A review of basic probability theory

Math 485

August 27, 2015

1 Probability and Events

1.1 Events and their properties

Consider an experiment where we toss a coin twice. All the possible outcomes are

$$\{TT\}, \{HH\}, \{TH\}, \{HT\}.$$

We call these (elementary) events. The events have the following properties:

a. The union of two events is an event:

$$\{TT\} \cup \{TH\} = \{\text{First toss is } T\}.$$

b. The intersection of two events is an event:

$$\{\text{First toss is } T\} \cap \{\text{Second toss is } T\} = \{TT\}.$$

c. The complement of an event is an event:

$$\{TT\}^c = \{ \text{ At least one of the toss is } H\}.$$

Note: In everyday language, union corresponds to OR, intersection corresponds to AND, complement corresponds to NOT.

Suppose we toss a coin n times. It is not difficult to see that that more generally we have the followings:

a'. The union of finitely many events is an event: The event $\{\text{First toss is } T\}$ is the union of finitely many events where each of them has the form $\{T \cdots\}$.

b'. The intersection of finitely many events is an event: The event $\{\text{All tosses are } T\}$ is the intersection of n events where each of them has the form $\{ \text{ The nth toss is } T\}$.

Suppose we toss a coin indefinitely. Then we have the followings:

a". The union of (countably) infinitely many events is an event: The event $\{$We eventually see a $T\}$ is the (countable) union of events of the form $\{$ The nth toss is $T, n = 1, 2\cdots\}$.

b". The intersection of (countably) infinitely many events is an event: The event $\{$All the even toss is $T\}$ is the (countable) intersection of events of the form $\{$ The nth toss is $T, n = 2, 4, 6\cdots\}$.

Terminology: When two events have nothing in common (their intersection is $\emptyset$, the empty set) we say they are *mutually exclusive*. For example, the two events $\{$First toss is $H\}$ and $\{$First toss is $T\}$ are mutally exclusive.

Abstractly, we use capital letters at the beginning of the alphabet: A, B or $E_1, E_2\cdots$ to denote an event. We also see that in the examples above, an outcome (or an elementary event) is an event that has no sub-event contained in it (in other words, a smallest possible event).

1.2 Probability

The union of all possible outcomes is an event, (the *universal* event, also called the *sample space*), which we denote by Ω. Then all events are subsets of Ω. We assign a probability, which is a number between 0 and 1, on each event. The probability then is nothing but a mapping from the set of events to the interval $[0, 1]$. Intuitively, this mapping should satisfy the following property:

a. The probability of the union of all outcomes is 1: $P(\Omega) = 1$.

b. The probability of the empty set is 0: $P(\emptyset) = 0$.

c. The probability of the union of two mutually exclusive events is the sum of the individual probability of each event: If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.

From c, we have the following inclusion - exclusion principle: For any events A, B (not necessarily mutually exclusive)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Exercise: Prove the inclusion - exclusion principle.

Using a,b,c one can come up with more probability identity, for example $P(A^c) = 1 - P(A)$ etc.

When assigning probability, besides a,b,c, we also use the following "common-sense" principle: outcomes that are equally likely have the same probability. For example, if a coin is *fair*, then all outcomes $\{TT\}, \{HH\}, \{TH\}, \{HT\}$ are equally likely. Now applying a and c, we see easily that each of them should have probability equals 1/4.

1.3 Examples

Example 1.1. *We toss a coin twice. The probability that we get at least 1 tail is*

$$P(\{TT\} \cup \{TH\} \cup \{HT\}) = \frac{3}{4}.$$

The probability that we get no tail is

$$P(\{HH\}) = \frac{1}{4}.$$

Example 1.2. *Combinatorics Suppose an urn has 2 white balls and 3 red balls. We pick out (without replacement) 2 balls. What is the probability that the 2 balls are red?*

Ans: Here we need to see what the sample space is. It is all possible ways we can pick out 2 balls from the urn. What is the event of interest? It is all possible ways we can pick 2 red balls form the urn. Since each outcome from our pick is equally likely (by equally likely outcome here we mean suppose we number all the balls from 1 to 5, then the possibility we pick out balls 1,2 is the same as the possibiltiy we pick out balls 4,5), the probability of interest is just the ratio of the size of the event with the size of the sample space.

