CLARK UNIVERSITY

Summary of basic probability theory
Math 218, Mathematical Statistics

D Joyce, Spring 2016

Sample space. A *sample space* consists of a *underlying* set Ω, whose elements are called *outcomes*, a collection of subsets of Ω called *events*, and a function P on the set of events, called a *probability* function, satisfying the following axioms.

1. The probability of any event is a number in the interval $[0, 1]$.

2. The entire set Ω is an event with probability $P(\Omega) = 1$.

3. The union and intersection of any finite or countably infinite set of events are events, and the complement of an event is an event.

4. The probability of a disjoint union of a finite or countably infinite set of events is the sum of the probabilities of those events,

$$P(\bigcup_i E_i) = \sum_i P(E_i).$$

From these axioms a number of other properties can be derived including these.

5. The complement $E^c = \Omega - E$ of an event E is an event, and

$$P(E^c) = 1 - P(E).$$

6. The empty set $\emptyset$ is an event with probability $P(\emptyset) = 0$.

7. For any two events E and F,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F),$$

therefore

$$P(E \cup F) \leq P(E) + P(F).$$

8. For any two events E and F,

$$P(E) = P(E \cap F) + P(E \cap F^c).$$

9. If event E is a subset of event F, then $P(E) \leq P(F)$.

10. Statement 7 above is called the *principle of inclusion and exclusion*. It generalizes to more than two events.

$$\begin{aligned} P(\bigcup_{r=1}^{n} E_r) &= \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) \\ &+ \sum_{i<j<k} P(E_i \cap E_j \cap E_k) - \cdots \\ &+ (-1)^{n-1} P(E_1 \cap E_2 \cap \cdots \cap E_n) \end{aligned}$$

In words, to find the probability of a union of n events, first sum their individual probabilities, then subtract the sum of the probabilities of all their pairwise intersections, then add back the sum of the probabilities of all their 3-way interections, then subtract the 4-way intersections, and continue adding and subtracting k-way intersections until you finally stop with the probability of the n-way intersection.

Random variables notation. In order to describe a sample space, we frequently introduce a symbol X called a *random variable* for the sample space. With this notation, we can replace the probability of an event, $P(E)$, by the notation $P(X \in E)$, which, by itself, doesn't do much. But many events are built from the set operations of complement, union, and intersection, and with the random variable notation, we can replace those by logical operations for 'not', 'or', and 'and'. For instance, the probability $P(E \cup F^c)$ can be written as $P(X \in E \text{ but } X \notin F)$.

Also, probabilities of finite events can be written in terms of equality. For instance, the probability of a singleton, $P(\{a\})$, can be written as $P(X{=}a)$, and that for a doubleton, $P(\{a, b\}) = P(X{=}a \text{ or } X{=}b)$.

One of the main purposes of the random variable notation is when we have two uses for the same

sample space. For instance, if you have a fair die, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$ where the probability of any singleton is $\frac{1}{6}$. If you have two fair dice, you can use two random variables, X and Y, to refer to the two dice, but each has the same sample space. (Soon, we'll look at the joint distribution of (X, Y), which has a sample space defined on $\Omega \times \Omega$.

Random variables and cumulative distribution functions. A sample space can have any set as its underlying set, but usually they're related to numbers. Often the sample space is the set of real numbers $\mathbf{R}$, and sometimes a power of the real numbers $\mathbf{R}^n$.

The most common sample space only has two elements, that is, there are only two outcomes. For instance, flipping a coin as two outcomes—Heads and Tails; many experiments have two outcomes—Success and Failure; and polls often have two outcomes—For and Against. Even though these events aren't numbers, it's useful to replace them by numbers, namely 0 and 1, so that Heads, Success, and For are identified with 1, and Tails, Failure, and Against are identified with 0. Then the sample space can have $\mathbf{R}$ as its underlying set.

When the sample space does have $\mathbf{R}$ as its underlying set, the random variable X is called a *real random variable*. With it, the probability of an interval like $[a, b]$, which is $P([a, b])$, can then be described as $P(a \le X \le b)$. Unions of intervals can also be described, for instance $P((-\infty, 3) \cup [4, 5])$ can be written as $P(X < 3 \text{ or } 4 \le X \le 5)$.

