Political Science 100a/200a

Fall 2001

**Probability, part I**[1]

The following is a list of the major concepts that will be introduced in these lectures:

- probability numbers, probability measures, probability distributions, samples spaces, interpretation of probability
- probability axioms and deductions from these
- counting
- independent events, marginal distributions
- conditional probability, Bayes' rule
- random variables, discrete v. continuous distributions
- cumulative distributions and density functions
- expectations, mean, variance, moments, and hazard rates
- some specific distributions

# 1 Introduction

- Recall problem of sampling variability: If some of the causes of the things we want to explain in social science are random, or "random for all we could know," then how do know if an observed association between our DV and IVs is due to the IVs, in accord with our theories, or is actually due to chance?
- Answer was to ask what is the probability that we would observe this degree of association if in fact there were no systematic association between IV and DV? Recall coin flip example: To answer "Is this coin biased?" we ask about the probability that we would observe the data ($x$ heads) if the coin were fair. Or, to answer "Is joint democracy systematically related to interstate peace?" we ask about the probability that we would observe this distribution of wars across dyads if there were no relationship.
- To do this we need some probability theory.

[1]Notes by James D. Fearon, Dept. of Political Science, Stanford University, October 2001.

## 2 Definitions and interpretation of "probability"

Will start with a technical, mathematical definition or treatment of probability first, and then proceed to how to interpret it.

**Def$^{\underline{n}}$:** The *sample space* is a set $S$ composed of all the possible outcomes of an "experiment," or all the things that might hypothetically occur. (Think of the "experiment" as the social, political, or economic *process* about which we are trying to draw inferences; e.g., the process that generates interstate wars, or life expectancy in a country.)

- e.g.: If we flip a coin twice the sample space might be taken as $S = \{hh, ht, th, tt\}$.
- e.g.: The sample space for the outcomes of the 2000 US presidential election might be taken as $S = \{Bush, Gore, Nadar, Buchanan\}$.
- e.g: The sample space might be taken as a set of numbers that might be killed due to civil war in a country in a given year, thus $S = \{0, 1, 2, 3, \dots\}$.
- For the most part, we will be assuming that $S$ is *finite* or at least countably infinite (i.e., not uncountably infinite like the set of real numbers). Sometimes, though we may pose the sample space is uncountably infinite, such as representing the sample space of life expectancies $S = [0, 150]$ (which means *all* the numbers between 0 and 150.)

**Def$^{\underline{n}}$:** An *event* is a subset of the sample space, or, $A \subset S$ is an "event."

- e.g.: $A = \{hh\}$ would be the event that heads occur twice when a coin is flipped twice ($S = \{hh, ht, th, tt\}$). $A = \{hh, tt\}$ would be the event that either two heads or two tails result.
- e.g.: $A = \{Gore, Bush\} \subset S = \{Gore, Bush, Nader, Buchanan\}$ would be the event that a major party candidate wins the election.
- e.g.: $A = \{72.3, 80\}$ could be the event that the population of a country has life expectancy of 72.3 or 80 years, and $A = [72, 80]$ could be understood as the event that a country's life expectancy is between 72 and 80, inclusive.

**Def$^{\underline{n}}$:** Let $2^S$ represent the *power set* of $S$, which is the set of all subsets of $S$. (Note that this is a well-defined finite set in the case that $S$ is finite.)

- e.g.: If $S = \{h, t\}$, the the power set $2^S$ is $\{\emptyset, \{h\}, \{t\}, \{h, t\}\}$.

**Def$^{n}$:** A *probability measure* is function $P(\cdot)$ that assigns every element of $2^S$ (that is, every subset or "event" in $S$) to a number in the [0, 1] interval, and which satisfies the three axioms given below.

**Def$^{n}$:** A *probability number* is a particular number between 0 and 1 that is assigned to a particular subset ("event") in the sample space $S$. (Probability numbers should be distinguished from the idea of a probability measure.)

The three axioms which, if satisfied, constitute a probability measure are:

1. For every event $A \subset S$, the value of the function is a non-negative real number, i.e., $P(A) \geq 0$. (Read $P(A)$ as "the probability of event $A$.") (This says that there is no such thing as negative probability.)
2. For any two disjoint sets $A, B \subset S$, (i.e., $A \cap B = \emptyset$), the probability number assigned to their union $A \cup B$ is equal to the sum of the probability numbers for $A$ and $B$. Thus, when $A \cap B = \emptyset$,
$$P(A \cup B) = P(A) + P(B).$$
3. The value of the function for $S$ equals 1, i.e., $P(S) = 1$.

