Basic Probability

1 Elementary facts

Combinatorics The number of ways to arrange n objects in order is:

$$n! = n(n-1)(n-2)\cdots 1 \text{ (and } 0! = 1).$$

The number of ways to choose r objects from n objects is:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

For $n_1 + n_2 + \ldots n_r = n$, the number of ways to choose n_1 objects of type 1, n_2 objects of type 2, up to n_r objects of type r, is

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

Definitions These are the basic definitions for talking about probability.

The *sample space* Ω is the set of outcomes in an experiment.

An *event* is a subset E of Ω such that $P(E)$ is defined (an event is also sometimes called a *measurable* subset). Events will always be defined so that if A is an event, then 1) the complement of A, denoted A^C is an event, and 2) if $A_1, A_2, \ldots$ are a sequence of events, then $\cup_{i=1}^{\infty} A_i$ will also be an event. (Any set of events satisfying 1) and 2) is called a σ-algebra or σ-field.)

P is a function that given an event A, tells the probability that the outcome lies in A.

The events A and B are *disjoint* or *mutually exclusive* if $A \cap B = \emptyset$.

Measures A probability is a special type of measure that obeys the following three rules:

Axiom 1: $0 \leq P(B) < \infty$ (probabilities are finite positive real numbers)

Axiom 2: $P(\Omega) = 1$ (the probability that something occurs is 1).

Axiom 3: For $B_1, B_2, \ldots$ disjoint events,

$$P\left(\cup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i).$$

Simple facts Some basic facts follow from these axioms.

Prop: $0 \leq P(A) \leq 1$.

Prop: $P(A^C) = 1 - P(A)$.

Prop: $P(A \cup B) = P(A) + P(B) - P(AB)$

Prop: $P(\emptyset) = 0$.

A word about intersection For sets A and B, the intersection of A an B can be denoted $A \cap B$, AB, or A, B. All of these notations mean the same thing:

$$A \cap B := \{x : x \in A \text{ and } y \in B\}.$$

Conditional probabilities If $P(B) > 0$, the conditional probability of A given B is

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

Bayes' Formula If $F_1, \ldots, F_n$ are disjoint and $\cup_{i=1}^n F_i = \Omega$, then

$$P(F_i \mid A) = \frac{P(A \mid F_i)P(F_i)}{P(A \mid F_1)P(F_1) + \ldots P(A \mid F_n)P(F_n)}.$$

Random variables A *random variable* is a function of the outcome. The values the random variable can take on are called *states*, and lie in the *state space*. In other words, a random variable is a function from the sample space to the state space.

For a discrete random variable X, the expected value of X is

$$E[X] = \sum_{x:p(x)>0} xp(x) = \sum_{\omega \in \Omega} X(\omega)P(\{\omega\}).$$

For any two random variables X and Y,

$$E[X + Y] = E[X] + E[Y].$$

Independence Two events A and B are *independent* if

$$P(AB) = P(A)P(B) \Leftrightarrow P(A \mid B) = P(A).$$

Two random variables X and Y are independent if for any event $X \in A$ and $Y \in B$,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

2 A short guide to solving probability problems

Equally likely outcomes. If all outcomes are equally likely,

$$P(E) = \frac{\text{number of outcomes in } E}{\text{total number of outcomes}}.$$

Trick #1: Use complements. It is often easier to find $P(A^C)$ then $P(A)$, remember

$$P(A) = 1 - P(A^C).$$

Trick #2: Use independence to turn intersections into products. If we want the probability of the intersection of $A_1, \ldots, A_n$, then we can break it apart only when the events are independent:

$$P(A_1 \cdots A_n) = P(A_1)P(A_2) \cdots P(A_n).$$

Trick #3: Use disjointness to turn unions into sums. If the events $A_1, \ldots, A_n$ are disjoint,

$$P(A_1 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \ldots P(A_n).$$

Trick #4: Use Principle of In/Ex to deal with any union. We can *always* break apart unions of events $A_1 \dots A_n$ using the Principle of Inclusion/Exclusion, which we use most often when $n = 2$:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1A_2).$$

Its easier to say the Principle of Inclusion/Exclusion in words than symbols: the probability of any event occurring is the sum of the probabilities that one event occurs minus the sum of the probabilities that 2 events occur plus the sum of the probabilities that 3 events occur etcetera until we reach the probability that all events occur.

