Econometrics I

Lecture 1: Probability

Fall 2025

Roadmap

1. Probability spaces
2. Random variables
3. Distribution and density functions
4. Moments of random variables, mean and variance
5. Conditional expectations
6. Multivariate distributions
7. Independence
8. Bayes's Theorem
9. Law of Iterated Expectations

Basic Definitions

To discuss probability we first need a few basic definitions:

- An outcome is something we can observe but may not know in advance

  Example: For a coin flip, H (heads) is an outcome

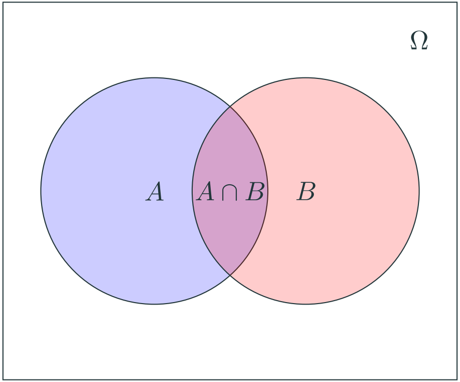
  Example: The wage of a randomly sampled worker
- A sample space, Ω, is a set of all possible outcomes

  Example: For a coin flip, $\{H, T\}$ is the sample space

  Example: For two coin flips, $\{HH, HT, TH, TT\}$ is the sample space
- An event is any subset of the sample space

  Example: For two coin flips, $\{HH, TT\}$ is the event "getting the same side both times"
- A probability is a function from S, the set of all events, to $[0, 1]$ such that
  1. $P(E) \in [0, 1]$ for any event, E
  2. $P(\Omega) = 1$
  3. $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$

Outcomes and Events as a Venn Diagram



- A and B are events in the sample space, Ω
- The intersection, $A \cap B$, is the purple part
- The union, $A \cup B$, would be everything that isn't white

Random Variables

- A random variable assigns numeric values to outcomes:

$$X : \Omega \to \mathbb{R}$$

- We can define a probability distribution on this random variable in terms of our original probability space:

$$\Pr(X = x) = \Pr\left(\bigcup_{\omega : X(\omega) = x} \omega\right)$$

- Example: If rolling two dice, probability of the sum being 11 is given by:

$$\begin{aligned} P(D_1 + D_2 = 11) &= P((D_1 = 5, D_2 = 6) \cup (D_1 = 6, D_2 = 5)) = \\ &= P(D_1 = 5, D_2 = 6) + P(D_1 = 6, D_2 = 5)) \\ &= 1/36 + 1/36 = 1/18 \end{aligned}$$

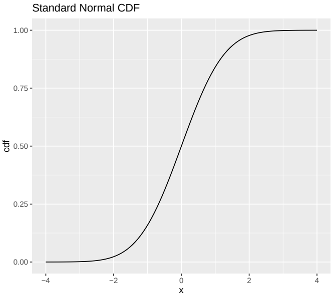
Continuous Random Variables

- For a continuous random variable X the probability that $X = x$ is "0" since any one outcome happens with vanishingly small probability
- Instead we think about sets like $\Pr(a < X < b)$
- Define the Cumulative Distribution Function of X to be:

$$F(x) = \Pr(\omega : X(\omega) \leq x)$$

- n.b.: CDF's are always weakly increasing and live in $[0, 1]$
- Actually, any non-decreasing, right-continuous function F with $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$ is a CDF.

Standard Normal (Gaussian) CDF



if(!(require(ggplot2)))install.packages('ggplot2'

ggplot(data.frame(x = c(-4, 4)), aes(x = x))

+

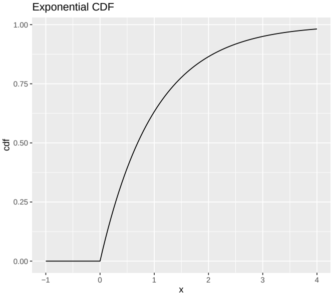
$\text{stat}_f unction(fun = pnorm) + ylab(\text{''}cdf\text{''}) +$

$ggtitle(\text{''}StandardNormalCDF\text{''})$

$$F\left(x\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

n.b.: $\mu = 0$, $\sigma = 1$, plotted here, is what we call the standard normal

Exponential CDF



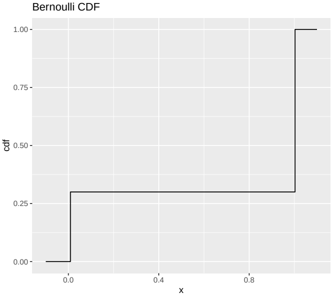
if(!(require(ggplot2)))install.packages('ggplot2'
ggplot(data.frame(x = c(-1, 4)), aes(x = x))
+ stat$_{f}unction(fun = pexp) + ylab(''cdf'') +$
$ggtitle(''ExponentialCDF''))$

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

where $\lambda > 0$ is the rate parameter, plotted here for $\lambda = 1$.

