Probability and Statistics: MA6.101

Tutorial 11

Topics Covered: Probability Inequalities, Statistics (Classical and Bayesian)

Probability Inequalities

Q1: A biased coin, which lands heads with probability 1/10 each time it is flipped, is flipped 200 times consecutively. Give an upper bound on the probability that it lands heads at least 120 times.

A: The number of heads is a binomially distributed random variable, X, with parameters $p = 1/10$ and $n = 200$.
Thus, the expected number of heads is

$$E(X) = np = 200 \cdot (1/10) = 20.$$

By **Markov Inequality**, the probability of at least 120 heads is

$$P(X \geq 120) \leq \frac{E(X)}{120} = \frac{20}{120} = \frac{1}{6}.$$

Q2: Show that convergence in mean square implies convergence in probability.

A:
We can apply the Markov inequality to a generic term of the sequence $\{(X_n - X)^2\}$:

$$P\big((X_n - X)^2 \geq \epsilon^2\big) \leq \frac{\mathbb{E}\big[(X_n - X)^2\big]}{\epsilon^2}$$

For any strictly positive real number ϵ, taking the square root of both sides of the left-hand inequality, we obtain:

$$P(|X_n - X| \geq \epsilon) \leq \frac{\mathbb{E}\big[(X_n - X)^2\big]}{\epsilon^2}$$

Taking limits on both sides, we get:

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\mathbb{E}\big[(X_n - X)^2\big]}{\epsilon^2} = \frac{\lim_{n\to\infty} \mathbb{E}\big[(X_n - X)^2\big]}{\epsilon^2} = 0$$

Since,

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0$$

And by the definition of probability,

$$P(|X_n - X| \geq \epsilon) \geq 0$$

Then it must be that also:

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

Q3: Chebyshev Inequality states that:
If X is any random variable, then for any $b > 0$ we have:

$$P\big(|X - EX| \geq b\big) \leq \frac{Var(X)}{b^2}.$$

Let $X \sim Binomial(n, p)$. Using Chebyshev's inequality, find an upper bound on $P(X \geq \alpha n)$, where $p < \alpha < 1$. Evaluate the bound for $p = \frac{1}{2}$ and $= \frac{3}{4}$. Calculate using Markov's inequality for the similar parameters, and comment on the betterness of the bounds obtained.

A: We can write this bound as

$$\begin{aligned}P(X \geq \alpha n) &= P(X - np \geq \alpha n - np)\\ &\leq P\big(|X - np| \geq n\alpha - np\big)\\ &\leq \frac{Var(X)}{(n\alpha - np)^2}\\ &= \frac{p(1-p)}{n(\alpha - p)^2}.\end{aligned}$$

Substituting the values for p and α we get:

$$P(X \geq \frac{3n}{4}) \leq \frac{4}{n}.$$

On applying Markov's inequality, with $E[X] = np$

$$\begin{aligned}P(X \geq \alpha n) &\leq \frac{E[X]}{\alpha n}\\ &= \frac{pn}{\alpha n} = \frac{p}{\alpha}\end{aligned}$$

Substituting the values for p and α we get:

$$P(X \geq \frac{3n}{4}) \leq \frac{2}{3}.$$

We can see that Chebyshev gives a better and stronger bound than Markov. It is constant and does not change as n increases.

Statistics

Classical/Frequentist Methods

Q1: Prove that maximising the likelihood is the same as maximising the log likelihood. [Hint: Start by taking 2 different values of the parameter; one maximises the likelihood function, and the other maximises the log-likelihood function].

A: Let $L(\theta)$ = the likelihood function for parameter θ given data D: $L(\theta) = p(D|\theta)$ The log-likelihood function,

$$\ell(\theta) = \log L(\theta) = \log p(D|\theta)$$

Proof by Contradiction: Assume there exist two different parameter values θ_1 and θ_2 such that:

$$\begin{gathered}L(\theta_1) > L(\theta_2)\\ l(\theta_2) > \ell(\theta_1)\end{gathered}$$

The logarithm function $\log(x)$ is monotonically increasing. Hence if

$$L(\theta_1) > L(\theta_2)$$

then

$$\log L(\theta_1) > \log L(\theta_2)$$

which implies

$$\ell(\theta_1) > \ell(\theta_2)$$

Note: $\log(x)$ is monotonically increasing (can be verified from derivative) From our assumption,

$$L(\theta_1) > L(\theta_2) \implies \ell(\theta_1) > \ell(\theta_2)$$

However, this contradicts the initial assumption that $\ell(\theta_2) > \ell(\theta_1)$.

