Math 180A: Introduction to Probability

Suhas Arehalli

December 6, 2015

1 Preliminaries

First, fundamental definitions:

Experiment A procedure that can lead to certain outcomes.

Sample Space A set Ω of all possible outcomes of an experiment. The "image" of an experiment.

Event A subset of the sample space Ω.

Now, using this, we can define **Kolmogorov's Axioms of Probability**, which tell us that a function $P : \Omega \to [0, 1]$ is a probability if and only if

1. $\forall A \subset \Omega, 0 \leq P(A) \leq 1$. Probabilities must be positive and less that 1.

2. $P(\Omega) = 1$. The probability of having an outcome must be 1.

3. Given a disjoint A and B, $P(A \cup B) = P(A) + P(B)$. The probability of either 2 mutually exclusive events occurring is the sum of their individual probabilities.

4. Given an infinite sequence $A_1, A_2, ...$ of pairwise disjoint ($A_i \cap A_j = \emptyset$) events,
$$P\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} P(A_i)$$
This is extends Axiom 3 to the infinite case.

We also have several useful properties that come out of this:

Complementary Probabilities $P(A^c) = 1 - P(A)$.
The probability of the complement of an event A is $P(A^c) = P(\Omega \setminus A)$, and since $(\Omega \setminus A) \cup A = \Omega$, $P(A^c) + P(A) = 1$ and thus $P(A^c) = 1 - P(A)$. Calculating complementary probabilities can sometimes be easier that calculating probabilities directly.

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

We know

$$P(A) = P(A \cap B) + P(A \cap B^c)$$
$$P(B) = P(A \cap B) + P(A^c \cap B)$$

Thus

$$P(A) + P(B) = 2P(A \cap B) + P(A \cap B^c) + P(A^c \cap B)$$

and since

$$P(A \cap B^c) + P(A^c \cap B) + P(A \cap B) = P(A \cup B)$$

we can rearrange to obtain the desired equality.

Monotonicity If $A \subset B$, the $P(A) \leq P(B)$.

if $A \subset B$, then $B \cap A^c \neq \emptyset$ and $A \cap B = A$, and thus $P(B) - P(A) = P(A^c \cap B) \geq 0$, which shows the necessary inequality.

Now we can discuss independence:

Independence Two events, A and B, are independent if and only if

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

That's pretty quick. Now to random variables and distributions.

Random Variable A random variable is a function $X : \Omega \to \mathbb{F}$ for some field $\mathbb{F}$ ($\mathbb{Z}$ or $\mathbb{N}$ for discrete random variables and $\mathbb{R}$ for continuous) that maps numbers to outcomes of an experiment. Common choices are the sums of the faces of 2 rolled dice, then number of heads after a series of coin flips, etc. You're really never *that* formal about it.

Distribution (Discrete) A distribution of a random variable X is a probability $P(X = x)$ for all x in the image of X. These also obey the probability axioms, with the sample space being the image of X.

Expectation (Discrete) The expectation, or expected value of a discrete random variable X is

$$E[X] = \sum_x xP(X = x)$$

consider it an average of outcomes (made numerical by our random variable) weighted by their probabilities.

Expectation is Linear. Thus, for scalars a_i and random vars X_i

$$E\left[\sum_i a_i X_i\right] = \sum_i a_i E[X_i]$$

Moments The expectation is a specific case of a moment, particularly the 1st moment. the kth moment of a random variable is defined as

$$E[X^k] = \sum_x x^k P(X = x)$$

Variance We define the variance as the centered 2nd moment, namely

$$var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

And that's the basics.

2 Combinatorial Probability

The following rules allow you to count the outcomes of nearly any experiment

Multiplication Rule Given m experiments with the ith experiment having n_k outcomes, we have $\prod_{i=0}^{m} n_k$ total outcomes

Ordering Given a set of n objects, we can arrange them in $n!$ unique orders.

Combinations The number of ways to choose k objects from a set of n is the binomial coefficient

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

Partitioning The number of ways to partition a set into m groups of sizes $n_1, ..., n_m$ with $\sum_i n_i = n$ is the multinomial coefficient

$$\binom{n}{n_1, ..., n_m} = \frac{n!}{\prod_i (n_i!)}$$

With these counting tools, we can calculate probabilities of complex events

Urn Problems The probability of choosing k items from a subset of size m from a set of n items is

$$\frac{\binom{m}{k}}{\binom{n}{k}}$$

Binomial Distribution Suppose a random variable $X \sim binom(n, p)$, or is distributed according to a binomial distribution with parameters n and p. Then

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

which represents the probability of having exactly k successes after n trials of an experiment with a success rate of p.

