

MIT OpenCourseWare
<http://ocw.mit.edu>

14.30 Introduction to Statistical Methods in Economics
Spring 2009

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

14.30 Introduction to Statistical Methods in Economics

Lecture Notes 2

Konrad Menzel

February 5, 2009

1 Probability of Events

So far, we have only looked at definitions and properties of events - some of them very unlikely to happen (e.g. Schwarzenegger being elected 44th president), others relatively certain - but we haven't said anything about the *probability* of events, i.e. how likely an event is to occur relative to the rest of the sample space.

Formally, a probability P is defined as a function from a collection of events $\mathcal{A} = \{A_1, A_2, \dots\}$ in S^1 to the real numbers, i.e.

$$P : \begin{cases} \mathcal{A} & \longrightarrow & \mathbb{R} \\ A & \longmapsto & P(A) \end{cases}$$

In order to get a useful definition of a probability, we require any probability function P to satisfy the following axioms:

(P1) $P(A) \geq 0$ for any event $A \in \mathcal{A}$.

(P2) $P(S) = 1$ - i.e. "for sure, *something* is going to happen"

(P3) For any sequence of *disjoint* sets A_1, A_2, \dots ,

$$P\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} P(A_i)$$

As a mathematical aside, in order for these axioms (and our derivations of properties of $P(A)$ next lecture) to make sense, it must in fact be true that the collection \mathcal{A} in fact contains the event S , and the complements and unions of its members, and this is what constitutes a sigma-algebra as defined in the footnote on the previous page. But again, for this class, we'll take this as given without any further discussion, so you can ignore this fine print for now.

¹For a consistent definition of a probability, this collection of events must satisfy the following properties

(S1) $S \in \mathcal{A}$

(S2) If $A \in \mathcal{A}$, then its complement $A^C \in \mathcal{A}$

(S3) Any countable union of events A_1, A_2, \dots is in \mathcal{A} , i.e. $A_1 \cup A_2 \cup \dots \in \mathcal{A}$

Such a collection of events is called a *sigma-algebra* on S . For the purposes of this class, this is not important, and we'll take it as given that the problem at hand satisfies these axioms.

Definition 1 A probability distribution on a sample space S is a collection of numbers $P(A)$ which satisfies the axioms (P1)-(P3).

Note that the axioms (P1)-(P3) do not pin down a *unique* assignment of probabilities to events. Instead, these axioms only give *minimal* requirements which any probability distribution should satisfy in order to be consistent with our basic intuitions of what constitutes a probability (we'll actually check that below). In principle any function $P(\cdot)$ satisfying these properties constitutes a valid probability, but we'd have to see separately whether it's actually a good description of the random experiment at hand, which is always a hard question. In part 5 of this class ("Special Distributions"), we'll discuss a number of popular choices of $P(\cdot)$ for certain standard situations.

2 Some Properties of Probabilities

Now we still have to convince ourselves that the axioms (P1)-(P3) actually are sufficient to ensure that our probability function has the properties we would intuitively expect it to have, i.e. (1) the probability that an event happens plus the probability that it doesn't happen should sum to one, (2) the probability that the impossible event, \emptyset , happens should equal zero, (3) if an event B is contained in an event A , its probability can't be greater than $P(A)$, and (4) the probability for any event should be in the interval $[0, 1]$. We'll now prove these properties from the basic axioms.

Proposition 1

$$P(A^C) = 1 - P(A)$$

PROOF: By the definition of the complement A^C ,

$$1 \stackrel{(P2)}{=} P(S) \stackrel{\text{Defn. } A^C}{=} P(A \cup A^C) \stackrel{(P3)}{=} P(A) + P(A^C)$$

where the last step uses that $A \cap A^C = \emptyset$, i.e. that A and its complement are disjoint. Rearranging this, we get

$$P(A^C) = 1 - P(A)$$

which is what we wanted to show \square

Proposition 2

$$P(\emptyset) = 0$$

PROOF: Since $\emptyset^C = S$, we can use the previous proposition to show that

$$P(\emptyset) = P(S^C) \stackrel{\text{Prop.1}}{=} 1 - P(S) \stackrel{(P2)}{=} 1 - 1 = 0$$

Proposition 3 If $B \subset A$, then $P(B) \leq P(A)$.