Q2: Let $X_1, X_2, \dots, X_n$ be a random sample from a *Geometric*(p) distribution. Suppose we observe the data $\{x_1, x_2, x_3, x_4\} = \{2, 1, 7, 3\}$. The probability mass function of the Geometric distribution is given by:

$$P_X(x;p) = p(1-p)^{x-1}$$

Find the maximum likelihood estimate (MLE) of p

A: The likelihood function for the sample $\{x_1, x_2, \dots, x_n\}$ is:

$$L(x_1, \dots, x_n; p) = \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{\sum_{i=1}^{n} x_i - n}.$$

The MLE of p, denoted as $\hat{p}_{ML}$, is given by:

$$\hat{p}_{ML} = \arg\max_p L(x_1, \dots, x_n; p)$$

$$= \arg\max_p \log L(x_1, \dots, x_n; p) \text{ [Since, log is monotonically increasing]}$$

$$= \arg\max_p \left(n\log(p) + \left(\sum_{i=1}^{n} x_i - n \right) \log(1-p) \right)$$

To find the MLE, we differentiate $\log L(p)$ with respect to p, set the derivative to zero, and solve:

$$\frac{d}{dp} \log L(p) = \frac{n}{p} - \frac{\sum_{i=1}^{n} x_i - n}{1-p} = 0.$$

Simplifying:

$$\frac{n}{p} = \frac{\sum_{i=1}^{n} x_i - n}{1-p},$$

$$n(1-p) = p\left(\sum_{i=1}^{n} x_i - n \right),$$

$$n = p\sum_{i=1}^{n} x_i.$$

Thus, the MLE is:

$$\hat{p}_{ML} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{4}{13}$$

Q3: In the Cilantro experiment, assume 55 out of 100 people said Cilantro tastes like soap. Find the maximum likelihood estimate for p, the true proportion of people who feel that way.

A: The likelihood function for p is

$$L(p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

Because the log function turns multiplication into addition, it is often convenient to use the log of the likelihood function:

$$\text{log likelihood} = \ln(\text{likelihood}) = \ln(P(\text{data} \mid p)).$$

In our example:

$$\text{Log likelihood } l(p) = \ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p).$$

Now we can set the derivative of $l(p)$ to 0 to find the MLE:

$$l'(p) = \frac{55}{p} - \frac{45}{1-p} = 0.$$

This is easy to solve for p. We get $\hat{p} = 0.55$.

Adding a hat is a standard way of indicating an estimate, i.e., $\hat{p}$ is an estimate of the unknown parameter p.

Q4: Let $\mathcal{D} = \{x_1, ..., x_n\}$ denote i.i.d samples from a uniform random variable $U[0, a]$, where a is unknown. Find an MLE estimate for the unknown parameter a.

A:

Since we know that the samples are drawn from the uniform distribution $U[0, a]$, we have that the PDF of a sample x_i is given by

$$f(x_i) = \begin{cases} \frac{1}{a} & \text{if } 0 \leq x_{min} \leq x_{max} \leq a \\ 0 & \text{otherwise} \end{cases}$$

Now, the likelihood function $\mathcal{L}$ is given by

$$\mathcal{L}(x_1, ..., x_n; a) = P(x_1, ..., x_n; a)$$

Since we have that the samples are i.i.d.,

$$P(x_1, ..., x_n; a) = \prod_{i=1}^{n} P(x_i; a)$$

$$\Rightarrow \mathcal{L}(x_1, ..., x_n; a) = \prod_{i=1}^{n} P(x_i; a) = \frac{1}{a^n}$$

$$\Rightarrow \mathcal{L}(x; a) = \frac{1}{a^n} \quad \forall 0 \leq x \leq a$$

On calculating the derivative of $\mathcal{L}(x; a)$ with respect to a and setting it to 0 to maximize $\mathcal{L}(a)$, we get the following expression:

$$\frac{d\mathcal{L}}{da} = \frac{-n}{a^{n+1}} \neq 0$$

So, we analyze the expression on the RHS of the likelihood function and we can infer that $\frac{1}{a^n}$ is a decreasing function. Hence, the maximum value of the likelihood function will be at the minimum possible value of a, i.e., $\max(L) = \min(a)$.

We also have to note the constraint on a that $0 \leq x_{min} \leq x_{max} \leq a$ and hence, the maximum likelihood estimate for the unknown parameter a is given by

$$a_{MLE} = \max(x_1, ..., x_n)$$

Q5: Assume our data $\mathbf{Y} = (y_1, y_2, ..., y_n)^T$ given X is independently identically distributed, i.i.d. $Y|X = x \sim \mathit{Exponential}(\lambda = x)$, and we chose the prior to be $X \sim \mathit{Gamma}(\alpha, \beta)$.

a. Find the likelihood of the function, $L(\mathbf{Y}; X) = f_{\mathbf{Y}|X}(y_1, y_2, ..., y_n|x)$.

b. Using the likelihood function of the data, show that the posterior distribution is $\mathit{Gamma}(\alpha + n, \beta + \sum_{i=1}^n y_i)$.

c. Write out the PDF for the posterior distribution, $f_{X|\mathbf{Y}}(x|\mathbf{y})$.

A:

(a) Since $Y_i|X = x \sim \mathit{Exp}(x)$, We have that

$$f_{Y_i|X}(y|x) = \begin{cases} xe^{-xy} & \text{for } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood function is thus:

$$\begin{aligned} L(Y; x) &= f_{Y_1,Y_2,\ldots,Y_n|X}(y_1, y_2, \ldots, y_n|x) \\ &= \prod_{i=1}^n f_{Y_i|X}(y_i|x) \quad \text{(by independence)} \\ &= \prod_{i=1}^n xe^{-xy_i} \\ &= x^n e^{-x\sum_{i=1}^n y_i}. \end{aligned}$$

