50 Pen and Paper Exercises on Statistical Science (with Solutions)

By Tanujit Chakraborty (Webpage)

For UG and PG students in Mathematics, Statistics, Economics, Data Science, and Computer Science

Topics to be covered (Exercises and Solutions):

1. Transformation of Random Variables and Sampling Distributions
2. Limit Theorems and Order Statistics
3. Theory of Point Estimation
4. Confidence Intervals
5. Bayesian Inference
6. Testing of Hypothesis
7. Statistical Simulation using R
8. Linear and Nonlinear Regression Analysis
9. Multicollinearity and Shrinkage Methods
10. Logistic Regression

Acknowledgment: These problems and their solutions are largely based on some extraordinary books on Statistics and several outstanding university courses available online, with some additions by the author. A non-exhaustive list is given below. Happy Learning!

Note: To err is human! There can be typos and unwanted mistakes in the solutions. If you find one, please email me at ctanujit@gmail.com.

References (Books):

1. John A. Rice. Mathematical Statistics and Data Analysis, 2007.
2. Draper, Norman R., and Harry Smith. Applied Regression Analysis, 1998.
3. Shao, Jun. Mathematical statistics, 2003.

References (Online Resources):

1. STAT 200: Introduction to Statistical Inference at Stanford University.
2. NPTEL Regression Analysis Course at IIT Kharagpur.
3. ISI MSTAT and IITJAM MS Previous Papers.

Statistical Science - Pen and Paper Exercises (with Solutions)

Problem 1. (Bivariate and Multivariate Distributions)

(a) **(Accounting for voter turnout)** Let N be the number of people in the state of Iowa. Suppose pN of these people support Hillary Clinton, and $(1-p)N$ of them support Donald Trump, for some $p \in (0,1)$. N is known (say $N = 3,000,000$) and p is unknown.

(i) Suppose that each person in Iowa randomly and independently decides, on election day, whether or not to vote, with probability 1/2 of voting and probability 1/2 of not voting. Let V_{Hillary} be the number of people who vote for Hillary and V_{Donald} be the number of people who vote for Donald. Show that

$$\mathbb{E}\left[V_{\text{Hillary}}\right] = \frac{1}{2}pN, \quad \mathbb{E}\left[V_{\text{Donald}}\right] = \frac{1}{2}(1-p)N.$$

What are the standard deviations of V_{Hillary} and V_{Donald}, in terms of p and N? Explain why, when N is large, we expect the fraction of voters who vote for Hillary to be very close to p.

(ii) Now, suppose there are two types of voters - "passive" and "active." Each passive voter votes on election day with probability 1/4 and doesn't vote with probability 3/4, while each active voter votes with probability 3/4 and doesn't vote with probability 1/4. Suppose that a fraction q_H of the people who support Hillary are passive and $1-q_H$ are active, and a fraction q_D of the people who support Donald are passive and $1-q_D$ are active. Show that

$$\mathbb{E}\left[V_{\text{Hillary}}\right] = \frac{1}{4}q_H pN + \frac{3}{4}\left(1-q_H\right)pN, \quad \mathbb{E}\left[V_{\text{Donald}}\right] = \frac{1}{4}q_D(1-p)N + \frac{3}{4}\left(1-q_D\right)(1-p)N.$$

What are the standard deviations of V_{Hillary} and V_{Donald} in terms of p, N, q_H, and q_D? If we estimate p by $\hat{p}$ using a simple random sample of $n = 1000$ people from Iowa. Explain why $\hat{p}$ might not be a good estimate of the fraction of voters who will vote for Hillary.

(iii) We do not know q_H and q_D. However, suppose that in our simple random sample, we can observe whether each person is passive or active, in addition to asking them whether they support Hillary or Donald. Suggest estimators $\hat{V}_{\text{Hillary}}$ and $\hat{V}_{\text{Donald}}$ for $\mathbb{E}\left[V_{\text{Hillary}}\right]$ and $\mathbb{E}\left[V_{\text{Donald}}\right]$ using this additional information. Show, for your estimators, that

$$\mathbb{E}\left[\hat{V}_{\text{Hillary}}\right] = \frac{1}{4}q_H pN + \frac{3}{4}\left(1-q_H\right)pN, \quad \mathbb{E}\left[\hat{V}_{\text{Donald}}\right] = \frac{1}{4}q_D(1-p)N + \frac{3}{4}\left(1-q_D\right)(1-p)N.$$

Solution: (i) Recall that a Binomial(n,p) random variable has mean np and variance $np(1-p)$. A total of pN people support Hillary, each voting independently with probability $\frac{1}{2}$, so $V_{\text{Hillary}} \sim \text{Binomial}\left(pN, \frac{1}{2}\right)$. Then

$$\mathbb{E}\left[V_{\text{Hillary}}\right] = \frac{1}{2}pN, \quad \operatorname{Var}\left[V_{\text{Hillary}}\right] = \frac{1}{4}pN,$$

and the standard deviation of V_{Hillary} is $\sqrt{\frac{1}{4}pN}$. Similarly, as $(1-p)N$ people support Donald, $V_{\text{Donald}} \sim \text{Binomial}\left((1-p)N, \frac{1}{2}\right)$, so

$$\mathbb{E}\left[V_{\text{Donald}}\right] = \frac{1}{2}(1-p)N, \quad \operatorname{Var}\left[V_{\text{Donald}}\right] = \frac{1}{4}(1-p)N,$$

and the standard deviation of V_{Donald} is $\sqrt{\frac{1}{4}(1-p)N}$. The fraction of voters who voted for Hillary is

$$\frac{V_{\text{Hillary}}}{V_{\text{Hillary}} + V_{\text{Donald}}} = \frac{V_{\text{Hillary}}/N}{V_{\text{Hillary}}/N + V_{\text{Donald}}/N}.$$

As $\mathbb{E}\left[V_{\text{Hillary}}/N\right]=\frac{1}{2}p$ (a constant) and $\operatorname{Var}\left[V_{\text{Hillary}}/N\right]=\frac{1}{4}p/N\to 0$ as $N\to\infty$, V_{Hillary}/N should be close to $\frac{1}{2}p$ with high probability when N is large. Similarly, as $\mathbb{E}\left[V_{\text{Donald}}/N\right]=\frac{1}{2}(1-p)$ and $\operatorname{Var}\left[V_{\text{Donald}}/N\right]=\frac{1}{4}(1-p)/N\to 0$ as $N\to\infty$, V_{Donald}/N should be close to $\frac{1}{2}(1-p)$ with high probability when N is large. Then the fraction of voters for Hillary should, with high probability, be close to

$$\frac{\frac{1}{2}p}{\frac{1}{2}p+\frac{1}{2}(1-p)}=p.$$

(The above statement "close to with high probability" may be formalized using Chebyshev's inequality, which states that a random variable is, with high probability, not too many standard deviations away from its mean.)

(ii) Let $V_{\text{H,p}}$ and $V_{\text{H,a}}$ be the number of passive and active voters who vote for Hillary, and similarly define $V_{\text{D,p}}$ and $V_{\text{D,a}}$ for Donald. There are $q_H pN$ passive Hillary supporters, each of whom votes independently with probability $\frac{1}{4}$, so

$$V_{\text{H,p}}\sim\text{Binomial}\left(q_H pN,\frac{1}{4}\right).$$

Similarly,

$$\begin{aligned}V_{\text{H,a}}&\sim\text{Binomial}\left(\left(1-q_H\right)pN,\frac{3}{4}\right),\\ V_{\text{D,p}}&\sim\text{Binomial}\left(q_D(1-p)N,\frac{1}{4}\right),\\ V_{\text{D,a}}&\sim\text{Binomial}\left(\left(1-q_D\right)(1-p)N,\frac{3}{4}\right),\end{aligned}$$

and these four random variables are independent. Since $V_{\text{Hillary}}=V_{\text{H,p}}+V_{\text{H,a}}$,

$$\begin{aligned}\mathbb{E}\left[V_{\text{Hillary}}\right]&=\mathbb{E}\left[V_{\text{H,p}}\right]+\mathbb{E}\left[V_{\text{H,a}}\right]=\frac{1}{4}q_H pN+\frac{3}{4}\left(1-q_H\right)pN,\\ \operatorname{Var}\left[V_{\text{Hillary}}\right]&=\operatorname{Var}\left[V_{\text{H,p}}\right]+\operatorname{Var}\left[V_{\text{H,a}}\right]=\frac{3}{16}q_H pN+\frac{3}{16}\left(1-q_H\right)pN=\frac{3}{16}pN,\end{aligned}$$

and the standard deviation of V_{Hillary} is $\sqrt{\frac{3}{16}pN}$. Similarly,

$$\begin{aligned}\mathbb{E}\left[V_{\text{Donald}}\right]&=\mathbb{E}\left[V_{\text{D,p}}\right]+\mathbb{E}\left[V_{\text{D,a}}\right]=\frac{1}{4}q_D(1-p)N+\frac{3}{4}\left(1-q_D\right)(1-p)N,\\ \operatorname{Var}\left[V_{\text{Donald}}\right]&=\operatorname{Var}\left[V_{\text{D,p}}\right]+\operatorname{Var}\left[V_{\text{D,a}}\right]=\frac{3}{16}q_D(1-p)N+\frac{3}{16}\left(1-q_D\right)(1-p)N\\ &=\frac{3}{16}(1-p)N,\end{aligned}$$

and the standard deviation of V_{Donald} is $\sqrt{\frac{3}{16}(1-p)N}$.

The quantity $\hat{p}$ estimates p, but in this case, p may not be the fraction of voters who vote for Hillary: By the same argument as in part (a), the fraction of voters who vote for Hillary is given by

$$\begin{aligned}\frac{V_{\text{Hillary}}}{V_{\text{Hillary}}+V_{\text{Donald}}}&=\frac{V_{\text{Hillary}}/N}{V_{\text{Hillary}}/N+V_{\text{Donald}}/N}\\ &\approx\frac{\frac{1}{4}q_H p+\frac{3}{4}\left(1-q_H\right)p}{\frac{1}{4}q_H p+\frac{3}{4}\left(1-q_H\right)p+\frac{1}{4}q_D(1-p)+\frac{3}{4}\left(1-q_D\right)(1-p)},\end{aligned}$$

where the approximation is accurate with high probability when N is large. When $q_H\neq q_D$, this is different from p: For example, if $q_H=0$ and $q_D=1$, this is equal to $\frac{p}{p+(1-p)/3}$ which is greater than p, reflecting the fact that Hillary supporters are more likely to vote than are Donald supporters.

(iii) Let $\hat{p}$ be the proportion of the 1000 surveyed people who support Hillary. Among the surveyed people supporting Hillary, let $\hat{q}_H$ be the proportion who are passive. Similarly, among the surveyed people supporting

Donald, let $\hat{q}_D$ be the proportion who are passive. (Note that these are observed quantities computed from our sample of 1000 people.) Then we may estimate the number of voters for Hillary and Donald by

$$\begin{aligned}\hat{V}_{\text{Hillary}} &= \frac{1}{4}\hat{q}_H\hat{p}N + \frac{3}{4}\left(1-\hat{q}_H\right)\hat{p}N \\ \hat{V}_{\text{Donald}} &= \frac{1}{4}\hat{q}_D(1-\hat{p})N + \frac{3}{4}\left(1-\hat{q}_D\right)(1-\hat{p})N.\end{aligned}$$

$\hat{q}_H\hat{p}$ is simply the proportion of the 1000 surveyed people who both support Hillary and are passive. Hence, letting $X_1, \ldots, X_{1000}$ indicate whether each surveyed person both supports Hillary and is passive, we have

$$\hat{q}_H\hat{p} = \frac{1}{n}\left(X_1 + \ldots + X_n\right).$$

Each $X_i \sim \text{Bernoulli}\left(q_H p\right)$, so linearity of expectation implies $\mathbb{E}\left[\hat{q}_H\hat{p}\right] = q_H p$. Similarly, $\left(1-\hat{q}_H\right)\hat{p}, \hat{q}_D(1-\hat{p})$, and $\left(1-\hat{q}_D\right)(1-\hat{p})$ are the proportions of the 1000 surveyed people who support Hillary and are active, support Donald and are passive, and support Donald and are active, so the same argument shows $\mathbb{E}\left[\left(1-\hat{q}_H\right)\hat{p}\right] = \left(1-q_H\right)p, \mathbb{E}\left[\hat{q}_D(1-\hat{p})\right] = q_D(1-p)$, and $\mathbb{E}\left[\left(1-\hat{q}_D\right)(1-\hat{p})\right] = \left(1-q_D\right)(1-p)$. Then applying linearity of expectation again yields

$$\mathbb{E}\left[\hat{V}_{\text{Hillary}}\right] = \mathbb{E}\left[V_{\text{Hillary}}\right], \quad \mathbb{E}\left[\hat{V}_{\text{Donald}}\right] = \mathbb{E}\left[V_{\text{Donald}}\right].$$

(b) Let X and Y be continuous random variables with joint density function

$$f(x,y) = \begin{cases} \frac{y^3}{2}e^{-y(x+1)} & \text{for } x>0, y>0 \\ 0 & \text{otherwise.}\end{cases}$$

(i) Find the marginal pdf of X.

(ii) Find the marginal pdf of Y and $E[Y]$.

Solution: (i) Integrating the joint pdf over y gives the marginal pdf of X. For $x>0$, we have

$$f_X(x) = \int_0^\infty \frac{y^3}{2}e^{-y(x+1)}dy = \frac{\Gamma(4)}{2(x+1)^4}\int_0^\infty \frac{(x+1)^4}{\Gamma(4)}y^3e^{-y(x+1)}dy = \frac{3}{(x+1)^4},$$

as the integrand in the last integral over y is a Gamma$(4, x+1)$ pdf, and so integrates (over y) to 1. For $x \le 0, f_X(x) = 0$.

(ii) Integrating the joint pdf over x gives the marginal pdf of Y. For $y>0$, we have

$$f_Y(y) = \int_0^\infty \frac{y^3}{2}e^{-y(x+1)}dx = \frac{y^2}{2}e^{-y}\int_0^\infty ye^{-yx}dx = \frac{y^2}{2}e^{-y}$$

as the integrand in the last integral over x is an Exponential (y) pdf, and so integrates (over x) to 1 . For $y \le 0, f_Y(y) = 0$. We can recognize the marginal distribution of Y to be Gamma$(3,1)$, and so $E[Y] = 3/1 = 3$.

(c) **(Existence of multivariate normal)**

(i) Suppose $\left(X_1, \ldots, X_k\right) \sim \mathcal{N}(0, \Sigma)$ for a covariance matrix $\Sigma \in \mathbb{R}^{k\times k}$. Let $Y_1, \ldots, Y_m$ be linear combinations of $X_1, \ldots, X_k$, given by

$$Y_j = a_{j1}X_1 + \ldots + a_{jk}X_k$$

for each $j = 1, \ldots, m$ and some constants $a_{j1}, \ldots, a_{jk} \in \mathbb{R}$. Consider the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mk}\end{pmatrix}.$$

By computing the means, variances, and covariances of $Y_1, \ldots, Y_m$, show that

$$\left(Y_1, \ldots, Y_m\right) \sim \mathcal{N}\left(0, A\Sigma A^T\right).$$

(ii) Let $A \in \mathbb{R}^{k \times k}$ be any matrix, let $\Sigma = AA^T$, and let $\mu \in \mathbb{R}^k$ be any vector. Show that there exist random variables $Y_1, \ldots, Y_k$ such that $(Y_1, \ldots, Y_k) \sim \mathcal{N}(\mu, \Sigma)$. (Hint: Let $X_1, \ldots, X_k \overset{IID}{\sim} \mathcal{N}(0,1)$, and let each Y_j be a certain linear combination of $X_1, \ldots, X_k$ plus a certain constant.)

Solution: (i) Any linear combination of $Y_1, \ldots, Y_m$ is a linear combination of $X_1, \ldots, X_k$, so $(Y_1, \ldots, Y_m)$ is multivariate normal. Using linearity of expectation and bilinearity of covariance, we compute

$$\begin{aligned} \mathbb{E}[Y_i] &= a_{i1}\mathbb{E}[X_1] + \ldots + a_{ik}\mathbb{E}[X_k] = 0, \\ \operatorname{Cov}[Y_i, Y_j] &= \operatorname{Cov}[a_{i1}X_1 + \ldots + a_{ik}X_k, a_{j1}X_1 + \ldots + a_{jk}X_k] \\ &= \sum_{r=1}^{k}\sum_{s=1}^{k} a_{ir}a_{js} \operatorname{Cov}[X_r, X_s] = \sum_{r=1}^{k}\sum_{s=1}^{k} a_{ir}a_{js}\Sigma_{rs} = a_i \Sigma a_j^T, \end{aligned}$$

where a_i and a_j denote rows i and j of the matrix A. (The computation of covariance is valid for both $i \neq j$ and $i = j$; in the latter case this yields $\operatorname{Var}[Y_i] = \operatorname{Cov}[Y_i, Y_i]$.) As $a_i \Sigma a_j^T = \left(A\Sigma A^T\right)_{ij}$, this implies by definition $Y \sim \mathcal{N}\left(0, A\Sigma A^T\right)$.

(ii) Take $X_1, \ldots, X_k \overset{iid}{\sim} \mathcal{N}(0,1)$, define $Z_j = a_{j1}X_1 + \ldots + a_{jk}X_k$ for each $j = 1, \ldots, k$, and let $Y_j = Z_j + \mu_j$. As $(X_1, \ldots, X_k)$ have the multivariate normal distribution $\mathcal{N}(0, I)$ where I is the $k \times k$ identity matrix, and as $AIA^T = AA^T = \Sigma$, part (a) implies $(Z_1, \ldots, Z_k) \sim \mathcal{N}(0, \Sigma)$. Then $(Y_1, \ldots, Y_k) \sim \mathcal{N}(\mu, \Sigma)$ (since adding the vector $\mu = (\mu_1, \ldots, \mu_k)$ does not change the variances and covariances of $Y_1, \ldots, Y_k$ but shifts their means by $\mu_1, \ldots, \mu_k$).

(d) (i) Let (X, Y) be a random point uniformly distributed on the unit disk $\{(x,y) : x^2 + y^2 \leq 1\}$. Show that $\operatorname{Cov}[X, Y] = 0$, but that X and Y are not independent.

(ii) Let $X, Y \sim \mathcal{N}(0,1)$ be independent. Compute $\mathbb{P}[X + Y > 0 \mid X > 0]$. (Hint: Use the rotational symmetry of the bivariate normal PDF.)

Solution: (i) X has the same distribution as $-X$, so $\mathbb{E}[X] = \mathbb{E}[-X] = -\mathbb{E}[X]$, hence $\mathbb{E}[X] = 0$. Similarly $\mathbb{E}[Y] = 0$. Also (X, Y) has the same joint distribution as $(-X, Y)$, so $\mathbb{E}[XY] = \mathbb{E}[-XY] = -\mathbb{E}[XY]$, hence $\mathbb{E}[XY] = 0$. Then $\operatorname{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$. On the other hand, conditional on $X = x, Y$ is uniformly distributed on the interval $\left[-\sqrt{1-x^2}, \sqrt{1-x^2}\right]$. As this depends on x, X and Y are not independent.

(ii) We have

$$\mathbb{P}[X + Y > 0 \mid X > 0] = \frac{\mathbb{P}[X + Y > 0, X > 0]}{\mathbb{P}[X > 0]}.$$

Since X has the same distribution as $-X, \mathbb{P}[X > 0] = \mathbb{P}[-X > 0] = \mathbb{P}[X < 0]$, so $\mathbb{P}[X > 0] = \frac{1}{2}$. To compute $\mathbb{P}[X + Y > 0, X > 0]$, note that this is the integral of the bivariate normal PDF $f_{X,Y}(x,y)$ in the region to the right of the y-axis and above the line $y = -x$. The integral of $f_{X,Y}(x,y)$ over all of $\mathbb{R}^2$ must equal 1; hence by rotational symmetry of $f_{X,Y}(x,y)$ around the origin, the integral over any wedge formed by two rays extending from the origin is $\theta/(2\pi)$ where θ is the angle formed by these rays. For the above region, this angle is $3\pi/4$, so $\mathbb{P}[X + Y > 0, X > 0] = 3/8$. Then $\mathbb{P}[X + Y > 0 \mid X > 0] = 3/4$.

(e) (i) A family of Yellow-Faced (YF) gophers consisting of 2 parents and 3 children are kept in a laboratory. In addition to these, a family of YF gophers with 2 parents and 4 children, a family of Big Pocket (BP) gophers with 2 parents and 5 children, and a family of BP gophers with 1 mother and 4 children are also kept in the laboratory. A sample of 4 gophers is selected randomly from all the gophers in the laboratory. What is the probability that the sample consists of one adult female, one adult male, and 2 children, with both adults of the same genus (either both YF or both BP)?

(ii) Let X and Y be arbitrary random variables, and let $g(\cdot)$ and $h(\cdot)$ be arbitrary real valued functions defined on $\mathbb{R}$. For each statement, say whether it is TRUE or FALSE. If TRUE, prove it, and if FALSE, give a counterexample.

(A) If X and Y are uncorrelated then so are $g(X)$ and $h(Y)$ for any g and h.

(B) If $g(X)$ and $h(Y)$ are uncorrelated for all g and h then X and Y are uncorrelated.

Solution: (i) Let X_1 be the number of male YF gophers, X_2 the number of female YF gophers, X_3 the number of male BP gophers, X_4 the number of female BP gophers, and X_5 the number of child gophers in the sample. The sample size is $n = 4$, and the total number of gophers in the laboratory is 23. The total numbers of YF male, YF female, BP male, BP female, and children gophers are $2, 2, 1, 2$ and 16 respectively. The joint distribution of $(X_1, X_2, X_3, X_4, X_5)^T$ is Multivariate Hypergeometric and the desired probability is

$$\begin{aligned}
&P\left(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0, X_5 = 2\right) + P\left(X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 2\right)\\
&= \frac{\binom{2}{1}\binom{2}{1}\binom{1}{0}\binom{2}{0}\binom{16}{2}}{\binom{23}{4}} + \frac{\binom{2}{0}\binom{2}{0}\binom{1}{1}\binom{2}{1}\binom{16}{2}}{\binom{23}{4}}\\
&= \frac{2 \times 2 \times 1 \times 1 \times 120}{8855} + \frac{1 \times 1 \times 1 \times 2 \times 120}{8855} = \frac{720}{8855} \approx 0.0813.
\end{aligned}$$

(ii) (A) FALSE. Take X to have a $N(0, 1)$ distribution. Then $E[X] = 0$ and $E\left[X^3\right] = 0$, and take $Y = X^2$. Then $\mathrm{Cov}(X, Y) = \mathrm{Cov}\left(X, X^2\right) = E\left[X^3\right] - E[X]E\left[X^2\right] = 0$. Take $g(X) = X^2$ and $h(Y) = Y$. Then $\mathrm{Cov}(g(X), h(Y)) = \mathrm{Cov}\left(X^2, X^2\right) = \mathrm{Var}\left(X^2\right) = 2 > 0$, where $\mathrm{Var}\left(X^2\right)$ is the variance of a χ^2_1 distribution.

(B) TRUE. Let A and B be arbitrary subsets of $\mathbb{R}$, and take $g(X) = I_A(X)$ and $h(Y) = I_B(Y)$. Since $\mathrm{Cov}(g(X), h(Y)) = 0$ we have $E\left[I_A(X)I_B(Y)\right] = E\left[I_A(X)\right]E\left[I_B(Y)\right]$, which is equivalent to $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$. Since A and B were arbitrary this implies that X and Y are independent. But then X and Y are uncorrelated.

Problem 2. (Transformation of Random Variables)

(a) Let X_1, X_2, X_3 be independent $N(0, 1)$ random variables. Find the probability density function of $U = X_1^2 + X_2^2 + X_3^2$.

Solution: We know from class that X_i^2 has the χ^2_1 distribution, i.e., the Gamma$(1/2, 1/2)$ distribution. Since X_1, X_2, X_3 are independent, so are X_1^2, X_2^2, X_3^2, so $U = X_1^2 + X_2^2 + X_3^2$ is a sum of 3 independent Gamma$(1/2, 1/2)$ random variables, and as such has the Gamma$(3/2, 1/2)$ distribution (or χ^2_3 distribution). Thus

$$f_U(u) = \begin{cases} \frac{(1/2)^{3/2}}{\Gamma(3/2)} u^{1/2} e^{-(1/2)u} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The expression can be simplified by noting that $\Gamma(3/2) = (1/2)\Gamma(1/2) = \frac{\sqrt{\pi}}{2}$.

(b) Suppose that the random vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$ is uniformly distributed on the sphere of radius 1 centered at the origin; that is, $\boldsymbol{Y}$ has joint probability density function (pdf)

$$f_{\boldsymbol{Y}}\left(y_1, y_2, y_3\right) = \begin{cases} \frac{3}{4\pi} & \text{if } (y_1, y_2, y_3) \in S \\ 0 & \text{otherwise,} \end{cases}$$

where $S = \left\{(y_1, y_2, y_3) \in \mathbb{R}^3 : y_1^2 + y_2^2 + y_3^2 \le 1\right\}$ is the sphere of radius 1 centred at $(0, 0, 0)$. Let $V = Y_1^2 + Y_2^2 + Y_3^2$. Find the probability density function of V and find $E[V]$.

Solution: We can find the pdf of V by first computing the CDF of V. For $v \in [0, 1]$ we have

$$F_V(v) = P(V \le v) = P\left(Y_1^2 + Y_2^2 + Y_3^2 \le (\sqrt{v})^2\right) = P\left((Y_1, Y_2, Y_3) \in S_{\sqrt{v}}\right),$$

where $S_{\sqrt{v}}$ is the sphere of radius $\sqrt{v}$. Thus, for $v \in [0, 1]$

$$\begin{aligned}
F_V(v) &= \iiint_{S_{\sqrt{v}}} f_Y\left(y_1, y_2, y_3\right) dy_1 dy_2 dy_3 = \iiint_{S_{\sqrt{v}}} \frac{3}{4\pi} dy_1 dy_2 dy_3\\
&= \frac{3}{4\pi} \times \text{ Volume of } S_{\sqrt{v}}\\
&= \frac{3}{4\pi} \times \frac{4\pi(\sqrt{v})^3}{3} = v^{3/2}.
\end{aligned}$$

We also have $F_V(v) = 0$ for $v < 0$ and $F_V(v) = 1$ for $v > 1$. Differentiating, we obtain the pdf of V as

$$f_V(v) = F_V'(v) = \begin{cases} \frac{3\sqrt{v}}{2} & \text{for } 0 \leq v \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E[V] = \frac{3}{2}\int_0^1 v^{3/2} dv = \frac{3}{2}\left[\frac{2}{5}v^{5/2}\right]_0^1 = \frac{3}{5}.$$

(c) Let X and Y be independent and identically distributed uniform random variables on the interval $(0, 1)$. Define $U = \frac{1}{2}(X + Y)$ to be the average and define $V = X$.

(i) Find the joint probability density function of (U, V) and draw the sample space of (U, V).

(ii) Find the marginal probability density function of U.

