Probability: An Extremely Concise Review

1. A scalar-valued random variable X is completely characterized by its distribution function

$$F_X(u) = \mathrm{P}\left(X \leq u\right).$$

This is also called the **cumulative distribution function** (cdf). $F_X(u)$ is monotonically increasing in u, goes to one as $u \to \infty$ and goes to zero as $u \to -\infty$.

2. If F_X is continuously differentiable, then we can also characterize X using its **probability density function**

$$f_X(x) = \left.\frac{\mathrm{d}F_X(u)}{\mathrm{d}u}\right|_{u=x}$$

The density has the properties $f_X(x) \geq 0$ and

$$\int_{-\infty}^{\infty} f_X(x)\ \mathrm{d}x = 1.$$

Events of interest are subsets[1] of the real line — given such an event/subset $\mathcal{E}$, we can compute the probability of $\mathcal{E}$ occurring as

$$\mathrm{P}\left(\mathcal{E}\right) = \int_{x\in\mathcal{E}} f_X(x)\ \mathrm{d}x.$$

[1] Technically, it must be a subset of the real line that can be written as some combination of countable unions, countable intersections, and complements of intervals. You really have to know something about real analysis to construct a set that does not meet this criteria.

3. The **expectation** of a function $g(X)$ of a random variable is

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \ \mathrm{d}x.$$

This is the "average value" of $g(X)$ in that given a series of realizations $X = x_1, X = x_2, \ldots,$ of X,

$$\frac{1}{M}\sum_{m=1}^{M} g(x_m) \to \mathrm{E}[g(X)], \quad \text{as } M \to \infty.$$

This fact is known as the **(weak) law of large numbers**.

4. The **moment** of X of degree p is the expectation of the monomial $g(x) = x^p$. The zeroth moment is always 1:

$$\mathrm{E}[X^0] = \mathrm{E}[1] = \int_{-\infty}^{\infty} f_X(x) \ \mathrm{d}x = 1,$$

and the first moment is the **mean**:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x) \ \mathrm{d}x.$$

The **variance** is the second moment minus the mean squared:

$$\mathrm{var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mathrm{E}[(X - \mathrm{E}[X])^2].$$

This is sometime referred to as the "variation around the mean". Aside from the zeroth moment, there is nothing that says that the integrals above must converge; it is easy to construct examples of well-defined random variables where $\mathrm{E}[X] = \infty$.

5. A pair of random variables (X, Y) are completely described by their **joint distribution function** (joint cdf)[2]

$$F_{X,Y}(u, v) = \mathrm{P}\left(X \leq u, Y \leq v\right).$$

Again, if $F_{X,Y}$ is continuously differentiable, (X, Y) is also characterize by the density

$$f_{X,Y}(x, y) = \left.\frac{\partial F_{X,Y}(u, v)}{\partial u\, \partial v}\right|_{(u,v)=(x,y)}.$$

In this case, events of interest correspond to regions in the plane $\mathbb{R}^2$, and the probability of an event occurring is the integral of the density over this region.

6. From the joint pdf $f_{X,Y}(x, y)$, we can recover the individual **marginal pdfs** for X and Y using

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\ \mathrm{d}y,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\ \mathrm{d}x.$$

The pair of densities $f_X(x)$, $f_Y(y)$ tell us how X and Y behave individually, but not how they *interact*.

7. If X and Y do interact in a meaningful way, then observing one of them affects the distribution of the other. If we observe $X = x$, then with this knowledge, the density for Y becomes

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

[2] For fixed $u, v \in \mathbb{R}$, the notation $\mathrm{P}\left(X \leq u, Y \leq v\right)$ should be read as "the probability that X is $\leq u$ and Y is $\leq v$.

This is a density over y; it is easy to check that it is positive everywhere and that it integrates to one. $f_X(y|X = X)$ is called the **conditional density** for Y given $X = x$.

8. We call X and Y **independent** if observing X tells us nothing about Y (and vice versa). This means

$$f_Y(y|X = x) = f_Y(y), \quad \text{for all } x \in \mathbb{R},$$

and

$$f_X(x|Y = y) = f_X(x), \quad \text{for all } y \in \mathbb{R}.$$

(If one of the statements above is true, then the other follows automatically.) Equivalently, independence means that the joint pdf is *separable*:

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y).$$

9. We can always factor the joint pdf in two different ways:

$$f_X(x)f_Y(y|X = x) = f_{X,Y}(x, y) = f_Y(y)f_X(x|Y = y).$$

At this point, we should be comfortable enough with what is going on that we can use $f_Y(y|x)$ as short-hand notation for $f_Y(y|X = x)$. Then we can rewrite the above in its more common form as

$$f_X(x)f_Y(y|x) = f_{X,Y}(x, y) = f_Y(y)f_X(x|y).$$

This factorization also gives us a handy way to compute the marginals:

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y)f_X(x|y)\ \mathrm{d}y.$$

It also yields **Bayes' equation**

$$f_X(x|y) = \frac{f_Y(y|x) f_X(x)}{f_Y(y)},$$

which is a fundamental relation for statistical inference.

10. All of the above extends in the obvious way to more than two random variables. A **random vector**

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}$$

is completely characterized by the density $f_X(\boldsymbol{x}) = f_X(x_1, \ldots, x_D)$ on $\mathbb{R}^D$. In general, we can factor the joint pdf as

$$f_X(\boldsymbol{x}) = f_{X_1}(x_1)\, f_{X_2}(x_2|x_1)\, f_{X_3}(x_3|x_2, x_1) \cdots f_{X_D}(x_D|x_1, \ldots, x_{D-1}).$$

11. The pth moment of a random vector X that maps into $\mathbb{R}^D$ is the collection of expectations of all monomials of order p. The mean of a random vector is a vector of length D:

$$\mathrm{E}[X] = \begin{bmatrix} \mathrm{E}[X_1] \\ \vdots \\ \mathrm{E}[X_D] \end{bmatrix},$$

the second moment is the $D \times D$ matrix of all correlations between entries:

$$\mathrm{E}[XX^{\mathrm{T}}] = \begin{bmatrix} \mathrm{E}[X_1^2] & \mathrm{E}[X_1X_2] & \cdots & \mathrm{E}[X_1X_D] \\ \vdots & & \ddots & \vdots \\ \mathrm{E}[X_DX_1] & \cdots & & \mathrm{E}[X_D^2] \end{bmatrix},$$

the third moment is the $D \times D \times D$ tensor $\mathrm{E}[X \otimes X \otimes X]$, where

$$\left(\mathrm{E}[X \otimes X \otimes X]\right)(i,j,k) = \mathrm{E}[X_i X_j X_k],$$

and so on.