Staff Solutions to In-Class Problems Week 14, Mon.

STAFF NOTE: Sampling & Law of Large Numbers Ch. 20.4–5

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 $\frac{675}{1928} \approx 0.350$. 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?

COMMENTS:

- CP.gallup.poll
- from: F05 cp14f, S08 cp13f
- revised a lot by ARM 12/5/09
- edited 5/11/14

keywords = [pairwise_independent sampling variance estimate confidence probability uniform independent]

Solution.

\[
\frac{1}{4}
\]

Let $H$ be an indicator variable with probability $p$ being 1 and probability $q$, which is $(1 - p)$, being 0. By Lemma 20.3.2, $\text{Var}[H] = pq = p(1 - p)$.

STAFF NOTE: Here’s a proof starting from the definition of variance:

\[
\begin{align*}
\text{Var}[H] &= \text{Ex}[(H - p)^2] \\
&= \text{Ex}[H^2 - 2pH + p^2] \\
&= \text{Ex}[H^2] - 2p \text{Ex}[H] + p^2 \\
&= p - p^2 \\
&= pq
\end{align*}
\]

Noting that the derivative of $p(1 - p)$ with respect to $p$ is $2p - 1$. This derivative is zero when $p = 1/2$; it follows that $p(1 - p)$ is maximal when $p = 1/2$, and so the maximum value of $\text{Var}[H]$ 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 Theorem, the probability that a sample of size $n = 1928$ is further than $x = 0.03$ of the actual fraction is at most

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

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

(e) 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. The variance $\sigma^2$ of a single sample might really be as bad as 1/4 (this happens when $p = 1/2$) and the variance of $n$ samples might be as bad as $n\sigma^2$, so there is no better mileage to be gotten out of the Chebyshev bound.

Instead, Gallup could use the fact that the sample has a binomial distribution $B_{1928,p}$, where $p$ is the unknown quantity to be estimated. The tails of a binomial distribution decrease much more rapidly than arbitrary distributions (see Section 19.3.4), so confidence degrees calculated using this distribution will be higher than calculations based solely on variance.

So Gallup 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 binomial random variable differs from its mean will be largest when $p = 1/2$. This is in fact the case. So Gallup will calculate

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

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 percentage 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.
Let \( G_1, G_2, G_3, \ldots \) be an infinite sequence of pairwise independent random variables with the same expectation \( \mu \) 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:

\[
Q_0 \ Q_1 \ Q_2 \ Q_3 \ f(n, \epsilon) \geq 1 - \delta
\]

where \( Q_0, Q_1, Q_2, Q_3 \) is a sequence of four quantifiers from among:

\[
\forall n, \ \exists n, \ \forall n \geq n_0, \ \exists n \geq n_0.
\]

\[
\forall n_0, \ \exists n_0, \ \forall n_0 \geq n, \ \exists n_0 \geq n.
\]

\[
\forall \delta, \ \exists \delta, \ \forall \delta > 0, \ \exists \delta > 0.
\]

\[
\forall \epsilon, \ \exists \epsilon, \ \forall \epsilon > 0, \ \exists \epsilon > 0.
\]

Here the \( n, n_0 \) range over nonnegative integers, and \( \delta, \epsilon \) range over nonnegative real numbers.

Write out the proper sequence \( Q_0, Q_1, Q_2, Q_3 \).

**COMMENTS:**

- CP_large_numbers_quantifiers
- same as FP_large_numbers_quantifiers but here epsilon is NOT given
- from S11,F03,S02.final

**keywords = [ quantifiers law_of_large_numbers ]

Solution.

\[
\forall \epsilon > 0 \ \forall \delta > 0 \ \exists n_0 \ \forall n \geq n_0.
\]

Problem 3.
The proof of the Pairwise Independent Sampling Theorem 20.4.1 was given for a sequence \( R_1, R_2, \ldots \) of pairwise independent random variables with the same mean and variance.

The theorem generalizes straightforwardly 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 \( \epsilon > 0 \),

\[
\Pr[|A_n - \mu_n| \geq \epsilon] \leq \frac{b}{\epsilon^2} \cdot \frac{1}{n}.
\] (1)
(a) Prove the Generalized Pairwise Independent Sampling Theorem.

**COMMENTS:**
- CP.pairwise_independent_theorem
- from: S09.cp14t, F07.rec15t
- revised 5/9/14

**keywords** = [ pairwise_independent sampling variance law of large numbers ]

**Solution.** Essentially identical to the proof of Theorem 20.4.1 in the text, except that $S_n/n$ gets replaced by $A_n$, the $x$ gets replaced by $\epsilon$, the constant $b$ replaces $\text{Var}[G_i]$, and the equality before the first occurrence of $b$ gets replaced by an inequality ($\leq$).

**STAFF NOTE:** It is OK for students to refer to the proof in the text and simply comment on where changes must be made. It is also OK to have students write down the revised proof.

