Lecture 3: Interpretations of probability; Decision theory

Christopher A. Sims
Princeton University
sims@princeton.edu

September 7, 2020

Sample distribution

- The AK dataset you have worked with contains $N = 329,509$ observations on 5 variables.
- We can treat the observations as a state space, giving each observation (row of the AK data matrix) the probability $1/N$.
- We can construct subsets of the data matrix — for example, the sset A of all those observations with `logwage` between 4.9 and 5.2 — and define the number of observations in A divided by N to be $P[A]$.
- This gives us a $\{S, \mathcal{F}, P\}$ that satisfies all the properties of a probability. It's called the sample distribution of the data set.

A smaller $\mathcal{F}$

- If we want to characterize just sets defined by the values taken on by the five random variables defined by the columns of the AK data, we get a much smaller family of subsets of S to consider.

- You might think, since `logwage` is a real number, that every observation would probably have a distinct value for it. But this is far from true. There are fewer than 27,000 distinct values of `logwage` among the N observations in the data set.

- In fact more than 39,000 of the observations have exact matches for all 5 variables with some other observation.

- So if we define our probability only on subsets of observations that can be characterized in terms of values of the 5 variables, we get a substantially smaller $\mathcal{F}$.

This P has nothing to do with "randomness"

- We're just counting points in the dataset, which is already available to us and not uncertain.
- But often we do want to connect a P to some notion of "randomness".

Randomness as based on uncertainty and repetition

- We think of fair coin flips as uncertain in advance with equal chance of head or tail, but as predictably converging toward equal proportions of heads and tails over long sequences of flips.

- Another example like this is the number of points on a throw of two dice, which can deliver a number between 1 and 12.

- We can also think of the AK data as having been produced by something like a sequence of coin flips or throws of dice.

Sampling distributions

- (Note: the sample distribution is not a sampling distribution.)
- We imagine that there is an actual, finite population of millions of men born in 1930-1939.
- We think of the AK data as generated by giving each person in that population an index number, then using some "random mechanism" to draw index numbers until we have N people in the sample.
- It makes things easier if we think of the drawing as "with replacement", meaning that the population we draw from is always the same, and thus that it is possible for the same person to be drawn twice.

AK data as a "random sample"

- If we think of the data as generated this way, the whole AK data sample is like a single coin flip or a single throw of dice.

- We don't know what it will be beforehand, but if we knew the distribution of the population, we could put probabilities on different possible draws of N people from the population.

- For example, if the proportion of the population with `logwage` less than .5 is .44, we would expect that across many draws of N people from the population, the proportion of the N draws with `logwage` below .5 would average out to .44, just as the proportion of heads in a long sequence of coin flips averages out to .5.

A second type of sampling distribution

- What we have described above is a probability distribution based on what has been called "design-based inference". The *only* randomness in it is the randomness introduced by the creator of the sample, so that if the methods by which the sample is constructed are fully explicit, that's all we need to understand completely the randomness.

- But for many purposes, especially if we are estimating something (e.g. "returns to education") that we think is useful for decision-making outside the context of just summarizing the properties of this population, it is useful to think of the finite population itself as generated as a random sample from an underlying distribution. Then the current sample of size N can be thought of as a random sample from the underlying distribution. This way of thinking about it is called "model based inference".

Probability as reflecting beliefs

- Consider a game in which when it is one's turn, one draws a card that contains a question with three mutually exclusive possible answers, like, "Is the population of Davenport, IA, less than 50,000, between 50,000 and 250,000, or more than 250,000? Answer to be determined by checking Wikipedia.

- The population of Davenport is not random in the sense of sampling probability. It is a number. The player just doesn't know it (probably). The player then must quote odds on the three choices, say 5:2, 1:6 and 1:12, e.g. The player's opponent then can divide up $\$10$ in betting for or against the three alternatives, at the quoted odds.

- Then the answer is looked up and the payment made.

How to avoid being a sure loser

- The player who has to quote odds obviously should think about what likely answers are in deciding on the odds quote. There's no way to avoid losing to an opponent who knows the answer, or even, in expectation, to an opponent who is uncertain, but has better information than you do.

- You should convert any odds you quote to probabilities (5:2 becomes 5/7, 1:6 becomes 1/7, etc.). Then you should check that the odds you qre quoting satisfy the properties of a probability.

- If not, then your opponent has a "Dutch book" available: She can, without any knowledge about the true answer, guarantee that whatever the truth is, she wins money from the odds-quoter.

What's "random" here?

