Stat 5101 Lecture Slides: Deck 1

Probability and Expectation on Finite Sample Spaces

Charles J. Geyer
School of Statistics
University of Minnesota

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (`http://creativecommons.org/licenses/by-sa/4.0/`).

Sets

In mathematics, a *set* is a collection of objects thought of as one thing.

The objects in the set are called its *elements*.

The notation $x \in S$ says that x is an element of the set S.

The notation $A \subset S$ says that the set A is a *subset* of the set S, that is, every element of A is an element of S.

Sets (cont.)

Sets can be indicated by listing the elements in curly brackets $\{1, 2, 3, 4\}$.

Sets can collect anything, not just numbers

$$\{1, 2, \pi, \text{cabbage}, \{0, 1, 2\}\}$$

One of the elements of this set is itself a set $\{0, 1, 2\}$.

Most of the sets we deal with are sets of numbers or vectors.

Sets (cont.)

The *empty set* $\{\}$ is the only set that has no elements.

Like the number zero, it simplifies a lot of mathematics, but isn't very interesting in itself.

The empty set has its own special notation $\varnothing$.

Sets (cont.)

Some very important sets also get their own special notation.

- $\mathbb{N}$ denotes the natural numbers $\{0, 1, 2, \ldots\}$.
- $\mathbb{Z}$ denotes the integers $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$.
- $\mathbb{R}$ denotes the real numbers.

Sets (cont.)

Another notation for sets is the *set builder* notation:

$$\{\, x \in S : \text{some condition on } x \,\}$$

denotes the set of elements of S that satisfy the specified condition and

$$\{\, h(x) : x \in S \,\}$$

denotes the image (also called the range) of the function h having domain S.

For example,

$$\{\, x \in \mathbb{R} : x > 0 \,\}$$

is the set of positive real numbers.

Intervals

Another important special kind of set is an *interval*. We use the notation

$$(a, b) = \{\, x \in \mathbb{R} : a < x < b \,\} \tag{1}$$
$$[a, b] = \{\, x \in \mathbb{R} : a \leq x \leq b \,\} \tag{2}$$
$$(a, b] = \{\, x \in \mathbb{R} : a < x \leq b \,\} \tag{3}$$
$$[a, b) = \{\, x \in \mathbb{R} : a \leq x < b \,\} \tag{4}$$

which assumes a and b are real numbers such that $a < b$.

(1) is called the *open* interval with endpoints a and b; (2) is called the *closed* interval with endpoints a and b; (3) and (4) are called *half-open* intervals.

Intervals (cont.)

We also use the notation

$$(a, \infty) = \{\, x \in \mathbb{R} : a < x \,\} \tag{5}$$
$$[a, \infty) = \{\, x \in \mathbb{R} : a \leq x \,\} \tag{6}$$
$$(-\infty, b) = \{\, x \in \mathbb{R} : x < b \,\} \tag{7}$$
$$(-\infty, b] = \{\, x \in \mathbb{R} : x \leq b \,\} \tag{8}$$
$$(-\infty, \infty) = \mathbb{R} \tag{9}$$

which assumes a and b are real numbers.

(5) and (7) are *open* intervals. (6) and (8) are *closed* intervals. (9) is both open and closed.

Functions

A mathematical *function* is a rule that for each point in one set called the *domain* of the function gives a point in another set called the *codomain* of the function. Functions are also called *maps* or *mappings* or *transformations*.

Functions are often denoted by single letters, such as f, in which case the rule maps points x in the domain to values $f(x)$ in the codomain.

f is a function, $f(x)$ is the value of this function at the point x.

Functions (cont.)

If X is the domain and Y the codomain of the function f, then to indicate this we write

$$f : X \to Y$$

or

$$X \xrightarrow{f} Y$$

Functions (cont.)

To define a function, we may give a formula

$$f(x) = x^2, \qquad x \in \mathbb{R}.$$

Note that we indicate the domain in the formula.

The same function can be indicated more simply by $x \mapsto x^2$, read "x maps to x^2."

This "maps to" notation does not indicate the domain, which must be indicated some other way.

Functions (cont.)

If the domain is a small finite set, we can just give a table

x	1	2	3	4
$f(x)$	1/10	2/10	3/10	4/10

Functions can map any set to any set

x	red	orange	yellow	green	blue
$f(x)$	tomato	orange	lemon	lime	corn

Functions (cont.)

You probably aren't used to being careful about domains of functions. You will have to start now.

What is wrong with saying a function maps numbers to numbers?

Define f by

$$f(x) = \sqrt{x}, \qquad x \geq 0.$$

Without the domain indicated, the definition makes no sense.

Functions (cont.)

Functions can be indicated by notations other than letters

$$\mathbb{R} \xrightarrow{\exp} (0, \infty)$$

is the exponential function, which has values $\exp(x)$. This function can also be denoted $x \mapsto e^x$.

$$(0, \infty) \xrightarrow{\log} \mathbb{R}$$

is the logarithmic function, which has values $\log(x)$.

These functions are inverses of each other

$$\log\Big(\exp(x)\Big) = x, \qquad \text{for all } x \text{ in the domain of exp}$$
$$\exp\Big(\log(x)\Big) = x, \qquad \text{for all } x \text{ in the domain of log}$$

Functions (cont.)

If you are used to distinguishing between base e and base 10 logarithms, calling one $\ln(x)$ and the other $\log(x)$, forget it.

In this course, $\log(x)$ always means the base e logarithm, also called natural logarithm.

Base 10 logarithms are used in probability and statistics only by people who are confusing themselves and everyone else.

Functions (cont.)

Two kinds of functions that simplify a lot of mathematics, but aren't very interesting in themselves are constant functions and identity functions.

For any constant c, the function $x \mapsto c$ can be defined on any set and is called a *constant* function.

The function $x \mapsto x$ can be defined on any set and is called the *identity* function for that set.

Functions (cont.)

We never say x^2 is a function. We must always write $x \mapsto x^2$ to indicate the squaring function.

If you are in the habit of calling x^2 a function, then how can you describe identity and constant functions? Would you say x is a function? Would you say 2 is a function?

