Solutions to In-Class Problems Week 14, Fri.

Problem 1. A prison warden has two holding cells, Cell 1 and Cell 2, for his prisoners. Unfortunately, many pairs of these prisoners are *incompatible*, and there will be some trouble if an incompatible pair are in the same cell. The warden would like to minimize trouble by not having too many incompatible pairs in the same cell. Unfortunately, the warden has no idea how to split the up prisoners, and so he decides to go with a random assignment: he will assign prisoners to one cell or the other by successive (independent) flips of a fair coin.

For any two incompatible prisoners, a, b, let $T_{a,b}$ be 1 if a and b are placed in the same cell, and 0 otherwise.

(a) What is the expected value of $T_{a,b}$?

Solution. $1/2$ ■

Suppose there are n incompatible sets of prisoners, a, b. The total trouble, T, of an assignment of prisoners to cells is the sum of $T_{a,b}$ where the sum is over the n sets of incompatible prisoners $a \neq b$. So T is the total number of incompatible sets of two prisoners that are in the same cell. The warden hopes to minimize the total trouble, T.

(b) What is the expected value of T?

Solution. $\mathrm{E}[T] = n\,\mathrm{E}[T_{a,b}] = n/2$. ■

(c) Explain why there must be a split of the prisoners between cells that separates at least half the incompatible pairs.

Solution. There must be a way of splitting up the prisoners so that the total trouble is at most the expected trouble for a random split. That is, there is an assignment of prisoners to cells that causes no more trouble than the average $n/2$.

This is a *probabilistic proof* of the existence of a pretty good (no worse than average) assignment to cells. We used a similar argument to prove there was a satisfying truth assignment for some Boolean formulas in a previous class problem. ■

(d) Suppose a, b, c are different prisoners, where a and b are incompatible, and a and c are also incompatible. Explain why $T_{a,b}$ is independent of $T_{a,c}$. Conclude that set of all the $T_{a,b}$'s is pairwise independent.

Solution. For incompatible prisoners a, b and $i, j \in \{1, 2\}$, let $S(a, b, i)$ be the event that a and b are both in cell i.

$$\begin{aligned}
&\Pr\{T_{a,b} = 1 \text{ and } T_{a,c} = 1\} \\
&= \Pr\{S(a,b,1) \cap S(a,c,1)\} + \Pr\{S(a,b,2) \cap S(a,c,2)\} && (\Pr\{\text{disjoint events}\}) \\
&= \Pr\{a, b, c \text{ in cell } 1\} + \Pr\{a, b, c \text{ in cell } 2\} \\
&= (1/2)^3 + (1/2)^3 && (\text{the cells of } a, b, c \text{ are independent}) \\
&= (1/2) \cdot (1/2) \\
&= \Pr\{T_{a,b} = 1\} \cdot \Pr\{T_{a,c} = 1\},
\end{aligned}$$

which proves that $[T_{a,b} = 1]$ and $[T_{a,c} = 1]$ are independent events. But indicator variables are independent iff the events they indicate are independent, so the random variables $T_{a,b}$ and $T_{a,c}$ are independent.

Also, it's obvious that $T_{a,b}$ and $T_{c,d}$ are independent if a, b and c, d don't overlap. So any two $T_{a,b}$'s are independent. ■

(e) Are the $T_{a,b}$'s mutually independent?

Solution. No: if a, b are in opposite cells and b, c are in opposite cells, then a, c are in the same cell. ■

(f) What is the variance of T?

Solution. The variance of $T_{a,b}$ is $1/2 - 1/4 = 1/4$ since it is a Bernoulli variable with bias $1/2$. Since the $T_{a,b}$'s are pairwise independent, the variance of T is

$$\sum_{a,b \text{ incompatible}} 1/4 = n/4.$$

■

(g) Suppose among 1000 prisoners, about a fifth of the pairs, say 100,000 pairs, turn out to be incompatible. Show that there is at most a 10% chance that the warden's random assignment differs by more than 1% from the expected number of incompatible pairs in the same cell. *Hint:* Chebyshev.

Solution. Letting $\mu = \mathrm{E}[T]$, Chebyshev says

$$\Pr\{|T - \mu| > \epsilon\mu\} \leq \frac{\mathrm{Var}[T]}{(\epsilon\mu)^2}$$

For $\epsilon\mu$ to be 1% of μ we have $\epsilon = 0.01$. Using the facts that $\mu = 50,000$ and $\mathrm{Var}[T] = 25,000$, we conclude that the probability that the total trouble differs by more than 1% from the expected 50,000 incompatible pairs is at most

$$\frac{\mathrm{Var}[T]}{(\epsilon\mu)^2} = \frac{25,000}{(0.01(50,000))^2} = \frac{25,000}{500^2} = \frac{1}{10}.$$

So there is at most a 10% chance the warden's split will cause 1% more (or less) trouble than expected. ■

Problem 2. Now we look at the situation in the previous problem in more detail. Suppose there are levels of conflict between incompatible prisoners, where the conflict level of two prisoners who may hate each other's guts is 1 if they wouldn't actually touch each other, 2 if they might hurt each other but wouldn't cause a trip to the hospital, and 3 if having them in the same cell would be *really bad*. Suppose we model the situation by assuming that a random conflict level $w_{a,b}$ equal to 1,2, or 3 is assigned to every two incompatible prisoners, a, b, uniformly and independently of all other conflict levels.

