Gov 2000 - 3. Multiple Random Variables

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

WHERE ARE WE? WHERE ARE WE GOING?

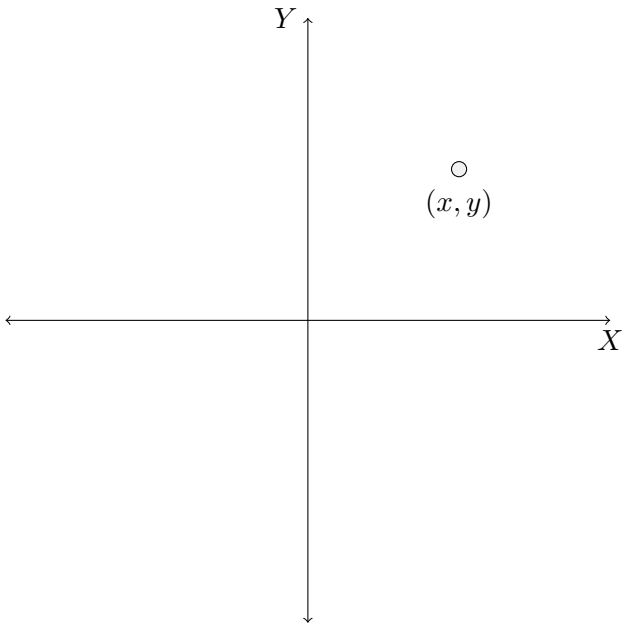
We described a formal way to talk about uncertain outcomes, probability. We've talked about how to use that framework to characterize and summarize the uncertainty in one random variable. This is fine if we want to talk about the mean of some variable or the variance of some variable, but for regression we are going to want to know what the relationships are between variables. To understand those relationships, we need a few more concepts about how to deal with multiple random variables at the same time.

Remember that right now what we are doing is defining things about probability distribution that we might want to learn about. In the coming weeks, we are going to turn to estimating these features of the probability distribution.

JOINT DISTRIBUTIONS

- Remember when we defined probability we talked a lot about joint probabilities of events—what was the probability of A and B occurring: $\mathbb{P}(A \cap B)$. We also talked about the conditional probability of A given that B occurred.
- It turns out that a very important part of statistical inference is thinking about more than one r.v. at the same time. This will be crucial to regression when we want to think about the how the distribution of one variable changes under different values of another variable.

- For instance, we might want to know if changing the number of negative ads in elections changes the distribution of turnout. In order to answer these sorts of questions, we need to understand how to relate two r.v.s.
- Generally, the **joint distribution** of two (or more) variables describes the pairs of observations (one for each covariate) that we are more or less likely to see.
- We're going to think about two r.v.s now, X and Y, each defined on the real line, $\mathbb{R}$. This means that if we think about them together, (X, Y) then one draw from the distribution of this pair will be in a subset of two-dimensional space, or $\mathbb{R} \times \mathbb{R}$:



- Imagine the joint distribution like this: imagine we are throwing darts onto this two-dimensional board. The joint distribution tells us where the darts are more likely to land and where we should see the highest density of darts.
- You should also note that our distributions might be limited to a subset of the real line. If you think about two Uniform variables, then they can only be between 0 and 1 in either dimension. Discrete r.v.s also are typically only defined on the integers. The key is that with two r.v.s, there are now two dimensions to deal with.

Discrete r.v.s

- Let X and Y both be discrete random variables. Just as before we talked about the joint probability of two events occurring, we can also think about the joint probability of these two variables taking certain values.
- **Definition**: The **joint distribution** of (X, Y) can be fully described by the **joint probability mass function**:

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(\{X = x\} \cap \{Y = y\})$$

  To be clear, this is the probability that $X = x$ and that $Y = y$. The second formulation ties this explicitly to the joint probability of two events.

- Given the nature of probabilities, we know a few restrictions on this function: $f_{X,Y}(x, y) \geq 0$ and $\sum_x \sum_y f_{X,Y}(x, y) = 1$. The first is saying that the joint probabilities can't be negative and the second is saying that something must happen—the sum of the probabilities across all pairs of outcomes is 1.
- With discrete r.v.s this is very similar to thinking about a cross-tab, with frequencies/probabilities in the cells instead of raw numbers.

