CSE 312 – Section 9 Solutions

Spring 2026

Review of Main Concepts

Law of Total Expectation

- **Conditional Expectation**: Let X and Y be **discrete** random variables. Then, the conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X|Y=y] = \sum_x x \cdot p_{X|Y}(x|y) = \sum_x x \cdot \mathbb{P}(X=x|Y=y).$$

  Note that linearity of expectation still applies to conditional expectation:

$$\mathbb{E}[X+Y|A] = \mathbb{E}[X|A] + \mathbb{E}[Y|A].$$

- **Discrete Law of Total Expectation (event version):** Let $A_1 \dots, A_n$ be a partition of the sample space. Then

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i]\,\mathbb{P}(A_i).$$

- **Discrete Law of Total Expectation (r.v. version):** Let X and Y be two random variables. Then

$$\mathbb{E}[X] = \sum_{y \in \Omega_Y} \mathbb{E}[X|Y=y] \cdot \mathbb{P}(Y=y).$$

- **Continuous Law of Total Expectation:**

$$\mathbb{E}[X] = \int_{y \in \Omega_Y} \mathbb{E}[X|Y=y]\, f_Y(y)\, \mathrm{d}y$$

- **Expected value of X conditioned on r.v. Y:** Suppose that Y is a random variable that takes values $y_1, \dots, y_k$. Then $\mathbb{E}[X|Y]$ is the following random variable

$$\mathbb{E}[X|Y] = \begin{cases} \mathbb{E}[X|Y=y_1] & \text{with probability } \mathbb{P}(Y=y_1) \\ \mathbb{E}[X|Y=y_2] & \text{with probability } \mathbb{P}(Y=y_2) \\ \dots & \\ \mathbb{E}[X|Y=y_k] & \text{with probability } \mathbb{P}(Y=y_k) \end{cases}$$

- **Law of total expectation (rewritten):** Given the above definition, we can write

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \sum_{i=1}^{k} \mathbb{E}[X|Y=y_i] \cdot \mathbb{P}(Y=y_i).$$

- **Covariance:** We may not get to this in class, but it's important to know about. To find out more, check out section 5.4 in the Tsun book. And now the definition: For any two random variables X, Y the *covariance* is defined as

$$\operatorname{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right].$$

  It can also be shown that

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

- **Conditional distributions:** We are not explicitly covering this topic in class, but it is **highly** recommended that you study it. Much of the above can be more appropriately rewritten in terms of conditional distributions. See Tsun, Section 5.3.

	Discrete	**Continuous**
Conditional PMF/PDF	$p_{X\|Y}(x\|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Conditional Expectation	$\mathbb{E}[X\|Y=y] = \sum_x x p_{X\|Y}(x\|y)$	$\mathbb{E}[X\|Y=y] = \int_{-\infty}^{\infty} x f_{X\|Y}(x\|y)\,\mathrm{d}x$

Maximum Likelihood Estimation

We assume access to samples $X_1, X_2, \ldots, X_n$ from a distribution (e.g. Bernoulli, Geometric, Normal, etc) with unknown parameters. Our goal is to estimate this parameters. When the parameter is unknown we denote it by θ.

- **Realization/Sample**: A realization/sample x of a random variable X is the value that is actually observed.
- **Likelihood**: Let $X_1 = x_1, \ldots X_n = x_n$ be iid samples of a random variable with probability mass function $p_X(\mathrm{x}\,;\theta)$ (if X discrete) or density $f_X(\mathrm{x}\,;\theta)$ (if X continuous), where θ is a parameter (or a vector of parameters). We define the likelihood function to be the probability of seeing the data (in the discrete case). If X is discrete:

$$L(X_1 = x_1, \ldots X_n = x_n\,;\,\theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta)$$

  If X is continuous:

$$L(X_1 = x_1, \ldots X_n = x_n\,;\,\theta) = \prod_{i=1}^{n} f_X(x_i\,;\,\theta)$$

- **Log-Likelihood**: We define the log-likelihood as the natural logarithm of the likelihood function. Since the logarithm is a strictly increasing function, the value of θ that maximizes the likelihood will be exactly the same as the value that maximizes the log-likelihood. If X is discrete:

$$\ln L(x_1, \ldots, x_n\,;\,\theta) = \sum_{i=1}^{n} \ln p_X(x_i\,;\,\theta)$$

  If X is continuous:

$$\ln L(x_1, \ldots, x_n\,;\,\theta) = \sum_{i=1}^{n} \ln f_X(x_i\,;\,\theta)$$

- **Maximum Likelihood Estimator (MLE)**: We denote the MLE of θ as $\hat{\theta}_{\text{MLE}}$ or simply $\hat{\theta}$, the parameter (or vector of parameters) that maximizes the likelihood function (probability of seeing the data) or equivalently, maximizes the log-likelihood. For a fixed set of realizations/samples $x_1, \dots, x_n$

$$\hat{\theta}_{\text{MLE}}(x_1, \dots, x_n) = \arg\max_{\theta} L(x_1, \dots, x_n\,;\,\theta) = \arg\max_{\theta} \ln L(x_1, \dots, x_n\,;\,\theta)$$

  Viewing the estimator $\hat{\theta}_{\text{MLE}}$ as a function of the random variables $X_1, \dots, X_n$, the estimator $\hat{\theta}_{\text{MLE}}(X_1, \dots, X_n)$ is itself a random variable.

- **Bias**: The bias of an estimator $\hat{\theta}$ for a true parameter θ is defined as

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}(X_1, \dots, X_n)] - \theta.$$

  Here, $\hat{\theta}$ is treated as a *random variable*, computed from i.i.d. random variables $X_1, \dots . X_n$.

- **Unbiased Estimator**: An estimator $\hat{\theta}$ of θ is unbiased iff $\text{Bias}(\hat{\theta}, \theta) = 0$, or equivalently $\mathbb{E}[\hat{\theta}(X_1, \dots, X_n)] = \theta$.

- **Steps to find the maximum likelihood estimator $\hat{\theta}$:**

  a) Find the likelihood and log-likelihood of the data.

  b) Take the derivative of the log-likelihood

  c) Set it to 0 to find a candidate for the MLE, $\hat{\theta}$. (note: at this step, we change from the θ to the $\hat{\theta}$ because in this step we are solving for the *maximum* likelihood estimator for θ)

  d) Take the second derivative and show that $\hat{\theta}$ indeed is a maximizer, that $\frac{\partial^2 L}{\partial \theta^2} < 0$ at $\hat{\theta}$. Also ensure that it is the global maximizer: check points of non-differentiability and boundary values.

