

Solutions to In-Class Problems — Week 14, Mon

Problem 1. The Weak Law of Large Numbers in [Notes 13-14](#) was given for a sequence G_1, G_2, \dots of pairwise independent random variables with the same mean and variance. We can generalize the Law to sequences of pairwise independent random variables, possibly with *different* means and variances, as long as their variances are bounded by some constant.

Theorem (Generalized Pairwise Independent Sampling). Let X_1, X_2, \dots 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

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

Then for every $\epsilon > 0$,

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

(a) Prove the Generalized Pairwise Independent Sampling Theorem. *Hint:* The proof of the Pairwise Independent Sampling Theorem from the Notes is repeated in the Appendix.

Solution. Essentially identical to the proof attached, except that $\text{Var}[G_i]$ gets replaced by b , and the equality becomes \leq where the b is first used. ■

(b) Conclude

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

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

Solution.

$$\begin{aligned} \Pr\{|A_n - \mu_n| > \epsilon\} &= 1 - \Pr\{|A_n - \mu_n| \leq \epsilon\} \\ &\geq 1 - b/(n\epsilon^2) \end{aligned} \tag{by (1)},$$

and for any fixed ϵ , this last term approaches 1 as n approaches infinity. ■

Problem 2. Attack of the Clones

We want to determine the percentage, r , of people who intend to watch the *Episode II: Attack of the Clones*. Since we are still working on our jedi mind tricks, we decide to conduct a poll.

Suppose that we poll 800 people. Use the Pairwise Sampling Theorem to bound the confidence level that our estimated value is within ± 0.04 of the actual value r .

Solution. The Pairwise Sampling Theorem gives us a bound on the probability that the average of a sum of independent, identically distributed random variables differs from their mean by more than a certain amount ϵ . Call this probability δ . Our confidence level, $c = 1 - \delta$. From the theorem we find,

$$\Pr \left\{ \left| \frac{S_n}{n} - r \right| \geq \epsilon \right\} \leq \frac{\sigma^2}{n\epsilon^2}$$

Recall that σ^2 is the variance of the random variables we are summing. In this case, these are Bernoulli random variables, so the variance is pq , where $q := 1 - p$. Unfortunately, p is also the value we are trying to compute! However, we can take advantage of the fact that the variance is maximized when $p = 1/2$. This gives us the following bound on δ .

$$\begin{aligned} \delta &\leq \frac{p(p-1)}{n\epsilon^2} \\ &\leq \frac{0.5(0.5-1)}{n\epsilon^2} \\ &\leq \frac{1}{4n\epsilon^2}. \end{aligned}$$

Plugging in 800 for n , and 0.04 for ϵ we find $\delta \leq 0.195$. Our confidence level is $1 - \delta$, so $c \geq 1 - 0.195$, or 80%. ■

Problem 3. Explaining Sampling to a Jury

You just calculated the confidence level, based on a poll of 800 people, of an estimate being within ± 0.04 of the fraction of people in the US who intend to watch *Star Wars Episode II: The Attack of the Clones*. The lecture showed that to achieve a 95% confidence level for such an estimate, a poll of 3125 randomly selected people is sufficient.

Suppose you were going to serve as an expert witness in a trial. How would you explain to a jury why it is reasonable to model people as independent coin tosses? How would you explain why the number of people necessary to poll *does not depend on the population size*?

Solution. This was intended to be a thought-provoking, conceptual question, and it was. In fact, it became apparent from the extended discussions at nearly all tables that, although most of the class had cranked through the formulas for poll size and confidence levels last week, they couldn't articulate, and indeed didn't really understand the answers to the questions raised.

Here's a way to explain why we model polling people about Star Wars as independent coin tosses that a jury might be able to follow:

Of the approximately 250,000,000 people in the US, there are some unknown number, say 100,000,000, who plan to see Star Wars. So in this case, the *fraction* of people planning to see Star Wars would be $100,000,000/250,000,000 = 0.4$.

To estimate this unknown fraction, we randomly select one person from the 250,000,000 in such a way that *everyone has an equal chance of being picked*. For example, we might get computer files from the Census Bureau listing all 250,000,000 people in the US. Then we would generate a number between 1 and 250,000,000 by some physical or computational process that generated each number with equal probability, and then we would interview the person whose number came up. In this way, we can be sure that the probability that a person we select will be planning to see the movie is exactly the unknown fraction who plan to see the movie.

After we have picked a person and learned their movie plans, we perform the procedure again, making sure that everyone is equally likely to be picked the second time, and so on, for picking a third, fourth, *etc.* person. On each pick the probability of getting a person who plans to see the movie is the same fraction, so we describe picking a person in this way as “flipping a coin” that has this fraction as probability of coming up “Heads”—meaning that the person selected plans to see the movie.

