Probability Review

1 Probability space

Definition of probability space requires 3 parts

- Sample space S containing all outcomes of the experiment
- A collection Ω of the subsets of S, which obey the following,
  - If $A, B \in \Omega$ then $A \cup B \in \Omega$
  - If $A \in \Omega$ then $S \backslash A \in \Omega$.
  - Ω must contain both S and $\emptyset$.

  This collection is should be thought of as how refined in information are the observations of the random variable. For instance, if the experiment is the values of the stock market then observations on the first day of the month are not able to uncover the values of the stock on the second day of the month.
- A probability measure is a function $\mathbb{P} : \Omega \to \mathbb{R}$ so that
  - For all $A \in \Omega$, $0 \leq \mathbb{P}(A) \leq 1$.
  - For $(E_i)_{i=1}^{\infty}$ a subcollection of disjoint sets of Ω we have $\mathbb{P}(\cup_i E_i) = \sum_i \mathbb{P}(E_i)$.
  - $P(\emptyset) = 0$ and $P(S) = 1$.

1.1 Random Variables

A random variable X is a function on S, which may take vales, for example in $\mathbb{R}^n$ for $n \geq 1$. Let us write $X : S \to \mathcal{R}$ in this case, which means X is a function with Domain S and range $\mathcal{R}$.

For 'any' subset A of $\mathcal{R}$, the set $X^{-1}(A) \in \Omega$. (Here 'any' should actually be limited for technical reasons that do not concern us).

Thus in the most rigorous sense, to define a random variable we need a trio $(S, \Omega, \mathbb{P})$ and a function $X : S \to \mathcal{R}$.

Example: (Bernoulli Trial)

■

Recall a Bernoulli trial is a model of a (weighted) coin flip. For pedagogical reasons, lets go through the full above construction. Lets call $S = \{H, T\}$, and $\Omega = \{\emptyset, \{T\}, \{H\}, \{H, T\}\}$, which in this case is the full power set of S. Define $P(T) = q$ and $P(H) = p$. Let $X : S \to \{0, 1\}$, defined by $X(T) = 0$ and $X(H) = 1$.

Thus the probability X is 1 is given by

$$\mathbb{P}(X = 1) = P(X^{-1}(1)) = P(\{H\}) = p$$

■

If you have the feeling that we just went through alot of trouble to get nowhere, for the sake of this example, you are right. In this example and in many random variables we are aquainted with, it is easier to think in terms of $S = \mathcal{R}$, then we simply think of 1 itself as having mass p and 0 having mass q.

Density function To complete the discussion, we define the density function for the random variable.

In the discrete case, the density is defined on $\mathcal{R}$ the image of S under X,

$$p(x) = \mathbb{P}(X^{-1}(x)) = \sum_{s:X(s)=x} \mathbb{P}(s)$$

We can define the cumulative distribution,

$$F_X(x) = \mathbb{P}(X^{-1}(-\infty, x]).$$

If the variable is continuously distributed and F_X is differentiable then,

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Example: (Bernoulli Trial - part 2)

■

We can define the cumulative and density distribution of X. The density, $f_X(0) = q$ and $f_X(1) = p$. The cummulative distribution,

$$F_X(x) = \begin{cases} 0 & x < 0, \\ q & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

■

1.2 Why worry about Ω?

Consider a series of random variables $X_1, X_2, ..$ which are, say, the value of a given stock each day. Then X_{i+1} depends in some way on X_i.

At each time step some new information is uncovered. Thus at time t, the events that are of the form $\{X_0 = x_0, .., X_t = x_t\}$ is zero or one. Then at time t we condition on such an event to find the probability of an event of the form $\{X_0 = x_0, .., X_t = x_t, X_{t+1} = x_{t+1}\}$

Thus we think of $\Omega(= \Omega_{t+1})$ at time $t + 1$ to be larger (ie contain more information) than the $\Omega(= \Omega_t)$ at time t.