  Similar steps apply when the distribution has multiple parameters, e.g. a normal distribution with unknown mean and variance. In this case, maximizing the likelihood or log-likelihood involves maximizing a function of multiple variables (the different parameters). See, for example, problem 11 on this worksheet.

Tail bounds ++

- **Union Bound:** The union bound is a simple application of inclusion/exclusion and says that for any two events A and B, it holds that

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

  The union bound is used to bound probabilities when, e.g. $\mathbb{P}(A \cap B)$ is difficult to compute.

  We may or may not get to cover the following tail bounds.

- **Markov's Inequality**: Let X be a non-negative random variable, and $\alpha > 0$. Then,

$$\mathbb{P}\left(X \geq \alpha\right) \leq \frac{\mathbb{E}\left[X\right]}{\alpha}$$

- **Chebyshev's Inequality**: Suppose Y is a random variable with $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2$. Then, for any $\alpha > 0$,
$$\mathbb{P}(|Y - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$$
- **(Multiplicative) Chernoff Bound**: Let $X_1, X_2, ..., X_n$ be *independent* Bernoulli random variables. Let $X = \Sigma_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X]$. Then, for any $0 \leq \delta \leq 1$,
  - $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{3}}$
  - $\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2}}$

Section plan

- Content review I
- Problem 7
- Problem 8
- Problem 11
- If time, Problem 12

1 Content Review I

a) **Law of Total Expectation.** Let X and Y be discrete random variables. Which of the following represents the Law of Total Expectation for X?

- ○ $\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y] p_X(x)$
- ○ $\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y] p_Y(y)$
- ○ $\mathbb{E}[X] = \sum_x \mathbb{E}[X \mid Y = y] p_Y(y)$
- ○ $\mathbb{E}[X] = \sum_y \mathbb{E}[X] p_{X|Y}(x|y)$

Answer: $\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y] p_Y(y)$
The Law of Total Expectation states that the overall expected value of X is the weighted average of the conditional expectations of X given $Y = y$, weighted by the probability of $Y = y$ occurring ($p_Y(y)$). *(Note: Option 1 uses the wrong PMF weight. Option 3 sums over the wrong variable. Option 4 makes no mathematical sense.)*

b) **Linearity of Conditional Expectation.** True or False: Linearity of expectation does **not** apply to conditional expectations. That is, $\mathbb{E}[X + Y \mid A]$ is generally NOT equal to $\mathbb{E}[X \mid A] + \mathbb{E}[Y \mid A]$.

- ○ True
- ○ False

Answer: False
Linearity of expectation holds unconditionally *and* conditionally. As long as the expectations exist, $\mathbb{E}[X + Y \mid A] = \mathbb{E}[X \mid A] + \mathbb{E}[Y \mid A]$ is always true.

c) Suppose X and Y are random variables and A is an event. Given that $\mathbb{E}[X|A] = 4$ and $\mathbb{E}[Y|A] = 10$, what is $\mathbb{E}\left[2X + \frac{Y}{2} \mid A\right]$?

- ○ 14
- ○ 18
- ○ 9
- ○ 13

Choice d is the correct answer since the linearity of expectation still applies to conditional expectation:

$\mathbb{E}[2X + Y/2|A] = \mathbb{E}[2X|A] + \mathbb{E}[Y/2|A] = 2\mathbb{E}[X|A] + \mathbb{E}[Y|A]/2 = 2 \cdot 4 + 10/2 = 8 + 5 = 13.$

d) **True or False**: We get a different answer if we maximize log-likelihood rather than likelihood, but because it is close enough and easier to compute, we use it for our estimate of θ.

False: Since the logarithm is a strictly increasing function, the value of θ that maximizes the likelihood will be exactly the same as the value that maximizes the log-likelihood.

e) **True or False**: [Notation question] $\hat{\theta}$ is the true parameter and θ is our estimate.

False: It is the other way around. Remember to switch to $\hat{\theta}$ when you set your equation to zero!

f) **True or False**: An estimator is unbiased if $\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] - \theta = 0$ or equivalently $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

True by definition.

g) You flip a coin 10 times and observe HHHTHHTHHH (8 heads, 2 tails). What is the MLE of θ, where θ is the true probability of seeing tails?

- ○ $\hat{\theta} = .2$
- ○ $\hat{\theta} = .25$
- ○ $\hat{\theta} = .8$
- ○ $\hat{\theta} = .3$

Option 1: $\hat{\theta} = .2$

h) **True or False**: The Union Bound always gives a result in $[0, 1]$.

> False. Consider X and Y, which are independent indicator random variables. Suppose $p_X(x) = \begin{cases} 0.75 & x = 0 \\ 0.25 & x = 1 \end{cases}$ and $p_Y(y) = \begin{cases} 0.75 & y = 0 \\ 0.25 & y = 1 \end{cases}$. Then we may apply the Union Bound to place a bound on $P(X = 0 \cup Y = 0)$:
>
> $$P(X = 0 \cup Y = 0) \leq P(X = 0) + P(Y = 0) = 0.75 + 0.75 = 1.5.$$
>
> In these cases, the Union Bound tells us very little, since the probability of any event occurring is at most 1.

2 Content Review II (material not properly covered in class)

a) Suppose C and D are discrete random variables. Then $\mathbb{E}[C|D = d] =$

- ○ $\sum_d d p_{D|C}(d|c)$
- ○ $\sum_c c p_{C|D}(c|d)$
- ○ $\int_{-\infty}^{\infty} c f_{C|D} \, \mathrm{d}x$
- ○ $\frac{\mathbb{E}[C]}{\mathbb{E}[D]}$

> Choice b is the correct answer from the definition of conditional expectation for discrete random variables.

b) True or false: Markov's Inequality always gives a non-negative result.

> True. Markov's Inequality is
>
> $$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$$
>
> as long as X is a non-negative random variable and $\alpha > 0$. Since X is a non-negative random variable, $\mathbb{E}[X] \geq 0$, so $\frac{\mathbb{E}[X]}{\alpha} \geq 0$.

c) True or false: Chebyshev's Inequality can best be described as giving an upper bound on the distribution's right tail.

> False. Chebyshev's Inequality gives an upper bound on the sum of the probabilities of the left and right tails of the distribution.

3 Trapped Miner

A miner is trapped in a mine containing 3 doors.

- D_1: The 1^{st} door leads to a tunnel that will take him to safety after 3 hours.

- D_2: The 2^{nd} door leads to a tunnel that returns him to the mine after 5 hours.
- D_3: The 3^{rd} door leads to a tunnel that returns him to the mine after a number of hours that is Binomial with parameters $(12, \frac{1}{3})$.