Solution: (i) The inverse transformation is $X = V$ and $Y = 2U - V$. The matrix of partial derivatives of the inverse transformation is

$$\begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}$$

with determinant -2. The joint pdf of (X, Y) is

$$f_{XY}(x, y) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

So the joint pdf of (U, V) is

$$f_{UV}(u, v) = \begin{cases} 2 & \text{for } (u, v) \in S_{UV} \\ 0 & \text{otherwise} \end{cases}$$

where S_{UV} is the sample space of (U, V), determined by the constraints $0 \leq v \leq 1$ and $0 \leq 2u - v \leq 1$, or $0 \leq v \leq 1$ and $v/2 \leq u \leq (v + 1)/2$. The sample space of (U, V) is plotted in Fig. 1.

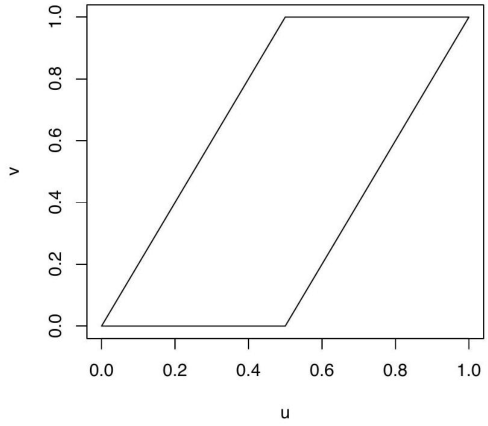


Figure 1: Sample space of (U, V).

(ii) To get the marginal pdf of U we integrate the joint pdf $f_{UV}(u, v)$ over the variable v from $-\infty$ to ∞. From Figure 1, we see that the limits of integration are different depending on whether $u \in [0, 0.5]$ or $u \in (0.5, 1]$. For $u \in [0, 0.5]$ we have

$$f_U(u) = \int_0^{2u} (2) dv = 4u.$$

For $u \in (0.5, 1]$ we have

$$f_U(u) = \int_{2u-1}^{1} (2) dv = 2(1 - (2u - 1)) = 4(1 - u)$$

Clearly, for $u \notin [0,1]$ we have $f_U(u) = 0$. To summarize,

$$f_U(u) = \begin{cases} 4u & \text{for } u \in [0, 0.5] \\ 4(1-u) & \text{for } u \in (0.5, 1] \\ 0 & \text{otherwise.} \end{cases}$$

(d) Let $X = (X_1, X_2)^T$ be uniformly distributed on the positive quadrant intersected with the disk of radius 1 centered at the origin; i.e., X has joint pdf $f_X(x_1, x_2) = \frac{4}{\pi} I_{S_X}(x_1, x_2)$, where

$$S_X = \left\{(x_1, x_2) \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0, x_1^2 + x_2^2 \leq 1\right\}.$$

Let $Y_1 = X_1^2$ and $Y_2 = X_2^2$. Find the joint pdf of $(Y_1, Y_2)^T$ and the marginal pdf of Y_1.

Solution: We use the bivariate change of variable formula to get the joint pdf of (Y_1, Y_2). The inverse transformation is $X_1 = Y_1^{1/2}$ and $X_2 = Y_2^{1/2}$. The Jacobian is

$$\mathbf{J} = \det \begin{bmatrix} \frac{1}{2\sqrt{y_1}} & 0 \\ 0 & \frac{1}{2\sqrt{y_2}} \end{bmatrix} = \frac{1}{4\sqrt{y_1 y_2}}$$

The support of (Y_1, Y_2) is

$$S_Y = \left\{(y_1, y_2) \in \mathbb{R}^2 : y_1 \geq 0, y_2 \geq 0, y_1 + y_2 \leq 1\right\}$$

By the change of variable formula, the joint pdf of (Y_1, Y_2) is given by

$$f_Y(y_1, y_2) = f_X\left(\frac{1}{2\sqrt{y_1}}, \frac{1}{2\sqrt{y_2}}\right)|\mathbf{J}| = \frac{4}{\pi}\frac{1}{4\sqrt{y_1 y_2}} I_{S_Y}(y_1, y_2) = \frac{1}{\pi\sqrt{y_1 y_2}} I_{S_Y}(y_1, y_2).$$

The marginal pdf of Y_1 is obtained by integration. For $y_1 \in [0,1]$,

$$\begin{aligned} f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_Y(y_1, y_2)\, dy_2 &= \int_0^{1-y_1} \frac{1}{\pi\sqrt{y_1 y_2}} dy_2 \\ &= \frac{1}{\pi\sqrt{y_1}} \left[2\sqrt{y_2}\right]_0^{1-y_1} \\ &= \frac{2\sqrt{1-y_1}}{\pi\sqrt{y_1}}, \end{aligned}$$

and $f_{Y_1}(y_1) = 0$ otherwise.

(e) Let X_1, X_2, X_3 be independent and identically distributed Exponential (λ) random variables (the Exponential (λ) distribution has pdf $f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$ and $f_X(x) = 0$ for $x \leq 0$, and df $F_X(x) = 1 - e^{-\lambda x}$ for $x > 0$ and $F_X(x) = 0$ for $x \leq 0$). Find $P\left(X_1 + X_2 + X_3 \leq \frac{3}{2}\right)$. (Write down the appropriate 3-dimensional integral and evaluate it).

Solution: The appropriate 3-dimensional integral is (by symmetry, every order in which the integration is done is equally easy)

$$P\left(X_1 + X_2 + X_3 \leq \frac{3}{2}\right) = \int_0^{3/2} \int_0^{3/2-x_3} \int_0^{3/2-x_2-x_3} f_X(x_1) f_X(x_2) f_X(x_3)\, dx_1 dx_2 dx_3$$

$$\begin{aligned} &= \int_0^{3/2} \int_0^{3/2-x_3} f_X(x_2) f_X(x_3) \left(1 - e^{-\lambda(3/2-x_2-x_3)}\right) dx_2 dx_3 \\ &= \int_0^{3/2} \lambda e^{-\lambda x_3} \int_0^{3/2-x_3} \left[\lambda e^{-\lambda x_2} - \lambda e^{-\lambda(3/2-x_3)}\right] dx_2 dx_3 \\ &= \int_0^{3/2} \lambda e^{-\lambda x_3} \left[1 - e^{-\lambda(3/2-x_3)} - \lambda\left(\frac{3}{2} - x_3\right) e^{-\lambda(3/2-x_3)}\right] dx_3 \\ &= 1 - e^{-3\lambda/2} - \frac{3\lambda}{2} e^{-3\lambda/2} - \left(\frac{3\lambda}{2}\right)^2 e^{-3\lambda/2} + \lambda^2 e^{-3\lambda/2} \frac{(3/2)^2}{2} \\ &= 1 - e^{-3\lambda/2}\left[1 + \frac{3\lambda}{2} + \frac{1}{2}\left(\frac{3\lambda}{2}\right)^2\right]. \end{aligned}$$

Problem 3. (Moment-generating function)

(a) Let $X \sim \operatorname{Binomial}(n, p)$. Find the moment-generating function of X in terms of n and p. (Hint: X is the sum of n independent Bernoulli random variables.)

Solution: X is the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$, each with moment generating function

$$M_{X_i}(t) = \mathbb{E} \exp\left(tX_i\right) = pe^t + (1-p).$$

Combining these and applying independence yields

$$M_X(t) = \mathbb{E}\exp(tX) = \mathbb{E}\exp\left(t\left(X_1 + \ldots + X_n\right)\right) = \prod_{i=1}^{n} \mathbb{E}\exp\left(tX_i\right) = \left(1 - p + pe^t\right)^n.$$

(b) Let $X_1, X_2, \ldots$ be independent and identically distributed random variables, each with a Poisson distribution with mean 1. Let $S_n = X_1 + \ldots + X_n$ for $n \geq 1$ and let $M_n(t)$ be the moment generating function of S_n.

(i) Find the smallest n such that $P\left(M_n\left(S_n\right) > 1\right) \geq 0.99$ using the exact probability.

(ii) Find the smallest n such that $P\left(M_n\left(S_n\right) > 1\right) \geq 0.99$ using the central limit theorem.

Solution: (i) First, the common moment generating function of the X_i is

$$M(t) = E\left[e^{tX_i}\right] = \sum_{k=0}^{\infty} e^{tk}\frac{1}{k!}e^{-1} = e^{-1}\sum_{k=0}^{\infty}\frac{\left(e^t\right)^k}{k!} = e^{-1}e^{e^t} = e^{e^t - 1}.$$

Then $M_n(t) = M(t)^n = e^{n\left(e^t - 1\right)}$ and $M_n\left(S_n\right) = e^{n\left(e^{S_n} - 1\right)}$. So,

$$P\left(M_n\left(S_n\right) > 1\right) = P\left(e^{n\left(e^{S_n}-1\right)} > 1\right) = P\left(n\left(e^{S_n} - 1\right) > 0\right) = P\left(e^{S_n} > 1\right) = P\left(S_n > 0\right).$$

The exact distribution of S_n is Poisson(n), so $P\left(S_n > 0\right) = 1 - P\left(S_n = 0\right) = 1 - e^{-n}$. Setting this to .99 gives $e^{-n} = .01$, or $n = \ln 100 = 4.605$. So $n = 5$ is the smallest n.

(ii) By the central limit theorem, the distribution of $\left(S_n - n\right)/\sqrt{n}$ is approximated by a $N(0,1)$ distribution, so that (from part i),

$$P\left(M_n\left(S_n\right) > 1\right) = P\left(S_n > 0\right) = P\left(\left(S_n - n\right)/\sqrt{n} > -\sqrt{n}\right) \approx 1 - \Phi(-\sqrt{n}),$$

where $\Phi(\cdot)$ is the standard normal cdf. Setting this to 0.99 gives $\Phi(-\sqrt{n}) = 0.01$, or (from the table) $-\sqrt{n} = -2.33$, or $n = 5.429$. So $n = 6$ is the smallest n according to this approximation.

(c) Let X have a Gamma $(3,3)$ distribution. Conditional on $X = x$ let Z have a normal distribution with mean x and variance 2. Finally, let $Y = e^Z$. Find $E[Y]$ and $\operatorname{Var}(Y)$.

Solution: Conditioned on $X = x, E[Y]$ is the moment generating function of a $N(x, 2)$ distribution evaluated at 1 . Letting $M_x(\cdot)$ denote this mgf, we have $M_x(t) = e^{xt+t^2}$, and so $E[Y \mid X = x] = M_x(1) = e^{x+1}$. Similarly, $E\left[Y^2 \mid X = x\right] = E\left[e^{2Z} \mid X = x\right] = M_x(2) = e^{2x+4}$. Then by the law of total expectation, $E[Y] = E\left[e^{X+1}\right] = eM_X(1)$ and $E\left[Y^2\right] = E\left[e^{2X+4}\right] = e^4 M_X(2)$, where $M_X(\cdot)$ is the moment generating function of X. Since X has a Gamma $(3,3)$ distribution we have $M_X(t) = \left(\frac{3}{3-t}\right)^3$, and so

$$E[Y] = e\left(\frac{3}{2}\right)^3 = \frac{27e}{8} \approx 9.174 \quad \text{and} \quad E\left[Y^2\right] = 27e^4.$$

Then

$$\operatorname{Var}(Y) = E\left[Y^2\right] - E[Y]^2 = 27e^4 - \left(\frac{27}{8}\right)^2 e^2 \approx 1390.$$

(d) Let X be a random variable with Exponential(λ) distribution. Recall that the moment generating function of X is $M_X(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$.

(i) Find $E[X^n]$, where n is any positive integer. You may use the mgf or compute $E[X^n]$ directly.

(ii) Find $M_Y(t)$, the mgf of $Y = \ln X$, for $t > -1$.

Solution: (i) Using the mgf, we write the mgf as

$$M_X(t) = \frac{1}{1 - t/\lambda} = \sum_{n=0}^{\infty} \left(\frac{t}{\lambda}\right)^n$$

which is valid for $|t| < \lambda$. Comparing to $M_X(t) = \sum_{n=0}^{\infty} \frac{E[X^n]}{n!} t^n$ (which is the MacLaurin series expansion of $M_X(t)$ when it is valid (it is valid for all $t < \lambda$)), we get that $E[X^n] = \frac{n!}{\lambda^n}$. Computing $E[X^n]$ directly, we have

$$E[X^n] = \int_0^{\infty} x^n \lambda e^{-\lambda x} dx = \lambda \frac{\Gamma(n+1)}{\lambda^{n+1}} \int_0^{\infty} \frac{\lambda^{n+1}}{\Gamma(n+1)} x^{(n+1)-1} e^{-\lambda x} dx = \frac{n!}{\lambda^n},$$

since the integrand in the last integral is a Gamma($n+1, \lambda$) density and so the integral is 1, and $\Gamma(n+1) = n!$.

(ii) Write the mgf of Y as $M_Y(t) = E\left[e^{tY}\right] = E\left[e^{t \ln X}\right] = E[X^t]$. Computing $E[X^t]$ directly, we get an integral similar to the one in part(a) with n replaced by t. For $t > -1$ we get a proper Gamma($t+1, \lambda$) density in the final integral and so we get

$$M_Y(t) = E\left[X^t\right] = \frac{\Gamma(t+1)}{\lambda^t}.$$

(e) Let $X_1, \ldots, X_n$ be independent Poisson(1) random variables and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Let $k > 1$ be given. Show that

$$P(\bar{X} \geq k) \leq \left(\frac{e^{k-1}}{k^k}\right)^n.$$

Solution: We find the mgf of $\bar{X}$ and then apply Chernoff's bound. The mgf of each X_i is

$$M_{X_i}(t) = E\left[e^{tX_i}\right] = \sum_{j=0}^{\infty} e^{tj} \frac{1}{j!} e^{-1} = e^{-1} e^{e^t} = e^{e^t - 1}.$$

Then the mgf of $\bar{X}$ is given by

$$M_X(t) = E\left[e^{t\bar{X}}\right] = M_{X_i}\left(\frac{t}{n}\right)^n = e^{n\left(e^{t/n} - 1\right)}.$$

Chernoff's bound gives

$$P(\bar{X} > k) \leq \min_{t>0} \frac{e^{n\left(e^{t/n}-1\right)}}{e^{tk}} = e^{n\left(e^{t/n}-1\right) - tk}. \qquad (1)$$

Differentiating the exponent with respect to t gives $e^{t/n} - k$. Setting this to 0 and solving for t gives $t = n \ln k$. Plugging this into the RHS of Eqn. (1) then gives

$$P(\bar{X} > k) \leq e^{n(k-1) - \ln k^{kn}} = \frac{e^{n(k-1)}}{k^{kn}} = \left(\frac{e^{k-1}}{k^k}\right)^n,$$

as desired.

Problem 4. (Order Statistics)

(a) Let $X_1, \ldots, X_n$ be a sequence of independent random variables uniformly distributed on the interval $(0,1)$, and let $X_{(1)}, \ldots, X_{(n)}$ denote their order statistics. For fixed k, let $g_n(x)$ denote the probability density function of $nX_{(k)}$. Find $g_n(x)$ and show that

$$\lim_{n\to\infty} g_n(x) = \begin{cases} \frac{x^{k-1}}{(k-1)!}e^{-x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

which is the Gamma$(k,1)$ density.

Solution: The pdf of $X_{(k)}$ is

$$f_k\left(x_k\right) = \begin{cases} \frac{n!}{(k-1)!(n-k)!}x_k^{k-1}\left(1-x_k\right)^{n-k} & \text{for } 0 < x_k < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let $X = nX_{(k)}$. Then the set of possible values of X (the sample space of X) is $\{x : 0 < x < n\}$. So for $x \in (0,n)$, by the (1-dimensional) change of variable formula the pdf of X is given by

$$g_n(x) = f_k\left(\frac{x}{n}\right)\frac{1}{n} = \frac{(n-1)!}{(k-1)!(n-k)!}\left(\frac{x}{n}\right)^{k-1}\left(1-\frac{x}{n}\right)^{n-k}$$

and $g_n(x) = 0$ for $x \notin (0,n)$. Now fix $x > 0$ and let $n > x$. Then

$$\begin{aligned} g_n(x) &= \frac{(n-1)!}{(k-1)!(n-k)!}\left(\frac{x}{n}\right)^{k-1}\left(1-\frac{x}{n}\right)^{n-k} \\ &= \frac{(n-1)\times\ldots\times(n-k+1)}{n^{k-1}}\left(1-\frac{x}{n}\right)^{-k}\frac{1}{(k-1)!}x^{k-1}\left(1-\frac{x}{n}\right)^{n} \\ &\to \frac{1}{(k-1)!}x^{k-1}e^{-x}, \end{aligned}$$

as desired, since

$$\begin{aligned} \frac{(n-1)\times\ldots\times(n-k+1)}{n^{k-1}} &= \left(\frac{n-1}{n}\right)\times\ldots\times\left(\frac{n-k+1}{n}\right) \\ &\to 1\times\ldots\times 1 = 1, \end{aligned}$$

$(1-x/n)^{-k}$ clearly converges to 1 as $n\to\infty$ (for fixed k), and $(1-x/n)^n \to e^{-x}$ as $n\to\infty$ (from calculus). Since $g_n(x) = 0$ for $x < 0$ it is obvious that $g_n(x)\to 0$ as $n\to\infty$ if $x < 0$.

(b) (i) Let $X_1, \ldots, X_6$ be independent random variables uniformly distributed on the interval $(0,1)$. Find the pdf of $U = \min\left\{\max\left(X_1, X_2\right), \max\left(X_3, X_4\right), \max\left(X_5, X_6\right)\right\}$.

(ii) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with a Uniform distribution on the interval $(0,1)$. Let $X = \min\left(X_1, \ldots, X_n\right)$ and $Y = \max\left(X_1, \ldots, X_n\right)$. Find $P\left(X < \frac{1}{2} < Y\right)$ and $E\left[X^3\right]$.

Solution: (i) Let $Y_1 = \max\left(X_1, X_2\right), Y_2 = \max\left(X_3, X_4\right)$, and $Y_3 = \max\left(X_5, X_6\right)$. Then Y_1, Y_2, Y_3 are independent and identically distributed random variables, each with cdf

$$F_Y(y) = P\left(X_1 \leq y, X_2 \leq y\right) = \left[P\left(X_1 \leq y\right)\right]^2 = \begin{cases} y^2 & \text{for } 0 < y < 1 \\ 0 & \text{for } y \leq 0 \\ 1 & \text{for } y \geq 1 \end{cases}$$

and pdf

$$f_Y(y) = \begin{cases} 2y & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then $U = \min\left(Y_1, Y_2, Y_3\right)$ has cdf

$$\begin{aligned} F_U(u) &= 1 - P\left(\min\left(Y_1, Y_2, Y_3\right) > u\right) \\ &= 1 - P\left(Y_1 > u, Y_2 > u, Y_3 > u\right) \\ &= 1 - \left(1 - F_Y(u)\right)^3 \\ &= \begin{cases} 1 - \left(1-u^2\right)^3 & \text{for } 0 < u < 1 \\ 0 & \text{for } u \leq 0 \\ 1 & \text{for } u \geq 1. \end{cases} \end{aligned}$$

By differentiation, then, U has pdf given by

$$f_U(u) = \begin{cases} 6u\left(1-u^2\right)^2 & \text{for } 0 < u < 1 \\ 0 & \text{otherwise.} \end{cases}$$

(ii) The joint pdf of (X, Y) is

$$f(x,y) = \begin{cases} n(n-1)(y-x)^{n-2} & \text{for } 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} P\left(X < \frac{1}{2} < Y\right) &= \int_{1/2}^{1}\int_{0}^{1/2} n(n-1)(y-x)^{n-2}dxdy \\ &= \int_{1/2}^{1} n\left[-\left.(y-x)^{n-1}\right|_0^{1/2}\right]dy \\ &= \int_{1/2}^{1} ny^{n-1}dy - \int_{1/2}^{1} n\left(y-\frac{1}{2}\right)^{n-1}dy \\ &= \left.y^n\right|_{1/2}^{1} - \left.\left(y-\frac{1}{2}\right)^n\right|_{1/2}^{1} = 1 - \frac{1}{2^n} - \frac{1}{2^n} = 1 - \frac{1}{2^{n-1}}. \end{aligned}$$

The marginal pdf of X is

$$f_X(x) = \begin{cases} n(1-x)^{n-1} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$E\left[X^3\right] = \int_0^1 x^3 n(1-x)^{n-1}dx = n\int_0^1 x^3(1-x)^{n-1}dx$$

The integral on the right is $B(4, n)$, where $B(\cdot,\cdot)$ is the beta function. So

$$E\left[X^3\right] = nB(4,n) = n\frac{\Gamma(4)\Gamma(n)}{\Gamma(n+4)} = \frac{6n}{(n+3)(n+2)(n+1)n} = \frac{6}{(n+3)(n+2)(n+1)}.$$

(c) Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be independent Uniform $(0,1)$ random variables. We form n rectangles, where the i-th rectangle has adjacent sides of length X_i and Y_i, for $i = 1, \ldots, n$. Let A_i be the area of the i th rectangle, $i = 1, \ldots, n$, and define $A_{\max} = \max\left(A_1, \ldots, A_n\right)$. Find the pdf of $A_{\max}$.

Solution: First note that $A_i = X_iY_i$ for $i = 1, \ldots, n$, and $A_1, \ldots, A_n$ are independent since the vectors $\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right)$ are independent. Also, the A_i are identically distributed, each with support $(0,1)$. We first find $P\left(A_i \leq a\right)$ for any $a \in (0,1)$, which is $P\left(X_iY_i \leq a\right)$. If $Y_i \leq a$ then the limits for X_i are from 0 to 1. If $Y_i > a$ then the limits for X_i are from 0 to a/Y_i. The joint pdf of $\left(X_i, Y_i\right)$ is $f(x,y) = 1$ for $0 < x, y < 1$. Thus we have

$$\begin{aligned} P\left(A_i \leq a\right) = P\left(X_iY_i \leq a\right) &= \int_0^a\int_0^1 dxdy + \int_a^1\int_0^{a/y} dxdy \\ &= a + a\int_a^1 \frac{1}{y}dy = a + a(\ln 1 - \ln a) = a(1-\ln a). \end{aligned}$$

Next, we get the cdf of $A_{\max}$ as

$$F(a) = P\left(A_{\max} \leq a\right) = P\left(A_1 \leq a, \ldots, A_n \leq a\right) = P\left(A_1 \leq a\right)^n = (a(1-\ln a))^n,$$

for $0 < a < 1$. Finally, we obtain the pdf f of $A_{\max}$ by differentiation:

$$f(a) = F'(a) = n(a(1-\ln a))^{n-1}[1-\ln a - 1] = -n\ln a(a(1-\ln a))^{n-1},$$

which is valid for $a \in (0,1)$, and $f(a) = 0$ for $a \leq 0$ or $a \geq 0$.

(d) Let $X_1, \ldots, X_n$ be independent and identically distributed Exponential (λ) random variables. Compute $E\left[X_{(1)} e^{-\lambda X_{(2)}}\right]$, where $X_{(1)}$ and $X_{(2)}$ are the first and second order statistics of $X_1, \ldots, X_n$.

Solution: The joint pdf of $X_{(1)}$ and $X_{(2)}$ is

$$f_{12}\left(x_1, x_2\right)=\begin{cases} n(n-1)\lambda^2 e^{-\lambda x_1} e^{-\lambda x_2}\left(e^{-\lambda x_2}\right)^{n-2} & \text{for } 0<x_1<x_2<\infty \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} E\left[X_{(1)} e^{-\lambda X_{(2)}}\right] &= n(n-1)\lambda^2 \int_0^{\infty} \int_{x_1}^{\infty} x_1 e^{-\lambda x_2} e^{-\lambda x_1} e^{-\lambda(n-1) x_2} dx_2 dx_1 \\ &= (n-1)\lambda \int_0^{\infty} x_1 e^{-\lambda x_1} e^{-\lambda n x_1} dx_1 \\ &= \frac{n-1}{n+1} \int_0^{\infty} x_1 \lambda(n+1) e^{-\lambda(n+1) x_1} dx_1 \\ &= \frac{n-1}{\lambda(n+1)^2} \end{aligned}$$

since the final integral is the mean of an Exponential$(\lambda(n+1))$ distribution.

(e) Let X_1, X_2, X_3 be independent Uniform $(0,1)$ random variables, and let $X_{(1)}, X_{(2)}, X_{(3)}$ denote their order statistics. Let X be the area of the square with side length $X_{(2)}$ and let Y be the area of the rectangle with side lengths $X_{(1)}$ and $X_{(3)}$. Find $P(X>Y)$, $E[X]$, and $E[Y]$.

Solution: We wish to compute $P\left(X_{(2)}^2 > X_{(1)} X_{(3)}\right)$. The joint pdf of $\left(X_{(1)}, X_{(2)}, X_{(3)}\right)$ is $f_{123}\left(x_1, x_2, x_3\right)=6$ for $0<x_1<x_2<x_3<1$ and equals 0 otherwise. The solution is computed by integrating $f_{123}\left(x_1, x_2, x_3\right)$ over the region $A=\left\{\left(x_1, x_2, x_2\right): x_2^2-x_1 x_3>0\right\}$. If we set the inner integral over x_2 and note that for given values of $0<x_1<x_3<1$, the possible values of x_2 are $\sqrt{x_1 x_3}<x_2<x_3$. The integral can be written as

$$\begin{aligned} \iiint_A f_{123}\left(x_1, x_2, x_3\right) dx_1 dx_2 dx_3 &= 6 \int_0^1 \int_0^{x_3} \int_{\sqrt{x_1 x_3}}^{x_3} dx_2 dx_1 dx_3 \\ &= 6 \int_0^1 \int_0^{x_3}\left(x_3-x_1^{1/2} x_3^{1/2}\right) dx_1 dx_3 \\ &= 6 \int_0^1\left(x_3^2-x_3^{1/2} \frac{2}{3} x_3^{3/2}\right) dx_3 \\ &= 2 \int_0^1 x_3^2 dx_3=\frac{2}{3}. \end{aligned}$$

The marginal pdf of $X_{(2)}$ and the joint pdf of $\left(X_{(1)}, X_{(3)}\right)$ are given by

$$f_2\left(x_2\right)=\begin{cases} 6 x_2\left(1-x_2\right) & 0<x_2<1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_{13}\left(x_1, x_3\right)=\begin{cases} 6\left(x_3-x_1\right) & 0<x_1<x_3<1 \\ 0 & \text{otherwise} \end{cases}$$

respectively. Then

$$\begin{aligned} E[X] &= E\left[X_{(2)}^2\right]=\int_0^1 x_2^2 6 x_2\left(1-x_2\right) dx_2=\frac{6}{4}-\frac{6}{5}=\frac{3}{10} \\ E[Y] &= E\left[X_{(1)} X_{(3)}\right]=\int_0^1 \int_0^{x_3} x_1 x_3 6\left(x_3-x_1\right) dx_1 dx_3=6 \int_0^1\left(\frac{x_3^4}{2}-\frac{x_3^4}{3}\right) dx_3=\frac{1}{5}. \end{aligned}$$

Problem 5. (Convergence of Random Variables and Limit Theorems)

(a) Let X and $X_1, X_2, \ldots$ be random variables each having a $N(0,1)$ distribution. Suppose (X_n, X) has a bivariate normal distribution for each n and the correlation between X_n and X is $\rho(X_n, X) = \rho_n$, for $n \geq 1$.

(i) Show that X_n converges to X in distribution as $n \to \infty$ (for arbitrary correlations ρ_n).

(ii) If $\rho_n \to 1$ as $n \to \infty$, show that X_n converges to X in probability as $n \to \infty$.