- For example, given $S = \{h, t\}$, $P(\emptyset) = 0$, $P(\{h\}) = P(\{t\}) = 1/2$, $P(\{h, t\}) = 1$ is a probability measure. $P(\{t\} = 1/2$ is a probability number.
- So is $P(\emptyset) = 0$, $P(\{h\} = .31$, $P(\{t\}) = .69$, $P(\{h, t\}) = 1$ a probability measure.
- Note that so far this is a purely mathematical set up and definition without any developed interpretation connecting it to real world things. In this sense, a probability measure is just a rule that assigns numbers to sets, or a particular sort of way of "measuring" sets. But this definition is in fact motivated by thinking about a specific sort of real world problems, and originally problems connected with gambling.

**Def$^{n}$:** Suppose $S$ is finite and has elements $\{x_1, x_2, ..., x_n\}$. Let $p_i = P(\{x_i\})$. We will refer to the list $(p_1, p_2, ..., p_n)$ as a *probability distribution.*

- Since $P(\cdot)$ is a probability measure, a probability distribution must satisfy $p_i \geq 0$ for all $i$ and $\sum_{i=1}^{n} p_i = 1$. (how does disjointness come in here?)
- Note that a probability measure implies a probability distribution, and a probability distribution implies a probability measure.
  - e.g.: If $A = \{x_1, x_6, x_8\}$, then if we have the probability distribution we know that $P(A) = p_1 + p_6 + p_8$ using axiom 2, since $A$ is the union of the disjoint sets $\{x_1\}$,

$\{x_6\}$, $\{x_8\}$. The probability distribution is just the measure 'broken down' on the constituent elements of $S$.

- Important: Note that a relative frequency histogram of a variable can be interpreted as a probability distribution. Illustrate with **graph lifeexp ,bin(12)**. What is $S$ here?
- Illustrate again with **tab lifeexp**. Any variable $X = (x_1, x_2, x_3, \ldots, x_n\}$ can be thought of as a probability distribution on the "sample space" $X$ with $P(\{x_i\}) = 1/n$.
- In which case, the probability of (drawing) a life expectancy value between, say, 72 and 80 is just $\sum_{\mathbf{lifeexp} \in [\mathbf{72},\mathbf{80}]} 1/n$, or the number of observations with **lifeexp** greater than 72 and less than 80, divided by the total number of observations.

- There is an important intuition/rule built into the definition of a probability measure, often known as the "additive law of probability." First a definition.

  **Def$^{\underline{n}}$:** Two events are *mutually exclusive* if they cannot both happen – the occurrence of one precludes the occurrence of the other. In formal terms this means that two events $A$ and $B$ are mutually exclusive if $A \cap B = \emptyset$.

- Next a claim:

  *If events $A$ and $B$ are mutually exclusive then the probability that either $A$ or $B$ happens is equal to the sum of the probabilities of events $A$ and $B$.* Formally, if $A \cap B = \emptyset$, then (by axiom 2) $P(A \cup B) = P(A) + P(B)$.

- Examples: dice, life expectancy.

**Interpreting "probability"**

- How to interpret a probability measure and probability numbers? Take a simple, motivating case. Suppose we have a typical coin, and we flip it. It is natural to set the sample space here as $S = \{h, t\}$, where the elements refer to heads and tails respectively. Consider the event $A = \{h\}$. Intuitively, we interpret the probability of $A$, $P(A)$, a something like the *likelihood* that heads will be the outcome of the "experiment."
- This is one of *three* intuitions about how to think about what a probability number is, which form the grounds of three competing schools of philosophical thought on the question

  1. *Frequentist Interpretation*: The probability of an event in the sample space is the *relative frequency* that the event would occur if the experiment or trial were run many, many times, over and over again. Thus, $P(A)$ in the example above is understood in this approach as the proportion of heads that would appear if we flipped the coin over and over again for a very long time.

- Problems: (a) What is a "very long time"? (b) The circumstances of the coin tossing must be presumed to vary randomly from trial to trial, but then we may be smuggling in a notion of what we are trying to define. (c) And no matter how many times you flipped the coin, the exact proportion would always vary a little bit, even if the coin was perfectly fair, and there is no way to talk about likely variation around one half without again introducing the idea we are trying to define. (d) Finally, there is the big problem that it is hard or impossible to think about a lot of events in the world as being repeated many time under random circumstances. Frequentists have a hard time with questions like "What is the probability that Colin Powell will run for president in the year 2004?"

2. *Classical Interpretation*: The probability of an event is the ratio of the number of cases favorable to that event to the total number of cases, provided that all these are *equally likely*. This definition comes from thinking about things like the flip of a coin or the roll of a die. In the die case, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$, and given a presumption of equal likelihood for each event $\{k\}$, $k \in \{1, 2, 3, 4, 5, 6\}$, we have a probability measure in the sense given above. E.g., the probability of any particular face turning up is $1/6$, while the event $\{2, 5\}$, has probability $2/6 = 1/3$ under this definition.
   - Problem: this 'definition' uses the notion of probability ("equally likely") in the definition of probability.
   - But keep this approach in mind because the idea of "equally likely" events drives (in a sort of axiomatic sense) a lot of basic thinking about probability, and is very helpful in working through problems.
3. *Subjective Interpretation*: A probability number is a subjective (or personal) estimate of the likelihood that a particular outcome will obtain.
   - Problem: Basically gives up and just used the intuitive notion of "likelihood" as a theoretical primitive, thinks of probability as a sort of constructive formalization of this intuition. Rejects the idea of trying to define probability as an objective feature of the world, in some way.