Better to be pedantically correct and say $x \mapsto x^2$ so we can also say $x \mapsto x$ and $x \mapsto 2$.

Probability Models

Probability model, also called *probability distribution*, basic idea of probability theory.

Saying you have a probability model or distribution, doesn't say exactly how it is specified.

Probability Models (cont.)

Several ways to specify

- probability mass function (PMF)
- probability density function (PDF)
- distribution function (DF)
- probability measure
- expectation operator
- function mapping from one probability model to another

Probability Mass Functions

A *probability mass function* (PMF) is a function

$$S \xrightarrow{f} \mathbb{R}$$

whose domain S, which can be any nonempty set, is called the *sample space*, whose codomain is the real numbers, and which satisfies the following conditions: its values are nonnegative

$$f(x) \geq 0, \qquad x \in S$$

and sum to one

$$\sum_{x \in S} f(x) = 1.$$

Probability Mass Functions (cont.)

If we write the sample space as $\{x_1, \ldots, x_n\}$, then we could write the PMF as

$$\{x_1, \ldots, x_n\} \xrightarrow{g} \mathbb{R}$$

and rewrite the conditions

$$g(x_i) \geq 0, \qquad i = 1, \ldots, n$$

and

$$\sum_{i=1}^{n} g(x_i) = 1.$$

Probability Mass Functions (cont.)

Mathematical content is the same whatever notation is used.

Mathematics is invariant under changes of notation.

A PMF is a function whose values are nonnegative and sum to one. This concept can be expressed in many different notations, but the underlying concept is always the same.

Learn the concept not the notation. We will use these notations and more for PMF. You must learn to recognize the concept clothed in any notation.

Interpretation

An element of the sample space is called an *outcome*. The value $f(x)$ of the PMF at an outcome x is called the *probability* of that outcome.

We will say more about interpretation later. For now, a casual notion will do. Probability 0.3 means whatever a weatherperson means in saying there is a 30% chance of snow tomorrow.

In this course, we *never* express probabilities in percents.

Forget percents. They yield only confusion; they just help you make mistakes.

Interpretation (cont.)

For any outcome x, we have $f(x) \geq 0$ by definition of PMF.

For any outcome x, we have $f(x) \leq 1$ because

$$f(x) = 1 - \sum_{\substack{y \in S \\ y \neq x}} f(y)$$

and the right-hand side is less than or equal to one because all the terms in the sum are nonnegative.

Probabilities are between zero and one, inclusive.

Interpretation (cont.)

Probability zero means whatever a weather forecast of 0% chance of snow tomorrow would mean.

Probability one means whatever a weather forecast of 100% chance of snow tomorrow would mean.

Probability zero means "can't happen" or at least the possibility is ignored in the forecast.

Probability one means "certain to happen" or at least the possibility of it not happening is ignored in the forecast.

Finite Probability Models

A probability model is *finite* if its sample space is a finite set.

For a few weeks we will do only finite probability models.

Example 1

A sample space cannot be empty. The smallest possible has one point, say $S = \{x\}$. Then $f(x) = 1$.

This probability model is of no interest in applications. It is just the simplest of all probability models.

Example 2

The next simplest possible probability model has a sample space with two points, say $S = \{x_1, x_2\}$. Say $f(x_1) = p$. Then we know that $0 \leq p \leq 1$. Also from $f(x_1) + f(x_2) = 1$ it follows that $f(x_2) = 1 - p$.

The PMF f is determined by one real number p

$$f(x) = \begin{cases} p, & x = x_1 \\ 1 - p, & x = x_2 \end{cases}$$

For each different value of p, we get a different probability model.

The Bernoulli Distribution

Our first "brand name" distribution.

Any probability distribution on the sample space $\{0, 1\}$ is called a *Bernoulli distribution*. If $f(1) = p$, then we use the abbreviation Ber(p) to denote this distribution.

A Bernoulli distribution can represent the distribution on any two point set.

If the actual sample space of interest is $S = \{\text{apple}, \text{orange}\}$, then we map this to a Bernoulli distribution by "coding" the points. Let 0 represent apple and 1 represent orange.

Statistical Models

A *statistical model* is a family of probability models.

We often say, in a rather sloppy use of terminology, the "Bernoulli distribution" when we really mean the Bernoulli family of distributions, the set of all Ber(p) distributions for $0 \leq p \leq 1$.

The PMF of the Ber(p) distribution can be defined by

$$f_p(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

We can think of the Bernoulli statistical model as this family of PMF's

$$\{ f_p : 0 \leq p \leq 1 \}.$$

Statistical Models (cont.)

f_p is a different function for each different p.

We say that x is the argument of the function f_p.

p is not the argument of the function f_p. We need a term for it, and the standard term is *parameter*.

p is the parameter of the Bernoulli family of distributions.

Statistical Models (cont.)

The set of allowed parameter values is called the *parameter space* of a statistical model.

For the Bernoulli statistical model (family of distributions) the parameter space is the interval $[0, 1]$.

For any $p \in [0, 1]$ there is a PMF f_p of a Bernoulli distribution.

Example 3

The next simplest possible probability model has a sample space with three points, say $S = \{x_1, x_2, x_3\}$. Say $f(x_1) = p_1$ and $f(x_2) = p_2$. Now from the condition that probabilities sum to one we derive $f(x_3) = 1 - p_1 - p_2$.

The PMF f is determined by two parameters p_1 and p_2

$$f(x) = \begin{cases} p_1, & x = x_1 \\ p_2, & x = x_2 \\ 1 - p_1 - p_2, & x = x_3 \end{cases}$$

Example 3 (cont.)

Instead of saying we have two parameters p_1 and p_2, we can say we have a two-dimensional parameter vector $\mathbf{p} = (p_1, p_2)$.

The set of all pairs of real numbers (all two-dimensional vectors) is denoted $\mathbb{R}^2$. For this model the parameter space is

$$\{\, (p_1, p_2) \in \mathbb{R}^2 : p_1 \geq 0 \text{ and } p_2 \geq 0 \text{ and } p_1 + p_2 \leq 1 \,\}$$

Discrete Uniform Distribution

Our second "brand name" distribution.

