Probability: 510 Notes

Moulinath Banerjee

November 12, 2020

1 Sample Space, Sigma Field, Probability Measure

From an abstract perspective, probability starts with a sample space Ω (with a generic element in this set denoted ω) which is viewed as the set of all possible outcomes of a *random* experiment. The simplest interesting example is $\Omega = \{H, T\}$ where H is the outcome Heads and T the outcome Tails when a coin is flipped. Here, one is viewing the coin flipping experiment as *random*: one cannot say for certainty which of these outcomes will materialize when a coin is flipped, and describes the experiment in terms of an intrinsic chance (or probability) of the coin landing heads.
Probabilities are assigned to subsets of Ω. For technical reasons (which are best understood at a more advanced level) one does not assign a probability to every subset A of Ω but rather to special classes of subsets which satisfy some structural properties. Such classes are called sigma-fields. Formally, a sigma-field $\mathcal{A}$ on Ω is a class of subsets satisfying:

(i) $\Omega \in \mathcal{A}$

(ii) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$.

(iii) if $\{A_n\}$ is a sequence of events in $\mathcal{A}$ then $\cup_n A_n \in \mathcal{A}$.

Any member of $\mathcal{A}$ is called an event. A probability measure P on $(\Omega, \mathcal{A})$ is a map from $\mathcal{A}$ to $[0, 1]$ such that (i) $P(\Omega) = 1$ and (ii) if $\{A_n\}$ is a mutually disjoint sequence of events in $\mathcal{A}$, then $P(\cup_n A_n) = \sum_n P(A_n)$.
The second property is called the *countable additivity* property for obvious reasons. We talk of the triplet $(\Omega, \mathcal{A}, P)$ as a probability space.

Example 1.1. [Discrete sample space]: *Consider $\Omega = \{H, T\}$, $\mathcal{A} = 2^\Omega$ and P defined by $P(H) := P(\{H\}) = p$ for $0 \leq p \leq 1$. Check that $P(H)$ uniquely (and consistently) defines P for all subsets of Ω and distinct p's produce distinct P's.*
In general for a discrete sample space, i.e. Ω is finite or countable, say $\Omega = \{x_1, x_2, x_3, \ldots\}$, we take $\mathcal{A} = 2^\Omega$, and each P is uniquely characterized by the vector $\{p_1, p_2, p_3, \ldots\}$ where $p_i \geq 0$ and $\sum p_i = 1$ via the identification where $p_j = P(\{x_j\})$.

Example 1.2. [Continuous sample space]: *Continuous sample spaces only arise when the sample space is uncountable. There are various different kinds of uncountable spaces but for our purposes we will only be concerned with spaces of the type $\mathbb{R}^d$ where d is a positive integer or $\mathbb{R}^{\mathcal{N}}$ which is the space of all real-valued sequences. For the moment, confine to $[0,1]$. Most of the interesting probability calculations are done by equipping this with its Borel-sigma-field $\mathcal{B}_{[0,1]}$, which is defined as the* smallest sigma-field containing all closed sub-intervals of $[0,1]$. *This can also be defined as the intersection of all sigma-fields on $[0,1]$ that contain the closed sub-intervals, since an arbitrary intersection of sigma-fields also gives a sigma-field. The construction of a probability measure on the Borel-sigma-field is usually done via an* extension based procedure. *For example, the uniform probability on $([0,1], \mathcal{B}_{[0,1]})$ is defined as the unique probability P such that $P([a,b]) = b - a$. It is possible to show that such a P uniquely exists.*

Exercise 1.1. *If $\{A_n\}$ is a sequence of events in $\mathcal{A}$ then $\cap_n A_n \in \mathcal{A}$.*

Exercise 1.2. *If $\{A_n\}$ is a sequence of growing events, i.e. $A_j \subset A_{j+1}$, and $A := \cup_n A_n$, then $P(A_n) = \lim_n P(A_n)$. Formulate and establish an analogous proposition for a decreasing sequence of events.*

Lemma 1.1. *If A_1 and A_2 are two events $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$.*

Proof. $A_1 \cup A_2 = (A_1 \cap A_2) \cup (A_1 \cap A_2^c) \cup (A_2 \cap A_1^c)$, this being a disjoint union. ($A_1 \cap A_2^c$ is also written as $A_1 - A_2$.) Since P is countably additive, it is also finitely additive (why? show it). Hence $P(A_1 \cup A_2) = P(A_1 \cap A_2) + P(A_1 \cap A_2^c) + P(A_2 \cap A_1^c)$ which readily evaluates to $P(A_1) + P(A_2) - P(A_1 \cap A_2)$. □

Exercise 1.3. *For a sequence of events $A_1, A_2, \ldots, A_n$, show that $P(\cap_{j=1}^n A_j) \geq \sum_{j=1}^n P(A_j) - (n-1)$.*

We can develop an exact expression for $P(\cup_{j=1}^n A_j)$ in terms of all possible intersections of events from the family $\{A_j\}_{j=1}^n$. This is an extension of Lemma 1.1 and stated below.

Lemma 1.2. *For a sequence of events $A_1, A_2, \ldots, A_n$*

$$P(\cup_{i=1}^n A_i) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) .$$

Proof. We have proved it for $n = 2$. Assume the equality for a general n. We then prove that the expression holds for $n+1$. We then have:

$$\begin{aligned} P(\cup_{i=1}^{n+1} A_i) &= P(\cup_{i=1}^n A_i) + P(A_{n+1}) - P(\cup_{i=1}^n (A_i \cap A_{n+1})) \\ &= \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) + P(A_{n+1}) \end{aligned}$$

$$
\begin{aligned}
& - \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k} \cap A_{n+1}) \\
= & \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) + P(A_{n+1}) \\
& - \sum_{k'=2}^{n+1} (-1)^{k'} \sum_{1 \le i_1 < i_2 < \ldots < i_{k'} = n+1} P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i'_k}) \\
= & \sum_{i=1}^{n+1} P(A_i) + \sum_{k=2}^{n+1} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \ldots < i_k \ne n+1} P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) \\
& + \sum_{k'=2}^{n+1} (-1)^{k'-1} \sum_{1 \le i_1 < i_2 < \ldots < i_{k'} = n+1} P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_{k'}}) \, .
\end{aligned}
$$

But this is clearly the desired expression for $P(A_1 \cup \ldots \cup A_{n+1})$. □

In the case that the sample space is finite, the above proposition can be proved directly using a simple combinatorial argument. This is presented below.

Alternative proof: Let $\Omega = \{1, 2, \ldots, N\}$ and let P be a probability such that for any i, $P(i) = 1/N$. Then, clearly for any subset A, $P(A) = \#(A)/N$. Noting that $(-1)^{k-1} = (-1)^{k+1}$, proving the above proposition boils down to establishing that,

$$
\begin{aligned}
\#(\cup A_i) \;=\; & \sum \#(A_i) - \sum_{i<j} \#(A_i \cap A_j) + \sum_{i<j<k} \#(A_i \cap A_j \cap A_k) \\
& - \ldots\ldots + (-1)^{n+1} \#(\cap_{i=1}^{n} A_i) \, .
\end{aligned}
$$

So consider some element s belonging to $\cup A_i$. We need to show that s is counted exactly once on the right side of the above expression. Suppose that s belongs to k of the n sets. Then, on the right side of the above expression s is counted

$$
k - \binom{k}{2} + \binom{k}{3} - \ldots + (-1)^{k+1} \binom{k}{k}
$$

times. Call this number m. Then,

$$
\begin{aligned}
m \;&=\; \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} \\
&=\; 1 - 1 + \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j}
\end{aligned}
$$

$$
\begin{aligned}
&= 1 - \left(1 + \sum_{j=1}^{k} (-1)^j \binom{k}{j}\right) \\
&= 1 - \left(\sum_{j=0}^{k} \binom{k}{j} (-1)^j (1)^{k-j}\right) \\
&= 1 - (1-1)^k \\
&= 1\,.
\end{aligned}
$$

This finishes the proof.

1.1 Counting in discrete sample spaces

As the combinatorial proof above suggests, counting is an important technique for eliciting probabilities of events in discrete sample spaces. In this section, we discuss a number of modes of counting that prove to be quite useful in a variety of probability calculations.
Consider n distinct balls numbered 1 through n. We are interested in drawing a subset of r distinct balls from this set. Such a subset is called a subsample of size r *drawn without replacement* from the population and arises frequently in statistical problems. We ask the following questions.

Ordered subsamples of size r without replacement: How many such subsamples can be constructed? By ordered, we mean that we keep track of the arrangements. So if $r = 2$, and 1 and 2 are drawn sequentially in that order, we differentiate it from the subsample $(2, 1)$ that could also be drawn. To obtain the correct number, note that the first ball can be drawn in n ways, the second in $n - 1$ ways, and so on and so forth, with the r'th having $n - r + 1$ possibilities. The total number of ways of drawing an ordered sample is therefore $n \times (n-1) \times \ldots \times (n-r+1)$ which is denoted by the symbol $n_{P_r} = n!/(n-r)!$.

Unordered subsamples of size r without replacement: As the name suggests, here we do not care about the order. Thus, we are only interested in the number of distinct proper subsets of size r. This can be derived easily from the number of ordered subsamples. Each distinct subset of size r is counted in the above number $r!$ times since this is the number of arrangements of r distinct objects. Hence, the answer is $n_{C_r} = n!/(r!(n-r)!)$. And clearly, $n_{C_r} = n_{C_{n-r}}$.

Ordered subsamples of size r drawn with replacement: When we draw with replacement, an ordered subsample is a sequence of integers of length r where each integer can vary between 1 and n but integers could be repeated. So there are n possibilities for

the first slot, n again for the second slot, and so on and so forth, resulting in n^r possible subsamples.

Unordered subsamples of size r drawn with replacement: An unordered subsample of size r drawn without replacement is obtained by ignoring the order of the integers obtained in an ordered subsample of size r drawn with replacement. It is therefore characterized only by the number of 1's, number of 2's, ..., number of n's in the subsample. (It is important to note that when we draw with replacement r could be strictly larger than n.) So if $n = 3$ and $r = 3$, then the ordered subsample $3, 2, 2$ and the ordered subsample $2, 3, 2$ are counted as the same unordered subsample that comprises one 3 and two 2's. To determine the number of distinct unordered subsamples, we note that each subsample has a one-one correspondence with a vector of integers $(k_1, k_2, \ldots, k_n)$ where $k_i \geq 0$ and $\sum_i k_i = r$. Here k_i is the number of times i appears in the subsample. The total number of such vectors is given by the coefficient of x^r in the polynomial expansion $(1 + x + x^2 + \ldots + x^r)^n$.

Exercise 1.4. *Show that the coefficient of x^r in the above expansion is $(n + r - 1)_{C_r}$.*

Exercise 1.5. *Consider r identical balls and n urns marked 1 through n. Each ball is picked and assigned to an urn at random. Find the total number of ways in which the balls can be distributed into the urns.*

Example 1.3. Banach's Matchbox Problem: *A mathematician has two matchboxes in their pocket each containing n matchsticks. Every time they light a cigar they draw a matchstick from one pocket picked at random. Find the chance that when they reach into a pocket to find it empty the first time, the other has r matches left.*
Solution: *For this event to happen, the draw where they find their pocket empty the first time has to be the $2n - r + 1$'st draw, and the event can happen with the last draw being from the right pocket or the left pocket, giving two disjoint sub-events. By symmetry, each sub-event has the same probability, so let's compute the chance of the $2n-r+1$'st draw being from the right pocket and producing no matchstick. There are 2^{2n-r+1} possible sequences of symbols L and R of length $2n-r+1$: an L represents a draw from the left pocket and an R from the right. Of these, the number of sequences that end up producing no matchstick from the right pocket at the $2n-r+1$'st draw is $(2n-r)_{C_n}$ (since the $2n-r+1$'st draw is fixed at R and we need to distribute n R's and $n-r$ L's in $2n-r$ slots). Since all sequences of L and R are equally likely, the probability of this sub-event is $(2n-r)_{C_n}/2^{2n-r+1}$. Since the other sub-event (the one ending with L) has the same probability, we end up with $(2n-r)_{C_n}/2^{2n-r}$ as our final answer.*

Example 1.4. Matching problem: *N women put their hats in a box and then each draws a hat at random. Find the probability that there are no matches, i.e. no woman gets her own hat.*

Solution: *Suppose woman i wears a hat numbered i (hats are considered unidentical). Let H_i denote the hat drawn by the i'th woman. Then $(H_1, H_2, \ldots, H_N)$ is a permutation of the integers 1 through N and all permutations are to be considered equally likely. We seek to evaluate the probability of the event $E := \{H_i \neq i \forall i\}$. Consider the event $E^c = \cup_{i=1}^n \{H_i = i\}$. We can Lemma 1.2 to find $P(E^c)$:*

$$P(E^c) = \sum_{i=1}^{N} P(H_i = i) - \sum_{i_1 < i_2} P(\{H_{i_1} = i_1\} \cap \{H_{i_2} = i_2\}) - \ldots ,$$

which evaluates to

$$N \frac{1}{N} - \binom{N}{2} \frac{(N-2)!}{N!} + \binom{N}{3} \frac{(N-3)!}{N!} + \ldots + (-1)^{N+1} \binom{N}{N} \frac{1}{N!} ,$$

or

$$1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^{N+1} \frac{1}{N!} .$$

Hence

$$P(E) = \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^{N} \frac{1}{N!} .$$

Note that the number of permutations with no matches is therefore $N! \times \left(\sum_{j=0}^{N} (-1)^j \frac{1}{j!} \right)$ where $0!$ is interpreted as 1.

Exercise 1.6. *What is the probability of exactly k matches?*
Hint: *We need to count the number of permutations with exactly k matches. First, fix a group of k spots from 1 to N for the k matches. In the remaining $N - k$ slots we have to put the remaining $N - k$ integers so that there is no match. But this can be answered from the above example. This number needs to be multiplied by $\binom{N}{k}$. To get the probability divide by $N!$.*

1.2 Independent Events

Definition 1.1. *A finite sequence of events $A_1, A_2, \ldots, A_n$ is said to be mutually independent if, for every subcollection $\{A_{i_1}, A_{i_2}, \ldots A_{i_k}\}$ from this sequence,*

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j}) .$$

An infinite sequence of events $\{A_i\}_{i=1}^{\infty}$ is said to be mutually independent if every finite subcollection is independent in the above sense.

Therefore, for events A, B independence simply means $P(A \cap B) = P(A)P(B)$. Note that A is independent of itself if and only if it has probability 1 or 0.

Lemma 1.3. *If $A_1, A_2, \ldots A_n$ are mutually independent events, then so are $B_1, B_2, \ldots, B_n$ where each B_i is A_i or A_i^c.*

Proof. The properties of independence do not depend on the ordering of the events, so w.l.o.g. one may suppose that only the first k are complemented, where $k \geq 1$. Now, assume that the result is true for k (note that we know that the result is true for $k = 0$ vaccuously). We will prove that it holds for $k+1$. The events involved are $A_1^c, A_2^c, \ldots, A_k^c, A_{k+1}^c, A_{k+2}, \ldots, A_n$. We only need verify the product property for sub-collections that include A_{k+1}^c. So, consider such a sub-collection $C_1, C_2, \ldots, C_l, A_{k+1}^c, D_1, D_2, \ldots, D_m$ where the C_i's are a sub-collection of the first k events and the D_j's are a sub-collection of the last $n - (k+1)$ events. Now,

$$\begin{aligned}
P((\cap_1^l C_i) \cap A_{k+1}^c \cap (\cap_1^m D_j)) &= P((\cap_1^l C_i) \cap (\cap_1^m D_j)) - P((\cap_1^l C_i) \cap A_{k+1} \cap (\cap_1^m D_j)) \\
&= \left(\prod_1^l P(C_i)\right)\left(\prod_1^m P(D_j)\right) - \left(\prod_1^l P(C_i)\right) P(A_{k+1}) \left(\prod_1^m P(D_j)\right) \\
&= \left(\prod_1^l P(C_i)\right)\left(\prod_1^m P(D_j)\right)(1 - P(A_{k+1})) \\
&= \left(\prod_1^l P(C_i)\right)\left(\prod_1^m P(D_j)\right) P(A_{k+1}^c)\,.
\end{aligned}$$

□

The notion of independence is a powerful one and plays an important role in building models in probability and statistics. If two random events do not have any obvious causal connections, the probability of joint occurrence is modeled as the product of the individual probabilities of occurrences. In order to put this on a formal footing, it is useful to introduce the notion of product probability spaces with which we will dwell upon briefly. There are subtle intricacies we will gloss over as these are better covered in a measure-theoretic probability course (Stat 621).

Product Spaces: Let $(\Omega_1, \mathcal{A}_1, P_1), (\Omega_2, \mathcal{A}_2, P_2), \ldots, (\Omega_k, \mathcal{A}_k, P_k))$ be k probability spaces. The *product probability space* $(\prod_i \Omega_i, \prod_i \mathcal{A}_i, \prod_i P_i)$ is the probability space given by the cartesian product of the individual sample spaces $(\prod_i \Omega_i)$, with events being given by the smallest σ-field generated by all sets of the form $A_1 \times A_2 \times \ldots A_k$ where each $A_i \in \mathcal{A}_i$, denoted $\prod_i \mathcal{A}_i$ and probability measure P denoted $\prod_i P_i$ on $\prod_i \mathcal{A}_i$, defined uniquely by the requirement that:

$$P(A_1 \times A_2 \times \ldots \times A_k) = P(A_1) \times P(A_2) \times \ldots P(A_k)\,.$$

That such a P exists uniquely can be established by more advanced techniques but this need not bother us. Rather, what we will deal with is how such a construction enables us to go from simple random experiments to more complex ones by this sort of concatenation.

