Masters Bridge Program **Summer 2021**

Session 6: Dependence between random variables

Instructor: Bryan Liu *Contact: runjing_liu@berkeley.edu*

Supplemental reading: Pitman, *Probability* Chapter 6.

6.1 Conditional densities

Recall that for a discrete random variable X, the conditional distribution of Y given $X = x$ is

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}.$$

However, if X is a continuous random variable, then the denominator $X = x$ has probability 0, which results in an undefined expression.

In order to define conditional distributions for continuous random variables X and Y, we express the conditional distribution in terms of densities.

Conditional density of Y given $X = x$. For random variables X and Y with joint density $f(x, y)$, for each x with marginal density $f_X(x) > 0$, the conditional density of Y given $X = x$ is the probability density function

$$f_Y(y|X = x) = \frac{f(x, y)}{f_X(x)}. \tag{6.1}$$

Then for any set B,

$$\mathbb{P}(g(Y)|X = x) = \int_B f_Y(y|X = x)\, dy$$

Exercise 1 (Uniform on a triangle)**.** Let X, Y uniformly distribution on the region $\{(x, y) : x \geq 0, y \geq 0, x + y \leq 2\}$. Find $\mathbb{P}(Y > 1|X = x)$.

We can re-arrange Equation (6.1) to arrive at the multiplication rule for densities,

$$f(x, y) = f_X(x) f_Y(y|X = x).$$

Exercise 2 (Gamma and uniform)**.** Suppose X has a Gamma$(2, \lambda)$ distribution, and that given $X = x$, Y is uniform on $(0, x)$.

Here, the Gamma density is

$$f_X(x) = \lambda^2 x e^{-\lambda x}$$

if $x > 0$ and 0 otherwise.

(a) Find the joint density of (X, Y).

(b) Find marginal density of Y.

6.2 Conditional expectations

Let g be a function that maps $\mathbb{R} \mapsto \mathbb{R}$. The **conditional expectation of** $g(Y)$ **given** $X = x$ is

$$\mathbb{E}(g(Y)|X = x) = \int g(y) f_Y(y|X = x)\, dx.$$

Exercise 3. Show,

(a) **Average conditional probability**:

$$\mathbb{P}(Y \in B) = \int \mathbb{P}(Y \in B|X = x) f_X(x)\, dx$$

(b) **The tower property**:

$$E(Y) = \int \mathbb{E}(Y|X = x) f_X(x)\, dx.$$

We can succinctly write the tower property as $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X))$.

Exercise 4 (Expectation of a product)**.** Let X and Y be random variables, and h a function of X. Show that

$$\mathbb{E}(h(X)Y) = \mathbb{E}(h(X)\mathbb{E}(Y|X)).$$

Exercise 5 (Predictions)**.** In a previous session, we saw that if we want to predict Y by a constant b, in order to minimize the MSE

$$\text{MSE}(b) = \mathbb{E}((Y - b)^2),$$

we should set $b = \mathbb{E}(Y)$.

Now suppose the data come in pairs (X, Y), and we get to observe X in order to predict Y. We want to construct a predictor g, which is a function that takes input X and uses $g(X)$ to predict Y.

If we again want to minimize the MSE,

$$\text{MSE}(g(X)) = \mathbb{E}((Y - g(X))^2), \tag{6.2}$$

argue that the optimal g is $g(x) = \mathbb{E}(Y|X = x)$.

6.3 Covariance and correlation

Given two random variables X and Y, their *covariance* measures how these variables move together.

For random variables X and Y, their **covariance** is defined as

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)),$$

where $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$.

Exercise 6 (Alternative formula for covariance)**.** Show that we can compute the covariance as $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

Generally, $\text{Cov}(X, Y)$ is positive if above-average values of X tend to be associated above values of Y. $\text{Cov}(X, Y)$ is negative if above-average values tend to be associated with below-average values of Y, and vice-versa.

