CS 70 Discrete Mathematics and Probability Theory
Summer 2025 Tate **HW 5**

1 Sumanth and the High E Note

Note 15
Note 19

Every morning, Sumanth practices his flute. On any given day, he has a fixed probability p of successfully hitting the high E note. Each day's attempt is independent of the others.

(a) Let X be the number of days until he hits the high E note for the first time. What is $\mathbb{E}[X]$? (You may cite a known result.)

(b) Now suppose Sumanth vows to quit flute forever if he doesn't get it within the first k days. What is the expected number of days he will practice in this case?

(c) Suppose Sumanth plays two different flutes, one in the morning and one in the evening. Each gives him an independent chance p of hitting the high E. What is the expected number of days until he hits the note on either flute?

(d) Now suppose Sumanth plays n different flutes each day, each giving him an independent probability p of hitting the high E note. What is the expected number of days until he hits it on at least one flute?

Solution:

(a) This is the classic geometric distribution. Let $X \sim \text{Geom}(p)$, representing the number of trials until the first success.

$$\mathbb{E}[X] = \frac{1}{p}$$

(b) Let's denote the expected number of days Sumanth practices as:

$$\mathbb{E}[X_k] = \sum_{i=1}^{k} i \cdot \mathbb{P}[\text{first success on day } i] + k \cdot \mathbb{P}[\text{no success in } k \text{ days}]$$

$$= \sum_{i=1}^{k} i \cdot (1-p)^{i-1} p + k \cdot (1-p)^k$$

This accounts for all successful outcomes on days 1 through k, and adds k again for the case that he never succeeds.

To find a closed form for $\mathbb{E}[X_k]$, notice that

$$\begin{aligned}\mathbb{E}[X_k] &= \sum_{i=1}^{k} i \cdot \mathbb{P}[\text{first success on day } i] + k \cdot \mathbb{P}[\text{no success in } k \text{ days}] \\ &= \mathbb{E}[X \mid X \leq k]\,\mathbb{P}[X \leq k] + k\,\mathbb{P}[X > k]\end{aligned}$$

where X is defined in part (a).

From the law of total expectation:

$$\mathbb{E}[X] = \mathbb{E}[X \mid X \leq k]\,\mathbb{P}[X \leq k] + \mathbb{E}[X \mid X > k]\P[X > k]$$

Solving for $\mathbb{E}[X \mid X \leq k]\,\mathbb{P}[X \leq k]$ gives us

$$\mathbb{E}[X \mid X \leq k]\,\mathbb{P}[X \leq k] = \mathbb{E}[X] - \mathbb{E}[X \mid X > k]\,\mathbb{P}[X > k]$$

Substituting this into our original expression for $\mathbb{E}[X_k]$ gives us

$$\begin{aligned}\mathbb{E}[X_k] &= \mathbb{E}[X \mid X \leq k]\,\mathbb{P}[X \leq k] + k\,\mathbb{P}[X > k] \\ &= \mathbb{E}[X] - \mathbb{E}[X \mid X > k]\,\mathbb{P}[X > k] + k\,\mathbb{P}[X > k] \\ &= \mathbb{E}[X] - (k - \mathbb{E}[X \mid X > k])\,\mathbb{P}[X > k] \\ &= \frac{1}{p} - (k - (k + \frac{1}{p}))(1-p)^k \\ &= \frac{1}{p} - (\frac{1}{p})(1-p)^k \\ &= \frac{1-(1-p)^k}{p}\end{aligned}$$

where going from line 3 to line 4 uses $\mathbb{E}[X \mid X > k] = k + \frac{1}{p}$, which follows from X's memoryless property. (The added k term accounts for the fact that there have been k unsuccessful days already.) Another way to view is is to define $Y = X - k$, the number of days until success starting after the kth day. Then with $X = Y + k$, $\mathbb{E}[X \mid X > k] = \mathbb{E}[Y + k \mid Y + k > k] = \mathbb{E}[Y + k \mid Y > k] = \mathbb{E}[Y + k] = k + \mathbb{E}[Y]$, and Y is geometric by memoryless property.