Check properties of exponential CDF

Bernoulli CDF



if(!(require(ggplot2)))install.packages('ggplot2' sfun0 <- stepfun(0:1, c(0., .3, 1.), f = 0) x = seq(-.1, 1.1, length.out = 100) df = data.frame(x = x, y = sfun0(x)) ggplot(df, aes(x,y)) + geom$_s tep() + ylab("cdf") + ggtitle("BernoulliCDF")$

The Bernoulli distribution describes a random variable $X \in \{0, 1\}$.
Plotted is Bernoulli distribution with probability of success $\Pr(X = 1) = p = .7$

CDF: draw your own!

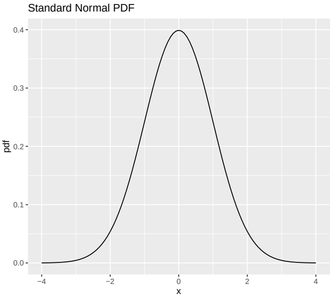
Probability Density Functions

- A random variable's Probability Density Function is the derivative of its CDF:

$$f(x) = \frac{d}{dx} F(x)$$

- PDFs are well-defined when the CDF is absolutely continuous (which requires continuity and almost-everywhere differentiability)

Standard Normal (Gaussian) PDF

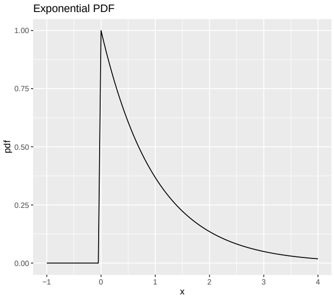


if(!(require(ggplot2)))install.packages('ggplot2'
ggplot(data.frame(x = c(-4, 4)), aes(x = x))
$+$
stat$_f unction(fun = dnorm) + ylab("pdf") +$
$ggtitle("StandardNormalPDF")$

$$f\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

n.b.: $\mu = 0$, $\sigma = 1$, plotted here, is what we call the standard normal

Exponential PDF



if(!(require(ggplot2)))install.packages('ggplot2'
ggplot(data.frame(x = c(-1, 4)), aes(x = x))
+ stat$_f unction(fun = dexp) +$
$ylab("pdf") + ggtitle("ExponentialPDF")$

$$f\left(x\right) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

where $\lambda > 0$ is the rate parameter, plotted here for $\lambda = 1$.

- Note that for discrete distributions like the Bernoulli distribution, we have discontinuous jumps in the CDF, and the PDF is not defined.
- Instead of a PDF, we can define a probability mass function:

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

- The set of points x such that $\Pr(X = x) > 0$ are call the support of X. Similarly, the support of a continuously distributed random variable can be defined as those points x where the pdf $f(x) > 0$.

Expectations

- For a continuously distributed random variable:

$$\mathbb{E}_F[X] \equiv \int x f(x) dx$$

- For a discretely distributed random variable:

$$\mathbb{E}_p[X] \equiv \sum_{x \in Supp(X)} x p(x)$$

- These expectations are also called the mean or first moment of X. Often, $\mu \equiv \mathbb{E}[X]$.
- The F subscript denotes the distribution used to take the expectation. We will often omit it when there is no ambiguity.

Expectations of functions of random variables

- For a continuously distributed random variable:

$$\mathbb{E}_F\left[g\left(X\right)\right] \equiv \int g\left(x\right) f(x)dx$$

- For a discretely distributed random variable:

$$\mathbb{E}_p\left[g\left(X\right)\right] \equiv \sum_{x \in Supp(X)} g\left(x\right) p\left(x\right)$$

- Note that functions of random variables are random variables themselves, so we're not adding much here.

Jensen's Inequality
If g is convex, $\mathbb{E}\left[g\left(X\right)\right] \geq g\left(\mathbb{E}\left[X\right]\right)$.

If g is concave, $\mathbb{E}\left[g\left(X\right)\right] \leq g\left(\mathbb{E}\left[X\right]\right)$.

