Given Ω is chosen, a probability law on Ω is a mapping $\mathbb{P}$ that assigns a number to every event (i.e. to every subset of Ω) such that:

1) $\mathbb{P}(A) \geq 0$ for every event A
2) $\mathbb{P}(\Omega) = 1$ (normalization)
3) Additivity Rules
   i) If $A \cap B = \emptyset$ (A, B are events) then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
   ii) If $A_1, A_2, A_3, \ldots$ is a countable sequence of mutually disjoint events (i.e. $A_i \cap A_j = \emptyset \ \forall i \neq j$), then $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$

$\bigcup_{n=1}^{\infty} A_n$ ← union of all of them

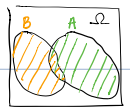
So, given an event $A \subset \Omega$, $\mathbb{P}(A)$ is a model for "the likelihood that the outcome of the uncertain experiment is in A"

"Event A occurs" means "outcome of experiment is in A"

Given: Ω and $\mathbb{P}$ and two events $A, B \subset \Omega$, define
$\mathbb{P}(A|B)$ = "Probability of A given B"

Idea: Given event B occurs, what's the likelihood that A occurs?

Knowledge that B occurs is gonna effect the likelihood of A's occurrence.



Intuitively: $A|B$ is "fraction of B's $\mathbb{P}$-mass that also lies in A"

Motivates the definition

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\text{⧅})}{\mathbb{P}(/\!/\!/)}$$

Undefined for $\mathbb{P}(B) = 0$

Observation: Given $B \subset \Omega$ w/ $\mathbb{P}(B) > 0$, as A runs over all events, $\mathbb{P}(A|B)$ defines a new probability law on Ω

- $\mathbb{P}(A|B) \geq 0 \ \forall A \subset \Omega$
- $\mathbb{P}(\Omega) = 1$ (by 2nd reality check above)
- If $A_1 \cap A_2 = \emptyset$, then (disjoint)

$$(A_1 \cup A_2) \cap B = (A_1 \cap B) \cup (A_2 \cap B)$$

divide both sides by $\mathbb{P}(B)$! $\Rightarrow \mathbb{P}((A_1 \cup A_2) \cap B) = \mathbb{P}(A_1 \cap B) + \mathbb{P}(A_2 \cap B)$

$$\mathbb{P}(A_1 \cup A_2 | B) = \mathbb{P}(A_1|B) + \mathbb{P}(A_2|B)$$

P_X is defined as follows: for every possible value of $x \in X$,

$$P_X(x) = \mathbb{P}(A_x) \text{ where } A_x = \{s \in \Omega : X(s) = x\}$$

i.e. $P_X(x)$ = the probability that the r.v. X takes on the specific value x

Book uses $\mathbb{P}(\{X = x\})$ or $\mathbb{P}(X = x)$ ← misuse of notation compared to above to refer to $\mathbb{P}(A_x)$, where $A_x = \{s \in \Omega : X(s) = x\}$

Things to note about P_X:
- $P_X(x) \geq 0$ for all possible values of X (why? cause for any x, $P_X(x)$ is $\mathbb{P}$(an event) ≥ 0!)
- If V is any finite or countably infinite set of possible values of X, then if we set
  B = the event "$X \in V$"
  i.e. $B = \{s \in \Omega : X(s) \in V\}$
  then
  $$\mathbb{P}(B) = \sum_{x \in V} P_X(x)$$

21:44

Product Rule: If $D_1 \supset D_2 \supset D_3 \supset \cdots \supset D_n$ is a nested decreasing sequence of positive-probability events, then

$$\mathbb{P}(D_n) = \mathbb{P}(D_1)\,\mathbb{P}(D_2|D_1)\,\mathbb{P}(D_3|D_2) \cdots \mathbb{P}(D_n|D_{n-1})$$

If $A_1, A_2, \ldots, A_n$ is a partition of Ω, then $C_1, C_2, \ldots, C_n$ partitions B, where

$$C_k = B \cap A_k \quad 1 \leq k \leq n$$

Idea: $B = B \cap \Omega = B \cap (A_1 \cup \cdots \cup A_n)$
$= (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$

By additivity,

$$\mathbb{P}(B) = \mathbb{P}(C_1) + \cdots + \mathbb{P}(C_n)$$
$$\mathbb{P}(B) = \mathbb{P}(B \cap A_1) + \cdots + \mathbb{P}(B \cap A_n)$$

LAW OF TOTAL PROBABILITY

$$\boxed{\mathbb{P}(B) = \mathbb{P}(B|A_1)\,\mathbb{P}(A_1) + \mathbb{P}(B|A_2)\,\mathbb{P}(A_2) + \cdots + \mathbb{P}(B|A_n)\,\mathbb{P}(A_n)}$$

Given $\Omega, \mathbb{P}$, if $A_1, \ldots, A_n$ are events that partition Ω and have nonzero probability, then for any event B, and for any k: $1 \leq k \leq n$,

$$\mathbb{P}(A_k|B) = \frac{\mathbb{P}(B|A_k)\,\mathbb{P}(A_k)}{\mathbb{P}(B|A_1)\,\mathbb{P}(A_1) + \cdots + \mathbb{P}(B|A_n)\,\mathbb{P}(A_n)}$$

"PROOF:" Numerator = $\mathbb{P}(B \cap A_k)$ by definition of conditional probability $\frac{\mathbb{P}(A_k \cap B)}{\mathbb{P}(B)} = \mathbb{P}(A_k|B)$!

