Academic notes: 1A Probability

J. Mak (September 5, 2015) [From notes of I. M. MacPhee, Durham][*]

I. THE CLASSICAL MODEL AND AN AXIOMATIC FORMULATION

A. The classical model

Suppose we are dealing with events that occur producing an outcome that we cannot predict. We can potentially list all the possible outcomes and, assuming all outcomes are equally likely, work out the likelihood of the outcome(s) occurring. For example, we may have:

- drawing a card (or cards) from a shuffled deck;
- flipping a coin some number of times with outcomes H or T;
- rolling a dice.

With m possible outcomes $\Omega = \{\omega_1, \omega_2, \dots \omega_m\}$, an event is a subset $A \subseteq \Omega$. Then we may assign a probability $P(A)$ on the likelihood of the event occurring.

Suppose we make k successive selections where each event is independent of each other. If m_i is the number of possibilities for $i \in \mathbb{N}$, then the total number of permutations for k objects is

$$m_1 \times m_2 \times \cdots m_k = \prod_{i=1}^{k} m_i.$$

If, instead we have a set of m objects and we select $r \leq m$ in order without replacement (i.e., ordering matters), then the number of possible selections of r objects is

$$m(m-1)(m-2)\cdots(m-r+1) = \frac{m!}{(m-r)!}.$$

Example Work out the following:

1. The probability of being dealt a diamond royal flush ($\diamondsuit$10, J, Q, K and A in any order).

   The number of possible ways to choose five cards from 52 is $52!/(52-5)! = 52!/47!$. Thus we have

   $$P(\diamondsuit 10 \text{ to A}) = \frac{5!/0!}{52!/47!} = \frac{5!47!}{52!}.$$

2. The number of combinations for four digit PINs, where valid PINs are not allowed to have the same four digits (e.g., 1111), or the digits being consecutive increasing or decreasing (e.g., 0123, 9876, but 7890 is ok).

   There are 10^4 possible sequences. The combination $dddd$ occurs $10 \cdot 1 \cdot 1 \cdot 1 = 10$ times, while consecutive sequences occurs $2(7 \cdot 1 \cdot 1 \cdot 1) = 14$ times, so we have $10000 - (10 + 14) = 9976$ possible admissible combinations.

3. With N people in the same room, what is the change that two (or more) people have a common birthday?

   With N people, the number of possibilities for birthdays is (discounting 29th Feb) 365^N. Let $\overline{B}$ be the even that no one has the same birthday, i.e., $365 \cdot 364 \cdots (365 - (N-1))$, and so

   $$P(\overline{B}) = \frac{365 \cdot 364 \cdots (365 - (N-1))}{365^N} = 1\left(1 - \frac{1}{365}\right)\cdots\left(1 - \frac{N-1}{365}\right) \qquad \Rightarrow \qquad P(B) = 1 - P(\overline{B}).$$

[*] julian.c.l.mak@googlemail.com

Let $\alpha = 1/365$. We know that $1 - x < \mathrm{e}^{-x}$ for all x values, so

$$P(\overline{B}) < 1 \cdot \mathrm{e}^{-\alpha}\mathrm{e}^{-2\alpha} \cdots \mathrm{e}^{-(N-1)\alpha} = \exp\left[-\sum_{j=1}^{N-1}(j\alpha)\right] = \exp\left[-\frac{N(N-1)\alpha}{2}\right].$$

We then have the following values:

N	$\exp(-N(N-1)\alpha/2)$	$1-\exp(-N(N-1)\alpha/2)$
5	0.9730	0.0270
10	0.8840	0.1160
15	0.7500	0.2500
20	0.5942	0.4058
23	0.5000	0.5000
42	0.0945	0.9055

So with $N = 23$, the probability of two (or more) people having the same birthday is around more than a half, whilst with $N = 42$, the probability of two (or more) people having the same birthday is around $9/10$.

Sometimes we are interested in cases where the order of the events is not important. In general, if we are choosing r objects from m , the number of combinations are

$$\frac{m!}{(m-r)!r!} = \binom{m}{r} = C^m_r.$$

Note that $C^m_r = C^m_{m-r}$.

Example Work out the probability of getting five cards such that four are of the same suit.

The amount of possibilities are C^{52}_5, of which $4C^{13}_4 \times 3C^{13}_1$ choices are the ones we want (choose four that are of the same suit, with four suits to choose from, and choose one from the remaining three suits). So $P(A) = 12C^{13}_4 C^{13}_1 / C^{52}_5 \approx 0.0429$.

The number of ways to arrange m objects into k groups with sizes $r_1, r_2 \ldots r_k$, where $\sum_{i=1}^k r_i = m$ is

$$\frac{m!}{r_1! r_2! \cdots r_k!} = \binom{m}{r_1, r_2, \ldots r_k}.$$

B. Axiomatic formulation

We generalise the classical model, and as before the trial takes place and its outcome is unpredictable. We say the set of all possible outcomes is the sample space, denotes Ω. The event that happens is a subset $A \subseteq \Omega$.

We observe that:

- the set with no outcome is the empty set $\emptyset$;
- A or B is $A \cup B$;
- A and B is $A \cap B$;
- not A is $A' \equiv \Omega \setminus A$;
- A but not B is $A \setminus B \equiv A \cap B'$.

We have $A \subseteq B$ when $A \cap B = A$. We say A and B are incompatible if $A \cap B = \emptyset$, i.e., A and B cannot happen both at once.

Example For a deck of cards, drawing an eight is $E = \{\spadesuit 8, \heartsuit 8, \diamondsuit 8, \clubsuit 8\}$, and drawing a spade is $S = \{\spadesuit A, \ldots \spadesuit K\}$. Then, drawing an eight and a spade is $E \cap S = \{\spadesuit 8\}$, while drawing an eight and not a spade is $E \setminus S = \{\heartsuit 8, \diamondsuit 8, \clubsuit 8\}$.

A probability distribution on Ω is a collection of numbers $P(A)$ defined for each event $A \subseteq \Omega$ obeying the following axioms:

1. $P(A) \geq 0$ for all A;
2. $P(\Omega) = 1$;
3. $A \cap B = \emptyset$ implies that $P(A \cup B) = P(A) + P(B)$;
4. $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ iff, for all i and j, $A_i \cap A_j = \emptyset$.

(This does not apply to uncountable collections of incompatible events.) We can consider the axioms as similar to the unit length, unit area etc., where the length/area represents the probability.

Proposition I.1 *As a consequence of the axioms, we have the following:*

- *for all A and B, $P(A \setminus A) = P(B) - P(A \cap B)$;*
- *for all A, $P(A') = 1 - P(A)$;*
- $P(\emptyset) = 0$*;*
- *for all A, $P(A) \leq 1$;*
- *if $A \subseteq B$, then $P(A) \leq P(B)$;*
- *for all A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;*
- *for all i and j, if $A_i \cap A_j = \emptyset$ and $i \neq j$, we have $P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i)$;*
- *for not necessarily disjoint sets, we have $P\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P(A_i)$.*

Example Consider the following examples.

1. Jimmy's dice has values $\{2, 2, 2, 2, 5, 5\}$, while yours has values $\{1, 1, 4, 4, 4, 4\}$. Rolling the dice once and the highest value wins. Let J be the event that Jimmy winds, what is $P(J)$?

   Let F be the even that Jimmy rolls a 5, and A is when you roll a 1. We see that $J = A \cup F$, and that $P(A) = P(F) = 1/3$, and that $P(A \cap F) = (1/3)(1/3) = 1/9$, so $P(J) = 1/3 + 1/3 - 1/9 = 5/9$.