Let $\{x_1, \ldots, x_n\}$ denote the sample space. The word "uniform" means all outcomes have equal probability, in which case the requirement that probabilities sum to one implies

$$f(x_i) = \frac{1}{n}, \qquad i = 1, \ldots, n$$

defines the PMF.

Later we will meet another uniform distribution, the continuous uniform distribution. The word "discrete" is to distinguish this one from that one.

Discrete Uniform Distribution (cont.)

Applications of the discrete uniform distribution are coin flips and dice rolls.

A coin flip is modeled by the uniform distribution on a two-point sample space. The two possible outcomes, usually denoted "heads" and "tails" are generally considered equally probable, although magicians can flip whatever they want.

The roll of a die (singular die, plural dice) is modeled by a uniform distribution on a six-point sample space. The six possible outcomes, 1, 2, 3, 4, 5, 6, are generally considered equally probable, but loaded dice won't have those probabilities.

Supports

More generally, if S is the sample space of a probability distribution and f is the PMF, then we say the *support* of this distribution is the set

$$\{\, x \in S : f(x) > 0 \,\},$$

that is, $f(x) = 0$ except for x in the support.

We also say the distribution is *concentrated* on the support.

Supports (cont.)

Since points not in the support "can't happen" it does not matter if we remove such points from the sample space.

On the other hand it may be mathematically convenient to leave such points in the sample space.

In the Bernoulli family of distributions, all of the distributions have support $\{0, 1\}$ except the distribution for the parameter value $p = 0$, which is concentrated at 0, and the distribution for $p = 1$, which is concentrated at 1.

Events and Measures

A subset of the sample space is called an *event*.

If f is the PMF, then the *probability* of an event A is defined by

$$\Pr(A) = \sum_{x \in A} f(x).$$

By convention, a sum with no terms is zero, so $\Pr(\varnothing) = 0$.

This defines a function Pr called a *probability measure* that maps events to real numbers $A \mapsto \Pr(A)$.

Events and Measures (cont.)

Functions $A \mapsto \Pr(A)$ whose arguments are sets are a bit fancy for a course at this level. We will not develop tools for dealing with such functions as functions, leaving that for more advanced courses.

It is important to understand that each different probability model has a different measure. The notation $\Pr(A)$ means different things in different probability models.

When there are many probability models under consideration, we decorate the notation with the parameter, as we did with PMF.

$\Pr_\theta$ is the probability measure for the parameter value θ.

Example

Consider the probability model with PMF

x	1	2	3	4
$f(x)$	1/10	2/10	3/10	4/10

and sample space $S = \{1, 2, 3, 4\}$.

What is the probability of the events

$$A = \{x \in S : x \geq 3\}$$
$$B = \{x \in S : x > 3\}$$
$$C = \{x \in S : x > 4\}$$

Events and Measures (cont.)

PMF and probability measures determine each other.

$$\Pr(A) = \sum_{x \in A} f(x), \qquad A \subset S$$

goes from PMF to measure, and

$$f(x) = \Pr(\{x\}), \qquad x \in S$$

goes from measure to PMF.

Note the distinction between the outcome x and the event $\{x\}$.

Interpretation Again

For any event A, we have $\Pr(A) \geq 0$ because all the terms in the sum in

$$\Pr(A) = \sum_{x \in A} f(x)$$

are nonnegative.

For any event A, we have $\Pr(A) \leq 1$ because all the terms in the sum in

$$\Pr(A) = 1 - \sum_{\substack{x \in S \\ x \notin A}} f(x)$$

are nonnegative.

Interpretation Again (cont.)

This gives the same conclusion as before.

Probabilities are between zero and one, inclusive.

So probabilities of events obey the same rule as probabilities of outcomes.

Random Variables and Expectation

A real-valued function on the sample space is called a *random variable*.

If f is the PMF, then the *expectation* of a random variable X is defined by

$$E(X) = \sum_{s \in S} X(s) f(s).$$

This defines a function E called an *expectation operator* that maps random variables to real numbers $X \mapsto E(X)$.

Random Variables and Expectation (cont.)

Functions $X \mapsto E(X)$ whose arguments are themselves functions are a bit fancy for a course at this level. We will not develop tools for dealing with such functions as functions, leaving that for more advanced courses.

It is important to understand that each different probability model has a different expectation operator. The notation $E(X)$ means different things in different probability models.

When there are many probability models under consideration, we decorate the notation with the parameter, as we did with PMF and probability measures.

E_θ is the expectation operator for the parameter value θ.

Sets Again: Cartesian Product

The *Cartesian product* of sets A and B, denoted $A \times B$, is the set of all pairs of elements

$$A \times B = \{ (x, y) : x \in A \text{ and } y \in B \}$$

We write the Cartesian product of A with itself as A^2.

In particular, $\mathbb{R}^2$ is the space of two-dimensional vectors or points in two-dimensional space.

Sets Again: Cartesian Product (cont.)

Similarly for triples

$$A \times B \times C = \{\, (x, y, z) : x \in A \text{ and } y \in B \text{ and } x \in C \,\}$$

We write $A \times A \times A = A^3$.

In particular, $\mathbb{R}^3$ is the space of three-dimensional vectors or points in three-dimensional space.

Sets Again: Cartesian Product (cont.)

Similarly for n-tuples

$$A_1 \times A_2 \times \cdots \times A_n = \{\, (x_1, x_2, \ldots, x_n) : x_i \in A_i,\ i = 1,\ \ldots,\ n \,\}$$

We write $A \times A \times \cdots \times A = A^n$ when there are n sets in the product.

In particular, $\mathbb{R}^n$ is the space of n-dimensional vectors or points in n-dimensional space.

Random Variables and Expectation (cont.)

Any function of random variables is a random variable.

If g is a function $\mathbb{R} \to \mathbb{R}$ and X is a random variable, then

$$s \mapsto g\Big(X(s)\Big),$$

which we write $g(X)$, is also a random variable.

If g is a function $\mathbb{R}^2 \to \mathbb{R}$ and X and Y are random variables, then

$$s \mapsto g\Big(X(s), Y(s)\Big),$$

which we write $g(X, Y)$, is also a random variable.