We see therefore that we need many Ω, one for each time step. We will develop these ideas with examples later in the course.

1.3 Important examples of random variables

Here is a list, please refer to your favorite probability text for details.

- Discrete/continuous uniform

- Bernoulli trial
- Binomial
- Geometric
- Normal/Gaussian
- Poisson
- Exponential

1.4 Expectation

The most important quantity to measure a random variable is the expectation. We'll define this in the case of $\mathcal{R} \subset \mathbb{R}$.

Lets use the probability space definition to define the expectation,

$$\mathbb{E}(X) = \sum_{s \in S} X(s)\mathbb{P}(s).$$

On the other hand, if we use a density distribution,

$$\mathbb{E}(X) = \sum_{x \in \mathcal{R}} xp(x)$$

In the continuous case:

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x) dx$$

Example: (Continuous 'triangular' distribution)

■

Consider X given by density $f_X(x) = 2x$ for $0 \leq x \leq 1$. Find the expectation,

$$\mathbb{E}(X) = \int_0^1 x f_X(x) dx = \int_0^1 x(2x) dx = \frac{2x^3}{3}\Big|_0^1 = \frac{2}{3}$$

■

1.4.1 Tail sum formula

If X is a nonnegative random variable with cummulative distribution F_X then

$$F_X(x) = 0$$

for all $x < 0$.

If X is continuously distributed then $f_X(x) = 0$ for all $x < 0$.

If X is a non negative random variable we can calculate the expectation with the following formula,

$$\mathbb{E}(X) = \int_{t=0}^{\infty} \mathbb{P}(X > t) dt = \int_{t=0}^{\infty} [1 - F_X(t)] dt$$

This follows in the case that X is continously distributed from,

$$\mathbb{E}(X) = \int_{t=0}^{\infty} [1 - F_X(t)] dt = \int_{t=0}^{\infty} \int_{x=t}^{\infty} f_X(x) dx dt = \int_{x=0}^{\infty} \int_{t=0}^{x} f_X(x) dt dx = \int_{x=0}^{\infty} x f_X(x) dx$$

This formula is valid for discrete random variables as well.

Example: (Geometric distribution)

■

Suppose $p + q = 1$ and $\mathbb{P}(X = k) = q^{k-1}p$. So

$$\mathbb{P}(X \leq k) = p + qp + ... + q^{k-1}p = \frac{1 - q^k}{1 - q}p = 1 - q^k$$

So $\mathbb{P}(X > k) = q^k$ Now we have,

$$\sum_{k=0}^{\infty} \mathbb{P}(X > k) = 1 + q + q^2 + \cdots = \frac{1}{1 - q} = \frac{1}{p}.$$

■

2 Random Variables with joint distribution

Suppose Z takes on values in $\mathbb{R}^2$, then we can write $Z = (X, Y)$ where there is some joint density function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}^+$ where $\mathbb{R}^+ = [0, \infty)$. Then for $A \subset \mathbb{R}^2$, we have

$$\mathbb{P}(Z \in A) = \int \int_{(x,y)\in A} f_{X,Y}(x, y)$$

Example: (Uniform distribution on a triangle)

■

Consider $Z = (X, Y) \sim \mathcal{U}(A)$ where A is the triangle $A = \{(x, y) : 0 \leq y \leq x \leq 1\}$ then Z has the density $f_Z(z) = 2\chi_{z\in A}$. Equivalently, (X, Y) has joint density $f_{X,Y}(x, y) = 2\chi_{(x,y)\in A}$ ■

Note here we use notation $\chi_{P(z)} = 1$ when condition $P(z)$ is true and $\chi_{P(z)} = 0$ otherwise.

