2 + u_2 x_2)) \end{aligned}$$

We have used the fact that $e^{a|t|} \leq e^{at} + e^{-at}$ for all $a > 0$ and all $t \in \mathbb{R}$in choosing a dominating function G. Since $\delta_0 < \epsilon$ and $\delta_1^2 + u_2^2 < \epsilon^2$, we have $\int_{\mathbb{R}^2} G(\underline{x}) dP_{\underline{X}}(\underline{x}) = M[\psi(\delta_1, u_2) + \psi(-\delta_1, u_2)] < \infty$. By interchange of differentiation and integral, $\dfrac{\partial}{\partial u_1}\psi(\underline{u}) = \dfrac{\partial}{\partial u_1} \int g(\underline{x}, u_1) dP_{\underline{X}}(\underline{x}) = \int \dfrac{\partial}{\partial u_1} g(\underline{x}, u_1) dP_{\underline{X}}(\underline{x}) = \int x_1 \exp[u_1 x_1 + u_2 x_2] dP_{\underline{X}}(\underline{x})$. Hence, $\dfrac{\partial \psi}{\partial u_1} \mid_{\underline{u}=0} = \int x_1 dP_{\underline{X}}(\underline{x}) = E[X_1]$. This shows one component of the equation $\nabla\psi(\underline{0}) = E[\underline{X}]$, and the others follow similarly. A similar argument shows $\dfrac{\partial^2 \psi}{\partial u_i \partial u_j} = \int x_i x_j \exp[\underline{u}^T \underline{x}] dP_{\underline{X}}(\underline{x})$, so the Hessian at $\underline{u} = \underline{0}$ is $E[\underline{X}\underline{X}^T]$ □

A slight extension of the argument above can be used to prove the following theorem, which has many applications in statistics. If $z = x + iy$ is a complex number with $x \in \mathbb{R}$ and $y \in \mathbb{R}$, then $x = \text{Real}[z]$ and $y = \text{Imag}[z]$ are called the real and imaginary parts, respectively. If $\underline{z} \in \mathbb{C}^n$, i.e. $\underline{z}$ is an n-tuple of complex numbers (or an n–vector with complex components), say $\underline{z} = (z_1, \cdots, z_n)$, then $\text{Real}[\underline{z}] = (\text{Real}[z_1], \cdots, \text{Real}[z_n])$ is the vector of real parts, and similarly for $\text{Imag}[\underline{z}]$. Recall that for $D \subset \mathbb{R}^n$, the interior of D is $\text{int}[D] = \{\underline{x} \in D : B(\underline{x}, \epsilon) \subset D, \epsilon > 0\}$, i.e. the points $\underline{x}$ in D for which an entire neighborhood $B(\underline{x}, \epsilon)$ of $\underline{x}$ (otherwise known as an ϵ ball centered at $\underline{x}$) is contained in D. One can easily show that $\text{int}[D]$ is the largest open subset of D.

Now we briefly review some complex analysis. A complex valued function g of a complex variable (i.e. $g : \mathbb{C} \to \mathbb{C}$) is analytic at $z \in \mathbb{C}$ iff it is differentiable in a neighborhood of z. One remarkable result from complex analysis is that a function which is analytic in an open set of $\mathbb{C}$ is in fact infinitely differentiable in that open set. (See e.g. Ahlfors, Complex Analysis, pp. 120-122.) (Here, an open subset of $\mathbb{C}$ is the same as an open subset of $\mathbb{R}^2$ when we identify $\mathbb{C}$ with $\mathbb{R}^2$ via $x + iy \to (x, y)$. We will mainly consider a "strip" of the form $\{x + iy : -\epsilon < x < \epsilon, -\infty < y < \infty\} = \{z \in \mathbb{C} : |Real(z)| < \epsilon\}$

Theorem 2.2.2. *Suppose $f : \Omega \to \mathbb{C}$ is any bounded Borel function on a measure space $(\Omega, \mathcal{F}, \mu)$. Let $\underline{T} : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}_n)$ and let $\underline{\theta} \in \mathbb{C}^n$. Let*

$$B(\underline{\theta}) = \int_{\Omega} f(\omega) \exp[\underline{\theta}^T \underline{T}(\omega)] d\mu(\omega)$$

For $1 \leq j \leq n$ and $(\xi_1, \cdots, \xi_{j-1}, \xi_{j+1}, \cdots, \xi_n) \in \mathbb{R}^j \times \mathbb{R}^{n-j-1}$, define the set

$$W_j(\xi_1, \cdots, \xi_{j-1}, \xi_{j+1}, \cdots, \xi_n) = \{\xi_j \in \mathbb{R} : \int_\Omega f(\omega) \exp[\underline{\xi}^T \underline{T}(\omega)] d\mu(\omega) < \infty\}$$

where $\underline{\xi} = (\xi_1, \cdots, \xi_{j-1}, \xi_j, \xi_{j+1}, \cdots, \xi_n) \in \mathbb{R}^n$ in the above. If $\theta_k \in W$ are fixed for $k \neq j$, then B is an analytic function in $\{\theta_j : Real[\theta_j] \in int[W_j]\}$, where W_j is sa given above with $\xi_k = Real[\theta_k]$ for $k \neq j$. Further, any order partial derivative of B can be computed by differentiation under the integral sign

Remark. 1. The fact that $\text{Real}[\theta_j] \in \text{int}[W_j]$ allows us to use a dominating function as in the proof of Theorem 2.2.1 (b) above.

2. Another remarkable fact from complex analysis is the following: Suppose f and g are both analytic functions in the open strip $\{z \in \mathbb{C} : |Real(z)| < \epsilon\}$, and that $\{z_n : n \in \mathbb{N}\}$ is an infinite sequence of distinct values which converges to a limit in the strip, say $z_n \to z$ with $|Real(z)| < \epsilon$. Then if $g(z_n) = f(z_n)$ for all n, we have $f = g$ everywhere on the strip. Now suppose X is a r.v. with $\psi_X(u) < \infty$ for all $|u| < \epsilon$. Then ψ_X can be extended to an analytic function in the strip $\{z : |Real(z)| < \epsilon\}$, which contains the imaginary axis. (This is an example of analytic continuation, which is discussed at length in Ahlfors, p. 285 ff.) Hence, $\phi_X(u) = \psi_X(iu)$ by the previous theorem. Note that under these conditions, it is possible to obtain a stronger uniqueness condition than in Theorem 2.2.1 (d), namely if both $\psi_X(u) < \infty$ and $\psi_Y(u) < \infty$ for all $|u| < \epsilon$, and $\psi_X(z_n) = \psi_Y(z_n)$ for any distinct sequence of complex numbers in the strip $\{z : |Real(z)| < \epsilon\}$ with a limit in that strip, then $Law[X] = Law[Y]$.

3. Another useful fact about analytic functions is that they can be expanded in power series, i.e. suppose g is a complex function of a complex variable and $\rho > 0$ is such that g is analytic in the disk (or "ball") $\{z \in \mathbb{C} : |z - z_0| < \rho\}$. Then

$$g(z) = \sum_{j=0}^{\infty} \frac{1}{j!} g^{(j)}(z_0)(z - z_0)^j, \text{ for } |z - z_0| < \rho$$

Further, derivatives of g may be computed by differentiating under the summation sign, for $|z - z_0| < \rho$. Using this fact along with Theorem 2.2.2, one can show that if X is a r.v. with $\psi_X(u) < \infty$ for all $|u| < \epsilon$, then

$$\psi_X(u) = \sum_{r=0}^{\infty} \frac{d^r \psi_X}{du^r}(0) \frac{1}{r!} u^r = \sum_{r=0}^{\infty} \frac{1}{r!} E[X^r] u^r$$

Thus, we can read off the moments of X from the power series expansion of the m.g.f.

Now we consider the multivariate case with a random n-vector $\underline{X}$. First, we will introduce some notations that will make it easier to present the material. An n–vector $\underline{r}$ with nonnegative integer components is called a **multi–index** i.e.$\underline{r} = (r_1, r_2, \cdots, r_n)$ with each $r_i \in \mathbb{N}$. We can use a "vector" exponential notation for a monomial as in

$$\underline{x}^{\underline{r}} = \prod_{j=1}^{n} x_j^{r_j}$$

where $\underline{x} \in \mathbb{R}$. Thus, by analogy with the univariate case, we may call

$$\mu_{\underline{r}} = E[\underline{X}^{\underline{r}}] = E\left[\prod_{j=1}^{n} x_j^{r_j}\right]$$

the r-th *moment* of the random n–vector $\underline{X}$. The "multi-index" factorial is defined by

$$\underline{r}! = \prod_{j=1}^{n} r_j!$$

The *order* of the multi-index $\underline{r}$ is

$$|\underline{r}| = \sum_{j=1}^{n} r_j$$

We can also define an $\underline{r}$-th *order derivative* by

$$D^{\underline{r}} = \frac{\partial^{|\underline{r}|}}{\prod_{j=1}^{n} \partial u_j^{r_j}}$$

Note that this is a partial differential operator of order $|\underline{r}|$. With these notations, one can show that the power series expansion about 0 for a complex function g of n complex variables which is analytic in each variable is given by

$$g(\underline{z}) = \sum_{\underline{r}} \frac{1}{\underline{r}!} D^{\underline{r}} g(\underline{0}) \underline{z}^{\underline{r}}$$

Thus, if X is a random n-vector with $\psi_{\underline{X}} < \infty$ for all $||\underline{u}|| < \epsilon$, then

$$g(\underline{z}) = \sum_{\underline{r}} \frac{1}{\underline{r}!} D^{\underline{r}} \psi_{\underline{X}}(\underline{0}) \underline{u}^{\underline{r}} = \sum_{\underline{r}} \frac{1}{\underline{r}!} E[\underline{X}^{\underline{r}}] \underline{u}^{\underline{r}}$$

where the series converges in a neighborhood of $u = 0$.

4. Let $\underline{X}$ and $\psi_{\underline{X}}$ be as in part (b). Consider the **cumulant generating function** given by

$$K(\underline{u}) = \log \psi_{\underline{X}}(\underline{u})$$

Then the $\underline{r}$-th cumulant of $\underline{X}$ is

$$\kappa_{\underline{r}} = \frac{\partial^{|\underline{r}|} K}{\prod_{j=1}^{n} \partial u_j^{r_j}}(\underline{0}) = D^{\underline{r}} K(\underline{0})$$

One can show by comparison of the terms of the power series that if $n = 1$, then

$$\kappa_0 = 0, \kappa_1 = E[X], \kappa_2 = Var[X]$$

For higher dimensional random vectors, we still have $\kappa_{\underline{0}} = 0$, and

$$E[X_i] = \kappa_{\underline{r}}, \text{ with } r_i = 1 \text{ and } r_j = 0 \text{ if } j \neq i.$$

Also, if $V = Cov[\underline{X}]$, then

$$V_{ij} = \kappa_{\underline{r}}, \text{ with } r_i = r_j = 1 \text{ and } r_k = 0 \text{ if } k \neq i \text{ or } k \neq j$$

2.3 Common Distributions Used in Statistics

2.3.1 Statistical Models:

We assume our data (typically a matrix) are obtained as a realization of a r.v. X, called the **observable**, taking values in a space Ξ (typically a finite dimensional vector space). The **model** is a family of probability measures $\mathbb{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, which are the possible distributions for X. Here, Ξ is called the observation space. If, for example, X is an $n \times m$ matrix of n observations with m variables each, then $\Xi = \mathbb{R}^{(n\times m)}$ (Note our notation for the linear space of $n \times m$ matrices.) We will often treat it as an nm dimensional vector, e.g., by "stacking" the columns or rows. Θ is called the **parameter space**. In fact, this framework includes so-called nonparametric models. For example, if $X_1, \cdots, X_n$ are assumed to be i.i.d. univariate with unknown c.d.f. F, then we may take Θ as the family of c.d.f.s and P_θ is the n-fold measure product of the distribution determined by the c.d.f. θ. Think of θ as just a label for a possible distribution of X.

Definition 2.3.1. With this kind of mapping, $\theta \to \mathbb{P}_\theta$, called **parametrization**, we will generally require our model to be **identifiable** by which we mean that if $\theta_1 \neq \theta_2$, then $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$.

With this formulation, we think of a true parameter value θ, i.e. the one which actually generates the data. We will typically denote this as θ^*. Then "statistical inference" is about making inferences about θ^*.

Note that $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$ means for some measurable set A, $\mathbb{P}_{\theta_1}(A) \neq \mathbb{P}_{\theta_2}(A)$. In general, we want to use only identifiable parameterizations. If the parameter is not identifiable there will be differences in parameter values which are not statistically meaningful since we cannot determine them from the distribution of the observable. In general, if we have a nonidentifiable parameterization, we will reparametrize to obtain identifiability.

