Part II, Day 3: Probability and Statistics

Jacob Bradt[1]
jbradt@g.harvard.edu
Harvard Kennedy School

Harvard Public Policy, Health Policy, and Business Administration PhD Math Camp
19 August 2022

[1]These notes are a slight adaptation of previous versions by Robert French, Rebecca Sachs, Todd Gerarden, Wonbin Kang, Sam Richardson, and Jonathan Borck. I am deeply indebted to previous Math Camp TF's for their work on these notes.

Outline

Outline

Random variables

- A random variable is a function that maps outcomes from a sample space to real number
    - The sample space is the set of all possible outcomes of a random variable
- Intuitively, random variables assign a number to each possible outcome of an uncertain event
- Random variables:
    - are usually denoted by an upper-case letter, and
    - can be discrete (e.g., heads or tails) or continuous (e.g., time to event).

Random variables

- Random variables are characterized by a **probability density function (pdf)**, $f_X(x)$, and a **cumulative distribution (density) function (cdf)**, $F_X(x)$:

$$f_X(x) = \Pr(X = x) \qquad\qquad F_X(x) = \Pr(X \leq x)$$

- For discrete random variables, the pdf is also referred to as a **probability mass function**.

Properties of the CDF

The **cdf** F_X has the following properties:

1. for $x_1 \leq x_2$,
$$F_X(x_2) - F_X(x_1) = \Pr(x_1 < X \leq x_2)$$
2. $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to \infty} F_X(x) = 1$
3. $F_X(x)$ is non-decreasing
4. $F_X(x)$ is right continuous: $\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$

Discrete random variables

- If F_x is constant except at a countable number of points (i.e., F_x is a step function), then we say that X is a **discrete random variable**
- The size of the jump in the cdf at x_i is the probability that X takes on the value x_i:

$$p_i = F_X(x_i) - \lim_{x \to x_i^-} F_X(x) = \Pr(X = x_i)$$

- The cdf or probability mass function (pmf) of X is defined as:

$$f_X(x) = \begin{cases} p_i & \text{if } x = x_i, i = 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

- We can write

$$\Pr(x_1 < X \leq x_2) = \sum_{x_1 < x \leq x_2} f_X(x)$$

Continuous random variables

- If F_x can be written as

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

  where $f_X(x)$ satisfies

$$f_X(x) \geq 0, \forall x \in \mathbb{R}$$
$$\int_{-\infty}^{\infty} f_X(t)dt = 1$$

  then we say that X is a **continuous random variable**

- Continuous RVs have a nice relationship between the cdf and pdf; by the Fundamental Theorem of Calculus, where f_X is continuous:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Continuous random variables

- Some (potentially) helpful notation: We call

$$S_X = \{x : f_X(x) > 0\}$$

  the **support** of X

- Note that for $x_2 \geq x_1$,

$$\begin{aligned}\Pr(x_1 < X \leq x_2) &= F_X(x_2) - F_X(x_1) \\ &= \int_{x_1}^{x_2} f_X(t)dt\end{aligned}$$

- Interestingly, for a continuous RV X,

$$\Pr(X = x) = 0$$

  due to the infinitely large support of X

Joint distributions

- Let X, Y be two scalar random variables; the **joint cumulative distribution function** of X, Y is

$$F_{X,Y}(x, y) = \Pr(X \leq x, Y \leq y)$$

- For discrete random variables X, Y, the joint cdf is

$$F_{X,Y}(x, y) = \sum_{u \leq x} \sum_{c \leq y} f_{X,Y}(u, v)$$

where $f_{X,Y} = \Pr(X = x, Y = y)$ is the joint pmf of X, Y

Joint distributions

- For continuous random variables X, Y, the joint pdf is

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{\infty}^{y} f_{X,Y}(u,v) dvdu$$

  where again $f_{X,Y}(x,y)$ is the joint pdf of X, Y

- As in the univariate case,

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

  at the points of continuity of $F_{X,Y}$

Independence

- Suppose X and Y are discrete random variables. Then X and Y are said to be independent if and only if

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$$

- Suppose X and Y are random variables. Then X and Y are said to be independent if and only if

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

Independent and identically distributed

- Suppose X and Y are random variables

- We say that X and Y are **independent and identically distributed (i.i.d. or IID)** if X and Y are independent

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

  and have the same distribution

$$f_X(y) = f_Y(y) \text{ and } F_X(x) = F_Y(y)$$

- *i.i.d.* will come up frequently when doing asymptotic statistics in econometrics

Marginal distributions

- We can recover **marginal distributions** from joint distributions

- From the joint cdf of (X, Y), we can recover the **marginal cdf** of X

$$
\begin{aligned}
F_X(x) &= \Pr(X \leq x) \\
&= \Pr(X \leq x, Y \leq \infty) \\
&= \lim_{y \to \infty} F_{X,Y}(x, y)
\end{aligned}
$$

- We can also recover the **marginal pdfs** from the joint pdf using

$$
f_X(x) = \sum_y f_{X,Y}(x, y) \text{ if discrete}
$$

$$
f_X(x) = \int_{S_y} f_{X,Y}(x, y) dy \text{ if continuous}
$$

Conditional distributions: Discrete random variables

- Consider the discrete random variables (X, Y) and let x be such that $f_X(x) > 0$