- A probability arises here without any reference to repetition or averaging.
- But it does correspond to putting weights on a state space reflecting the nature of the player's uncertainty about an unknown quantity.
- The population of Davenport is random to the player who does not know it, until it is looked up, after which it is not random. Randomness here is a synonym for uncertainty.
- (If your opponent quotes 5:2, 1:6 and 1:7 odds, you have a Dutch book available. What property of a probability distribution is violated, and how should you allocate your $\$10$ to guarantee a win?)

The "market measure"

- Consider an asset market for contingent claims, where agents can buy securities that make payments contingent on arbitrary subsets of a state space.

- In such a market, arbitrage implies that the asset prices, normalized so that a riskless asset paying 1 in all states has price 1, define a probability.

- We could also define a probabiility, if this market repeats with the same state space, by setting the price of an asset equal to its average return over repeated markets.

- It is one of the lessons of asset pricing theory that the sampling

distribution "price" is not the market price in the presence of risk aversion.

Decision theory

- This generalizes the "Davenport, IA" example.
- There is again a state space S, with two random vectors X and Y defined on it.
- There is a utility function $U(Y, X, d)$, where d is a vector that we get to choose. (d for "decision")
- We choose d after seeing the value of X, but without seeing the value of Y.

Admissibility

- Suppose someone proposes that for a particular value ξ of X we choose $\delta(\xi) = d$, but there is another choice d^* such that $U(Y, \xi, d^*) > U(Y, \xi, d)$ for every possible value of Y.

- Then it is clearly better to use d^* than d. No probabilities are involved.

- If there is a d^* such that $U(Y, \xi, d^*) \geq U(Y, \xi, d)$ for every value of Y, and the inequality is strict for some value of Y, we say d is **inadmissible**. A d that is not inadmissible is **admissible.**

The complete class theorem

- Every admissible d maximizes $E[U(Y, \xi, d)]$ with respect to d under some probability measure $P[| \ \xi]$ on $Y \mid \xi$.

- We have tp add some regularity assumptions. In a finite state space, all we have to add is that the set of $U(\quad, \xi, d)$ functions obtainable as we vary d, keeping $= xi$ fixed, is convex.

- Decision rules of this form are called **Bayes decision rules**.

- When the state space is infinite, the statement of the theorem and its regularity conditions become more complicated. For example, some admissible d's may only be limits of sequences of Bayes decision rules.

Proving the complete class theorem

- We'll prove it in a two-state framework, where we can do it with 2-d plots.

- More generally, though, the math is exactly the same as the familiar theorem that the efficiency frontier of a production technology, under suitable convexity assumptions, has the property that every efficient point minimizes cost for some price vector.

- The probability in the complete class theorem plays the role of the price vector in the efficiency result.