2. In an indefinitely long sequence of coin tosses, let A_i be the first H on the i^{th} coin toss. We have $P(A_i) = 2^{n-i}/2^n = 2^{-i}$ (all the preceding tosses are T).

   Let H^* be the even that we get H eventually, then $H^* = \bigcup_{i=1}^{\infty} A_i$. Noting that if a_i is a sequence of positive numbers, then $\sum_{i=1}^{\infty} a_i = \lim_{n \to \infty} \sum_{i=1}^{n} a_i$. With this, then

   $$P(H^*) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} 2^{-i} = \frac{1}{2}\left(1 + \frac{1}{2} + \frac{1}{4} + \dots\right) = \frac{1/2}{1 - 1/2} = 1.$$

   So it is certain that we will eventually get a H.

II. CONDITIONAL PROBABILITY

Probability quantifies out uncertainty about unpredictable events. Acquiring information denotes out uncertainty so we need rules for how probability changes given some information.

The conditional probability of event A given event B is

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

assuming $P(B) > 0$.

Example A family has two children, and all combinations of sexes are equally likely. Let G be the even both are girls, and A be the even that at least one is a girl, then

$$P(G|A) = \frac{1/4}{1/3} = \frac{1}{3}, \qquad P(A|G) = \frac{1/4}{1/4} = 1.$$

The first one is that, given at least one is a girl, what is the probability that both are girls. The second is obvious, that given both are girls, the probability that at least one of the is a girl should have probability 1.

For $P(E) > 0$, $P(A|E)$ follows the axioms of probability:

1. $P(A|E) \geq 0$;
2. $P(\Omega|E) = 1$;
3. $P(A \cup B|E) = P(A|E) + P(B|E)$ if $A \cap B = \emptyset$;
4. for $A_i \cap A_j = \emptyset$ for all i and j,
$$P\left(\bigcup_{i=1}^{\infty} A_i|E\right) = \sum_{i=1}^{\infty} P(A_i|E);$$
5. $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.

Theorem II.1 (Partition theorem) *Events $B_1, B_2, \dots B_n$ form a partition of Ω if $B_i \cap B_j = \emptyset$ for all i and j, and that $\bigcup_{i=1}^{\infty} B_i = \Omega$. For any event A, we have*

$$A = \bigcup_{i=1}^{n} (A \cap B_i), \qquad P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Example Consider the bloaty head syndrome which affects 1 in $1,000$. Let D be the event that a person tested has a disease, and T be the event that the test is positive. Suppose we know that $P(T|D) = 0.9$ and $P(T|D') = 0.03$. We are really interested in obtaining $P(D|T)$, so we observe that since $P(D|T)P(T) = P(T|D)P(D)$, and $P(T) = P(T|D)P(D) + P(T|D')P(D')$, we have

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{0.9 \times 0.001}{P(T|D)P(D) + P(T|D')P(D')} = \frac{0.9 \times 0.001}{0.9 \times 0.001 + 0.03 \times 0.999} \approx 0.03,$$

so about 3% of the people tested positive has the disease. If $P(D) = 0.2$, then $P(D|T) \approx 90\%$.

Theorem II.2 (Bayes' theorem)

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

Example Murder on the island. There are $n+1$ person on an island and a murdered person. Everybody is equally likely to have committed the murder. Somebody is arrested and their DNA matches the evidence at the scene. Let G by the even the person was actually guilty, and E be the event that the evidence of the given genotype was left. Let $P(E|G') = p$, i.e., the probability that evidence matches someone else (we know that $P(E|G|) = 1$). We want to find $P(G|E)$. To use Baye's theorem, we need $P(E)$; we see that

$$P(E) = P(E|G)P(G) + P(E|G')P(G') = 1 \times \frac{1}{n+1} + p \times \frac{n}{n+1} = \frac{1+np}{n+1}.$$

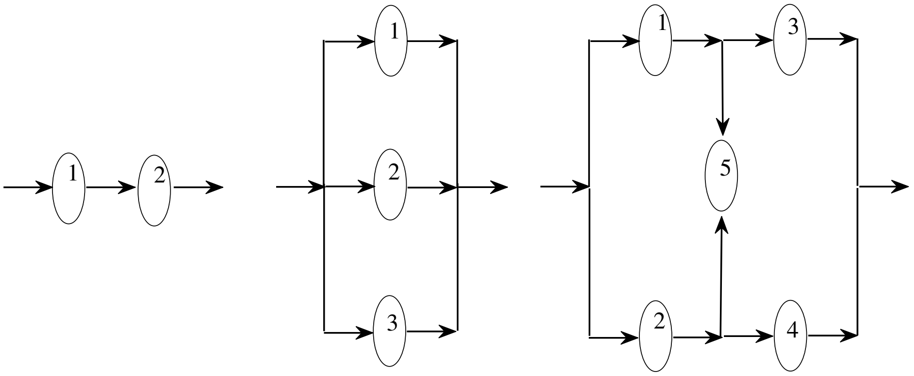
So then

$$P(G|E) = \frac{P(E|G)P(G)}{P(E)} = \frac{1 \times 1/(n+1)}{p(n+1)/(1+n)} = \frac{1}{1+np}.$$

We see that if p is small, i.e., the probability of someone else having similar DNA is small, then $P(G|E)$ is close to 1, and so the test will be reasonably accurate.

An event A is independent of B when $P(A|B) = P(A)$. We can also show independence occurs mutually, i.e., $P(B|A) = P(B)$.

Example Consider the following circuits:



The circuit works if we can enter on the left and leave on the right. Let ω_i be the event that component i works, $p_i = P(\omega_i)$, and S the even that the system works. Fin d$P(S)$ for the three circuits assuming failure of nodes are independent of each other.

- $S = \omega_1 \cap \omega_2$, so $P(S) = P(\omega_1|\omega_2)P(\omega_2)$. By independence, we have $P(S) = P(\omega_1)P(\omega_2) = p_1p_2$.
- $S = \omega_1 \cup \omega_2 \cup \omega_3$, so $S' = \omega_1' \cap \omega_2' \cap \omega_3'$, and $P(S') = P(\omega_1')P(\omega_2')P(\omega_3')$, hence $P(S) = 1-(1-p_1)(1-p_2)(1-p_3)$.
- Let $B_1 = \omega_1 \cap \omega_3$, $B_2 = \omega_2 \cap \omega_4$, and $B_3 = (\omega_1 \cap \omega_2' \cap \omega_3' \cap \omega_4 \cap \omega_5) \cup (\omega_1' \cap \omega_2' \cap \omega_3' \cap \omega_4' \cap \omega_5')$. $S = B_1 \cup B_2 \cup B_3$, and $B_3 \cap B_1 = B_3 \cap B_2 = \emptyset$, and $P(S) = P(B_1) + P(B_2) - P(B_1 \cap B_2) + P(B_3)$. We see all ω_i are incompatible events, so $P(B_3) = p_1(1-p_2)(1-p_3)p_4p_5 + (1-p_1)p_2p_3(1-p_4)p_5$, and hence

$$P(S) = p_1p_3 + p_2p_4 - p_1p_2p_3p_4 + P(B_3).$$

Hardy–Weinburg equilibrium In genetics, offspring get an allele for each gene on a standard chromosome from each parent randomly selected from the parents' two alleles and are independent. A particular result known as the Hardy–Weinburg equilibrium states that the ratio of alleles remains constant throughout generations, i.e., even dominant genes that confer advantages remain constant throughout generations. To see this, it is important to understand what happens to neutral alleles, those that do not confer advantages or disadvantages.

