

1 The basics of probability theory

E.R. Morey: BasicsOfProbabilityTheory.tex and .pdf September 8, 2010

1.1 What does probability mean?

1.1.1 What is the probability of some event?

In the "street sense" of the word, probability of an event is a *measure of likeliness*: how likely is it that the event will happen - the more likely it is that the event will occur, the higher the probability number. That is, if the likelihood of an event occurring increases, its probability increases. Note that *probability* is from *probable*.

The guy on the street would likely put a few restrictions on this measure of likelihood. He would probably add:

1. The probability of an event cannot be less than zero
2. The probability of an event cannot be greater than 1 (greater than 100%)
3. The probability of something happening is 1
4. And, the probability of nothing happening is zero.

Us statisticians agree with the guy on the street that probability has these four properties.

But, these are just conventions. For example, there is no reason "likelihood" could not be measured on a scale from -3 to 1.5 where -3 corresponds to certainty. On Mars they might have a different range on probability.

Note a few things about this street definition of probability. It is an **ex ante** concept, not an **ex post** concept: once something has happened we know what happened (what event occurred). Once an event has occurred, it is certain; it is not a random variable, and does not have a probability of occurrence.

Sampling and probability are two side of the same thing: Consider the question of whether it will rain in Boulder tomorrow as a function of what the weather in Boulder is today and the weather elsewhere. Let's assume that the weather tomorrow is a function of the both of these observables, plus some random component.

A sample is a realization of an experiment with uncertain outcomes, the realization of a stochastic data-generating process.

That is, tomorrow's weather is, today, a RV with some probability density function where that probability density function depends on the weather today both in Boulder and elsewhere. Tomorrow's weather will be a draw from that distribution. Tomorrow, we might draw (sample) a sunny day, or we might draw (sample) a rainy day. Today, all we can ask is what is the probability that tomorrow will be sunny given today is bla, bla, bla.

For example, consider a discrete distribution with three spikes: one for the probability of rain, one for sun, and one for snow, where the height of each spike depends on whether today was rain, sun or snow.

(possibly insert graph here, maybe create in excel, and use the excel macro that converts stuff to .tex)

What if the issue was tomorrow's temperature. Imagine we assumed temperature tomorrow would be a random draw from a normal distribution whose mean was today's temperature, or whose mean was a weighted average of today and yesterday's temperatures.

An observation, what happens on a day, is a draw from the population of possible weathers. Or said differently, there is a underlying process with a random component that will generate tomorrow's weather, and tomorrow's actual weather is one of many possible outcomes. Tomorrow we will sample the weather (we are forced to sample it) and see what we get.

We want to determine, or estimate, the probability that an observation in a sample will be event A . For example, the probability that it will rain tomorrow (rain is an event).

The underlying process and what happens tomorrow can all be thought of in terms of randomly drawing a colored ball from an urn. For example, based on today's weather, the urn for tomorrow's weather holds two white balls (white means snow), five red balls (red means sunny) and three grey balls (grey means

rain).¹ Tomorrow's weather is a draw from the urn. It will be either snow, rain, or sun.² Once the ball is drawn the outcome is known. Beforehand, we want to know the likelihood (probability) of each outcome.

For example, in my research I am often interested in the probability that individual i will choose alternative j from some finite number of alternatives J . The probability of choosing each alternative will depend on its characteristics and the characteristics of the other alternatives in the choice set. Guys in marketing and transportation love these models.

This type of model is called a discrete-choice model:

I am going skiing, which area will I choose.

Every day, unmarried people "decide" whether to get married or not, every day, married people decide whether to get divorced.

Coke or Pepsi. McDonalds, Wendys or Burger King.

In terms of urns, every possible configuration of characteristics for the J alternatives is represented by a different urn - they could be hundred or thousand of different configurations of alternatives, so that number of urns.

Each of the different urns will contain numbered balls, the number on the ball corresponding to the number of the alternative. Different urns will have different proportions of the different numbered balls. An observed choice is a draw from the urn that represents the current configuration of the alternatives in terms of their characteristics.

For example, if the problem is determining the probability of where to ski as a function of the number of ski areas, their locations, and their characteristics, one could imagine an urn for each possible configuration of ski areas. To simplify, assume there are always 5 ski areas and their locations are fixed. Assume that what varies from day to day, and across ski areas, is snow conditions (good or bad) and temperature (cold, nice, and too warm). How many urns would be needed.

¹If today's weather were different, the composition of the balls in this urn would be different, or, thinking of it another way, there could be a different urn for every possible state of today's weather.

In such a world, the weather predictor might own three urns: the urn for if it rained today, the urn for if it snowed today, and the sun urn), be able to observe today's weather, and, at midnight predict tomorrow's weather by drawing a ball from the appropriate urn. every morning. The draw is the weather forecast for the next day.

Note that the weather-person's draw will not necessarily be tomorrow's weather. That is determined by a draw by the weather god.

²Note the assumption that only three things can happen: snow, rain, or sun, so weather is assumed a discretely-distributed random variable. It can take only three values.

In each urn there would be balls of five colors, a different color for each ski area. The proportion of balls by color would vary across urns.

One wakes up in the morning, finds the urn that corresponds to today's conditions. The probability of choosing the red ski area is then the probability of drawing a red ball from that urn. Once the ball is drawn your choice has been made.