Any of these ways will mainly do fine for you, at least here; 2 is perhaps especially intuitive (cf. Friedman, who works from the classical definition).

## 3 Deductions from the axioms

Some theorems:

**Th$^{\underline{\mathbf{m}}}$ 1:** For all $A \subset S$, $P(A) + P(A^c) = 1$.

In words: the probability that any event $A$ either occurs or doesn't occur is one.

Proof: $A$ and $A^c$ are disjoint, and $A \cup A^c = S$, so from Axioms 2 and 3 we have $P(A)+P(A^c) = P(A \cup A^c) = P(S) = 1$. QED.

**Th$\underline{\mathbf{m}}$ 2:** For any set $A \subset S$, $P(A) \leq 1$.

In words: Nothing can be more than certain (have probability 1), and all probability numbers are bounded between 0 and 1.

Proof: Theorem 1 implies that for any $A \subset S$, $P(A) = 1 - P(A^c)$, and by axiom 1 $P(A^c) \geq 0$, so

$$P(A) = 1 - P(A^c) \leq 1.$$

QED.

Question: Which is greater, (a) the probability that U.S. forces will capture/kill OBL, or (b) the probability that U.S. forces will capture/kill OBL as a result of a tip from an Afghan?

**Th$\underline{\mathbf{m}}$ 3:** For any two sets such that $A \subset B$, $P(A) \leq P(B)$ and $P(B - A) = P(B) - P(A)$.

Proof: Try it yourself.

**Th$\underline{\mathbf{m}}$ 4:** ("finite additivity") For any finite number of disjoint sets,

$$P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ...P(A_n).$$

Proof: From induction using axiom 2.

The next one generalizes this a bit.

**Th$\underline{\mathbf{m}}$ 5:** ("Boole's inequality") For any finite number of sets $A_1, ..., A_n$,

$$P(A_1 \cup ... \cup A_n) \leq P(A_1) + ... + P(A_n).$$

Proof: Do for case of two sets $A$ and $B$. Note that $A \cup B$ can be written as the union of two disjoint sets as follows:

$$A \cup B = A \cup (A^c \cap B).$$

Thus

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

But since $A^c \cap B \subset B$, by theorem 3 above $P(A^c \cap B) \leq P(B)$, implying the result. This proves it for the case of two sets; the general case is done by induction. QED.

Intuitively, if there are nonempty intersections among the sets, these only lower the probability of the union of these sets below the sum of their separate probabilities.

The next theorem, which is an important one to understand, says exactly how much it is lowered.

**Th$^{\underline{\mathbf{m}}}$ 6:** For any two events $A$ and $B$,

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

Proof: Show intuitively from Venn diagram.

Intuitively, we subtract the probability of the elements $A \cap B$ once so as not to count them twice, as we would if just had $P(A \cap B) = P(A) + P(B)$.

But the important special case is the following general rule: If two events $A$ and $B$ are mutually exclusive, $P(A \cup B) = P(A) + P(B)$.

- e.g.: A fair coin is flipped twice. What is the probability of getting a "head" on at least one of the two tosses? Are these events mutually exclusive? (Do via sample space ...)
- What is the probability of getting a head on exactly one of the two tosses?
- if we draw a country at random, what is the probability that its population's life expectancy is greater than 72 or less than 45? What is the probability that life expectancy will be greater than 60 but less than 72, or greater than 68 but less than 75?

## 4 Independence

**Def$^{\underline{\mathbf{n}}}$:** Two events $A, B \subset S$ are *independent* when $P(A \cap B) = P(A)P(B)$.

- Intuition: Two events, $A$ and $B$ are *independent* of each other when the occurence of one in no way influences the likelihood of the other occuring. $P(A \cap B)$ is the probability that both $A$ *and* $B$ happen.

  If $A$ happens about $x\%$ of the time, and $B$ happens about $y\%$ of the time, then of the $x\%$ occasions that $A$ occurs, in $y\%$ of these, $B$ will also occur. Thus, $A$ *and* $B$ occur about $y\%$ of $x\%$ of the time, or $\frac{x}{100}\frac{y}{100} = P(A)P(B)$. Note that it is important to this

conclusion that the probability of $B$ does not depend in any way on the occurence or nonoccurence of $A$.

