

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

$$1) P(A) \geq 0 \text{ for every event } A$$

$$2) P(\Omega) = 1 \text{ (normalization)}$$

3) Additivity Rules

$$i) \text{ If } A \cap B = \emptyset \text{ (A,B are events) then } P(A \cup B) = P(A) + P(B)$$

$$ii) \text{ If } A_1, A_2, A_3, \dots \text{ is a countable sequence of mutually disjoint events (i.e. } A_i \cap A_j = \emptyset \forall i \neq j\text{), then } P(\bigcup A_n) = \sum P(A_n)$$

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

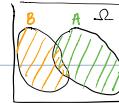
"Event A occurs" means "outcome of experiment is in A ".

Given: Ω and P , and two events $A, B \subseteq \Omega$, define $P(A|B)$ = "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 P -mass that also lies in A ".



Motivates the definition:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\text{shaded})}{P(\text{all})} \quad \text{for } P(B) > 0$$

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

- $P(A|B) \geq 0 \quad \forall A \subseteq \Omega$
- $P(\Omega|B) = 1$ (by 2nd reality check above)
- If $A_1, A_2, \dots = \emptyset$, then disjoint
- $(A_1 \cup A_2) \cap B = (A_1 \cap B) \cup (A_2 \cap B)$
- $\xrightarrow{\text{divide both sides by } P(B)} \Rightarrow P((A_1 \cup A_2) \cap B) = P(A_1 \cap B) + P(A_2 \cap B)$
- $\Rightarrow P(A_1 \cap B) = P(A_1|B) + P(A_2|B)$

Counting Principles

Re-count # of subsets of $\Omega = \{S_1, \dots, S_n\}$. 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"

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

Comment: Amazing that this is an integer!

Can also write

$${n \choose k} = \frac{n!}{k!(n-k)!}$$

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 simulate via flip in this case.

- MT possible outcome at each stage - if $P(\text{flip}) = p$, $P(\text{not flip}) = 1-p$ and you perform experiment n times, for each k , we know

$$P(\text{sequence you get}) = p^k (1-p)^{n-k}, \text{ where } k = \# \text{ of heads}$$

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

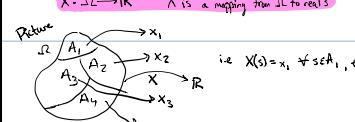
$P(\text{you get } k \text{ H's}) = \#(\text{ways of picking } k \text{ spots in } n \text{ available spots})$

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

Next BIG TOPIC: DISCRETE RANDOM VARIABLES

Start with Ω and 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. $X: \Omega \rightarrow \mathbb{R}$ "X is a mapping from Ω to \mathbb{R} "



Conditional Independence: Ω, P ; say events A and B are conditionally independent given (event) C when

$$P(A \cap B|C) = P(A|C)P(B|C)$$

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

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

To see this just play w/ formulas

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

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

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

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

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

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

Book uses

$$P(X=x) \text{ or } P(X=x) \xrightarrow{\text{misses of notation}} \text{compared to above}$$

to refer to $P(A_x)$, where $A_x = \{s \in \Omega : X(s) = x\}$

Things to note about P_X :

$$- P_X(x) \geq 0 \text{ for all possible values of } X$$

(why? cause for any x , $P_X(x)$ is P (an event) ≥ 0 !)

$$- \text{If } V \text{ is any finite or countably infinite set of possible values of } X, \text{ then if we set}$$

$$B = \text{the event " } X \in V \text{"}$$

$$\text{i.e. } B = \{s \in \Omega : X(s) \in V\}$$

$$\text{then } P(B) = \sum_{x \in V} P_X(x)$$

Binomial RV

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

$$P_k(x) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{if } k \leq n$$

This could arise from

a) fall H-T sequence of length n

b) Many separate $p \in (0, 1)^n$

like the outcome of n -independent flips of a possibly unfair coin.

$$X(\text{any sequence}) = \#(\text{Heads}) \text{ in sequence}$$

Geometric Random Variable

$$Geometric(p) = p(1-p)^{k-1}, \quad k \in \mathbb{N}$$

A possible Ω, P ? You have a coin with $P(H) = p$; $\Omega = \text{set of all infinite sequences of H-T}$; $X(s) = \text{index of first head for all } s \in \Omega$.

