Simon Fraser University, Department of Economics
Econ 798, Introduction to Mathematical Economics
Prof. B. Antoine
(based on the notes of Prof. A. Karaivanov)

Part 5: Probability theory and Statistics

1 Probability space and random variables

1.1 Introduction

- **Definition** (Probability space)
  The triple $(\Omega, \mathcal{F}, P)$ where

  - Ω is a set of all possible outcomes ω of some experiment (observation), e.g. tosses of a coin;
  - $\mathcal{F}$ is a σ-algebra, i.e., a collection of subsets of Ω such that:
    (i) $\Omega \in \mathcal{F}$.
    (ii) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ where A^c denotes the complement of A in Ω.
    (iii) if $A_1, A_2, \cdots \in \mathcal{F}$ then $\cup_i A_i \in \mathcal{F}$ for countably many sets A_i.
  - P is a probability measure (see below).

  We call the subsets of Ω **events**. The largest σ-algebra on Ω is the set of all subsets of Ω, the smallest one is $\{\emptyset, \Omega\}$. Think why! Can you define other valid σ-algebras if for example Ω consists of all outcomes from throwing a die?

  We have a space but we need to somehow quantify the "size" of subsets in it, i.e. we need some "measure". This is given by the function P defined below.

- **Definition** (Probability measure)
  A probability measure is a real-valued function P defined over $\mathcal{F}$ and such that:

  (i) for any $A \in \mathcal{F}$, $P(A) \geq 0$.
  (ii) $P(\Omega) = 1$, $P(\emptyset) = 0$.
  (iii) (countable additivity) If $\{A_j\}$ is a countable collection of disjoint sets in $\mathcal{F}$, i.e., $A_{j_1} \cap A_{j_2} = \emptyset$, then $P(\cup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$.

  If A is an event (which entails that some specific ω's are being realized), then $P(A)$ is called the probability of event A.

- **Definition** (Support)
  A support of P is any set $A \in \mathcal{F}$ s.t. $P(A) = 1$.

  Note: suppose you have a die, i.e., $\Omega = \{1, 2, 3, 4, 5, 6\}$ which has been rigged so that it never throws 6. Then $P(\{1, 2, 3, 4, 5\}) = 1$ and $A = \{1, 2, 3, 4, 5\}$ is a support of P.

1.2 Properties of probability measures

1. **Monotonicity:** Let P be a probability measure and $A, B \in \mathcal{F}$.
   If $A \subset B$ then $P(A) \leq P(B)$.

2. **Inclusion/exclusion:** Let $A_1, \cdots, A_k \in \mathcal{F}$.

   $$P(\cup_{k=1}^{n} A_k) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

   This property implies that
   $$P(A^c) = 1 - P(A)$$

3. **Continuity:** If $\{A_n\} \in \mathcal{F}$ and $A \in \mathcal{F}$ and $A_n \uparrow A$ then $P(A_n) \to P(A)$.
   note: this means that $\{A_n\}$ is an increasing sequence of sets A_n ($A_j \subseteq A_{j+1}$ $\forall j$) that converges to A.
   The same property holds for a decreasing sequence of sets converging to A.

4. **Countable sub-additivity:** If $A_1, A_2, \cdots$ and $\cup_{k=1}^{\infty} A_k$ are all elements of $\mathcal{F}$
   then $P(\cup_k A_k) \leq \sum_k P(A_k)$.
   note: here A_k are not necessarily disjoint.

1.3 Random variables and distributions

- **Definition** (Random variable)
  A random variable (RV) on a probability space $(\Omega, \mathcal{F}, P)$ is a real-valued function $X = X(\omega)$, $\omega \in \Omega$ such that for any $x \in \mathbb{R}$, the set $\{\omega \in \Omega \; s.t. \; X(\omega) < x\} \in \mathcal{F}$. When a function satisfies the latter condition we say it is $\mathcal{F}$-measurable.

- Note: the reason we require $X(\omega)$ to be $\mathcal{F}$ -measurable is that we want to be able to assign a probability measure to any set $\{\omega \in \Omega \; s.t. \; X(\omega) < x\}$.

- **Definition**: a random vector is simply a vector of random variables (i.e., vector of functions).