Suppose for some neutral gene, genotypes AA, Aa and aa occur in partitions u, $2v$ and w respectively ($u + 2v + w = 1$). Assume that people choose mates randomly with respect to this gene and consider the first generation of the offspring. Let AF, AM be the evens that a given child gets allele A from Father and Mother respectively. Let F_{AA}, F_{Aa}, F_{aa} be the evens the Father has the respective genotypes. Using this as a partition we get

$$P(AF) = P(AF|F_{AA})P(F_{AA}) + P(AF|F_{Aa})P(F_{Aa}) + P(AF|F_{aa})P(F_{aa}) = u + \frac{1}{2}2v + 0 = p.$$

$p = u + v$ denotes the proportion of A alleles and $q = 1 - p$ denotes the proportion of a alleles. The same calculation gives

$$P(AM) = p, \qquad P(aM) = P(aF) = q.$$

Then with C_{AA} the event that the child has genotype AA and etc., we have

$$P(C_{AA}) = P(AF \cap AM) = p^2, \qquad P(C_{aA}) = P((aF \cap AM) \cup (aM \cap AF)) = qp + pq = 2pq, \qquad P(C_{aa}) = q^2.$$

For a large population, the proportions with the various genotypes will be close to these probabilities.

Let $u_1 = p^2$, $2v_1 = 2pq$, $w_1 = q^2$ be the proportions in generation 1. Proportion of A alleles in generation 2 is

$$u_1 + v_1 = p^2 + pq = p(p+q) = p$$

In fact, this is true for a alleles, for all generations. Hence proportion of neutral alleles in a population remains constant.

Example The above scenario does not apply to sex-link genes such as X and Y chromosomes for example. Suppose in an example a gene has alleles A and a, and F_{aa} and $M_{?a}$ are unhealthy, i.e., men only need an a allele to be unhealthy. Suppose Jane is healthy, but has an unhealthy male cousin, but all other relatives including two brothers are healthy. Find the probability that Jane is of genotype Aa, assuming independence.

Let J be the event that Jane is of genotype Aa. The father is for sure AA since he is healthy; the mother could be either AA or Aa. Let M be the event that the mother is of genotype Aa, and M' the event that the mother is of genotype AA. Let B the event that both brothers are healthy. We want $P(J|B)$; we need $P(J \cap B)$ and $P(B)$. We have, by independence,

$$P(B) = P(B|M)P(M) + P(B|M')P(M') = \left(\frac{1}{2} \times \frac{1}{2}\right)\frac{1}{2} + (1 \times 1)\frac{1}{2} = \frac{5}{8},$$

and

$$P(J \cap B) = P(J \cap B|M)P(M) + P(J \cap B|M')P(M') = \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right)\frac{1}{2} + (0 \times 1 \times 1)\frac{1}{2} = \frac{1}{16},$$

so

$$P(J|B) = \frac{P(J \cap B)}{P(B)} = \frac{1/16}{10/16} = \frac{1}{10}.$$

III. RANDOM VARIABLES AND DISTRIBUTIONS

We can now calculate probability for events, a collection of outcomes for some unpredictable trial. We are usually interested in numerical values based on these outcomes.

For instance, if we throw two dice, there are 36 possible outcomes. Let

$$X(\omega) = \omega_1 + \omega_2,$$

where $\omega = (i, j)$ are the values of the respective dices. X here is a random variable. In general, a random variable is a rule for attaching numerical values to outcomes.

A. Discrete random variables

Suppose that $\Omega = \{\omega\}$ is countable, then we say X is a discrete random variable.

Example Throwing three random coins, we have

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Let X be the number of heads, then $X(1) = 1$ for $\omega \in \{TTH, THT, HTT\}$. We can use $X = 1$ to denote the even that one head shows up, so, in this case, $P(X = 1) = 3/8$. Similarly, $P(X \geq 2) = 1/2$.

For any random variable X, the function $p(x) = P(X = x)$ gives us probabilities for a particular partition of Ω. We call $p(x)$ the probability density function (pdf) of x.

We see that $(X = x) \cap (X = y) = \emptyset$ when $x \neq y$, so, by axiom,

$$P(x \in A) = \sum_{x \in A} p(x),$$

where A is a set of values. If X takes values in $x_1, x_2, \ldots x_n \ldots$, then, by axiom,

$$\sum_{i \geq 1} p(x_i) = 1.$$

B. Continuous random variables

A continuous random variable takes values in a continuous set (usually $\mathbb{R}$ or a subset of it), and has a probability distribution defined by a non-negative pdf f. We set

$$P(X \in A) = \int_A f(x)\,\mathrm{d}x, \qquad A \subseteq \Omega.$$

Generally, $P(a < X < b) = \int_a^b f(x)\,\mathrm{d}x$, and $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$.

Example Suppose we have a pdf

$$f(x) = \begin{cases} k(1+x), & -1 < x < 0, \\ k(2-x), & 0 \le x < 2, \end{cases}$$

and zero otherwise. Then

$$1 = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = k\int_{-1}^{0}(1+x)\,\mathrm{d}x + k\int_0^2 (2-x)\,\mathrm{d}x = \frac{5k}{2},$$

so $k = 2/5$.

C. Binomial distribution

Suppose there are only two outcomes possible, then, with n trials, there are 2^n outcomes. A particular outcome with x successes has $(n-x)$ failures associated with it. For $p = P(\text{success})$, the probability of this outcome is

$$p^x(1-p)^{n-x}.$$

Each trial is independent, the event $X = x$ contains many outcomes; in fact, C^n_x such outcomes. Hence

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x \in [0, n].$$

Example 105 people have bought tickets for a flight. They independently miss the flight with chance 0.04, so we say $X \sim \mathrm{B}(105, 0.04)$. Then

$$P(X=0) = 0.96^{105} \approx 0.01376, \qquad P(X=1) = \binom{105}{1}(0.04)(0.96)^{104} \approx 0.06018,$$

and so on. We also have

$$P(X \ge 3) = \sum_{x=3}^{105} P(x) = 1 - \sum_{x=0}^{2} P(x) \approx 0.79867.$$

D. Poisson distribution

Suppose events occur randomly at rate r over some time $(t, t+h)$, then

$$P(0\text{ events}) \approx 1 - rh, \qquad P(1\text{ event}) \approx rh, \qquad P(\ge 2\text{ events}) \le h^2,$$

and the number of events in distinct time intervals are independent. The distribution of the number of events, T, in a period of s time units has the Poisson distribution $T \sim \mathrm{Po}(\lambda)$, with parameter $\lambda = rs$. We can find $P(T = t)$ as a limit of binomial probabilities. Consider an interval of length $h = s/n$, where n is big and h is small. Let A_n be the even that all intervals have less than 1 event. Then

$$P(A_n) \ge (1-h^2)^n = \left(1 - \frac{sh}{n}\right)^n \to \mathrm{e}^{-sh}$$

as $n \to \infty$. Note that $\mathrm{e}^{-sh} \to 1$ as $h \to 0$.

For any fixed $t \in \mathbb{Z}^+$ and for $n > t$,

$$P(T = t) \approx P(X_n = t) = \binom{n}{t} \left(\frac{\lambda}{n}\right)^t \left(1 - \frac{\lambda}{n}\right)^{n-t} = \frac{n!}{n^t(n-t)!} \frac{\lambda^t}{t!} \left(1 - \frac{\lambda}{n}\right)^{n-t} = \frac{n^{\tilde{t}}}{n^t} \left(1 - \frac{\lambda}{n}\right)^{n-t} \frac{\lambda^t}{t!},$$

where

$$\frac{n^{\tilde{t}}}{n^t} = 1\left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{t-1}{n}\right) \to 1^t = 1$$

for $n \to \infty$. So then

$$P(T = t) \approx (X_n = t) \to \mathrm{e}^{-\lambda} \frac{\lambda^t}{t!} \qquad (t \in \mathbb{Z}^+)$$

as $n \to \infty$. This limit is an useful approximation for $P(X_n = t)$ with quite moderate n when λ is small. As a rule of thumb, with $p \leq 0.05$, we find that $n = 20$ gives a reasonable approximation of the Poisson distribution to the binomial distribution. With bigger n the approximation is even better.

