Probability Refresher

Keith A. Lewis

June 20, 2020

Abstract

This note collects salient facts about probability theory.

Probability is an extension of logic. Instead of propositions being either true or false a degree of belief can be specified for events occurring, or not occuring. All probabilities are conditional on models of available information.

Probability Model

A probability model specifies a sample space and a probability measure. Algebras of events model partial information.

Sample Space

A sample space is what can happen in a model: heads or tails as the outcome of a coin toss, the integers from 1 to 6 as the outcomes of rolling a single die, the set of all sequences of not more than 280 characters as a model of possible Twitter tweets.

An *event* is a subset of a sample space.

People seem to be surprised probabilities are modeled using sets. Sets have no structure, they are just a bag of things (*elements*).

Probability Measure

A *probability measure* assigns a number between 0 and 1 to events. If Ω is a sample space and P is a probability measure then the measure of the union of sets is the sum of the measure of each set minus the measure of the intersection: $P(E\cup F) = P(E)+P(F)-P(E\cap F)$ for events E and F. This is the mathematical way to say measures do not double count.

A probability measure must also satisfy $P(\emptyset) = 0$ and $P(\Omega) = 1$.

Exercise. If Q is a measure with $Q(\emptyset) = a$ and $Q(\Omega) = b$, show $(Q - a)/(b - a)$ is a probability measure.

Algebra

An *algebra of sets*, or *algebra*, on Ω is a collection of subsets (events), $\mathcal{A}$, that is closed under complement and union. This lets us talk about an event not happening and whether event A or B occured.

We also assume the empty set belongs to $\mathcal{A}$, hence also Ω. By De Morgan's Laws an algebra is also closed under intersection. The *power set* of Ω, $\mathcal{P}\Omega = \{E : E \subseteq \Omega\}$, clearly satisfies these conditions.

The set $2^\Omega = \{\xi : \Omega \to \{0, 1\}\}$ is isomomorphic to the power set. The function 1_A selects subsets of Ω: the set $A \subset \Omega$ corresponds to the function $1_A(\omega)$ where $1_A(\omega)$ is 1 if $\omega \in A$ and 0 if $\omega \notin A$.

An *atom* of an algebra is a member, A, of the algebra such that if $B \subseteq A$ and B is in the algebra, then either $B = A$ or B is the empty set.

Partition

A *partition* of a set is a collection of pairwise disjoint subsets whos union is the entire set.

Exercise. If an algebra is finite its atoms are a partition.

Hint: Show $A_\omega = \cap\{B \in \mathcal{A} : \omega \in B\}$ is an atom for all $\omega \in \Omega$.

This shows there is a one-to-one correspondence between finite partitions and finite algebras of sets. A partition is the mathematical way of specifying partial information. Knowing the outcome, $\omega \in \Omega$, corresponds to complete knowledge. Knowing which atom the outcome belongs to corresponds to partial knowledge. For example, the partition $\{\{1, 3, 5\}, \{2, 4, 6\}\}$ corresponds to knowing whether the roll of a die is odd or even.

The coarsest partition, $\{\Omega\}$, corresponds to no knowledge while the finest partition $\{\{\omega\} : \omega \in \Omega\}$ corresponds to complete knowledge.

Measurable

A function $X \colon \Omega \to \mathbf{R}$ is $\mathcal{A}$-*measureable* if the sets $X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\}$ belong to $\mathcal{A}$ for $x \in \mathbf{R}$.

Exercise: If $\mathcal{A}$ is finite, show that a function is measurable if and only if it is constant on atoms of $\mathcal{A}$.

In this case $X \colon \mathcal{A} \to \mathbf{R}$ is indeed a function on the atoms.

Random Variable

A *random variable* is a variable, a symbol that can be used in place of a number, with additional information: the probability of the values it can take on.

Cumulative Distribution Function

The *cumulative distribution function* of the random variable X is $F(x) = F^X(x) = P(X \le x)$. It tells you everything there is to know about X. For example, $P(a < X \le b) = F(b) - F(a)$.

Exercise. Show $P(a \le X \le b) = \lim_{x\uparrow a} F(b) - F(x)$.

Hint: $[a,b] = \cap_n (a - 1/n, b]$.

In general, $P(X \in A) = \int_A dF(x)$ for sufficiently nice $A \subset \mathbf{R}$ where we are using Riemann–Stieltjes integration.

Exercise: Show for any cumulative distribution function, F, that F is non-decreasing, $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$, and F is right continuous with left limits.

Every such function is the cumulative distribution function of a random variable.

