Lecture Notes

This lecture is devoted to one rather general probability question. Let $A_1, A_2, \ldots, A_N$ be mutually independent events over the same sample space. Let the random variable T be the number of these events that occur. What is the probability that at least k events occur? That is, what is $\Pr(T \geq k)$?

This question comes up often. For example, suppose we want to know the probability that at least k heads come up in N tosses of a coin. Here A_i is the event that that the coin is heads on the i-th toss, T is the total number of heads, and $\Pr(T \geq k)$ is the probability that at least k heads come up.

As a second example, suppose that we want the probability of a student answering at least k questions correctly on an exam with N questions. In this case, A_i is the event that the student answers the i-th question correctly, T is the total number of questions answered correctly, and $\Pr(T \geq k)$ is the probability that the student answers at least k questions correctly.

There is an important difference between these two examples. The first example is a special case in that all events A_i have equal probability; that is, the coin is as likely to come up heads on one flip as on another. In particular, suppose that the coin comes up heads with probability p. Then all events A_i have probability p, and T has the now-familiar binomial distribution:

$$\Pr(T = k) = \binom{n}{k} p^k (1-p)^{n-k} = f_{n,p}(k)$$

We studied the binomial distribution extensively two weeks ago. Therefore, we already have an answer to the question posed above in the special case where all events A_i have equal probability.

In the second example, however, some exam questions might be more difficult than others. If question 1 is easier than question 2, then the probability of event A_1 is greater than the probability of event A_2. This lecture focuses on questions of this more general type in which the events A_i may have different probabilities.

1 The Probability of at Least One Event

Before computing the probability that at least k events occur, we will consider another special case, namely the probability that at least one event occurs. That is, we will determine $\Pr(T \geq 1)$.

1.1 Markov's Inequality

We can obtain a weak bound on the probability that at least one of the events $A_1, A_2, \ldots, A_N$ occurs using Markov's Theorem:

$$\Pr(T \geq 1) \leq \frac{\text{Ex}(T)}{1}$$

Markov's Theorem applies because the random variable T is non-negative.

All that remains is to compute $\text{Ex}(T)$, the expected number of events that occur. We can do this by introducing an indicator variable T_i for each event A_i; that is, $T_i = 1$ if event A_i occurs, and $T_i = 0$ if event A_i does not occur. Then T, the total number of events that occur, is the sum of these indicator variables:

$$T = T_1 + T_2 + \ldots + T_N$$

We can now compute the expected value of T as follows:

$$\begin{aligned} \text{Ex}(T) &= \text{Ex}(T_1 + T_2 + \ldots + T_N) \\ &= \text{Ex}(T_1) + \text{Ex}(T_2) + \ldots + \text{Ex}(T_N) \\ &= \Pr(A_1) + \Pr(A_2) + \ldots \Pr(A_N) \end{aligned}$$

In the first step, we take the expected value of both sides of the preceding equation. The second step uses linearity of expectation. In the last step, we use the fact that the expected value of an indicator variable is the probability that the indicator is 1.

Substituting the expression for $\text{Ex}(T)$ into the inequality from Markov's Theorem gives the following bound on the probability that some event occurs.

Fact 1 Let $A_1, A_2, \ldots, A_N$ be events over the same sample space, and let the random variable T be the number of these events that occur. Then

$$\Pr(T \geq 1) \leq \sum_{i=1}^{n} \Pr(A_i).$$

This is a formula well worth remembering. In words, the probability that at least one event occurs is at most the sum of the event probabilities.

This result is weak for events with sizable probabilities. For example, if we flip a fair coin three times, then we can conclude that the probability of flipping heads at least once is at most $\frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}$, which is not front-page news.

Nevertheless, Fact 1 is useful if the events A_i all have relatively low probability. A key advantage is that the events A_i need not be independent. (The only tools that we used in the derivation were

Markov's Theorem and linearity of expectation, and neither of these requires independence.) This makes the result easy to apply.

