Gov 2001 Section 2, Feb. 14, 2025

1. Rmarkdown: click the "New File" icon at the upper left corner of RStudio and select "R Markdown". Choose "PDF" as the output Format, then click "OK". In the created Rmd file, you can write math expressions and format as in Overleaf. To do coding tasks, click the green "+c" icon at the upper right to add an R chunk. After finishing the coding, click "Knit" and then choose "Knit to PDF".

2. Please simplify your result. However, you do not need to expand the binomial coefficient $\binom{n}{k}$ (unless the final expression can obviously be further simplified). Also there is no need to calculate the exact numbers, especially when there are difficult factorials like 365! - you can just leave it there.

3. To falsify a statement, you only need one counter-example. To prove a statement, one useful strategy is proof by contradiction: first assume that the statement is false (i.e., its converse is true), and then derive a contradiction with the given conditions or axioms.

4. Intersections are commutative: $P(A \cap B) = P(B \cap A)$, and $P(A|B, C) = P(A|C, B)$.

5. We have $P(A|B) = 1 - P(A^c|B)$. Be careful: $P(A|B) \neq P(B) - P(A^c|B)$, and $P(A|B) \neq 1 - P(A|B^c)$.

6. If $A \perp B$, $P(A|B) = P(A|B^c) = P(A)$. If $A \perp B \mid C$, $P(A|B, C) = P(A|B^c, C) = P(A|C)$.

Random Variable: A function that assigns a number to each possible outcome in our sample space - or we may say, it "crystallizes" an event (outcome) to a value. In calculations, always be careful to distinguish random variables with constants!

Support: $\{x_1, x_2, \ldots, x_n, \ldots\}$ is the set of all values a random variable X can be.

PMF: $P(X = k)$ for some k in the support of the r.v. X. *It must be nonnegative and sum to* 1. We usually find the PMF in two ways:

(1) By recognizing the story of a named distribution and citing the PMF that we have derived in class (always check this first, since it's easier!).

(2) By calculating $P(X = k)$ for every k in the support using naive definition of probability, multiplication rule, counting, etc.

CDF: $F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$.

Expectation:

$$E(X) \;=\; \sum_k k\, P(X = k)$$

Also, using LOTUS (Law of the Unconscious Statistician) for a function $g(X)$:

$$E\big(g(X)\big) \;=\; \sum_k g(k)\, P(X = k).$$

Variance:

$$\mathrm{Var}(X) \;=\; E[(X - E[X])^2] \;=\; E\big(X^2\big) \;-\; \big[E(X)\big]^2.$$

Jensen's inequality: Let X be a random variable. If g is a convex function, like $g(x) = 2^x$, then $E[g(X)] \geq g(E[X])$. If g is a concave function, like $g(x) = \log x$, then $E[g(X)] \leq g(E[X])$.

Exercise Questions

1. True or false:

   (a) If discrete random variables X and Y have the same CDF, they have the same PMF;

   (b) If X and Y have the same CDF, they have the same expectation.

   (c) If X and Y have the same CDF, they must be dependent.

   (d) If X and Y have the same CDF, we must have $P(X < Y) \leq 1/2$.

   True (it actually also holds for continuous random variables); True (but the reverse is not true); False (we can have i.i.d.); False (only holds for i.i.d. random variables).

2. Find the expected number of birthday pairs in a class of k students and the expected number of days in a year on which at least two of these k people were born.

   Hint: use the fundamental bridge and linearity of expectation to solve this problem. They are very powerful!

   Solution to the second problem (more complicated; the first one follows similar steps):

   Let X be the number of days (out of 365) on which at least two of the 50 people share a birthday. Define the indicator variable

   $$I_d = \begin{cases} 1, & \text{if at least two people are born on day } d, \\ 0, & \text{otherwise.} \end{cases}$$

   Then

   $$X = \sum_{d=1}^{365} I_d \quad \text{and} \quad \mathbb{E}[X] = \sum_{d=1}^{365} \mathbb{E}[I_d].$$

   Because each day is equally likely for each birthday,

   $$\mathbb{E}[I_d] = P(\text{at least two people share day } d) = 1 - \left(\frac{364}{365}\right)^k - k \cdot \frac{1}{365}\left(\frac{364}{365}\right)^{k-1}.$$

   Hence,

   $$\mathbb{E}[X] = 365 \times \left[1 - \left(\frac{364}{365}\right)^k - k \cdot \frac{1}{365}\left(\frac{364}{365}\right)^{k-1}\right].$$

3. A random binary sequence of length n is generated, with $n \geq 3$. Each digit is 1 with probability p and 0 with probability $q = 1 - p$, independently. Let X be the number of occurrences of the pattern 110. For example, 11111000011010001101111 has $X = 3$.