Definition 2.3.2. We call a model a **dominated family** if there is a σ-finite measure μ such that for all $\theta \in \Theta$, $P_\theta \ll \mu$. We denote the densities (w.r.t. μ) by $f_\theta(x)$ or $f(x|\theta)$ equals to $\frac{d\mathbb{P}_\theta}{d\mu}(x)$. If we treat the parameter θ as a realization of a r.v., denoted ϑ, as suggested by the notation $f(x|\theta)$, then P_θ is (a version of) the conditional distribution of X given $\vartheta = \theta$.

If x is the observed value of X, then $L(\theta|x) = f(x|\theta)$ is the likelihood. Much of statistical inference is based on likelihoods. Most of the methods we will develop depend on the likelihood - that is, data sets that give the same likelihood will give the same inference. Often the data will take the form (x_i, y_i) where we are interested in developing a "predictive model" which is a function of x that can be used to predict a y when we observe a new x-value. In this case, the statistical model will typically treat the y's as random and the x's as nonrandom (or having a degenerate distribution at the observed value within each pair). This is the framework of regression modeling in statistics. We will then think of our inference problem as not about the distribution of the Y's but the conditional distribution of a Y given $X = x$.

2.3.2 Exponential Families

Definition 2.3.3. A dominated family $\mathbb{P} = \{\mathbb{P}_\theta : \theta \in \Theta\} \ll \mu$ with μ σ–finite is called an exponential family iff the densities w.r.t. μ can be written in the form

$$f_\theta(x) = \exp[\underline{\eta}(\theta)^T \underline{T}(x) - B(\theta)]h(x)$$

where

$$\begin{aligned}
&\underline{\eta} : \Theta \to \mathbb{R}^p \\
&B : \Theta \to \mathbb{R} \\
&\underline{T} : (\xi, \mathcal{G}) \to (\mathbb{R}^p, \mathcal{B}_n) \\
&h : (\xi, \mathcal{G}) \to (\mathbb{R}, \mathcal{B})
\end{aligned}$$

and $B(\theta)$ is the **normalizing constant** in logarithmic form $B(\theta) = \log \int_\Xi \exp[\underline{\eta}(\theta)^T \underline{T}(x)] h(x) d\mu(x)$

Example 2.3.1. Let $\mathbb{P}$ be the normal family on $\mathbb{R}$, $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$. Then $\mu = m$, Lebesgue measure, may be taken as the dominating measure and this is σ–finite. The density may be written in the form

$$\begin{aligned}
f_{\mu,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \\
&= \exp\left[-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right]\left(\frac{1}{\sqrt{2\pi}}\right)
\end{aligned}$$

where

$$\begin{aligned}
\underline{\eta}(\mu, \sigma^2) &= \left(\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right) \\
B(\mu, \sigma^2) &= \frac{\mu^2}{2\sigma^2} + \log\sigma \\
\underline{T}(\mu, \sigma^2) &= (x^2, x) \\
h(x) &= \frac{1}{\sqrt{2\pi}}
\end{aligned}$$

Definition 2.3.4. Given $\underline{\mu} \in \mathbb{R}^n$ and V an $n \times n$ nonnegative definite matrix, let $N(\underline{\mu}, V)$ denote the Borel probability measure on $\mathbb{R}^n$ with m.g.f $\psi(\underline{u} = \exp\left[\underline{\mu}^T \underline{u} + \frac{1}{2}\underline{u}^T V \underline{u}\right]$

Example 2.3.2. (Gamma Distribution): Gamma(α, β), $\alpha, \beta > 0$, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^x dx$

$$\begin{aligned}
f_{\alpha,\beta}(x) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) \\
&= \exp\left\{\alpha\log(x) + \frac{-1}{\beta} - [\alpha\log\beta + \log\Gamma(\alpha)]\right\}[x^{-1} I_{(0,\infty)}(x)]
\end{aligned}$$

$$\begin{aligned}
\underline{\eta}(\alpha, \beta) &= \left(\alpha, \frac{-1}{\beta}\right) \\
B(\alpha, \beta) &= \alpha\log\beta + \log\Gamma(\alpha) \\
\underline{T}(x) &= (x, \log x) \\
h(x) &= x^{-1} I_{(0,\infty)}(x)
\end{aligned}$$

Gamma$(1, \beta)$ is exponential distribution $Exp(\beta)$, also an exponential subfamily of the Gamma family.

Remark. 1. A dominated family $\mathbb{P} = \{\mathbb{P}_\theta : \theta \in \Theta\} \ll \mu$ with μ σ–finite. $(X_1, \cdots, X_n)$ are i.i.d. with common distribution from $\mathbb{P}$, then $\underline{X} = (X_1, \cdots, X_n)$ has a distribution from an exponential family dominated by μ^n with densities

$$f_\theta(\underline{x}) = \exp[\underline{\eta}(\theta)^T \underline{\tilde{T}}(\underline{x}) - \tilde{B}(\theta)]\tilde{h}(\underline{x})$$

where

$$\tilde{B}(\theta) = nB(\theta)$$
$$\underline{\tilde{T}}(\underline{x}) = \sum_{i=1}^{n} T(x_1)$$
$$\tilde{h}(\underline{x}) = \prod_{i=1}^{n} h(x_i)$$

2. By changing dominating measures, we can take $h(x) \equiv 1$. Define a new dominating measure ν by $\frac{d\nu}{d\mu} = h$. We claim the ν is σ-finite. Fix $\theta \in \Theta$, define $B_m = \left\{x : \exp[\underline{\eta}(\theta)^T \underline{T}(x) - B(\theta)] \geq \frac{1}{m}\right\}$. We have $\bigcup_m B_m = \Xi$, so

$$\begin{aligned}\nu(B_m) &= \int_{B_m} h(x)d\mu(x) \leq m \cdot \int_{B_m} \exp[\underline{\eta}(\theta)^T \underline{T}(x) - B(\theta)]h(x)d\mu(x) \\ &\leq m\mathbb{P}_\theta(B) \leq m\end{aligned}$$

Now, for all θ and A measurable,

$$\begin{aligned}\mathbb{P}_\theta(A) &= \int_A \exp[\underline{\eta}(\theta)^T \underline{T}(x) - B(\theta)]h(x)d\mu(x) \\ &= \int_A \exp[\underline{\eta}(\theta)^T \underline{T}(x) - B(\theta)]d\mu(x)\end{aligned}$$

We have $\frac{d\mathbb{P}_\theta}{d\nu}(x) = \exp[\underline{\eta}(\theta)^T \underline{T}(x) - B(\theta)]$ which is an exponential family with $h \equiv 1$. For convenience, we will often delete the factor $h(x)$ when writing the density in an exponential family. We need to determine the new dominating measure ν in the previous examples which causes h to disappear.

3. Note that the density $\exp[\underline{\eta}(\theta)^T \underline{T}(x) - B(\theta)]$ is strictly positive, so we conclude that in general, the region where the density is positive is not dependent on the parameter θ. A family we consider below which is related to the Exp(β) family is the Exp$[a, b]$ distribution has Lebesgue density given by $f_{a,b}(x) = \frac{1}{a}\exp\left(\frac{-(x-b)}{a}\right), x \geq b$. Here the parameter a is required to be positive and b is an arbitrary real number. Since the density is positive exactly on the set $[b, \infty)$, it follows that this family is not an exponential family.

4. Let $\Lambda = \underline{\eta}(\Theta)$, then $\Lambda \subset \mathbb{R}^p$ and $\underline{\eta} \in \Lambda$ may be used as parameter rather than θ since the actual probability measure only depends on $\underline{\eta}^T \underline{T}(x)$. We have the **canonical form**

$$f_{\underline{\eta}}(x) = \exp[\underline{\eta}^T \underline{T}(x) - A(\underline{\eta})]$$

. $A(\underline{\eta}) = \log\left(\int_\Xi \exp[\underline{\eta}^T \underline{T}(x)]d\mu(x)\right)$ is a normalizing constant. The new parameter $\underline{\eta}$ is called the natural parameter. If the natural parameterization is used, then we call the family a natural parameter exponential family or a canonical form exponential family. It is of course required that $A(\underline{\eta})$ be finite to define a probability density, but it is obviously also sufficient. The natural parameter space Λ_0 is the largest possible parameter space for the natural parameter

$$\begin{aligned}\Lambda_0 &= \{\underline{\eta} \in \mathbb{R}^p : 0 < \int_\Xi \exp[\underline{\eta}^T \underline{T}(x)]h(x)d\mu(x) < \infty\} \\ &= \{\underline{\eta} : -\infty < A(\underline{\eta}) < \infty\}\end{aligned}$$

An exponential family in canonical form with the **natural parameter space** is called a **natural exponential family**.

5. The canonical form representation is not unique. Indeed, let D be any nonsingular $p \times p$ matrix and put $\tilde{\underline{\eta}} = (D^{-1})^T \underline{\eta}$, $\tilde{\underline{T}}(x) = D\underline{T}(x)$, then $\underline{\eta}^T \underline{T} = \underline{\eta}^T D^{-1} D\underline{T} = [(D^{-1})^T \underline{\eta}]^T (D\underline{T}) = \tilde{\underline{\eta}}^T \tilde{\underline{T}}$ and we may use the new parameter $\tilde{\underline{\eta}}$ in place of $\underline{\eta}$ provided we switch from $\underline{T}$ to $\tilde{\underline{T}}(x)$

6. If the parameter space $\Lambda \subset M$ where M is a linear manifold in $\mathbb{R}^p$ with $\dim(M) = q < p$, then the natural parameter satisfies $p - q$ independent linear constraints. (To wit, $C^T \underline{\eta} = C^T \underline{\zeta}$ where C is a $p \times (p-q)$ matrix with columns orthogonal to $M - \underline{\zeta}$ and $\underline{\zeta}$ is any element of M. Note that $M - \underline{\zeta}$ is a q-dimensional linear subspace since it contains 0, and is the unique such subspace parallel to M.) Then there is a $p \times q$ matrix B such that for any $\underline{\zeta} \in M$, there is for each $\underline{\eta} \in \Lambda$ a unique $\tilde{\underline{\eta}} \in \mathbb{R}^q$ such that $\underline{\eta} = B\tilde{\underline{\eta}} + \zeta$, and we will denote by $\tilde{\Lambda}$ the set of all such $\tilde{\underline{\eta}}$. (Here, B may be taken as any matrix whose columns span $M - \zeta$, and then the entries in $\tilde{\underline{\eta}}$ are just the coefficients in the expansion of $\underline{\eta} - \underline{\zeta}$ using the basis consisting of the columns of M.) Then $\underline{\eta}^T \underline{T} = \tilde{\underline{\eta}}^T (B^T \underline{T}) + \underline{\zeta}^T \underline{T}$. so $f_{\underline{\eta}}(x) = \exp[\underline{\eta}^T \underline{T}(x) - A(\underline{\eta})] = \exp[\tilde{\underline{\eta}}^T \tilde{\underline{T}}(x) - \tilde{A}(\tilde{\underline{\eta}})]\tilde{h}(x)$ where $\tilde{\underline{T}}(x) = B^T \underline{T}(x) \in \mathbb{R}^q$, $\tilde{A}(\tilde{\underline{\eta}}) = A(B\tilde{\underline{\eta}} + \underline{\zeta})$ and $\tilde{h}(x) = \exp[\underline{\zeta}^T \underline{T}(x)]$. Note that $\tilde{\underline{\eta}}$ does not appear in $\tilde{h}(x)$. Thus, we may reparametrize and reduce the dimension of $\underline{\eta}$ and $\underline{T}$ so that Λ does not belong to any proper linear manifold.

   Similarly, suppose $\underline{T}$ satisfies $p - q$ linear constraints, i.e. if there is a q-dimensional linear manifold M of dimension $q < p$, $\mathbb{P}_\eta\{x : T(x) \in M\} = 1$, there is a $p \times (p-q)$ matrix C and a $\underline{\zeta} \in \mathbb{R}^p$ s.t. $\mathbb{P}_\eta\{x : C^T \underline{T}(x) = C^T \underline{\zeta}\} = 1$. Note that if this happens for one $\underline{\eta}$ then it happens for all $\underline{\eta}$ since the set where $f_\eta > 0$ doesn't depend on η. Now let B be a $p \times q$ matrix with columns spanning M and $\underline{\tau} \in M$, then $T = B\tilde{T} + \tau$ for a unique $\tilde{T} \in \mathbb{R}^q$ and we may reparametrize with $\tilde{\eta} = B^T \eta$ and reduce dimensionality again and $\tilde{T}$ will not satisfy any linear constraints (i.e. not be confined to a proper linear manifold in $\mathbb{R}^q$). Note that even though η was not constrained here, we have lost nothing since if $(\eta_1 - \eta_2)$ is orthogonal to $M - \tau$, we have $\eta_1^T T = \eta_2^T T$ μ–a.e. where μ is the dominating measure, i.e. the original parameterization was not identifiable.