- e.g.: We roll a die twice in succession. What is the probability of observing the outcome (3, 3)?

  First, what is the sample space $S$ here?

$$S = \begin{array}{ccccccc} & & & & & & \text{“}marginals\text{”} \\ (1,1) & (1,2) & (1,3) & ... & (1,6) & & 1/6 \\ (2,1) & (2,2) & (2,3) & ... & (2,6) & & 1/6 \\ (3,1) & (3,2) & (3,3) & ... & (3,6) & & 1/6 \\ . & . & . & . & . & & 1/6 \\ (6,1) & . & . & . & . & & 1/6 \\ & & & & & & \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & & \end{array}$$

  - If we assume that all outcomes in $S$ are equally likely, and thus each $\{x_i\} \subset S$ has probability 1/36.
  - Draw events $A$ and $B$. Note that $P(A) = 1/36+1/36+1/36+1/36+1/36+1/36 = 1/6$, and likewise for $P(B)$.
  - $A \cap B = \{(3,3)\}$, and $P(A \cap B) = 1/36 = (1/6)(1/6)$, so the definition of independence is satisfied here.
  - Intuitively, provided that rolling the die the first time does not affect the distribution on possible outcomes the second time, these events will be independent.

- Important concept: The *marginal probabilities* or *marginals* in the above table are the probabilities associated with a row or a column, that is, with a particular outcome on the first (row) or second (column) roll of the die. Another way of saying that two events are independent is to say that the probability of the event $\{s_i\} \subset S$ is equal to the product of the relevant marginals.

  - illustrate with: **tab ally cowwar if bicontig == 1 ,row col** ...

- Examples:

  1. Four cards are drawn at random from a pack of cards. What is the probability that all four are Aces?
  2. One student comes to class 60% of time, and another comes 80% of the time. Their decisions are made independently. What is the probability that both will be in class on a given day? At least one of them? Just the better attender?

3. Suppose that each member of an 9-person jury has a .9 chance of voting "Guilty" when the defendant is guilty and a .9 chance of voting "Innocent" when the defendant is innocent. Assume (implausibly) that their votes are independent. What is the probability that they collectively make a mistake if the defendant is guilty? If innocent?
4. Isaac throws six dice and wins if he gets at least one ace (a one). Sam throws twelve dice and wins if he gets at least two aces. Who has the better odds of winning?

## 5 Counting methods and binomial coefficients

Question 1: Suppose you have 3 shirts, 2 sweaters, and 2 pairs of pants. How many "outfits" can you form from these? 3+2+2? $3^{2^2}$?

- These are both wrong. Draw a picture ...
- Counting the possible paths gives the answer: 12.
- Also suggests a rule: Since for each of the three shirts there are two possible choices of sweaters, there are $3 \times 2 = 6$ possible combinations of these. For each of these 6, there are two pants possibilities, so there are $3 \times 2 \times 2 = 12$ total possibilities.

**Fact:** ("the fundamental rule of counting) A number of multiple choices are to be made. There are $m_1$ possibilities for the first choice, $m_2$ for the second, and so on. If these choices can be combined freely, then the total number of possibilities for the whole set of choices is $m_1 m_2 m_3 \cdots$.

- e.g.: How many ways can six dice appear when rolled in sequence (or roll the same die six times, recording the outcome in each case)?
- e.g.: Ten survey respondents place themselves on a 7-point "liberal to conservative" scale of political preferences. How many preference profiles are there for the group?
- e.g.: You have a series of binary independent variables coding features of dyad-years, including contiguity, alliance status, joint democracy, and the existence of a border dispute. How many "types" of dyad-years do you have?
- e.g.: We repeatedly flip a coin, $n$ times in a row, recording the heads or tails outcome each flip. How many possible sequences are there in $n$ flips? (e.g., a sequence might be denoted 01000111010100110..., where 1 represents heads and 0 tails. How many of these are there?)

Question 2: You have a survey questionaire with $n$ questions on it. How many ways are there to order the $n$ questions?

- There are $n$ ways to choose the first question, but after deciding this one, there are only $n-1$ ways to choose the second, $n-2$ ways to choose the third, and so on.
- Thus, the number is $n(n-1)(n-2)\cdots 2\cdot 1$, which we call $n$ *factorial*, or $n!$ for short.
- The $n!$ different arrangments are called *permutations*.
- The urn model generalizes these arguments:
  1. Imagine drawing $r$ balls from an urn with $n$ numbered balls *with replacement*: Then there are $n$ possibilities on each draw, and thus $n^r$ possible samples.
  2. Now imagine drawing $r$ balls *without replacement*: As above, there are $n$ possibilities for the first draw, $n-1$ for the second, yielding $n(n-1)(n-2)\cdots(n-r+1)$, which is sometimes called a *continued product* and written $(n)_r$.
- e.g.: How many ways are there for this class to sit in this classroom?