Example 1.5. Bernoulli and Binomial random experiments: *Let $\Omega_1 = \{H, T\}$, $0 \leq p \leq 1$ and define $P(H) = p$ and $P(T) = 1 - p = q$. This completely defines a valid probability on 2^Ω. This is the Bernoulli(p) probability: we think of tossing a coin with intrinsic probability p of landing H and q of landing T (coin cannot land on rim). But now suppose we think of n consecutive tosses of this coin where one toss has no connection whatever to another toss. How do we assign probabilities to a sequence of H and T's? Intuitively, the tosses are independent and therefore we should multiply the probabilities of outcome at each stage. The notion of the product space above helps in formalizing this intuition. The k'th toss corresponds to the probability space $\Omega_k = \Omega_1$ with the same σ-field and P measure on it. So the product space Ω is clearly all sequences comprising H's and T's that have length n, the product σ-field is all subsets of Ω (this is formally not diffcult to check as we are dealing with finite sample spaces). The product probability P_{prod} then works in the following manner: Let E_i denote the event $\{H\}$ or $\{T\}$ depending on what transpired at the i'th toss.*

$$P_{prod}(E_1 \times E_2 \times \ldots \times E_n) = P(E_1) \times P(E_2) \times P(E_n) = p^m q^{n-m}\,,$$

where m is the number of E_i's that are H. Thus, every sequence of length n has a probability assigned to it. The product probability P_{prod}, which we lazily subsequently also refer to as P and expect the differentiation to be clear from the context, corresponds to what we call a Binomial *experiment because it leads to what we call the Binomial distribution for the total number of heads. This will be clarified below.*
We can ask the question: what is the probability of H in the 1'st toss and T in the last toss? This should be $p \times q$ since we multiply probabilities across different tosses. To view this in terms of the product space, we are merely asking for the product probability of the event $\{H\} \times \Omega_2 \times \Omega_{n-1} \times \{T\}$ and

$$P(\{H\} \times \Omega_2 \times \Omega_{n-1} \times \{T\}) = P(H) \times (1)^{n-2} \times P(T) = pq\,.$$

Another natural question to ask is: what is the probability that m heads are obtained in n tosses? Let $A(m)$ denote the subset of Ω comprising all sequences with m H's. We are interested in $P(A(m))$. Note that each $\omega \in A(m)$ has probability $p^m q^{n-m}$. Then

$$P(A(m)) = \sum_{\omega \in A(m)} P(\omega) = |A(m)| \times p^m\, q^{n-m} = \binom{n}{m} p^m q^{n-m}\,.$$

Thus, the terms of the binomial expansion of $(p+q)^m$ give the probabilities corresponding to different numbers of heads.

Example 1.6. Geometric and Negative Binomial Experiments: *These random experiments are rather closely allied with the binomial experiments. In the binomial experiments, a coin is tossed a number of times, and one records the number of times H shows up. In the negative-binomial experiments, one fixes the number of H's in advance and keeps tossing the coin till this pre-fixed number is realized. The geometric is the simplest when the number of H's is 1.*
The geometric probability space is most easily described by $\Omega_g = \{H, TH, TTH, TTTH, \ldots\}$. This is the set of all possible outcomes. Probabilities are assigned by the product rule to each element in the sample space where $P(H) = p$ and $P(T) = q$. Writing $\omega_n = T^{(n-1)}H$, the element with T's in the first $n-1$ slots followed by H, we have $P(\omega_n) = q^{n-1}p$. This is the Geometric(p) experiment.
The $NB(r, p)$ experiment, where NB stands for negative-binomial, p the intrinsic probability of H on a single toss, and r the number of H's till which time we keep tossing the coin, can be described by the space of all finite sequences of H's and T's that terminate in an H and have a total number of H's. We can therefore ask the question: what is the probability that we need n trials to obtain r H's, where obviously $n \geq r$? Consider any such sequence of length n. This has probability $q^{n-r}p^r$ by allocation (reflecting independent coin-flips). How many such sequences exist? The last element is fixed at H, so the total number is simply $\binom{n-1}{r-1}$. The probability desired is therefore $\binom{n-1}{r-1}p^r q^{n-r}$. It is easy to see that the probability that x T's precede the r'th H is simply $\binom{x+r-1}{r-1}p^r q^{n-r}$, where $x \geq 0$.

1.3 Conditional probabilities

Definition 1.2. *The conditional probability of a generic event A* given *a generic event B with $P(B) > 0$ written $P(A|B)$ is defined as $P(A|B) := P(A \cap B)/P(B)$*

At this point, we will not deal with conditioning on an event B that has probability 0. However, it will become necessary later on, to define conditional probabilities given probability 0 events. These are tackled in conceptually different ways. The most rigorous treatment arises via the notion of the conditional expectation but needs measure-theoretic probability and will not be dealt with in this course. In an important class of problems, conditioning on an event of probability 0 can be defined in terms of a limit, by taking a natural sequence of events of positive probability shrinking to the zero probability event and conditioning on those. The notion of conditional probability is of paramount importance as it provides a way of updating probabilities of outcomes when knowledge about a related outcome *is available*. Indeed, most interesting problems in statistics, in some shape or form, deal with modeling conditional probabilities.

Example 1.7. *Consider the discrete space $\{1, 2, \ldots, N\}$ with the uniform probability assignment: $P(i/N) = 1/N$, corresponding to the classical random experiment where an*

element of this set is chosen in a way that gives no preference to which of the members in the set can appear. In this case:

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

In other words, the chance of A given B is simply the proportion of elements in $A \cap B$ to the number of elements in B.

Note that $P(\cdot|B)$ is a proper probability on the sample space. If A and B do not intersect, then the conditional probability is of course 0.

Now, note that

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

The above identity is referred to as **Bayes Rule**. Simple as it looks, its ramifications are profound. Bayes Rule gives us a way of obtaining the conditional probability of one event given another, in terms of the conditional probability of the second event given the first, and the marginal probabilities. Sometimes conditional probabilities are easy to elicit in one direction, but not necessarily in the other, even though the other direction may be of much interest. Below, we state a more generalized version of Bayes Rule.

Lemma 1.4. *Suppose that $\{B_j\}$ is a (finite/infinite) partition of Ω and A is any event with $P(A) > 0$. Also suppose that $P(B_i) > 0$ for each i (without any loss of generality). Then,*

$$P(B_j \mid A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A \mid B_j)\, P(B_j)}{\sum P(A \mid B_i)\, P(B_i)}\,.$$

The proof is immediate.

Example 1.8. *There are three cabinets, A, B and C, each of which has two drawers. Each drawer has one coin. A has two gold coins, B has two silver coins and C has one gold and one silver coin. A cabinet is chosen at random; one drawer is opened and a silver coin found. What is the chance that the other drawer also has a silver coin ?*
Solution: *To fit this in the framework of Lemma 1.4, take the B_j's to be the events that A is chosen, B is chosen and C is chosen. For brevity we shall denote the events by A, B and C. Note that these events are indeed a partition of the sure event that a cabinet is chosen, and furthermore*

$$P(A) = P(B) = P(C) = 1/3\,.$$

Let $S1$ denote the event that the opened drawer of the chosen cabinet has a silver coin. Clearly, we are required to find $P(B \mid S1)$. Now, using Bayes Rule, we have

$$\begin{aligned} P(B \mid S1) &= \frac{P(S1 \mid B)\, P(B)}{P(S1 \mid A)\, P(A) + P(S1 \mid B)\, P(B) + P(S1 \mid C)\, P(C)} \\ &= \frac{1 \times 1/3}{0 \times 1/3 + 1 \times 1/3 + 1/2 \times 1/3} \\ &= \frac{1}{1 + 1/2} \\ &= 2/3\,. \end{aligned}$$

Example 1.9. *A box has R identical red balls and B identical black balls. Two balls are chosen at random in succession without replacement.*
(a) Find the probability that both balls are red.
(b) Find the probability that one is red and the other is black.

Solution: *First consider (a). There are two ways of arriving at the answer. First, let's look at the more bare-bones argument. There are $\binom{r+b}{r}$ distinct ways of arranging the balls and under the sampling scheme, all arrangements are equally likely. The chance that both balls are red is therefore the number of distinct arrangements that the first two elements are both R divided by the total number of sequences. The numerator is simply $\binom{r-2+b}{r-2}$. Thus,*

$$P(RR) = \frac{\binom{r-2+b}{r-2}}{\binom{r+b}{r}} = \frac{r(r-1)}{(r+b)(r+b-1)}\,.$$

The same solution can be arrived in somewhat simpler fashion using conditional probabilities. Letting $R1$ be the event that red appears in the first draw and $R2$ being defined analogously,

$$P(RR) = P(R1) \times P(R2|R1) = \frac{r}{b+r} \times \frac{r-1}{b+r-1}\,,$$

which coincides with the above answer. Here we use the sequential nature of the drawing to elicit the conditional probabilities.

We can use conditional probabilities to work out (b) [you can try the more bare-bones argument yourself]. We want

$$P(BR \cup RB) = P(B1)P(R2|B1) + P(R1)P(B2|R1) = \frac{b}{r+b}\frac{r}{r+b-1} + \frac{r}{r+b}\frac{b}{r+b-1} = \frac{2rb}{(r+b)(r+b-1)}\,.$$

The notion of conditional probabilities allows the following general decomposition of the probability of intersection of a number of events as can be readily verified:

$$P(\cap_{i=1}^{n} A_i) = P(A_1) \times \prod_{j=2}^{n} P(A_j \mid \cap_{i=1}^{j-1} A_i)\,.$$

1.4 The infinite coin tossing experiment and the Uniform $[0,1]$ distribution

Visualizing discrete sample spaces and constructing probability measures on them (to be precise, their power set) is easy. Constructing probability measures on 'continuous' (uncountable) sample spaces is more involved, since uncountable spaces are much larger and one cannot really assign probabilities to all possible subsets. Here, we make an effort to visualize a canonical probability measure on a canonical uncountable sample space, namely the uniform probability measure on $[0,1]$. Once the uniform probability on $[0,1]$ has been constructed, it turns out one can manufacture any probability on the real line (i.e. where the sample space is the real line) via a simple trick – the inverse distribution function technique which we will consider later.

We will consider the infinite coin-tossing experiment. A fair coin, i.e. one for which $P(H) = P(T) = 1/2$ is flipped infinitely many times, where each flip is independent of the others and the outcome recorded. Now, of course, this is a hypothetical experiment but let's see what transpires. The sample space $\Omega = \{\omega = (e_1\, e_2\, e_3\, \ldots) : e_i = 1 \text{ or} 0\}$ where we identify H with 1 and $T = 0$. We can immediately assign a probability to every sample point here: for every $m \geq 1$: since,

$$P(\{\omega\}) \leq P(e_1\, e_2\, \ldots, e_m\, \star\, \star\star) = \frac{1}{2^m}\,,$$

we conclude that $P(\{\omega\}) = 0$. Thus, every sample point has 0 probability under this particular experiment. However, not all subsets of Ω have 0 probability. Let $\Omega_m := \{e_1^0\, e_2^0, \ldots, e_m^0\, \star\, \star\star\}$, i.e. all sequences where the first m elements take pre-fixed values but the remaining ones can vary as they please. Then $P(\Omega_m)$ is only determined by the first m flip outcomes and therefore equals $1/2^m$. This is one *critical difference* between probabilities on discrete spaces from probabilities on uncountable ones (indeed Ω is uncountable, as we will see shortly): with uncountable spaces, individual points can all have 0 probability but non-empty uncountable subsets can have positive probability (indeed, the set Ω_m is uncountable). This is no contradiction, since the law of additivity of probability only applies to *countable unions of events*.

The space Ω has a natural (and almost a 1-1 correspondence) with the interval $[0,1]$. We first define the set of dyadic rationals as follows. For each $m \geq 1$, define

$$\mathcal{D}_m := \{k/2^m : 0 \leq k \leq 2^m\}\,.$$

Now, define the set of dyadic rationals $\mathcal{D}$ to be $\cup_{m=1}^{\infty} \mathcal{D}_m$. Note that the $\mathcal{D}_m$'s form a sequence of increasing finite subsets of $[0,1]$. While $\mathcal{D}_m$ chops up $[0,1]$ into a uniform grid of width $1/2^m$, $\mathcal{D}_{m+1}$ chops each sub-interval induced by $\mathcal{D}_m$ into two further sub-intervals of equal length. Now, consider an arbitrary point x in $[0,1]$. We will determine an element in Ω that corresponds to x. First check if $x \geq 1/2$ or less. If $x \geq 1/2$, assign $e_1^{(x)} = 1$, otherwise, let

$e_1^{(x)} = 0$. Suppose $e_1^{(x)} = 0$, indicating that $x < 1/2$. Now check whether $x \geq 1/4$ or $x < 1/4$. If the former, assign $e_2^{(x)} = 1$, otherwise, let $e_2^{(x)} = 0$. Now, continue this process to generate the sequence $e_1^{(x)}\, e_2^{(x)}\, e_3^{(x)} \ldots\ldots$, which, by construction, is well-defined. It is easily seen that $|x - \sum_{j=1}^{m} e_j^{(x)} 2^{-j}| \leq 1/2^m$ and therefore

$$x = \sum_{j=1}^{\infty} \frac{e_j^{(x)}}{2^j}\,.$$

By this assignment the point $x = 1/2$ is mapped to the sequence 1000000..... However, note that there is another sequence in Ω whose infinite series adds up to $1/2$, and this is the sequence 011111.... This prevents a pure one-one correspondence. More generally, every dyadic rational other than 0 or 1 has two equivalent modes of representation: either in terms of a sequence ending in an infinite sequence of 0's, or an infinite sequence of 1's. To avoid this dichotomy, we throw out all sequences ending in infinitely many 0's from the sample space. This does not matter as this set is countable, and so we've thrown away a chunk of probability measure 0 (negligible). Continue to denote the reduced space by Ω. Now

$$x \leftrightarrow \sum_{j=1}^{\infty} \frac{e_j^{(x)}}{2^j}$$

establishes a one-one correspondence. This is the *binary representation* of x.
We next establish that under this correspondence $P((0,x]) = x$ for all $x \in (0,1)$. We first start with x being a dyadic rational and therefore of the form $k/2^m$ for some m. Also, choose the smallest m such that it can be represented in this way. Then k is necessarily odd, and the sequence $e_1^{(x)}\, e_2^{(x)} \ldots$ must have a 1 in the m'th position and 0's ever after. Let $1 \leq i_1 < i_2 < \ldots i_p = m$ denote the positions of the 1's in the sequence. Clearly

$$x = \sum_{l=1}^{p} \frac{1}{2^{i_l}}\,.$$

Now, consider the set of all sequences in Ω that correspond to numbers no larger than x. This can be decomposed into the disjoint union $S_1 \cup S_2 \cup \ldots S_p \cup S_{p+1}$ where

$$S_1 = \{e_1 = e_2 = \ldots = e_{i_1} = 0, \star\star\star\ldots\}$$

$$S_2 = \{e_1 = e_2 = \ldots = e_{i_1-1} = 0, e_{i_1} = 1, e_{i_1+1} = \ldots = e_{i_2} = 0, \star\star\star\ldots\}\,,$$

$$S_3 = \{e_1 = \ldots = e_{i_1-1} = 0, e_{i_1} = 1, e_{i_1+1} = \ldots = e_{i_2-1} = 0, e_{i_2} = 1, e_{i_2+1} = \ldots = e_{i_3} = 0, \star\star\star\ldots\}\,,$$

and so on and so forth, and $S_{p+1} = \{x\}$. Thus,

$$P((0,x]) = P(S_1) + P(S_2) + \ldots + P(S_p) + P(\{x\}) = \frac{1}{2^{i_1}} + \frac{1}{2^{i_2}} + \ldots + \frac{1}{2^{i_p}} + 0 = x\,.$$

To assign a probability to a general $(0, x]$ (which is not a dyadic rational), take a sequence of x_n's increasing to x. This is readily obtained from the binary representation of x. Then $(0, x) = \cup_n (0, x_n]$ and this is an increasing sequence of intervals, so $P((0, x)) = \lim P((0, x_n]) = \lim x_n = x = P((0, x])$. We have therefore assigned probabilities to all intervals of the form $(0, x]$ and it is follows that $P((x_1, x_2]) = x_2 - x_1$.
With sample space $[0, 1]$, the natural σ-field is the Borel σ-field generated by all intervals of the form $(a, b]$. We have naturally assigned probabilities to all intervals, such that

$$P(\text{interval}) = \text{length of interval}\,.$$

It takes some advanced techniques to show that this P can be consistently and uniquely extended to all sets in the σ-field above. Thus, the uniform probability on $[0, 1]$ is well-defined. Under the uniform probability assignment, volume (which is length) is identical to probability. The uniform probability is also known as *Lebesgue measure* on $[0, 1]$.

2 Random Variables and Distribution Functions

Given a probability space $(\Omega, \mathcal{A}, P)$, a random variable X is simply a function from Ω to $\mathbb{R}$ such that for all $B \in \mathcal{B}^1$, $X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\}$ is a member of $\mathcal{A}$.

For this course, we will not be much concerned about verifying the condition in the above definition. For discrete sample spaces where $\mathcal{A}$ is the power set, it is automatic. For uncountable sample spaces, like $[0,1]$ where $\mathcal{A}$ is the Borel σ-field, it is indeed possible to construct functions that don't satisfy the inclusion condition, but such functions will not arise in our considerations.

We start first with a discrete sample space $\Omega = \{\omega_1, \omega_2, \ldots\}$ and $X : \Omega \mapsto \mathbb{R}$. Then X assumes values in a countable set $\mathcal{X} = \{x_1, x_2, \ldots\}$ where, of course, multiple ω's may be mapped to the same x_j. The p.m.f. or probability mass function of X is a map $p : \mathcal{X} \mapsto [0,1]$ such that $p(x_j) = P(X = x_j)$. Note that

$$P(X = x_j) = \sum_{i:X(\omega_i)=x_j} P(\{\omega_i\})\,.$$

Example 2.1. *First consider the random experiments considered in Example 1.5. We focus on the Binomial experiment. Recall that the sample space that describes n flips of a coin is the space Ω of all n-long sequences of H and T. We consider the product probability as considered previously. Define $X(\omega) = m$ if ω contains m H's. Now, X takes values in the set of integers $\{0, 1, 2, \ldots, n\}$ and the p.m.f of X is*

$$p(m) = P(X = m) = P(A(m)) = \binom{n}{m} p^m q^{n-m}\,.$$

Next, consider Example 1.6. Start with the Geometric experiment as described there. If $T^{(j-1)}H$ for $j \geq 1$ denotes the element of Ω_g that has $j-1$ T's preceding the terminal H, define $\tilde{X}(T^{(j-1)}H) = j$. Thus, the r.v. $\tilde{X}$ denotes the number of trials needed to get the first H. One can easily write down its p.m.f. $\tilde{p}$ as:

$$\tilde{p}(j) = P(\tilde{X} = j) = P(T^{(j-1)}H) = q^{j-1}p\,.$$

Continuing with the Negative Binomial experiment, consider a generic ω in the corresponding sample space. This is a sequence of H's and T's with r H's that ends with an H. Define $Y(\omega)$ to be the length of the sequence. Thus Y denotes the number of trials needed to obtain r H's. Denoting its p.m.f. by p_Y, we have

$$p_Y(n) = P(Y = n) = \binom{n-1}{r-1} p^r q^{n-r} \quad \text{for } n = r, r+1, r+2, \ldots\,.$$

[1] $\mathcal{B}$ is the Borel σ-field on $\mathbb{R}$

The above example focuses on *discrete random variables* which assume values in a finite or countable set. Another important class of random variables is the class of *continuous random variables*. A random variable X is called continuous if $P(X = x) = 0$ for all real x. Section 1.4 allows us to construct such a continuous random variable quite easily. So consider $\Omega = [0, 1]$ equipped with its Borel σ-field and P the uniform probability constructed on it. Define $X : \Omega \mapsto [0, 1]$ as $X(\omega) = \omega$. Then clearly, $P(X = x) = P(\{x\}) = 0$ as shown in Section 1.4 and X is said to follow the Uniform$(0, 1)$ distribution. An interesting fact that we will later see is that all random variables can be generated starting with a uniform random variable X.