For example, suppose that we want to compute the probability of a meltdown at MIT's nuclear plant on Albany Street. First, we enumerate all events A_i that could lead to a meltdown, e.g. operator error, earthquake, terrorist attack, meteor strike, etc. Next, we somehow determine the probability of each type of catastrophe. Then Fact 1 says that the probability of a meltdown is at most the sum of these event probabilities. Thankfully, we can totally ignore tricky independence issues such as, what if the terrorist team is wiped out by a meteor strike?

1.2 An Alternate Proof

We can obtain Fact 1 in another way. Recall Boole's Inequality, which says that for any events $A_1, A_2, \ldots, A_N$, we have:

$$\Pr(A_1 \cup A_2 \cup \ldots \cup A_N) \quad \leq \quad \Pr(A_1) + \Pr(A_2) + \ldots + \Pr(A_N)$$

The left side of the inequality is the probability that at least one event A_i occurs, and the right side is the sum of the event probabilities. Therefore, Boole's Inequality states that the probability that at least event occurs is at most the sum of the event probabilities. This is precisely the conclusion reached above!

1.3 A Lower Bound Using Independence

The previous method gives an upper bound on $\Pr(T \geq 1)$. We can get a lower bound by computing the maximum of $\Pr(A_i)$ over all i.

Fact 2 Let $A_1, A_2, \ldots, A_N$ be events over the same sample space, and let the random variable T be the number of these events that occur. Then

$$\Pr(T \geq 1) \geq \max_{1 \leq i \leq n} \Pr(A_i).$$

The left side is $\Pr(A_1 \cup A_2 \cup \ldots \cup A_n)$. Since every event A_i is a subset of the union $A_1 \cup A_2 \cup \ldots \cup A_n$, we have $\Pr(A_1 \cup A_2 \cup \ldots \cup A_n) \geq \Pr(A_i)$ for all i. In particular, this holds for the i which maximizes $\Pr(A_i)$.

We can prove a stronger result if the events A_i are mutually independent. In this case, if we expect more than one event to occur, then almost certainly at least one event will occur.

Theorem 1.1 Let $A_1, A_2, \ldots A_N$ be independent events, and let T be the number of these events that occur. The probability that none of the events occur is at most $e^{-\text{Ex}(T)}$.

The proof of the theorem is surprisingly simple. The main trick is to use the fact that $1 - x \leq e^{-x}$ for all x. This inequality follows from the Taylor expansion $e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \ldots$.

Proof.

$$\begin{aligned}
\Pr(T = 0) &= \Pr(\bar{A}_1 \cap \bar{A}_2 \cap \ldots \cap \bar{A}_N) \\
&= \prod_{i=1}^{n} \Pr(\bar{A}_i) \\
&= \prod_{i=1}^{n} 1 - \Pr(A_i) \\
&\leq \prod_{i=1}^{n} e^{-\Pr(A_i)} \\
&= e^{-\sum_{i=1}^{n} \Pr(A_i)} \\
&= e^{-\mathrm{Ex}(T)}
\end{aligned}$$

In the first step, we reason that if zero of the events A_i occur, then all of the complementary events $\bar{A}_i$ must occur. Equality holds in the second step because the events are independent. The third step uses the fact that the probability of an event is one minus the probability of the complementary event. The fourth step relies on the inequality $1 - x \leq e^{-x}$ mentioned above. The remainder is simplification. ■

Interestingly, this theorem does not depend on N, the number of events. The theorem gives the same bound whether there are 100 events each with probability 0.1 or 1000 events each with probability 0.01. In both cases, the expected number of events is 10, and so the probability of no event occurring is at most e^{-10} or about 1 in 22,000. (Though the bounds given by the theorem are equal, the actual probabilities are somewhat different in these two cases.)

This theorem can be interpreted as a sort of "Murphy's Law": if we expect some things to go wrong, then something probably will. For example, suppose that we are building a microprocessor, and the fabrication process is such that each transistor is faulty mutually independently with a probability of one in a million. This sounds good. However, microprocessors now contain about ten million transistors, so the expected number of faulty transistors is 10 per chip. Since we expect some things to go wrong, something probably will. In fact, Theorem 1.1 implies that the probability of a defect-free a chip is less than 1 in 22,000!

