Stat 217 Week 2 Assignment Spring 2008

Reading

Tuesday 22 Jan

– *Probability Explained* pages 55–72

– Review Appendices B, S, and F if needed.

Thursday 24 Jan

– None

Homework Due at Noon on Fri 1 Feb

Do the problems starting on the next page. Note that there are only 6 problems to do on this homework. I have included two problems ("Infinitely Often", "Run Lengths. . .") for reference, which I've marked as such, because you will need the ideas contained in those problems. You should read and think about these two, but you do not need to turn in a solution.

There are many words in these exercises, but the tasks you have to do and calculations you need to make are not complex. The real challenge is setting up the situation so that you know what you need to find.

1

Jaggers's Best Bet

In the Jaggers story, suppose slots 7, 8, 9, 17, 18, 19, 22, 28, 39 showed an "upward bias," meaning that across many spins, these slots came up more often than 1/38th of the time. Define the set $\mathcal{J} = \{7, 8, 9, 17, 18, 19, 22, 28, 39\}$ and let p_s be the proportion of times in the long run that slot s comes up on repeated spins, as defined on page 21. Assume that

$$p_s > \frac{1}{38} + \epsilon \text{ if } s \in \mathcal{J}$$
$$p_s \leq \frac{1}{38} - \delta \text{ if } s \notin \mathcal{J},$$

where $\epsilon > 0$ represents the upward bias of the wheel in Jaggers's favor and $\delta > 0$ is completely determined by the value of ϵ. Here, we are assuming for simplicity that the bias is the same for all slots in $\mathcal{J}$.

(a) Express δ in terms of ϵ. Recall that $\sum_{s=-1}^{36} p_s = 1$.

(b) Define the random variable $B_{\mathcal{J}}$ to be the payoff of a combined bet consisting of \$1 Straight Plays on slots 7, 8, 9, 17, 18, 19, 22, 28, and 39. Write $B_{\mathcal{J}}$ explicitly as a function on the outcome space defined in (1.6). (Note: You can use previously defined random variables in expressing $B_{\mathcal{J}}$.)

(c) Using equation (1.7) or (1.8), find $\mathsf{E}(B_{\mathcal{J}})$ as a function of ϵ.

Then find the smallest value of ϵ – the smallest bias in the wheel – such that Jaggers will make money in the long run, i.e., such that $\mathsf{E}(B_{\mathcal{J}}) > 0$. Does your answer make the Jaggers story seem more or less plausible?

(d) What is Joseph Jaggers optimal bet in this situation among those allowed by the casino? That is, you need to find a bet B combining official bets (as $B_{\mathcal{J}}$ does) that maximizes $\mathsf{E}(B)$ over all bets on sets of slots. Support your answer.

2

Infinitely Often

THIS EXERCISE IS FOR REFERENCE. READ THIS AND THINK ABOUT IT, BUT YOU DO NOT NEED TO TURN IN A SOLUTION.

As in the coin flipping model, we will not be shy about considering experiments with an infinite sequence of actions. It allows us to model some complicated situations more easily. When dealing with an infinite sequence of measurements, there are some events that can only be described with reference to the entire sequence.

One example is the event that infinitely many events in some sequence occurs. Let $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \ldots$ be a sequence of events. We define the event $\{A_i \text{ infinitely often}\}$, which we abbreviate $\{A_i \text{ i.o.}\}$, by

$$\{\mathcal{A}_i \text{ infinitely often}\} = \bigcap_{k=1}^{\infty} \bigcup_{i \geq k} \mathcal{A}_i.$$

Notice that the index i in $\{\mathcal{A}_i \text{ i.o.}\}$ is a dummy variable that implicitly goes over all the indices of the $\mathcal{A}_i$s; we could write it equivalently as $\{\mathcal{A}_j \text{ i.o.}\}$ or $\{\mathcal{A}_n \text{ i.o.}\}$ or whatever.

(a) Interpret the right hand side of this definition in several complete sentences. Keep in mind what union and intersection mean.

(b) Explain why the following holds: if $\omega \in \{\mathcal{A}_i \text{ i.o.}\}$, then the sequence of numbers $1_{\mathcal{A}_i}(\omega)$ for $i \geq 1$ contains an infinite number of 1s.