1.2 There are three notions of probability (actually there are four)

Classical probability

Frequency probability

Axiomatic probability

and "subjective" probability

Historical development of probability theory: Classical \rightarrow Frequency \rightarrow Axiomatic

The Axiomatic definition encompasses the Classical and Frequency definitions of probability, and subjective probability

Note that in a number of places in these notes I use *axiom* as a synonym for *assumption*. We don't prove axioms - they are the "givens"

1.3 Subjective probability

Is a concept familiar to the guy on the street, but not one that we will not deal with in Econ 7818.

When we talk about probability, we will restrict ourselves to discussing the probabilities associated with the outcomes of experiments that are, at least in theory, repeatable - experiment that have a stable data-generating process)

Subjective probability, to the extent that I understand, I am not sure I do, is the likelihood of an event occurring when the process that generated the outcome cannot be repeated.

For example, Wanda is going out on a first date and she, and her date, might both think about the probability that Wanda will get "lucky." While the likelihood of this is of interest to Wanda, it is not to us (at least not in 7818) because what happens on the date is not the realization of what happens on a random draw from a "constant" random process. One can't repeat a "first date."

That said, I have a little trouble understanding when the probability is, and is not subjective. I understand the part about a stable process, but why can't the outcome of Wanda's date be the outcome from one run of a stable process.

The probability that George loves me is a subjective probability. Subjective probabilities are important, just not in econometrics. Feller, in his book *Introduction to Probability Theory: Volume I* (page 4), uses the example "Paul is *probably* [*ital* added] a happy man." noting that "In a rough way we may characterize this concept by saying that probabilities do not refer to judgements but to possible outcomes of a conceptual experiment."

Maybe some experiments are just not repeatable. Maybe before I married Wanda Sue it was reasonable to think about whether we would marry as the outcome of some definable stochastic process, but once we are married and then divorced - stuff has changed - whether we marry again is not determined by the same process. In other words, **the draw changes the contents of the urn.**

1.4 Classical probability

Definition 1 *Classical probability (MBG):* If a random experiment (process with an uncertain outcome) can result in n mutually exclusive and equally likely outcomes, and if n_A of these outcomes has an attribute A , then the probability of A is the fraction (n_A/n) .³

Let $\Pr[A]$ be the abbreviation for the "probability of A ."

This notion of probability had its conception in the study of games of chance; in particular, fair games of chance.

In explanation, many games are "fair" in the sense that each outcome has an equal chance of occurring. To win such a game one needs to figure out the probabilities associated with events of interest. For example, in poker, what is the probability that your opponent has three-of-a-kind.

E.g. if one tosses a coin, there are two mutually exclusive outcomes: head or tail. Of these two outcomes, one is associated with the attribute heads; one is associated with the attribute tails. If the coin is fair, each outcome is equally likely. In which case, $\Pr[\text{head}] = \frac{n_A}{n} = \frac{1}{2}$, where $n = 2$ and n_A is the number of possible outcomes associated with a head (1).

Consider some other examples:

1. The roll of a die: There are 6 equally likely outcomes. The probability of each is $1/6$.
2. Draw a card from a deck: There are 52 equally likely outcomes.
3. The roll of two die: There are 36 equally likely outcomes (6×6): 6 possibilities for the first die, and 6 for the second.⁴ The probability of each outcome is $1/36$.
4. Drawing (with replacement) four balls from an urn with an equal number of red, white, and blue balls: There are 81 possible outcomes ($3 \times 3 \times 3 \times 3 = 3^4$). For example, {red, white, white, blue} is an outcome which is a different outcome from {white, white, red, blue}. The probability associated with each outcome is $1/81$.
5. The toss of two coins: The four possible outcomes are (H, H) , (H, T) , (T, H) and (TT) . The probability of each is $1/4$.
6. The draw of two cards: There are 52^2 possible outcomes.

³ n and n_A must both be finite numbers and, of course, $n_A \leq n$.

⁴Note that 5, 1 is a different outcome than 1, 5.

Terms to note in the definition of classical probability are *random*, *n*, *mutually exclusive*, and *equally likely*.

Axiom 2 *A Basic assumption in the definition of classical probability is that n is a finite number; that is, there is only a finite number of possible outcomes.*

If there is an infinite number of possible outcomes, the probability of an outcome is not defined in the classical sense.

Definition 3 *mutually exclusive: The random experiment result in the occurrence of only one of the n outcomes. E.g. if a coin is tossed, the result is a head or a tail, but not both. That is, the outcomes are defined so as to be mutually exclusive.*

Definition 4 *equally likely: Each outcome of the random experiment has an equal chance of occurring.*

Definition 5 *random experiment: A random experiment is a process leading to at least two possible outcomes with uncertainty as to which will occur.*

Definition 6 *sample space: The collection of all possible outcomes of an experiment. If I had to guess, I would say it is called "sample space" because it is the collection (set) of all possible samples.*

As important thing to note is that classical probabilities can be **deduced** from knowledge of the sample space and the assumptions. Nothing has to be observed in terms of outcomes to deduce the probabilities. **No estimation is required** to determine classical probabilities. Probability of an outcome is just the number of times that outcome can occur divided by the total number of possible outcomes.

For example, the probability of drawing a queen from a fair deck of cards is $\frac{4}{52} = \frac{1}{13}$, and the probability of drawing a queen from a fair deck after one of the queens has been removed is $\frac{3}{51}$.

In classical probability, probabilities are deduced, not estimated: no sampling is required.