Example From 1979 to 1989, 1103 Bristol postmen reported 245 dog biting incidents. 191 were bitten, 145 of those just once. Are the bites occurring at random or are some postmen more prone to dog bites?

Taking each incident as a trial, where each postman has a $1/1103$ chance of being attacked. The number of attacks X suffered by a single postman is $X \sim \mathrm{B}(245, 1/1103)$. Hence

$$P(X = 0) = \left(1 - \frac{1}{1103}\right)^{245} \approx \mathrm{e}^{-\lambda}, \qquad P(X = 1) = \frac{245}{1103}\left(1 - \frac{1}{1103}\right)^{244} \approx \lambda \mathrm{e}^{-\lambda}$$

with $\lambda = 245/1103$. We have that

$$P(X \geq 2) \approx 0.02, \qquad P(X = 0) \approx 0.8, \qquad P(X = 1) \approx 0.18.$$

Comparing this to data, we have that

$$P(X_{\text{data}} = 0) = \frac{1103 - 191}{1103} \approx 0.83, \qquad P(X_{\text{data}} = 1) = \frac{45}{1103} \approx 0.13, \qquad P(X_{\text{data}} \geq 2) \approx 0.04.$$

Maybe some postmen are more likely to be attacked. We can use a test for goodness of fit.

E. Uniform distribution

Suppose $a < b$, $a, b \in \mathbb{R}$, and $f(x) = 1/(b-a)$ for $x \in (a, b)$ and zero otherwise. We write the random variable following this probability distribution by $X \sim \mathrm{U}(a, b)$. For $a < c < d < b$, we have

$$P(c < X < d) = \frac{d - c}{b - a},$$

while for $c < a < d < b$ (i.e., some of the range lies outside of the support),

$$P(c < X < d) = \frac{d - a}{b - a}.$$

F. Exponential distribution

Consider some random events occurring at rate β (like the Poisson set up). Let T denote the time to the first event occurring, and consider the event $\{T > t\} = \{\text{no choices in } (0, t)\}$. We know the probability of no choices in $(0, t)$, which is

$$P(\text{no choices in } (0, t)) = P(X = 0) = \mathrm{e}^{-\beta t}, \qquad t > 0$$

since $X \sim \mathrm{Po}(\beta t)$. Hence, for $t > 0$,

$$P(T > t) = \mathrm{e}^{-\beta t}, \qquad P(T \leq t) = 1 - \mathrm{e}^{-\beta t},$$

and

$$\int_0^x \beta \mathrm{e}^{-\beta y}\, \mathrm{d}y = 1 - \mathrm{e}^{-\beta x},$$

so T has pdf $f(t) = \beta \mathrm{e}^{-\beta t}$ for $t > 0$. Any random variable with X with this pdf and $f(t) = 0$ for $t < 0$ is written $X \sim \mathrm{Exp}(\beta)$.

G. Normal (Gaussian) distribution

This distribution has two parameters, $\sigma > 0$ and μ. We say $X \sim \mathrm{N}(\mu, \sigma^2)$ when its pdf is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \geq 0, \qquad x \in \mathbb{R}.$$

This function is symmetric about μ.

H. Cumulative distribution function (cdf)

For any random variable X, its cumulative distribution function (cdf) is

$$F(x) = \begin{cases} \int_{-\infty}^{x} f(y)\,\mathrm{d}y, & X \text{ continuous}, \\ \sum_{y \leq x} f(y), & X \text{ discrete}. \end{cases}$$

The cdf and pdf of X is related by $F'(x) = f(x)$. It is enough to know either F or f to be able to find $P(X \in A)$ for any set A.

Example For the pdf

$$f(x) = \begin{cases} (2/5)(1+y), & -1 \leq x < 0, \\ 1/5 + (2/5)(2-y), & 0 \leq x \leq 2, \end{cases}$$

the cdf is

$$F(x) = \begin{cases} 0, & x < -1, \\ (x+1)^2/5, & -1 \leq x < 0, \\ 1-(x-2)/5, & 0 \leq x \leq 2, \\ 1, & x > 2. \end{cases}.$$

While $f(x)$ is discontinuous, $F(x)$ is continuous.

Example The standard normal distribution has parameters $\mu = 0$, $\sigma = 1$. The cdf is denoted

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-x^2/2}\,\mathrm{d}x,$$

and values are tabulated in many books. From symmetry, $\Phi(\mu) = 1/2$, and $\Phi(-z) = 1 - \Phi(z)$. For instance,

$$P(z < -1.04) = 1 - \Phi(1.04) = 1 - (0.6 \times 0.841 + 0.4 \times 0.864) = 0.150,$$

and

$$P(-1.04 < z < 2.09) = \Phi(2.09) - \Phi(1.04) = (0.1 \times 0.977 + 0.9 \times 0.982) - 0.150 = 0.831.$$

I. Transformation of random variables

Suppose X is a random variable and $g : \mathbb{R} \to \mathbb{R}$. Then $g(X)$ is also a random variable, and is found by composing g with X. X assigns numerical values to trial outcomes, and g transforms this to other numerical values.

Example Consider $X \sim \mathrm{Exp}(\beta)$. Find the distribution of aX, $a > 0$.

$P(aX \leq y) = P(X \leq y/a) = 1 - \mathrm{e}^{-\beta y/a}$, as $P(X \leq y/a)$ is found by the cdf of $aX \sim \mathrm{Exp}(\beta)$. Since aX has cdf $1 - \mathrm{e}^{-(\beta/a)y}$, $aX \sim \mathrm{Exp}(\beta/a)$.

The probability integral transform Suppose $U \sim \mathrm{U}(0,1)$. Consider any random variable X with cdf $F(x)$ such that it is strictly increasing on (a,b). Hence there exists F^{-1}, and F^{-1} increases from a to b on $(0,1)$. Consider the distribution $F^{-1}(u)$. We have

$$P(F^{-1}(u) \leq X) = P(F(F^{-1}(u)) \leq F(x)) = P(u \leq F(x)),$$

so for $X \in (a,b)$, we have $F(x) \in (0,1)$. Thus $P(U \leq F(x)) = F(x)$, and hence $F^{-1}(u)$ has the same distribution as $F(u)$.

Suppose $X \sim \mathrm{Exp}(\beta)$, so that $F(x) = 1 - \mathrm{e}^{-\beta x}$, $x \geq 0$. Let $u = 1 - \mathrm{e}^{-\beta x}$, so that

$$F^{-1}(u) = -\frac{1}{\beta}\log(1-u),$$

and hence $-(1/\beta)\log(1-U) \sim \mathrm{Exp}(\beta)$ when $U \sim \mathrm{U}(0,1)$.

If $X \sim \mathrm{N}(\mu,\sigma^2)$ and $Z \sim \mathrm{N}(0,1)$, then $(X-\mu)/\sigma \sim \mathrm{N}(0,1)$, and $\mu + \sigma Z \sim \mathrm{N}(\mu,\sigma^2)$, since

$$\left(\frac{x-\mu}{\sigma} \leq Z\right) \Leftrightarrow (X \leq \mu + \sigma z)\,,$$

so that

$$P(X \leq \mu + \sigma z) = P(Z \leq z)$$

by letting $z = (x-\mu)/\sigma$.