As an aside, cognitive psychologists found out that even though this rule seems very intuitive, people often violate it in everyday probabilistic reasoning.² PROOF: In order to be able to use the probability

²E.g. in a study by the psychologists Daniel Kahneman and Amos Tversky, several people were given the following description of Linda:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Individuals who were asked to give the probability that Linda was a bank teller tended to state a lower figure than those asked about the probability that she was a *feminist* bank teller.

axioms, it is useful to partition the event A using properties of unions and intersections

$$A = A \cap S = A \cap (B \cup B^C) = (A \cap B) \cup (A \cap B^C) = B \cup (A \cap B^C)$$

where the last step uses that $B \subset A$ implies that $A \cap B = B$. Now in order to be able to use axiom (P3), note that B and $A \cap B^C$ are disjoint:

$$B \cap (A \cap B^C) = B \cap (B^C \cap A) = (B \cap B^C) \cap A = \emptyset \cap A = \emptyset$$

Therefore, we can conclude

$$P(A) = P(B) + P(A \cap B^C) \geq P(B)$$

by axiom (P1) \square

Proposition 4 For any event A , $0 \leq P(A) \leq 1$.

PROOF: $0 \leq P(A)$ is axiom (P1). For the second inequality, note (P1) also implies that $P(A^C) \geq 0$. Therefore by proposition 1

$$P(A) = 1 - P(A^C) \leq 1$$

Proposition 5

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

PROOF: Note that, as in the proof of proposition 3, we can partition the events A and B into

$$A = A \cap S = A \cap (B \cup B^C) = (A \cap B) \cup (A \cap B^C)$$

and in the same fashion

$$B = (B \cap A) \cup (B \cap A^C)$$

You can easily check that these are in fact partitions, i.e. each of the two pairs of sets is disjoint. Therefore, we can see from axiom (P3) that

$$P(A) = P((A \cap B) \cup (A \cap B^C)) = P(A \cap B) + P(A \cap B^C)$$

and

$$P(B) = P(B \cap A) + P(B \cap A^C)$$

Therefore

$$P(A) + P(B) = P(A \cap B) + [P(A \cap B^C) + P(B \cap A) + P(B \cap A^C)] = P(A \cap B) + P(A \cup B)$$

by (P3) since $(A \cap B), (A \cap B^C), (B \cap A^C)$ is a partition of $A \cup B$ (figure 6 gives a graphical illustration of the idea). Rearranging the last equation gives the desired result \square

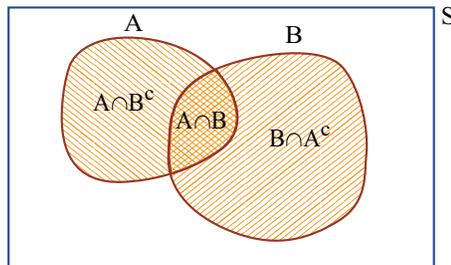


Image by MIT OpenCourseWare.

Figure 1: Partition of $A \cup B$ into disjoint events

3 Example: “Simple” Probabilities

Suppose we have a finite sample space with outcomes which are ex ante symmetric in the sense that we have no reason to believe that either outcome is more likely to occur than another. If we let $n(C)$ denote the number of outcomes in an event C , we can define a probability

$$P(A) := \frac{n(A)}{n(S)}$$

i.e. the probability equals the fraction of all possible outcomes in S that are included in the event A . This distribution is called the “simple” probability distribution or also the “logical” probability. A randomization device (e.g. a coin or a die) for which each outcome is equally likely is said to be *fair*. Now let’s check whether the three axioms are in fact satisfied:

(P1): $P(A) \geq 0$ follows directly from the fact that $n(\cdot)$ only takes (weakly) positive values.

