ST2131 (Probability)

Jia Cheng

January 2021

1 Definitions and Formula

Either symbol S, Ω denotes a sample space. Either $s \in S, \omega \in \Omega$ denotes an outcome.

Elementary event $\{s\}$ (singleton)

2 Observations

Ways of defining outcomes When defining outcomes regarding sequences or selections, it may be wise to specify 2 important properties.

- Distinguishable/Indistinguishable
- Ordered/unordered selection

The most **basic outcome** is usually an ordered selection of distinguishable elements. For e.g. when we want to select r balls from n red balls and m blue balls, the most basic way to define an outcome is to consider the n + m balls to be labelled, and that our order of selection matters.

Here, we define "**basic outcome**" as above.
Our discussion now assumes that each basic outcome has equal probability. In a finite sample space, the consequence of this is that $P(E) = \frac{|E|}{|S|}$.
However, sometimes we want to consider events of equal probabilities (or equal number of outcomes).
I like to think of these as equivalence classes with equal cardinalities.
It is then convenient to define these as outcomes instead.
For e.g. we define an outcome to be an unordered selection of r distinct elements, since each unordered selection can be considered as an equivalence class of $rPr = r!$ ordered selections of r distinct elements.

From now on, I will use the notion of an "equivalence class" to refer to an event or a collection of basic outcomes which may then be used to define new outcomes.

Ensuring outcomes are of equal probability The above process may lead to mistakes if the equivalence classes are not actually of equal size.

1. Suppose we want to distribute 20 (all distinct) items between 2 people, such that each person gets 10 items. 10 items are of type A, 10 items are of type B.
   Here, there are $\binom{20}{10,10}$ basic outcomes. Suppose I now define the events $E_{(a_1,a_2,\dots,a_{10})}$ as follows, $E_{(a_1,a_2,\dots,a_{10})}$ is the event where person 1 gets i-th item of type A if $a_i = 1$, person 2 gets i-th item of type A if $a_i = 0$.
   We can easily check that $|E_{(1,1,\dots,1)}|$ is much smaller than $|E_{(1,1,1,1,1,0,0,0,0,0)}|$. Hence, if we define such events as outcomes, then the resulting outcomes would not be of equal probability.

This is especially relevant when it comes to conditional probability, as we will discuss below.

3 Conditional Probability

Consider example 2e in the textbook. Suppose that an urn contains 8 red balls and 4 white balls. We draw 2 balls from the urn without replacement.
(a) If we assume that at each draw, each ball in the urn is equally likely to be chosen, what is the probability that both balls drawn are red?

The traditional way to do this would to define outcomes as the unordered selection of 2 balls. (this is valid since each unordered selection corresponds to 2! ordered selections.) Hence, $P(R_1R_2) = \frac{|R_1R_2|}{|S|} = \frac{\binom{8}{2}}{\binom{12}{2}} = \frac{14}{33}$

Another way to do this is via conditional probability. We claim that $P(R_2|R_1) = \frac{7}{11}$ and hence $P(R_1R_2) = P(R_1)P(R_2|R_1) = \frac{8}{12}\frac{7}{11} = \frac{14}{33}$.

But why this claim that $P(R_2|R_1) = \frac{7}{11}$ justified? Remember that conditional probability is merely a definition, that $P(A|B) = \frac{P(AB)}{P(B)}$, and definitions do not contribute to the underlying theory.

This is because we are implicitly defining new outcomes based on events. Precisely stated, we now relabel the remaining 11 balls from 1 to 11. Let E_i be the event where the first ball chosen is a red ball and the 2nd ball chosen is ball i. So how many basic outcomes lie in each event E_i? Since the first ball chosen could have been any red ball, $|E_i| = 8 \forall 1 \leq i \leq 11$. Since all the events are equally large, we can consider these as outcomes.
Out of these 11 outcomes, only 7 result in the 2nd ball chosen being red. Hence, $\frac{P(R_1R_2)}{P(R_1)} = \frac{|R_1R_2|}{|R_1|} = \frac{7}{11}$.
And by definition of conditional probability, $P(R_2|R_1) = \frac{7}{11}$.

In conclusion, when we make claims about conditional probability in such problems, we are actually implicitly redefining the underlying outcomes and then making use of the traditional method.

Taking advantage of symmetries Suppose event B is the disjoint union of sets, where $B = \cup_i B_i$ and that $P(B_i)$ is constant for all i. Suppose also that for event A, we have $P(A \cap B_i)$ constant for all i.

Then,

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

This can be the case when we see statements like these:

- An ordinary deck of 52 playing cards is randomly divided into 4 piles of 13 cardseach. Compute the probability that each pile has exactly 1 ace.(Example 2g of textbook)
  Let B be the event where the ace of spades and the ace of hearts are in different piles . Note that here, it is not specified exactly which pile the 2 aces are in. However, using the symmetrical nature of things, we can specify a special case $B_{1,2}$ and let ace of spades reside in pile 1, ace of hearts reside in pile 2.

Defining the sample space/outcomes This uses example 3o from chapter 3 as an example. A crime has been committed by a solitary individual, who left some DNA at the scene of the crime. Forensic scientists who studied the recovered DNA noted that only five strands could be identified and that each innocent person, independently, would have a probability of of having his or her DNA match on all five strands. The district attorney supposes that the perpetrator of the crime could be any of the 1 million residents of the town. Ten thousand of these residents have been released from prison within the past 10 years; consequently, a sample of their DNA is on file. Before any checking of the DNA file, the district attorney thinks that each of the 10,000 ex-criminals has probability of being guilty of the new crime, whereas each of the remaining 990,000 residents has probability where (That is, the district attorney supposes that each recently released convict is times as likely to be the crime's perpetrator as is each town member who is not a recently released convict.) When the DNA that is analyzed is compared against the database of the 10,000 ex-convicts, it turns out that A. J. Jones is the only one whose DNA matches the profile. Assuming that the district attorney's estimate of the relationship between and is accurate, what is the probability that A. J. is guilty?