The cdf $F(x) = \max\{0, \min\{1, x\}\}$ defines the uniformly distributed random variable, U, on the interval $[0,1]$. For $0 \le a < b \le 1$, $P(a < U \le b) = b - a$.

Two random variables, X and Y, have the same *law* if they have the same cdf.

Exercise. If X has cdf F, then X and $F^{-1}(U)$ have the same law.

Exercise. If X has cdf F, then $F(X)$ and U have the same law.

This shows a uniformly distributed random variable has sufficient randomness to generate any random variable. There are no random, random variables.

The mathematician's definition of a random variable is that it is a measurable function $X\colon \Omega \to \mathbf{R}$. Its cumulative distribution function is $F(x) = P(X \le x) = P(\{\omega \in \Omega \mid X(\omega) \le x\})$. Given a cdf F we can define $X\colon \mathbf{R} \to \mathbf{R}$ to be the identity function and let P be the probability measure defined by F: $P(A) = \int_A dF(x)$.

Expected Value

The *expected value* of a random variable is defined by $EX = \int_{-\infty}^{\infty} x\,dF(x)$. The expected value of any function of a random variable is $Ef(X) = \int_{-\infty}^{\infty} f(x)\,dF(x)$.

If $X = \sum a_i 1_{A_i}$ where $a_i \in \mathbf{R}$ and A_i are events, the *expected value* of X is $EX = \sum_i a_i P(A_i)$.

Exercise. Show that if $\sum_i a_i 1_{A_i} = 0$ then $\sum_i a_i P(A_i) = 0$.

Hint: Replace the A_i by disjoint B_j so $b_j = 0$ for all j.

This shows expected value is well-defined.

Exercise. Show $P(\cup_i A_i) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \cdots$.

Hint: Use $(1_A - 1_{A_1})\cdots(1_A - 1_{A_n}) = 0$, where $A = \cup_{k=1}^n A_k$.

Moments

The *moments* of a random variable, X, are $m_n = E[X^n]$, $n = 0, 1, 2, \ldots$. They don't necessarily exist for all n, except for $n = 0$. They also cannot be an arbitrary sequence of values.

Suppose all moments of X exist, then for any complex numbers, (c_i), $0 \leq E|\sum_i c_i X^i|^2 = E\sum_{j,k} c_j \bar{c_k} X^{j+k} = \sum_{j,k} c_j \bar{c_k} m_{j+k}$. This says the Hankel matrix, $M = [m_{j+k}]_{j,k}$, is positive definite. The converse is also true: if the Hankel matrix is positive definite there exists a random variable with the corresponding moments. This is not a trivial result and the random variable might not be unique.

Cumulant

The *cumulant* of a random variable, X, is $\kappa(s) = \kappa^X(s) = \log E\exp(sX)$. The *cumulants*, (κ_n), are the coefficients of the power series expansion $\kappa(s) = \sum_{n>0} \kappa_n s^n/n!$.

It is easy to see $\kappa_1 = EX$ and $\kappa_2 = \operatorname{Var} X$. The third and fourth cumulants are related to skew and kurtosis. We will see the exact relationship below.

If c is a constant then $\kappa^{cX}(s) = \kappa^X(cs)$ so $\kappa_n^{cX} = c^n\kappa_n^X$. If X and Y satisfy $Ee^{sX}e^{sY} = Ee^{sX}E^{sY}$ then $\kappa^{X+Y}(s) = \kappa^X(s) + \kappa^Y(s)$ and $\kappa_n^{X+Y} = \kappa_n^X + \kappa_n^Y$\$

Bell Polynomial

The relationship between moments and cumulants is given by *Bell polynomials*. They are defined by $\exp(\sum_{n=1}^\infty a_n s^n/n!) = \sum_0^\infty B_n(a_1, \ldots, a_n)s^n/n!$. Taking the derivative with respect to s and equating powers of s shows $B_0 = 1$ and $B_{n+1}(a_1, \ldots, a_{n+1} = \sum_{k=0}^n \binom{n}{k} B_{n-k}(a_1, \ldots, a_{n-k})a_{k+1}$.

Bell polynomials connect moments and cumulants of a random variable. Since $E\exp(sX) = \sum_0^\infty EX^n s^n/n! = \sum_0^\infty m_n s^n/n!$ where m_n is the n-th moment and $E\exp(sX) = \exp(\kappa(s)) = \exp(\sum_{n=1}^\infty \kappa_n s^n/n!)$.

Exercise: Show $m_n = \sum_{k=1}^n B_k(\kappa_1, \ldots, \kappa_n)$.

Exercise: Find the first five Bell polynomials.