Another way to find the closed form for $\mathbb{E}[X_k]$ is to apply the tail sum formula. Because Sumanth always stops within the first k days, $\mathbb{P}[X_k \geq i] = 0$ for any $i > k$. As a result, the tail sum formula becomes

$$\mathbb{E}[X_k] = \sum_{i=1}^{\infty} \mathbb{P}[X_k \geq i] = \sum_{i=1}^{k} \mathbb{P}[X_k \geq i] = \sum_{i=1}^{k} (1-p)^{i-1}$$

This sum looks like a geometric series, except that it has finitely many terms. We can express it as the difference of two geometric series as follows:

$$\begin{aligned}\sum_{i=1}^{k}(1-p)^{i-1} &= \sum_{i=1}^{\infty}(1-p)^{i-1} - \sum_{i=k+1}^{\infty}(1-p)^{i-1} \\ &= \sum_{i=1}^{\infty}(1-p)^{i-1} - (1-p)^k \cdot \sum_{i=k+1}^{\infty}(1-p)^{i-k-1} \\ &= \sum_{i=1}^{\infty}(1-p)^{i-1} - (1-p)^k \cdot \sum_{j=1}^{\infty}(1-p)^{j-1}\end{aligned}$$

In the last line, we just re-indexed by $j = i-k$. Both of the geometric series in the expression above evaluate to $\frac{1}{p}$, so our final answer is

$$\mathbb{E}[X_k] = \frac{1}{p} - (1-p)^k \cdot \frac{1}{p} = \frac{1-(1-p)^k}{p}$$

(c) Now, each day Sumanth gets two independent tries, each with success probability p. The probability that he fails both flutes in a day is:

$$(1-p)^2$$

So the probability that he succeeds on at least one flute is:

$$1-(1-p)^2$$

Now this is a geometric distribution with success probability $1-(1-p)^2$, so:

$$\mathbb{E}[X] = \frac{1}{1-(1-p)^2} = \frac{1}{2p-p^2}$$

(d) Same idea as Part (c), but generalized to n flutes.

Each day, the probability that all n flutes fail is $(1-p)^n$

So the probability of at least one success is:

$$1-(1-p)^n$$

Therefore, the expected number of days until success is:

$$\mathbb{E}[X] = \frac{1}{1-(1-p)^n}$$

2 Ritwik's NVIDIA Gauntlet

Note 15
Note 19

Ritwik is applying to an NVIDIA internship. There are three interviews: Coding (C), Systems (S), and HR (H). Define the following random variables with possible values $\{0,1\}$:

- $C = 1$ if he passes coding, with $\mathbb{P}[C = 1] = 0.8$.
- $S = 1$ if he passes systems. Given $C = 1$, $\mathbb{P}[S = 1] = 0.6$; given $C = 0$, $\mathbb{P}[S = 1] = 0.1$.
- $H = 1$ if he passes HR. Given $S = 1$, $\mathbb{P}[H = 1] = 0.9$; given $S = 0$, $\mathbb{P}[H = 1] = 0.3$.

Define random variables $X = C + S + H$ and $Y = 2C + 3S + H$.

(a) Compute the joint distribution table for (C, S) and (S, H).

(b) Compute $\mathbb{E}[X]$.

(c) Which is more likely: Ritwik passes all three rounds, or only Systems and HR (i.e., $C = 0, S = 1, H = 1$)? Justify.

(d) Compute $\mathbb{E}[Y]$.

Solution:

(a) **Compute joint distributions**

We are given:

$$\begin{aligned} \mathbb{P}[C = 1] &= 0.8 & \Rightarrow \mathbb{P}[C = 0] &= 0.2 \\ \mathbb{P}[S = 1 \mid C = 1] &= 0.6 & \Rightarrow \mathbb{P}[S = 0 \mid C = 1] &= 0.4 \\ \mathbb{P}[S = 1 \mid C = 0] &= 0.1 & \Rightarrow \mathbb{P}[S = 0 \mid C = 0] &= 0.9 \end{aligned}$$

Using the chain rule $\mathbb{P}[C, S] = \mathbb{P}[C] \cdot \mathbb{P}[S \mid C]$:

C \S	**S = 1**	**S = 0**
$C = 1$	$0.8 \cdot 0.6 = \mathbf{0.48}$	$0.8 \cdot 0.4 = \mathbf{0.32}$
$C = 0$	$0.2 \cdot 0.1 = \mathbf{0.02}$	$0.2 \cdot 0.9 = \mathbf{0.18}$

Now for (S, H). First compute $\mathbb{P}[S = 1] = 0.48 + 0.02 = 0.5$, and $\mathbb{P}[S = 0] = 0.32 + 0.18 = 0.5$.