Example

Consider the probability model with PMF

x	1	2	3	4
$f(x)$	1/10	2/10	3/10	4/10

and sample space S.

What are

$$E(X)$$
$$E\{(X-3)^2\}$$

Averages and Weighted Averages

The *average* of the numbers $x_1, \ldots, x_n$ is

$$\frac{1}{n} \sum_{i=1}^{n} x_i$$

The *weighted average* of the numbers $x_1, \ldots, x_n$ with the *weights* $w_1, \ldots, w_n$ is

$$\sum_{i=1}^{n} w_i x_i$$

The weights in a weighted average are required to be nonnegative and sum to one.

Random Variables and Expectation (cont.)

As always, we need to learn the concept beneath the notation.

Expectation and weighted averages are the same concept in different language and notation. In expectation we sum

$$\sum \text{values of random variable} \cdot \text{probabilities}$$

in weighted averages we sum

$$\sum \text{arbitrary numbers} \cdot \text{weights}$$

but weights are just like probabilities (nonnegative and sum to one) and the values of a random variable can be defined arbitrarily (whatever we please) and are numbers.

Random Variables and Expectation (cont.)

So "expectation of random variables" and "weighted averages" are the same concept clothed in different woof and different notation.

In both cases you have a sum and each term is the product of two things. One of those things is arbitrary, the values of the random variable in the case of expectation. One of those things is nonnegative and sums to one, the probabilities in the case of expectation.

Averages and Weighted Averages (cont.)

An ordinary average is the special case of a weighted average when the weights are all equal.

This corresponds to the case of expectation in the model where the probabilities are all equal, which is the discrete uniform distribution.

Ordinary averages are like expectations for the discrete uniform distribution.

Random Variables and Expectation (cont.)

When using f for the PMF, S for the sample space, and x for points of S, if $S \subset \mathbb{R}$, then we often use X for the identity random variable $x \mapsto x$. Then

$$E(X) = \sum_{x \in S} x f(x) \tag{10}$$

and

$$E\{g(X)\} = \sum_{x \in S} g(x) f(x) \tag{11}$$

(10) is the special case of (11) where g is the identity function.

Don't need to memorize two formulas if you understand this specialization.

Random Variables and Expectation (cont.)

Don't need to memorize any formulas if you understand the concept clothed in the notation.

You always have a sum (later on integrals too) in which each term is the product of the random variable in question — be it denoted $X(s)$, x or $g(x)$, or $(x-6)^3$ — times the probability — be it denoted $f(s)$ or $f(x)$ or $f_\theta(x)$ or whatever.

Probability of Events and Random Variables

Suppose we are interested in $\Pr(A)$, where A is an event involving a random variable

$$A = \{\, s \in S : 4 < X(s) < 6 \,\}.$$

A convenient shorthand for this is $\Pr(4 < X < 6)$.

The explicit subset A of the sample space the event consists of is not mentioned. Nor is the sample space S explicitly mentioned.

Since X is a function $S \to \mathbb{R}$, the sample space is implicitly mentioned.

Sets Again: Set Difference

The *difference* of sets A and B, denoted $A \setminus B$, is the set of all points of A that are not in B

$$A \setminus B = \{\, x \in A : x \notin B \,\}$$

Functions Again: Indicator Functions

If $A \subset S$, the function $S \to \mathbb{R}$ defined by

$$I_A(x) = \begin{cases} 0, & x \in S \setminus A \\ 1, & x \in A \end{cases}$$

is called the *indicator function* of the set A.

If S is the sample space of a probability model, then $I_A : S \to \mathbb{R}$ is a random variable.

Indicator Random Variables

Any indicator function I_A on the sample space is a random variable.

Conversely, any random variable X that takes only the values zero or one (we say zero-or-one-valued) is an indicator function. Define

$$A = \{\, s \in S : X(s) = 1 \,\}$$

Then $X = I_A$.

Probability is a Special Case of Expectation

If Pr is the probability measure and E the expectation operator of a probability model, then

$$\Pr(A) = E(I_A), \qquad \text{for any event } A$$

Philosophy

Philosophers and philosophically inclined mathematicians and scientists have spent centuries trying to say exactly what probability and expectation are.

This project has been a success in that it has piled up an enormous literature.

It has not generated agreement about the nature of probability and expectation.

If you ask two philosophers what probability and expectation are, you will get three or four conflicting opinions.

Philosophy (cont.)

This is not a philosophy course. It is a mathematics course. So we are much more interested in mathematics than philosophy.

However, a little philosophy may possibly provide some possibly helpful intuition.

Although there are many, many philosophical theories about probability and expectation, only two are commonly woofed about in courses like this: frequentism and subjectivism.

We will discuss one more: formalism.

Frequentism

The *frequentist* theory of probability and expectation holds that they are objective facts about the world.

Probabilities and expectations can actually be measured in an infinite sequence of repetitions of a random phenomenon, if each repetition has no influence whatsoever on any other repetition.

Let X_1, X_2, ... be such an infinite sequence of random variables and for each n define

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

then $\overline{X}_n$ gets closer and closer to $E(X_i)$ — which is assumed to be the same for all i because each X_i is the "same" random phenomenon — as n goes to infinity.

Frequentism (cont.)

The assertion that $\overline{X}_n$ gets closer and closer to $E(X_i)$ as $n \to \infty$, is actually a theorem of mathematical probability theory, which we will soon prove.

But when one tries to build philosophy on it, there are many problems.

What does it mean that repetitions have "no influence whatsoever" on each other?

What does it mean that repetitions are of "the same random phenomenon"?

Theories that try to formalize all this are much more complicated than conventional probability theory.

Frequentism (cont.)

Worse, if probability and expectation can only be defined with respect to infinite sequences of repetitions of a phenomenon, then it has no real-world application. Such sequences don't exist in the real world.

Thus no one actually uses the frequentist philosophy of probability, although many — not understanding what that theory actually is — claim to do so.

As we shall see next semester, one of the main methodologies of statistical inference is called "frequentist" even though it has no necessary connection with the frequentist philosophy.

