1. Week 12 Supplementary material

Remark 1.1 (Interpretation of parameters appearing in the p.d.f. of a Continuous RV). In the examples of continuous RVs discussed in this course, we have seen that certain parameters appear in the description of p.d.fs. If we specify the values of these parameters, then we obtain a specific example of distribution from a family of possible distributions. In certain cases, we have already been able to interpret them in terms of properties of the distribution of the RV. For example, if $X \sim N(\mu, \sigma^2)$, then $\mu = \mathbb{E}X$ and $\sigma^2 = Var(X)$. We list some interpretation of these parameters.

(a) (Location parameter) If we have a family of p.d.f.s $f_\theta, \theta \in \Theta$, where θ is a real valued parameter (i.e. $\Theta \subseteq \mathbb{R}$) and if $f_\theta(x) = f_0(x - \theta), \forall x \in \mathbb{R}$, then we say that θ is a location parameter for the family of distributions given by the p.d.f.s f_θ. In this case, the family is called a location family and the p.d.f. f_0 is free of θ, i.e. does not depend on θ. We can restate this fact in terms of the corresponding RVs X_θ as follows: the p.d.f./distribution of $X_\theta - \theta$ does not depend on θ.

(b) (Scale parameter) If we have a family of p.d.f.s f_θ, where θ is a real valued parameter (i.e. $\Theta \subseteq \mathbb{R}$) and if $f_\theta(x) = \frac{1}{\theta} f_1(\frac{x}{\theta}), \forall x \in \mathbb{R}$, then we say that θ is a scale parameter for the family of distributions given by the p.d.f.s f_θ. In this case, the family is called a scale family and the p.d.f. f_1 is free of θ, i.e. does not depend on θ. We can restate this fact in terms of the corresponding RVs X_θ as follows: the p.d.f./distribution of $\frac{1}{\theta} X_\theta$ does not depend on θ.

(c) (Location-scale parameter) If we have a family of p.d.f.s $f_{\mu,\sigma}$ with $\sigma > 0$ and if $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) = f_{0,1}(x), \forall x \in \mathbb{R}$, then we say that (μ, σ) is a location-scale parameter for the family of distributions given by the p.d.f.s $f_{\mu,\sigma}$. In this case, the family is called a location-scale family and the p.d.f. $f_{0,1}$ is free of (μ, σ), i.e. does not depend on (μ, σ). We can restate this fact in terms of the corresponding RVs $X_{\mu,\sigma}$ as follows: the p.d.f./distribution of $\frac{X_{\mu,\sigma} - \mu}{\sigma}$ does not depend on (μ, σ).

(d) (Shape parameter) Some family of p.d.f.s also has a shape parameter, where changing the value of the parameter affects the shape of the graph of the p.d.f..

Example 1.2. (a) The family of RVs $X_{\mu,\theta} \sim Cauchy(\mu, \theta), \mu \in \mathbb{R}, \theta > 0$ with the p.d.f.

$$f_{\mu,\theta}(x) = \frac{\theta}{\pi} \frac{1}{\theta^2 + (x-\mu)^2}, \forall x \in \mathbb{R}$$

is a location-scale family with location parameter μ and scale parameter θ.

(b) For the family of RVs $X_\alpha \sim Gamma(\alpha, 1), \alpha > 0$ with the p.d.f.

$$f_\alpha(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \text{ if } x > 0, \\ 0, \text{ otherwise} \end{cases}$$

α is a shape parameter.