(iii) Show that if $\rho_n = 1 - a^n$ for some constant $a \in (0,1)$, then X_n converges to X with probability 1 as $n \to \infty$. Do you get convergence with probability 1 if $a = 0$? If $a = 1$? Prove your answers.

Solution: (i) If Φ is the standard normal cdf and F_n is the cdf of X_n for $n \geq 1$, then $F_n(x) = \Phi(x)$ for all x (this is given). Therefore, $F_n(x) \to \Phi(x)$ as $n \to \infty$ (trivially). It is also given that $\Phi(x)$ is the cdf of X. Therefore, X_n converges to X in distribution.

(ii) Let $\epsilon > 0$ be given. Both X_n and X are zero mean, so $X_n - X$ also has zero mean. So by Chebyshev'e inequality,

$$P(|X_n - X| > \epsilon) \leq \frac{\operatorname{Var}(X_n - X)}{\epsilon^2}. \tag{2}$$

But

$$\operatorname{Var}(X_n - X) = \operatorname{Var}(X_n) + \operatorname{Var}(X) - 2\operatorname{Cov}(X_n, X) = 1 + 1 - 2\rho_n = 2(1 - \rho_n). \tag{3}$$

So if $\rho_n \to 1$ as $n \to \infty$ then $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$. Thus, X_n converges to X in probability.

(iii) Plugging $\rho_n = 1 - a^n$ into (3) we have $\operatorname{Var}(X_n - X) = 2a^n$. Plugging this into Chebyshev's inequality in (2), we have

$$P(|X_n - X| > \epsilon) \leq \frac{2a^n}{\epsilon^2}.$$

If $a \in (0,1)$, the sum on the right converges, and so X_n converges to X with probability 1 by the sufficient condition from class. If $a = 0$, the sum is again convergent (it is equal to 0), and so X_n converges to X with probability 1. If $a = 1$ then $\rho_n = 0$ so that X_n is independent of X for any n (since (X_n, X) is bivariate normal). In this case $X_n - X$ has a $N(0,2)$ distribution and so $P(|X_n - X| > \epsilon) = 2(1 - \Phi(\epsilon/\sqrt{2}))$, where Φ is the standard normal cdf. This is some positive constant for every n and so $P(|X_n - X| > \epsilon)$ does not converge to 0 as $n \to \infty$. Therefore, X_n does not converge to X in probability. This then implies that X_n does not converge to X with probability 1.

(b) Suppose 80 points are uniformly distributed in the ball in $\mathbb{R}^3$ centered at the origin with radius 1. Let D_i be the Euclidean distance from the origin of the i th point, $i = 1, \ldots, 80$, and let $\bar{D} = \frac{1}{80}\sum_{i=1}^{80} D_i$. Use the central limit theorem to find a value c satisfying $P(|\bar{D} - E[\bar{D}]| \leq c) = 0.95$.

Solution: We first find the distribution of D_i. For $x \in (0,1), D_i \leq x$ if and only if the i-th point in the ball is in the ball in $\mathbb{R}^3$ centred at the origin of radius x. That is, $P(D_i \leq x) = \frac{4\pi x^3/3}{4\pi/3} = x^3$. From this we have that the pdf of D_i is

$$f_D(x) = \begin{cases} 3x^2 & \text{for } x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

The mean, second moment, and variance are then easily computed to be

$$E[D_i] = \frac{3}{4}, \quad E\left[D_i^2\right] = \frac{3}{5}, \quad \operatorname{Var}(D_i) = \frac{3}{80}.$$

By the central limit theorem

$$\begin{aligned} P(|\bar{D} - 3/4| \leq c) &= P\left(\frac{\sqrt{80}|\bar{D} - 3/4|}{\sqrt{3/80}} \leq \frac{c\sqrt{80}}{\sqrt{3/80}}\right) \\ &\approx P\left(|Z| \leq \frac{c\sqrt{80}}{\sqrt{3/80}}\right), \end{aligned}$$

where Z has a $N(0,1)$ distribution. In order for this probability to be equal to 0.95 we need $\frac{c\sqrt{80}}{\sqrt{3/80}}$ equal to 1.96. Solving for c gives

$$c = \frac{1.96\sqrt{3}}{80} \approx 0.0424.$$

(c) (i) Suppose that $\{X_n\}$ is a sequence of zero-mean random variables and X is a zero mean random variable, and suppose that $E\left[(X_n - X)^2\right] \leq C/n^p$ for every n, for some constants C and $p > 1$. Show that $X_n \to X$ almost surely.

(ii) Suppose that $\{X_n\}$ is a sequence of non-negative random variables. Show that $E[X_n] \to 0$ as $n \to \infty$ implies that $X_n \to 0$ in probability, but that the converse is false in general.

Solution: (i) Let $\epsilon > 0$ be given. Since $E[X_n - X] = 0$ we have by Chebyshev's inequality that

$$P(|X_n - X| > \epsilon) \leq \frac{\mathrm{Var}(X_n - X)}{\epsilon^2} = \frac{E\left[(X_n - X)^2\right]}{\epsilon^2} \leq \frac{C}{\epsilon^2 n^p}.$$

Therefore,

$$\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) \leq \sum_{n=1}^{\infty} \frac{C}{\epsilon^2 n^p} < \infty$$

if $p > 1$. By a sufficient condition from class, this implies that $X_n \to X$ almost surely.

(ii) Suppose that $E[X_n] \to 0$. Let $\epsilon > 0$ be given. Then by Markov's inequality

$$P(|X_n - 0| > \epsilon) \leq \frac{E[X_n]}{\epsilon} \to 0$$

as $n \to \infty$. Hence, $X_n \to 0$ in probability. To prove that the reverse implication is false in general, we give a counterexample. Let the distribution of X_n be given by

$$X_n = \begin{cases} 0 & \text{with probability } 1 - \frac{1}{n} \\ n & \text{with probability } \frac{1}{n}. \end{cases}$$

Then if $\epsilon > 0$ is given $P(|X_n - 0| > \epsilon) = P(X_n = n) = \frac{1}{n} \to 0$ as $n \to \infty$, so $X_n \to 0$ in probability. However, $E[X_n] = n\left(\frac{1}{n}\right) = 1$ for all n, which does not converge to 0.

(d) Suppose that $\{X_n\}$ and $\{Y_n\}$ are sequences of random variables and X and Y are random variables such that $X_n \to X$ in distribution and $Y_n \to Y$ in distribution. Give an example where it is not true that $X_n + Y_n$ converges to $X + Y$ in distribution. [Hint: Consider $Y = -X$]

Solution: Let W and Z be independent $N(0,1)$ random variables. Let $X_n = W$ for all n and $Y_n = Z$ for all n. Let $X = W$ and $Y = -W$. It is easy to see that X and Y are both $N(0,1)$, and it is clear that $X_n \to X$ in distribution and $Y_n \to Y$ in distribution. But $X + Y = 0$ whereas $X_n + Y_n$ is distributed as $N(0,2)$ for all n. Therefore, it is not true that $X_n + Y_n$ converges to $X + Y$ in distribution.

(e) Consider the probability

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} F(X_i) - \frac{1}{2}\right| \geq \frac{1}{\sqrt{3n}}\right).$$

Use Chebyshev's inequality to bound this probability. Use the central limit theorem to approximate this probability for large n.

Solution: The random variable $\frac{1}{n}\sum_{i=1}^{n} F(X_i)$ has mean $\frac{1}{2}$ and variance $\frac{1}{12n}$. Applying Chebyshev's inequality, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} F(X_i) - \frac{1}{2}\right| \geq \frac{1}{\sqrt{3n}}\right) \leq \frac{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} F(X_i)\right)}{(1/\sqrt{3n})^2} = \frac{1/12n}{1/3n} = \frac{1}{4}.$$

By the central limit theorem, $Z_n = \frac{\frac{1}{n}\sum_{i=1}^n F(X_i) - 1/2}{\sqrt{1/12n}}$ has approximately a $N(0,1)$ distribution for large n. Therefore,

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^n F(X_i) - \frac{1}{2}\right| \geq \frac{1}{\sqrt{3n}}\right) &= P\left(\left|\frac{1}{n}\sum_{i=1}^n F(X_i) - \frac{1}{2}\right|\sqrt{12n} \geq \frac{\sqrt{12n}}{\sqrt{3n}}\right) \\ &= P\left(|Z_n| \geq 2\right) \approx 0.0456. \end{aligned}$$

Problem 6. (Sampling Distributions)

(a) Let $T \sim t_1$ (the t distribution with 1 degree of freedom). Explain why T has the same distribution as $\frac{X}{|Y|}$ where $X, Y \overset{IID}{\sim} \mathcal{N}(0,1)$, and hence why T also has the same distribution as $\frac{X}{Y}$.

[Hints: The distribution of $\frac{X}{Y}$ when $X, Y \overset{IID}{\sim} \mathcal{N}(0,1)$ is also called the Cauchy distribution (the t distribution with 1 degree of freedom). You may check that it has a PDF:

$$f(x) = \frac{1}{\pi}.\frac{1}{1+x^2}.$$

You may use this result without proof in part (b).]

Solution: By definition of the t distribution, T has the same distribution as $\frac{X}{\sqrt{U}}$, where $X \sim \mathcal{N}(0,1), U \sim \chi_1^2$, and X and U are independent. By definition of the χ^2 distribution, U has the same distribution as Y^2 where $Y \sim \mathcal{N}(0,1)$. Therefore T has the same distribution as $\frac{X}{\sqrt{Y^2}} = \frac{X}{|Y|}$. To show $\frac{X}{|Y|}$ has the same distribution as $\frac{X}{Y}$: Suppose, for any $t \in \mathbb{R}$,

$$\mathbb{P}\left[\frac{X}{|Y|} \leq t\right] = \mathbb{P}[X \leq t|Y|] = \mathbb{P}[X \leq tY,\ Y > 0] + \mathbb{P}[X \leq -tY,\ Y < 0],$$

and

$$\mathbb{P}\left[\frac{X}{Y} \leq t\right] = \mathbb{P}[X \leq tY, Y > 0] + \mathbb{P}[X \geq tY,\ Y < 0] = \mathbb{P}[X \leq tY,\ Y > 0] + \mathbb{P}[-X \leq -tY,\ Y < 0].$$

Since (X, Y) has the same distribution as $(-X, Y)$, the above implies

$$\mathbb{P}\left[\frac{X}{|Y|} \leq t\right] = \mathbb{P}\left[\frac{X}{Y} \leq t\right].$$

So the CDFs of $\frac{X}{|Y|}$ and $\frac{X}{Y}$ are the same.

(b) t_1 is an example of an extremely "heavy-tailed" distribution: For $T \sim t_1$, show that $\mathbb{E}[|T|] = \infty$ and $\mathbb{E}\left[T^2\right] = \infty$. If $T_1, \ldots, T_n \overset{IID}{\sim} t_1$, explain why the Law of Large Numbers and the Central Limit Theorem do not apply to the sample mean $\frac{1}{n}(T_1 + \ldots + T_n)$.

Solution: Noting $f(x) = f(-x)$, we compute

$$\begin{aligned} \mathbb{E}[|T|] &= \int_{-\infty}^{\infty} |x| f(x) \mathrm{d}x = 2\int_0^\infty x \frac{1}{\pi}\frac{1}{x^2+1}\,\mathrm{d}x \\ &= \frac{1}{\pi}\int_0^\infty \frac{1}{x^2+1}\,\mathrm{d}\left(x^2+1\right) = \frac{1}{\pi}\ln\left(x^2+1\right)\Big|_0^\infty = \infty. \end{aligned}$$

Also,

$$\mathbb{E}\left[T^2\right] = \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x = 2\int_0^\infty x^2 \frac{1}{\pi}\frac{1}{x^2+1}\,\mathrm{d}x = \frac{2}{\pi}\int_0^\infty \frac{x^2}{x^2+1}\,\mathrm{d}x,$$

which equals ∞ since $\frac{x^2}{x^2+1} \to 1$ as $x \to \infty$. So, T does not have a well-defined (finite) mean or variance, and the LLN and CLT both do not apply. (In fact, it may be shown that $\frac{1}{n}(T_1 + \ldots + T_n)$ does not converge to 0 but rather $\frac{1}{n}(T_1 + \ldots + T_n) \sim t_1$ for any n.)

(c) Let $U_n \sim \chi_n^2$. Show that $\frac{1}{\sqrt{\frac{1}{n}U_n}} \to 1$ in probability as $n \to \infty$. (Hint: Apply the Law of Large Numbers and the Continuous Mapping Theorem.)

[**Continuous Mapping Theorem:** If random variables $\{X_n\}_{n=1}^{\infty}$ converge in probability to $c \in \mathbb{R}$ (as $n \to \infty$), and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then $\{g(X_n)\}_{n=1}^{\infty}$ converge in probability to $g(c)$.]

Solution: We may write $U_n = \sum_{i=1}^{n} X_i^2$, where $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0,1)$. Then the LLN implies $\frac{1}{n}U_n \to \mathbb{E}\left[X_i^2\right] = 1$ in probability as $n \to \infty$. The function $x \mapsto 1/\sqrt{x}$ is continuous (at every $x > 0$), so by the Continuous Mapping Theorem

$$\frac{1}{\sqrt{\frac{1}{n}U_n}} \to 1$$

in probability as $n \to \infty$.

(d) Using Slutsky's lemma, show that if $T_n \sim t_n$ for each $n = 1, 2, 3, \ldots$, then as $n \to \infty$, $T_n \to Z$ in distribution where $Z \sim \mathcal{N}(0,1)$. (This formalizes the statement that "the t_n distribution approaches to the standard normal distribution as n gets large".)

[**Slutsky's lemma:** If sequences of random variables $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$ satisfy $X_n \to c$ in probability for a constant $c \in \mathbb{R}$ and $Y_n \to Y$ in distribution for a random variable Y, then $X_n Y_n \to cY$ in distribution.]

Solution: We may write $T_n = \frac{Z_n}{\sqrt{\frac{1}{n}U_n}}$ where $Z_n \sim \mathcal{N}(0,1)$, $U_n \sim \chi_n^2$, and Z_n and U_n are independent. By part (c), $\frac{1}{\sqrt{\frac{1}{n}U_n}} \to 1$ in probability. Clearly, $Z_n \to \mathcal{N}(0,1)$ in distribution, since the distribution of Z_n does not change with n. Then $T_n \to \mathcal{N}(0,1)$ in distribution by Slutsky's lemma.

(e) If ξ_p is the p-th quantile of $F_{m,n}$ distribution (F distribution with (m,n) degree of freedoms) and ξ_p' the p-th quantile of $F_{n,m}$. Show that, $\xi_p \xi_{1-p}' = 1$.

Solution:

$$F = \frac{\chi_m^2/m}{\chi_n^2/n} \Rightarrow \frac{1}{F} = \frac{\chi_n^2/n}{\chi_m^2/m} \sim F_{n,m}.$$

$$p = P\left[F_{m,n} \leq \xi_p\right] = P\left[\frac{1}{F_{m,n}} \geqslant \frac{1}{\epsilon_p}\right] = P\left[F_{n,m} \geqslant \frac{1}{\xi_p}\right]$$

$$\Rightarrow P\left[F_{n,m} \leq \frac{1}{\xi_p}\right] = 1 - p.$$

But, $p\left[F_{n,m} \leq \xi_{1-p}'\right] = 1 - p \Rightarrow \frac{1}{\xi_p} = \xi_{1-p}'$.

Problem 7. (Methods of Estimation)

(a) Suppose $X_1, \ldots, X_n \overset{\text{IID}}{\sim}$ Geometric(p), where Geometric(p) is the geometric distribution on the positive integers $\{1, 2, 3, \ldots\}$ defined by the probability mass function (PMF)

$$f(x \mid p) = p(1-p)^{x-1},$$

with a single parameter $p \in [0,1]$. Compute the method-of-moments estimate of p, as well as the MLE of p. For large n, what approximately is the sampling distribution of the MLE? (You may use, without proof, the fact that the Geometric(p) distribution has mean $1/p$.)

Solution. The method of moments estimator sets the population mean, $1/p$, equal to the sample mean,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Inverting to solve for p gives

$$\hat{p}_{\text{MOM}} = \frac{1}{\bar{X}}.$$

For the maximum likelihood estimator, the likelihood is

$$L\left(p \mid X_1, \ldots, X_n\right) = \prod_{i=1}^{n} \left(p(1-p)^{X_i - 1}\right) = p^n (1-p)^{n(\bar{X}-1)}$$

and the log-likelihood is therefore

$$\ell(p) = n \log p + n(\bar{X} - 1)\log(1-p).$$

The derivative is

$$\ell'(p) = \frac{n}{p} - \frac{n(\bar{X}-1)}{1-p}.$$

Setting equal to zero and solving for p gives

$$\hat{p}_{\text{MLE}} = \frac{1}{\bar{X}}.$$

We must also check that $\ell(p)$ achieves a maximum at $\bar{X}^{-1}$; this may be verified by checking that $\ell'(p)$ takes positive for $p < \bar{X}^{-1}$ and negative values for $p > \bar{X}^{-1}$.

Now, we can get the asymptotic distribution using the delta method. We have from the central limit theorem that

$$\sqrt{n}(\bar{X} - 1/p) \sim \mathcal{N}\left(0, \frac{1-p}{p^2}\right).$$

Taking $g(\theta) = 1/\theta$ gives $\left(g'(\theta)\right)^2 = \theta^{-4}$, which for $\theta = 1/p$ is $\left(g'(\theta)\right)^2 = p^4$. Hence

$$\sqrt{n}\left(\hat{p}_{\text{MLE}} - p\right) = \sqrt{n}(1/\bar{X} - p) = \sqrt{n}(g(X) - g(1/p)) \Rightarrow \mathcal{N}\left(0, p^2(1-p)\right).$$

Alternatively, we could obtain the variance using the Fisher information:

$$\sqrt{n}\left(\hat{p}_{\text{MLE}} - p\right) \Rightarrow \mathcal{N}\left(0, \frac{1}{I(p)}\right),$$

where $I(p)$ is the Fisher information for a single observation. We compute

$$\begin{aligned} I(p) = -\mathbf{E}_p\left[\ell''(p)\right] &= -\mathbf{E}_p\left[\frac{\partial^2}{\partial^2 p}(\log p + (X-1)\log(1-p))\right] \\ &= -\mathbf{E}_p\left[\frac{\partial}{\partial p}\left(\frac{1}{p} - \frac{X-1}{1-p}\right)\right] = -\mathbf{E}_p\left[-\frac{1}{p^2} - \frac{X-1}{(1-p)^2}\right] \\ &= \frac{1}{p^2(1-p)} \end{aligned}$$

$$\text{So,} \quad \sqrt{n}\left(\hat{p}_{\text{MLE}} - p\right) \sim \mathcal{N}\left(0, p^2(1-p)\right).$$

(b) Let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$. We showed in class that the MLEs for μ and σ^2 are given by

$$\hat{\mu} = \bar{X} \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

(i) By computing the Fisher information matrix $I\left(\mu, \sigma^2\right)$, derive the approximate joint distribution of $\hat{\mu}$ and $\hat{\sigma}^2$ for large n. (Hint: Substitute $v = \sigma^2$ and treat v as the parameter rather than σ.)

(ii) Suppose it is known that $\mu = 0$. Compute the MLE $\tilde{\sigma}^2$ in the one-parameter sub-model $\mathcal{N}\left(0, \sigma^2\right)$. The Fisher information matrix in part (i) has off-diagonal entries equal to 0 when $\mu = 0$ and n is large. What does this tell you about the standard error of $\tilde{\sigma}^2$ as compared to that of $\hat{\sigma}^2$?

Solution. (i) Denote $v = \sigma^2$. Then

$$f(X \mid \mu, v) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{1}{2v}(X-\mu)^2}$$

and

$$\ell(\mu, v) = \log f(X \mid \mu, v) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log v - \frac{1}{2v}(X-\mu)^2.$$

To obtain the Fisher information matrix $I(\mu, v)$, we must compute the four second-order partial derivatives of $\ell(\mu, v)$. These quantities are

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{1}{v},$$
$$\frac{\partial^2 \ell}{\partial v^2} = \frac{1}{2v^2} - \frac{1}{v^3}(X-\mu)^2,$$
$$\frac{\partial^2 \ell}{\partial \mu \partial v} = \frac{\partial^2 \ell}{\partial v \partial \mu} = -\frac{X-\mu}{v^2}.$$

Then

$$I(\mu, v) = -\mathbf{E}_{\mu,v}\left[\begin{array}{cc} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial v} \\ \frac{\partial^2 \ell}{\partial v \partial \mu} & \frac{\partial^2 \ell}{\partial v^2} \end{array}\right] = \left[\begin{array}{cc} 1/v & 0 \\ 0 & 1/2v^2 \end{array}\right].$$

This matrix has an inverse

$$I(\mu, v)^{-1} = \left[\begin{array}{cc} v & 0 \\ 0 & 2v^2 \end{array}\right].$$

Substituting back $v = \sigma^2$, we have

$$I\left(\mu, \sigma^2\right)^{-1} = \left[\begin{array}{cc} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{array}\right],$$

which we conclude is the asymptotic variance of the maximum likelihood estimate. In other words,

$$\sqrt{n}\left(\left[\begin{array}{c} \bar{X} \\ S^2 \end{array}\right] - \left[\begin{array}{c} \mu \\ \sigma^2 \end{array}\right]\right) \sim \mathcal{N}\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{array}\right]\right).$$

(ii) The joint log-likelihood in this one-parameter sub-model is given by

$$\ell(v) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log v - \frac{1}{2v}\sum_{i=1}^{n} X_i^2,$$

where again $v = \sigma^2$. Then

$$\ell'(v) = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n} X_i^2,$$

and setting equal to zero and solving for v gives

$$\tilde{v} = \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Since the off-diagonals of the inverse Fisher information matrix are zero, the sample mean and standard deviation are asymptotically uncorrelated, and so $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ have the same asymptotic standard error.

(c) Let $X_1, \ldots, X_n \overset{IID}{\sim}$ Uniform$(0, \theta)$ for a single parameter $\theta > 0$ having PDF

$$f(x \mid \theta) = \frac{1}{\theta}\mathbb{1}\{0 \leq X \leq \theta\}.$$

(i) Compute the MLE $\hat{\theta}$ of θ. (Hint: Note that the PDFs $f(x \mid \theta)$ do not have the same support for all $\theta > 0$, and they are also not differentiable with respect to θ you will need to reason directly from the definition of MLE.)

(ii) If the true parameter is θ, explain why $\hat{\theta} \leq \theta$ always, and hence why it cannot be true that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to $\mathcal{N}(0, v)$ for any $v > 0$.

Solution. (i) The likelihood is

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}\left\{0 \leq X_i \leq \theta\right\}.$$

Now, notice that the expression

$$\prod_{i=1}^{n} \mathbb{1}\left\{0 \leq X_i \leq \theta\right\},$$

taken as a function of θ, is the same as

$$\mathbb{1}\left\{\theta > X_i \text{ for all } i\right\} = \mathbb{1}\left\{\theta \geq \max_i X_i\right\}.$$

This means that the likelihood can be written as

$$L(\theta) = \frac{1}{\theta^n} \mathbb{1}\left\{\theta \geq \max_i X_i\right\},$$

and that the maximum likelihood estimate is the value of θ that maximizes $1/\theta^n$ on the interval $[\max_i X_i, \infty)$. Since $1/\theta^n$ is a decreasing function, this maximum occurs at the left end point, so

$$\hat{\theta} = \max_i X_i.$$

(ii) The true parameter θ must satisfy $\theta \geq X_i$ for all $i = 1, \ldots, n$, since the range of X_i is bounded above by θ. Hence $\theta \geq \max_i X_i = \hat{\theta}$ as well. This means that for any value of n, $\sqrt{n}(\hat{\theta} - \theta)$ takes on positive values with probability zero, so $\sqrt{n}(\hat{\theta} - \theta)$ cannot be asymptotically normally distributed.

(d) Consider a parametric model $\{f(x \mid \theta) : \theta \in \mathbb{R}\}$ of the form

$$f(x \mid \theta) = e^{\theta T(x) - A(\theta)} h(x),$$

where T, A, and h are known functions.

(i) Show that the Poisson(λ) model is of this form, upon reparametrizing by $\theta = \log \lambda$. What are the functions $T(x), A(\theta)$, and $h(x)$?

(ii) For any model of the above form, differentiate the identity

$$1 = \int e^{\theta T(x) - A(\theta)} h(x) dx$$

with respect to θ on both sides, to obtain a formula for $\mathbb{E}_\theta[T(X)]$, where $\mathbb{E}_\theta$ denotes expectation when $X \sim f(x \mid \theta)$. Verify that this formula is correct for the Poisson example in part (i).

(iii) The following procedure defines the generalized method-of-moments estimator: For a fixed function $g(x)$, compute $\mathbb{E}_\theta[g(X)]$ in terms of θ, and take the estimate $\hat{\theta}$ to be the value of θ for which

$$\mathbb{E}_\theta[g(X)] = \frac{1}{n} \sum_{i=1}^{n} g\left(X_i\right).$$

The method-of-methods estimator discussed in class is the special case of this procedure for $g(x) = x$. Let $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$, where $f(x \mid \theta)$ is of the given form, and consider the generalized method-of-moments estimator using the function $g(x) = T(x)$. Show that this estimator is the same as the MLE. (You may assume that the MLE is the unique solution to the equation $0 = l'(\theta)$, where $l(\theta)$ is the log-likelihood.)

Solution. (i) A Poisson random variable has a mass function

$$f(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} e^{x \log \lambda - \lambda},$$

for $x = 0, 1, 2, \ldots$. Reparametrizing by $\theta = \log \lambda$, we obtain

$$f(x \mid \theta) = \frac{1}{x!} e^{\theta x - e^{\theta}},$$

which is of the form in question. The functions are given by

$$T(x) = x, \quad A(\theta) = e^{\theta}, \quad \text{and} \quad h(x) = \frac{1}{x!}.$$

(ii) The derivative of the right hand side is

$$\frac{d}{d\theta} \int e^{\theta T(x) - A(\theta)} h(x) dx = \int \frac{d}{d\theta} e^{\theta T(x) - A(\theta)} h(x) dx \quad = \int \left(T(x) - A'(\theta)\right) e^{\theta T(x) - A(\theta)} h(x) dx.$$

Since the derivative of the left-hand side is 0, we have

$$0 = \int \left(T(x) - A'(\theta)\right) e^{\theta T(x) - A(\theta)} h(x) dx$$

which implies

$$\underbrace{\int T(x) e^{\theta T(x) - A(\theta)} h(x) dx}_{\mathrm{E}_\theta[T(X)]} = A'(\theta) \underbrace{\int e^{\theta T(x) - A(\theta)} h(x) dx}_{1}.$$

Using the identities noted above, we obtain the formula

$$\mathbf{E}_\theta[T(X)] = A'(\theta).$$

(Note: By replacing integrals with sums, the identity holds for discrete models as well.)
In the Poisson model, $A(\theta) = e^{\theta}$, so $A'(\theta) = e^{\theta}$ as well, and $T(X) = X$. This means

$$\mathbf{E}_\theta[X] = e^{\theta} = \lambda,$$

which we know to be true.