At all times, he is equally likely to choose any one of the doors. What is the expected number of hours for this miner to reach safety?

Let T = number of hours for the miner to reach safety. (T is a random variable)
Let D_i be the event the i^{th} door is chosen. $i \in \{1, 2, 3\}$. Finally, let T_3 be the time it takes to return to the mine in the third case only (a random variable). Note that the expectation of T_3 is $12 * \frac{1}{3}$ because it is binomially distributed with parameters $n = 12, p = \frac{1}{3}$. By Law of Total Expectation, linearity of expectation, and by applying the conditional expectations given by the problem statement:

$$\begin{aligned}
\mathbb{E}[T] &= \mathbb{E}[T|D_1]\,\mathbb{P}(D_1) + \mathbb{E}[T|D_2]\,\mathbb{P}(D_2) + \mathbb{E}[T|D_3]\,\mathbb{P}(D_3) \\
&= 3 \cdot \frac{1}{3} + (5 + \mathbb{E}[T]) \cdot \frac{1}{3} + (\mathbb{E}[T_3 + T]) \cdot \frac{1}{3} \\
&= 3 \cdot \frac{1}{3} + (5 + \mathbb{E}[T]) \cdot \frac{1}{3} + (\mathbb{E}[T_3] + \mathbb{E}[T]) \cdot \frac{1}{3} \\
&= 3 \cdot \frac{1}{3} + (5 + \mathbb{E}[T]) \cdot \frac{1}{3} + (4 + \mathbb{E}[T]) \cdot \frac{1}{3}
\end{aligned}$$

Solving this equation for $\mathbb{E}[T]$, we get $\mathbb{E}[T] = 12$.

Therefore, the expected number of hours for this miner to reach safety is 12.

4 The Compound Server

A web server receives requests according to a Poisson process with a rate of λ requests per minute. Each incoming request is independently classified as a "database query" with probability p, or a "static asset request" with probability $1 - p$. Let N be the total number of requests in a minute, and X be the number of database queries in a minute.

a) Derive the marginal distribution of X. Also, write down the marginal distribution of Y, defined as the number of static asset requests in a minute. (You only need to prove your answer for X.)

b) Show that X and Y are independent.

c) Given that exactly k database queries were received in a particular minute, what is the expected *total* number of requests $\mathbb{E}[N \mid X = k]$ received during that minute?

a) **Marginal Distribution:** We are given $N \sim \text{Poisson}(\lambda)$ and X conditioned on the

value of N is Binomial(N, p). Using the Law of Total Probability:

$$\begin{aligned}\mathbb{P}(X = x) &= \sum_{n=x}^{\infty} \mathbb{P}(X = x \mid N = n)\mathbb{P}(N = n) \\ &= \sum_{n=x}^{\infty} \binom{n}{x} p^x (1-p)^{n-x} \frac{e^{-\lambda}\lambda^n}{n!} \quad .\end{aligned}$$

Rearranging terms and factoring out constants relative to the sum:

$$\mathbb{P}(X = x) = \frac{p^x e^{-\lambda}}{x!} \sum_{n=x}^{\infty} \frac{(1-p)^{n-x}\lambda^n}{(n-x)!} \quad .$$

Let $j = n - x$:

$$\begin{aligned}\mathbb{P}(X = x) &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{j=0}^{\infty} \frac{(\lambda(1-p))^j}{j!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{\lambda(1-p)} \quad . \\ &= \frac{(\lambda p)^x e^{-\lambda p}}{x!}\end{aligned}$$

This proves $X \sim$ Poisson(λp), a classic result known as Poisson thinning. Similarly $Y \sim$ Poisson$(\lambda(1-p))$.

b) **Independence:** To show that X and Y are independent, we must show that their joint PMF is equal to the product of their marginal PMFs: $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$ for all integers $x, y \geq 0$.

We can find the joint PMF by relating X and Y back to N. Note that the event $\{X = x \cap Y = y\}$ is equivalent to the event $\{X = x \cap N = x + y\}$.

$$\begin{aligned}\mathbb{P}(X = x, Y = y) &= \mathbb{P}(X = x, N = x + y) \\ &= \mathbb{P}(X = x \mid N = x + y)\mathbb{P}(N = x + y) \\ &= \binom{x+y}{x} p^x (1-p)^y \frac{e^{-\lambda}\lambda^{x+y}}{(x+y)!}\end{aligned}$$

Expanding the binomial coefficient, substituting $\lambda^{x+y} = \lambda^x \lambda^y$, and splitting the $e^{-\lambda}$ term into $e^{-\lambda p} e^{-\lambda(1-p)}$:

$$\begin{aligned}\mathbb{P}(X = x, Y = y) &= \frac{(x+y)!}{x!y!} p^x (1-p)^y \frac{e^{-\lambda p} e^{-\lambda(1-p)} \lambda^x \lambda^y}{(x+y)!} \\ &= \left(\frac{(\lambda p)^x e^{-\lambda p}}{x!}\right)\left(\frac{(\lambda(1-p))^y e^{-\lambda(1-p)}}{y!}\right) \\ &= \mathbb{P}(X = x)\mathbb{P}(Y = y)\end{aligned}$$

Since the joint PMF factors perfectly into the product of the marginal PMFs for all valid x and y, X and Y are independent.

c) **Conditional Expectation:** Let $Y = N - X$ be the number of static asset requests. By the same thinning logic, $Y \sim \text{Poisson}(\lambda(1-p))$, and crucially, X and Y are independent. Because $N = X + Y$:

$$\begin{aligned}\mathbb{E}[N \mid X = k] &= \mathbb{E}[X + Y \mid X = k] \\ &= \mathbb{E}[k + Y \mid X = k]\end{aligned}.$$

Due to the independence of the thinned processes, $\mathbb{E}[Y \mid X = k] = \mathbb{E}[Y] = \lambda(1-p)$.

$$\mathbb{E}[N \mid X = k] = k + \lambda(1-p) \quad .$$

5 Lemonade Stand

Suppose I run a lemonade stand, which costs me \$100 a day to operate. I sell a drink of lemonade for \$20. Every person who walks by my stand either buys a drink or doesn't (no one buys more than one). If it is raining, n_1 people walk by my stand, and each buys a drink independently with probability p_1. If it isn't raining, n_2 people walk by my stand, and each buys a drink independently with probability p_2. It rains each day with probability p_3, independently of every other day. Let X be my profit over the next week. In terms of n_1, n_2, p_1, p_2 and p_3, what is $\mathbb{E}[X]$?