- The **conditional pmf** of Y given $X = x$ is

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

  which satisfies the properties necessary for a well-defined pmf for a discrete random variable,

$$f_{Y|X}(y|x) \geq 0 \qquad\qquad \sum_y f_{Y|X}(y|x) = 1$$

- The **conditional cdf** of Y given $X = x$ is then

$$F_{Y|X}(y|x) = \Pr(Y \leq y|X = x) = \sum_{v \leq y} f_{Y|X}(v|x)$$

Conditional distributions: Continuous random variables

- Consider the analogous case of continuous random variables (X, Y)
- For any x such that $f_X(x) > 0$, the **conditional pdf** of Y given $X = x$ is

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

  which is well-defined provided that $f_X(x) > 0$

- The **conditional cdf** is

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(v|x)dv$$

Transformations of random variables: The discrete case

- Let X be a random variable with cdf F_X and define the random variable $Y = h(X)$, where h is a one-to-one function whose inverse h^{-1} exists

- If X is discrete and takes on value $x_1, ..., x_n$ then Y is also discrete and takes on the values $y_i = h(x_i)$ for $i = 1, ..., n$

- The pmf of Y is given by

$$\Pr(Y = y_i) = \Pr(X = h^{-1}(x_i))$$
$$f_Y(y_i) = f_X(h^{-1}(y_i))$$

Transformations of random variables: The continuous case

- If X is continuous and h is increasing, we have that

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(X \leq h^{-1}(y)) = F_X(h^{-1}(y)) \end{aligned}$$

It follows directly that

$$f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}$$

- When h is decreasing, it follows analogously that

$$f_Y(y) = \frac{dF_Y(y)}{dy} = -f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}$$

Random variables: Example problem

Example: Suppose you are a basketball player shooting two free throws. You are an 80% free throw shooter. Assuming independence of your free throw attempts, what is the probability that you make both free throws?

By independence of free throws, we have

$$\mathbb{P}(\text{Make both free throws}) = \mathbb{P}(\text{Make first})\mathbb{P}(\text{Make second}) = .8 \times .8 = .64$$

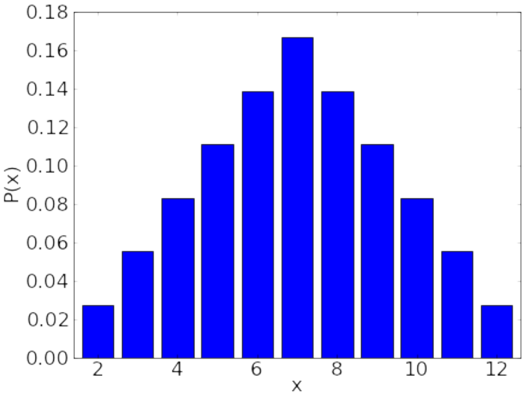
You are still a basketball player shooting two free throws. The probability that you make the first free throw is 80%. If you make the first free throw, the probability that you make the second is 85%. If you do not make the first free throw, the probability that you make the second is 70%. What is the probability that you make both free throws?

Here we can make use of the law of total probability:

$$\begin{aligned}\mathbb{P}(\text{Make both}) &= \mathbb{P}(\text{Make both}|\text{Made } 1^{st})\mathbb{P}(\text{Made } 1^{st}) \\ &\quad + \mathbb{P}(\text{Make both}|\text{Missed } 1^{st})\mathbb{P}(\text{Missed } 1^{st}) \\ &= .85 \times .8 + 0 \times .2\end{aligned}$$

Random variables: Dice problem

Example: You throw two fair dice and construct a random variable, X, equal to the sum of the numbers on the two faces. Fill in a table with the values of the pmf of X, $f_X(x)$, considering all possible outcomes of the random variable. Draw a graph of the $f_X(x)$.



Random variables: Dice problem (continued)

Suppose you throw one die and then the other, and denote the scores by the random variables X and Y. What is the conditional probability that $Y = 4$ given $X = 3$, or $\Pr(Y = 4|X = 3)$? Use the definition of conditional probability.

$$\mathbb{P}(Y = 4|X = 3) = \frac{\mathbb{P}(Y = 4) \cap X = 3)}{\mathbb{P}(X = 3)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

What about $\Pr(Y = 5|X = 3)$? What is the pmf of the conditional distribution $f_{Y|X}(y|3)$?

$$\mathbb{P}(Y = 5|X = 3) = \frac{\mathbb{P}(Y = 5) \cap X = 3)}{\mathbb{P}(X = 3)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

The conditional distribution $f_{Y|X}(y|3)$ is given by

$$f_{Y|X}(y|3) = \begin{cases} \frac{1}{6}, & \text{for } y \in \{4, 5, 6, 7, 8, 9\} \\ 0 & \text{otherwise} \end{cases}$$

Does the conditional distribution depend on the first roll? Why or why not? Yes!

Random variables: Marginal distribution problem

Example: Find the marginal probability density functions, $f_X(x)$ and $f_Y(y)$, of the bivariate distribution characterized by the pdf

$$f_{X,Y}(x,y) = (x+y)1_{0\le x\le 1}1_{0\le y\le 1},$$

where $1_{0\le x\le 1}$ represents the indicator function, which takes on the value one for $0 \le x \le 1$ and zero otherwise. Are X and Y independent random variables?