We can define cumulative density as usual $F_{X,Y}(x, y) = \mathbb{P}(X \leq x;\ Y \leq y)$.
If the density $f_{X,Y}$ exists then,

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v)dudv$$

so

$$\frac{d^2}{dxdy}F_{X,Y} = f_{X,Y}$$

Example: (Two day stock price)

■

Suppose the value of a stock on the first day X_1 is uniformly distributed on $[1, 2]$, suppose the value on the second day X_1 is uniformly distributed over an interval of lenght one centered at the value of the first day.

We can write independent variables $U_i \sim \mathcal{U}[0, 1]$ for $i = 1, 2$. Then $X_1 = 1 + U_1$, and $X_2 = X_1 - 1/2 + U_2$.

The shape of this object in $\mathbb{R}^2$ is a parallelogram with vertices $(1, 0.5), (1, 1.5), (2, 1.5)$ and $(2, 2.5)$. $f_{X_1,X_2}(x_1, x_2) = 1$ for (x_1, x_2) on the interior of the parallelogram and 0 otherwise. ■

Notice X_2 however is distributed from 0.5 to 2.5. Is it uniformly distributed?

2.1 Marginals

The formula for marginal density of bivariate discrete random variable:

$$f_X(x) = \sum_{y:(x,y)\in S} \mathbb{P}[(X,Y) = (x,y)]$$

similarly for the continuous case, where $f_{X,Y}$ is joint density

$$f_X(x) = \int_{y\in\mathbb{R}} f_{X,Y}(x,y)dy$$

2.2 Sums

We wish to consider the special examples of sums of random variables.

Example: (Two day stock price – part 2)

■

We continue the example above. Let us find the density of X_2.

$$f_{X_2}(x_2) = \int_{y\in\mathbb{R}} f_{X_1,X_2}(y,x_2)dy = \begin{cases} 0 & 2.5 < x_2 \\ -x_2 + 2.5 & 1.5 < x_2 \leq 2.5 \\ x_2 - 0.5 & 0.5 < x_2 \leq 1.5 \\ 0 & x_2 \leq 0.5 \end{cases}$$

■

Let us consider the simple example of the sum of a die roll.

Example: (Sum of die roll)

■

Suppose $X, Y \sim \mathcal{U}\{1,2,..,6\}$ that is, both are the outcomes of a dice roll. Let $W = X + Y$, what is the distribution of W ?

We have

$$\mathbb{P}(W=2) = \frac{1}{36},\ ...,\mathbb{P}(W=7) = \frac{6}{36},\ ...,\mathbb{P}(W=12) = \frac{1}{36}.$$

■

Now consider a continuous example,

Example: (Sum of coordinates in a triangle)

■

Suppose $(X,Y) \sim \mathcal{U}(A)$, where $A = \{(x,y) : 0 \leq y \leq x \leq 1\}$. What is the density?

The cummulative can be found by simply considering the area of the subset, we leave the reader to verify,

$$\mathbb{P}(W \leq w) = \begin{cases} w^2/2 : & 0 \leq w \leq 1 \\ 1 - (2-w)^2/2 : & 1 \leq w \leq 2 \end{cases}$$

so the density is

$$f_W(w) = \begin{cases} w : & 0 \leq w \leq 1 \\ 2 - w : & 1 \leq w \leq 2 \end{cases}$$

■

Notice these density distributions are 'tent-shaped', the moral of these examples for now is that sums of random variables tend to concentrate.

2.3 Independent Random Variables

Variables X and Y on $\mathbb{R}$ are independent if, for all $x, y \in \mathbb{R}$ we have

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x; Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y) \leq F_X(x)F_Y(y)$$

If F is differentiable and we can define the density it then follow that the density is multiplicative, ie,

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

A corollary of this representation is that for X, Y independent random variables we have

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

3 Derived random Variables

Given random variable X, define probability in the case of $Y = g(X)$.