Proposition 1.3. *Let $X_1, \cdots, X_n$ be a random sample of continuous RVs with the common DF F and the common p.d.f. f. The joint p.d.f. of $(X_{(1)}, \cdots, X_{(n)})$ is given by*

$$g(y_1, \cdots, y_n) = \begin{cases} n! \prod_{i=1}^n f(y_i), \text{ if } y_1 < \cdots < y_n, \\ 0, \text{ otherwise.} \end{cases}$$

Further the marginal p.d.f. of $X_{(r)}$ is given by

$$g_{X_{(r)}}(y) = \frac{n!}{(r-1)!(n-r)!} (F(y))^{r-1} (1 - F(y))^{n-r} f(y), \forall y \in \mathbb{R}.$$

Proof. Observe that a sample value $(y_1, \cdots, y_n)$ of $(X_{(1)}, \cdots, X_{(n)})$ is related to a sample $(x_1, \cdots, x_n)$ of $(X_1, \cdots, X_n)$ in the following way

$$(y_1, \cdots, y_n) = (x_{(1)}, \cdots, x_{(n)}),$$

$x_{(r)}$ being the r-th smallest of $x_1, \cdots, x_n$. Note that $y_r = x_{(r)}$.
Now, the actual values $x_1, \cdots, x_n$ may have been arranged in a different order than $x_{(1)}, \cdots, x_{(n)}$. In fact, the values $x_{(1)}, \cdots, x_{(n)}$ arise from one of the $n!$ permutations of the values $x_1, \cdots, x_n$. But, any such transformation/permutation is obtained by the action of a permutation matrix on the vector $(x_1, \cdots, x_n)$. For example, if $x_1 < x_2 < \cdots < x_{n-2} < x_n < x_{n-1}$, then $x_{(1)} = x_1, \cdots, x_{(n-2)} = x_{n-2}, x_{(n-1)} = x_n, x_{(n)} = x_{n-1}$ which interchanges the $n-1$ and n-th values, i.e. x_{n-1} and x_n.

Hence, the Jacobian matrix for this transformation is the same as the corresponding permutation matrix and the Jacobian determinant is ± 1.
Since $X_1, \cdots, X_n$ are i.i.d., the joint p.d.f. of $(X_1, \cdots, X_n)$ is given by

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = f(x_1) \times \cdots \times f(x_n), \forall (x_1, \cdots, x_n) \in \mathbb{R}^n.$$

Therefore, we have the joint p.d.f. of $(X_{(1)}, \cdots, X_{(n)})$ is given by

$$g(y_1, \cdots, y_n) = \begin{cases} n! \prod_{i=1}^n f(y_i), \text{ if } y_1 < \cdots < y_n, \\ 0, \text{ otherwise.} \end{cases}$$

The marginal p.d.f. of $X_{(r)}$ can now be computed for $y \in \mathbb{R}$,

$$g_{X_{(r)}}(y)$$
$$= \int_{y_{r-1}=-\infty}^{y} \int_{y_{r-2}=-\infty}^{y_{r-1}} \cdots \int_{y_1=-\infty}^{y_2} \int_{y_{r+1}=y}^{\infty} \int_{y_{r+2}=y_{r+1}}^{\infty} \cdots \int_{y_n=y_{n-1}}^{\infty} n! \prod_{i=1}^{n} f(y_i)\, dy_n dy_{n-1} \cdots dy_{r+1} dy_1 dy_2 \cdots dy_{r-1}$$

The above integral simplifies to the result stated above. □

Example 1.4. Let X_1, X_2, X_3 be a random sample from $Uniform(0, 1)$ distribution. The common p.d.f. here is given by

$$f(x) = \begin{cases} 1, \text{ if } x \in (0, 1) \\ 0, \text{ otherwise.} \end{cases}$$

By the above result, the joint p.d.f. of $(X_{(1)}, X_{(2)}, X_{(3)})$ is given by

$$g(y_1, y_2, y_3) = \begin{cases} 6, \text{ if } 0 < y_1 < y_2 < y_3 < 1, \\ 0, \text{ otherwise.} \end{cases}$$

and the marginal p.d.f. of $X_{(1)}$ is

$$g(y_1) = \begin{cases} 3(1 - y_1)^2, \text{ if } y_1 \in (0, 1) \\ 0, \text{ otherwise.} \end{cases}$$

