CS 70 Discrete Mathematics and Probability Theory
Fall 2021 DIS 14B

# 1 LLSE

We have two bags of balls. The fractions of red balls and blue balls in bag $A$ are $2/3$ and $1/3$ respectively. The fractions of red balls and blue balls in bag $B$ are $1/2$ and $1/2$ respectively. Someone gives you one of the bags (unmarked) uniformly at random. You then draw 6 balls from that same bag with replacement. Let $X_i$ be the indicator random variable that ball $i$ is red. Now, let us define $X = \sum_{1 \leq i \leq 3} X_i$ and $Y = \sum_{4 \leq i \leq 6} X_i$.

(a) Compute $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

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

(c) Compute $\mathrm{cov}(X,Y)$. (*Hint*: Recall that covariance is bilinear.)

(d) Compute $L(Y \mid X)$, the best linear estimator of $Y$ given $X$. (*Hint*: Recall that

$$L(Y \mid X) = \mathbb{E}[Y] + \frac{\mathrm{cov}(X,Y)}{\mathrm{Var}(X)}\big(X - \mathbb{E}[X]\big).$$

)

**Solution:** Although the indicator random variables are not independent, we can still apply linearity of expectation. By symmetry, we also know that each indicator follows the same distribution.

(a)

$$\mathbb{E}[X] = \mathbb{E}[Y] = 3 \cdot \mathbb{E}[X_1] = 3 \cdot \mathbb{P}(X_1 = 1) = 3 \cdot \left(\frac{1}{2} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{1}{2}\right) = \frac{7}{4}.$$

(b)

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{cov}\Big(\sum_{1 \leq i \leq 3} X_i, \sum_{1 \leq j \leq 3} X_j\Big) \\
&= 3 \cdot \mathrm{Var}(X_1) + 6 \cdot \mathrm{cov}(X_1, X_2) \\
&= 3\big(\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2\big) + 6 \cdot \frac{1}{144} \\
&= 3\left[\frac{7}{12} - \left(\frac{7}{12}\right)^2\right] + 6 \cdot \frac{1}{144} = \frac{111}{144}.
\end{aligned}$$

(c)

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \operatorname{cov}\Big(\sum_{1\le i\le 3} X_i, \sum_{4\le j\le 6} X_j\Big) \\
&= 9\cdot \operatorname{cov}(X_1,X_4) \\
&= 9\cdot \big(\mathbb{E}[X_1X_4] - \mathbb{E}[X_1]\cdot\mathbb{E}[X_4]\big) \\
&= 9\cdot \big(\mathbb{P}(X_1=1,X_4=1) - \mathbb{P}(X_1=1)^2\big) \\
&= 9\cdot\left(\left[\frac{1}{2}\cdot\left(\frac{2}{3}\right)^2 + \frac{1}{2}\cdot\left(\frac{1}{2}\right)^2\right] - \left[\frac{1}{2}\cdot\left(\frac{2}{3}\right) + \frac{1}{2}\cdot\left(\frac{1}{2}\right)\right]^2\right) = \frac{9}{144}.
\end{aligned}$$

(d)

$$L(Y\mid X) = \frac{7}{4} + \frac{9}{111}\left(X - \frac{7}{4}\right) = \frac{3}{37}X + \frac{119}{74}.$$

## 2 Balls in Bins Estimation

We throw $n > 0$ balls into $m \ge 2$ 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 $\operatorname{Var}(X)$.

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

(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 $i$th 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 $i$th ball falls into bin 2.

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

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 $i$th ball is independent of the indicator for the $j$th ball when $i \neq j$. Hence, $\operatorname{cov}(X,Y) = n(-1/m^2) = -n/m^2$.

(f)

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

# 3 Continuous LLSE

Suppose that $X$ and $Y$ are uniformly distributed on the shaded region in the figure below.

That is, $X$ and $Y$ have the joint distribution:

$$f_{X,Y}(x,y) = \begin{cases} 1/2, & 0 \le x \le 1, 0 \le y \le 1 \\ 1/2, & 1 \le x \le 2, 1 \le y \le 2 \end{cases}$$

(a) Do you expect $X$ and $Y$ to be positively correlated, negatively correlated, or neither?

(b) Compute the marginal distribution of $X$.

(c) Compute $L[Y \mid X]$, the best linear estimator of $Y$ given $X$.

