14.32 Econometric Data Science
Professor Anna Mikusheva
September, 2019

Lectures 1-2

Review of Probability

Random variables and their characteristics

Random variable X is a function from some (unspecified) outcome space to $\mathbb{R}$, sometimes can write it as $X(\omega)$. Random variable is fully characterized by a set of possible values and probabilities with which these values are taken. We will usually talk about discrete and continuous random variables (but that does not cover all possible random variables).

Discrete random variable takes a finite or countable number of values that can be written as a list $x_1, x_2, ...$ with probabilities $p_1, p_2, ...$ Here $p_k = P(X = x_k) > 0$ and $\sum_k p_k = 1$.

Example 1. Bernoulli random variable takes values 1 and 0 with probabilities p and $q = 1 - p$.

Continuous random variable can take any value in an interval and has differentiable *cumulative distribution function* (cdf) $F(x) = P\{X \le x\}$. We can define cdf for any random variable (continuous or not), it is a non-decreasing function with values between 0 and 1. Derivative of cdf for a continuous random variable is called *probability density function* (pdf), denote it $f(x)$. For continuous random variable:

$$P\{a < X \le b\} = F(b) - F(a) = \int_a^b f(s)ds.$$

Example 2. Uniform variable on interval [0,1] takes any value in interval [0,1], with pdf $f(x) = \mathbb{I}\{a < x \le b\}$. Here $\mathbb{I}$ is an indicator function (takes value 1 if the condition in parenthesis is true and zero otherwise).

Expected value for discrete random variable is defined as $EX = \sum_k x_k p_k$, for continuous random variable as $EX = \int x f(x)dx$.

Example 3. For Bernoulli $EX = 1 \cdot p + 0 \cdot q = p$, for uniform $EX = \int_0^1 x dx = 1/2$.

Property of the expectation. It is linear: if for two constants (constants mean non-random) a and b and a random variable X we construct a new random variable $Y = a + bX$, then $EY = a + bEX$.

Variance is another characteristic of distribution, and it measures the spread:

$$Var(X) = E(X - EX)^2 = E(X^2 - 2XEX + (EX)^2) = EX^2 - 2EX \cdot EX + (EX)^2 = EX^2 - (EX)^2.$$

Example 4. For Bernoulli $EX^2 = p$, $Var(X) = p - p^2 = p(1 - p)$.

Property of variance.

$$Var(a + bX) = b^2 Var(X).$$

Gaussian distribution

Notation $X \sim N(\mu, \sigma^2)$ reads random variable X has Gaussian distribution with mean μ and variance σ^2, this means a random variable with pdf:

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

Properties:

Cite as: Anna Mikusheva, course materials for 14.32 Econometric Data Science, Fall 2019. Massachusetts Institute of Technology.

- $\int_{-\infty}^{+\infty} f(x)dx = 1$;
- $\int_{-\infty}^{+\infty} xf(x)dx = \mu$;
- $\int_{-\infty}^{+\infty} (x - mu)^2 f(x)dx = \sigma^2$;
- If $X \sim N(\mu, \sigma^2)$, and $Y = a + bX$, then $Y \sim N(a + b\mu, b^2\sigma^2)$.
- If $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0, 1)$.

Random variable $N(0, 1)$ is called standard normal, its cd is denoted as Φ, it is tabulated, and students are responsible for knowing how to use it. If $X \sim N(\mu, \sigma^2)$, then

$$P\{a < X \leq b\} = P\{\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\} = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Multivariate distributions

We may consider several random variables realized together. Their distribution is known as joint distribution. All concepts below can be written in a meaningful way for all combinations of discrete/ continuous variables, we will write them here as for two continuous variables. All concepts are easy to generalize for more than two dimensions as well.

Joint cdf is connected to pdf in the following way:

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

Joint distribution contains within itself *marginal* distributions, that is, distributions of each variable considered separately:

$$F_X(x) = P\{X \leq x, -\infty < Y < \infty\} = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u, v)dvdu = \int_{-\infty}^{x} f_X(u)du,$$

where $f_X(u) = \int_{-\infty}^{\infty} f_{X,Y}(u, v)dv$ is the marginal pdf.

Conditional distribution of Y given that $X = x$ is given by pdf:

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

It is a pdf, that is $f_{Y|X=x}(y|x) \geq 0$ and $\int f_{Y|X=x}(y|x)dy = 1$. It has the following interpretation: we expected to see two random variables jointly distributed according to pdf $f_{X,Y}(x, y)$, we do see realization of $X = x$, how do we think Y is distributed given this information?