They are not independent random variables since the product of their marginal distributions does not equal their joint distribution. To see this:

$$\begin{aligned} f_X(x) &= \int_0^1 (x+y)1_{0\le x\le 1}dy \\ &= 1_{0\le x\le 1}(x+\frac{1}{2}) \end{aligned}$$

and by symmetry, $f_Y(y) = 1_{0\le y\le 1}(y+\frac{1}{2})$. By inspection, $f_Y(y)f_X(x) \neq f_{X,Y}(x,y)$.

Outline

Expected value

- The **expected value** can be thought of as a measure of central tendency

- In the scalar case, it is a weighted average of all possible realizations of a random variable X (weighted by the probability of each realization)

- If X is a discrete random variable with J possible realizations, the expected value of X is defined as:

$$\mathbb{E}[X] = \sum_j x_j f_X(x_j) = \sum_{j=1}^{J} x_j \Pr(X = x_j)$$

  if $\sum_x |x| f_X(x) < \infty$, otherwise we say its expectation does not exist

- If X is a continuous random variable, its expectation is defined as

$$\mathbb{E}[X] = \int_{S_x} x f_X(x) dx$$

  if $\int_{S_X} |x| f_X(x) dx < \infty$, otherwise its expectation is said to not exist

Expected value

- Furthermore, the expected value of a function $g(X)$, for discrete X, is given by:

$$\mathbb{E}[g(X)] = \sum_{j=1}^{J} g(x_j) \Pr(X = x_j)$$

- Similarly, the expected value of a function $g(X)$, for continuous X, is given by:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Properties of expected value

1. For any constant c, $\mathbb{E}[c] = c$.
2. For any constants a and b, $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.
3. For any constants $\{a_1, a_2, \cdots, a_n\}$ and random variables $\{X_1, X_2, \cdots, X_n\}$,

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i].$$

4. When $a_i = 1 \ \forall i$, Property 3 reduces to

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Properties 2, 3, and 4 are examples of a general property of the expectation operator known as **linearity of expectation.**

Expected value and the indicator function

- We can use the expectation operator to express probabilities
- An **indicator function**, $\mathbb{1}(A)$ is a function that is equal to one if condition A is true and zero otherwise
- Consider the case where X is a random variable; then

$$\mathbb{1}(X \leq x) = \begin{cases} 1 & \text{if } X \leq x \\ 0 & \text{otherwise} \end{cases}$$

Note that for the continuous case:

$$\begin{aligned} \mathbb{E}[\mathbb{1}(X \leq x)] &= \int_{-\infty}^{\infty} \mathbb{1}(X \leq x) f_X(x) dx \\ &= \int_{-\infty}^{x} f_X(x) dx \\ &= F_X(x) = \Pr(X \leq x) \end{aligned}$$

This useful result comes up often in (applied) econometrics!

Expected value: A few more useful results

- Suppose X, Y are random variables with joint density $f_{X,Y}(x, y)$; if $g(x, y) : \mathbb{R}^2 \to \mathbb{R}$, then we have that

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dydx$$

- By linearity of expectation, $\forall a, b \in \mathbb{R}$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

- If X, Y are independent, then for any $h_1(\cdot)$ and $h_2(\cdot)$

$$\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)]$$

Expected value: Example problem #1

Example: Calculate the expected value of a roll of one die. If you win \$2 times the face value of your toss, what is your expected winning? *Hint: You just learned a property that can be used to simplify the second part of this question.*

Let X be a random variable denoting the value of the roll. We can use the linearity of expectation to calculate the expected value of the dice roll if you win \$2 times the face value of a single toss:

$$\mathbb{E}[2 \times X] = 2 \times \mathbb{E}[X] = 2 \times \left(\frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \right) = 7$$

Expected value: Example problem #2

Example: $f_X(x) = 2x$ for $0 \leq x \leq 1$ and $f_X(x) = 0$ otherwise. Calculate $\mathbb{E}[X]$.

$$\begin{aligned}\mathbb{E}[X] &= \int_0^1 xf(x)dx \\ &= \int_0^1 2x^2 dx \\ &= \frac{2}{3}x^3\Big|_0^1 \\ &= \frac{2}{3}\end{aligned}$$

Expected value: Example problem #3

Example: Upon graduation, you expect to earn \$50,000 with probability 0.2, \$60,000 with probability 0.5, and \$80,000 with probability 0.3. Your utility function for money is $U = \ln(w)$, where w = your wealth (earnings). Calculate your expected earnings and your expected utility upon graduation.

