Probability Definitions

Probability Space (or **Sample Space**): ordered triple $(\Omega, \mathcal{F}, P)$.

- Ω is a set (of **elementary** outcomes).

- $\mathcal{F}$ is a family of subsets (**events**) of Ω which is a σ-field (or Borel field or σ-algebra):

  1. Empty set $\emptyset$ and Ω are members of $\mathcal{F}$.

  2. $A \in \mathcal{F}$ implies $A^c = \{\omega \in \Omega : \omega \notin A\} \in \mathcal{F}$

  3. $A_1, A_2, \cdots$ all in $\mathcal{F}$ implies
  $$A = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \,.$$

- P a function, domain $\mathcal{F}$, range a subset of $[0,1]$ satisfying:

  1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.

  2. **Countable additivity**: $A_1, A_2, \cdots$ **pairwise disjoint** $(j \neq k \implies A_j A_k = \emptyset)$

  $$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Axioms guarantee can compute probabilities by usual rules, including approximation, without contradiction.

Consequences:

1. **Finite additivity** $A_1, \cdots, A_n$ pairwise disjoint:

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)\,.$$

2. For any event A $P(A^c) = 1 - P(A)$.

3. If $A_1 \subset A_2 \subset \cdots$ are events then

$$P(\bigcup_1^\infty A_i) = \lim_{n\to\infty} P(A_n)\,.$$

4. If $A_1 \supset A_2 \supset \cdots$ then

$$P(\bigcap_1^\infty A_i) = \lim_{n\to\infty} P(A_n)\,.$$

Most subtle point is σ-field, $\mathcal{F}$. Needed to avoid some contradictions which arise if you try to define $P(A)$ for every subset A of Ω when Ω is a set with uncountably many elements.

Random Variables:

Vector valued random variable: function X, domain Ω, range in $\mathbb{R}^p$ such that

$$P(X_1 \le x_1, \ldots, X_p \le x_p)$$

is defined for any constants $(x_1, \ldots, x_p)$. Notation: $X = (X_1, \ldots, X_p)$ and

$$X_1 \le x_1, \ldots, X_p \le x_p$$

is shorthand for an event:

$$\{\omega \in \Omega : X_1(\omega) \le x_1, \ldots, X_p(\omega) \le x_p\}$$

X function on Ω so X_1 function on Ω.

For this course I assume you know:

Definitions and uses of *joint*, *marginal* and *conditional* **densities** and **probability mass functions** or **discrete densities**.

Definitions and uses of *joint* and *marginal* distribution functions.

How to go back and forth between distributions and densities.

Change of variables formula.

Independence

Events A and B **independent** if

$$P(AB) = P(A)P(B)\,.$$

Events A_i, $i = 1, \ldots, p$ are **independent** if

$$P(A_{i_1} \cdots A_{i_r}) = \prod_{j=1}^{r} P(A_{i_j})$$

for any set of distinct indices $i_1, \ldots, i_r$ between 1 and p.

Example: $p = 3$

$$\begin{aligned} P(A_1A_2A_3) &= P(A_1)P(A_2)P(A_3) \\ P(A_1A_2) &= P(A_1)P(A_2) \\ P(A_1A_3) &= P(A_1)P(A_3) \\ P(A_2A_3) &= P(A_2)P(A_3) \end{aligned}$$

Need all equations to be true for independence!

Example: Toss a coin twice. If A_1 is the event that the first toss is a Head, A_2 is the event that the second toss is a Head and A_3 is the event that the first toss and the second toss are different. then $P(A_i) = 1/2$ for each i and for $i \neq j$

$$P(A_i \cap A_j) = \frac{1}{4}$$

but

$$P(A_1 \cap A_2 \cap A_3) = 0 \neq P(A_1)P(A_2)P(A_3)\,.$$

Def'n: Rvs $X_1, \ldots, X_p$ are **independent** if

$$P(X_1 \in A_1, \cdots, X_p \in A_p) = \prod P(X_i \in A_i)$$

for any choice of $A_1, \ldots, A_p$.

Theorem 1 *1. If X and Y are independent and discrete then*

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

for all x, y

2. If X and Y are discrete and

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

for **all** *x, y then X and Y are independent.*

Theorem 2 *If $X_1, \ldots, X_p$ are independent and $Y_i = g_i(X_i)$ then $Y_1, \ldots, Y_p$ are independent. Moreover, $(X_1, \ldots, X_q)$ and $(X_{q+1}, \ldots, X_p)$ are independent.*

Conditional probability

Important modeling and computation technique:

Def'n: $P(A|B) = P(AB)/P(B)$ if $P(B) \neq 0$.