- **Example**: Consider throwing a die. The event space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. A possible σ-algebra is $\mathcal{F} = \{\Omega, \emptyset, \{1, 2, 3, 4\}, \{5, 6\}\}$. A possible probability measure is to assign

$P = 1/6$ to each possible outcome ω_i, $i = 1, \cdots, 6$ and extend it (using the rules above) to all sets in $\mathcal{F}$. Notice that the probability measure needs to be defined on all sets in $\mathcal{F}$! Now, consider the function:

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \in \{1, 2, 3\} \\ 0 & \text{otherwise} \end{cases}$$

a) Is $X(\omega)$ a random variable?
b) What if we had $\omega \in \{1, 2, 3, 4\}$ instead in the definition of $X(.)$ in the first row above?

- **The Borel σ-algebra**
  Let $\Omega = \mathbb{R}$ and assume we are interested in defining probabilities over open intervals in $\mathbb{R}$, including $(-\infty, +\infty)$. It turns out that we can construct a σ-algebra of all open intervals on the real line, which is called the Borel σ-algebra on $\mathbb{R}$ and denoted by $\mathcal{B}^1$. Note that by the definition of σ-algebra, $\mathcal{B}^1$ must contain also all closed, all semi-open intervals, and all singletons in $\mathbb{R}$ (think why!)

  Next, we proceed by characterizing random variables in more detail.

- **Definition** (cumulative distribution function)
  The cumulative distribution function (cdf), $F(x)$ of the random variable $X(\omega)$ is defined as:

$$F(x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$$

  We often write as a shorthand $P(X \leq x)$ instead of $P(\{\omega : X(\omega) \leq x\})$.

- **Properties of the cdf**

  1. $F(-\infty) = 0$ and $F(\infty) = 1$.
  2. F is non-decreasing and continuous from the right.
  3. $F(x^-) \equiv \lim_{t \uparrow x} F(t)$ and $F(x^+) \equiv \lim_{t \downarrow x} F(t)$ exist and are finite.
  4. F is continuous at x iff $F(x^-) = F(x^+) = F(x)$.
  5. The only possible discontinuities in F are jumps up.
  6. The set of discontinuity points of F is countable.

- **Probability and the cdf:**
  We have the following relationships between the probability measure P and the cdf F of a RV X

  1. $P(X > x) = 1 - F(x)$
  2. $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$

3. $P(X < x) = F(x^-)$
4. $P(X = x) = F(x^+) - F(x^-)$ (it is positive only if there is a jump at x and zero otherwise).

1.4 Types of random variables

- **Definition** (discrete random variable) A discrete RV, X is a RV which can take only finitely or countably infinitely many values $x_1, x_2, \cdots$ with probabilities $p_1, p_2, \cdots$ with $p_j > 0$ and $\sum p_j = 1$. Its cdf is called a discrete distribution function and is a step function with jumps at $x_1, x_2, \cdots$. We can also define the probability function for X as $f(x) = P(X = x)$. Clearly, $f(x_i) \geq 0$ and $\sum f(x_i) = 1$.

- **Definition** (continuous random variable)
  A continuous RV is one that has a continuous cdf. Then there exists a function f s.t. $\forall x, y$ with $x < y$, $F(y) - F(x) = \int_x^y f(t)dt$ and F has a derivative equal to f almost everywhere (except on a set of measure 0). The function f is called the probability density function (pdf) of X and, as we see, it is only defined up to a set of measure zero, i.e., it is not unique. Unlike the discrete RVs there is no direct connection between $f(x)$ and $P(X = x)$ as the latter is always zero for continuous RVs.

  Finally, a mixed random variable is for example one that has a continuous part and mass at some point.

- **Functions of random variables** Let X is a RV and let us look at the random variable $Y = g(X)$. Suppose we are interested in the distribution of Y; i.e., we want to find $f(y)$:

  - Case 1: X is discrete with pf $f_x(x)$. We have:
  $$f_y(y) = P(g(x) = y) = \sum_{x:g(x)=y} f_x(x)$$
  - Case 2: X is continuous RV with pdf $f_x(x)$ and $P(a < x < b) = 1$. Let g be a continuous and strictly monotonic function and let $a < x < b$ iff $\alpha < y < \beta$. Then the pdf of $Y = g(X)$, $f_y(y)$ is:
  $$f_y(y) = \begin{cases} f_x(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & \text{for } \alpha < y < \beta \\ 0 & \text{otherwise} \end{cases}$$

1.5 Expectation

- **Definition: Expectation of a continuous random variable**
  The expectation of a continuous RV with pdf $f(x)$ is defined as:
  $$E(X) = \int x f(x) dx$$

if the integral exists. The integration is performed on the support of X.