(b) Since $X \sim \mathit{Gamma}(\alpha, \beta)$, We have that $f_X(x) \propto x^{\alpha-1}e^{-\beta x}$ for $x > 0$ and $f_X(x) = 0$ otherwise. Therefore, for $x > 0$,the posterior is:

$$\begin{aligned} f_{X|Y_1,Y_2,\ldots,Y_n}(x|y_1, y_2, \ldots, y_n) &\propto f_{Y_1,Y_2,\ldots,Y_n|X}(y_1, y_2, \ldots, y_n|x)f_X(x) \\ &= L(Y; x)f_X(x) \\ &\propto x^n e^{-x\sum_{i=1}^n y_i} x^{\alpha-1}e^{-\beta x} \\ &= x^{\alpha+n-1}e^{-x\left(\sum_{i=1}^n y_i + \beta\right)}, \end{aligned}$$

while for $x \leq 0$, $f_{X|Y_1,Y_2,\ldots,Y_n}(x|y_1, y_2, \ldots, y_n) = 0$. Now, since α and n are greater than 0, so too is $\alpha + n$. Further, since it is assumed that $Y_i|X \sim \mathit{Exp}(X)$, it is implicit that all y_is are greater than 0, and since $\beta > 0$, so too is $\sum_{i=1}^n y_i + \beta$. Therefore, we see that, up to a normalizing constant, the posterior has the functional form of a $\mathit{Gamma}(\alpha + n, \sum_{i=1}^n y_i + \beta)$ distribution, so that $X|Y = y \sim \mathit{Gamma}(\alpha + n, \sum_{i=1}^n y_i + \beta)$.

(c) The posterior PDF is given by:

$$f_{X|Y_1,Y_2,\ldots,Y_n}(x|y_1, y_2, \ldots, y_n) = \begin{cases} \frac{(\sum_{i=1}^n y_i + \beta)^{\alpha+n} x^{\alpha+n-1} e^{-x(\sum_{i=1}^n y_i + \beta)}}{\Gamma(\alpha+n)} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Q6: Let $X_1, \ldots, X_n$ be a random sample from a Poisson(λ) distribution.

(a) Find the likelihood function, $L(x_1, \dots, x_n; \lambda)$

(b) Find the log-likelihood function and use that to obtain the MLE for λ, $\hat{\lambda}_{\text{ML}}$.

A:

(a)

$$\begin{aligned} L(x_1, \dots, x_n; \lambda) &= \prod_{i=1}^{n} P_{X_i}(x_i; \lambda) \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \frac{1}{x_i!} \end{aligned}$$

(b) The log-likelihood, $\mathcal{L}$, is:

$$\begin{aligned} \mathcal{L}(\lambda) &= \ln L(x_1, \dots, x_n; \lambda) \\ &= -n\lambda + \sum_{i=1}^{n} x_i \ln \lambda - \sum_{i=1}^{n} \ln x_i! \end{aligned}$$

Differentiating this with respect to λ, and setting it equal to zero, we have:

$$0 = -n + \frac{1}{\hat{\lambda}_{\text{ML}}} \sum_{i=1}^{n} x_i$$

Solving for the maximum likelihood estimate, we have that:

$$\hat{\lambda}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

That is, the maximum likelihood estimate of λ is simply the sample mean.

Q7: Suppose you are trying to fit a distribution

$$P(X = x|\theta) = (\tfrac{\theta}{2})^{|x|}(1 - \theta)^{1-|x|}$$

where the support set is $\{-1, 0, 1\}$ and θ is a real number in the range $[0, 1]$.

(a) Define an estimator as

$$T(X) = 2 \text{ if } x = 1 \text{ otherwise } T(X) = 0.$$

Show that $T(X)$ is an unbiased estimator of θ.

(b) Now define another estimator as $G(X) = |X|$. Show that $G(x)$ is also an unbiased estimator of θ.

(c) Which of the estimators is better for θ? Justify your answer.

A:

(a) Note that the distribution for $T(X)$ is given by:

$$T(X) = \begin{cases} 2 & \text{w.p. } \frac{\theta}{2} \\ 0 & \text{w.p. } 1 - \frac{\theta}{2} \end{cases}$$

So, the expectation of $T(X)$ is

$$E[T(X)] = 2\left(\frac{\theta}{2}\right) + 0\left(1 - \frac{\theta}{2}\right) = \theta$$

So, $T(X)$ is an unbiased estimator of θ.

(b) Note that the distribution for $G(X)$ is given by:

$$G(X) = \begin{cases} 1 & \text{w.p. } \theta \\ 0 & \text{w.p. } 1-\theta \end{cases}$$

$G(X)$ is hence a Bernoulli random variable. So, we get $E[G] = \theta$. So, G is also an unbiased estimator of θ.

(c) Since both estimators are unbiased, we need to compare their variances.
For $T(X)$,the variance is $\sigma_T^2 = (2-\theta)\theta$.
For $G(X)$, it is a Bernoulli random variable so we get $\sigma_G^2 = (1-\theta)\theta$.
Clearly, the variance of G is smaller than T so we can say that G is a better estimator than T, even though both are unbiased estimators.

Q8: Using a rod of length μ, you lay out a square plot whose length of each side is μ. Thus, the area of the plot will be μ^2 (unknown). Based on n independent measurements $X_1, X_2, \ldots, X_n$ of the length, estimate μ^2. Assume that each X_i has mean μ and variance σ^2.

(a) Show that $\overline{X}^2$ is not an unbiased estimator for μ^2.

(b) For what value of k is the estimator $\overline{X}^2 - kS^2$ unbiased for μ^2?