When the sample space is $\mathbf{R}$, the probability function P is determined by a cumulative distribution function (c.d.f.) F as follows. The function $F : \mathbf{R} \to \mathbf{R}$ is defined by

$$F(x) = P(X \le x) = P((-\infty, x]).$$

Then, from F, the probability of a half-open interval can be found as

$$P((a, b]) = F(b) - F(a).$$

Also, the probability of a singleton $\{b\}$ can be found as a limit

$$P(\{b\}) = \lim_{a \to b} (F(b) - F(a)).$$

From these, probabilities of unions of intervals can be computed. Sometimes, the c.d.f. is simply called the *distribution*, and the sample space is identified with this distribution.

Discrete distributions. Many sample distributions are determined entirely by the probabilities of their outcomes, that is, the probability of an event E is

$$P(E) = \sum_{x \in E} P(X{=}x) = \sum_{x \in E} P(\{x\}).$$

The sum here, of course, is either a finite or countably infinite sum. Such a distribution is called a *discrete distribution*, and when there are only finitely many outcomes x with nonzero probabilities, it is called a *finite distribution*.

A discrete distributions is usually described in terms of a probability mass function (p.m.f.) f defined by

$$f(x) = P(X{=}x) = P(\{x\}).$$

This p.m.f. is enough to determine this distribution since, by the definition of a discrete distribution, the probability of an event E is

$$P(E) = \sum_{x \in E} f(x).$$

In many applications, a finite distribution is *uniform*, that is, the probabilities of its outcomes are all the same, $1/n$, where n is the number of outcomes with nonzero probabilities. When that is the case, the field of combinatorics is useful in finding probabilities of events. Combinatorics includes various principles of counting such as the multiplication principle, permutations, and combinations.

Continuous distributions. When the cumulative distribution function F for a distribution is differentiable function, we say it's a *continuous distribution*. Such a distribution is determined by a probability density function f. The relation between F and f is that f is the derivative F' of F, and F is the integral of f.

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

Conditional probability and independence. If E and F are two events, with $P(F) \neq 0$, then the *conditional probability* of E given F is defined to be

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Two events, E and F, neither with probability 0, are said to be *independent*, or *mutually independent*, if any of the following three logically equivalent conditions holds

$$\begin{aligned} P(E \cap F) &= P(E)\,P(F) \\ P(E|F) &= P(E) \\ P(F|E) &= P(F) \end{aligned}$$

Bayes' formula. This formula is useful to invert conditional probabilities. It says

$$\begin{aligned} P(F|E) &= \frac{P(E|F)\,P(F)}{P(E)} \\ &= \frac{P(E|F)\,P(F)}{P(E|F)\,P(F) + P(E|F^{\mathsf{c}})\,P(F^{\mathsf{c}})} \end{aligned}$$

where the second form is often more useful in practice.

Expectation. The *expected value* $E(X)$, also called the *expectation* or *mean* μ_X, of a random variable X is defined differently for the discrete and continuous cases.

For a discrete random variable, it is a weighted average defined in terms of the probability mass function f as

$$E(X) = \mu_X = \sum_x x f(x).$$

For a continuous random variable, it is defined in terms of the probability density function f as

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

There is a physical interpretation where this mean is interpreted as a center of gravity.

Expectation is a linear operator. That means that the expectation of a sum or difference is the difference of the expectations

$$E(X + Y) = E(X) + E(Y),$$

and that's true whether or not X and Y are independent, and also

$$E(cX) = c\,E(X)$$

where c is any constant. From these two properties it follows that

$$E(X - Y) = E(X) - E(Y),$$

and, more generally, expectation preserves linear combinations

$$E\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i E(X_i).$$

Furthermore, when X and Y are independent, then $E(XY) = E(X)\,E(Y)$, but that equation doesn't usually hold when X and Y are not independent.

Variance and standard deviation. The *variance* of a random variable X is defined as

$$\mathrm{Var}(X) = \sigma_X^2 = E((X - \mu_X)^2) = E(X^2) - \mu_X^2$$

where the last equality is provable. Standard deviation, σ, is defined as the square root of the variance.

Here are a couple of properties of variance. First, if you multiply a random variable X by a constant c to get cX, the variance changes by a factor of the square of c, that is

$$\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X).$$

That's the main reason why we take the square root of variance to normalize it—the standard deviation of cX is c times the standard deviation of X. Also, variance is translation invariant, that is, if you add a constant to a random variable, the variance doesn't change:

$$\mathrm{Var}(X + c) = \mathrm{Var}(X).$$

In general, the variance of the sum of two random variables is *not* the sum of the variances of the two random variables. But it is when the two random variables are independent.

Moments, central moments, skewness, and kurtosis. The k^{th} *moment* of a random variable X is defined as $\mu_k = E(X^k)$. Thus, the mean is the first moment, $\mu = \mu_1$, and the variance can be found from the first and second moments, $\sigma^2 = \mu_2 - \mu_1^2$.