In conclusion, we can always reduce an exponential family in canonical form so that neither the parameter η nor the T satisfies any linear constraints. We say the family is **minimal**. If the parameter space of a minimal exponential family (in canonical form) has nonempty interior (i.e. the parameter space contains a nonempty open set, such as an open rectangle $(a_1, b_1) \times \cdots \times (a_p, b_p)$ where $a_i < b_i$ for $1 \le i \le p$),then the family is said to be of **full rank**.

Definition 2.3.5. The **Frobenius norm** of A is $||A|| = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij}^2 \right)^{1/2}$

Proposition 2.3.1. *For any matrices A and B, $||AB|| \le ||A||||B||$, provided AB is defined. In particular, if x is a **vector of appropriate dimension**, $||Ax|| \le ||A||||x||$.*

Proposition 2.3.2. *Suppose $\{f_\eta : \eta \in \Lambda_0\}$ is a natural exponential family which is minimal.*

1. *The natural parameter space Λ_0 is a convex subset of $\mathbb{R}^p$ and the family is full rank.*

   Proof. Assume $\eta_1, \eta_2 \in \Lambda_0$ and put $\eta = \alpha\eta_1 + (1-\alpha)\eta_2$ for some $\alpha \in [0,1]$. The exponential function is convex, so $\exp[\alpha\eta_1^T T + (1-\alpha)\eta_2^T T] \le \alpha \exp[\eta_1^T T] + (1-\alpha)\exp[\eta_2^T T]$. Taking integrals w.r.t. the dominating measure (and noting that the integrands are positive, so the integrals exist) gives

$$\int_\Xi \exp[\eta^T T(x)] d\mu(x) \le \alpha \int_\Xi \exp[\eta^T T] d\mu(x) + (1-\alpha) \int_\Xi \exp[\eta^T T] d\mu(x)$$

   Thus, finiteness of the two integrals on the r.h.s. implies finiteness of the integral on the l.h.s., i.e. that η is in Λ_0 and hence that Λ_0 is convex.

   To show that the family is full rank, it is only necessary to show that the natural parameter space has nonempty interior (since we know by minimality that T does not satisfy any linear constraint).

Since the canonical parameter η does not satisfy any linear constraints, we know that Λ_0 does not lie in a lower dimensional linear manifold. Thus, we can find $p+1$ vectors $\underline{\eta}_0, \cdots, \underline{\eta}_p$ such that $\{\underline{\eta}_1 - \underline{\eta}_0, \underline{\eta}_2, -\underline{\eta}_0, \cdots, \underline{\eta}_p, -\underline{\eta}_0\}$ forms a linear independent set of p-dimensional vectors. We will assume without loss of generality that $\underline{\eta}_0 = 0$ by subtracting $\underline{\eta}_0$ from every other $\underline{\eta}$.

$$K = \left\{ \sum_{j=0}^{p} a_j \underline{\eta}_j : a_j \geq 0, 0 \leq j \leq p \& \sum_{j=0}^{p} a_j = 1 \right\}$$

$$\tilde{\underline{\eta}} = (p+1) \sum_{j=0}^{p} a_j \underline{\eta}_j$$

Of course, $\tilde{\underline{\eta}} \in K$, and our goal is to show that for some $\epsilon > 0$, $||\underline{\eta} - \tilde{\underline{\eta}}|| < \epsilon$ implies $\underline{\eta} \in K$, i.e. that $\tilde{\underline{\eta}}$ has a neighborhood contained in K, so K has nonempty interior. Now any $\underline{\eta} \in \mathbb{R}^p$ can be written as $\underline{\eta} = \tilde{\underline{\eta}} + \sum_{j=1}^{p} b_j \underline{\eta}_j$, where $\underline{b} = (b_1, \cdots, b_p)$ can be found by solving $A\underline{b} = \underline{\eta} - \tilde{\underline{\eta}}$ where A is the $p \times p$ matrix with j-th column equal to $\underline{\eta}_j$. We know that A is invertible, so by Proposition 2.3.1, $||b|| \leq ||A^{-1}||||\underline{\eta} - \tilde{\underline{\eta}}||$. Now in order to guarantee that $\underline{\eta}$ is in K we need that $a_j = (p+1)^{-1} + b_j, 1 \leq j \leq p$, and $a_0 = (p+1)^{-1} - \sum_{j=1}^{p} b_j$ are nonnegative since they already sum to 1. For this it suffices that $\max_{1 \leq j \leq p} |bj| \leq (p+1)^{-1}$ and $\sum_{j=1}^{p} |b_j| \leq (p+1)^{-1}$. Now $\max_{1 \leq j \leq p} |bj| \leq ||\underline{b}||$ and using Cauchy-Schwartz, it is easy to see that $\sum_{j=1}^{p} |b_j| \leq p^{1/2}||\underline{b}||$. Hence, as long as we make $||\underline{b}|| < \min\{p^{-1/2}(p+1)^{-1}, (p+1)^{-1}\} = p^{-1/2}(p+1)^{-1}$, then we will satisfy our requirements on $\underline{b}$ so that $\underline{\eta} \in K$. Thus, it suffices to take $\epsilon = ||A^{-1}||^{-1} p^{-1/2}(p+1)^{-1}$ □

2. *If η_0 is an interior point of Λ_0 (i.e. there is some open ball $B(\eta_0, \epsilon) \subset \Lambda_0$, where the radius $\epsilon > 0$), then the m.g.f. ψ_{η_0} of $Law_{\eta_0}[T(X)]$ is finite in a neighborhood of 0 and is given by*

$$\psi_{\eta_0}(u) = \exp[A(\eta_0 + u) - A(\eta_0)]$$

and the cumulant generating function is $\kappa_{\eta_0}(u) = A(\eta_0+u) - A(\eta_0)$. In particular, $E_{\eta_0}[T(X)] = \nabla A(\eta_0)$ and $Cov_{\eta_0}[T(X)] = D^2 A(\eta_0)$. Furthermore, $A(\eta)$ is a strictly convex function on the interior of Λ_0.

Proof. For the m.g.f. calculation, we have

$$\begin{aligned}
\psi_{\eta_0}(u) = E_{\eta_0}[\exp(u^T T(X))] &= \int_{\Xi} \exp[u^T T(X) + \eta_0^T T(x) - A(\eta_0)] d\mu(x) \\
&= \int_{\Xi} \exp[(u+\eta_0)^T T(X) - A(u+\eta_0)] d\mu(x) \exp[A(u+\eta_0) - A(\eta_0)] \\
&= \exp[A(u+\eta_0) - A(\eta_0)]
\end{aligned}$$

where this is valid provided $u + \eta_0$ is in Λ_0. Since there is a neighborhood of η_0 contained in Λ_0, it follows that there is a neighborhood of 0 such that if u is in this neighborhood of 0, then $u + \eta_0$ is in the neighborhood of η_0 contained in Λ_0, and everything in the last displayed calculation is finite, i.e. ψ_{η_0} is finite in a neighborhood of 0. The formula for *kappa* is immediate and the formulae for the first two moments of T under η_0 follows by an elementary calculation. Since the family is minimal, T is not almost surely confined to some proper linear manifold of $\mathbb{R}^p$, so by Proposition 2.1.8, the covariance is full rank, i.e. positive definite. This shows that $A(\eta)$ is strictly convex by the second derivative test.

$$\begin{aligned}
E_{\eta_0}[T(X)] &= \nabla\psi_{\eta_0}(0) = \nabla_u \exp[A(u+\eta_0) - A(\eta_0)] \mid_{u=0} \\
&= (\nabla A(u+\eta_0)\exp[A(u+\eta_0) - A(\eta_0)]) \mid_{u=0} = \nabla A(\eta_0) \\
E_{\eta_0}[T(X)T(X)^T] &= D^2\psi_{\eta_0}(0) = D\{DA(u+\eta_0)\exp[A(u+\eta_0) - A(\eta_0)]\} \mid_{u=0} \\
&= (D^2A(u+\eta_0))\exp[A(u+\eta_0) - A(\eta_0)] \mid_{u=0} \\
&+ [\nabla A(u+\eta_0)][\nabla A(u+\eta_0)]^T \exp[A(u+\eta_0) - A(\eta_0)] \mid_{u=0} \\
&= D^2A(\eta_0) + (\nabla A(\eta_0))(\nabla A(\eta_0))^T = D^2A(\eta_0) + E[T(X)]E[T(X)]^T \\
\implies & Cov_{\eta_0}[T(X)] = D^2A(\eta_0)
\end{aligned}$$

□

3. *Under the same hypotheses, if $\phi : \Xi \to \mathbb{R}$ is such that $E_\eta[\phi(X)] < \infty$, then the function $h(\eta) = E_\eta[\phi(X)]$ is finite in a neighborhood of η_0. Furthermore, h is infinitely differentiable and the derivatives may be computed by interchange of differentiation and integration.*

   Proof. We apply Theorem 2.2.2. For the bounded function $f(x) = I_{[0,\infty)}(\phi(x)) - I_{(-\infty,0)}(\phi(x))$ and for the measure $\tilde{\mu}$ in Theorem 2.2.2, use $d\tilde{\mu}(x) = |\phi(x)d\mu(x)|$. Note that $f(x)|\phi(x)| = \phi(x)$. Then apply that theorem to $B(\eta) = \int_\Xi f(x)\exp[\eta^T T(X)]d\tilde{\mu}(x)$. Infinite differentiability of B at an interior point of Λ_0 implies the same for h, and the interchangeability of the differentiation and integration operators follows as well. □

Example 2.3.3. Gamma distribution Gamma(α, β)

$$\begin{aligned}
f_{\alpha,\beta}(x) &= \frac{x^{\alpha-1}}{\beta^\alpha\Gamma(\alpha)}\exp(-x/\beta)I_{[0,\infty)}(x) \\
&= \exp\left[\frac{-1}{\beta}x + \alpha\log x - (\alpha\log\beta + \log\Gamma(\alpha))\right]h(x)
\end{aligned}$$

$\eta = (-1/\beta, \alpha), T = (X, \log X), A(\eta) = -\eta_2\log(-\eta_1) + \log\Gamma(\eta_2)$, thus,

$$\begin{aligned}
\psi_\eta(u_1,u_2) &= E_\eta[\exp\{u_1X + u_2\log X\}] \\
&= \exp[-(\eta_2+u_2)\log(-(\eta_1+u_1)) + \log\Gamma(\eta_2+u_2) + \eta_2\log(-\eta_1) - \log\Gamma(\eta_2)] \\
&= \exp\left[\eta_2\log\left(\frac{\eta_1}{\eta_1+u_1}\right) + u_2\log(-(\eta_1+u_1)) + \log\frac{\Gamma(\eta_2+u_2)}{\Gamma(\eta_2)}\right] \\
&= \exp\left[\alpha\log\left(\frac{-1/\beta}{-1/\beta+u_1}\right) + \alpha\log\left(\frac{1}{\beta} - u_1\right) + \log\left(\frac{\alpha+u_2}{\alpha}\right)\right] \\
&= \exp\left[-\alpha\log(1-\beta u_1) + \alpha\log\left(\frac{1-\beta u_1}{\beta}\right) + \log\left(1+\frac{u_2}{\alpha}\right)\right]
\end{aligned}$$

$E_\eta[T(X)] = \nabla A(\eta) = \left(-\frac{\eta_2}{\eta_1}, -\log(-\eta_1) + \frac{\Gamma'(\eta_2)}{\Gamma(\eta_2)}\right) = (\alpha\beta, \log\beta + \psi(\alpha))$ where $\psi(\alpha)$ is the digamma function. $Cov[T(X)] = \begin{bmatrix} \frac{\eta_2}{\eta_1^2} & \frac{-1}{\eta_1} \\ \frac{-1}{\eta_1} & \psi_{(1)}(\eta_2) \end{bmatrix} = \begin{bmatrix} \alpha\beta^2 & \beta \\ \beta & \psi_{(1)}(\alpha) \end{bmatrix}$, where $\psi_{(1)}$ is the trigamma function.