[Divyaraj]

A:

(a) Note that:

$$\mathbb{E}(\overline{X}^2) = \text{Var}(\overline{X}) + [\mathbb{E}(\overline{X})]^2 = \frac{\sigma^2}{n} + \mu^2.$$

Thus, the bias of the estimator $\overline{X}^2$ is:

$$\mathbb{E}(\overline{X}^2 - \mu^2) = \frac{\sigma^2}{n}.$$

Therefore, $\overline{X}^2$ tends to overestimate μ^2.

(b) Consider the estimator $\overline{X}^2 - kS^2$. Its expected value is:

$$\mathbb{E}(\overline{X}^2 - kS^2) = \mathbb{E}(\overline{X}^2) - k\mathbb{E}(S^2).$$

Substituting the known expectations:

$$\mathbb{E}(\overline{X}^2) = \mu^2 + \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2) = \sigma^2,$$

we get:

$$\mathbb{E}(\overline{X}^2 - kS^2) = \mu^2 + \frac{\sigma^2}{n} - k\sigma^2.$$

To make the estimator unbiased, set:

$$\mu^2 + \frac{\sigma^2}{n} - k\sigma^2 = \mu^2.$$

Simplifying:

$$\frac{\sigma^2}{n} - k\sigma^2 = 0 \implies k = \frac{1}{n}.$$

Hence, with $k = \frac{1}{n}$, the estimator $\overline{X}^2 - kS^2$ is unbiased for μ^2.

Q9: Consider an experiment with two possible outcomes:

- **{Success}**, with probability p
- **{Failure}**, with probability $1 - p$

The probability of success, p, is known to lie within the interval $\left[\frac{1}{10}, \frac{1}{5}\right]$. Suppose the experiment is repeated independently n times. The estimator for p is defined as:

$$\widehat{p} = \frac{\text{Successes obtained}}{\text{Total experiments performed}}.$$

Our objective is to determine the minimum value of n such that the standard deviation of the estimator $\widehat{p}$ is guaranteed to be less than $\frac{1}{100}$ for all $p \in \left[\frac{1}{10}, \frac{1}{5}\right]$. [Divyaraj]

A:
The estimator $\widehat{p}$ can be written as:

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

where n is the number of repetitions of the experiment, and X_i are n independent random variables, each following a Bernoulli distribution with parameter p.
The variance of $\widehat{p}$ is given by:

$$\operatorname{Var}(\widehat{p}) = \frac{\operatorname{Var}(X_i)}{n} = \frac{p(1-p)}{n}.$$

Since $p \in \left[\frac{1}{10}, \frac{1}{5}\right]$, the maximum value of $p(1-p)$ occurs at $p = \frac{1}{5}$:

$$p(1-p) = \frac{1}{5}\left(1 - \frac{1}{5}\right) = \frac{4}{25}.$$

Thus,

$$\operatorname{Var}(\widehat{p}) \leq \frac{4}{25n}.$$

To ensure that the standard deviation of $\widehat{p}$ is less than $\frac{1}{100}$, we need:

$$\sqrt{\operatorname{Var}(\widehat{p})} \leq \frac{1}{100}.$$

Squaring both sides:

$$\operatorname{Var}(\widehat{p}) \leq \frac{1}{10,000}.$$

Substituting the maximum variance:

$$\frac{4}{25n} \leq \frac{1}{10,000}.$$

Simplifying:

$$4 \cdot 10,000 \leq 25n \implies n \geq \frac{40,000}{25} = 1,600.$$

Thus, the minimum number of experiments required is:

$$\boxed{n = 1,600}.$$

Q10: Let $X_1, X_2, \dots, X_n$ be a random sample from the following distribution:

$$f_X(x) = \begin{cases} \theta\left(x - \frac{1}{2}\right) + 1 & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in [-2, 2]$ is an unknown parameter. We define the estimator $\hat{\Theta}_n$ as:

$$\hat{\Theta}_n = 12\overline{X} - 6$$

to estimate θ.

(a) Is $\hat{\Theta}_n$ an unbiased estimator of θ?

(b) Is $\hat{\Theta}_n$ a consistent estimator of θ?

(c) Find the mean squared error (MSE) of $\hat{\Theta}_n$.

A:

(a) To see this, we write:

$$\mathbb{E}[\hat{\Theta}_n] = \mathbb{E}[12\overline{X} - 6] = 12\mathbb{E}[\overline{X}] - 6.$$

Since:

$$\mathbb{E}[\overline{X}] = \mathbb{E}[X] = \frac{12 \cdot \theta + 6}{12},$$

it follows that:

$$\mathbb{E}[\hat{\Theta}_n] = 12 \cdot \frac{\theta + 6}{12} - 6 = \theta.$$

Thus, $\hat{\Theta}_n$ **is an unbiased estimator of** θ.

(b) To show that $\hat{\Theta}_n$ is a consistent estimator of θ, we need to show:

$$\lim_{n \to \infty} P(|\hat{\Theta}_n - \theta| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0.$$

Since:

$$\hat{\Theta}_n = 12\overline{X} - 6 \quad \text{and} \quad \mathbb{E}[\overline{X}] = \mathbb{E}[X],$$

we conclude:

$$P(|\hat{\Theta}_n - \theta| \geq \epsilon) = P(12|\overline{X} - \mathbb{E}[X]| \geq \epsilon) = P(|\overline{X} - \mathbb{E}[X]| \geq \epsilon/12),$$

which goes to zero as $n \to \infty$ by the law of large numbers. Therefore, $\hat{\Theta}_n$ is a **consistent estimator of** θ.