Application to even moments: If $g(x) = x^{2k}$, and X is a random variable, then g is convex as

$$\frac{d^2 g}{dx^2}(x) = 2k(2k-1)x^{2k-2} \geq 0 \quad \forall x \in \mathbb{R}$$

and so

$$g(\mathbb{E}[X]) = (\mathbb{E}[X])^{2k} \leq \mathbb{E}\left[X^{2k}\right]$$

What does this mean if $k = 1$?

Jensen's Inequality
If g is convex, $\mathbb{E}\left[g\left(X\right)\right] \geq g\left(\mathbb{E}\left[X\right]\right)$.

If g is concave, $\mathbb{E}\left[g\left(X\right)\right] \leq g\left(\mathbb{E}\left[X\right]\right)$.

Application to even moments: If $g(x) = x^{2k}$, and X is a random variable, then g is convex as

$$\frac{d^2 g}{dx^2}(x) = 2k(2k-1)x^{2k-2} \geq 0 \quad \forall x \in \mathbb{R}$$

and so

$$g(\mathbb{E}[X]) = (\mathbb{E}[X])^{2k} \leq \mathbb{E}\left[X^{2k}\right]$$

What does this mean if $k = 1$?

Variance

- The variance of a random variable:

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

  where $\mu = \mathbb{E}\left[X\right]$

- Usually, $\sigma^2 \equiv Var\left[X\right]$

- $\mathbb{E}\left[X^k\right]$ is called the kth (uncentered) moment of X

- Note that knowing first two moments is equivalent to knowing mean and variance.

- Pareto distribution:

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1 - \left(\frac{1}{x}\right)^2 & \text{if } x \geq 1 \end{cases}$$

- What are the mean and variance of this distribution?

Moment-generating function

- The moment generating function of X is

$$m_X(t) = \mathbb{E}\left[\exp(tX)\right],$$

  as long as this expectation is defined for t in a neighborhood of zero.

- $\frac{d^k m_X(0)}{dt^k}$ equals the kth moment of X.

- If two random variables have the same moment generating function, they have the same distribution. [We can use this later].

Multivariate distributions

- Econometric models are usually concerned with several random variables.
- Denote a collection of random variables:

$$\{X_T\}_{t=1}^{T}.$$

  Note that t here does not necessarily have anything to do with time.
- Examples:
  - One of the variables is “dependent” (denoted Y) and the others “explanatory”
  - A measurement observed repeatedly over time (a time series), e.g.,
    - the price of a stock
  - A measurement observed for different individuals (a cross section), e.g.,
    - wages or income
  - Most commonly, we will have a combination of these things: multiple observations of several variables. (panel data)

- We can now define

$$\mathrm{X}_T : \Omega \to \mathbb{R}^T,$$

where $\mathrm{X}_T(\omega) = (X_1(\omega), \ldots, X_T(\omega))$

Joint and marginal distributions

- The joint CDF is a natural extension to the CDF for a single random variable:

$$F_{\mathrm{X}}(\mathrm{x}) = \Pr\left(\omega : X_1(\omega) \leq x_1, \ldots, X_T(\omega) \leq x_T\right)$$

  where $\mathrm{x} = (x_1, \ldots, x_T) \in \mathbb{R}^T$.

- The joint PDF can be defined as $f(\mathrm{x}) = \frac{\partial^T F_{\mathrm{X}}(\mathrm{x})}{\partial x_1 \partial x_2 \ldots \partial x_T}$.
- Marginal distributions refer to the CDF of the individual X_t variables,

$$F_{X_t}(x) = \Pr\left(\omega : X_t(\omega) \leq x\right)$$

  Formally, marginal CDFs can be obtained from joint distributions by integrating over the other variables.

- Covariance is an important property of the joint distribution of random variables:

$$Cov\left(X_1, X_2\right) = \mathbb{E}\left[\left(X_1 - \mathbb{E}\left[X_1\right]\right)\left(X_2 - \mathbb{E}\left[X_2\right]\right)\right]$$

- And more generally for a random vector

$$Cov\left(\mathrm{X}_T\right) = \mathbb{E}\left[\left(\mathrm{X}_T - \mathbb{E}\left[\mathrm{X}_T\right]\right)\left(\mathrm{X}_T - \mathbb{E}\left[\mathrm{X}_T\right]\right)'\right],$$

  which is a $T \times T$ matrix.
- Note that $Cov\left(X, X\right) = Var\left(X\right)$.