Remark 1.5. For random samples from discrete distributions, there is no general formula or result which helps in computing the joint distribution of the order statistics. Usually they are done by a case-by-case analysis. Let X_1, X_2, X_3 be a random sample from $Bernoulli(p)$ distribution, for some $p \in (0,1)$. The common p.m.f. here is given by

$$f(x) = \begin{cases} p, \text{ if } x = 1 \\ 1-p, \text{ if } x = 0 \\ 0, \text{ otherwise.} \end{cases}$$

Note that $X_{(1)}$ is also a $\{0,1\}$-valued RV with $X_{(1)} = \min\{X_1, X_2, X_3\} = 1$ if and only if $X_1 = X_2 = X_3 = 1$. Then using independence,

$$\mathbb{P}(X_{(1)} = 1) = \mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 1) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1)\mathbb{P}(X_3 = 1) = p^3$$

and $\mathbb{P}(X_{(1)} = 0) = 1 - \mathbb{P}(X_{(1)} = 1) = 1 - p^3$. Therefore, $X_{(1)} \sim Bernoulli(p^3)$. Similarly, $X_{(3)} \sim Bernoulli(1-(1-p)^3)$. The distribution of $X_{(2)}$ is left as an exercise in the problem sets.

Note 1.6 (Moments do not determine the distribution of an RV)**.** Let $X \sim N(0,1)$ and consider $Y = e^X$. The distribution of Y is usually called the lognormal distribution, since $\ln Y = X \sim N(0,1)$. Using standard techniques, we can compute the p.d.f. of Y:

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} y^{-1} \exp\left[-\frac{(\ln y)^2}{2}\right], \text{ if } y > 0, \\ 0, \text{ otherwise.} \end{cases}$$

It can be shown that the continuous RVs $X_\alpha, \alpha \in [-1,1]$ with the p.d.fs

$$f_{X_\alpha}(y) = f_Y(y)\left[1 + \alpha \sin(2\pi \ln y)\right], \forall y \in \mathbb{R}$$

has the same moments as Y. However, the distributions are different. This shows that the moments of an RV do not determine the distribution. (see the article 'On a property of the lognormal distribution' by C.C. Heyde, published in Journal of the Royal Statistical Society: Series B, volume 29 (1963).)

Note 1.7 (Operations on DFs)**.** Recall that a DF $F : \mathbb{R} \to [0, 1]$ is characterized by the properties that it is right continuous, non-decreasing and $\lim_{x\to\infty} F(x) = 1, \lim_{x\to-\infty} F(x) = 0$. Given two DFs $F, G : \mathbb{R} \to [0, 1]$ and $\alpha \in [0, 1]$, we make the following observations.

(a) (Convex combination of DFs) The function $H : \mathbb{R} \to [0, 1]$ defined by $H(x) := \alpha F(x) + (1 - \alpha)G(x), \forall x \in \mathbb{R}$ has the relevant properties and hence is a DF.

(b) (Product of DFs) The function $H : \mathbb{R} \to [0, 1]$ defined by $H(x) := F(x)G(x), \forall x \in \mathbb{R}$ has the relevant properties and hence is a DF. In particular, F^2 is a DF, if F is so.

In fact, a general DF can be written as a convex combination of discrete DFs, absolutely continuous DFs and singular continuous DFs. We do not discuss such results in this course.

Definition 1.8 (Conditional Expectation, Conditional Variance and Conditional Covariance)**.** Let $X = (X_1, X_2, \cdots, X_{p+q})$ be a $p + q$-dimensional random vector with joint p.m.f./p.d.f. f_X. Let the joint p.m.f./p.d.f. $Y = (X_1, X_2, \cdots, X_p)$ and $Z = (X_{p+1}, X_{p+2}, \cdots, X_{p+q})$ be denoted by f_Y and f_Z, respectively. Let $h : \mathbb{R}^p \to \mathbb{R}$ be a function. Let $z \in \mathbb{R}^q$ be such that $f_Z(z) > 0$.