2 A Mind-Reading Trick

Theorem 1.1 is the basis for a neat card trick.

For this trick, every card has a "value". The value of a number card less than 10 is just the number, and the value of a 10, face card, ace, or joker is 1. For example, 8♣ has value 8, Q♠ has value 1, and 10♢ has value 1.

A volunteer shuffles the deck and secretly chooses a number from 1 to 9. She deals this number of cards minus one from the top of the deck and looks at the new top card. For example, if she chooses the number 4, then she deals off 3 cards and looks at the 4th card.

She then takes the value of this top card, deals off that number of cards minus one, and again looks at the new top card. For example if she sees a $6\diamondsuit$, then she deals off 5 more cards and looks at the next one.

She repeats this procedure until she runs out of cards. "Her card" is the last card that she looked at to determine the number to deal off. For example, suppose that the deck is as follows:

$$Q\heartsuit\ 4\spadesuit\ \underline{2\spadesuit}\ J\clubsuit\ \underline{K\spadesuit}\ \underline{5\heartsuit}\ 3\heartsuit\ A\clubsuit\ 7\diamondsuit\ 7\heartsuit\ \underline{3\clubsuit}\ K\clubsuit$$

If the volunteer chooses 3 as her secret number, then she deals off $3-1=2$ cards and looks at the $2\spadesuit$. Since the value of this card is 2, she deals off $2-1=1$ more card and looks at the $K\spadesuit$. This card has value 1, so she deals off $1-1=0$ more cards and looks at the $5\heartsuit$. Since this card has value 5, she deals of $5-1=4$ more cards and looks at the $3\clubsuit$. This card has value 3, so she tries to deal off $3-1=2$ more cards. However, she then runs out of cards, so "her card" is $3\clubsuit$.

After the volunteer completes the procedure, the magician is able to name her card!

The Secret

The secret is that the magician performs the same procedure as the volunteer, starting with the number 1. When he guesses "her card", he is actually naming "his card"! Remarkably, the two are almost certain to be the same! The reason is that if the magician and volunteer ever land on the same card, then they will finish with the same card. In the example, the magician lands on the underlined sequence of cards:

$$\underline{Q\heartsuit}\ \underline{4\spadesuit}\ 2\spadesuit\ J\clubsuit\ K\spadesuit\ \underline{5\heartsuit}\ 3\heartsuit\ A\clubsuit\ 7\diamondsuit\ 7\heartsuit\ \underline{3\clubsuit}\ K\clubsuit$$

Even though the volunteer and the magician start with different numbers, they both coincidentally land on the $5\heartsuit$. After this point, they both continue in the same way and both end up with $3\clubsuit$.

All that remains is to prove that the magician and the volunteer are likely to land on the same card at some point. An exact analysis of this probability is difficult, but Theorem 1.1 gives some intuition.

Let A_i be the event that on his i-th jump the magician lands on a card also visited by the volunteer. If the volunteer's next card is 1 away, then the magician has a $\frac{1}{3}$ chance of landing on it (12 face cards, 3 10's, and 2 jokers gives $\frac{18}{54}=\frac{1}{3}$). If the volunteer's next card is more than 1 away, then the magician has a $\frac{2}{27}$ chance of landing on it (4 correct number cards gives $\frac{4}{54}=\frac{2}{27}$).

Since the magician takes 10-15 jumps through the deck, the sum of these probabilities is rather high; that is, $\text{Ex}(T)$ is around 3 or so. Consequently, Theorem 1.1 suggests that the magician is likely to eventually jump on a card visited by the volunteer. (As a ballpark estimate, the probability might be around $1 - e^{-3}$.) In this case, the magician and volunteer end up with the same final card and the trick works.

Of course, Theorem 1.1 does not really apply, because the events A_i are not at all independent. For example, event A_i implies event A_{i+1}; once the magician and volunteer match, they always agree. This means that our probability estimates are bogus. However, computer tests suggest that for a full deck with jokers, the trick works more than 90% of the time.