Denominator = $\mathbb{P}(B)$ by total probability theorem

Next Circle of Ideas: Independence

Given $\Omega, \mathbb{P}$. Say two events $A, B \subset \Omega$ are independent when $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

Same as $\mathbb{P}(A|B) = \mathbb{P}(A)$ when $\mathbb{P}(B) > 0$ ← Knowing B occurred says nothing more about whether A occurred
$\mathbb{P}(B|A) = \mathbb{P}(B)$ when $\mathbb{P}(A) > 0$
Knowing that A occurred says nothing more about whether B occurred.

CAUTION: Independence is "about" not only the events and how they sit in Ω, but also (more crucially) about $\mathbb{P}$!!

Common error: if "$A \cap B = \emptyset$, then A and B are independent"
WRONG (whenever $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$)

Conditional Independence: $\Omega, \mathbb{P}$; say events A and B are conditionally independent given (event) C when

$$\mathbb{P}(A \cap B | C) = \mathbb{P}(A|C)\,\mathbb{P}(B|C)$$

$\mathbb{P}(A|B \cap C) = \mathbb{P}(A|C)$; $\mathbb{P}(B|A \cap C) = \mathbb{P}(B|C)$

knowledge of B gives no functional info about probability of A on top of knowledge of C

To see this just play w/ formulas

$$\mathbb{P}(A|B \cap C) = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)} = \frac{\mathbb{P}(A \cap B|C)\,\mathbb{P}(B|C)}{\mathbb{P}(B|C)} = \mathbb{P}(A|C)$$

Given a discrete r.v. X w/ $P_X(x)$ pmf, define the expected value (or expectation)

$$\mathbb{E}(X) = \sum_{x \in X} x\,P_X(x)$$

Binomial RV

Given positive integer n for some $p \in [0,1]$, the Binomial (n,p) pmf defined as follows:

$$P_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & 0 \leq k \leq n \\ 0, & \text{all other } k \end{cases}$$

This could arise from
Ω = {all H-T sequences of length n}
$\mathbb{P}$(any sequence) = $p^{\#H}(1-p)^{n-\#H}$
like the outcome of n-independent flips of a possibly unfair coin.
X(any sequence) = #(Heads) in sequence

Geometric Random Variable

Geometric(p):

$$P_X(k) = p(1-p)^{k-1}, \quad 1 \leq k < \infty$$

A possible $\Omega, \mathbb{P}$? You have a coin where $\mathbb{P}(H) = p$;
Ω = set of all semi-infinite $(1^{st}, 2^{nd}, \ldots)$ of H,T; $X(\omega)$ = index of first head for all $\omega \in \Omega$

Reality check: recall $\sum_{x \in X} P_X(x) = 1$ for any discrete r.v. X.

Verify this for geometric r.v.s:

$$\sum_{x \in X} P_X(x) = \sum_{k=1}^{\infty} p(1-p)^{k-1} = p \sum_{n=0}^{\infty} (1-p)^n = p \frac{1}{1-(1-p)} = 1 \ \checkmark$$

Counting Principle:

Re-count # of subsets of $\Omega = \{s_1, \ldots, s_n\}$.
We can view each subset as arising from a multi stage building process.

Stage 1: Choose whether to put s_1 in the subset
- either yes or no ($n_1 = 2$)

Stage 2: Choose whether to put s_2 in the subset
- either yes or no ($n_2 = 2$)

⋮

Stage n: Choose whether to put s_n in the subset
- either yes or no ($n_n = 2$)

$\#(\text{Subsets}) = n_1 n_2 \cdots n_n = 2^n$

"n choose k"

$$\binom{n}{k} := \frac{n(n-1)\cdots(n-k+1)}{k!}$$

Comment: Amazing that this is an integer!
Can also write
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

This comes up in situations involving independent trials

Idea: Perform a random experiment repeatedly + independently

If experiment has 2 outcomes, call them Bernoulli Trials. Generally visualize coin flips in this case - H,T

possible outcomes at each stage - if $\mathbb{P}(\{H\}) = p$, $\mathbb{P}(\{T\}) = 1-p$ and you perform experiment n times, for each k, $0 \leq k \leq n$,

$\mathbb{P}$(sequence you get) = $p^k(1-p)^{n-k}$, where $k = \#(H)$ you flip

This is true REGARDLESS of the order each H,T appears in the sequence of flips

$\mathbb{P}$(you get k H's) = #(ways of picking k spots in a available spots) $\cdot$

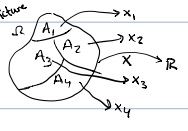
$$= \binom{n}{k} p^k (1-p)^{n-k} \quad 0 \leq k \leq n$$

Next BIG TOPIC: DISCRETE RANDOM VARIABLES

Start with Ω and $\mathbb{P}$: a discrete random variable (r.v.) is a real valued function with domain Ω that takes on only finite or countably infinite number of different values.

i.e. $X: \Omega \to \mathbb{R}$ "X is a mapping from Ω to reals"

Picture



i.e. $X(s) = x_i \ \forall s \in A_i$, etc

Next, multiple discrete rvs and joint pmfs, etc
Given $\Omega, \mathbb{P}$ and two discrete rvs X, Y defined in Ω, define the joint pmf of X and Y via

$$P_{X,Y}(x,y) = \mathbb{P}(\underbrace{\{X = x\}}_{A_x} \cap \underbrace{\{Y = y\}}_{B_y})$$

$\forall x \in X, y \in Y$.

