AIM Qualifying Review Exam: Probability and Discrete Mathematics

January 6, 2025

There are five (5) problems in this examination.

There should be sufficient room in this booklet for all your work. But if you use other sheets of paper, be sure to mark them clearly and staple them to the booklet.

Problem 1

Find, as function of n, the sum

$$1 + \frac{1}{2}\binom{n}{1} + \frac{1}{3}\binom{n}{2} + \cdots + \frac{1}{n+1}\binom{n}{n}.$$

Hint: try integrating a series for $(1+x)^n$.

Solution outline

Brualdi exercise 5.7.16, p. 156.

For $n \geq 0$, we have $(1+x)^n = 1 + \binom{n}{1}x + \binom{n}{2}x^2 + \cdots \binom{n}{n}x^n$. Integrating, we get $x + \frac{1}{2}\binom{n}{1}x^2 + \frac{1}{3}\binom{n}{2}x^3 + \cdots + \frac{1}{n+1}\binom{n}{n}x^{n+1}$. Our series S_n is the result of substituting $x = 1$. Thus, the sum of the series is $\int_{x=0}^{1}(1+x)^n\,dx = x + \frac{n}{2}x^2 \cdots \Big|_0^1$, since the antiderivative vanishes at the lower bound of integration, $x = 0$. Putting $u = x + 1$ so $du = dx$, we get $\int_{u=1}^{2} u^n\,du = \frac{1}{n+1}u^{n+1}\Big|_1^2 + 1 = \frac{2^{n+1}-1}{n+1}$. This can be checked directly for small n:

n	S_n
0	1
1	3/2

Mathematical concepts: binomial coefficients, generating functions

Problem 2

An r-combination of a multiset M is an unordered collection of r items in M.

(a) Let T^* be the multiset $\{\infty \cdot a, \infty \cdot b, \infty \cdot c\}$ (infinitely-many a's, b's, and c's). Determine the the number of 10-combinations of T^*. (So, for example, $\overbrace{aaaaaaaaa}^{9}b$ and $b\overbrace{aaaaaaaaa}^{9}$ are different names for the same valid 10-combination, since they consist of a, b, c but have the same number of each letter and differ only in the order.)

(b) Determine the number of 10-combinations in the (size 12) multiset $T = \{3 \cdot a, 4 \cdot b, 5 \cdot c\}$, consisting of three a's, four b's, and five c's.

(c) Determine the number of 10-combinations in $T = \{3 \cdot a, 4 \cdot b, 8 \cdot c\}$, consisting of three a's, four b's, and eight c's.

Solution outline

Brualdi example, page 169.

(a) The "balls and bars" formulation asks for the number of ways to arrange 10 balls and 2 bars in 12 slots. For example, $\bullet\bullet\,|\,\bullet\bullet\bullet\bullet\bullet\bullet\bullet\,|\,\bullet$ corresponds to 2 a's, 7 b's, and 1 c. That number is $\binom{10+3-1}{3-1} = 66$, since the number, 2, of bars is one less than the number, 3, of types of letters, i.e. groups of balls.

(b) Let T^* be the multiset $\{\infty \cdot a, \infty \cdot b, \infty \cdot c\}$. Let A be the set of 10-combinations of T^* with more than 3 a's, let B be the set of 10-combinations of T^* with more than 4 b's, and let C be the set of 10-combinations of T^* with more than 5 c's. Let S be the set of all 10-combinations of T^*.

Using the Inclusion-Exclusion principle, we want

$$\begin{aligned} |\overline{A} \cap \overline{B} \cap \overline{C}| &= |S| \\ &\quad -(|A| + |B| + |C|) \\ &\quad +(|A \cap B| + |B \cap C| + |C \cap A|) \\ &\quad -|A \cap B \cap C|. \end{aligned}$$

Above we computed $|S| = 66$. Then A counts 10-combinations having 4 a's plus any 6-combination (i.e., $10-(3+1)$)-combination, reading 10 and 3 from the given information of T^*, etc., so $|A| + |B| + |C| = \binom{10-(3+1)+3-1}{3-1} + \binom{10-(4+1)+3-1}{3-1} + \binom{10-(5+1)+3-1}{3-1} = 28 + 21 + 15 = 64$. And $A \cap B$ counts 4 a's and 5 b's and any 1-combination, etc., so $|A \cap B| + |A \cap C| + |B \cap C| = \binom{10-(3+1+4+1)+3-1}{3-1} + \binom{10-(3+1+5+1)+3-1}{3-1} + \binom{10-(4+1+5+1)+3-1}{3-1} = 3 + 1 + 0$, and $|A \cap B \cap C| \leq |B \cap C| = 0$. We get $66 - 64 + 4 - 0 = 6$.

Alternatively, from the $12 = 3 + 4 + 5$ items, toss two, in $\binom{2+3-1}{3-1} = 6$ ways: aa, ab, ac, bb, bc, cc. That is, the number of 10-combinations is the same as the number of 2-combinations from the $10 + 2 = 12$ element multiset, and each type of item numbers at least 2. So the Inclusion-Exclusion principle is not required (and, arguably, not simplest).