(c) To find the mean squared error (MSE) of $\hat{\Theta}_n$, we write:

$$\text{MSE}(\hat{\Theta}_n) = \text{Var}(\hat{\Theta}_n) + B(\hat{\Theta}_n)^2.$$

Since $\hat{\Theta}_n$ is unbiased, the bias term $B(\hat{\Theta}_n)$ is 0. Thus:

$$\text{MSE}(\hat{\Theta}_n) = \text{Var}(\hat{\Theta}_n).$$

Now:

$$\text{Var}(\hat{\Theta}_n) = \text{Var}(12\overline{X} - 6) = 144\text{Var}(\overline{X}).$$

Using:

$$\text{Var}(\overline{X}) = \frac{\text{Var}(X)}{n},$$

we get:

$$\mathrm{Var}(X) = \frac{12 - \theta^2}{12},$$

and so:

$$\mathrm{Var}(\hat{\Theta}_n) = \frac{144(12 - \theta^2)}{12n} = \frac{12 - \theta^2}{n}.$$

Therefore:

$$\mathrm{MSE}(\hat{\Theta}_n) = \frac{12 - \theta^2}{n}.$$

Note that this gives us another way to argue that $\hat{\Theta}_n$ is a consistent estimator of θ. In particular, since:

$$\lim_{n \to \infty} \mathrm{MSE}(\hat{\Theta}_n) = 0,$$

we conclude that $\hat{\Theta}_n$ is a consistent estimator of θ.

Q11: Let $X_1, X_2, X_3, ..., X_n$ be a random sample from a distribution with mean $E[X_i] = \theta$ and variance $Var(X_i) = \sigma^2$. Consider the following two estimators for θ:

1. $\hat{\Theta}_1 = X_1$ 2. $\hat{\Theta}_2 = \overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$

Find $MSE(\hat{\Theta}_1)$ and $MSE(\hat{\Theta}_2)$ and show that for $n > 1$ we have $MSE(\hat{\Theta}_1) > MSE(\hat{\Theta}_2)$.

A:

We have

$$MSE(\hat{\Theta}_1) = E[(\hat{\Theta}_1 - \theta)^2] = E[(X_1 - EX_1)^2] = Var(X_1) = \sigma^2.$$

To find $MSE(\hat{\Theta}_2)$, we can write

$$MSE(\hat{\Theta}_2) = E[(\hat{\Theta}_2 - \theta)^2] = E[(\overline{X} - \theta)^2] = Var(\overline{X} - \theta) + (E[\overline{X} - \theta])^2.$$

Since $E[Y^2] = Var(Y) + (E[Y])^2$, where $Y = \overline{X} - \theta$. Now, note that $Var(\overline{X} - \theta) = Var(\overline{X})$ since θ is a constant. Also, $E[\overline{X} - \theta] = 0$. Thus, we conclude

$$MSE(\hat{\Theta}_2) = Var(\overline{X}) = \frac{\sigma^2}{n}.$$

Therefore, for $n > 1$, we have

$$MSE(\hat{\Theta}_1) > MSE(\hat{\Theta}_2).$$

Bayesian Inference

Q12: Let $X \sim Uniform(0, 1)$. Suppose that we know

$$Y|X = x \sim Geometric(x).$$

Find the posterior density of X given $Y = 2$, $f_{X|Y}(x|2)$.

A:

Using Bayes' rule we have

$$f_{X|Y}(x|2) = \frac{P_{Y|X}(2|x) f_X(x)}{P_Y(2)}.$$

We know $Y|X = x \sim Geometric(x)$, so

$$P_{Y|X}(y|x) = x(1-x)^{y-1}, \quad \text{for } y = 1, 2, ...$$

Therefore,

$$P_{Y|X}(2|x) = x(1-x).$$

To find $P_Y(2)$, we can use the law of total probability

$$P_Y(2) = \int_{-\infty}^{\infty} P_{Y|X}(2|x) f_X(x) dx$$

$$= \int_0^1 x(1-x) \cdot 1 dx$$

$$= \frac{1}{6}.$$

Therefore, we obtain

$$f_{X|Y}(x|2) = \frac{x(1-x) \cdot 1}{\frac{1}{6}}$$

$$= 6x(1-x), \quad \text{for } 0 \le x \le 1.$$

Q13: Let X be a continuous random variable with the following PDF:

$$f_X(x) = \begin{cases} 3x^2 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Also, suppose that $Y|X = x \sim Geometric(x)$. Find the MAP estimate of X given $Y = 5$.

A: From Bayes' rule, we know that the posterior density for $0 \le x \le 1$ is:

$$f_{X|Y}(x|5) \propto P_{Y|X}(5|x) f_X(x) = 3x^2 \cdot x(1-x)^4 = 3x^3(1-x)^4$$

(and 0 otherwise), where the symbol $\propto$ means proportional to as a function of x. Therefore, the MAP estimate is given by

$$\hat{x}_{MAP} = \arg\max_x \{3x^3(1-x)^4\}$$

which can be found by setting the derivative of the argument equal to zero and solving for x:

$$0 = 9\hat{x}_{MAP}^2(1-\hat{x}_{MAP})^4 - 12\hat{x}_{MAP}^3(1-\hat{x}_{MAP})^3$$

$$\Rightarrow \hat{x}_{MAP} = \frac{3}{7}$$

This value is indeed in the interval $[0, 1]$.