The expected value of your earnings is:

$$\mathbb{E}[w] = \$50,000 \times .2 + \$60,000 \times .5 + \$80,000 \times .3 = \$64,000$$

We can use the Law of the Unconscious Statistician (LOTUS) to find the expected utility:

$$\mathbb{E}[u(w)] = \ln(\$50,000) \times .2 + \ln(\$60,000) \times .5 + \ln(\$80,000) \times .3$$

Conditional expectations

- Given pair of random variables, (X, Y), with joint density $f_{X,Y}(x, y)$, we can define the **conditional expectation** of Y given $X = x$ as

$$\mathbb{E}[Y|X = x] = \int_{S_Y} y \cdot f_{Y|X}(y|x)dy$$

  in the continuous case and

$$\mathbb{E}[Y|X = x] \sum_j y_j \times \Pr(Y = y_j|X = x)$$

  Note that in both cases this is a function of this is a function of x

- You may sometimes see the conditional expectation denoted as $\mu_Y(x)$

- You have likely seen this before: conditional expectation can sometimes be referred to as the **regression function**

Properties of conditional expectations

1. For any function $c(X)$, $\mathbb{E}[c(X)|X = x] = c(x)$.
2. For functions $a(X)$ and $b(X)$, $\mathbb{E}[a(X)Y + b(X)|X] = a(X)\mathbb{E}[Y|X] + b(X)$.
3. If X and Y are independent, then $\mathbb{E}[Y|X] = \mathbb{E}[Y]$.
4. The Law of Iterated Expectations: $\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y]$.

The Law of Iterated Expectations

- This *may* be the most important thing we cover in math camp

- **The Law of Iterated Expectations:**

$$\mathbb{E}_Y[Y] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[Y]\right] \qquad\qquad \mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right]$$

- *Proof:* Using the definitions of the conditional expectation and the conditional pdf:

$$\begin{aligned}
\mathbb{E}_X\left[\mathbb{E}_{Y|X}[Y]\right] &= \int \left(\int y f_{Y|X}(y|x) dy \right) f_X(x) dx \\
&= \int \int y f_{Y|X}(y|x) f_X(x) dy dx \\
&= \int y \left(\int f_{X,Y}(x,y) dx \right) dy \\
&= \int y f_Y(y) dy = \mathbb{E}[Y]
\end{aligned}$$

Moments

- Consider a random variable X
- The k-**th moment** of X is defined as $\mathbb{E}[X^k]$
  - The first moment of X is its mean: $\mathbb{E}[X]$
  - The k-**th centered moment** of X is $\mathbb{E}[(X - \mathbb{E}[X])^k]$
  - The second centered moment of X is its **variance**: $V(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$

Variance and standard deviation

- In contrast to the expectation, which is a measure of central tendency, variance and standard deviation are measures of variability or spread

- As we have seen, the **variance** of the random variable X is:

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

- Variance is also denoted by σ_X^2 or σ^2

- It is frequently useful to express the variance as:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

  $\rightarrow$ We can derive this using our properties of expectations operators!

- The **standard deviation** of a random variable X is the (positive) square root of its variance:

$$\mathrm{sd}(X) = \sigma_X = \sigma = \sqrt{\mathrm{Var}(X)}$$

Properties of variance and standard deviation

1. For any constant c, $\mathrm{Var}(c) = 0$.
2. For any constants a and b, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
3. For any constants a and b, $\mathrm{sd}(aX + b) = |a|\sqrt{\mathrm{Var}(X)}$.
4. For any X, $\mathrm{Var}(X) \geq 0$ and $\mathrm{sd}(X) \geq 0$.

Variance: Examples

Example: $f_X(x) = 2x$ for $0 \le x \le 1$ and $f_X(x) = 0$ otherwise. Calculate $\mathrm{Var}(X)$.
We have already calculated $\mathbb{E}[x] = \frac{2}{3}$. We can calculate $\mathbb{E}[x^2]$ using LOTUS:

$$\mathbb{E}[x^2] = \int_0^1 x^2 f(x)dx = \int_0^1 2x^3 dx = (\frac{2}{4}x^4)\Big|_0^1 = \frac{1}{2}$$

Then, the variance of x is given by $\mathbb{E}[x^2] - \mathbb{E}[x]^2 = \frac{1}{18}$.

Variance: Examples

Example: Calculate the variance and standard deviation of your earnings upon graduation as described in the previous example

We already have that

$$\mathbb{E}[w] = \$50,000 \times .2 + \$60,000 \times .5 + \$80,000 \times .3 = \$64,000$$

Using LOTUS, we have that

$$\mathbb{E}[w^2] = \$50,000^2 \times .2 + \$60,000^2 \times .5 + \$80,000^2 \times .3 = 4,220,000,000$$

Then $\mathrm{Var}(w) = 4,220,000,000 - 64,000^2 = 124,000,000$. To convert the variance measure back to a dollar value, we find the standard deviation:

$$\sigma_w = \sqrt{124,000,060} = \$11,136$$

Covariance

- If X and Y are random variables (discrete or continuous), $\mu_X = \mathbb{E}[X]$, and $\mu_Y = \mathbb{E}[Y]$, then their **covariance** is:

$$\mathrm{Cov}(X, Y) = \sigma_{XY} = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$$

- Exercise: It is often useful to express the covariance as $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mu_X \mu_Y$. Derive this result from the definition of covariance.