Example: What is the probability that at least two people in a room of $n$ people have the same birthday?

- Let $A$ be the event in question. Because there are so many ways that $A$ could happen, a good trick to ask if it is easier to figure out $P(A^c)$, the probability that no two people in the room have the same birthday.
- $A^c$ can happen in $365\cdot 364\cdot 363\cdots(365-n+1)$ ways. (see why?). There are $365^n$ possible "rooms," so the probability of $A^c$ is $(365)_n/365^n$, and the probability of $A$ is $1-(365)_n/365^n$. (show Stata results)

Question 3: You have a sample of $n$ states. How many dyads are there?

- This is *almost* like drawing twice from an urn filled with country names without replacement, in which case the answer would be $n(n-1)$. But (U.S., Britain) is the *same* "dyad" as (Britain, U.S.)!
- We call samples or subsets where the order of the draws matters *ordered samples*, and samples or subsets where the order doesn't matter *unordered samples*.
- Our question is thus, How many unordered samples of size two are there from an urn with $n$ countries? There are two ways to order each sample of size two, so we have

$$\text{total \# samples } (n(n-1)) = \text{\# unordered samples} \times \text{\# ways to order sample, or}$$

$$\text{\# unordered samples } = \frac{\text{total \# samples}}{\text{\# ways to order sample}}$$

- Thus, the number of distinct dyads you can form from a list of $n$ countries is $n(n-1)/2$.
- More generally, the number of unordered samples of size $r \leq n$ you can draw from a population of $n$ objects is $(n)_r/r!$. See why?
- Note that $(n)_r = \frac{n!}{(n-r)!}$. The number of unordered samples of size $r$,

$$\binom{n}{r} = \frac{(n)_r}{r!} = \frac{n!}{(n-r)!r!}$$

  is called a *binomial coefficient*.
- e.g.: If I choose three "committees" at random from a set of 8 students, how many possible committees are there?
- e.g.: If there are 455 members of Congress, and a particular congressional committee can have 25 people on it, how many possible committee memberships are there (hypothetically)?

Question 4: A fair coin is flipped $n$ times. How many sequences are there with exactly $r$ heads?

- It is not immediately obvious, but this is the same as the last question.
- Imagine we have an urn with $n$ numbered balls. We are going to draw out $r$ balls, and assign each one to a "place" between 1 and $n$. (This is like assigning, say, four "heads" to spots in a sequence of 10 flips.) We want to know how many ways we can do this.
- There are $n$ spots for the first draw, $n-1$ for the second, and so on, thus $(n)_r$ ways to assign the $r$ draws in the sequence.
- *But*, note that we get the same result whether, say, the first draw is assigned to the third spot and the fifth to the 8th spot, or vice versa.
- How many different ways are there to order the assignment of $r$ balls to the $r$ chosen spots? $r!$.

- So we conclude that the number of ways to get $r$ heads in $n$ tosses of a coin is (again) the binomial coefficient

$$\binom{n}{r} \equiv \frac{(n)_r}{r!} = \frac{n!}{(n-r)!r!}.$$

- A handy way of computing binomial coefficients for small values of $m$ is to use Pascal's triangle, which looks like this:

| flip # | Pascal's $\triangle$ | # sequences |
|---|---|---|
| 1 | 1 1 | 2 |
| 2 | 1 2 1 | 4 |
| 3 | 1 3 3 1 | 8 |
| 4 | 1 4 6 4 1 | 16 |
| 5 | 1 5 10 10 5 1 | 32 |
| 6 | 1 6 15 20 15 6 1 | 64 |
| . | ... | . |

and so on. Do you see the logic of the triangle?

Question 5: Now we can finally ask, What is the probability that if you flip a fair coin $n$ times, exactly $r$ will be heads?

- There are how many sequences of flips total?
- In how many of these are there exactly $r$ heads?
- Thus the probability is

$$\frac{\binom{n}{r}}{2^n} = \frac{n!}{(n-r)!r!}\frac{1}{2^n}.$$

- use stata to illustrate ten coin tosses ...
  - **set obs 11**, **range r 0 10**, **help functions**, **gen binom = comb(10,r)**/$2^{10}$.
  - Suppose you observe 7 heads, and you were wondering if the coin might be biased. We let our *alternative hypothesis* be

$$H_1 = \text{the coin is biased (i.e., } Pr(head) \neq .5).$$

And we let our *null hypothesis* be

$$H_0 = \text{the coin is fair (i.e., } Pr(head) = .5).$$

Then we ask, What is probability you would get *at least 7 or no more than 3 heads* if $H_0$ were true, i.e., the coin were in fact fair?

