Each question is worth 10 points. Explain your answers clearly.

1. You are given one sample that is either Uniform$(-1,1)$ if $\Theta = 0$ or Uniform$(0,4)$ if $\Theta = 1$. Your prior on Θ is equally likely ($\mathrm{P}(\Theta = 0) = \mathrm{P}(\Theta = 1) = 1/2$).

   (a) What is the MAP estimator for Θ from this sample?

   Solution: Let the sample be x. By Bayes' rule,

   $$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta)\,\mathrm{P}(\theta)$$

   Since $\mathrm{P}(\Theta = 0) = \mathrm{P}(\Theta = 1) = 1/2$, the MAP estimator maximizes $f_{X|\Theta}(x|\theta)$, which gives $MAP = 0$ if $x < 1$ and $MAP = 1$ otherwise.

   (b) What is the error probability of the MAP estimator in part (a)?

   Solution: $\mathrm{P}(MAP \neq \Theta) = \mathrm{P}(\Theta = 0, X \geq 1) + \mathrm{P}(\Theta = 1, X < 1) = 0 + 1/2 \cdot 1/4 = 1/8$.

2. You observe the following samples of a normal random variable with unknown mean and variance: $-1.5, -0.8, 1.9$.

   (a) What is the adjusted sample standard deviation S^2?

   Solution:

   $$\overline{X} = 1/3(-1.5 - 0.8 + 1.9) = -2/15$$

   $$S^2 = \frac{1}{2}\big((-1.5 + 2/15)^2 + (-0.8 + 2/15)^2 + (1.9 + 2/15)^2\big) \approx 3.223$$

   (b) Give a 95% confidence interval for the actual standard deviation. Justify the use of your formula for confidence intervals.

   Solution: $(n-1)S^2/\sigma^2$ is a $\chi^2(n-1)$ random variable. A 95% confidence interval for $\chi^2(2)$ is $[0.051, 7.378]$. Therefore the 95%-confidence interval for σ is $[\sqrt{\frac{(n-1)S^2}{7.378}}, \sqrt{\frac{(n-1)S^2}{0.051}}]$ which evaluates to $[0.93, 11.24]$.

3. A fair n-sided die with equally likely face values $1, 2, \ldots, n$ is tossed five times.

   (a) You observe the outcomes $1, 2, 1, 1, 5$. What is the maximum likelihood estimate for n?

   Solution: The joint PMF of the five samples given n is:

   $$\mathrm{P}(1, 2, 1, 1, 5|n) = \frac{1}{n^5},$$

   where n must be at least five. $1/n^5$ is maximized at $n = 5$ so the ML estimate is 5.

   (b) Let MAX be the largest of the five outcomes. Is $[MAX, 2MAX]$ a 95% confidence interval for n? Justify your answer.

Solution: Yes. Let $X_1, X_2, \cdots, X_5$ be the five samples. The CDF of MAX is:

$$\mathrm{P}(MAX \le t) = \mathrm{P}(X_1 \le t, X_2 \le t, \cdots, X_5 \le t) = \left(\frac{t}{n}\right)^5, t = 1, 2, \cdots, n$$

so the confidence level is

$$\begin{aligned}\mathrm{P}(MAX \le n \le 2MAX) &= \mathrm{P}(n/2 \le MAX \le n) \\ &= \mathrm{P}(MAX \le n) - \mathrm{P}(MAX \le n/2) \\ &= 1 - \mathrm{P}(MAX \le n/2).\end{aligned}$$

The probability of $MAX \le n/2$ is $(1/2)^5$ when n is even and slightly smaller when n is odd, so the event $MAX \le n \le 2MAX$ has probability at least $1 - (1/2)^5 = 96.875\%$.

4. A random variable has PMF $f(-1) = f(1) = \theta$, $f(0) = 1 - 2\theta$, where θ is unknown ($0 \le \theta \le \frac{1}{2}$).

   (a) What is the actual standard deviation σ of the random variable?

   Solution: The mean μ is zero, so the standard deviation is $\sigma^2 = (-1)^2 \cdot \theta + 1^2 \cdot \theta + 0^2 \cdot (1 - 2\theta) = 2\theta$.

   (b) What is the PMF of the adjusted sample standard deviation S^2 for two samples?

   Solution: $S^2 = (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 = (X_1 - X_2)^2/2$ takes value 0 when $X_1 = X_2$, value 2 when one of the samples is 1 and the other is -1, and value $1/2$ otherwise. Therefore

   $$\mathrm{P}(S^2 = 0) = \mathrm{P}(X_1 = X_2) = 2\theta^2 + (1 - 2\theta)^2$$

   and

   $$\mathrm{P}(S^2 = 4) = \mathrm{P}(X_1 = 1, X_2 = -1) + \mathrm{P}(X_1 = -1, X_2 = 1) = 2\theta^2.$$

   This gives the following PMF:

v	0	1/2	2
$\mathrm{P}(S^2 = v)$	$2\theta^2 + (1 - 2\theta)^2$	$4\theta(1 - 2\theta)$	$2\theta^2$

Each question is worth 10 points. Explain your answers clearly.

1. X is a Normal$(0, \Theta)$ random variable, where the prior PMF of the parameter Θ is $\mathrm{P}(\Theta = 1/2) = 1/2$, $\mathrm{P}(\Theta = 1) = 1/2$. You observe the following three independent samples of X: $1.0, 1.0, -1.0$.