The cumulative here is defined as

$$F_Y(y) = \int_{x:g(x)\leq y} f_X(x)dx$$

Suppose for every y an interval of $(y-\epsilon, y+\epsilon)$ exists so that there are local functions (x_i) so that $g(x_i(y')) = y'$ for $y' \in (y-\epsilon, y+\epsilon)$. Then the derivative of the cummulative density,

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \sum_i f_X(x_i(y)) \left| \frac{d}{dy}[x_i(y)] \right|$$

Example: (Square of uniform)

■

Let $X \sim \mathcal{U}([-2,2])$ and $Y = X^2$. Calculate the CDF and PDF.

Then for $0 \leq y \leq 4$,

$$F_Y(y) = \mathbb{P}(X : X^2 \leq y) = \int_{-\sqrt{y}}^{\sqrt{y}} (1/4)dx = \frac{2\sqrt{y}}{4}.$$

Let $x_\pm(y) = \pm\sqrt{y}$, then

$$f_Y(y) = f_X(+\sqrt{y})\,|(+\sqrt{y})'| + f_X(-\sqrt{y})\,|(-\sqrt{y})'| = \frac{1}{4\sqrt{y}}$$

for $0 < y < 4$, and 0 otherwise. ■

3.1 Moments

The moments of the random variables are the expectations of powers of the random variables The k^{th} moment of X is

$$\mathbb{E}(X^k) = \int_{\mathcal{R}} x^k f_X(x)dx$$

The k^{th} *central* moment is

$$\mathbb{E}([X - \mathbb{E}(X)]^k) = \sum_i (-1)^i \binom{k}{i} \mathbb{E}(X^i)[\mathbb{E}(X)]^{k-i}$$

3.1.1 Variance

Of course the most important central moment is the second. It is known as the variance,

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

The square root of the variance is the Standard deviation

$$\sigma_X = \sqrt{var(X)}$$

Recall these helpfull properties of variance:

1 For $a, b \in \mathbb{R}$, we have $var(aX + b) = a^2 var(X)$.

2 For X_1, X_2 independent random variables $var(X_1 + X_2) = var(X_1) + var(X_2)$

3 For X_1, X_2 independent random variables $var(X_1X_2) = var(X_1)var(X_2) - \mu_{X_1}^2 var(X_2) - \mu_{X_2}^2 var(X_1)$. As a corrolary, if $\mu_{X_1} = \mu_{X_2} = 0$ then $var(X_1X_2) = var(X_1)var(X_2)$.

Example: (Two day stock price)

■

Let us find the variance of X_1 and X_2. Define $U_i \sim \mathcal{U}[0, 1]$ for $i = 1, 2$ - the U_i are uniformly distributed between 0 and 1. Here $X_1 = 1 + U_1$ and $X_2 = X_1 - \frac{1}{2} + U_2 = \frac{1}{2} + U_1 + U_2$.

$$var(X_1) = var(U_1) = 1/12$$

having used property 1 and *that for* $U \sim \mathcal{U}[a, b]$ *that* $var(U) = [b - a]/12$.
On the other hand the variance for X_2 is,

$$var(X_2) = var(\frac{1}{2} + U_1 + U_2) = var(U_1) + var(U_2) = 1/6.$$

having used property 1 and 2. ■

4 Covariance

Suppose X and Y have a joint CDF F_{XY}, then

$$cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

for $\mu_X = \mathbb{E}X$ and $\mu_Y = \mathbb{E}Y$.

We can easily derive $cov(X, Y) = \mathbb{E}(XY) - \mu_X\mu_Y$. Notice as well $cov(X, Y) = cov(Y, X)$ and $var(X) = cov(X, X)$.

Define the covariance matrix

$$\Sigma = \Sigma_{X,Y} = \begin{pmatrix} var(X) & cov(X, Y) \\ cov(X, Y) & var(Y) \end{pmatrix}$$

Let $Z = aX + bY$ then (*varify this*)

$$var(Z) = cov(aX + bY, aX + bY) = a^2 varX + 2ab\ cov(X, Y) + b^2 var(Y) = \begin{pmatrix} a & b \end{pmatrix} \Sigma \begin{pmatrix} a \\ b \end{pmatrix}$$

In particular, for $Z = X + Y$ *we have* $var(Z) = var(X) + 2cov(X, Y) + var(Y)$

That is Σ has the property that for all $v \in \mathbb{R}^2$ we have $v^T \Sigma v \geq 0$. This property is known as positive semidefinite.