Def'n: For discrete rvs X, Y conditional pmf of Y given X is

$$\begin{aligned} f_{Y|X}(y|x) &= P(Y = y|X = x) \\ &= f_{X,Y}(x,y)/f_X(x) \\ &= f_{X,Y}(x,y)/\sum_t f_{X,Y}(x,t) \end{aligned}$$

IDEA: used as both computational tool and modelling tactic.

Specify joint distribution by specifying "marginal" and "conditional".

Modelling:

Assume $X \sim$ Poisson(λ).

Assume $Y|X \sim$Binomial(X, p).

Let $Z = X - Y$. Joint law of Y, Z?

$$
\begin{aligned}
&P(Y = y, Z = z) \\
&\quad = P(Y = y, X - Y = z) \\
&\quad = P(Y = y, X = z + y) \\
&\quad = P(Y = y|X = y + z)P(X = y + z) \\
&\quad = \binom{z+y}{y} p^y (1-p)^z e^{-\lambda} \lambda^{z+y}/(z+y)! \\
&\quad = \exp\{-p\lambda\}\frac{(p\lambda)^y}{y!} \exp\{(1-p)\lambda\}\frac{\{(1-p)\lambda\}^z}{z!}
\end{aligned}
$$

So: Y, Z independent Poissons.

Expected Value

Undergraduate definition of E: integral for absolutely continuous X, sum for discrete. But: $\exists$ rvs which are neither absolutely continuous nor discrete.

General definition of E.

A random variable X is **simple** if we can write

$$X(\omega) = \sum_1^n a_i 1(\omega \in A_i)$$

for some constants $a_1, \ldots, a_n$ and events A_i.

Def'n: For a simple rv X we define

$$E(X) = \sum a_i P(A_i)$$

For positive random variables which are not simple we extend our definition by approximation:

Def'n: If $X \geq 0$ (almost surely, $P(X \geq 0) = 1$) then

$$E(X) = \sup\{E(Y) : 0 \leq Y \leq X, Y \text{ simple}\}$$

Def'n: We call X **integrable** if

$$E(|X|) < \infty \,.$$

In this case we define

$$E(X) = E(\max(X, 0)) - E(\max(-X, 0))$$

Facts: E is a linear, monotone, positive operator:

1. **Linear**: $E(aX+bY) = aE(X)+bE(Y)$ provided X and Y are integrable.

2. **Positive**: $P(X \geq 0) = 1$ implies $E(X) \geq 0$.

3. **Monotone**: $P(X \geq Y) = 1$ and X, Y integrable implies $E(X) \geq E(Y)$.

Major technical theorems:

Monotone Convergence: If $0 \le X_1 \le X_2 \le \cdots$ a.s. and $X = \lim X_n$ (which exists a.s.) then

$$E(X) = \lim_{n\to\infty} E(X_n)$$

Dominated Convergence: If $|X_n| \le Y_n$ and $\exists$ rv X st $X_n \to X$ a.s. and rv Y st $Y_n \to Y$ with $E(Y_n) \to E(Y) < \infty$ then

$$E(X_n) \to E(X)$$

Often used with all Y_n the same rv Y.

Fatou's Lemma: If $X_n \ge 0$ then

$$E(\liminf X_n) \le \liminf E(X_n)$$

Theorem: With this definition of E if X has density $f(x)$ (even in $\mathbb{R}^p$ say) and $Y = g(X)$ then

$$E(Y) = \int g(x)f(x)dx\,.$$

(This could be a multiple integral.)

Works even if X has density but Y doesn't.

If X has pmf f then

$$E(Y) = \sum_x g(x)f(x)\,.$$

Def'n: r^{th} moment (about origin) of a real rv X is $\mu'_r = E(X^r)$ (provided it exists). Generally use μ for $E(X)$. The r^{th} central moment is

$$\mu_r = E[(X-\mu)^r]$$

Call $\sigma^2 = \mu_2$ the variance.

Def'n: For an $\mathbb{R}^p$ valued rv X $\mu_X = E(X)$ is the vector whose i^{th} entry is $E(X_i)$ (provided all entries exist).

Def'n: The $(p \times p)$ variance covariance matrix of X is

$$Var(X) = E\left[(X-\mu)(X-\mu)^t\right]$$

which exists provided each component X_i has a finite second moment. More generally if $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ both have all components with finite second moments then

$$\text{Cov}(X,Y) = \text{E}\left[(X-\mu_X)(Y-\mu_Y)^T\right]$$

We have

$$\text{Cov}(AX+a, BY+b) = A\text{Cov}(X,Y)B^T$$

for general (conforming) matrices A, B and vectors a and b.

Moments and probabilities of rare events are closely connected.