J. Discrete jointly distributed random variable

Suppose (X,Y) takes values (x_i, y_i). The joint pdf of (X,Y) is $P(x_i,y_i) < P(X = x_i, Y = y_i)$. As before, for any set of outcome pairs A, we calculate

$$P((X,Y) \in A) = \sum_{(x,y)\in A} p(x,y).$$

In particular,

$$P(X = y) = \sum_y p(x,y) = p_x(x), \qquad P(Y = y) = \sum_x p(x,y) = p_y(y),$$

the pdf of X and Y respectively.

Example Suppose that we have two random variables X and Y that are jointly distributed, with joint distribution density given by

$p(x,y)$	1	2	3	4	$p_X(x)$
0	0	0	0	1/4	1/4
1	0	1/4	1/4	0	1/2
2	1/4	0	0	0	1/4
$p_Y(y)$	1/4	1/4	1/4	1/4	

We see that, in fact, $X \sim \mathrm{B}(2,1/2)$ and $Y \sim \mathrm{U}(1,4)$.

For any bivarite random variable (X,Y), the conditional distribution of X given $Y = y$ is

$$P(x|y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x,y)}{p_Y(y)}.$$

Then, for any event A which we describe just using X,

$$P(A) = \sum_{(x,y)\in A} p(x,y) = \sum_{(x,y)\in A} P(x|y)p_Y(y) = \sum_x\sum_y P(x|y)p_Y(y).$$

This is the bivarite version of the partition theorem.

We say X is independent of Y when $P(X = x) = p_X(x) = P(x|y)$, i.e., $p_X(x)p_Y(y) = p(x,y)$ for all (x,y). The previous example is thus not independent.

Consider $g : \mathbb{R}^2 \to \mathbb{R}$, so, for example, $X+Y$, $X^2 - Y^2$, e^{XY}. Although this can be done, it may be complicated to work out the joint distribution of $g(X,Y)$.

K. Continuous jointly distributed random variable

For a pair (x, y), we describe their probability distribution as the integral of a joint distribution $f(x, y)$. For any rectangle $[a, b] \times [c, d]$, we define

$$P((x,y) \in A) = \int_a^b \int_c^d f(x,y)\,\mathrm{d}y\,\mathrm{d}x, \qquad P(a < x < b) = \int_a^b \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

To obtain marginal densities of X and Y, we integrate $f(x, y)$ over the other variable, i.e.,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y, \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x.$$

Then the conditional densities of X given Y is

$$f(x|y) = \begin{cases} f(x,y)/f_Y(y), & f_Y(t) > 0, \\ 0, & f_Y(y) = 0. \end{cases}$$

The law of total probability then has a continuous version given by

$$\begin{aligned} P(a < x < b) &= \int_a^b \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_a^b \int_{-\infty}^{\infty} f(x|y) f_Y(y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \left[\int_a^b f(x|y)\,\mathrm{d}x \right] f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} P(a < x < b | Y = y) f_Y(y)\,\mathrm{d}y. \end{aligned}$$

Example Suppose (X, Y) has joint density

$$f(x,y) = \begin{cases} x + y, & (x,y) \in [0,1]^2, \\ 0, & \text{otherwise}. \end{cases}$$

f is non-negative and is seen to integrate to 1. The marginal densities are then

$$f_X(x) = \int_0^1 f(x,y)\,\mathrm{d}y = x + \frac{1}{2}, \qquad f_Y(y) = \int_0^1 f(x,y)\,\mathrm{d}x = y + \frac{1}{2}$$

for $0 \le x \le 1$ and $0 \le y \le 1$. We have

$$P\left(\frac{1}{4} < x < \frac{3}{4},\ 0 < y < \frac{1}{2}\right) = \int_{1/4}^{3/4} \int_0^{1/2} (x+y)\,\mathrm{d}y\,\mathrm{d}x = \frac{3}{16},$$

$$P(X < Y) = \int_0^1 \int_0^y (x+y)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2},$$

$$P(X^2 < Y) = \int_0^1 \int_0^{\sqrt{y}} (x+y)\,\mathrm{d}x\,\mathrm{d}y = \frac{13}{20},$$

and so on. For the conditional probabilities,

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{x+y}{x+1/2}, \qquad f(y|x) = \frac{f(x,y)}{f_Y(y)} = \frac{x+y}{y+1/2}.$$

We see then X and Y are not independent, since

$$f(x,y) = x + y \neq (x+1/2)(y+1/2) = f_X(x) f_Y(y).$$

As a further example, we have

$$P(Y > 1/2 \mid X = x) = \int_{1/2}^1 \frac{x+y}{x+1/2}\,\mathrm{d}y = \frac{x+3/4}{2x+1}.$$

L. Generating functions

For any random variable X, the function $M_X : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ defined by

$$M_X(t) = \mathbb{E}\left(\mathrm{e}^{tX}\right)$$

is called the moment generation function of X. We have

$$M_X(t) = \begin{cases} \sum_x p(x)\mathrm{e}^{tx}, & X \text{ discrete}, \\ \int_{-\infty}^{\infty} f(x)\mathrm{e}^{tx}\,\mathrm{d}x, & X \text{ continuous}. \end{cases}$$

Example 1. If $X \sim \mathrm{B}(1,0)$, then

$$M_X(t) = (1-p)\mathrm{e}^0 + p\mathrm{e}^{t\cdot 1} = p\mathrm{e}^t + (1-p).$$

2. If $Y \sim \mathrm{Po}(\lambda)$, then

$$M_Y(t) = \mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda\mathrm{e}^t)^k}{k!} = \mathrm{e}^{\lambda(\mathrm{e}^t-1)}.$$

3. If $U \sum \mathrm{U}(a,b)$, then

$$M_U(t) = \int_a^b \mathrm{e}^{tu}\frac{1}{b-a}\,\mathrm{d}u = \begin{cases} \dfrac{\mathrm{e}^{bt}-\mathrm{e}^{at}}{t(b-a)}, & t \neq 0, \\ 1, & t = 0. \end{cases}$$

Moment generating functions have the following special properties:

1. Moment. For $r \in \mathbb{Z}^+$,

$$\mathbb{E}(X^r) = \left.\frac{\mathrm{d}^r M_X}{\mathrm{d}t^r}\right|_{t=0}.$$

2. Uniqueness. $M_X(t)$ uniquely determines the probability distribution of f provided $M_X(t)$ is finite in $(-h,h)$, $h > 0$.

3. Scaling. If X has $M_X(t)$ and $Y = aX + b$, then

$$M_Y(t) = \mathrm{e}^{bt}M_X(at).$$

4. Multiplicative. Suppose $X_1, \ldots X_n$ are independent random variables, and let $Y = \sum_{i=1}^n X_i$, then

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t).$$

5. Convergence. Suppose $Y_1, \ldots$ is an infinite sequence of random variables, and Y is a further random variable. Suppose $M_Y(t)$ is finite for $|t| < a$, $a > 0$, and $M_{Y_n}(t) \to M_Y(t)$ as $n \to \infty$ for all $t \in (-a,a)$, then

$$P(Y_n \leq c) \to P(Y \leq c)$$

as $n \to \infty$ for all c such that $P(Y = c) = 0$.