The k^{th} *central moment* is defined as $E((X-\mu)^k$. Thus, the variance is the second central moment.

A third central moment of the standardized random variable $X^* = \dfrac{X-\mu}{\sigma}$,

$$\beta_3 = E((X^*)^3) = \frac{E((X-\mu)^3)}{\sigma^3}$$

is called the *skewness* of X. A distribution that's symmetric about its mean has 0 skewness. (In fact all the odd central moments are 0 for a symmetric distribution.) But if it has a long tail to the right and a short one to the left, then it has a positive skewness, and a negative skewness in the opposite situation.

A fourth central moment of X^*,

$$\beta_4 = E((X^*)^4) = \frac{E((X-\mu)^4)}{\sigma^4}$$

is called *kurtosis*. A fairly flat distribution with long tails has a high kurtosis, while a short tailed distribution has a low kurtosis. A bimodal distribution has a very high kurtosis. A normal distribution has a kurtosis of 3. (The word kurtosis was made up in the early 19th century from the Greek word for curvature.)

The moment generating function. There is a clever way of organizing all the moments into one mathematical object, and that object is called the *moment generating function*. It's a function $m(t)$ of a new variable t defined by

$$m(t) = E(e^{tX}).$$

Since the exponential function e^t has the power series

$$e^t = \sum_{k=0}^{\infty} \frac{t^k}{k!} = 1 + t + \frac{t^2}{2!} + \cdots + \frac{t^k}{k!} + \cdots,$$

we can rewrite $m(t)$ as follows

$$m(t) = E(e^{tX})1 + \mu_1 t + \frac{\mu_2}{2!}\, t^2 + \cdots + \frac{\mu_k}{k!}\, t^k + \cdots.$$

That implies that $m^{(k)}(0)$, the k^{th} derivative of $m(t)$ evaluated at $t = 0$, equals the k^{th} moment μ_k of X. In other words, the moment generating function generates the moments of X by differentiation.

For discrete distributions, we can also compute the moment generating function directly in terms of the probability mass function $f(x) = P(X{=}x)$ as

$$m(t) = E(e^{tX}) = \sum_x e^{tx} f(x).$$

For continuous distributions, the moment generating function can be expressed in terms of the probability density function f_X as

$$m(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx.$$

The moment generating function enjoys the following properties.

Translation. If $Y = X + a$, then

$$m_Y(t) = e^{ta} m_X(t).$$

Scaling. If $Y = bx$, then

$$m_Y(t) = m_X(bt).$$

Standardizing. From the last two properties, if

$$X^* = \frac{X-\mu}{\sigma}$$

is the standardized random variable for X, then

$$m_{X^*}(t) = e^{-\mu t/\sigma} m_X(t/\sigma).$$

Convolution. If X and Y are independent variables, and $Z = X + Y$, then

$$m_Z(t) = m_X(t)\, m_Y(t).$$

The primary use of moment generating functions is to develop the theory of probability. For instance, the easiest way to prove the central limit theorem is to use moment generating functions.

The median, quartiles, quantiles, and percentiles. The *median* of a distribution X, sometimes denoted $\tilde{\mu}$, is the value such that $P(X \le \tilde{\mu}) = \frac{1}{2}$. Whereas some distributions, like the Cauchy distribution, don't have means, all continuous distributions have medians.

If p is a number between 0 and 1, then the p^{th} *quantile* is defined to be the number θ_p such that

$$P(X \le \theta_p) = F(\theta_p) = p.$$

Quantiles are often expressed as percentiles where the p^{th} quantile is also called the $100p^{\text{th}}$ *percentile*. Thus, the median is the 0.5 quantile, also called the 50^{th} percentile.

The *first quartile* is another name for $\theta_{0.25}$, the 25^{th} percentile, while the *third quartile* is another name for $\theta_{0.75}$, the 75^{th} percentile

Joint distributions. When studing two related real random variables X and Y, it is not enough just to know the distributions of each. Rather, the pair (X, Y) has a joint distribution. You can think of (X, Y) as naming a single random variable that takes values in the plane $\mathbf{R}^2$.