Note that if one draws, without replacement, a random sample of size N from a population with a finite number of members, M , Classical probability has relevance: there is a finite number of possible samples and each is equally likely. How many? The answer depends on whether you define two sample with the same elements but drawn in a different order, the same or a different sample. (see the asides, at the end of these notes)

1.5 Frequency Probability

What if N is not finite? In that case, the classical definition is not applicable. What if the outcomes are not equally likely? Again, the classical definition of probability is not applicable.

In such cases, how might we define the probability of an outcome that has attribute A .

We might take a random sample from the population of interest and identify the proportion of the sample with attribute A . That is, calculate Relative freq of A in the sample

$$= \frac{\text{relative frequency of } A \text{ in the sample}}{\text{number of observ. in sample with attribute } A} \\ = \frac{\text{number of observ. in sample with attribute } A}{\text{number of observ. in sample}}$$

Axiom 7 *Relative freq of A in the sample is an estimate of $\Pr[A]$*

The foundation of this approach is that there is some $\Pr[A]$ - we assume it exists. We cannot deduce it, as in Classical probability, but we can estimate it.

For example, one tosses a coin, which might or might not be fair, 100 times and observes heads on 52 of the tosses. One's estimate of the probability of a head is .52. Frequency probability allows one to estimate probabilities when Classical probability provides no insight.

Another example: assume that there is some probability of dying of toe fungus, $\Pr[D_{tf}]$ and by digging up dead people one can determine whether they died of toe fungus. $\Pr[D_{tf}]$ is not known and cannot be determined, but one can estimate it by taking a sample of dead people and seeing what proportion died of toe fungus. This proportion is your estimate of $\Pr[D_{tf}]$.

1.6 Axiomatic Approach to Probability

Put simply, the axiomatic approach to probability builds up probability theory from a number of assumptions (axioms). One makes assumption about what properties *probabilities* and *probability functions* (neither yet defined) should have, and then one derives theorems about probability and probability functions from those assumptions.

Think of it as specifying a model: one defines a bunch of terms, makes a bunch of assumption, and then derives predictions (hypotheses, "if...then" statements). These predictions follow logically from the definitions and assumptions.

While you are free to invent whatever kind of probability theory you want, everyone else on this planet pretty much agrees on the axioms/assumptions in "axiomatic probability theory"

Axiom 8 *There is some sample space, Ω : the collection of all possible outcomes of an experiment*

For example, if the experiment is tossing a coin once $\Omega = \{H, T\}$. If the experiment is rolling a die once $\Omega = \{1, 2, 3, 4, 5, 6\}$. If the experiment is tossing two coins $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$. If the experiment is taking a sample of one from the U.S. population, $\Omega = \{\text{population of the U.S.}\}$. If the experiment is drawing a random integer, Ω is the set of all integers (there are an infinite number of these).

You should try and come up with some more interesting and insightful examples.

Let ω denote an element in Ω

Imagine that one takes a random sample of 100 individuals from the U.S. population. How many elements are there in the sample space: how many different samples are possible?

When answering, consider whether the order in which individuals are chosen is important. Note that if the sampling process is random, each possible sample is equally likely. We talk about random samples, but what we really mean is that the sampling process is random (the process is such that every sample has an equal chance of being drawn).

Your answer also depends on whether you are sampling with or without replacement.

1.6.1 Events

Definition 9 *Event A (MGB): a subset of the sample space. The set of all events associated with an experiment is defined as the "event space", \mathcal{A}*

Put roughly, events are defined in terms of the properties of outcomes (e.g., the outcome is an odd number, or the outcome/sample includes someone named Shirley, *oddness* and *Shirleyness* are properties that a set might have).

Put simply, event A is the set of samples that have the property associated with A , and \mathcal{A} is the set of all events: a set of sets, where each set is associated with a different event.

It helps me to think of events as things one might bet on. For example, if a coin is to be flipped twice one might bet that the outcome of the two flips will be the same (you would then win if the outcome were HH or TT). Or one might bet whether a random sample of five students taken from this class will have at least three women. In the coin example, the event is "two of a kind;" in the second example, the event is "at least three women."

If the outcome of the experiment/draw has the specified property of a set, that event is realized.

Note that the above "formal definition" of A is not a complete definition: it identifies a necc. condition of something to be an event, but not the necc. and suff. conditions.

In explanation, being a subset of Ω ($A \subset \Omega$) is a necessary but not a sufficient condition for being an event.

That is, every event is a subset of the sample space, but some subsets of some sample space are not events.

If the number of outcomes in Ω is finite, then all subsets of Ω are events, and $A \subset \Omega$ is necc. and suff. for A to be an event. However, if Ω is not finite (has an infinite number of elements) it is possible to have a subset of Ω that is not an event. Why?⁵

⁵MGB give an "answer" on page 23 of the third edition. I will return to this issue after we have introduced the concept of a probability function. Reading Feller, *An Introduction to Probability Theory and Its Applications, Vol 1, Third edition*, Feller seems to say, disagreeing with MGB, that a subset of the sample space and an event are the same things. I will return to this "inconsistency" between MGB and Feller, also after probability functions have been defined.