Consider that there are $\binom{n}{k}$ ways of choosing where the k successes occur, and for each of these trials, the trial occurs with probability $p^k(1-p)^{n-k}$. Thus the total probability is as given.

Expected Value: Rephrase in terms of indicator random variables X_i, which are 1 if the ith trial succeeds and 0 otherwise.:

$$X = \sum_{k=1}^{n} X_i$$
$$E[X] = E\left[\sum_{k=1}^{n} X_i\right] = \sum_{k=1}^{n} E[X_i]$$

and since $E[X_i]$ is simply p,

$$E[X] = \sum_{k=1}^{n} p = np$$

Variance: As above,

$$var[X] = var\left[\sum_{k=1}^{n} X_i\right] = \sum_{k=1}^{n} var[X_i] = n(var[X_i])$$

to find $var[X_i]$,

$$E[X_i] = (1)(p) + (0)(1-p)$$
$$E[X_i^2] = (1^2)p + (0^2)(1-p)$$
$$E[X_i^2] = p$$
$$var[X_i] = E[X_i^2] - E[X_i]^2$$
$$var[X_i] = p - p^2 = p(1-p)$$

thus,

$$var[X] = np(1-p)$$

These can easily extend to experiments with more than one outcome (using the multinomial distribution), but this distrbution is higher dimensional, so don't worry too much about the distribution.

The binomial distribution isn't always easy to calculate, especially when n becomes large. However, we can use the **Poisson Distribution** to approximate the binomial distribution for large n

Poisson Distribution Suppose $X \sim poisson(\lambda)$. Then

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k \in \mathbb{N}$$

Expected Value: Note that $e^x = \sum_{k=0} \frac{x^k}{k!}$. Taylor series.

$$\begin{aligned}
E[X] &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{(k-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} e^{\lambda} = \lambda
\end{aligned}$$

Variance: Here, note that $E[X^2] = E[X(X-1)] + E[X]$

$$\begin{aligned}
E[X(X-1)] &= e^{-\lambda} \sum_{i=0}^{\infty} (k)(k-1) \frac{\lambda^k}{k!} \\
&= \lambda^2 e^{-\lambda} \sum_{i=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \\
&= \lambda^2 \\
var[X] &= E[X^2] - E[X]^2 = E[X(X-1)] + E[X] - E[X]^2 \\
&= \lambda^2 + \lambda - \lambda^2 = \lambda
\end{aligned}$$

Poisson Approximation Theorem Suppose $S_n \sim binom(n, p_n)$ and $X \sim poisson(\lambda)$. If, $\lim_{n\to\infty} p_n = 0$ and $\lim n \to \infty np_n = \lambda$, then

$$\lim_{n\to\infty} P(S_n = k) = P(X = k)$$

Thus, if we have a large n in a binomial distribution, we can allow $\lambda = np$ and approximate it using the Poisson distribution.

3 Conditional Probability

Conditional Probability The probability of A given B is defined as

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

Note that if we let $F(A) = P(A|B)$, $F : \Omega \to [0, 1]$ is a probability as defined by Kolmogorov's axioms. It can also be rephrased as

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

Which tells us that A and B are independent if and only if

$$P(A|B) = P(A)$$

Law of Total Probability If we create a **partition** of Ω $B_1, ..., B_n$, meaning $\bigcup_i B_i = \Omega$ and $B_i \cap B_j = \emptyset, \forall i \neq j$, then

$$P(A) = \sum_{i=0}^{n} P(A|B_i)P(B_i)$$

Bayes' Rule Using the above, we can obtain

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$

Now we can consider joint probabilities: the probability that 2 or more random variable both equal specific values.