3 The Chernoff Bound

We now address the general question posed at the beginning of lecture. Given mutually independent events $A_1, A_2, \ldots, A_N$, what is the probability that at least k of the events occur? We will prove that the number of events that occur is almost always very close to the expectation. This result is called the Chernoff Bound.

For example, if we toss N coins, the expected number of heads is $\frac{N}{2}$ heads. The Chernoff Bound implies that for sufficiently large N, the number of heads is almost always very close to $\frac{N}{2}$.

A nice feature of the Chernoff Bound is that we do not event need to know the probability of each event A_i or even the number of events N; rather, we need only the expected number of events that occur and the fact that the events are mutually independent.

The theorem is stated below in terms of Bernoulli variables instead of events. However, we can regard T_i as an indicator for the event A_i.

Theorem 3.1 (Chernoff Bound) Let $T_1, T_2, \ldots, T_N$ be mutually independent Bernoulli variables, and let $T = T_1 + T_2 + \ldots + T_N$. Then for all $c \geq 1$, we have

$$\Pr(T \geq c\mathrm{Ex}(T)) \leq e^{-\alpha c \mathrm{Ex}(T)}$$

where $\alpha = \ln c + \frac{1}{c} - 1$.

The formula for α is a little awkward. The situation is simpler when $c = e = 2.718\ldots$, as in the following corollary.

Corollary 3.2 Let $T_1, T_2, \ldots, T_N$ be mutually independent Bernoulli variables, and let $T = T_1 + T_2 + \ldots + T_N$.

$$\Pr(T \geq e\mathrm{Ex}(T)) \leq e^{-\mathrm{Ex}(T)}$$

Proof. Setting $c = e$ in the Chernoff Bound gives:

$$\begin{aligned} \Pr(T \geq e\mathrm{Ex}(T)) &\leq e^{-(\ln e + \frac{1}{e} - 1)e\mathrm{Ex}(T)} \\ &= e^{-\mathrm{Ex}(T)} \end{aligned}$$

■

We will prove the Chernoff Bound shortly. First, let's see an example of how it is used.

3.1 Pick 4

There is a lottery game called Pick 4. In this game, each player picks 4 digits, defining a number in the range 0 to 9999. A winning number is drawn each week. The players who picked the winning number win some cash. A million people play the lottery, so the expected number of winners each week is $\frac{1}{10{,}000} \cdot 1,000,000 = 100$. However, on some lucky day thousands of people might all pick the winning number, costing the lottery operators loads of money. How likely is this?

Assume that all players pick numbers uniformly and mutually independently. Let T_i be an indicator variable for the event that the i-th player picks the winning number. Let $T = T_1 + T_2 + \ldots + T_N$. Then T is the total number of players that pick the winning number. As noted above, an average of 100 people win each week, so $\text{Ex}(T) = 100$. We can use Corollary 3.2 to bound the probability that number of winners is greater than 272 as follows:

$$\begin{aligned} \Pr(T \geq 272) &\leq \Pr(T \geq e\text{Ex}(T)) \\ &\leq e^{-Ex(T)} \\ &= e^{-100} \end{aligned}$$

The probability of 272 or more people winning is absurdly small! It appears that the lottery operators should not worry that ten thousand people will pick correctly one day!

But there is a catch. The assumption that people choose Pick 4 numbers uniformly and mutually independently is empirically false; people choose certain "favorite" numbers far more frequently than others.

Chernoff used this fact in devising a scheme to actually make money on the lottery. In this case, a fraction of all money taken in by the lottery was divided up equally among the winners. A bad strategy would be to pick a popular number. Then, even if you pick the winning number, you must share the cash with many other players. A better strategy is to pick a lot of unpopular numbers. You are just as likely to win with an unpopular number, but will not have to share with anyone. Chernoff found that peoples' picks were so highly correlated that he could actually turn a 7% profit by picking unpopular numbers!