Concretely, the number of ways we can pick 2 balls out of 5 balls is $\binom{5}{2} = 10$*. The number of ways we can pick 2 red balls is* $\binom{3}{2} = 3$*. So the probability is* $\frac{3}{10}$*.*

2 Conditional probability and independent events

2.1 Conditional probability

2.1.1 Motivating example

Suppose we toss a coin twice. What is the probability that we get 2 tails? From the above, it's $\frac{1}{4}$. Suppose, however, that you know the additional information that the first toss is a tail. We ask the same question: what is the probability that we get 2 tails? Clearly it's no longer $\frac{1}{4}$, because for you, the set of *all possible events* have changed. Namely, the outcomes $\{HH\}, \{HT\}$ are no longer possible.

Concretely, the set of all possible outcomes now are:

$$\{TT\}, \{TH\}.$$

Thus the probability that you get 2 tails is $\frac{1}{2}$. We say: the probability that we get 2 tails, *conditioned on* the first toss being a tail, is $\frac{1}{2}$.

2.1.2 Conditional probability

Definition 2.1. *Let A, B be events. If $P(A) > 0$, the probability of B conditioned on A, or B given A, denoted $P(B|A)$, is defined as:*

$$P(B|A) = \frac{P(B \cap A)}{P(A)}.$$

The interpretation is that we have already had the knowledge that A happened. So the probability of the event B happening, given that A has happened, should be calculated as given in the definition.

Remark 2.2. *If $P(A) = 0$ then we cannot use the above formula to define $P(B|A)$. There is a way around it, using the measure theoretic definition of conditional expectation, and the notion of regular conditional probability. We'll discuss this later on in Lecture 1b. See also the discussion on conditional density in Lecture 1b.*

Example 2.3. *We toss a die. What is the probability that we get a 6, given that we know the toss is even?*

Ans: Let A be the event that we get an even toss, B the event that we get a 6 (when you get used to this, you don't have to explicitly name out the events). Then $P(A) = 1/2$, $P(A \cap B) = P(B) = 1/6$. Thus $P(B|A) = 1/3$.

2.1.3 Bayes' rule

From the definition of conditional probability, we have

$$P(B|A)P(A) = P(B \cap A).$$

It is clear that

$$P(A|B) = \frac{P(B \cap A)}{P(B)}.$$

Therefore, we conclude

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

This formula is called the Baye's rule. At first glance this is pure mathematical manipulation. But it has an important implication: that of switching what we conditioned on. An example would illustrate what this means.

It is well-known that medical test is not 100% reliable. That is suppose you test for a disease, which has 1% chance of happening, then even if the test comes out negative, it doesn't mean you have 0% of contracting the disease. Instead, with a very small probability, it could be a false negative. Concretely, suppose that if you indeed have the disease, then there is 98% chance that the test comes out positive, and 2% negative. However, suppose you don't have the disease, there is 95% chance the test comes out negative, and 5% chance it comes out positive. Now you go for the test, and it comes out negative. What is the probability that you contract the disease?

Ans: Let A be the event that you contract the disease and B be the event that the test is positive. Then we have

$$P(B|A) = .98, P(B^c|A) = .02, P(B|A^c) = .05, P(B^c|A^c) = .95.$$

The question asks for $P(A|B^c)$. Thus you see how Bayes' rule is appropriate for the situation. Can you figure out what it is?

2.2 Independent Events

Definition 2.4. *Two events A and B, are said to be independent if $P(A|B) = P(A)$ and $P(B|A) = P(B)$.*

Remarks: If $P(A|B) = P(A)$ then $P(B|A) = \frac{P(B\cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} = P(B)$. Thus we actually need one of the two equalities given above for the definition of 2 independent events.