Some examples of event sets:

1. the toss of a single coin: $\Omega = \{H, T\}$. The possible events are H , T , neither H nor T , and either H or T , so four in total.
2. the toss of two coins: $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$. How many events are there? (H, H) , (H, T) , (T, H) , (T, T) , "one of the these outcomes", "none of these outcomes", "at least one head" ((H, H) , (H, T) or (T, H)), "at least one tail" ((H, T) , (T, H) or (T, T)), "a head first" ((H, H) , (H, T)), "a head second" ((H, H) , (T, H)), "a tail first" ((T, H) , (T, T)), and "a tail second" ((H, T) , (T, T))Can you think of others? There are 16 in total.
3. drawing 4 cards from a deck of cards: Events include all spades, sum of the 4 cards is ≤ 20 (assuming face cards have a value of zero), a sequence of integers, a hand with a 2, 3, 4 and 5. There are many events. How many?
4. An example in MGB: randomly choose a light bulb to observe and record time it takes to burn out. $\Omega = \{x : x \geq 0\}$ ⁶
Define event $A(k, m)$ as $A \equiv \{x : k \leq x \leq m\}$ where this event is the light bulb lasts between k and m hours.

⁶Note that there are an infinite number of outcomes, and events

In econometrics you will be concerned with the probability of events. For example, you might be concerned with the probability that the winner is a female or that the sample mean is less than 5 given that the population mean is 0.

We need only define those events we care about. (Why bother to define events we don't care about?)

Typically, one uses capital letters to represent events. E.g. A , B , C ; or A_1 , A_2 , A_3 , etc. As noted above, I will use \mathcal{A} to represent the set of all possible events.

Counting the number of events when the number of outcomes in Ω is finite If there are M elements in Ω , M finite, then Ω has 2^M subsets, including the event \emptyset and the event Ω .

Since events are subsets of Ω , there cannot be more than 2^M events, and typically there will be 2^M events. That is, \mathcal{A} has 2^M elements if there are M elements in Ω .⁷

Note that we are typically not interested in all possible events. We want only to determine the probability of observing those events that are of interest.

⁷One wants to count all of the possible subsets of Ω , and, for this, the binomial coefficient $\binom{M}{n}$ proves useful. $\binom{M}{n}$ counts all of the subsets of M that have n members (order immaterial). Therefore, there are $\binom{M}{1} = M$ subsets (elementary events (each possible sample)); and there are $\binom{M}{3}$ subsets with three elements. So, the number of subsets is $\sum_{n=0}^M \binom{M}{n}$, and this equals 2^M , as I now show. The binomial theorem is $(a + b)^M = \sum_{n=0}^M \binom{M}{n} a^n b^{M-n}$. If $a = b = 1$ this simplifies to $2^M = \sum_{n=0}^M \binom{M}{n}$.

Assume that event space has the following properties:

Axiom 10 $\Omega \in \mathcal{A}$. That is, the event that one of the outcomes occurs is an event. Note that $\Pr[\Omega] = 1$.

Axiom 11 if $A \in \mathcal{A}$ then $\bar{A} \in \mathcal{A}$ where \bar{A} is the compliment of A . That is, if A is an event, then not A is an event

Axiom 12 if A_1 and $A_2 \in \mathcal{A}$ then $A_1 \cup A_2 \in \mathcal{A}$. That is, either event happening is an event

Any collection of events that fulfills the above three assumptions/axioms is called a Boolean algebra.⁸

These three axioms/assumptions about event space along with the definitions of events and events space imply the following:

1. $\emptyset \in \mathcal{A}$ This follows from the first two Axioms. Why?
2. if A_1 and $A_2 \in \mathcal{A}$, then $A_1 \cap A_2 \in \mathcal{A}$
3. if $A_1, A_2, \dots, A_n \in \mathcal{A}$ then $\bigcup_{i=1}^n A_i$ and $\bigcap_{i=1}^n A_i \in \mathcal{A}$

Can you convince your fellow students that these three theorems follow logically from the three axioms?

⁸According to MathWorld, The Boolean algebra of a set S is the set of subsets of S that can be obtained from S by means of a finite number of the set operations \cap \cup and \setminus . For more details, see see??

Wikipedia says, or used to say, "a Boolean algebra is an algebraic structure (a collection of elements and operations on them obeying defining axioms) that captures essential properties of both set operations and logic operations. Specifically, it deals with the set operations of intersection, union, complement; and the logic operations of AND, OR, NOT." See see??

For example if A_1 and $A_2 \in \mathcal{A}$, then by the third axiom $A_1 \cup A_2 \in \mathcal{A}$. Then

by the second axiom $\overline{A_1 \cup A_2} \in \mathcal{A}$. But from De Morgans law we know that $\overline{A \cup B} = \overline{A} \cap \overline{B}$. So $\overline{A_1 \cup A_2} = \overline{A_1} \cap \overline{A_2} \in \mathcal{A}$.

Try and answer the following question from the first set of review questions

1. The experiment is that a coin is flipped twice. How many outcomes are in the sample space and what are they? Now define and enumerate the event space.

1.6.2 The probability function, $\Pr[A]$

Paraphrasing MGB, recollect that Ω denotes the sample space and that \mathcal{A} is the set of events, an algebra of events, of interest from some random experiment.