	Favor Gay Marriage $Y = 1$	Oppose Gay Marriage $Y = 0$	Marginal
Female $X = 1$	0.3	0.21	0.51
Male $X = 0$	0.22	0.27	0.49
Marginal	0.52	0.48	1

- One question you probably have is how we relate this joint distribution to the distributions over a single variable, which we sometimes call the **marginal distribution**. It turns out there is a simple way to recover the marginal from the joint, a process we call **marginalization** (we marginalize over X in this case).
- **Definition**: If (X, Y) have a joint distribution with mass function $f_{X,Y}$, then the **marginal mass function** for Y is defined as:

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = x) = \sum_x f_{X,Y}(x, y)$$

- Why does this work? You can think of the marginal probability of a particular value (say, supporting gay marriage) as being the union of a bunch of disjoint sets: supporting gay marriage *and* having a certain gender. We know how to deal with the probability of the union of disjoint sets! We add them together! To get the probability that someone, man or woman, would favor gay marriage,

we simply add the probability of a woman supporting gay marriage ($\mathbb{P}(Y = 1, X = 1)$) and the probability that a man supports gay marriage ($\mathbb{P}(Y = 1, X = 0)$).

Continuous r.v.s

- Now, let's think about the case where X and Y are continuous.
- **Definition**: For two continuous r.v.s X and Y, the **joint probability density function** (or joint PDF) $f_{X,Y}(x, y)$ is a function such that:

1. $f_{X,Y}(x, y) \geq 0$ for all values of (x, y),
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$,
3. for any real numbers, a, b, c, d,

$$\mathbb{P}(a < X < b, c < Y < d) = \int_c^d \int_a^b f_{X,Y}(x, y)dxdy$$

- We can recover the marginal PDF of one of the variables by integrating over the distribution of the other variable:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$$

Joint c.d.f.s

- For two r.v.s X and Y, the *joint cumulative distribution function* or joint c.d.f. $F_{X,Y}(x, y)$ is a function such that for finite values x and y,

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

- Deriving p.d.f. from c.d.f.: $f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$
- Deriving c.d.f. from p.d.f: $F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(r, s)drds$

PROPERTIES OF JOINT DISTRIBUTIONS

Independence

- **Definition**: two r.v.s Y and X are independent (which we write $X \perp\!\!\!\perp Y$) if for all sets A and B:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

- Basically, two r.v.s are independent if the probabilities of various events for the two r.v.s are independent. Here's the intuition behind independence: if I tell you the value of one of the variables, it gives you no information about the value of the other variable.
- From this general definition of independence, we can derive the follow property for independent r.v.s (discrete or continuous):

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

  where $f_X(x)$ and $f_Y(y)$ are the marginal p.m.f.s or p.d.f.s for X and Y.
- **Theorem** If X and Y are independent r.v.s, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

- Proof for discrete X and Y:

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_x \sum_y xy \; f_{X,Y}(x,y) \\
&= \sum_x \sum_y xy \; f_X(x) f_Y(y) \\
&= \left(\sum_x x \; f_X(x) \right) \left(\sum_y y \; f_Y(y) \right) \\
&= \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

- Independence of r.v.s is massively important to statistics and regression. If we want to know if there is a causal effect of a treatment (X) on an outcome, we need to worry if the treatment is independent of any background characteristics that might cause a spurious relationship. It turns out that randomizing a treatment will make it independent of any of these backgrounds characteristics.

Covariance

- When we have a joint distribution, we often want to measure the strength of the relationship between the variables. That is, how dependent are they on one another? There are two measures of dependence that we will commonly use, covariance and correlation.
- The covariance between two r.v.s is exactly what it sounds like: it's a measure of how they covary. Do high values of one variable tend to occur with high values of the other? Or do low values of one tend to go with high values of the other?