Example 1. Suppose $Z \sim \mathrm{N}(0,1)$, then

$$\begin{aligned} M_Z(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{tx}\mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{-(x^2-2tx+t^2)/2+t^2/2}\,\mathrm{d}x \\ &= \frac{\mathrm{e}^{t^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{(x-t)^2/2}\,\mathrm{d}x \\ &= \mathrm{e}^{t^2/2}. \end{aligned}$$

Then we have

$$M_Z'(t) = t\mathrm{e}^{t^2/2} \qquad \Rightarrow \qquad \mathbb{E}(X) = M_Z'(0) = 0,$$
$$M_Z''(t) = (t^2+1)\mathrm{e}^{t^2/2} \qquad \Rightarrow \qquad \mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = M_Z''(0) - [M_Z'(0)]^2 = 1 - 0 = 1.$$

For general normal distributions, let $W = \sigma Z + \mu$, then $W \sim \mathrm{N}(\mu\sigma^2)$ and the moment generation function is transformed to

$$M_W(t) = \mathrm{e}^{\mu t} M_Z(\sigma t) = \mathrm{e}^{\mu t + (\sigma t^2)/2}.$$

2. Suppose $X_1, \dots X_n$ are independent and $X_i \sim \mathrm{Po}(\lambda_i)$ for all i. Then

$$M_i(t) = \sum_{x=0}^{\infty} \mathrm{e}^{tx}\mathrm{e}^{-\lambda_i}\frac{\lambda_i^x}{x!} = \sum_{x=0}^{\infty} \mathrm{e}{-}\lambda_i \frac{(\mathrm{e}^t\lambda_i)^x}{x!} = \mathrm{e}^{-\lambda_i}\mathrm{e}^{\lambda_i \mathrm{e}^t} = \mathrm{e}^{\lambda_i(\mathrm{e}^t-1)}.$$

Uniqueness implies that $M_Y(t) = \mathrm{e}^{\lambda(\mathrm{e}^t-1)}$ iff $\mathrm{Po}(\lambda)$, while $M_Z(t) = \mathrm{e}^{\mu t + (\sigma t)^2/2}$ iff $Z \sim \mathrm{N}(\mu, \sigma^2)$. We can show that the sum of Poisson and normal random variables are still Poisson and normal respectively.

IV. STATISTICS

A. Expected values

For a random variable X we define its expected value $\mathbb{E}(X)$ as

$$\mathbb{E}(X) = \begin{cases} \sum_i x_i p(x_i), & X \text{ discrete,} \\ \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x, & X \text{ continuous.} \end{cases}$$

Example 1. For $X \sim \mathrm{B}(3, 1/2)$, $p(x) = 1/8, 3/8, 3/8, 1/8$ for $x = 0, 1, 2, 3$, we have

$$\mathbb{E}(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}.$$

2. For $X \sim \mathrm{U}(1, n)$, $p(x) = 1/n$, and

$$\mathbb{E}(x) = \frac{1}{n}\sum_i = 1^n i = \frac{1}{n}\frac{(n+1)n}{2} = \frac{n+1}{2}.$$

3. X has density $f(x) = x/2$ for $x \in (0, 2)$, so

$$\mathbb{E}(X) = \int_0^2 \frac{x^2}{2}\,\mathrm{d}x = \frac{4}{3}.$$

We then also have

$$\mathbb{E}(g(X)) = \begin{cases} \sum_{j=1}^{\infty} p(x_j) g(x_j), & X \text{ discrete,} \\ \int_{-\infty}^{\infty} g(x) p(x)\,\mathrm{d}x, & X \text{ continuous.} \end{cases}$$

Example 1. Suppose X takes integer values between -2 and 3, and $P(X = x) = 1/6$. Then

$$\mathbb{E}(X^2) = \frac{1}{6}\sum_{x=-2}^{3} x^2 = \frac{19}{6},$$

and

$$\mathbb{E}(\sin(\pi x/4)) = \frac{1}{6}\left(-1 - \frac{1}{\sqrt{2}} + 0 + \frac{1}{\sqrt{2}} + 1 + \frac{1}{\sqrt{2}}\right) = \frac{1}{6\sqrt{2}}.$$

2. Suppose $X \sim U(-1,1)$ with $p(x) = 1/2$ and zero otherwise, then

$$\mathbb{E}(X^2) = \frac{1}{2}\int_{-1}^{1} x^2 \, \mathrm{d}x = \frac{1}{3}.$$

Since sums and integrals are linear and additive operations, we see that

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

and, as a special case,

$$\mathbb{E}(a + bX) = a + b\mathbb{E}(X).$$

B. Variance

$\mathbb{E}(X)$ is a description of the average of X. We also need a measure of the spread of X's pdf, and we use the variance as it is the easiest one, defined as

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

In applications Var(X) has the wrong units, so we use the standard deviation $\sigma(X)$. By transformation, this is

$$\text{Var}(X) = \begin{cases} \sum_{i=1}^{\infty}(x_i - \mu)^2 p(x_i), & X \text{ discrete,} \\ \int_{-\infty}^{\infty}(x_i - \mu)^2 p(x)\,\mathrm{d}x, & X \text{ continuous,} \end{cases}$$

where $\mu = \mathbb{E}(X)$.

Example 1. If X takes values $0, 10, 20$ with $P(X = x) = 1/3$, then

$$\mathbb{E}(X) = 10, \qquad \text{Var}(X) = \frac{1}{3}[(0-10)^2 + (10-10)^2 + (20-10)^2]\frac{200}{3}.$$

2. If $X \sim \text{N}(0,1)$, then $p(x) = (1/\sqrt{2\pi})\mathrm{e}^{-x^2/2}$, and we notice that $xp(x)$ is an odd function so is has zero mean. On the other hand,

$$\text{Var}(X) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 \mathrm{e}^{-x^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x\frac{\mathrm{d}}{\mathrm{d}x}\left(-\mathrm{e}^{-x^2/2}\right)\mathrm{d}x = 1$$

after an integration by parts.

For the variance,

$$\text{Var}(a + bX) = \mathbb{E}([a + bX - a - b\mathbb{E}(X)]^2) = b^2\mathbb{E}([X - \mathbb{E}(X)]^2) = b^2\text{Var}(X).$$

Further,

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2, \qquad \text{Var}(-X) = (-1)^2\text{Var}(X) = \text{Var}(X).$$

C. Expected values for bivariate random variables

Suppose $g : \mathbb{R} \to \mathbb{R}^2$, then $g(x,y)$ is a random variable, with

$$\mathbb{E}(g(X,Y)) = \begin{cases} \sum_i \sum_j p(x_i, y_j) g(x_i, y_j), & (X,Y) \text{ discrete,} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)g(x,y)\,\mathrm{d}x\,\mathrm{d}x, & (X,Y) \text{ continuous.} \end{cases}$$

If (X,Y) is a bivariate random variable, with finite $\mathbb{E}(X)$ and $\mathbb{E}(Y)$, then

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

Example

1. Tickets in a raffle cost 1G. There are 2,000 tickets and some prizes totalling 1,000G. Let X denote the amount you win from buying a ticket. Every prize is won with a change of $1/2000$, and so we have

$$\mathbb{E}(X_1) = \sum_{i=1}^{2000} \frac{v_i}{2000},$$

where v_i is the value of the i^{th} prize. So $\mathbb{E}(X_1) = 1000/2000 = 1/2$G.

Suppose you buy 20 tickets instead. Let $T = X_1 + \ldots X_{20}$ be the total you win. Then $E(T) = E(X_1) + \ldots \mathbb{E}(X_{20}) = 20(1/2) = 10$G. We can see that if we buy 2,000 tickets, we recover the 1,000G, as expected.