Consider the 2 sample spaces that can be defined. S1: For an ex-con, the sample space has 2 outcomes guilty, not guilty. Based on the qn, $P(\{guilty\}) = \alpha$ and $P(\{notguilty\} = 1 - \alpha$.

S2: For the population of the whole town, we can label each of the one million residents from 1 to 1 million. The sample space then has 1 million outcomes $\{p1, p2, \ldots, p_{1000000}\}$, where outcome p_i indicates the i-th person is guilty.

Now for a qn: What is the probability that none of the ex-convicts are guilty? Is it $(1-\alpha)^{10000}$ or $1 - 10000\alpha$?
Ans: $1 - 10000\alpha$. Since the event E that the guilty is a member of the ex-convicts is the subset of S2 containing all the ex-convicts. Hence $P(E) = 10000\alpha$.
It would be wrong to multiply together $(1-\alpha)^{10000}$ since the event that person i is not guilty is not independent from the event that person j is not guilty.

3.1 Principle of Inclusion Exclusion

The upper and lower bounds of $P(\cup_{i=1}^n E_i)$ are an excellent exercise in manipulating summation indices.

Notation: Given 2 sets E, F, define $EF = E \cap F$.

$$P(\cup_{i=1}^n E_i) = \sum_{1 \le j \le n} (-1)^{j+1} \sum_{1 \le i_1 < \cdots < i_j \le n} P(E_{i_1} \ldots E_{i_j})$$

Claim: For odd k,

$$P(\cup_{i=1}^n E_i) \le \sum_{1 \le j \le k} (-1)^{j+1} \sum_{1 \le i_1 < \cdots < i_j \le n} P(E_{i_1} \ldots E_{i_j})$$

For even k,

$$P(\cup_{i=1}^n E_i) \ge \sum_{1 \le j \le k} (-1)^{j+1} \sum_{1 \le i_1 < \cdots < i_j \le n} P(E_{i_1} \ldots E_{i_j})$$

Here, we shall go from the odd k to even $k+1$. We first note that

$$P(\cup_{i=1}^n E_i) = \sum_{1 \le i \le n} P(E_i) - \sum_{1 \le i \le n} P(\cup_{1 \le j < i} E_j E_i)$$

Fix some $1 \leq i \leq n$.Applying the inequality for odd k, we have

$$\begin{aligned} P(\cup_{1 \leq j < i} E_j E_i) &\leq \sum_{1 \leq j \leq k} (-1)^{j+1} \sum_{1 \leq i_1 < \cdots < i_j \leq n} P(E_{i_1} E_i E_{i_2} E_i \ldots E_{i_j} E_i) \\ &= \sum_{1 \leq j \leq k} (-1)^{j+1} \sum_{1 \leq i_1 < \cdots < i_j \leq n} P(E_{i_1} E_{i_2} \ldots E_{i_j} E_i) \end{aligned}$$

Hence,

$$\begin{aligned} P(\cup_{i=1}^n E_i) &\geq \sum_{1 \leq i \leq n} P(E_i) - \sum_{1 \leq i \leq n} \sum_{1 \leq j \leq k} (-1)^{j+1} \sum_{1 \leq i_1 < \cdots < i_j \leq n} P(E_{i_1} E_{i_2} \ldots E_{i_j} E_i) \\ &= \sum_{1 \leq i \leq n} P(E_i) - \sum_{1 \leq j \leq k} (-1)^{j+1} \sum_{1 \leq i_1 < \cdots < i_j < i_{j+1} \leq n} P(E_{i_1} E_{i_2} \ldots E_{i_j} E_{i_{j+1}}) \\ &= \sum_{1 \leq j \leq k+1} (-1)^{j+1} \sum_{1 \leq i_1 < \cdots < i_{j+1} \leq n} P(E_{i_1} E_{i_2} \ldots E_{i_{j+1}}) \end{aligned}$$

Going from even k to odd $k+1$ is similar.

$P(\cdot|E)$ is a probability function Suppose $P : \mathcal{P}(S) \to [0,1]$ is a probability function on sample space S. Let $E \subseteq S$ such that $P(E) > 0$. Define function

$$\begin{aligned} Q : \mathcal{P}(S) &\to [0,1] \\ F &\mapsto P(F|E) \end{aligned}$$

Then, we can check that Q upholds all 3 probability axioms. Hence Q is a probability function on the **same** sample space as P.

Note that it is a misconception to say that Q is a probability function on E. When we say reduced sample space E, we really mean that all probabilities associated with any outcomes in $S - E$ have be set to 0 under Q.

4 Expectation

Proofs using Iverson's notation.

For random variable $X : S \to \mathbb{R}$, define its range, $range(X) := \{r \in \mathbb{R} : P(X = r) > 0\}$. For a discrete r.v., the range is at most countable.