- **Definition: Expectation of a discrete random variable**
  The expectation of a discrete RV with pf $f(x)$ is defined as:

$$E(X) = \sum_i x_i f(x_i)$$

  if the sum exists.

  Notice that the above implies that in some cases the expectation of a RV does not exist. For example take a Cauchy distributed random variable, pdf $f(x) = 1/[\pi(1+x^2)]$. We have,

$$\int_{-\infty}^{\infty} x f(x) dx = \frac{1}{\pi} \ln(1+x^2)\Big|_0^{\infty} = \infty$$

- Example: compute the expectation of the uniformly distributed random variable on $[a, b]$ with $F(x) = (x-a)/(b-a)$ for any $x \in [a, b]$.

- **Properties of expectations**

  1. integrability: $\int x f(x) dx < \infty$ iff $\int |x| f(x) dx < \infty$.
  2. linearity: $E(aX + bY) = aE(X) + bE(Y)$.
  3. If there exists a constant a s.t. $P(X \geq a) = 1$ then $E(X) \geq a$.
  4. If $P(X \leq Y) = 1$ then $E(X) \leq E(Y)$.
  5. If $P(X = Y) = 1$ then $E(X) = E(Y)$.
  6. If $P(X \geq 0) = 1$ and $E(X) = 0$ then $P(X = 0) = 1$.
  7. If $P(X = 0) = 1$ then $E(X) = 0$.
  8. If c is a constant, $E(c) = c$.

- **Expectation of a function of RV** Let X be a RV with pdf $f(x)$. Suppose we want to find $E(Y)$, where $Y = g(X)$ is a random variable which is a function of X. Then the following is true:

$$E(Y) = E(g(X)) = \int g(x) f(x) dx$$

1.6 Moments of random variables

The expectation of a RV is just one of the many possible characteristics of its distribution. In general, we can define the so-called moments of the distribution of a given RV as the expectations of powers of X or $X - E(X)$. In particular, for $r \in \mathbb{N}$, we call $E(X^r)$ the r-th raw moment of

X and $E((X-E(X))^r)$ the r-th central moment. The r-th raw moment exists if $E(|X|^r) < \infty$. Also if $E(|X|^k) < \infty$ for some k, then $E(|X|^j) < \infty$ for any $j < k$.
Let us look at some frequently used moments.

1. Variance (2-nd central moment):
   The variance of a RV is given by $V(X) = E[(X-E(X))^2] = E(X^2) - (E(X))^2$. We also call $V(X)$ the standard deviation of X.
   The following are some useful properties of the variance:

   (i) For any a constant, $V(a) = 0$.

   (ii) For any a, b constants, $V(aX+b) = a^2V(X)$.

2. Skewness (3-rd central moment), $E((X-E(X))^3)$.

3. Kurtosis (4-th central moment), $E((X-E(X))^4)$.

A related concept is the mean square error (MSE) of a RV defined as:

$$MSE(X) = E(X-c)^2 = Var(X) + (c-E(X))^2$$

for some constant c. The interpretation is that it 'measures' the average deviation from c. Clearly the MSE is minimized at $c = E(X)$.

1.7 Some useful inequalities (for reference)

- **Theorem** (Jensen's Inequality)
  Let X be a RV and h is a concave function. Then:

$$E(h(X)) \leq h(E(X))$$

- **Theorem** (Markov's Inequality)
  Let Y be a non-negative RV, i.e. $P(Y < 0) = 0$ and let k be a positive constant. Then:

$$P(Y \leq k) \leq \frac{E(Y)}{k}$$

- **Theorem** (Chebyshev's Inequality #1)
  Let X be a RV, c is a constant and d is a positive constant. Then:

$$P(|X-c| \geq d) \leq \frac{E(X-c)^2}{d^2}$$

- **Theorem** (Chebyshev's Inequality #2)
  Let X be a RV with expectation $E(X) = \mu$ and variance $V(X) = \sigma^2$ and d is a positive constant. Then:

$$P(|X-\mu| \geq d) \leq \frac{\sigma^2}{d^2}$$

1.8 Random vectors

In this section we study vectors of random variables and their distribution functions. Some new concepts need to be defined in this context.