2. $X \sim \mathrm{B}(n,p)$, and let $S_i = 1$ that trial i is a success and 0 when it is a failure. Then for $X = \sum_i S_i$,

$$\mathbb{E}(X) = \sum_i \mathbb{E}(S_i) = n(1 \cdot p + 0 \cdot (1-p)) = np.$$

D. Expectation / probability inequalities

1. If a random variable satisfies $X \geq a$, $P(X \geq a) = 1$, then $\mathbb{E}(X) \geq a$.
2. For any random variable X, Var$(X) \geq 0$.
3. Markov's inequality: If $P(X \geq 0) = 1$, then, for $a > 0$,

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

This follows because

$$\mathbb{E}(X) = \int_0^\infty x f(x)\,\mathrm{d}x \geq \int_a^\infty a f(x)\,\mathrm{d}x = aP(X = a)$$

for $a > 0$.

4. Chebyshev's inequality: For any random variable,

$$P(|X - \mathbb{E}(X)| \geq a) \leq \frac{\text{Var}(X)}{a^2}.$$

To see this, applying Markov's inequality to $[X - \mathbb{E}X]^2$ at a^2 gives

$$P([X - \mathbb{E}(X)]^2 \geq a^2) \leq \frac{\mathbb{E}([X - \mathbb{E}(X)]^2)}{a^2} = \frac{\text{Var}(X)}{a^2},$$

and the result follows.

Example If (X, Y) is bivariate with finite $\mathbb{E}(X)$ and $\mathbb{E}(Y)$, then shown $\mathbb{E}(Y) \leq \mathbb{E}(X)$ if $P(X \geq Y) = 1$.

Defining $g(X,Y) = X - Y$, we see this is non-negative, so by Markov's inequality, $0 \leq \mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y)$, and result follows.

Expected values and probabilities are closed related. Consider any set A, then for the characteristic function $\chi_{x \in A}$ (1 if $x \in A$ and 0 otherwise), we have

$$\mathbb{E}(\chi_{x \in A}) = \sum_i \chi_{x_i \in A} p(x_i) = \sum_{x \in A} p(x_i) = P(x \in A).$$

Example Suppose there are 2,000 tickets and 1,000G worth of prizes. Let X denote the amount we win with one ticket. By Markov's inequality,

$$P(X \geq 1) \leq \frac{1/2}{1} = \frac{1}{2}.$$

Suppose there are 1,000 1G prizes, then $P(X \geq 1) = 1/2$, and by Markov's inequality,

$$P(X \geq 1,000) \leq \frac{1/2}{1000} = \frac{1}{2000}.$$

Suppose instead there is only one 1,000G prize, then $P(X \geq 1000) = 1/2000$. This shows the bounds are not sharp.

If X and Y are independent, then $p(x,y) = p_X(x)p_Y(y)$ for all (x,y). Suppose we can write $f(x,y) = g(x)h(y)$ for all (x,y), then

$$\mathbb{E}(f(x,y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)p(x,y)\,\mathrm{d}x\,\mathrm{d}y = \left(\int_{-\infty}^{\infty} g(x)p_X(x)\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} h(y)p_Y(y)\,\mathrm{d}y\right) = \mathbb{E}(g(X))\mathbb{E}(h(Y)).$$

Further more, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ if X and Y are independent.

E. Conditional expectation

Suppose X is a discrete variable which takes values in Ω and suppose $A \subseteq \Omega$. Then the condition expectation of X given A is

$$\mathbb{E}(X \mid A) = \sum_{x\in A} xP(X = x \mid A).$$

Example Suppose there is a 500G prize and five 100G prizes. Let A be the event we win the 500G prize, so $P(X = 500 \mid A) = 1$, so $\mathbb{E}(X \mid A) = 500 \cdot 1 = 500$. On the other hand,

$$P(X = 5 \mid A') = \frac{5}{1999}, \quad P(X = 0 \mid A') = \frac{1994}{1999}, \qquad \Rightarrow \qquad \mathbb{E}(X = x \mid A') = 100 \cdot \frac{5}{1999} = \frac{500}{1999}.$$

Theorem IV.1 (Partition theorem) *If events $A_1, \dots A_n$ form a partition of Ω, then*

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X \mid A_i)P(A_i).$$

Proof Suppose X has possible values $x_i, \dots x_k \dots$. By definition of $\mathbb{E}(X)$ and the partition theorem,

$$\mathbb{E}(X) = \sum_{j=1}^{\infty} x_j P(X = x) = \sum_{j=1}^{\infty} x_j \sum_{i=1}^{n} P(X = x_j \mid A_i)P(A_i).$$

Re-arranging the order of summation,

$$\mathbb{E}(X) = \sum_{i=1}^{n}\left(\sum_{j=1}^{\infty} x_j P(X = x_j \mid A_i)\right)P(A_i) = \sum_{i=1}^{n} E(X \mid A_i)P(A_i).$$

Example With the above example,

$$\mathbb{E}(X) = 500 \cdot \frac{1}{2000} + \frac{500}{1999}\frac{1999}{2000} = \frac{1}{2}.$$

Consider a bivarite random variable (X,Y), and we seek $\mathbb{E}(X \mid Y = y)$. The conditional probability $P(X = x \mid A)$ is denoted by $p(x \mid y)$, hence

$$\mathbb{E}(X \mid Y = y) = \begin{cases} \sum_x xp(x \mid y), & (x,y) \text{ discrete,} \\ \int_{-\infty}^{\infty} xp(x \mid y)\,\mathrm{d}x, & (x,y) \text{ continuous.} \end{cases}$$

We can write $g(y) = \mathbb{E}(X \mid Y = y)$, and here

$$\mathbb{E}(g(Y)) = \sum_y p_Y(y)g(y) = \mathbb{E}(X), \qquad \Rightarrow \qquad \mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}(X).$$

This is because

$$\mathbb{E}(\mathbb{E}(X \mid Y)) = \sum_y p_Y(y)g(y) = \sum_y p_Y(y)\left(\sum_x xp(x \mid y)\right) = \sum_x x\sum_y p_Y(y)p(x \mid y) = \sum_x\sum_y p(x,y) = \sum_x p_X(x) = \mathbb{E}(X).$$

Example A shop gets N customers a day, with $\mathbb{E}(N) = 800$. Each customer independently spends X_i, with $\mathbb{E}(X_i) = 25$. Let $T = \sum_{i=1}^n X_i$ be the amount of turnover of the shop in the day. Taking a partition of N, we know that

$$\mathbb{E}(T \mid N = n) = \sum_{i=1}^n \mathbb{E}(X_i \mid N = n) = \sum_{i=1}^n \mathbb{E}(X_i) = 25n.$$

So then

$$\mathbb{E}(T) = \sum_{n=1}^\infty \mathbb{E}(T \mid N = n)P(N = n) = \sum_{n=1}^\infty 25np(N = n) = 25\sum_{n=1}^\infty np(N = n) = 25\mathbb{E}(N) = 20000,$$

which we would have expected from intuition anyway.