Example 2.3.4. Multinomial Family: Suppose Ω is partitioned into $A_1, \cdots, A_k$. Let p_i be the probability of the A_i has nonnegative entries which sum to 1. Let $\underline{X} = (I_{A_1}, \cdots, I_{A_k})$ be the random indicator

k-vector. Then the i-th entry of X is 1 if the outcome is in A_i, and the other entries of X are 0. Now let $\underline{X}_1, \cdots, \underline{X}_n$ be i.i.d. with the same distribution as $\underline{X}$, and let $\underline{Y} = \sum \underline{X}_i$. Thus, Y_i is the number of times A_i occurs in the n trials. Then $\underline{Y}$ has a multinomial distribution with parameters $Mult(n, \underline{p})$. As the parameter n is always known (since $\sum Y_i = n$), we only show $\underline{p}$ in the probabilities, etc. $P_{\underline{p}}[\underline{Y} = \underline{y}] = \binom{n}{\underline{y}} \underline{p}^{\underline{y}}$ where $\underline{y}$ is a k-**multi-index** satisfying $\sum_{i=1}^{k} y_i = n$. Also, $\binom{n}{\underline{y}} = \dfrac{n!}{\underline{y}!} = \dfrac{n!}{\prod_{i=1}^{k} y_i!}$ is a multinomial coefficient and $\underline{p}^{\underline{y}} = \prod_{i=1}^{k} p_i^{y_i}$ is the monomial. Now if we take as dominating measure the discrete measure

$$\mu = \sum_{\underline{y}} \binom{n}{\underline{y}} \delta_{\underline{y}}$$

then the density of $Mult(n, \underline{p})$ is $f_{\underline{p}}(\underline{y}) = \underline{p}^{\underline{y}} = \exp\left[\sum_{i=1}^{k} y_i \log(p_i)\right]$, provided it is known $p_i \neq 0$ for all i. This is an exponential family with natural parameters $\eta_i = \log(p_i)$ and $\underline{T} = \underline{y}$, but $\underline{T}$ satisfies the linear constraint $\sum_{i=1}^{k} y_i = n$ and the η_i satisfy the nonlinear constraint $\sum \exp[\eta_i] = 1$. There are many ways of eliminating this indeterminacy, but the most common is to use $\underline{T} = (\underline{y}_1, \ldots, \underline{y}_{k-1})$ (i.e. leave off the last component which is determinable from the other components and multinomial coefficient), and form the **multinomial logit**

$$\eta_i = \log\left(\frac{p_i}{1 - \sum_{j=1}^{k-1} p_j}\right), 1 \leq i \leq (k-1)$$

Note that given any probability vector $\underline{p}$ one can obtain a $(k-1)$ vector $\underline{\eta}$, and conversely given any $\underline{\eta} \in \mathbb{R}^{k-1}$, one can obtain the corresponding probability vector through

$$p_k = \frac{1}{1 + \sum_{j=1}^{k-1} \exp[\eta_j]}, p_i = p_k \exp[\eta_i], 1 \leq i \leq (k-1)$$

Note that the multinomial logit $\underline{\eta}$ is an unconstrained vector in $\mathbb{R}^(k-1)$ whereas the probability vector p is a k-vector which satisfies the constraints of nonnegativity and $\sum p_i = 1$. The multinomial density can be written as

$$\begin{aligned} f_{\underline{p}}(\underline{y}) &= \exp\left[\sum_{i=1}^{k-1} y_i \log(p_i) + \left(n - \sum_{i=1}^{k-1} y_i\right) \log\left(1 - \sum_{j=1}^{k-1} p_j\right)\right] \\ &= \exp\left[\sum_{i=1}^{k-1} y_i \left(\log(p_i) - \log\left(1 - \sum_{j=1}^{k-1} p - j\right)\right) + n \log\left(1 - \sum_{j=1}^{k-1} p_j\right)\right] \\ &= \exp\left[\sum_{i=1}^{k-1} y_i \eta_i - n \log\left(1 + \sum_{j=1}^{k-1} \exp[\eta_i]\right)\right] \end{aligned}$$

which is an exponential family in canonical form with

$$A(\underline{\eta}) = n \log\left(1 + \sum_{j=1}^{k-1} \exp[\eta_i]\right)$$

. But, $(\log p_1, \cdots, \log p_k)$ does not satisfy the linear constraint and then the family is not minimal. From this in conjunction with Proposition 2.3.2, we have for $1 \leq i < k$,

$$E_{\underline{p}}[Y_i] = \frac{n \exp[\eta_i]}{1 + \sum_{j=1}^{k-1} \exp[\eta_i]} = np_i$$

and since $Y_k = n - \sum_{j=1}^{k-1} Y_j$, $E_{\underline{p}}[Y_k] = n - \sum_{j=1}^{k-1} E_{\underline{p}}[Y_j] = n - \sum_{j=1}^{k-1} np_j = n\left[1 - \sum_{j=1}^{k-1} p_j\right] = np_k$. Also, if $1 \leq i < j \leq k$ and $1 \leq i < k$ respectively, then

$$Cov_{\underline{p}}[Y_i, Y_j] = \frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\underline{\eta}) = \frac{-n \exp[\eta_i] \exp[\eta_j]}{\left(1 + \sum_{m=1}^{k-1} \exp[\eta_m]\right)^2} = np_i p_j$$

$$Var_{\underline{p}}[Y_i] = \frac{\partial^2}{\partial^2 \eta_i} A(\underline{\eta}) = \frac{n\left[\exp[\eta_i]\left(1 + \sum_{j=1}^{k-1} \exp[\eta_j]\right) - \exp[2\eta_i]\right]}{\left(1 + \sum_{m=1}^{k-1} \exp[\eta_m]\right)^2} = n(p_i - p_i^2) = np_i(1 - p_i)$$

2.3.3 Location-Scale Families

Definition 2.3.6. Let $\mathbb{P}$ be a Borel p.m. on $\mathbb{R}$.

1. The location family generated by $\mathbb{P}$ is $\{P_b : b \in \mathbb{R}\}$ where $P_b(A) = A - b$ and $A - b = \{x - b : x \in A\}$. Note that if $\tau_b^{-1} : \mathbb{R} \to \mathbb{R}$ is **transition** by b i.e. $\tau_b(x) = x + b$, then $P_b = \mathbb{P} \circ \tau_b^{-1}$, i.e. if $Z \sim \mathbb{P}$ then $\tau_b(Z) = Z + b \sim P_b$

2. The scale family generated by $\mathbb{P}$ is $\{P_a : a > 0\}$ where $P_a(A) = P(a^{-1}A)$ and $a^{-1}A = \{a^{-1}x : x \in A\}$. Note that if $\varsigma_b^{-1} : \mathbb{R} \to \mathbb{R}$ is **multiplication** by a i.e. $\varsigma_a(x) = ax$, then $P_a = \mathbb{P} \circ \varsigma_a$, i.e. if $Z \sim \mathbb{P}$ then $\varsigma_a^{-1}(Z) = aZ \sim P_a$

3. The location-scale family generated by $\mathbb{P}$ is $\{P_{ab} : a > 0 \text{ and } b \in \mathbb{R}\}$ where $P_{ab}(A) = P(a^{-1}(A - b))$. Note that if $Z \sim \mathbb{P}$ then $\tau_b(\varsigma_a(Z)) = aZ + b \sim P_{ab}$ i.e. $P_{ab} = \mathbb{P} \circ \varsigma_a^{-1} \circ \tau_b^{-1}$

4. If $\mathbb{P} \ll m$ with Lebesgue density f, then $f_{ab}(x) = \frac{dP_{ab}}{dm}(x) = \frac{1}{a} f\left(\frac{x - b}{a}\right)$

Example 2.3.5.

1. The $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ is a location scale family generated by $N(0, 1)$. (We would use the parametrization (σ, μ) to be consistent with the above.)

2. The $\{Unif(a, b) : -\infty < a < b < \infty\}$ family is a location-scale family generated by $Unif(0, 1)$. The location parameter is a and the scale parameter is $b - a$.

3. The above example is also an example of what is known as a **truncation family**. Let $g : \mathbb{R} \to [0, \infty)$ be a Borel function satisfying $0 < \int_a^b g(x)dx < \infty$ for all $-\infty < a < b < \infty$. Then we put $f_{ab}(x) = \frac{g(x) I_{[a,b](x)}}{\int_a^b g(y)dy}$. Clearly the uniform family is a truncation family with constant g. Such truncation families have little if any application in practice, although they seem to play an important role in mathematical statistics textbooks.

4. Let $P = Exp(1)$, the exponential distribution with Lebesgue density $f(x) = \exp(-x), x > 0$. The location–scale family generated by P is called the **shifted exponentials**, and a member thereof will be denoted $Exp[a, b]$ and has Lebesgue density $f_{ab}(x) = a^{-1}\exp[-(x-b)/a]I[b,\infty)(x)$. Note that the support $[b, \infty)$ depends on the parameter. The scale family of exponential distributions $Exp[a, 0], a > 0$ (called the family of exponential distributions and not the exponential family) is perhaps more fundamental and is frequently used as a model for observations which must be positive, such as lifetimes or masses. Note that the shifted exponential family $\{Exp[\beta, b] : \beta > 0 \text{ and } b \in \mathbb{R}\}$ is not a subfamily of Gamma, and it is also not an exponential family since the support depends on the parameter b. See Remark 2.3.1 (c) above.

5. The location family of distributions $Exp[1, b]$ where b is an arbitrary real number has little application in practice. It is however an example of a **left truncation family** in exercise.

6. We often generate families not from a single distribution but a family. For example, the Weibull(α, β) distribution has Lebesgue density $f_{\alpha\beta}(x) = \dfrac{\alpha x^{\alpha-1}}{\beta^\alpha}\exp[-(x/\beta)^\alpha]I_{(0,\infty)}(x)$. β is a scale parameter. α is a shape parameter.

2.3.4 Group Families

Definition 2.3.7. 1. A class of transformations T on $(\Xi, \mathcal{G})$ is called a **transformation group** iff the following hold:

   (a) Every $g \in T$ is measurable $g : (\Xi, \mathcal{G}) \to (\Xi, \mathcal{G})$.

   (b) T is closed under composition, i.e. if g_1 and g_2 are in T the so is $g_1 \circ g_2$

   (c) T is closed under taking inverses, i.e. if $g \in T$ then $g^{-1} \in T$.

   If $g_1 \circ g_2 = g_2 \circ g_1$ for all g_1 and g_2 in T, then T is called commutative.

2. If T is a transformation group and $\mathbb{P}_0$ is a family of probability measures on $(\Xi, \mathcal{G})$, then the group family generated by $\mathbb{P}_0$ under T is $\mathbb{P}_0 \circ T^{-1} = \{\mathbb{P} \circ g^{-1} : \mathbb{P} \in \mathbb{P}_0 \text{ and } g \in T\}$. Note that if $Z \sim P$,then $g(Z) \sim \mathbb{P} \circ g^{-1}$.

Example 2.3.6. Consider the observation space $(\mathbb{R}^n, \mathcal{B}_n)$. For an $n \times n$ nonsingular matrix A and $\underline{b} \in \mathbb{R}^n$, define the transformation $g_{A,\underline{b}}(x) = Ax + \underline{b}$. The family of transformations $\mathbf{T} = \{g_{A,\underline{b}} : A \text{ is an } n \times n \text{ nonsingular matrix and } \underline{b} \in \mathbb{R}^n\}$ is called the **affine group**. We verify that $\mathbf{T}$ is indeed a transformation group by checking the three defining properties.

1. $g_{A,\underline{b}} : \mathbb{R}^n \to \mathbb{R}^n$ is Borel measurable since it is continuous.

2. Given $g_{A_1,\underline{b}_1}$ and $g_{A_2,\underline{b}_2}$, we have by some simple algebra $(g_{A_1,\underline{b}_1} \circ g_{A_2,\underline{b}_2})(\underline{x}) = (A_1A_2)\underline{x} + (A_1\underline{b}_2 + \underline{b}_1)$, i.e. $(g_{A_1,\underline{b}_1} \circ g_{A_2,\underline{b}_2}) = g_{A,\underline{b}}$ where $A = A_1A_2$ and $\underline{b} = A_1\underline{b}_2 + \underline{b}_1$. This shows $\mathbf{T}$ is closed under taking composition.

3. Given $g_{A,\underline{b}}$ and $\underline{x} \in \mathbb{R}^n$ consider solving for $\underline{y}$ in $g_{A,\underline{b}}(\underline{y}) = \underline{x}$, which gives $\underline{y} = A^{-1}\underline{x} + (-A^{-1}\underline{b})$, i.e. ($g^{-1}_{A,\underline{b}}$ is an affine transformation with matrix A^{-1} and shift $-A^{-1}\underline{b}$. This shows $\mathbf{T}$ is closed under taking inverses.

