1. (10 points) **Test Statistics:** Suppose you have a random sample of size n from a $N(\mu, \sigma^2)$ PDF. The Neyman-Pearson Lemma gives the following 6 statistics (or slight variations thereof) as the best uses of the data under various assumptions:

$$(a)\ \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \qquad (b)\ \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \qquad (c)\ \frac{(n-1)\,S^2}{\sigma_0^2}$$

$$(d)\ \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \qquad (e)\ \frac{\bar{X}-\mu}{S/\sqrt{n}} \qquad (f)\ \frac{(n-1)\,S^2}{\sigma^2}.$$

   Explain when and how you should use these statistics appropriately for testing and reporting confidence intervals.

   Hint: State the purpose and assumptions about the normal parameters.

2. (10 points) **Equivalence of Tests and Confidence Intervals:** Suppose you have a random sample of size n from a $N(\mu, \sigma^2)$ PDF with unknown mean and variance. You wish to test the null hypothesis $H_0 : \mu = \mu_0$ versus the two-sided alternative $H_0 : \mu \neq \mu_0$.
   (a) (2 points) Using the results of Section 7.3.2, what is the form of the best critical region at the 5% level? What is the form of the 95% confidence interval?
   (b) (4 points) Suppose you rejected H_0 in part (a). What is the relationship between the location of μ_0 and the confidence interval?
   (c) (4 points) On the other hand, suppose you failed to reject H_0 in part (a). Now what is the relationship between the location of μ_0 and the confidence interval?

   Hint: Your answer should be both algebraic and graphical.

3. (10 points) A Two-Sample Confidence Interval

   In 1963, Dr. Natrella of the National Bureau of Standards (now NIST) used two methods to measure the latent heat of fusion of ice. Data were collected measuring the change in total heat from ice at $-0.72°C$ to water $0°C$ in calories per gram of mass:

Method I (x)	Method II (y)
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	

You may assume the data are normally distributed for each method, and that the variances are the same for each method. However, a casual inspection of the data casts some doubt that the methods are both properly calibrated. Construct a 95% confidence interval for $\mu_y - \mu_x$. What is your conclusion?

Hints: $n_x = 13$, $n_y = 8$, $\bar{x} = 80.02077$, $\bar{y} = 79.97875$,

$$\sum(x_i - \bar{x})^2 = 0.006892308, \text{ and } \sum(y_i - \bar{y})^2 = 0.006887500\,.$$

4. (15 points) Suppose we have a random sample $\{X_1, X_2, \ldots, X_n\}$ from a Poisson PMF, $p(x|m)$, that is, a single parameter $\theta = m$.
   (a) (2 points) Write out $p(x|m)$, $\log p(x|m)$, and the log-likelihood $\ell(m|x_1, x_2, \ldots, x_n)$?
   (b) (3 points) Find the MLE $\widehat{m}$? (You do not need to check the second derivative.)
   (c) (3 points) Compute the mean and variance of your $\widehat{m}$?
   (d) (3 points) What is the Fisher Information for a single sample?
   (e) (3 points) For any unbiased estimator $\hat{\theta}$ of m, what is the Cramér-Rao Lower Bound on the variance of $\hat{\theta}$?
   (f) (1 point) Is your MLE 100% efficient?

5. (10 points) χ^2-**Goodness-of-Fit Test:**
   (a) (5 points) The number of births at the Cedar Rapids, Iowa hospital in 2019 was

$$\boldsymbol{o} = (244, 238, 257, 261)$$

   for the 4 quarters. Test the null hypothesis that the number of births is constant for each quarter at the 5% level.
   (b) (5 points) What is the p-value?

6. (10 points) χ^2-**Test of Independence:** A web designer wonders if sales at her new site vary by the day of the week. After her web site is stable, she collects data for one week:

Observed	M	T	W	T	F	S	S	Total
no purchase	399	261	284	263	393	531	502	2633
single purchase	119	72	97	51	143	145	150	777
multiple purchases	39	50	20	15	41	97	86	348
Total	557	383	401	329	577	773	738	3758

   (a) (8 points) Test the null hypothesis that the number of purchases is the same for all days of the week, versus the alternative that some are different. Use $\alpha = 1\%$.
   (b) (2 points) What is the p-value?

