

Probability Basics

Richard Lockhart

Simon Fraser University

STAT 870 — Summer 2011



Purposes of Today's Lecture

- Run through basic definitions of probability theory
- Define Probability space, random variables.
- Define expected value, moments.
- Present basic convergence theorems.
- Discuss conditional expectation.



Probability Definitions

Probability Space (or **Sample Space**): ordered triple (Ω, \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, \dots all in \mathcal{F} implies

$$A = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$



Probability Measure Defined

- 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, \dots **pairwise disjoint**
($j \neq k \implies A_j A_k = \emptyset$)

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

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



Consequences

- 1 **Finite additivity** A_1, \dots, A_n pairwise disjoint:

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

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

- 3 If $A_1 \subset A_2 \subset \dots$ are events then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \rightarrow \infty} P(A_n).$$

- 4 If $A_1 \supset A_2 \supset \dots$ then

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n \rightarrow \infty} P(A_n).$$



Consequences

- 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.
- Classic example uses uniform distribution and axiom of choice.



Random Variables

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

$$P(X_1 \leq x_1, \dots, X_p \leq x_p)$$

is defined for any constants (x_1, \dots, x_p) .

- Notation: $X = (X_1, \dots, X_p)$ and

$$X_1 \leq x_1, \dots, X_p \leq x_p$$

is shorthand for an event:

$$\{\omega \in \Omega : X_1(\omega) \leq x_1, \dots, X_p(\omega) \leq x_p\}$$

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.



Densities

- If X takes values in \mathbb{R}^p then X has density f if and only if

$$P(X \in A) = \int_A f(x) dx.$$

We say X has an *absolutely continuous* distribution.

- If there is a countable set $C = \{x_1, x_2, \dots\}$ such that

$$P(X \in C) = 1$$

then we say X has a *discrete* distribution.

- In this case we define the discrete density of X by

$$f(x) = P(X = x).$$



Independence

- Events A and B **independent** if

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

- Events A_i , $i = 1, \dots, p$ 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, \dots, i_r between 1 and p .

- Example: $p = 3$

$$P(A_1 A_2 A_3) = P(A_1)P(A_2)P(A_3)$$

$$P(A_1 A_2) = P(A_1)P(A_2)$$

$$P(A_1 A_3) = P(A_1)P(A_3)$$

$$P(A_2 A_3) = P(A_2)P(A_3)$$

Need all equations to be true for independence!



Example

- Toss a coin twice.
- A_1 is the event that the first toss is a Head
- A_2 is the event that the second toss is a Head
- 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

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



Independence extended

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

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

for any choice of A_1, \dots, A_p .

Def'n: We say σ -fields $\mathcal{F}_1, \dots, \mathcal{F}_p$ are independent if and only if

$$P(A_1 \cdots A_p) = P(A_1) \cdots P(A_p)$$

for all $A_i \in \mathcal{F}_i$.

Def'n: These definitions extend to infinite collections of events and σ -fields by requiring them to hold for each finite sub-collection.



Conditions for independence of rvs

Theorem

- 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

If X_1, \dots, X_p are independent and $Y_i = g_i(X_i)$ then Y_1, \dots, Y_p are independent. Moreover, (X_1, \dots, X_q) and (X_{q+1}, \dots, 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 $X \sim \text{Poisson}(\lambda)$.
- Assume $Y|X \sim \text{Binomial}(X, p)$.
- Let $Z = X - Y$.
- Joint law of Y, Z ?

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

- So: Y, Z independent Poissons.



Expected Value – simple rvs

- 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, \dots, a_n and events A_i .

- **Def'n**: For a simple rv X we define

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



Expected value – non-negative rvs

- 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

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



Properties of E

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 $E(X) \geq E(Y)$.

Jargon: If $P(A) = 1$ we say A happens almost surely. Almost everywhere is the corresponding concept for Lebesgue measure. A measure ν is like a probability but $\nu(\Omega)$ might not be 1.



Major technical theorems

- **Monotone Convergence:** If $0 \leq X_1 \leq X_2 \leq \dots$ a.s. and $X = \lim X_n$ (which exists a.s.) then

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

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

$$E(X_n) \rightarrow E(X)$$

Often used with all Y_n the same rv Y .

- **Fatou's Lemma:** If $X_n \geq 0$ then

$$E(\liminf X_n) \leq \liminf E(X_n)$$



Conditions for independence of rvs

Theorem

- ① *If X and Y are independent and (X, Y) has density $f(x, y)$ then X has a density, say g and Y has a density, say h and for all x, y*

$$f(x, y) = g(x)h(y)$$

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

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

- ② *If X and Y are independent and have densities g and h respectively then (X, Y) has a density $f(x, y) = g(x)h(y)$.*
- ③ *If there are functions $g(x)$ and $h(y)$ which have the property that $f(x, y) = g(x)h(y)$ is a density of (X, Y) then X and Y are independent and both X and Y have densities given by multiples of g and h respectively.*

Relation to undergraduate definitions

Theorem

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.

Theorem

If X has pmf f then

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



Moments

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).



Variance-covariance matrices

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

$$\text{Var}(X) = E \left[(X - \mu)(X - \mu)^T \right]$$

- this 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

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

- 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 .



Inequalities

- Moments and probabilities of rare events are closely connected.
- Markov's inequality ($r = 2$ is Chebyshev's inequality):

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

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



Moments and independence

Theorem

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

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



Iterated integrals – Tonelli's Theorem

- **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

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

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



Fubini's Theorem

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

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

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

$$E(Y|X).$$

- $Y|X = x \sim \text{Binomial}(x, p)$. Since mean of a Binomial(n, p) is np we find

$$E(Y|X = x) = px$$

and

$$E(Y|X) = pX$$

Notice you simply replace x by X .



Properties of conditional expectation

Here are some properties of the function

$$E(Y|X = x)$$

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

$$E(A(X)Y|X = x) = A(x)E(Y|X = x)$$

and so

$$E(A(X)Y|X) = A(X)E(Y|X)$$

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

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

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



Properties of conditional expectation

3 Additivity

$$E(Y + Z|X) = E(Y|X) + E(Z|X)$$

4 Putting the first two items together gives

$$\begin{aligned} E\{E(A(X)Y|X)\} &= \\ E\{A(X)E(Y|X)\} &= E(A(X)Y) \end{aligned} \tag{1}$$



General conditional expectations

- Definition of $E(Y|X)$ when X and Y are not assumed to be discrete:
- $E(Y|X)$ is a random variable which is a measurable function of X satisfying (1).
- Existence is a measure theory problem.
- Properties: all 4 properties still hold.



Relation to undergraduate ideas

Theorem

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

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

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

Theorem

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

$$E(Y|X = x) = E(Y|X^* = g(x))$$

and

$$E(Y|X) = E(Y|X^*)$$



Interpretation

- Formula is “obvious”.
- 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:

$$E(Y|X = 17) = E(Y|X^* = 3)$$



Interpretation

- In fact for a general k and n

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

so

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

- At the same time

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

so

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

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