Reality check: recall $\sum_{k=0}^{\infty} P_k(x) = 1$ for any discrete r.v. X .

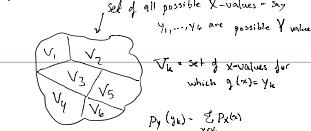
Verify this for geometric r.v.:

$$\sum_{k=0}^{\infty} P_k(x) = \sum_{k=0}^{\infty} p(1-p)^{k-1} = p \sum_{k=0}^{\infty} (1-p)^k = p \frac{1}{1-(1-p)} = 1$$

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

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

Why?



Thus,

$$E(Y) = \sum_{y \in Y} y P_Y(y)$$

$$= \sum_{y \in Y} y \sum_{x \in X} P_X(x) \quad (\text{for } x \in f^{-1}(y))$$

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

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

$$P(x) = \sum_{i=1}^n P(A_i) P(x|A_i)$$

Given two r.v.s on same Ω, P - say X and Y - define

$$P_{XY}(x,y) = \frac{P((x,y) \cap \{(x,y)\})}{P((x,y))} = \frac{P(x,y)}{P(X \times Y)} = \frac{P(X \times Y)}{P(X)P(Y)}$$

For fixed y , $P_{XY}(y)$ defines a prof over x -values -

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

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

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

$$P_{XY}(x,y) = P_X(x)P_{X|Y}(x|y) + x \times y$$

OR

$$P_{XY}(x,y) = P_Y(y)P_{Y|X}(y|x) + x \times y$$

This expresses joint in terms of marginals + conditionals. Also have a total-probability rule of sorts:

$$P_{XY}(x,y) = \frac{P((x,y) \cap A)}{P(A)} = \frac{P((x,y))}{P(A)} \quad \text{where } A = \{s \in \Omega : X(s) = x, Y(s) = y\}$$

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

$$P_{XY}(x,y) = \sum_{x \in X} P_{XY}(x,y)$$

$$P_{XY}(y) = \sum_{x \in X} P_{XY}(x,y)$$

Here's a fact that's similar to (and follows directly from) the Total Probability Thm: If events A_1, \dots, A_n partition Ω , and $P(A_i) > 0$ for $1 \leq i \leq n$, then for any discrete r.v. X on Ω ,

$$P_{XY}(x) = \sum_{i=1}^n P(A_i) P_{XY}(x|A_i) \quad \text{where } B = \text{the event } \{x\}$$

More often, encounter conditional prof of X given some other r.v. Y (defined on same Ω, P): given X, Y defined on Ω, P , conditional prof of X given Y is defined for all x & and for all y with $P(Y=y) > 0$ as

$$P_{XY}(x|y) = P_{XY}(x,y) \quad \text{Same as } P_{XY}(x,y)$$

$$P_Y(y) \quad \text{where } A = \{x : Y(x) = y\}$$

Standard Notation

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

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

Given S , P , and $X: S \rightarrow \mathbb{R}$ a rv. Say X is a continuous rv when there exists a function $f_X(x)$ - called the probability density function (pdf) of X - such that "any" $\forall x \in \mathbb{R}$,

$$P(\{x \in S\}) = \int_S f_X(x) dx \quad \text{if } f_X(x) \text{ has to be reasonable enough for integrals to make sense}$$

$- f_X(x) \geq 0 \forall x$ (need this to ensure $P(\{x \in S\}) \geq 0 \forall S \subset \mathbb{R}$)

$$- \lim_{R \rightarrow \infty} \int_R^\infty f_X(x) dx = 1 \rightarrow \int_0^\infty f_X(x) dx + P(X \in (-\infty, 0)) = 1$$

- Given $x \in \mathbb{R}$, $f_X(x)$ is NOT $P(\text{some event})$ - in particular, $f_X(x) \neq P(\{X=x\})$

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

Expected Value

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

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx \quad \text{Caution: NOT always defined - integral may fail to exist}$$

then

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

Could also derive marginal formulas as follows:

$$\forall V \subset \mathbb{R}, \quad P(\{X \in V\}) = P(\{(X,V) \in (-\infty, \infty)^2\})$$

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

Then $f_X(x) = \int_V^{\infty} f_{X,Y}(x,y) dy$

mean is $f_X(x)$, integrate over V to get $P(X \in V)$

Other stuff,

$$- \int_{-\infty}^{+\infty} dy (f_{X,Y}(x,y)) = 1$$

- Joint CDF: $F_{X,Y}(x,y) = P(\{(X,Y) \in (-\infty, x) \times (-\infty, y)\})$

$$- f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

Conditional Stuff For Continuous Random Variables

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

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

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

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

$$\text{F}_{X|A}(x) = P(\{X \in (-\infty, x)|A\})$$

- Then take derivative to get $f_{X|A}(x)$

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

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

How does this arise?