Joint and marginal probability mass functions. Let's consider the discrete case first where both X and Y are discrete random variables. The probability mass function for X is $f_X(x) = P(X{=}x)$, and the p.m.f. for Y is $f_Y(y) = P(Y{=}y)$. The joint random variable (X, Y) has its own p.m.f. denoted $f_{(X,Y)}(x, y)$, or more briefly $f(x, y)$:

$$f(x, y) = P((X, Y){=}(x, y)) = P(X{=}x \text{ and } Y{=}y),$$

and it determines the two individual p.m.f.s by

$$f_X(x) = \sum_y f(x, y), \qquad f_Y(y) = \sum_x f(x, y).$$

The individual p.m.f.s are usually called *marginal probability mass functions*.

For example, assume that the random variables X and Y have the joint probability mass function given in this table.

		Y			
		-1	0	1	2
X	-1	0	1/36	1/6	1/12
	0	1/18	0	1/18	0
	1	0	1/36	1/6	1/12
	2	1/12	0	1/12	1/6

By adding the entries row by row, we find the the marginal function for X, and by adding the entries column by column, we find the marginal function for Y. We can write these marginal functions on the margins of the table.

		Y				f_X
		-1	0	1	2	
X	-1	0	1/36	1/6	1/12	5/18
	0	1/18	0	1/18	0	1/9
	1	0	1/36	1/6	1/12	5/18
	2	1/12	0	1/12	1/6	1/3
f_Y		5/36	1/18	17/36	1/3	

Discrete random variables X and Y are independent if and only if the joint p.m.f is the product of the marginal p.m.f.s

$$f(x, y) = f_X(x) f_Y(y).$$

In the example above, X and Y aren't independent.

Joint and marginal cumulative distribution functions. Besides the p.m.f.s, there are joint and marginal cumulative distribution functions. The c.d.f. for X is $F_X(x) = P(X \leq x)$, while the c.d.f. for Y is $F_Y(y) = P(Y \leq y)$. The joint random variable (X, Y) has its own c.d.f. denoted $F_{(X,Y)}(x, y)$, or more briefly $F(x, y)$:

$$F(x, y) = P(X \leq x \text{ and } Y \leq y),$$

and it determines the two marginal p.m.f.s by

$$F_X(x) = \lim_{y \to \infty} F(x, y), \qquad F_Y(y) = \lim_{x \to \infty} F(x, y).$$

Joint and marginal probability density functions. Now let's consider the continuous case where X and Y are both continuous. The last paragraph on c.d.f.s still applies, but we'll have marginal probability density functions $f_X(x)$ and $f_Y(y)$, and a joint probability density function $f(x, y)$ instead of probability mass functions. Of course, the derivatives of the marginal c.d.f.s are the density functions

$$f_X(x) = \frac{d}{dx} F_X(x) \qquad f_Y(y) = \frac{d}{dy} F_Y(y)$$

and the c.d.f.s can be found by integrating the density functions

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt \qquad F_Y(y) = \int_{-\infty}^{y} f_Y(t)\, dt.$$

The joint probability density function $f(x, y)$ is found by taking the derivative of F twice, once with respect to each variable, so that

$$f(x, y) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} F(x, y).$$

(The notation ∂ is substituted for d to indicate that there are other variables in the expression that are held constant while the derivative is taken with respect to the given variable.) The joint cumulative distribution function can be recovered from the joint density function by integrating twice

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\, dt\, ds.$$

Furthermore, the marginal density functions can be found by integrating joint density function.

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \quad f_Y(x) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

Continuous random variables X and Y are independent if and only if the joint density function is the product of the marginal density functions

$$f(x, y) = f_X(x) f_Y(y).$$

Covariance and correlation. The *covariance* of two random variables X and Y is defined as

$$\text{Cov}(X, Y) = \sigma_{XY} = E((X - \mu_X)(Y - \mu_Y)).$$

It can be shown that

$$\text{Cov}(X, Y) = E(XY) - \mu_X \mu_Y.$$

When X and Y are independent, then $\sigma_{XY} = 0$, but in any case

$$\text{Var}(X + Y) = \text{Var}(X) + 2\ \text{Cov}(X, Y) + \text{Var}(Y).$$

Covariance is a bilinear operator, which means it is linear in each coordinate

$$\begin{aligned} \text{Cov}(X_1 + X_2, Y) &= \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y) \\ \text{Cov}(aX, Y) &= a\ \text{Cov}(X, Y) \\ \text{Cov}(X, Y_1 + Y_2) &= \text{Cov}(X, Y_1) + \text{Cov}(X, Y_2) \\ \text{Cov}(X, bY) &= b\ \text{Cov}(X, Y) \end{aligned}$$

The *correlation*, or *correlation coefficient*, of X and Y is defined as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Correlation is always a number between -1 and 1.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`