So many statisticians say they are "frequentists" without having commitment to any particular philosophy.

Subjectivism

The *subjectivist* theory of probability and expectation holds that they are all in our heads, a mere reflection of our uncertainty about what will happen or has happened.

Consequently, subjectivism is *personalistic*. You have your probabilities, which reflect or "measure" your uncertainties. I have mine. There is no reason we should agree, unless our information about the world is identical, which it never is.

Subjectivism (cont.)

Hiding probabilities and expectations inside the human mind, which is incompletely understood, avoids the troubles of frequentism, but it makes it hard to motivate any properties of such a hidden, perhaps mythical, thing.

Subjectivism (cont.)

The best attempt to motivate mathematical probability from subjectivism imagines each person as a bookie, who is obligated to take bets for or against any possible event in the sample space of a random phenomenon.

The bookie must formulate odds on each event and must offer to take bets for or against the occurrence of the event at the same odds.

It can be shown (we won't bother) that the odds offered must derivable from a probability measure or else there is a combination of bets on which the bookie is guaranteed to lose money.

Subjectivism (cont.)

The technical term for odds on events derived from a probability measure, so there is no way the bookie is certain to lose money, is *coherent*.

Subjectivists often say everyone else is incoherent.

But this claim is based on (1) already having accepted subjectivism and (2) accepting the picture that all users of probability and statistics are exactly like the philosophical bookie.

Since both (1) and (2) are debatable, the "incoherence" label is just as debatable.

Subjectivism (cont.)

As we shall see next semester, one of the main methodologies of statistical inference is called "Bayesian" after one of the first proponents, Thomas Bayes. Bayesian inference is often connected with subjectivist philosophy, although not always. There are people who claim to be objective Bayesians, even though there is no philosophical theory backing that up.

Many statisticians say they are "Bayesians" without having commitment to any particular philosophy.

Formalism

The mainstream philosophy of all of mathematics — not just probability theory — of the twentieth century and the twenty-first, what there is of it so far, is formalism.

> Mathematics may be defined as the subject in which we never know what we are talking about, nor whether what we are saying is true
>
> — Bertrand Russell

Formalism (cont.)

Formalists only care about the form of arguments, that theorems have correct proofs, conclusions following from hypotheses and definitions by logically correct arguments.

It does not matter what the hypotheses and definitions "really" mean ("we never know what we are talking about") nor whether they are "really" true ("nor whether what we are saying is true").

Hence we don't know whether the conclusions are true either.

We know that *if* the hypotheses and definitions are true *then* the conclusions are true. But we don't know about the "if".

Formalism (cont.)

The famous mathematician David Hilbert, who among many other things did a modern formalization of geometry (updating Euclid) said that if "point," "line," and "plane" are replaced by "table," "chair," and "beer mug" the mathematical theory does not change.

Similarly, if we replace "probability" and "expectation" by "Pooh" and "Eeyore" the mathematical theory does not change.

Formalism avoids hopeless philosophical problems about what things "really" mean and allows mathematicians to get on with doing mathematics.

Everyday Philosophy

How statisticians really think about probability and expectation.

You've got two kinds of variables: random variables are denoted by capital letters like X and ordinary variables are denoted by lower case letters like x.

A random variable X doesn't have a value yet, because you haven't seen the results of the random process that generates it. After you have seen it, it is either a number or an ordinary variable x standing for whatever number it is.

Everyday Philosophy (cont.)

In everyday philosophy, a random variable X is a mysterious thing.

It is just like an ordinary variable x except that it doesn't have a value yet, and some random process must be observed to give it a value.

Mathematically, X is a function on the sample space.

Philosophically, X is a variable whose value depends on a random process.

Everyday Philosophy (cont.)

For any random variable X, its expectation $E(X)$ is the best guess as to what its value will be when observed.

As in the joke about the average family with 1.859 children, this does not mean that $E(X)$ is a possible value of X.

It only means that $E(X)$ is a number that is closest (on average) to the observed value of X for some definition of "close" (more on this idea later).

If you have to pick one number to represent X before its value is observed, $E(X)$ is (arguably) it.

Everyday Philosophy (cont.)

When the sample space is a subset of the real numbers, the identity function $x \mapsto x$ is a random variable.

Mathematically, it is just a random variable like any other.

Philosophically, it feels different. So in everyday philosophy we distinguish between the "original" random variable, which is the identity function on the sample space, and all other random variables, which are functions of it.

Everyday Philosophy (cont.)

We say f is the PMF of a random variable X meaning

$$\Pr(X = x) = f(x), \qquad x \in S,$$

and

$$E(X) = \sum_{x \in S} x f(x),$$

where S is the sample space.

But mathematically, X is just the identity function on S.

Change of Variable

Suppose f_X is the PMF of a random variable X having sample space S, and $Y = g(X)$ is another random variable.

If we want to consider Y as the "original" random variable rather than X, then we need to determine its PMF f_Y.

This is a function on the codomain of g, call that T, given by

$$f_Y(y) = \mathsf{Pr}(Y = y), \qquad y \in T.$$

and

$$\begin{aligned} \mathsf{Pr}(Y = y) &= \mathsf{Pr}\{g(X) = y\} \\ &= \sum_{\substack{x \in S \\ g(x) = y}} f_X(x) \end{aligned}$$

Change of Variable (cont.)

Thus we have derived the *change-of-variable* formula for discrete probability distributions.