(iii) From part (ii), $\mathbf{E}_\theta[T(X)] = A'(\theta)$, so the generalized method-of-moments estimator is the value of θ satisfying

$$A'(\theta) = \frac{1}{n} \sum_{i=1}^{n} T\left(X_i\right).$$

We now compute the maximum likelihood estimator. The log-likelihood is

$$\theta \sum_{i=1}^{n} T\left(X_i\right) - nA(\theta) + \sum_{i=1}^{n} \log h\left(X_i\right),$$

which has derivative

$$\sum_{i=1}^{n} T\left(X_i\right) - nA'(\theta).$$

Setting equal to zero, we see that the MLE must satisfy

$$A'(\theta) = \frac{1}{n} \sum_{i=1}^{n} T\left(X_i\right),$$

which is the same as the GMM estimator for $g(x) = T(x)$.

(e) Suppose that X is a discrete random variable with

$$\begin{aligned}\mathbb{P}[X=0] &= \frac{2}{3}\theta \\ \mathbb{P}[X=1] &= \frac{1}{3}\theta \\ \mathbb{P}[X=2] &= \frac{2}{3}(1-\theta) \\ \mathbb{P}[X=3] &= \frac{1}{3}(1-\theta)\end{aligned}$$

where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from such a distribution: $\{3,0,2,1,3,2,1,0,2,1\}$. (For parts (i) and (ii), feel free to use any asymptotic approximations you wish, even though $n=10$ here is rather small.)

(i) Find the method of moments estimate of θ and compute an approximate standard error of your estimate using asymptotic theory.

(ii) Find the maximum likelihood estimate of θ and compute an approximate standard error of your estimate using asymptotic theory. (Hint: Your formula for the log-likelihood based on n observations $X_1,\ldots,X_n$ should depend on the numbers of 0's, 1's, 2's, and 3's in this sample.)

Solution. (i) The expectation of X is

$$\mathbb{E}[X] = \frac{2}{3}\theta \cdot 0 + \frac{1}{3}\theta \cdot 1 + \frac{2}{3}(1-\theta)\cdot 2 + \frac{1}{3}(1-\theta)\cdot 3 = \frac{7}{3} - 2\theta.$$

For an IID sample $X_1,\ldots,X_n$, equating $\frac{7}{3}-2\theta$ with the sample mean $\bar{X}$ and solving for θ, the method-of-moments estimate is $\hat{\theta} = \frac{1}{2}\left(\frac{7}{3}-\bar{X}\right)$. For the 10 given observations, $\hat{\theta}=0.417$.
The variance of X is

$$\begin{aligned}\operatorname{Var}[X] &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ &= \frac{2}{3}\theta\cdot 0^2 + \frac{1}{3}\theta\cdot 1^2 + \frac{2}{3}(1-\theta)\cdot 2^2 + \frac{1}{3}(1-\theta)\cdot 3^2 - \left(\frac{7}{3}-2\theta\right)^2 \quad = \frac{2}{9} + 4\theta - 4\theta^2\end{aligned}$$

Then $\operatorname{Var}[\hat{\theta}] = \frac{1}{4}\operatorname{Var}[\bar{X}] = \frac{1}{4n}\operatorname{Var}[X] = \frac{1}{4n}\left(\frac{2}{9}+4\theta-4\theta^2\right)$.
An estimate of the standard error is $\sqrt{\frac{1}{4n}\left(\frac{2}{9}+4\hat{\theta}-4\hat{\theta}^2\right)}$, which for the 10 given observations is 0.173.
(An alternative estimate of the standard error is given by $\frac{1}{4n}$ times the sample variance of $X_1,\ldots,X_n$, which for the 10 given observations is 0.171.)

(ii) For an IID sample $X_1,\ldots,X_n$, let N_0,N_1,N_2,N_3 be the total numbers of observations equal to $0,1,2,$ and 3 . Then the log-likelihood is

$$\begin{aligned}l(\theta) &= \log\left(\prod_{i=1}^{n}\left(\frac{2}{3}\theta\right)^{\mathbb{1}\{X_i=0\}}\left(\frac{1}{3}\theta\right)^{\mathbb{1}\{X_i=1\}}\left(\frac{2}{3}(1-\theta)\right)^{\mathbb{1}\{X_i=2\}}\left(\frac{1}{3}(1-\theta)\right)^{\mathbb{1}\{X_i=3\}}\right) \\ &= N_0\log\frac{2}{3}\theta + N_1\log\frac{1}{3}\theta + N_2\log\frac{2}{3}(1-\theta) + N_3\log\frac{1}{3}(1-\theta).\end{aligned}$$

To compute the MLE for θ, we set

$$0 = l'(\theta) = \frac{N_0}{\theta} + \frac{N_1}{\theta} - \frac{N_2}{1-\theta} - \frac{N_3}{1-\theta}$$

and solve for θ, yielding $\hat{\theta} = (N_0+N_1)/(N_0+N_1+N_2+N_3) = (N_0+N_1)/n$. For the 10 given observations, $\hat{\theta}=0.5$. The total probability that $X=0$ or $X=1$ is θ, so $N_0+N_1 \sim \text{Binomial}(n,\theta)$. Then $\operatorname{Var}[\hat{\theta}] = \frac{1}{n^2}\operatorname{Var}[N_0+N_1] = \frac{\theta(1-\theta)}{n}$.

(Alternatively, we may compute

$$\frac{\partial^2}{\partial \theta^2} \log f(x \mid \theta) = \begin{cases} -\frac{1}{\theta^2} & x = 0 \text{ or } x = 1 \\ -\frac{1}{(1-\theta)^2} & x = 2 \text{ or } x = 3, \end{cases}$$

so the Fisher information is $I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X \mid \theta)\right] = \frac{1}{\theta(1-\theta)}$. This shows that the variance of $\hat{\theta}$ is approximately $\frac{\theta(1-\theta)}{n}$ for large n.)

Thus, an estimate of the standard error is $\sqrt{\hat{\theta}(1-\hat{\theta})/n}$, which for the 10 given observations is 0.158.

(f) Let $(X_1, \ldots, X_k) \sim \text{Multinomial}\,(n, (p_1, \ldots, p_k))$. (This is not quite the setting of n IID observations from a parametric model, although you can think of $(X_1, \ldots, X_k)$ as a summary of n such observations $Y_1, \ldots, Y_n$ from the parametric model Multinomial $(1, (p_1, \ldots, p_k))$, where Y_i indicates which of k possible outcomes occurred for the i^{th} observation.) Find the MLE of p_i.

Solution. The log-likelihood is given by

$$l\,(p_1, \ldots, p_k) = \log\left(\binom{n}{X_1, \ldots, X_k} p_1^{X_1} \ldots p_k^{X_k}\right) = \log\binom{n}{X_1, \ldots, X_k} + \sum_{i=1}^{k} X_i \log p_i,$$

and the parameter space is

$$\Omega = \{(p_1, \ldots, p_k) : 0 \leq p_i \leq 1 \text{ for all } i \text{ and } p_1 + \ldots + p_k = 1\}.$$

To maximize $l\,(p_1, \ldots, p_k)$ subject to the linear constraint $p_1 + \ldots + p_k = 1$, we may use the method of Lagrange multipliers: Consider the Lagrangian

$$L\,(p_1, \ldots, p_k, \lambda) = \log\binom{n}{X_1, \ldots, X_k} + \sum_{i=1}^{k} X_i \log p_i + \lambda\,(p_1 + \ldots + p_k - 1)$$

for a constant λ to be chosen later. Clearly, subject to $p_1 + \ldots + p_k = 1$, maximizing $l\,(p_1, \ldots, p_k)$ is the same as maximizing $L\,(p_1, \ldots, p_k, \lambda)$. Ignoring momentarily the constraint $p_1 + \ldots + p_k = 1$, the unconstrained maximizer of L is obtained by setting for each $i = 1, \ldots, k$

$$0 = \frac{\partial L}{\partial p_i} = \frac{X_i}{p_i} + \lambda,$$

which yields $\hat{p}_i = -X_i/\lambda$. For the specific choice of constant $\lambda = -n$, we obtain $\hat{p}_i = X_i/n$ and $\sum_{i=1}^{n} \hat{p}_i = \sum_{i=1}^{n} X_i/n = 1$, so the constraint is satisfied. As $\hat{p}_i = X_i/n$ is the unconstrained maximizer of $L\,(p_1, \ldots, p_k, -n)$, this implies that it must also be the constrained maximizer of $L\,(p_1, \ldots, p_k, -n)$, so it is the constrained maximizer of $l\,(p_1, \ldots, p_k)$. So the MLE is given by $\hat{p}_i = X_i/n$ for $i = 1, \ldots, k$.

Problem 8. (Properties of Estimator)

(a) Suppose X_1, X_2, X_3 are independent normally distributed random variables with mean μ and variance σ^2. However, instead of X_1, X_2, X_3, we only observe $Y_1 = X_2 - X_1$ and $Y_2 = X_3 - X_2$. Which of the following statistics is sufficient for σ^2?

(i) $Y_1^2 + Y_2^2 - Y_1 Y_2$ (ii) $Y_1^2 + Y_2^2 + 2Y_1 Y_2$ (iii) $Y_1^2 + Y_2^2$ (iv) $Y_1^2 + Y_2^2 + Y_1 Y_2$.

Solution. Given $X_1, X_2, X_3 \overset{\text{iid}}{\sim} N\left(\mu, \sigma^2\right)$ and $Y_1 = X_2 - X_1$ and $Y_2 = X_3 - X_2$. Now,

$$\begin{aligned} E\,(Y_1) &= E\,(X_2) - E\,(X_1) = \mu - \mu = 0 \\ E\,(Y_2) &= E\,(X_3) - E\,(X_2) = \mu - \mu = 0 \\ V\,(Y_1) &= V\,(X_2) + V\,(X_1) - 2\,\text{Cov}\,(X_2, X_1) = \sigma^2 + \sigma^2 - 0 \quad = 2\sigma^2 \\ V\,(Y_2) &= V\,(X_3) + V\,(X_2) - 2\,\text{cov}\,(X_3, X_2) = 2\sigma^2 \end{aligned}$$

$$\operatorname{Cov}(Y_1, Y_2) = \operatorname{Cov}(X_2 - X_1, X_3 - X_2) = \operatorname{Cov}(X_2, X_3) - \operatorname{Cov}(X_2, X_2) - \operatorname{Cov}(X_1, X_3) + \operatorname{Cov}(X_1, X_2) = -\sigma^2.$$

$$\operatorname{Corr}(Y_1, Y_2) = \frac{\operatorname{Cov}(Y_1, Y_2)}{\sqrt{V(Y_1)\,V(Y_2)}} = \frac{-\sigma^2}{\sqrt{2\sigma^2 \cdot 2\sigma^2}} = -\frac{1}{2}.$$

Since Y_1, Y_2 are linear combinations of Normal random variables, their joint distribution is also Normal

$$(Y_1, Y_2) \sim N_2\left(0, 0, 2\sigma^2, 2\sigma^2, -1/2\right)$$

$$f_{Y_1,,Y_2}(y_1, y_2) = \frac{1}{2\sqrt{3}\pi\sigma^2} e^{-\frac{2}{3}\left[\frac{y_1^2}{2\sigma^2} + \frac{y_2^2}{2\sigma^2} + \frac{y_1 y_2}{2\sigma^2}\right]} = \frac{1}{2\sqrt{3}\pi\sigma^2} e^{-\left[y_1^2 + y_2^2 + y_1 y_2\right]/3\sigma^2} = g_{\sigma^2}(T(y_1, y_2))\,h(y_1, y_2),$$

where $g_\sigma^2(T(y_1, y_2)) = \frac{1}{2\sqrt{3}\pi\sigma^2} e^{-\left[y_1^2+y_2^2+y_1y_2\right]/3\sigma^2}$ and $h(y_1, y_2) = 1$ and $T(y_1, y_2) = y_1^2 + y_2^2 + y_1 y_2$.
$\therefore$ By Factorization Theorem, $Y_1^2 + Y_2^2 + Y_1 Y_2$ is sufficient for σ^2.

(b) Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter p; $0 \le p \le 1$. The bias of the estimator $\frac{\sqrt{n}+2\sum_{i=1}^n X_i}{2(n+\sqrt{n})}$ for estimating p is equal to

(i) $\frac{1}{\sqrt{n}+1}\left(p - \frac{1}{2}\right)$ (ii) $\frac{1}{n+\sqrt{n}}\left(\frac{1}{2} - p\right)$ (iii) $\frac{1}{\sqrt{n}+1}\left(\frac{1}{2} + \frac{p}{\sqrt{n}}\right) - p$ (iv) $\frac{1}{\sqrt{n}+1}\left(\frac{1}{2} - p\right)$.

Solution. Given $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \text{Ber}(p)$.

$$\text{Let,} \quad T = \frac{\sqrt{n} + 2\sum_{i=1}^n X_i}{2(n+\sqrt{n})}.$$

$$\begin{aligned}
\text{Bias}(T) &= E(T) - p \\
&= E\left(\frac{\sqrt{n} + 2\sum_{i=1}^n X_i}{2(n+\sqrt{n})}\right) - p = \frac{\sqrt{n} + 2E\left(\sum X_i\right)}{2(n+\sqrt{n})} - p \\
&= \frac{\sqrt{n} + 2np}{2(n+\sqrt{n})} - p \quad \left[\; since \quad \textstyle\sum_{i=1}^N X_i \sim \text{Bin}(n, p)\,. \;\right] \\
&= \frac{\sqrt{n} + 2np - 2np - 2\sqrt{n}p}{2(n+\sqrt{n})} = \frac{\sqrt{n}(1-2p)}{2\sqrt{n}(1+\sqrt{n})} \\
&= \frac{1}{\sqrt{n}+1}\left(\frac{1-2p}{2}\right) = \frac{1}{\sqrt{n}+1}\left(\frac{1}{2} - p\right).
\end{aligned}$$

(c) Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with the probability density function;

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}, & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 0$. Derive the Cramér-Rao lower bound for the variance of any unbiased estimator of θ. Hence, prove that $T = \frac{1}{n}\sum_{i=1}^n X_i$ is the uniformly minimum variance unbiased estimator of θ.

Solution. $X_1, \ldots, X_n \overset{IID}{\sim} \text{Exp(mean } \theta)$

$$f(x|\theta) = \frac{1}{\theta} e^{-x/\theta} \quad ;\; x > 0.$$

Let $\tau(\theta) = \theta$ which is to be estimated.
Now, if $T(\underset{\sim}{x})$ is any unbiased estimator of $\tau(\theta)$ then by Cramer-Rao Inequality

$$V(T(\underset{\sim}{x})) \geqslant \frac{(\tau'(\theta))^2}{nE\left(\frac{\partial}{\partial\theta} \ln f(x|\theta)\right)^2}$$

provided the distribution of X follows the regularity conditions.
Since $X \sim \operatorname{Exp}($ mean $\theta)$, thus X belongs to One Parameter Exponential family and hence satisfies the regularity conditions

$$\begin{aligned} f(x \mid \theta) &= \frac{1}{\theta} e^{-x/\theta} \\ \ln f(x \mid \theta) &= -\ln \theta - \frac{x}{\theta} \\ \frac{\partial}{\partial \theta} \ln f(x|\theta) &= -\frac{1}{\theta} + \frac{x}{\theta^2} \\ E\left(\frac{\partial}{\partial \theta} \ln f(x|\theta)\right)^2 &= E\left(\frac{x}{\theta^2} - \frac{1}{\theta}\right)^2 = E\left(\frac{x^2+\theta^2-2\theta x}{\theta^4}\right) = \frac{2\theta^2+\theta^2-2\theta^2}{\theta^4} = \frac{1}{\theta^2} \end{aligned}$$

$$\text{Since, } \tau(\theta) = \theta \Rightarrow \tau'(\theta) = 1 \therefore V(T(\underset{\sim}{x})) \geqslant \frac{1}{n \cdot \left(\frac{1}{\theta^2}\right)} = \frac{\theta^2}{n}.$$

Thus, the Cramer-Rao Lower Bound $= \frac{\theta^2}{n}$
Now for $T = \frac{1}{n} \sum_{i=1}^n X_i$,

$$E(T) = \frac{1}{n} \sum E\left(X_i\right) = \theta \quad \text{and} \quad V(T) = \frac{1}{n^2} \sum V\left(X_i\right) = \frac{\theta^2}{n}.$$

$\therefore T$ is an Unbiased Estimator of θ and it attains CRLB and T is the UMVUE of θ.

(d) Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N\left(\mu, \sigma^2\right)$ distribution, where both μ and σ^2 are unknown. Find the value of b that minimizes the mean squared error of the estimator $T_b = \frac{b}{n-1} \sum_{i=1}^n \left(X_i - \bar{X}\right)^2$ for estimating σ^2, where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Solution.

$$\text{Given} \quad T_b = \frac{b}{n-1} \sum_{i=1}^n \left(X_i - \bar{X}\right)^2 = bS^2 \qquad (\text{where } S^2 \text{ is the Sample Variance})$$

$$\begin{aligned} \operatorname{MSE}\left(T_b\right) &= \operatorname{Var}\left(T_b\right) + \left(E\left(T_b\right) - \sigma^2\right)^2 = \operatorname{Var}\left(bS^2\right) + \left(E\left(bS^2\right) - \sigma^2\right)^2 \\ &= b^2 \operatorname{Var}\left(S^2\right) + \left(\sigma^2(b-1)\right)^2 = b^2 \cdot \frac{2\sigma^4}{n-1} + \sigma^4(b-1)^2 \\ &= b^2 \cdot \frac{2\sigma^4}{n-1} + \sigma^4\left(b^2+1-2b\right) = \left(\frac{2\sigma^4}{n-1} + \sigma^4\right) b^2 - 2\sigma^4 b + \sigma^4 \\ &= \left(\frac{n+1}{n-1}\right) \sigma^4 \cdot b^2 - 2\sigma^4 b + \sigma^4, \quad \text{which is a quadratic polynomial in } b. \end{aligned}$$

$$\begin{aligned} \text{Let,} \quad p(b) &= \left(\frac{n+1}{n-1}\right) \sigma^4 b^2 - 2\sigma^4 b + \sigma^4. \\ p'(b) &= \left(\frac{n+1}{n-1}\right) \sigma^4 \cdot 2b - 2\sigma^4 = 0 \quad \Rightarrow \quad b = \frac{n-1}{n+1}. \\ p''(b) &= 2\sigma^4 \left(\frac{n+1}{n-1}\right) > 0 \quad \forall \quad b \in \mathbb{R}. \end{aligned}$$

$\therefore p(b)$ is minimum at $b = \frac{n-1}{n+1}$ and for $b = \frac{n-1}{n+1}$, the $\operatorname{MSE}\left(T_b\right)$ is minimum for estimating σ^2.

(e) Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution with the probability density function;

$$f(x; \lambda) = \begin{cases} \frac{2x}{\lambda} e^{-\frac{x^2}{\lambda}}, & \text{if } x > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$. Find the maximum likelihood estimator of λ and show that it is sufficient and an unbiased estimator of λ.

Solution. Given $X_1, X_2 \ldots, X_n$ is a random sample form

$$f(x;\lambda) = \frac{2x}{\lambda} e^{-\frac{x^2}{\lambda}};\ x > 0$$

The likelihood function is

$$\begin{aligned}
L(\lambda \mid \underset{\sim}{x}) &= \prod_{i=1}^{n} \frac{2x_i}{\lambda} e^{-\frac{x_i^2}{\lambda}} = \frac{2^n}{\lambda^n} \prod_{i=1}^{n} x_i e^{-\frac{\sum x_i^2}{\lambda}};\ \ x_i\ >\ 0\,\forall\, i = 1, \ldots, n \\
\ln L(\lambda \mid \underset{\sim}{x}) &= -\frac{\sum x_i^2}{\lambda} - n\ln\lambda + \text{ terms independent of } \lambda \\
\frac{\partial}{\partial\lambda} \ln L(\lambda \mid \underset{\sim}{x}) &= \frac{\sum x_i^2}{\lambda^2} - \frac{n}{\lambda} = 0 \\
&\Rightarrow \frac{\sum x_i^2 - n\lambda}{\lambda^2} = 0 \\
&\Rightarrow \lambda = \frac{1}{n} \sum_{i=1}^{n} x_i^2
\end{aligned}$$

$\therefore \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$ is the MLE of λ. Now,

$$\begin{aligned}
E\left(X^k\right) &= \int_0^\infty x^k \cdot \frac{2x}{\lambda} e^{-\frac{x^2}{\lambda}dx};\quad \text{substituting}\quad \frac{x^2}{\lambda} = t \\
&= 2\int_0^\infty \frac{(\sqrt{\lambda t})^{k+1}}{\lambda} e^{-t} \cdot \frac{\lambda}{2\sqrt{\lambda t}} dt \\
&= \int_0^\infty (\sqrt{\lambda t})^k e^{-t} dt = \lambda^{k/2} \int_0^\infty t^{k/2+1-1} e^{-t} dt = \lambda^{k/2}\Gamma(k/2+1) \\
\therefore E(X) &= \lambda^{1/2}\Gamma(1/2+1) = \sqrt{\lambda} \cdot \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi\lambda}}{2} \\
E\left(X^2\right) &= \lambda\Gamma(2) = \lambda \\
\therefore E(\hat{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) = E\left(X_i^2\right) = \lambda
\end{aligned}$$

$\therefore \hat{\lambda}$ s the unbiased estimator of λ.

The joint pdf of $X_1, \ldots, X_n$ is

$$\begin{aligned}
f(\underset{\sim}{x} \mid \lambda) &= \prod_{i=1}^{n} \frac{2x_i}{\lambda} e^{-x_i^2/\lambda} = \frac{2^n \prod x_i}{\lambda^n} e^{-\sum_{i=1}^{n} x_i^2/\lambda} \\
&= \left(\frac{1}{\lambda^n} e^{-\frac{\sum x_i^2}{\lambda}}\right)\left(2^n \prod_1^n x_i\right) = g_\lambda\left(\sum_{i=1}^{n} x_i^2\right) h(\underset{\sim}{x})
\end{aligned}$$

$\therefore$ By Neyman Fisher Factorization Theorem (NFFT), $\sum_{i=1}^{n} X_i^2$ is a sufficient statistic for λ.

$\therefore \hat{\lambda} = \frac{1}{n}\sum X_i^2$ being a one-one function of $\sum X_i^2$ is also a sufficient statistic of λ.

Problem 9. (Sufficiency Principle)

(a) Let X be a single observation from a population belonging to the family $\{f_0(x), f_1(x)\}$, where

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \text{ and } f_1(x) = \frac{1}{\pi\left(1+x^2\right)};\quad x \in \mathbb{R}.$$

Find a non-trivial sufficient statistic for the family of distribution.

Solution. Writing the family as $\{f_\theta(x) : \theta \in \Omega = \{0, 1\}\}$. [Here, the parameter θ is called the labeling parameter.]

Define

$$I(\theta) = \begin{cases} 0 & \text{if } \theta = 0 \\ 1 & \text{if } \theta = 1 \end{cases}$$

The pdf of X is

$$\begin{aligned} f_\theta\left(x\right) &= \left\{f_0\left(x\right)\right\}^{1-I(\theta)}\left\{f_1\left(x\right)\right\}^{I(\theta)} = \left\{\frac{f_1\left(x\right)}{f_0\left(x\right)}\right\}^{I(\theta)} .f_0\left(x\right) = \left\{\frac{\frac{1}{\pi(1+x^2)}}{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}}\right\}^{I(\theta)} \cdot \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \\ &= g\left(T\left(x\right);\theta\right)h\left(x\right) \text{ where } h(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \text{ and } T(x) = x^2 \text{ or } |x|. \end{aligned}$$

Hence, X^2 or $|X|$ is sufficient for the family of distribution.

(b) Let $X_1, X_2, \ldots, X_n$ be a random sample from the following pdf. Find the non-trivial sufficient statistic in each case: [Hints: If in the range of X_i, there is the parameter of the distribution present then we have to use the concept of Indicator function $\left(X_{(1)} \text{ or } X_{(n)}\right)$ or $\min_i\{X_i\}$ or $\max_i\{X_i\}$.]

(i) $f(x;\theta) = \begin{cases} \theta x^{\theta-1} & ; \text{ if } 0 < x < 1 \\ 0 & ; \text{ otherwise.} \end{cases}$

Solution. The joint pdf of $X_1, X_2, \ldots, X_n$ is

$$\begin{aligned} f\left(\underset{\sim}{x}\right) &= \theta^n\left(\prod_{i=1}^n x_i\right)^{\theta-1} \\ &= g_\theta\left\{\prod_{i=1}^n x_i\right\}\cdot h\left(\underset{\sim}{x}\right), \text{ where } h\left(\underset{\sim}{x}\right) = 1 \text{ and } T\left(\underset{\sim}{x}\right) = \left(\prod_{i=1}^n x_i\right) \end{aligned}$$

$\therefore$ By Neyman-Fisher Factorization criterion, $T\left(\underset{\sim}{X}\right) = \prod_{1=1}^n X_i$ is sufficient for θ.

(ii) $f\left(x;\mu\right) = \frac{1}{|\mu|\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)}{2\mu^2}}; x \in \mathbb{R}$

Solution. We know if $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$:

$$f\left(x;\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)}{2\sigma^2}}.$$

So, in the given problem $X \sim \mathrm{N}\left(\mu, \mu^2\right)$, where $\mu \neq 0$. Hence, $T(\underset{\sim}{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is sufficient for μ.

(iii) $f(x;\alpha,\beta) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathbf{B}(\alpha,\beta)} & ; \text{ if } 0 < x < 1;\ \alpha,\ \beta\ >\ 0 \\ 0 & ; \text{ otherwise.} \end{cases}$

Solution. The joint pdf of $X_1 \ldots, X_n$ is

$$\begin{aligned} f(\underset{\sim}{x}) &= \left[\frac{1}{\mathbf{B}(\alpha,\beta)}\right]^n\left(\prod_{i=1}^n x_i\right)^{\alpha-1}\left(\prod_{i=1}^n (1-x_i)^{\beta-1}\right) \\ &= g\left(T\left(\underset{\sim}{x}\right);\alpha,\beta\right)h\left(\underset{\sim}{x}\right), \text{ where } h\left(\underset{\sim}{x}\right) = 1 \text{ and } T(\underset{\sim}{x}) = \left(\prod_{i=1}^n x_i, \prod_{i=1}^n (1-x_i)\right). \end{aligned}$$

Hence, $T(\underset{\sim}{X}) = \left(\prod_{i=1}^n X_i, \prod_{i=1}^n (1-X_i)\right)$ is jointly sufficient for (α,β).

(iv) $f(x;\mu,\lambda)=\begin{cases}\frac{1}{\lambda}e^{-\frac{(x-\mu)}{\lambda}} & ; \text{ if } x>\mu\\ 0 & ; \text{ otherwise.}\end{cases}$

Solution. The joint PDF of $\underset{\sim}{X}$ is

$$\begin{aligned} f\left(\underset{\sim}{x}\right) &= \frac{1}{\lambda^n}\cdot e^{\frac{-\sum_{i=1}^n(x_i-\mu)}{\lambda}} \text{ if } x_i>\mu \\ &= \frac{1}{\lambda^n}\cdot\exp\left\{\frac{-\sum_{i=1}^n x_i - n\mu}{\lambda}\right\}\cdot I\left(x_{(1)},\mu\right) \qquad \text{where } I(a,b)=\begin{cases}1 & \text{if } a\ge b\\ 0 & \text{otherwise.}\end{cases}\\ &= g\left(\sum_{i=1}^n x_i, x_{(1)};\lambda,\mu\right)\times h(\underset{\sim}{x}),\ \text{ where } h(\underset{\sim}{x})=1.\end{aligned}$$

Thus, $X_{(1)}$ and $\sum_{i=1}^n X_i$ are jointly sufficient statistic for μ and λ.