- **Definition: joint distribution function**
  Let (X, Y) be a random vector. The joint distribution function $F(x, y)$ of (X, Y) is:

  $$F(x, y) \equiv P(\{\omega \in \Omega \ s.t. \ X(\omega) \leq x \wedge Y(\omega) \leq y\}) \quad \text{for any } x, y \in \mathbb{R}$$

  If X and Y are both discrete RVs then they also have a discrete joint prob. function and we can define it as

  $$f(x, y) = P(X = x \wedge Y = y)$$

  with $f(x_i, y_j) \geq 0$ and $\sum_{i,j} f(x_i, y_j) = 1$. The jdf is just the "overall" distribution function of the vector (X, Y) and is thus analogous to the cdf in the univariate case. If instead we want to look at the components of the random vector one at a time we need to define the following marginal distribution.

- **Definition: marginal distribution function**
  Let (X, Y) be a random vector. The marginal distribution function of X is:

  $$F(x) = \lim_{y \to \infty} F(x, y) = \lim_{y \to \infty} P(X \leq x \wedge Y \leq y)$$

  The interpretation is that the marginal of X is its distribution function (summed) over all possible values for Y .

  For discrete RVs X and Y taking values $\{x_i\}$ and $\{y_j\}$ respectively, the above limit can be computed as simply adding up the joint probabilities $f(x_i, y_j)$ over all possible y_j for any given x_i, i.e.,

  $$f(x_i) = \sum_j f(x_i, y_j) \quad \text{for any } i$$

- **Definition: joint and marginal probability density functions** (X, Y) have a continuous joint distrib. function if there exists a non-negative function $f(x, y)$ called the joint probability density function (joint pdf) of (X, Y) such that for any set $A \subset \mathcal{B}^2$:

  $$P((X, Y) \in A) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(s, t) ds dt$$

  The joint pdf satisfies the following:

  $$f(x, y) \geq 0, \quad \int_{\mathbb{R}} \int_{\mathbb{R}} f(s, t) ds dt = 1, \quad f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

  whenever the derivative exists.

  The marginal probability density function of X is $f(x) = \int_{\mathbb{R}} f(x, t) dt$. In the discrete case it is called a marginal probability function and is given by $f(x) = P(X = x) = \sum_y f(x, y)$.

- **Exercise:** Suppose that a point (X, Y) is chosen at random from the rectangle $S = \{(x, y) : 0 \leq x \leq 2; 1 \leq y \leq 4\}$. Determine the joint cdf and pdf of X and Y, the marginal cdf and pdf of X and the marginal cdf and pdf of Y.
  Answers: $F(x, y) = (x/2)(y - 1)/3$ on $[0, 2]x[1, 4]$ (think how you'd define it outside that area). The joint pdf is then $\partial^2 F(x, y)/\partial x \partial y = 1/6$. We also have $F(x) = x/2$ and $F(y) = (y - 1)/3$.

1.9 Conditional distribution

The marginal distributions defined above dealt with the distribution of one of the variables in a random vector for all possible values of the other. Next we study the distribution of one of the variables given some fixed value for the other.

- **Definition: conditional probability function**
  Let X, Y be discrete RVs. The conditional probability function of X given $Y = y$ is defined as:
  $$f(x|y) = \begin{cases} \frac{f(x,y)}{f(y)} & \text{for } f(y) > 0 \\ 0 & \text{otherwise} \end{cases}$$
  where $f(x, y)$ is the joint pf of (X, Y) and $f(y)$ is the marginal pf of Y.

- Definition: **conditional probability density function**
  Let X, Y be continuous RVs. The conditional probability density function of X given $Y = y$ is defined as:
  $$f(x|y) = \begin{cases} \frac{f(x,y)}{f(y)} & \text{for } f(y) > 0 \\ 0 & \text{otherwise} \end{cases}$$
  where $f(x, y)$ is the joint pdf of (X, Y) and $f(y)$ is the marginal pdf of Y.