(Discrete r.v.) Breaking down into elementary events Claim: $E[X] = \sum_{s \in S} X(s) P(\{s\})$. We denote $p(s) := P(\{s\})$.

$$
\begin{aligned}
E[X] &= \sum_{x \in range(X)} x \cdot P(X = x) \\
&= \sum_{x \in range(X)} x \cdot \sum_{s \in X^{-1}(x)} p(s) \\
&= \sum_{x \in range(X)} x \cdot \sum_{s \in S} [s \in X^{-1}(x)] p(s) \\
&= \sum_{s \in S} p(s) \cdot \sum_{x \in range(X)} x[s \in X^{-1}(x)] \\
&= \sum_{s \in S} p(s) \cdot \sum_{x \in range(X)} x[x = X(s)] \\
&= \sum_{s \in S} p(x) X(s) = \sum_{s \in S} X(s) p(s)
\end{aligned}
$$

Note that $[s \in X^{-1}(x)] = [X(s) = x] = [x = X(s)]$. While for a discrete r.v., $range(X)$ is at most countable and hence can be summed over, the sample space S may not be countable. *Hence, this part assumes S to be at most countable.*

The above formulation of expected values is useful in proving linearity of expectation (in the special case where sample space is at most countable).

Lemma For a discrete r.v. X whose range is a subset of the natural numbers, we have

$$E[X] = \sum_{i \geq 0} P(X > i)$$

Lemma For a continuous r.v. X whose range is a subset of the non-negative real numbers, we have

$$E[X] = \int_0^\infty P(X > x) dx$$

5 Discrete Distributions

5.1 Poisson distribution

Applications of Poisson r.v. The Poisson r.v. can be applied in the following situations.

- When an experiment consists of n independent trials, each trial i with probability of success p_i, the number of successes is approximately $X \sim Poi(\lambda = \sum_{1 \leq i \leq n} p_i)$
- In particular, when $\forall i, p_i = p$, we have a binomial distribution, with $X \sim Poi(\lambda = np)$. The approximation is more valid when n is large and p is small.
- When an experiment consists of trials that are weakly dependent. For e.g. let E_i, E_j be events representing the success of trial i, j respectively. Then we may have weak dependence if $P(E_i | E_j) \approx P(E_i)$ for all choices of i, j.

5.2 Geometric distribution

Infinite series We give an analytical justification for 2 derivations of expected values of a geometric r.v.

Let $X \sim Geom(p)$.

First derivation.

$$E[X] = \sum_{1\leq i} i \cdot (1-p)^{i-1}p = \sum_{1\leq i}(i-1)\cdot(1-p)^{i-1}p + \sum_{1\leq i}(1-p)^{i-1}p = (1-p)E[X]+1$$

The second equality is justified since we can use the root test to show that $\sum_{1\leq i}(i-1)\cdot(1-p)^{i-1}p$ converges. Since both $\sum_{1\leq i}(i-1)\cdot(1-p)^{i-1}p$ and $\sum_{1\leq i}(1-p)^{i-1}p$ converges, we have the second equality.

Second derivation.

$$E[X] = \sum_{1\leq i} i \cdot (1-p)^{i-1}p = \sum_{1\leq i}\sum_{1\leq j\leq i}(1-p)^{i-1}p = \sum_{1\leq j}\sum_{j\leq i}(1-p)^{i-1}p$$

The switching of summations is justified by the following lemma. For each $n, m \in \mathbb{Z}^+$, let real numbers $a_{n,m} \geq 0$. Then $\sum_{m\geq 1}\sum_{n\geq 1} a_{n,m}$ converges iff $\sum_{n\geq 1}\sum_{m\geq 1} a_{n,m}$ converges, and when this is so, they are equal.
Root test says that $\sum_{1\leq i} i\cdot(1-p)^{i-1}p$ converges, hence $\sum_{1\leq i}\sum_{1\leq j\leq i}(1-p)^{i-1}p$ converges. We can then apply the lemma and switch the order of summation.
Alternatively, we can first compute $\sum_{1\leq j}\sum_{j\leq i}(1-p)^{i-1}p$ and show that it converges, and then use the lemma to show that $\sum_{1\leq i}\sum_{1\leq j\leq i}(1-p)^{i-1}p = \sum_{1\leq j}\sum_{j\leq i}(1-p)^{i-1}p$ holds.

Similar arguments apply for the analytic derivation of $E[X^2]$.

5.3 Negative Binomial Distribution

Negative binomial is finite with probability 1 Let $X \sim NB(r,p)$, then $X = \sum_{1\leq i\leq r} Y_i$, where each $Y_i \sim Geom(p)$ is the r.v. on the number of trials needed to attain the ith success after the $i-1$-th success.

We use the "continuity" property of a probability function on a monotonic set sequence, i.e. if (E_i) is a monotonic sequence of events, where either $\forall i, E_i \subseteq E_{i+1}$ or $\forall i, E_i \supseteq E_{i+1}$, then $\lim_{n\to\infty} P(E_i) = P(\lim_{n\to\infty} E_i)$, where in the case of a monotonically increasing sequence, $\lim_{n\to\infty} E_i := \cup_{i\in\mathbb{N}} E_i$ and the case of a monotonically decreasing sequence, $\lim_{n\to\infty} E_i := \cap_{i\in\mathbb{N}} E_i$.
This property shows that $P(Y_i \in \mathbb{Z}^+) = 1$, by considering the monotonic sequence $E_i := \{Y_i \leq i\} = \{\omega \in S : Y_i(\omega) \leq i\}$.