Definition 13 *The axiomatic definition of probability: Given our definition of events and definition of event space. And given the three axioms imposed on event space, a probability function $\Pr[\cdot]$ will exist that maps events in \mathcal{A} onto the $0, 1$ interval⁹ That is, there exists some function that identifies the probability associated with any event in \mathcal{A} , $\Pr[A]$.*

Assume it has the following three properties:

Axiom 14 $\Pr[A] \geq 0 \forall A \in \mathcal{A}$

Axiom 15 $\Pr[\Omega] = 1 \quad \Omega \in \mathcal{A}$

Axiom 16 *If A_1, A_2, \dots, A_n is a sequence of mutually exclusive events, ($A_i \cap A_j = \emptyset \forall i, j \ i \neq j$) and if $\cup_{i=1}^n A_i \in \mathcal{A}$ then $\Pr[\cup_{i=1}^n A_i] = \sum_{i=1}^n \Pr[A_i]$.*

From this definition of a probability function (with the three axioms) and the earlier definitions and axioms, it is possible to deduce a bunch of additional properties that a probability function must have (MGB). Including

- $\Pr[\emptyset] = 0$
- $\Pr[\bar{A}] = 1 - \Pr[A]$
- If A and $B \in \mathcal{A}$ and $A \subseteq B$ then $\Pr[A] \leq \Pr[B]$
- If $A_1, A_2, \dots, A_n \in \mathcal{A}$ then $\Pr[A_1 \cup A_2 \cup \dots \cup A_n] \leq \Pr[A_1] + \Pr[A_2] + \dots + \Pr[A_n]$. This is called Boole's inequality.
- If A and $B \in \mathcal{A}$ then $\Pr[A] = \Pr[A \cap B] + \Pr[A \cap \bar{B}]$.¹⁰ That is, $\Pr[A] = \Pr[AB] + \Pr[A\bar{B}]$
- If A and $B \in \mathcal{A}$ then $\Pr[A \cap \bar{B}] = \Pr[A] - \Pr[A \cap B]$. That is, $\Pr[A - B] = \Pr[A] - \Pr[AB]$

Put simply, the axiomatic probability theory builds up the notion of probability from a number of assumptions/axioms.

The axiomatic approach subsumes and incorporates the traditions of the the Classical and Frequency approaches.

⁹Note that \mathcal{A} consists of events not numbers. That is, the domain of the function is events (collections of sets), so, formally speaking, $\Pr[A]$ is a *set function*.

You are likely more familiar with function where the domain of the function is not a collection of sets, but rather all or part of the real line, or all or part of Euclidian N space.

¹⁰Note that $\Pr[A \cap B] \equiv \Pr[AB]$ and $\Pr[A \cap \bar{B}] \equiv \Pr[A - B]$

Now that we have defined the probability function $\Pr[A]$ we can return to the earlier assertion that all events are subsets of the sample space, but that for sample spaces of infinite size, some subsets of the sample space are not events.

MBG, page 23, explain this as follows.

In our definition of event and \mathcal{A} , a collection of events, we stated that \mathcal{A} cannot always be taken to be the collection of all subsets of Ω . The reason for this is that for "sufficiently large" Ω the collection of all subsets of Ω is so large that it is impossible to define a probability function consistent with the above axioms (referring to the three axioms above).

I am not sure I completely understand, but here is a shot at further explanation (maybe your other stats books have a better explanation).

MGB are defining events as being something different from a subset of the sample space: being a subset of the sample is necc. but not sufficient.

This raises the question of what are the other required conditions? Reading between the lines, MGB are saying that an event is something that has a probability associated with it, and that probability must be consistent with: $\Pr[A] \geq 0 \forall A \in \mathcal{A}$, $\Pr[A] \leq 1 \forall A \in \mathcal{A}$, and If A_1, A_2, \dots, A_n is a sequence of mutually exclusive events, ($A_i \cap A_j = \phi \forall i, j \ i \neq j$) and if $\cup_{i=1}^n A_i \in \mathcal{A}$ then $\Pr[\cup_{i=1}^n A_i] = \sum_{i=1}^n \Pr[A_i]$.

So, if I read MGB correctly, they are saying if there is a large enough number of subsets of the sample space, there is not a probability function that maps all of those subsets into probabilities and fulfills the above axioms.

As an aside this is not typically a problem when one works with infinite sample space because, typically, one is only interested in a unlarge number of events. Consider, for example, a sample of one observation when the the sample space is the real line from zero to one, including the end points. This sample space consists of an infinite number of samples/points. There is an infinite number of events. That said, we can, for example, consider only two events: $< .5$ and $\geq .5$, and easily define the probability function for both of these events, in a way that is consistent with the above axioms.

Thinking about his raises the following question in my mind: if the sample space is $\Omega = \{\omega : 0 \leq \omega \leq 1\}$, each number in this range is a possible sample (a point in the sample space), and also an event. Some of these samples can only be identified with an infinite number of digits, e.g. .5721.....666.....175.....763.... One can explicitly define events that are ranges of numbers, but I am not sure about some of the individual numbers. There is also the question in my mind of the probability of drawing a specific number. Is it zero= $\frac{1}{\infty}$?

Returning to Feller's book, page 14

(ital. added by Feller)The sample space provides the model of an ideal experiment in the sense that that, by definition, every thinkable outcome of the experiment is completely describe by one, and one one, sample point. It is meaningful to talk about an event A only when it is clear for every outcome of the experiment whether the event A has or has not occured. The collection of all those sample points representing outcomes where A has occured completely describes the event. Conversely, any given aggregate A containing one or more sample points can be called an event: this event does, or does not, occur according as the outcome of the experiment is, or is not, represented by a point of the aggregate A . We therefore define the word event to mean the same as an aggregate of sample points. We shall say that an event A consists of (or contains) certain points, namely those representing outcomes of the ideal experiment in which A occurs.