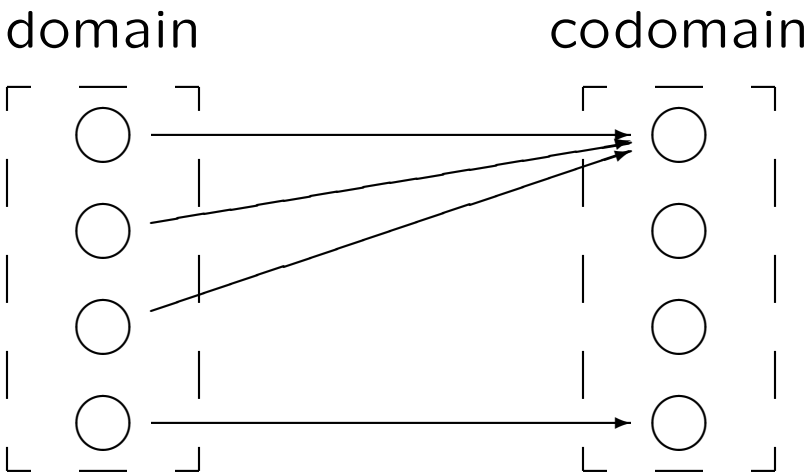
$$f_Y(y) = \sum_{\substack{x \in S \\ g(x) = y}} f_X(x), \qquad y \in T. \tag{$*$}$$

The probability distribution with PMF f_Y is sometimes called the *image distribution* of the distribution with PMF f_X because its support is the *image* of the support of X under the function g

$$g(S) = \{\, g(x) : x \in S \,\}$$

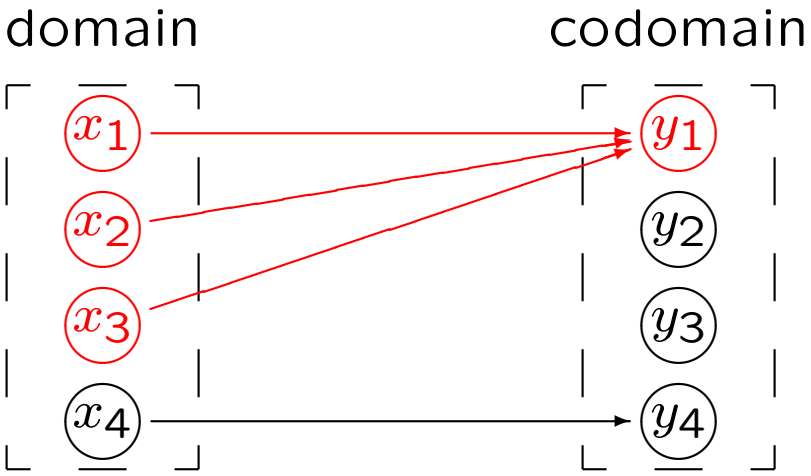
(if S is the support of X). But $(*)$ works even if S is larger than the support of X.

Change of Variable (cont.)



A picture of a function. Arrows go from x in the domain to $g(x)$ in the codomain.

Change of Variable (cont.)



$$f_Y(y_1) = f_X(x_1) + f_X(x_2) + f_X(x_3)$$

$$f_Y(y_2) = 0$$

$$f_Y(y_3) = 0$$

$$f_Y(y_4) = f_X(x_4)$$

Change of Variable (cont.)

Suppose the random vector (X, Y) has the uniform distribution on the set

$$S = \{\, (x, y) \in \mathbb{Z}^2 : 0 \leq y \leq x \leq 4 \,\}$$

What are the distributions induced by the natural projection maps

- $(x, y) \mapsto x$ and
- $(x, y) \mapsto y$?

Change of Variable (cont.)

The easy case is when the function g is *one-to-one* (maps each point of the domain to a different point of the codomain). Then we just have

$$f_Y(y) = f_X(x), \qquad \text{when } y = g(x)$$
$$f_Y(y) = 0, \qquad \text{when } y \neq g(x) \text{ for all } x$$

Otherwise, we say g is *many-to-one*, and you have to use the general change-of-variable formula, which means you have to figure out which points map to which points. A picture may help.

Change of Variable (cont.)

Suppose the random vector X has the uniform distribution on the set

$$S = \{ x \in \mathbb{Z} : -4 \leq x \leq 4 \}$$

What is the distribution induced by the map

- $x \mapsto x^3$?

Change of Variable (cont.)

We check that the function f_Y defined by the change-of-variable formula actually satisfies the conditions for a PMF. The first condition is obvious: $f_Y(y) \geq 0$ because the sum of nonnegative terms is nonnegative. The second condition is obvious from a picture: every point of the domain goes to exactly one point of the codomain, so

$$\begin{aligned}\sum_{y \in T} f_Y(y) &= \sum_{y \in T} \sum_{\substack{x \in S \\ g(x) = y}} f_X(x) \\ &= \sum_{x \in S} f_X(x) \\ &= 1\end{aligned}$$

Change of Variable (cont.)

Change-of-variable is another way of specifying a probability model.

Any function on the sample space of a probability model defines a new probability model (the image distribution).

We will use this change-of-variable formula (and other change-of-variable formulas for continuous distributions) a lot. Important!

The PMF of a Random Variable

A random variable is a function on the sample space. Hence it induces an image distribution by the change-of-variable formula.

We say two random variables X and Y having different probability models (possibly different sample spaces and different PMF's) are *equal in distribution* or *have the same distribution* if they have the same image distribution.

What is the distribution that they have? Apply the change-of-variable formula.

The PMF of a Random Variable (cont.)

The trick of allowing the sample space to be bigger than the support allows us to define the PMF of a random variable on the whole real line.

If X is a random variable, then

$$f_X(x) = \mathsf{Pr}(X = x), \qquad x \in \mathbb{R},$$

extends the PMF to all of $\mathbb{R}$.

Although $\mathbb{R}$ is an infinite set, the support of f_X is finite, so sums defining probabilities make sense (by convention the sum of any number of zeros is zero, even an infinite number of them).

The PMF of a Random Variable (cont.)

So X and Y are equal in distribution if and only if the image distributions they induce have the same PMF, that is $f_X = f_Y$ or

$$\Pr(X = r) = \Pr(Y = r), \qquad \text{for all } r \in \mathbb{R}$$

If two random variables are equal in distribution, we often say they have the same distribution, not worrying about them being defined in different probability models.

The PMF of a Random Variable (cont.)

If probability theory is to make sense, it had better be true that if $Y = g(X)$ and f_X and f_Y are the PMF's of X and Y, then

$$\begin{aligned} E(Y) &= \sum_{y \in T} y f_Y(y) \\ &= E\{g(X)\} \\ &= \sum_{x \in S} g(x) f_X(x) \end{aligned}$$

for any function $g : S \to T$, where S is the sample space for X and T is the sample space for Y.

Proving this is a homework problem.