- Note that covariance measure linear dependence between two r.v.s. It may not pick up non-linear dependencies between variables.
- **Definition**: The **covariance** between two r.v.s, X and Y is defined as:

$$\text{Cov}[X, Y] = \mathbb{E}\Big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\Big]$$

- What is the basic intuition here? Basically, if we X is above its mean when Y is also above its mean and vice versa, then the covariance will be positive. This is because $(X - \mathbb{E}[X])$ and $(Y - \mathbb{E}[Y])$ will either both be positive or both be negative and, thus, their product will be positive.
- When high values of X tend to occur with low values of Y, then when X is above its mean $(X - \mathbb{E}[X]) > 0$, then Y will be below its mean $(Y - \mathbb{E}[Y]) < 0$ and thus the products will be negative. This will lead to negative covariance.
- We can show that $\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
- What should $\text{Cov}[X, Y]$ be when $X \perp\!\!\!\perp Y$? Zero! Why?

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0\end{aligned}$$

- Thus, independence of two r.v.s implies that there is no covariance between them. This makes sense—if there was covariance then we could use one to predict the other and that would violate the intuition behind independence.
- Does $\text{Cov}[X, Y] = 0$ imply that $X \perp\!\!\!\perp Y$? No!
- Let's say that we have $X \in \{-1, 0, 1\}$ with equal probability and $Y = X^2$. Are X and Y independent? No,

$$\mathbb{P}[Y = 1 | X = 1] = 1 \neq \frac{2}{3} = \mathbb{P}[Y = 1]$$

- But notice that $\text{Cov}[X, Y] = 0$:

$$\text{Cov}[X, Y] = \mathbb{E}[XX^2] - \mathbb{E}[X]\mathbb{E}[X^2] = \mathbb{E}[X^3] - 0 \cdot \mathbb{E}[X^2] = \mathbb{E}[X] = 0.$$

- Why can we have zero covariance, but no independence in this example? Because this is an example of non-linear dependence, which covariance doesn't capture. But it is still dependence.

- Properties of covariances:

1. $\text{Cov}[aX + b, cY + d] = ac\text{Cov}[X, Y]$.
2. $\text{Cov}[X, X] = \mathbb{V}[X]$

- Properties of variances that we can state now that we know covariance:

1. $\mathbb{V}[aX + bY + c] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y] + 2ab\text{Cov}[X, Y]$
2. If X and Y independent, $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.

Correlation

- Notice that while it is easy to interpret the sign of the covariance, the magnitude will depend on the scales of X and Y. So it is hard to compare covariances across different sets of r.v.s. A bigger covariance just might be an indication of a different scale, not any "stronger" relationship.
- Correlation is a scale-free measure of linear dependence, so that it is the same regardless of how we might rescale a variable.
- **Definition**: The **correlation** between two r.v.s X and Y is defined as:

$$\rho = \rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}} = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}$$

- Basically what we are doing here is taking the covariance and dividing out the scales of the respective variables.
- The correlation is always between -1 and 1. We call those variables that have correlation of 1 or -1 as **perfectly correlated**, since this implies a perfect, deterministic linear relationship: $Y = a + bX$.

CONDITIONAL DISTRIBUTIONS

- A very important type of distribution is the conditional distribution, which tells us how what the distribution of a variable is given that we know the outcome of another variable.
- **Definition**: The **conditional probability mass function** or conditional pmf of Y conditional of X is

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

- For instance, what's the probability of supporting gay marriage conditional on being a man?

$$f_{Y|X}(y = 1|x = 0) = \frac{\mathbb{P}(\{X = 0\} \cap \{Y = 1\})}{\mathbb{P}(X = 0)} = \frac{0.22}{0.22 + 0.27} = 0.44$$

- We might ask how that differs from the probability of supporting gay marriage conditional on being a woman $\mathbb{P}(Y = 1|X = 1)$. First, why isn't this just going to be $1 - \mathbb{P}(Y = 1|X = 0)$? Here's how we do it:

$$f_{Y|X}(y = 1|x = 1) = \frac{\mathbb{P}(\{X = 1\} \cap \{Y = 1\})}{\mathbb{P}(X = 1)} = \frac{0.3}{0.3 + 0.21} = 0.59$$

- **Definition**: the conditional pdf of a continuous random variable is

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

assuming that $f_X(x) > 0$. Then, we have the following:

$$\mathbb{P}(a < Y < b|X = x) = \int_a^b f_{Y|X}(y|x)dy.$$

Conditional Expectation

- When we were looking at univariate/marginal distribution, we wanted summarize those distributions with a couple of numbers—the mean and variance. Because we are going to be using the conditional distributions a lot in this class, we'll want to do the same for those distribution. This leads us to think about the conditional expectation or conditional mean.