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

Total Probability Theorem in context of f_X(x):

If X is a continuous rv and A_1, \dots, A_n are events of positive probability that partition \mathbb{R} , then

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

To see this: go via cdfs.

$$F_{X|A_k}(x) = \frac{P(\{X \in (-\infty, x)|A_k\})}{P(A_k)}$$

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

By Total Probability Theorem,

$$F_X(x) = P(\{X \in (-\infty, x)\}) = \sum_{k=1}^n F_{X|A_k}(x) P(A_k) \xrightarrow{\text{def}} \sum_{k=1}^n f_{X|A_k}(x) P(A_k) = f_X(x)$$

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

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{What you integrate over for any } x \in \mathbb{R} \text{ to get } P(\{X \in V|Y=y\})$$

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 pdfs 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

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

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

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

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

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

Conditional Expected Values

Given $X|A$

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

Given $X|Y$

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

Expected Value Rule

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

Recall the Total Probability-type results

- If events A_1, A_2, \dots, A_n have >0 probability and partition \mathbb{R} , then

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

From these follows

Total Expectation Theorems

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

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

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

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

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

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

$$\text{If } Z \sim \text{Gaussian}(0,1) \\ X = \sigma Z + \mu, \text{ then} \\ \text{If } Y = aX + b, \text{ then} \\ Y \sim \text{Gaussian}(a\mu + b, a^2\sigma^2)$$

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

↳ requires proof

X, Y independent $\Leftrightarrow F_{X,Y}(x,y) = F_X(x) F_Y(y) \quad \forall x, y$

↳ implies (take ∂_x, ∂_y)

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

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

Comment: When X, Y independent, we have

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

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

$$- \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

$$f_{X,Y}(x,y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x) f_X(x)}{\int_{-\infty}^{+\infty} f_{Y|X}(y|x) f_X(x) dx} \quad \left. \begin{array}{l} \text{Continuous} \\ \text{Baye's} \\ \text{Rule} \end{array} \right\}$$

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

Flip Baye's Rule.

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

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

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

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

② When X is a continuous rv, $F_X(x)$ is continuous in x and differentiable "almost everywhere" (comes in fact 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 .

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

- 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 $(g(a), g(b))$ or $(g(b), g(a))$
- 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

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

Covariance

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

$$\text{Cov}(X, Y) = \text{IE}((X - \text{IE}(X))(Y - \text{IE}(Y)))$$

Note:

$$\text{Cov}(X, X) = \text{IE}((X - \text{IE}(X))^2) = \text{Var}(X)$$

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

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

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

If statements
not iff: Independent $\Rightarrow \text{IE}(XY) = \text{IE}(X)\text{IE}(Y) \Rightarrow \text{Cov}(X, Y) = 0$

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: $\rho = 0 \Leftrightarrow X, Y$ uncorrelated and

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

Conditional Expectation Revisited

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

Question: What is $\text{IE}(\text{IE}(X|Y))$?

Fact: Law of iterated expectations

$$\text{IE}(X) = \text{IE}(\text{IE}(X|Y))$$

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

Thus

$$\text{IE}(\text{IE}(X|Y)) = \text{IE}(g(Y)) \quad \leftarrow \text{use expected value rule to get this}$$

$$\text{IE}(g(Y)) = \sum_{y \in Y} g(y) P_Y(y)$$

$$\text{IE}(g(Y)) = \int_{-\infty}^{+\infty} g(y) f_Y(y) dy$$

OR

$$= \sum_{y \in Y} \text{IE}(X|Y=y) P_Y(y)$$

law of total expectation \rightarrow

$$= \text{IE}(X)$$

$$= \int_{-\infty}^{+\infty} \text{IE}(X|Y=y) f_Y(y) dy$$

$$= \text{IE}(X)$$

Law of Iterated Expectations: $\text{IE}(\text{IE}(X|Y)) = \text{IE}(X)$

For any function $h(Y)$,

$$\bullet \text{IE}(\text{IE}(h(Y)|Y)) = h(Y)$$

$$\bullet \text{IE}(\text{IE}(h(Y)X|Y)) = h(Y)\text{IE}(X|Y)$$

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

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

$$\bullet \text{IE}(Y|Y) = \text{IE}(Y) \text{ by law of iterated expectations}$$