Correlation

- The **correlation coefficient** between X and Y is:

$$\mathrm{Corr}(X, Y) = \rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

- The correlation coefficient is effectively a normalization of covariance
- This is closely related to the underlying algebraic formulation of the coefficient recovered from a linear regression of Y on X (or vice versa)
- Whereas the covariance can take on any numerical value, the value of the correlation coefficient always lies between -1 and 1

Properties of covariance and correlation

1. If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. This implies $\mathrm{Corr}(X, Y) = 0$. However, the converse is not true: zero covariance (or a correlation coefficient of zero) does not imply independence.
2. For any constants a_1, a_2, b_1, and b_2, $\mathrm{Cov}(a_1 X + b_1, a_2 Y + b_2) = a_1 a_2 \mathrm{Cov}(X, Y)$.
3. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
4. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
5. $\mathrm{Cov}(X + Y, W + V) = \mathrm{Cov}(X, W) + \mathrm{Cov}(X, V) + \mathrm{Cov}(Y, W) + \mathrm{Cov}(Y, V)$.
6. The Cauchy-Schwartz Inequality implies: $|\mathrm{Cov}(X, Y)|^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y)$.
7. $-1 \leq \mathrm{Corr}(X, Y) \leq 1$. *Note: This is implied by Property 6.*
8. $\mathrm{Corr}(X, Y)$ does not depend on the units of X and Y.

Covariance and correlation: Examples

Example: Consider two independent rolls of a fair die and let X and Y denote the values of the first and second rolls. What is $\mathrm{Cov}(X+Y, X-Y)$? Are the random variables $X+Y$ and $X-Y$ independent? If not, provide an example showing otherwise.

We can manipulate the expression $\mathrm{Cov}(X+Y, X-Y)$ to simplify our problem:

$$\begin{aligned}\mathrm{Cov}(X+Y, X-Y) &= \mathrm{Cov}(X,X) + \mathrm{Cov}(X,-Y) + \mathrm{Cov}(Y,X) + \mathrm{Cov}(Y,-Y) \\ &= \mathrm{Var}(X) - \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Y) - \mathrm{Var}(Y) \\ &= V(X) - V(Y) = 0\end{aligned}$$

Since X and Y are random variables reporting the role of two identical die, we know that their variances must be equal.

By inspection though, we know that if $X+Y=12$, then $X=Y=6$, and so that value of $X-Y$ is fully determined; it's equal to 0. The random variables $X+Y$ and $X-Y$ are not independent.

Covariance and correlation: Examples

Example: Let X and Y be discrete random variables with the joint probability mass function

$$f_{X,Y}(x,y) = \frac{1}{4}, \quad (x,y) \in \{(0,0),(1,1),(1,-1),(2,0)\}$$

and zero otherwise.

1. Find the marginal probability mass function, expectation, and variance of X.
2. Find the marginal probability mass function, expectation, and variance of Y.
3. Compute $\mathbb{E}[XY]$ and use this to find the covariance of X and Y.
4. Are X and Y independent?

Covariance and correlation: Examples

1. Find the marginal probability mass function, expectation, and variance of X.

Let's first find the probability mass function:

$$\mathbb{P}(X=x)=\begin{cases}\sum_{(0,y)\in\{(0,0),(1,1),(1,-1),(2,0)\}} f(0,y)=\frac{1}{4}, \text{ for } x=0\\ \sum_{(1,y)\in\{(0,0),(1,1),(1,-1),(2,0)\}} f(1,y)=\frac{1}{4}+\frac{1}{4}, \text{ for } x=1\\ \sum_{(2,y)\in\{(0,0),(1,1),(1,-1),(2,0)\}} f(2,y)=\frac{1}{4}, \text{ for } x=2\end{cases}$$

The expectation is given by

$$\mathbb{E}[X]=\sum_{x\in\{0,1,2\}} x\mathbb{P}(X=x)0\times\frac{1}{4}+1\times\frac{1}{2}+2\times\frac{1}{4}=1$$

To find the variance, it's easiest to first calculate

$$\mathbb{E}[X^2]=\sum_{x\in\{0,1,2\}} x^2\mathbb{P}(X=x)=0\times\frac{1}{4}+1\times\frac{1}{2}+4\times\frac{1}{4}=\frac{3}{2}$$

Then the variance is $\frac{3}{2}-1=\frac{1}{2}$.

Covariance and correlation: Examples

2. Find the marginal probability mass function, expectation, and variance of Y.

We can take the exact same approach as above, and calculate that

$$\mathbb{P}(Y = y) = \begin{cases} \frac{1}{4}, \text{ for } y = -1 \\ \frac{1}{2}, \text{ for } y = 0 \\ \frac{1}{4}, \text{ for } y = 1 \end{cases}$$

We also have that,

$$\mathbb{E}[Y] = 0$$
$$\mathbb{E}[Y^2] = \frac{1}{2}$$
$$Var(Y) = \frac{1}{2}$$

Covariance and correlation: Examples

3. Compute $\mathbb{E}[XY]$ and use this to find the covariance of X and Y.

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x,y} xyf(x,y) \\
&= \sum_{(x,y)\in\{(0,0),(1,1),(1,-1),(2,0)\}} xy \times \frac{1}{4} \\
&= 0 \times \frac{1}{4} + 1 \times \frac{1}{4} - 1 \times \frac{1}{4} + 0 \times \frac{1}{4} \\
&= 0
\end{aligned}$$

4. Are X and Y independent? No!

Variance of sums of random variables

1. For scalar random variables X and Y and any constants a and b, $\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)$.
2. If X and Y are uncorrelated—that is, if $\mathrm{Cov}(X, Y) = 0$—then
$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$
and
$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$
3. For any constants $\{a_1, a_2, \cdots, a_n\}$ and independent random variables $\{X_1, X_2, \cdots, X_n\}$,
$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i).$$
4. When $a_i = 1\ \forall i$, Property 3 reduces to
$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Outline