- **Example:** suppose we have a fair coin, which we toss twice. Assign 0 if tails occur and 1 if heads. Define the RV X as the sum of the outcomes of the two throws and Y as the difference between the first and second throw outcome.

  Q1 what are the possible values for X, for Y?

  Q2 what is the joint probability distribution of X and Y.

  Q3 what is the marginal probability function of X? of Y?

  Q4 what is the conditional probability function of X given $Y = 0$? What is the conditional pf of Y given $X = 0$?

1.10 Covariance

So far we only studied the characteristics of the distribution of a single random variable. Often, however it is interesting to know how two RVs are related. We can define the covariance of two RVs X and Y as:

$$cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

Notice the closeness with the concept of variance (just put $Y = X$ above). The covariance measures how 2 RVs are **linearly** related to each other. Notice that if X and Y are such that if X goes up then Y goes up too, they will have a positive covariance, whereas if they are negatively related they will have a negative covariance.

- **Properties:**

  1. $cov(X,Y) = E(XY) - E(X)E(Y)$.
  2. $cov(X+Y,Z) = cov(X,Z) + cov(Y,Z)$
  3. $cov(aX,Y) = a(cov(X,Y))$ for a constant.
  4. $cov(X,a) = 0$ for a constant.
  5. If $X_1, X_2, \cdots, X_n$ are RVs with $Var(X_i) < \infty$ fo all i, then:

$$Var(\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Var(X_i) + \sum_{i,j=1,\cdots,n,i\neq j} cov(X_i, X_j)$$

1.11 Correlation

A related concept to that of covariance is the **correlation** of two RVs defined as:

$$\rho(X,Y) \equiv \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \quad \text{for } Var(X), Var(Y) < \infty$$

Notice that the correlation is simply the covariance normalized by the product of standard deviations of the two variables. Because of this normalization we can prove the following.

- **Theorem**
  The correlation between two RVs is always less or equal to 1 in absolute value, i.e. $\rho(X,Y) \in [-1,1]$.

- If $\rho(X,Y) = 1$ we say that the RVs are perfectly positively correlated.

- If $\rho(X,Y) = -1$ we call them perfectly negatively correlated.

- If $\rho(X,Y) = 0$ we say that the RVs are uncorrelated.

- **Exercise**: Suppose that X is uniformly distributed on the interval $[-1, 1]$ and $Y = X^2$. Show that X and Y are uncorrelated.
  Solution: It will be enough to show that $cov(X, Y) = 0$ (show that the variances are finite first). We have $E(X) = 0$, $F(x) = (x + 1)/2$ and $f(x) = 1/2$. Now, we can compute the expectation of Y:
  $$E(Y) = \int_{-1}^{1} \left(x^2 \times \frac{1}{2} \right) dx = \left(\frac{1}{6} + \frac{1}{6} \right) = \frac{1}{3}$$
  Thus, we have:
  $$cov(X, Y) = E[(X - 0)(X^2 - 1/3)] = E(X^3 - \frac{X}{3})) = E(X^3) - 0$$
  Finally, we have
  $$E(X^3) = \int_{-1}^{1} x^3 \frac{1}{2} dx = (1/8) - (1/8) = 0$$
  which means that $cov(X, Y) = 0$. Thus, the two RVs are indeed uncorrelated even though one is a function of the other!

  ** Recall that the correlation/cov only captures **linear dependence**.

1.12 Conditional expectation

Often we may be interested in the expectation of some RV X given that some other variable Y takes some fixed value.

- **Definition: conditional expectation**

  (a) The conditional expectation of a discrete RV X given a discrete RV Y, is the **random variable** denoted $E(X|Y)$ taking values $E(X|Y = y_j)$ with respective probabilities $f_y(y_j)$ - the marginal pf of Y - where:
  $$E(X|Y = y_j) = \sum_x x \times f(x|Y = y_j)$$
  and where $f(x|Y = y_j)$ is the conditional pf of X given $Y = y_j$.

  (b) The conditional expectation of a continuous RV X given a continuous RV Y is the RV $E(X|Y)$ defined analogously as in part (a) but with
  $$E(X|Y = y_j) = \int x \times f(x|Y = y_j) dx$$

  Related to the concept of conditional expectation we have the following important result.

- **Exercise:** compute $E(X)$, $E(Y)$ and the conditional expectations for some fixed X or Y in the coin toss example above.