2.1 The Distribution Function

Associated with every random variable X is its distribution function F (sometimes written F_X to indicate its association with X defined as $F(t) = P(X \leq t)$ for every real t.
The following properties characterize a distribution function F.

(i) F is non-decreasing, i.e. for $x < y$, $F(x) \leq F(y)$.

(ii) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

(iii) F is an RCLL function, i.e. F is right continuous with left limits.
Right continuity means that if x_n converges to x with $x_n \geq x$ for all n, then $F(x_n) \to F(x)$. The property of possessing left-limits means that if x_n converges to x from the left then $\lim_n F(x_n)$ exists and is invariant to x_n.

Establishing these properties is not difficult if one writes F in terms of its underlying random variable X. That F is non-decreasing follows from the observation that $\{X \leq x\} \subset \{X \leq y\}$ whenever $x < y$, so $F(x) = P(X \leq x) \leq P(X \leq y) = F(y)$. We skip the proof of (ii) but provide a proof of (iii).
Consider a sequence x_n converging to x from the right. Because F is a monotone function, the limit of $F(x_n)$ exists if and only if the limit of any decreasing subsequence exists and the two limits must coincide in that case. So w.l.o.g. consider the case that x_n decreases to x. Then the events $\{X \leq x_n\}$ decrease to the event that $\{X \leq x\}$, i.e. $\{X \leq x_{n+1}) \subset \{X \leq x_n)$ and $\{X \leq x\} = \cap_n \{X \leq x_n\}$. By Exercise 1.2, $\lim_n P(X \leq x_n) = P(X \leq x)$.
To show the LL property, consider a sequence x_n converging to x from the left. Now, it suffices to look at an increasing subsequence, so we assume that x_n's are increasing. Now, note that $\{X \leq x_n\} \subset \{X \leq x_{n+1}\}$ and $\cup_n \{X \leq x_n\} = \{X < x\}$. By applying Exercise 1.2, we conclude that $\lim_n P(X \leq x_n) = P(X < x)$. The quantity $P(X < x)$ is denoted $F(x-)$ and called the left-hand limit of F at x. It is not difficult to see that $F(x-) = \sup_{y<x} F(y)$.

Remark 2.1. *Since $P(X = x) = F(x) - F(x-)$ it follows that X is a continuous random variable if and only if its distribution function F is continuous (i.e. has no left jump-discontinuities, which are the only kind of discontinuities possible).*

Exercise 2.1. *(a) Graph the distribution function of $U \sim$ Uniform$(0,1)$. (b) Graph the distribution function of $X \sim$ Bernoulli(p).*

Quantiles and Inverse Distribution Functions: For any $0 < p < 1$, the p'th quantile of F is any number x_p such that

$$F(x_p-) \equiv \lim_{y \to x_p-} F(y) = P(X < x_p) \leq p \leq P(X \leq x_p) = F(x_p)\,.$$

Clearly, the p'th quantile need not be unique. However, if F is a strictly increasing continuous function, in which case it is one-one, it's inverse function F^{-1} is well defined on $(0,1)$. For any $0 < p < 1$, $F^{-1}(p)$ is the unique number x, such that $F(x) = p$, and $F^{-1}(p)$ is the unique p'th quantile of F. When $p = 0.5$, we refer to the quantile as the median.

Fortunately, there is a neat way to define the inverse function F^{-1} even when F is not strictly increasing and continuous, as is the case with discrete random variables. For any $0 < t < 1$, we set,

$$F^{-1}(t) = \text{smallest element of } \{x : F(x) \geq t\}\,.$$

The fact that the above set does indeed have a smallest element can be shown. To be more rigorous, one would define F^{-1} as:

$$F^{-1}(t) = \inf\{x : F(x) \geq t\}\,.$$

Exercise 2.2. *It is now easy to see that $F^{-1}(t)$ is indeed a t'th quantile of F, though not necessarily the unique one.*

Note that,

$$F^{-1}(t) \leq x \Leftrightarrow F(x) \geq t\,. \tag{1}$$

This leads to the following crucial theorem: the Inverse Distribution Function Technique.

Theorem 2.1. *Let X be a random variable with distribution function F. Let U be a Uniform random variable on $(0,1)$. Then $Y = F^{-1}(U)$ is also a random variable and its distribution function is F.*

Proof. We have,

$$P(F^{-1}(U) \leq x) = P(F(x) \geq U) = P(U \leq F(x)) = F(x)\,.$$

Thus, by knowing F, and hence in principle F^{-1}, one can generate a random variable with distribution function F, provided one can generate from a $U(0,1)$ distribution. Another related theorem follows. □

A closely related theorem, an almost-converse though not quite, is the following.

Theorem 2.2. *If X is a continuous random variable, then $F(X)$ has the uniform distribution.*

The proof of this theorem, in the case that F is strictly increasing and continuous is not difficult and left as an exercise. The general case is omitted.

Exercise 2.3. *(a) Let $F(x) = (1 - exp(-\lambda x))\, 1(x \geq 0)$. Assume you have a uniform random number generator. How would you generate a random variable with distribution function F? (b) Consider a random variable X that assumes values in the set $\{1, 2, 3, \ldots\}$ with $p(j) = P(X = x_j)$. Write down its F in terms of the $p(j)$'s and use the Inverse Distribution Function technique to give an explicit way of generating a random variable $\tilde{X}$ that has the same distribution as X, given a realization of $U \sim Uniform(0, 1)$.*

Probability density functions: Continuous random variables cannot have a p.d.f (probability density function) since the probability of every point is 0. However, a large class of continuous random variables, called *absolutely continuous random variables* have the property that their distribution function can be written in terms of the integral of a non-negative function, called its density.

The random variable X is said to have density function f if $f(u) \geq 0$ for all u and $F(x) = \int_{-\infty}^{x} f(u)du$ for all x. Note that, it follows immediately that

$$P(X \in (a, b]) = \int_{(a,b]} f(u)du\,;$$

indeed, it can be shown that if the above is true, $P(X \in A) = \int_A f(u)du$ for all sets A in the Borel σ-field. If f is continuous at the point x_0, then $F'(x_0) = f(x_0)$. For a given F, the density f need not be unique but two densities for F can only differ on a very 'small' set of points.

Example 2.2. *Let $X \sim Uniform(0, 1)$ and let F be its distribution function. Then*

$$F(x) = x\, 1(0 \leq x \leq 1) + 1(x > 1)\,,$$

and $f(t) = 1(0 < t < 1)$ is easily seen to be a p.d.f. for X.

We note that any non-negative function g that integrates to 1 on the real line produces a valid continuous distribution function via the equation $F(t) = \int_{(-\infty,x]} f(u)du$, and therefore one can find a random variable X with this particular distribution function F via the inverse distribution function technique. Here are two important examples.

Example 2.3. The exponential and normal distributions: *The class of exponential distributions is described by a parameters $\lambda \in (0, \infty)$. Formally, X is*

said to follow the Exponential(λ) distribution if its distribution function has the form $F(x) = (1 - e^{-\lambda x})1(x \geq 0)$. It is not difficult to see that $f(x) = \lambda e^{-\lambda x}1(x \geq 0)$ is a density function for F.

The class of normal distributions is described by a pair of parameters (μ, σ^2) where $\mu \in \mathbb{R}$ and $\sigma > 0$. We say that $X \sim N(\mu, \sigma^2)$ if it has a density function given by:

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Exercise 2.4. *If $Z \sim N(0,1)$, show that X has the same distribution as $\mu + \sigma Z$.*

Transforming Random Variables: Recall that a random variable, X, discrete or continuous, is formally a map from a probability space $(\Omega, \mathcal{A}, P)$ to $(\mathbb{R}, \mathbf{B})$ such that $X^{-1}(B)$ (define) is an event for any Borel set B. The probability P on $\mathcal{A}$ *induces* a probability P_X on $\mathbf{B}$ via the equation

$$P_X(B) = P(X^{-1}(B)).$$

We call P_X the *distribution* of X. If $Y = g(X)$ for a *measurable* function $g : (\mathbb{R}, \mathbf{B}) \to (\mathbb{R}, \mathbf{B})$, i.e. $g^{-1}(B) \in \mathcal{B}$ for all $B \in \mathcal{B}$, then Y is also a random variable defined on $(\Omega, \mathcal{A}, P)$ and its distribution P_Y is given by:

$$P_Y(B) = P(Y^{-1}B) = P(g(X)^{-1}B) = P(X^{-1}g^{-1}(B)).$$

Two random variables X and X' are said to have the *same* distribution if P_X and $P_{X'}$ are identical. Note that the distribution function of X, say F_X is precisely $P_X((-\infty, t])$ for all t.

We will next deal with how to deduce properties of a random variable Y that is produced by transforming a random variable X with known properties. We consider, first, the case when X is discrete.

Theorem 2.3. *Suppose X is a discrete random variable assuming values $\{x_1, x_2, x_3, \ldots\}$ with probabilities $\{p_1, p_2, p_3, \ldots\}$. Let $Y = g(X)$, where g is a given function and let $\{y_1, y_2, y_3, \ldots\}$ be the values assumed by Y. Then the p.m.f. of Y is given by:*

$$P(Y = y_i) \;=\; \sum_{j:g(x_j)=y_i} p_j\,.$$

The proof is immediate.

Example 2.4. *Let X be a discrete random variable assuming values in $\{\pm m, m = 0, 1, 2, \ldots\}$ with*

$$P(X = 0) = \frac{1}{2} \;\; \textit{and}\; P(X = m) = \frac{2}{3}\,\frac{1}{2^{m+1}} \;\; \textit{and} \;\; P(X = -m) = \frac{1}{3}\,\frac{1}{2^{m+1}}.$$

Find the p.m.f. of $Y = |X|$.
Solution: $P(Y = 0) = P(X = 0) = 1/2$, *and for* $m > 0$,

$$P(Y = m) = \sum_{j \in \mathbb{Z}: |j| = m} P(X = j) = P(X = m) + P(X = -m) = \frac{1}{2^{m+1}}\,.$$

For continuous random variables with a density f, the **Jacobian Theorem** gives us the distribution of a transformed variable under some regularity conditions on the transformation g. This is the content of the below theorem.

Theorem 2.4. Change of variable theorem: *Let* X *be a continuous random variable with density function* f *and* g *be a real-valued function defined on some open interval* I*, such that* $P(X \in I) = 1$. *Assume further that* g *is continuously differentiable and that* $g^{'}(x) \neq 0$ *for* $x \in I$ *(these assumptions actually entail that* g *is a strictly monotone transformation on* I*). Let* $Y = g(X)$. *Then the density function of* Y *can be computed as,*

$$f_Y(y) = f(g^{-1}(y)) \mid \frac{d}{dy}\, g^{-1}(y) \mid = f(g^{-1}(y)) \left| \frac{1}{g^{'}(g^{-1}(y)} \right| \quad y \in g(I)\,,$$

and 0 otherwise.

Proof: Suppose $I = (a, b)$ and g is increasing. Then $a < X < b$ with probability 1 and the density f can be taken to be identically 0, outside of I. Also $g(a) < Y \equiv g(X) < g(b)$ with probability 1 and the density f_Y of Y can be taken to be identically 0 outside of $g(I)$. Let F_Y denote the distribution function of Y. Let $g(a) < y < g(b)$. Then,

$$F_Y(y) = P(Y \leq y) = P(g(a) \leq Y \equiv g(X) \leq y) = P(a \leq g^{-1}(Y) \leq g^{-1}(y))\,.$$

Since, $X \equiv g^{-1}(Y)$,

$$F_Y(y) = P(a \leq X \leq g^{-1}(y)) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))\,.$$

On differentiating the above, with respect to y, we obtain,

$$f_Y(y) = f(g^{-1}(y))\, \frac{d}{d\, y}\, g^{-1}(y) = f(g^{-1}(y)) \left| \frac{1}{g^{'}(g^{-1}(y)} \right|\,,$$

where the last equality follows from the chain rule and the fact that g and g^{-1} have positive non-vanishing derivatives. □

The equality in the above display clearly implies that for any subset S of I:

$$\int_S f(x)\, dx = \int_{g(S)} f(g^{-1}(y)) \mid \frac{1}{g^{'}(g^{-1}(y)} \mid dy\,. \tag{2}$$

In fact, the above formula is valid very generally; f does not need to be a density function. Any non-negative function, or for that matter, any function, with a well-defined (finite) integral will do. Let's see how we can apply this formula to compute an integral, that one might come across in calculus. Suppose we wish to compute,

$$\mathcal{S} = \int_0^1 u \, sin \, u^2 \, du \, .$$

We write

$$\int_0^1 u \, sin \, u^2 \, du = \int_0^1 \frac{1}{2} \, 2u \, sin(u^2) \, du \, ,$$

and then set $w = u^2$, noting that as u runs from 0 to 1, so does w in the same direction. But $w = u^2$ means that $dw = 2u \, du$ and thus

$$\int_0^1 \frac{1}{2} \, 2u \, sin(u^2) \, du = \int_0^1 \frac{1}{2} \, sin \, w \, dw \, ,$$

by direct substitution and we obtain,

$$\mathcal{S} = \frac{1}{2} \, (1 - cos \, 1) \, .$$

Basically, what we have done here is use (2) informally. To see this, let

$$f(x) = \frac{1}{2} \, 2x \, sin \, x^2 \quad I = (0, 1) \;\; \text{and} \;\; g(x) = x^2 \, .$$

Clearly, g is continuously differentiable on I and $g^{'}$ is non-vanishing (since $g^{'}(x) = 2x$). Now, the inverse transformation g^{-1} from $g(I) = (0, 1)$ to I is

$$g^{-1}(y) = \sqrt{y} \, .$$

Also,

$$f(g^{-1}(y)) = \frac{1}{2} \, 2\sqrt{w} \, sin \, w \, .$$

Thus,

$$\begin{aligned}
\mathcal{S} &= \int_I f(x) \, dx \\
&= \int_{g(I)} f(g^{-1}(y)) \; | \; \frac{1}{g^{'}(g^{-1}(y)} \; | \, dy \\
&= \int_0^1 \frac{1}{2} \, 2\sqrt{y} \, sin \, y \, \frac{1}{2\sqrt{y}} \\
&= \frac{1}{2} \int_0^1 sin \, y \, dy \\
&= \frac{1}{2} \, (1 - cos \, 1) \; .
\end{aligned}$$

Exercise 2.5. *(i) If $U \sim$ Uniform$(0,1)$ find the density function of $W := (-1/\lambda)\log U$. (ii) If F is a strictly increasing distribution function with density function f and $f(x) \neq 0$ for any x, find the distribution of $F(X)$ using Theorem 2.4. (iii) Let Y be a random variable defined as $Y = X1(X \leq M) + M1(X > M)$, where M is a positive number and $X \sim$ Exponential(λ). Write down the c.d.f of Y and use it to generate a random variable with its distribution on the computer.*

Note that the change of variable theorem as described above deals with one-to-one monotone transformations. However, there are natural transformations that do not fall in this category but are frequently encountered, e.g. $g(x) = x^2$ considered as a function from $\mathbb{R}$ to $[0,\infty)$. Clearly g is not one-to-one, but it is one-to-one when restricted to the positive axis or the negative axis. An extension of the above change-of-variable theorem allows us to address such scenarios.

Theorem 2.5. *Let $Y = g(X)$ where $P(X \in B) = 1$ for some open set B on which it has a p.d.f f. Let $P(Y \in \mathcal{Y}) = 1$ for some open set $\mathcal{Y}$. Consider a partition $A_0, A_1, A_2, \ldots, A_k$ of S such that for each i, $g_{|A_i}$ is a monotone one-to-one transformation between A_i and $\mathcal{Y}$ with a non-vanishing derivative. Call this map (on the restricted domain) g_i and let g_i^{-1} denote the (differentiable) inverse map. Then, for any $y \in \mathcal{Y}$,*

$$f_Y(y) = \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right| ,$$

and $f_Y(y) = 0$ for any $y \notin \mathcal{Y}$.

An extension: *Next suppose that apart from a probability 0 set, $\mathcal{Y}$ can be partitioned as $B_1 \cup B_2 \cup \ldots B_l$ such that for each B_j there is a sub-collection, denoted $\mathcal{A}_j := \{A_{ij}\}$ of $\{A_1, A_2, \ldots, A_k\}$ with the property that $g_{|A_{ij}}$ is a one-to-one transformation between A_{ij} and B_j, with corresponding inverse denoted by g_{ij}^{-1}. Note that the sub-collection $\{A_{ij}\}$ cannot have any element in common with $\{A_{ij'}\}$ for $j \neq j'$. Then, for each j, and for $y \in B_j$,*

$$f_Y(y) = \sum_{i:A_{ij} \in \mathcal{B}_j} f_X(g_{ij}^{-1}(y)) \left| \frac{d}{dy} g_{ij}^{-1}(y) \right| ,$$

and $f_Y(y) = 0$ for any $y \notin \mathcal{Y}$.

The following exercise illustrates applications of Theorem 2.5.

Exercise 2.6. *(i) Suppose $X \sim N(0,1)$. Find the density of $Y = X^2$. (ii) Let $\Theta \sim$ Uniform$(0, 2\pi)$. Let $Y = tan(\Theta)$. Find the distribution of $tan(\Theta)$.*

Hint to (ii): Partition $(0, 2\pi)$ into $(0, \pi/2) \cup (\pi/2, \pi) \cup (\pi, 3\pi/2) \cup (3\pi/2, 2\pi)$ apart from a 0 probability set. Then the tan function is a one-one function (continuously differentiable) from each of these sub-intervals to either $(0,\infty)$ (for the first and third intervals) or $(-\infty, 0)$ (for the second and fourth intervals) and the extension of the theorem above can be invoked directly to show that $tan(\Theta)$ has the Cauchy density.

3 Expectation, Median, Variance

3.1 Expectation

Definition 3.1. *For a discrete random variable X:*

$$E[g(X)] = \sum_x g(x)p(x)dx\,,\ \textit{provided}\ \ E[|g(X)|] < \infty\,.$$

For a continuous X with density f (i.e. absolutely continuous)

$$E(g(X)) = \int g(x)p(x)dx\ \ \textit{provided}\ \ E(|g(X)|) < \infty\,.$$

Basically, the idea is that any random variable X can be written as $X^+ - X^-$ where $X^+ = X \vee 0$ and $X^- = -(X \wedge 0)$. Next, $E(X^+)$ and $E(X^-)$ always exist in the sense that either they are finite or equal to ∞. Since $|X| = X^+ + X^-$, $E(|X|)$ is finite if and only if $E(X^+)$ and $E(X^-)$ are both finite and in this case we define $E(X) = E(X^+) - E(X^-)$.