So $T_{a,b}w_{a,b}$ is 0 if a, b are placed in different cells and is $w_{a,b}$ otherwise. Define the *total conflict* to be

$$C ::= \sum_{a,b \text{ incompatible}} T_{a,b}w_{a,b},$$

that is, the sum of the levels of conflicting pairs in which the members are assigned to the same cell. We would like the total conflict to be small.

(a) What is the expected value of $w_{a,b}$ and $T_{a,b}w_{a.b}$?

Solution.

$$\mathrm{E}[w_{a,b}] = 1/3 + 2/3 + 3/3 = 2.$$

Also, since $w_{a,b}$ and $T_{a,b}$ are obviously independent,

$$\mathrm{E}[T_{a,b}w_{a,b}] = \mathrm{E}[T_{a,b}]\,\mathrm{E}[w_{a,b}] = (1/2)2 = 1.$$

■

(b) What is the variance of $w_{a,b}$ and $T_{a,b}w_{a,b}$?

Solution. $\text{Var}[w_{a,b}] = (1^2 + 2^2 + 3^2)/3 - 2^2 = 2/3$, and $\text{Var}[T_{a,b}w_{a,b}] = 4/3$ similarly. ■

(c) What is the expected value of C?

Solution. $n\,\text{E}[T_{a,b}w_{a,b}] = n$. ■

(d) Are the $T_{a,b}w_{a,b}$'s pairwise independent? ...mutually independent?

Solution. Yes. No. As in the previous problem. ■

(e) What is the variance of C?

Solution. Since the $T_{a,b}w_{a.b}$'s are pairwise independent, the variance of C is $n\,\text{Var}[T_{a,b}w_{a.b}] = 4n/3$. ■

(f) What does Chebyshev's inequality give for a bound on $\Pr\{|C - \text{E}[C]|\} > n/4$?

Solution. $64/3n$. ■

(g) Suppose someone complains about our modeling the situation as choosing a random conflict level of 1,2, or 3, and agrees only that conflict levels range between 1 and 3. So then the w's for different incompatible pairs may have different distributions, but we still assume they are independent. Could we still use Chebyshev's inequality to say something about the probability of deviating from the mean? *Hint:* : What is the maximum possible variance for a random variable with values between 0 and 3?

Solution. Since $0 \le T_{a,b}w_{a,b} \le 3$, it follows that $|\text{E}[T_{a,b}w_{a,b}] - T_{a,b}w_{a,b}| \le 3$ and so

$$\text{Var}[T_{a,b}w_{a,b}] = \text{E}\left[(T_{a,b}w_{a,b} - T_{a,b}w_{a,b})^2\right] \le 9.$$

The variance of C is then at most $9n$. Then $\Pr\{|C - \text{E}[C]| > n/4\}$ is at most $9n/(n/4)^2 = 144/n$. So as long as $n > 144$, we can say something nontrivial. ■

Problem 3. A recent Gallup poll found that 35% of the adult population of the United States believes that the theory of evolution is "well-supported by the evidence". Gallup polled 1928 people and claims a margin of error of 3 percentage points.

Let's check Gallup's claim. Suppose that there are m adult Americans, of whom pm believe in evolution; this means that $(1-p)m$ Americans do not believe in evolution. Gallup polls 1928 Americans selected uniformly and independently at random. Of these, 675 believe in evolution, leading to Gallup's estimate that the fraction of Americans who believe in evolution is within 0.03 of $675/1928 \approx 0.350$.

(a) Explain how to use the Binomial Sampling Theorem (available in the Appendix) to determine the confidence level with which Gallup can make his claim. You do not actually have to do the calculation, but are welcome to if you have the means.

Solution. We let $\epsilon = 0.03$ and $n = 1928$ in the expression on the righthand side of equation (1) in the Binomial Sampling Theorem. Evaluating the expression in Scheme (see Appendix), we see that the probability that the error is 0.03 or more is less than $0.009983 < 0.01$, which means that 99% of the time[1] the fraction p will lie within the specified range 0.35 ± 0.03. ■

(b) If we accept all of Gallup's polling data and calculations, can we conclude that there is a high probability that the number of adult Americans who believe in evolution is 35 ± 3 percent?

Solution. No. As explained in Notes and a class problem last week, the assertion that fraction p is in the range 0.35 ± 0.03 is an assertion of fact that is either true or false. The number p is an *constant* whose value we don't know, so we don't know if the asserted fact is true or false, but there is nothing probabilistic about its truth or falsehood.