Let R be the event it rains. Let X_i be how many drinks I sell on day i for $i = 1, ..., 7$. We are interested in $X = \sum_{i=1}^{7} (20X_i - 100)$. We have $X_i|R \sim \mathsf{Binomial}(n_1, p_1)$, so $\mathbb{E}[X_i|R] = n_1 p_1$. Similarly, $X_i|R^C \sim \mathsf{Binomial}(n_2, p_2)$, so $\mathbb{E}[X_i|R^C] = n_2 p_2$. By the law of total expectation,

$$\mu = \mathbb{E}[X_i] = \mathbb{E}[X_i|R]\,\mathbb{P}(R) + \mathbb{E}[X_i|R^C]\,\mathbb{P}(R^C) = n_1 p_1 p_3 + n_2 p_2 (1 - p_3)$$

Hence, by linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{7} (20X_i - 100)\right] = 20 \sum_{i=1}^{7} \mathbb{E}[X_i] - 700 = 140\mu - 700$$

$$= 140 \cdot (n_1 p_1 p_3 + n_2 p_2 (1 - p_3)) - 700.$$

6 The Dice and Coin Cascade

Alice repeatedly rolls a fair six-sided die until she rolls a 6. Let N be the total number of rolls she makes (including the final roll of 6). Once Alice is finished, Bob flips a fair coin exactly N times. Let H be the total number of Heads Bob flips.

a) Compute the exact probability that Bob flips zero Heads, $\mathbb{P}(H = 0)$.

b) Compute the expected value $\mathbb{E}[H]$ and variance $\text{Var}(H)$.

a) **Probability of Zero Heads:** We have $N \sim \text{Geometric}(1/6)$ and H conditioned on N

is Binomial$(N, 1/2)$.

$$\begin{aligned}\mathbb{P}(H=0) &= \sum_{n=1}^{\infty} \mathbb{P}(H=0 \mid N=n)\mathbb{P}(N=n) \\ &= \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n \left(\frac{5}{6}\right)^{n-1} \left(\frac{1}{6}\right)\end{aligned} .$$

Factoring out constants to create a clean geometric series:

$$\begin{aligned}\mathbb{P}(H=0) &= \frac{1/6}{5/6} \sum_{n=1}^{\infty} \left(\frac{5}{12}\right)^n \\ &= \frac{1}{5}\left(\frac{5/12}{1-5/12}\right) . \\ &= \frac{1}{5}\left(\frac{5/12}{7/12}\right) = \frac{1}{7}\end{aligned}$$

b) **Expectation and Variance:** Using the Law of Total Expectation:

$$\begin{aligned}\mathbb{E}[H] &= \mathbb{E}[\mathbb{E}[H \mid N]] \\ &= \sum_{n=1}^{\infty} \mathbb{E}[H|N=n]\mathbb{P}(N=n) \\ &= \sum_{n=1}^{\infty} \frac{n}{2}\mathbb{P}(N=n) \\ &= \frac{1}{2}\mathbb{E}[N] = \frac{1}{2}(6) = 3\end{aligned} .$$

To get the variance, we can use the Law of Total Variance (which we have not covered in class):

$$\begin{aligned}\mathrm{Var}(H) &= \mathbb{E}[\mathrm{Var}(H \mid N)] + \mathrm{Var}(\mathbb{E}[H \mid N]) \\ &= \mathbb{E}\left[N\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right] + \mathrm{Var}\left(\frac{N}{2}\right) \\ &= \frac{1}{4}\mathbb{E}[N] + \frac{1}{4}\mathrm{Var}(N)\end{aligned} .$$

For $N \sim \text{Geometric}(p)$, $\mathrm{Var}(N) = \frac{1-p}{p^2} = \frac{5/6}{1/36} = 30$. Therefore,

$$\mathrm{Var}(H) = \frac{1}{4}(6) + \frac{1}{4}(30) = \frac{36}{4} = 9 \quad .$$

7 Nested Uniforms

Let $X \sim \mathsf{Unif}(0, 1)$. We draw a random variable Y such that, conditioned on $X = x$, $Y \sim \mathsf{Unif}(0, x)$. What is the expected value of Y, $\mathbb{E}[Y]$?

We can compute this using the Continuous Law of Total Expectation:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \mathbb{E}[Y \mid X = x] \, f_X(x) \, \mathrm{d}x$$

Because $X \sim \mathsf{Unif}(0, 1)$, its PDF is $f_X(x) = 1$ on the interval $[0, 1]$ and 0 elsewhere. Therefore, we can restrict our bounds of integration to 0 and 1.

We are given that $Y \mid X = x \sim \mathsf{Unif}(0, x)$. The expected value of a continuous uniform random variable on (a, b) is $\frac{a+b}{2}$, so the conditional expectation is:

$$\mathbb{E}[Y \mid X = x] = \frac{0 + x}{2} = \frac{x}{2}$$

Substituting this back into the Law of Total Expectation integral:

$$\begin{aligned} \mathbb{E}[Y] &= \int_0^1 \frac{x}{2} \cdot 1 \, \mathrm{d}x \\ &= \frac{x^2}{4}\Big|_0^1 \\ &= \frac{1}{4} - 0 = \frac{1}{4} \end{aligned}$$

8 A Red Poisson

Suppose that $x_1, \ldots, x_n$ are i.i.d. samples from a Poisson(θ) random variable, where θ is unknown. Find the MLE for θ.

Because each Poisson RV is i.i.d., the likelihood of seeing that data is just the PMF of the Poisson distribution multiplied together for every x_i. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for $\hat{\theta}$.

$$\begin{aligned} L(x_1, \ldots, x_n \,;\, \theta) &= \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} \\ \ln L(x_1, \ldots, x_n \,;\, \theta) &= \sum_{i=1}^{n} \left[-\theta - \ln(x_i!) + x_i \ln(\theta) \right] \\ \frac{\partial}{\partial \theta} \ln L(x_1, \ldots, x_n \,;\, \theta) &= \sum_{i=1}^{n} \left[-1 + \frac{x_i}{\theta} \right] \\ -n + \frac{\Sigma_{i=1}^{n} x_i}{\hat{\theta}} &= 0 \\ \hat{\theta} &= \frac{\Sigma_{i=1}^{n} x_i}{n} \, . \end{aligned}$$

Notice that the $-\ln(x_i!)$ term disappears since it is a constant relative to θ, of which we take the derivative.