Now we prove the claim: $P(X \in \mathbb{Z}^+) = 1$. Since $Y_1 \in \mathbb{Z}^+ \wedge Y_2 \in \mathbb{Z}^+ \wedge \cdots \wedge Y_r \in \mathbb{Z}^+ \implies X \in \mathbb{Z}^+$, we have $\{X \in \mathbb{Z}^+\} \supseteq \{Y_1 \in \mathbb{Z}^+ \wedge \cdots \wedge Y_r \in \mathbb{Z}^+\} = \cap_{1\leq i\leq r}\{Y_i \in \mathbb{Z}^+\}$. By axioms of probability, $1 \geq P(X \in \mathbb{Z}^+) \geq P(\cap_{1\leq i\leq r}\{Y_i \in \mathbb{Z}^+\}) = \prod_{1\leq i\leq r} P(Y_i \in \mathbb{Z}^+) = 1$, where the first equality follows by the independence of the events $\{Y_i \in \mathbb{Z}^+\}$. For example, $P(Y_2 \in \mathbb{Z}^+ | Y_1 \in \mathbb{Z}^+) = 1 = P(Y_2 \in \mathbb{Z}^+)$.

Hence $P(X \in \mathbb{Z}^+) = 1$.

5.4 Properties of c.d.f.

We claim that the cumulative distribution function of a discrete r.v., defined by $F = b \mapsto P(X \leq b)$ is right continuous but not left continuous in general.

Similar to elsewhere in this document, for a predicate Q, we let $\{Q(X)\} = \{s \in S : Q(X(s))\}$.

Let (b_n) be a monotonically decreasing sequence s.t. $\lim_{n\to\infty} b_n = b$. Then,

$$\begin{aligned}
\lim_{n\to\infty} F(b_n) &= \lim_{n\to\infty} P(X \le b_n) \\
&= P(\lim_{n\to\infty} \{X \le b_n\}) \\
&= P(\bigcap_{n\in\mathbb{N}} \{X \le b_n\}) \\
&= P(\{\wedge_{n\in\mathbb{N}} X \le b_n\}) \\
&= P(\{\wedge_{n\in\mathbb{N}} X \in (-\infty, b_n]\}) \\
&= P(\{X \in \bigcap_{n\in\mathbb{N}} (-\infty, b_n]\}) \\
&= P(\{X \in (-\infty, b]\}) \\
&= P(X \le b) = F(b)
\end{aligned}$$

It is interesting to see how the intersection operation turns into a logical AND operation upon moving into a set, then the logical operation again turns into an intersection operation.

We observe that for a monotonically increasing sequence (a_n) such that $\lim_{n\to\infty} a_n = b$, in the event that a is not in the range of the sequence (i.e. $\forall n, a_n < a$), then $\bigcup_{n\in\mathbb{N}}(-\infty, a_n] = (-\infty, b)$. Furthermore, if $P(X = b) > 0$, then we must have $\lim_{n\to\infty} F(a_n) = P(X < b) < P(X \le b) = F(b)$. Hence, in general, F is not left continuous.

Remark It is interesting to see that in our proof of continuity, we did not need to use any epsilon-delta analysis common in mathematical analysis. Why is that so? I am guessing that this is due to probability axiom 3, that for countably many disjoint events E_i, $P(\cup_{i\in\mathbb{N}} E_i) = \sum_{i\in\mathbb{N}} P(E_i)$. This axiom is used to prove the continuity of probability functions over monotonic set sequences. So, this axiom subsumes the usual analytical part.

Note that for continuous r.v., we always have $P(X \le b) = P(X < b)$.

Expectation

- A binomial r.v. $X \sim B(n, p)$ can be viewed as the sum of n independent Bernoulli r.v. $X_i, 1 \le i \le n$, where X_i is the indicator variable of the occurrence of the ith trial.
- A negative binomial r.v. $X \sim NB(r, p)$ can be viewed as the sum of r independent geometric r.v. $X_i, 1 \le i \le r$, where X_i is the number of trials needed since the $i-1$-th successful trial to achieve the i-th success.
- The coupon collection problem is similar to the negative binomial problem. If X is the number of coupons needed to collect at least one of each time, then we can split X into a sum of geometric r.v. X_i, where X_i is the number of coupons needed since the $i-1$-th new coupon to obtain the i-th new coupon. The difference is that for different i, the value of p changes.
- A hypergeometric r.v. $X \sim HG(N \text{ total}, m \text{ special}, n \text{ sample size})$ can be viewed as the sum of n indicator r.v. $X_i, 1 \le i \le n$, where X_i is the indicator variable of the "success" of the ith trial. While the trials are non-independent, linearity of expectation does not care about this, so the expected value of a HG r.v. is equal to the expected value of a binomial r.v. with $p = \frac{m}{N}$.

6 Continuous distributions

cdf and pdf

At the level of this textbook (A First Course in Probability, Sheldon Ross), it is assumed that for a

continuous r.v., a p.d.f. always exists such that the cdf can be evaluated by integrating using the pdf. It also seems that almost all instances of pdf encountered in the textbook are continuous, and by the fundamental theorem of calculus, the cdf is differentiable, with the derivative of the cdf = pdf.
However, see math.stackexchange, this is not always true. In general, cdf $b \mapsto F(b)$ is always well-defined as the probability function P on the set $(-\infty, b]$, and the pdf is defined by the derivative of the cdf. When the cdf does not behave "nicely" enough to be almost everywhere differentiable, then the pdf does not exist.
In particular, it is worth noting that the more general approach is to move from cdf to pdf, rather from pdf to cdf as in the textbook.