Then use:

$$\begin{aligned} \mathbb{P}[H = 1 \mid S = 1] = 0.9 &\Rightarrow \mathbb{P}[H = 0 \mid S = 1] = 0.1 \\ \mathbb{P}[H = 1 \mid S = 0] = 0.3 &\Rightarrow \mathbb{P}[H = 0 \mid S = 0] = 0.7 \end{aligned}$$

S \H	**H = 1**	**H = 0**
$S = 1$	$0.5 \cdot 0.9 = \mathbf{0.45}$	$0.5 \cdot 0.1 = \mathbf{0.05}$
$S = 0$	$0.5 \cdot 0.3 = \mathbf{0.15}$	$0.5 \cdot 0.7 = \mathbf{0.35}$

(b) **Compute** $\mathbb{E}[X] = \mathbb{E}[C+S+H]$

By linearity of expectation:

$$\begin{aligned} \mathbb{E}[C] &= 0.8 \\ \mathbb{E}[S] &= 0.5 \quad \text{(from joint above)} \\ \mathbb{E}[H] &= \mathbb{P}[H=1] = 0.45+0.15 = 0.6 \\ \Rightarrow \mathbb{E}[X] &= 0.8+0.5+0.6 = \boxed{1.9} \end{aligned}$$

(c) **Compare two scenarios:**

- Case 1: Ritwik passes all three:

$$\mathbb{P}[C=1,S=1,H=1] = \mathbb{P}[C=1,S=1] \cdot \mathbb{P}[H=1 \mid S=1] = 0.48 \cdot 0.9 = \boxed{0.432}$$

- Case 2: He fails coding but passes the rest:

$$\mathbb{P}[C=0,S=1,H=1] = 0.02 \cdot 0.9 = \boxed{0.018}$$

So Ritwik is **much more likely** to pass all three rounds than only S and H.

(d) **Compute** $\mathbb{E}[Y] = \mathbb{E}[2C+3S+H]$

Use linearity:

$$\begin{aligned} \mathbb{E}[Y] &= 2\cdot\mathbb{E}[C] + 3\cdot\mathbb{E}[S] + \mathbb{E}[H] \\ &= 2(0.8)+3(0.5)+0.6 = 1.6+1.5+0.6 = \boxed{3.7} \end{aligned}$$

3 Anikait's Completion Saga

Note 16
Note 19

Anikait, a huge anime fan recovering from a basketball injury, has been spending his recovery time collecting limited-edition anime figurines. Each booster pack contains **one** random figurine, and there are n different figurines in total.

Let's define the following scenario:

Anikait opens booster packs one at a time:

- The probability that a pack contains a figurine he **hasn't** collected yet is proportional to the number of figurines he hasn't seen yet.
- Let T_i represent the number of packs he needs to open to get the i^{th} new figurine.
- Let $S_n = \sum_{i=1}^{n} T_i$ be the total number of packs needed to collect all n figurines.
- You may use the following approximation for harmonic numbers: $\sum_{j=a}^{b} \frac{1}{j} \approx \ln b - \ln a$. In particular, the n-th harmonic number is defined as $H_n = \sum_{j=1}^{n} \frac{1}{j} \approx \ln n + \gamma$, where γ is the Euler–Mascheroni constant (you may ignore γ for approximation purposes).

(a) Show that $\mathbb{E}[T_i] = \frac{n}{n-i+1}$ (you should identify the type of distribution T_i has).

(b) Compute $\mathbb{E}[S_n]$.

(c) Compute $\mathrm{Var}(S_n)$.

(d) Suppose Anikait is only aiming to collect k distinct figurines instead of all n. Find $\mathbb{E}[S_k]$ and explain how it scales with k and n.

