Econ 140 - Spring 2016
Section 1

GSI: Fenella Carpena

January 21, 2016

In this section note, we will review basic probability and statistics concepts. Please be advised that this note is not intended to be a comprehensive review of the textbook/lecture, since the textbook contains a lot more material than we have time to cover. However, I have tried to focus on the concepts which I believe are necessary to be successful in our class.

1 Random variables and their probability distributions

- A **random variable** represents the outcome of a random process. They can be either **discrete** (i.e., can take on only a discrete set of values) or **continuous** (i.e., can take on a continuous set of values).[1]
  - *Notation: Capital letters X, Y, Z etc. are used to denote random variables. We will also use "r.v." as an abbreviation for "random variable."*
  - *Notation: Lower case letters indicate a generic possible value of the random variable.*
- Every random variable is associated with two functions: a **probability distribution function (pdf)** and a **cumulative distribution function (cdf)**. The pdf is concerned with "point probabilities" of random variables (i.e., $f(x) = P(X = x)$), while the cdf is concerned with the probability that the random variable takes on a value less than or equal to a particular number (i.e., $F(x) = P(X \leq x)$).
  - *Notation: The pdf is denoted with lower case f. $P(X = x)$ is read as the "probability that X equals x," where x is a generic value.*
  - *Notation: The cdf is denoted with upper case F. $P(X \leq x)$ is read as the "probability that X is less than or equal to x," where x is a generic value.*
  - For a discrete random variable, the pdf can be presented in a tabular format, listing all the possible outcomes of the random variable and the probability at which each outcome will occur.
  - For a continuous random variable, the pdf also summarizes information about the possible outcomes the variable can take. But since a continuous random variable can take on a infinite number of values, the probability that it is equal to a particular value is infinitesimally zero. Hence, we use the pdf of a continuous random variable only to compute the probability that X lies within a range of values, e.g. to compute $P(a \leq X \leq b)$. Note that $P(a \leq X \leq b)$ is the area under the pdf between points a and b.
- **Important properties of pdf**:
  1. The probabilities $P(X = x)$ are all non-negative, since probabilities can never be less than 0.
  2. The probabilities sum up to 1 (for discrete) or integrate to 1 (for continuous), which happens because the pdf must cover all possible states of the world.

[1] Mixed random variables exist as well but we are not going to cover them in this course.

- **Important properties of cdf**:
  1. The cdf goes to 0 at negative infinity, and goes to 1 at positive infinity.
  2. The cdf is a non-decreasing function.
- **Important properties when working with probability:**
  1. For any constant c, $P(X > c) = 1 - P(X \leq c)$.
  2. For any constants $a < b$, $P(a < X \leq b) = P(X \leq b) - P(X \leq a)$.
  3. In the case where X is a continuous r.v., it does not matter whether the inequality in the previous two points are strict or not, i.e., $P(X \geq c) = P(X > c)$ and $P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b)$. This is because as previously mentioned, for a continuous r.v., the probability that it is equal to a particular value is infinitesimally zero.

2 Joint Distributions and Independence

- The **joint probability distribution** of two discrete random variables X and Y gives the probability for simultaneous outcomes, $P(X = x, Y = y)$. If X and Y are both continuous, we can also define their joint distribution, however this case is not covered in the textbook so we will not discuss it here.
- Two discrete random variables X and Y are said to be **independent** if and only if $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ for all possible x and y. Intuitively, independence means that knowing the value of X gives us no information about the value of Y, and vice versa. If the random variables are not independent, they are said to be **dependent**.

3 Features of Probability Distributions and Joint Distributions

- The **expected value** of a random variable X, denoted $E(X)$, is a weighted average of all possible values of X. Its precise definition depends on whether the variable X is discrete or continuous. We will focus on the discrete case.
  - For a discrete r.v. X, $E(X) \equiv \sum_{i=1}^{N} x_i \cdot P(X = x_i)$, where N is the # of possible values of X.
  - *Notation: $E(X)$ may also be denoted as μ. In some cases, we may also write μ with its r.v. as a subscript, e.g. μ_X emphasizes that μ refers the mean of X and not any other r.v.*
- The **variance** of a random variable X, denoted $var(X)$, is the expected value of the squared deviations from the mean. It is a measure of how much the distribution of X is tightly centered around is mean. Note that variance is always non-negative.
  - $var(X) \equiv E[(X - \mu)^2] = E(X^2) - \mu^2$
  - *Notation: We also denote $var(X)$ as σ^2 or σ_X^2.*
- The **standard deviation** of a random variable, denoted $sd(X)$, is the positive square root of the variance.
  - $sd(X) \equiv \sqrt{var(X)}$.
  - *Notation: We also denote $sd(X)$ as σ or σ_X*
- The **covariance** between two random variables X and Y, denoted $cov(X, Y)$, is defined as the expected value of the product of $(X - \mu_X)(Y - \mu_Y)$. It measures the amount of *linear* dependence between the two random variables X and Y.
  - $cov(X, Y) \equiv E(X - \mu_X)(Y - \mu_Y) = E(XY) - \mu_X\mu_Y$.

- *Notation: cov(X,Y) may also be denoted as* σ_{XY}

- The **correlation** between two random variables X and Y, denoted $corr(X, Y)$, is a scale-free measure of linear dependence between these variables.
  - $corr(X,Y) \equiv \frac{cov(X,Y}{sd(X)sd(Y)}$
  - *Notation: corr(X,Y) may also be denoted as* ρ *or* ρ_{XY}

- **Important properties of expectation, variance, sd, covariance, and correlation:** In what follows, let X, Y and Z be random variables, and let a, b, c, d be constants.
  1. $E(a) = a$
  2. $E(aX + b) = aE(X)$
  3. $E(X + Y) = E(X) + E(Y)$
  4. $var(a) = 0$, i.e. the variance of any constant is zero. Conversely, if a random variable has zero variance, then it is essentially a constant.
  5. $var(aX + b) = a^2Var(X)$
  6. $var(aX + bY) = a^2var(X) + b^2var(Y) + 2abcov(X,Y)$
  7. $sd(c) = 0$
  8. $sd(aX + b) = |a|sd(X)$
  9. $cov(a,X) = 0$
  10. $cov(aX + c, bY + d) = a \cdot b \cdot cov(X,Y)$
  11. $cov(X + Y, Z) = cov(X,Z) + cov(Y,Z)$
  12. $cov(X,X) = var(X)$
  13. $corr(X,Y)$ must be between -1 and 1, i.e. $-1 \leq corr(X,Y) \leq 1$. Values of $corr(X,Y)$ closer to 1 or -1 indicate a stronger linear relationship between X and Y.
  14. If X and Y are independent, then:
      - $cov(X,Y) = 0$. However, the converse is not true. That is, if $cov(X,Y) = 0$, this does *not* imply X and Y are independent.
      - $E(XY) = E(X)E(Y)$
      - $var(X + Y) = var(X) + var(Y)$
  15. If $cov(X,Y) \neq 0$, then X and Y are not independent.