The normal distribution

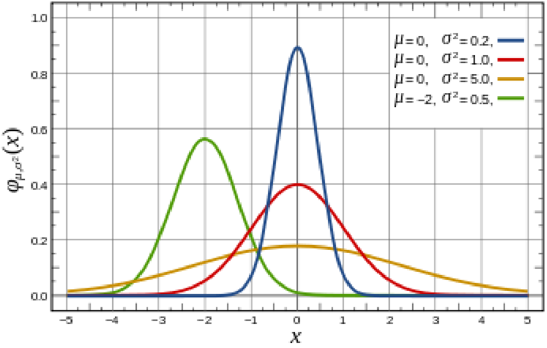
- The normal distribution is pervasive in economics and econometrics
    - In econometrics, this comes up time and again because many econometric results appeal to the Central Limit Theorem (CLT)

- **The normal distribution:**

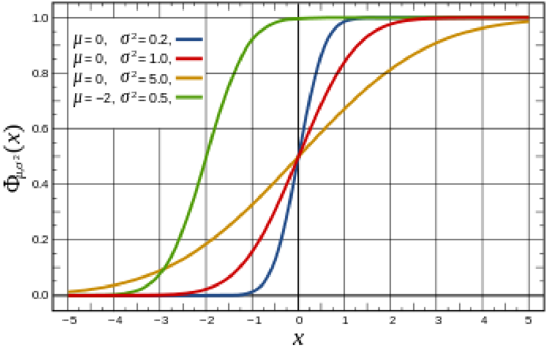
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \quad , \quad -\infty < x < \infty$$

The normal distribution



(a) Normal pdf



(b) Normal cdf

Properties of the Normal Distribution

1. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.
2. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
3. If $\{Y_1, Y_2, \cdots, Y_N\}$ are independent RVs and $Y_i \sim \mathcal{N}(\mu, \sigma^2)\ \forall i$, then $\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$.

Variations on the normal distribution

- **Standard Normal Distribution:** This is a special case of the normal distribution where $\mu = 0$ and $\sigma^2 = 1$:

$$Z \sim \mathcal{N}(0, 1)$$

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-(z)^2}{2}\right] \quad , \quad -\infty < z < \infty$$

- **Lognormal Distribution:** If X is a positive random variable and $Y = \log(X)$ has a normal distribution, X has a lognormal distribution. This distribution is sometimes used to model non-negative economic variables, such as income and market entry costs (i.e., fixed costs).

χ^2 distribution

- The chi-squared distribution is obtained directly from independent, standard normal random variables

- Let $Z_1, Z_2, \cdots, Z_k$ be k independent random variables, each distributed as standard normal

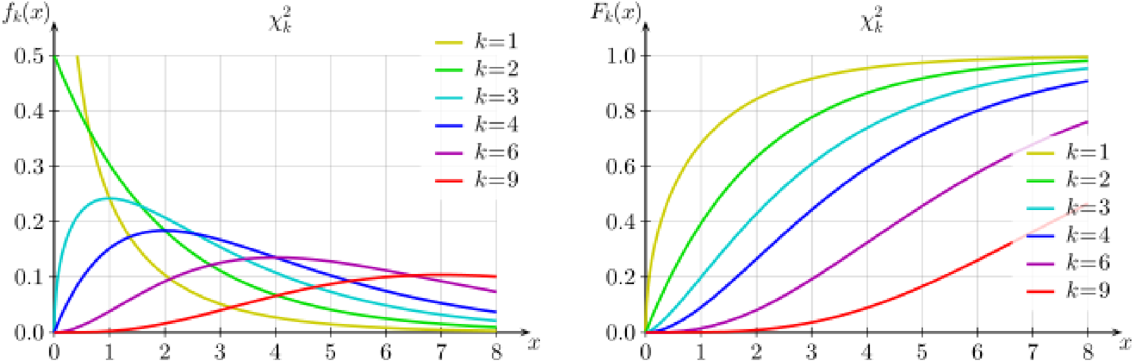
- Define a new variable X as the sum of the squares of the Z_i:

$$X = \sum_{i=1}^{k} Z_i^2$$

  We say that X has a **chi-squared distribution** with k degrees of freedom: $X \sim \chi_k^2$

- *Example:* Wald test statistic

χ^2 distribution



(a) Chi-squared pdf (b) Chi-squared cdf

t distribution

- The t distribution is obtained from a standard normal and a chi-square random variable
- Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_k$ and assume that X and Z are independent. Define a new random variable T:
$$T = \frac{Z}{\sqrt{X/k}}$$
This variable T is distributed according to the t distribution with k degrees of freedom: $T \sim t_k$
- *Example:* Historically, the t distribution has been used for statistical inference with small sample sizes. This is becoming less common in modern empirical analysis due to larger datasets, but it still comes up on occasion.