- The conditional expectation is just the mean of some variable given that we know the value of another variable. It might be the mean number of coups given a particular type of political institution or it might be the conditional expectation of ideology given a particular income level. The basic idea behind regression is that we want to understand how these change as change the value of the conditioning variable. Are richer people more conservative than poorer people, on average?

- **Definition**: The **conditional expectation** of Y conditional on $X = x$ is:

$$\mathbb{E}[Y|X = x] = \begin{cases} \sum_y y \; f_{Y|X}(y|x) & \text{discrete } Y \\ \int_{-\infty}^{\infty} y \; f_{Y|X}(y|x)dy & \text{continuous } Y \end{cases}$$

- **Theorem** If X and Y are independent r.v.s, then

$$\mathbb{E}[Y|X] = \mathbb{E}[Y].$$

- This second theorem encodes the basic intuition about independent r.v.s: knowing about X gives us no information about the mean of Y, if they are independent.

- How does this work intuitively? Basically all we are doing here is using the definitions of expectation, plugging in the conditional distribution of Y given X in place of the marginal distribution of Y. Everything else is the same.

- Let's use the gay marriage example from earlier. What's the conditional expectation of support for gay marriage Y given someone is a man $X = 1$? Well, let's just plug things into the formula:

$$\mathbb{E}[Y|X = 1] = 0 \times f(y = 0|x = 0) + 1 \times f(y = 1|x = 0) = 0.44$$

- Notice here that the conditional expectation of the binary variable Y is the conditional probability of $Y = 1$ given a value of X. Thus, the intuitive connection between the mean and the proportion in binary variables continues to be true in the case of conditional expectations.

- For a particular value of X, say x, we can calculate the conditional expectation for that value. But we can also think about $\mathbb{E}[Y|X]$ as we allow X to take on different values. If we define $\mu(X) = \mathbb{E}[Y|X]$, then obviously $\mu(X)$ is a random variable because X is also a random variable.

- Why is this a random variable? Let's say that X is a binary r.v. Then, before we observe X, the conditional expectation is a random variable with two possible values:

$$\mathbb{E}[Y|X] = \begin{cases} \mathbb{E}[Y|X = 0] & \text{with prob. } \mathbb{P}(X = 0) \\ \mathbb{E}[Y|X = 1] & \text{with prob. } \mathbb{P}(X = 1) \end{cases}$$

- Because it's a r.v., the conditional expectation has a mean, $\mathbb{E}[\mathbb{E}[Y|X]]$. This is the average of the conditional means. It also has a variance, $\mathbb{V}[\mathbb{E}[Y|X]]$, which how much the conditional expectation varies between different values of X. If $\mathbb{E}[Y|X=1]$ and $\mathbb{E}[Y|X=0]$ are very different, then $\mathbb{V}[\mathbb{E}[Y|X]]$ will be high.
- As before we wanted some way to relate conditional distributions to marginal distribution, we want to do the same with conditional expectation and the marginal mean. The law of iterated expectations does just that. Intuitively, it says that the average of the conditional expectations is just the overall (marginal) expectation on Y.
- **Theorem** (The Law of Iterated Expectations): If the expectation exist,

$$\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right] = \begin{cases} \sum_x \mathbb{E}[Y|X=x] f_X(x) & \text{discrete } X \\ \int_{-\infty}^{\infty} \mathbb{E}[Y|X=x] f_X(x) dx & \text{continuous } X \end{cases}$$

- Let's go to the gay marriage example. We already calculated the conditional means, $\mathbb{E}[Y|X=1] = 0.59$ and $\mathbb{E}[Y|X=0] = 0.44$. From the above table, we also know the marginal distribution of gender (0.49 female with $X=1$ and 0.51 male with $X=0$). Thus, we can plug these in:

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[Y|X=0] f_X(0) + \mathbb{E}[Y|X=1] f_X(1) \\ &= 0.44 \times 0.51 + 0.59 \times 0.49 \\ &= 0.51 \end{aligned}$$