9 Independent Shreds, You Say?

You are given 100 independent samples $X_1 = x_1, X_2 = x_2, \dots, X_{100} = x_{100}$ from Bernoulli(θ), where θ is unknown. (Each sample is either a 0 or a 1). These 100 samples sum to 30. We saw in class that the maximum likelihood estimate for θ is

$$\hat{\theta}(x_1, \dots, x_{100}) = \frac{1}{100} \sum_{i=1}^{100} x_i = \frac{30}{100}.$$

Is $\hat{\theta}(X_1, \dots, X_{100}) = \frac{\sum_{i=1}^{100} X_i}{100}$ an unbiased estimator of θ?

An estimator is unbiased if the expectation of the estimator (as a function of the random variables $X_1, \dots, X_{100}$ is equal to the original parameter, i.e., $E[\hat{\theta}(X_1, \dots, X_{100})] = \theta$. Setting up the expectation of our estimator and plugging it in for the generic case, we get the following:

$$\begin{aligned}
\mathbb{E}[\hat{\theta}(X_1, \dots, X_{100})] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{100} X_i\right] \\
&= \frac{1}{100} \sum_{i=1}^{100} \mathbb{E}[X_i] \\
&= \frac{1}{100} \cdot 100\theta = \theta.
\end{aligned}$$

So it is unbiased.

10 Laplace MLE

Suppose $x_1, \dots, x_{2n}$ are iid realizations from the Laplace density (double exponential density): for $x \in \mathbb{R}$,

$$f_X(x \mid \theta) = \frac{1}{2} e^{-|x-\theta|}$$

Find the MLE for θ. For this problem, you need not verify that the MLE is indeed a maximizer. You may find the **sign** function useful:

$$\operatorname{sgn}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

We begin by setting up the likelihood as we do in any case. Since these are i.i.d. realizations, we can multiply all their PDFs together. From there, take the log-likelihood, then the first derivative, where we notice that the derivative of $-\ln 2 - |x_i - \theta|$ is just the sign function of

$x_i - \theta$. Then, set that equal to 0 and solve for $\hat{\theta}$.

$$\begin{aligned} L\left(x_1, \ldots, x_{2n} \mid \theta\right) &= \prod_{i=1}^{2n} \frac{1}{2} e^{-|x_i - \theta|} \\ \ln L\left(x_1, \ldots, x_{2n} \mid \theta\right) &= \sum_{i=1}^{2n} \left[-\ln 2 - |x_i - \theta|\right] \\ \frac{\partial}{\partial \theta} \ln L\left(x_1, \ldots, x_{2n} \mid \theta\right) &= \sum_{i=1}^{2n} \operatorname{sgn}\left(x_i - \theta\right) = 0 \\ \hat{\theta} &= \text{ any value in } [x_n^{'}, x_{n+1}^{'}] \end{aligned}$$

where $x_i^{'}$ is the i^{th} order statistic: the i^{th} smallest observation (see 5.10 in the textbook for more details). If you wanted to argue that this is a global maximizer, note that the log likelihood is the sum of concave functions, so every critical point is a global maximizer.

11 Bird Watching

You are an ornithologist studying a rare species of birds in a nature reserve. Over a period of 50 days, you record the number of sightings of this bird (you see $x_1, x_2, ..., x_{50}$ birds on each day). Your research has shown that the number of sightings of this species depends on the average number of monkeys in the reserve, θ_1, and the average number of eagles in the reserve, θ_2. After years of studying this rare species in other environments, you've found that the number of birds observed on a particular day follows the following distribution:

$$p_X(k) = \frac{1}{k!}(\theta_1^k \cdot e^{-\theta_1} \cdot \theta_2^k \cdot e^{-3\theta_2})$$

Find the MLE for θ_1 and θ_2 (i.e., find $\hat{\theta_1}$ and $\hat{\theta_2}$).

a) What is the likelihood function?

Once again, the likelihood of seeing the above samples is just their PDFs multiplied together.

$$L(x; \theta_1, \theta_2) = \prod_{i=1}^{50} \left(\frac{1}{x_i!} (\theta_1^{x_i} e^{-\theta_1} \cdot \theta_2^{x_i} e^{-3\theta_2}) \right)$$

b) What is the log-likelihood function?

We take the log of the above and simplify:

$$\ln(L(x;\theta_1,\theta_2)) = \ln(\prod_{i=1}^{50}(\frac{1}{x_i!}(\theta_1^{x_i}\cdot e^{-\theta_1}\cdot\theta_2^{x_i}\cdot e^{-3\theta_2}))) \tag{1}$$

$$= \sum_{i=1}^{50}(\ln(\frac{1}{x_i!}) + \ln(\theta_1^{x_i}) + \ln(e^{-\theta_1}) + \ln(\theta_2^{x_i}) + \ln(e^{-3\theta_2})) \tag{2}$$

$$= \sum_{i=1}^{50}(\ln(\frac{1}{x_i!}) + x_i\ln(\theta_1) - \theta_1 + x_i\ln(\theta_2) + -3\theta_2) \tag{3}$$

c) We want to find values of θ_1 and θ_2 that maximize the likelihood function. To do this, we will take the partial derivative with respect to each of these parameters and solve for the values that make them both zero. First, take the partial derivative of the likelihood function with respect to θ_1.

When taking the partial derivative with respect to a certain variable, we take the derivative as usual, but treat other variables as constants! So here, we treat θ_2 as a constant and differentiate with respect to θ_1.

$$\frac{\partial}{\partial\theta_1}(\ln(L(x;\theta_1,\theta_2))) = \sum_{i=1}^{50}(\frac{x_i}{\theta_1} - 1) \tag{4}$$

$$= \sum_{i=1}^{50}(\frac{x_i}{\theta_1}) - 50 \tag{5}$$

d) Now, take the partial derivative with respect to θ_2.

Now, we treat θ_1 as a constant and take the derivative with respect to θ_2.

$$\frac{\partial}{\partial\theta_2}(\ln(L(x;\theta_1,\theta_2))) = \sum_{i=1}^{50}(\frac{x_i}{\theta_2} - 3) \tag{6}$$

$$= \sum_{i=1}^{50}(\frac{x_i}{\theta_2}) - 150 \tag{7}$$

e) Set both these partial derivatives to 0, and solve for $\hat{\theta_1}$ and $\hat{\theta_2}$.