(P2): $P(S) = \frac{n(S)}{n(S)} = 1$.

(P3): For two disjoint events A and B ,

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A) + n(B) + n(A \cap B)}{n(S)} = \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} = P(A) + P(B)$$

For more than two sets, the argument is essentially identical.

Example 1 Suppose a fair die is rolled once. Then the sample space equals $S = \{1, 2, \dots, 6\}$, so $n(S) = 6$. What is the probability of rolling a number strictly greater than 4? - since the event is $A = \{5, 6\}$, $n(A) = n(\{5, 6\}) = 2$. Hence $P(A) = \frac{n(A)}{n(S)} = \frac{2}{6} = \frac{1}{3}$.

If a die is rolled twice, what is the probability that the sum of the numbers is less than or equal to 4? Let’s check: $S = \{(1, 1), (1, 2), \dots, (2, 1), (2, 2), \dots, (6, 6)\}$, so that $n(S) = 6^2 = 36$. The event

$$B = \text{“Sum of Dice } \leq 4\text{”} = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)\}$$

so that $P(B) = \frac{n(B)}{n(S)} = \frac{6}{36} = \frac{1}{6}$.

In a minute, we’ll look at more sophisticated techniques for enumerating outcomes corresponding to certain events.

4 Counting Rules

The examples we looked at so far were relatively simple in that it was easy to enumerate the outcomes in A and S , respectively. If S has many elements, and an event A is sufficiently complex, it may be very tedious and impractical to obtain $n(A)$ and $n(S)$ by going down the full list of outcomes. Today, we'll look at *combinatorics* part of which gives simple rules for counting the number of combinations or permutations of discrete objects (outcomes) corresponding to a different pattern (event).

Example 2 *The famous chess player Bobby Fischer (who died 3 weeks ago) eventually got bored of playing "classic" chess and proposed a variant in which only the 8+8 pawns are set up as usual, but the other pieces (1 king, 1 queen, 2 bishops, 2 knights, 2 rooks) are put in random positions on the first rank, where each white piece faces the corresponding black piece on the other side. As further restrictions, (1) one bishop must be on a black, the other on a white square, and (2) the king must start out between the two rooks (in order to allow for castling).*

The idea behind this is that since chess players tend to use standard game openings which work well only for the standard starting positions, the new variant forces them to play more creatively if there are enough possible ways of setting up a game to make it impossible for players to memorize openings for any constellation. But how many possible starting positions are there?

We'll actually do the calculation later on in the lecture today using some of the counting techniques introduced in this class. If you get bored, you can already start figuring out a (preferably elegant) way of attacking the problem.

For now, we will not explicitly talk about random experiments or probabilities, but digress on methods to enumerate and count outcomes which we will put to use later on in the lecture.

4.1 Composed Experiments

Rule 1 (Multiplication Rule): *If an experiment has 2 parts, where the first part has m possible outcomes, and the second part has n possible outcomes regardless of the outcome in the first part, then the experiment has $m \cdot n$ outcomes.*

Example 3 *If a password is required to have 8 characters (letters or numbers), then the corresponding experiment has 8 parts, each of which has $2 \cdot 26 + 10 = 62$ outcomes (assuming that the password is case-sensitive). Therefore, we get a total of 62^8 (roughly 218 trillion) distinct passwords. Clearly, counting those up by hand would not be a good idea.*

Example 4 *The standard ASCII character set used on most computer systems has 127 characters (excluding the space): each character is attributed 1 byte = 8 bit of memory. For historical reasons, the 8th bit was used as a "parity" bit for consistency checks to detect transmission or copying errors in the code. Therefore, we have an experiment of 7 parts, each of which has outcomes from $\{0, 1\}$, so we have a total number of $2^7 = 128$ distinct characters.*

Example 5 *A card deck has 52 cards, so if we draw one card each from a blue and a red deck, we get $52 \cdot 52 = 2704$ possible combinations of cards (if we can't tell ex post which deck each card came from, we get a smaller number of distinguishable outcomes). If, on the other hand, we draw two cards from the same deck without putting the first card back on the stash, regardless of which card we drew first, only 51 cards will remain for the second draw. Of course which 51 cards are left is going to depend on which card was drawn at first, but notice that this doesn't matter for the multiplication rule. Therefore, if we draw two cards from the same deck, we'll have $52 \cdot 51 = 2652$ possible combinations.*

The last example illustrates two types of experiments that we want to describe more generally: *sampling with replacement* versus *sampling without replacement*, each of which has a different counting rule.