Example 5. Similar concepts can be introduced for discrete random variables. Imagine that for every person there are two variables Y = 'person is employed' and X = 'person has college degree'. For the US labor force's population over 25 years of age in 2012, the distribution was: The column and row with title 'total'

	$Y = 0$	$Y = 1$	total
$X = 0$	0.053	0.586	0.639
$X = 1$	0.015	0.346	0.361
total	0.068	0.932	1

are marginal distributions. If a person has a college diploma, then the probability that s/he is employed is

$$P\{Y = 1|X = 1\} = \frac{P\{X = 1, Y = 1\}}{P\{X = 1\}} = \frac{0.346}{0.361} = 0.958.$$

If a person does not work then the probability of being employed is

$$P\{Y=1|X=0\} = \frac{P\{X=0, Y=1\}}{P\{X=0\}} = \frac{0.586}{0.639} = 0.917.$$

Thus, a person with college education is more likely to be employed than one without.

If the conditional distribution does not depend on the value of the variable one conditions, then two variables are *independent*. An equivalent definition is: for all possible values of x, y:

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

for continuous variables and

$$P(X=x, Y=y) = P(X=x)P(Y=y)$$

for discrete.

Since $f_Y|X(y|x)$ is a distribution of a random variable for any value of x for which $f_X(x) \neq 0$, we can talk about mean and variance of this distribution (assuming they are finite). We will denote expectation as

$$E(Y|X=x) = \int y f_{Y|X}(y|x)dy.$$

This gives a number for any value of x from the range of variable X. Notation $E(Y|X)$ is a random variable, where one plugs random variable X in the previous formula.

A measure of linear association between two random variables is its covariance:

$$cov(X,Y) = E\left[(X-EX)(Y-EY)\right] = E[XY] - E[X]E[Y]$$

Properties of covariance:

- $cov(X,X) = Var(X)$;
- $cov(a+bX, Y) = bcov(X,Y)$;
- $Var(X+Y) = Var(X) + Var(Y) + 2cov(X,Y)$;
- If X and Y are independent then $cov(X,Y) = 0$.

Since covariance depends on the scale, we also introduce unitless version, known as correlation:

$$corr(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

The correlation is bounded by one in absolute value: $|corr(X,Y)| \leq 1$ (due to Cauchy-Schwartz inequality). If correlation equals 1 or -1, then there exists a perfect linear relation between two random variables, that is, for some constants a, b and c we have $aX + bY = c$.

Multivariate normal distribution

Bivariate gaussian distribution is given by pdf

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \exp\left\{-\frac{1}{2(1-\rho_{xy}^2)}\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{1}{2(1-\rho_{xy}^2)}\left(\frac{y-\mu_y}{\sigma_y}\right)^2 - \frac{\rho_{xy}}{(1-\rho_{xy}^2)}\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right\}.$$

Marginals for this joint distribution are: $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$, the correlation between two components is ρ_{xy}.

In general for any n-dimensional vector μ and positive definite $n\times n$ matrix Σ we can define n-dimensional gaussian random vector X by its pdf:

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}.$$

One can show that it is a pdf (integrates to 1), its marginals are gaussian with means collected in μ. Matrix Σ contains all variance and covariance.

Limit Theorems

Variables $X_1, X_2, ..., X_n$ are called i.i.d. (independent and identically distributed) if they all have the same marginal distribution and are independent, that is (in a case of continuous random variables),

$$f_{X_1,...,X_n}(x_1,...,x_n) = f(x_1)...f(x_n).$$

Assume that the marginal distribution has finite absolute mean ($E|X| < \infty$) and $EX_i = \mu$, then due to linearity of expectation we have

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

Here, quantity $\frac{1}{n}\sum_{i=1}^n X_i$ is called a sample average. The Law of Large Numbers claims that the sample average converges to the mean when the sample size increases to infinity:

$$\frac{1}{n}\sum_{i=1}^n X_i \to^p \mu \text{ as } n\to\infty.$$

Here, $\to^p$ stays for convergence in probability and means that for any $\varepsilon > 0$ we have

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| > \varepsilon\right\} = 0.$$

One way of proving the Law of Large Numbers (under a stronger assumption of finite variance) is to check that $Var(\frac{1}{n}\sum_{i=1}^n X_i) \to 0$.

$$Var\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{\sigma^2}{n} \to 0.$$

The Central Limit Theorem characterizes how much the sample average deviates from the mean in a large sample. In particular, if $X_1, X_2, ..., X_n$ are i.i.d. random variables with $EX_i = \mu$ and $Var(X_i) = \sigma^2$, then

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right) \Rightarrow N(0,1) \text{ as } n\to\infty.$$

Sign $\Rightarrow$ denotes weak convergence (convergence in distribution), and essentially says that the cdf of the left-hand-side variable will be very close to the cdf of the standard normal distribution when n is large. In particular, since standard gaussian falls in interval [-1,1] with probability approximately 68%, into interval [-2,2]- approximately 95%, and into interval [-3,3] with probability 99%; in large samples the sample average falls within $\frac{\sigma}{\sqrt{n}}$ of the mean with probability approximately 68%, within $\frac{2\sigma}{\sqrt{n}}$ with probability approximately 95%, and within $\frac{3\sigma}{\sqrt{n}}$ with probability approximately 99%.