Math for CS 2015/2019 solutions to "In-Class Problems Week 14, Mon. (Session 34)"

https://github.com/spamegg1

November 22, 2022

Contents

1 Problem 1

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 Americans selected uniformly and independently at random. Of these, 675 asserted belief in evolution, leading to Gallup's estimate that the fraction of Americans who believe in evolution is $675/1928 \approx 0.350$. Gallup claims a margin of error of 3 percentage points, that is, he claims to be confident that his estimate is within 0.03 of the actual percentage.

1.1 (a)

What is the largest variance an indicator variable can have?

Proof. Answer: 1/4.

Assume H is an indicator variable. By Corollary 19.3.2 we have $\text{Var}[H] = p - p^2$. Taking its derivative with respect to p, we have $\dfrac{d(p-p^2)}{dp} = 1 - 2p$ is zero when $p = 1/2$. So it follows that the maximum value of $p - p^2$ must be at $p = 1/2$, so the maximum value of $\text{Var}[H]$ is $(1/2) - (1/2)^2 = 1/4$. □

1.2 (b)

Use the Pairwise Independent Sampling Theorem to determine a confidence level with which Gallup can make his claim.

Proof. By the Pairwise Independent Sampling, the probability that a sample of size $n = 1928$ is further than $x = 0.03$ of the actual fraction is at most

$$\left(\frac{\sigma}{x}\right)^2 \cdot \frac{1}{n} \leq \left(\frac{1}{4(0.03)^2} \cdot \frac{1}{1928}\right) \leq 0.144$$

so we can be confident of Gallup's estimate at the 85.6% level. □

1.3 (c)

Gallup actually claims greater than 99% confidence in his estimate. How might he have arrived at this conclusion? (Just explain what quantity he could calculate; you do not need to carry out a calculation.)

Proof. Gallup's sample has a binomial distribution $B_{1928,p}$ for an unknown p he estimates to be about 0.35. So he wants an upper bound on

$$Pr\left[\left|\frac{B_{1928,p}}{1928} - p\right| > 0.03\right]$$

By part (a), the variance of $B_{n,p}$ is largest when $p = 1/2$, which suggests that the probability that a sample average differs from the actual mean will be largest when $p = 1/2$. This is in fact the case. So Gallup will calculate

$$\begin{aligned} Pr\left[\left|\tfrac{B_{1928,(1/2)}}{1928} - \tfrac{1}{2}\right| > 0.03\right] &= Pr\left[\left|B_{1928,(1/2)} - \tfrac{1928}{2}\right| > (0.03)(1928)\right] \\ &= Pr[906 \leq B_{1928,(1/2)} \leq 1021] \\ &= \left[\sum_{i=906}^{1021} \binom{1928}{i}\right] / (2^{1928}) \\ &\approx 0.9912 \end{aligned}$$

Mathematica will actually calculate this sum exactly. There are also simple ways to use Stirling's formula to get a good estimate of this value. □

1.4 (d)

Accepting the accuracy of all of Gallup's polling data and calculations, can you conclude that there is a high probability that the percentage of adult Americans who believe in evolution is 35 ± 3 percent?

Proof. No. As explained in Notes and lecture, 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 a constant. We don't know its value, and we don't know if the asserted fact is true or false, but there is nothing probabilistic about the fact's 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. □

2 Problem 2

Let $B_1, B_2, \ldots, B_n$ be mutually independent random variables with a uniform distribution on the integer interval $[1, d]$. Let D equal the number of events $[B_i = B_j]$ that happen where $i \neq j$. It was observed in Section 16.4 (and proved in Problem 18.2) that $\Pr[B_i = B_j] = 1/d$ for $i \neq j$ and that the events $B_i = B_j$ are pairwise independent.

2.1 (a)

What are $\text{Ex}[E_{i,j}]$ and $\text{Var}[E_{i,j}]$ for $i \neq j$?

Proof. ??? □

2.2 (b)

What are $\text{Ex}[D]$ and $\text{Var}[D]$?