We note that $\mathbf{T}$ is not commutative, even when $n = 1$. There are two interesting transformation subgroups, i.e. subsets of the affine group which are also closed under composition and taking inverses. One is the **general linear group** $\{g_{A,\underline{0}} : A \text{ is an } n\times n \text{ nonsingular matrix}\}$, which is simply the group of all nonsingular linear transformations on $\mathbb{R}^n$. It is sometimes denoted $\mathbf{GL}(n)$. The other subgroup of interest is the

translation subgroup $\{g_{I,\underline{b}} : \underline{b} \in \mathbb{R}^n\}$, where I is an $n \times n$ identity matrix. We now generate a group family under the affine group. Let $\mathbb{P}_0$ consist of the single p.m. $N(\underline{0}, I)$, i.e. the standard normal distribution on $\mathbb{R}^n$. If $\underline{Z} \sim N(\underline{0}, I)$, then $A\underline{Z} + \underline{b} \sim N(\underline{b}, AA^T)$. Further, any nonsingular normal distribution on IRn can be so generated. We do have the following problem if we use the parameterization $(\underline{b}, A)$: two different A's can give rise to the same normal distribution (i.e. if $AA^T = A_1A_1^T$). Thus, the parameter does not uniquely define the distribution, i.e. the parameter is not identifiable in the terminology of Definition 2.3.2. To avoid this problem, we will use instead the parameters $(\underline{b}, V)$ where $V = AA^T$ is the covariance.

Definition 2.3.8. Let $\mathbf{T}$ be a transformation group and let $\mathbb{P}$ be a family of probability measures on $(\Xi, \mathcal{G})$. We say that $\mathbb{P}$ is $\mathbf{T}$-invariant iff $\mathbb{P} \circ \mathbf{T}^{-1} = \mathbb{P}$

Proposition 2.3.3. *If $\mathbb{P}$ is a group family (generated by some $\mathbb{P}_0$) under $\mathbf{T}$, then $\mathbb{P}$ is $\mathbf{T}$-invariant.*

Example 2.3.7. Let $\mathbb{P}$ be the multivariate location family generated by the $N(\underline{0}, I)$ on $\mathbb{R}^n$, i.e. $\mathbb{P} = \{N(\underline{\mu}, I) : \underline{\mu}\mathbb{R}^n\}$. Then of course $\mathbb{P}$ is translation invariant by the last Proposition. It turns out that the family is also **spherically invariant**, whereby we mean that it is invariant under the group of orthogonal transformations $\mathbf{O}(n) \equiv \{U : U \text{ is an } n \times n \text{ orthogonal matrix}\}$. To see this, note that if $X \sim N(\underline{\mu}, I)$ and U is orthogonal then $UX \sim N(U\underline{\mu}, UIU^T)$ and $UIU^T = UU^T = I$, so $UX \sim N(U\underline{\mu}, I)$.

2.3.5 (Generalized) Regression Models

Often our data are represented as (x_i, Y_i) where the Y variable (usually, but not always, univariate) is a "response" variable and x_i is a "predictor" or "explanatory" variable. The x's are treated as fixed (non-random) whereas the Y's are modeled as random variables. The interest is in the conditional distribution $P_{Y|X}(\cdot|x)$. We typically assume independence of observations. Possibly the data are realizations of i.i.d. pairs (X_i, Y_i), but we are not interested in the marginal distribution P_X. In a Generalized Regression Model (GRM) we assume the conditional distributions take the form $P_{\theta(x)}$ where P_θ is a typical parametric family and $\theta(x)$ is restricted to some family of functions mapping the space of x values to the parameter space. The real parameter for the GRM is the function $\theta(x)$. For example, the Normal Linear Model: we have p predictor variables $x_1, \cdots, x_p$ which we put in a vector $\vec{x}$. Our model is specified by $Y_i = \alpha + \beta^T\vec{x}_i + \epsilon_i, \epsilon \sim N(0, \sigma^2)$. The parameters for the GRM are $\alpha \in \mathbb{R}, \beta \in \mathbb{R}$, and $\sigma > 0$. Another example is the logistic regression model where the $Y_i \sim B(1, \pi(\vec{x}_i))$ where $\log \dfrac{\pi(\vec{x})}{1 - \pi(\vec{x})} = \alpha + \beta^T\vec{x}$. A Non-Parametric Regression (NPR) model might take the form $Y_i = \mu(x_i) + \epsilon_i$, $\epsilon_i \sim$ i.i.d. with c.d.f. F with $E[\epsilon_i] = 0$. Here, the parameter would be $\mu(\cdot)$ and F. We may make restrictions on μ, e.g., that it satisfy "smoothness" properties, and on F, e.g., finite 2nd moments.

2.4 Distributional Calculations

2.4.1 Lebesgue Densities and Transformations

In conjunction with the change of variables theorem (Theorem 1.2.10), it was mentioned that one often encounters a Jacobian in actually computing the induced measure, which we now explain. First, some more review of advanced calculus on $\mathbb{R}^n$. Let U be an open subset of $\mathbb{R}$ and $h : U \to \mathbb{R}$ have continuous partial derivatives $\dfrac{\partial h_i}{\partial x_j}$ of all component functions, $1 \le i \le k, 1 \le j \le n$. The derivative $Dh(x)$ is the $k \times n$ matrix with (i, j) entry $\dfrac{\partial h_i}{\partial x_j}(x)$. $Dh(x)$ is sometimes called the Jacobian matrix. It is a matrix valued function of x. Also, $Dh(x)$ may be used for local linear approximation of h in the sense that

$h(y) = h(x) + Dh(x)(y-x) + Rem(x,y)$ where the remainder term satisfies $\lim_{y\to x} \frac{||Rem(x,y)||}{||y-x||} = 0$ This last equation states that the "linear" function $h(x) + Dh(x)(y-x)$ as a function of y tends to be a good approximation to $h(y)$ for y close to x. If $U \subset \mathbb{R}^n$ and $h : U \to \mathbb{R}^n$, then $Dh(x)$ is a square $n \times n$ matrix, so its determinant $detDh(x) = J(x)$ is defined and is sometimes called the Jacobian (determinant). The Inverse Function Theorem (p. 221 of Rudin, Principles of Mathematical Analysis) states that under these conditions, if $J(a) \neq 0$ at some $a \in U$, then h is invertible in a neighborhood of a and h^{-1} has derivative $[D(h^{-1})](y) = [(Dh)(h^{-1}(y))]^{-1} = [((Dh) \circ h^{-1})(y)]^{-1}$ at a point y in this neighborhood of $h(a)$. Part of the conclusion is that this inverse matrix exists in the neighborhood of $h(a)$. Also, if $J(x) \neq 0$ for all $x \in U$, then $h(V)$ is an open set for any open set $V \subset U$. This latter fact (V open implies $h(V)$ open) implies that h^{-1} is measurable, if it exists on all of $h(U)$.

Remark. If $h : \mathbb{R}^n \to \mathbb{R}$, then the derivative Dh as defined above is an $n \times 1$ matrix, i.e. a "row vector," whereas the gradient ∇h is a $1 \times n$ "column vector." Note that $Dh = (\nabla h)^T$. $h(y) = h(x) + (\nabla h(x))^T(y - x) + Rem(x, y)$

Theorem 2.4.1. *Suppose $\Omega \subset \mathbb{R}^n$ is open and $h : \Omega \to \mathbb{R}^n$ is a one to one mapping with nonvanishing Jacobian (i.e. $J(x) \neq 0 \forall x \in \Omega$). Let $\Lambda = h(\Omega)$, and let ν be Lebesgue measure restricted to Ω. Then $\nu \circ h^{-1}$ is a Borel measure on Λ, $\nu \circ h^{-1} \ll m$, and*

$$\frac{d(\nu \circ h^{-1})}{dm}(y) = \begin{cases} |detD(h^{-1})(y)| & \text{if } y \in \Lambda \\ 0 & \text{otherwise} \end{cases} \quad m - a.e.$$

To check that det $Dh(x) \neq 0$ for all x, it suffices to show that det $D(h^{-1})(y) \neq 0$ for all y by the Inverse Function Theorem applied to h^{-1}. A relation between the Jacobian of h^{-1} and h is given by $D(h^{-1})(y) = [Dh(h^{-1}(y))]^{-1}$

Proposition 2.4.2. *Suppose $\mathbb{P}$ is a Borel p.m. on $\mathbb{R}^n$ which has Lebesgue density f. Let $h : \Omega \to \Lambda$ be as in Theorem 2.4.1 where $\Lambda = h(\Omega)$ and suppose $\mathbb{P}(\Omega) = 1$. Then $\mathbb{P} \circ h^{-1}$ has Lebesgue density g given by $g(y) = f(h^{-1}(y))|detD(h^{-1})(y)|, \forall y \in \Lambda$. Put otherwise, if Law$[X] = \mathbb{P}$ and $Y = h(X)$, then Law$[Y]$ has Lebesgue density given by g above.*

Proof. Let $J(y) = detD(h^{-1})(y)$, and let $B \subset \Lambda$ be a Borel set. Then

$$(\mathbb{P} \circ h^{-1})(B) = \mathbb{P}(h^{-1}(B)) = \int_{h^{-1}(B)} f(x)dx = \int_\Omega I_{h^{-1}(B)}(x)f(x)dx = \int_\Omega I_B(h(x))f(x)dx$$

where the last equality follows since $x \in h^{-1}(B)$ iff $h(x) \in B$. Now put

$$\beta(y) = I_B(y)(f \circ h^{-1})(y)$$

and since $(f \circ h^{-1})(h(x)) = f(x)$, we have

$$(\mathbb{P} \circ h^{-1})(B) = \int_\Omega \beta((h(x))dm^n(x) = \int_\Lambda \beta(y)d(m^n \circ h^{-1})(y)$$

where the last equation follows from the change of variables theorem (Theorem 1.2.8). By Proposition 1.4.2 (a) and the previous theorem,

$$\begin{aligned}(\mathbb{P} \circ h^{-1})(B) &= \int_\Lambda \beta(y)|J(y)|dm^n(y) = \int_\Lambda I_B(y)(f \circ h^{-1})(y)|J(y)|dy \\ &= \int_B (f \circ h^{-1})(y)|J(y)|dy = \int_B g(y)dy\end{aligned}$$

Since the above result holds for arbitrary Borel $B \subset \Lambda$, it follows that $d(\mathbb{P} \circ h^{-1})/dm^n$ exists and equals g, by the uniqueness part of the Radon-Nikodym Theorem. □

Example 2.4.1. 1. **Log-normal distribution** Suppose $X \sim N(\mu, \sigma^2)$ and $Y = \exp[X]$. Then Y is said to have a log-normal distribution with parameters μ and σ^2. Perhaps we should say Y has an "exponential-normal distribution" as it is the exponential of a normal r.v., but the terminology "log-normal" is standard. It presumably arose from something like the statement, "The logarithm is normally distributed." Now we derive the Lebesgue density using the previous theorem. Now $\Omega = \mathbb{R}$ and $\Lambda = (0, \infty)$. Of course, $h(x) = \exp[x]$ and $h^{-1}(y) = \log y$, so $D(h^{-1})(y) = 1/y$. Hence, letting f be the $N(\mu, \sigma^2)$ density we have

$$\begin{aligned} g(y) &= f(\log y)\frac{1}{y}, y > 0, \\ &= \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log y - \mu)^2\right], y > 0. \end{aligned}$$

Next we consider the problem of computing the mean and variance of Y . One approach would be to compute $\int_0^\infty y^m dy$ for $y = 1, 2$. However, one should always consider all options in computing expectations via the law of the unconscious statistician. Now $E[Y^m] = E[\exp(mX)]$ which is the m.g.f. of X evaluated at m. Recalling the m.g.f. of a univariate normal distribution $\psi_{N(\mu,\sigma^2)}(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]$ we have $E[Y^m] = \exp\left[\mu m + \frac{1}{2}\sigma^2 m^2\right]$ and so $E[Y] = \exp\left[\mu + \frac{1}{2}\sigma^2\right]$, $Var[Y] = E[Y^2] - E[Y]^2 = \exp\left[2\mu + 2\sigma^2\right] - \exp\left[2\mu + \sigma^2\right] = \exp\left[2\mu + \sigma^2\right]\exp\left[\sigma^2 - 1\right]$

2. **Student's t-distribution** Suppose X and Y are independent r.v.'s with the following distributions: $\text{Law}[X] = N(0, 1)$, $\text{Law}[Y] = \chi^2_n$, i.e. the Lebesgue densities are given by

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \\ f_Y(y) &= \frac{y^{n/2-1}e^{-y/2}}{\Gamma(n/2)2^{n/2}} I_{(0,\infty)}(y) \end{aligned}$$

