

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 $\epsilon > 0$,

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

(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.

(b) Conclude

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

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

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 .

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*?

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 $\Pr \{X + Y = k\}$?
- (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).

(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.

A Appendix

Lemma (Stirling's Approximation).

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

The *probability density function (pdf)* for a random variable, R , is the function $f_R : \text{range}(R) \rightarrow [0, 1]$ defined by:

$$f_R(x) ::= \Pr \{R = x\}.$$

Random variables R_1, R_2, \dots are *mutually independent* iff

$$\Pr \left\{ \bigcap_i [R_i = x_i] \right\} = \prod_i \Pr \{R_i = x_i\},$$

for all $x_1, x_2, \dots \in \mathbb{R}$. They are *k-wise independent* iff $\{R_i \mid i \in J\}$ are mutually independent for all subsets $J \subset \mathbb{N}$ with $|J| = k$.

The Binomial Distribution with parameters n, p is the distribution of the sum $B_{n,p} ::= \sum_{i=1}^n X_i$ where X_1, \dots, X_n are mutually independent Bernoulli (indicator) variables with $\Pr\{X_i = 1\} = p$. It has pdf

$$f_{n,p}(k) ::= \binom{n}{k} p^k q^{n-k}$$

where $q ::= 1 - p$.

$$\begin{aligned} \mathbb{E}[B_{n,p}] &= np, \\ \text{Var}[B_{n,p}] &= npq. \end{aligned}$$

Theorem (Expectation of a Product). If R_1, R_2, \dots, R_n are mutually independent, then

$$\mathbb{E}[R_1 \cdot R_2 \cdots R_n] = \mathbb{E}[R_1] \cdot \mathbb{E}[R_2] \cdots \mathbb{E}[R_n].$$

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

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

Variance can also be equivalently defined as:

$$\text{Var}[R] ::= \mathbb{E}[R^2] - \mathbb{E}^2[R],$$

Lemma. For $a, b \in \mathbb{R}$,

$$\text{Var}[aR + b] = a^2 \text{Var}[R] \tag{3}$$

Theorem 4.1. If R_1, R_2, \dots, R_n are pairwise independent random variables, then

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

Theorem (Markov's Theorem). If R is a nonnegative random variable, then for all $x > 0$

$$\Pr\{R \geq x\} \leq \frac{\mathbb{E}[R]}{x}.$$

An alternative formulation is

$$\Pr\{R \geq x \mathbb{E}[R]\} \leq \frac{1}{x}.$$

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

$$\Pr\{|R - \mathbb{E}[R]| \geq x\} \leq \frac{\text{Var}[R]}{x^2}. \tag{4}$$

An alternative formulation is

$$\Pr\{|R - \mathbb{E}[R]| \geq x\sigma_R\} \leq \frac{1}{x^2},$$

where $\sigma_R ::= \sqrt{\text{Var}[R]}$ is the standard deviation of R .

Theorem (Pairwise Independent Sampling). *Let*

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

where G_1, \dots, G_n are pairwise independent random variables with the same mean, μ , and deviation, σ . Then

$$\Pr \{|A_n - \mu| > x\} \leq \left(\frac{\sigma}{x}\right)^2 \cdot \frac{1}{n}. \quad (5)$$

Proof. By linearity of expectation,

$$\mathbb{E}[A_n] = \frac{\mathbb{E}[\sum_{i=1}^n G_i]}{n} = \frac{\sum_{i=1}^n \mathbb{E}[G_i]}{n} = \frac{n\mu}{n} = \mu.$$

Since the G_i 's are pairwise independent, their variances will also add, so

$$\begin{aligned} \text{Var}[A_n] &= \left(\frac{1}{n}\right)^2 \text{Var}\left[\sum_{i=1}^n G_i\right] && \text{(by (3))} \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^n \text{Var}[G_i] && \text{(by Theorem 4.1)} \\ &= \left(\frac{1}{n}\right)^2 n\sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Now letting R be A_n in Chebyshev's Bound (4) yields (5), as required.

□