– Now suppose you observe 7 heads, but you were wondering whether the coin is biased *favor of heads*. Here we might let our *alternative hypothesis* be

$$H_1 = \text{the coin is biased in favor of heads (i.e., } Pr(head) > .5).$$

And we let our *null hypothesis* be

$$H_0 = \text{the coin is fair (i.e., } Pr(head) = .5).$$

Then we ask, What is the probability that you would get *at least 7 heads* in 10 flips if $H_0$ were in fact true?

– These two problems illustrate so-called *one-tailed versus two-tailed tests*. Discuss ...

Question 6: Suppose that presidential approval ratings are determined by a random process that produces, for a given president, daily ratings that follow a normal distribution fairly well within a short period of time (a few months). If we were to randomly sample 5 approval ratings within such a period, what is the probability that 2 of them would be more than two standard deviation from the mean?

- First ask, What is the probability that a single "draw" is two sds or more from the mean?
- So the "experiment" is exactly like flipping a "coin" five times that has a .05 chance of coming up heads each time. What is the probability of a sequence of draws that has exactly 2 heads with this coin? (work through to answer)
- This example illustrates how we generalize the fair coin case discussed above. *Many* problems can be represented as follows: There is a sequence of $n$ *Bernoulli trials* or "experiments," which can result in either "success" (e.g., a head turns up) or "failure." The probability of success on each trial is the same, $p$, and the trials are independent. What is the probability of $x$ successes in $n$ trials?
- Generalizing the above argument, the answer is

$$B(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

which is called a *binomial distribution with parameters n and p*, because it assigns a probability number to each possible outcome $x = 0, 1, 2, 3, ..., n$.

- e.g.: Suppose that there are 100 other voters, besides you, and each other votes for candidate A with probability $p$. What is the probability that exactly 50 of the others will vote for A, so that your vote would be decisive? (calc with stata, using $p = .5$ and $p = .6$)
- Note: The binomial distribution with parameters $n$ and $p$ gives the probability of getting $x$ "successes" in $n$ *Bernoulli trials* with parameter $p$. A Bernoulli trial follows (what is usually called) *Bernoulli distribution*, which is simply

$$Pr(X) = \begin{cases} p & \text{if } X = 1 \\ 1 - p & \text{if } X = 0 \\ 0 & \text{for all other } X \end{cases}$$

## 6 Conditional probability

Question 7: Working for the U.S. Census Bureau, you knock on the door of a house of a family that you know to have two children. A boy answers the door. What is the probability that the other child is a boy (assuming boys and girls are equally likely in the population, and that probabilities are independent across births within a family)?

Question 8: The next house you will visit also has two children, and you know (only) that the first child is a boy. What is the probability that the second child will also be a boy?

- work out with sample space diagram ...
- This is an example of *conditional probability*. Intuitively, conditional probability is in play when we are asking questions of the form What is the probability of A conditional on (a different) event B having occurred?
- Let $P(A|B)$ read as *the probability of A given B* or *conditional on B* or *provided that B occurs*.
- e.g.: Let $A$ = rain tomorrow and $B$ = clouds today. What is $P(A|B)$? What relation would you expect between $P(A|B)$ and $P(A)$.
- e.g.: Let $y_i = 1$ if dyad-year $i$ has a war, and 0 if not. Let $x_i = 1$ if dyad-year $i$ is jointly democratic, and 0 if not. What is $P(1|1)$? $P(1|0)$? $P(1|0) + P(1|1)$? $P(0|1) + P(1|1)$ in words?
- e.g.: Let $y_i$ be life expectancy in country $i$, measured in (integer) years, and $x_i$ be log of per capita income in country $i$. What is $P(y_i|x_i)$?

- The underlying or meta- issue implicit in conditional probability is: how should we revise our beliefs/hypotheses in light of new data?
  - e.g.: Suppose you have a hypothesis $H$, and a prior belief about whether the hypothesis is true. For example, consider the hypothesis $H$ = states' regime types (e.g., democracy or nondemocracy) do not affect their international behavior. Now suppose that you observe some data, $D$ = democracies don't or very rarely fight wars against each other. It is natural to be interested in the probability of the hypothesis being true conditional on observing this data, $P(H|D)$.
  - In general, if $H$ = hypothesis, $D$ = observed data, then we call $P(H)$ the *prior belief* or *prior probability* that $H$ is true, and $P(H|D)$ the *posterior belief* or *posterior probability* that $H$ is true. (I.e., posterior to observing the data $D$.)