Q14: Suppose $D = \{x_1, ..., x_n\}$ is a data set consisting of independent samples of a Bernoulli random variable with unknown parameter θ, i.e., $f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$ for $x_i \in \{0, 1\}$. We are also given that $\theta \sim U[0, 1]$. Obtain an expression for the posterior distribution on θ. Using this, obtain $\hat{\theta}_{MAP}$ and the conditional expectation estimator $\hat{\theta}_{CE}$.

(Hint: $\int_0^1 \theta^m(1-\theta)^r d\theta = \frac{m!r!}{(m+r+1)!}$)

A:

$$\begin{aligned} f(\theta|x_1,...,x_n) &= \frac{f(x_1,...,x_n,\theta)}{f(x_1,...,x_n)} \\ &= \frac{f(x_1,...,x_n|\theta)p(\theta)}{\int_0^1 f(x_1,...,x_n|\theta)p(\theta)d\theta} \\ &= \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}d\theta} \end{aligned}$$

Now using the fact that for integral values m and r

$$\int_0^1 \theta^m(1-\theta)^r d\theta = \frac{m!r!}{(m+r+1)!}$$

and letting $s = \sum_{i=1}^n x_i$, the expression for the posterior becomes:

$$f(\theta|x_1,...,x_n) = \frac{(n+1)!\theta^s(1-\theta)^{n-s}}{s!(n-s)!}$$

Now, solving for $\hat{\theta}_{MAP}$:

$$\begin{aligned} \hat{\theta}_{MAP} &= \arg\max_\theta P(x_1,\ldots,x_n|\theta) \\ &= \arg\max_\theta \theta^s(1-\theta)^{n-s} \end{aligned}$$

Taking the derivative and setting to 0, we get:

$$\begin{aligned} s\theta^{s-1}(1-\theta)^{n-s} - (n-s)\theta^s(1-\theta)^{n-s-1} &= 0 \\ \theta^{s-1}(1-\theta)^{n-s-1}[s(1-\theta)-(n-s)\theta] &= 0 \end{aligned}$$

Since θ cannot take the value of 0 or 1 for maximizing the posterior, the second part of the above expression must go to 0.

$$\begin{aligned} s(1-\theta)-(n-s)\theta &= 0 \\ s - s\theta - n\theta + s\theta &= 0 \\ \theta &= \frac{s}{n} \end{aligned}$$

We got that the MAP estimate for θ is $\frac{s}{n}$, which is just the sample mean $\bar{X}$ of the data. For θ_{CE}, we have:

$$\begin{aligned} \theta_{CE} &= E[\theta|x_1,\ldots,x_n] \\ &= \int_0^1 \theta f(\theta|x_1,\ldots,x_n)d\theta \\ &= \frac{(n+1)!}{s!(n-s)!}\int_0^1 \theta^{1+s}(1-\theta)^{n-s}d\theta \\ &= \frac{(n+1)!}{s!(n-s)!}\frac{(1+s)!(n-s)!}{(n+2)!} \\ &= \frac{s+1}{n+2} \end{aligned}$$

Q15: Suppose we have a prior $\theta \sim \mathcal{N}(4, 8)$, and the likelihood function is $\phi(x_1|\theta) \sim \mathcal{N}(\theta, 5)$. Suppose also that we have one measurement $x_1 = 3$. Show that the posterior distribution is normal.

[Divyaraj]

A:

(a) **Prior:**

$$f(\theta) = c_1 e^{-(\theta-4)^2/16}$$

(b) **Likelihood:**

$$\phi(x_1|\theta) = c_2 e^{-(x_1-\theta)^2/10} = c_2 e^{-(3-\theta)^2/10}$$

(c) **Posterior:**
We multiply the prior and likelihood to get the posterior:

$$f(\theta|x_1) = c_3 e^{-(\theta-4)^2/16} e^{-(3-\theta)^2/10} = c_3 \exp\left(-\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10}\right)$$

We complete the square in the exponent:

$$-\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10} = \frac{-5(\theta-4)^2}{80} + \frac{-8(3-\theta)^2}{80}$$
$$= \frac{-130\theta^2 + 88\theta + 152}{80} = \frac{-\theta^2 + \frac{88}{13}\theta + \frac{152}{13}}{80/13}$$

Therefore, the posterior is:

$$f(\theta|x_1) = c_4 e^{-(\theta - \frac{44}{13})^2 \frac{80}{13}}$$

This has the form of the probability density function for $\mathcal{N}(\frac{44}{13}, \frac{40}{13})$.

Q16: Suppose that the signal $X \sim N(0, \sigma_X^2)$, is transmitted over a communication channel. Assume that the received signal is given by:

$$Y = aX + bW \quad \text{where} \quad W \sim N(0, \sigma_W^2)$$

and W is independent of X.

(a) Find the ML estimate of X, given Y=y is observed.

(b) Find the MAP estimate of X, given Y=y is observed.