I take this to mean that Feller is defining an event and a subset of the sample space as the same thing, contradicting MGB.

However, on page 18 Feller says, "In this volume we shall consider only discrete sample spaces." in which case, I think, MGB and Feller would agree that events and subsets of the sample space are one and the same.¹¹

1.6.3 Note that at this point, we know what an axiomatic probability function is, but don't know, in general, how to determine the probability of an event.

But, we know how to calculate the probabilities if the number of outcomes are finite and equally likely (the classical world of probability). In this case $\Pr[A] = \frac{N(A)}{N(\Omega)}$ where $N(\Omega)$ is the number of elements in Ω (number of possible outcomes), and $N(A)$ is the number of elements in A .

If the number of outcomes are finite but all outcomes are not equally likely, one can determine the probability of an event if one knows the probability associated with each element in Ω , $\Pr[\omega_j]$: $\Pr[A] = \sum_{j \in A} \Pr[\omega_j]$.¹²

Consider a case where the sample space has an infinite number of elements, not necessarily all equally likely. In this case, one can estimate $\Pr[A]$ using the frequency definition of probability. Probabilities estimated using the frequency approach fulfill all the required properties (axioms) of the axiomatic definition of $\Pr[A]$, so an axiomatic approach to probability.

Consider the case, where the number of outcomes is finite, they are not equally likely and one does not know the $\Pr[\omega_j]$. In this case one can also estimate $\Pr[A]$ using the frequency definition of probability.¹³

Much of what we do in econometrics to estimate probabilities is in the spirit of the frequency approach.

¹¹Feller defines a discrete sample space as follows: "A sample space is called discrete if it contains only finitely many points or infinitely many points which can be arranged into a simple sequence..." MGB would say such a sample space is countable.

¹²Not that equally likely is the special case where $\Pr[\omega_j] = 1/N(\Omega) \forall j$

¹³Unless one has a large sample space, one will probably want to sample with replacement

1.7 But don't I have to learn about how many possible samples there are if the population is of size M , where M is a finite number, and the sample size is N , $N \leq M$?

It would seem so. Books on probability or statistics cover, or assume, this topic.

Determining probabilities is often all about how many different things can happen and how many of them will have some property.

The answer to how many things can happen is simple if the population size is infinite. In that case, there are an infinite number of samples and sometimes an infinite number of ways an event can occur.

When M is finite and $N \leq M$, the answer depends on whether one samples with or without replacement and depends on whether one is talking about *ordered* or *un-ordered* samples

For the moment, we will limit ourselves to considering sampling without replacement (what economists typically do)

Consider two possible "samples of 3 observations: (2, 55, 17), and (55, 2, 17) where each observation in the sample is identified by a number, so in the first sample the first observation is population member #2.

These are two different outcomes

Are these two different samples or are they the same sample? The answer depends on whether the order is important. If ordering is important, there are two distinct samples. If ordering is not important, they are the same sample.

Consider poker, 5-card hands: Each possible hand is a different outcome. But in terms of the game, order is not important; for example, a hand with 3 kings and 2 jacks has the same meaning/value, independent of the order in which you were dealt the cards. So, 3 kings and 2 jacks is an event, an event generated by a number of different outcomes. How many different outcomes?

1.7.1 With the above in mind, how many outcomes (ordered samples) are there if one is sampling without replacement, the population is of size M and the sample size is N ?

$$\begin{aligned}
 & M(M-1)(M-2)\dots(M-N+1) \\
 = & \frac{M(M-1)(M-2)\dots(M-N+1)(M-N)(M-N-1)(M-N-2)\dots 1}{(M-N)(M-N-1)(M-N-2)\dots 1} \\
 = & \frac{M!}{(M-N)!} \equiv (M)_N
 \end{aligned}$$

where $X! = X(X-1)(X-2)(X-3)\dots 1$

Why is it $(M)_N$? On the first draw there are M possibilities, on the next draw $M-1$, the next $M-2$, and on until N observations are drawn.

For example if the population of interest, Ph.D. students in economics, is 56 and one randomly samples 10 individuals from this population, there are $\frac{(56)!}{(56-10)!} = 1.2921 \times 10^{17}$ possible samples, if the order in which the individuals were drawn matters.

1.7.2 How many ways are there if order is not important?

$N!$ is the number of different ways one can order the same N observations (why?) so, the answer is

$$\frac{M!}{(M-N)!N!} \equiv \frac{(M)_N}{N!} \equiv \binom{M}{N}$$

For our students, the answer is $\frac{(56)!}{(56-10)!10!} = 3.5607 \times 10^{10}$, many fewer than when order is important. but still a big number.

$\binom{M}{N}$ is called the *binomial coefficient*.¹⁴ It is also known as a *combination* or *combinatorial number*. An alternative notation is ${}_M C_N$ and can be read as " M choose N ."