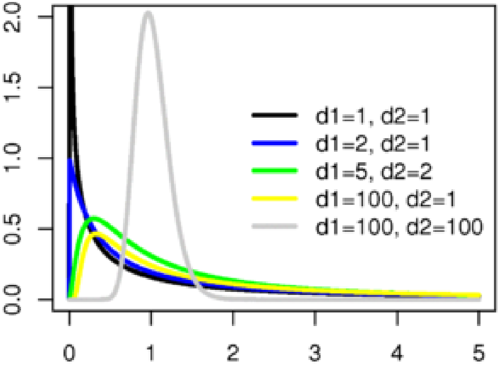
F distribution

- The F distribution is obtained from two chi-square random variables
- Let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$. Furthermore, assume that X_1 and X_2 are independent
- Define a new random variable F:

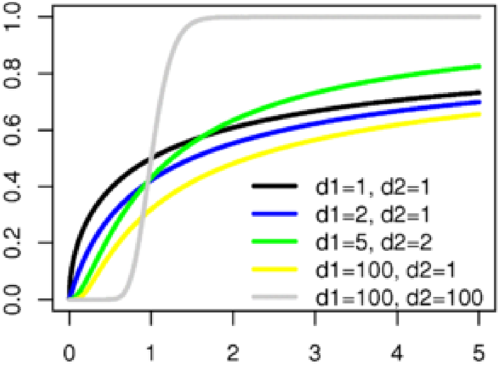
$$F = \frac{X_1 / k_1}{X_2 / k_2}.$$

  We say that F has an F distribution with $(k1, k2)$ degrees of freedom: $F \sim F_{k_1,k_2}$
  - k_1 is often called the degrees of freedom of the numerator, k_2 is often called the degrees of freedom of the denominator
- *Example:* The F-statistic is used in econometrics for testing instruments.

F distribution



(a) F-distribution pdf



(b) F-distribution cdf

Outline

Basic sampling theory

- In your econometrics/statistics courses, you will try to learn something about a **population** of subjects (which could be individuals, firms, cities, states, etc.) by analyzing a **sample** of that population

- Let's call the population parameter whose value we want to know θ

- Mathematical statistics is largely about designing an estimator for θ

Basic sampling theory: Some terminology

- **Estimand:** describes what is to be estimated based on the question of interest; using our notation, this is θ

- **Estimator:** a *rule* (think formula or function) that assigns each possible outcome of the sample a particular value (or distribution of values); call the estimator W.

- **Estimate:** an actual value generated using the estimator based on a particular sample; we call the estimate $\hat{\theta}$

Three praiseworthy properties of estimators

- **Unbiasedness:** $\mathbb{E}[W] = \theta$.
- **Efficiency:** An estimator W_1 is efficient relative to W_2 if $\mathrm{Var}(W_1) < \mathrm{Var}(W_2)$.
- **Consistency:** An estimator W is consistent if $\lim_{n\to\infty} \Pr(|W_n - \theta| \geq \epsilon) = 0$ as the sample size (n) increases to infinity.

Example problem

Example: Suppose we want to characterize the distribution of heights of women in the United States. Assume the height of women in the United States is normally distributed:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

If we choose appropriate estimators and collect some data, we can construct estimates of the parameters (μ, σ^2) of this distribution, conditional on our data. Suppose you hired a research assistant to randomly sample heights, and now you have the following data: $\{X_1, X_2, \cdots, X_n\} = \{60, 62, 64, 66, 68\}$, all in inches.

First, consider the average height. One **estimator** for μ is the sample mean: $W_\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$. What is our **estimate** $\hat{W}_\mu$?

Now, consider the variance of heights. One **estimator** for the population variance, σ^2, is the biased sample variance: $W_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{W}_\mu)^2$. What is our **estimate** $\hat{W}_{\sigma^2}$?

Example problem

Example: Let Y_1, Y_2, Y_3, Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\bar{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$ denote the average of these four random variables.[2]

1. What are the expected value and variance of $\bar{Y}$ in terms of μ and σ^2?
2. Now, consider a different estimator of μ: $W = \frac{Y_1}{8} + \frac{Y_2}{8} + \frac{Y_3}{4} + \frac{Y_4}{2}$. Show that W is also an unbiased estimator of μ. Find the variance of W.
3. Based on your answers to parts (i) and (ii), which estimator of μ do your prefer, $\bar{Y}$ or W? Why?
4. Now, consider a different estimator of μ: $W_a = a_1 Y_1 + a_2 Y_2 + a_3 Y_3 + a_4 Y_4$. What condition is needed on the a_i for W_a to be an unbiased estimator of μ?

[2] Source: Wooldridge 2000, Appendix C, Problem C1.

Example problem

Example: Let Y_1, Y_2, Y_3, Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\bar{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$ denote the average of these four random variables.

1. What are the expected value and variance of $\bar{Y}$ in terms of μ and σ^2?

$$\mathbb{E}[\bar{Y}] = \frac{1}{4}\sum_{i=1}^{4}\mathbb{E}[Y_i] = \frac{1}{4} \times 4\mu = \mu$$

Because Y_i are independent, we can write the variance of the sample mean as,

$$\begin{aligned} Var(\bar{Y}) &= \frac{1}{16}\sum_{i=1}^{4} Var(Y_i) \\ &= \frac{\sigma^2}{4} \end{aligned}$$

Example problem

Example: Let Y_1, Y_2, Y_3, Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\bar{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$ denote the average of these four random variables.