• The estimation error $X - \text{IE}(X|Y)$ is uncorrelated w/ the estimate $\text{IE}(X|Y)$ — in fact, $X - \text{IE}(X|Y)$ is uncorrelated with Y — More generally, w/ any function $h(Y)$

Taking $\text{IE}(X|Y) \approx$ 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): $\text{IE}(X|Y)$ is the function of Y that minimizes $\text{IE}((X - h(Y))^2)$ over ALL functions $h(Y)$

Conditional Variance

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

$$\text{Var}(X|Y) = \text{IE}((X - \text{IE}(X|Y))^2 | Y)$$

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

• Given g , compute

$$\text{Var}(X|Y=y) = \text{IE}[(X - \text{IE}(X|Y=y))^2 | Y=y]$$

- 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 — 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 $\text{IE}(X|Y=y)$

- Variance of that is

$$\text{Var}(X - \text{IE}(X|Y) | Y=y) = \begin{cases} \int_x (\text{IE}[X|Y=y] - x)^2 f_{X|Y}(x|y) dx \\ \text{OR} \\ \sum_x (\text{IE}[X|Y=y] - x)^2 P_{X|Y}(x|y) \end{cases}$$

Law of Total Variance: (a sometimes useful identity)

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

Moment Generating Function (MGF)

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

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

$M_X(s)$ 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:

$$\left. \begin{array}{l} f_X(x) \\ P_X(x) \end{array} \right\} \xrightarrow{\text{one-to-one correspondence}} M_X(s)$$

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

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

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

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

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

- first find $M_N(s) \leftarrow Y \text{ depends on } N$
- then plug $M_N(s)$ everywhere you see e^s in $M_N(s)$

Let

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

Since

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

Chebyshev Inequality:

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\text{Var}(X)}{c^2} \quad \forall c > 0$$

From this, it follows that

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

Consequence

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

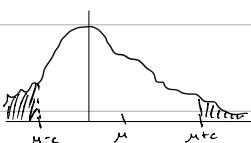
Known as the Weak Law of Large Numbers (WLLN)

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

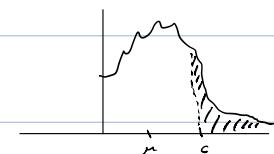
$$\mathbb{P}(\{X \geq c\}) \leq \frac{\mathbb{E}(X)}{c}$$

Think of these both as quantitative bounds on tail probabilities

Chebyshev



Markov



Another consequence of Chebyshev:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \text{"probability of being k std deviation away from mean" \leq \frac{1}{k^2}}$$

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 $n \rightarrow \infty$ when $\forall \epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|Y_n - a| > \epsilon) = 0$$

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

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

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

Can show mean square convergence \Rightarrow convergence in probability

Central Limit Theorem

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

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

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

$$Z_n = \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \quad \left. \begin{array}{l} \text{check that this} \\ \text{works!} \end{array} \right\} \text{set } Z_n = \frac{1}{\sigma} (M_n - \mu)$$

Fact: this is the Central Limit Theorem!

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

Converging in the sense that for every z ,

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

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

Consider the sequence $\{Y_n : n > 0\}$.

Given some random variable Y ,

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

is an event - need to refer back to S_n, P_i , etc.

Say $Y_n \rightarrow Y$ with probability 1 (w.p. 1)

$$Y_n \xrightarrow{\text{w.p. 1}} Y$$

$$Y_n \xrightarrow{\text{a.s.}} Y \quad \text{a.s.} \Rightarrow \text{almost surely}$$

when this event has probability 1.

Turns out: Convergence with probability 1 \Rightarrow convergence in probability

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

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