Exercise 3.1. *Suppose X follows the Cauchy density, i.e. $f(x) = (1/\pi)\,1/(1+x^2)$ for $x \in \mathbb{R}$. Show that $E(X^+) = E(X^-) = \infty$.*

A general definition of expectation is difficult to provide in this course, for example, for r.v.'s that may be continuous and *NOT* have a density. But that will not be a big issue. Based on the above definitions, we can define an expectation for a *mixed* random variable.

We will call X a 'mixed' random variable if there exists a set $\mathcal{X} = \{x_1, x_2, x_3, \ldots\}$, a (Borel) set S, a function $p : \mathcal{X} \to (0,1)$, and a function $f : S \to [0,\infty)$ such that $P(X = x_j) = p(x_j)$, and for any (Borel) subset of S, say A, $P(X \in A) = \int_A f(u)du$.

For a mixed X, we define:

$$E(g(X)) = \sum_{x_j} g(x_j)p(x_j) + \int_S g(x)\,f(x)dx\,,$$

provided $E(|g(X)|) < \infty$, where this expectation is defined analogously to the above display.

Example 3.1. *As a simple example, consider a random variable Y with density $f_Y(y) = \lambda\, e^{-\lambda|y|}/2$ for $y \in \mathbb{R}$. This is the so-called double exponential density. Define $\tilde{X} = M1(Y \geq M) + (-M)1(Y \leq -M) + Y1(-M < Y < M)$. Then, it is not difficult to check that $\tilde{X}$ is mixed in the above sense with $\mathcal{X} = \{-M, M\}$ with $p(M) = e^{-\lambda M/2}$ and $p(-M) = e^{-\lambda M/2}$, while $S = (-M, M)$ and on S we have a density $f(s) = e^{-\lambda|y|}/2$ for $s \in S$. Also,*

$$E(\tilde{X}) = (-M)\,p(-M) + M\,p(M) + \int_{(-M,M)} ye^{-\lambda|y|/2}dy = 0\,,$$

as can be readily checked.

Exercise 3.2. *Once again, consider Y following the double-exponential density (above). Define $X = Y1(Y > 0) + [(\lfloor Y \rfloor) + 1]\,1(Y \leq 0)$. Find $E(Y)$.*

For $k \geq 1$, the k'th moment of X is defined as $E(X^k)$ (with the usual caveat that it exists finitely) and denoted μ_k. For certain distributions, calculating a slightly different kind of moment, called the *factorial moment* often facilitates calculations.

Factorial moment: For a random variable X, it's k'th factorial moment denoted $E[(X)_k]$ is defined as $E(X(X-1)(X-2)\dots(X-k+1)] \equiv E(X_{P_k})$. The k'th moment $E(X^k)$ can be expressed as

$$E(X^k) = \sum_{r=0}^{k} S(k,r)E[(X)_r] \quad \text{where} \quad S(k,r) = \frac{1}{r!}\sum_{j=0}^{r}(-1)^{r-j}\binom{r}{j}j^k\,.$$

Let's calculate the k'th factorial moment for $X \sim Bin(n,p)$. We have:

$$\begin{aligned}
E[(X)_k] &= \sum_{x=0}^{n} x(x-1)\dots(x-k+1)\binom{n}{k}p^x(1-p)^{n-x} \\
&= \sum_{x=k}^{n} x(x-1)\dots(x-k+1)\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum_{x=k}^{n} x(x-1)\dots(x-k+1)\,\frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x} \\
&= \sum_{x=k}^{n} \frac{n!}{(x-k)!}\frac{1}{(n-x)!}p^x(1-p)^{n-x} \\
&= \sum_{y=0}^{n-k} \frac{n!}{y!(n-k-y)!}\,p^{y+k}(1-p)^{n-k-y} \\
&= \frac{n!}{(n-k)!}p^k\sum_{y=0}^{n-k}\binom{(n-k)}{y}p^y(1-p)^{n-k-y} = \frac{n!}{(n-k)!}p^k\,.
\end{aligned}$$

Exercise 3.3. *Calculate the general factorial moment of a Poisson(λ) random variable.*

The Gamma family of distributions: For $\alpha \in (0,\infty)$ the Gamma function is defined as:

$$\Gamma(\alpha) = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)}\,e^{-x}x^{\alpha-1}\,dx\,.$$

Important properties of the Gamma function are: (a) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, (b) for $n \geq 1$, $\Gamma(n) = (n-1)!$, and (c) $\Gamma(1/2) = \sqrt{\pi}$.

For positive parameter (α, λ) the $Gamma(\alpha, \lambda)$ density function (which defines a unique distribution function) is given by:

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} 1(x > 0) .$$

It is not difficult to see that the above expression integrates to 1 on $(0, \infty)$ by making a change of variable $y = \lambda x$ and then using the definition of the Gamma function. A random variable X following the Gamma(α, λ) density is non-negative with probability 1. We will calculate the k'th moment of a Gamma random variable which is given by a nice closed form expression.

$$\begin{aligned} E(X^k) &= \int_0^\infty x^k \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-\lambda x} x^{\alpha+k-1} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}} \\ &= \frac{\alpha(\alpha+1)\ldots(\alpha+k-1)}{\lambda^k} . \end{aligned}$$

Change of Variable result for Expectations: We defined the expectation of a general transformation $(g(X))$ of a random variable (X) in terms of the distribution of X (albeit in restricted but useful settings like discrete, continuous with density, and mixed). The change of variable result for expectations (which we will not deal with in full generality as it requires some measure theory ideas) basically says that this definition is consistent, in the sense that if we define $Y = g(X)$ and compute the expectation of Y with respect to its distribution i.e. *without viewing it as a transformation of* X, we will get the same result. We first illustrate with discrete random variables.

Let X be a discrete random variable assuming values in $\mathcal{X}$ and let $Y = g(X)$ for a general transformation g. Then Y is a discrete random variable assuming values in the set $\mathcal{Y}$. Note that $\mathcal{X} = \cup_{y\in\mathcal{Y}} g^{-1}(\{y\})$, and that the union is disjoint.[2] If p_X and p_Y denote the p.m.f.'s of X and Y respectively, note that $p_Y(y) = \sum_{x\in g^{-1}(y)} p_X(x)$, and for every $x \in g^{-1}(\{y\}), g(x) = y$. Now,

$$E[g(X)] = \sum_{x\in\mathcal{X}} g(x)p_X(x) = \sum_{y\in\mathcal{Y}} \sum_{g^{-1}(\{y\})} g(x)p_X(x) = \sum_{y\in\mathcal{Y}} y \sum_{g^{-1}(\{y\})} p_X(x) = \sum_{y\in\mathcal{Y}} y\, p_Y(y) .$$

This shows that the expectation is well-defined irrespective of whether one uses the p.m.f of Y or views Y as a function of X and uses the p.m.f. of X.

[2]Recall that $g^{-1}(y) = \{x \in \mathcal{X} : g(x) = y\}$.

Let's deal with the version of this result in the continuous case for functions satisfying the regularity conditions of Theorem 2.4. We'll stick to the setting of a one-to-one transformation g. So, as in that theorem let $P(X \in (a,b)) = 1$ and g map (a,b) to $(c,d) = (g(a), g(b))$, i.e. we assume g is strictly increasing. Let f_X and f_Y denote the densities of X and Y respectively. Then,

$$E(Y) = \int_{(g(a),g(b))} y f_Y(y) dy = \int_{(g(a),g(b))} y\, f_X(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))}\, dy\,.$$

We now recompute the integral on the right using the change of variable $y = g(x)$, so that $x = g^{-1}(y)$, $dy = g'(x)dx$, and the domain in terms of x is (a,b). This gives:

$$E(Y) = \int_{(a,b)} g(x) f_X(x) \frac{1}{g'(x)}\, g'(x) dx = \int_{(a,b)} g(x) f_X(x) dx = E(g(X))\,,$$

by definition.
The change of variable result for expectations is valid in full generality as noted before and we will use it with that understanding in subsequent calculations.

The next result documents important properties of expectations and is essentially Theorem 2.2.5 of Casella and Berger.

Theorem 3.1. *Let X be a random variable and g_1 and g_2 be functions such that $E(|g_1(X)|)$ and $E(|g_2(X)|)$ are both finite and let a, b, c be constants. Then:*

(i) $E\left[a g_1(X) + b g_2(X) + c\right] = a E(g_1(X)) + b E(g_2(X)) + c\,.$

(ii) If $g_1(X) \geq 0$, $E[g_1(X)] \geq 0$ and equals 0 if and only if $P(g_1(X) = 0) = 1$.

(iii) If $g_1(X) \geq g_2(X)$ on a set of probability 1, $Eg_1(X)) \geq E(g_2(X))$. As a corollary, if $a \leq g_1(X) \leq b$, then $a \leq E(g_1(X)) \leq b$.

Exercise 3.4. *Show that $E(X)$ minimizes $E[(X-a)^2]$ over $a \in \mathbb{R}$.*

3.2 Additional Material on the notion of Expectation

We formally define and validate the notion of the average or expectation of a non-negative random variable. We start with $(\Omega, \mathcal{A}, P)$ a probability space and a non-negative random variable from $X : \Omega \to \mathbb{R}$. [Generally one considers extended real valued non-negative random variables that can assume the value ∞ but we stay away from this.]
By a (non-negative) simple function, we mean a random variable S of the form $S(\omega) :=$

$\sum_{i=1}^{m} \alpha_i 1_{A_i}(\omega)$ where α_i's are non-negative numbers, the A_i's form a partition of Ω, and m is some natural number. We define:

$$E(S) := \sum_{i=1}^{m} \alpha_i P(A_i)\,.$$

Once the expectation has been defined for a simple functions, the following properties are easy to verify from basic principles.

(i) If S_1 and S_2 are two non-negative simple functions then $E(S_1 + S_2) = E(S_1) + E(S_2)$.

(i) $E(\beta S) = \beta E(S)$ for any non-negative β and simple function S.

(iii) If $S_1 \geq S_2$ are simple functions, then so is $S_1 - S_2$ and $E(S_1 - S_2) = ES_1 - ES_2$.

An easy consequence of the above is the following:

$$E\left[\sum_{j=1}^{l} \lambda_j S_j\right] = \sum_{j=1}^{l} \lambda_j\, E(S_j)\,.$$

where the S_j are simple functions and the λ_j's are non-negative numbers.

For a general non-negative X **define** E(X) as follows: Take any sequence of non-negative simple functions X_n that *increase* to X. Set $E(X) = \lim_n E(X_n)$.

For this definition to make sense, we need to establish two things: (a) That there exists at least one sequence of simple functions increasing to X, and, (b) The limit is invariant to *any* choice of increasing sequence. We first prove (b) and then demonstrate that at least one such sequence exists.

So let X_n and Y_n be two sequences of simple functions that both increase to X. Fix a $0 < t < 1$, arbitrary. Fix an $m \geq 1$ and for each n, define

$$A_n := \{\omega : X_n(\omega) \geq tY_m(\omega)\}\,.$$

It is clear that by the monotonicity of X_n, the A_n's are increasing sets and therefore $P(A_n)$ is increasing. Also, for any ω, $X_n(\omega)$ is eventually strictly greater than $tY_m(\omega)$ and therefore $\cup A_n = \Omega$. It follows that $P(A_n)$ increases to 1. Then,

$$E(X_n) \geq E(X_n 1_{A_n}) \geq E(tY_m 1_{A_n}) = tE(Y_m\, 1_{A_n})\,.$$

Note that all functions involved in the above expression are simple functions and we have therefore only used the properties of expectations of simple functions. Taking limits as $n \to \infty$, we conclude that

$$\lim_n E(X_n) \geq t \lim_{n\to\infty} tE(Y_m 1_{A_n})\,.$$

Let $Y_m = \sum_l \gamma_l 1_{B_l}$ for a finite partition $\{B_l\}$ of Ω and non-negative numbers γ_l. Then,

$$E(Y_m 1_{A_n}) = \sum_l \gamma_l P(B_l \cap A_n) \to_n \sum_l \gamma_l P(B_l) = E(Y_m)\,,$$

since, for each l, $B_l \cap A_n$ increases to B_l with increasing n. We conclude that

$$\lim_n E(X_n) \geq tE(Y_m)\,.$$

Now taking limit on the right side as m goes to ∞, we get $\lim_n E(X_n) \geq t \lim_m E(Y_m)$ for all $0 < t < 1$. Taking a limit now as t increases to 1, we obtain $\lim_n E(X_n) \geq \lim_m E(Y_m)$. Interchanging the roles of the X_n and Y_m gives the reverse inequality, which shows the invariance of the limit to the sequence of increasing functions chosen.

To address point (a) above, define

$$X_k(\omega) = \sum_{j=0}^{k2^k-1} \frac{j}{2^k} 1\{X(\omega) \in [j/2^k, (j+1)/2^k)\} + k\, 1(X(\omega) \geq k)\,.$$

Check by direct verification that X_k increases to X.

Finally, it can be checked that EX as defined above has an alternative simple characterization: Let $\mathcal{S}_X$ denote the class of simple functions S such that $0 \leq S \leq X$. Then

$$EX = \sup_{\mathcal{S}_X} ES\,.$$

To see this, show that given any simple function S that is no larger than X, you can construct a sequence of simple functions $S_1 \leq S_2 \leq \ldots$ increasing to X, such that $S = S_1$. Use a construction similar to the construction of the X_k's above.

Finally, for a general random variable X, write $X = X^+ - X^-$ and define EX provided both EX^+ and EX^- are finite, and in that case set it as

$$EX = EX^+ - EX^-\,.$$

3.3 Medians and percentiles

Median as minimizer of L_1 distance: Consider $E(|X-a|)$ and let m be a median of the distribution of X. Consider the situation that $a \leq m$. We can write

$$E(|X-a|) \;=\; \int_{(-\infty,a)} (a-x)\, dF(x) + \int_{[a,\infty)} (x-a)\, dF(x)$$

$$
\begin{aligned}
&= aF(a-) - \int_{(-\infty,a)} x\,dF(x) + \int_{[a,\infty)} x\,dF(x) - a(1-F(a-)) \\
&= a - 2a(1-F(a-)) - \mu + 2\int_{[a,\infty)} x\,dF(x) \\
&= a + 2\int_{[a,\infty)} (x-a)\,dF(x) - \mu\,.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
E(|X-a|) - E(|X-m|) &= a - m + 2\int_{[a,\infty)} (x-a)\,dF(x) - 2\int_{[m,\infty)} (x-m)\,dF(x) \\
&= (a-m) + 2\int_{[a,m)} (x-a)\,dF(x) \\
&\qquad +2\int_{[m,\infty)} (x-a)\,dF(x) - 2\int_{[m,\infty)} (x-m)\,dF(x) \\
&= (a-m) + 2\int_{[a,m)} (x-a)\,dF(x) + 2\int_{[m,\infty)} (x-m)\,dF(x) \\
&\quad +2\int_{[m,\infty)} (m-a)\,dF(x) - 2\int_{[m,\infty)} (x-m)\,dF(x) \\
&= (a-m) + 2\int_{[a,m)} (x-a)\,dF(x) + 2\int_{[m,\infty)} (m-a)\,dF(x) \\
&= 2\,(m-a)[(1-F(m-)) - 1/2] + 2\int_{[a,m)} (x-a)\,dF(x) \\
&\geq 0,
\end{aligned}
$$

since $(1-F(m-)) \geq 1/2$ and $m \geq a$.

3.4 Variance

While the mean and the median capture the so-called 'notions of central tendency', the variance and its (less popular) counterparts capture the notion of the spread of a distribution, i.e. the extent of variability of the values assumed by a random variable. We will introduce two such notions, but deal primarily with the latter one.

Definition 3.2. *The Mean Absolute Deviation (MAD) of a random variable X is defined as $E(|X-\mu|)$. The reason behind the nomenclature is fairly clear.*

Definition 3.3. *The variance of a random variable X is defined as $Var(X) := E[(X-\mu)^2]$.*

Thus, MAD measures the average absolute distance of a random variable from its mean whilst the variance measures the average squared distance. The s.d. or standard deviation of X is simply the square root of the variance and therefore has the same units as that of the random variable. In the above definitions, we of course assume all relevant moments to exist.

Exercise 3.5. *Using the fact that the variance of a random variable is always non-negative, show that the s.d. is always at least as large as the MAD.* **Hint***: We can write var*$(Y) = E(Y^2) - (EY)^2$*. Why?)*

Exercise 3.6. *Let* $MSE(X, a) := E(X - a)^2$ *denote the mean squared error of* X *about the point* a*. Show that the best possible* a*, i.e. the best constant approximation to the mean is* EX*.*

Exercise 3.7. *If* X *is a random variable assuming values in* $[c, d]$ *then Var*$(X) \leq (d-c)^2/4$*.* **Hint:** $|X - (c + d)/2| \leq (d - c)/2$*.*

The following property is easy to prove but extremely useful:

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)\,.$$

4 Convergence in Distribution, Monotone and Dominated Convergence Theorems, Moment Generating Functions

The notion of convergence in distribution formally quantifies the idea of a sequence of random variables (which could be updated estimates of some feature of a population based on increasingly more available data) whose distribution functions start increasingly resembling the distribution of a fixed random variable. This is reflected in the following definition.

Definition 4.1. *A sequence* X_n *of random variables is said to converge in distribution to a (limiting) random variable* X *if for every* x *such that* $F(x-) = F(x)$ *where* F *is the cdf of* X *(such* x*'s, recall, are precisely the continuity points of* X*) we have* $\lim_n F_n(x) = F(x)$*, where* F_n *is the cdf of* X_n*.*

The important thing to note here is that the convergence is only desired for the continuity points of F. At points where F is discontinuous (and these cannot be too many, only a countable set at most) the point wise convergence of F_n may not hold! Indeed, this *failure* is essential to make distributional convergence meaningful as a notion. Here is an exercise that illustrates this fact.

Exercise 4.1. *Suppose that the sequence of real numbers x_n converges to x. Let X_n be the random variable that assumes the value x_n with probability 1, and let X be defined likewise. Show that X_n converges in distribution to X but that $F_n(x)$ might not converge to $F(x)$.*

A sequence of discrete random variables may converge to a continuous random variable in distribution and vice-versa.

