In-Class Problems Week 14, Mon.

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.

(a) What is the largest variance an indicator variable can have?

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

(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.)

(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?

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 to 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.

Let $E_{i,j}$ be the indicator variable for the event $[B_i = B_j]$.

(a) What are $\mathrm{Ex}[E_{i,j}]$ and $\mathrm{Var}[E_{i,j}]$ for $i \neq j$?

(b) What are $\mathrm{Ex}[D]$ and $\mathrm{Var}[D]$?

(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 $|D - \mathrm{Ex}[D]| < 6$ IFF $D \in [12, 23]$.

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:

$$\begin{array}{cccccc} \forall n & \exists n & \forall n_0 & \exists n_0 & \forall n \geq n_0 & \exists n \geq n_0 \\ & \forall \delta > 0 & \exists \delta > 0 & \forall \delta \geq 0 & \exists \delta \geq 0 & \end{array}$$

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

Write out the proper sequence $Q_1 Q_2 \ldots$

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/

Supplementary Problems

Problem 5.

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.

Problem 6.

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.

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

$$A_n ::= \frac{X_1 + X_2 + \cdots + X_n}{n},$$
$$\mu_n ::= \text{Ex}[A_n].$$

Then for every $\epsilon > 0$,

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

(a) Prove the Generalized Pairwise Independent Sampling Theorem.

(b) Conclude that the following holds:

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

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