(c) Similar to the above, we want

$$\begin{aligned} \binom{10+3-1}{3-1} &- \left[\binom{10-(3+1)+3-1}{3-1} + \binom{10-(4+1)+3-1}{3-1} + \binom{10-(8+1)+3-1}{3-1}\right] \\ &+ \binom{10-(3+1+4+1)+3-1}{3-1} + \binom{10-(3+1+8+1)+3-1}{3-1} + \binom{10-(4+1+8+1)+3-1}{3-1} \\ &- \binom{10-(3+1+4+1+8+1)+3-1}{3-1}, \end{aligned}$$

which is $66 - [28 + 21 + 3] + [3 + 0 + 0] - 0 = 17$.

In this case, from a multiset of size $3+4+8=15$, we need to toss $15-10=5$ items, which is more than the number, 3, of a's and more than the number, 4, of b's. So the Inclusion-Exclusion principle is non-trivially useful either to count the surviving combinations or to count the ways to toss elements. Still, the numbers are smaller and the non-zeros are fewer if we count ways to toss. The number of ways to toss 5 elements is

$$\begin{aligned}
&\binom{5+3-1}{3-1} \quad - \quad \left[\binom{5-(3+1)+3-1}{3-1}+\binom{5-(4+1)+3-1}{3-1}+\binom{5-(8+1)+3-1}{3-1}\right] \\
&\qquad + \quad \binom{5-(3+1+4+1)+3-1}{3-1}+\binom{5-(3+1+8+1)+3-1}{3-1}+\binom{5-(4+1+8+1)+3-1}{3-1} \\
&\qquad - \quad \binom{5-(3+1+4+1+8+1)+3-1}{3-1}.
\end{aligned}$$

This is $21-[3+1+0]+[0+0+0]-0=17$.

Mathematical concepts: Combinations, Inclusion-Exclusion principle

Problem 3 Let $f(x,y)=24xy$ on the triangular region $0\le x,y\le x+y\le 1$.

(a) Show that f is a joint probability density function.

(b) Find $E[X]$, where X is the random variable associated with x under f.

(c) Find $E[Y]$, where Y is the random variable associated with y under f. Solve using the above without any new computation.

(d) Are X and Y independent?

Solution outline

Ross 6.21, page 288.

(a) We can readily check that $f\ge 0$. The integral on the given region R is

$$\begin{aligned}
\iint_R 24xy &= 24\int_{x=0}^{1} x\int_{y=0}^{1-x} y\,dy\,dx \\
&= 24\int_{x=0}^{1} x\frac{(1-x)^2}{2}dx \\
&= 12\int_0^1 x-2x^2+x^3\,dx \\
&= 6x^2-8x^3+3x^4\big|_0^1 \\
&= 6-8+3=1.
\end{aligned}$$

(b) The expectation is $\iint_R xf(x,y)$. That is

$$\begin{aligned}\iint_R 24x^2y &= 24\int_{x=0}^{1} x^2 \int_{y=0}^{1-x} y dy dx \\ &= 24\int_{x=0}^{1} x^2 \frac{(1-x)^2}{2} dx \\ &= 12\int_0^1 x^2 - 2x^3 + x^4\, dx \\ &= 4x^2 - 6x^3 + \frac{12}{5}x^4\Big|_0^1 \\ &= 4 - 6 + \frac{12}{5} = \frac{2}{5}.\end{aligned}$$

(c) The density (inside and outside of R) is symmetric in the sense that it is unchanged under exchange of x and y, so $24xy = 24yx$ and $x + y = 1$ iff $y + x = 1$, etc., in the definition of R. So X and Y are identical and have the same expectation, $\frac{2}{5}$.

(d) The variables are dependent. If $X = 1$, then $Y \equiv 0$, whereas, if $X = \frac{1}{2}$, then Y has density that is a normalized version of $24(\frac{1}{2})y$, or $4y$, for $0 \le y \le \frac{1}{2}$. (The function f is continuous, so the above also holds approximately conditioned on the non-zero probability events $X \approx 1$ and $X \approx \frac{1}{2}$.)

Mathematical concepts: Continuous random variables, joint distributions, expectation, independence

Problem 4 A surveyer is knocking on doors in Ann Arbor, collecting answer to the sensitive question, "are you rooting for Ohio State?". Respondents are to follow this protocol:

- Flip a coin C_1 with heads probability p.
- If C_1 is heads, answer YES or NO truthfully.
- If C_1 is tails, flip another coin, C_2, with heads probability $\frac{1}{2}$, and answer YES or NO according to C_2.

Suppose n people are surveyed and Suppose $cn \le n$ people have true answer YES. Let X denote the total number of YES answers given to the survey (which is often not exactly cn).