Note that X has the standard normal distribution and Y has a chi-squared distribution with n degrees of freedom. Let $T = \dfrac{X}{\sqrt{Y/n}}$. Then T is said to have Student's t-distribution with n degrees of freedom. We will derive the Lebesgue density for T. By Proposition 1.4.3, the joint density for X and Y is $f_{XY}(x, y) = f_X(x)f_Y(y)$. Letting $\Omega = \mathbb{R} \times [0, \infty) = supp(Law[X, Y])$ (this last equality follows from Exercise 1.4.19) and $h(x, y) = (x/\sqrt{y/n}, y), (x, y) \in \Omega$, then $h(\Omega) = \Omega$ and h is one to one on Ω since $h(x, y) = (t, u)$ iff $x = t\sqrt{u/n}$ and $y = u$, and this gives the inverse function $h^{-1}(t, u) = (t\sqrt{u/n}, u)$. Now the Jacobian matrix for h^{-1} is

$$Dh^{-1}(t, u) = \begin{bmatrix} \sqrt{u/n} & t/(2\sqrt{un}) \\ 0 & 1 \end{bmatrix}$$

with Jacobian $detDh^{-1}(t, u) = \sqrt{u/n}$ which is nonvanishing for all $(t, u) \in \Omega$. Hence, the joint density of (T, U) is by Theorem 2.4.1

$$\begin{aligned} f_{TU}(t, u) &= f_{XY}(h^{-1}(t, u))|detDh^{-1}(t, u)| \\ &= \left[\frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2 u}{2n})\right] \left[\frac{u^{n/2-1}e^{-u/2}}{\Gamma(n/2)2^{n/2}} I_{(0,\infty)}(u)\right] \sqrt{u/n} \\ &= \frac{1}{\sqrt{\pi}\Gamma(n/2)2^{(n+1)/2}n^{1/2}} u^{(n-1)/2} \exp\left[-\frac{u(1 + t^2)}{2n}\right] I_{(0,\infty)}(u) \end{aligned}$$

To get the marginal density for T, we apply Proposition 1.4.4 to obtain

$$f_T(t) = \int f_{TU}(t, u)du = \frac{1}{\sqrt{\pi}\Gamma(n/2)2^{(n+1)/2}n^{1/2}} \int_0^\infty u^{(n-1)/2} \exp\left[-\frac{u(1 + t^2)}{2n}\right] du$$

In the last integral make the change of variables $v = (1 + t^2/n)u$, so that $du = \frac{dv}{1 + t^2/n}$. This gives

$$f_T(t) = \frac{1}{\sqrt{\pi}\Gamma(n/2)2^{(n+1)/2}n^{1/2}} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \int_0^\infty v^{(n+1)/2-1}e^{-v/2}dv$$
$$= \frac{1}{\sqrt{\pi}\Gamma(n/2)2^{(n+1)/2}n^{1/2}} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \Gamma\left(\frac{n+1}{2}\right) 2^{(n+1)/2}$$

where the last line follows since the integrand in the previous line is the $\chi^2_{(n+1)}$ density without the normalizing constant. In summary,

$$f_T(t) = \frac{\Gamma(n+1)/2}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

This is the (Lebesgue) density of Student's t–distribution with n degrees of freedom.

The preceding example is typical of how the method gets used when one wishes to obtain the Lebesgue density for a real valued random variable Y that is a function of a random vector $\underline{X}$: one must extend Y to a vector $\underline{Y}$ of the same dimension as $\underline{X}$ to obtain a one to one transformation with nonsingular Jacobian and then apply mariginalization to get the desired density. Sometimes, it is not possible to compute the marginal density in a neat closed form and one must be satisfied with an integral expression or something similar.

2.4.2 Applications of Conditional Distributions

Theorem 2.4.3. ***Jensen's Inequality for Conditional Expectation:*** *Let $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda, \mathcal{G})$ be a random element and $\underline{X}$ a random n–vector defined on the same probability space. Assume there is a convex Borel set $K \subset \mathbb{R}_n$ such that $\mathbb{P}[\underline{X} \in K] = 1$. Let $g : K \times \Lambda \to \mathbb{R}$ be a measurable function on $(K \times \Lambda, \mathcal{B}_K \times \mathcal{G})$ where $\mathcal{B}_K$ denotes the Borel subsets of K. Assume that $g(\cdot, y)$ is a convex function on K for each fixed $y \in \Lambda$ and that $E|g(X, Y)| < \infty$. Then*

$$E[g(\underline{X}, Y)|Y = y] \geq g(E[\underline{X}|Y = y]), Law[Y] - a.s. \tag{2.3}$$

Furthermore, if for $Law[Y]$ almost all $y \in \Lambda$, $Law[X|Y = y]$ is nondegenerate, and if $g(\cdot, y)$ is strictly convex, then strict inequality holds.

Proof. By (2.3), $E[g(\underline{X}, Y)|Y = y] = \int_K g(\underline{x}, y)dP_{\underline{X}|Y}(\underline{x}|y), Law[Y] - a.s.$ where the integral may be taken over K since $I_K(\underline{X}) = 1$ a.s. by assumption. Applying the ordinary Jensen's inequality to $P_{\underline{X}|Y}(\cdot|Y = y)$ and the convex function $g(\cdot, y)$ on the r.h.s. of the last displayed equation we have $E[g(\underline{X}, Y)|Y = y] \geq g(\int_K \underline{x}dP_{\underline{X}|Y}(\underline{x}|y)), Law[Y] - a.s.$ □

Example 2.4.2. Let # be counting measure on $\mathbb{N} = \{0, 1, \cdots\}$ and let m be Lebesgue measure on $\mathbb{R}$. Suppose (X, Y) is a random 2–vector having joint density w.r.t. $\# \times m$, $f(x, y) = \frac{e^{-2y}y^x}{x!} I_{(0,\infty)}(y) = C(y)\frac{y^x}{x!}$ where $C(y)$ doesn't depend on x. $\frac{y^x}{x!}$ depends on x, it is the density (w.r.t. #) of a Poisson r.v. with mean y. $Law[X|Y = y] = Poisson(y)$. Since $\sum_{x=0}^{\infty} \frac{y^x}{x!} = e^y$, we see $f_Y(y) = e^{-y}I_{(0,\infty)}(y)$ is an exponential distribution with mean 1. $E[X|Y] = 1$ a.s. and $Var[X|Y] = 1$ a.s. Similarly, the functional dependence of $f(x, y)$ on y can be concentrated in a factor $e^{-2y}y^x I_{(0,\infty)}(y)$, which is a $Gamma(x + 1, 1/2)$ density except for a normalizing constant $1/(\Gamma(x + 1)(1/2)^{(x+1)})$, so $Law[Y|X] = Gamma(x + 1, 1/2)$. Notice that we did not compute the marginal density of X w.r.t. # to obtain this conditional distribution.

2.5 Order Statistics

Sometimes,$X_1, X_2, \cdots, X_n$ are referred to as a random sample from P_X. Here, n is the sample size or the number of trials. We can construct an "estimator" for P_X given by $\hat{P}_n - \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$. Now for a fixed ω (which gives observed values $X_1(\omega), \cdots, X_n(\omega)$ which are fixed real numbers), $\hat{P}_n$ is a probability measure. Indeed, $\hat{P}_n$ with the random X_i's is a random probability measure on $\mathbb{R}$. If h is a real valued function on $\mathbb{R}$ then $\int h(x)d\hat{P}_n(x) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$, i.e. integration w.r.t. $\hat{P}_n$ amounts to averaging the function h over the sample. This is also a random variable. Thus, $E[\int h(x)d\hat{P}_n(x)] = E[h(X)]$. This equation says that $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$is an **unbiased estimator** of $E[h(X)]$.

If h is a given Borel function of a real variable, then the map $H(P_X) = \int h(x)dP_X(x)$ is a functional defined on all Borel p.m.'s P_X for which the integral exists and is finite (e.g. $h(x) = x$ gives the mean functional). Another functional we may wish to estimate is the minimal α quantile $(0 \leq \alpha \leq 1)$ $F^-(\alpha) = \inf\{x : F(x) \geq \alpha\}$, where F is the c.d.f. of X. For fixed α, $F \to F^-(\alpha)$ is a functional on all distributions on $\mathbb{R}$. Replacing F in the definition by the empirical distribution function $\hat{F}_n$ gives the minimal α sample quantile $\hat{F}_n^-(\alpha)$. In general, we don't have the relation that $E[\hat{F}_n^-(\alpha)] = F^-(\alpha)$. That is, the sample quantile is generally a biased estimator of the true quantile. See Example 2.5.1 below. But the estimate is still a very natural one, and the bias is generally quite small.

Above we spoke of $\hat{P}_n$ as being a "random probability measure". Such a random object is not well defined at this point because we have not introduced a σ-field on the set of probability measures on $\mathbb{R}$. Also, we don't know what it means for $\hat{F}_n$ to be a "random distribution function" since we haven't introduced a σ-field on the set of cumulative distribution functions. However, $\hat{P}_n$ has a very special form since it is discrete, $supp[\hat{P}_n]$ has at most n points (exactly n points if all values in the sample are distinct), and the amount of probability mass at each point is a positive integer times $1/n$ (exactly $1/n$ if the points are distinct). Thus, we can think of the subset of such probability measures, which is "isomorphic" with a Euclidean space. Put less technically, we only need a finite number of numbers to determine $\hat{P}_n$, e.g. n numbers where n is the sample size, since if we know all n observed values then we know $\hat{P}_n$. Similar remarks hold for $\hat{F}_n$. However, the mapping from $\mathbb{R}^n$ to discrete probability measures given by $p(x_1, \cdots, x_n) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}, x_i \in \mathbb{R}, 1 \leq i \leq n$ is not one to one since we can't reconstruct the order of the observations from $\hat{P}_n$. For instance, if π is a permutation of $\{1, \cdot, n\}$, then $p(x_{\pi(1)}, \cdots, x_{\pi(n)})p(x_1, \cdots, x_n)$. Recall that a permutation of $\{1, \cdot, n\}$ is a one to one correspondence (bijective map) of the finite set with (into) itself. Thus, if π is a permutation of $\{1, \cdot, n\}$, then $\{\pi(1), \cdot, \pi(n)\}$ is simply a reordering of $\{1, \cdot, n\}$. This last displayed equation merely states the obvious fact that if we reorder the observations, then we get the same empirical probability.

2.5.1 Order Statistics Introduction

Let $X = (X_1, X_2, \cdots, X_n)$ denote the vector of all observations. Consider the subset of IRn given by $\mathbb{P}^n = \{x \in \mathbb{R}^n : x_1 \leq \cdots \leq x_n\}$. On $\mathbb{P}^n$, the mapping p above is a one to one correspondence, so we can identify the set of possible empirical probability distributions (or empirical c.d.f.'s) with $\mathbb{P}^n$, since given an element of $\mathbb{P}^n$, we can associate a unique empirical probability distribution, and vice versa.

The mapping which "orders" our sample $X = (X_1, \cdots, X_n)$ so that it becomes a random vector taking values in $\mathbb{P}_n$ will be denoted **Sort**, i.e. $\underline{Y} = \mathbf{Sort}(\underline{X})$ means $Y \in \mathbb{P}^n$ and there is a permutation π of $\{1, \cdot, n\}$ such that $Y_i = X_{\pi(i)}$ for all i. $\underline{Y}$ is known as the vector of **order statistics**. Two notations for the i'th component Y_i of $\underline{Y}$ that are frequently used are $X_{(i)}$ and $X_{i:n}$. Intuitively, if we believe the components of X are i.i.d., then $\mathbf{Sort}(\underline{X})$ contains "as much information" about the unknown probability distribution

as the original vector of observations $\underline{X}$. We will show below in fact that given $\mathbf{Sort}(\underline{X})$, it is possible to "reconstruct" $\underline{X}$ in random vector with the same distribution.

Let $\mathbf{Perm}(n)$ denote the set of all permutations of $\{1, \cdots, n\}$, then $\#\mathbf{Perm}(n) = n!$. Note that $\mathbf{Perm}(n)$ has the following properties:

1. If π_1 and π_2 are in $\mathbf{Perm}(n)$ then so is $\pi_1 \circ \pi_2$
2. There is an element $\iota \in \mathbf{Perm}(n)$ s.t. $\iota \circ \pi = \pi \circ \iota$ for every $\pi \in \mathbf{Perm}(n)$
3. For every $\pi \in \mathbf{Perm}(n)$, there is an element $\pi^{-1} \in \mathbf{Perm}(n)$ s.t. $\pi \circ \pi^{-1} = \pi^{-1} \circ \pi = \iota$

These three properties make $\mathbf{Perm}(n)$ into a group under the (group) operation of composition (i.e. $\circ$). Note that $\mathbb{R}$ is a group under $+$, and both $\mathbb{R} - \{0\}$ and $(0, \infty)$ are groups under multiplication. Now define $\zeta \circ \mathbf{Perm}(n) = \{\zeta \circ \pi : \pi \in \mathbf{Perm}(n)\}$. One can show that for all $\zeta \in \mathbf{Perm}(n)$, $\zeta \circ \mathbf{Perm}(n) = \mathbf{Perm}(n)$. We will write $\mathbf{Perm}$ when n is clear from context. To each permutation $\pi \in \mathbf{Perm}(n)$ there corresponds a unique linear transformation $\tilde{\pi}$ on $\mathbb{R}^n$ which reorders the components of a vector $\tilde{\pi}(y_1, \cdots, y_n) = (y_{\pi(1)}, \cdots . y_{\pi(n)})$. One can easily see that the $n \times n$ matrix corresponding to $\tilde{\pi}$ is A where $A_{ij} = 1$ if $\pi(j) = i$ and otherwise $A_{ij} = 0$. Note that there is a single 1 in every row and in every column of A, and the remaining entries are 0. Such a matrix is called a **permutation matrix**. Also, one can show that $A^{-1} = A^T$, i.e. A is an orthogonal matrix.