- n draws from a group of size N *with replacement*:

$$\underbrace{N \cdot N \cdot \dots \cdot N}_{n \text{ times}} = N^n$$

possible outcomes.

- n draws from a group of size N *without replacement* (where $N \geq n$):

$$P_{N,n} := N(N-1)(N-2)\dots(N-(n-1)) = \frac{N(N-1)(N-2)\dots 3 \cdot 2 \cdot 1}{(N-n)(N-(n+1))\dots 3 \cdot 2 \cdot 1} = \frac{N!}{(N-n)!}$$

possible outcomes, where $k! := 1 \cdot 2 \cdot \dots \cdot (k-1)k$ (read as "k-factorial"), and we define $0! = 1$.

In fact, both of these enumeration rules derive from the multiplication rules, but since they are very prominent in statistics, we treat them separately.

4.2 Permutations

Example 6 *A shuffled deck of cards is a permutation of an ordered deck of cards: it contains each card exactly once, though the order will in most cases be different.*

Definition 2 *Any ordered rearrangement of objects is called a permutation.*

Note that generating permutations corresponds to drawing N out of a group of N without replacement.

Example 7 *Dodecaphony is a composition scheme in modern classical music in which each piece is based on a tone row in which each of the twelve notes of the chromatic scale (C, C sharp, D, D sharp, etc. up to B) appears exactly once. Therefore, each tone row is a permutation of the chromatic scale, and we could in principle count the number of all possible distinct "melodies" (about 479 million).*

Example 8 *The famous traveling salesman problem looks at a salesman who has to visit, say, 15 towns in an arbitrary order, and given distances between towns, we are supposed to find the shortest route which passes through each town on the list (at least) once. Using our formula for drawing 15 out of a group of 15, we can calculate that there are $15!$, which is about 1.3 trillion, different paths, so this is a complicated problem, and we won't solve it.*

We could imagine that in each town, the salesman has to visit 5 customers. If we consider all possible paths from customers to customers, we get $(15 \cdot 5)!$ permutations (that's a lot!). However, it may seem sensible to restrict our search to travel plans according to which the salesman meets with all 5 customers once he is in town (in an order to be determined). There are $5!$ possible orders in which the salesman can see customers in each town, and $15!$ possible orders in which he can visit towns, so we can use the multiplication rule to calculate the number of permutations satisfying this additional restriction as

$$\underbrace{(5!5! \dots 5!)}_{15 \text{ times}} 15! = (5!)^{15} 15!$$

*which is still an insanely high number, but certainly much less than the $(15 \cdot 5)!$ unrestricted permutations.*³

³Few people, if any, have a good intuition for the scale of factorials since $k!$ grows extremely fast in k . Stirling's

4.3 Combinations

Example 9 *If we want to count how many different poker hands we can draw from a single deck of cards, i.e. 5 cards drawn from a single deck without replacements, we don't care about the order in which cards were drawn, but rather whether each of the card was drawn at all.*

Definition 3 *Any unordered collection of elements is called a combination.*

A combination constitutes a draw without replacement from a group, but since we now do not care about the order of elements, we don't want to double-count series of draws which consist of the same elements, only in different orders. For a collection of n elements, there are $n!$ different orders in which we could have drawn them (i.e. the number of permutations of the n elements). Therefore the number of different combinations of n objects from N objects is

$$C_{N,n} = \frac{\# \text{ outcomes from drawing } n \text{ out of } N \text{ without replacement}}{\# \text{ orders in which can draw } n \text{ elements}} = \frac{N!}{(N-n)!n!}$$

This number is also known as the *binomial coefficient*, and often denoted as

$$\binom{N}{n} := \frac{N!}{(N-n)!n!}$$

Note that, even though we look at a ratio of factorials, the binomial coefficient always takes integer values (as it should in order for the number of combinations to make sense).