Observe Σ is real and symmetric so that $\Sigma = ODO^T$ where the matrices O are orthogonal and D is diagonal - with entries being the eigenvalues of Σ. As Σ is positve definite, the eigenvalues are nonnegative.

Of course this entire discussion generalizes to a set of n jointly distributed random variables $X_1, ..., X_n$.

Correlation The covariance of two random variables may be positive or negative. Random variables which have positive covariance have the property that as one increases then *on average* the second increases.

On the otherhand if the covariance is negative then as one increases on average the second decreases.

It is usefull to normalize the covariance to a number ρ between -1 and 1. Where $|\rho| = 1$ indicates 'perfect' correlation between the two random variables. Let

$$\rho = \rho_{XY} = corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Example: (Two day stock price – part 3)

■

We continue the example above. Let us find the covariance matrix and the correlation of X_1 and X_2.

First find the covariance of X_1 and X_2. Find the expectation of the product,

$$\begin{aligned}
\mathbb{E}(X_1 X_2) &= \mathbb{E}((1 + U_1)(1/2 + U_1 + U_2)) \\
&= \mathbb{E}\{(3/2 + [U_1 - 1/2])(3/2 + [U_1 - 1/2] + [U_2 - 1/2])\} \\
&= (3/2)^2 + (3/2)\mathbb{E}([U_1 - 1/2] + [U_2 - 1/2]) + (3/2)\mathbb{E}([U_1 - 1/2]) + \\
&+ \mathbb{E}((U_1 - 1/2)^2) + \mathbb{E}(U_1 - 1/2)(U_2 - 1/2) \\
&= (3/2)^2 + (1/12) + 0 = (3/2)^2 + (1/12)
\end{aligned}$$

But $\mathbb{E}(X_1)\mathbb{E}(X_2) = (3/2)^2$ so

$$cov(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) = 1/12$$

■

4.1 Gaussians

One dimensional Gaussians $X \sim N(\mu, \sigma^2)$ have the PDF

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

Recall Gaussians are stable, that is for $X_i \sim N(\mu_i, \sigma_i^2)$ we have $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

For any covariance matrix Σ (that is a positive definite real symmetric matrix) in $\mathbb{R}^{n \times n}$ (real $n \times n$ matrices) and vector $\overline{\mu} \in \mathbb{R}^n$ we can define a Multivariate Gaussian distribution X with PDF

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}(x-\overline{\mu})^T \Sigma^{-1} (x-\overline{\mu})}$$

Notice if Σ is the identity matrix we find X_1, X_2 normal random variables which are not independent but the correlation is zero.

5 Conditionals

Given a probability experiment, one often would like to *condition* on some partial knowledge of the outcome.

Example: (Die roll)

■

Suppose we roll two die, one red and one green. The sample space is the set of ordered pairs $S = \{(i,j) : i,j \in \{1,..,6\}\}$. Consider the probability that one of the faces shows a 1.

$$\mathbb{P}(\text{At least one dice shows 1}) = 1 - \mathbb{P}(\text{Neither dice shows a 1}) = 1 - (5/6)^2 = 11/36$$

But if we have partial knowledge of the outcome the probability may change, suppose we know the sum is 4.

$$\mathbb{P}(\text{At least one dice shows 1}|\text{The sum of the die is 4}) = 2/3$$

This can be seen by deduction, by considering the three possible outcomes where the sum of the roll is 4: $\{(1,3),(2,2),(3,1)\}$, two of three of these have a face with 1 showing. ■

5.1 Conditional probabilities

We will define the conditional probabilities, recall Bayes rule: let $A, B \subset S$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{5.1}$$

this gives the probability of event A *given* event B.