(e) Suppose Anikait collects figurines his friend Pranav, and both open one pack per hour independently until either of them completes the full set of n figurines. Let $X \sim S_n$ be the number of packs Anikait needs, and let $Y \sim S_n$ be the number of packs Pranav needs. Compute or bound $\mathbb{E}[\min(X,Y)]$ using symmetry.

Hint 1: Recall the identity for any two random variables X, Y:

$$\min(X,Y) = \frac{X+Y-|X-Y|}{2}.$$

This implies:

$$\mathbb{E}[\min(X,Y)] = \mu - \frac{1}{2}\mathbb{E}[|X-Y|],$$

where $\mu = \mathbb{E}[S_n] = nH_n$.

Hint 2: Use the Cauchy-Schwarz inequality to bound $\mathbb{E}[|X-Y|]$. For any random variables A, B,

$$\mathbb{E}[|AB|] \le \sqrt{\mathbb{E}[A^2]\cdot\mathbb{E}[B^2]}.$$

Solution:

(a) Since $T_i \sim \mathrm{Geom}(p_i)$ and the expectation of a geometric distribution is $\mathbb{E}[T_i] = \frac{1}{p_i}$, we have:

$$\mathbb{E}[T_i] = \frac{1}{\frac{n-(i-1)}{n}} = \frac{n}{n-i+1}.$$

(b) We know that $S_n = \sum_{i=1}^n T_i$, and expectation is linear even for dependent variables:

$$\mathbb{E}[S_n] = \sum_{i=1}^n \mathbb{E}[T_i] = \sum_{i=1}^n \frac{n}{n-i+1} = n\sum_{j=1}^n \frac{1}{j} = nH_n,$$

where H_n is the n-th harmonic number.

(c) We're given the variance of T_i as:

$$\mathrm{Var}(T_i) = \frac{1-p_i}{p_i^2}$$

with $p_i = \frac{n-i+1}{n}$. Plugging in:

$$\mathrm{Var}(T_i) = \frac{1-\frac{n-i+1}{n}}{\left(\frac{n-i+1}{n}\right)^2} = \frac{\frac{i-1}{n}}{\left(\frac{n-i+1}{n}\right)^2} = \frac{(i-1)n}{(n-i+1)^2}.$$

Since $S_n = \sum T_i$ and the T_i are independent:

$$\text{Var}(S_n) = \sum_{i=1}^{n} \text{Var}(T_i) = \sum_{i=1}^{n} \frac{(i-1)n}{(n-i+1)^2}.$$

(d) We want to compute $\mathbb{E}[S_k]$, the expected number of packs to collect k unique figurines (with $k < n$).

Using the same logic as before:

$$\mathbb{E}[S_k] = \sum_{i=1}^{k} \frac{n}{n-i+1} = n \sum_{j=n-k+1}^{n} \frac{1}{j}.$$

This grows roughly like:

$$n \ln\left(\frac{n}{n-k}\right)$$

for $k \ll n$, which is slower than full collection ($n \ln n$).

(e) Let $X \sim S_n$, and let $Y \sim S_n$, both i.i.d. random variables representing the number of packs each person (Anikait and his friend) needs to complete their collection.

They stop as soon as either finishes, i.e., total time is $\min(X,Y)$.

Use the identity:

$$\min(X,Y) = \frac{X+Y-|X-Y|}{2} \Rightarrow \mathbb{E}[\min(X,Y)] = \mu - \frac{1}{2}\mathbb{E}[|X-Y|],$$

where $\mu = \mathbb{E}[S_n] = nH_n$.

Since X and Y are identically distributed and independent: - $\mathbb{E}[|X-Y|]$ is symmetric around 0, and - computing it exactly is tricky without more info, but:

We know that:

$$\mathbb{E}[|X-Y|] \leq \sqrt{\mathbb{E}[(X-Y)^2]} = \sqrt{2\text{Var}(S_n)}$$

So we can bound:

$$\mathbb{E}[\min(X,Y)] \geq \mu - \frac{1}{2}\sqrt{2\text{Var}(S_n)}$$

Thus, they can expect to finish faster than one person alone — this makes intuitive sense since they're working in parallel.

4 Double-Check Your Intuition Again

Note 16

(a) You roll a fair six-sided die and record the result X. You roll the die again and record the result Y.

