

ST2131 (Probability)

Jia Cheng

January 2021

1 Definitions and Formula

2 Observations

2.1 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.

2.1.1 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.

2.2 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 be to define outcomes as the unordered selection of 2 balls. (this is valid since each unordered selection corresponds to 2! ordered selections.) Hence, $P(R_1 R_2) = \frac{|R_1 R_2|}{|S|} = \frac{\binom{8}{2}}{\binom{12}{2}} = \frac{14}{33}$

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

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, $\frac{P(R_1 R_2)}{P(R_1)} = \frac{|R_1 R_2|}{|R_1|} = \frac{7}{11}$. 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.

2.2.1 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 cards each. 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.

2.3 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 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 $\{p_1, p_2, \dots, p_{1000000}\}$, 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.

2.4 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 $EF = E \cap F$.

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

Claim: For odd k ,

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

For even k ,

$$P(\cup_{i=1}^n E_i) \geq \sum_{1 \leq j \leq k} (-1)^{j+1} \sum_{1 \leq i_1 < \dots < i_j \leq n} P(E_{i_1} \dots E_{i_j})$$

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

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

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

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

Hence,

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

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

3 Distributions

3.1 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 .

3.2 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 $E[X^2]$.

3.3 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 i th 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 $\lim_{n \rightarrow \infty} P(E_i) = P(\lim_{n \rightarrow \infty} E_i)$, where in the case of a monotonically increasing sequence, $\lim_{n \rightarrow \infty} E_i := \cup_{i \in \mathbb{N}} E_i$ and the case of a monotonically decreasing sequence, $\lim_{n \rightarrow \infty} E_i := \cap_{i \in \mathbb{N}} E_i$.

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 $Y_1 \in \mathbb{Z}^+ \wedge Y_2 \in \mathbb{Z}^+ \wedge \dots \wedge Y_r \in \mathbb{Z}^+ \implies X \in \mathbb{Z}^+$, we have $\{X \in \mathbb{Z}^+\} \supseteq \{Y_1 \in \mathbb{Z}^+ \wedge \dots \wedge Y_r \in \mathbb{Z}^+\} = \cap_{1 \leq i \leq r} \{Y_i \in \mathbb{Z}^+\}$. 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 $P(X \in \mathbb{Z}^+) = 1$.

4 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

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

for the following section.

4.1 Laws of large numbers

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

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

Strong Law

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

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 $\lim_{n \rightarrow \infty} \overline{X_n}(\omega) = \mu$. In fact, even if for all $\omega \in \Omega$, $\lim_{n \rightarrow \infty} \overline{X_n}(\omega) \neq \mu$, 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 $n \rightarrow \infty$. (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, $\lim_{n \rightarrow \infty} \overline{X_n}(\omega) = \mu$. In other words, the set $\{\omega \in \Omega : \lim_{n \rightarrow \infty} \overline{X_n}(\omega) \neq \mu\}$ is essentially negligible with respect to the measure imposed on the probability space.

5 Some Problems

5.1 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

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

5.2 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 $c \leq b, d \geq a$.

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.