3.2 Proof of the Chernoff Bound

The Chernoff Bound has the most difficult proof of the course. The intuition is in the first two steps; the rest is algebraic chicanery.

$$\begin{aligned} \Pr(T \geq c\text{Ex}(T)) &= \Pr(c^T \geq c^{c\text{Ex}(T)}) \\ &\leq \frac{\text{Ex}(c^T)}{c^{c\text{Ex}(T)}} \end{aligned}$$

The first step is a shocker; we raise both sides of the inequality in the probability to the power c. Since the new inequality describes the same event as the old, the probabilities are the same. The second step uses Markov's Theorem.

Recall that Markov's Theorem sometimes gives weak bounds and sometimes gives tight bounds. The motivation for the first step is to alter the distribution of the random variable T to hit the "sweet spot" of Markov's Theorem. That is, Markov's Theorem gives a tighter bound on the random variable c^T than on the original variable T.

All that remains is to evaluate $\operatorname{Ex}(c^T)$. This leads to a lot of algebra.

$$\begin{aligned}
\operatorname{Ex}(c^T) &= \operatorname{Ex}(c^{T_1+T_2+\ldots+T_N}) \\
&= \operatorname{Ex}(c^{T_1}) \cdot \operatorname{Ex}(c^{T_2}) \cdots \operatorname{Ex}(c^{T_N}) \\
&= \prod_{i=1}^{N} \operatorname{Ex}(c^{T_i})
\end{aligned}$$

In the first step, we expand T according to its definition. The second step uses linearity of expectation, and in the third step we shift from $\ldots$ notation to product notation. To simplify the product, we can rewrite $\operatorname{Ex}(c^{T_i})$ as follows.

$$\begin{aligned}
\operatorname{Ex}(c^{T_i}) &= c^1 \Pr(T_i = 1) + c^0 \Pr(T_i = 0) \\
&= c^1 \Pr(T_i = 0) + c^0 (1 - \Pr(T_i = 1)) \\
&= 1 + (c-1) \Pr(T_i = 0) \\
&\leq e^{(c-1)\Pr(T_i=1)} \\
&= e^{(c-1)\operatorname{Ex}(T_i)}
\end{aligned}$$

The first step uses the definition of expected value. The second step uses the fact that the probability of an event is one minus the probability of the complement. The third step is simplification. The fourth line uses the inequality $1+x \leq e^x$ which follows from the Taylor series $e^x = 1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots$. The last line relies on the fact that the expected value of a Bernoulli variable is equal to the probability that the variable is 1.

We can now substitute this bound on $\operatorname{Ex}(c^{T_i})$ into the product for $\operatorname{Ex}(c^T)$. The motivation for the approximation $1 + x \leq e^x$ used above is that terms of the form e^x are easy to work with inside a product.

$$\begin{aligned}
\operatorname{Ex}(c^T) &\leq \prod_{i=1}^{N} e^{(c-1)\operatorname{Ex}(T_i)} \\
&= e^{\sum_{i=1}^{N} (c-1)\operatorname{Ex}(T_i)} \\
&= e^{(c-1)\sum_{i=1}^{N} \operatorname{Ex}(T_i)} \\
&= e^{(c-1)\operatorname{Ex}\left(\sum_{i=1}^{N} T_i\right)} \\
&= e^{(c-1)\operatorname{Ex}(T)}
\end{aligned}$$

In the first step, we substitute the bound on $\text{Ex}(c^{T_i})$ into the expression for $\text{Ex}(c^T)$. The second equality follows from an exponent rule. In the third step, we pull the constant $(c-1)$ out of the summation. The fourth step uses linearity of expectation, and the last step uses the definition of the random variable T.