Trick #5: Use De Morgan's Laws to covert unions and intersections. Convert back and forth between union and intersection using De Morgan's Laws:

$$(A_1A_2 \cdots A_n)^C = A_1^C \cup A_2^C \cdots \cup A_n^C,$$

$$(A_1 \cup A_2 \cup \cdots \cup A_n)^C = A_1^C A_2^C \cdots A_n^C.$$

Trick #6: Use Bayes' Formula to reverse conditional probabilities. If you know $P(A \mid F_i)$ for all i as well as $P(F_i)$, and want $P(F_i \mid A)$, then use Bayes' Formula.

Trick #7: Acceptance/Rejection Method 1 Suppose that we perform a trial which if successful, has outcomes $A_1, \dots, A_n$. If we fail, then we try again until one of A_1 through A_n occur. Then

$$P(A_i \text{ occurs on final trial }) = P(A_i \text{ on first trial}) \mid \text{first trial a success}) = \frac{P(A_i \text{ on first trial})}{P(\text{first trial a success})}.$$

Trick #8: Acceptance/Rejection Method 2 The other way to tackle acceptance rejection problem is using infinite series. Remember, when $|r| < 1$,

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}.$$

Common errors Try to avoid making these errors! Events use complements, unions, and intersections. A statement like $P(A)^C$ doesn't make sense, since $P(A)$ is a number. What was probably meant was $P(A^C)$. Similarly, use +, - and times for numbers like probabilities, and never for sets. We haven't defined A + B, what was probably intended was $P(A)+P(B)$.

Steps to a problem: If you don't know how to get started on a problem, the following steps usually can get you going:

(1) Write down the sample space. Even if you can't write down the whole sample space, write down some of the outcomes. Make up symbols, like H for head or T for tails or W for win and L for a loss to make writing outcomes easier.

(2) Write down the events that you are given probabilities for, and the event that you are trying to find the probability of (the target event).

(3) See if you can express the target event in terms of union, intersection, or complements of the events that you are given (here is where the five tricks come into play).

Simple checks on an answer: Make sure that your final probabilities lie between 0 and 1. If you know that a set of probabilities must add to 1, then check by actually adding them. If you have a simple intuitive reason to believe that A is more likely than B, check that $P(A) > P(B)$.

3 A short guide to counting

Order matters When order matters, then there are $n!$ ways to order n objects.

Thinking about n choose k. There are several ways of thinking about $\binom{n}{k}$, all of which are equivalent.

(1) It's the number of the ways to choose a subset of size k from a set of size n.

(2) It's the number of ways to order a group of letters $A \dots AB \dots B$ where A appears k times and B appears $n-k$ times.

(3) Given n spaces, it's the number of ways to mark k of those spaces in some way.

(4) It's the number of ways of choosing k out of n trials to be successful.

Multichoosing Now $\binom{n}{n_1,\dots,n_r}$ is similar, in that it generalizes $\binom{n}{k}$. This is because $\binom{n}{k} = \binom{n}{k,n-k}$. The number n multichoose $n_1, n_2, \dots, n_r$ counts the following.

(1) It's the number of the ways to choose a partition of a set of size n where the first subset has size n_1, the second n_2, etcetera.

(2) It's the number of ways to order a group of letters $A_1 \dots A_1 A_2 \dots A_2 \dots A_r \dots A_r$ where A_i appears n_i times.

(3) Given n spaces, it's the number of ways to mark n_1 of those spaces with a 1, n_2 spaces with a 2, up to n_r spaces with n_r.

(4) Suppose each trial has r different outcomes. Then its the number of ways of choosing n_1 trials to have outcome 1, n_2 trials to have outcome 2, up to n_r trials having outcome r.

When all else fails. Almost any problem can be written as a problem with ordering. If you are uncomfortable with n choose r or can't figure out what should be ordered and what shouldn't then give everything in your problem a number and order everything.

For example, what's the probability of choosing a given five card hand from a set of 52 cards? One way: number of outcomes is 1, total number of outcomes is $\binom{52}{5}$, so

$$P(\text{hand}) = \frac{1}{\binom{52}{5}}.$$

Another way: number all the cards $1, \dots, 52$ and order them in any one of 52! ways. Then any outcome where the five cards we are interested in appear first in the ordering of cards works. There are 5! ways to order these cards and (52 - 5)! ways to order the remaining 47 cards, so the total number of outcomes is 5!(47!), so

$$P(\text{hand}) = \frac{5!47!}{52!},$$

which is the same answer as the other way.

