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 number of adult Americans who believe in evolution is 35 ± 3 percent?

Problem 2.
Yesterday, the programmers at a local company wrote a large program. To estimate the fraction, b, of lines of code in this program that are buggy, the QA team will take a small sample of lines chosen randomly and independently (so it is possible, though unlikely, that the same line of code might be chosen more than once). For each line chosen, they can run tests that determine whether that line of code is buggy, after which they will use the fraction of buggy lines in their sample as their estimate of the fraction b.

The company statistician can use estimates of a binomial distribution to calculate a value, s, for a number of lines of code to sample which ensures that with 97% confidence, the fraction of buggy lines in the sample will be within 0.006 of the actual fraction, b, of buggy lines in the program.

Mathematically, the *program* is an actual outcome that already happened. The *sample* is a random variable defined by the process for randomly choosing s lines from the program. The justification for the statistician's confidence depends on some properties of the program and how the sample of s lines of code from the program are chosen. These properties are described in some of the statements below. Indicate which of these statements are true, and explain your answers.

- The probability that the ninth line of code in the *program* is buggy is b.
- The probability that the ninth line of code chosen for the *sample* is defective, is b.
- All lines of code in the program are equally likely to be the third line chosen in the *sample*.
- Given that the first line chosen for the *sample* is buggy, the probability that the second line chosen will also be buggy is greater than b.
- Given that the last line in the *program* is buggy, the probability that the next-to-last line in the program will also be buggy is greater than b.
- The expectation of the indicator variable for the last line in the *sample* being buggy is b.
- Given that the first two lines of code selected in the *sample* are the same kind of statement —they might both be assignment statements, or both be conditional statements, or both loop statements,. . . —the probability that the first line is buggy may be greater than b.
- There is zero probability that all the lines in the *sample* will be different.

Problem 3.
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 4.
An *International Journal of Epidemiology* has a policy that they will only publish the results of a drug trial when there were enough patients in the drug trial to be sure that the conclusions about

the drug's effectiveness hold at the 95% confidence level. 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 astonished and embarrassed to learn that *every one* of the 20 drug trial results they published during the year was wrong. This happened even though the editors and reviewers had carefully checked the submitted data, and every one of the trials was *properly performed and reported* in the published paper.

The editors thought the probability of this was negligible (namely, $(1/20)^{20} < 10^{-25}$). Explain what's wrong with their reasoning and how it could be that all 20 published studies were wrong.

Appendix

Variance

The *variance*, $\text{Var}[R]$, of a random variable, R, is:

$$\text{Var}[R] ::= \text{E}\left[(R - \text{E}[R])^2\right].$$

It is easy to show that

$$\text{Var}[R] = \text{E}\left[R^2\right] - \text{E}^2[R].$$

[Variance of an index variable], I, with $\Pr\{I = 1\} = p$:

$$\text{Var}[I] = pq$$

where $q ::= 1 - p$.

[Variance and constants] For constants, a, b,

$$\text{Var}[aR + b] = a^2 \text{Var}[R].$$

[Variance Additivity] If $R_1, R_2, \ldots, R_n$ are *pairwise* independent variables, then

$$\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{Var}[R_1] + \text{Var}[R_2] + \cdots + \text{Var}[R_n]$$

Chebyshev' s Bound

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

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

Pairwise Independent Sampling

Theorem. *Let*

$$A_n ::= \frac{\sum_{i=1}^{n} R_i}{n}$$

where $R_1, \ldots, R_n$ are pairwise independent random variables with the same mean, μ, and deviation, σ. Then

$$\Pr\{|A_n - \mu| > x\} \leq \left(\frac{\sigma}{x}\right)^2 \cdot \frac{1}{n}.$$

6.042J / 18.062J Mathematics for Computer Science
Spring 2010