Massachusetts Institute of Technology
6.042J/18.062J, Spring '10: Mathematics for Computer Science
Prof. Albert R. Meyer

May 10
revised May 10, 2010, 677 minutes

Solutions to 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?

Solution.

$$\frac{1}{4}$$

By Lemma 21.4.2, $\text{Var}\left[H\right] = pq$.

Noting that $d\,p(1-p)/dp = 2p - 1$ is zero when $p = 1/2$, it follows that the maximum value of $p(1-p)$ must be at $p = 1/2$, so the maximum value of $\text{Var}\left[H\right]$ is $(1/2)(1 - (1/2)) = 1/4$. ■

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

Solution. 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. ■

(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.)

Solution. 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

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

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. ■

(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?

Solution. 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. ■

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.

1. The probability that the ninth line of code in the *program* is buggy is b.

Solution. False.

The program has already been written, so there's nothing probabilistic about the buggyness of the ninth (or any other) line of the program: either it is or it isn't buggy, though we don't know which. You could argue that this means it is buggy with probability zero or one, but in any case, it certainly isn't b. ■

2. The probability that the ninth line of code chosen for the *sample* is defective, is b.

Solution. True.

The ninth line sampled is equally likely to be any line of the program, so the probability it is buggy is the same as the fraction, b, of buggy lines in the program. ■

3. All lines of code in the program are equally likely to be the third line chosen in the *sample*.

Solution. True.

The meaning of "random choices of lines from the program" is precisely that at each of the s choices in the sample, in particular at the third choice, each line in the program is equally likely to be chosen. ■

4. 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.

Solution. False.

The meaning of "*independent* random choices of lines from the program" is precisely that at each of the s choices in the sample, in particular at the second choice, each line in the program is equally likely to be chosen, independent of what the first or any other choice happened to be. ■

5. 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.

Solution. False.

As noted above, it's zero or one. ■

6. The expectation of the indicator variable for the last line in the *sample* being buggy is b.

Solution. True.

The expectation of the indicator variable is the same as the probability that it is 1, namely, it is the probability that the sth line chosen is buggy, which is b, by the reasoning above. ■

7. 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.

Solution. True.

We don't know how prone to bugginess different kinds of statements may be. It could be for example, that conditionals are more prone to bugginess than other kinds of statements, and that there are more conditional lines than any other kind of line in the program. Then given that two randomly chosen lines in the sample are the same kind, they are more likely to be conditionals, which makes them more prone to bugginess. That is, the conditional probability that they will be buggy would be greater than b. ■

8. There is zero probability that all the lines in the *sample* will be different.

Solution. False.

We know the length, r, of the program is larger than the "small" sample size, s, in which case the probability that all the lines in the sample are different is

$$\frac{r}{r} \cdot \frac{r-1}{r} \cdot \frac{r-2}{r} \cdots \frac{r-(s-1)}{r} = \frac{r!}{(r-s)!\, r^s} > 0.$$

Of course it would be true by the Pigeonhole Principle if $s > r$. ■

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.

Solution. 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.

■

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.

Solution. 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 –period. 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. ■