A:

(a) Proving the distribution using MGF:
The moment generating function of $W \sim N(0, \sigma_W^2)$ is:

$$M_W(t) = \exp\left(\frac{t^2\sigma_W^2}{2}\right).$$

The received signal is:

$$Y = aX + bW.$$

Conditional on $X = x$, Y is a linear transformation of W:

$$Y \mid X = x = ax + bW.$$

The MGF of $Y \mid X = x$ is:

$$M_{Y|X=x}(t) = \mathbb{E}[\exp(t(ax + bW))].$$

Substitute $Y = ax + bW$:

$$M_{Y|X=x}(t) = \exp(tax) \cdot M_W(bt).$$

Since $M_W(t) = \exp(t^2\sigma_W^2/2)$:

$$M_{Y|X=x}(t) = \exp(tax) \cdot \exp\left(\frac{(bt)^2\sigma_W^2}{2}\right).$$

Simplify:

$$M_{Y|X=x}(t) = \exp\left(tax + \frac{b^2t^2\sigma_W^2}{2}\right).$$

This is the MGF of a normal distribution:

$$Y \mid X = x \sim N(ax, b^2\sigma_W^2).$$

(b) The log-likelihood function is:

$$\log f_{Y|X}(y \mid x) = -\frac{1}{2}\log(2\pi b^2\sigma_W^2) - \frac{(y-ax)^2}{2b^2\sigma_W^2}.$$

Ignoring the constant terms, this reduces to:

$$\mathcal{L}(x) = -\frac{(y-ax)^2}{2b^2\sigma_W^2}.$$

To maximize the log-likelihood, minimize the squared error term:

$$(y-ax)^2.$$

Differentiating with respect to x and setting the derivative to zero:

$$\frac{d}{dx}(y-ax)^2 = -2a(y-ax) = 0 \implies x = \frac{y}{a}.$$

Thus, the ML estimate is:

$$\hat{X}_{\text{ML}} = \frac{y}{a}.$$

(c) Using Bayes' theorem:

$$f_{X|Y}(x \mid y) \propto f_{Y|X}(y \mid x) f_X(x),$$

where:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{x^2}{2\sigma_X^2}\right),$$

and:

$$f_{Y|X}(y \mid x) \propto \exp\left(-\frac{(y-ax)^2}{2b^2\sigma_W^2}\right).$$

Thus:

$$f_{X|Y}(x \mid y) \propto \exp\left(-\frac{(y-ax)^2}{2b^2\sigma_W^2} - \frac{x^2}{2\sigma_X^2}\right).$$

The log-posterior is:

$$\log f_{X|Y}(x \mid y) = -\frac{(y-ax)^2}{2b^2\sigma_W^2} - \frac{x^2}{2\sigma_X^2} + \text{constant}.$$

Simplify:

$$\log f_{X|Y}(x \mid y) = -\frac{a^2}{2b^2\sigma_W^2}x^2 + \frac{ay}{b^2\sigma_W^2}x - \frac{y^2}{2b^2\sigma_W^2} - \frac{1}{2\sigma_X^2}x^2 + \text{constant}.$$

Group the quadratic terms:

$$\log f_{X|Y}(x \mid y) = -\frac{1}{2}\left(\frac{a^2}{b^2\sigma_W^2} + \frac{1}{\sigma_X^2}\right)x^2 + \frac{ay}{b^2\sigma_W^2}x + \text{constant}.$$

The posterior is maximized at:

$$x_{\text{MAP}} = -\frac{B}{2A},$$

where:

$$A = \frac{1}{2}\left(\frac{a^2}{b^2\sigma_W^2} + \frac{1}{\sigma_X^2}\right), \quad B = \frac{ay}{b^2\sigma_W^2}.$$

Simplify:

$$x_{\text{MAP}} = \frac{\frac{ay}{b^2\sigma_W^2}}{\frac{a^2}{b^2\sigma_W^2} + \frac{1}{\sigma_X^2}} = \frac{ay\sigma_X^2}{a^2\sigma_X^2 + b^2\sigma_W^2}.$$

Thus, the MAP estimate is:

$$\hat{X}_{\text{MAP}} = \frac{ay\sigma_X^2}{a^2\sigma_X^2 + b^2\sigma_W^2}.$$

Q17: Let Θ be a continuous random variable with pdf as $\frac{1}{6}$ for $\theta \in [4, 10]$ and 0 elsewhere. And we know that $X = \Theta + U[-1, 1]$. Find the conditional expectation estimator.

A:

We aim to derive the conditional expectation estimator $\mathbb{E}[\Theta \mid X = x]$ for the random variables Θ and X, given:

1. $\Theta \sim \text{Uniform}[4, 10]$, with pdf:

$$f_\Theta(\theta) = \begin{cases} \frac{1}{6}, & 4 \le \theta \le 10, \\ 0, & \text{otherwise.} \end{cases}$$

2. $X = \Theta + U$, where $U \sim \text{Uniform}[-1, 1]$, and U is independent of Θ.