   (a) What is the posterior PMF of Θ?

   Solution: By Bayes' rule

   $$f_{\Theta|X_1X_2X_3}(\theta|1.0, 1.0, -1.0) \propto f_{X_1X_2X_3|\Theta}(1.0, 1.0, -1.0|\theta)\,\mathrm{P}(\Theta = \theta) \propto \frac{1}{\theta^3} e^{-3/2\theta^2}\,\mathrm{P}(\Theta = \theta).$$

   As Θ is equally likely to take values $1/2$ and 1, the posterior PMF is

   $$f_{\Theta|X_1X_2X_3}(1/2|1.0, 1.0, -1.0) = \frac{8e^{-6}}{8e^{-6} + e^{-3/2}} \quad f_{\Theta|X_1X_2X_3}(1|1.0, 1.0, -1.0) = \frac{e^{-3/2}}{8e^{-6} + e^{-3/2}}.$$

   (b) What is the MAP estimate of Θ?

   Solution: As $e^{-3/2} \approx 0.2231$ is larger than $8e^{-6} \approx 0.0198$, the MAP estimate is 1.

2. The true fraction of employees in some company that support longer lunch breaks is 80%. Ten employees are polled about their support for longer lunch breaks (randomly with repetition). What is the probability that at least 70% of the polled employees support longer lunch breaks?

   Solution: The number of polled employees T supporting longer lunch breaks is a Binomial$(10, 0.8)$ random variable. We are looking for the probability that such a random variable takes value 7 or more. Using this online calculator we find $\mathrm{P}(\text{Binomial}(10, 0.8) \geq 7) \approx 0.879$.

3. In a random 50 participant survey about favorite colors, 20 choose "blue", 15 choose "red", 10 choose "green", and 5 choose "yellow".

   (a) Give 95% confidence intervals for the popularity of blue and green among the general population using the "simplified" formula for confidence intervals.

   Solution: Let p_{color} be the popularity of a color, which can be considered as the parameter for an indicator random variable. From the survey, the sample mean for color blue is 0.4 and the sample mean for color green is 0.2. The "simplified" formula gives a 95% confidence interval of an indicator by

   $$\left[\overline{X} - z\sqrt{\overline{X}(1-\overline{X})/n}, \overline{X} + z\sqrt{\overline{X}(1-\overline{X})/n}\right],$$

   where $z = 1.96$. Given that, the 95% confidence interval for p_{blue} is $[0.264, 0.536]$ and $[0.089, 0.311]$ for p_{green}.

(b) Based on your calculation in part (a), what is your confidence level for the claim "blue is more popular than green"? Justify your answer.

Solution: As the two intervals are not disjoint we cannot say that the blue is popular than green with 90% or more confidence. This would be an acceptable answer, but if we want to get some confidence level for our claim we would look for the largest value of z for which the two intervals are disjoint, namely the largest z so that $0.2+z\sqrt{0.2 \cdot 0.8/50} < 0.4-z\sqrt{0.4 \cdot 0.6/50}$, which is $z \approx 1.589$. As $\mathrm{P}(-z \leq \mathrm{Normal}(0,1) \leq z)$ is about 89%, we can be 89% confident that the popularity of green is within $(0.110, 0.290)$ and 89% confident that the popularity of green is within $(0.290, 0.510)$, and therefore 78% confident that blue is more popular than green.

4. A random variable X is Normal$(1,1)$ with probability p and Normal$(-1,1)$ with probability $1-p$, where the parameter p is unknown.

(a) What is the maximum likelihood estimator of p from a single sample X?

Solution: Let Θ be the indicator that X is Normal$(1,1)$. Then by total probability theorem,

$$\begin{aligned} f_X(x) &= f_{X|\Theta}(x|1)f_\Theta(1) + f_{X|\Theta}(x|0)f_\Theta(0) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} \cdot p + \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+1)^2}{2}} \cdot (1-p) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2+1}{2}} \left(p(e^x - e^{-x}) + e^{-x}\right) \end{aligned}$$

The maximum likelihood estimator finds the p between 0 and 1 for which $f_X(x)$ is maximized. Since $f_X(x)$ is a linear function in p with positive slope iff $e^x - e^{-x} > 0$, i.e. when $x > 0$, the ML estimate is

$$ML = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

(b) Is the estimator in part (a) unbiased? Justify your answer.

Solution: The expected value of ML (for a fixed value of the parameter p) is

$$\begin{aligned} \mathrm{E}[ML] = \mathrm{P}(X > 0) &= p\,\mathrm{P}(X > 0|\Theta = 1) + (1-p)\,\mathrm{P}(X > 0|\Theta = 0) \\ &= p\,\mathrm{P}(\mathrm{Normal}(1,1) > 0) + (1-p)\,\mathrm{P}(\mathrm{Normal}(-1,1) > 0) \\ &\approx p \cdot 0.841 + (1-p) \cdot 0.158. \end{aligned}$$

This is not an unbiased estimator of p: For example when $p = 1$, $\mathrm{E}[ML] \approx 0.841$. In fact, you can conclude the answer is no without doing any calculation: $\mathrm{E}[ML] = \mathrm{P}(X > 0)$ is always strictly smaller than 1, so it cannot be an unbiased estimator when $p = 1$.