Note: for any set V of possible value pairs for X, Y, we have

$$\sum_{(x,y) \in V} P_{X,Y}(x,y) = \mathbb{P}(\text{event that } (X,Y) \in V)$$

Since X, Y are discrete rvs, they have their own pmfs P_X, P_Y. These are determined as follows from the joint pmf $P_{X,Y}(x,y)$:

① $P_X(x) = \sum_y P_{X,Y}(x,y) \quad \forall x$

② $P_Y(y) = \sum_x P_{X,Y}(x,y) \quad \forall y$

Why are these true?

$P_X(x) = \mathbb{P}(\{X = x\}) = \mathbb{P}(A_x) = \sum_y \mathbb{P}(A_x \cap B_y) = P_{X,Y}(x,y) \ \forall x$ ← by Total Probability Thm

These definitions generalize in an obvious way to >2 rv's defined on same $\Omega, \mathbb{P}$.

KEY THING: - joint pmf determines the marginals
- marginals do NOT determine the joint

Recall the expected value rule: Given X, $P_X(x)$, $Y = g(X)$, have

$$\mathbb{E}(Y) = \sum_x g(x) P_X(x)$$ ← good to save us from computing $P_Y(y)$ first before $\mathbb{E}(Y)$

Similarly, given X, Y w/ joint pmf $P_{X,Y}(x,y)$ and some real valued function $z = g(X,Y)$, we have

$$\mathbb{E}(z) = \sum_{x \in X} \sum_{y \in Y} g(x,y) P_{X,Y}(x,y)$$ ← don't need $P_Z(z)$ to get $\mathbb{E}(z)$

Next, Conditional Stuff

Given $\Omega, \mathbb{P}$ and a discrete rv X defined on Ω and an event $A \subset \Omega$ w/ $\mathbb{P}(A) > 0$, and a possible value x for X, the conditional pmf of X given A is defined as

$$P_{X|A}(x) = \frac{\mathbb{P}(\{X = x\} \cap A)}{\mathbb{P}(A)}$$ ← "$\mathbb{P}(B|A)$ where B is the event $\{X = x\}$"

Observe that for any A w/ $\mathbb{P}(A) > 0$, $P_{X|A}(x)$ as x ranges over X's value space defines a pmf - i.e. $P_{X|A}(x) \geq 0 \ \forall x$ and $\sum_{x \in X} P_{X|A}(x) = 1$

Here's a fact that's similar to (and follows directly from) the Total Probability Thm: If events $A_1, A_2, \ldots$ partition Ω, and $\mathbb{P}(A_k) > 0$ for $1 \leq k \leq n$, then for any discrete rv X on Ω,

$$P_X(x) = \sum_{k=1}^{n} P_{X|A_k}(x)\,\mathbb{P}(A_k)$$

More often, encounter conditional pmf of X given some other rv Y (defined on same $\Omega, \mathbb{P}$) - given X, Y defined on $\Omega, \mathbb{P}$, conditional pmf of X given Y is defined for all x and for all y with $\mathbb{P}(\{Y = y\}) = P_Y(y) > 0$ as

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$$ ← Same as $P_{X|A}(x)$ where $A = \{Y = y\}$ ← Standard Notation

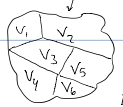
Note that for any y w/ $P_Y(y) > 0$, $P_{X|Y}(x|y)$ as x ranges over X values defines a pmf

i.e. $P_{X|Y}(x|y) \geq 0$ and $\sum_{x \in X} P_{X|Y}(x|y) = 1$

Given X, P_X, and $Y = g(X)$,

$$\mathbb{E}(Y) = \sum_{x \in X} g(x) P_X(x)$$

Why?