If I_j are indicators, we usually shorten the notation for the event $\{I_j = 1 \text{ i.o.}\}$ by writing $\{I_j \text{ i.o.}\}$.

(c) Using what you have learned about the coin flipping experiment, indicate whether you think the event $\{H_i \text{ i.o.}\}$ (i) must occur, (ii) can occur, (iii) cannot occur, and explain your answer.

3

Run Lengths and Borel-Cantelli 0

THIS EXERCISE IS FOR REFERENCE. READ THIS AND THINK ABOUT IT, BUT YOU DO NOT NEED TO TURN IN A SOLUTION.

Let $I_1, I_2, \ldots$ be indicator random variables and define $I = 1_{\{I_j \text{ i.o.}\}}$ to be the indicator that infinitely many of the I_js equal 1.

Two important results, called the Borel-Cantelli Lemmas, allow us to compute $\mathsf{E}(I)$ if we know $\mathsf{E}(I_j)$ for every j. The first of these is what I call Borell-Cantelli 0. (The standard name for these are the First and Second Borel-Cantelli Lemma, which in my view has no mnemonic value. Why not label the results by their conclusions? I call them Borel-Cantelli 0 and Borel-Cantelli 1, see below for the latter.) Borell-Cantelli 0 states that if

$$\sum_{j=1}^{\infty} \mathsf{E}(I_j) < \infty,$$

then $\mathsf{E}(I) = 0$. As we will see in the next chapter when we discuss probability, this is a strong conclusion: $\mathsf{E}(I) = 0$ means that $\{I_j \text{ i.o.}\}$ does *not* occur. The name Borel-Cantelli 0 is intended to remind you of the zero in the conclusion.

We will use Borel-Cantelli 0 to study the lengths of runs of heads in a sequence of coin flips, what are often called "run lengths."

Suppose we have positive integers $\ell_1, \ell_2, \ldots$, and define random variables R_n to be the indicator that there is a run of heads of length ℓ_n beginning with the nth flip. We can write R_n in terms of the H_is as follows:

$$R_n = H_n H_{n+1} \cdots H_{n+\ell_n-1},$$

where the product equals 1 only if all the Hs listed are 1.

(a) Show that $\mathsf{E}(R_n) = 2^{-\ell_n}$ by equation (1.28).

(b) Assume that $\ell_n \geq an + b$ for constants $a > 0$ and , b Show that $\mathsf{E}\left(1_{\{R_n \text{ i.o.}\}}\right) = 0$.

(c) Assume that $\ell_n \geq a \log_2 n + b$ for constants $a > 1$ and b. Show that $\mathsf{E}\left(1_{\{R_n \text{ i.o.}\}}\right) = 0$.

Describe in a sentence or two what you can conclude from this.

NOTE: You may use the fact that $\sum_n 1/n^a < \infty$ when $a > 1$.

(d) What can you conclude when ℓ_n equals a constant b for all n?

4 Independent Indicators

In later Chapters, we will say that two random variables X and Y are *independent* if knowing the value of one of them gives you no useful information for predicting the value of the other. This idea will take a while to develop fully, but in this exercise I describe a useful special case – independent indicator random variables.

Two indicator random variables I_1 and I_2 are *independent* if

$$\mathsf{E}(I_1 I_2) = \mathsf{E}(I_1)\mathsf{E}(I_2).$$

The intuition for this, which we will develop in the next chapter, relates to the Multiplication Rule for counting (see page 589 in Appendix B and equation 1.25).

If we have more than two indicators, the same relationship generalizes. Indicators $I_1, \ldots, I_n$ are *independent* if

$$\mathsf{E}(I_1 I_2 \cdots I_n) = \mathsf{E}(I_1)\mathsf{E}(I_2)\cdots\mathsf{E}(I_n).$$

(a) Suppose we have two standard six-sided dice, one red and one blue. We run a random experiment where we roll those two dice once. Describe a simple outcome space for this experiment and define the following random variables as functions on that outcome space:
– R is the number of pips (dots) on the upward face of the red die.
– B is the number of pips (dots) on the upward face of the red die.
– $S = R + B$ is the sum of the pips showing on both dice.