- **Theorem** (the Law of iterated expectations or LIE)
  Let X, Y be two random variables with finite expectations. Then:

$$E(X) = E(E(X|Y))$$

  where the outer E (sum or integral) is taken over the support of Y (i.e., using f_y) while the inside E is as in the previous definition.

  Intuition: $E(Y)$ is the expectation about Y without any information about X, whereas $E(Y|X)$ is the expectation about Y knowing X. Thus, in general, $E(Y|X) \neq E(Y)$ but on average - i.e. $E(E(Y|X))$ - it must be equal to $E(Y)$ otherwise the latter would be wrong.

- There is also a similar result for the variance (known as the variance decomposition):

$$Var(X) = Var(E(X|Y)) + E(Var(X|Y))$$

1.13 Independence

- **Definition: independence**
  The random variables $\{X_j\}_{j=1}^n$ are independent if and only if

$$P(\cap_{j=1}^n (X_j \in A_j)) = \Pi_{j=1}^n P(X_j \in A_j) \quad \text{for any } A_j \in \mathcal{B}^1 \text{ with } j = 1, \cdots, n$$

  The following results are often used to verify the independence of RVs.

- **Theorem**
  The random variables $X_1, X_2, \cdots, X_n$ are independent if and only if for all $x_1, \cdots, x_n$

  (i) $F(x_1, x_2, \cdots, x_n) = \Pi_{j=1}^n F(x_j)$ where $F(X)$ is the cdf of X.

  (ii) $f(x_1, \cdots, x_n) = \Pi_{j=1}^n f(x_j)$ if the marginal pdfs exist.

1.14 Independence vs. uncorrelatedness

- **Theorem**

  (i) If $X_1, \cdots, X_n$ be n independent RVs defined on $(\Omega, \mathcal{F}, P)$ whose respective raw moment of order $r_1, \cdots, r_n$ exists. Then:

$$E(X_1^{r_1} X_2^{r_2} \cdots X_n^{r_n}) = E(X_1^{r_1}) E(X_2^{r_2}) \cdots E(X_n^{r_n})$$

  (ii) If X and Y are 2 independent RVs, then for any measurable functions f and g, we have:

$$E(f(X)g(Y)) = E(f(X))E(g(Y))$$

  The above result can be extended to n RVs.

- **Corollary**
  If the RVs X and Y are independent then they are also uncorrelated.

  Proof: The above follows directly from the previous theorem, since $E(XY) = E(X)E(Y)$ - with f and g taken as identity functions - which implies that $cov(X, Y) = 0$. However, it is not true that if two variables are uncorrelated, they are also independent. Take a look at the following counter-example.

- **Example 1:** (uncorrelatedness does not imply independence)
  Remember our fair coin, which we toss twice. Assign 0 if "tails" occur and 1 if "heads". Define the RV X as the sum of the outcomes of the two throws and Y as the difference between the first and second outcome. The joint probability distribution of X and Y is provided in the table below:

		X		
		0	**1**	**2**
Y	**-1**	0	1/4	0
	0	1/4	0	1/4
	1	0	1/4	0

  We have $E(X) = 1$, $E(Y) = 0$, $E(XY) = 0$ i.e., $cov(X, Y) = 0$, thus X and Y are uncorrelated (verify those claims). However they are not independent since $f(X = 1, Y = 0) = 0$ but $f(X = 1) = 1/2$ and $f(Y = 0) = 1/2$ i.e., $f(1, 0) \neq f_x(1)f_y(0)$.

  NOTE: another counter-example is one from above with $Y = X^2$. Show it.

- **Example 2:** Suppose that (X, Y) are uniformly distributed on the square $S = \{f(x, y) : -6 \leq x \leq 6,\ -6 \leq y \leq 6\}$. Show that X and Y are independent and hence uncorrelated.
  Solution: we have $F(x; y) = (x + 6)/12 \times (y + 6)/12$ on $[-6; 6] \times [-6; 6]$ (draw a picture, this is about the area of a rectangle...). But it is also true that the marginal distributions $F(x) = lim_{y\to\infty}F(X, Y) = (x + 6)/12$ and similarly for Y.
  However, under some special assumptions about X and Y independence and uncorrelatedness may be equivalent.