We end up with the equations (notice we added the hats to the thetas at this point - since we set these derivatives to 0, we are now solving for the *maximum* likelihood estimator):

$$\sum_{i=1}^{50}(\frac{x_i}{\hat{\theta_1}}) - 50 = 0$$

$$\sum_{i=1}^{50}(\frac{x_i}{\hat{\theta_2}}) - 150 = 0$$

We now solve this system of equations for $\hat{\theta_1}$ and $\hat{\theta_2}$. The first equation gives us:

$$\sum_{i=1}^{50}(\frac{x_i}{\hat{\theta_1}}) = 50$$

$$(\frac{\sum_{i=1}^{50} x_i}{\hat{\theta_1}}) = 50$$

$$\hat{\theta_1} = (\frac{\sum_{i=1}^{50} x_i}{50})$$

With similar steps, we get that:

$$\hat{\theta_2} = (\frac{\sum_{i=1}^{50} x_i}{150})$$

Remark: Similar to the single-variable case, we treat the points where both partial derivatives are 0 (also known as the critical point) directly as the maximum of the function. This is just a convenient simplification for 312. For general functions, we need to do the second derivative test to confirm whether a critical point is indeed a local maximum/minimum. For multi-variable functions, the second derivative test is much more complicated - it is not enough to take the partial second derivatives, you would have to check the Hessian matrix.

12 A biased estimator

In class, we showed that the maximum likelihood estimate of the variance θ_2 of a normal distribution (when both the true mean μ and true variance σ^2 are unknown) is what's called the *population variance*. That is

$$\hat{\theta}_2 = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}_1)^2\right)$$

where $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the MLE of the mean. Is $\hat{\theta}_2$ unbiased?

Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then the estimator $\hat{\theta_2}$ as a random variable can be written as

$$\hat{\theta_2}\,(X_1, \ldots, X_n) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)$$

Therefore,

$$E(\hat{\theta}_2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) - E\left(\frac{2}{n}\overline{X}\sum_{i=1}^{n}X_i\right) + E(\overline{X}^2)$$
$$= \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) - 2E(\overline{X}^2) + E(\overline{X}^2)$$
$$= \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) - E(\overline{X}^2). \qquad (**)$$

The second equality follows from the first by linearity of expectation (and distributing the sum). We know that for any random variable Y, since $Var(Y) = E(Y^2) - (E(Y))^2$ it holds that

$$E(Y^2) = Var(Y) + (E(Y))^2.$$

Also, we have $E(X_i) = \mu,\ Var(X_i) = \sigma^2\ \forall i$ and $E(\overline{X}) = \mu,\ Var(\overline{X}) = \frac{\sigma^2}{n}$. Combining these facts, we get

$$E(X_i^2) = \sigma^2 + \mu^2 \quad \forall i \quad \text{and} \quad E(\overline{X}^2) = \frac{\sigma^2}{n} + \mu^2.$$

Substituting these equations into (**) we get

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) - E(\overline{X}^2) = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= \left(1 - \frac{1}{n}\right)\sigma^2.$$

Thus $\hat{\theta}_2$ is not unbiased.

13 It Means Nothing

Suppose $x_1, x_2, \ldots, x_n$ are samples from a normal distribution whose mean is known to be μ but the variance is unknown. How does the maximum likelihood estimator for the variance differ from the maximum likelihood estimator when both mean and variance are unknown? Which, if any, is unbiased?

Begin with the same derivation as before, however we now use μ instead of the mean of 0, which gets us:

$$
\begin{aligned}
L(x_1, \dots, x_n\,;\,\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_i-\mu)^2}{2\sigma^2} \\
\ln L(x_1, \dots, x_n\,;\,\sigma^2) &= \sum_{i=1}^{n} -\ln\sqrt{2\pi\sigma^2} - \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} \\
\frac{\partial}{\partial\sigma^2}\ln L(x_1, \dots, x_n\,;\,\sigma^2) &= -\frac{n}{2\sigma^2} + \frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4} \\
-\frac{n}{2\hat{\sigma}^2} + \frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{2\hat{\sigma}^4} &= 0 \\
\frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{2\hat{\sigma}^4} &= \frac{n}{2\hat{\sigma}^2} \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}
$$

Then, we do the same but with two parameters, θ_1, θ_2, the former being the mean, and the latter being the variance. We can take the derivative with respect to θ_2, and do effectively the same as before.

$$
\begin{aligned}
L(x_1, \dots, x_n\,;\,\theta_1, \theta_2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} \exp \frac{-(x_i-\theta_1)^2}{2\theta_2} \\
\ln L(x_1, \dots, x_n\,;\,\theta_1, \theta_2) &= \sum_{i=1}^{n} -\ln\sqrt{2\pi\theta_2} - \frac{(x_i-\theta_1)^2}{2\theta_2} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi\theta_2 - \frac{(x_i-\theta_1)^2}{2\theta_2} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\theta_2 - \frac{(x_i-\theta_1)^2}{2\theta_2} \\
&= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\theta_2 - \frac{\Sigma_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2} \\
\frac{\partial}{\partial\theta_2}\ln L(x_1, \dots, x_n\,;\,\theta_1, \theta_2) &= -\frac{n}{2\theta_2} + \frac{\Sigma_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2^2} \\
-\frac{n}{2\hat{\theta_2}} + \frac{\Sigma_{i=1}^{n}(x_i-\hat{\theta_1})^2}{2\hat{\theta_2}^2} &= 0 \\
\frac{\Sigma_{i=1}^{n}(x_i-\hat{\theta_1})^2}{2\hat{\theta_2}^2} &= \frac{n}{2\hat{\theta_2}}
\end{aligned}
$$

$$\hat{\theta_2} = \qquad\qquad \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta_1})^2$$

Now, we just need to find if both estimators are biased or unbiased. We do this by seeing if their expected value is equal to the original parameter or not. Let's start with the former. We move the expectation in with linearity of expectation, and then can identify that the remaining expectation is just the definition of variance (expected deviation from the mean squared) and see that it is unbiased.

$$E[\hat{\sigma}^2] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n}n\sigma^2 = \sigma^2$$

For the second estimator, following the same computation in Task 12, we obtain

$$E[\hat{\theta_2}] = \left(1 - \frac{1}{n}\right)\sigma^2$$

suggesting that $\hat{\theta_2}$ is not unbiased.

14 Maximum Likelihood Estimators

a) Let $x_1, \ldots, x_n$ i.i.d. samples from a random variable that follows a Rademacher distribution with unknown parameter $p \in [0, 1]$, i.e., a distribution from the family

$$\mathbb{P}[x; p] = \begin{cases} p & \text{if } x = +1 \\ 1 - p & \text{if } x = -1 \end{cases},$$

What is the maximum likelihood estimator for p? Write this as a function of the x_i's and n.