For any fixed $j, k \in [1\,..\,6]$, show that the two indicators $1_{B=j}$ and $1_{\{R=k\}}$ are independent.

Are $1_{\{B=j\}}$ and $1_{\{B=k\}}$ independent, where $j, k \in [1\,..\,6]$? Support your answer.

Are $1_{\{B=j\}}$ and $1_{\{S=\ell\}}$ independent, where $j \in [1\,..\,6]$ and $\ell \in [2\,..\,12]$? Support your answer.

(b) In the coin flipping experiment, define random variables $T_i = 1 - H_i$ for $i \in \mathbb{Z}_+$. This is the indicator that the coin comes up tails on the ith flip.

For positive integers i, j with $i \neq j$, compute $\mathsf{E}(H_i H_j)$ via equation (1.28). Compute $\mathsf{E}(T_i H_j)$ and $\mathsf{E}(T_i T_j)$ as well, though you may not need to use equation 1.28 in these cases.

Use these results to determine whether the following pairs of random variables are independent: (i) H_i and H_j, (ii) T_i and H_j, and (iii) T_i and T_j. If H_i and H_j are independent, it would say that knowing whether we got heads on the ith flip tells us nothing about whether we will get heads on the jth flip.

(c) Give a rough argument to support the claim that for any distinct $i_1, \ldots, i_n \in \mathbb{Z}_+$, $H_{i_1}, \ldots, H_{i_n}$ are independent indicators and similarly if any H's are replaced by Ts.

OPTIONAL: Make this argument formal using equation (1.28).

(d) Consider a pattern `HHTH` and define

$$Y_n = H_n H_{n+1} T_{n+2} H_{n+3},$$

for integer $n \geq 1$. This Y_n is an indicator. Describe it in a sentence.

Show that Y_n and Y_{n+1} are not independent but that Y_n and Y_{n+4} *are* independent. You may assume the claim made in the previous part. NOTE: The same basic argument works here for any finite pattern.

5 Monkeys, Typewriters, and Borel-Cantelli 1

The claim that a monkey banging randomly on a typewriter will eventually produce *Hamlet* word for word is no mere urban legend. It's true, at least given the right assumptions.

Here you will demonstrate this using the second of the Borel-Cantelli Lemmas referred to in exercise "Run Lengths..." above, which I call Borel-Cantelli 1.

Again let $I_1, I_2, \ldots$ be indicator random variables and define $I = 1_{\{I_j \text{ i.o.}\}}$ to be the indicator that infinitely many of the I_js equal 1. Borel-Cantelli 1 states that if the indicators I_j are *independent* (see Exercise "Independent Indicators") and if

$$\sum_{j=1}^{\infty} \mathsf{E}(I_j) = \infty,$$

then $\mathsf{E}(I) = 1$. This means that the I_js will certainly be 1 infinitely often.

Instead of *Hamlet*, we will use a simpler pattern `HHTH`, but I promise you that if you encode *Hamlet* as a string of 0s and 1s and do this analysis, you will get the same result.

Using the coin-flipping experiment, let $S_n = H_n H_{n+1} T_{n+2} H_{n+3}$, where $T_i = 1 - H_i$ as above.

Show that $\mathsf{E}\left(1_{\{S_n \text{ i.o.}\}}\right) = 1$ and explain what this means in terms of monkeys and typewriters.

6 Capture-Recapture

Suppose that we want to count the number of fish in a lake, which we denote by n. (The parameter n here is some fixed but unknown value.) Consider the following method, called the method of capture-recapture:

A. Capture n_1 fish (sampling without replacement), tag them all, and release the fish back into the lake.
B. Wait a while for the captured fish to mix thoroughly with the uncaptured fish.
C. Capture n_2 fish (sampling without replacement) and determine the number of fish T in the second sample that have tags on them. This is a random variable.

We take n_1 and n_2 to be non-random constants that have been specified before any sampling is performed. Usually in practice, n_2 is substantially smaller than n_1.

We make the following assumptions about the situation:

A. The population of fish is closed (none enter or leave), so that n is also constant.
B. In the first sample (capture), all samples of size n_1 from the n fish are equally likely. In particular, every fish is equally likely to be captured in the first sample.
C. Whether a particular fish was captured in the first sample tells you nothing about whether it is recaptured in the second sample.
D. In the second sample (recapture), all samples of size n_2 from the n fish are equally likely.