- Basically, here we are taking the conditional expectation for each outcome of X and weighting them by the probability that this outcome occurs.
- Properties of conditional expectation:

1. $\mathbb{E}[c(X)|X] = c(X)$ for any function $c(X)$. (Basically, any function of X is a constant with regard to the conditional expectation. If we know X, then we also know X^2, for instance.)
2. If $\mathbb{E}[Y^2] < \infty$ and $\mathbb{E}[g(X)^2] < \infty$ for some function g, then

$$\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X], \leq \mathbb{E}[(Y - g(X))^2|X]$$
$$\mathbb{E}[(Y - \mathbb{E}[Y|X])^2] \leq \mathbb{E}[(Y - g(X))^2].$$

- The second property is quite important. It says that the conditional expectation is the function of X that minimized the squared prediction error for Y across any possible function of X.

Conditional Independence

- **Definition**: Two r.v.s X and Y are **conditionally independent** given Z (written $X \perp\!\!\!\perp Y|Z$) if
$$f_{X,Y|Z}(x, y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z).$$
- Implication:
$$\mathbb{E}[Y|X = x, Z = z] = \mathbb{E}[Y|Z = z].$$
- Basically, sometimes variables might only be independent conditional on some other variable. For instance, if X is the number of swimming-related accidents and Y is the number of ice cream cones sold, then these might only be independent conditional on the temperature on a given day Z.

Conditional Variance

- In addition to the conditional expectation, we'll also want to know the conditional variance. Remember, the conditional distribution of Y given X is basically like any other probability distribution, so we are going to want to summarize the center and spread.
- **Definition**: The **conditional variance** of Y given $X = x$ is defined as:
$$\mathbb{V}[Y|X = x] = \begin{cases} \sum_y (y - \mathbb{E}[Y|X = x])^2 \ f_{Y|X}(y|x) & \text{discrete } Y \\ \int_{-\infty}^{\infty} (y - \mathbb{E}[Y|X = x])^2 \ f_{Y|X}(y|x)dy & \text{continuous } Y \end{cases}$$
- Again, $\mathbb{V}[Y|X]$ is a random variable and a function of X, just like $\mathbb{E}[Y|X]$. With a binary X:
$$\mathbb{V}[Y|X] = \begin{cases} \mathbb{V}[Y|X = 0] & \text{with prob. } \mathbb{P}(X = 0) \\ \mathbb{V}[Y|X = 1] & \text{with prob. } \mathbb{P}(X = 1) \end{cases}$$
- We can also relate the marginal variance to the conditional variance and the conditional expectation.
- **Theorem** (Law of Total Variance/EVE's law):
$$\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]]$$
- The total variance can be decomposed into the average of the within group variance ($\mathbb{E}[\mathbb{V}[Y|X]]$) and how much the average varies between groups ($\mathbb{V}[\mathbb{E}[Y|X]]$).

SUMS OF NORMAL DISTRIBUTIONS

- First off, let's remember what the Normal distribution is. A variable that has a Normal distribution with mean μ and variance σ^2, which we write $X \sim N(\mu, \sigma^2)$, has the following pdf:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \Big\}$$

- The mean and the variance completely describe a Normal distribution so that if we know those two quantities, then we can answer lots of questions about the distribution.

- Last time we talked about the **standard Normal** distribution which has mean 0 and variance 1. It turns out that we can related any Normal to a standard Normal by subtracting its mean and dividing by the standard deviation. Thus, if we have $X \sim N(\mu, \sigma^2)$, then we also know that $Z = (X - \mu)/\sigma$ has a standard Normal distribution, $Z \sim N(0, 1)$.

- There is another fact about Normals that we can exploit: recentering and rescaling the variable perserves the Normal distribution. Suppose that $X \sim N(\mu, \sigma^2)$, then:

1. $Y = aX + b$ has a Normal distribution with $Y \sim N(\mu + b, a^2\sigma^2)$.

- Now that we know these things, we can also talk about the distribution of sums of independent Normal r.v.s. Suppose that $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ and they are independent, then the sum of the two r.v.s is also normal.

$$(X + Y) \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- You can prove all of these facts using the above properties of means and variances.