Another example: given a random ordering of $MIIIISSSSPP$, what's the probability that it spells $MISSISSIPPI$? Think about numbering every symbol, so we are ordering $x_1x_2x_3x_4x_5x_6x_7x_8x_9x_{10}x_{11}$, where $x_1 = M$, x_2 through x_5 equal I, etc. Then the total number of outcomes is 11!. The number of outcomes that are successful? Well x_1 has to be in first position, x_2, x_3, x_4 and x_5 have to occupy positions $2, 5, 8$, and 10 (which they can do in 4! ways, there are 4! ways to order the x_i that equal S and 2! ways to order the x_i that equal P. So

$$P(MISSISSIPPI) = \frac{1!4!4!2!}{11!}.$$

4 How to find $E[X]$

Step 1 Find the values that X can take on (this is called the *positive support* of X). If X is discrete, this will be either a finite number of values $\{x_0, x_1, \ldots, x_n\}$ or a countable number of values $\{x_0, x_1, x_2, \ldots\}$. If X is continuous, it could be an interval or union of intervals, like $(0, \infty)$ or $(3, 4) \cup [10, 15)$.

Step 2 Use the right formula. If $X \in \{x_0, \ldots, x_n\}$, then

$$E[X] = \sum_{x:p(x)>0} xp(x) = \sum_{i=1}^{n} x_i P(X = i).$$

If $X \in \{x_0, x_1, \ldots\}$, then

$$E[X] = \sum_{x:p(x)>0} xp(x) = \sum_{i=1}^{\infty} x_i P(X = i).$$

If X is continuous in set B, then

$$E[X] = \int_B xf(x)\ \mathrm{d}x.$$

If X takes on values $0, 1, 2, 3, \ldots$, then

$$E[X] = \sum_{i=0}^{\infty} P(X > i).$$

If X is continuous and nonnegative (which means the positive support of X is a subset of $(0, \infty)$) then

$$E[X] = \int_{x=0}^{\infty} P(X > x)\ \mathrm{d}x.$$

Note: If we wish to find $E[g(X)]$ then use

$$E[g(X)] = \sum_{x:p(x)>0} g(x)p(x) = \sum_{i=1}^{\infty} g(x_i)P(X = i),$$

and

$$E[g(X)] = \int_B g(x)f(x)dx.$$

For uncorrelated random variables, $E[XY] = E[X]E[Y]$. Independent random variables are uncorrelated, but uncorrelated random variables might not be independent.

5 How to find Var(X)

Method 1: Use

$$\text{Var}(X) = E[X^2] - (E[X])^2.$$

Method 2: Use

$$\text{Var}(X) = E[(X - E[X])^2].$$

For uncorrelated random variables, $\text{Var}(X+Y) = \text{Var}(X)+\text{Var}(Y)$. Independent random variables are uncorrelated.

6 Distributions

The *distribution* of a random variable is a complete listing of $P(X \in A)$ for all sets A of interest. The distribution also referred to as the law of X, and denoted $\mathcal{L}(X)$. When X and Y have the same distribution, this is denoted

$$X \sim Y, \text{ or } \mathcal{L}(X) = \mathcal{L}(Y).$$

The *distribution function* of a random variable X (also known as the cumulative distribution function) is

$$F(a) = P(X \leq a).$$

This is a function that is bounded, that is, it always lies between 0 and 1. It is also right continuous, that is if $a_1, a_2, a_3, \ldots$ decrease and their limit is a, then limit of $F(a_1), F(a_2), \ldots$ equals $F(a)$.

Because of a theorem from measure theory called the Carathéodory Extension Theorem, knowing F allows computation of $P(X \in A)$ for any A of interest. In particular, if $A = (a, b]$, then $P(X \in A) = F(b) - F(a)$. (Looks a bit like the fundamental theorem of calculus, which is one reason why F is always capitalized when used for the distribution function.)

More precisely, if F_X is the distribution function of X and F_Y is the distribution function of Y, then

$$\mathcal{L}(X) = \mathcal{L}(Y) \iff F_X(a) = F_Y(a)\ \forall a.$$

If X is discrete then the graph of $F(a)$ will have jumps, if X is continuous then $F(a)$ will be continuous. Some more formulas that come in handy:

$$\begin{aligned}
P(a < X \leq b) &= F(b) - F(a) \\
P(a < X < b) &= F(b) - F(a) - P(X = b) \\
P(a \leq X < b) &= F(b) - F(a) - P(X = b) + P(X = a) \\
P(a \leq X \leq b) &= F(b) - F(a) + P(X = a).
\end{aligned}$$

Remember that for continuous random variables $P(X = s) = 0$ for any s, so the right hand side of these formula just becomes $F(b) - F(a)$. Also for continuous X,

$$f(a) = \frac{dF(a)}{da}$$

and

$$F(a) = \int_{-\infty}^{a} f(a)da,$$

where $f(x)$ is the *probability density function* (sometimes just called the density) of X.