(i) What is $\text{cov}(X+Y, X-Y)$?

(ii) Prove that $X+Y$ and $X-Y$ are not independent.

For each of the problems below, if you think the answer is "yes" then provide a proof. If you think the answer is "no", then provide a counterexample.

(b) If X is a random variable and $\text{Var}(X)=0$, then must X be a constant?

(c) If X is a random variable and c is a constant, then is $\text{Var}(cX)=c\,\text{Var}(X)$?

(d) If A and B are random variables with nonzero standard deviations and $\text{Corr}(A,B)=0$, then are A and B independent?

(e) If X and Y are not necessarily independent random variables, but $\text{Corr}(X,Y)=0$, and X and Y have nonzero standard deviations, then is $\text{Var}(X+Y)=\text{Var}(X)+\text{Var}(Y)$?

(f) If X and Y are random variables then is $\mathbb{E}[\max(X,Y)\min(X,Y)]=\mathbb{E}[XY]$?

(g) If X and Y are independent random variables with nonzero standard deviations, then is

$$\text{Corr}(\max(X,Y),\min(X,Y))=\text{Corr}(X,Y)?$$

Solution:

(a) (i) Using bilinearity of covariance, we have

$$\begin{aligned}\text{cov}(X+Y,X-Y)&=\text{cov}(X,X)+\text{cov}(X,Y)-\text{cov}(Y,X)-\text{cov}(Y,Y)\\&=\text{cov}(X,X)-\text{cov}(Y,Y),\\&=0\end{aligned}$$

where we use that $\text{cov}(X,Y)=\text{cov}(Y,X)$ to get the second equality.

(ii) Observe that $\mathbb{P}[X+Y=7,X-Y=0]=0$ because if $X-Y=0$, then the sum of our two dice rolls must be even. However, both $\mathbb{P}[X+Y=7]$ and $\mathbb{P}[X-Y=0]$ are nonzero, so $\mathbb{P}[X+Y=7,X-Y=0]\neq\mathbb{P}[X+Y=7]\cdot\mathbb{P}[X-Y=0]$.

(b) Yes. If we write $\mu=\mathbb{E}[X]$, then $0=\text{Var}(X)=\mathbb{E}[(X-\mu)^2]$ so $(X-\mu)^2$ must be identically 0 since perfect squares are non-negative. Thus $X=\mu$.

(c) No. We have $\text{Var}(cX)=\mathbb{E}[(cX-\mathbb{E}[cX])^2]=c^2\,\mathbb{E}[(X-\mathbb{E}[X])^2]=c^2\,\text{Var}(X)$ so if $\text{Var}(X)\neq 0$ and $c\neq 0$ or $c\neq 1$ then $\text{Var}(cX)\neq c\,\text{Var}(X)$. This does prove that $\sigma(cX)=c\sigma(X)$ though.

(d) No. Let $A=X+Y$ and $B=X-Y$ from part (a). Since A and B are not constants then part (b) says they must have nonzero variances which means they also have nonzero standard deviations. Part (a) says that their covariance is 0 which means they are uncorrelated, and that they are not independent.

Recall from lecture that the converse is true though.

(e) Yes. If $\text{Corr}(X,Y)=0$, then $\text{cov}(X,Y)=0$. We have $\text{Var}(X+Y)=\text{cov}(X+Y,X+Y)=\text{Var}(X)+\text{Var}(Y)+2\,\text{cov}(X,Y)=\text{Var}(X)+\text{Var}(Y)$.

(f) Yes. For any values x,y we have $\max(x,y)\min(x,y) = xy$. Thus, $\mathbb{E}[\max(X,Y)\min(X,Y)] = \mathbb{E}[XY]$.

(g) No. You may be tempted to think that because $(\max(x,y),\min(x,y))$ is either (x,y) or (y,x), then $\mathrm{Corr}(\max(X,Y),\min(X,Y)) = \mathrm{Corr}(X,Y)$ because $\mathrm{Corr}(X,Y) = \mathrm{Corr}(Y,X)$. That reasoning is flawed because $(\max(X,Y),\min(X,Y))$ is not always equal to (X,Y) or always equal to (Y,X) and the inconsistency affects the correlation. It is possible for X and Y to be independent while $\max(X,Y)$ and $\min(X,Y)$ are not.