6.1 Normal distribution

Approximation to the binomial distribution (This is a special case of the central limit theorem.) Let $X \sim B(n,p)$, where $\mu = np, \sigma = \sqrt{np(1-p)}$. Then

$$\lim_{n \to \infty} P(a \leq \frac{X - \mu}{\sigma} \leq b) = \phi(b) - \phi(a)$$

where ϕ is the p.d.f. of the standard normal variable.

The approximation is good when $np(1-p) \geq 10$.

6.2 Exponential distribution

Let $X \sim Exp(\lambda)$.

Recurrence relation of moments

$$E[X^n] = \frac{n}{\lambda} E[X^{n-1}]$$

As mentioned below, the gamma distribution is a generalization of the exponential distribution. Using integration by parts, we can show that the nth moment of the gamma r.v. is $E[X^n] = \frac{\Gamma(\alpha+n)}{\lambda^n \Gamma(\alpha)} = \frac{(\alpha+n-1)^{\underline{n}}}{\lambda^n}$. Specialized to the exponential r.v., we get $\frac{(1+n-1)^{\underline{n}}}{\lambda^n} = \frac{n!}{\lambda^n}$

Relation between exponential and poisson r.v Consider an event whose rate of occurrence is given by λ, i.e. over a period of time t, the number of occurrences $N(t) \sim Poi(\lambda t)$.

The amount of time T till the first event has the following formula. For $t > 0$, $P(T > t) = P(N(t) = 0) = e^{-\lambda t}$.
Hence the cdf of T, $F_T(t) = P(T \leq t) = 1 - e^{-\lambda t}$. Differentiating gives $f_T(t) = \lambda e^{-\lambda t}$, which is indeed the exponential pdf.

The generalization to this is the Gamma r.v. Let T_n be the amount of time till n events occur. Note that $T = T_1$. Then $P(T_n \leq t) = P(N(t) \geq n) = \sum_{0 \leq i \leq n} P(N(t) = i)$ and we can use this to derive the c.d.f. of T_n and then differentiate to get the p.d.f.
In particular, $Exp(\lambda) = Gamma(\alpha = 1, \lambda)$.

7 Joint distributions

Joint pdf Given random variables $X_1, X_2, \ldots, X_n$ and joint pdf f, and a "nice" set $C \subseteq \mathbb{R}^n$, $P((X_1, \ldots, X_n) \in C) = \int_C f(x_1, \ldots, x_n) d(x_1, \ldots, x_n)$.

In particular, if Q is a predicate involving $X_1, \ldots, X_n$, then $P(\{s \in S : Q(X_1(s), \ldots, X_n(s))\}) = P(Q(X_1, \ldots, X_n)) = P((X_1, \ldots, X_n) \in C)$ where $C = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : Q(x_1, \ldots, x_n)\}$. Hence,

$$P(Q(X_1, \ldots, X_n)) = \int_C f = \int_{\mathbb{R}^n} [Q]f$$

We note that Q is the characteristic function of region C.

Characterization of independent random variables Let X, Y be 2 random variables. Then the following are equivalent. And when this is the case, the random variables are said to be independent.

- $\forall A, B \subseteq \mathbb{R}, P(X \in A \wedge Y \in B) = P(X \in A)P(Y \in B)$
- $\forall a, b \in \mathbb{R}, F_{X,Y}(a, b) = F_X(a)F_Y(b)$
- If X, Y are both discrete, $\forall a, b \in \mathbb{R}, p_{X,Y}(a, b) = p_X(a)p_Y(b)$
- If X, Y are both continuous, $\forall a, b \in \mathbb{R}, f_{X,Y}(a, b) = f_X(a)f_Y(b)$
- If X, Y are both continuous, $\exists$ functions h, g, $\forall a, b \in \mathbb{R}, f_{X,Y}(a, b) = h(a)g(b)$, i.e. the joint density function $f_{X,Y}$ is separable.

Here, F is the cumulative distribution function, p is the probability mass function, f is the probability density function.

Sum of independent random variables

- $X \sim Gamma(\alpha_1, \lambda), Y \sim Gamma(\alpha_2, \lambda) \implies X + Y \sim Gamma(\alpha_1 + \alpha_2, \lambda)$
- $X \sim Normal(\mu_1, \sigma_1^2), Y \sim Normal(\mu_2, \sigma_2^2) \implies X + Y \sim Normal(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- $X \sim Poi(\lambda_1), Y \sim Poi(\lambda_2) \implies X + Y \sim Poi(\lambda_1 + \lambda_2)$
- $X \sim B(n_1, p), Y \sim B(n_2, p) \implies X + Y \sim B(n_1 + n_2, p)$

Conditional r.v. Let X, Y be random variables. For some fixed $y \in \mathbb{R}$, consider the function $Q = \mathcal{P}(S) \to [0, 1], E \mapsto P(E|Y = y)$, i.e. Q is the probability function that gives the conditional probability of any event $E \subseteq S$ given $Y = y$. Note that Q fulfils all 3 probability axioms and hence is a probability function.

Then the pmf $x \mapsto Q(X = x) = P(X = x|Y = y) = \frac{P(X=x \wedge Y=y)}{P(Y=y)}$ maps each x to the probability of the conditionally distributed $X|Y = y$. Note that $X|Y = y$ is also a random variable using Q as a probability function.