Example: (Die roll 2)

■

Suppose we roll two die, one red and one green. The sample space is the set of ordered pairs $S = \{(i,j) : i,j \in \{1,..,6\}\}$. Let $X(i,j) = i + j$ the sum of the faces of the die. Suppose we know that the sum is greater than 7, ie $X > 7$. What is the probabilty X is greater than or equal to 10?

Let $A = \{X \geq 10\}$ and $B = \{X > 7\}$.

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) = 6/36.$$

on the other hand $\mathbb{P}(B) = 15/36$ so

$$\mathbb{P}(A|B) = \frac{6/36}{15/36} = \frac{2}{5}$$

Of course, we notice $\frac{2}{5} \neq \frac{6}{36}$ ■

Example: (Die roll 3)

■

Let us continue the previous example. Let C be the event that the green dice is even. Find

$$\mathbb{P}(A|C)$$

Notice the elements of $A \cap C$ have the green dice is either 4 or 6. If the green dice is 4 the red dice is 6. If the green dice is 6 the red dice is 4,5, or 6.

$$\mathbb{P}(A \cap C) = 4/36 = 1/9$$

Of course $\mathbb{P}(C) = 1/2$. So

$$\mathbb{P}(A|C) = \frac{1/9}{1/2} = \frac{2}{9}$$

■

Finally let us return to the stock price example.

Example: (Two day stock price – part 4)

■

We continue the example above Suppose the event B is that the value of the stock on the second day is 5/3 ie $B = \{X_2 = 5/3\}$. Notice $\mathbb{P}(B) = 0$ can we condition on it?!

Of course, if we think in terms of our usual notion of taking things in calculus we can define and then take limits.

That is let $B_\epsilon = \{|X_2 - 5/3| < \epsilon\}$, and define for an event A, $\mathbb{P}(A|B) = \lim_{\epsilon \to 0} \mathbb{P}(A|B_\epsilon)$.

Let A be the event that $X_1 < 3/2$. Show

$$\mathbb{P}(A|B) = 2/5$$

■

5.2 Conditional Random Variable

We condition the random variable Y on X with joint PDF $f_{X,Y}$ with the function,

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

where f_X is the marginal of X defined in Section 2.1.

In the discrete setting the conditional distribution is

$$P_Y(y|X = x) = \frac{P_{X,Y}(x, y)}{P_X(x)}$$

where P_X is again the marginal of X.

Example: (Two day stock price – part 5)

■

Again let B be the value of the stock on the second day is 5/3 ie $B = \{X_2 = 5/3\}$. Show that $f_{X_2}(x) = 6/5$ for $x \in (7/6, 2)$ and zero otherwise. ■

5.3 Conditional expectation

As the name indicates this is expectation conditioned on some function on the probability space. Formally, for jointly distributed random variables X and Y,

$$\mathbf{E}(Y|X) = \sum_y \mathbb{P}(Y = y|X).$$

Notice this is a function of X. As a matter of fact:

$$\mathbb{E}(Y) = \mathbb{E}_X[\mathbf{E}(Y|X)]$$

where $\mathbb{E}_X$ indicates taking expectation with respect to X.

Example: (unfair coin flips)

■

Suppose 3 unfair coins are flipped. The outcome of each flip is $F_i \in \{h, t\}$, $i = 1, 2, 3$. Let p be the probability of a head, $\mathbb{P}(F_i = h) = p$. Let X map the coin flip to $\{0, 1\}$ ie $X(h) = 1$ and $X(t) = 0$, define $X_i = X(F_i)$ The total number of heads is H,

$$H = X_1 + X_2 + X_3.$$

Have we seen H before? Is it one of the 'Important examples of random variables'?

Find $\mathbf{E}(H|X_1)$ and $\mathbf{E}(X_1|H)$.