If π is any permutation, then clearly $\mathbf{Sort}(\tilde{\pi}\underline{x}) = \mathbf{Sort}(\underline{x})$, i.e. if we permute the components of $\underline{x}$ and then rearrange permuted components into ascending order, we obtain the same result as if we didn't permute the components before ordering them. Thus, we say Sort is invariant under coordinate permutations, or simply permutation invariant. Now, we characterize some measurability properties of the mapping $\mathbf{Sort}$: $\mathbb{R}^n \to \mathbb{P}^n$

Theorem 2.5.1. *1. A Borel set B is $\sigma(\boldsymbol{Sort})$ measurable iff it satisfies the following symmetry property: $x \in B$ imples $\tilde{\pi}\underline{x} \in B$ for all $\pi \in \boldsymbol{Perm}$. A function $h : \mathbb{R}^n \to \mathbb{R}$ is $\sigma(\boldsymbol{Sort})$ measurable iff it is invariant under permutations of the variables, i.e. $h \circ \tilde{\pi} = h$ for all $\pi \in \boldsymbol{Perm}$.*

Proof. We claim that for $A \subset \mathbb{P}^n$, $\mathbf{Sort}^{-1}(A) = \bigcup_{\pi \in \mathbf{Perm}} \tilde{\pi}^{-1}(A)$. Let $\underline{x} \in \mathbb{R}^n$ and let $\underline{y} = \mathbf{Sort}(\underline{x})$. If $\mathbf{Sort}(x) \in A$, then $\tilde{\pi}x \in A$ for some permutation π, and hence $\mathbf{Sort}^{-1}(A) \subset \bigcup_{\pi}\{\underline{x} : \tilde{\pi}\underline{x} \in A\}$. If $\tilde{\pi}\underline{x} \in A$ for some π, then $\tilde{\pi}\underline{x} = \mathbf{Sort}(\underline{x})$ since $A \subset \mathbb{P}^n$ and $\mathbf{Sort}(\underline{x})$ is the unique element of $\mathbb{P}^n$ that can be obtained by permuting the components of $\underline{x}$, so $\mathbf{Sort}(x) \in A$, and we have shown that $\bigcup_{\pi}\{\underline{x} : \tilde{\pi}\underline{x} \in A\} \subset \mathbf{Sort}^{-1}(A)$.

Suppose $B \in \sigma(\mathbf{Sort})$, $B = \bigcup_{\pi} \tilde{\pi}^{-1}A$ for some Borel set $A \subset \mathbb{P}^n$. If $\zeta \in \mathbf{Perm}$ then $\tilde{\zeta}^{-1}B = \bigcup_{\pi \in \mathbf{Perm}} \tilde{\zeta}^{-1}\tilde{\pi}^{-1}A = \bigcup_{\pi}(\tilde{\pi} \circ \tilde{\zeta})^{-1}A = \bigcup_{\pi} \tilde{\pi}^{-1}A$. This shows that B is symmetric. Conversely, if B is symmetric, then it is easy to see that $B = \bigcup_{\pi} \tilde{\pi}^{-1}A$ with $A = B \cap \mathbb{P}^n$, and hence B is $\sigma(\mathbf{Sort})$ measurable.

By Theorem 1.5.1, h is $\sigma(\mathbf{Sort})$ measurable iff there is a $g : \mathbb{P}^n \to \mathbb{R}$ s.t. $h = g \circ (\mathbf{Sort})$. It follows that if h is $\sigma(\mathbf{Sort})$ measurable then $h \circ \tilde{\pi} = g \circ (\mathbf{Sort}) \circ \tilde{\pi} = g \circ (\mathbf{Sort}) = h$ since $(\mathbf{Sort}) \circ \tilde{\pi} = \mathbf{Sort}$ for any $\pi \in \mathbf{Perm}$. Conversely, suppose $h = h \circ \tilde{\pi}$ for all $\pi \in \mathbf{Perm}$. Now for every $x \in \mathbb{R}^n$, $\mathbf{Sort}(\underline{x})$ is obtained by a permutation of the components of $\underline{x}$. So $h(\underline{x}) = h(\mathbf{Sort}(\underline{x}))$ and h is $\sigma(\mathbf{Sort})$ measurable by Prop 1.2.3 □

2. Suppose $\underline{X}$ is a random n-vector with i.i.d. components and continuous one dimensional marginal c.d.f. Then $P[\boldsymbol{Sort}(\underline{X}) \in D] = n!P[\underline{X} \in D]$, for $D \subset \mathbb{P}^n$. In particular, if X_1 has a Lebesgue density f,

then under the i.i.d. assumption, $\underline{Y} = \boldsymbol{Sort}(\underline{X})$ *has a Lebesgue density on* $\mathbb{R}^n$ *given by*

$$f_{\underline{Y}}((\underline{y}) = \begin{cases} n! \prod_{i=1}^{n} f(y_i), & \text{if } \underline{y} \in \mathbb{P}^n \\ 0, & \text{if } \underline{y} \notin \mathbb{P}^n \end{cases}$$

Proof. If $D \subset \mathbb{P}^n$ then $P[\mathbf{Sort}(X) \in D] = P[\underline{X} \in \mathbf{Sort}^{-1}(D)] = P[\underline{X} \in \bigcup_{\pi} \tilde{\pi}^{-1}(D)]$, where the union is over all $\pi \in \mathbf{Perm}$. Claim that if $\pi \neq \zeta$, then $P[\underline{X} \in \tilde{\pi}^{-1}(D) \cap \tilde{\zeta}^{-1}(D)] = 0$. Assuming the claim is true, it follows that the sets in the union "essentially disjoint" and hence $P[\underline{X} \in \bigcup_{\pi} \tilde{\pi}^{-1}(D)] = \sum_{\pi} P[\underline{X} \in \tilde{\pi}^{-1}(D)]$. By "essentially disjoint" we mean that the intersection has probability measure 0.

We hope that these claims are fairly obvious, but for the sake of mathematical formalism, we will show that $I_{\bigcup_{\pi} \tilde{\pi}^{-1} D}(\underline{x}) = \sum_{\pi \in \mathbf{Perm}} I_{\tilde{\pi}^{-1} D}(\underline{x})$, for $Law[\underline{X}]$ almost all $\underline{x}$. Taking expectations (i.e. integrating w.r.t. the distribution of $\underline{X}$) of both sides of $P[\underline{X} \in \bigcup_{\pi} \tilde{\pi}^{-1}(D)] = \sum_{\pi} P[\underline{X} \in \tilde{\pi}^{-1}(D)]$. Given $\underline{x}$ the sum on $I_{\bigcup_{\pi} \tilde{\pi}^{-1} D}(\underline{x})$ is the number of sets $\tilde{\pi}^{-1} D$ to which $\underline{x}$ belongs, and $\underline{x}$ belongs to two or more $\tilde{\pi}^{-1} D$ iff there is a pair of distinct permutations π and ζ such that $x \in (\tilde{\pi}^{-1} D) \cap (\tilde{\zeta}^{-1} D)$. However, this means that $\underline{x} = \tilde{\pi}\underline{y} = \tilde{\zeta}\underline{y}$ for some $\underline{y} \in D$, where π and ζ are dinstinct permutations. However, when π and ζ are distinct permutations it is true that $(\tilde{\pi}^{-1} D) \cap (\tilde{\zeta}^{-1} D) \subset \{\underline{x} : x_i = x_j, i \neq j\}$

Note that π and ζ being distinct permutations implies $\pi(k) \neq \zeta(k)$. Suppose $x \in (\tilde{\pi}^{-1} D) \cap (\tilde{\zeta}^{-1} D)$ where π and ζ are distinct permutations. This means $\underline{x} = \tilde{\pi}\underline{y} = \tilde{\zeta}\underline{y}$ for some $\underline{y} \in D$. But because π and ζ are distinct permutations, it follows that $y_{\pi(k)} = y_{\zeta(k)}$, and takin $i = \pi(k)$ and $j = \zeta(k)$, we have $i \neq j$ but $x_i = y_{\pi(k)} = x_j = y_{\zeta(k)}$,and hence $x \in \{x : x_i = x_j, i \neq j\}$. $P_{\underline{X}}[(\tilde{\pi}^{-1} D) \cap (\tilde{\zeta}^{-1} D)] \leq P_{\underline{X}}(\{x : x_i = x_j i \neq j\} = 0$. This equality follows by the assumption that the common c.d.f. of the X_i is continuous. This implies that $P[X_i = x] = 0$ for every $x \in \mathbb{R}$. □

3. *If* $\underline{X}$ *has i.i.d. components with continuous c.d.f., as in part (b), then* $Law[\underline{X} | \boldsymbol{Sort}(\underline{X}) = \underline{y}] = \frac{1}{n!} \sum_{\pi \in \boldsymbol{Perm}} \delta_{\tilde{\pi}\underline{y}}$. *Hence,* $E[h(\underline{X}) | \boldsymbol{Sort}(\underline{X})] = \frac{1}{n!} \sum_{\pi \in \boldsymbol{Perm}} h(\tilde{\pi}\underline{X})$

Proof. For $B \in \mathcal{B}_n$ and $\underline{y} \in \mathbb{P}^n$,let $p(B, \underline{y}) = \frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} \delta_{\tilde{\pi}\underline{y}}(B)$. For fixed $\underline{y}$, $p(\cdot, \underline{y})$ is clearly a p.m. Thus, we need to show $P[\underline{X} \in B | \mathbf{Sort}(\underline{X}) = \underline{y}] = \frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} I_B(\tilde{\pi}\underline{y})$, i.e. that $p(B, \underline{y})$ is a version of $P[\underline{X} \in B | \mathbf{Sort}(\underline{X}) = \underline{y}]$. $\frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} I_B(\tilde{\pi}\underline{y})$ is a Borel function of $\underline{y}$. Thus, we need to check that if