Exercise 4.2. *Let U_n be a discrete random variable that assumes the values $\{1/n, 2/n, \ldots, 1\}$ each with probability $1/n$. Then U_n converges to the uniform distribution on $[0, 1]$. Draw a diagram illustrating this convergence.*

Exercise 4.3. *This result is often called* **Polya's theorem***: Suppose X_n converges in distribution to X and that X is continuous. Then show that F_n converges uniformly to F, i.e $\sup_{x \in R} |F_n(x) - F(x)| = 0$.*

4.1 Convergence theorems

The following two convergence theorems are of central importance in probability and statistics but proving them in a general sense is outside the scope of this course. Nevertheless, we shall use these results. We first present the monotone convergence theorem (MCT).

Theorem 4.1. *Let X_n be a sequence of non-negative random variables increasing a.s. (almost surely) to a random variable X (which, in particular, could take the value ∞ on a set of non-zero probability).*[3]*. Then EX_n increases to EX which could possibly be ∞.*

The MCT leads to the Dominated Convergence Theorem (DCT).

Theorem 4.2. *Let X_n be a sequence of random variables converging a.s. to a random variable X, and let Z be such that with probability 1 $|X_n| \leq Z$ for all n (and therefore $|X| \leq Z$ with probability 1). If $EZ < \infty$, then $EX_n \to EX$ and all these expectations are finite.*

Discussion on the monotone convergence theorem: Based on the discussion in Section 3.2, the proof of the monotone convergence theorem is relatively easy. Since X_n's increase to X, it is clear that $\lim E(X_n) \leq EX$. We would like to establish the converse inequality.

So let Y be any simple function that lies below X. Fix a $0 < t < 1$, arbitrary. For each n, define

$$A_n := \{\omega : X_n(\omega) \geq tY(\omega)\}.$$

[3]The notation a.s. means that on a set of probability 1, $X_n(\omega)$ increases to $X(\omega)$.

It is clear that by the monotonicity of X_n, the A_n's are increasing sets and therefore $P(A_n)$ is increasing. Also, for any ω, $X_n(\omega)$ is eventually strictly greater than $tY(\omega)$ and therefore $\cup A_n = \Omega$. It follows that $P(A_n)$ increases to 1. Then,

$$E(X_n) \geq E(X_n 1_{A_n}) \geq E(tY 1_{A_n}) = tE(Y\, 1_{A_n})\,.$$

Taking limits as $n \to \infty$, we conclude that

$$\lim_n E(X_n) \geq t \lim_{n\to\infty} tE(Y 1_{A_n})\,.$$

Let $Y = \sum_l \gamma_l 1_{B_l}$ for a finite partition $\{B_l\}$ of Ω and non-negative numbers γ_l. Then,

$$E(Y 1_{A_n}) = \sum_l \gamma_l P(B_l \cap A_n) \to_n \sum_l \gamma_l P(B_l) = E(Y)\,,$$

since, for each l, $B_l \cap A_n$ increases to B_l with increasing n. We conclude that $\lim_n E(X_n) \geq tE(Y)$. As this is true for all $0 < t < 1$, we conclude that $\lim_n E(X_n) \geq E(Y)$. But as EX is the supremum of the expectations of all simple functions Y that lie below X, it follows that $\lim_n E(X_n) \geq E(X)$. This completes the proof.

4.2 Moment generating functions

Given a random variable X, its moment generating function is defined as $M_X(t) = E[e^{tX}]$ for $t \in R$. We say that the m.g.f (moment generating function) of X exists if $M_X(t)$ is finite in an open interval $(-h_0, h_0)$ around 0. Note that $M_X(0) = 1$ for any random variable X. In terms of the distribution function F:

$$M_X(t) = E(e^{tX}) = \int_{\mathbb{R}} e^{tx}\, dF(x)\,.$$

Note that for any $0 < t < h_0$,

$$e^{t|X|} = e^{tX} 1(X > 0) + e^{-tX} 1(X \leq 0)$$

and as the right side has finite expectation, we conclude that $E(e^{t|X|}) < \infty$. Note that this inequality is trivially true for all negative t.

Moment generating functions do not exist for large families of random variables, because the existence of mgf can be shown to be tantamount to the random variable having very little probability mass or being thin 'in the tails': in fact, all random variables that possess an mgf are *sub-exponential* which roughly means that for large enough $b > 0$, the chance that the random variable is larger in absolute value than b is smaller than the chance that some exponential random variable is larger than b in absolute magnitude. As you can

imagine this is quite stringent. Nevertheless, large families of random variables which arise in various problems in probability, statistics and machine learning have these properties and moment generating functions can then be used to derive various powerful results about the suprema of families of sub-exponential random variables. The modern theory of concentration inequalities relies heavily on moment generating functions. Apart from that, moment generating functions when they exist have a one to one correspondence with distribution functions and can be used to characterize distributional convergence [results below].

Theorem 4.3. *If X has m.g.f $M_X(t)$, then*

$$E(e^{tX}) = \sum_{j=0}^{\infty} \frac{t^j E(X^j)}{j!} .$$

Furthermore,

$$E(X^n) = M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t) \,|_{t=0} .$$

Proof: Let $S_n = \sum_{j=0}^{n} t^j X^j/j!$. Then $S_n \to e^{tX}$, and $|S_n| \leq e^{|t||X|}$ with $E(|t||X|) < \infty$. By the DCT $E(S_n) \to E(e^{tX})$. But $\lim E(S_n) = \sum_{j=0}^{\infty} t^j E(X^j)/j!$
To prove the second part, note that if

$$\frac{d^n}{dt^n} M_X(t) \equiv \frac{d^n}{dt^n} E(e^{tX}) = E\left[\frac{d^n}{dt^n} e^{tX}\right] \equiv E[X^n e^{tX}] ,$$

then, the result follows by evaluating the n'th derivative at $t = 0$. The only step that requires justification here is that the differentiation can be moved to within the expectation, and this will be accomplished by an application of the DCT. We demonstrate this with $n = 1$, and the general proof follows along similar lines. So, consider any sequence $h_n \to 0$ and note that

$$\frac{d}{dt} E(e^{tX}) = \lim_{n\to\infty} \frac{E(e^{(t+h_n)X}) - E(e^{tX})}{h_n} = \lim_{n\to\infty} E\left[e^{tX} \frac{e^{h_n X} - 1}{h_n}\right] .$$

Now,

$$\lim_n e^{tX} \frac{e^{h_n X} - 1}{h_n} = X e^{tX} .$$

Since h_n goes to 0, $|h_n|$ eventually less than $0 < c_0$ where c_0 is chosen sufficiently small such that $c_0 + |t| < h_0$. Thus, for all sufficiently large n,

$$\begin{aligned} \left| e^{tX} \frac{e^{h_n X} - 1}{h_n} \right| &\leq e^{|t||X|} \left| \frac{e^{h_n X} - 1}{h_n} \right| \\ &= \left| X + \frac{h_n X^2}{2!} + \frac{h_n^2 X^3}{3!} + \ldots \right| e^{|t||X|} \end{aligned}$$

$$
\begin{aligned}
&\leq \left(|X| + \frac{c_0\,|X|^2}{2!} + \frac{c_0^2|X|^3}{3!} + \ldots \right| e^{|t||X|} \\
&= \left| \frac{e^{c_0|X|}-1}{c_0} \right| e^{|t||X|} \\
&= \frac{e^{(|t|+c_0)X} + e^{|t|X}}{c_0}\,.
\end{aligned}
$$

The bounding random variable on the right side of the equality sign in the last step is integrable (i.e. has finite expectation), and therefore by DCT,

$$
\frac{d}{dt}\,E(e^{tX}) = \lim_{n\to\infty}\; E\left[e^{tX}\,\frac{e^{h_nX}-1}{h_n} \right] = E(Xe^{tX})\,.
$$

Some standard MGF computations: We start with $N(\mu,\sigma^2)$. Suppose $Z \sim N(0,1)$. Then $Y = \mu + \sigma Z$. First compute m.g.f. of Z. Then use the fact that if $Y = a + bX$, $M_Y(t) = e^{ta}\,M_X(bt)$ to finish the calculation. So,

$$
\begin{aligned}
E(e^{tZ}) &= \int_{-\infty}^{\infty} \frac{1}{2\pi}\,e^{tz}e^{-t^2/2}dz \\
&= e^{t^2/2}\int_{-\infty}^{\infty} \frac{1}{2\pi}\,e^{-(z-t)^2/2}dz \;\; ;\text{completing the square} \\
&= e^{t^2/2}\,.
\end{aligned}
$$

It follows that $E(e^{tY}) = \exp(t\mu + \sigma^2t^2/2)$.

Exercise 4.4. *Calculate the m.g.f's of the Binomial (n,p), Poisson (λ) and $\Gamma(\alpha,\lambda)$ random variables.*

Let's do the calculation for the Gamma. If $X \sim \Gamma(\alpha,\lambda)$,

$$
\begin{aligned}
E(e^{tX}) &= \int_0^{\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)}\,e^{tx}\,e^{-\lambda x}\,x^{\alpha-1}\,dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^{\infty} e^{-(\lambda-t)x}\,x^{\alpha-1}dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)}\,\frac{\Gamma(\alpha)}{(\lambda-t)^\alpha} \\
&= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha},
\end{aligned}
$$

showing that the mgf exists for all $t \in (-\lambda,\lambda)$.
As indicated above, MGFs are quite useful for characterizing distributions and deducing

distributional convergence. Indeed, we will prove (a restricted) version of the CLT using moment generating functions later on. Theorems 2.3.11 and 2.3.12 from the text are critical in this regard.

Theorem 4.4. Equivalent of Theorem 2.3.11 (CB): *Suppose X and Y are two random variables whose moment-generating functions exist in an open neighborhood around 0. Then $F_X = F_Y$ if and only if $M_X(t) = M_Y(t)$ for all t in this open neighborhood.*

We will not prove this theorem as it requires some heavy duty analytical tools which will not be particularly useful for this course, but the following corollary is easy to establish.

Corollary 4.5. *Let X and Y be bounded random variables, so that their m.g.f.s exist on the entire line. Then X and Y have the same distribution if all of their moments coincide.*

We avoid writing a formal proof, but as X and Y are bounded random variables, for all $t \in R$,

$$E(e^{tX}) = \sum_{j=0}^{\infty} \frac{t^j}{j!} E(X^j) \text{ and } E(e^{tY}) = \sum_{j=0}^{\infty} \frac{t^j}{j!} E(Y^j) .$$

From the assumption that $E(X^j) = E(Y^j)$ for all $j \geq 1$, it follows that the mgf's of X and Y are the same function and hence X and Y must have the same distribution by the preceding theorem.

The next result asserts that convergence of moment generating functions implies convergence of distributions.

Theorem 4.6. *Let X_n be a sequence of random variables and let X be a fixed random variable. Assume that the mgf's of all these random variables exist finitely in an open interval around 0, and further that for each t in some open interval around 0 (typically smaller than the previous neighborhood) $M_n(t) \to M(t)$, where M_n is the mgf of X_n and M that of X. Then $X_n \to_d X$.*

DeMoivre-Laplace theorem: We can use the above result to deduce the DeMoivre-Laplace theorem which asserts that the number of heads in n independent coin flips converges to a normal distribution under adequate normalization. So, let $S_n \sim \text{Bin}(n, p)$. Then

$$\tilde{S}_n := \frac{S_n - np}{\sqrt{npq}} \to_d N(0, 1) .$$

Proof: Let $M_n(t) := E(e^{t\tilde{S}_n})$. Then

$$M_n(t) \;=\; E\left[\exp\left[t\left(\frac{S_n - np}{\sqrt{np(1-p)}}\right)\right]\right]$$

$$
\begin{aligned}
&= E\left[\exp\left(\frac{tS_n}{\sqrt{np(1-p)}} - \frac{npt}{\sqrt{np(1-p)}}\right)\right] \\
&= \left(e^{-tp/\sqrt{np(1-p)}}\right)^n \left[(1-p) + p\, e^{t/\sqrt{np(1-p)}}\right]^n \text{ Binomial mgf} \\
&= \left[pe^{(t/\sqrt{n})\sqrt{(1-p)/p}} + (1-p)e^{(-t/\sqrt{n})\sqrt{p/(1-p)}}\right]^n .
\end{aligned}
$$

Now, expanding this last expression in an infinite series, we get:

$$
p\left(1 + \frac{t}{\sqrt{n}}\sqrt{\frac{1-p}{p}} + \frac{t^2}{2n}\,\frac{1-p}{p} + \frac{t^3}{3!\,n^{3/2}}\,\left(\frac{1-p}{p}\right)^{3/2} + \ldots\right)
$$

$$
+(1-p)\left(1 - \frac{t}{\sqrt{n}}\sqrt{\frac{p}{1-p}} + \frac{t^2}{2n}\,\frac{p}{1-p} - \frac{t^3}{3!\,n^{3/2}}\,\left(\frac{p}{1-p}\right)^{3/2} + \ldots\right)
$$

which simplifies to

$$
\left\{p + \frac{t}{\sqrt{n}}\sqrt{p(1-p)} + \frac{t^2}{2n}(1-p) + (1-p) - \frac{t}{\sqrt{n}}\sqrt{p(1-p)} + \frac{t^2}{2n}p + O(1/n^{3/2})\right\}^n .
$$

But this is simply

$$
\left(1 + \frac{t^2}{2n} + o(1/n)\right)^n \to_{n\to\infty} e^{t^2/2}
$$

and as the limit is the mgf of $N(0,1)$ this completes the proof.

This result permits an extension which is stated below and can be established by similar techniques.

Exercise 4.5. *Suppose* $S_n \sim Bin(n, p_n)$ *such that* $np_n(1-p_n) = Var(S_n) \to \infty$. *Then* $(S_n - np_n)/\sqrt{np_n(1-p_n)} \to_d N(0,1)$.

A different 'Poissonian' behavior is obtained when the variance of the Binomial stabilizes in the limit. So, let $S_n \sim Bin(n, p_n)$ where $np_n \to \lambda$ and therefore so does $\mathrm{Var}(S_n)$. This limiting regime can be used to model situations where we see an enormous number of independent events, but only a very small proportion of them are interesting and the average number of such interesting events is a modest number: for example, thousands of individuals being tested for a very rare (and consequential) genetic mutation and maybe out of a million people 7 will carry the mutation. Then we have that $S_n \to_d V$ where $V \sim \mathrm{Poisson}(\lambda)$. As before, we show this by calculating the limiting mgf of S_n.

We have:

$$\begin{aligned} M_n(t) &= E(e^{tS_n}) \\ &= [(1-p_n)+p_n e^t]^n \\ &= \left[1+\frac{x_n}{n}\right]^n , \end{aligned}$$

where $x_n = np_n(e^t-1) \to \lambda(e^t-1)$. Therefore

$$\lim_n M_n(t) = \exp(\lambda(e^t-1))$$

which is precisely the mgf of Poisson(λ).

5 Various Distributions

Pages 85–98 of CB's book discuss a variety of one-dimensional distribution. Here, we discuss the Hypergeometric Distribution. The setting of the Hypergeometric ties naturally to what is called *sampling without replacement* from a finite population.
Consider a box with a total number of N balls out of which M are red and the remaining $N-M$ are green. We will pick K distinct balls from the box at random. The way to interpret this is as any set of K balls from the box being as equally likely to be drawn as any other. Let X be the number of red balls. We wish to find the probability mass function of X: what is $P(X=x)$ for $0 \le x \le K \wedge M$?
Think of the balls as being marked1 through N. Now, the total number of distinct ways in which you can pick K distinct balls out of N is $\binom{N}{K}$. For X to equal x, x balls must be red, and there are $\binom{K}{x}$ distinct groups, whilst for the remaining $K-x$ balls there are. $\binom{(N-M)}{(K-x)}$ distinct groups. Thus there are $\binom{K}{x}\binom{(N-M)}{(K-x)}$ distinct groups of K balls that entail x red balls. Since all groups of K balls are equally likely, it follows that

$$P(X=x) = \frac{\binom{K}{x}\binom{(N-M)}{(K-x)}}{\binom{K}{x}} .$$

Instead of the box analogy, we can also think of an election analogy where the box is the total number of eligible voters with M Democrats and $N-M$ Republicans. A sample of K distinct individuals are picked at random from the population. Then X is the random number of D's in the collected sample. This quantity is useful to estimate proportion of D's. Our estimate of $p := M/N$ is given by.

$$\hat{p} = \frac{X}{K} .$$

Exercise 5.1. *Show that* $E\hat{p} = p$.

As you can see, all one needs to do is calculate EX based on its pmf. There are other ways of approaching this problem by viewing X as the sum of identically distributed random variables which gives another perspective on the problem but for that we'll have to wait until we learn about joint distributions.

Exercise 5.2. *If* X *follows the Hypergeometric* (N, M, K) *distribution, show that*

$$EX = \frac{KM}{N} \quad \textit{and} \quad Var(X) = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}\,.$$

Think of an asymptotic setting where N and M are both allowed to increase so that M/N converges to a number $0 < p < 1$, and K is *fixed*. In this case, the Hypergeomteric (N, M, K) distribution converges to the $\text{Bin}(K, p)$ distribution. Verify this directly by showing that

$$P(X = m) \to \binom{K}{m} p^m (1-p)^{K-m}\,.$$

Capture-Recapture and Hypergeometric Distribution: Here is a simple way of estimating a species of fish in a lake. Suppose N is the total number of such fish. A total of M such fish are captured and tagged at various locations, and released back into the lake. After a short time, a total number of K such fish are recaptured (say $K < M$) at random and the number X of tagged fish are counted. Since $EX = KM/N$, we can estimate N as KM/X.