(a) The conditional expectation of $h(Y)$ given $Z = z$, denoted by $\mathbb{E}(h(Y) \mid Z = z)$, is the expectation of $h(Y)$ under the conditional distribution of Y given $Z = z$.

(b) The conditional variance of $h(Y)$ given $Z = z$, denoted by $Var(h(Y) \mid Z = z)$, is the variance of $h(Y)$ under the conditional distribution of Y given $Z = z$.

(c) Let $1 \leq i \neq j \leq p$. The conditional covariance between X_i and X_j given $Z = z$, denoted by $Cov(X_i, X_j \mid Z = z)$, is the covariance between X_i and X_j under the conditional distribution of (X_i, X_j) given $Z = z$.

Notation 1.9. On $\{z \in \mathbb{R}^q : f_Z(z) > 0\}$, consider the function, $g_1(z) := \mathbb{E}(h(Y) \mid Z = z)$. We denote the RV $g_1(Z)$ by $\mathbb{E}(h(Y) \mid Z)$. Similarly, define the RVs $Var(h(Y) \mid Z)$ and $Cov(X_1, X_2 \mid Z)$

Proposition 1.10. *The following are properties of Conditional Expectation, Conditional Variance and Conditional Covariance. Here, we assume that the relevant expectations exist.*

(a) $\mathbb{E}h(Y) = \mathbb{E}(\mathbb{E}(h(Y) \mid Z)$.

(b) $Var(h(Y)) = Var(\mathbb{E}(h(Y) \mid Z)) + \mathbb{E}Var(h(Y) \mid Z)$.

(c) $Cov(X_1, X_2) = Cov(\mathbb{E}(X_1 \mid Z), \mathbb{E}(X_2 \mid Z)) + \mathbb{E}Cov(X_1, X_2 \mid Z)$.

Proof. We only prove the first statement under a simple assumption. The general case and other statements can be proved using appropriate generalization.

Take $p = q = 1$ and let $X = (Y, Z)$ be a 2-dimensional continuous random vector. Then,

$$\begin{aligned}\mathbb{E}(\mathbb{E}(h(Y) \mid Z) &= \int_{-\infty}^{\infty} \mathbb{E}(h(Y) \mid Z = z) f_Z(z)\, dz \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} h(y) f_{Y|Z}(y \mid z)\, dy\right] f_Z(z)\, dz \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(y) f_{Y,Z}(y, z)\, dydz \\ &= \mathbb{E}h(Y).\end{aligned}$$

□

Example 1.11. Suppose $\binom{X_1}{X_2} \sim N_2\left(\binom{\mu_1}{\mu_2}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$. Hence, $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$, and ρ is the correlation between X_1 and X_2. Assume that $|\rho| < 1$. Here, the joint p.d.f. is given by

$$\begin{aligned}&f_{X_1,X_2}(x_1, x_2) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right],\end{aligned}$$

for all $(x_1, x_2) \in \mathbb{R}^2$. The conditional distribution of X_1 given $X_2 = x_2 \in \mathbb{R}$ is described by the conditional p.d.f.

$$\begin{aligned}f_{X_1|X_2}(x_1 \mid x_2) &= \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sigma_1^2(1-\rho^2)}\left\{x_1 - \left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\right)\right\}^2\right], \forall x_1 \in \mathbb{R}\end{aligned}$$

and hence $X_1 \mid X_2 = x_2 \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2))$. Similarly, for $x_1 \in \mathbb{R}$, $X_2 \mid X_1 = x_1 \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2))$.