2. Now, consider a different estimator of μ: $W = \frac{Y_1}{8} + \frac{Y_2}{8} + \frac{Y_3}{4} + \frac{Y_4}{2}$. Show that W is also an unbiased estimator of μ. Find the variance of W.

   We can show that W is unbiased:

   $$\mathbb{E}[W] = \frac{\mu}{8} + \frac{\mu}{8} + \frac{\mu}{4} + \frac{\mu}{2} = \mu$$

   The variance of W is given by (again by independence of Y):

   $$\begin{aligned} Var(W) &= \frac{\sigma^2}{64} + \frac{\sigma^2}{64} + \frac{\sigma^2}{16} + \frac{\sigma^2}{4} \\ &= \frac{\sigma^2}{64} + \frac{\sigma^2}{64} + \frac{4\sigma^2}{64} + \frac{16\sigma^2}{64} \\ &= \frac{5\sigma^2}{16} \end{aligned}$$

Example problem

Example: Let Y_1, Y_2, Y_3, Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\bar{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$ denote the average of these four random variables.

3. Based on your answers to parts (i) and (ii), which estimator of μ do your prefer, $\bar{Y}$ or W? Why?
   The sample mean $\bar{Y}$ is more efficient than W. Since they are both unbiased, we should prefer $\bar{Y}$.
4. Now, consider a different estimator of μ: $W_a = a_1 Y_1 + a_2 Y_2 + a_3 Y_3 + a_4 Y_4$. What condition is needed on the a_i for W_a to be an unbiased estimator of μ?
   Taking the expectation of W_a, we have

$$\begin{aligned} \mathbb{E}[W_a] =& a_1 \mathbb{E}[Y_1] + a_2 \mathbb{E}[Y_2] + a_3 \mathbb{E}[Y_3] + a_4 \mathbb{E}[Y_4] \\ =& \mu(a_1 + a_2 + a_3 + a_4) \end{aligned}$$

   And so, we much have that $a_1 + a_2 + a_3 + a_4 = 1$.

Asymptotic statistics

- We often do not want to make strict parametric assumptions in our econometric models — can we sill say something about the behavior of our estimators without these assumptions?

- Yes! In large samples
  - We ask: how would my estimator behave in very large samples ($n \to \infty$)
  - Use the limiting behavior of our estimator in infinite samples to approximate finite sample behavior

- Pro: In infinite samples, most estimators will have simple behaviors

- Con: This is only an approximation!

- Bottom line: Much of econometrics revolves around asymptotic approximations

Convergence

- For a given outcome ω in the sample space Ω, the sequence of random variables $\{X_n\}$ **converges to the random variable** X **almost surely** if

$$\Pr(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$$

  which we can write as $X_n \stackrel{a.s.}{\to} X$

- The sequence of random variables $\{X_n\}$ **converges to the random variable** X **in probability** if $\forall \epsilon > 0$

$$\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) \to 0$$

  which we can write as $X_n \stackrel{p}{\to} X$

Convergence

- The sequence of random variables $\{X_n\}$ **converges to the random variable** X **in mean** if

$$\lim_{n \to \infty} \mathbb{E}[|X_n - X|] = 0$$

  which we can write as $X_n \xrightarrow{m} X$

- Let $\{X_n\}$ be a sequence of random variables and $F_n(\cdot)$ is the cdf of X_n, and let X be a random variable with cdf $F(\cdot)$. $\{X_n\}$ converges in distribution to X if

$$\lim_{n \to \infty} F_n(x) = F(x)$$

  for all points x at which $F(x)$ is continuous. We can write this as $X_n \xrightarrow{d} X$

Convergence

$$\begin{array}{ccc} \text{Almost surely} & & \\ \Downarrow & & \\ \text{In probability} & \Leftarrow & \text{In mean} \\ \Downarrow & & \\ \text{In distribution} & & \end{array}$$

Two main building blocks of asymptotics

1. Law(s) of large numbers
2. Central limit theorem(s)

Laws of large numbers

- Laws of large numbers (LLNs) show that sample averages converge to expectations under certain conditions

- **Weak law of large numbers:** Let $X_1, ...X_n$ be a sequence of random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$ for all i and $\mathrm{Cov}(X_i, X_j) = 0$ forall $i \neq j$. Then,

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

- **Strong law of large numbers:** If $X_1, ..., X_n$ are *iid* with $\mathbb{E}[X_i] = \mu < \infty$, then

$$\bar{X}_n \xrightarrow{a.s.} \mu$$

Central limit theorem

- Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Furthermore, let $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Then, as $n \to \infty$,

$$\sqrt{(}n)\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

- In words, as $n \to \infty$, the mean of a sample is approximately normally distributed *regardless of the underlying distribution of the sample*. We call this **asymptotic normality**.

Outline

Key takeaways

- Be familiar with random variables and the differences/relationships between cdfs, pdfs, joint distributions, marginal distributions, and conditional distributions
- Know how to define expected value and work with expectations operators
- Be familiar with the definitions of cariance, standard deviation, and covariance
- Understand the *intuition* behind basic sampling theory (LLN, CLT)

Additional resources

- Simon and Blume, Appendix 5
- Wooldridge (2000), appendices B and C
- Ashesh Rambachan's 2018 Harvard Econ Math Camp, Probability Notes