Example 10 *For poker, we can use this formula to calculate that there are $\binom{52}{5} = 2598960$ possible hands.*

Example 11 *A functional study group should not have more than, say, 5 members (there is no pedagogical justification for this number, but I just want to keep the math from becoming too complicated). There are currently 28 students registered for this class. How many possibilities for viable study groups (including students working on their own) would be possible? We'll have to calculate the number of study groups for each group size 1, 2, 3, 4, 5 and add them up, so that (if I didn't make any mistakes) there are*

$$S = \binom{28}{1} + \binom{28}{2} + \binom{28}{3} + \binom{28}{4} + \binom{28}{5} = 28 + 378 + 3,276 + 20,475 + 98,280 = 122,437$$

possible study groups.

Now back to our "challenge problem" from the beginning of the class:

approximation, which works quite well for "high" values of k is

$$k! \approx \sqrt{2k\pi} \left(\frac{k}{e}\right)^k$$

In the pop-sci literature a common comparison to illustrate extremely large numbers involves the estimated total number of atoms in the observable universe, which is about 10^{80} (well, I don't have an intuition for that either!). In terms of factorials, $10^{80} \approx 59!$. The number $75!$ can be expressed as roughly $2.5 \cdot 10^{30}$ (two and a half million trillion trillion) times the number of atoms in the universe.

Since you'd want to avoid calculations involving such high numbers, note that for most purposes, we only have to deal with ratios of factorials, so we should first see which terms cancel, e.g. $\frac{98!}{94!} = \frac{98 \cdot 97 \cdot 96 \cdot 95 \cdot 94!}{94!}$.

Example 12 *Back to Fischer Random Chess: Let's first ignore the restrictions (1) and (2) about the rooks and bishops, i.e. allow for any allocations of the pieces on the bottom rank of the board. We have to allocate the 8 white (or black, this doesn't matter) pieces onto the 8 squares in the first rank of the chessboard. Notice that this is a permutation, so that we have 8! possible orderings. However, rooks, knights and bishops come in pairs and for the rules of the game the "left" and the "right" piece are equivalent. Therefore, there are $2 \cdot 2 \cdot 2$ possible ways of generating each starting position by exchanging the two rooks/knights/bishops with each other, respectively. Hence, the number of distinct games is*

$$\# \text{Games} = \frac{8!}{8} = 7! = 5040$$

As we said earlier, the actual rules for Fischer Random Chess impose furthermore that (1) one bishop is placed on a black, and the other on a white square, and (2) that the king has to start out between the two rooks. For this variant, we can use the multiplication rule if we are a little clever about the order in which we fill up the row: I propose that we first allocate the two bishops, one on a random white, the other on a random black square, so there are $4 \cdot 4$ possibilities. Next, the queen takes one out of the remaining 6 squares (6 possibilities, obviously). Now we put the two knights on any of the five squares that are left.

This is a combination, so there are $\binom{5}{2} = \frac{120}{6 \cdot 2} = 10$ ways of allocating the knights. Because of the restriction on the king and the rooks, there is always exactly one way of setting up the three pieces onto the three remaining free fields. In sum, we have

$$\# \text{Games} = 4 \cdot 4 \cdot 6 \cdot 10 \cdot 1 = 960$$

potential "games" to be played.