V_k = set of x-values for which $g(x) = y_k$

$P_Y(y_k) = \sum_{x \in V_k} P_X(x)$

Thus,

$$\mathbb{E}(Y) = \sum_{y_k} y_k P_Y(y_k)$$
$$= \sum_{y_k} y_k \sum_{x \in V_k} P_X(x)$$ ← for all $x \in V_k$, $g(x) = y_k$ (for $1 \leq k \leq \ell$)
$$= \sum_{y_k} \sum_{x \in V_k} g(x) P_X(x)$$
$$= \sum_{x \in X} g(x) P_X(x)$$

Total probability rule: If $A_1, \ldots, A_n$ partition Ω, then

$$P_X(x) = \sum_{k=1}^{n} P_{X|A_k}(x)\,\mathbb{P}(A_k)$$

Given two rvs on same $\Omega, \mathbb{P}$ - say X and Y - define

$$P_{X|Y}(x|y) = \frac{\mathbb{P}(\{X = x\} \cap \{Y = y\})}{\mathbb{P}(\{Y = y\})} = \frac{P_{X,Y}(x,y)}{P_Y(y)} \ ; \ \forall x, y \text{ w/ } P_Y(y) > 0$$

For fixed y, $P_{X|Y}(x|y)$ defines a pmf "over x-values" - i.e.

$$P_{X|Y}(x|y) \geq 0 \ \forall x \quad \text{and} \quad \sum_{x \in X} P_{X|Y}(x|y) = 1$$

"$P_{X|Y}(x|y)$ = conditional pmf of X given $Y = y$"

Like the conditional "event-centered story", have a product rule of sorts

$$P_{X,Y}(x,y) = P_Y(y) P_{X|Y}(x|y) \quad \forall \ \begin{matrix} x \in X \\ y \in Y \end{matrix}$$

OR

$$P_{X,Y}(x,y) = P_X(x) P_{Y|X}(y|x) \quad \forall \ \begin{matrix} x \in X \\ y \in Y \end{matrix}$$

This expresses joint in terms of marginals + conditional(s). Also have a total-probability rule of sorts:

$$P_X(x) = \sum_{y \in Y} P_Y(y) P_{X|Y}(x|y)$$

$$P_Y(y) = \sum_{x \in X} P_X(x) P_{Y|X}(y|x)$$

Given $\Omega, \mathbb{P}$, and $X: \Omega \to \mathbb{R}$ a r.v., say X is a **continuous r.v.** when there exists a function $f_X(x)$ — called the probability density function (pdf) of X — such that "any" $V \subseteq \mathbb{R}$,

$$\mathbb{P}(\{X \in V\}) = \int_V f_X(x)\,dx$$

← $f_X(x)$ has to be reasonable enough for integrals to make sense

- $f_X(x) \geq 0 \quad \forall x$ (need this to ensure $\mathbb{P}(\{X \in V\}) \geq 0 \ \forall V \subseteq \mathbb{R}$)
- $\lim_{R\to\infty} \int_{-R}^{R} f_X(x)\,dx = 1 \to \int_{-\infty}^{\infty} f_X(x)\,dx = \mathbb{P}(\{X \in (-\infty,\infty)\}) = 1$
- Given $x \in \mathbb{R}$, $f_X(x)$ is **NOT** $\mathbb{P}$(some event) — in particular,

$$f_X(x) \neq \mathbb{P}(\{X = x\})$$

Turns out $\mathbb{P}(\{X=x\}) = 0 \ \forall x \in \mathbb{R}$ when X is a continuous random variable

Expected Value

The expected value of a continuous rv X w/ pdf $f_X(x)$:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\,dx$$

← Caution: NOT always defined — integral might fail to exist

Expected Value Rule

Given X w/ pdf $f_X(x)$ and $Y = g(X)$, we have

$$\mathbb{E}[Y] = \int_{-\infty}^{+\infty} g(x) f_X(x)\,dx$$

← enables $\mathbb{E}[Y]$ computation w/out finding $f_Y(y)$ or $P_Y(y)$

Variance

Variance of continuous rv:

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

By expected value rule, we have

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^2 f_X(x)\,dx$$

Also, as before,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Next, define — for **ANY** rv X (discrete OR continuous) — the cumulative distribution function (cdf) by

$$F_X(x) = \mathbb{P}(\{X \leq x\}) \quad \forall x \in \mathbb{R}$$

Observation: If X is a continuous rv w/ pdf $f_X(x)$, then since

$$\mathbb{P}(\{X \leq x\}) = \int_{-\infty}^{x} f_X(t)\,dt$$

we have

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \quad \text{and} \quad f_X(x) = \frac{d}{dx} F_X(x)$$

Discrete version: If X is a discrete rv w/ pmf $p_X(x)$ we

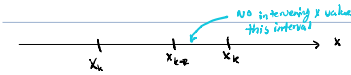
$$F_X(x) = \sum_{\{k : x_k \leq x\}} p_X(x_k)$$

← set of all possible X-values that don't exceed x

Can invert this formula to get $p_X(x)$ in terms of $F_X(x)$:

$$p_X(x_k) = F_X(x_k) - F_X(x_{k-1})$$

where x_{k-1} is the "NEXT largest value" of X below x_k.



General Properties of CDFs

① $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$

② When X is a continuous rv, $F_X(x)$ is continuous in x and differentiable "almost everywhere" (corners in $F_X(x)$ correspond to jumps in $f_X(x)$)

③ X is a discrete iff $F_X(x)$ is a piecewise constant.