**Def$^{\underline{n}}$:** For sets $A, B \subset S$, the conditional probability $P(A|B)$ is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

- Illustrate with boy-girl example ...
- Intuitively, $P(A|B)$ is just the probability of event $A$ *when event $B$ is treated as the new sample space.* Illustrate with Venn diagrams ...
- Dividing out $P(B)$ renormalizes so that the the sum of the probabilities of mutually exclusive events within $B$ equals one. Venn diagrams ..
- e.g.: Suppose that the sample space $S$ refers to three possible mechanical conditions of a used car you might buy. Let $S = \{s_1, s_2, s_3\}$, where $s_1$ = "car is bad", $s_2$ = "car is ok" and $s_3$ = "car is good." Further, suppose that you initially judge these three outcomes as equally likely (this is your prior probability distribution on the space $S$). Let $A = \{s_3\}$ be the event that the car is good, and $B = \{s_2, s_3\}$ be the event that a mechanic tells you that the car is "not bad." What is $P(A|B)$?
- Recall that events $A$ and $B$ are independent if $P(A \cap B) = P(A)P(B)$. Suppose this is the case for two events $A$ and $B$, and let's ask about $P(A|B)$. By the formula this is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

- So for independent events, the conditional probability of one given the other is just the prior probability of the event. For independent events, the occurence of one has no bearing on the occurrence of the other.

- We can rewrite the above expression, solving for $P(A \cap B)$

$$P(A \cap B) = P(B)P(A|B).$$

This is sometimes called *the multiplicative law of probability.* Note that it generates $P(A/capB) = P(A)P(B)$ for independent events.

## 7 Bayes' Rule

- There is another way to write the expression for conditional probability that is very useful in a broad range of applications, called Bayes' rule or theorem.
- First, since $P(A|B) = \frac{P(A \cap B)}{P(B)}$ and $P(B|A) = \frac{P(A \cap B)}{P(A)}$, then we can solve the second of these for $P(A \cap B)$ as follows: $P(A \cap B) = P(B|A)P(A)$. Then we can substitute this into the expression for $P(A|B)$ to yield

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

- Next, we need the following proposition:
  **Th$\underline{\mathbf{m}}$ :** Suppose $S = \sum_{i=1}^{n} A_i$ and the $A_i$ are disjoint sets (that is, the sets $A_i$ form a *partition* of the state space $S$). Then for any set $B \subset S$,

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

"Prove" by means of Venn diagram ... , considering $A$, $A^c$.

- Using this, and the fact noted earlier, we can rewrite the definition of conditional probability as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

since the numerator is the same as $P(A \cap B)$ and the denominator is just a way of writing $P(B)$ in terms of probabilities conditioned on $A$ and $A^c$.

- More generally, suppose we are interested in the probability of a variety of possible events $A_1, A_2, A_3, A_4, \ldots, A_m$, where $Pr(A_1 \cup A_2 \cup \ldots \cup A_m) = 1$, *conditional on* another event $B$ occuring. Then Bayes' rule says that

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{m} P(B|A_j)P(A_j)}$$

- This expression for conditional probability is called *Bayes' rule* or *Bayes' theorem*, after the Reverend Thomas Bayes who used it in a short paper published in 1763.
- It is especially convenient for problems of *inference. Bayes' rule provides a normative standard for how to update one's beliefs based on new data or information.*
- e.g.: Let $H$ = hypothesis, $P(H)$ = our prior belief that the hypothesis is true, $D$ = the (new) data, and $P(H|D)$ = our updated or posterior belief that the hypothesis is true. Then from Bayes' rule we have:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D|H)P(H) + P(D| \sim H)P(\sim H)}$$

  where $\sim H$ means "the hypothesis is not true." Note that $P(\sim H) = 1 - P(H)$.
- Thus, if you have a prior belief about the hypothesis, and beliefs about the relative likelihood that you would observe the data if the hypothesis were true or false ($P(D|H)$ and $P(D| \sim H)$, then Bayes' rule says how optimally to form your posterior belief.
- e.g.: Let $H$ be the proposition that the Soviet Union was fundamentally aggressive and expansionist, and let $D$ be the observation that Gorbachev "let Eastern Europe" go. Suppose that you had $P(H) = .7$ (you were a bit of a hawk), $P(D|H) = .1$ (thus think it not very likely to observe this if hypothesis were true, although it might occur, say, because they were forced by circumstance), and $P(D| \sim H) = .8$ (thus you think it more likely that you would observe the data if the S.U. were not fundamentally aggressive). Cranking all this through Bayes' rule gives

$$P(H|D) = \frac{.1(.7)}{.1(.7) + .8(.3)} = \frac{.07}{.31} = .23$$

  Notice that this is significantly less than the prior belief of $P(H) = .7$, so observing this data would lead a rational observer to decrease his or her belief about the Soviet Union's disposition.
- There is a somewhat nicer way of expressing Bayes' rule in terms of *odds ratios.* If the probability of an event occuring is $p$, then the *odds ratio* for this event is $\frac{p}{1-p}$. For instance, if you think the probability that space aliens run the government is .7, then you think the odds are $\frac{.7}{.3} = \frac{7}{3}$, or, colloquially, "7 to 3."
- In words, then Bayes' rule may be expressed as follows: *The posterior odds equal the prior odds times the likelihood ratio.* The ratio $\frac{P(D|H)}{P(D|\sim H)}$ is called the likelihood ratio, because it gives the relative likelihood that you would observe this data in the events the hypothesis is true or false.

