Probability

Keith A. Lewis

August 15, 2019

Abstract

This short note collects salient facts about probability theory.

Contents

Probability is an extension of logic. Instead of propositions being either true or false a degree of belief can be specified for events occurring. All probabilities are based on information and different people may have different degrees of belief. However, as the amount of information increases, probabilities should converge.

A *probability model* specifies a *sample space* and a *probability measure*. The sample space is just a set of what can possibly happen: heads or tails as the outcome of a coin toss, the numbers 1 through 6 as the outcomes of rolling a singled die, the set of all sequences of not more than 280 characters as a model of possible Twitter tweets.

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

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

Let $1_A(\omega) = 1$ if $\omega \in A$ and $= 0$ if $\omega \notin A$. If $X = \sum a_i 1_{A_i}$ where $a_i \in \mathbf{R}$ and A_i are events, Define the *expected value* of X by $EX = \sum_i a_i P(A_i)$.

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

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 $A = \cup_{k=1}^n A_k$.

Exercise. (Inclusion-Exclusion principal) Let S be a finite set and let f be any function defined on subsets of S. Define $\phi f(T) = \sum_{U \supseteq T} f(U)$ and $\psi g(T) = \sum_{U \supseteq T} (-1)^{|U|-|T|} g(T)$. These are both operators from $2^S \to \mathbf{R}$. Show $\phi\psi g = g$ and $\psi\phi f = f$.

Hint: Group the sum by $|Y| - |T|$.

Algebras of Sets

An *algebra of sets* on Ω is a collection of subsets, $\mathcal{A}$, that is closed under complement and union. We also assume the empty set belongs to $\mathcal{A}$. By De Morgan's Laws an algebra is also closed under intersection and Ω belongs to $\mathcal{A}$. The *power set* of Ω, $2^\Omega = \{E \subseteq \Omega\}$, clearly satisfies these conditions.

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.

A *partition* of a set is a collection of pairwise disjoint subsets who's union is equal to the set.

Exercise. If an algebra on Ω is finite its atoms form a partition of Ω.

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

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.

Measurablity

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 Variables

The physicist's definition of a random variable that it 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. The *cumulative distribution function* is $F(x) = F^X(x) = P(X \leq x)$ It tells you everything there is to know about X. For example, $P(a < X \leq b) = F(b) - F(a)$. A nice subset of $\mathbf{R}$ can be approximated by a disjoint union of intervals.

In general, $P(X \in A) = E1_A = \int 1_A(x)\,dF(x)$ for sufficiently nice $A \subset \mathcal{R}$. Here we use the Riemann–Stieltjes integral.

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 $F(x) = P(X \leq x) = P(\{\omega \in \Omega \mid X(\omega) \leq x\})$. Given a cdf F we can define recover the physicists random variable by defining $X\colon \mathbf{R} \to \mathbf{R}$ to be the identity function and let P be the probability measure defined by F: $P(A) = \int 1_A(x)\,dF(x)$.

Two random variables, X and Y, are defined by their *joint distribution*, $F(x,y) = F^{X,Y}(x,y) = P(X \leq x, Y \leq y)$. For example (X,Y) is in the square $(a,b] \times (c,d]$ with probability $P(a < X \leq b, c < Y \leq d) = P(X \leq b, Y \leq d) - P(X \leq a) - P(Y \leq c) + P(X \leq a, Y \leq c)$. A nice subset of $\mathbf{R}^2$ can be approximated by a disjoint union of squares.

This is where the mathematician's definition is ???

One probablity measure, and two (or any number) of function.

Expected Value

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

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), $0 \le E|\sum_i c_i X^i|^2 = E\sum_{j,k} c_j\bar{c_k}X^{j+k} = \sum_{j,k} c_j\bar{c_k}m_{j+k}$. This says the Hankel matrix, $M = [m_{j+k}]_{j,k}$, is positive definite. The converse is true: if the Hankel matrix is positive definite there exists a random variable with the corresponding moments.

% Dunford Schwartz Volume 2 pg 1251. % Extending unbounded symmetric operators. Deficiency index.

Cumulants

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

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 are independent then $\kappa^{X+Y}(s) = \kappa^X(s) + \kappa^Y(s)$ so $\kappa_n^{X+Y} = \kappa_n^X + \kappa_n^Y$$

Characteristic Function

The *characteristic function* of a random variable, X, is $\xi(t) = \kappa(it)$.

Fourier Transform.

Examples

If X is normal then $E\exp(X) = \exp(EX + \operatorname{Var}(X)/2)$ so $\kappa_1 = EX$, $\kappa_2 = \operatorname{Var}(X)$, and $\kappa_n = 0$ for $n > 2$.

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.

Bell Polynomials

The relationship between moments and cumulants is given by 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 is the excess kurtosis.

Examples

Uniform

every rv comes from uniform

Normal

Poisson

Infinitely Divisible

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

Characteristic function …

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. In particular, $P(A|A) = 1$. Since $P(A|B) = P(A \cap B)/P(B)$ we have $P(A|B) = P(B|A)P(A)/P(B)$. 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 $E[X|A] = E[X1_A]/P(A)$ for any random variable X. If $X = 1_B$ then this coincides with the definition of conditional expectation above.

If we write this as $E[X|A]P(A) = E[X1_A]$ then defining $E[X|\mathcal{A}]$ by $E[X|\mathcal{A}]P|_{\mathcal{A}} = (XP)_{\mathcal{A}}$ agrees on atoms of $\mathcal{A}$.

moments, Hamburger moment problem.

cumulants, Bell polynomials

Normal

Poisson

Infinitely Divisible

Stochastic Processes

A *stochastic process* is ...

Brownian Motion

reflection

Le'vy Processes

Remarks

Cheval de Mere

Pascal

Bernoulli(s)

Kolmogorov

Willy Feller