- On all outcomes $s \in S$, it is clear that $(X|Y = y)(s) = X(s)$, i.e. $X|Y = y$ and X agree on all outcomes in the sample space.
- However, what is different is in the probability. $Q(X = x) = P(X = x|Y = y)$ can be different from $P(X = x)$, since the probability associated with the event $X = x|Y = y$ only considers the reduced sample space of $\{s \in S : Y(s) = y\}$.
- For example, for $s \in S$ where $Y(s) \neq y$, suppose $X(s) = x_0$. Then s will contribute a non-zero probability to the set $\{X = x_0\}$ if $P(\{s\}) > 0$. However, s will not contribute any probability to the set $\{X = x_0|Y = y\}$ (even though $(X|Y = y)(s) = x_0$) since $Q(\{s\}) = 0$. Therefore, viewing $X|Y = y$ as a random variable under probability function Q is consistent with probability axiom 3.

- Note that however, it is incorrect to view $X|Y = y$ as a random variable under probability function P as in the previous example $s \in \{s \in S : (X|Y = y)(s) = x_0\} = \{X = x_0|Y = y\}$ and assuming $P(\{s\}) > 0$, s contributes a positive probability, which shouldn't be the case.

Conditional mass and density functions The following are some formula derived or defined in the text. It is important to note that some of these are consequences of existing definitions and axioms, whereas others are truly definitions.

Discrete X, discrete Y

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

This is trivially a result from the definition of conditional probability, where $p_{X|Y}(x|y) := P(X = x|Y = y)$.

Continuous X, continuous Y

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

This is a definition. The reason is that $P(Y = y) = 0$ and it cannot be a quotient using the usual definition of conditional probability.

Motivation: By viewing $f_{X|Y}(x|y)dx = P(x < X < x{+}dx|y < Y < y{+}dy) = \frac{P(x<X<x+dx \wedge y<Y<y+dy)dxdy}{P(y<Y<y+dy)dy} = \frac{f(x,y)dxdy}{f_Y(y)dy}$ and cancelling dx on both sides. Obviously, this is not anywhere rigorous and is meant for motivation purposes only.

Continuous X, discrete N

$$f_{X|N}(x|n) = \frac{P(N = n|X = x)}{P(N = n)} f(x)$$

This is a definition.

8 Properties of expectation

This section introduces expectation of a multi-variable function of random variables. In the chapter on joint distributions, while cdf's and pdf's of a function on multiple r.v. have been studied (in particular sum $X + Y$), expected values were not discussed. Note that it is indeed possible to evaluate something like $E[g(X,Y)]$ using the pdf of $g(X,Y)$. (for e.g. in the previous chapter, it is derived that the pdf of $X + Y$ in the continuous case is $\int_{-\infty}^{a} f_X(a-y) f_Y(y) dy$)
However, analogous to the one variable case, we see that we can find $E[g(X,Y)]$ only using the joint pmf/pdf of X, Y.

Discrete case

$$E[g(X,Y)] = \sum_{x,y} g(x,y)p(x,y)$$

where x, y run over the ranges of X, Y respectively. The proof of this should be similar to the one variable case. For each element $z \in range(g)$, consider the preimage $g^{-1}(\{z\}) = \{(x,y) \in \mathbb{R}^2 :$

$g(x, y) = z\}$. We then partition $g^{-1}(\{z\})$ into distinct (x, y) tuples. The value of $g(X(s), Y(s))$ evaluated when $X(s) = x, Y(s) = y$ is then $z = g(x, y)$.

Linearity of expectation in the discrete case then follows from this formula. In the textbook, the earliest occurring proof of linearity of expectation made use of the summation $\sum_{s \in S}(X+Y)(s)P(\{s\})$, however, this requires S to be at most countable, which may not even be true in the discrete case. Hence using $E[g(X, Y)] = \sum_{x,y} g(x, y)p(x, y)$ where $g(X, Y) = X + Y$ is the more general way to prove LoE.

Continuous case

$$E[g(X, Y)] = \int_{\mathbb{R}^2} g(x, y)f(x, y)d(x, y)$$

With this, we can prove linearity of expectation for continuous r.v. as well. Then, we can evaluate $E[X + Y] = E[X] + E[Y]$ without having to evaluate f_{X+Y}.

Expectation of moments The text gives a useful formula for calculating moments of expectation of a certain type of r.v.

Given a number of events $A_i, 1 \leq i \leq n$, let X be the random variable that counts the number of events that occur. Then

$$E\left[\binom{X}{k}\right] = \sum_{1 \leq i_1 < \cdots < i_k \leq n} P(A_{i_1}A_{i_2} \ldots A_{i_k})$$

The proof is by indicator random variables.
The expectation of moments $E[X^k]$ can be evaluated recursively, since $\binom{X}{k} = \frac{X^k}{k!}$ can be expressed in terms of powers up to k.
The value of this method is that we do not even need to evaluate the pmf of X.

Correlation coefficient The correlation coefficient between 2 variables X, Y is defined as

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Covariance can be seen as a measure of how positively related or negatively related 2 variables are. Correlation coefficient can be seen as a "normalized" form of covariance, scaled to the interval $[-1, 1]$.

As described in the text, in a multinomial distribution, letting N_i be the number of trials that result in outcome i, then if $i \neq j$, $Cov(N_i, N_j) < 0$, which is intuitively true as we expect larger N_i to suggest smaller N_j and vice versa.

By considering $Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right)$ (resp. $Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)$), we can show that when $\rho(X, Y) = -1$ (resp. 1), then $Y = aX + b$ for some constants a, b.

The converse is also true, if $Y = aX + b$ where $a \neq 0$, then $\rho(X, Y) = \frac{a}{|a|} = sgn(a) = \pm 1$.