Now we all understand that if we keep flipping a coin with a certain probability of coming up Heads, then the more we flip, the closer the fraction of Heads flipped will be to that probability. Mathematical theory lets us calculate us how many time to flip coins to make the fraction of Heads very likely close to the right fraction, but we won’t go into those details.

It’s also clear, that if two different coins have the same probability of coming up Heads, it makes no difference in our experiments which coin we use: the number of flips we need for the fraction of Heads flipped very likely to be close to the probability of a Head will be the same for either coin. So whether the coin had probability of heads, say, 0.4, because 100,000,000 out of 250,000,000 people intended to see Star Wars, or 10,000 out of 25,000, or 2 out of 5—the same number of flips will allow us to estimate the probability of Heads, and hence to estimate the fraction of the population aiming to see the movie. So the *number of people we need to poll is the same*, whether we are selecting from a large population or a small population, as long as the fraction planning to see the movie is the same in the small population as in the large one.



NOTE: We didn’t get to either of the following two problems.

Problem 4. Let X, Y be independent Binomial random variables with parameters (n, p) and (m, p) , respectively.

(a) What is $\Pr\{X + Y = k\}$?

Solution. The pdf of X is the probability of tossing k Heads out of n independent flips of a coin with bias p . Likewise for Y and m flips. Since, X and Y are independent, the pdf of $X + Y$ corresponds to $n + m$ independent flips, *i.e.*, $X + Y$ is a Binomial variable with parameters $(n + m, p)$. Hence,

$$\binom{m+n}{k} p^k (1-p)^{m+n-k}.$$

■

(b) Prove that if $p = 1/2$ and n is even, then

$$\Pr \left\{ X = \frac{n}{2} \right\} \sim \sqrt{\frac{2}{n\pi}}. \quad (2)$$

Hint: Use Stirling's approximation (in the appendix).

Solution. The probability of m heads is $\binom{n}{m}/2^n$.

Using Stirling for an approximation:

$$\begin{aligned} \binom{n}{m} &\sim \frac{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n}{\sqrt{2\pi m} \left(\frac{m}{e}\right)^m \sqrt{2\pi(n-m)} \left(\frac{n-m}{e}\right)^{n-m}} \\ &= \frac{\sqrt{n} \left(\frac{n}{e}\right)^m \left(\frac{n}{e}\right)^{n-m}}{\sqrt{2\pi m(n-m)} \left(\frac{m}{e}\right)^m \left(\frac{n-m}{e}\right)^{n-m}} \\ &= \sqrt{\frac{n}{2\pi m(n-m)}} \left(\frac{n}{m}\right)^m \left(\frac{n}{n-m}\right)^{n-m}. \end{aligned} \quad (3)$$

Setting $m = n/2$ in (3) yields

$$\binom{n}{n/2} \sim \frac{2^n}{\sqrt{n\pi/2}}, \quad (4)$$

so

$$\Pr \left\{ X = \frac{n}{2} \right\} = \frac{\binom{n}{\lceil n/2 \rceil}}{2^n} \sim \sqrt{\frac{2}{n\pi}}, \quad (5)$$

proving (2).

If n is odd, the same asymptotic bound also holds (with $n/2$ replaced by $\lceil n/2 \rceil$). ■

(c) Estimate the probability that the number of heads in 400 flips of a fair coin will be between 195 and 205, and likewise that in 10,000 flips it will be between 4980 and 5020.

Also, discuss writing a program to calculate the exact answer.

Solution. The numbers are all too large to calculate exactly by hand. Computing $400!$, $190!$, and $210!$ exactly will overflow on any calculator, so an exact computation by calculator would also be impractical. In fact, it will cause numerical overflow errors in most programming languages. Real Scheme has infinite precision rational arithmetic and can handle $400!$, but typically overflows when computing $10,000!$, so exact computation in the 10,000 case is unlikely to work in any language—remember the answer will have around 40,000 digits¹. But there are no applications that require anywhere near such accuracy.

So we aim for an approximate solution using Stirling's approximation for $n!$. But there are still pitfalls: $(400/e)^{400}$ will overflow floating point arithmetic on calculators and essentially all computers, so the simplified version (3) would be needed; even using this version, the calculation of the powers of n/m and $n/(n-m)$ must be properly interleaved with divisions by powers of 2 to avoid overflow. But since floating point is usually good for 8 to 10 places on most machines, getting six place accuracy should be manageable.

A calculator will yield the value of the righthand side of equation (2) at $n = 400$; it is ≈ 0.0399 . But all the probabilities for 195 to 205 heads are about the same, so an offhand estimate would be $11 \times 0.0399 \approx 0.439$. The actual answer is 0.418.

Similarly, for $n = 10,000$, the value of (2) is about 0.00798, and all the probabilities for 4980 to 5020 heads are about the same, so an offhand estimate would be $41 \times 0.00798 \approx 0.327$. The actual answer is 0.318. ■

¹Though on machines with enough RAM and a gigahertz+ processor, Scheme actually can handle exact calculations with 40,000 digit integers.