- **Theorem**
  If X and Y are **jointly normally distributed** $BN(\mu, \Sigma)$ with $\mu = [\mu_1\ \mu_2]'$ and $\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_y^2 \end{pmatrix}$ that is their joint pdf is

$$f(x, y) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{1}{2}(z - \mu)'\Sigma^{-1}(z - \mu)\right) \quad \text{where} \quad z = \begin{pmatrix} x \\ y \end{pmatrix}$$

  then X and Y being independent is equivalent to X and Y being uncorrelated.

2 Convergence of random variables

We consider sequences of random variables $\{X_n\}_{n=1}^{\infty}$ and we are interested in the convergence of such sequences. Remember that random variables are functions so this is related to the general theory on convergence of functions. Serving different purposes, the following three possible definitions of convergence of random variables exist. They are not equivalent but some of them are implied by the others as we will see later on.

- **Definition** (almost sure convergence)
  We say that $\{X_n\}$ converges almost surely to X (written as $X_n \xrightarrow{a.s.} X$) if and only if:

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

  The above definition requires that $\{X_n\}$ converges to X for all $\omega \in \Omega$ except possibly on a set of measure 0.

  usually, the ω's disappear from the definition, and we simply write:

  $$P(\lim_{n\to\infty} X_n = X) = 1$$

- **Definition** (convergence in probability)
  We say that $\{X_n\}$ converges in probability (written as $X_n \xrightarrow{P} X$) if and only if:

  $$\forall \epsilon > 0 \ \lim_{n\to\infty} P(\{\omega \in \Omega \ |X_n(\omega) - X(\omega)| > \epsilon\}) = 0$$

  where $|a|$ denotes the absolute value (in general, for random vectors the Euclidean norm of such vector).

  once again, the ω's often disappear from the definition, and we simply write:

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

- **Definition** (convergence in distribution)
  We say that $\{X_n\}$ converges in distribution to X (written as $X_n \xrightarrow{d} X$) if and only if:

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

  where F_n is the cdf of X_n and F is the cdf of X.

  The following theorem gives the relationships between the above convergence concepts.

- **Theorem** (Relationship between convergence concepts)
  Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables such that the above convergence concepts are well defined. Then:

(i) the following relationships hold: $X_n \xrightarrow{a.s.} X \;\Rightarrow\; X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{d} X$

(ii) If $X_n \xrightarrow{d} c$ (where c is a constant), then $X_n \xrightarrow{P} c$.

We see that $X_n \xrightarrow{d} X$ is the weakest convergence concept.

We now recall three well-known convergence theorems that are very often used in econometrics with various applications.

- **Theorem** (Continuous mapping theorem or Mann-Wald)

  (i) Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function.

  $$X_n \xrightarrow{d} X \;\Rightarrow\; g(X_n) \xrightarrow{d} g(X)$$

  (ii) Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function.

  $$X_n \xrightarrow{P} X \;\Rightarrow\; g(X_n) \xrightarrow{P} g(X)$$

  (iii) Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function.

  $$X_n \xrightarrow{a.s.} X \;\Rightarrow\; g(X_n) \xrightarrow{a.s.} g(X)$$

- **Theorem** (Slutsky)
  Let X_n and Y_n be 2 scalar RVs s.t. $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ (with c a constant), then:

  (i) $(X_n, Y_n) \xrightarrow{d} (X, c)$.

  (ii) $X_n + Y_n \xrightarrow{d} X + c$.

  (iii) $X_n \times Y_n \xrightarrow{d} X \times c$.

  (iv) $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$ if $c \neq 0$.

  The following result is very useful for finding limiting distributions of functions of random variables.

- **Theorem** (Delta method)
  Let $\alpha_n(X_n - b) \xrightarrow{d} X$ where α_n is a sequence of real numbers s.t. $\alpha_n \to \infty$ and b is a constant that does not depend on n. Let g be a function (possibly of other variables too) differentiable at b. Then:

  $$\alpha_n(g(X_n) - g(b)) \xrightarrow{d} \frac{\partial g(b)}{\partial b} X$$

3 Laws of large numbers

Let X_i, $i = 1, 2, \cdots$ be a collection of random variables with finite expectations given by $\mu_i = E(X_i)$. Let $\overline{X}_n$ be the average of the first n X_i in this collection, $\overline{X}_n = \sum_{i=1}^{n} X_i/n$.
A Law of large numbers (LLN) provides conditions on $\{X_i\}_{i=1}^{\infty}$ such that the derived sequence of partial averages $\{\overline{X}_n\}_{n=1}^{\infty}$ converges in probability (for a Weak LLN) or almost surely (for a Strong LLN) to some limit.
** notice that $\{\overline{X}_n\}_{n=1}^{\infty}$ is a sequence of RVs itself.