6 Random Vectors and Multivariate Distributions

As before let Ω be the sample space of all possible outcomes, with a generic outcome being denoted ω, equipped with a sigma-field $\mathcal{A}$, and a probability P.
A d-dimensional random vector $\underline{X} = (X_1, X_2, \ldots, X_d)$ is a measurable map from Ω to R^d in the sense that for every Borel set $B \in R^d$, $\underline{X}^{-1}(B) \in \mathcal{A}$. The Borel sigma field on R^d is the (smallest) σ-field containing all the open d-dimensional rectangles of the form $(a_1, b_1) \times (a_2, b_2) \times \ldots \times (a_d, b_d)$. The condition that the inverse image of all Borel sets under $\underline{X}$ lie in $\mathcal{A}$ is equivalent to the condition that the inverse images of all the open d-dimensional rectangles above lie in $\mathcal{A}$.
Consider now the distributional measure of X, denoted P_X by the relation $P_X(B) = P(X^{-1}(B))$ for all B in $B \in R^d$. Then, P_X is determined completely by F_X, the multivariate distribution function of X, which is defined on R^d as:

$$F_x(x_1, x_2, \ldots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)\,.$$

Elementary Example: Consider a dice thrown by someone, twice in succession. Here $\Omega = \{(i,j) : 1 \leq i,j \leq 6\}$. As this is a discrete state space, the σ-field $\mathcal{A}$ is the set of all subsets of Ω. Define the random vector (X_1, X_2) as $X_1(\omega) = i+j$ and $X_2(\omega) = |i-j|$, where $\omega = (i,j)$. Or we can define $\tilde{X}_1(\omega) = i$ and $\tilde{X}_2(\omega) = j$.
As far as P is concerned any function $p : \{1,\ldots,6\}\times\{1\ldots 6\} \mapsto [0,1]$ such that $\sum_{i,j}, p(i,j) = 1$ gives a valid probability on the state space. One particular, and natural choice is $p(i,j) = 1/36$ which, identifying $P(\{i,j\})$ with $p(i,j)$ means that all possible outcomes are equally likely. Let's investigate the distribution of P_{X_1,X_2} under this choice of p.

$$P_{X_1,X_2}\{11,0\} = P\left[(X_1,X_2)^{-1}(11,0)\right] = P\left[\{(i,j) : i+j = 11, |i-j| = 0\}\right] = 0\,.$$

In fact the probability mass of any $(m,0)$ where m is an odd number is 0 under P_{X_1,X_2}. Let $\mathcal{S}_{X_1,X_2}$ denote the support of P_{X_1,X_2}, i.e. all pairs in the range of (X_1,X_2) that have non-zero probability mass under P_{X_1,X_2}. Then,

$$\mathcal{S}_{X_1,X_2} = \{(i+j,|i-j|) : 1 \leq i,j \leq 6\}\,.$$

Let $\mathcal{S}_{X_1}$ denote the support of X_1, i.e. all integers k such that $P(X_1 = k) \neq 0$. Let P_{X_1} denote the distribution of X_1. We call this the *marginal distribution of* X_1. Let's compute the marginal p.m.f of this distribution, which we write as p_{X_1}.

$$\begin{aligned} p_{X_1}(8) &= P\{(i,j) : i+j = 8\} \\ &= P\{(2,6),(6,2),(3,5),(5,3),(4,4)\} \\ &= \frac{5}{36}\,. \end{aligned}$$

A full computation of the marginal distribution is easy but tedious. Similarly, the marginal of X_2 can be computed. The function p_{X_1,X_2} from S_{X_1,X_2} to $[0,1]$ is called the joint p.m.f of (X_1,X_2) if for each $(x_1,x_2) \in S_{X_1,X_2}$,

$$p_{X_1,X_2}(x_1,x_2) = P(X_1 = x_1, X_2 = x_2) = P_{X_1,X_2}\{x_1,x_2\}\,.$$

Of course the joint pmf p_{X_1,X_2} can be considered as defined on the larger set $\{1,2,\ldots,6\}^2$ in which cases it takes the value 0 for all pairs of integers outside its support.

Another important notion is that of the conditional distribution of one random variable given the other: for example, in this case, the distribution of X_1 given $X_2 = x_2$. This is captured by the notion of conditional pmfs. Thus, the conditional pmf of X_1 given $X_2 = x_2$ is:

$$p_{X_1|X_2=x_2}(x_1) = P(X_1 = x_1 \mid X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)}\,.$$

Let's compute.

$$\begin{aligned} p_{X_1|X_2=2}(8) &= P(X_1 = 8 \mid X_2 = 2) \\ &= \frac{P(X_1 = 8, X_2 = 2)}{P(X_2 = 2)} \\ &= \frac{2/36}{8/36} = \frac{1}{4}\,. \end{aligned}$$

To see that the denominator is $8/36$ note that there are 8 pairs (i, j) such that $|i - j| = 2$: $\{(1,3),(3,1),(2,4),(4,2),(5,3),(3,5),(4,6),(6,4)\}$. Only two pairs out of these eight satisfy $i + j = 8$.
The marginal p.m.f's X_1 and X_2 can be recovered systematically from the joint. So, for $x_1 \in \mathcal{S}_{X_1}$:

$$\begin{aligned} p_{X_1}(x_1) &= P(X_1 = x_1) \\ &= \sum_{\mathcal{X}_2} P(X_1 = x_1, X_2 = x_2) \\ &= \sum_{\mathcal{X}_2} p_{X_1,X_2}(x_1, x_2) \\ &= \sum_{x_2:(x_1,x_2)\in\mathcal{S}_{X_1,X_2}} p_{X_1,X_2}(x_1, x_2)\,. \end{aligned}$$

Independent random variables: Consider the random vector $\underline{X} = (X_1, X_2, \ldots, X_d)$. We say that $X_1, X_2, \ldots, X_d$ are *mutually independent* if:

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_d \in A_d) = \prod_{i=1}^{d} P(X_i \in A_i)\,,$$

where each A_i is a Borel subset of $\mathbb{R}$.
Denoting by $P_{\underline{X}}$ the joint distributional measure of $(X_1, \ldots, X_d)$, this can be written as:

$$P_{\underline{X}}(A_1 \times A_2 \times \ldots \times A_d) = P_{X_1}(A_1) \times \ldots \times P_{X_d}(A_d)\,.$$

When the X_i's are all discrete random variables, this is equivalent to:

$$p_{\underline{X}}(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p_{X_i}(x_i)\,,$$

where $x_i \in \mathcal{X}_i$, $\mathcal{X}_i$ being the set of numbers in which X_i assumes values. This is quite easy to prove: let's look at the the simplest case when $d = 2$. The factoring of the joint pmf as

the product of the marginal pmfs of the X_i's when the X_i's are independent according to the definition we started with is immediate (for any d). For the converse, consider sets A_2 and A_2 contained in $\mathcal{X}_1$ (the range of X_1) and $\mathcal{X}_2$ (the range of X_2) respectively. We have:

$$\begin{aligned}
P(X_1 \in A_1, X_2 \in A_2) &= \sum_{(x_1,x_2)\in A_1\times A_2} p_{X_1,X_2}(x_1,x_2) \\
&= \sum_{(x_1,x_2)\in A_1\times A_2} p_{X_1}(x_1)p_{X_1}(x_1)p_{X_2}(x_2) \\
&= \sum_{x_1\in A_1)\sum_{\{x_1}\times A_2} p_{X_1}(x_1)p_{X_2}(x_2) \\
&= \sum_{x_1\in A_1)} p_{X_1}(x_1)\left(\sum_{x_2\in A_2} p_{X_2}(x_2)\right) \\
&= P_{X_1}(A_1)P_{X_2}(A_2)\,.
\end{aligned}$$

Multinomial Distribution: The Multinomial Distribution is a natural generalization of the binomial distribution to random experiments where each run of the experiment produces a finite number of discrete outcomes. So, consider a random experiment that, in any particular realization produces $K \geq 2$ distinct outcomes. An easy example is a random throw of an evenly loaded die where $K = 6$. Let p_j denote the probability of the j'th outcome materializing (for the die example $1 \leq j \leq 6$ and $p_j = 1/6$). Then $0 < p_j < 1$ and $\sum_{j=1}^{K} p_j = 1$.

Consider n independent runs of the random experiment. At each stage the sample space $\mathcal{X} = \{1, 2, \ldots, K\}$ with $P(\{j\}) = p_j$ for each j. The sample space corresponding to the n independent runs, say $\mathcal{X}$, is the set of all sequences of length K of the form $(\eta_1, \eta_2, \ldots, \eta_n)$ where each η_i assumes values between 1 and K, and therefore has cardinality K^n. The probability of any generic sequence is given by the product probability of the individual elements in the sequence, i.e. we are looking at the n fold product probability space:

$$P(\{\eta_1, \eta_2, \ldots, \eta_n\}) = \prod_{i=1}^{n}\left(\sum_i 1(\eta_i = j)p_j\right)\,. \tag{3}$$

Now, define random variables $N_1, N_2, \ldots, N_K$ where

$$N_j(\{\eta_1, \eta_2, \ldots, \eta_n\}) = \sum_{i=1}^{n} 1(\eta_i = j)\,,$$

i.e. N_j is the number of times j appears in the sequence.
The random vector $(N_1, N_2, \ldots, N_K)$ is said to follow the *multinomial distribution* with

parameters $(n, p_1, p_2, \ldots, p_K)$. We will now compute the joint probability mass function of the multinomial random vector, i.e. compute

$$P(N_1 = n_1, N_2, = n_2, \ldots N_K = n_K)$$

for a generic K tuple of non-negative integers $(n_1, n_2, \ldots, n_K)$ that sum up to n. Now, note that for any sample point $(\eta_1, \eta_2, \ldots, \eta_n)$ for which $N_j = n_j$, by Equation (3),

$$P(\{\eta_1, \eta_2, \ldots, \eta_n\}) = \prod_{i=1}^{k} p_j^{n_j} .$$

We next need to count how many such distinct sequences there are. This is simply the total number of distinct permutation of the symbols $1, 2, \ldots, K$ that are of length n, and it follows by the law of additivity of probabilities that:

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_K = n_K) = \frac{n!}{\prod_{j=1}^{K} n_j!} \prod_{i=1}^{k} p_j^{n_j} .$$

Observe that we can write:

$$(N_1, N_2, \ldots, N_k) := \left(\sum_{i=1}^{n} N_{i,1}, \sum_{i=1}^{n} N_{i,2}, \ldots, \sum_{i=1}^{n} N_{i,K} \right) ,$$

where $(N_{i,1}, N_{i,2}, \ldots, N_{i,K})$ is the random vector that gives the counts for the K different outcomes in the i'th replication of the random experiment. Clearly, only one of the $N_{i,j}$'s can be one, and the rest have to be equal to 0. For each i, $(N_{i,1}, N_{i,2}, \ldots, N_{i,K})$ follows Multinomial$(1, p_1, p_2, \ldots, p_K)$, and these random vectors are independent across the i's.

How do we find *lower dimensional marginals of the multinomial distribution*? Without loss of generality, what is the distribution of $(N_1, N_2, \ldots, N_M)$ where $M < K$? Notice that as there are a totality of n outcomes $\tilde{N}_{M+1} = n - \sum_{j=1}^{M} N_j$ must be the number of outcomes of a type different from the first M, say the complementary outcome. At each run of the experiment, either one of the first M outcomes must occur with the j'th having probability p_j, or the complementary outcome with probability $\tilde{p}_{M+1} \equiv 1 - p_1 - \ldots - p_M$ must transpire. But then this is clearly a multinomial experiment with a reduced number of outcomes, and we conclude that:

$$(N_1, N_2, \ldots, N_M, \tilde{N}_{M+1}) \sim \text{Multinomial } (n, p_1, p_2, \ldots, p_M, \tilde{p}_{M+1}) .$$

Finally, let's talk about the *conditional distribution* of $(N_{M+1}, \ldots, N_K)$ conditional on $(N_1, N_2, \ldots, N_M)$ or equivalently $(N_1, N_2, \ldots, N_M, \tilde{N}_{M+1})$. This is simply a matter of finding

a conditional mass function. We seek:

$$P\left(N_{M+1}=n_{M+1},\ldots,N_K=n_K \mid N_1=n_1,\ldots N_M=n_M, \tilde{N}_{M+1}=n-\sum_{j=1}^{M} n_j\right).$$

The conditional probability is of course 0 if $\sum_{j=M+1}^{K} n_j \neq n-\sum_{j=1}^{M} n_j$, so we will assume below that $\sum_{j=M+1}^{K} n_j \neq n-\sum_{j=1}^{M} n_j$. Then, the above conditional probability is:

$$\frac{\frac{n!}{\prod_{j=1}^{K} n_j!}\, p_1^{n_1}\ldots p_M^{n_M} p_{M+1}^{n_{M+1}}\ldots p_K^{n_K}}{\frac{n!}{\left(\prod_{j=1}^{M} n_j!\right)\,(n-\sum_{j=1}^{M} n_j)!} p_1^{n_1}\ldots p_M^{n_M}\,(1-p_1-\ldots-p_M)^{n-\sum_{j=1}^{M} n_j}}.$$

This simplifies to:

$$\frac{(n-(n_1+n_2+\ldots n_M))!}{n_{M+1}!n_{M+2}!\ldots n_K!}\left(\frac{p_{M+1}}{1-p_1-\ldots-p_M}\right)^{n_{M+1}}\ldots\ldots\left(\frac{p_K}{1-p_1-\ldots-p_M}\right)^{n_K},$$

which is again clearly a multinomial distribution.

Continuous random vectors: The basics of dealing with a continuous two-dimensional random vector (X,Y) are well articulated in Section 4.1 of Casella and Berger (which also deals with the basics of two-dimensional discrete random vectors) and have been discussed in class. So, let (X,Y) be a 2–dimensional random vector with joint density $f(x,y)$. Then, for any Borel subset A of $\mathbb{R}^2$,

$$P((X,Y)\in A)=\int_A f(x,y)dxdy$$

where the above integral is to be interpreted as a bivariate integral in general but can be computed iteratively as well. The bivariate distribution function is related to the bivariate density function via:

$$F_{X,Y}(x,y)\equiv P(X\leq x, Y\leq y)=\int_{(-\infty,x]\times(-\infty,y])} f(u,v)dudv=\int_{-\infty}^{x}\left(\int_{-\infty}^{y} f(u,v)dv\right)du\,.$$

The marginal densities of X and Y respectively are:

$$f_X(x)=\int_{\mathbb{R}} f(x,y)dy \;\text{ and }\; f_Y(y)=\int_{\mathbb{R}} f(x,y)dx\,.$$

We next *seek to define conditional probabilities of the type* $P(Y\in(a,b]\mid X=x)$. The problem with the usual definition that uses the Bayes' formula is that we end up writing

$P(Y \in (a,b], X = x)/P(X = x)$, but this quantity is undefined as both numerator and denominator are 0. To arrive at a meaningful definition, we use a limiting argument.

Consider $h > 0$ and define:

$$\begin{aligned}\mathbb{P}(h,y,x) &= P(Y \le y \,|\, X \in (x-h, x+h]) \\ &= \frac{P(Y \le y, X \in (x-h,x+h])}{P(X \in (x-h,x+h])} = \frac{\int_{x-h}^{x+h}\int_{-\infty}^{y} f(u,v)\,dv\,du}{\int_{x-h}^{x+h} f_X(u)\,du}\,,\end{aligned}$$

where $f_X(u) = \int f(u,v)\,dv$ is the marginal density of X. Define $G(u,y) = \int_{-\infty}^{y} f(u,v)\,dv$, and consider:

$$\lim_{h\to 0}\mathbb{P}(h,y,x) = \frac{\lim_{h\to 0}\int_{x-h}^{x+h} G(u,y)\,du}{\lim_{h\to 0}\int_{x-h}^{x+h} f_X(u)\,du}\,.$$

If q is a continuous function, then

$$\lim_{h\to 0}\frac{\int_{x-h}^{x+h} q(u)\,du}{2h} = q(x)\,,$$

since, in terms of $Q(x) = \int_{-\infty}^{x} q(t)\,dt$, the above display can be written as:

$$\begin{aligned}\lim_{h\to 0}\frac{Q(x+h)-Q(x-h)}{2h} &= \lim_{h\to 0}\frac{1}{2}\frac{Q(x+h)-Q(x)}{h} + \lim_{h\to 0}\frac{1}{2}\frac{Q(x-h)-Q(x)}{-h} \\ &= \frac{1}{2}q(x) + \frac{1}{2}q(x) = q(x)\,,\end{aligned}$$

using the definition of the derivative as the limit of difference quotients and the fact that q is the derivative of Q. Using the above fact, we conclude that:

$$\mathbb{P}(y,x) \equiv \lim_{h\to 0}\mathbb{P}(h,y,x) = \frac{\lim_{h\to 0}(1/2h)\int_{x-h}^{x+h} G(u,y)\,du}{\lim_{h\to 0}(1/2h)\int_{x-h}^{x+h} f_X(u)\,du} = \frac{G(x,y)}{f_X(x)} = \int_{-\infty}^{y}\frac{f(x,v)}{f_X(x)}\,dv\,.$$

Thus, when x is fixed, $\mathbb{P}(y,x)$, viewed as a function of y is a valid distribution function. We call this the conditional distribution function of Y given $X = x$ and write it as $F_{Y\,|\,x}(y)$. The conditional distribution has a density function $f_{Y\,|\,x}(y) = f(x,y)/f_X(x)$ which is obtained by differentiating $F_{Y\,|\,x}(y)$.

Characterization of independence in terms of densities: The random variables X and Y are independent if and only if $f_{X,Y} = f_X(x)f_Y(y)$ for all (x,y) except on a set of volume (which is area in $\mathbb{R}^2$) 0.

The if part is quite easy. If the joint density splits as a product of the marginal densities, then it is an easy verification that:

$$F_{X_Y}(x,y) = P(X \le x, Y \le y) = P(X \le x)P(Y \le y) = F_X(y)F_Y(y)$$

which is equivalent to the fact that for all real Borel sets A and B,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)\,.$$

Conversely, if X and Y are independent, then

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y) = \int_{(-\infty,x]\times(-\infty,y]} (f_X(u)f_Y(v))dudv\,,$$

which shows that $f_X(u)f_Y(v)$ is a valid candidate for the joint density.

Note that X and Y are independent if and only if $f_{Y|x}(y) = f_Y(y)$ or $f_{X|Y}(x) = f_X(x)$.

Lemma 4.2.7 of CB gives another useful necessary and sufficient condition for independence of X and Y: The random variables X and Y are mutually independent if and only if $f(x,y) = g(x)h(y)$ for functions g and h. In this case g and h are multiples of $f_X(x)$ and $f_Y(y)$. If $g(x) = C\, f_X(x)$, then necessarily $h(y) = C^{-1}\, f_Y(y)$.