The intuitive motivation for this sampling plan is that the proportion of tagged fish in the second sample should be approximately equal to the proportion of fish captured in the first sample. That is, we should have

$$\frac{T}{n_2} \approx \frac{n_1}{n},$$

where $\approx$ indicates that the two sides should typically be close to each other. To the degree that this is a good approximation, this suggests

that

$$n \approx \frac{n_1 n_2}{T},$$

and therefore, we use the right-hand side as an estimate of n. Specifically, we define a random variable N by

$$N = \frac{n_1 n_2}{T}.$$

The (random) value of N is our estimate of the unknown number n. Note that T is a random variable, so our estimate N, being a function of a random variable, is also a random variable.

To determine N gives good estimates of n, we need to assess how much T varies across the outcome space.

(a) Describe the elementary outcomes for this experiment. (What goes on each ticket?)

(b) Can the random variable T be negative? Can T be bigger than n_1? Bigger than n_2? Why or why not? What values can T take?

(c) What are the biggest and smallest values that N can take? What values can N take?

(d) How many (see Appendix B pages 586 and 589) elementary outcomes are contained in the event $\{T = k\}$ for non-negative integer k. (First ask yourself what this event means; describe the event in a sentence.)
HINT: How many samples contain k tagged fish and $n_2 - k$ untagged fish?

(e) Assuming that expected values (long-run averages) in this model equal averages over the outcome space, as we've seen in previous examples, find $\mathsf{E}(1_{T=k})$ for each non-negative integer k.

ASIDE. The capture-recapture estimator, as well as a few more sophisticated methods based on the same idea, are actually used in a variety of fields. For example, besides being used to estimate mammal and marine populations, these methods also played an important role in efforts to adjust the United States Census in 1990 and 2000. The U.S. Census is a highly complex task and provides a far from perfect count. People are missed (not counted), or are counted twice, and sometimes non-existent people are included in the count. This might not be

so serious a problem if these mistakes were made uniformly across the population. However, certain subsets of the population have been undercounted consistently more than others, with at least a resulting loss of representation and federal funding. This fact motivates attempts by the Census bureau to adjusts the Census counts using statistical methods. Capture-recapture estimates play a central role; they are, in effect, used to guess the number of people in different groups that have been missed by the Census. The Census corresponds to the first sample (the capture), and the so-called Post Enumeration Survey (PES) corresponds to the second (recapture). The PES is a much smaller survey carried out shortly after the Census. Adjustment of the Census has been contested, and a ruling by the Supreme Court and Congressional opposition makes it unlikely that the adjusted counts will be used in next Census.

7

Comparing Random Variable

In the coin flip experiment, define the following random variables:

- Let N be the number of flips required until the pattern `THH` first appears on three consecutive flips (including the last three flips).
- Let Q be the pattern `HHHHH` first appears (including the last five flips).
- Let R be the length of the longest run of consecutive heads until `THH` first appears (and including these three flips).

(a) What values can each of these random variables take? the random variable N take? Can any of them be infinite?

(b) Give two distinct elementary outcomes that lead to the same value of N.

(c) Which of the following is true and why: (i) $R < N$, (ii) $R \geq N$, (iii) neither $R < N$ nor $R \geq N$. Try reasoning ω by ω.

(d) Which of the following is true and why: (i) $Q < N$, (ii) $Q \geq N$, (iii) neither $Q < N$ nor $Q \geq N$.

(e) If Q and N were algebraic variables (i.e., standing for numbers) would it be possible for $Q < N$ and $Q \geq N$ to both be false? Why is it possible with random variables?

HINT: The relationship $Q < N$ is a relationship between *functions*.

(f) What do we know about R on the event $\{Q < N\}$?
NOTE: The question is asking about values of $R(\omega)$ for $\omega \in \{Q < N\}$.

(g) Describe the random variable $\min(Q, N)$ in a sentence.
NOTE: This is the random variable that maps $\omega \in \Omega$ to $\min(Q(\omega), N(\omega))$.