(v) $f(x;\mu,\sigma)=\begin{cases}\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(\ln x-\mu)^2} & ; \text{ if } x>0\\ 0 & ; \text{ otherwise.}\end{cases}$

Solution. The joint PDF of $\underset{\sim}{X}$ is

$$\begin{aligned} f\left(\underset{\sim}{x}\right) &= \frac{1}{\left(\prod_{i=1}^n x_i\right)\sigma^n\left(\sqrt{2\pi}\right)^n}\cdot\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(\ln x_i-\mu)^2\right\}; \qquad \text{if } x_i>0\\ &= \frac{1}{\sigma^n\left(\sqrt{2\pi}\right)^n}\cdot\exp-\left\{\frac{\sum_{i=1}^n(\ln x_i)^2}{2\sigma^2}-\frac{\mu\sum_{i=1}^n\ln x_i}{\sigma^2}+\frac{n\mu^2}{2\sigma^2}\right\}\cdot\frac{1}{\prod_{i=1}^n x_i}\\ &= T\left(\sum_{i=1}^n\ln x_i,\sum_{i=1}^n(\ln x_i)^2;\mu,\sigma\right)\cdot h\left(\underset{\sim}{x}\right);\ \text{where } h\left(\underset{\sim}{x}\right)=\frac{1}{\prod_{i=1}^n x_i}.\end{aligned}$$

Hence, $T\left(\underset{\sim}{X}\right)=\left(\sum_{i=1}^n\ln X_i,\sum_{i=1}^n(\ln X_i)^2\right)$ is sufficient for μ and σ.

(vi) $f(x;\alpha,\theta)=\begin{cases}\frac{\theta\alpha^\theta}{x^{\theta+1}} & ; \text{ if } x>\alpha\\ 0 & ; \text{ otherwise.}\end{cases}$

Solution. The joint PDF of $\underset{\sim}{X}$ is

$$\begin{aligned} f(\underset{\sim}{x}) &= \theta^n\frac{\left(\alpha^\theta\right)^n}{\prod_{i=1}^n\left(x_i^{\theta+1}\right)} \text{ if } x_i>\alpha\\ &= \left(\theta\alpha^\theta\right)^n\frac{1}{\prod_{i=1}^n\{x_i\}^{\theta+1}}I\left(x_{(1)},\alpha\right) \qquad \text{if } x_{(1)}>\alpha,\ \text{and } I(a,b)=\begin{cases}1 & \text{if } a>b\\ 0 & \text{otherwise.}\end{cases}\\ &= g\left(\prod_{i=1}^n x_i,x_{(1)};\theta,\alpha\right)\cdot h\left(\underset{\sim}{x}\right);\ \text{where } h\left(\underset{\sim}{x}\right)=1 \text{ and } T\left(\underset{\sim}{x}\right)=\left(\prod_{i=1}^n x_i,x_{(1)}\right).\end{aligned}$$

Hence, $T\left(\underset{\sim}{X}\right)=\left(\prod_{i=1}^n X_i,X_{(1)}\right)$ is sufficient for $(\theta,\,\alpha)$.

(vii) $f(x;\theta)=\begin{cases}\frac{2(\theta-x)}{\theta^2} & ; \text{ if } 0<x<\theta\\ 0 & ; \text{ otherwise.}\end{cases}$

Solution. The joint PDF of $\underset{\sim}{X}$ is

$$f\left(\underset{\sim}{x}\right) = \frac{2^n}{\theta^{2n}} \prod_{i=1}^{n} (\theta - x_i)\,; 0 < x_i < \theta$$
$$= \left(\frac{2}{\theta^2}\right)^n \cdot (\theta - x_1)(\theta - x_2)\ldots(\theta - x_n)\,; 0 < x_i < \theta.$$

These cannot be expressed in the form of factorization criterion. So, $(X_1, X_2, \ldots, X_n)$ or $\left(X_{(1)}, X_{(2)}, \ldots X_{(n)}\right)$ are trivially sufficient for θ. Hence there is no non-trivial sufficient statistic.

(c) If $f_\theta(x) = \frac{1}{2}$; $\theta - 1 < x < \theta + 1$, then show that $X_{(1)}$ and $X_{(n)}$ are jointly sufficient for θ, $X_i \sim U(\theta - 1, \theta + 1)$.

Solution. The joint PDF of $\underset{\sim}{X}$ is

$$f\left(\underset{\sim}{x}\right) = \left(\frac{1}{2}\right)^n$$
$$= \frac{1}{2^n} \cdot I\left(\theta - 1, x_{(1)}\right) I\left(x_{(n)}, \theta + 1\right); \quad \theta - 1 < x_{(1)} < x_{(n)} < \theta + 1 \text{ where } I(a,b) = \begin{cases} 1 & \text{if } a < b \\ 0 & \text{if } a \geq b \end{cases}$$
$$= g\left(T\left(\underset{\sim}{x}\right); \theta\right) h\left(\underset{\sim}{x}\right); \text{ where } h\left(\underset{\sim}{x}\right) = \frac{1}{2^n} \text{ and } T\left(\underset{\sim}{x}\right) = \left(x_{(1)}, x_{(n)}\right).$$

Hence $T\left(\underset{\sim}{X}\right) = \left(X_{(1)}, X_{(n)}\right)$ is jointly sufficient for θ.

(d) If a random sample of size $n \geqslant 2$ is drawn from a Cauchy distribution with PDF

$$f_\theta(x) = \frac{1}{\pi\left[1 + (x - \theta)^2\right]},$$

where $-\infty < \theta < \infty$, is considered. Then can you have a single sufficient statistic for θ?

Solution.

$$\prod_{i=1}^{n} f(x_i; \theta) = \frac{1}{\pi^n \left\{\prod_{i=1}^{n}\left[1 + (x_i - \theta)^2\right]\right\}}.$$

Note that $\displaystyle\prod_{i=1}^{n}\left\{1 + (x_i - \theta)^2\right\} = \left\{1 + (x_1 - \theta)^2\right\}\left\{1 + (x_2 - \theta)^2\right\}\ldots\left\{1 + (x_n - \theta)^2\right\}$

$$= 1 + \text{ term involving one } x_i + \text{ term involving two } x_i's + \ldots + \text{ term involving all } x_i's$$
$$= 1 + \sum_i (x_i - \theta)^2 + \sum_{i \neq j} (x_i - \theta)^2 (x_j - \theta)^2 + \ldots + \prod_{i=1}^{n} (x_i - \theta)^2.$$

Clearly, $\prod_{i=1}^{n} f(x_i; \theta)$ cannot be written as $g\left(T\left(\underset{\sim}{x}\right), \theta\right) \cdot h\left(\underset{\sim}{x}\right)$ for a statistic other than the trivial choices $(X_1, \ldots, X_n)$ or $\left(X_{(1)}, \ldots, X_{(n)}\right)$. Hence there is no non-trivial sufficient statistic. Therefore, in this case, no reduction in space is possible. Thus, the whole set $(X_1, \ldots, X_n)$ is jointly sufficient for θ.

Problem 10. (Uniformly Minimum Variance Unbiased Estimator)

(a) Let $X_1, X_2, \ldots, X_n$ be a random sample from $f(x;p) = \begin{cases} p(1-p)^x & ; x = 0, 1, \ldots \\ 0 & ; \text{ otherwise.} \end{cases}$
Show that unbiased estimator of p based on $T = \sum_{i=1}^n X_i$ is unique. Hence or otherwise, find the UMVUE of p.

Solution. $T = \sum_{i=1}^n X_i \sim \text{NB}(n, p)$.

To solve for $h(T)$ such that

$$\begin{aligned}
& E\{h(T)\} = p \quad \forall \quad p \in (0,1) \\
\Rightarrow & \sum_{t=0}^{\infty} h(t) \binom{t+n-1}{n-1} p^n q^t = p \quad \forall \quad p \\
\Rightarrow & \sum_{t=0}^{\infty} h(t) \binom{t+n-1}{n-1} q^t = p^{-(n-1)} = (1-q)^{n-1} \\
\Rightarrow & \sum_{t=0}^{\infty} h(t) \binom{t+n-1}{n-1} q^t = \sum_{t=0}^{\infty} \binom{n-1+t-1}{t} q^t, \quad \text{as } 0 < q < 1.
\end{aligned}$$

By the uniqueness property of Power Series, we get,

$$h(t)\binom{t+n-1}{n-1} = \binom{n+t-2}{t}, \; t = 0, 1, 2, \ldots.$$

Hence, $h(T) = \frac{n-1}{t+n-1}$ is the only solution of $E\{h(T)\} = p \quad \forall \quad p$.

Thus, $h(T)$ is the only UE of p based on T.

It can be shown that $T = \sum_{i=1}^n X_i$ is sufficient.

By Rao-Blackwell theorem, UMVUE is a function of T.

As there is only one UE of p based on T, then UE $h(T)$ is the UMVUE of p.

(b) Let X_1 and X_2 be two independent random variables having the same mean θ. Suppose that $E(X_1 - \theta)^2 = 1$ and $E(X_2 - \theta)^2 = 2$. For estimating θ, consider the estimators $T_\alpha(X_1, X_2) = \alpha X_1 + (1-\alpha)X_2, \alpha \in [0,1]$. The value of $\alpha \in [0,1]$, for which the variance of $T_\alpha(X_1, X_2)$ is minimum, equals

(i) $\frac{2}{3}$ (ii) $\frac{1}{2}$ (iii) $\frac{1}{4}$ (iv) $\frac{3}{4}$

Solution. Given X_1 and X_2 are independent with

$$E(X_1) = E(X_2) = \theta \text{ and}$$
$$V(X_1) = E(X_1 - \theta)^2 = 1, V(X_2) = E(X_2 - \theta)^2 = 2$$

Given $T_\alpha(X_1, X_2) = \alpha X_1 + (1-\alpha)X_2, \; \alpha \in [0,1]$

$$\begin{aligned}
V(T_\alpha(X_1, X_2)) =& V(\alpha X_1 + (1-\alpha)X_2) \\
=& \alpha^2 V(X_1) + (1-\alpha)^2 V(X_2) + 2\alpha(1-\alpha)\operatorname{Cov}(X_1, X_2) \quad \begin{bmatrix} \operatorname{Cov}(X_1, X_2) = 0 \\ \because X_1, X_2 \text{ are independent} \end{bmatrix} \\
=& \alpha^2 + 2(1-\alpha)^2 = \alpha^2 + 2\left(1 - 2\alpha + \alpha^2\right) = 3\alpha^2 - 4\alpha + 2
\end{aligned}$$

which is a quadratic polynomial in α.

$$\begin{aligned}
\text{Let} \quad p(\alpha) &= 3\alpha^2 - 4\alpha + 2, \; \alpha \in [0,1] \\
p'(\alpha) &= 6\alpha - 4 = 0 \Rightarrow \alpha = 4/6 = 2/3 \\
p''(\alpha) &= 6 > 0 \quad \forall \quad \alpha \in [0,1]
\end{aligned}$$

$\therefore p(\alpha)$ is minimum at $\alpha = 2/3$. $\therefore$ For $\alpha = 2/3$, $V(\mathrm{T}_\alpha(X_1, X_2))$ is minimum.

(c) Let $X_1, X_2, \ldots, X_n$ be a random sample from

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)}, & \text{if } x > \theta \\ 0, & \text{otherwise.} \end{cases}$$

Show that $T = X_{(1)}$ is a complete sufficient statistic. Hence find the UMVUE of θ.

Solution. The PDF of $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ is

$$\begin{aligned}
\prod_{i=1}^{n} f(x_i;\theta) &= \begin{cases} e^{-\sum_{i=1}^{n}(x_i-\theta)} & , \text{ if } x_i > \theta,\ \forall i = 1(1)n \\ 0 & , \text{ otherwise} \end{cases} \\
&= \begin{cases} e^{-\sum_{i=1}^{n}(x_i-\theta)} & , \text{ if } x_{(1)} > \theta \\ 0 & , \text{ otherwise} \end{cases} \\
&= e^{-\sum_{i=1}^{n}(x_i-\theta)} \cdot I\left(x_{(1)}, \theta\right), \text{ where } I(a,b) = \begin{cases} 1 & , \text{ if } a > b \\ 0 & , \text{ if } a < b. \end{cases} \\
&= e^{\theta} \cdot I\left(x_{(1)}, \theta\right) \cdot e^{-\sum_{i=1}^{n} x_i} \\
&= g\left(T\left(\underset{\sim}{x}\right), \theta\right) \cdot h\left(\underset{\sim}{x}\right) \text{ with } T\left(\underset{\sim}{x}\right) = x_{(1)}.
\end{aligned}$$

By factorization criterion $T = X_{(1)}$ is sufficient.

Let

$$E\{h(T)\} = 0 \quad \forall\theta$$

$$\Rightarrow \int_{-\infty}^{\infty} h(t) \cdot f_T(t) dt = 0 \quad \forall\theta \qquad \left[\begin{array}{rl} F_T(t) & = 1 - P[T > t] \\ & = 1 - P\left[X_{(1)} > t\right] \\ & = 1 - \{P[X_1 > t]\}^n \\ & = 1 - \left\{\int_t^{\infty} e^{-(x_1-\theta)} dx_1\right\}^n \text{ if } t > \theta \\ & = 1 - e^{-n(t-\theta)} \text{ if } t > \theta \\ \therefore f_T(t) & = \begin{cases} ne^{-n(t-\theta)} & , \text{ if } t > \theta \\ 0 & , \text{ otherwise .} \end{cases} \end{array}\right]$$

$$\Rightarrow \int_{-\theta}^{\infty} h(t) \cdot ne^{-n(t-\theta)} dt = 0 \quad \forall\,\theta$$

$$\Rightarrow \int_{\theta}^{\infty} h(t) \cdot e^{-nt} dt = 0 \quad \forall\,\theta$$

Differentiating w.r.t. θ,

$$0 - h(\theta) \cdot e^{-n\theta} = 0 \quad \forall \quad \theta \quad \Rightarrow \quad h(\theta) = 0 \quad \forall \quad \theta \text{ as } e^{-n\theta} > 0.$$

Hence, $h(T) = 0$, with probability 1, $\forall\,\theta \Rightarrow T$ is complete.

Now,

$$\begin{aligned}
E(T - \theta) &= \int_{-\infty}^{\infty} (t-\theta) f_T(t) dt = \int_0^{\infty} (t-\theta) ne^{-n(t-\theta)} dt \\
&= \frac{1}{n} \int_0^{\infty} ue^{-u} du, \text{ where } u = n(t-\theta) = \frac{1}{n} \cdot \Gamma(2) = \frac{1}{n}
\end{aligned}$$

Thus, $E\left(T - \frac{1}{n}\right) = \theta$. By Lehmann–Scheffé Theorem, $h(T) = T - \frac{1}{n} = X_{(1)} - \frac{1}{n}$ is the UMVUE of θ.

(d) Is the following families of distribution regular in the sense of Cramer & Rao? If so, find the lower bound for the variance of an unbiased estimator of θ based on a sample of size n. Also, find the UMVUE of θ for the PDF:

$$f(x,\theta) = \frac{e^{-\frac{x^2}{2\theta}}}{\sqrt{2\pi\theta}} \quad ; -\infty < x < \infty, \quad \theta > 0.$$

Solution. As we know '=' holds in CR inequality whenever the family of distributions is OPEF. The given pdf is OPEF, satisfying the regularity conditions for CR inequality. That is, it is regular in the sense of Cramer-Rao.

By CR inequality, for an unbiased estimator T of θ,

$$\operatorname{Var}(T) \geqslant \frac{1}{I_n(\theta)} = CRLB.$$

$$\text{Here, } f(x,\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}} \quad ;\ x \in \mathbb{R},\ \theta > 0 \Rightarrow \ln f(x,\theta) = -\frac{1}{2}\ln(2\pi\theta) - \frac{x^2}{2\theta}.$$

$$\text{Differentiating w.r.t } \theta \text{ we get, } \frac{\partial}{\partial\theta}\ln f(x_1,\theta) = -\frac{1}{2\theta} + \frac{x_1^2}{2\theta^2} \text{ and } \frac{\partial^2}{\partial\theta^2}\ln f(x_1,\theta) = \frac{1}{2\theta^2} - \frac{x_1^2}{\theta^3}.$$

$$I_n(\theta) = n \cdot I_1(\theta) = n \cdot E\left(-\frac{\partial^2}{\partial\theta^2}\ln f(x_1,\theta)\right) = n \cdot \left\{-\frac{1}{2\theta^2} + \frac{E\left(X_1^2\right)}{\theta^3}\right\} = n\left\{-\frac{1}{2\theta^2} + \frac{\theta}{\theta^3}\right\} = \frac{n}{2\theta^2}.$$

Hence, $\operatorname{Var}(T) \geqslant \frac{2\theta^2}{n} = CRLB$. The MVBUE, if exists for θ, is given by

$$T = \psi(\theta) \pm \frac{\psi'(\theta)}{\mathrm{I_n}(\theta)} \cdot \frac{\partial}{\partial\theta}\ln L(\underset{\sim}{x},\theta) = \theta \pm \frac{1}{\frac{n}{2\theta^2}} \cdot \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\ln f(x_i,\theta) = \theta \pm \frac{2\theta^2}{n}\left\{-\frac{n}{2\theta} + \frac{\sum x_i^2}{2\theta^2}\right\}$$

$$= \theta + \frac{2\theta^2}{n}\left\{-\frac{n}{2\theta} + \frac{\sum x_i^2}{2\theta^2}\right\}, \text{ taking +ve sign only. } = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

Hence, $T = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ attains CRLB and $\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)$ is the UMVUE of θ.

(e) Based on a random sample $X_1, X_2, \ldots, X_n$ from Gamma(α). Obtain an estimator of $\psi_\alpha = \frac{\partial}{\partial\alpha}\ln\Gamma(\alpha)$ which attains CRLB and its variance.

Solution. The pdf of $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ is

$$f\left(\underset{\sim}{x},\alpha\right) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} e^{-x_i} \cdot x_i^{\alpha-1} = \frac{1}{\{\Gamma(\alpha)\}^n} e^{-\sum_{i=1}^{n} x_i} \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1}, \text{ if } x_i > 0,\ \forall\ i = 1(1)n.$$

$$\Rightarrow \ln f\left(\underset{\sim}{x},\alpha\right) = -n\ln\Gamma(\alpha) - \sum_{i=1}^{n} x_i + (\alpha-1)\sum_{i=1}^{n}\ln x_i, \qquad \text{if } x_i > 0$$

$$\frac{\partial}{\partial\alpha}\ln f\left(\underset{\sim}{x},\alpha\right) = -n\frac{\partial}{\partial\alpha}\ln\Gamma(\alpha) + \sum_{i=1}^{n}\ln x_i$$

$$\text{and } \frac{\partial^2}{\partial\alpha^2}\ln f\left(\underset{\sim}{x},\alpha\right) = -n\frac{\partial^2}{\partial\alpha^2}\ln\Gamma(\alpha)$$

$$\text{Here, } I(\alpha) = E\left(-\frac{\partial^2}{\partial\alpha^2}\ln f\left(\underset{\sim}{x},\alpha\right)\right) = n\frac{\partial^2}{\partial\alpha^2}\ln\Gamma(\alpha).$$

A UE which attains CRLB, if it exists, is given by,

$$T = \psi(\alpha) \pm \frac{\psi'(\alpha)}{I(\alpha)} \cdot \frac{\partial}{\partial\alpha}\ln f\left(\underset{\sim}{x},\alpha\right) = \frac{\partial}{\partial\alpha}\ln\Gamma(\alpha) \pm \frac{\frac{\partial^2}{\partial\alpha^2}\ln\Gamma(\alpha)}{n\frac{\partial^2}{\partial\alpha^2}\ln\Gamma(\alpha)}\left\{-n\frac{\partial}{\partial\alpha}\ln\Gamma(\alpha) + \Sigma\ln x_i\right\}.$$

$$= \frac{\partial}{\partial\alpha}\ln\Gamma(\alpha) \pm \left\{-\frac{\partial}{\partial\alpha}\ln\Gamma(\alpha) + \frac{1}{n}\sum_{i}\ln x_i\right\} = \frac{1}{n}\sum_{i=1}^{n}\ln x_i, \text{ taking positive sign only}$$

$$= \ln G, \text{ where } G = \left(\prod_{i=1}^{n} X_i\right)^{1/n} \text{ is the GM of } X_1, X_2, \ldots, X_n.$$

$$\text{Clearly, } \operatorname{Var}(T) = \text{ CRLB } = \frac{\{\psi'(\alpha)\}^2}{I(\alpha)} = \frac{\left\{\frac{\partial^2}{\partial\alpha^2}\ln\Gamma(\alpha)\right\}^2}{n \cdot \frac{\partial^2}{\partial\alpha^2}\ln\Gamma(\alpha)} = \frac{\frac{\partial^2}{\partial\alpha^2}\ln\Gamma(\alpha)}{n}.$$

Problem 11. (Finding Confidence Intervals)

(a) Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{U}(0, \theta),\ \theta > 0$. Find a confidence interval for θ with confidence coefficient $(1-\alpha)$, based on $X_{(n)}$.

(b) Consider a random sample of size n from the rectangular distribution

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

If Y be the sample range then ξ is given by $\xi^{n-1}\left[n - (n-1)\xi\right] = \alpha$. Show that, Y and $Y\xi^{-1}$ are confidence limit to θ with confidence coefficient $(1-\alpha)$.

(c) Consider a random sample of size n from an exponential distribution, with PDF

$$f_X(x) = \begin{cases} \exp[-(x-\theta)] & \text{if } \theta < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Suggest a $100(1-\alpha)\%$ confidence interval for θ.

Solution.

(a) The pdf of $X_{(n)}$ is

$$f_{X_{(n)}}(x) = \begin{cases} \frac{nx^{n-1}}{\theta^n} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

The pdf $\psi\left(X_{(n)}, \theta\right) = \frac{X_{(n)}}{\theta} = T$ is

$$g(t) = \begin{cases} nt^{n-1} & \text{if } 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}$$

which is independent of θ.

$$\begin{aligned} &\text{Now, } P\left[c < \psi\left(X_{(n)}, \theta\right) < 1\right] = 1 - \alpha. \\ &\Rightarrow \int_c^1 nt^{n-1} dt = 1 - \alpha, \ \text{where } c \text{ is the critical region.} \\ &\Rightarrow 1 - c^n = 1 - \alpha, \ \text{i.e. } c = \alpha^{1/n}. \end{aligned}$$

$$\begin{aligned} \text{Note that, } \alpha^{1/n} < \psi(X_{(n)}, \theta) = \frac{X_{(n)}}{\theta} < 1 \\ \Rightarrow \alpha^{-1/n} > \frac{\theta}{X_{(n)}} > 1 \\ \text{i.e., } X_{(n)} < \theta < \alpha^{-1/n} X_{(n)}. \end{aligned}$$

Hence, $\left[X_{(n)},\ \alpha^{-1/n} X_{(n)}\right]$ is a confidence interval for θ with confidence coefficient $(1-\alpha)$.

(b) Here, $Y = X_{(n)} - X_{(1)}$. The pdf of Y, is

$$f_Y(y) = \begin{cases} n(n-1)y^{n-2}(1-y) & \text{if } 0 < y < \theta \\ 0 & \text{otherwise.} \end{cases}$$

The pdf of $\psi(Y, \theta) = U$, is

$$f_U(u) = \begin{cases} n(n-1)u^{n-2}(1-u) & \text{if } 0 < u < 1 \\ 0 & \text{otherwise} \end{cases}$$

which is independent of θ.

$$\begin{aligned}&\text{Now, } P[\xi \leqslant U \leqslant 1] = 1-\alpha\\ &\Rightarrow \int_{\xi}^{1} n(n-1)u^{n-2}(1-u)du = 1-\alpha\\ &\Rightarrow n(n-1)\int_{\xi}^{1}\left[u^{n-2}-u^{n-1}\right]du = 1-\alpha\\ &\Rightarrow \xi^{n-1}[n-(n-1)\xi] = \alpha.\end{aligned}$$

Note that $\{\xi \leqslant U \leqslant 1\} = \left\{\xi \leqslant \frac{Y}{\theta} \leqslant 1\right\} = \left\{Y \leqslant \theta \leqslant \frac{Y}{\xi}\right\}$.

Hence, $\left(Y,\, Y\xi^{-1}\right)$ is a random confidence interval for θ with confidence coefficient $1-\alpha$, where ξ is such that $\xi^{n-1}[n-(n-1)\xi] = \alpha$.

(c) The distribution function of $X_{(1)}$ is

$$\begin{aligned}F_{X_{(1)}}(x_{(1)}) &= P\left[X_{(1)} \leqslant x_{(1)}\right] = 1-P\left[X_{(1)} > x_{(1)}\right] = 1-\left\{P\left[X > x_{(1)}\right]\right\}^n\\ &= 1-\left\{e^{-(x_{(1)}-\theta)}\right\}^n = 1-e^{-n\left(x_{(1)}-\theta\right)} \text{ if } x_{(1)} > \theta.\end{aligned}$$

Hence, $U = e^{-n\left(X_{(1)}-\theta\right)} = 1-F_{X_{(1)}}\left(x_{(1)}\right) \sim U(0,1)$.

We know the p.d.f. of $X_{(1)}$ is $f_{X_{(1)}}(x_{(1)}) = \dfrac{d}{dx_{(1)}}F_{X_{(1)}}(x_{(1)}) = ne^{-n(x_{(1)}-\theta)} \quad \text{if } x_{(1)} > \theta$

Let $u = e^{-\left(x_{(1)}-\theta\right)} \Rightarrow \log u = -(x_{(1)}-\theta) \Rightarrow \dfrac{1}{u}\cdot du = -dx_{(1)} \Rightarrow J = \left|\dfrac{dx_{(1)}}{du}\right| = \dfrac{1}{u}$.

$$\therefore f_U(u) = \begin{cases} nu^{n-1} & \text{if } 0 < u < 1\\ 0 & \text{otherwise.}\end{cases}$$

$$\text{Now,}\quad 1-\alpha = P[c \leqslant U \leqslant 1] = \int_c^1 nu^{n-1}du \quad \left|\quad \begin{aligned}&\alpha = P[0 \leqslant U \leqslant c]\\ &\Rightarrow c = \alpha^{1/n}\end{aligned}\right.$$

$$\Rightarrow 1-\alpha = 1-c^n \Rightarrow c = \alpha^{1/n}.$$

$$\begin{aligned}\text{Note that,}\quad &\alpha^{1/n} \leqslant u \leqslant 1 \Rightarrow \alpha^{1/n} \leqslant e^{-\left(x_{(1)}-\theta\right)} \leqslant 1\\ &\Rightarrow \frac{1}{n}\log\alpha \leqslant -(x_{(1)}-\theta) \leqslant 0 \Rightarrow x_{(1)}+\frac{1}{n}\log\alpha \leqslant \theta \leqslant x_{(1)}.\end{aligned}$$

Problem 12. (A Heteroskedastic Linear Model)

Consider observed response variables $Y_1,\ldots,Y_n \in \mathbb{R}$ that depend linearly on a single covariate $x_1,\ldots,x_n$ as follows:

$$Y_i = \beta x_i + \varepsilon_i.$$

Here, the ε_i's are independent Gaussian noise variables, but we do not assume they have the same variance. Instead, they are distributed as $\varepsilon_i \sim \mathcal{N}\left(0,\sigma_i^2\right)$ for possibly different variances $\sigma_1^2,\ldots,\sigma_n^2$. The unknown parameter of interest is β.