Proof. ??? □

2.3 (c)

In a 6.01 class of 500 students, the youngest student was born 15 years ago and the oldest 35 years ago. Show that more than half the time, there will be will be between 12 and 23 pairs of students who have the same birth date. (For simplicity, assume that the distribution of birthdays is uniform over the 7305 days in the two decade interval from 35 years ago to 15 years ago.)

Hint: Let D be the number of pairs of students in the class who have the same birth date. Note that $|D - Ex[D]| < 6$ IFF $D \in [12, 23]$.

Proof. ??? □

3 Problem 3

Let $G_1, G_2, G_3, \ldots$, be an infinite sequence of pairwise independent random variables with the same expectation, μ and the same finite variance. Let

$$f(n, \epsilon) ::= Pr\left[\left|\frac{\sum_{i=1}^{n} G_i}{n} - \mu\right| \leq \epsilon\right]$$

The Weak Law of Large Numbers can be expressed as a logical formula of the form:

$$\forall \epsilon > 0 \;\; Q_1 Q_2 \ldots [f(n, \epsilon) \geq 1 - \delta]$$

where $Q_1, Q_2, \ldots$ is a sequence of quantifiers from among:

$$\forall n \;\; \exists n \;\; \forall n_0 \;\; \exists n_0 \;\; \forall n \geq n_0 \;\; \exists n \geq n_0 \;\; \forall \delta > 0 \;\; \exists \delta > 0 \;\; \forall \delta \geq 0 \;\; \exists \delta \geq 0$$

Here the n and n_0 range over nonnegative integers, and δ and ϵ range over real numbers.

Write out the proper sequence $Q_1 Q_2 \ldots$

Proof. $(\forall \delta > 0)(\exists n_0)(\forall n \geq n_0)$ □

4 Problem 4

An International Journal of Epidemiology has a policy of publishing papers about drug trial results only if the conclusion about the drug's effectiveness (or lack thereof) holds at the 95% confidence level. The editors and reviewers carefully check that any trial whose results they publish was properly performed and accurately reported. They are also careful to check that trials whose results they publish have been conducted independently of each other.

The editors of the Journal reason that under this policy, their readership can be confident that at most 5% of the published studies will be mistaken. Later, the editors are embarrassed, and astonished, to learn that every one of the 20 drug trial results they published during the year was wrong. The editors thought that because the trials were conducted independently, the probability of publishing 20 wrong results was negligible, namely, $(1/20)^{20} < 10^{-25}$.

Write a brief explanation to these befuddled editors explaining what's wrong with their reasoning and how it could be that all 20 published studies were wrong.

Hint: xkcd comic: "significant" xkcd.com/882/

Proof. The editors have confused the statistical confidence level with probability. It's a mistake to think that because the conclusion of particular drug trial submitted to the journal holds at the 95% confidence level, this means its conclusion is wrong with probability only 1/20.

The conclusion of the particular submitted drug trial is right or wrong. An assertion of 95% confidence means that if very many trails were carried out, we expect that

close to 95% of the trials would yield a correct conclusion. So if the results of all the many trials were all submitted for publication, and the editors selected 20 of these at random to publish, then they could reasonably expect that only one of them would be wrong.

But that's not what happens: not all the trials are written up and submitted, so the confidence level of the trial is not specially relevant. For example, there may be more than 400 worthless "alternative" drugs being tried by proponents who are genuinely honest, even if misguided. When they conduct careful trials with a 95% confidence level, we can expect that in 1/20 of the 400 trials, worthless, even damaging, drugs will look helpful. The remaining 19/20 of the 400 trials would not be submitted for publication by honest proponents because the trials did not show positive results at the 95% level. But the 20 that mistakenly showed positive results might well all be submitted with no intention to mislead.

This is why, unless there is an explanation of why a therapy works, scientists and doctors usually doubt results claiming to confirm the efficacy of some mysterious therapy at a high confidence level. □

5 Problem 5 (Supplementary Problem)

A defendent in traffic court is trying to beat a speeding ticket on the grounds that, since virtually everybody speeds on the turnpike, the police have unconstitutional discretion in giving tickets to anyone they choose. (By the way, we don't recommend this defense :-).)

