Data 102 (Fall 2022) **Names:**
Discussion Worksheet 1
08/26/2022

---

1. (Data 140 Exercise 15.6.1) Let X have density given by

$$f(x) = \begin{cases} c(1-x^2), & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find:

(a) c

Solution: We must have $\int f(x)dx = 1$. So $c\int_{-1}^{1}(1-x^2)dx = 1$. Since $\int_{-1}^{1}(1-x^2)dx = (x - \frac{x^3}{3})\Big|_{-1}^{1} = \frac{4}{3}$, we have $c = \frac{3}{4}$.

(b) the cdf of X

Solution: The cdf is the integral of the density. As computed above, this is $\frac{3}{4}(x - x^3/3) + A$ for some constant A. Since the cdf must be 0 at $x = -1$, we have $\frac{3}{4} \cdot \frac{-2}{3} + A = 0$, or $A = \frac{1}{2}$. So the cdf is

$$F(x) = \frac{1}{2} + \frac{3}{4}x - \frac{1}{4}x^3$$

for $-1 < x < 1$. (It is 0 for $x \leq -1$ and 1 for $x \geq 1$.)

(c) $P(|X| > 0.5)$

Solution: This can be written as $P(X < -0.5) + P(X > 0.5) = F(-0.5) + 1 - F(0.5)$. To simplify algebra, we note that $F(x) - F(-x) = \frac{3}{2}x - \frac{1}{2}x^3$. So we get $1 - \frac{3}{4} + \frac{1}{16} = \frac{5}{16}$.

2. Suppose X and Y are independent random variables. Which of the following are necessarily true?

(a) $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

(b) $\mathbb{E}[e^{X+Y}] = \mathbb{E}[e^X]\mathbb{E}[e^Y]$

(c) $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

(d) $\mathrm{Var}[XY] = \mathrm{Var}[X]\,\mathrm{Var}[Y]$

Solution: (a), (b), and (c) are true while (d) is false.
For (a) and (b), recall that independence implies that $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$ for all functions f, g. We get (a) from letting $f(x) = x, g(y) = y$ and we get (b) from letting $f(x) = e^x$, $g(y) = e^y$ and recalling that $e^{x+y} = e^x e^y$.
For (c), this is the law of total expectation: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X, Y]$, and the final term is zero because X and Y are independent.
For (d), a simple counterexample: if X is 0 or 1 with equal probability, and Y is also, then $\mathrm{Var}[X] = \mathrm{Var}[Y] = \frac{1}{4}$, but $\mathrm{Var}[XY]$ is $\frac{3}{16}$ (since it is 1 with probability $1/4$ and 0 otherwise).

3. Consider the following linear regression model:

$$\hat{y}_i = \theta_0 + \theta_1 x_{i,1} + \theta_2 x_{i,2}$$

Suppose that we observe the data:

$$\begin{aligned} y_1 = 1,&\ \ x_1 = (2, 1) \\ y_2 = 2,&\ \ x_2 = (2, -1) \\ y_3 = 3,&\ \ x_3 = (0, -1) \\ y_4 = 4,&\ \ x_4 = (0, 1) \end{aligned}$$

(a) What is the least-squares estimate for θ?

Solution: We form the vector $\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$ and the matrix $\mathbf{X} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{bmatrix}$.

Here the first column is to handle the constant term for θ_0. Now, the least squares estimate for θ can be calculated as $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. We have

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 4 & 4 & 0 \\ 4 & 8 & 0 \\ 0 & 0 & 4 \end{bmatrix} = 4 \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ and } (\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{4} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and also $\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 10 \\ 6 \\ 0 \end{bmatrix}$.

Multiplying together gives the estimate $\hat{\theta} = \begin{bmatrix} \frac{2 \cdot 10 - 6}{4} \\ \frac{-10+6}{4} \\ 0 \end{bmatrix} = \begin{bmatrix} 3.5 \\ -1 \\ 0 \end{bmatrix}$.

(b) What is the predicted value $\hat{y}$ when $x = (1, 1)$?

Solution: $3.5 + (-1) \cdot 1 + 0 \cdot 1 = 2.5$

(c) What is the RMSE (root mean-squared error)?

Solution: The prediction $\hat{\mathbf{y}}$ is equal to $\mathbf{X}\hat{\theta}$, or

$$\hat{\mathbf{y}} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3.5 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 1.5 \\ 3.5 \\ 3.5 \end{bmatrix}.$$

The RMSE is the squareroot of the mean squared error between $\hat{\bar{y}}$ and $\bar{y}$, or

$$\sqrt{\frac{((1.5-1)^2 + (1.5-2)^2 + (3.5-3)^2 + (3.5-4)^2}{4}} = 0.5.$$