The PMF of a Random Variable (cont.)

The preceding slide states an important fact.

If X and Y are equal in distribution, then

$$E\{g(X)\} = E\{g(Y)\}$$

for all functions g.

All expectations and probabilities — probability being a special case of expectation — depend on the distribution of a random variable but not on anything else.

The PMF of a Random Vector

For any random variable X taking values in a finite subset S of $\mathbb{R}$ and any random variable Y taking values in a finite subset T of $\mathbb{R}$ define

$$f(x,y) = \mathsf{Pr}(X = x \text{ and } Y = y), \qquad (x,y) \in S \times T.$$

By the change-of-variable formula, $f : S \times T \to \mathbb{R}$ is the PMF of the two-dimensional random vector (X, Y).

The PMF of a Random Vector (cont.)

For any random variables X_1, X_2, ..., X_n taking values in finite subsets S_1, S_2, ..., S_n of $\mathbb{R}$, respectively, define

$$f(x_1, x_2, \ldots, x_n) = \mathsf{Pr}(X_i = x_i,\ i = 1, \ldots, n),$$
$$(x_1, x_2, \ldots, x_n) \in S_1 \times S_2 \times \cdots \times S_n.$$

By the change-of-variable formula, $f : S_1 \times S_2 \times \cdots \times S_n \to \mathbb{R}$ is the PMF of the n-dimensional random vector $(X_1, X_2, \ldots, X_n)$.

The PMF of a Random Vector (cont.)

As with random variables, so with random vectors.

If $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are equal in distribution, then

$$E\{g(X_1, \ldots, X_n)\} = E\{g(Y_1, \ldots, Y_n)\}$$

for all functions g.

All expectations and probabilities — probability being a special case of expectation — depend on the distribution of a random vector but not on anything else.

Independence

The only notion of independence used in probability theory, sometimes called *statistical independence* or *stochastic independence* for emphasis, but the adjectives are redundant.

Random variables X_1, ..., X_n are *independent* if the PMF f of the random vector $(X_1, \ldots, X_n)$ is the product of the PMF's of the component random variables

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i), \qquad (x_1, \ldots, x_n) \in S_1 \times \cdots \times S_n$$

where

$$f_i(x_i) = \Pr(X_i = x_i), \qquad x_i \in S_i$$

Terminology using the Word Independence

In elementary mathematics, we say in $y = f(x)$ that x is the independent variable and y is the dependent variable. Unless your career plans include teaching elementary school math, forget this terminology!

In probability theory, it makes no sense to say one variable is independent. A set of random variables $X_1, \ldots, X_n$ is (stochastically) independent or not, as the case may be.

It also makes no sense to say one variable is dependent. A set of random variables $X_1, \ldots, X_n$ is (stochastically) *dependent* if they are not independent.

Interpretation of Independence

When we are thinking of $X_1, \ldots, X_n$ as variables whose values we haven't observed yet — data that are yet to be observed — then independence is the property that these variables have no effect whatsoever on each other.

When we are thinking mathematically — random variables are functions on the sample space — then independence has the mathematical definition just given.

Don't get the two notions — informal and formal — mixed up.

In applications, we say random variables (functions of observable data) are independent if they have no effect whatsoever on each other. In mathematics, we use the formal definition.

Independence (cont.)

Random variables $X_1, \ldots, X_n$ are *independent* if and only if the PMF f of the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ satisfies the following properties.

The support of $\mathbf{X}$ is a Cartesian product $S_1 \times \cdots \times S_n$.

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} h_i(x_i), \qquad (x_1, \ldots, x_n) \in S_1 \times \cdots \times S_n$$

where the h_i are any (strictly) positive-valued functions.

Independence (cont.)

Proof of the assertion on the preceding slide.

The distribution of the random variable X_k has PMF

$$\begin{aligned} f_k(x_k) &= \sum_{x_1 \in S_1} \cdots \sum_{x_{k-1} \in S_{k-1}} \sum_{x_{k+1} \in S_{k+1}} \cdots \sum_{x_n \in S_n} \prod_{i=1}^{n} h_i(x_i) \\ &= c_1 \cdots c_{k-1} c_{k+1} \cdots c_n h_k(x_k) \end{aligned}$$

where

$$c_i = \sum_{x_i \in S_i} h_i(x_i)$$

So each h_i is proportional to the PMF f_i of X_i. That proves one direction.

Independence (cont.)

Conversely, if S_i is the support of the distribution of X_i and the components of $\mathbf{X}$ are independent, then

$$\Pr(X_i = x_i, i \in 1, \ldots, n) = \prod_{i=1}^{n} \Pr(X_i = x_i)$$

and the right hand side is nonzero if and only if each term is nonzero, which is if and only if $(x_1, \ldots, x_n) \in S_1 \times \cdots \times S_n$.

Independence (cont.)

With our simplified criterion it is simple to check independence of the components of a random vector.

Is the support of the random vector a Cartesian product?

Is the PMF of the distribution of the random vector a product of functions of one variable?

If yes to both, then the components are independent. Otherwise, not.

Independence (cont.)

$\mathbf{X} = (X_1, \ldots, X_n)$ has the uniform distribution on S^n.

Are the components independent? Yes, because (1) the support of $\mathbf{X}$ is a Cartesian product and (2) a constant function of the vector $(x_1, \ldots, x_n)$ is the product of constant functions of each variable.

Independence (cont.)

$\mathbf{X} = (X_1, X_2)$ has the uniform distribution on

$$\{\, (x_1, x_2) \in \mathbb{N}^2 : x_1 \leq x_2 \leq 10 \,\}$$

Are the components independent? No, because (1) the support of $\mathbf{X}$ is *not* a Cartesian product, and hence we don't need to check (2).

Counting

How many ways are there to arrange n distinct things?

You have n choices for the first.

After the first is chosen, you have $n-1$ choices for the second.

After the second is chosen, you have $n-2$ choices for the third.

There are $n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$ arrangements, which is read "n factorial".

Counting (cont.)

n factorial can also be written

$$n! = \prod_{i=1}^{n} i$$

The $\prod$ sign is like $\sum$, except $\prod$ means product where $\sum$ means sum.