Interpretation: Intuitively, two events are independent if the knowledge of one event already happened does not influence the probability of the other happening, hence the definition.

Alternatively, one can define A and B to be independent if $P(A\cap B) = P(A)P(B)$. You should check that this is equivalent to the condition $P(A|B) = P(A)$ given in the definition. So in fact one have two possible ways to define what it means for 2 events to be independent. The interpretation of the equality $P(A \cap B) = P(A)P(B)$ is not very clear (at least to me) so I prefer to use the other equality for definition of independence.

3 Random variables

3.1 Definition

In an experiment, we have (random) outcomes. We can give them names (for example tossing a coin twice, we can get $HH, TT\cdots$). Each of these have some weight attached to them, i.e. their probability (in the coin toss example, 1/4 for each). However, we cannot do computations with these outcomes unless we give them some numerical values. A *random variable* is a way to *quantify* the random outcomes in a meaningful manner. We use capital letters at the end of the alphabet: X, Y, Z, to denote random variables.

Formally, a random variable (*from now on abbreviated as* RV) X is a mapping from the set of outcomes to the real line ($\mathbb{R}$) such that all sets of the form $\{X \in [a, b]\}$ are events. That is, we can assign probability to these sets.

Example 3.1. *Let X be a random variable corresponding to a coin toss. That is $X = 1$ is the coin turns up H and $X = 0$ if the coin turns up T. Then we can see that $P(X = 1) = P(X = 0) = 1/2$.*

Note: There is no reason why 1 has to be assigned to H and 0 assigned to T. One can assign a different value to these outcomes and get a different variable, as suited one's purpose. For example, the RV Y such that $Y = 1$ if the coin is H nd $Y = -1$ if the coin is T is also an example of a RV.

Example 3.2. *Let X be a random variable that corresponds to the time one has to wait at the Hill Center's bus stop before one can catch a bus to College Ave. Suppose that the bus arrives every 15 minutes, and they arrive uniformly during any time frame. The we see that $P(a < X < b) = \frac{b-a}{15}$, for $0 \leq a \leq b \leq 15$. Also one should observe that $P(X = a) = 0$ for any $a \in [0, 15]$ (the probability that one waits exactly 7 minutes before the bus arrives is 0).*

3.2 Discrete versus continuous RVs

In probability theory, one distinguishes between discrete and continuous RVs (note that these are not the only types of RVs there are. One can have a mixed RV as well). Roughly speaking, a discrete RV takes values on a discrete set (for example, the natural numbers is a discrete set, so is $\{1, 2, 3, 4, 5\}$). Moreover, if X is a discrete RV then $P(X = x) > 0$, where x is in the range of X. Examples of discrete RVs that you may have learned are: the Binomial, the Geometric, the Hypergeometric, the Poisson.

A continuous RV, on the other hand, takes values on an interval (or several intervals). Moreover, if X is a continuous RV then $P(X = x) = 0$, even if x is in the range of X. Examples of continuous RVs that you may have learned are: the Exponential, the Normal, the Uniform, the Gamma, the Cauchy.

3.3 Probability distribution, pdf, cdf

3.3.1 Discrete RV

To characterize a discrete RV, we use the probability distribution function. It gives the formula for the probability that the RV takes some specific value. For example, if X has Bionimial(n,p) distribution, then $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ is the distribution function of X.

3.3.2 Continuous RV

To characterize a continuous RV, we use the probability density function (pdf). The pdf does not give a probability itself, but it is connected to a probability via the following formula:

$$P(X \leq x) = \int_{-\infty}^{x} f_X(u) du,$$

where f_X above is the pdf of the RV X.

3.4 cdf

Both continuous and discrete RVs can also be described via the cumulative distribution function, which gives the formula for the probability that the RV is less than or equal to some value:

$$F_X(x) = P(X \leq x).$$

Note that if X is a continuous RV, then F_X is differentiable, and its derivative is the density function f_X.