To support his argument, the defendent arranged to get a random sample of trips by 3,125 cars on the turnpike and found that 94% of them broke the speed limit at some point during their trip. He says that as a consequence of sampling theory (in particular, the Pairwise Independent Sampling Theorem), the court can be 95% confident that the actual percentage of all cars that were speeding is $94 \pm 4\%$.

The judge observes that the actual number of car trips on the turnpike was never considered in making this estimate. He is skeptical that, whether there were a thousand, a million, or 100,000,000 car trips on the turnpike, sampling only 3,125 is sufficient to be so confident.

Suppose you were were the defendent. How would you explain to the judge why the number of randomly selected cars that have to be checked for speeding does not depend on the number of recorded trips? Remember that judges are not trained to understand formulas, so you have to provide an intuitive, nonquantitative explanation.

Proof. This was intended to be a thought-provoking, conceptual question. In past terms, although most of the class could follow the derivations and crank through the formulas to calculate sample size and confidence levels, many students 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 sampling cars as independent coin tosses that might work, though we aren't sure about this.

Of the approximately 36,000,000 recorded turnpike trips by cars in 2009, there were some unknown number, say 35,000,000, that broke the speed limit at some point during their trip. So in this case, the fraction of speeders is 35,000,000/36,000,000 which is a little over 0.97.

To estimate this unknown fraction, we randomly select some trip from the 36,000,000 recorded in such a way that every trip has an equal chance of being picked. Picking a trip to check for speeding this way amounts to rolling a pair dice and checking that double sixes were not rolled —this has exactly the same probability as picking a speeding car.

After we have picked a car trip and checked if it ever broke the speed limit, make another pick, again making sure that every recorded trip is equally likely to be picked the second time, and so on, for picking a bunch of trips. Now each pick is like rolling the dice and checking against double sixes.

Now everyone understands that if we keep rolling dice looking for double sixes, then the longer we roll, the closer the fraction of rolls that are double sixes will be to 1/36, since only 1 out of the 36 possible dice ouotcomes is double six. Mathematical theory lets us calculate us how many times to roll the dice to make the fraction of double sixes very likely close to 1/36, but we needn't go into the details of the calculation.

Now suppose we had a different number of recorded trips, but the same fraction were speeding. Then we could simply use the same dice in the same way to estimate the speeding fraction from this different set of trip records.

So the number of rolls needed does not depend on how many trips were recorded, it just depends on the fraction of recorded speeders. □

6 Problem 6 (Supplementary Problem)

The proof of the Pairwise Independent Sampling Theorem 19.4.1 was given for a sequence $R_1, R_2, \ldots$ of pairwise independent random variables with the same mean and variance in the course textbook.

The theorem generalizes straighforwardly to sequences of pairwise independent random variables, possibly with different distributions, as long as all their variances are bounded by some constant.

Theorem 1. *(Generalized Pairwise Independent Sampling). Let $X_1, X_2, \ldots$ be a sequence of pairwise independent random variables such that $Var[X_i] \leq b$ for some $b \geq 0$ and all $i \geq 1$. Let*

$$A_n ::= \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$\mu_n ::= Ex[A_n]$$

Then for every $\epsilon > 0$,

$$Pr[|A_n - \mu_n| \geq \epsilon] \leq \frac{b}{\epsilon^2} \cdot \frac{1}{n}$$

6.1 (a)

Prove the Generalized Pairwise Independent Sampling Theorem.

Proof. Essentially identical to the proof of the Pairwise Independent Sampling (19.4.1) in the text, except that G gets replaced by X and Var $[G_i]$ by b, with the equality where the b is first used becoming $\leq$. □

6.2 (b)

Conclude that the following holds:

Corollary 1. *(Generalized Weak Law of Large Numbers). For every* $\epsilon > 0$,

$$\lim_{n \to \infty} Pr[|A_n - \mu_n| \leq \epsilon] = 1$$

Proof. $\Pr[|A_n - \mu_n| \leq \epsilon] = 1 - \Pr[|A_n - \mu_n| > \epsilon] \geq 1 - b/(n\epsilon)^2$.

For any fixed ϵ this last term approaches 1 as $n \to \infty$. □