④ $F_X(x)$ is **monotonically** increasing in x.

$$x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$$

Say X, Y rvs defined in same $\Omega, \mathbb{P}$ are jointly continuous w/ joint pdf $f_{X,Y}(x,y)$ when

$$\mathbb{P}((X,Y) \in V) = \iint_V f_{X,Y}(x,y)\,dx\,dy \quad \forall V \subseteq \mathbb{R}^2$$

Special case of a V: $[a_1,b_1] \times [a_2,b_2]$

Then

$$\mathbb{P}((X,Y) \in V) = \int_{a_1}^{b_1} dx \int_{a_2}^{b_2} dy \left(f_{X,Y}(x,y)\right)$$

Again, have marginals

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy \quad ; \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx$$

An official way to get this: Get $F_X(x)$ first then take $\frac{d}{dx} F_X(x)$

$$F_X(x) = \mathbb{P}(\{X \leq x\}) = \mathbb{P}(\{(X,Y) \in (-\infty, x] \times (-\infty,\infty)\}) = \int_{-\infty}^{x} dt \int_{-\infty}^{+\infty} dy \left(f_{X,Y}(t,y)\right)$$

then

$$\frac{d}{dx} F_X(x) = \int_{-\infty}^{+\infty} dy \left(f_{X,Y}(x,y)\right)$$

Could also derive marginal formulas as follows:

$$\forall V \subseteq \mathbb{R}, \ \mathbb{P}(\{X \in V\}) = \mathbb{P}(\{(X,Y) \in V \times (-\infty,\infty)\}) = \int_V dx \int_{-\infty}^{+\infty} dy \left(f_{X,Y}(x,y)\right) = \int_V \left(\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy\right) dx$$

← must be $f_X(x)$: integrate over V to get $\mathbb{P}(\{X \in V\})$

Other stuff

- $\int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy \left(f_{X,Y}(x,y)\right) = 1$
- Joint CDF: $F_{X,Y}(x,y) = \mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}) = \int_{-\infty}^{x} ds \int_{-\infty}^{y} dt \left(f_{X,Y}(s,t)\right)$
- $f_{X,Y}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y} F_{X,Y}(x,y)$

Conditional Stuff For Continuous Random Variables

Given a continuous rv X on $\Omega, \mathbb{P}$ and some event $A \subseteq \Omega$, the conditional pdf of X given A "defined" as follows:

For any $V \subseteq \mathbb{R}$, we have

$$\mathbb{P}(\{X \in V\} | A) = \int_V f_{X|A}(x)\,dx$$

In general, no decent formula for $f_{X|A}(x)$ in terms of $f_X(x)$. One way to compute it:

- First get conditional cdf of x given A

$$F_{X|A} = \mathbb{P}(\{X \leq x\} | A)$$

- Then take d/dx to get $f_{X|A}(x)$

However, if A is an event of the form $\{X \in W\}$, and $\mathbb{P}(A) > 0$, we have

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{\mathbb{P}(\{X \in W\})}, & \text{when } x \in W \\ 0, & \text{otherwise} \end{cases}$$

How does this arise?

$$\mathbb{P}(\{X \in V\} | A) = \frac{\mathbb{P}(\{X \in V \cap W\})}{\mathbb{P}(\{X \in W\})} = \frac{\int_{V\cap W} f_X(x)\,dx}{\mathbb{P}(\{X \in W\})}$$

Total Probability Theorem in context of $f_{X|A}$:

If X is a continuous rv and $A_1, \ldots, A_n$ are events of positive probability that partition Ω, then

$$f_X(x) = \sum_{k=1}^{n} f_{X|A_k}(x)\,\mathbb{P}(A_k)$$

To see this: go via cdfs.

$$F_{X|A_k} = \frac{\mathbb{P}(\{X \leq x\} \cap A_k)}{\mathbb{P}(A_k)}$$

$$\frac{d}{dx} F_{X|A_k}(x) = f_{X|A_k}(x)$$

By Total Probability Theorem,

$$F_X(x) = \mathbb{P}(\{X \leq x\}) = \sum_{k=1}^{n} F_{X|A_k}(x)\,\mathbb{P}(A_k) \xrightarrow{\frac{d}{dx}} \sum_{k=1}^{n} f_{X|A_k}(x)\,\mathbb{P}(A_k) = f_X(x)$$

Comment: this holds when A_k aren't of the special form $\{X \in W\}$!

Bottom line: conditional pdf of X given $Y = y$ is

$$f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

← What you integrate over for any $x \in V$ to get $\mathbb{P}(\{X \in V\} | \{Y = y\})$

Expected value rule for joints.

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy\, g(x,y) f_{X,Y}(x,y)$$

As for conditional pmfs in discrete-world, can use conditional pdfs to compute joints, marginals, etc., in situations most naturally expressed in conditional terms.

e.g.