Using the conditional distributions obtained above, we conclude

$$\mathbb{E}[X_1 \mid X_2 = x_2] = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2),$$

$$Var[X_1 \mid X_2 = x_2] = \sigma_1^2(1-\rho^2),$$
$$\mathbb{E}[X_2 \mid X_1 = x_1] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1),$$
$$Var[X_2 \mid X_1 = x_1] = \sigma_2^2(1-\rho^2).$$

Example 1.12. Let $X_1, X_2, \cdots$ be i.i.d. $Uniform(0,\theta)$ RVs, for some $\theta > 0$. The sequence $\{X_n\}_n$ being i.i.d. means that the collection $\{X_n : n \geq 1\}$ is mutually independent and that all the RVs have the same law/distribution. Here, the common p.d.f. and the common DF are given by

$$f(x) = \begin{cases} \frac{1}{\theta}, \text{ if } x \in (0,\theta), \\ 0, \text{ otherwise} \end{cases} \quad , \quad F(x) = \begin{cases} 0, \text{ if } x < 0, \\ \frac{x}{\theta}, \text{ if } 0 \leq x < \theta, \\ 1, \text{ if } x \geq \theta. \end{cases}$$

Consider $X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$. Using Proposition 1.3, we have the marginal p.d.f. of $X_{(n)}$ is given by

$$g_{X_{(n)}}(x) = \begin{cases} \frac{n}{\theta^n}x^{n-1}, \text{ if } x \in (0,\theta), \\ 0, \text{ otherwise.} \end{cases}$$

Then,

$$\mathbb{E}X_{(n)} = \int_0^\theta x\,\frac{n}{\theta^n}x^{n-1}\,dx = \frac{n}{n+1}\theta, \quad \mathbb{E}X_{(n)}^2 = \int_0^\theta x^2\,\frac{n}{\theta^n}x^{n-1}\,dx = \frac{n}{n+2}\theta^2$$

and

$$Var(X_{(n)}) = \theta^2\left[\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right] = \theta^2\frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} = \theta^2\frac{n}{(n+2)(n+1)^2}.$$

Now, $\lim_n \mathbb{E}X_{(n)} = \theta$ and $\lim_n Var(X_{(n)}) = 0$. Hence, $\{X_{(n)}\}_n$ converges in 2nd mean to θ and also in probability.

Remark 1.13 (Convergence in probability does not imply convergence in r-th mean). Consider a sequence of discrete RVs $\{X_n\}_n$ with $X_n \sim Bernoulli(\frac{1}{n}), \forall n$. Consider $Y_n := nX_n, \forall n$. Then Y_n's

are also discrete with the p.m.f.s given by

$$f_{Y_n}(y) = \begin{cases} 1 - \frac{1}{n}, \text{ if } y = 0, \\ \frac{1}{n}, \text{ if } y = n, \\ 0, \text{ otherwise.} \end{cases}$$

For all $\epsilon > 0$, we have $\mathbb{P}(|Y_n| \geq \epsilon) = \frac{1}{n} \xrightarrow{n\to\infty} 0$ and hence $Y_n \xrightarrow[n\to\infty]{P} 0$. But, for any $r > 1$, $\mathbb{E}|Y_n|^r = n^{r-1}, \forall n$. Here, $\{Y_n\}_n$ does not converge to 0 in r-th mean.

Example 1.14. Let $\{X_n\}_n$ be i.i.d. RVs with the common distribution $Bernoulli(p)$ for some $p > 0$. Here, we may visualize X_n's as a sequence of coin tosses with probability of success (obtaining head) as p. By the WLLN, $\bar{X}_n \xrightarrow[n\to\infty]{P} \mathbb{E}X_1 = p$, i.e. for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - p| \geq \epsilon) = 0$. This supports the intuitive notion that by tossing a coin, with unknown p, a large number of times we can make an educated guess about the value of p.

Example 1.15. Continuing with the discussion of the previous example, we can justify the working methodology of assigning probabilities by a relative frequency approach. Suppose we repeat a random experiment n times and observe whether an event E occurs or not in each trial. For $i = 1, 2, \cdots, n$, we consider an RV X_i to be 1 if E occurs and 0 otherwise. Here, $X_i \sim Bernoulli(p)$, where $p = \mathbb{P}(E)$. If p is unknown, then by the WLLN we have $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{P} p$, i.e., the observed relative frequency $\frac{1}{n}\sum_{i=1}^n X_i$ in first n trials approximates p in probability, for large n.