First we find,

$$\mathbf{E}(H|X_1) = 0\mathbb{P}(H = 0|X_1) + 1\mathbb{P}(H = 1|X_1) + 2\mathbb{P}(H = 2|X_1) = 3\mathbb{P}(H = 3|X_1)$$

consider fixing X_1,

$$\mathbb{E}(H|X_1 = 0) = 2pq + 2p^2 = 2p$$

and

$$\mathbb{E}(H|X_1 = 1) = 1 \cdot q^2 + 2 \cdot 2pq + 3 \cdot p^2 = 1 + 2pq + 2p^2 = 1 + 2p$$

consider making this a function of X_1 (note above each are of the form of the value of X_1 plus the expectation of $X_2 + X_3$)

$$\begin{aligned} \mathbf{E}(H|X_1) &= \mathbf{E}[X_1 + X_2 + X_3|X_1] \\ &= X_1 + \mathbb{E}[X_2 + X_3] \\ &= X_1 + 2p \end{aligned}$$

notice this is a function of X_1 only - everything else has been integrated. On the other hand,

$$\begin{aligned} \mathbb{E}(X_1|H = 0) &= \mathbb{P}(X_1 = 1|H = 2) = 0 \\ \mathbb{E}(X_1|H = 1) &= \mathbb{P}(X_1 = 1|H = 1) = 1/3 \\ \mathbb{E}(X_1|H = 2) &= \mathbb{P}(X_1 = 1|H = 2) = 2/3 \\ \mathbb{E}(X_1|H = 3) &= \mathbb{P}(X_1 = 1|H = 3) = 1 \end{aligned}$$

clearly a good function is

$$\mathbf{E}(X_1|H) = H/3.$$

An alternative derivation is due to each coin having equal probability of turning up heads, so any of the heads has equal probability to be the first coin. ■

The first part of the example illustrates that

$$\mathbf{E}(f(X) + g(Y)|X) = f(X) + \mathbf{E}(g(Y)|X).$$

Similar:

$$\mathbf{E}(f(X)g(Y)|X) = f(X)\mathbf{E}(g(Y)|X)$$

And if Y is independent of X, $\mathbf{E}(g(Y)|X) = \mathbb{E}(g(Y))$ which is a number - no longer a function.

The above can be generalized to the case conditioning on several random variables.

Example: (biased coin flips)

■

Again suppose $\mathbb{P}(X_i = h) = p$, and $\mathbb{P}(X_i = t) = 1-p = q$. If $H = X_1 + X_2 + X_3$. Then the conditional expectation is

$$\mathbf{E}(H|X_1, X_2) = X_1 + X_2 + p.$$

or

$$\mathbf{E}(X_1|H) = H/3$$

Now suppose $H_m = X_1 + \cdots + X_m$ so for $m < n$ we have

$$\mathbf{E}(H_m|H_n) = \frac{m}{n} H_n$$

on the other hand,

$$\mathbf{E}(H_n|H_m) = H_m + \mathbb{E}(X_{m+1} + \cdots + X_n) = H_n + p(n-m).$$

■

In the above example we can let $\mathcal{F}_i$ contain 'all the information obtained from the first i coins.' This is just notation, so we will write,

$$\mathbf{E}(H_n|\mathcal{F}_m) \equiv \mathbf{E}(H_n|X_1, \cdots, X_m).$$

Example: (Sum of i.i.d.)