In particular $m_1 = \kappa_1$ and $m_2 = \kappa_1^2 + \kappa_2$ so κ_1 is the mean and κ_2 is the variance. If the mean is 0 and the variance is 1, then κ_3 is the skew and κ_4 is the excess kurtosis.

Conditional Expectation

The *conditional expectation* of an event B given an event A is $P(B|A) = P(B \cap A)/P(A)$. In some sense, this reduces the sample space to A since $P(A|A) = 1$.

We also have $P(A|B) = P(A \cap B)/P(B)$ so $P(A|B) = P(B|A)P(A)/P(B)$. This is the simplest form of Bayes Theorem. It shows how to update your degree of belief based on new information. Every probability is conditional on given information.

Define the conditional expectation of the random variable X with respect to the event A by $E[X|A] = E[X1_A]/P(A)$. If $X = 1_B$ then this coincides with the definition of conditional expectation above.

Define the conditional expectation of X with respect to the algebra $\mathcal{A}$, $E[X|\mathcal{A}] : \mathcal{A} \to \mathbf{R}$, by $E[X|\mathcal{A}](A) = E[X|A]$ for A an atom of $\mathcal{A}$.

Joint Distribution

Two random variables, X and Y, are defined by their *joint distribution*, $H(x,y) = P(X \le x, Y \le y)$. For example, the point (X,Y) is in the square $(a,b] \times (c,d]$ with probability $P(a < X \le b, c < Y \le d) = P(X \le b, Y \le d) - P(X \le a) - P(Y \le c) + P(X \le a, Y \le c)$.

The *marginal distbutions* are $F(x) = H(x,\infty)$ and $G(y) = H(\infty,y)$, where F and G are the cumulative distributions of X and Y respectively.

In general, the joint distribution of X_1, ..., X_n is $F(x_1,\dots,x_n) = P(X_1 \le x_1, \dots, X_n \le x_n)$.

Independent

The random variables X and Y are *independent* if $H(x,y) = F(x)G(y)$ for all x and y. This is equivalent to $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for any sets A and B.

We also have that $Ef(X)g(Y) = Ef(X)Eg(Y)$ for any functions f and g whenever all expected values exist.

Exercise: Prove this for the case $f = \sum_i a_i 1_{A_i}$ and $g = \sum_j b_j 1_{B_j}$.

In general, X_1, ..., X_n are independent if $F(x_1,\dots,x_n) = F_1(x_1)\cdots F_n(x_n)$, where F_j is the law of X_j.

Copula

A *copula* is the joint distribution of uniformly distributed random variables on the unit interval. The copula of X and Y is the joint distribution of $F^{-1}(X)$ and $G^{-1}(Y)$ where F and G are the cumulative distributions of X and Y respectively: $C(u,v) = C^{X,Y}(u,v) = P(F^{-1}(X) \le u, G^{-1}(Y) \le v)$.

Exercise: Show $C(u,v) = H(F(u),G(v))$ where and H is the joint distribution of X and Y and F and G are the cumulative distribution of X, and Y.

Exercise: Show $H(x,y) = C(F^{-1}(x), G^{-1}(y))$

This shows how to use the copula and marginal distributions to recover the joint distribution.

An equivalent definition is a copula is a probability measure on $[0,1]^2$ with uniform marginals.

Exercise: Prove this.

If U and V are independent, uniformly distributed random variables on the unit interval then $C(u,v) = uv$.

If $V = U$ then their joint distribution is $C(u,v) = P(U \leq u, V \leq v) = P(U \leq u, U \leq v) = P(U \leq \min\{u,v\}) = \min\{u,v\} = M(u,v)$.

If $V = 1-U$ then their joint distribution is $C(u,v) = P(U \leq u, V \leq v) = P(U \leq u, 1-U \leq v) = P(1-v \leq U \leq u) = \max\{u-(1-v), 0\} = \max\{u+v-1, 0\} = W(u,v)$

Exercise: (Fréchet-Hoeffding) For every copula, C, $W \leq C \leq M$.

Hint: For the upper bound use $H(x,y) \leq F(x)$ and $H(x,y) \leq G(y)$. For the lower bound note $0 \leq C(u_1,v_1) - C(u_1,v_2) - C(u_2,v_1) + C(u_2,v_2)$ for $u_1 \geq u_2$ and $v_1 \geq v_2$.

Examples

Discrete

A *discrete* random variable, X, is defined by $x_i \in \mathbf{R}$ and $p_i > 0$ with $\sum p_i = 1$. The probability the random variable takes on value x_i is P(X = x_i) = p_i.