$A \subset \Omega$ which is $\sigma(\mathbf{Sort}(\underline{X}))$ measurable, then

$$\begin{aligned}
\int_A I_B(\underline{X})dP &= \int_A p(B, \mathbf{Sort}(\underline{X}))dP = \int_C p(B, \mathbf{Sort}(\underline{X}))dP_{\underline{X}}(\underline{x}) \qquad (A = \underline{X}^{-1}(C) \text{ for some } C \in \sigma(\mathbf{Sort})) \\
&= \int_{\bigcup_\pi \tilde{\pi}^{-1}D} p(B, \mathbf{Sort}(\underline{X}))dP_{\underline{X}}(\underline{x}) \text{ (by Change of variables and } C = \bigcup_\pi \tilde{\pi}^{-1}D \text{ for some Borel } D \subset \mathbb{P}^n) \\
&= \sum_{\pi \in \mathbf{Perm}} \int_{\tilde{\pi}^{-1}D} p(B, \mathbf{Sort}(\underline{X}))dP_{\underline{X}}(\underline{x}) \qquad (I_{\bigcup_\pi \tilde{\pi}^{-1}D}(\underline{x}) = \sum_{\pi \in \mathbf{Perm}} I_{\tilde{\pi}^{-1}D}(\underline{x})) \\
&= \sum_{\pi \in \mathbf{Perm}} \int_D p(B, \mathbf{Sort}(\tilde{\pi}^{-1}\underline{w}))dP_{\tilde{\pi}\underline{X}}(\underline{w}) \\
&\quad \text{(Change of variables and } \underline{W} = \tilde{\pi}\underline{X},\ Law[\underline{W}] = Law[\tilde{\pi}\underline{X}] = Law[\underline{X}] \\
&\quad \text{since } P_{\tilde{\pi}\underline{X}} = \prod_{i=1}^n P_{X_{\pi(i)}} = \prod_{i=1}^n P_{X_1} = P_{\underline{X}} \text{ because } X_1, \cdots, X_n \text{ all have the same marginal distribution.)} \\
&= n! \int_D p(B, \mathbf{Sort}(\underline{x}))dP_{\underline{X}}(\underline{x}) \qquad (\mathbf{Sort} \text{ is permutation invariant}) \\
&= n! \int_D \left[\frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} I_B(\tilde{\pi}\mathbf{Sort}(\underline{x})) \right] dP_{\underline{X}}(\underline{x}) = \sum_{\pi \in \mathbf{Perm}} \int_D I_B(\tilde{\pi}\mathbf{Sort}(\underline{x}))dP_{\underline{X}}(\underline{x}) \\
&= \int_D \left[\sum_{\pi \in \mathbf{Perm}} I_B(\tilde{\pi}\mathbf{Sort}(\underline{x})) \right] dP_{\underline{X}}(\underline{x}) \\
&= \int_D \left[\sum_{\pi \in \mathbf{Perm}} I_B(\tilde{\pi}\underline{x}) \right] dP_{\underline{X}}(\underline{x}) \qquad (\text{For a given } \underline{x}\ \tilde{\pi}\mathbf{Sort}(\underline{x}) \text{ ranges over the same collection of values as } \tilde{\pi}(\underline{x})) \\
&= \sum_{\pi \in \mathbf{Perm}} \int_D I_B(\tilde{\pi}\underline{x})dP_{\underline{X}}(\underline{x}) = \sum_{\pi \in \mathbf{Perm}} \int_{\tilde{\pi}^{-1}D} I_B(\underline{w})dP_{\tilde{\pi}\underline{X}}(\underline{w}) \sum_{\pi \in \mathbf{Perm}} \int_{\tilde{\pi}^{-1}D} I_B(\underline{w})dP_{\underline{X}}(\underline{w}) \\
&= \int_{\bigcup_\pi \tilde{\pi}^{-1}D} I_B(\underline{x})dP_{\underline{X}}(\underline{x}) = \int_C I_B(\underline{x})dP_{\underline{X}}(\underline{x}) = \int_A I_B(\underline{X})dP
\end{aligned}$$

□

Remark. Note that for each fixed $\underline{y} \in \mathbb{P}^n$, $Law[\underline{X}|\mathbf{Sort}(\underline{X} = \underline{y}]$ is a p.m. on $\mathbb{R}^n$. Given the order statistics, each of the $n!$ possible permutations of the data is equally likely.

2.5.2 Applications

$\underline{X}$ i.i.d. and let $X_{(i)}$ denote the i'th order statistic, the c.d.f. of $X_{(i)}$ is $P[X_{(i)} \le x] = P[\text{at least } i \text{ of } \underline{X} \text{ are} \le x] = P[\sum_{j=1}^n I_{(-\infty,x]}(X_j) \ge i] = \sum_{j=1}^n \binom{n}{j} F(x)^j[1-F(x)]^{n-j}$, similarly the $Bin(n, F(x))$. Assuming the X_i's have Lebesgue density $f_{X_{(i)}}(x) = i\binom{n}{i}F(x)^{i-1}[1-F(x)]^{n-i}f(x) = n\binom{n-1}{i-1}F(x)^{i-1}[1-F(x)]^{n-i}f(x)$

Example 2.5.1. Let X_i be i.i.d. $Unif(0,1)$. Lebesgue density for i'th order statistics is $f_i(x) = \dfrac{n!}{(i-1)!(n-1)!}x^{i=1}(1-x)^{n-i} = \dfrac{\Gamma(n+1)}{\Gamma(i)\Gamma(n+1-i)}x^{i=1}(1-x)^{n-i}$. This is a $Beta(n+1, i)$ distribution. $E[X_{(i)}] = \dfrac{\Gamma(i+1)\Gamma(n-i)}{\Gamma(i)\Gamma(n+1-i)}$.
In particular, we know that $X_{(i)} = \hat{F}_n^-(\alpha)$ for $\alpha \in ((i-1)/n, i/n]$, so $\hat{F}_n^-(\alpha)$ is an unbiased estimator of α only for the particular value $\alpha = i/(n+1)$.

Example 2.5.2. Let X_i be i.i.d. $Exp(1)$. In this setting one achieves a particularly nice result for the joint distribution of the spacings defined by $Y_1 = X_{(1)}$ and $Y_i = X_{(i)} - X_{(i-1)}$, and employ the transformation theory to derive the order statistics joint and marginal distribution based on Jacobians. The inverse transformation is $x_{(i)} = \sum_{j=1}^{i} y_j$. $\frac{d\mathbf{Sort}((\underline{x})}{d\underline{y}}$ is a unit lower triangle matrix, and $|det(\frac{d\mathbf{Sort}((\underline{x})}{d\underline{y}})| = 1$. The joint Lebesgue density of $X_{(i)}$ is $f(x_{(1),\cdots,x_(n)} = n! \exp\left[-\sum_{i=1}^{n} x_{(i)}\right]$, $\sum_{i=1}^{n} x_{(i)} = \sum_{i=1}^{n}\sum_{j=1}^{i} y_j = \sum_{j=1}^{n}\sum_{i=j}^{n} y_j = \sum_{j=1}^{n}(n-j+1)y_j$, so the Lebesgue density of Y_i's $f(\underline{y}) = n! \exp\left[-\sum_{j=1}^{n}(n-j+1)y_j\right] = \prod_{j=1}^{n}(n-j+1)\exp[-(n-j+1)y_j]$ with $Law[Y_i] = Exp\left(\frac{1}{n-i+1}\right)$

Example 2.5.3. Let X_i be i.i.d. $N(\mu, \sigma^2)$ $Range(\underline{X}) = \max(\underline{X}) - \min(\underline{X}) = X_{(n)} - X_{(1)}$. In quality control, it is common to estimate the standard deviation σ by a multiple of the sample range, i.e. to use the estimator $\hat{\sigma}_R = C_n Range(\underline{X})$ where C_n is chosen to have $E[\hat{\sigma}_R] = \sigma$. $X_i = Z_i\sigma + \mu$ are i.i.d. $N(\mu, \sigma^2)$. $Range(\underline{X}) = \sigma Range(\underline{Z})$, $E[C_n Range(\underline{X})] = C_n\sigma E[Range(\underline{Z})]$. If we take $C_n^{-1} = E[Range(\underline{Z})]$, then $E[C_n Range(\underline{X})] = \sigma$. $E[Range(\underline{Z})] = E[\max(\underline{Z})] - E[\min(\underline{Z})] = E[\max(\underline{Z})] + E[\max(-\underline{Z})] = E[\max(\underline{Z})] + E[\max(\underline{Z})] = 2E[\max(\underline{Z})]$. The Lebesgue density for $Z_{(n)}$ is $f(z) = n\Phi(z)^{n-1}\phi(z)$, $E[\max(\underline{Z})] = n\int_{\mathbb{R}} z\Phi(z)^{n-1}\phi(z)dz$

Example 2.5.4. Let $\underline{U}$ be i.i.d. $Unif[0,1]$. If $F(x)$ is a given c.d.f., then $X_i = F^-(U_i)$ with i.i.d. marginal c.d.f. F (Proposition 1.2.4). If $\underline{V} = \mathbf{Sort}(\underline{U})$, then $\underline{Y} = \mathbf{Sort}(\underline{X}) = (F^-(V_1), \cdots, F^-(V_n))$. Assuming $Law[X_i]$ has a Lebesgue density $f(x)$, one can show that $\frac{dv_i}{dy_i} = f(y_i)$. If $i < j$, then $f_{Y_i,Y_j}(y_i, y_j) = f_{V_i,V_j}(F(y_i), F(y_j))f(y_i)f(y_j)$. Now to compute a bivariate marginal Lebesgue density for V_i and V_j with $i < j$, we will use the integration formulae

$$\int_0^{v_i}\cdots\int_0^{v_2} dv_1\cdots dv_{i-1} = \frac{1}{(i-1)!}v_i^{i-1}$$
$$\int_{v_j}^{1}\cdots\int_{v_{n-1}}^{1} dv_n\cdots dv_{j+1} = \frac{1}{(n-j)!}(1-v_j)^{n-j}$$
$$\int_{v_i}^{v_j}\cdots\int_{v_i}^{v_{i+2}} dv_{i+1}\cdots dv_{j-1} = \frac{1}{(j-i+1)!}(v_j - v_i)^{j-i+1}$$

From these it follows that

$$f_{V_i,V_j}(v_i, v_j) = \frac{n!}{(i-1)!(j-i+1)!(n-j)!}v_i^{i-1}(F(y_j) - F(y_i))^{j-i+1}(1-F(y_j))^{n-j}$$
$$f_{Y_i,Y_j}(y_i, y_j) = \frac{n!}{(i-1)!(j-i+1)!(n-j)!}F(y_i)_i^{i-1}(F(y_j) - F(y_i))^{j-i+1}(1-F(y_j))^{n-j}f(y_i)f(y_j), y_i < y_j$$

2.5.3 Further Result

Let $\underline{U}_n$ be a random vector with i.i.d. $Unif[0,1]$, and $\underline{V}_n = \mathbf{Sort}(\underline{U}_n)$. For$1 \le i \le j \le n$, $\underline{V}_n[i:j] = (V_{i,n}, V_{i+1,n}, \cdots, V_{j,n})$. If $i > j$, then $\underline{V}_n[i:j]$ is an empty vector with no component. We also deine $V_{0,n} = 0$ and $V_{n+1,n} = 1$. To determine $Law[\underline{V}_n[i:j]|\underline{V}_n[1:(i-1)], \underline{V}_n[(j+1):n]]$, we need the Lebesgue density of $\underline{V}_n$, $f_{\underline{V}_n}(\underline{v}) = n!$ for $0 \le v_1 \le \cdots \le v_n \le 1$. Thus, we could have

$$f_{\underline{V}_n[i:j]|\underline{V}_n[1:(i-1)],\underline{V}_n[(j+1):n]}(\underline{v}[i:j]|\underline{v}[1:(i-1)], \underline{v}[(j+1):n]) = \frac{f_{\underline{V}_n}(\underline{v})}{f_{\underline{V}_n[1:(i-1)]|\underline{V}_n[(j+1):n]}(\underline{v}[1:(i-1)]|\underline{v}[(j+1):n])}$$

The numerator is constant in the region where density of $\underline{V}_n$ is positive, so as a function of $\underline{v}[i:j]$, the conditional density is constant, i.e. it is a uniform density on the region of $\mathbb{R}^{j-i+1}$ where it is positive. Thus, it is only necessary to determine the region where it is positive, which clearly is $v_{i-1} \leq v_i \leq v_{i+1} \leq \cdots \leq v_j \leq v_{j+1}$. Note however that this is the Lebesgue density of the order statistics of $j-i+1$ i.i.d. random variables with the uniform distribution on $[v_{i-1}, v_{j+1}]$. Thus,

$$Law[\underline{V}_n[i:j]|\underline{V}_n[1:(i-1)] = \underline{v}[1:(i-1)]\&\underline{V}_n[(j+1):n] = \underline{v}[(j+1):n]] = Law[(v_{j+1}-v_{i-1})\underline{V}_{j-i+1} + v_{i-1}]$$

Using above equation and Example 2.5.4 related to uniform random variable transform, one can show that if $\underline{X}$ has i.i.d. components with Lebesgue density f, then denoting the order statistics by $X_{(1)} \leq \cdots \leq X_{(n)}$, we have for instance $Law[X_{(2)}, \cdots, X_{(n-1)}|X_{(1)} = x_{(1)}, X_{(n)} = x_{(n)}]$ has a Lebesgue density on $\mathbb{R}^2_n$ given by

$$f(x_{(2)}, \cdots, x_{(n-1)}|x_{(1)}, x_{(n)}) = \frac{(n-2)! \prod_{i=1}^{n-1} f(x_{(i)})}{[F(x_{(n)}) - F(x_{(1)})]^{n-2}}, \qquad \text{for } x_{(1)} \leq \cdots \leq x_{(n)}$$

Proposition 2.5.2. *Let $\underline{X}$ be as in Theorem 2.5.1 (2). Then $\boldsymbol{Rank}(\underline{X})$ has a uniform distribution on $\boldsymbol{Perm}(n)$, i.e. $P[\boldsymbol{Rank}(\underline{X}) = \pi] = \dfrac{1}{n!}$ for all $\pi \in \boldsymbol{Perm}(n)$.*

References

[1] D.D. Cox , "Mathematical Statistics for Data Scientist." Chapter 2.

[2] J. Shao , "Mathematical Statistics." Chapter 1, 2.

[3] E.L. Lehmann and G. Caasella, "Theory of Point Estimation." Chapter 1.