We can use the solution from class for Bernoulli random variables (which we used for the case of a magic coin with unknown bias). We just denote as n^+ the number of $+1$'s. We then now that the MLE $\hat{p}$ of p is $\hat{p} = n^+/n$. Note that

$$\sum_{i=1}^{n} x_i = n^+ - (n - n^+) = 2n^+ - n\,,$$

or equivalently,

$$n^+ = \frac{n + \sum_{i=1}^{n} x_i}{2}\,.$$

And thus,

$$\hat{p} = \frac{n + \sum_{i=1}^{n} x_i}{2n} = \frac{1}{2} + \frac{\sum_{i=1}^{n} x_i}{2n}\,.$$

b) Let $x_1, \ldots, x_n$ be i.i.d samples that follow a $\mathsf{Two}(\theta)$ distribution with unknown parameter

$\theta \in [0,1]$, where the probabilities from the family are given by

$$\mathbb{P}[x;\theta] = \begin{cases} (1-\theta)^2 & x=0 \\ 2\theta(1-\theta) & x=1 \\ \theta^2 & x=2 \end{cases}$$

Suppose that in the sample there are n_0 0's, n_1 1's, and n_2 2's. What is the maximum likelihood estimator for θ in terms of n, n_0, n_1, n_2?

The likelihood is

$$\begin{aligned} \mathcal{L}(x_1, x_2, \dots, x_n\,;\theta) &= (1-\theta)^{2\cdot n_0} \cdot (2\theta\,(1-\theta))^{n_1} \cdot \theta^{2n_2} \\ &= (1-\theta)^{2n_0} \cdot (1-\theta)^{n_1} \cdot 2^{n_1} \cdot \theta^{n_1} \cdot \theta^{2n_2} \\ &= (1-\theta)^{n_1+2n_0} \cdot 2^{n_1} \cdot \theta^{2n_2+n_1}\,. \end{aligned}$$

Therefore the log-likelihood is

$$\begin{aligned} \ln \mathcal{L}(x_1, x_2, \dots, x_n\,;\theta) &= \ln((1-\theta)^{n_1+2n_0}) + \ln(2^{n_1}) + \ln(\theta^{2n_2+n_1}) \\ &= (n_1+2n_0)\ln(1-\theta) + n_1 \ln 2 + (2n_2+n_1)\ln\theta\,. \end{aligned}$$

We take the derivative of the log-likelihood with respect to the parameter θ:

$$\frac{\partial}{\partial\theta} \ln \mathcal{L}(x_1, x_2, \dots, x_n\,;\theta) = \frac{(2n_2+n_1)}{\theta} - \frac{(n_1+2n_0)}{1-\theta}\,.$$

Now, we set the derivative to 0 and solve (here we replace θ with $\hat{\theta}$):

$$\begin{aligned} \frac{(2n_2+n_1)}{\hat{\theta}} - \frac{(n_1+2n_0)}{1-\hat{\theta}} &= 0 \\ \frac{(2n_2+n_1)}{\hat{\theta}} &= \frac{(n_1+2n_0)}{1-\hat{\theta}} \\ (2n_2+n_1)(1-\hat{\theta}) &= (n_1+2n_0)\hat{\theta} \\ (2n_2+n_1) &= 2(n_0+n_1+n_2)\hat{\theta} \end{aligned}$$

Therefore

$$\boxed{\hat{\theta} = \frac{n_1+2n_2}{2(n_0+n_1+n_2)} = \frac{n_1+2n_2}{2n}}\,.$$

c) Let $x_1, \dots, x_n$ be i.i.d. samples from a random variable that follows a so-called Borel distribution with unknown parameter θ, i.e., a distribution from the family

$$\mathbb{P}[k;\theta] = \frac{e^{-\theta k}(\theta k)^{k-1}}{k!}\,,$$

where $0 < \theta \leq 1$ is a real number, and $k \geq 1$ is an integer. What is the maximum likelihood estimator for θ?

The likelihood is

$$\mathcal{L}(x_1, x_2, \dots, x_n\,; \theta) = \prod_{i=1}^{n} \frac{e^{-\theta x_i}(\theta x_i)^{x_i - 1}}{x_i!}\ .$$

Therefore the log-likelihood is

$$\ln \mathcal{L}(x_1, x_2, \dots, x_n\,; \theta) = \sum_{i=1}^{n} \left[-\theta x_i + (x_i - 1)(\ln(\theta) + \ln(x_i)) - \ln(x_i!)\right]$$

We take the derivative of the log-likelihood with respect to the parameter θ:

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, x_2, \dots, x_n\,; \theta) = \sum_{i=1}^{n} \left(-x_i + \frac{x_i - 1}{\theta}\right)\ .$$

Now, we set the derivative to 0 and solve (here we replace θ with $\hat{\theta}$:

$$\sum_{i=1}^{n} \left(-x_i + \frac{x_i - 1}{\theta}\right) = 0$$

$$\frac{1}{\hat{\theta}} \sum_{i=1}^{n} (x_i - 1) = \sum_{i=1}^{n} x_i$$

Therefore

$$\boxed{\hat{\theta} = \frac{\sum_{i=1}^{n}(x_i - 1)}{\sum_{i=1}^{n} x_i} = 1 - \frac{n}{\sum_{i=1}^{n} x_i}}\ .$$

Check that it's a global maximum (not required):

$$\frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(x_1, x_2, \dots, x_n\,; \theta) = -\sum_{i=1}^{n} \frac{x_i - 1}{\theta^2} < 0$$

so $\ln \mathcal{L}(x_1, x_2, \dots, x_n\,; \theta)$ is concave downward everywhere.

d) If the samples from the Borel distribution are 5, 7, 10, 2, 7, 5, 12, 13, 11, what is the maximum likelihood estimator for θ? Give an exact answer as a simplified fraction.

$$\hat{\theta} = 1 - \frac{9}{5 + 7 + 10 + 2 + 7 + 5 + 12 + 13 + 11} = 1 - \frac{1}{8} = \boxed{\frac{7}{8}}\ .$$

The remaining problems cover material we may not get to this quarter.

15 Tail bounds

Suppose $X \sim \mathsf{Binomial}(6, 0.4)$. We will bound $\mathbb{P}(X \geq 4)$ using the tail bounds we've learned, and compare this to the true result.

a) Give an upper bound for this probability using Markov's inequality. Why can we use Markov's inequality?

We know that the expected value of a binomial distribution is np, so: $\mathbb{P}(X \geq 4) \leq \frac{\mathbb{E}[X]}{4} = \frac{2.4}{4} = 0.6$. We can use it since X is nonnegative.

b) Give an upper bound for this probability using Chebyshev's inequality. You may have to rearrange algebraically and it may result in a weaker bound.