Finally, say that $X_1, X_2, \ldots$ are independent identically distributed, or i.i.d., if they are independent and all have the same distribution.

6.1 Discrete distributions

A random variable is *discrete* if it only takes on a finite or countable number of values. The distribtuion of a discrete random variable is also called discrete in this instance. Discrete r.v.s have probability mass functions, where $p(X) = P(X = i)$.

Uniform Written: $U\{1, \ldots, n\}$ or dn. What its like: rolling a fair die with n sides.

$$\begin{aligned} P(X = i) &= \frac{1}{n} 1_{\{1,\ldots,n\}} \\ E[X] &= \frac{n+1}{2} \\ \text{Var}(X) &= \frac{(n-1)(n+1)}{12} \end{aligned}$$

Bernoulli Written: $Bern(p)$. What its like: flipping a coin once that comes up heads with probability p and counting the number of heads. Also, the number of successes in a single trial where the trial is a success with probability p.

$$\begin{aligned} P(X = 1) &= p \\ P(X = 0) &= 1 - p \\ E[X] &= p \\ \text{Var}(X) &= p(1-p). \end{aligned}$$

Binomial Written: $Bin(n, p)$. What it's like: flipping i.i.d coins n times where the probability of heads is p and counting the number of heads. Also, the number of successes in a single trial where the trial is a success with probability p.

Also $X = X_1 + X_2 + \ldots X_n$, where X_i are i.i.d. and distributed as $Bern(p)$.

$$\begin{aligned} P(X = i) &= \binom{n}{i} p^i (1-p)^{n-i} 1_{\{0,\ldots,n\}} \\ E[X] &= np \\ \text{Var}(X) &= np(1-p). \end{aligned}$$

Geometric Written: $Geo(p)$. What it's like: flipping i.i.d. coins with probability p of heads and counting the number of flips needed for the first head. Also, the number of trials needed for 1 success when the probability of success at each trial is p and each trial is independent.

$$\begin{aligned} P(X = i) &= (1-p)^{i-1} p 1_{\{0,1,\ldots\}} \\ E[X] &= \frac{1}{p} \\ \text{Var}(X) &= \frac{1-p}{p^2}. \end{aligned}$$

Negative Binomial Written: $NB(r,p)$. What it's like: flipping i.i.d. coins with probability p of heads and counting the number of flips needed for r heads to arrive. Also, the number of trials needed for r successes when the probability of success at each trial is p and each trial is independent.

Also $X = X_1 + X_2 + \dots X_r$, where X_i are i.i.d. and distributed as $Geo(p)$.

$$\begin{aligned} P(X=i) &= \binom{i-1}{r-1} p^r (1-p)^{i-1} 1_{\{0,1,\dots\}} \\ E[X] &= \frac{r}{p} \\ \mathrm{Var}(X) &= r\frac{1-p}{p^2}. \end{aligned}$$

Hypergeometric Written: $HG(N,m,n)$. What it's like: drawing n balls from an urn holding m green balls and $N-m$ blue balls and counting the number of green balls chosen.

$$\begin{aligned} P(X=i) &= \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} 1_{\{0,1,\dots,n\}} \\ E[X] &= \frac{nm}{N} \\ \mathrm{Var}(X) &= \frac{N-n}{N-1} np(1-p). \end{aligned}$$

Zeta Written: $Zeta(\alpha)$. A.k.a. Zipf or power law. What it's like: things like city sizes and incomes have Zeta distributions.

$$\begin{aligned} P(X=i) &= \frac{C}{i^{\alpha+1}} 1_{\{1,2,\dots\}} \\ E[X] &= ???? \\ \mathrm{Var}(X) &= ????. \end{aligned}$$

Special notes: Except for special values of α like 1, we do not have a closed form solution for the value of C, the normalizing constant. Choose C so that $\sum_{i=1}^{\infty} P(X=i) = 1$. Similarly, there are no closed form solutions for $E[X]$ or $\mathrm{Var}(X)$. These must be evaluated numerically. When $\alpha < 1$, $E[X]$ does not exist (or is considered infinite). Similarly, when $\alpha < 2$, $\mathrm{Var}(X)$ does not exist (or can be considered infinite).