(a) Suppose that the error variances $\sigma_1^2,\ldots,\sigma_n^2$ are all known. Show that the MLE $\hat{\beta}$ for β, in this case, minimizes a certain weighted least-squares criterion, and derive an explicit formula for $\hat{\beta}$.

(b) Show that the estimate $\hat{\beta}$ in part (a) is unbiased, and derive a formula for the variance of $\hat{\beta}$ in terms of $\sigma_1^2,\ldots,\sigma_n^2$ and $x_1,\ldots,x_n$.

(c) Compute the Fisher information $I_{\mathbf{Y}}(\beta) = -\mathbb{E}_\beta\left[l''(\beta)\right]$ in this model (still assuming $\sigma_1^2, \ldots, \sigma_n^2$ are known constants). Show that the variance of $\hat{\beta}$ that you derived in part (b) is exactly equal to $I_{\mathbf{Y}}(\beta)^{-1}$.

In the remaining parts of this question, denote by $\tilde{\beta}$ the usual (unweighted) least-squares estimator for β, which minimizes $\sum_i (Y_i - \beta x_i)^2$. In practice, we might not know the values of $\sigma_1^2, \ldots, \sigma_n^2$, so we might still estimate β using $\tilde{\beta}$.

(d) Derive an explicit formula for $\tilde{\beta}$, and show that it is also an unbiased estimate of β.

(e) Derive a formula for the variance of $\tilde{\beta}$ in terms of $\sigma_1^2, \ldots, \sigma_n^2$ and $x_1, \ldots, x_n$. Show that when all error terms have the same variance σ_0^2, this coincides with the general formula $\sigma_0^2 \left(X^T X\right)^{-1}$ for the linear model.

(f) Using the Cauchy-Schwarz inequality $\left(\sum_i a_i^2\right)\left(\sum_i b_i^2\right) \geq \left(\sum_i a_i b_i\right)^2$ for any positive numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, compare your variance formulas from parts (b) and (e) and show directly that the variance of β is always at least the variance of $\hat{\beta}$. Explain, using the Cramer-Rao lower bound, why this is to be expected given your finding in (c).

Solution. (a) The log-likelihood is

$$\ell(\beta) = -\frac{n}{2}\log 2\pi + \sum_{i=1}^{n} \sigma_i - \sum_{i=1}^{n} \frac{1}{2\sigma_i^2}\left(Y_i - \beta x_i\right)^2,$$

so that the MLE $\hat{\beta}$ minimizes the weighted sum of squares

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left(Y_i - \beta x_i\right)^2.$$

Taking derivatives and solving for zero gives the explicit solution

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i/\sigma_i^2}{\sum_{i=1}^{n} x_i^2/\sigma_i^2}.$$

(b) The estimator has mean

$$\mathrm{E}[\hat{\beta}] = \frac{1}{\sum_{i=1}^{n} x_i^2/\sigma_i^2}\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2}\mathrm{E}\left[Y_i\right] = \frac{1}{\sum_{i=1}^{n} x_i^2/\sigma_i^2}\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2}\beta x_i = \beta,$$

so it is unbiased.
The variance is

$$\mathbf{V}[\hat{\beta}] = \frac{\sum_{i=1}^{n} x_i^2 \mathbf{V}\left[Y_i\right]/\sigma_i^4}{\left(\sum_{i=1}^{n} x_i^2/\sigma_i^2\right)^2} = \frac{\sum_{i=1}^{n} x_i^2\sigma_i^2/\sigma_i^4}{\left(\sum_{i=1}^{n} x_i^2/\sigma_i^2\right)^2} = \frac{1}{\sum_{i=1}^{n} x_i^2/\sigma_i^2}, \text{ using } \mathbf{V}\left[Y_i\right] = \sigma_i^2.$$

(c) Taking derivatives, we have

$$\ell'(\beta) = \sum_{i=1}^{n}\frac{x_i}{\sigma_i^2}\left(Y_i - \beta x_i\right)$$

$$\ell''(\beta) = -\sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2}.$$

Therefore, $I_{\mathbf{Y}}(\beta) = -\mathbf{E}_\beta\left[\ell''(\beta)\right] = \sum_{i=1}^{n} x_i^2/\sigma_i^2$. From part (b), this is exactly $1/\mathbf{V}[\hat{\beta}]$.

(d) Taking the derivative of $\sum_i \left(Y_i - \beta x_i\right)^2$ and solving gives

$$\tilde{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2},$$

and its expectation is

$$\mathbf{E}[\tilde{\beta}] = \frac{\sum_{i=1}^n x_i \mathbf{E}[Y_i]}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n \beta x_i^2}{\sum_{i=1}^n x_i^2} = \beta.$$

(e) The variance of $\tilde{\beta}$ is given by

$$\mathbf{V}[\tilde{\beta}] = \frac{\sum_{i=1}^n x_i^2 \mathbf{V}[Y_i]}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sum_{i=1}^n x_i^2 \sigma_i^2}{\left(\sum_{i=1}^n x_i^2\right)^2}.$$

If $\sigma_i^2 = \sigma_0^2$, then this reduces to

$$\mathbf{V}[\tilde{\beta}] = \frac{\sigma_0^2}{\sum_{i=1}^n x_i^2}.$$

In our case the model matrix X is a single vector $(x_1, \ldots, x_n)^T$, so that $\left(X^T X\right)^{-1} = 1/\sum_{i=1}^n x_i^2$. Hence the variance formula above is consistent with the general formula $\sigma_0^2 \left(X^T X\right)^{-1}$.

(f) Applying the Cauchy-Schwarz inequality with $a_i = |x_i \sigma_i|$ and $b_i = |x_i/\sigma_i|$, we obtain the inequality

$$\left(\sum_{i=1}^n x_i^2 \sigma_i^2\right)\left(\sum_{i=1}^n \frac{x_i^2}{\sigma_i^2}\right) \geq \left(\sum_{i=1}^n x_i^2\right)^2,$$

and hence, rearranging terms,

$$\mathbf{V}[\tilde{\beta}] = \frac{\sum_{i=1}^n x_i^2 \sigma_i^2}{\left(\sum_{i=1}^n x_i^2\right)^2} \geq \frac{1}{\sum_{i=1}^n x_i^2/\sigma_i^2} = \mathbf{V}[\hat{\beta}].$$

The Cramèr-Rao lower bound states that any unbiased estimator has a variance no smaller than the inverse Fisher information. Since the variance of $\hat{\beta}$ attains the lower bound by part (c), and $\tilde{\beta}$ is unbiased, the above inequality is expected.

Problem 13. (The delta method for two samples)

Let $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(p), and let $Y_1, \ldots, Y_m \overset{IID}{\sim}$ Bernoulli(q), where the X_i's and Y_i's are independent. For example, $X_1, \ldots, X_n$ may represent, among n individuals exposed to a certain risk factor for a disease, which individuals have this disease, and $Y_1, \ldots, Y_m$ may represent, among m individuals not exposed to this risk factor, which individuals have this disease. The odds-ratio

$$\frac{p}{1-p} \Big/ \frac{q}{1-q}$$

provides a quantitative measure of the association between this risk factor and this disease. The log-odds-ratio is the (natural) logarithm of this quantity,

$$\log\left(\frac{p}{1-p} \Big/ \frac{q}{1-q}\right).$$

(a) Suggest reasonable estimators $\hat{p}$ and $\hat{q}$ for p and q, and suggest a plugin estimator for the log-odds-ratio.

(b) Using the first-order Taylor expansion

$$g(\hat{p}, \hat{q}) \approx g(p, q) + (\hat{p} - p)\frac{\partial g}{\partial p}(p, q) + (\hat{q} - q)\frac{\partial g}{\partial q}(p, q)$$

as well as the Central Limit Theorem and independence of the X_i's and Y_i's, derive an asymptotic normal approximation to the sampling distribution of your plugin estimator in part (a).

(c) Give an approximate 95% confidence interval for the log-odds-ratio $\log \frac{p}{1-p} \Big/ \frac{q}{1-q}$. Translate this into an approximate 95% confidence interval for the odds-ratio $\frac{p}{1-p} \Big/ \frac{q}{1-q}$. (You may use a plugin estimate for the variance of the normal distribution you derived in part (b).)

Solution. (a) We may estimate p by $\hat{p} = \bar{X}$, and q by $\hat{q} = \bar{Y}$. The plugin estimator for the log-odds-ratio is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\bigg/\frac{\hat{q}}{1-\hat{q}}\right).$$

(b) Let

$$g(p,q) = \log\left(\frac{p}{1-p}\bigg/\frac{q}{1-q}\right) = \log p - \log(1-p) - \log q + \log(1-q).$$

Applying a first-order Taylor expansion to g,

$$g(\hat{p},\hat{q}) \approx g(p,q) + \frac{\hat{p}-p}{p(1-p)} + \frac{\hat{q}-q}{q(1-q)}.$$

$\hat{p}$ and $\hat{q}$ are independent, and by the Central Limit Theorem, $\sqrt{n}(\hat{p}-p) \to \mathcal{N}(0, p(1-p))$ and $\sqrt{m}(\hat{q}-q) \to \mathcal{N}(0, q(1-q))$. Hence, for large m and n, $g(\hat{p},\hat{q})$ is approximately distributed as $\mathcal{N}(g(p,q), v)$ where

$$v = \frac{p(1-p)}{n} \times \frac{1}{p^2(1-p)^2} + \frac{q(1-q)}{m} \times \frac{1}{q^2(1-q)^2} = \frac{1}{np(1-p)} + \frac{1}{mq(1-q)}.$$

(c) Let $\hat{v} = \frac{1}{n\hat{p}(1-\hat{p})} + \frac{1}{m\hat{q}(1-\hat{q})}$ be the plugin estimate of v. As $m,\ n \to \infty$, $\hat{v}/v \to 1$ in probability, so by part (b) and Slutsky's lemma,

$$\mathbb{P}\left[-Z_{0.025} \leq \frac{g(\hat{p},\hat{q}) - g(p,q)}{\sqrt{\hat{v}}} \leq Z_{0.025}\right] \approx 0.95$$

for large m and n. Rearranging yields a 95% confidence interval for $g(p,q)$ given by

$$g(\hat{p},\hat{q}) \pm Z_{0.025}\sqrt{\hat{v}} = \log\left(\frac{\hat{p}}{1-\hat{p}}\bigg/\frac{\hat{q}}{1-\hat{q}}\right) \pm Z_{0.025}\sqrt{\frac{1}{n\hat{p}(1-\hat{p})} + \frac{1}{m\hat{q}(1-\hat{q})}}.$$

Denoting this interval by $[L(\hat{p},\hat{q}), U(\hat{p},\hat{q})]$, we may exponentiate to obtain the confidence interval $\left[e^{L(\hat{p},\hat{q})}, e^{U(\hat{p},\hat{q})}\right]$ for the odds-ratio $\frac{p}{1-p}\Big/\frac{q}{1-q}$.

Problem 14. (Laplace distribution and Bayesian Analysis)

The double-exponential distribution with mean μ and scale b is a continuous distribution over $\mathbb{R}$ with PDF

$$f(x \mid \mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right).$$

It is sometimes used as an alternative to the normal distribution to model data with heavier tails, as this PDF decays exponentially in $|x-\mu|$ rather than in $(x-\mu)^2$.

(a) What are the MLEs $\hat{\mu}$ and $\hat{b}$ given data $X_1, \ldots, X_n$? Why is this MLE $\hat{\mu}$ more robust to outliers than the MLE $\hat{\mu}$ in the $\mathcal{N}\left(\mu, \sigma^2\right)$ model?

You may assume that n is odd and that the data values $X_1, \ldots, X_n$ are all distinct. (Hint: The log-likelihood is differentiable in b but not in μ. To find the MLE $\hat{\mu}$, you will need to reason directly from its definition.)

Solution. The joint log-likelihood is

$$\ell(\mu, b) = -n\log(2b) - \frac{1}{b}\sum_{i=1}^{n}|X_i - \mu|.$$

The likelihood is differentiable in b, so differentiating with respect to b gives

$$\frac{\partial \ell}{\partial b} = -\frac{n}{b} + \frac{1}{b^2}\sum_{i=1}^{n}|X_i - \mu|.$$

Setting this equal to 0 , substituting in the MLE $\hat{\mu}$ for μ, and solving gives the MLE for b as

$$\hat{b} = \frac{1}{n} \sum_{i=1}^{n} \left| X_i - \hat{\mu} \right| .$$

We can see that the MLE $\hat{\mu}$ is the value of μ that minimizes the total absolute deviations $K(\mu) = \sum_{i=1}^{n} |X_i - \mu|$. Without loss of generality assume that the $X_1, \ldots, X_n$ are ordered. We shall see that the minimizer is the sample median $\hat{\mu} = X_m$, where $m = (n+1)/2$. We see that $K(\mu)$ is continuous everywhere (it is the sum of absolute value functions), and furthermore, it is decreasing for $\mu < X_m$ and increasing for $\mu > X_m$.

Therefore the minimizer is given by $\hat{\mu} = X_m$. This estimator is more robust to outliers because it only depends on the middle few ordered values, so a few data points with extreme values won't change the median, whereas the mean depends on all data points.

(b) Suppose it is known that $\mu = 0$. In a Bayesian analysis, let us model the scale parameter as a random variable B with prior distribution $B \sim \text{InverseGamma}(\alpha, \beta)$, where $\alpha, \beta > 0$. If $X_1, \ldots, X_n \sim$ Laplace $(0, b)$ when $B = b$, what are the posterior distribution and posterior mean of B given the data $X_1, \ldots, X_n$?
(Hints: The InverseGamma (α, β) distribution is a continuous distribution on $(0, \infty)$ with PDF

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$$

and with mean $\frac{\beta}{\alpha-1}$ when $\alpha > 1$.)

Solution. If $\mu = 0$ and $B \sim \text{InverseGamma}(\alpha, \beta)$, then the posterior density is given by

$$\begin{aligned} f\left(B \mid \alpha, \beta, X_1, \ldots, X_n\right) &\propto f\left(X_1, \ldots, X_n \mid B\right) f(B \mid \alpha, \beta) \\ &= \frac{1}{(2B)^n} \exp\left\{-\frac{1}{B} \sum_{i=1}^{n} |X_i|\right\} \frac{\beta^\alpha}{\Gamma(\alpha)} B^{-\alpha-1} e^{-\beta/B} \\ &\propto B^{-(\alpha+n)-1} \exp\left\{-\frac{1}{B}\left(\beta + \sum_{i=1}^{n} |X_i|\right)\right\}, \end{aligned}$$

where we have dropped any normalizing constants into the proportionality term. From here, we can see that the posterior distribution of B follows an InverseGamma $(\alpha + n, \beta + \sum |X_i|)$ distribution, and therefore has posterior mean $\left(\beta + \sum |X_i|\right) / (\alpha + n - 1)$.

(c) Still supposing it is known that $\mu = 0$, what is the MLE $\hat{b}$ for b in this sub-model? How does this compare to the posterior mean from part (b) when n is large?

Solution. The MLE for b when $\mu = 0$ is

$$\hat{b} = \frac{1}{n} \sum_{i=1}^{n} |X_i| .$$

We can write the posterior mean as a weighted average

$$\underbrace{\frac{\beta + \sum_{i=1}^{n} |X_i|}{\alpha + n - 1}}_{\text{posterior mean}} = \frac{\alpha - 1}{\alpha + n - 1} \underbrace{\frac{\beta}{\alpha - 1}}_{\text{prior mean}} + \frac{n}{\alpha + n - 1} \underbrace{\frac{1}{n} \sum_{i=1}^{n} |X_i|}_{\text{MLE}}$$

of the prior mean and the MLE, from which we see that the posterior mean tends to the MLE as $n \to \infty$.

Problem 15. (Bayesian inference for Multinomial Proportions)

The Dirichlet $(\alpha_1, \dots, \alpha_K)$ distribution with parameters $\alpha_1, \dots, \alpha_K > 0$ is a continuous joint distribution over K random variables $(P_1, \dots, P_K)$ such that $0 \le P_i \le 1$ for all $i = 1, \dots, K$ and $P_1 + \dots + P_K = 1$. It has (joint) PDF

$$f\left(p_1, \dots, p_k \mid \alpha_1, \dots, \alpha_K\right) \propto p_1^{\alpha_1 - 1} \times \dots \times p_K^{\alpha_K - 1}.$$

Letting $\alpha_0 = \alpha_1 + \dots + \alpha_K$, this distribution satisfies

$$\mathbb{E}\left[P_i\right] = \frac{\alpha_i}{\alpha_0}, \quad \operatorname{Var}\left[P_i\right] = \frac{\alpha_i\left(\alpha_0 - \alpha_i\right)}{\alpha_0^2\left(\alpha_0 + 1\right)}.$$

(a) Let $(X_1, \dots, X_6) \sim \text{Multinomial}\left(n, (p_1, \dots, p_6)\right)$ be the numbers of 1's through 6's obtained in n rolls of a (possibly biased) die. Let us model $(P_1, \dots, P_6)$ as random variables with prior distribution Dirichlet $(\alpha_1, \dots, \alpha_6)$. What is the posterior distribution of $(P_1, \dots, P_6)$ given the observations $(X_1, \dots, X_6)$? What is the posterior mean and variance of P_1?

(b) How might you choose the prior parameters $\alpha_1, \dots, \alpha_6$ to represent a strong prior belief that the die is close to fair (meaning $p_1, \dots, p_6$ are all close to $1/6$)?

(c) How might you choose an improper Dirichlet prior to represent no prior information? How do the posterior mean estimates of $p_1, \dots, p_6$ under this improper prior compare to the MLE?

Solution.

(a) The posterior distribution has density proportional to

$$P_1^{\alpha_1 - 1} \times \dots \times P_6^{\alpha_6 - 1} \times P_1^{X_1} \times \dots \times P_6^{X_6} = P_1^{\alpha_1 + X_1 - 1} \times \dots \times P_6^{\alpha_6 + X_6 - 1}.$$

So the posterior distribution of $(P_1, \dots, P_6)$ given $(X_1, \dots, X_6)$ is Dirichlet$(\alpha_1 + \ X_1, \dots, \alpha_6 + X_6)$. The posterior mean and variance are given by

$$\mathbf{E}\left[P_i \mid X_1, \dots, X_6\right] = \frac{\alpha_i + X_i}{\alpha_0 + n\bar{X}} \; ; \quad \mathbf{V}\left[P_i \mid X_1, \dots, X_6\right] = \frac{\left(\alpha_i + X_i\right)\left(\alpha_0 + n\bar{X} - \alpha_i - X_i\right)}{\left(\alpha_0 + n\bar{X}\right)^2\left(\alpha_0 + n\bar{X} + 1\right)}.$$

(b) We would like to select the parameters α_i such that

- The prior mean is $1/6$ for each i,
- The prior variance is small.

Since

$$\mathbf{E}\left[P_i\right] = \frac{\alpha_i}{\sum_{j=1}^{6} \alpha_j},$$

a prior mean of $1/6$ can be achieved by setting $\alpha_i = \alpha$ for each i. Then the variance is given by

$$\mathbf{V}\left[P_i\right] = \frac{\alpha(6\alpha - \alpha)}{(6\alpha)^2(\alpha + 1)} = \frac{5}{36(\alpha + 1)},$$

from which we see that a large value of α achieves a small variance. (The stronger our prior belief that the die is fair, the larger we would set α.)

(c) From the posterior mean calculated in part (a), we can interpret the parameters α_i as "prior counts," so an uninformative prior sets $\alpha_i = 0$. Then the posterior mean is

$$\mathbf{E}\left[P_i \mid X_1, \dots, X_6\right] = \frac{X_i}{\sum_{j=1}^{n} X_j} = \frac{X_i}{n},$$

which is the same as the MLE.

Problem 16. (Basics of Testing)

(a) Let Z be a random variable with probability density function $f(z) = \frac{1}{2}e^{-|z-\mu|}, z \in \mathbb{R}$ with parameter $\mu \in \mathbb{R}$. Suppose, we observe $X = \max(0, Z)$.

i. Find the constant c such that the test that "rejects when $X > c$" has size 0.05 for the null hypothesis $H_0 : \mu = 0$.

ii. Find the power of this test against the alternative hypothesis $H_1 : \mu = 2$.

(b) Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, and that the following hypotheses are to be tested:

$$\begin{aligned} H_0 : &\quad \theta \geq 2, \\ H_1 : &\quad \theta < 2. \end{aligned}$$

Let $Y_n = \max\{X_1, \ldots, X_n\}$, and consider a test procedure such that the critical region contains all the outcomes for which $Y_n \leq 1.5$.

i. Determine the power function of the test.

ii. Determine the size of the test.

(c) Let X be a single observation of an $\text{Exp}(\lambda)$ random variable, which has PDF

$$f_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Consider testing $H_0 : \lambda \geq \lambda_0$ versus $H_1 : \lambda < \lambda_0$.

i. Find the power function of the hypothesis test that rejects H_0 if and only if $X \geq c$.

ii. Let $0 < \alpha < 1$. Find a value of c such that the test in part (i) has size α.

iii. For what true values of λ is P_λ (type II error) $\geq 1/2$ for the test in part (i) with size α as in (ii)?

(d) Let $X_1, X_2 \overset{\text{IID}}{\sim} \text{Bin}(1, \theta)$, and consider testing $H_0 : \theta = 1/3$ versus $H_1 : \theta < 1/3$.

i. Find a test that has size 2/9 exactly. Note: It does not have to be a sensible test.

ii. Find the power function of the test from part (i), and use it to explain why this test is not a good test of these hypotheses.

(e) Suppose X is a random variable on $\{0, 1, 2, \ldots\}$ with unknown PMF $p(x)$. To test the hypothesis $H_0 : X \sim \text{Poisson}(1/2)$ against $H_1 : p(x) = 2^{-(x+1)}$ for all $x \in \{0, 1, 2, \ldots\}$, we reject H_0 if $x > 2$. The probability of type-II error for this test is

(A) $\frac{1}{4}$; (B) $1 - \frac{13}{8}e^{-1/2}$; (C) $1 - \frac{3}{2}e^{-1/2}$; (D) $\frac{7}{8}$.

Solution.

(a) i. Given that $P_{H_0}(X > c) = 0.05$. Now, under H_0, $\mu = 0$. So, we have the pdf of Z as $f(z) = \frac{1}{2}e^{-|z|}$. As the support of Z is $\mathbb{R}$, we can partition it in $\{Z \geq 0, Z < 0\}$. Now, let's condition based on this partition. So, we have:

$$P_{H_0}(X > c) = P_{H_0}(X > c, Z \geq 0) + P_{H_0}(X > c, Z < 0) = P_{H_0}(X > c, Z \geq 0) = P_{H_0}(Z > c).$$

$$\text{Thus, we have } P_{H_0}(X > c) = P_{H_0}(Z > c) = \int_c^\infty \frac{1}{2}e^{-|z|}dz = \frac{1}{2}e^{-c}.$$

Equating $\frac{1}{2}e^{-c}$ with 0.05, we get $c = \ln 10$.

ii. Power of test against H_1 is given by:

$$P_{H_1}(X > \ln 10) = P_{H_1}(Z > \ln 10) = \int_{\ln 10}^{\infty} \frac{1}{2} e^{-|z-2|} dz = \frac{e^2}{20}.$$

(b) i. The power function of the test is

$$\text{Power}(\theta) = P_\theta\left(Y_n \leq 1.5\right) = P_\theta\left(\max_{1\leq i\leq n} X_i \leq 1.5\right) = \prod_{i=1}^{n} P_\theta\left(X_i \leq 1.5\right) = \left[P_\theta\left(X_1 \leq 1.5\right)\right]^n = \left(\frac{1.5}{\theta}\right)^n.$$

for $\theta \geq 1.5$, and $\text{Power}(\theta) = 1$ for $\theta < 1.5$.

ii. $\text{Power}(\theta)$ is a non-increasing function of θ, so

$$\sup_{\theta \geq 2} \text{Power}(\theta) = \text{Power}(2) = \left(\frac{1.5}{2}\right)^n = \left(\frac{3}{4}\right)^n.$$

Thus, the size of the test is $(3/4)^n$.

(c) i. $\text{Power}(\lambda) = P_\lambda(X \geq c) = \int_c^\infty f_\lambda(x)dx = \exp(-\lambda c)$.

ii. $\text{Power}(\lambda)$ is a non-increasing function of λ, so

$$\sup_{\lambda \geq \lambda_0} \text{Power}(\lambda) = \text{Power}\left(\lambda_0\right) = \exp\left(-\lambda_0 c\right).$$

Thus, the size of the test is $\exp\left(-\lambda_0 c\right)$. Then the test has size α if and only if

$$\exp\left(-\lambda_0 c\right) = \alpha \iff c = -\frac{\log \alpha}{\lambda_0}$$

noting that $\log \alpha$ is negative since $0 < \alpha < 1$.

iii. P_λ (type II error) $\geq 1/2$ if and only if both $\lambda < \lambda_0$ and $\text{Power}(\lambda) \leq 1/2$. The test in part (i) with size α as in (ii) has power function

$$\text{Power}(\lambda) = \exp\left[-\lambda\left(-\frac{\log \alpha}{\lambda_0}\right)\right] = \alpha^{\lambda/\lambda_0},$$

and hence

$$\text{Power}(\lambda) \leq 1/2 \iff \lambda \geq -\frac{\lambda_0 \log 2}{\log \alpha},$$

again noting that $\log \alpha$ is negative. Thus, P_λ (type II error) $\geq 1/2$ if and only if

$$-\frac{\lambda_0 \log 2}{\log \alpha} \leq \lambda < \lambda_0.$$

(Note that if $\alpha \geq 1/2$, then there are no such values of λ.)

(d) i. Note that there are only four possible values of (X_1, X_2), i.e., the sample space consists of only four points. If $\theta = 1/3$, then

$$(X_1, X_2) = \begin{cases} (0,0) & \text{with probability } 4/9 \\ (0,1) & \text{with probability } 2/9 \\ (1,0) & \text{with probability } 2/9 \\ (1,1) & \text{with probability } 1/9 \end{cases}$$

Thus, the only tests with size $2/9$ exactly are the test that rejects H_0 if and only if $(X_1, X_2) = (0,1)$ and the test that rejects H_0 if and only if $(X_1, X_2) = (1,0)$.

ii. $\text{Power}(\theta) = \theta(1-\theta)$ for both of the tests from part (i). Note that $\text{Power}(1/3) > \text{Power}(\theta)$ for all $\theta < 1/3$. Thus, these tests are more likely to reject H_0 if it is true than if it is false, which is exactly the opposite of what a good hypothesis test should do.