Joint Probability Distribution We define a joint probability distribution for random variables X and Y as a distribution

$$P(X = x, ..., X_n = x_n) = P(X = x \cap ... \cap X_n = x_n)$$

Marginal Distribution Given a joint distribution $P(X_1, ...X_n)$, we can find a marginal distribution $P(X_i = k)$ since

$$P(X_i = k) = \sum_{x_1} ... \sum_{x_{i-1}} \sum_{x_{i+1}} ... \sum_{x_n} P(X_1 = x_1, ..., X_i = k, ..., X_n = x_n)$$

Basically, sum over everything *but* the random variable you want a distribution for. With 2 random variables.

$$P(X = k) = \sum_{y} P(X = k, Y = y)$$

Independence Redux We can also say 2 random variables are independent if

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

We can now talk about conditional distributions. Again, exactly what you'd expect.

Conditional Distribution

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x, Y = y)}{\sum_u P(X = x, Y = y)}$$

Conditional Expectation

$$E[X|Y = y] = \sum_{x} xP(X = x|Y = y)$$

4 Continuous Distributions

So far we've discussed distributions that take discrete values. Now we can move to distributions of random variables that take continuous values.

Continuous Random Variable A function from $X : \Omega \to \mathbb{R}$ that maps outcomes to continuous, real values. See above for more info on Random Variables.

Probability Density Function A function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) \geq 0$$
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

is a Probability Density Function for a continuous random variable X if

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Expected Value (Continuous) For a continuous random variable X, the Expected Value of X is

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

And further, for some function $r : \mathbb{R} \to \mathbb{R}$

$$E[r(X)] = \int_{-\infty}^{\infty} r(x)f(x)dx$$

Cumulative Distribution Function We define the distribution function $F : \mathbb{R} \to \mathbb{R}$ for continuous random variable X as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

Note the parallels to the antiderivative. Namely

$$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$$

And now let's give examples of common continuous distributions

Uniform Distribution let $X \sim uniform(a, b)$.

PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & otherwise \end{cases}$$

Expected Value

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_{a}^{b} (x) \left(\frac{1}{b-a} \right) dx \\ &= \left(\frac{1}{b-a} \right) \int_{a}^{b} x dx \\ &= \left(\frac{1}{b-a} \right) \left[\frac{x^2}{2} \right]_{a}^{b} \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{b+a}{2} \end{aligned}$$

CDF

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(y) dy \\ &= \int_{a}^{x} \left(\frac{1}{b-a} \right) dy, x \in [a, b] \\ &= \frac{x-a}{b-a} \\ F(x) &= \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x \geq b \end{cases} \end{aligned}$$

Exponential Distribution Let $X \sim exp(\lambda)$.

PDF

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

Expected Value

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} f(x)dx \\ &= \int_{0}^{\infty} \lambda e^{-\lambda x} dx \\ &= \lambda \int_{0}^{\infty} e^{-\lambda x} dx \\ &= \lambda \left[-\frac{1}{\lambda} e^{-\lambda x} dx \right]_{0}^{\infty} \\ &= 0 - (-1) = 1 \end{aligned}$$

CDF

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(x)dx \\ &= \int_{0}^{x} \lambda e^{-\lambda x} dx \\ &= \left[-e^{-\lambda x} \right]_{0}^{x} \\ &= -e^{-\lambda x} - (-1) \\ &= 1 - e^{-\lambda x} \end{aligned}$$

Now lets discuss a few transformations that are useful theoreticall and practically.

Reduction to the Uniform . Let X have some arbitrary continuous distribution. Then $Y = F(X)$ is uniform on $(0, 1)$.

Define the inverse of F, F^{-1}, as

$$F^{-1} = min\{x : F(x) \geq y\}$$

Since a distribution is not necessarily bijective, we just use the minimum satisfying value. Now we have,

$$\begin{aligned} P(F(X) < y) &= P(X < F^-1(y)) \\ &= F(F^-1(y)) = y \end{aligned}$$

Which is the CDF of the uniform distribution on $[0, 1]$, thus,

$$Y \sim uniform(0, 1)$$

Construction from Uniform Let $X \sim uniform(0, 1)$. Then $Y = F^{-1}(X)$ has distribution function F.

