1. Review of Main Concepts

(a) Realization/Sample: A realization/sample $x$ of a random variable $X$ is the value that is actually observed.

(b) Likelihood: Let $x_1, \ldots, x_n$ be iid realizations from probability mass function $p_X(x; \theta)$ (if $X$ discrete) or density $f_X(x; \theta)$ (if $X$ continuous), where $\theta$ is a parameter (or a vector of parameters). We define the likelihood function to be the probability of seeing the data.

If $X$ is discrete:

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

If $X$ is continuous:

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

(c) Maximum Likelihood Estimator (MLE): We denote the MLE of $\theta$ 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).

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

(d) 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 $\theta$ 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:

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

(e) Bias: The bias of an estimator $\hat{\theta}$ for a true parameter $\theta$ is defined as $\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta$. An estimator $\hat{\theta}$ of $\theta$ is unbiased iff $\text{Bias}(\hat{\theta}, \theta) = 0$, or equivalently $\mathbb{E}[\hat{\theta}] = \theta$.

(f) 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 and set it to 0 to find a candidate for the MLE, $\hat{\theta}$.

(c) 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.

(g) A discrete-time stochastic process (DTSP) is a sequence of random variables $X_0, X_1, X_2, \ldots$, where $X_t$ is the value at time $t$. For example, the temperature in Seattle or stock price of TESLA each day, or which node you are at after each time step on a random walk on a graph.

(h) A Markov Chain is a DTSP, with the additional following three properties:
I. ...has a finite (or countably infinite) **state space** $S = \{s_1, \ldots, s_n\}$ which it bounces between, so each $X_t \in S$.

II. ...satisfies the **Markov property**. A DTSP satisfies the Markov property if the future is (conditionally) independent of the past given the present. Mathematically, it means, $P(X_{t+1} = x_{t+1} | X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = x_t) = P(X_{t+1} = x_{t+1} | X_t = x_t)$.

III. ...has **stationary transition probabilities**. Meaning, if we are at some state $s_i$, we transition to another state $s_j$ with probability independent of the current time. Due to this property and the previous, the transitions are governed by $n^2$ probabilities: the probability of transitioning from one of $n$ current states to one of $n$ next states. These are stored in a square $n \times n$ **transition probability matrix** (TPM) $P$, where $P_{ij} = P(X_{t+1} = s_j | X_t = s_i)$ is the probability of transitioning from state $s_i$ to state $s_j$ for any/every value of $t$.

2. **312 Grades**

Suppose Professor Karlin loses everyone’s grades for 312 and decides to make it up by assigning grades randomly according to the following probability distribution, and hoping the $n$ students won’t notice: give an A with probability $0.5$, a B with probability $\theta$, a C with probability $2\theta$, and an F with probability $0.5 - 3\theta$. Each student is assigned a grade independently. Let $x_A$ be the number of people who received an A, $x_B$ the number of people who received a B, etc, where $x_A + x_B + x_C + x_F = n$. Find the MLE for $\theta$, $\hat{\theta}$.

3. **A Red Poisson**

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

4. **Independent Shreds, You Say?**

You are given 100 independent samples $x_1, x_2, \ldots, x_{100}$ from Bernoulli($\theta$), where $\theta$ is unknown. (Each sample is either a 0 or a 1). These 100 samples sum to 30. You would like to estimate the distribution’s parameter $\theta$. Give all answers to 3 significant digits.

   (a) What is the maximum likelihood estimator $\hat{\theta}$ of $\theta$?

   (b) Is $\hat{\theta}$ an unbiased estimator of $\theta$?

5. **Y Me?**

Let $y_1, y_2, \ldots, y_n$ be i.i.d. samples of a random variable with density function

$$f_Y(y|\theta) = \frac{1}{2\theta} \exp\left(-\frac{|y|}{\theta}\right).$$

Find the MLE for $\theta$ in terms of $|y_i|$ and $n$.

6. **A biased estimator**

In class, we showed that the maximum likelihood estimate of the variance $\theta_2$ of a normal distribution (when both the true mean $\mu$ and true variance $\sigma^2$ 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?
7. Faulty Machines
You are trying to use a machine that only works on some days. If on a given day, the machine is working it will break down the next day with probability $0 < b < 1$, and works on the next day with probability $1 - b$. If it is not working on a given day, it will work on the next day with probability $0 < r < 1$ and not work the next day with probability $1 - r$.

(a) In this problem we will formulate this process as a Markov chain. First, let $X_t$ be a random variable that denotes the state of the machine at time $t$. Then, define a state space $S$ that includes all the possible states that the machine can be in. Lastly, for all $A, B \in S$ find $P(X_{t+1} = A \mid X_t = B)$ ($A$ and $B$ can be the same state).

(b) Suppose that on day 1, the machine is working. What is the probability that it is working on day 3?

(c) As $n \to \infty$, what does the probability that the machine is working on day $n$ converge to? To get the answer, solve for the stationary distribution.

8. Three tails
You flip a fair coin until you see three tails in a row. Model this as a Markov chain with the following states:
- $S$: start state, which we are only in before flipping any coins.
- $H$: We see a heads, which means no streak of tails currently exists.
- $T$: We’ve seen exactly one tail in a row so far.
- $TT$: We’ve seen exactly two tails in a row so far.
- $TTT$: We’ve accomplished our goal of seeing three tails in a row and stop flipping.

(a) Write down the transition probability matrix.

(b) Write down the system of equations whose variables are $D(s)$ for each state $s \in \{S, H, T, TT, TTT\}$, where $D(s)$ is the expected number of steps until state $TTT$ is reached starting from state $s$. Solve this system of equations to find $D(S)$.

(c) Write down the system of equations whose variables are $\gamma(s)$ for each state $s \in \{S, H, T, TT, TTT\}$, where $\gamma(s)$ is the expected number of heads seen before state $TTT$ is reached. Solve this system to find $\gamma(S)$, the expected number of heads seen overall until getting three tails in a row.

9. Another Markov chain
Suppose that the following is the transition probability matrix for a 4 state Markov chain (states 1,2,3,4).

\[
P = \begin{bmatrix}
0 & 1/2 & 1/2 & 0 \\
1/3 & 0 & 0 & 2/3 \\
1/3 & 1/3 & 0 & 1/3 \\
1/5 & 2/5 & 2/5 & 0
\end{bmatrix}
\]

(a) What is the probability that $X_2 = 4$ given that $X_0 = 4$?

(b) Write down the system of equations that the stationary distribution must satisfy and solve them.
10. Law of Total Probability Review

(a) (Discrete version) Suppose we flip a coin with probability $U$ of heads, where $U$ is equally likely to be one of $\Omega_U = \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ (notice this set has size $n + 1$). Let $H$ be the event that the coin comes up heads. What is $P(H)$?

(b) (Continuous version) Now suppose $U \sim \text{Uniform}(0,1)$ has the continuous uniform distribution over the interval $[0, 1]$. What is $P(H)$?

(c) Let’s generalize the previous result we just used. Suppose $E$ is an event, and $X$ is a continuous random variable with density function $f_X(x)$. Write an expression for $P(E)$, conditioning on $X$.

11. Poisson CLT practice

Suppose $X_1, \ldots, X_n$ are iid Poisson($\lambda$) random variables, and let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, the sample mean. How large should we choose $n$ to be such that $P(\frac{\lambda}{2} \leq \bar{X}_n \leq \frac{3\lambda}{2}) \geq 0.99$? Use the CLT and give an answer involving $\Phi^{-1}(\cdot)$. Then evaluate it exactly when $\lambda = 1/10$ using the $\Phi$ table on the last page.