(e) (D) Given $x \sim p(x)$. Reject H_0 if $x > 2$.
P(Type-II error) = P(Accepting H_0 when it is false) = 1 - P(Reject H_O | H_1 is true)
= 1 - P_{H_1}(Reject H_0) = 1 - $P_{H_1}(X > 2)$. Now, we compute $P_{H_1}(X > 2)$ as follows:

$$\begin{aligned} P(x > 2) &= 1 - P(x \leq 2) \\ &= 1 - [P(x=0) + P(x=1) + P(x=2)] \\ &= 1 - \left[2^{-1} + 2^{-2} + 2^{-3}\right] \\ &= 1 - \frac{7}{8} \\ &= \frac{1}{8}. \end{aligned}$$

Problem 17. (Likelihood ratio tests)

(a) For data $X_1, \ldots, X_n \in \mathbb{R}$ and two fixed and known values $\sigma_0^2 < \sigma_1^2$, consider the following testing problem:

$$\begin{aligned} H_0 &: X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}\left(0, \sigma_0^2\right) \\ H_1 &: X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}\left(0, \sigma_1^2\right) \end{aligned}$$

What is the most powerful test for testing H_0 versus H_1 at level α? Letting $\chi_n^2(\alpha)$ denote the $1-\alpha$ quantile of the χ_n^2 distribution explicitly describes the test statistic T and the rejection region for this test.

(b) What is the distribution of this test statistic T under the alternative hypothesis H_1? Using this result and letting F denote the CDF of the χ_n^2 distribution, provide a formula for the power of this test against H_1 in terms of $\chi_n^2(\alpha), \sigma_0^2, \sigma_1^2$, and F. Keeping σ_0^2 fixed, what happens to the power of the test as σ_1^2 increases to ∞ ?

(c) Consider two probability density functions on $[0,1]: f_0(x) = 1$ and $f_1(x) = 2x$. Among all tests of the null hypothesis $H_0 : X \sim f_0(x)$ versus the alternative $H_1 : X \sim f_1(x)$ with significance level $\alpha = 0.10$, how large can the power possibly be?

(d) Let X_1 and X_2 be a random sample from a distribution having the probability density function $f(x)$. Consider the testing of $H_0 : f(x) = f_0(x)$ against $H_1 : f(x) = f_1(x)$ based on X_1 and X_2, where

$$f_0(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_1(x) = \begin{cases} 4x^3, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases}$$

(i) For a given level, show that the critical region of the most powerful test for testing H_0 against H_1 is of the form

$$\{(x_1, x_2) : \ln x_1 + \ln x_2 > c\}$$

for some constant c.
(ii) Determine c in terms of a suitable cutoff point of a Chi-square distribution when the level is α.

(e) Let $X_1, \ldots, X_n$ be i.i.d. observation from the density,

$$f(x) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right), x > 0$$

where $\mu > 0$ is an unknown parameter. Consider the problem of testing the hypothesis $H_o : \mu \leq \mu_o$ against $H_1 : \mu > \mu_o$.

Show that the test with critical region $\left[\bar{X} \geq \mu_o \chi^2_{2n,1-\alpha}/2n\right]$, where $\chi^2_{2n,1-\alpha}$ is the $(1-\alpha)^{th}$ quantile of the χ^2_{2n} distribution, has size α. Give an expression of the power in terms of the c.d.f. of the χ^2_{2n} distribution.

Solution.

(a) The joint PDF under H_0 is

$$f_0\left(x_1, \ldots, x_n\right)=\left(\frac{1}{\sqrt{2 \pi \sigma_0^2}}\right)^n \exp \left(-\frac{\sum_{i=1}^n x_i^2}{2 \sigma_0^2}\right)$$

The joint PDF under H_1 is

$$f_1\left(x_1, \ldots, x_n\right)=\left(\frac{1}{\sqrt{2 \pi \sigma_1^2}}\right)^n \exp \left(-\frac{\sum_{i=1}^n x_i^2}{2 \sigma_1^2}\right)$$

So the likelihood ratio statistic is

$$L\left(X_1, \ldots, X_n\right)=\frac{f_0\left(X_1, \ldots, X_n\right)}{f_1\left(X_1, \ldots, X_n\right)}=\left(\frac{\sigma_1}{\sigma_0}\right)^n \exp \left(\frac{\sigma_0^2-\sigma_1^2}{2 \sigma_0^2 \sigma_1^2} \sum_{i=1}^n x_i^2\right)$$

Since $\sigma_0^2<\sigma_1^2$, L is a decreasing function of $T:=\sum_{i=1}^n X_i^2$. Then rejecting for small values of L is the same as rejecting for large values of T.

Since under H_0, $\sum_{i=1}^n\left(\frac{X_i}{\sigma_0}\right)^2 \sim \chi_n^2$, we have $\frac{1}{\sigma_0^2} T \sim \chi_n^2$, so $T \sim \sigma_0^2 \chi_n^2$. Then the rejection threshold should be $c=\sigma_0^2 \chi_n^2(\alpha)$, and the most powerful test rejects H_0 when $T>c$.

(b) Under $H_1, \sum_{i=1}^n\left(\frac{X_i}{\sigma_1}\right)^2 \sim \chi_n^2$, so $T \sim \sigma_1^2 \chi_n^2$. Then the probability of type II error is

$$\begin{aligned}
\beta=\mathbb{P}_{H_1}\left[\text{accept } H_0\right] & =\mathbb{P}_{H_1}\left[T \leq \sigma_0^2 \chi_n^2(\alpha)\right] \\
& =\mathbb{P}_{H_1}\left[\frac{T}{\sigma_1^2} \leq \frac{\sigma_0^2}{\sigma_1^2} \chi_n^2(\alpha)\right]=F\left(\frac{\sigma_0^2}{\sigma_1^2} \chi_n^2(\alpha)\right) .
\end{aligned}$$

where F is the χ_n^2 CDF. The power of the test is then

$$\text{Power }=1-\beta=1-\mathrm{F}\left(\frac{\sigma_0^2}{\sigma_1^2} \chi_{\mathrm{n}}^2(\alpha)\right)$$

As $\sigma_1^2 \rightarrow \infty$, $\beta \rightarrow F(0)=0$ and the power of the test $\rightarrow 1$.

(c) The likelihood ratio statistic is

$$L(X)=\frac{f_0(X)}{f_1(X)}=\frac{1}{2 X}$$

The condition $L(X)<c$ is then equivalent to $X>\tilde{c}$, where $\tilde{c}=\frac{1}{2 c}$. Under the hypothesis $H_0, X \sim \operatorname{Uniform}(0,1)$, so the rejection threshold $\tilde{c}$ should be $1-0.1=0.9$, i.e. the most powerful tests rejects H_0 when $X>0.9$. Under the hypothesis $H_1, X \sim f_1(x)=2 x$. Then the type II error probability is

$$\beta=\mathbb{P}_{H_1}\left[\text{ accept } H_0\right]=\mathbb{P}_{H_1}[X \leq 0.9]=\int_0^{0.9} 2 x d x=0.81 .$$

Thus the power of the test is

$$\text{Power }=1-\beta=0.19 .$$

This is the maximum power that can be achieved: According to the Neyman-Pearson lemma, for any other test of H_0 with a significance level at most 0.1, its power against H_1 is at most 0.19.

(d) (i) By NP Lemma, the Most Powerful Critical Region is given by

$$\left\{\underset{\sim}{x} \mid \frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})}>k\right\}, \text{ where } k \text{ is a constant that depends on level requirements.}$$

$$\text{Now,}\quad \frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} = \frac{\prod_{i=1}^{2} 4x_i^3}{\prod_{i=1}^{2} 1} \Rightarrow\ 16(x_1x_2)^3 > k$$

$$\Rightarrow\ (x_1x_2)^3 > k_1 \Rightarrow\ 3\ln(x_1x_2) > k_2 \Rightarrow\ \ln x_1 + \ln x_2 > c.$$

$\therefore\ W = \{(x_1, x_2) \mid \ln x_1 + \ln x_2 > c\}$ for some constant c.

(ii) Given level α,

$$\mathrm{P_{H_0}}(\text{Reject } H_0) = \alpha \Rightarrow \mathrm{P_{H_0}}(\ln x_1 + \ln x_2 > c) = \alpha.$$

Under $H_0,\ X_1, X_2 \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$,

$$\begin{aligned}
&\therefore\ -2\ln X_1 \sim \chi^2_{(2)}\\
&\therefore\ -2\ln X_1 - 2\ln X_2 \sim \chi^2_{(4)}\\
&\therefore\ \mathrm{P_{H_0}}(-2\ln X_1 - 2\ln X_2 < -2c) = \alpha\\
&\Rightarrow \mathrm{P_{H_0}}(\chi^2_{(4)} < -2c) = \alpha\\
&\therefore\ -2c = \chi^2_{4;\alpha}\\
&\Rightarrow c = -\frac{1}{2}\chi^2_{4;\alpha}.
\end{aligned}$$

$\therefore\ c = -\frac{1}{2}\chi^2_{4;\alpha}$ is the required value of c.

(e) Hence, the Likelihood function of the μ for the given sample is, $L(\mu \mid \vec{X}) = \left(\frac{1}{\mu}\right)^n \exp\left(-\frac{\sum_{i=1}^n X_i}{\mu}\right), \mu > 0$, also observe that sample mean of $\vec{X}$ is the MLE of μ. So, the Likelihood Ratio statistic is,

$$\begin{aligned}
\lambda(\vec{x}) &= \frac{\sup_{\mu \le \mu_o} L(\mu \mid \vec{x})}{\sup_{\mu} L(\mu \mid \vec{x})}\\
&= \begin{cases} 1 & \mu_0 \ge \bar{X} \\ \frac{L(\mu_0 \mid \vec{x})}{L(\bar{X} \mid \vec{x})} & \mu_0 < \bar{X} \end{cases}
\end{aligned}$$

So, our test function is,

$$\phi(\vec{x}) = \begin{cases} 1 & \lambda(\vec{x}) < k \\ 0 & \text{otherwise} \end{cases}$$

We, reject H_0 at size α, when $\phi(\vec{x}) = 1$, for some k, $\mathrm{E}_{H_0}(\phi) \le \alpha$. Hence, $\lambda(\vec{x}) < k$

$$\begin{aligned}
&\Rightarrow L(\mu_0 \mid \vec{x}) < kL(\bar{X} \mid \vec{x})\\
&\ln k_1 - \frac{1}{\mu_0}\sum_{i=1}^{n} X_i < \ln k - n\ln\bar{X} - \frac{1}{n}\\
&n\ln\bar{X} - \frac{n\bar{X}}{\mu_0} < K*
\end{aligned}$$

for some constant, $K*$.

Let $g(\bar{x}) = n\ln\bar{x} - \frac{n\bar{x}}{\mu_0}$, and observe that g is, decreasing function of $\bar{x}$ for $\bar{x} \ge \mu_o$ Hence, there exists a c such that $\bar{x} \ge c$, we have $g(\bar{x}) < K*$. So, the critical region of the test is of form $\bar{X} \ge c$, for some c such that, $P_{H_o}(\bar{X} \ge c) = \alpha$, for some $0 \le \alpha \le 1$, where α is the size of the test. Now, our task is to find c, and for that observe, if $X \sim$ Exponential(θ), then $\frac{2X}{\theta} \sim \chi^2_2$.
Hence, in this problem, since the X_i 's follows Exponential (μ), hence, $\frac{2n\bar{X}}{\mu} \sim \chi^2_{2n}$, we have,

$$\begin{aligned}
&P_{H_o}(\bar{X} \ge c) = \alpha\\
&P_{H_o}\left(\frac{2n\bar{X}}{\mu_o} \ge \frac{2nc}{\mu_o}\right) = \alpha\\
&P_{H_o}\left(\chi^2 2n \ge \frac{2nc}{\mu_o}\right) = \alpha
\end{aligned}$$

which gives $c = \frac{\mu_o \chi^2_{2n;1-\alpha}}{2n}$, Hence, the rejection region is indeed, $\left[\bar{X} \geq \frac{\mu_o \chi^2_{2n;1-\alpha}}{2n}\right]$. Hence Proved!

Now, we know that the power of the test is,

$$\beta = E_\mu(\phi) = P_\mu(\lambda(\bar{x}) > k) = P\left(\bar{X} \geq \frac{\mu_o \chi^2_{2n;1-\alpha}}{2n}\right)$$
$$\beta = P_\mu\left(\chi^2_{2n} \geq \frac{\mu_o}{\mu}\chi^2_{2n;1-\alpha}\right)$$

Hence, the power of the test is of the form of a cdf of the chi-squared distribution.

Problem 18. (MP tests and GLRT test)

(a) Let $X_1, \ldots, X_{10}$ be a random sample of size 10 from a population having a probability density function

$$f(x \mid \theta) = \begin{cases} \frac{\theta}{x^{\theta+1}}, & \text{if } x > 1, \\ 0, & \text{otherwise} \end{cases}$$

where $\theta > 0$. For testing $H_0 : \theta = 2$ against $H_1 : \theta = 4$ at the level of significance $\alpha = 0.05$, find the most powerful test. Also, find the power of this test.

(b) Let $X_1, \ldots, X_n$ be a random sample from the population having probability density function

$$f(x, \theta) = \begin{cases} \frac{2x}{\theta^2} e^{-\frac{x^2}{\theta^2}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Obtain the most powerful test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ $(\theta_1 < \theta_0)$.

(c) Let $X_1, X_2, \ldots, X_5$ be a random sample from a $N\left(2, \sigma^2\right)$ distribution, where σ^2 is unknown. Derive the most powerful test of size $\alpha = 0.05$ for testing $H_0 : \sigma^2 = 4$ against $H_1 : \sigma^2 = 1$.

(d) Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where both μ and σ^2 are unknown. Consider the problem of testing

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0$$

Show that the generalized likelihood ratio test statistic for this problem simplifies to

$$\Lambda\left(X_1, \ldots, X_n\right) = \left(\frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2 + n\bar{X}^2}\right)^{n/2}.$$

Letting $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2$ and $T = \sqrt{n}\bar{X}/S_X$ (the usual one-sample t-statistic for this problem), show that $\Lambda\left(X_1, \ldots, X_n\right)$ is a monotonically decreasing function of $|T|$, and hence the generalized likelihood ratio test is equivalent to the two-sided t-test which rejects for large values of $|T|$.

Solution.

(a) By NP Lemma, the Most Powerful Critical Region is given by reject H_0 if $\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} > k$ where k is a constant that depends on level requirements.

$$\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} = \frac{\prod_{i=1}^{10} \frac{4}{x_i^5}}{\prod_{i=1}^{10} \frac{2}{x_i^3}} = \frac{\frac{4^{10}}{(\prod_{i=1}^{10} x_i)^5}}{\frac{2^{10}}{(\prod_{i=1}^{10} x_i)^3}} = \frac{2^{10}}{(\prod_{i=1}^{10} x_i)^2} > k$$
$$\Rightarrow \prod_{i=1}^{10} x_i < k' \quad \Rightarrow \sum_{i=1}^{10} \ln x_i < c.$$

Now, level $\alpha = 0.05$, we have $\mathrm{P}_{\mathrm{H}_0}(\text{Reject } H_0) = 0.05 \Rightarrow \mathrm{P}_{\mathrm{H}_0}(\sum_{i=1}^{10} \ln x_i < c) = 0.05$.

If X_i has the given pdf then $\ln X_i \sim \mathrm{Exp}\left(\frac{1}{theta}\right)$. Under H_0,

$$\ln X_i \sim \mathrm{Exp}\left(\frac{1}{2}\right)$$
$$\therefore 4 \ln X_i \sim \mathrm{Exp}\,(2)$$
$$\therefore \mathrm{P}\left(\sum_{i=1}^{10} 4 \ln X_i < 4c\right) = 0.05 \Rightarrow \quad \mathrm{P}\left(\chi^2_{20} < 4c\right) = 0.05 \Rightarrow \mathrm{P}\left(\chi^2_{20} > 4c\right) = 0.95$$
$$\therefore 4c = \chi^2_{20,0.95}$$
$$\Rightarrow c = \frac{1}{4}\chi^2_{20,0.95} = \frac{10.85}{4} = 2.7125.$$

$\therefore$ Reject H_0 if $\sum_{i=1}^{10} \ln X_i < 2.7125$.

Power of the test, $\mathrm{P}_{H_1}(\text{Reject } H_0) = \mathrm{P}_{H_1}(\sum_{i=1}^{10} \ln X_i < 2.7125)$.
Under H_1,

$$\ln X_i \sim \mathrm{Exp}\left(\frac{1}{4}\right)$$
$$\therefore 8 \ln X_i \sim \mathrm{Exp}\,(2)$$
$$\therefore \mathrm{P}_{H_1}\left(\sum_{i=1}^{10} 8 \ln X_i < 8(2.7125)\right) = \mathrm{P}\left(\chi^2_{20} < 21.7\right) = 1 - \mathrm{P}\left(\chi^2_{20} > 21.7\right) = 1 - 0.36 = 0.64.$$

(b) By NP Lemma, the Most Powerful Critical Region is given by reject H_0 if $\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} > k$ where k is a constant that depends on level requirements.

$$\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} = \frac{\prod_{i=1}^{n} \frac{2x_i}{\theta_1} e^{-\frac{x_i^2}{\theta_1^2}}}{\prod_{i=1}^{n} \frac{2x_i}{\theta_0} e^{-\frac{x_i^2}{\theta_0^2}}} = \frac{2^n \frac{\prod_{i=1}^{n} x_i}{\theta_1^n} e^{-\sum_{i=1}^{n} \frac{x_i^2}{\theta_1^2}}}{2^n \frac{\prod_{i=1}^{n} x_i}{\theta_0^n} e^{-\sum_{i=1}^{n} \frac{x_i^2}{\theta_0^2}}}$$
$$\Rightarrow \left(\frac{\theta_0}{\theta_1}\right)^n e^{-\sum_{i=1}^{n} x_i^2 \left(\frac{1}{\theta_1^2} - \frac{1}{\theta_0^2}\right)} > k$$
$$\Rightarrow -\sum_{i=1}^{n} x_i^2 \left(\frac{1}{\theta_1^2} - \frac{1}{\theta_0^2}\right) > k'$$
$$\Rightarrow \sum_{i=1}^{n} x_i^2 < c \left[\begin{array}{c} \because \theta_0 < \theta_1 \\ \therefore \frac{1}{\theta_1^2} > \frac{1}{\theta_0^2} \\ \therefore \frac{1}{\theta_1^2} - \frac{1}{\theta_0^2} > 0 \end{array}\right]$$

Now, if X has the given distribution then $X^2 \sim \mathrm{Exp}\left(mean = theta^2\right)$

$$\therefore \frac{2X^2}{\theta^2} \sim \mathrm{Exp}\,(mean = 2)\,.$$

Under H_0,

$$\frac{2X_i^2}{\theta_0^2} \sim \text{Exp}\,(mean = 2)$$
$$\therefore \sum_{i=1}^{n} \frac{2X_i^2}{\theta_0^2} \sim \chi^2_{2n}$$

Let the test be a level α test

$$\therefore \text{P}_{H_0}\left(\sum_{i=1}^{n} X_i^2 < c\right) = \alpha$$
$$\Rightarrow \text{P}_{H_0}\left(\sum_{i=1}^{n} \frac{2X_i^2}{\theta_0^2} < \frac{2c}{\theta_0^2}\right) = \alpha$$
$$\Rightarrow \text{P}_{H_0}\left(\chi^2_{2n} < \frac{2c}{\theta_0^2}\right) = \alpha$$
$$\Rightarrow \text{P}_{H_0}\left(\chi^2_{2n} > \frac{2c}{\theta_0^2}\right) = 1 - \alpha$$
$$\therefore \frac{2c}{\theta_0^2} = \chi^2_{2n,1-\alpha}$$
$$\therefore c = \frac{\theta_0^2}{2}\,\chi^2_{2n,1-\alpha}\,.$$

$\therefore$ Reject H_0 if $\sum_{i=1}^{n} X_i^2 < \frac{\theta_0^2}{2}\,\chi^2_{2n,1-\alpha}\,.$

(c) Given $X_1, X_2, \ldots, X_5 \sim \text{N}(2, \sigma^2)$.
To test,

$$H_0 : \sigma^2 = 4 \text{ vs } H_1 : \sigma^2 = 1.$$

By NP Lemma, the Most Powerful Critical Region is given by W $= \left\{\underset{\sim}{x} \mid \frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} > k\right\}$ where $k \in$ R such that the test is of size α.

$$\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} = \frac{\prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-2)^2}}{\prod_{i=1}^{5} \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x_i-2)^2}{2}}} > k$$
$$\Rightarrow \frac{2^5 e^{-\frac{1}{2}\sum(x_i-2)^2}}{e^{-\frac{1}{2}\sum\frac{(x_i-2)^2}{2}}} > k \Rightarrow\; 2^5 e^{-\frac{1}{4}\sum(x_i-2)^2} > k$$
$$\Rightarrow \sum_{i=1}^{5} (x_i - 2)^2 < c.$$

$\therefore$ Reject H_0 if $\sum_{i=1}^{5} (x_i - 2)^2 < c$.
Now, $\text{P}_{H_0}\left(\sum_{i=1}^{5} (x_i - 2)^2 < c\right) = 0.05$. Under H_0,

$$X_1, X_2, \ldots, X_5 \overset{\text{IID}}{\sim} \text{N}(2, 4)$$
$$\therefore \sum_{i=1}^{5} \left(\frac{x_i - 2}{2}\right)^2 \sim \chi^2_5$$
$$\therefore \text{P}_{H_0}\left(\frac{\sum_{i=1}^{5} (x_i - 2)^2}{4} < \frac{c}{4}\right) = 0.05$$
$$\Rightarrow \text{P}_{H_0}\left(\chi^2_5 < \frac{c}{4}\right) = 0.05$$
$$\Rightarrow \text{P}_{H_0}\left(\chi^2_5 > \frac{c}{4}\right) = 0.95$$
$$\Rightarrow \frac{c}{4} = \chi^2_{5,0.95} \Rightarrow\; c = 4\,\chi^2_{5,0.95}.$$

$\therefore$ The MP test of size $\alpha = 0.05$ is reject H_0 if $\sum_{i=1}^{5} (x_i - 2)^2 < 4\,\chi^2_{5,0.95}$.

(d) The log-likelihood for the full model is

$$-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right)^2,$$

and the MLEs for μ and σ are

$$\hat{\mu}=\bar{X} \quad \text{and} \quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)^2.$$

Under the submodel defined by $\mu=0$, the log-likelihood is

$$-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}X_i^2$$

and the MLE for σ^2 is

$$\tilde{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2.$$

Therefore, the GLRT statistic is given by

$$\begin{aligned}
\frac{\sup_{\sigma^2}\ell\left(0,\sigma^2\right)}{\sup_{\mu,\sigma^2}\ell\left(\mu,\sigma^2\right)} &= \frac{\left(2\pi\tilde{\sigma}^2\right)^{-n/2}\exp\left\{-\frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}X_i^2\right\}}{\left(2\pi\hat{\sigma}^2\right)^{-n/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)^2\right\}} = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sum_{i=1}^{n}X_i^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2+n\bar{X}^2}\right)^{n/2}.
\end{aligned}$$

We can rewrite this as

$$\begin{aligned}
\Lambda\left(X_1,\ldots,X_n\right) &= \left(\frac{(n-1)S_X^2}{(n-1)S_X^2+n\bar{X}^2}\right)^{n/2} \\
&= \left(\frac{n-1}{n-1+n\bar{X}^2/S_X^2}\right)^{n/2} \\
&= \left(\frac{n-1}{n-1+T^2}\right)^{n/2},
\end{aligned}$$

which is a decreasing function of T^2 and hence of $|T|$ as well.

Problem 19. (Sign Test)

Consider data $X_1,\ldots,X_n \overset{IID}{\sim} f$ for some unknown probability density function f, and the testing problem

$$\begin{aligned}
&H_0 : f \text{ has median } 0 \\
&H_1 : f \text{ has median } \mu \text{ for some } \mu>0
\end{aligned}$$

(a) Explain why the Wilcoxon signed rank statistic does not have the same sampling distribution under every $P\in H_0$. Draw a picture of the graph of a density function f with median 0, such that the Wilcoxon signed rank statistic would tend to take larger values under f than under any density function g that is symmetric about 0.

Solution.

For a density function f with median 0 but is skewed right, such as in Fig. (a), positive values of $X_1,\ldots,X_n$ would tend to have a higher rank than negative values, so the Wilcoxon signed rank statistic would tend to take larger values under f than under any density function g that is symmetric about 0.

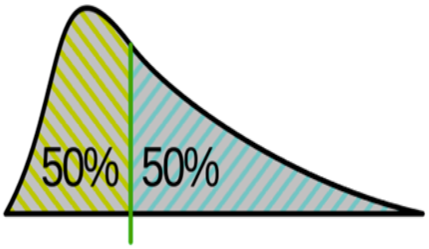


Figure 2: f with median 0.

(b) Consider the sign statistic S, defined as the number of values in $X_1, \ldots, X_n$ that are greater than 0. Explain why S has the same sampling distribution under every $P \in H_0$. How would you conduct a level-α test of H_0 Vs. H_1 using the test statistic S? (Describe explicitly the rejection threshold; you may assume that for $X \sim \text{Binomial}\left(n, \frac{1}{2}\right)$, there exists an integer k such that $\mathbb{P}[X \geq k]$ is exactly α.)

Solution. Let $Y_i = 1$ if $X_i > 0$ and $Y_i = 0$ otherwise. Since f has median $0, \mathbb{P}\left[Y_i = 1\right] = \mathbb{P}\left[X_i > 0\right] = 1/2$. Then

$$S = \sum_{i=1}^{n} Y_i \sim \text{Binomial}\left(n, \frac{1}{2}\right).$$

This distribution is the same for any PDF f with median 0. A test of H_0 Vs. H_1 should reject for large values of S. To achieve level-α, it should reject when $S \geq k$, where k is a value such that $\mathbb{P}[S \geq k] = \alpha$ under H_0. This is exactly the value of k given in the problem statement.

(c) When n is large, explain why we may reject H_0 when $S > \frac{n}{2} + \sqrt{\frac{n}{4}} Z_\alpha$ where Z_α is the upper α point of $\mathcal{N}(0,1)$, instead of using the rejection threshold you derived in part (b).

Solution. Note $\mathbb{E}\left[Y_i\right] = 1/2, \operatorname{Var}\left[Y_i\right] = 1/4$, and $\frac{S}{n} = \bar{Y}$. Then by the CLT

$$\sqrt{4n}\left(\frac{S}{n} - \frac{1}{2}\right) \rightarrow \mathcal{N}(0,1)$$

in distribution as $n \rightarrow \infty$. So for large n,

$$\begin{aligned}
\alpha &\approx \mathbb{P}\left[\sqrt{4n}\left(\frac{S}{n} - \frac{1}{2}\right) > Z_\alpha\right] \\
&= \mathbb{P}\left[\frac{S}{n} > \frac{1}{2} + \frac{1}{\sqrt{4n}} Z_\alpha\right] \\
&= \mathbb{P}\left[S > \frac{n}{2} + \sqrt{\frac{n}{4}} Z_\alpha\right],
\end{aligned}$$

and we may take as an approximate rejection threshold $\frac{n}{2} + \sqrt{\frac{n}{4}} Z_\alpha$.