Remarks: The characterizations of independence discussed in terms of joint and conditional densities above and the decomposition of the joint density into the product of functions of the different variables hold in the discrete case as well. Furthermore, everything extends in a natural way to a continuous random vector in d dimensions. A continuous random vector $(X_1, X_2, \ldots, X_d)$ is said to have a density $f(x_1, x_2, \ldots, x_d)$ if

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_d \in A_d) = \int_{A_1\times A_2\times \ldots A_d} f(x_1, x_2, \ldots, x_d)\, dx_1\, dx_2, \ldots, dx_d\,,$$

for any d real Borel sets $A_1, A_2, \ldots A_d$. Equivalently, the A_i's can be taken to be intervals. The distribution function can therefore be written as:

$$F_{X_1,\ldots,X_d}(x_1, x_2, \ldots, x_d) = \int_{(-\infty,x_1]\times\ldots\times(-\infty,x_d]} f(x_1, x_2, \ldots, x_d)dx_1, dx_2, \ldots dx_d\,.$$

The marginal density of $(X_1, X_2, \ldots, X_k)$ for $k < d$ is given by:

$$f_{X_1,\ldots,X_k}(x_1, x_2, \ldots, x_k) = \int_{\mathbb{R}^{d-k}} f(x_1, x_2, \ldots, x_d)\, dx_{k+1}dx_{k+2}\ldots, dx_d\,.$$

The conditional density of $(X_{k+1}, X_{k+2}, \ldots, X_d)$ given $(X_1, \ldots, X_k)$ is given by:

$$f_{X_{[k+1:d]}|x_{[1:k]}}(x_{k+1}, \ldots, x_d) = \frac{f(x_1, x_2, \ldots, x_d)}{f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)}\,.$$

Independence of $(X_1, X_2, \ldots, X_k)$ from $(X_{k+1}, X_{k+2}, \ldots, X_d)$ is equivalent to

$$f(x_1, x_2, \ldots, x_d) = f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) \cdot f_{X_{k+1},\ldots,X_d}(x_{k+1}, \ldots, x_d)\,,$$

which is equivalent to the conditional density of one of the components given the other equalling its marginal density, which in turn is equivalent to

$$f(x_1, x_2, \ldots, x_d) = g(x_1, \ldots, x_k) \cdot h(x_{k+1}, \ldots, x_d)\,,$$

for some non-negative functions g, h.

Mutual independence of $(X_1, \ldots, X_{i_1}), (X_{i_1+1}, \ldots, X_{i_2}), \ldots$ is equivalent to the joint density factoring as a product of the m marginal densities [or m different functions, where the first is a function of the first i_1 co-ordinates, the second a function of the next $i_2 - i_1$ co-ordinates and so] where m is the number of groups into which the d variables are split:

$$f(x_1, x_2, \ldots, x_d) = g_1(x_1, \ldots, x_{i_1})\, g_2(x_{i_1+1}, \ldots, x_{i_2})\, \ldots, g_m(x_{i_{m-1}+1}, \ldots, x_{i_m})\,,$$

with $i_m = d$.

Conditional Expectation: We start with a pair of random variables (X, Y) having a joint pdf $f(x, y)$. We assume at the outset the finiteness of all integrals (expectations, conditional or otherwise) involved, so as to avoid technical complications. Recall that the marginal densities of X and Y, say f_1 and f_2 respectively, are:

$$f_1(x) = \int f(x, y)\, dy \;\text{ and }\; f_2(y) = \int f(x, y)\, dx\,.$$

We define:

$$\xi(t) \equiv E(Y \mid X = t) := \int y\, \frac{f(t, y)}{f_1(t)}\, dy\,.$$

By $E(Y \mid X)$ we denote the random variable $\xi(X)$. Then:

$$E(\xi(X)) = \int \xi(x)\, f_1(x)\, dx = \int \left(\int y\, \frac{f(x, y)}{f_1(x)}\, dy \right) f_1(x)\, dx = \int \int y\, f(x, y)\, dxdy = E(Y)\,. \tag{4}$$

This establishes:

$$E(Y) = E\, E(Y \mid X)\,.$$

For any function of Y, say $g(Y)$, we define

$$\xi_g(t) \equiv E(g(Y)|X=t) := \int g(y)\,\frac{f(t,y)}{f_1(t)}\,dy\,,$$

and set $E(Y|X) = \xi_g(X)$. It is then not difficult to verify that:

$$E(g(Y)) = E(E(g(Y)|X))\,.$$

The above ideas extend easily to defining conditional expectations of more general random variables, not just functions of Y but functions of (X,Y). So, consider a random variable $h(X,Y)$. As above,

$$\xi_h(t) \equiv E[h(X,Y)|X=t] = E[h(t,Y)|X=t) = \int h(t,y)\,\frac{f(t,y)}{f_1(t)}\,dy\,,$$

and set $E(h(X,Y)|X) = \xi_h(X)$. Once again, verify along the lines of (4) that $E[E(h(X,Y)|X)] = E(h(X,Y)]$. The following important (and useful) result is a direct corollary.

$$E[w(X)h(X,Y)|X] = w(X)\,E(h(X,Y)|X)\,. \tag{5}$$

In other words, when we take conditional expectations of a random variable that involves factors depending only on X, they can be taken *out of the conditional expectation*, since they behave like constants under the conditioning.

We next define the notion of the conditional variance for a general random variable $h(X,Y)$. For a general random variable Z, recall that

$$\mathrm{Var}(Z) = E[Z-EZ]^2 \equiv EZ^2 - (EZ)^2\,,$$

where the second equality follows by expanding the square in the middle term and taking expectations term by term. This is something you should verify. This leads to a natural definition of $\mathrm{Var}(h(X,Y)|X)$ as:

$$\mathrm{Var}(h(X,Y)|X) := E(h^2(X,Y)|X) - [E(h(X,Y)|X)]^2\,.$$

Equivalently, we can also write

$$\mathrm{Var}(h(X,Y)|X) = E[(h(X,Y) - E[h(X,Y)|X])^2|X]\,.$$

These two definitions coincide, as we would expect to. For convenience, we revert back to the ξ_h notation: recall that $\xi_h(X)$ is just alternative notation for $E(h(X,Y)|X)$. So consider

$$\begin{aligned}
E[(h(X,Y)-\xi_h(X))^2|X] &= E[(h^2(X,Y) - 2h(X,Y)\xi_h(X) + \xi_h(X)^2)|X] \\
&= E(h^2(X,Y)|X) - 2E[(h(X,Y)\xi_h(X))|X] + E(\xi_h(X)^2|X) \\
&= E(h^2(X,Y)|X) - 2\xi_h(X)E(h(X,Y)|X) + \xi_h(X)^2 \\
&= E(h^2(X,Y)|X) - [E(h(X,Y)|X)]^2\,,
\end{aligned}$$

For a general random variable Z of the form $h(X,Y)$, we next establish the formula:

$$\mathrm{Var}(Z) = E(\mathrm{Var}(Z|X)) + \mathrm{Var}(E(Z|X))\,.$$

Now

$$\mathrm{Var}(Z|X) = E(Z^2|X) - (E(Z|X))^2\,,$$

so

$$E(Z^2|X) = \mathrm{Var}(Z|X) + (E(Z|X))^2\,.$$

Now,

$$\begin{aligned}
\mathrm{Var}(Z) &= E(Z^2) - (EZ)^2 \\
&= E\left[E(Z^2|X)\right] - (E\,\xi_h(X))^2 \\
&= E\left[\mathrm{Var}(Z|X) + \xi_h^2(X)\right] - (E\,\xi_h(X))^2 \\
&= E(\mathrm{Var}(Z|X)) + E(\xi_h^2(X)) - (E\,\xi_h(X))^2 \\
&= E(\mathrm{Var}(Z|X)) + \mathrm{Var}(\xi_h(X)) = E(\mathrm{Var}(Z|X)) + \mathrm{Var}(E(Z|X))\,.
\end{aligned}$$

These ideas of course generalize to more than two random variables.

Exercise: A program searches through a list of n symbols sequentially to determine whether an object is present, and as soon as it finds the object it stops. The object is present in the list with probability p in which case its position is uniformly distributed among the n locations. What is the expected number of items the program has to search to find the object?
Solution: Let B be the random variable that is 1 if the item is present and 0 otherwise. Let N be the number of items searched. Then,

$$E(N) = E\,E(N \mid B) = E\,(N \mid B=1)\,p + E\,(N \mid B=0)\,(1-p) = p\,(n+1)/2 + (1-p)\,n = n - pn/2 + p/2\,.$$

Exercise: A fair coin is tossed n times and the number of heads H is noted. Having observed H, the coin is flipped H more times and the number of heads in the second trial is noted. Call this number N. The cumulative number of heads is then $H+N$. Find the mean and variance of $N+H$.

Solution: Because this experiment is formulated as a two stage experiment, where the parameters of the second stage experiment are determined *after* observing the outcome of the first stage experiment, i.e. *conditional on the parameters of the first stage experiment*, the conditional expectation and variance calculations come in extremely handy, as we shall

see below. One can imagine trying to solve it by first writing down the joint distribution of (H, N) and the basing the calculations off this joint pmf but this, as you are welcome to (and perhaps this is even instructive to realize the hassles involved) check, is a beastly calculation. We start with the observation that $H \sim \text{Bin}(n, /12)$ and $N|H \sim \text{Bin}(H, 1/2)$. We'll solve this by conditioning on H. So:

$$E[H+N] = EH+EN = (n/2)+E[E(N|H)] = (n/2)+E[H/2] = n/2+(1/2)(n/2) = 3n/4\,.$$

For the variance calculation:

$$\text{Var}(N+H) = E\,\text{Var}[(N+H)|H] + \text{Var}[E[(N+H)|H]]\,.$$

Now,

$$\text{Var}[(N+H)|H] = \text{Var}[N|H] = H/4 \;\text{ and }\; E[(N+H)|H] = H + E(N|H) = 3H/2\,.$$

Therefore,

$$\text{Var}(N+H) = E[H/4] + \text{Var}[3H/2] = (1/8) + (9/4)\text{Var}(H) = 1/8 + 9/16 = 11/16\,.$$

Mixed Joint Distributions: Joint distributions of multiple random variables can be built up by *combining* pre-specified marginal and conditional distributions. This 'combination' approach leads easily to joint distributions where some of the random variables involved may be discrete and the others continuous. Natural scenarios where joint distributions arise via separate specifications of marginals and conditionals are multi-stage random experiments of the kind described above with the coin-tossing experiment done in two stages in the exercise above, and also in the Bayesian formulation in statistics which we will say something about.

Given random variables (X, Y), we have: $f(x, y) = f(x)f(y|x)$: while, thus far, we have taken both pieces to be either probability density functions or probability mass functions, mixtures are also possible. Suppose that X is discrete, assuming countably many values in the space $\mathcal{X} \subset \mathbb{R}$, and given $X = x$, Y is a continuous random variables having a proper density function depending on the value x. Thus $f(x)$ is a pmf and $f(y|x)$ is a pdf. We want to compute the marginal of Y (this will be in terms of a probability density function) and the conditional of X given $Y = y$ which will be a discrete pmf. If $f(x, y)$ were a joint pdf that could be integrated over the plane, we would integrate out over the variable x to find the marginal of Y. However $f(x, y) = f(x)f(y|x)$ cannot be *integrated* with respect to x but can be indeed summed over the values $x \in \mathcal{X}$, so:

$$f_Y(y) = \sum_{x \in \mathcal{X}} f(x, y)\,.$$

The marginal of X given $Y = y$ is a p.m.f given via the usual formula:

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

Example 1: Suppose $X \sim \text{Exp}(\lambda)$, and $Y|X = x \sim \text{Poisson}(x)$. To generate this mechanism, one can first produce a realization of an exponential λ random variable by using the inverse distribution function technique, starting with a standard uniform random variable. Having observed the value of X, say x, one then generates a Poisson random variable Y with mean parameter x. The joint distribution of (X, Y) captures the variability in this random pair, as the two stage experiment is replicated many many times. [We will discuss later, how to generate a Poisson random variable with a given mean parameter.] We begin by writing down the (generalized) joint density of (X, Y) by multiplying the marginal of X by the conditional of $Y|X$. Thus,

$$f_{X,Y}(x,y) = \lambda\, e^{-\lambda x} 1(x > 0)\, \frac{e^{-x} x^y}{y!} 1(y \in \mathbb{Z}^+) = \frac{\lambda e^{-x(\lambda+1)} x^y}{y!}\, 1\{x > 0, y \in \mathbb{Z}^+\}\,.$$

Let's find the conditional distribution of X given $Y = y$. This is the joint density of (X, Y) divided by the marginal p.m.f of Y which is obtained by integrating out the joint density with respect to x. Thus:

$$f_{X|y}(x) = \frac{\frac{\lambda e^{-x(\lambda+1)} x^y}{y!}\, 1\{x > 0, y \in \mathbb{Z}^+\}}{\int_{[0,\infty)} \frac{\lambda e^{-x(\lambda+1)} x^y}{y!}\, dx}\,.$$

We will compute the integral in the denominator shortly, but we note that in this case (as in many other cases, when the functional forms under consideration are sufficiently nice) we don't require to actually compute the denominator explicitly if we are only interested in the conditional. Canceling the common factors in the numerator and the denominator and noting that the integral in the denominator is of the form $\psi(y, \lambda)$ we get:

$$f_{X|y}(x) = \frac{1}{\psi(y,\lambda)} e^{-x(\lambda+1)} x^{y+1-1} 1(x > 0)\,.$$

We note that apart from the constant in the denominator (constant as a function of x, recall we are conditioning on $Y = y$), this is the kernel of a Gamma$(y + 1, \lambda + 1)$ density [I am using the 'other' parametrization, not the one in CB]. And therefore the conditional density *has to be* this density and the constant $\psi(y, \lambda)$ has to be the right constant so that the expression, when integrated out over x gives 1.

Next, we find the marginal p.m.f of Y. We have:

$$\int_{[0,\infty)} \frac{\lambda e^{-x(\lambda+1)} x^y}{y!}\, dx = \frac{\lambda}{y!} \frac{\Gamma(y+1)}{(\lambda+1)^{y+1}} = \frac{\lambda}{(\lambda+1)^{y+1}}\,.$$

Example 2: Consider a pair independent random variables (Θ, N) where $\Theta \sim U(0,1)$ and $N \sim$ Poisson(λ). The random variable $H|(N,\Theta) = (n,\theta) \sim Bin(n,\theta)$. In other words, having observed the outcomes (θ, n) you flip a coin with probability θ of turning up Heads n times (independently) in succession and record the total number of Heads H. Then,

$$f_{\theta,N}(\theta,n) = 1\{0<\theta<1\}\,\frac{e^{-\lambda}\lambda^n}{n!}1\{n\geq 0\}$$

and

$$f_{H|(N,\Theta)}(h|(n,\theta)) = \binom{n}{h}\theta^h(1-\theta)^{n-h}1\{h\leq n\}\,.$$

The joint density is:

$$f(\theta,n,h) = \frac{e^{-\lambda}\lambda^n}{n!}\binom{n}{h}\theta^h(1-\theta)^{n-h}1\{0<\theta<1\}\,1\{0\leq h\leq n\}\,.$$

Let's find the marginal density of H. This is given by:

$$\begin{aligned}
f_H(h) &= \int_{[0,1]}\left[\sum_{n=h}^{\infty}\binom{n}{h}\theta^h(1-\theta)^{n-h}\,\frac{e^{-\lambda}\lambda^n}{n!}\right]d\theta \\
&= \int_{[0,1]}\frac{e^{-\lambda}\,(\lambda\theta)^h}{h!}\left[\sum_{n=h}^{\infty}\frac{(\lambda(1-\theta))^{n-h}}{(n-h)!}\right]d\theta \\
&= \int_{[0,1]}\frac{e^{-\lambda}\,(\lambda\theta)^h}{h!}\left[\sum_{l=0}^{\infty}\frac{(\lambda(1-\theta))^{l}}{l!}\right]d\theta \\
&= \int_{[0,1]}\frac{e^{-\lambda}\,(\lambda\theta)^h}{h!}e^{\lambda(1-\theta)}d\theta \\
&= \frac{\lambda^h}{h!}\int_{[0,1]}e^{-\lambda\theta}\,\theta^h\,d\theta\,.
\end{aligned}$$

Exercise: Compute EH and Var(H).

7 Prediction, Linear Regression and Bivariate Normal Distribution

Conditional Expectation as the Best Predictor and the Best Linear Predictor: Given a pair (Y,X) of random variables, the goal is to predict (the response) Y using the explanatory variable (called covariate) X. The optimal predictor, i.e. the one that minimizes least squares error is seen to be given by $\phi_{opt}(X) = E(Y|X)$. This is seen by noting that:

$$E[(Y-\phi(X))^2] = E[((Y-E(Y|X))+(E(Y|X)-\phi(X)))^2] = E[(Y-E(Y|X))^2]+E[(E(Y|X)-\phi(X))^2]\,.$$

Defining the residual $\epsilon := Y - E(Y|X)$, we note that $E(\epsilon|X) = 0$. The formulation:

$$Y = \mu(X) + \epsilon$$

where $\mu(X) = E(Y|X)$ and $E(\epsilon|X) = 0$ is called a *general regression model* with $E(Y|X)$ called the regression function. In statistical problems, the actual distribution of (X, Y) is not known (otherwise there would not be any need to collect data in the first place), and hence so isn't $E(Y|X)$. The statistician's job is to fit a model to $\mu(X)$, which essentially amounts to specifying a class of tenable distributions for (X, Y) with constraints on $E(Y|X)$ and attempt to find the best fitting distribution within the class.

One approach is to postulate that the conditional mean of Y given X is given by some nice smooth function of X depending on a few parameters. The simplest such incarnation is the *linear model*:

$$E(Y|X)_{postulated} = \alpha + \beta X\,,$$

where α, β are constants (parameters).

There are two possible cases. First, the model is **well-specified**, i.e. $E(Y|X)$ *indeed is linear*. Then

$$Y = \alpha_0 + \beta_0 X + \epsilon\,,$$

for some fixed constants (α_0, β_0) and $E(\epsilon|X) = 0$. The other case is when the model is **mis-specified**, in which case $E(\epsilon|X) \neq 0$.

We will talk about the well-specified case later, but for now, let's consider the general scenario where the model is mis-specified, and study how well the *postulated linear model* serves as an approximation. As a concrete example, suppose $Y = 1 + X + X^2 + \epsilon$ where $X \sim N(0, 1)$ and $\epsilon \sim N(0, 1/2)$ independent of X. Then $E(Y|X) = 1 + X + X^2$ which is *non-linear* in X. Nevertheless, we are still interested in prediciting Y by a linear function $\alpha + \beta X$. In what follows $\mu_Y, \mu(X)$ denote the means of Y and X respectively, and $\sigma_{XY}, \sigma_X^2, \sigma_Y^2$ and ρ_{XY} denote the covariance of X, Y, variance of X, variance of Y and correlation between X and Y respectively.