Markov's inequality ($r = 2$ is Chebyshev's inequality):

$$\begin{aligned} P(|X-\mu| \ge t) &= E[1(|X-\mu| \ge t)] \\ &\le E\left[\frac{|X-\mu|^r}{t^r} 1(|X-\mu| \ge t)\right] \\ &\le \frac{E[|X-\mu|^r]}{t^r} \end{aligned}$$

Intuition: if moments are small then large deviations from average are unlikely.

Moments and independence

Theorem: If $X_1, \ldots, X_p$ are independent and each X_i is integrable then $X = X_1 \cdots X_p$ is integrable and

$$E(X_1 \cdots X_p) = E(X_1) \cdots E(X_p)$$

Multiple Integration: Lebesgue integrals over $\mathbb{R}^p$ defined using Lebesgue measure on $\mathbb{R}^p$.

Iterated integrals wrt Lebesgue measure on $\mathbb{R}^1$ give same answer.

Theorem[Tonelli]: If $f : \mathbb{R}^{p+q} \mapsto \mathbb{R}$ is Borel and $f \geq 0$ almost everywhere then for almost every $x \in \mathbb{R}^p$ the integral

$$g(x) \equiv \int f(x,y)dy$$

exists and

$$\int g(x)dx = \int f(x,y)dxdy$$

RHS denotes $p+q$ dimensional integral defined previously.

Theorem[Fubini] If $f : \mathbb{R}^{p+q} \mapsto \mathbb{R}$ is Borel and integrable then for almost every $x \in \mathbb{R}^p$ the integral

$$g(x) \equiv \int f(x, y) dy$$

exists and is finite. Moreover g is integrable and

$$\int g(x) dx = \int f(x, y) dx dy \, .$$

Results true for measures other than Lebesgue.

Conditional distributions, expectations

When X and Y are discrete we have

$$\mathsf{E}(Y|X=x) = \sum_y yP(Y=y|X=x)$$

for any x for which $P(X=x)$ is positive.

Defines a function of x.

This function evaluated at X gives rv which is ftn of X denoted

$$\mathsf{E}(Y|X).$$

Example: $Y|X = x \sim$ Binomial(x,p). Since mean of a Binomial(n,p) is np we find

$$\mathsf{E}(Y|X=x) = px$$

and

$$\mathsf{E}(Y|X) = pX$$

Notice you simply replace x by X.

Here are some properties of the function

$$\mathsf{E}(Y|X=x)$$

1) Suppose A is a function defined on the range of X. Then

$$\mathsf{E}(A(X)Y|X=x) = A(x)\mathsf{E}(Y|X=x)$$

and so

$$\mathsf{E}(A(X)Y|X) = A(X)\mathsf{E}(Y|X)$$

2) Repeated conditioning: if X, Y and Z discrete then

$$\mathsf{E}\{\mathsf{E}(Z|X,Y)|X\} = \mathsf{E}(Z|X)$$
$$\mathsf{E}\{\mathsf{E}(Y|X)\} = \mathsf{E}(Y)$$

3) Additivity

$$\mathsf{E}(Y + Z|X) = \mathsf{E}(Y|X) + \mathsf{E}(Z|X)$$

4) Putting the first two items together gives

$$\mathsf{E}\left\{\mathsf{E}(A(X)Y|X)\right\} = \tag{1}$$
$$\mathsf{E}\left\{A(X)\mathsf{E}(Y|X)\right\} = \mathsf{E}(A(X)Y)$$

Definition of $\mathsf{E}(Y|X)$ when X and Y are not assumed to discrete:

$\mathsf{E}(Y|X)$ is rv which is measurable function of X satisfying (1).

Existence is measure theory problem.

Properties: all 4 properties still hold.

Theorem 3 *If X and Y have joint density and $f(y|x)$ is conditional density then*

$$E\Big\{g(Y)|X = x\Big\} = \int g(y)f(y|x)dy$$

provided $E(g(Y)) < \infty$.

Theorem 4 *If X is rv and $X^* = g(X)$ is a one to one transformation of X then*

$$\mathsf{E}(Y|X = x) = \mathsf{E}(Y|X^* = g(x))$$

and

$$\mathsf{E}(Y|X) = \mathsf{E}(Y|X^*)$$

Interpretation.

Formula is "obvious".

Example: Toss coin $n = 20$ times. Y is indicator of first toss is a heads. X is number of heads and X^* number of tails. Formula says:

$$\mathsf{E}(Y|X = 17) = \mathsf{E}(Y|X^* = 3)$$

In fact for a general k and n

$$\mathsf{E}(Y|X = k) = \frac{k}{n}$$

so

$$\mathsf{E}(Y|X) = \frac{X}{n}$$

At the same time

$$\mathsf{E}(Y|X^* = j) = \frac{n-j}{n}$$

so

$$\mathsf{E}(Y|X^*) = \frac{n - X^*}{n}$$

But of course $X = n - X^*$ so these are just two ways of describing the same random variable.