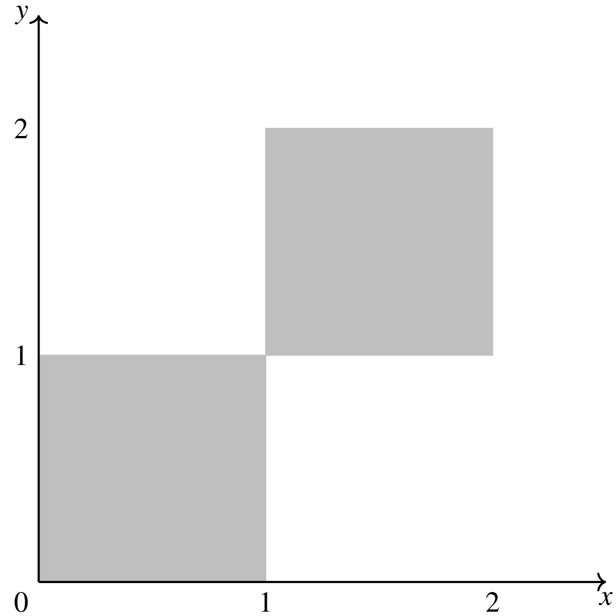


Figure 1: The joint density of $(X,Y)$ is uniform over the shaded region.

(d) What is $\mathbb{E}[Y \mid X]$?

**Solution:**

(a) Positively correlated, because high values of $Y$ correspond to high values of $X$.

(b) Intuitively, if we slice the joint distribution at any $x \in [0,2]$, then the probability is the same, so we should expect $X$ to be uniformly distributed on $[0,2]$. We verify this by explicit computation:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = 1\{0 \le x \le 1\} \int_0^1 \frac{1}{2}\,\mathrm{d}y + 1\{1 \le x \le 2\} \int_1^2 \frac{1}{2}\,\mathrm{d}y$$
$$= \frac{1}{2} 1\{0 \le x \le 2\}$$

(c) $\mathbb{E}[X] = \mathbb{E}[Y] = 1$ by symmetry. Since $X$ is uniform on $[0,2]$, $\mathrm{var}(X) = 4 \cdot 1/12 = 1/3$ (since the variance of a $U[0,1]$ random variable is $1/12$). We compute the covariance:

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1\int_0^1 xy \cdot \frac{1}{2}\,\mathrm{d}x\,\mathrm{d}y + \int_1^2\int_1^2 xy \cdot \frac{1}{2}\,\mathrm{d}x\,\mathrm{d}y$$
$$= \frac{1}{2}\left(\int_0^1 x\,\mathrm{d}x \int_0^1 y\,\mathrm{d}y + \int_1^2 x\,\mathrm{d}x \int_1^2 y\,\mathrm{d}y\right) = \frac{1}{2}\left(\frac{1}{4} + \frac{9}{4}\right) = \frac{5}{4}$$

So $\mathrm{cov}(X,Y) = 5/4 - 1 \cdot 1 = 1/4$. The LLSE is

$$L[Y \mid X] - 1 = \frac{1/4}{1/3}(X - 1)$$
$$L[Y \mid X] = \frac{3}{4}X + \frac{1}{4}$$

(d) The easiest way to solve this is to look at the picture of the joint density, and draw horizontal lines through middles of each of the two squares. Intuitively, $\mathbb{E}[Y \mid X]$ means "for each slice of $X = x$, what is the best guess of $Y$"? Slightly more formally, one can argue that conditioned on $X = x$ for $0 < x < 1$, $Y \sim U[0,1]$, so $\mathbb{E}[Y \mid X = x] = 1/2$ in this region. Conditioned on $X = x$ for $1 < x < 2$, $Y \sim U[1,2]$, so $\mathbb{E}[Y \mid X = x] = 3/2$ in this region. See Figure 2.

$$\mathbb{E}[Y \mid X = x] = \begin{cases} 1/2, & 0 \le x \le 1 \\ 3/2, & 1 \le x \le 2 \end{cases}$$

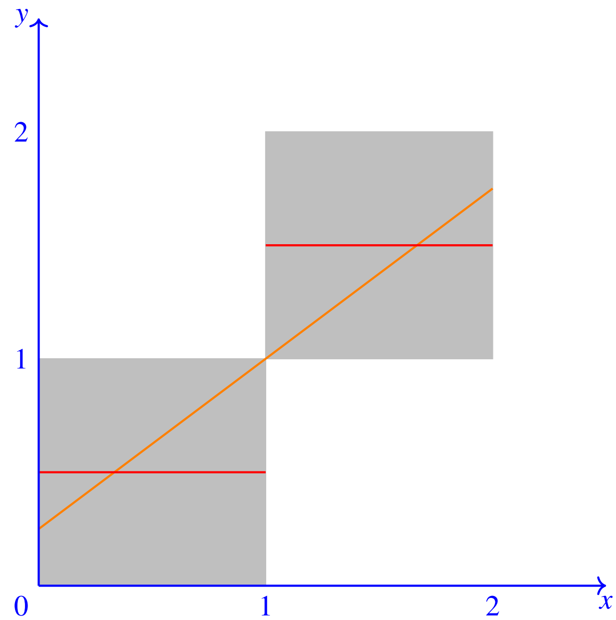


Figure 2: $L[Y \mid X]$ is the orange line. $\mathbb{E}[Y \mid X]$ is the red function.