Best Linear Predictor (BLP): We want to find the best linear approximation to Y. To this end, we minimize $E[(Y - \alpha - \beta X)^2]$ over (α, β). Setting

$$\nabla_\alpha\, E[(Y - \alpha - \beta X)^2] = 0 \;\text{ and }\; \nabla_\beta\, E[(Y - \alpha - \beta X)^2] = 0\,,$$

gives

$$E(Y - \alpha - \beta X) = 0 \;\text{ and }\; E[X(Y - \alpha - \beta X)] = 0\,.$$

The first equation gives $\mu_Y = \alpha + \beta\mu_X$ and $E(XY) = \alpha\mu_X + \beta E(X^2)$. From the first equation $\alpha = \mu_Y - \beta\mu_X$, and plugging this into the second, $\beta = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$. Hence $BLP = \alpha_0 + \beta_0 X$, where

$$\alpha_0 = \mu_Y - \beta_0\mu_X \quad\text{and}\quad \beta_0 = \sigma_{XY}/\mathrm{Var}(X) = \rho_{XY}\,(\sigma_Y/\sigma_X)\,.$$

Now, note that we are in the mis-specified setting, i.e. $\mu(X)$ is *not linear* in X. Therefore $\mu(X) \neq \alpha_0 + \beta_0 X$. If we consider *residual from the linear model*:

$$\tilde{\epsilon} := Y - \alpha_0 - \beta_0 X$$

then this is *not a proper residual* in the sense that $E(\tilde{\epsilon}|X) \neq 0$! This is easy to see:

$$E(\tilde{\epsilon}|X) = E[(\mu(X) + \epsilon - \alpha_0 - \beta_0 X)|X] = \mu(X) - \alpha_0 - \beta_0 X \neq 0\,.$$

However:

$$E(\tilde{\epsilon}) = \mu_Y - (\alpha_0 + \beta_0 \mu_X) = \mu_Y - (\mu_Y - \beta_0 \mu_X + \beta_0 \mu_X) = 0\,.$$

Further:

$$\mathrm{Cov}(X, \tilde{\epsilon}) = 0$$

as can be checked by a direct calculation (please do)!

It is interesting to calculate the total error incurred by the BLP: in other words, even if the BLP is wrong, it may still be a useful working approximation provided the total error is small. We have:

$$\begin{aligned} E[(Y - \alpha_0 - \beta_0 X)^2] &= E[((Y - \mu_Y) - \beta_0 (X - \mu_X))^2] \\ &= E[(Y - \mu_Y)^2] - 2\beta_0 \,\mathrm{Cov}(X, Y) + \beta_0^2 \mathrm{Var}(X) \\ &= \sigma_Y^2 - 2\rho_{XY} \frac{\sigma_Y}{\sigma_X} \rho_{XY} \sigma_Y \sigma_X + \rho_{XY}^2 \frac{\sigma_Y^2}{\sigma_X^2} \sigma_X^2 \\ &= \sigma_Y^2 (1 - \rho_{XY}^2)\,. \end{aligned}$$

Thus, the quality of the linear predictor depends on ρ_{XY} with the prediction approaching perfection as the correlation goes to 1 or -1. When the correlation is 1 or -1, Y is a *linear function* of X. Thus, the correlation is seen to be a direct measure of the linear association between X and Y. Also, note that when the correlation is close to 0, the slope of the best predicting line is close to 0, showing that a linear function of X does a poor job in explaining Y, and that Y is essentially equally well predicted in the absence of X by the constant μ_Y, its mean!

Note that in the *well-specified* case when $\mu(X)$ is indeed a linear function of X, it is given precisely by the best linear predictor. We next look at the scenario when the residual in a linear regression model is postulated to be independent of X: the classical signal plus noise model, and what transpires when the covariate and the residual have normal distributions. As we see, this leads to a two-dimensional distribution called the Bivariate Normal Distribution.

Exercise: Suppose $Y = \beta X + \epsilon$, where $E(X) = 0$ and $E(\epsilon) = 0$. Then $\text{Cov}(X, \epsilon) = 0$ if and only if $\beta = \beta_0$ where $\beta_0 X$, and $\beta_0 = \sigma_{XY}/\sigma_X^2$ is the slope of the BLP among all straight lines passing through the origin.

Linear Regression and the Bivariate Normal Distribution: Consider the classical linear regression model $Y = \alpha_0 + \beta_0 X + \varepsilon$ where ε is independent of X, $\varepsilon \sim N(0, \sigma^2)$ and $X \sim N(\mu_x, \sigma_x^2)$. Note that $E(Y \mid X) = \alpha_0 + \beta_0 X$. In fact $Y \mid X \sim N(\alpha_0 + \beta_0 X, \sigma^2)$. We can write down the joint density of (X, ε) as follows:

$$f_{X,\varepsilon}(x, \epsilon) = \frac{1}{2\pi\sigma_x\sigma} \exp\left[-\frac{1}{2\sigma_x^2}(x - \mu_x)^2 - \frac{1}{2\sigma^2}\epsilon^2\right].$$

Our goal is to write down the joint density of (X, Y) with the above as a starting point, which we accomplish using the change of variable theorem. The map $(X, \varepsilon) \mapsto (X, Y)$ is an invertible linear transformation with $\varepsilon = Y - \alpha_0 - \beta_0 X$. The Jacobian of the transformation that maps (X, Y) back into (X, ε) is:

$$J := \left|\det \frac{\partial(X, \epsilon)}{\partial(X, Y)}\right| = 1,$$

by an easy calculation. Hence, the joint density of (X, Y) can be written as:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma} \exp\left[-\frac{1}{2\sigma_x^2}(x - \mu_x)^2 - \frac{1}{2\sigma^2}(y - \alpha_0 - \beta_0 x)^2\right].$$

We will now re-write the above density in terms of the moment parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})$ where μ_y is the mean of Y, σ_y^2 is the variance of Y and σ_{xy}, the covariance between X and Y.

Note that $\sigma_{xy} := \text{Cov}(X, Y) = \beta_0\, \sigma_x^2$, so that

$$\beta_0 = \frac{\sigma_{xy}}{\sigma_x^2} = \rho\, \frac{\sigma_y}{\sigma_x},$$

where ρ is the correlation between X and Y. Also, $\mu_y = \alpha_0 + \beta_0\, \mu_x$, so $\alpha_0 = \mu_y - \beta_0\, \mu_x$. Using these relations, $f(x, y)$, after some algebra, can be easily re-written as:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma} \exp\left[-\frac{1}{2\sigma_x^2}(x - \mu_x)^2 - \frac{1}{2\sigma^2}\left[y - \left(\mu_y + \rho\, \frac{\sigma_y}{\sigma_x}(x - \mu_x)\right)\right]^2\right].$$

Next,

$$\sigma_y^2 = E[\text{var}(Y|X)] + \text{var}(E[Y|X]) = \sigma^2 + \beta_0^2\, \sigma_x^2,$$

so $\sigma^2 = \sigma_y^2\,[1 - \beta_0^2(\sigma_x^2/\sigma_y^2)] = \sigma_y^2\,(1-\rho^2)$. Using this:

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\,\exp\left[-\frac{1}{2\sigma_x^2}(x-\mu_x)^2 - \frac{1}{2\sigma_y^2(1-\rho^2)}\,\left[y-\left(\mu_y+\rho\,\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)\right]^2\right], \\ &= \frac{1}{\sqrt{2\pi}\,\sigma_x}\,\exp\left[-\frac{1}{2\sigma_x^2}(x-\mu_x)^2\right] \\ &\qquad \times \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2\sigma_y^2(1-\rho^2)}\,\left[y-\left(\mu_y+\rho\,\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)\right]^2\right]. \quad (6) \end{aligned}$$

Now, consider the expression that appears with the exponent of the above density:

$$-\frac{1}{2\sigma_x^2}(x-\mu_x)^2 - \frac{1}{2\sigma_y^2(1-\rho^2)}\,\left[y-\left(\mu_y+\rho\,\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)\right]^2.$$

This can be manipulated as:

$$-\frac{1}{2(1-\rho^2)}\,\left[(1-\rho^2)\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \frac{1}{\sigma_y^2}\,\left[(y-\mu_y)-\rho\,\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)\right]^2,$$

and on expanding the square, yields,

$$-\frac{1}{2(1-\rho^2)}\,\left[(1-\rho^2)\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\,\rho\,\left(\frac{x-\mu_x}{\sigma_x}\right)\,\left(\frac{y-\mu_y}{\sigma_y}\right) + \rho^2\,\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right].$$

This puts the joint density into a standardized form:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\,\exp\left[-\frac{1}{2(1-\rho^2)}\,\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\,\rho\,\left(\frac{x-\mu_x}{\sigma_x}\right)\,\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right]. \quad (7)$$

Definition: A bivariate random vector (X,Y) is said to follow the BVN$(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})$ distribution (where $\sigma_{xy}^2 < \sigma_x^2\,\sigma_y^2$) if it has a joint density given by:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\,\exp\left[-\frac{1}{2(1-\rho^2)}\,\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\,\rho\,\left(\frac{x-\mu_x}{\sigma_x}\right)\,\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right],$$

where $\rho = \sigma_{xy}/(\sigma_x\sigma_y)$.
The derivation above then immediately gives the moment interpretations of the five parameters and shows that the conditional distribution of Y given X follows a normal-distribution and that $Y = \alpha_0 + \beta_0 X + \epsilon$ for a pair of independent normals (X, ϵ) and

α_0, β_0 have the forms given above in terms of the parameters of the BVN distribution. Equivalently, the BVN distribution can also be written in terms of the correlation parameter as BVN$(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. We will switch between these parametrizations without warning.

Consider the representation in 6: here we are representing the joint as a product of the marginal of X times the conditional of Y given X: $f(x,y) = f_X(x)\, f_{\cdot|x}(y)$. Indeed, from the linear model representation that we started out with at the beginning, this is exactly what we should get: $f_X(x)$ is the density of $N(\mu_x, \sigma_x^2)$ and $f_{\cdot|x}(y)$ is the $N(\alpha_0 + \beta_0 x, \sigma^2)$ density.

But we can also represent the joint density as the marginal of Y times the conditional of X given Y. To do this, it is not necessary to do a new round of calculations. If we reverse the steps leading from (6) to (7), but now swapping the roles of x and y, it is not difficult to see that:

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sigma_y^2}(y-\mu_y)^2 - \frac{1}{2\sigma_x^2(1-\rho^2)}\left[x - \left(\mu_x + \rho\,\frac{\sigma_x}{\sigma_y}(y-\mu_y)\right)\right]^2\right], \\ &= \frac{1}{\sqrt{2\pi}\,\sigma_y} \exp\left[-\frac{1}{2\sigma_y^2}(y-\mu_y)^2\right] \\ &\qquad \times \frac{1}{\sqrt{2\pi}\sigma_x\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sigma_x^2(1-\rho^2)}\left[x - \left(\mu_x + \rho\,\frac{\sigma_x}{\sigma_y}(y-\mu_y)\right)\right]^2\right]. \quad (8) \end{aligned}$$

The second quantity in this product representation is the $N(\mu_x + \rho\,\frac{\sigma_x}{\sigma_y}(y-\mu_y), \sigma_x^2(1-\rho^2))$ density at the point x and the first quantity is the $N(\mu_y, \sigma_y^2)$ density at the point y: hence $Y \sim N(\mu_y, \sigma_y^2)$ and $X|Y=y$ follows $N(\tilde{\alpha}_0 + \tilde{\beta}_0 y, \tilde{\sigma}^2)$ where:

$$\tilde{\alpha}_0 = \mu_x - \tilde{\beta}_0\,\mu_y, \quad \tilde{\beta}_0 = \rho\,(\sigma_x/\sigma_y), \quad \tilde{\sigma}^2 = \sigma_x^2\,(1-\rho^2)\,.$$

Now, write:

$$X = \tilde{\alpha}_0 + \tilde{\beta}_0 Y + \tilde{\varepsilon}\,.$$

By using the change of variable theorem in 2 dimensions, where we start from variables (X,Y) and go to the variables $(\tilde{\varepsilon}, Y)$ where $\tilde{\varepsilon} := X - \tilde{\alpha}_0 - \tilde{\beta}_0 Y$, *show* that $\tilde{\varepsilon}$ and Y are *independent* and that $\tilde{\varepsilon} \sim N(0, \tilde{\sigma}^2)$.
We conclude that the linear model representation also holds *the other way*: in other words, when Y is expressed as a linear model in (the normal predictor variable) X with independent normal errors, X admits a similar representation in terms of Y. Indeed, a bi-directional representation of this sort *implies* that the joint distribution of (X,Y) has to be bivariate normal but this is difficult to prove.

Now:

$$Y \mid X = x \sim N(\mu_y + \rho\,\sigma_y\,(x - \mu_x)/\sigma_x, \sigma_y^2\,(1 - \rho^2))\,.$$

Similarly

$$X \mid Y = y \sim N(\mu_x + \rho\,\sigma_x\,(y - \mu_y)/\sigma_y, \sigma_x^2\,(1 - \rho^2))\,.$$

Note that the conditional variance of Y given $X = x$ is smaller than its unconditional variance: knowledge of X provides information about Y leading to less uncertainty in its distribution. The degree of reduction is captured by the correlation coefficient ρ. When $\rho = 0$, X and Y are independent, and the unconditional distributions coincide with the conditional distributions.

The above formulae are connected to the term 'regression' that is used abundantly in statistics. Statistical regression is concerned with understanding the effect of so-called 'predictor variables' X on a response variable Y. For example, X can be height and Y can be weight. The goal is to make good predictions of the value of Y given the value of X using some (optimal) function of X. The classical scenario from which the term 'regression' emanates concerns the joint distribution of father's height (X) and son's height (Y) in a population. Assume that the joint distribution is BVN with $\mu_x = \mu_y = \mu$, $\sigma_x^2 = \sigma_y^2 = \sigma^2$ and correlation ρ, which may be assumed positive. The best predictor of son's height given father's height is

$$E(Y \mid X = x) = \mu + \rho\,(x - \mu)\,.$$

So, if we take a father whose height is Δ units above the average height μ, the predicted height for his son is $\rho\,\Delta$ units above the average: since $\rho < 1$, the son's predicted height is dragged down towards the mean in comparison to the father's height, i.e. the son's height *regresses* towards the mean.

7.1 Matrix representation of Bivariate Normal Distribution

The class of bivariate normal distributions can also be conveniently written using matrix representation. Define the 2×2 positive definite matrix Σ where $\Sigma_{11} = \sigma_x^2$, $\Sigma_{22} = \sigma_y^2$ and $\Sigma_{12} = \Sigma_{21} = \sigma_{xy}$, and note that $|\Sigma| = \sigma_x^2\,\sigma_y^2\,(1 - \rho^2) > 0$. Then, the $BVN(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ density can be written as:

$$f(x, y) = \frac{1}{2\,\pi\,|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu_x, y - \mu_y)\,\Sigma^{-1}\,(x - \mu_x, y - \mu_y)^T\right]\,.$$

This is not difficult to establish using:

$$\Sigma^{-1} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}^{-1} = \frac{1}{\sigma_x^2\,\sigma_y^2\,(1-\rho^2)} \begin{pmatrix} \sigma_y^2 & -\sigma_{xy} \\ -\sigma_{xy} & \sigma_x^2 \end{pmatrix}.$$

Now, recall the following facts from linear algebra: (a) Since Σ is a p.d. matrix, there exists a symmetric invertible matrix B such that $B^2 = \Sigma$, and, (b) The determinant of the product of square matrices is the product of their determinants. We call B the *symmetric square root* of Σ and we denote it by $\Sigma^{1/2}$. We denote its inverse by $\Sigma^{-1/2}$.
Define:

$$(U,V)^T = \Sigma^{-1/2}\,(X-\mu_x, Y-\mu_y)^T,$$

so that:

$$(X,Y)^T = (\mu_x,\mu_Y)^T + \Sigma^{1/2}\,(U,V)^T.$$

Now, we can use the Jacobian (change of variable) theorem in two dimensions to calculate $f_{U,V}(u,v)$, the density of (U,V). Note that:

$$\left|\frac{\partial(x,y)}{\partial(u,v)}\right| = |\Sigma^{1/2}| = |\Sigma|^{1/2}$$

using fact (b) above. Following the steps of the Jacobian theorem, we conclude that:

$$f_{U,V}(u,v) = \frac{1}{2\,\pi}\,\exp\{-(u^2+v^2)/2\},$$

showing that (U,V) are i.i.d. $N(0,1)$. Thus, any bivariate normal distribution can be generated via a linear (affine) transformation of i.i.d. normals.

Conversely, let (U,V) be i.i.d. $N(0,1)$ and let B be any 2×2 non-singular matrix and define

$$(\tilde{X},\tilde{Y})^T = \mu + B\,(U,V)^T, \tag{9}$$

where $\mu = (\mu_1,\mu_2)^T$, the μ_i's being real numbers. Define $\tilde{\Sigma} = BB^T$ and note that it is a p.d. matrix. A direct application of the change of variable theorem shows that (X,Y) has the following density:

$$\tilde{f}(x,y) = \frac{1}{2\,\pi\,|\tilde{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu_1,y-\mu_2)\,\tilde{\Sigma}^{-1}\,(x-\mu_1,y-\mu_2)^T\right].$$

One now readily concludes that $(\tilde{X},\tilde{Y}) \sim BVN(\mu_1,\mu_2,\tilde{\Sigma}_{11},\tilde{\Sigma}_{22},\tilde{\rho})$ where $\tilde{\rho}^2 = \tilde{\Sigma}_{11}\tilde{\Sigma}_{22}(1-\tilde{\Sigma}_{12}^2)$ is the squared correlation between X and Y.
It follows that every pair (μ,Σ) where $\mu\in\mathbb{R}^2$ and Σ is a 2×2 p.d. matrix uniquely

determines a BVN distribution and a random vector following BVN(μ, Σ) can be generated by the affine transformation (9) for any B satisfying $BB^T = \Sigma$.

Observations:

[1] The random vector $(W_1, W_2) := ((X - \mu_x)/\sigma_x, (Y - \mu_y)/\sigma_y)$ follows $BVN(0, 0, 1, 1, \rho)$.

[2] The random vector (U, V) above follows $BVN(0, 0, 1, 1, 0)$.

[3] Let (X, Y) follow $BVN(\mu, \Sigma)$ and define

$$(P, Q)^T := \eta + A\,(X, Y)^T$$

where $\eta \in \mathbb{R}^2$ and $A_{2\times 2}$ is non-singular. Then $(P, Q) \sim BVN(\eta + A\,\mu, A\Sigma AT)$. Verify this by using the change of variable theorem. Thus, affine transformations preserve bivariate normality.

We end with a useful fact.

Fact: Let $\underline{W} \equiv (W_1, W_2, \ldots, W_p)^T$ be a random vector. Let Σ_W be its dispersion matrix: i.e. Σ_W is the $p \times p$ matrix such that $\Sigma_W(i, j)$ is the covariance between W_i and W_j. Define $\underline{V} = A\,\underline{W} + \xi$ where A is a $p \times p$ matrix and ξ is a $p \times 1$ vector. Then,

$$\Sigma_V = A\,\Sigma_W\,A^T \quad \text{and} \quad EV = AEW + \xi\,.$$

Note that when we write that $(X, Y) \sim BVN(\mu, \Sigma)$, $\mu = E[(X, Y)^T]$ and Σ is the dispersion matrix of $(X, Y)^T$. The mean and dispersion matrices of (P, Q) can be easily calculated from those of (X, Y) by using the Fact above.