Now we can substitute this bound on $\text{Ex}(c^T)$ into the expression derived from Markov's Theorem. All of the remaining steps are simplifications.

$$\begin{aligned}
\Pr(T \geq c\text{Ex}(T)) &= \frac{e^{(c-1)\text{Ex}(T)}}{c^{c\text{Ex}(T)}} \\
&= e^{(c-1)\text{Ex}(T) - c \ln c \text{Ex}(T)} \\
&= e^{-(\ln c + \frac{1}{c} - 1)c\text{Ex}(T)} \\
&= e^{-\alpha c \text{Ex}(T)} \qquad \text{where } \alpha = \ln c + \frac{1}{c} - 1
\end{aligned}$$

4 A Phone Network Problem

Suppose that there is a phone network that handles a billion calls a day. Some of these call are routed through a certain switch. The exact number of calls passing through this switch is somewhat random and fluctuates over time, but on average the switch handles a million calls a day.

Our problem is to set the capacity of the switch; that is, we must determine the number of calls that a switch is able to handle in a day. If we make the capacity too small, then some phone calls will not go through. On the other hand, if we make the capacity too large, then we are wasting money. Of course, we can not rule out a freak situation in which a huge number of calls are all coincidentally routed through the same switch, thus overloading it. However, we would like to guarantee that a switch is rarely overloaded.

Assume that each call has some probability of passing through a particular switch. In particular, let T_i be an indicator variable for the event that the i-th call passes through the switch. That is, $T_i = 1$ if the call is routed through the switch, and $T_i = 0$ if the call does not pass through the switch. Then the total call load on the switch is $T = T_1 + T_2 + \ldots + T_N$. We do not know the exact probability that the switch handles the i-th call, but we are given that the switch handles an average of a million calls a day; that is, $\text{Ex}(T) = 1{,}000{,}000$.

We will make the crucial assumption that the random variables T_i are mutually independent; that is, calls do or do not pass through the switch mutually independently.

4.1 How to Build One Switch

We can now compute the probability that the load on a switch fluctuates upwards by 1% due to the randomness of calling patterns. Substituting $c = 1.01$ and $\text{Ex}(T) = 1{,}000{,}000$ into the Chernoff Bound gives:

$$
\begin{aligned}
\Pr(\text{particular switch overloaded}) &= \Pr(T \geq 1.01 \cdot 1,000,000) \\
&\leq e^{-1.01(\ln 1.01 + \frac{1}{1.01} - 1) \cdot 1,000,000} \\
&< e^{-1.01(0.00004934) \cdot 1,000,000} \\
&< 2.3 \cdot 10^{-22}
\end{aligned}
$$

The probability that the load on the switch ever rises by even 1% is unbelievably small! (A terrorist attack on Albany Street during an earthquake is far more likely.) If we build the switch with capacity only 1% above the average load, then the switch will (almost) never be overloaded.

The strength of this result relies on the huge expected number of calls. For example, suppose that the average number of calls through the switch were 100 per day instead of 1,000,000. Then every million in the above calculation would be replaced by a hundred; no other numbers change. The final probability of overloading the switch would then be bounded above not by $2.3 \cdot 10^{-22}$, but by 0.995! If the switch handles only 100 calls on an average day, then the call load can very often fluctuate upward by 1% to 101 or more.

4.2 How to Build the Network

We now know that building 1% excess capacity into a switch ensures that it is effectively never overloaded. The next problem is to guarantee that no switch in the entire network is overloaded. Suppose that that are 1000 switches, and every switch handles an average of a million calls a day.

In Section 1 we saw that the probability that some event occurs is at most the sum of the event probabilities. In particular, the probability that some switch is overloaded is at most the sum of the probabilities that each of the 1000 switches is overloaded. Therefore, we have:

$$
\begin{aligned}
\Pr(\text{some switch overloaded}) &\leq 1000 \cdot \Pr(\text{particular switch overloaded}) \\
&< 2.3 \cdot 10^{-19}
\end{aligned}
$$

This means that building 1% excess capacity into every switch is sufficient to ensure that no switch is ever overloaded.

The above results are of limited practical value, however. One problem is that calls do not pass through a switch mutually independently. For example, after an earthquake on Albany Street, everyone would call through the eastern Cambridge switchboard to check on friends and family. Furthermore, there are many more phone calls on certain dates like Mother's Day. On such occasions, 1% excess capacity is insufficient.