By definition $0! = 1$. There is one way to order zero things. Here it is in this box $\boxed{}$. (There are zero things in the box, and they are in order.)

Counting (cont.)

How many ways are there to arrange k things chosen from n distinct things?

After the first is chosen, you have $n-1$ choices for the second.

After the second is chosen, you have $n-2$ choices for the third.

You stop when you have made k choices.

There are $(n)_k = n(n-1)(n-2)\cdots(n-k+1)$ arrangements, which is read "the number of permutations of n things taken k at a time".

Counting (cont.)

The number of permutations of n things taken k at a time can also be written

$$(n)_k = \prod_{i=n-k+1}^{n} i = \frac{n!}{(n-k)!}$$

The convention $0! = 1$ makes $(n)_n = n!$, which makes sense.

Counting (cont.)

How many ways are there to choose k things from n distinct things (the order of the k things chosen doesn't matter)?

There are $(n)_k$ ways to choose when order does matter.

Each choice can be arranged $k!$ ways.

Thus each choice is counted $k!$ times in the $(n)_k$ arrangements.

Thus the number of choices is

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

which is read "the number of combinations of n things taken k at a time".

Counting (cont.)

Note that

$$\binom{n}{k} = \binom{n}{n-k}$$

There are two ways to choose k things from n things.

You can just directly choose the k things.

Alternatively, you can choose the $n-k$ things that are left out.

$0! = 1$ comes into play here too. Since there is one way to chose n things from n things (take them all), there had better also be one way to choose zero things from n things (take none).

Counting (cont.)

Alternative notations for permutations

$$(n)_k = P(n,k) = {}_nP_k$$

Alternative notations for combinations

$$\binom{n}{k} = C(n,k) = {}_nC_k$$

For us, combinations are much more important than permutations and we will always use the notation $\binom{n}{k}$.

The $\binom{n}{k}$ are also called *binomial coefficients*.

Binomial Theorem

Expand $(a+b)^n$.

There are 2^n terms, each of the form $x_1 x_2 \cdots x_n$, where each x_i is either an a or a b.

Order doesn't matter because multiplication is commutative.

There are $\binom{n}{k}$ terms equal to $a^k b^{n-k}$ because there are that many ways to chose k slots to put a's in.

Hence (this is called the **binomial theorem**)

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

The Binomial Distribution

Let $X_1, \ldots, X_n$ be independent and identically distributed Bernoulli random variables. Identically distributed means they all have the same parameter value: they are all Ber(p) with the same p.

Define $Y = X_1 + \ldots + X_n$. The distribution of Y is called the *binomial distribution* for *sample size* n and *success probability* p, indicated Bin(n, p) for short.

For the special case $n = 1$ we have $Y = X_1$.

So Bin$(1, p)$ = Ber(p).

Binomial Distribution (cont.)

The possible values of Y clearly range from zero (when all the X_i are zero) to n (when all the X_i are 1).

Clearly $Y = k$ when exactly k of the X_i are equal to 1 and the rest are zero.

There are $\binom{n}{k}$ ways that exactly k of the X_i are equal to one. The rest have to be zero.

When $Y = k$ we have

$$\Pr(X_1 = x_1 \text{ and } \cdots \text{ and } X_n = x_n) = \prod_{i=1}^{n} \Pr(X_i = x_i) = p^k(1-p)^{n-k}$$

because the X_i are independent and because multiplication is commutative.

Binomial Distribution (cont.)

Hence the binomial distribution has PMF

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n$$

The sample space is $\{0, 1, \ldots, n\}$ and the parameter space is $[0, 1]$ just like for the Bernoulli distribution.

Addition Rules

We have now met another "brand name" distribution Bin(n, p).

We have also met our first "addition rule".

If $X_1, \ldots, X_n$ are independent and identically distributed (IID) Ber(p) random variables, then $Y = X_1 + \cdots + X_n$ is a Bin(n, p) random variable.

Binomial Distribution (cont.)

Suppose X is a Bin(n, p) random variable. What is $E(X)$?

$$
\begin{aligned}
E(X) &= \sum_{x=0}^{n} x f(x) \\
&= \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \cdot \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!\,(n-x)!} p^x (1-p)^{n-x}
\end{aligned}
$$

Binomial Distribution (cont.)

Continuing where we left off on the preceding slide

$$\begin{aligned} E(X) &= \sum_{x=1}^{n} \frac{n!}{(x-1)!\,(n-x)!} p^x (1-p)^{n-x} \\ &= np \cdot \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!\,(n-x)!} p^{x-1} (1-p)^{n-x} \\ &= np \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} \\ &= np \end{aligned}$$

In short, $E(X) = np$ when X has the Bin(n, p) distribution.

A Method of Calculating Expectations

The method used to calculate $E(X)$ here seems very tricky, but the principle is widely used and important to learn.

For many distributions — and the binomial is no exception — about the only relevant sums we know how to do are equivalent to the fact that the probabilities sum to one.

About the binomial distribution, we know

$$\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = 1$$

and this is the special case of the binomial theorem with $a = p$ and $b = 1 - p$. No theory we know tells how to do other sums involving binomial coefficients.

A Method of Calculating Expectations (cont.)

So if we can't use the fact that probabilities sum to one to do the expectation we are trying to do, then we can't do it at all.

Thus our decision to pull the factor np out of the sum and our decision to change the summation index from x to $k = x - 1$ were not unmotivated.

A Method of Calculating Expectations (cont.)

Once we saw the x in the numerator canceled with the x in the $x!$ in the denominator leaving $(x-1)!\,(n-x)!$ in the denominator, we asked ourselves what binomial coefficient has that denominator and answered $\binom{n-1}{x-1}$. Then we ask what binomial distribution has those coefficients and answered $\text{Bin}(n-1, p)$ with terms

$$\binom{n-1}{k} p^k (1-p)^{n-1-k} = \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x}$$

This trick will be used over and over, both with sums and later — when we define probabilities by integrals — also with integrals.

If you can't somehow use the fact that probabilities sum (or integrate) to one for every distribution in the family, then you probably can't do the sum (or integral) in question.