If a discrete random variable takes on a finite number of values, n, then if $p_i = 1/n$ for all i the variable is called *discrete uniform*.

Bernoulli

A *Bernoulli* random variable is a discrete random variable with $P(X=0) = p$, $P(X=1) = 1-p$.

Binomial

A *Binomial* random variable is a discrete random variable with $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, \ldots, n$.

Uniform

A *continuous uniform* random variable on the interval $[a,b]$ has density $f(x) = 1_{[a,b]}/(b-a)$.

Normal

The *standard normal* random variable, Z, has density function $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$.

If X is normal then $E\exp(N) = \exp(E[N] + \mathrm{Var}(N)/2)$ so the cumulants satisfy $\kappa_n = 0$ for $n > 2$.

This follows from

$$\begin{aligned}
E[e^N] &= E[e^{\mu+\sigma Z}] \\
&= \int_{-\infty}^{\infty} e^{\mu+\sigma z} e^{-z^2/2}\, dz/\sqrt{2\pi} \\
&= e^{\mu+\sigma^2/2} \int_{-\infty}^{\infty} e^{-(z-\sigma)^2/2}\, dz/\sqrt{2\pi} \\
&= e^{\mu+\sigma^2/2} \int_{-\infty}^{\infty} e^{-z^2/2}\, dz/\sqrt{2\pi} \\
&= e^{\mu+\sigma^2/2}
\end{aligned}$$

For any normal random variable, N, $E[e^N f(N)] = E[e^N]E[f(N + \mathrm{Var}(N)]$.

Exercise. Prove this by first showing $E[e^{\sigma Z} f(Z)] = e^{\sigma^2/2} E[f(Z + \sigma)]$.

If N, N_1, $\ldots$, are jointly normal then $E[e^N f(N_1, \ldots)] = E[e^N]E[f(N_1 + \mathrm{Cov}(N, N_1), \ldots)]$.

Poisson

A *Poisson* random variable with parameter λ is defined by $P(X = k) = e^{-\lambda}\lambda^k/k!$ for $k = 0, 1, \ldots$.

If X is Poisson with parameter λ then

$$\begin{aligned}
Ee^{sX} &= \sum_{k=0}^{\infty} e^{sk} e^{-\lambda}\lambda^k/k! \\
&= \sum_{k=0}^{\infty} (e^s\lambda)^k e^{-\lambda}/k! \\
&= \exp(\lambda(e^s - 1))
\end{aligned}$$

so $\kappa(s) = \lambda(e^s - 1)$ and $\kappa_n = \lambda$ for all n.

Infinitely Divisible

A random variable, X, is *infinitely divisible* if for any positive integer, n, there exist independent, identically distributed random variables $X_1, \ldots, X_n$ such that $X_1 + \cdots + X_n$ has the same law as X.

A theorem of Kolmogorov states for every infinitely divisible random variable the exists a number γ and a non-decreasing function G with

$$\kappa(s) = \log Ee^{sX} = \gamma s + \int_{-\infty}^{\infty} K_s(x)\,dG(x),$$

where $K_s(x) = (e^{sx} - 1 - sx)/x^2 = \sum_{n=2}^{\infty} x^{n-2}s^n/n!$. Note if $G(x) = 1_{(-\infty,0]}$ then $\kappa(s) = \gamma s + K_s(0) = \gamma s + s^2/2$ so the random variable is normal.

Note the cumulants of the random variable are $\kappa_1 = \gamma$ and $\kappa_n = \int_{-\infty}^{\infty} x^{n-2}\,dG(x)$ for $n \geq 2$.

If $G(x) = a^2 1_{(-\infty,a]}$ for $a \neq 0$ then

$$\begin{aligned}
\kappa(s) &= \gamma s + a^2 K_s(a) \\
&= \gamma s + a^2 \sum_{n=2}^{\infty} a^{n-2}s^n/n! \\
&= \gamma s + \sum_{n=2}^{\infty} a^n s^n/n! \\
&= \gamma s - as + \sum_{n=1}^{\infty} a^n s^n/n! \\
&= (\gamma - a)s + \sum_{n=1}^{\infty} a^n s^n/n!
\end{aligned}$$

so the random variable is Poisson with parameter a plus the constant $\gamma - a$.

This theorem states every infinitely divisible random variable can be approximated by a normal plus a linear combination of independent Poisson random variables.

If $X = \mu + \sigma Z + \sum_j \alpha_j a_j^2 P_j$ where P_j is Poisson with parameter a_j, then

$$\kappa(s) = \mu s + \sigma s^2/2 + \sum_j \alpha_j(e^{a_j s} - 1) - \alpha_j s$$