The crucial point about the order in which we place the pieces is to make sure that we can apply the multiplication rule, i.e. that the way we place the first pieces does not affect the number of possibilities we have left for the remaining pieces. As far as I can see, this only matters for the bishops: Say, we placed the rooks and the king first and then put up the bishops. Then we'd have to distinguish whether (a) all three pieces are on fields of the same color (so we'd have $1 \cdot 4 = 4$ possibilities of placing the bishops on fields of different colors), or (b) one of the three pieces stands on a field of a different color than the other two (leaving us with $2 \cdot 3 = 6$ possibilities for the bishops). As long as we place the bishops first and the king before the two rooks, it seems to be irrelevant in which order we proceed thereafter.

5 The Presidential Death Date Paradox (Recitation)

Conspiracy theories about living and dead presidents are typically built around "weird" coincidences. For example, for the two American presidents who were assassinated, i.e. Lincoln and Kennedy, one can find long lists of more or less remarkable commonalities between the two - e.g. Lincoln purportedly had a secretary named Kennedy who had warned him not to go the theater where he was shot, whereas Kennedy had a secretary named Evelyn Lincoln who had warned him not to go to Dallas before his assassination (well, that's at least what Wikipedia says...).

One particular coincidence is the fact that several of the 39 presidents who already died share the same death dates: Filmore and Taft both died on March 8. John Adams and Thomas Jefferson both died on July 4th in 1826, exactly 50 years after the signing of the declaration of independence, and James Monroe died exactly five years later, on July 4, 1831. Is this something to be surprised about?

Let's first look at the simple probability that two *given* presidents died on a *fixed* day, say February 6, assuming that probabilities equal the proportion of outcomes belonging to that event: we get that there is only one combination of the two presidents' death dates falls on February 6th, but by the multiplication

rule for enumerations, there are a total of 365^2 possible combinations of death dates. Hence the probability for this event is $1/365^2$ which is an extremely small number.

However, there is also a large number of pairs of presidents and days in the year which could be potential candidates for a double death date. Now, the probability of the event A that at least two out of the 39 presidents died on the same day can in principle be calculated as the ratio of the number of all possible combinations that have one pair of presidents with the same death date, two pairs, three presidents on one death date, and so forth. A more elegant way to attack this problem is by noting that, since $A \cup A^C = S$ and $A \cap A^C = \emptyset$, from the axioms P3 and P2 we have $P(A) = P(S) - P(A^C) = 1 - P(A^C)$. The event A^C can be formulated as "all 39 dead presidents have different birthdays." If there were only two dead presidents, there are 364 ways in which, given the death date of the first, the death date of the second could fall on a different date. You should now note that counting the number of possibilities of assigning *different* death dates to each of the n presidents corresponds to the experiment of drawing n out of 365 days *without replacement*, so the number of possibilities is $\frac{365!}{(365-n)!}$. The total number of possible assignments of death dates to presidents corresponds to drawing dates *without replacement*, so the number is 365^n .

Hence the probability of having at least one pair of presidents with the same death date is

$$P(A) = 1 - P(A^C) = 1 - \frac{365!}{365^n(365 - n)!}$$

which, for $n = 39$ is equal to about 87.82%. Using the formula, we can also calculate this probability for different numbers of presidents:

n	P(A)
1	0
2	0.27%
5	2.71%
10	11.69%
20	41.14%
30	70.63%
60	99.41%
366	100.00%

As we can see from the last line, one intuition for the probability increasing to one is that we eventually "exhaust" the number of potential separate death dates until eventually we have more dead presidents than death dates.

So, to resolve the paradox, while the event that two *given* presidents died on a *fixed* date would indeed be a great coincidence (i.e. has a very low probability), with an increasing number of presidents, there is a combinatorial explosion in the number of different constellations for such an event. In other words, while each individual outcome remains very unlikely, the "number of potential coincidences" increases very steeply, so that with a high probability at least *some* coincidences must happen.

The story behind the other conspiracy theories is presumably the same: people have been combing through zillions of details trying to find only a relatively tiny number of stunning parallels between Lincoln and Kennedy. In statistics, this search strategy is commonly called "data mining," and in this context we speak of those rare coincidences which actually occur as "false discoveries," which are typically not due to any systematic relationships, but simply result from the sheer numbers of potential relationships that we might investigate or test simultaneously.