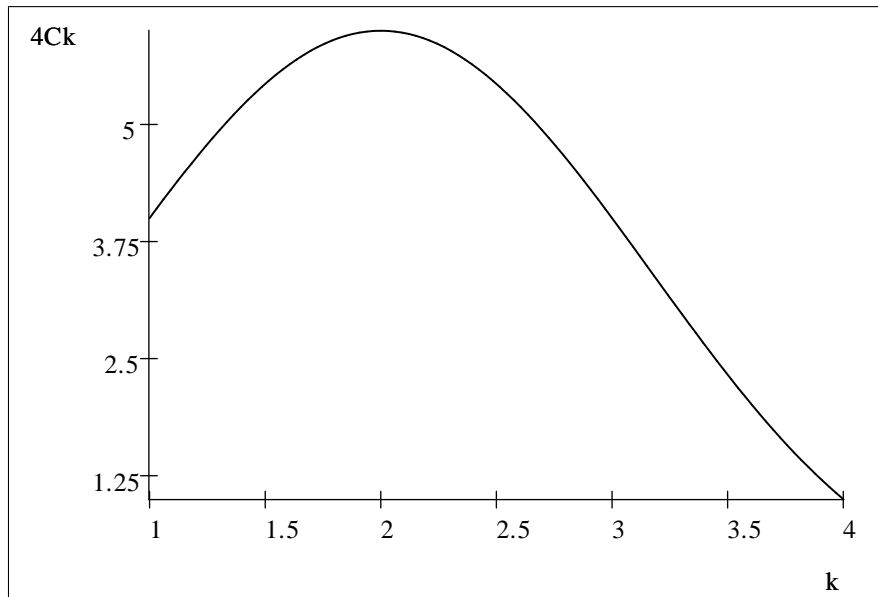
In terms of sets, where order does not matter, $\binom{M}{N}$ is the number of subsets of M that have N elements, so will be useful for counting events

¹⁴For a cool graph of the binomial coefficient, see <http://mathworld.wolfram.com/BinomialCoefficient.html>

You should be able to derive the above formulae, and explain what they mean

A graph the binomial assuming a population size of 4, $1 \leq k \leq 4$.

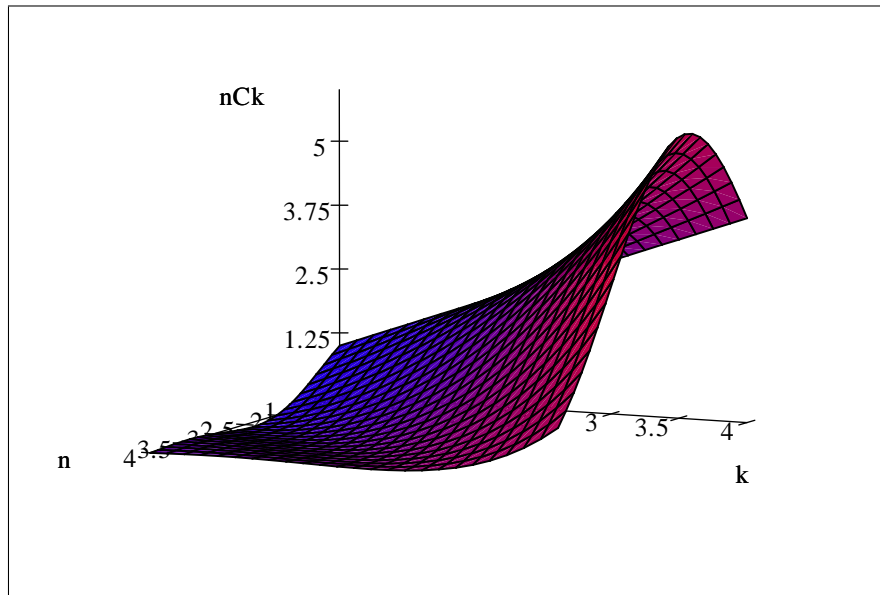
$$\binom{4}{k} \equiv {}_4C_k$$



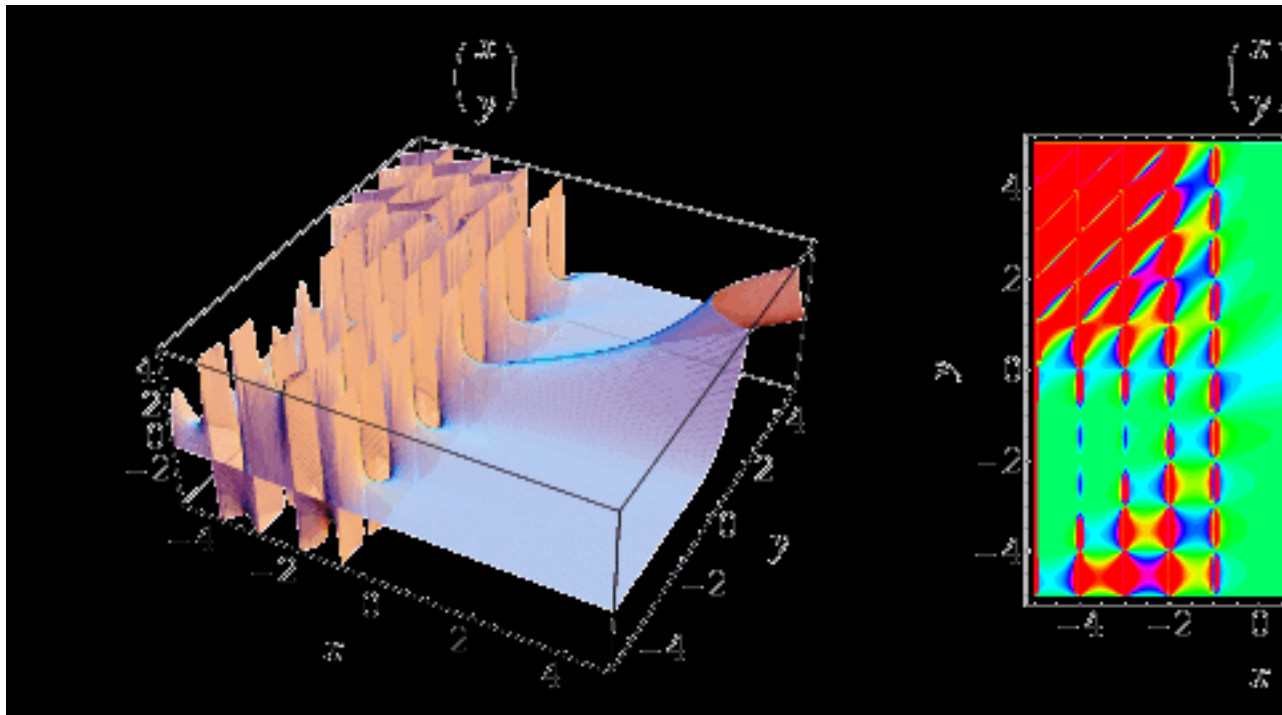
Looks correct: $\binom{4}{1} = 4.0$; there are 4 subsets of size 1; $\binom{4}{2} = 6.0$; there are 6 subset of size 2, $\binom{4}{3} = 4.0$, and $\binom{4}{4} = 1.0$