$\mathbb{P}(X \geq 4) = \mathbb{P}(X - 2.4 \geq 1.6) \leq \mathbb{P}(|X - 2.4| \geq 1.6)$ we can add those absolute value signs because that only adds more possible values, so it is an upper bound on the probability of $X - 2.4 \geq 1.6$. Then, using Chebyshev's inequality we get:
$\mathbb{P}(|X - 2.4| \geq 1.6) \leq \frac{Var(X)}{1.6^2} = \frac{1.44}{1.6^2} = 0.5625$

c) Give an upper bound for this probability using the Chernoff bound.

First, we solve for the values of δ that will allow us to use the Chernoff bound. We want $(1+\delta)\mathbb{E}[X] = (1+\delta)2.4 = 4$. Solving for δ here gives use $\delta = \frac{2}{3}$. Now, we can directly plug into the Chernoff bound. $\mathbb{P}(X \geq 4) = \mathbb{P}(X \geq (1+\frac{2}{3})2.4) \leq e^{-(\frac{2}{3})^2\mathbb{E}[X]/3} = e^{-4\times 2.4/27} \approx 0.7$

d) Give the exact probability.

Since X is a binomial, we know it has a range from 0 to n (or in this case 0 to 6). Thus, the possible values to satisfy $X \geq 4$ are 4, 5, or 6. We plug in the PMF for each to get: $\mathbb{P}(X \geq 4) = \mathbb{P}(X = 4) + \mathbb{P}(X = 5) + \mathbb{P}(X = 6) = \binom{6}{4}(0.4)^4(0.6)^2 + \binom{6}{5}(0.4)^5(0.6) + \binom{6}{6}0.4^6 \approx 0.1792$

16 Exponential Tail Bounds

Let $X \sim \text{Exp}(\lambda)$ and $k > 1/\lambda$.

a) Use Markov's inequality to bound $\mathbb{P}(X \geq k)$.

We can use Markov's inequality here because X is non-negative since it is an exponential distribution. We also know that $\mathbb{E}[X] = 1/\lambda$ because $X \sim \text{Exp}(\lambda)$. By Markov's inequality, we get that:

$$\mathbb{P}(X \geq k) \leq \frac{1}{\lambda k}$$

b) Use Markov's inequality to bound $\mathbb{P}(X < k)$.

From Markov's inequality (and our answer in (a)), we know that $\mathbb{P}(X \geq k) \leq \frac{1}{\lambda k}$. Then,

$$\begin{aligned}
\mathbb{P}(X \geq k) &\leq \frac{1}{\lambda k} \\
-\mathbb{P}(X \geq k) &\geq -\frac{1}{\lambda k} && \text{multiplying be a negative flips the inequality} \\
1 - \mathbb{P}(X \geq k) &\geq 1 - \frac{1}{\lambda k} \\
\mathbb{P}(X < k) &\geq 1 - \frac{1}{\lambda k} && \text{by definition of complement}
\end{aligned}$$

Note that because we took the complement and the sign flipped, we have now found a *lower* bound for $\mathbb{P}(X < k)$.

c) Use Chebyshev's inequality to bound $\mathbb{P}(X \geq k)$.

We rearrange algebraically to get into the form to apply Chebyshev's inequality. We then plug in the corresponding values and $Var(X) = \frac{1}{\lambda^2}$.

$$\mathbb{P}(X \geq k) = \mathbb{P}\left(X - \frac{1}{\lambda} \geq k - \frac{1}{\lambda}\right) \leq \mathbb{P}\left(\left|X - \frac{1}{\lambda}\right| \geq k - \frac{1}{\lambda}\right) \leq \frac{1}{\lambda^2(k - 1/\lambda)^2} = \frac{1}{(\lambda k - 1)^2}$$

d) What is the exact formula for $\mathbb{P}(X \geq k)$?

Using the CDF for an exponential distribution and the definition of complement:

$$\mathbb{P}(X \geq k) = 1 - \mathbb{P}(X \leq k) = 1 - (1 - e^{-\lambda k}) = e^{-\lambda k}$$

e) For $\lambda k \geq 3$, how do the bounds given in parts (a), (c), and (d) compare?

$$e^{-\lambda k} < \frac{1}{(\lambda k - 1)^2} < \frac{1}{\lambda k}$$

so Markov's inequality gives the worst bound.

17 Robbie's Late!

Suppose the probability Robbie is late to teaching lecture on a given day is at most 0.01. Do not make any independence assumptions.

a) Use a Union Bound to bound the probability that Robbie is late at least once over a 30-lecture quarter.

Let R_i be the event Robbie is late to lecture on day i for $i = 1, ..., 30$. Then, by the

union bound,

$$\begin{aligned}
\mathbb{P}(\text{late at least once}) &= \mathbb{P}(\bigcup_{i=1}^{30} R_i) \\
&\leq \sum_{i=1}^{30} \mathbb{P}(R_i) \quad \text{[union bound]} \\
&\leq \sum_{i=1}^{30} 0.01 \quad [\mathbb{P}(R_i) \leq 0.01] \\
&= 0.30
\end{aligned}$$

b) Use a Union Bound to bound the probability that Robbie is **never** late over a 30-lecture quarter.

As in the previous part, let R_i be the event Robbie is late to lecture on day i for $i = 1, ..., 30$. Then, by the union bound, we found that

$$\mathbb{P}(\text{late at least once}) \leq 0.30$$

The probability Robbie is never late is the complement of the probability he is late at least once over the 30 lectures. Taking the complement and doing algebra:

$$\begin{aligned}
\mathbb{P}(\text{late at least once}) &\leq 0.30 \\
-\mathbb{P}(\text{late at least once}) &\geq -0.30 \quad \text{[multiplying by negative flips the inequality]} \\
1 - \mathbb{P}(\text{late at least once}) &\geq 1 - 0.30 \\
\mathbb{P}(\text{never late}) &\geq 0.70
\end{aligned}$$

Note that we have now found a *lower* bound for this probability using the union bound because of taking the complement.

c) Use a Union Bound to bound the probability that Robbie is late at least once over a 120-lecture quarter.

Let R_i be the event Robbie is late to lecture on day i for $i = 1, ..., 120$. Then, by the union bound,

$$\begin{aligned}
\mathbb{P}(\text{late at least once}) &= \mathbb{P}(\bigcup_{i=1}^{120} R_i) \\
&\leq \sum_{i=1}^{120} \mathbb{P}(R_i) \quad \text{[union bound]} \\
&\leq \sum_{i=1}^{120} 0.01 \quad [\mathbb{P}(R_i) \leq 0.01] \\
&= 1.20
\end{aligned}$$

Notice that $\mathbb{P}(\text{late at least once}) \leq 1.20$ is not a very helpful bound since probabilities have to be at most 1 already.