**Proof.** We first observe that

$$\text{Var}[A_n] \leq \frac{b}{n}, \quad (2)$$

because

$$\begin{align*}
\text{Var}[A_n] &= \frac{1}{n^2} \text{Var} \left[ \sum_{i=1}^{n} X_i \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i] \quad \text{(pairwise independent additivity)} \\
&\leq \frac{1}{n^2} \cdot nb \\
&= \frac{b}{n}.
\end{align*}$$

Now, apply Chebyshev’s Theorem and conclude:

$$\begin{align*}
\text{Pr} \left[ |A_n - \mu_n| \geq \epsilon \right] &\leq \frac{\text{Var}[A_n]}{\epsilon^2} \\
&\leq \frac{b/n}{\epsilon^2} \quad \text{(by (2))} \\
&= \frac{b}{\epsilon^2 \cdot n}.
\end{align*}$$

(b) Conclude that the following holds:

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

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

\[
\Pr[|A_n - \mu_n| \leq \epsilon] = 1 - \Pr[|A_n - \mu_n| > \epsilon] \\
\geq 1 - \frac{b}{\epsilon^2} \cdot \frac{1}{n}, \quad \text{(by (1))}
\]

and for any fixed \(\epsilon\), this last bound approaches 1 as \(n\) approaches infinity. ■

Supplementary Problems

Problem 4.
Let \(R, S\) and \(T\) be mutually independent indicator variables.

In general, the event that \(S = T\) is not independent of \(R = S\). We can explain this intuitively as follows: suppose for simplicity that \(S\) is uniform, that is, equally likely to be 0 or 1. This implies that \(S\) is equally likely as not to equal \(R\), that is \(\Pr[R = S] = 1/2\); likewise, \(\Pr[S = T] = 1/2\).

Now suppose further that both \(R\) and \(T\) are more likely to equal 1 than to equal 0. This implies that \(R = S\) makes it more likely than not that \(S = 1\), and knowing that \(S = 1\), makes it more likely than not that \(S = T\). So knowing that \(R = S\) makes it more likely than not that \(S = T\), that is, \(\Pr[S = T \mid R = S] > 1/2\).

Now prove rigorously (without any appeal to intuition) that the events \([R = S]\) and \([S = T]\) are independent iff either \(R\) is uniform\(^1\), or \(T\) is uniform, or \(S\) is constant\(^2\).

COMMENTS:

- PS\_dependent_pairs
- sequel to PS\_equal_birthdays
- by ARM 4/28/11; corrected with complete soln 8/12/12
- revised to stand alone, ARM 11/21/13

keywords = [ random_variable independence mutual pairwise uniform distribution ]

Solution. Let \(r, s, t\) be the probabilities that \(R = 1, S = 1, T = 1\), respectively. Then

\[
\Pr[R = S] = rs + (1 - r)(1 - s) \\
\Pr[S = T] = st + (1 - s)(1 - t) \\
\Pr[R = S \text{ AND } S = T] = rst + (1 - r)(1 - s)(1 - t).
\]

So \([R = S]\) and \([S = T]\) are independent iff

\[
[rs + (1 - r)(1 - s)][st + (1 - s)(1 - t)] = rst + (1 - r)(1 - s)(1 - t). \quad (3)
\]

Subtracting the left from the right-hand side of this equation and factoring, we find that (3) holds iff

\[s(s - 1)(2r - 1)(2t - 1) = 0,\]

namely, iff \(s = 0\) or \(s = 1\) or \(r = 1/2\) or \(t = 1/2\). That is, independence holds iff \(S\) is constant or \(R\) or \(T\) is uniform. ■

\(^1\)That is, \(\Pr[R = 1] = 1/2\).

\(^2\)That is, \(\Pr[S = 1]\) is one or zero.
Problem 5.
A defendant 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 defendant 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 defendant. 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.

COMMENTS:

- CP_size_of_sample_vs_population
- from: S06.cp12f
- revised by ARM 10/6/09

keywords = [ sample confidence estimation probability ]

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.

De Veaux and Velleman (REF NEEDED) suggest thinking about sampling a well-stirred pot of soup.

>You can sip a teaspoon to see if it's too salty. Whether you have a small pot to serve two, or a large vat to serve a banquet, all you need sip is a teaspoon. You don’t need need to sip a whole ladle or a cup just because your pot of soup is large. The same applies to sampling from a population: we sample the same “teaspoon” amount no matter how big the population.

Here is an earlier explanation we initially proposed (without a lot of confidence) to help a jury understand that sampling cars is like rolling dice:

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.

We will explain why the number trips that need be sampled to estimate this fraction accurately does not depend on how many trips were recorded, it just depends on the fraction of speeders.

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 of 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 outcomes is double six. Mathematical theory lets us calculate 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 speeders.