Note that, as defined, $F^{-1}(y) \leq x$ iff $F(x) \geq y$. Then

$$\begin{aligned} P(F^{-1}(X) \leq y) &= P(U \leq F(y)) \\ &= F(x) \end{aligned}$$

(Note that $P(X \leq x)$ for $X \sim Uniform(0, 1)$)

And now for general functions of random variables

PDFs of r(X) Given a random variable X with PDF $f_X(x)$, the distribution function of $Y = r(X)$, $f_Y(y)$ is

$$f_Y(y) = f(r^{-1}(y))\frac{d}{dy}\left[r^{-1}(y)\right]$$

This follows from the chain rule. Calculate using the CDF and the theorems above, then derive and use the chain rule.

$$\begin{aligned} F_Y(y) &= P(r(X) < y) \\ &= P(X < r^{-1}(y)) \\ &= F(r^{-1}(y)) \\ \int_{-\infty}^{y} f_Y(t)dt &= \int_{-\infty}^{y} f_X(r^{-1}(t))dt \\ \frac{d}{dy}\left[\int_{-\infty}^{y} f_Y(t)dt\right] &= \frac{d}{dy}\left[\int_{-\infty}^{y} f_X(r^{-1}(t))dt\right] \\ f_Y(y) &= f_X(r^{-1}(y))\frac{d}{dy}\left[r^{-1}(y)\right] \end{aligned}$$

Now we discuss joint and conditional distributions in terms of Continuous Random Variables.

Joint Density Function A joint density function $f_{X,Y}$ for continuous random variables X, Y if for all $A \subset \mathbb{R}^2$

$$P((x,y) \in A) = \iint_A f_{X,Y}(x,y)dxdy$$

Also note that

$$\iint_{\mathbb{R}^2} f_{X,Y}dxdy = 1$$

Marginal Distributions (Continuous) Given a joint density function $F_{X,Y}$ for continuous random variables X, Y, we can calculate

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$cddy\left[r^{-1}(y)\right] f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Independence As you should expect, X and Y are independent iff

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \forall x, y \in \mathbb{R}$$

Conditional Distributions (Continuous) We define

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

and using the marginal distributions,

$$f_X(x|Y=y) = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx}$$

5 Limit Theorems

Here we describe the Normal Distribution and the convergence of sums to distributions to it.

Sums of Random Variables (Discrete) The distributions of the sums of independent discrete random variables can be calculated as the sum over all values of the 2 random variables that sum to the requested value.

$$P(X+Y=k) = \sum_x P(X=x)P(Y=(k-x))$$

Binomial
Let $X \sim binom(n,p)$ and $Y \sim binom(m,p)$ and independent. then

$$X + Y \sim binom(n+m, p).$$

Poisson
Let $X \sim poisson(\lambda)$ and $Y \sim poisson(\mu)$. Then

$$X + Y \sim poisson(\lambda + \mu)$$

$$\begin{aligned}
P(X+Y=k) &= \sum_{x=0}^{\infty} P(X=x)P(Y=(k-x)) \\
&= \sum_{x=0}^{\infty} \left(e^{-\lambda}\frac{\lambda^x}{x!}\right)\left(e^{-\mu}\frac{\mu^{k-x}}{(k-x)!}\right) \\
&= e^{-(\mu+\lambda)}\frac{1}{k!}\sum_{x=0}\left(\frac{k!}{x!(k-x)!}\right)\lambda^x\mu^{k-x} \\
&= e^{-(\mu+\lambda)}\frac{(\mu+\lambda)^k}{k!} \\
X+Y &\sim poisson(\lambda+\mu)
\end{aligned}$$

Sums of Random Variables (Continuous) As above, but integrating rather than summing.

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx$$

Examples are painful, so figure them out yourself (It's left as an exercise to the reader!).

Now let's talk about the mean and variance of these sums.

Expected Value of Sums Since expectation is linear,

$$\begin{aligned}
E[X+Y] &= E[X] + E[Y] \\
E\left[\sum_{i=0}^{n} X_i\right] &= \sum_{i=0}^{n} E[X_i]
\end{aligned}$$

Expected Value of Products If X and Y are independent,

$$\begin{aligned}
E[XY] &= \sum_{x,y} xyP(X=x, Y=y) \\
&= \sum_{x,y} xyP(X=x)P(Y=y) \\
&= \sum_{x} \left(xP(X=x) \left(\sum_{y} yP(Y=y) \right) \right) \\
&= \left(\sum_{y} yP(Y=y) \right) \left(\sum_{x} xP(X=x) \right) \\
&= E[X]E[Y]
\end{aligned}$$

