Exercise

(a) Provide a probability distribution for the random variable X_1

x	0	1
$P(x)$	$1-p$	p

We assume that $X_1, \dots, X_n$ have the same probability distributions. Moreover we assume that these observations are <u>indedent</u> meaning that knowledge about X_i does not tell us about X_j

(b) What is the expected value of X_1?

$$E[X_1] = 1 \cdot p + 0 \cdot (1-p)$$
$$= p$$

Note that p is the parameter for a 0-1 random variable.

(c) If we wanted to <u>minimize</u> the quantity

$$E[(X_1 - c)^2]$$

by choosing the number c, then show that

$$c = E[X_1]$$

$$\begin{aligned} E[(X_1-c)^2] &= p(1-c)^2 + (1-p)(0-c)^2 \\ &= p(1-2c+c^2) + (1-p)c^2 \\ &= p + (-2p)c + c^2 \\ &= (c-p)^2 \end{aligned}$$

To minimize we take $c = p = E[X_1]$

Alternatively we can use the derivative

$$\begin{aligned} 0 &= \frac{d}{dc}\left[p(1-c)^2 + (1-p)(0-c)^2\right] \\ &= p\frac{d}{dc}[(1-c)^2] + (1-p)\frac{d}{dc}[c^2] \\ &= -2p(1-c) + 2(1-p)c \\ &= -2p + 2c \end{aligned}$$

So again $c = p$.

(d) With this <u>data generation</u> we can use

$$\overline{X} = \frac{1}{n}(X_1 + \dots + X_n)$$

to estimate $E[X_1]$. Since $E[X_1] = p$ the estimate will help us to determine the parameter. Why is the mean a good estimate of the expectation?

Remember that by definition of expectation, the expectation of a summation is a summation of expectations. So

$$\begin{aligned} E[\overline{X}] &= E[\tfrac{X_1}{n}] + \dots + E[\tfrac{X_n}{n}] \\ &= \sum_{i=1}^{n} E[\tfrac{X_i}{n}] \\ &= \sum_{i=1}^{n} \frac{p}{n} = p \end{aligned}$$

So the expectation of the estimator $\overline{X}$ is the parameter.

Note that

$$\begin{aligned} Var(X_1+X_2) &= E[(X_1+X_2-E[X_1+X_2])^2] \\ &= E[((X_1-E[X_1])+(X_2-E[X_2]))^2] \end{aligned}$$

We have

$X_1 \backslash X_2$	1	0
1	$p \cdot p$	$p(1-p)$
0	$(1-p)p$	$(1-p)(1-p)$

So we compute

$$\begin{aligned} &= p^2((1-p)+(1-p))^2 + \\ &\quad 2p(1-p)((1-p)+(0-p))^2 + \\ &\quad (1-p)^2((0-p)+(0-p))^2 \\ &= 4p^2(1-p)^2 + 2p(1-p)(1-2p)^2 + \\ &\quad + 4p^2(1-p)^2 \\ &= 8p^2(1-p)^2 + 2p(1-p) - 8p^2(1-p)^2 \\ &= 2p(1-p) \end{aligned}$$

Therefore $Var(X_1+X_2) = Var(X_1) + Var(X_2)$.

In general variance of a summation is summation of variance for <u>independent</u> observations. So

$$\begin{aligned} Var(\overline{X}) &= Var(\tfrac{X_1}{n}) + \dots + Var(\tfrac{X_n}{n}) \\ &= \sum_{i=1}^{n} Var(\tfrac{X_i}{n}) \\ &= \sum_{i=1}^{n} (\tfrac{1}{n} - \tfrac{p}{n})^2 p + (0 - \tfrac{p}{n})^2(1-p) \\ &= \sum_{i=1}^{n} \frac{1}{n^2} p(1-p) \\ &= \frac{n}{n^2} p(1-p) = \frac{Var(X_1)}{n} \end{aligned}$$

Note that for large values of n the variance of $\overline{X}$ is small. So the larger the sample the closer the estimator will be to its expectation

$$E[\overline{X}] = p$$