For a concrete example, suppose X is either 0 or 1 with probability $1/2$ each and Y is independently drawn from the same distribution. Then $\mathrm{Corr}(X,Y) = 0$ because X and Y are independent. Even though X never gives information about Y, if you know $\max(X,Y) = 0$ then you know for sure $\min(X,Y) = 0$.

More formally, $\max(X,Y) = 1$ with probability $3/4$ and 0 with probability $1/4$, and $\min(X,Y) = 1$ with probability $1/4$ and 0 with probability $3/4$. This means

$$\mathbb{E}[\max(X,Y)] = 1\cdot\frac{3}{4} + 0\cdot\frac{1}{4} = \frac{3}{4}$$

and

$$\mathbb{E}[\min(X,Y)] = 1\cdot\frac{1}{4} + 0\cdot\frac{3}{4} = \frac{1}{4}.$$

Thus,

$$\begin{aligned}\mathrm{cov}(\max(X,Y),\min(X,Y)) &= \mathbb{E}[\max(X,Y)\min(X,Y)] - \frac{3}{16} \\ &= \frac{1}{4} - \frac{3}{16} = \frac{1}{16} \neq 0\end{aligned}$$

We conclude that $\mathrm{Corr}(\max(X,Y),\min(X,Y)) \neq 0 = \mathrm{Corr}(X,Y)$.

5 Achyut's Notre Dame Drone Challenge

Note 20

Achyut, a massive architect fan with a deep love for Notre Dame, is testing a new drone that scores points by successfully navigating through flying arches on a miniature replica of the cathedral.

Each **successful pass** earns a score drawn independently from a random variable X_i, where:

- $\mathbb{P}(X_i = 1) = 0.5$
- $\mathbb{P}(X_i = 2) = 0.3$
- $\mathbb{P}(X_i = 3) = 0.2$

He continues attempting passes until he accumulates at least 15 points. Let:

- N be the number of successful passes until the total score $S = \sum_{i=1}^{N} X_i \geq 15$
- The scores X_i are i.i.d., and independent of the stopping time N

(a) What is the expected score per successful pass, $\mathbb{E}[X_i]$?

(b) Using the approximation $\mathbb{E}[S] \approx \mathbb{E}[N] \cdot \mathbb{E}[X_i]$, estimate how many passes Achyut needs on average to reach 15 points.

(c) Let's define the MMSE (Minimum Mean Squared Error) estimator for S given $N = n$. Compute $\hat{S} = \mathbb{E}[S \mid N = n]$.

(d) Let T be the number of flights where Achyut scores exactly 3 points. Suppose n successful passes were made to reach total score s. Compute $\mathbb{E}[T \mid S = s]$ leaving your answer in the form n*(some conditional probability).

Solution:

(a)

$$\mathbb{E}[X_i] = 1 \cdot 0.5 + 2 \cdot 0.3 + 3 \cdot 0.2 = 0.5 + 0.6 + 0.6 = 1.7$$

(b) Let $\mathbb{E}[N] \approx \frac{15}{\mathbb{E}[X_i]} = \frac{15}{1.7} \approx 8.82$. So on average, Achyut will need about 9 passes to reach 15 points.

(c) By linearity of expectation and i.i.d. assumption:

$$\mathbb{E}[S \mid N = n] = \mathbb{E}[X_1 + \cdots + X_n \mid N = n] = n \cdot \mathbb{E}[X_i] = 1.7n$$

So the MMSE estimator is $\hat{S} = 1.7n$.