Covariance The covariance, a measure of how the variance of one variable interacts with the variance of the other, is defined as

$$\begin{aligned}
cov[X,Y] &= E[(X - E[X])(Y - E[Y])] \\
&= E[XY - E[X]y - xE[Y] - E[X]E[Y]] \\
&= E[XY] - E[X]E[Y] - E[X]E[Y] - E[X]E[Y] \\
&= E[XY] - E[X]E[Y]
\end{aligned}$$

Which shows that if X and Y are independent,

$$cov[X,Y] = E[X]E[Y] - E[X]E[Y] = 0$$

Variance of Sums By definition

$$\begin{aligned}
var[X+Y] &= E[((X+Y) - E[X+Y])^2] \\
&= E[((x - E[X]) + (Y - E[Y]))^2] \\
&= E[(x - E[X])^2] + 2E[(x - E[X])(y - E[Y])] + E[(y - E[Y])^2] \\
&= var[X] + var[Y] + 2cov[X,Y]
\end{aligned}$$

And if X and Y are independendent.

$$var\left[\sum_{i=0}^{n} X_i\right] = \sum_{i=0}^{n} var[X_i]$$

Now a few more related pieces

Chebyshev's Inequality Given $y \geq 0$.

$$P(|Y - E[Y]| \geq y) \leq \frac{var[Y]}{y^2}$$

Note that

$$\begin{aligned}E[Z] = \int_{\mathbb{R}} z f_Z(z)dz &\geq \int_k^\infty x f_Z(z)dz \\ &\geq \int_k^\infty k f_Z(z)dz \\ &\geq k\int_k^\infty f_Z(z)dz = kP(k \leq Z)\end{aligned}$$

If we let $Z = (Y - E[Y])^2$ and $k = y^2$,

$$\begin{aligned}var[Y] = E[(Y - var[Y])^2] &\geq y^2 P((Y - E[Y])^2 \geq y^2) \\ &\geq y^2 P(|Y - var[Y]| \geq y)\end{aligned}$$

and with rearrangement,

$$P(|Y - E[Y]| \geq y) \leq \frac{var[Y]}{y^2}$$

Now lets move to the 2 capstone results

Law of Large Numbers Given a series of independent and identically distributed (i.i.d.) random variables $X_1, X_2, ...$ with partial means $\bar{X}_n = \frac{1}{n}\sum_{i=0}^n X_i$, and $var[X] = \sigma^2$ and $E[X] = \mu$.

$$\lim_{n\to\infty} P(|\bar{X}_n - E[X]| \geq \varepsilon) = 0$$

$$\begin{aligned}var[\bar{X}_n] &= var\left[\frac{1}{n}\sum_{i=0}^n X_i\right] \\ &= \frac{1}{n^2}\cdot n(var[X_i]) \\ &= \frac{\sigma^2}{n} \\ P(|\bar{X}_n - E[X]| \geq \varepsilon) &\leq \frac{var[\bar{X}_n]}{\varepsilon^2} \\ P(|\bar{X}_n - \mu| \geq \varepsilon) &\leq \frac{\sigma^2}{n\varepsilon^2} \\ \lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \varepsilon) &\leq \lim_{n\to\infty}\frac{\sigma^2}{n\varepsilon^2} = 0 \\ \lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \varepsilon) &= 0\end{aligned}$$

Normal Distribution Given $X \sim N(\mu, \sigma)$ and $Y \sim N(\upsilon, \nu)$,

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

and

$$X + Y \sim N(a + b, u + v)$$

Proofs of these are trivial but painful.

Central Limit Theorem Suppose $X_1, X_2, \ldots$ are i.i.d. with $E[X_i] = \mu$ and $var[X_i] = \sigma^2 < \infty$. Let $S_n = \sum_{i=1}^{n} X_i$ and $\chi \sim N(0, 1)$. As $n \to \infty$,

$$\frac{S_n - \mu}{\sigma\sqrt{n}} \approx \chi$$

Histogram Correction Since the Central Limit Theorem is meant for continuous random variables, we offset this by treating an integer value k as the interval $[k - 0.5, k + 0.5]$. Thus, for a discrete random variable X,

$$P\left(\frac{X - \mu}{\sigma\sqrt{n}} \leq k\right) \approx P\left(\chi \leq k + 0.5\right)$$

$$P\left(\frac{X - \mu}{\sigma\sqrt{n}} \geq k\right) \approx P\left(\chi \geq k - 0.5\right)$$

And now we done, fam.