Properties of expectations and covariance I

- For constants a, b, you should be able to show that

$$\mathbb{E}\left[a + bX\right] = a + b \cdot \mathbb{E}\left[X\right]$$

- From this, it follows that

$$Cov\left[a + b \cdot X_1, X_2\right] = b \cdot Cov\left[X_1, X_2\right]$$

From this, it follows that

$$Var\left(a + b \cdot X\right) = b^2 \cdot Var\left(X\right)$$

Properties of expectations and covariance II

- For constants a, b, you should be able to show that

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

- From this, it follows that

$$Cov\left[aX + bY, Z\right] = a \cdot Cov\left[X, Z\right] + b \cdot Cov\left[Y, Z\right]$$

Independence

- Two events A, $B \subset \Omega$ are independent iff

$$\Pr(A \cap B) = \Pr(A)\Pr(B)$$

- A collection of random variables X is independent iff

$$F_{\mathrm{X}}(\mathrm{x}) = F_{\mathrm{X}_1}(x_1) F_{\mathrm{X}_2}(x_2) \dots F_{\mathrm{X}_T}(x_T)$$

- A collection of random variables X is independent iff for any (measurable) real-valued functions $g_1, g_2, \dots, g_T$,

$$\mathbb{E}\left(g_1(X_1)\, g_2(X_2) \dots g_T(X_T)\right) = \mathbb{E}\left[g_1(X_1)\right] \mathbb{E}\left[g_2(X_2)\right] \dots \mathbb{E}\left[g_T(X_T)\right]$$

- What does this imply about the relationship between $\mathbb{E}[XY]$ and $\mathbb{E}[X]\,\mathbb{E}[Y]$?

Conditional probability

- The conditional probability of event A given event B can be defined as

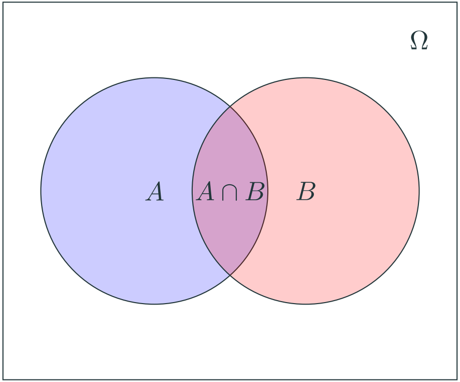
$$\Pr\left(A|B\right) = \frac{\Pr\left(A \cap B\right)}{\Pr(B)}$$

- The conditional CDF of random variable X conditional on $Y = y$ can be written

$$F_X\left(x_0|Y = y\right) = \int_{-\infty}^{x_0} \frac{f\left(x, y\right)}{f_Y\left(y\right)} dx$$

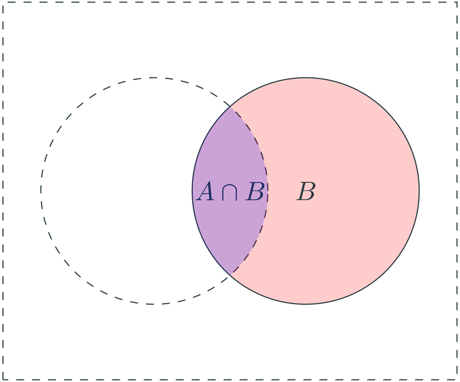
where $f_Y\left(y\right)$ is the marginal density of y

Conditional Probability in a Venn Diagram



- The probability of A, $\Pr(A)$ is the relative area of A (within Ω)
- But what if B definitely occurred?

Conditional Probability in a Venn Diagram



- The probability of A is STILL the relative area of A
- But we only take into account the part of A "inside" B

- From $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$, we have

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$$

- Bayes's Theorem follows:

$$\Pr(A|B) = \Pr(B|A)\frac{\Pr(A)}{\Pr(B)},$$

which is useful to frame Bayesian inference. If A is a theory (or parameter vector) and B is evidence, our updated (posterior) probability of the theory A depends on the probability of the observed evidence conditional on the theory.

Confusion Matrix

		True Value		
		Positive	Negative	Total
Estimate	Positive	TP	FP	P'=TP+FP
	Negative	FN	TN	N'=FN+TN
	Total	$TP + FN$	$FP + TN$	N

- Sensitivity: $= TP/(TP + FN)$ (good at identifying who has the disease)
- Specificity: $= TN/(TN + FP)$ (good at identifying who doesn't)

Imagine a scientist develops a test for a disease.

- The test has a false positive rate $\Pr(Test = + \mid Disease = -) = 5\%$.
- The test has a false negative rate $\Pr(Test = - \mid Disease = +) = 0\%$.