7. (10 points) **Correlation Test and Confidence Interval:** For the $n = 1{,}078$ Father and Son height data discussed in Chapter 1, the sample correlation is $R = 0.50134$. One theory of inheritance suggests their correlation should be 50%.

(a) (5 points) At the $\alpha = 5\%$ level, test for the hypothesis

$$H_0 : \rho = \frac{1}{2} \quad \text{versus} \quad H_1 : \rho \neq \frac{1}{2}.$$

What is the value of your test statistic? Your decision?

(b) (5 points) What is the 95% confidence interval for ρ?

8. (15 points) **Linear Regression:** Using the data $\{(x_i, y_i),\ i = 1, \ldots, 8\}$ from the 2011–2018 for the annual number of murders in Houston

Year	2011	2012	2013	2014	2015	2016	2017	2018
Number	198	216	214	241	303	302	269	279

we summarize

$$\sum_{i=1}^{8} x_i = 16{,}116 \qquad \sum_{i=1}^{8} (x_i - \bar{x})^2 = 42$$

$$\sum_{i=1}^{8} y_i = 2{,}022 \qquad \sum_{i=1}^{8} (x_i - \bar{x})\, y_i = 579$$

(a) (5 points) Find the best fitting least-squares straight line and add it to a scatter diagram with $2009 \le x \le 2020$. Give the constants $\hat{a}$ and $\hat{b}$, using the form given in Eqn (8.29).

(b) (5 points) For the years 2009-2020, add to your plot (1) the 95% confidence intervals for your predictions and (2) the 95% confidence interval for the actual number of murders?

(c) (5 points) What is your prediction for the number of murders in 2020 and a 95% confidence interval for it (population, not mean)?

Hints: In part (b), the first CI is for the conditional mean; the second CI is for the number of murders for each year. Do not worry about adjusting the α-level due to multiple testing (i.e., no Bonferroni adjustment to α). Also, you can check the identity $\sum(x_i - \bar{x})y_i = \sum(x_i - \bar{x})(y_i - \bar{y})$.

9. (10 points) **An MLE Problem:** Suppose we have a random sample $\{X_1, X_2, \ldots, X_n\}$ from the following PDF, defined piecewise on the closed interval $[0, \theta]$ as either

$$f(x|\theta) = \begin{cases} \frac{3}{2}c & 0 \le x < \frac{\theta}{2} \\ c & \frac{\theta}{2} \le x \le \theta \\ 0 & \text{elsewhere} \end{cases}$$

or as

$$f(x|\theta) = \begin{cases} \frac{3}{2}c & 0 \le x \le \frac{\theta}{2} \\ c & \frac{\theta}{2} < x \le \theta \\ 0 & \text{elsewhere.} \end{cases}$$

What is c so $\int f(x|\theta) = 1$? Find the MLE, $\hat{\theta}$, for the small (sorted) dataset

$$x = (.07, .47, .90, 1.64, 1.77)\,.$$

Hints: c will vary with $\hat{\theta}$. Recall the sorted data are denoted by the n order statistics

$$x_{(1)} < x_{(2)} < x_{(3)} < \cdots < x_{(n-1)} < x_{(n)}.$$

The sample interval given your estimate $\hat{\theta}$ is either

$$\left[0, \frac{\hat{\theta}}{2}\right) \uplus \left[\frac{\hat{\theta}}{2}, \hat{\theta}\right] \quad \text{or} \quad \left[0, \frac{\hat{\theta}}{2}\right] \uplus \left(\frac{\hat{\theta}}{2}, \hat{\theta}\right],$$

whichever gives a greater likelihood. (The solution is not as simple as previous versions of the uniform PDF. Make sketches.)