- Notice, then, that by Bayes' rule the posterior belief that $H$ is true should *increase* (decrease) if and only if it is judged more (less) likely that the data would be observed if it were true than that the data would be observed if the hypothesis were false.
- If you think you would be just as likely to observe the data whether the hypothesis was true or false (i.e., $P(D|H) = P(D| \sim H)$ so that the likelihood ratio is 1), then rationally your posteriors will equal your priors.
- For instance, suppose you ask the used car salesman if this car has major engine problems, and he says "No." Then, if you think he would be equally likely to say this whether or not the car had major engine troubles, then rationally this response should not lead you to change your prior beliefs about the car's engine.
- Or, in the Soviet Union example given above, we have this, using odds ratio form:

$$\text{Posterior odds SU is aggresive } = \frac{.1}{.8}\frac{.7}{.3} = .29.$$

  Note that since it is eight times more likely that you would observe the data (Gorbachev "letting Eastern Europe go") if the hypothesis that the S.U. was fundamentally aggressive were false, the posterior odds will, by Bayes' rule, fall by a factor of eight compared to the prior odds.
- e.g.: The Monte Hall Problem ...
- e.g.: The Bayesian approach to estimating parameters.

  bi
- Suppose we want to use social science data to estimate

  1. the proportion of voters who support the death penalty;
  2. the difference between African and Asian countries annual life expectancies;
  3. the difference in probability of war between jointly democratic and other interstate dyads;
  4. the probability that a particular coin lands heads when tossed.
- Let $\theta$ be such a parameter, and suppose that $\theta$ might take any of a (possibly large number) of discrete values.
- Suppose further that we observe some data in the form of a set of observations $y = (y_1, y_2, y_3, \ldots, y_n)$. e.g. ...
- From the Bayesian perspective, we have (or stipulate) a prior distribution $p(\theta)$ that gives probability numbers to different possible values of $\theta$.

- We want to use the observed data $y$ to update this prior belief. Using Bayes' Rule,

$$p(\theta_i|y) = \frac{p(y|\theta_i)p(\theta_i)}{\sum_{\forall j} p(y|\theta_j)p(\theta_j)}$$

- Note especially $p(y|\theta_i)$, which is the probability that you would observe this specific set of data if $\theta$ was in fact $\theta_i$. This probability is called *the likelihood* and it plays a very big role in more advanced data analysis methods.
- eg: This is all more comprehensible with an example. Suppose you find have a weird coin and from looking at it have absolutely no idea what is the probability that it would come up heads. You have no clue.
- Suppose you then flip the coin 10 times and observe 7 heads. What should you know believe about the probability that the coin turns heads when flipped? Bayes' Rule gives us a way to approach this question.
  - So let's let your prior belief be represented by the (uniform) distribution on the values $X = (0, .1, .2, .3, \ldots, .8, .9, 1)$:

$$p(x) = \begin{cases} 1/11 & \text{for all } x \in X \\ 0 & \text{for all other } x \end{cases}$$

  - (Why 1/11?)
  - To use Bayes' Rule as formulated above, we next need a likelihood function, $p(y|\theta)$. What is this in this context?
  - The probability that you would observe 7 heads if the coin's probability of turning heads on any given flip is $\theta$. But we know this one already:

$$p(7 \text{ heads}|\theta) = \binom{10}{7}\theta^7(1-\theta)^3,$$

    and more generally

$$p(y \text{ heads}|\theta) = \binom{10}{x}\theta^y(1-\theta)^{10-y}.$$

  - So, using Bayes' Rule, we now have:

$$p(\theta_i|y \text{ heads}) = \frac{p(y|\theta_i)p(\theta_i)}{\sum_{\forall j} p(y|\theta_j)p(\theta_j)} = \frac{\binom{10}{y}\theta_i^y(1-\theta_i)^{10-y}}{\sum_{\forall j}\binom{10}{y}\theta_j^y(1-\theta_j)^{10-y}}$$

  - For instance, your posterior belief that the probability that the coin is unbiased ($\theta = .5$) should be

$$p(\theta = .5|7 \text{ heads}) = \frac{\binom{10}{7}.5^7(1-.5)^3}{\sum_{\forall j}\binom{10}{7}\theta_j^7(1-\theta_j)^3} = \frac{.5^7(1-.5)^3}{\sum_{\forall j}\theta_j^7(1-\theta_j)^3}$$

– The difficult part here is evaluating the sum in the denominator, but it can done pretty easily in Stata. Demonstrate and discuss ...