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y)$$

Integrate over x or y to get

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X|Y}(x|y) f_Y(y)\,dy \quad \text{OR} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{Y|X}(y|x) f_X(x)\,dx$$

Conditional Expected Values

Given X, A

$$\mathbb{E}(X|A) = \int_{-\infty}^{+\infty} x f_{X|A}(x)\,dx$$

Given X, Y

$$\mathbb{E}(X|\{Y=y\}) = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)\,dx \quad \forall y$$

Expected Value Rule

$$\mathbb{E}(g(X)|A) = \int_{-\infty}^{+\infty} g(x) f_{X|A}(x)\,dx$$

$$\mathbb{E}(g(X)|Y=y) = \int_{-\infty}^{+\infty} g(x) f_{X|Y}(x|y)\,dx \quad \forall y$$

Recall the Total Probability-type results

- If events $A_1, A_2, \ldots, A_n$ have >0 probability and partition Ω then

$$f_X(x) = \sum_{k=1}^{n} f_{X|A_k}(x)\,\mathbb{P}(A_k)$$

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X|Y}(x|y) f_Y(y)\,dy$$

From these follows

Total Expectation Theorems

$$\mathbb{E}(X) = \sum_{k=1}^{n} \mathbb{E}(X|A_k)\,\mathbb{P}(A_k)$$

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} \mathbb{E}(X|Y=y) f_Y(y)\,dy$$

For **ANY** pair X, Y (continuous, discrete, or "one and one")

X, Y independent $\iff F_{X,Y}(x,y) = F_X(x) F_Y(y) \quad \forall x,y$ ← Requires proof

$\Downarrow$ implies (take $\partial/\partial x, \partial/\partial y$)

For X, Y **BOTH** continuous w/ densities $f_X(x), f_Y(y)$; $f_{X,Y}(x,y)$:

X, Y independent $\iff f_{X,Y}(x,y) = f_X(x) f_Y(y) \quad \forall x,y$

Comment: When X, Y independent, we have

- $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$
- $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$
- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x) f_X(x)}{\int_{-\infty}^{+\infty} f_{Y|X}(y|x) f_X(x)\,dx}$$

} Continuous Bayes's Rule

The **reverse** situation also arises. Observe some event A; infer about conditional rv Y — specifically $f_{Y|A}(y)$ — know $f_Y(y)$ and $\mathbb{P}(A|Y=y)$, $\mathbb{P}(A^c|Y=y)$.

Flip Baye's Rule.

$$\mathbb{P}(A|Y=y) = \frac{f_{Y|A}(y)\,\mathbb{P}(A)}{f_Y(y)} \rightsquigarrow f_{Y|A}(y) = \frac{\mathbb{P}(A|Y=y) f_Y(y)}{\mathbb{P}(A)}$$

Rewriting $\mathbb{P}(A)$ using Total Probability Theorem yields

$$f_{Y|A}(y) = \frac{\mathbb{P}(A|Y=y) f_Y(y)}{\int_{-\infty}^{+\infty} \mathbb{P}(A|Y=y) f_Y(y)\,dy}$$

$X \sim \mathrm{Uniform}[a,b]$

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in (a,b) \\ 0, & \text{else} \end{cases}$$

$$\mathbb{E}(X) = \frac{b+a}{2}, \ \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

$$F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & \text{else} \end{cases}$$

$X \sim \mathrm{Exponential}(\lambda)$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \lambda > 0 \\ 0, & \lambda < 0 \end{cases}$$

$$\mathbb{E}(X) = 1/\lambda, \ \mathrm{Var}(X) = 1/\lambda^2$$

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & \text{else} \end{cases}$$

$X \sim \mathrm{Gaussian}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}(X) = \mu, \ \mathrm{Var}(X) = \sigma^2$$

$$F_X(x) = \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

If $Z \sim \mathrm{Gaussian}(0,1)$

$X = \sigma Z + \mu$

If $Y = aX + b$, then

$Y \sim \mathrm{Gaussian}(a\mu + b, a^2\sigma^2)$

- X is a continuous rv whose density $f_X(x)$ is concentrated on a single interval $a<x<b$. $a=-\infty$ and/or $b=+\infty$ allowed.
- $Y=g(X)$, g strictly monotonic and differentiable, implying that $f_Y(y)$ is concentrated on $(g(a),g(b))$ (if increasing) OR $(g(b),g(a))$ (if decreasing)
- Let h be the inverse function of g – defined only on $(g(a),g(b))$ or $(g(b),g(a))$ — h is also strictly monotonic and differentiable on its domain of definition

Then

$$f_Y(y) = \begin{cases} \left|\frac{dh(y)}{dy}\right| f_X(h(y)) & , \ y\in(g(a),g(b)) \text{ OR } y\in(g(b),g(a)) \\ 0 & , \ \text{else} \end{cases}$$

(increasing / decreasing)

Covariance

Given any X,Y rvs (discrete, continuous, whatever) defined on **same** probability space, the **covariance** of X and Y defined as

$$\text{Cov}(X,Y) = \mathbb{E}\left((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right)$$

Note:

$$\text{Cov}(X,X) = \mathbb{E}\left((X-\mathbb{E}(X))^2\right) = \text{Var}(X)$$

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(\mathbb{E}(X)Y) - \mathbb{E}(X\mathbb{E}(Y)) + \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

Terminology: When $\text{Cov}(X,Y)=0$, say X and Y are uncorrelated.

Fact: If X,Y independent, then X,Y uncorrelated

(If statement; NOT iff)

Independent $\Rightarrow \mathbb{E}(XY)=\mathbb{E}(X)\mathbb{E}(Y) \Rightarrow \text{Cov}(X,Y)=0$

Converse is **NOT** true

Terminology: Given X,Y, the correlation coefficient of X and Y is

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Turns out, $|\rho|\leq 1$ – turns out this is a version of the Cauchy-Schwarz inequality.