Given a finite population (or sample space) Ω enumerable by $\{1, \ldots, n\}$, suppose the values of X, Y applied to each outcome i is $x_i = X(i)$ and $y_i = Y(i)$, then we can evaluate the population correlation coefficient as follows. Assume that the probability of the outcomes are uniform. Then,

$$\rho(X, Y) = \frac{\sum_{1 \leq i \leq n} \frac{1}{n} \cdot (x_i - E[X])(y_i - E[Y])}{\sqrt{\sum_{1 \leq i \leq n} \frac{1}{n} \cdot (x_i - E[X])^2 \sum_{1 \leq i \leq n} \frac{1}{n} \cdot (y_i - E[Y])^2}} = \frac{\sum_{1 \leq i \leq n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{1 \leq i \leq n}(x_i - \overline{x})^2 \sum_{1 \leq i \leq n}(y_i - \overline{y})^2}}$$

where $\overline{x} = E[X], \overline{y} = E[Y]$. This is exactly the formula used to determine the correlation coefficient of a set of data points. Of course, note that a set of data points is usually a proper subset of the sample space.

8.1 Equivalence of distributions

First, we define a new notation. Given random variables X, Y, where X is associated with probability function P and Y is associated with probability function Q, $X \overset{d}{=} Y$ if for all sets $A \subseteq \mathbb{R}$, $P(X \in A) = Q(Y \in A)$.
Furthermore, when this is the case, $E[X] = E[Y]$.

See https://math.stackexchange.com/questions/1306650/notation-for-two-random-variables-with-the-same-
I am unsure if this definition is mathematically standard or even sound, however I am motivated to do this due to conditional expectation, which I will elaborate upon below.

Equality case Given a sample space S equipped with probability function P. Let X, Y be 2 random variables on S using probability function P.

Suppose that $\forall s \in S, X(s) = Y(s)$. Then we can write $X = Y$, i.e. X and Y are truly identical in the mapping sense. Furthermore, they use the same probability function P. This implies that $X \overset{d}{=} Y$.

Non-equality, but i.i.d This discussion uses an example. Given 2 independent random variables X, Y both normally distributed with parameters $(0, 1)$, we can say that $X \overset{d}{=} Y$.

Different associated probability functions This last case is the motivation for my definition. Given sample space S, let X, Y be random variables on S, both using probability function P. Now, consider the conditional random variable $X|Y = y$, which uses the probability function $Q = E \mapsto \frac{P(E \wedge Y = y)}{P(Y = y)}$.

Then for some function g, we say that $(X|Y = y) \overset{d}{=} g(X)$ if $Q(X = x) = P(X = x|Y = y) = P(g(X) = x)$.

To make this more concrete, we use an example from the textbook.
Q: A miner is trapped in a mine containing 3 doors. The first door leads to a tunnel that will take him to safety after 3 hours of travel. The second door leads to a tunnel that will return him to the mine after 5 hours of travel. The third door leads to a tunnel that will return him to the mine after 7 hours. If we assume that the miner is at all times equally likely to choose any one of the doors, what is the expected length of time until he reaches safety?

We let X be the amount of time the miner takes to reach safety. We let E_i be the event that the miner proceeds through door i and let the r.v. Y be s.t. $\{Y = i\} = E_i$.

Then conditioning on Y, we have $E[X] = E[E[X|Y]] = \sum_{1 \leq i \leq 3} E[X|E_i]P(E_i) = \frac{1}{3} \sum_{1 \leq i \leq 3} E[X|E_i]$.

Now, we wish to express $(X|Y = i) = (X|E_i)$ purely in terms of X.

Is it correct to write $(X|E_2) = X + 5$? Formally speaking, no, consider an encoding $s = (2, 3, 2, 1)$ of the outcome door 2, door 3, door 2, door 1 (i.e. the miner first goes through door 2, then door 3, then door 2, then door 1 when he finally escapes. Then $(X|E_2)(s) = 5 + 7 + 5 + 3$, whereas $(X + 5)(s) = 5 + 7 + 5 + 3 + 5$, strictly greater.
The correct (imo) way of writing is to say $(X|E_2) \overset{d}{=} X + 5$. Similarly, $(X|E_3) \overset{d}{=} X + 7$. Also, $(X|E_1) \overset{d}{=} 3$ (the constant function 3).

To justify $(X|E_2) \overset{d}{=} X{+}5$, we form a trivial bijection from the reduced sample space $S_R = \{(i_1, i_2, \dots, i_n) : n \in \mathbb{Z}_{\geq 2} \wedge i_1 = 2 \wedge i_n = 1\}$ with $S = \{(i_1, i_2, \dots, i_n) : n \in \mathbb{Z}^+ \wedge i_n = 1\}$. Note that we can ignore any outcome in $s \in S - S_R$, since $Q(\{s\}) = P(\{s\}|E_2) = 0$.

Because they are distributed identically, their expectations are the same. Hence, $E[X] = \frac{1}{3}(3 + E[X] + 5 + E[X] + 7)$.

8.2 More applications of conditional expectation

Computing conditional probabilities Let A be an event, $X = [A]$. Let Y be a r.v.

If Y is discrete, (this is nothing new)

$$P(A) = \sum_y P(A|Y = y)p_Y(y)$$

a simple consequence of the conditional probability.

If Y is continuous,

$$P(A) = E[X] = E[E[X|Y]] = \int_{\mathbb{R}} E[X|Y = y]f_Y(y)dy = \int_{\mathbb{R}} P(A|Y = y)f_Y(y)dy$$

(this is a new result.)