The magnitude of the covariance may be hard to interpret. We can normalize the covariance as follows:

For random variables X and Y, their **correlation** is defined as

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X), \text{Var}(Y)}}. \tag{6.3}$$

We can apply the Cauchy-Schwartz inequality to check that the correlation is always between -1 and 1.

Exercise 7 (Empirical correlations)**.** Suppose we observe data pairs $(x_1, y_1), ..., (x_n, y_n)$. Use the plug-in principle, plugging in the empirical distribution for the correlation formula (6.3), to give a formula for the sample correlation.

When two random variables X and Y have zero correlation (or equivalently, zero covariance), we say that the random variables are *uncorrelated*.

Exercise 8 (Independence implies uncorrelated)**.** Show that if random variables X and Y are independence, then X and Y are uncorrelated.

Exercise 9 (Uncorrelated does not imply independence)**.** Let X take values $\{-1, 0, 1\}$ with equal probability. Let $Y = X^2$. Argue that X and Y are not independent but have $\text{Cov}(X, Y) = 0$. .

More generally, the correlation (or covariance) measures the *linear* relationship between two random variables. Two variables may have some other, nonlinear relationship, and using the correlation as a metric to measure their dependence may not be appropriate. See Figure 6.1 for an illustration.

Exercise 10 (Best linear predictor)**.** As in Exercise 5, suppose data comes in pairs (X, Y). In the previous exercise, we saw that the best predictor $g(X)$ was the conditional expectation $\mathbb{E}(Y|X)$.

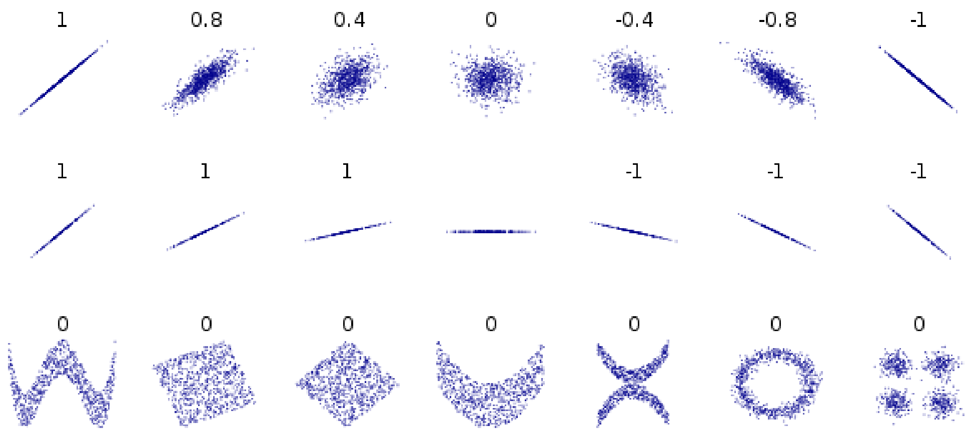

Figure 6.1: Plotted are scatter points of (x, y) pairs, with their empirical correlations (Exercise 7). Correlation measures the linear relationship between two random variables; however, zero correlation does not rule out other, nonlinear relationships. Figure taken from `https://en.wikipedia.org/wiki/Correlation`.

Here, we restrict ourselves to letting g be a linear function, that is, predictors of the form

$$g(X) = \beta X + \gamma$$

for $\beta, \gamma \in \mathbb{R}$.

We again want to minimize the MSE (6.2).

(a) Show that for fixed β, the unique γ which minimizes the MSE if $\gamma^*(\beta) = \mathbb{E}(Y) - \beta\mathbb{E}(X)$.

(b) Now plug in $\gamma = \gamma^*(\beta)$, and show that the optimal β is $\beta^* = \text{Cov}(X, Y)/\text{Var}(X)$.

(c) Apply the plug in principle by computing the variance/covariance under the empirical distribution, and find the sample formula for the best linear predictor. (In contrast, the solutions in (a) and (b) are *population* formulas).