Poisson Written: $Pois(\lambda)$. What it's like: Given occurences that happen at rate λ, it is the number of occurences that happen in 1 unit of time. Given an i.i.d. supply of exponential random variables with parameter λ, call them $X_1, X_2, \ldots$, it is

$$\max_i X_1 + X_2 + \ldots + X_i < 1.$$

$$\begin{aligned} P(X = i) &= e^{-\lambda}\frac{\lambda^i}{i!}1_{\{0,1,\ldots\}} \\ E[X] &= \lambda \\ \text{Var}(X) &= \lambda. \end{aligned}$$

6.2 Continuous Distributions

A random variable is *continuous* if $P(X = a) = 0$ for all a. The distribution of a continuous random variable is also called continuous.

Uniform Written: $U[a, b]$ Variations: $U(a, b), U(a, b], U[a, b)$ What it is: choosing randomly a real number from the interval (a, b).

$$\begin{aligned} f(x) &= \frac{1}{b-a}1_(a,b) \\ F(x) &= \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases} \\ E[X] &= \frac{b+a}{2} \\ \text{Var}(X) &= \frac{(b-a)^2}{12} \end{aligned}$$

Normal Written: $N(\mu, \sigma^2)$. What it is: the distribution that comes out of the Central Limit Theorem.

$$\begin{aligned} f(x) &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \\ F(x) &= \Phi\left(\frac{x-\mu}{\sigma}\right) \\ E[X] &= \mu \\ \text{Var}(X) &= \sigma^2 \end{aligned}$$

Addition of normals. Adding independent normal random variables gives back another normal random variable. If $X_i \sim N(\mu_i, \sigma_i^2)$, and $X = X_1 + X_2 + \ldots + X_n$, then

$$X \sim N(\sum_i \mu_i, \sum_i \sigma_i^2).$$

For X, Y independent $N(0, 1)$ random variables, the joint distribution of (X, Y) is rotationally invariant.

Normal random variables are symmetric around μ, and so $\Phi(x) = 1 - \Phi(-x)$.

Exponential Written: $Exp(\lambda)$. What it is: when events occur continuously over time at rate λ, this is the time you have to wait for the first event to occur.

$$
\begin{aligned}
f(t) &= \lambda e^{-\lambda t} 1_{(0,\infty)} \\
F(t) &= \begin{cases} 1 - e^{-\lambda t} & a \geq 0 \\ 0 & a < 0 \end{cases} \\
E[X] &= \frac{1}{\lambda} \\
\mathrm{Var}(X) &= \frac{1}{\lambda^2}
\end{aligned}
$$

7 How to use the Central Limit Theorem (CLT)

The CLT says that if $X_1, X_2, \ldots$ are identically distributed random variables and $Z_n = X_1 + \ldots X_n$, then

$$\lim_{n\to\infty} P\left(\frac{Z_n - E[Z_n]}{\sqrt{\mathrm{Var}(Z)}} \leq a\right) = \Phi(a).$$

We use it as an approximation tool for $Z = X_1 + \ldots X_n$:

$$P\left(\frac{Z - E[Z]}{\sqrt{\mathrm{Var}(Z)}} \leq a\right) \approx \Phi(a).$$

Often we are interested in approximating the probability of things like $P(Z \leq b)$ where $Z = X_1 + \ldots X_n$. This takes two steps.

Step 1 If Z is integral, apply the half integer correction. So instead of $P(Z = i)$ we write $P(i - 0.5 < Z < i + 0.5)$. This makes $P(Z \leq b)$ go to $P(Z \leq b + 0.5)$ (assuming b is also an integer).

Step 2 Subtract off $E[Z]$ and divide by the square root of $\mathrm{Var}(Z)$. So

$$P(Z \leq b + 0.5) = P\left(\frac{Z - E[Z]}{\sqrt{\mathrm{Var}(Z)}} \leq \frac{b + 0.5 - E[Z]}{\sqrt{\mathrm{Var}(Z)}}\right).$$

Step 3 Apply the CLT and say

$$P(Z \leq b) \approx \Phi\left(\frac{b + 0.5 - E[Z]}{\sqrt{\mathrm{Var}(Z)}}\right).$$