■

Let X_i for $i = 1, 2, \ldots$ be iid random variables $X_i \sim X$ with $\mathbb{E}(X) = \mu = 0$ and $Var(X) = \mathbb{E}(X^2) = \sigma^2$. Let $m < n$,

$$\begin{aligned}
\mathbf{E}[S_n^2|\mathcal{F}_m] &= \mathbf{E}[((S_n - S_m) + S_m)^2|\mathcal{F}_m] \\
&= \mathbf{E}[(S_n - S_m)^2 + 2(S_n - S_m)S_m + S_m^2|\mathcal{F}_m] \\
&= \mathbb{E}[(S_n - S_m)^2] + 2S_m\mathbb{E}[S_n - S_m] + S_m^2 \\
&= (n-m)\sigma^2 + S_m^2.
\end{aligned}$$

The third equality follows because $S_n - S_m$ is independent of $\mathcal{F}_m$ the information from the first n variables; the last equality follows because $\mu = 0$ and $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] = \mu^2 = 0$. ■

6 Asymptotic behavior

Let $(T_i)_{i\geq 1}$ be i.i.d. random variables, with distribution $F(x) = \mathbb{P}(T_i \leq x)$. We assume for all $i = 0, 1, 2, ...$ that $\mathbb{E}(T_i) < \infty$ and define $\mu := \mathbb{E}(T_1)$.

Law of large numbers (LLN) The law of large numbers states, for T_i that with probability 1 the average converges to the mean i.e.

$$\lim_{N\to\infty} \sum_{i=1}^{N} \frac{T_i}{N} \to \mu.$$

Notice the statement 'with probability 1', there are conceivable other sequences not averaging to the mean but they total zero in probability - like flipping infinite heads in a row. An easier to understand, but weaker, statement is

$$\mathbb{P}\left(\left|\frac{T_1+\cdots T_N}{N}-\mu\right|>\epsilon\right)\to 0 \text{ as } N\to\infty.$$

The moral is $T_1+\cdots+T_N\sim\mu N$, if we set $t=\mu N$ then $N=t/\mu$ so that

$$T_1+\cdots+T_{t/\mu}\sim t$$

Central Limit Theorem (CLT) We assume Var $(T_1)=\sigma^2$. The central limit theorem states, for $N(0,1)$ a normal variable with mean 0 and variance 1,

$$\frac{T_1+\cdots+T_N-N\mu}{\sigma\sqrt{N}}\to N(0,1)$$

by this we mean that

$$\mathbb{P}\left(\frac{T_1+\cdots+T_N-N\mu}{\sigma\sqrt{N}}\leq x\right)\to\Phi(x).$$

Where Φ is the cumulative distribution of a $N(0,1)$,

$$\Phi(x)=\int_{-\infty}^{x}e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}.$$

We write down the sum, (here $Z\sim N(0,1)$),

$$T_1+\cdots+T_N=S_N\sim N\mu+Z\sigma\sqrt{N}$$

Central Limit Theorem - for multivariate random variables (CLT) Suppose X_i are iid Random variables in $\mathbb{R}^d$ we may write

$$X_i=\begin{pmatrix}X_{i(1)}\\ \vdots\\ X_{i)(d)}\end{pmatrix}.$$

Suppose $\mathbb{E}(X_i)=\mu\in\mathbb{R}^d$ and $\Sigma\in\mathbb{R}^{d\times d}$ is the covariance matrix $\Sigma_{j,k}=cov(X_{i(j)},X_{i(k)})$.

Let us suppose that $det\Sigma\neq 0$ *(this is equivalent to saying that each $X_{i(j)}$ has some randomness not contained in the other $X_{i(k)}$ - that is one of the $X_{i(j)}$ is not a function of the other $X_{i(j)}$.)*

Then a sum of the X_i properly normalized approaches a Gaussian random variable. Let $S_n=X_1+\cdots+X_n$ That is for $A\in\mathbb{R}^d$ we have

$$\lim_{n\to infty}\mathbb{P}\left(\frac{S_n-n\mu}{\sqrt{n}}\in A\right)=\int\cdots\int_{x\in A}e^{-(x-\mu)^T\Sigma^{-1}(x-\mu)}\frac{dx_1\cdots dx_d}{\sqrt{(2\pi)^d\det\Sigma}}.$$

As a short hand we write $\frac{1}{\sqrt{n}}(S_n-n\mu)\xrightarrow{\mathcal{D}}Z$ where $Z\sim N(0,\Sigma)$.