(a) The random variable X has pdf determined by the convolution of $f_\Theta(\theta)$ and $f_U(u)$. First, compute the conditional pdf $f_{X|\Theta}(x \mid \theta)$. Since $X = \Theta + U$, and $U \sim \text{Uniform}[-1, 1]$, we have:

$$f_{X|\Theta}(x \mid \theta) = f_U(x - \theta) = \begin{cases} \frac{1}{2}, & \theta - 1 \le x \le \theta + 1, \\ 0, & \text{otherwise.} \end{cases}$$

The joint pdf $f_{X,\Theta}(x, \theta)$ is:

$$f_{X,\Theta}(x, \theta) = f_\Theta(\theta) f_{X|\Theta}(x \mid \theta).$$

Substituting the values:

$$f_{X,\Theta}(x, \theta) = \begin{cases} \frac{1}{12}, & 4 \le \theta \le 10 \text{ and } \theta - 1 \le x \le \theta + 1, \\ 0, & \text{otherwise.} \end{cases}$$

(b) To find $f_X(x)$, integrate $f_{X,\Theta}(x,\theta)$ over all θ:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,\Theta}(x,\theta)\, d\theta.$$

The range of θ depends on x because $x \in [\theta - 1, \theta + 1]$. Thus, $\theta \in [x - 1, x + 1]$. Additionally, since $\theta \in [4, 10]$, the valid limits for θ are:

$$\max(4, x - 1) \leq \theta \leq \min(10, x + 1).$$

Now, compute $f_X(x)$:

$$f_X(x) = \int_{\max(4,x-1)}^{\min(10,x+1)} \frac{1}{12}\, d\theta.$$

Evaluate the integral:

$$f_X(x) = \begin{cases} 0, & x \notin [3, 11], \\ \frac{\min(10,x+1)-\max(4,x-1)}{12}, & x \in [3, 11]. \end{cases}$$

Note: The cases in which the pdf gets divided can be seen from the max and min conditions. Check where the conditions change, those are your case boundaries! Simplify $f_X(x)$: - For $x \in [3, 5]$: $\min(10, x + 1) = x + 1$ and $\max(4, x - 1) = 4$,

$$f_X(x) = \frac{(x + 1) - 4}{12} = \frac{x - 3}{12}.$$

- For $x \in [5, 9]$: $\min(10, x + 1) = x + 1$ and $\max(4, x - 1) = x - 1$,

$$f_X(x) = \frac{(x + 1) - (x - 1)}{12} = \frac{2}{12} = \frac{1}{6}.$$

- For $x \in [9, 11]$: $\min(10, x + 1) = 10$ and $\max(4, x - 1) = x - 1$,

$$f_X(x) = \frac{10 - (x - 1)}{12} = \frac{11 - x}{12}.$$

Thus:

$$f_X(x) = \begin{cases} \frac{x-3}{12}, & 3 \leq x < 5, \\ \frac{1}{6}, & 5 \leq x \leq 9, \\ \frac{11-x}{12}, & 9 < x \leq 11, \\ 0, & \text{otherwise}. \end{cases}$$

(c) The conditional pdf is:

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)}.$$

For $\theta \in [\max(4, x - 1), \min(10, x + 1)]$:

$$f_{\Theta|X}(\theta \mid x) = \begin{cases} \frac{\frac{1}{12}}{f_X(x)}, & \max(4, x - 1) \leq \theta \leq \min(10, x + 1), \\ 0, & \text{otherwise}. \end{cases}$$

(d) Using the conditional pdf:

$$\mathbb{E}[\Theta \mid X = x] = \int_{-\infty}^{\infty} \theta f_{\Theta|X}(\theta \mid x)\, d\theta.$$

Substitute $f_{\Theta|X}(\theta \mid x)$:

$$\mathbb{E}[\Theta \mid X = x] = \frac{1}{f_X(x)} \int_{\max(4,x-1)}^{\min(10,x+1)} \frac{\theta}{12}\, d\theta.$$

Evaluate the integral:

$$\int \frac{\theta}{12}\, d\theta = \frac{\theta^2}{24}.$$

Thus:

$$\mathbb{E}[\Theta \mid X = x] = \frac{1}{f_X(x)} \left[\frac{\min(10, x+1)^2}{24} - \frac{\max(4, x-1)^2}{24} \right].$$

Q18: Show that the conditional expectation estimator is the best guess, in terms of minimizing mean square error, i.e. show the following -

$$E[(\theta - E[\Theta|X])^2|X] \leq E[(\theta - g(x))^2|X]$$

What would be the estimated value of conditional expectation?

A: Let us first address the simpler case of no conditioning. What is the value of c for which $E[(\theta - c)^2]$ is minimum? This is same as $E[\theta^2] + c^2 - 2cE[\theta]$. We can just differentiate this expression with c and set it to 0. This yeilds c = $E[\theta]$. We have just shown the following inequality -

$$E[(\theta - E[\Theta])^2] \leq E[(\theta - c)^2]$$

Now when we enter a conditional world our task is to find c such that $E[(\theta - c)^2|X]$ is minimized. The answer is just c = $E[\Theta|X = x]$). One can redo the calculation, but there is a clever way to see why this is obvious. Conditioning fundamentally means adding new information to the system. This may or may not revise our relative beliefs about the likelihood of the occurrence of events. If we know that X = x happened, our distribution of Θ will change. In which case the result would still hold for this new distribution, as it is true for all distributions of Θ! Just that expectation will now be over this newer distribution.

For expected value of conditional expectation/MMSE, we can see that

$$\begin{aligned} E[\hat{X}_M] &= E[E[X|Y]] \\ &= E[X] \quad \text{(by the law of iterated expectations).} \end{aligned}$$

To get a better understanding of the reasoning, refer to this.