(a) Suppose Alice is surveyed. Let f_{NO} be the probability mass function for X conditioned on Alice's true answer equal to NO and let f_{YES} be the probability mass function for X conditioned on Alice's true answer equal to YES. Find $b = \sum_x |f_{\text{YES}} - f_{\text{NO}}|$ that holds over all possibilities of others' answers. (This protects Alice's privacy.)

(b) Find $\mu = E[X]$. (This and the below insure that the data gathered is useful.)

(c) Find $\sigma^2 = \text{Var}[X] = E[(X - \mu)^2]$.

(d) The Chebyshev inequality says that, for any random variable Y with mean μ and standard deviation σ, and any non-negative k, we have $\Pr(|Y - \mu| > k\sigma) \le \frac{1}{k^2}$. Given b as above and if we want $\Pr(|X - E[X]| > \frac{1}{10}n) \le \frac{1}{100}$, find p and n to satisfy requirements, as guaranteed by Chebyshev.

Solution outline

(a) We can, without loss of generality, generate X by surveying Alice last, and consider the joint distribution $\vec{y}$ on all $n-1$ others. Conditioned on whatever the others respond, the joint distribution on all (including Alice) leads to $|f_{\text{YES}} - f_{\text{NO}}| = 1$ with probability p and, considering the two outcomes of C_2, leads to $|f_{\text{YES}} - f_{\text{NO}}| = 1$ with additional probability $\frac{1-p}{2}$. Summing for each x the outcome relevant to x, we get $b = \frac{1+p}{2}$.

(b) By linearity of expectation, sum the expectation of the respondents. The cn true-YES respondents contribute expectation $p \cdot 1 + \frac{1-p}{2}$ each, for total $cn\frac{1+p}{2}$, and the $(1-c)n$ true-NO respondents contribute $p \cdot 0 + \frac{1-p}{2}$ each, for total $(1-c)n\frac{1-p}{2}$. The grand total is $cnp + n\frac{1-p}{2}$. (In practice, c is unknown but n and p are known and so c can be isolated, to the extent that the empirical result equals or is close to the expected value.)

(c) The responses of distinct individuals are independent, so the variances add. Each of the cn true-YES respondents answers YES with probability $\frac{1+p}{2}$, so contributes variance $\frac{1+p}{2} \cdot \frac{1-p}{2}$, and, likewise, for the true-NO respondents, for total $n\frac{1-p^2}{4}$.

(d) We are given $b = \frac{1+p}{2}$ to protect privacy, so $p = 2b - 1$. The failure probability $\frac{1}{k^2} \leq \frac{1}{100}$ makes $k \geq 10$. The standard deviation is $\sigma = \frac{1}{2}\sqrt{n(1-p^2)}$. So we want $k\sigma = \frac{k}{2}\sqrt{n(1-p^2)} = \frac{n}{10}$, or $5k\sqrt{1-p^2} = \sqrt{n}$, or $n \geq 2500(1-p^2) = 2500(1-(2b-1)^2)$.

Mathematical concepts: Bernoulli random variables, mean, variance, tail bounds, probability sample spaces

Problem 5

The fruit orange used to be called norange in English like naranja in Spanish; after saying "a norange" many times, English shifted to "an orange."

Suppose we are given a string of letters without spaces, like anorange, and the goal is to insert spaces to maximize the sum of quality scores of the substrings. For example, in modern English, presumably quality(an) + quality(orange) is greater than either quality(a) + quality(norange) or quality(an) + quality(ora) + quality(nge). Note that the number of spaces is not fixed, but optimized, along with the placement of spaces. Qualities may be positive or negative. Ignore any consideration about whether the string of words makes sense; e.g., a/no/range consists of high quality individual words, even if the string of words is less plausible than an/orange.

Give an algorithm that takes a string of n symbols, has access to quality() as a unit-cost black box for single substrings (potential words), and, in time polynomial in n, finds a splitting that maximizes the sum of the qualities. Briefly show correctness and efficiency, including finding a in the runtime $O(n^a)$.

Solution outline

Kleinberg and Tardos, exercise 6.5, pages 316–17.

Use dynamic programming. Let `A[1..n]` denote the input string of letters and maintain the table $Q(i)$ of the best quality segmentation of the first i symbols `A[1..i]`.

```
Q(0) = 0 // handles empty string at start
For i = 1 to n
  Q(i) = maximum over 0 <= j < i of Q(j) + quality(A[j+1..i]).
```

That is, the best quality segmentation of `A[1..i]` consists of a segmentation of some possibly-empty prefix `A[1..j]` and some non-empty last word, `A[j+1..i]`. These are considered.

The maximization implies a loop of i iterations, nested inside an explicit `For` loop. The total number of iterations is about $\frac{n^2}{2}$, so the algorithm is quadratic, $O(n^2)$. (The number of iterations is at most n^2. For

the $n/2$ iterations where $i > n/2$, the maximum is over at least $n/2$ values of j, so there are at least $\frac{n^2}{4}$ iterations.)

Mathematical concepts: Dynamic Programming