3.5 The moments

3.5.1 Discrete RV

Let X be a discrete RV. Then its first moment, the Expectation, is defined as:

$$E(X) = \sum_n nP(X = n),$$

where the sum is understood to be taken over all values in the range of X.

It can be showed (note: not a definition) that for any function f, the expectation of the RV $f(X)$ is

$$E(f(X)) = \sum_n f(n)P(X = n).$$

In particular, we have the kth moment of X is $E(X^k) = \sum_n n^k P(X = n)$.

3.5.2 Continuous RV

For a continuous RV X, we define the expectation as:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx.$$

More generally, for any function g, we have

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

3.5.3 Variance, covariance, correlation

Let X be a RV. We then define its variance as

$$Var(X) = E\Big[(X - E(X))^2\Big] = E(X^2) - E^2(X).$$

The variance measures how "spread out" the RV is from its mean.

Let X, Y be RVs. We define their covariance as

$$Cov(X, Y) = E\Big[(X - E(X))(Y - E(Y))\Big] = E(XY) - E(X)E(Y).$$

The covariance measures how "correlated" two RVs are with respect to each other. There is a catch, two different pair of RVs may have the same degree of correlation, but their covariance may be very different. For example, it is clear that

$$Cov(X, X) = Var(X).$$

Intuitively, the degree of "correlation" between X and X, versus $100X$ and $100X$ should be the same (they are perfectly correlated in each case). However, you can easily check that $Cov(100X, 100X) = 10000Cov(X, X)$. Thus we need to introduce another quantity that measures only the correlation and not affected by scaling of the RVs. That is the correlation:

Let X, Y be RVs. We definte their correlation as

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

3.6 Joint distribution, joint pdf

When we have 2 RVs X, Y, besides describing each individual distribution of X, Y, we also need to know how they interact together. The joint distribution (in the discrete case) or the joint pdf (in the continuous case) gives us this information. In fact, to calculate $E(XY)$ in the Covariance formula we would need to use the joint distribution of X, Y.

a. Discrete: Let X, Y be discrete RVs. Then the joint distribution of X, Y is $P(X = x, Y = y)$.

b. Continuous: Let X, Y be continuous RVs. Then their joint pdf, denoted $f_{X,Y}(x, y)$ is such that

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

Some elementary properties:

a.

$$\sum_{x,y} P(X = x, Y = y) = 1.$$

b.

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u, v) du dv = 1.$$

c. Discrete:

$$E(XY) = \sum_{x,y} xy P(X = x, Y = y).$$

d. Continuous:

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} uv f_{X,Y}(u, v) du dv.$$

More generally

e. Discrete:

$$E(g(X,Y)) = \sum_{x,y} g(x,y)P(X=x,Y=y).$$

f. Continuous:

$$E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u,v)f_{X,Y}(u,v)dudv.$$

3.7 Independence

Two random variables X, Y are independent if all events they generated are independent. More specifically, X, Y are independent if for all x, y:

$$P(X \le x, Y \le Y) = P(X \le x)Y(\le y).$$

An easier criterion to check is if the joint distribution "splits", i.e.

$$P(X=x,Y=y) = P(X=x)P(Y=y)(\text{ discrete}), \text{ or}$$

$$f_{XY}(x,y) = f_X(x)f_Y(y) \text{ (continous) }.$$

An important property is that if X, Y are independent then $E(XY) = E(X)E(Y)$. Note that, the reverse implication is not generally true. That is $E(XY) = E(X)E(Y)$ does NOT imply that X, Y are independent. See the following example.

Example 3.3. *Let X have the following distribution: $P(X=1) = P(X=0) = P(X=-1) = 1/3$, and let $Y = X^2$. Then it is clear that X, Y are NOT independent (you should try to show this using the definition of independence). However, we can also easily check that*

$$E(XY) = E(X)E(Y) = 0.$$