Note that the function $\binom{n}{k}$ is defined for non integer values of k .

Now I will try to graph $\binom{n}{k}$, $1 \leq n, k \leq 4$



Plot of $\binom{n}{k}$, $1 \leq n, k \leq 4$



Note that the function $\binom{n}{k}$ is defined for non integer values of n and k .

It is much wilder looking if you include negative values for n and k

1.7.3 How often will you use these formulae (2^M , $(M)_N$ and $\binom{M}{N}$)?

A lot if you are a devotee of problems in classical probability (Ω is finite and each outcome is equally likely). Poker players should be devotees. These formulae give us shorthand ways to count, and classical probabilities (the probability of a full house in poker) are determined by counting

When M is small one can often intuit the number of times an event will occur without explicitly using the appropriate formula.

We will sometimes use them. For example, $(M)_N$ and $\binom{M}{N}$ appear in some discrete distribution functions, for example, the binomial distribution contains the term $\binom{M}{N}$.

Binomial distribution: a digression Continuing with this digression, a discrete random variable is said to have a binomial distribution if

$$\Pr[x : n] = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x = 0, 1, 2, \dots, n \\ 0 & \text{if otherwise} \end{cases}$$

where p is the probability of a "yes" so $\Pr[x : n]$ is the probability of observing x yes's out of n trials.

In explanation, $p^x (1-p)^{n-x}$ is the probability of **one way** of observing x yeses in n trials, but there are many ways of getting x yeses in n trials, specifically there are $\binom{n}{x}$ ways.

For a cool interactive graph of the binomial distribution, go to <http://demonstrations.wolfram.com/Binomial>

How might an econometrician use such a distribution? Let's say you want to model the probability of choosing beverage A over B as a function of the price of each, the amount of sugar in each, and whether the beverage is colored brown or not. One might assume that the probability of a yes to beverage A is¹⁵

$$p_A = \frac{e^{V_A}}{e^{V_A} + e^{V_B}}$$

where

$$V_j = \beta_s S_j + \beta_c C_j + \beta_p(\text{price}_j)$$

where S_j is the sugar content of beverage j , price_j is its price and $C_j = 1$ if its color is brown, and zero otherwise. Substituting these functions for the V_j , one has p_A as a function of the prices and characteristics of the two beverages, $p_A = f(\text{price}_A, \text{price}_B, S_A, S_B, C_A, C_B)$. Specifically,

$$p_A = \frac{f(\text{price}_A, \text{price}_B, S_A, S_B, C_A, C_B) e^{(\beta_s S_A + \beta_c C_A + \beta_p(\text{price}_A))}}{e^{(\beta_s S_A + \beta_c C_A + \beta_p(\text{price}_A))} + e^{(\beta_s S_B + \beta_c C_B + \beta_p(\text{price}_B))}}$$

Plugging $f(\text{price}_A, \text{price}_B, S_A, S_B, C_A, C_B)$ into the binomial distribution, one gets $\Pr[n_A : n]$, the probability that n_A of the n beverages you drink will be beverage A , as a function of the prices and characteristics of the two beverages.

$$\Pr[x : n] = \begin{cases} \binom{n}{x} p_A^x (1 - p_A)^{n-x} & \text{if } x = 0, 1, 2, \dots, n \\ 0 & \text{if otherwise} \end{cases}$$

And

$$\Pr[x : n] = \begin{cases} \binom{n}{x} (f(\text{price}_A, \text{price}_B, S_A, S_B, C_A, C_B))^x (1 - (f(\text{price}_A, \text{price}_B, S_A, S_B, C_A, C_B)))^{n-x} & \text{if } x = 0, 1, 2, \dots, n \\ 0 & \text{if otherwise} \end{cases}$$

One could then collect data from a bunch of pair-wise comparisons allowing the prices and other characteristics to vary across the pairs. One then could use the data to come up with estimates of β_s , β_c and β_p . Maximum likelihood estimation would be a good way to do this.

¹⁵I chose this functional form for the probability simply because it restricts p_A to be between zero and one, inclusive, and $p_A + p_B = 1$