(d) Let T be the number of flights where Achyut scores exactly 3 points. Suppose n successful passes were made to reach total score s. Compute $\mathbb{E}[T \mid S = s]$ leaving your answer in the form n*(some conditional probability).

$$I_i = \begin{cases} 1 & \text{if } X_i = 3 \\ 0 & \text{otherwise} \end{cases}$$

Then $T = \sum_{i=1}^{n} I_i$. We are interested in:

$$\mathbb{E}[T \mid S = s] = \mathbb{E}\left[\sum_{i=1}^{n} I_i \,\middle|\, S = s\right]$$

Using linearity of expectation:

$$\mathbb{E}[T \mid S = s] = \sum_{i=1}^{n} \mathbb{E}[I_i \mid S = s]$$

Since the X_i are i.i.d. and we're conditioning on the sum $S = s$, all the $\mathbb{E}[I_i \mid S = s]$ are equal. Denote this common value as p_s:

$$\mathbb{E}[T \mid S = s] = n \cdot p_s \quad \text{where } p_s = \mathbb{P}(X_i = 3 \mid S = s)$$

Thus, the final expression is:

$$\boxed{\mathbb{E}[T \mid S=s] = n \cdot \mathbb{P}(X_i = 3 \mid S=s)}$$

This is the best symbolic form we can give without further assumptions on the score distribution or a full enumeration of possible combinations of X_i values that sum to s.

6 Balls in Bins Estimation

Note 20

We throw $n > 0$ balls into $m \geq 2$ bins. Let X and Y represent the number of balls that land in bin 1 and 2 respectively.

(a) Calculate $\mathbb{E}[Y \mid X]$. [*Hint*: Your intuition may be more useful than formal calculations.]

(b) What is $L[Y \mid X]$ (where $L[Y \mid X]$ is the best linear estimator of Y given X)? [*Hint*: Your justification should be no more than two or three sentences, no calculations necessary! Think carefully about the meaning of the conditional expectation.]

(c) Unfortunately, your friend is not convinced by your answer to the previous part. Compute $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

(d) Compute $\mathrm{Var}(X)$.

(e) Compute $\mathrm{cov}(X,Y)$.

(f) Compute $L[Y \mid X]$ using the formula. Ensure that your answer is the same as your answer to part (b).

Solution:

(a) $\mathbb{E}[Y \mid X = x] = (n-x)/(m-1)$, because once we condition on x balls landing in bin 1, the remaining $n-x$ balls are distributed uniformly among the other $m-1$ bins. Therefore,

$$\mathbb{E}[Y \mid X] = \frac{n-X}{m-1}.$$

(b) We showed that $\mathbb{E}[Y \mid X]$ is a linear function of X. Since $\mathbb{E}[Y \mid X]$ is the best *general* estimator of Y given X, it must also be the best *linear* estimator of Y given X, i.e. $\mathbb{E}[Y \mid X]$ and $L[Y \mid X]$ coincide.

(c) Let X_i be the indicator that the ith ball falls in bin 1. Then, $X = \sum_{i=1}^{n} X_i$, and by linearity of expectation, $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n/m$, since there are n indicators and each ball has a probability $1/m$ of landing in bin 1. By symmetry, $\mathbb{E}[Y] = n/m$ as well.

(d) The number of balls that falls into the first bin is binomially distributed with parameters n and $1/m$. Hence the variance is $n(1/m)(1-1/m)$.

(e) Let X_i be as before, and let Y_i be the indicator that the ith ball falls into bin 2.

$$\mathrm{cov}(X,Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{cov}(X_i, Y_j)$$

We can compute $\operatorname{cov}(X_i,Y_i) = \mathbb{E}[X_iY_i] - \mathbb{E}[X_i]\,\mathbb{E}[Y_i] = 0 - (1/m)(1/m) = -1/m^2$ (note that $\mathbb{E}[X_iY_i] = 0$ because it is impossible for a ball to land in both bins 1 and 2). Also, we have $\operatorname{cov}(X_i,Y_j) = 0$ because the indicator for the ith ball is independent of the indicator for the jth ball when $i \neq j$. Hence, $\operatorname{cov}(X,Y) = n(-1/m^2) = -n/m^2$.

(f)

$$\begin{aligned}
L[Y \mid X] &= \mathbb{E}[Y] + \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}(X - \mathbb{E}[X]) \\
&= \frac{n}{m} + \frac{-n/m^2}{n(1/m)(1-1/m)}\left(X - \frac{n}{m}\right) \\
&= \frac{n}{m} - \frac{1}{m-1}\left(X - \frac{n}{m}\right) \\
&= \frac{mn - n - mX + n}{m(m-1)} = \frac{n-X}{m-1}
\end{aligned}$$