9 Limit Theorems

Since measure theory has not been taught yet, it suffices to have intuitive notions of measure theory like "almost everywhere". For now, we can simply understand an event or a set having "measure 0"/"probability 0" as being negligible in "size" with respect to the probability space.

Let E be an event. When we say $P(E) = 1$, all the outcomes $\omega \in E$ essentially "fill up" 99.9999...% of the sample space.

Let $X_i, i \in \mathbb{N}^+$ be identically distributed random variables.
Let $\mu = E[X_i], \sigma^2 = Var(X_i)$.
Denote Ω as the sample space. Denote

$$\overline{X_n} = \frac{X_1 + \cdots + X_n}{n}$$

for the following section.

9.1 Laws of large numbers

Weak Law For any fixed $\epsilon > 0$,

$$\lim_{n\to\infty} P(\{\omega \in \Omega : |\overline{X_n}(\omega) - \mu| \geq \epsilon\}) = 0$$

Strong Law

$$P(\{\omega \in \Omega : \lim_{n\to\infty} \overline{X_n}(\omega) = \mu\}) = 1$$

The strong law says that the random variable $\overline{X_n}$ converges almost surely to μ. This is the same as saying almost everywhere on Ω, the function $\overline{X_n}$ converges pointwise to the constant function μ.

Comparison between Weak and Strong Law The weak law is "weak" because it doesn't tell us anything about pointwise convergence of $\overline{X_n}$. For instance, if we fix any outcome $\omega \in \Omega$, the weak law does **not** tell us that $\lim_{n\to\infty} \overline{X_n}(\omega) = \mu$. In fact, even if for all $\omega \in \Omega$, $\lim_{n\to\infty} \overline{X_n}(\omega) \neq \mu$, this would still not contradict the weak law. All the weak law says is that, the proportion of Ω that lies ϵ close to μ tends to Ω if we let $n \to \infty$. (being a bit sloppy here, but you get the idea).
In contrast, the strong law says that if we fix any outcome $\omega \in \Omega$, then with probability 1, $\lim_{n\to\infty} \overline{X_n}(\omega) = \mu$. In other words, the set $\{\omega \in \Omega : \lim_{n\to\infty} \overline{X_n}(\omega) \neq \mu\}$ is essentially negligible with respect to the measure imposed on the probability space.

10 Some Problems

Maximums and minimums of dice throw Suppose we have N dice, for each $i \in \{1, \dots, N\}$, denote D_i as the outcome of the i-th dice roll. Note that $D_i \in \{1, \dots, 6\}$.

Define random variables $X = max\{D_i : 1 \leq i \leq N\}, Y = min\{D_i : 1 \leq i \leq N\}$.
Find $P(Y = j, X = k)$.

First we consider a simpler problem. What is $P(X = k)$?
$P(X \leq k) = (\frac{k}{6})^N$, and we can use this to find $P(X = k) = P(X \leq k) - P(X \leq k - 1)$

Similarly,

$$P(j \leq Y, X \leq k) = (\frac{k - j + 1}{6})^N$$

Then

$$P(j = Y, X = k) = \begin{cases} P(j \leq Y, X \leq k) - P(j + 1 \leq Y, X \leq k) \\ \quad - P(j \leq Y, X \leq k - 1) + P(j + 1 \leq Y, X \leq k - 1) & \text{if } k \geq j + 2 \\ P(j \leq Y, X \leq k) - P(j + 1 \leq Y, X \leq k) - P(Y = x = j) & \text{if } k = j + 1 \\ P(Y = X = j) & \text{if } k = j \end{cases}$$

Intersection of 2 intervals **Lemma** In general, given 2 intervals, $[a, b], [c, d]$, they intersect under the following 2 cases: $a \leq d, b \geq c$ and $c \leq b, d \geq b$.

Suppose we have 2 intervals I, J, with X the midpoint of I, Y the midpoint of J. Let $s, t \in \mathbb{R}^+$ such that $I = [X - s, X + s]$, $J = [Y - t, Y + t]$. Then

$$I \cap J \neq \emptyset \iff |X - Y| \leq s + t$$

The proof for this is as follows. We consider 2 cases. Suppose $X - s \leq Y + t$. Then by the above lemma we must have $X + s \geq Y - t$.
This implies $|X - Y| \leq s + t$.

Suppose $Y - t \leq X + s$. Then we must have $Y + t \geq X - s$.
This also implies $|X - Y| \leq s + t$.

Hence, in either case, we have $|X - Y| \leq s + t$, proving the result.

Separability of joint p.d.f. Let $f(x, y) = xy[0 < x < 1][0 < y < 1][0 < x + y < 1]$. Claim: $f(x, y)$ cannot be separated into a product $h(x)g(y)$.

We offer a more convincing proof of contradiction than what is shown in the text. Suppose not, such that there exists functions h, g such that $\forall x, y, f(x, y) = h(x)g(y)$.

- $h(0.1)g(0.8) = f(0.1, 0.8) = 0.1 \cdot 0.8 \neq 0$, hence $h(0.1), g(0.8) \neq 0$. In particular, $g(0.8) \neq 0$.
- $h(0.8)g(0.1) = f(0.8, 0.1) = 0.8 \cdot 0.1 \neq 0$, hence $h(0.8), g(0.1) \neq 0$. In particular, $h(0.8) \neq 0$.
- But $h(0.8)g(0.8) = f(0.8, 0.8) = 0$ since $[0 < 0.8 + 0.8 < 1] = 0$, which says that $h(0.8) = 0 \vee g(0.8) = 0$. This contradicts the above 2 statements.