   (a) Find $E(X)$.

   (b) For $n = 6$, find $\text{Var}(X)$.

Let I_j be the indicator variable that a pattern 110 starts at digit j. Obviously, $X = \sum_{j=1}^{n-2} I_j$. Since $E(I_j) = p^2q$, by linearity, we have $E(X) = \sum_{j=1}^{n-2} E[I_j] = (n-2)p^2q$.

Similar to (a), for $X = 2$, there is only one case: 110110, so the probability is p^4q^2. For $X = 1$, the 110 pattern can appear starting from the $1, 2, 3$ or 4th digit, while the other digits can be random number (but we must remove the $X = 2$ case). So the probability is $4p^2q - 2p^4q^2$. Therefore,

$$E[X^2] = 4p^2q - 2p^4q^2 + 4p^4q^2 = 2p^4q^2 + 4p^2q$$

From (a) we know that $E[X] = 4p^2q$. Therefore,

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = 4p^2q - 14p^4q^2$$

4. Let X be a random variable with support $\{1, 2, \ldots, n\}$, such that

$$P(X = j) = cj, \quad \text{for } j = 1, 2, \ldots, n,$$

where c is a normalizing constant. You may find the following sums useful:

$$\sum_{j=1}^{n} j = \frac{n(n+1)}{2}, \quad \sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6}, \quad \sum_{j=1}^{n} j^3 = \left(\frac{n(n+1)}{2}\right)^2.$$

(a) Find c.

(b) Find $E[X]$ and $\mathrm{Var}[X]$.

(c) Find the CDF of X for all real number x. You can use the floor function: let $\lfloor x \rfloor$ denote the greatest integer less than or equal to x.

(a) The PMF must sum to 1:

$$\sum_{j=1}^{n} cj = 1.$$

Using the formula for the sum of the first n integers,

$$c\sum_{j=1}^{n} j = c\frac{n(n+1)}{2} = 1.$$

Solving for c:

$$c = \frac{2}{n(n+1)}.$$

(b) Using the expectation formula:

$$\mathbb{E}[X] = \sum_{j=1}^{n} jP(X = j) = \sum_{j=1}^{n} j \cdot \frac{2j}{n(n+1)}.$$

Since

$$\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6},$$

we get

$$\mathbb{E}[X] = \frac{2}{n(n+1)} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{2n+1}{3}.$$

For variance, by LOTUS:

$$\mathbb{E}[X^2] = \sum_{j=1}^{n} j^2 P(X=j) = \frac{2}{n(n+1)} \sum_{j=1}^{n} j^3.$$

Using the formula:

$$\sum_{j=1}^{n} j^3 = \left(\frac{n(n+1)}{2}\right)^2,$$

we obtain:

$$\mathbb{E}[X^2] = \frac{2}{n(n+1)} \cdot \frac{n^2(n+1)^2}{4} = \frac{n(n+1)}{2}.$$

Thus, the variance is:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{n(n+1)}{2} - \left(\frac{2n+1}{3}\right)^2.$$

(c) For an integer k, the cumulative distribution function is:

$$F(k) = P(X \le k) = \sum_{j=1}^{k} P(X=j).$$

Using the sum formula:

$$F(k) = \sum_{j=1}^{k} \frac{2j}{n(n+1)} = \frac{2}{n(n+1)} \cdot \frac{k(k+1)}{2}.$$

Thus,

$$F(k) = \frac{k(k+1)}{n(n+1)}.$$

For all real x, we define:

$$F(x) = \begin{cases} 0, & x < 1, \\ \frac{\lfloor x \rfloor (\lfloor x \rfloor + 1)}{n(n+1)}, & 1 \le x < n, \\ 1, & x \ge n. \end{cases}$$