- If 1% of the population has the disease, what is the odds that someone with a positive test has the disease?
- If 0.05% have the disease, what is the odds that someone with a positive test has the disease?

"If you hear hoofbeats, think horses not zebras"

Conditional expectations

- Conditional expectations are extremely important in this course and econometrics broadly.
- The conditional expectation of $g(X)$ given $Y = y$ is

$$\mathbb{E}\left[g(X)\,|Y = y\right] = \int g(x)\,\frac{f(x,y)}{f_Y(y)}dx$$

- Note that this conditional expectation is a value for a given value y, but we can also treat conditional expectations as random variables. In other words, when the conditioning variable Y is a random variable

$$\mathbb{E}\left[g(X)\,|Y\right]$$

is a random variable.

- The Law of Iterated Expectations holds that

$$\mathbb{E}\left[\mathbb{E}\left[g\left(X\right)|Y\right]\right] = \mathbb{E}\left[g\left(X\right)\right]$$

- Define $\varepsilon = Y - \mathbb{E}\left[Y|X\right]$. What is $\mathbb{E}\left[\varepsilon X\right]$?

Linear Algebra Review

Linear Algebra Review

Basic Definitions

- A vector in $\mathbb{R}^n$ is a column of numbers $(x_1, x_2, ..., x_n)$.
- A matrix in $\mathbb{R}^{n \times m}$ is m columns of length n vectors (so n is the number of rows and m is the number of columns). We denote an element of a matrix by m_{ij} for row i and column j:

$$M = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$$

- For an entity on which there are many pieces of data, we store the data in a vector x_i

  Example: For the USA could have $x_{USA} = (GDP_{USA}, Population_{USA}, ...)$
- For many entities we can store all the data in a data matrix, X.

  Example: For two countries:

$$X = \begin{pmatrix} GDP_{USA} & Population_{USA} \\ GDP_{Canada} & Population_{Canada} \end{pmatrix}$$

Matrix Multiplication

- For two vectors of equal length define the dot product as $v \cdot w$ or $\langle v, w \rangle$:

$$v \cdot w = \sum_{i=1}^{n} v_i \times w_i$$

- For two matrices, A and B of sizes $n \times m$ and $m \times k$ define the matrix product $C = AB$ as the $n \times k$ matrix with entries $c_{ij} = \sum_{l=1}^{m} a_{il} b_{lj}$
  - Easy way to remember: $(i, j)^{th}$ element of product is dot product of i^{th} row and j^{th} column of A and B respectively.
  - Not all matrices can be multiplied: left matrix must have column length equal to right matrix's row length
  - Multiplication is NOT commutative: $AB \neq BA$ even if they both exist

https://eli.thegreenplace.net/2015/visualizing-matrix-multiplication-as-a-linear-combination/

Transposes

- Define the transpose of A as the matrix A' with elements $a'_{ij} = a_{ji}$ (reverse columns and rows)
- A matrix is symmetric if $A' = A$
- Important Properties:
  - The matrix $B = A'A$ is always a square matrix
  - The matrix $B = A'A$ is always symmetric
  - $(A')' = A$
  - Multiplication Rule: $(AB)' = B'A'$
  - Addition Rule: $(A + B)' = A' + B'$
- Note that dot products can also be written as an inner product: $v \cdot w = v'w$.
- Note that the outer product of two vectors is a matrix rather than a scalar: vw'.

Inverses

- The Identity Matrix, I, is a matrix with 1s on the diagonal and 0s elsewhere. Clearly $AI = A$.
- Define the left inverse of A to be the matrix A^{-1} such that $A^{-1}A = I$
  - Can analogously define right inverse
  - Right and left inverse will NOT be the same if A is not a square matrix
  - Right and left inverse WILL be equal if A is square (then we just say inverse)
- Important Properties:
  - Multiplication Rule: $(AB)^{-1} = B^{-1}A^{-1}$
  - Tranpose Rule: $(A')^{-1} = (A^{-1})'$
  - Dot Product: $v \cdot w = v'w$

- For a function $f : \mathbb{R}^n \to \mathbb{R}^m$ recall the definition of the derivative or Jacobian of f:

$$Df = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_m}{x_1} \\ \vdots & & \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_m}{x_n} \end{pmatrix}$$

- We WON'T be doing anything too complicated! But we can define two important functions given a vector x and a matrix A:
  - For Ax, $D(Ax) = A$ (as a line in 1-D calc)
  - For $x'Ax$, $D(x'Ax) = x'(A + A')$ (as a quadratic in 1-D calc)

The matrix cookbook https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf has a lot (more) helpful properties.