CS109 Lecture 13 Worksheet: Multinomial

The Multinomial Coefficient, Multinomial RVs, and Probabilistic Text Analysis

Summer 2026

Lecture 13 Worksheet: Multinomial

Topics: The multinomial coefficient (counting orderings with repeated types), the multinomial random variable and its joint PMF, marginal counts, modeling text as a multinomial ("bag of words"), Bayesian authorship/spam classification, and computing with log probabilities.

How to use this sheet. Work in small groups. A TA and the instructor will circulate to help. We will go over selected solutions on the board. You do not need to finish every part; aim for understanding over speed. Keep fractions exact where you can.

Problem 1 (Review of Lecture 12): A Two-Node Network

A monitoring system models a web server with a tiny Bayesian network: Overloaded $\to$ Slow, where $P(\text{Overloaded} = 1) = 0.1$, $P(\text{Slow} = 1 \mid \text{Overloaded} = 1) = 0.95$, and $P(\text{Slow} = 1 \mid \text{Overloaded} = 0) = 0.2$.

(a) Compute the joint probability $P(\text{Overloaded} = 1, \text{Slow} = 0)$.

(b) Compute $P(\text{Overloaded} = 1 \mid \text{Slow} = 1)$.

(c) If you instead estimated (b) by rejection sampling, which samples would you throw away?

Problem 2: Counting Orderings of a DNA Strand

A DNA fragment is a sequence of 12 nucleotides containing exactly 4 A's, 3 C's, 2 G's, and 3 T's.

(a) How many distinct sequences (orderings) are possible? Write your answer using the multinomial coefficient notation $\binom{12}{4,3,2,3}$, then evaluate it.

(b) A random sequence generator picks each of the 12 positions independently and uniformly from $\{A, C, G, T\}$. What is the probability it produces *exactly* 4 A's, 3 C's, 2 G's, and 3 T's (in any order)?

Problem 3: Load Balancing

A load balancer routes each incoming request independently to data center 1 with probability 0.5, data center 2 with probability 0.3, and data center 3 with probability 0.2. Ten requests arrive. Let X_1, X_2, X_3 count the requests routed to each data center.

(a) What is the joint PMF $P(X_1 = c_1, X_2 = c_2, X_3 = c_3)$? State the constraint the counts must satisfy.

(b) Compute $P(X_1 = 5, X_2 = 3, X_3 = 2)$.

Problem 4: Loot Boxes

In a video game, each loot box independently contains a *common* item with probability 0.6, a *rare* item with probability 0.3, or a *legendary* item with probability 0.1. You open 8 loot boxes.

(a) What is the probability you get exactly 4 commons, 3 rares, and 1 legendary?

(b) Now consider only the count of legendary items, ignoring the other two types. What is the distribution of this single count, and what is the probability of exactly 1 legendary? (Hint: each count in a multinomial, viewed on its own, is a familiar RV.)

Problem 5: Midterm Rooms (problem from the Winter 2025 Final)

For the CS109 Midterm, there are 300 students who are sent to one of three rooms based on their last name. Each student, independently, has a 30% chance of going to room 1, a 20% chance of going to room 2, and a 50% chance of going to room 3. What is the chance that exactly 90 students go to room 1, 60 students go to room 2, and 150 go to room 3? (An unevaluated expression is fine.)

Problem 6: Who Wrote the Pamphlet?

An anonymous pamphlet is attributed to one of two authors, H or M, with equal prior probability. You model the pamphlet as a multinomial over words. In the pamphlet, the word "upon" appears 2 times, "whilst" appears 1 time, and "commerce" appears 3 times. From each author's known writings you estimate:

word	$P(\text{word} \mid H)$	$P(\text{word} \mid M)$
upon	0.005	0.0005
whilst	0.0001	0.001
commerce	0.002	0.003

(a) Using Bayes' theorem, write an expression proportional to $P(H \mid \text{pamphlet})$ using the multinomial model. Why can you ignore the multinomial coefficient (and the words not shown) when *comparing* the two authors?

(b) Real documents have thousands of words, and each word probability is tiny. What goes wrong numerically when you multiply thousands of tiny probabilities, and how do logs fix it? State the log identity you rely on.

(c) Compute the log-scores $\sum_i n_i \ln P(\text{word}_i \mid \text{author})$ for both authors (a calculator is fine) and decide who more likely wrote the pamphlet. Roughly how many times more likely is that author?

Challenge Problem (optional): Multinomials Generalize Binomials

Let $(X_1, \ldots, X_r) \sim \text{Multinomial}(n; p_1, \ldots, p_r)$.

(a) Show that when $r = 2$, the joint PMF reduces exactly to the Binomial PMF for X_1.

(b) Argue (without heavy algebra) that any single count $X_i \sim \text{Bin}(n, p_i)$. Hint: recolor the outcomes as "type i" versus "everything else."