Note: $\rho=0 \overset{\text{iff}}{\iff} X,Y$ uncorrelated and

$\rho=\pm 1 \iff X$ and Y are "aligned" in sense $X=\alpha Y$ for some $\alpha\neq 0$

Conditional Expectation Revisited

Terminology: $\mathbb{E}(X|Y)$ = conditional expectation of X given Y

Question: What is $\mathbb{E}(\mathbb{E}(X|Y))$?

Fact: Law of iterated expectations

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

Idea: $\mathbb{E}(X|Y)=g(Y)$ for some function Y.

Thus

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(g(Y))$$ ← use expected value rule to get this

(Discrete)

$$\mathbb{E}(g(Y)) = \sum_{y\in Y} g(y)p_Y(y)$$

$$= \sum_{y\in Y} \mathbb{E}(X|Y=y)p_Y(y)$$

(law of total expectation)

$$= \mathbb{E}(X)$$

OR

(Continuous)

$$\mathbb{E}(g(Y)) = \int_{-\infty}^{+\infty} g(y)f_Y(y)\,dy$$

$$= \int_{-\infty}^{+\infty} \mathbb{E}(X|Y=y)f_Y(y)\,dy$$

$$= \mathbb{E}(X)$$

Law of Iterated Expectations: $\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$

For any function $h(Y)$,

- $\mathbb{E}(h(Y)|Y) = h(Y)$
- $\mathbb{E}(h(Y)X|Y) = h(Y)\mathbb{E}(X|Y)$

Can think of $\mathbb{E}(X|Y)$ as an estimator of X given Y.

In what sense does it "act like an estimator?"

- $\mathbb{E}(X|Y) = \mathbb{E}(X)$ by law of iterated expectations
- The estimation error $X-\mathbb{E}(X|Y)$ is uncorrelated w/ the estimate $\mathbb{E}(X|Y)$ – in fact, $X-\mathbb{E}(X|Y)$ is uncorrelated with Y – More generally, w/ any function $h(Y)$

Taking $\mathbb{E}(X|Y) \approx$ orthogonally projecting X onto "space" of functions of Y

Further justification of this geometric world-view:

- orthogonal projections of X onto space of $h(Y)$-functions should be the thing in that space "closest to X"

Standard notion of "closeness": mean-squared difference

Fact (Major): $\mathbb{E}(X|Y)$ is the function of Y that minimizes $\mathbb{E}((X-h(Y))^2)$ over ALL functions $h(Y)$

Conditional Variance

Given X,Y conditional variance of X given Y is the random variable

$$\text{Var}(X|Y) = \mathbb{E}\left((X-\mathbb{E}(X|Y))^2|Y\right)$$

A recipe similar to the "g-thing" for computing $\text{Var}(X|Y)$

- Given y, compute

$$\text{Var}(X|Y=y) = \mathbb{E}\left[(X-\mathbb{E}(X|Y=y))^2|Y=y\right]$$

  - Do this by finding conditional pmf $p_{X|Y}(x|y)$ or pdf $f_{X|Y}(x|y)$ and then computing variance of it

- This yields a function of y – $\gamma(y)$ – plug Y in for y; that yields

$$\text{Var}(X|Y) = \gamma(Y)$$

How to compute? In general

- Find conditional pdf/pmf $f_{X|Y}(x|y)$ / $p_{X|Y}(x|y)$
- Mean of that is $\mathbb{E}(X|Y=y)$
- Variance of that is

$$\text{Var}\left(X-\mathbb{E}(X|Y)\middle|Y=y\right) = \begin{cases} \int_x \left(X-\mathbb{E}[X|Y=y]\right)^2 f_{X|Y}(x|y)\,dx \\ \text{OR} \\ \sum_x \left(X-\mathbb{E}[X|Y=y]\right)^2 p_{X|Y}(x|y) \end{cases}$$

Law of Total Variance: (a sometimes useful identity)

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}[\mathbb{E}(X|Y)]$$

Moment Generating Function (MGF)

Given X, continuous or discrete, define MGF of X as

$$M_X(s) = \mathbb{E}(e^{sX}) \quad - \; s \text{ is "a variable"}$$

$M_X(s)$ is a function of s – need to be careful about its domain of definition

$$\mathbb{E}(X^k) = \frac{d^k}{ds^k} M_X(s)\Big|_{s=0}$$

Important Fact:

$f_X(x)$, $p_X(x)$ $\longleftrightarrow$ (one-to-one correspondence) $M_X(s)$

i.e $M_X(s)$ determines pmf or pdf of X completely.

Say X_1, X_2 are Gaussian (μ_1, σ_1^2) and (μ_2, σ_2^2) and independent. Have already seen

$$Y = X_1 + X_2$$

is also Gaussian.