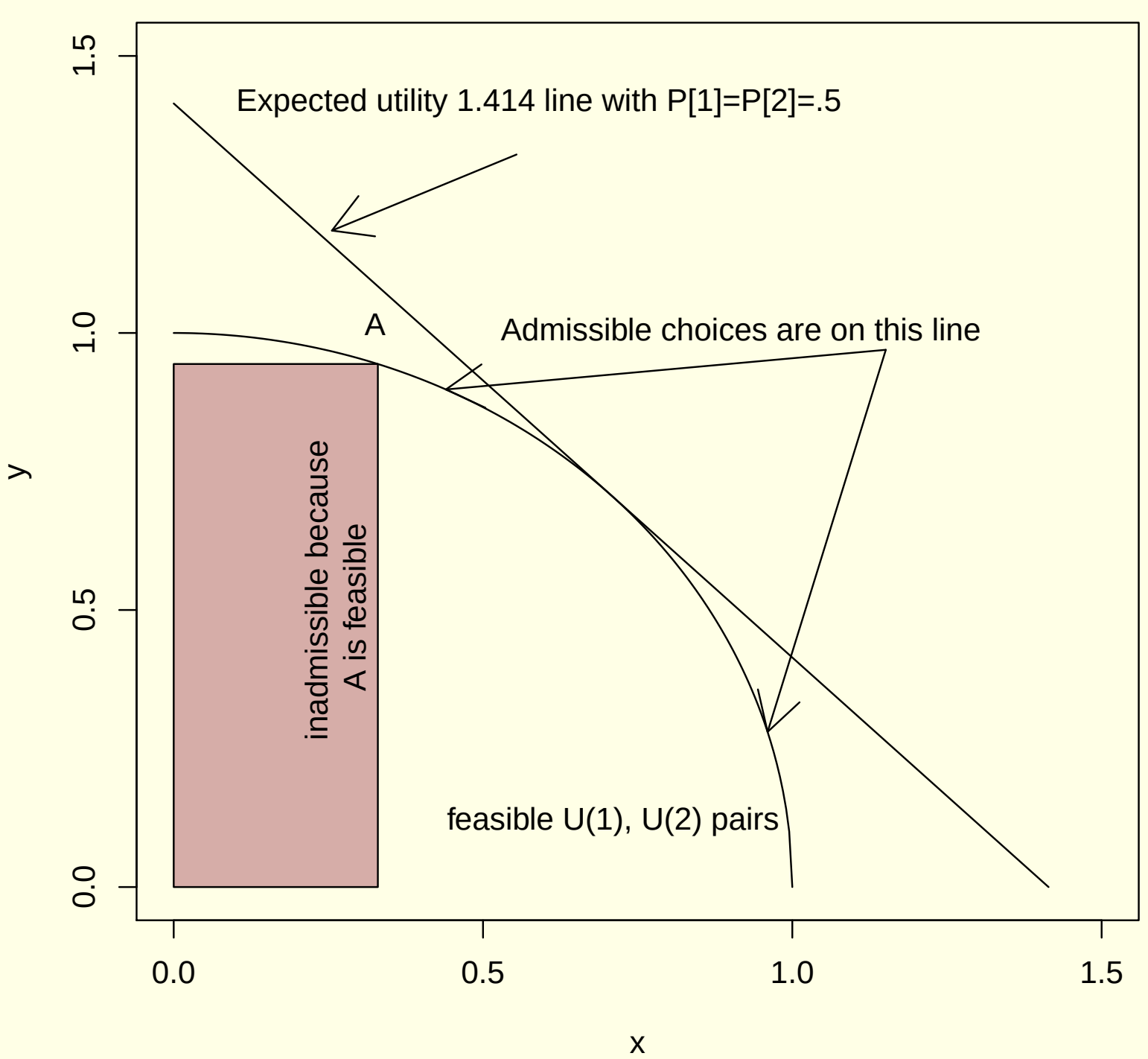
Why did we bother with X and ξ?

- The result holds X fixed at ξ, so it is in effect a separate result for each value of X. We could have just left X out.

- An admissible decision rule $d = \delta(X)$ must be admissible at each X value and thus must correspond, at each X, with a $Y \mid \xi$ probability.

- But the result doesn't require that we have a joint distribution for Y, X and form the conditional distribution of $Y \mid X$ from the joint distribution.

- However, we could generalize to a case where we have d_1 chosen based on seeing X_1, then d_2 after seeing X_1 and X_2. In that case we need a joint distribution over X_2, Y conditional on X_1, and admissibility requires that our distribution for $Y \mid X_1, X_2$ at the second stage be the conditional distribution implied by our initial $Y, X_2 \mid X_1$ distribution.


Two−state decision theory
Expected utility 1.414 line with P[1]=P[2]=.5
A
Admissible choices are on this line
inadmissible because
A is feasible
feasible U(1), U(2) pairs
y
x
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5

Frequentist version of decision theory

- There is a way of thinking about probability and inference that insists on sharply distinguishing "parameters" and "data".
- Data can be observed, and before they are observed, they are random, with distributions that may depend on the non-stochastic unknown parameters.
- In our framework above, Y plays the role of a parameter, since it is never observed, while X is data. To remind us of this, from here on in this discussion we replace "Y" by "β".
- A frequentist is willing to postulate a "model", that is a conditional distribution for $X \mid \beta$.

The gain function

- (Usually this theory uses a loss function equivalent to minus our previous $U(Y, X, d)$ and calls its expected value give β a "risk function".)

- We can form then $G(\beta, \delta(\,)) = E[U(\beta, X, \delta(X)) \mid \beta]$.

- Now we introduce a related, but different, notion of inadmissibility.

- The decision function $\delta(\,)$ is inadmissible if there is another decision function $\delta^*(\,)$ for which $G(\beta, \delta^*(\,)) \geq G(\beta, \delta(\,))$ for all β, with strict inequality for some β.

- Then if the set $\mathcal{G}$ of feasible $G(\beta, \delta(\,))$ functions of β is convex, every admissible $\delta()$ is a Bayes decision rule, meaning that there is a distribution for β such that $\delta(\,)$ maximizes over $\delta \in \mathcal{D}$ (the feasible δ's) $E[G(\beta, \delta(\,))]$.