We *can* say that either the assertion is true or else a 1-in-100 event occurred during the poll. Specifically, the unlikely event is that Gallup's random sample was unrepresentative. This may convince you that p is "probably" in the range 0.35 ± 0.03, but this informal "probably" is not a mathematical probability. ■

(c) Explaining Sampling to a Jury The calculation above revealed that, based on a poll of 1928 people, we can be highly confident that the per cent of people in the U.S. who believe in evolution is $35\% \pm 3\%$. Note that the actual population of the U.S. was never considered, because *it did not matter*!

Suppose you were going to serve as an expert witness in a trial. How would you explain to a jury why the number of people necessary to poll *does not depend on the population size*? (Begin by explaining why it is reasonable to model polling as independent coin tosses. Remember that juries do not understand algebra or equations; you might be ok using a little arithmetic.)

[1] An exact calculation shows that the 99% confidence level could have been achieved by polling only $n = 1863$ people. With 1928 people, the estimate actually holds at the 99.17% level.

Solution. This was intended to be a thought-provoking, conceptual question. In past terms, although most of the class could crank through the formulas for poll size and confidence levels, they couldn't articulate, and indeed didn't really believe that the derived sample sizes were actually adequate to produce reliable estimates.

Here's a way to explain why we model polling people about evolution (or anything else) as independent coin tosses that a jury might be able to follow:

> Of the approximately 250,000,000 people in the US, there are some unknown number, say 87,000,000, who believe evolution is well-supported. So in this case, the *fraction* of people believing in evolution would be
>
> $$87,000,000/250,000,000 = 0.35.$$
>
> To estimate this unknown fraction, we randomly select one person from the 250,000,000 in such a way that *everyone has an equal chance of being picked*. For example, we might get computer files from the Census Bureau listing all 250,000,000 people in the US. Then we would generate a number between 1 and 250,000,000 by some physical or computational process that generated each number with equal probability, and then we would interview the person whose number came up. In this way, we can be sure that the probability that a person we select believes in evolution is exactly the unknown fraction who believe in evolution.
>
> After we have picked a person and learned their beliefs, we perform the procedure again, making sure that everyone is equally likely to be picked the second time, and so on, for picking a third, fourth, *etc*. person. On each pick the probability of getting a person who believes in evolution is the same fraction, so we describe picking a person in this way as "flipping a coin" that has this fraction as probability of coming up "Heads"—meaning that the person selected believes in evolution.
>
> Now we all understand that if we keep flipping a coin with a certain probability of coming up Heads, then the more we flip, the closer the fraction of Heads flipped will be to that probability. Mathematical theory lets us calculate us how many time to flip coins to make the fraction of Heads very likely close to the right fraction, but we won't go into those details.
>
> It's also clear, that if two different coins have the same probability of coming up Heads, it makes no difference in our experiments which coin we use: the number of flips we need for the fraction of Heads flipped to be very likely to be close to the probability of a Head will be the same for either coin. So whether the coin had probability of heads, say, 0.35, because 87,000,000 out of 250,000,000 people believe in evolution, or 700 out of 2000, or 35 out of 100—the same number of flips will allow us to estimate the probability of Heads, and hence to estimate the fraction of the population believing in evolution. So the *number of people we need to poll is the same*, whether we are selecting from a large population or a small population, as long as the fraction believing in evolution is the same in the small population as in the large one.

■

1 Appendix

1.1 Binomial Sampling

Theorem. *Let $K_1, K_2, \ldots$, be a sequence of mutually independent 0-1-valued random variables with the same expectation, p, and let*

$$S_n ::= \sum_{i=1}^{n} K_i.$$

Then, for $1/2 > \epsilon > 0$,

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \geq \epsilon\right\} \leq \frac{1+2\epsilon}{2\epsilon} \cdot \frac{2^{-n(1-H((1/2)-\epsilon))}}{\sqrt{2\pi(1/4-\epsilon^2)n}}. \tag{1}$$

1.2 Scheme Code for Sampling Bounds

```
(define (pr n eps)
  (* (/ (+ 1 (* 2 eps)) (* 2 eps)
     (sqrt (* 2 pi (- 1/4 (* eps eps))))
     (expt 2 (* n (- 1 (h (- 1/2 eps)))))
     (sqrt n))))

(define (h a)
  (cond ((>= 0 a) 1)
        ((>= a 1) 1)
        (else (- (+ (* a (log2 a)) (* (- 1 a) (log2 (- 1 a))))))))

(define (log2 a) (/ (log a) (log 2)))

(define pi (* 4 (atan 1)))

(pr 1928 0.03)
;Value: 9.982587419699058e-3
```

1.3 Chebyshev's Theorem

Theorem (Chebyshev). *Let R be a random variable, and let x be a positive real number. Then*

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} \leq \frac{\mathrm{Var}[R]}{x^2}. \tag{2}$$