Can do this by

$$f_Y(y) = f_{X_1}(x) * f_{X_2}(x)$$

Alternative argument based on moment generating functions:
Based on the fact that when X_1, X_2 independent : $Y = X_1 + X_2$

$$M_Y(s) = M_{X_1}(s) M_{X_2}(s)$$

Reason: $M_Y(s) = \mathbb{E}(s^Y) = \mathbb{E}(e^{s(X_1+X_2)}) = \mathbb{E}(e^{sX_1} e^{sX_2}) = \mathbb{E}(e^{sX_1})\mathbb{E}(e^{sX_2})$

Convolution in pdf $\longleftrightarrow$ Multiplication in $M_X(s)$

we have a "recipe" for finding $M_Y(s) = \mathbb{E}(e^{sY})$:

- first find $M_N(s)$ ← Y depends on N
- then plug $M_X(s)$ everywhere you see e^s in $M_N(s)$

Let

$$M_n = \frac{X_1 + \cdots + X_n}{n}, \quad X_k \text{ iid}$$

Saw

$$\mathbb{E}(M_n) = \mu \;\; \forall n; \quad \mathrm{Var}(M_n) = \frac{\sigma^2}{n} \;\; \forall n$$

Chebyshev Inequality:

$$\mathbb{P}(\{|X-\mu| \geq c\}) \leq \frac{\mathrm{Var}(X)}{c^2} \quad \forall c$$

From this, it follows that

$$\mathbb{P}(\{|M_n - \mu| \geq \varepsilon\}) \leq \frac{\sigma^2}{n\varepsilon^2} \quad \forall \varepsilon > 0$$

Consequence

$$\mathbb{P}(\{|M_n - \mu| \geq \varepsilon\}) \to 0 \text{ as } n \to \infty \quad \forall \varepsilon > 0$$

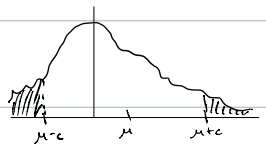
Known as the weak law of large numbers (WLLN)

Markov Inequality: If X is a nonnegative-valued rv, then for every $c > 0$,

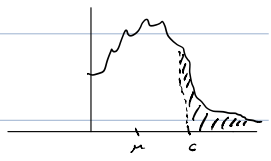
$$\mathbb{P}(\{X \geq c\}) \leq \frac{\mathbb{E}(X)}{c}$$

Think of these both as quantitative bounds on tail probabilities

Chebyshev



Markov



Another consequence of Chebyshev:

$$\mathbb{P}(\{|X-\mu| \geq k\sigma\}) \leq \frac{1}{k^2}$$

"probability of being k std deviation away from mean $\leq \frac{1}{k^2}$"

The kind of convergences taking place in WLLN **converges in probability**

Definition: Given a sequence of rv's Y_n, $1 \leq n < \infty$, and a number a, say Y_n converges in probability to a as $n \to \infty$ when $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\{|Y_n - a| \geq \varepsilon\}) = 0$$

WLLN just says "$M_n \to \mu$ in probability as $n \to \infty$"

Convergence in probability is a weakish kind of convergence; just one of many we encounter.

Another is, in the same context, say Y_n converges to a mean sequence when

$$\lim_{n\to\infty} \mathbb{E}[|Y_n - a|^2] = 0$$

Can show mean square convergence ⇒ Convergence in probability

Central Limit Theorem

Recall that if X_k iid w/ common μ, σ^2

$$M_n = \frac{X_1 + \cdots + X_n}{n} \Rightarrow \mathbb{E}(M_n) = \mu \text{ and } \mathrm{Var}(M_n) = \frac{\sigma^2}{n}$$

Form Z_n by renormalizing so $\mathbb{E}(Z_n) = 0$ and $\mathrm{Var}(Z_n) = 1 \; \forall n$.

$$Z_n = \frac{X + \cdots + X_n - n\mu}{\sqrt{n}\,\sigma}$$ (check that this works!) ← set $Z_n = \frac{\sqrt{n}}{\sigma}(M_n - \mu)$

Fact: this is the Central Limit Theorem!

In this context, Z_n converges to, as $n \to \infty$, a Gaussian with mean $\mu = 0$ and variance $\sigma^2 = 1$.

Converging in the sense that for every z,

$$F_{Z_n}(z) \longleftrightarrow \Phi(z)$$ ← standard normal CDF

Last thing: convergence w/ probability 1 of a sequence $Y_1, Y_2, \ldots$ of random variables.

Consider the sequence $\{Y_n : n > 0\}$.
Given some random variable Y,

$$\left\{ \lim_{n \to \infty} Y_n = Y \right\}$$

is an event – need to refer back to $\Omega, \mathbb{P}$, etc.

Say $Y_n \to Y$ with probability 1 (w.p.1)
$Y_n \xrightarrow{w.p.1} Y$
$Y_n \xrightarrow{a.s.} Y$ a.s. → almost surely

when this event has probability 1.

Turns out: Convergence with probability 1 ⇒ convergence in probability

Strong Law of Large Numbers: When X_n iid, common mean μ, common variance σ^2, and $M_n = \frac{X_1 + \cdots + X_n}{n}$, we have

$$M_n \xrightarrow{w.p.1} \mu \quad \text{as} \quad n \to \infty$$