F. Co-variance

Suppose X and Y are jointly distributed random variables. Then the co-variance of X and Y is

$$\text{Cov}(X, Y) = \mathbb{E}([X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]).$$

This can be re-written as

$$\text{Cov}(X, y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Example

$$\begin{aligned}\text{Var}(X + Y) &= \mathbb{E}\left([(X + Y) - \mathbb{E}(X + Y)]^2\right)\\ &= \mathbb{E}\left([(X - \mathbb{E}(X))^2 + (y - \mathbb{E}(Y))^2]\right)\\ &= \mathbb{E}\left([X - \mathbb{E}(X)]^2 + 2\mathbb{E}([X - \mathbb{E}(Y)][Y - \mathbb{E}(X)]) + [Y - \mathbb{E}(Y)]^2\right)\\ &= \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y).\end{aligned}$$

Some properties of the co-variance:

1. For all X, $\text{Var}(X) = \text{Cov}(X, X)$.
2. For all X and Y, $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
3. For all X, Y and Z,

$$\begin{aligned}\text{Cov}(aX, bY) &= ab\text{Cov}(X, Y),\\ \text{Cov}(X + Y, Z) &= \text{Cov}(X, Z) + \text{Cov}(Y, Z),\\ \text{Cov}(X, Y + Z) &= \text{Cov}(X, Y) + \text{Cov}(X, Z).\end{aligned}$$

4. If X and Y are independent, $\text{Cov}(X, Y) = 0$.

Example n people throw their hats up and catch it. Let H be the event the person catches their own hat. H is not binomial as trials are not independent. Let $X_i = 0$ if person i catches someone else's hat, and 1 is person i catches their own hat. Then $H =\sim_{i=1}^n X_i$. We know that

$$\mathbb{E}(X_i)^p = \frac{1}{n} \cdot 1^p + \frac{n-1}{n} \cdot 0^p = \frac{1}{n},$$

so then $\mathbb{E}(H) = \sum_{i=1}^n X_i = n/n = 1$. We also have

$$\begin{aligned}\text{Var}(H) &= \sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j),\\ \text{Var}(X_i) &= \mathbb{E}(X_i^2) - [\mathbb{E}(X_i)]^2\,\frac{1}{n} - \frac{1}{n^2},\\ \text{Cov}(X_i, X_j) &= \mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j),\\ \mathbb{E}(X_i, X_j) &= 1 \cdot P(X_i = 1, X_j = 1) = P(X_i = 1 \mid X_j = 1)P(X_j = 1) = \frac{1}{n-1}\frac{1}{n}.\end{aligned}$$

Together, this gives

$$\mathrm{Var}(H) = \frac{1}{n} - \frac{1}{n^2} + 2\frac{n(n-1)}{2}\frac{1}{n^2(n-1)} = 1$$

Example For $X \sim \mathrm{B}(n,p)$, $X = X_i + \dots X_n$, with $\mathbb{E}(X) = np$. For each trial i,

$$\mathrm{Var}(X_i) = \mathbb{E}(X_i^2) - p^2 = 1^2 \cdot p + 0^2(1-p) - p^2 = p(1-p).$$

Each trial is independent so $\mathrm{Var}(X) = \sum_{i=1}^n \mathrm{Var}(X_i) = np(1-p)$.

For $X \sim \mathrm{B}(n, 1/n)$,

$$\mathrm{Var}(X) = n\frac{1}{n}\left(1 - \frac{1}{n}\right) = \left(1 - \frac{1}{n}\right).$$

For large n, H (from the previous example) and X are approximately follows a Poisson distribution with $\lambda = 1$.

Chebyshev's inequality says that $P(|X - \mathbb{E}(X)| \geq t) \leq \mathrm{Var}(X)/t^2$, so for $X \sim \mathrm{B}(n,p)$ with $t = n\epsilon$, we have

$$P(|X - np| \geq n\epsilon) \leq \frac{np(1-p)}{n^2\epsilon^2} = \frac{p(1-p)}{\epsilon^2}\frac{1}{n}.$$

V. LIMIT THEOREMS

A. Stability of sample proportion

Suppose $X \sim \mathrm{B}(n,p)$ and $B_n = X/n$, which is the proportion of trials that are successes. We know that $\mathbb{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$ from the previous example. So from the properties of the expectation and the variance,

$$\mathbb{E}(B_n) = \frac{1}{n}\mathbb{E}(X) = p, \qquad \mathrm{Var}(B_n) = \frac{\mathrm{Var}(X)}{n^2} = \frac{p(1-p)}{n}.$$

By Chebyshev's inequality,

$$P(|B - n - p| \geq \epsilon) \leq \frac{p(1-p)}{\epsilon^2}\frac{1}{n} \to 0$$

as $n \to \infty$ for all t. So that $n \to \infty$, the sample proportion is within a small interval centred on p, the expected value of the trial.

B. Law of large numbers

Let $X_1, \dots X_n$ are n trials on some experiment that we know are independent, and $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ for all i. Considering $\overline{X_n}$, the sample mean random variable, then

$$\mathbb{E}(\overline{X_n}) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^n X_i\right)\frac{n\mu}{\nu} = \mu,$$

$$\mathrm{Var}(\overline{X_n}) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Hence, by Chebyshev's inequality, for all $t > 0$,

$$P(|\overline{X_n} - \mu| \geq t) \leq \frac{\sigma^2}{nt^2} \to 0$$

as $n \to \infty$. So the average random variable has a high probability of being near to μ for large n.

C. Normal approximation to binomial

If $X \sim \mathrm{B}(n,p)$, with n large and p not near 0 or 1, then X is approximately $\mathrm{N}(np, np(1-p))$. Then, for any c,

$$P(X \le c) \approx \Psi\left(\frac{c-\mu}{\sigma}\right) = \Psi\left(\frac{c-np}{\sqrt{np(1-np)}}\right).$$

Example A student must scale more than 30 on a multiple choice test with $n = 50$ questions. Suppose he has $p = 1/2$ change of answering correctly. Let X be the number of correct answers, so $X \sim \mathrm{B}(50, 1/2)$, and we wish to estimate $P(X \ge 30) = 1 - P(X \le 29) = 1 - P(X < 30)$. Using a continuity correction of $c = 29.5$ gives

$$P(X \ge 30) \approx 1 - \Phi\left(\frac{29.5-\mu}{\sigma}\right) = 1 - \Phi(1.273) = 0.102 \qquad \left(\mu = 25, \quad \sigma^2 = \frac{25}{2}\right).$$

D. Central Limit Theorem (CLT)

Suppose we have a sequence of random variables $X_1, \ldots$ with $\mathbb{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$ for all i. Let $S_n = \sum_{i=1}^n X_i$, then $\mathbb{E}(S_n) = n\mu$, $\mathrm{Var}(S_n) = n\sigma^2$. For large n, $S_n \sim \mathrm{N}(n\mu, n\sigma^2)$, i.e.,

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{X_i - \mu}{\sigma}.$$

Then $P(Z_n \le c) \to \Phi(c)$ as $n \to \infty$. For any set of n numbers the difference between their total t and their average $\overline{x}$ is expressed by $\overline{x} = t/n$.

CLT also applies to the sample average $\overline{X} = n^{-1}\sum_{i=1}^n X_i$. A useful form of the approximation is

$$P(\overline{X} \le \mu + c\sigma/\sqrt{n}) \approx \Phi(c).$$

Example Measures from an experiment have mean μ and $\sigma = 2.5$. Find the n such that $P(|\overline{X_n} - \mu| \le 0.1) \ge 0.95$.

We know that

$$\mathbb{E}(\overline{X_n}) = \mathbb{E}\left(\frac{1}{n}S_n\right) = \frac{n\mu}{n} = \mu, \qquad \mathrm{Var}(\overline{X_n}) = \mathrm{Var}\left(\frac{1}{n}S_n\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

With $P(\overline{X} \le \mu + c\sigma/\sqrt{n}) \approx \Phi(c)$, we have

$$P(|\overline{X_n} - \mu| \le c\sigma/n) \approx \Phi(c) - \Phi(-c), \qquad c > 0.$$

So we need $c\sigma/n \le 0.1$, $\Phi(c) - \Phi(-c) > 0.95$. Choosing $c = 1.96 \approx 2$, this gives

$$0.1 \le \frac{2.5(2\sigma)}{\sqrt{n}} \qquad \Rightarrow \qquad \sqrt{n} \ge 50 \qquad \Leftrightarrow \qquad n \ge 2,500.$$