- **Theorem** (Khintchin's weak LLN)
  Let X_i, $i = 1, 2, \cdots$ be independent and identically distributed (i.i.d) RVs and assume that $E(X_i) = \mu < \infty$. Then:
  $$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$$

- **Theorem** (Chebyshev's weak LLN)
  Let X_i, $i = 1, 2, \cdots$ satisfy $E(X_i) = \mu_i < \infty$, $Var(X_i) = \sigma_i^2 < \infty$ and $cov(X_i, X_j) = 0$ for all $i \neq j$.

  $$\text{If} \quad \lim_{n\to\infty} Var\left(\overline{X}_n\right) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 < \infty$$
  $$\text{Then} \quad \overline{X}_n - E(\overline{X}_n) \xrightarrow{P} 0$$

- **Theorem** (Kolmogorov's strong LLN)
  Suppose X_i, $i = 1, 2, \cdots$ are independent and with finite variances, given by σ_i^2.

  $$\text{If} \quad \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 < \infty$$
  $$\text{Then} \quad \overline{X}_n - E(\overline{X}_n) \xrightarrow{a.s} 0$$

  Note also that since almost sure convergence implies convergence in probability the strong LLN is also weak.

4 Central limit theorems

A Central Limit Theorem (CLT) gives conditions on a sequence of random variables $\{X_i\}_{i=1}^{\infty}$ such that the derived sequence of partial averages $\{\overline{X}_n\}$, when suitably normalized, converges in distribution to a standard normal distribution. As with the LLNs there are many possible CLTs of which the following are the most frequently used.

- **Theorem** (Lindberg-Levy CLT)
  Let X_i, $i = 1, 2, \cdots$ are i.i.d. random variables with $E(X_i) = \mu < \infty$ and $Var(X_i) = \sigma^2 < \infty$.

$$\frac{1}{\sqrt{\sigma^2/n}} \left[\overline{X}_n - \mu \right] \xrightarrow{d} \mathcal{N}(0,1)$$

  This theorem makes strong assumptions about homogeneity and independence of the RVs. Theorems with weaker homogeneity assumptions but stronger moments assumptions are given as follows.

- **Theorem** (Lyapunov CLT)
  Suppose X_i, $i = 1, 2, \cdots$ are independent random variables with $E(X_i) = \mu_i < \infty$ and $Var(X_i) = \sigma_i^2 < \infty$ and $\mu_i^3 = E(|X_i - \mu_i|^3) < \infty$.
  Let $B_n = (\sum_{i=1}^n \mu_i^3)^{1/3}$ and $C_n = (\sum_{i=1}^n \sigma_i^2)^{1/2}$.

$$\lim_{n\to\infty} \frac{B_n}{C_n} = 0 \quad \Rightarrow \quad \frac{(\overline{X}_n - E(\overline{X}_n))}{\sqrt{Var(\overline{X}_n)}} \xrightarrow{d} \mathcal{N}(0,1)$$

- **Theorem** (Lindberg-Feller CLT)
  Suppose X_i, $i = 1, 2, \cdots$ are independent random variables with distribution functions F_i, $E(X_i) = \mu_i < \infty$ and $Var(X_i) = \sigma_i^2 < \infty$. Let $C_n = |\sum_{i=1}^n \sigma_i^2|^{1/2}$.

$$\text{If} \quad \lim_{n\to\infty} \frac{1}{C_n^2} \sum_{i=1}^n \int_{|x-\mu_i|>\epsilon C_n} (x-\mu_i)^2 dF_i(x) = 0 \quad \forall \epsilon > 0$$

$$\text{Then} \quad \frac{(\overline{X}_n - E(\overline{X}_n))}{\sqrt{Var(\overline{X}_n)}} \xrightarrow{d} \mathcal{N}(0,1)$$