(d) In this problem, we'll study the power of this test against the specific alternative $\mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$, for a fixed constant $h > 0$ (say $h = 1$ or $h = 2$) and large n. If $X \sim \mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$, show that

$$\mathbb{P}[X > 0] = \Phi\left(\frac{h}{\sqrt{n}}\right);$$

where Φ is the CDF of the standard normal distribution $\mathcal{N}(0,1)$. Applying a first-order Taylor expansion of Φ around 0, show that for large n

$$\mathbb{P}[X > 0] \approx \frac{1}{2} + \frac{h}{\sqrt{2\pi n}}.$$

Solution. Given that $X \sim N\left(\frac{h}{\sqrt{n}}, 1\right)$, we have

$$\mathbb{P}[X > 0] = \mathbb{P}\left[X - \frac{h}{\sqrt{n}} > -\frac{h}{\sqrt{n}}\right] = 1 - \Phi\left(-\frac{h}{\sqrt{n}}\right) = \Phi\left(\frac{h}{\sqrt{n}}\right).$$

A first-order Taylor expansion for a differentiable function f suggests that

$$f(x+h) \approx f(x) + hf'(x)$$

Applying this to the above and noting $\Phi'(x)$ is the normal PDF $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$,

$$\Phi\left(\frac{h}{\sqrt{n}}\right) \approx \Phi(0) + \frac{h}{\sqrt{n}}\phi(0) = \frac{1}{2} + \frac{h}{\sqrt{2\pi n}}.$$

(e) Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$. In this case, show that $\sqrt{\frac{4}{n}}\left(S - \frac{n}{2}\right)$ has an approximate normal distribution that does not depend on n (but depends on h) – what is the mean and variance of this normal distribution? Using this result, derive an approximate formula for the power of the sign test against the alternative $\mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$, in terms of Z_α, h, and the CDF Φ.

Solution. The sign statistic S can be written as

$$S = \sum_i Y_i, \text{ where } Y_i \sim \text{Bernoulli}\left(\mathbb{P}\left[X_i > 0\right]\right).$$

By the CLT, $\sqrt{n}\left(\frac{S}{n} - \mathbb{E}\left[Y_i\right]\right)$ is approximately distributed as $\mathcal{N}\left(0, \text{Var}\left[Y_i\right]\right)$. Applying part (d), $\mathbb{E}\left[Y_i\right] \approx \frac{1}{2} + \frac{h}{\sqrt{2\pi n}}$, so

$$\sqrt{n}\left(\frac{S}{n} - \mathbb{E}\left[Y_i\right]\right) \approx \sqrt{n}\left(\frac{S}{n} - \frac{1}{2} - \frac{h}{\sqrt{2\pi n}}\right) = \frac{1}{\sqrt{n}}\left(S - \frac{n}{2}\right) - \frac{h}{\sqrt{2\pi}}.$$

For large n,

$$\text{Var}\left[Y_i\right] \approx \left(\frac{1}{2} + \frac{h}{\sqrt{2\pi n}}\right)\left(1 - \left(\frac{1}{2} + \frac{h}{\sqrt{2\pi n}}\right)\right) \approx \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

So $\frac{1}{\sqrt{n}}\left(S - \frac{n}{2}\right)$ is approximately distributed as $\mathcal{N}\left(\frac{h}{\sqrt{2\pi}}, \frac{1}{4}\right)$. Multiplying by 2, $\sqrt{\frac{4}{n}}\left(S - \frac{n}{2}\right)$ is approximately distributed as $\mathcal{N}\left(\frac{2h}{\sqrt{2\pi}}, 1\right)$. The power of the sign test against the alternative $\mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$ is given by

$$\mathbb{P}\left[S > \frac{n}{2} + \sqrt{\frac{n}{4}}Z_\alpha\right] = \mathbb{P}\left[\sqrt{\frac{4}{n}}\left(S - \frac{n}{2}\right) - \frac{2h}{\sqrt{2\pi}} > Z_\alpha - \frac{2h}{\sqrt{2\pi}}\right] \approx 1 - \Phi\left(Z_\alpha - \frac{2h}{\sqrt{2\pi}}\right) = \Phi\left(\frac{2h}{\sqrt{2\pi}} - Z_\alpha\right).$$

Problem 20. (Comparing Binomial proportions)

The popular search engine Google would like to understand whether visitors to a website are more likely to click on an advertisement at the top of the page than one on the side of the page. They conduct an "AB test" in which they show n visitors (group A) a version of the website with the advertisement at the top and m visitors (group B) a version of the website with the (same) advertisement at the side. They record how many visitors in each group clicked on the advertisement.

(a) Formulate this problem as a hypothesis test. (You may assume that visitors in group A independently click on the ad with probability p_A and visitors in group B independently click on the ad with probability p_B, where both p_A and p_B are unknown probabilities in $(0, 1)$.) What are the null and alternative hypotheses? Are they simple or composite?

Solution. Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}\left(p_A\right)$ and $Y_1, \ldots, Y_m \overset{IID}{\sim} \text{Bernoulli}\left(p_B\right)$ be indicators of whether each visitor clicked on the ad. We wish to test

$$H_0 : p_A = p_B$$
$$H_1 : p_A > p_B$$

Both hypotheses are composite, as they do not specify the exact value of p_A or p_B.

(b) Let $\hat{p}_A$ be the fraction of visitors in group A who clicked on the ad, and similarly for $\hat{p}_B$. A reasonable intuition is to reject H_0 when $\hat{p}_A - \hat{p}_B$ is large. What is the variance of $\hat{p}_A - \hat{p}_B$? Is this the same for all data distributions in H_0?

Solution. As $n\hat{p}_A \sim \text{Binomial}(n, p_A)$, we have $\text{Var}[n\hat{p}_A] = np_A(1-p_A)$ so $\text{Var}[\hat{p}_A] = \frac{p_A(1-p_A)}{n}$. Similarly $\text{Var}[\hat{p}_B] = \frac{p_B(1-p_B)}{m}$. Since $\hat{p}_A$ and $\hat{p}_B$ are independent,

$$\text{Var}[\hat{p}_A - \hat{p}_B] = \text{Var}[\hat{p}_A] + \text{Var}[-\hat{p}_B] = \text{Var}[\hat{p}_A] + \text{Var}[\hat{p}_B] = \frac{p_A(1-p_A)}{n} + \frac{p_B(1-p_B)}{m}.$$

Under $H_0, p_A = p_B = p$ for some $p \in (0,1)$, and this variance is $p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)$. This is not the same for all data distributions in H_0, as it depends on p. (So we cannot perform a test of H_0 directly using the test statistic $\hat{p}_A - \hat{p}_B$.)

(c) Describe a way to estimate the variance of $\hat{p}_A - \hat{p}_B$ using the available data, assuming H_0 is true – call this estimate $\hat{V}$. Explain heuristically why, when n and m are both large, the test statistic

$$T = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{V}}}$$

is approximately distributed as $\mathcal{N}(0,1)$ under any data distribution in H_0. (You may assume that when n and m are both large, the ratio of $\hat{V}$ to the true variance of $\hat{p}_A - \hat{p}_B$ that you derived in part (b) is very close to 1 with high probability.) Explain how to use this observation to perform an approximate level-α test of H_0 versus H_1.

Solution. One way of estimating the variance is to take

$$\hat{V} = \frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}.$$

Another way (since $p_A = p_B$ under H_0) is to first estimate a pooled sample proportion

$$\hat{p} = \frac{\hat{p}_A n + \hat{p}_B m}{n+m}$$

and then estimate the variance as

$$\hat{V} = \hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right).$$

(Both ways are reasonable under H_0.)

Under $H_0, p_A = p_B = p$ for some $p \in (0,1)$, so the CLT implies $\frac{\sqrt{n}(\hat{p}_A - p)}{\sqrt{p(1-p)}} \to \mathcal{N}(0,1)$ and $\frac{\sqrt{m}(\hat{p}_B - p)}{\sqrt{p(1-p)}} \to \mathcal{N}(0,1)$ in distribution as $n, m \to \infty$. So for large n and m, the distributions of $\hat{p}_A$ and $\hat{p}_B$ are approximately $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ and $\mathcal{N}\left(p, \frac{p(1-p)}{m}\right)$. Note that $\hat{p}_A$ and $\hat{p}_B$ are independent, so their difference is distributed approximately as $\mathcal{N}\left(0, p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)\right)$. We may write the test statistic T as

$$T = \frac{\sqrt{p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)}}{\sqrt{\hat{V}}} \frac{\hat{p}_A - \hat{p}_B}{\sqrt{p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)}}.$$

Since $\frac{\sqrt{p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)}}{\sqrt{\hat{V}}} \approx 1$ with high probability for large n and m, T is approximately distributed as $\mathcal{N}(0,1)$, and an asymptotic level-α test rejects H_0 for $T > Z_\alpha$.

Problem 21. (Covid-19 testing problem)

There are approximately 540 coronavirus testing locations in Abu Dhabi. At the beginning of the day, officials at each location record $Y =$ number of specimens tested to find the first positive case and assume Y follows a geometric distribution with probability p. The probability p satisfies $0 < p < 1$ and is unknown. In this application, we might also call p the "population prevalence" of the disease. Of course, careful thought should go into defining exactly what the "population" is here.

Suppose $Y_1, Y_2, \ldots, Y_{540}$ are iid geometric (p) random variables (one for each site) observed on a given day. Epidemiologists at the Department of Health would like to test

$$H_0 : p = 0.02$$

versus

$$H_a : p < 0.02.$$

(a) Show the likelihood function of p on the basis of observing $\mathbf{y} = (y_1, y_2, \ldots, y_{540})$ is given by

$$L(p \mid \mathbf{y}) = (1-p)^{\sum_{i=1}^{540} y_i - 540} p^{540}.$$

(b) Show the uniformly most powerful (UMP) level α test of H_0 versus H_a has a rejection region of the form

$$\text{RR} = \left\{ t = \sum_{i=1}^{540} y_i \geq k^* \right\}.$$

How would you choose k^* to ensure the test is level $\alpha = 0.05$? Hint: What is the sampling distribution of $T = \sum_{i=1}^{540} Y_i$ when H_0 is true?

Solution.

(a) Using the p.m.f of a geometric distribution, we can get

$$L(p \mid \mathbf{y}) = \prod_{i=1}^{540} \left\{ (1-p)^{y_i - 1} \times p \right\} = (1-p)^{\sum_{i=1}^{540} y_i - 540} p^{540}.$$

(b) We choose $H_a' : p = p_a$ to start.

$$\frac{L(p_0 \mid \mathbf{y})}{L(p_a \mid \mathbf{y})} = \left(\frac{1-p_0}{1-p_a} \right)^{\sum_{i=1}^{540} y_i - 540} \left(\frac{p_0}{p_a} \right)^{540}.$$

This ratio goes up when $T = \sum_{i=1}^{540} Y_i$ goes down, under $p_a < 0.02 = p_0$, namely $\frac{1-p_0}{1-p_a} < 1$. Therefore, by Neyman-Pearson Lemma, we can construct the most powerful test with its reject region like:

$$RR = \left\{ \frac{L(p_0 \mid \mathbf{y})}{L(p_a \mid \mathbf{y})} < k \right\} = \left\{ T = \sum_{i=1}^{540} Y_i > k^* \right\},$$

where k^* satisfies:

$$\alpha = P_{H_0}(RR) = P\left(\sum_{i=1}^{540} Y_i > k^* \mid H_0 \right).$$

When H_0 is true, we can argue $T = \sum_{i=1}^{540} Y_i \sim Neg - \text{Binomial}\,(540, p_0)$. (Here, I used the negative binomial in the version of "the number of trials that needed to get the first 540 success, i.e., the first 540 positive cases". There is another version of "the number of failures before the first 540 success". You just need to take care of the definition of your parameters.) Thus, we can choose k^* to be the (upper) quantile of the negative binomial distribution to achieve a level α test.

Now, we simply note that this rejection region does not depend on the value of p_a under H_a' (which we specified arbitrarily). Therefore, this test can be the uniformly most powerful (UMP) level α test. Note: To get an exact size α, you may refer to the method of making a randomized test.

Problem 22. (Epidemiological Modeling of Covid-19)

SEIR models are used by epidemiologists to describe COVID-19 disease severity in a population. The model consists of four different categories:

$\mathbf{S} =$ susceptible category; $\mathbf{E} =$ exposed category; $\mathbf{I} =$ infected category; $\mathbf{R} =$ recovered category.

The four categories are mutually exclusive and exhaustive among living individuals (SEIRD models do include a fifth category for those who have died from disease). A random sample of n individuals is selected from a population (e.g., residents of Abu Dhabi city), and the category status of each individual is identified. This produces the multinomial random vector

$$\mathbf{Y} \sim \text{mult}\left(n, \mathbf{p}; \sum_{j=1}^{4} p_j = 1\right),$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} \quad \text{and} \quad \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}.$$

The random variables Y_1, Y_2, Y_3, Y_4 record the number of individuals identified in the susceptible, exposed, infected, and recovered categories, respectively. Recall that the beta distribution is a conjugate prior for the binomial. Just as the multinomial distribution can be regarded as a generalization of the binomial (to more than two categories), we need a prior distribution for $\mathbf{p}$ that is a generalization of the beta. This generalization is the Dirichlet $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ distribution. Specifically, suppose $\mathbf{p}$ is best regarded as random with prior pdf

$$g(\mathbf{p}) = \begin{cases} \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3+\alpha_4)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} p_3^{\alpha_3-1} p_4^{\alpha_4-1}, & 0 < p_j < 1, \sum_{j=1}^{4} p_j = 1 \\ 0, & \text{otherwise} \end{cases}$$

where $\alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0$, and $\alpha_4 > 0$ are known.

(a) If $Y \mid \mathbf{p} \sim \text{mult}\left(n, \mathbf{p}; \sum_{j=1}^{4} p_j = 1\right)$ and $\mathbf{p} \sim g(\mathbf{p})$, show the posterior distribution $g(\mathbf{p} \mid \mathbf{y})$ is Dirichlet with parameters $\alpha_j^* = y_j + \alpha_j$, for $j = 1, 2, 3, 4$. Hint: The joint distribution of $\mathbf{Y}$ and $\mathbf{p}$ satisfies

$$f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p}) = f_{\mathbf{Y}|\mathbf{p}}(\mathbf{y} \mid \mathbf{p}) g(\mathbf{p})$$

and $g(\mathbf{p} \mid \mathbf{y})$ is proportional to $f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p})$.

(b) A special case of the Dirichlet distribution above arises when $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$, the so-called "symmetric Dirichlet distribution." This distribution would arise when one has no prior information to favor the count in one SEIR category over the other three. Do you think $g(\mathbf{p})$ with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ would be a reasonable prior model for covid-19 in Abu Dhabi city? Explain.

Solution:

(a)

$$\begin{aligned} g(\mathbf{p} \mid \mathbf{y}) &\propto f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p}) \\ &= f_{\mathbf{Y}|\mathbf{p}}(\mathbf{y} \mid \mathbf{p}) g(\mathbf{p}) \\ &= \left(\frac{n!}{\prod_{j=1}^{4} y_i!} \prod_{j=1}^{n} p_i^{y_i}\right) \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)}{\Gamma(\alpha_1) \Gamma(\alpha_2) \Gamma(\alpha_3) \Gamma(\alpha_4)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} p_3^{\alpha_3-1} p_4^{\alpha_4-1} \\ &\propto p_1^{\alpha_1+y_1-1} p_2^{\alpha_2+y_2-1} p_3^{\alpha_3+y_3-1} p_4^{\alpha_4+y_4-1} \\ &\sim \text{Dirichlet}\left(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*\right), \text{ where } \alpha_j^* = y_j + \alpha_j. \end{aligned}$$

(b) No, this prior might not be proper here. The probability for each category should not be all the same. (You can search for more information to argue whether it is reasonable or not, like historical data, expertise, doctors, or even your own observations, etc.)

Problem 23. (Mean referral waiting times)

We would like to compare the population mean referral waiting times for patients in Abu Dhabi and Dubai seeking care from a gastrointestinal specialist. By "referral waiting time", I mean the time it takes to see a gastrointestinal specialist once a referral has been made by another health professional (e.g., a primary care physician). We have independent random samples of patients from the two locations. Here are the corresponding waiting times and population-level models for them:

- Abu Dhabi: $X_1, X_2, \ldots, X_m \overset{\text{IID}}{\sim}$ Exponential (θ_1)
- Dubai: $Y_1, Y_2, \ldots, Y_n \overset{\text{IID}}{\sim}$ Exponential (θ_2).

The population parameters satisfy $\theta_1 > \theta$ and $\theta_2 > \theta$ and are unknown. The goal is to test

$$H_0 : \theta_1 = \theta_2$$

$$\text{versus}$$

$$H_a : \theta_1 \neq \theta_2.$$

(a) Preparing for an LRT derivation below, carefully describe the null parameter space Θ_0 and the entire parameter space Θ. Draw a picture of what both spaces look like.

(b) Show the likelihood function is given by

$$L(\theta \mid \mathbf{x}, \mathbf{y}) = L(\theta_1, \theta_2 \mid \mathbf{x}, \mathbf{y}) = \frac{1}{\theta_1^m} e^{-\sum_{i=1}^m x_i/\theta_1} \times \frac{1}{\theta_2^n} e^{-\sum_{j=1}^n y_j/\theta_2},$$

where $\theta = (\theta_1, \theta_2)$. This is just the likelihood from each sample multiplied together (because the two samples are independent).

(c) Show the (restricted) MLE of θ over the null parameter space Θ_0 is

$$\hat{\theta}_0 = \begin{pmatrix} \frac{m\bar{X}+n\bar{Y}}{m+n} \\ \frac{m\bar{X}+n\bar{Y}}{m+n} \end{pmatrix}.$$

(d) Show the (unrestricted) MLE of θ over the entire parameter space Θ is

$$\hat{\theta} = \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix}.$$

(e) Show the likelihood ratio test (LRT) statistic

$$\lambda = \frac{L\left(\hat{\theta}_0 \mid \mathbf{x}, \mathbf{y}\right)}{L(\hat{\theta} \mid \mathbf{x}, \mathbf{y})} = \frac{\bar{x}^m \bar{y}^n}{\left(\frac{m\bar{x}+n\bar{y}}{m+n}\right)^{m+n}}.$$

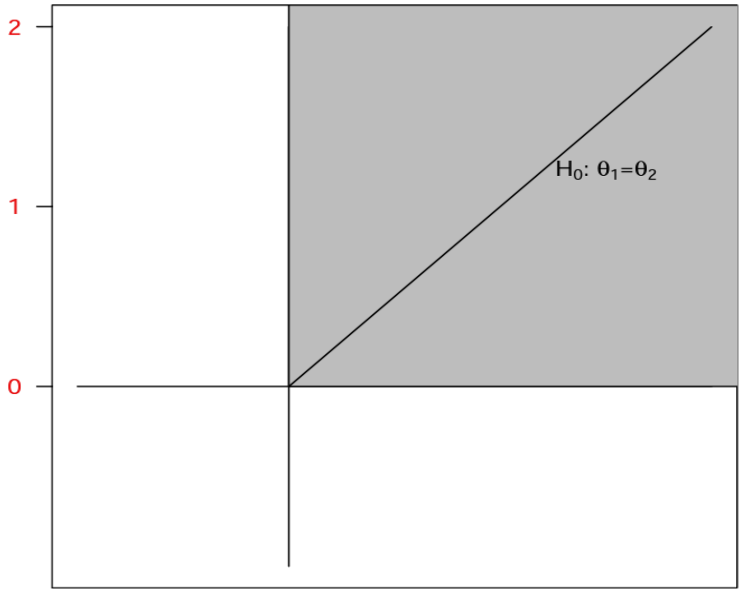
(f) Here are the observed data on the referral waiting times for both groups of patients:

Abu Dhabi		Dubai	
36	11	1	49
47	52	16	43
52	9	22	8
1	39	11	2
9	20	8	8
72	32	26	53
		12	24
		39	

Calculate $-2 \ln \lambda$ using the data above and implement a large-sample LRT to test H_0 versus H_a. What is your conclusion at $\alpha = 0.05$?

Solution.

(a) The entire space should be the region with all positive values for θ_1 and θ_2. The null space is then the line with $\theta_1 = \theta_2$.



(b) Plug in the p.d.f's to construct likelihood:

$$L(\theta \mid \mathbf{x}, \mathbf{y}) = L(\theta_1, \theta_2 \mid \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{m} \frac{1}{\theta_1} e^{-\frac{x_i}{\theta_1}} \prod_{j=1}^{n} \frac{1}{\theta_2} e^{-\frac{y_j}{\theta_2}}$$
$$= \frac{1}{\theta_1^m} e^{-\sum_{i=1}^m x_i/\theta_1} \times \frac{1}{\theta_2^n} e^{-\sum_{j=1}^n y_j/\theta_2}$$

(c) Under H_0, we assume $\theta_1 = \theta_2 = \theta$, then:

$$L(\theta \mid \mathbf{x}, \mathbf{y}) = \frac{1}{\theta^{(m+n)}} e^{-\frac{\sum_{i=1}^m x_i + \sum_{j=1}^n y_j}{\theta}}$$

log-Likelihood:

$$l = \log L\left(\theta_0 = (\theta, \theta)^T \mid \mathbf{x}, \mathbf{y}\right) = -(m+n)\ln\theta - \frac{\sum_{i=1}^m x_i + \sum_{j=1}^n y_j}{\theta}$$

Derivative:

$$\frac{\partial l}{\partial \theta} = -\frac{m+n}{\theta} + \frac{\sum_{i=1}^m x_i + \sum_{j=1}^n y_j}{\theta^2} \overset{set}{=} 0$$

Then, we can get

$$\hat{\theta} = \frac{\sum_{i=1}^m x_i + \sum_{j=1}^n y_j}{m+n} = \frac{m\bar{X} + n\bar{Y}}{m+n},$$

And therefore,

$$\hat{\theta}_0 = \begin{pmatrix} \hat{\theta} \\ \hat{\theta} \end{pmatrix} = \begin{pmatrix} \frac{m\bar{X}+n\bar{Y}}{m+n} \\ \frac{m\bar{X}+n\bar{Y}}{m+n} \end{pmatrix}.$$

(d) From the original likelihood:

$$L(\theta \mid \mathbf{x}, \mathbf{y}) = L(\theta_1, \theta_2 \mid \mathbf{x}, \mathbf{y}) = \frac{1}{\theta_1^m} e^{-\sum_{i=1}^m x_i/\theta_1} \times \frac{1}{\theta_2^n} e^{-\sum_{j=1}^n y_j/\theta_2},$$

we get log-likelihood:

$$l = \log L\left(\theta = (\theta_1, \theta_2)^T \mid \mathbf{x}, \mathbf{y}\right) = -m\ln\theta_1 - \frac{\sum_{i=1}^m x_i}{\theta_1} - n\ln\theta_2 - \frac{\sum_{j=1}^n y_j}{\theta_2}$$

Take partial derivatives:

$$\left\{\begin{array}{l} \frac{\partial l}{\theta_1} = -\frac{m}{\theta_1} + \frac{m\overline{X}}{\theta_1^2} \overset{\text{set}}{=} 0 \\ \frac{\partial l}{\theta_2} = -\frac{n}{\theta_2} + \frac{n\overline{Y}}{\theta_2^2} \overset{\text{set}}{=} 0 \end{array}\right. \Rightarrow \left\{\begin{array}{l} \hat{\theta}_1 = \bar{X} \\ \hat{\theta}_2 = \bar{Y} \end{array}\right.$$

Thus, we can get

$$\hat{\theta} = \left(\begin{array}{c} \bar{X} \\ \bar{Y} \end{array} \right).$$

(e) Plug in the results above, you will get:

$$\lambda = \frac{\max_{\theta\in\Theta_0} L(\theta \mid \mathbf{x}, \mathbf{y})}{\max_{\theta\in\Theta} L(\theta \mid \mathbf{x}, \mathbf{y})} = \frac{\frac{1}{\hat{\theta}^m} e^{-\sum_{i=1}^m x_i/\hat{\theta}} \times \frac{1}{\bar{\theta}^n} e^{-\sum_{j=1}^n y_j/\hat{\theta}}}{\frac{1}{\hat{\theta}_1^m} e^{-\sum_{i=1}^m x_i/\hat{\theta}_1} \times \frac{1}{\hat{\theta}_2^n} e^{-\sum_{j=1}^n y_j/\hat{\theta}_2}} = \frac{\frac{1}{\hat{\theta}^{m+n}} e^{-(m\bar{x}+n\bar{y})/\hat{\theta}}}{\frac{1}{\bar{x}^m} e^{-m\bar{x}/\bar{x}} \times \frac{1}{\bar{y}^n} e^{-n\bar{y}/\bar{y}}}$$

$$= \frac{\bar{x}^m \bar{y}^n e^{m+n}}{\hat{\theta}^{m+n} e^{m+n}} = \frac{\bar{x}^m \bar{y}^n}{\left(\frac{m\bar{x}+n\bar{y}}{m+n}\right)^{m+n}}$$

(f) We can use the large-sample result:

$$-2\ln\lambda \xrightarrow{d} \chi_1^2.$$

Here, the degree of freedom is 1 since we need to estimate two values under the entire parameter space, but under H_0, we just need 1 estimator. Thus, $\nu = \dim(\Theta) - \dim(\Theta_0) = 2 - 1 = 1$.

Plug in the values, we can get:
$-2\ln\lambda = 1.0158 < \chi^2_{1,0.95} = 3.841$, so we cannot reject the null hypothesis under the 0.05 significance level.

Problem 24. (Effect of a confounding factor)

To study the effectiveness of a drug that claims to lower blood cholesterol levels, we designed a simple experiment with n subjects in a control group and n (different) subjects in a treatment group. We administer the drug to the treatment group and a placebo to the control group, measure the cholesterol levels of all subjects at the end of the study and look at whether cholesterol levels are lower in the treatment group than in control. Let $X_1, \ldots, X_n$ be the cholesterol levels in the control group and $Y_1, \ldots, Y_n$ be those in the treatment group, and let

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

be the standard two-sample t-statistic where S_p^2 is the pooled variance.
Assume, throughout this problem, that the drug, in fact, is *not* effective and has the exact same effect as the placebo. However, suppose there are two types of subjects, high-risk and low-risk. (Approximately half of the human population is high-risk, and half is low-risk; assume that we cannot directly observe whether a person is high-risk or low-risk.) The cholesterol level for high-risk subjects is distributed as $\mathcal{N}\left(\mu_H, \sigma^2\right)$, and for low-risk subjects as $\mathcal{N}\left(\mu_L, \sigma^2\right)$.

(a) A carefully designed study randomly selects subjects for the two groups so that each subject selected for either group is (independently) with probability 1/2 high-risk and probability 1/2 low-risk. Explain why, in this case, $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are IID from a common distribution. What are $\mathbb{E}[X_i]$ and $\operatorname{Var}[X_i]$?

(b) Explain (using the CLT and Slutsky's lemma) why, when n is large, T is approximately distributed as $\mathcal{N}(0,1)$, and hence a test that rejects for $T > Z_\alpha$ is approximately level-α for large n.

(c) A poorly-designed study fails to properly randomize the treatment and control groups, so that each subject selected for the control group is with probability p high-risk and probability $1-p$ low-risk, and each subject selected for the treatment group is with probability q high-risk and probability $1-q$ low-risk. In this case, what are $\mathbb{E}[X_i], \operatorname{Var}[X_i], \mathbb{E}[Y_i]$, and $\operatorname{Var}[Y_i]$?

(d) In the setting of part (c), show that S_p^2 converges in probability to a constant $c \in \mathbb{R}$ as $n \to \infty$, and provide a formula for c. Show that T is approximately normally distributed, and provide formulas for the mean and variance of this normal. Is the rejection probability $\mathbb{P}[T > Z_\alpha]$ necessarily close to α? Discuss briefly how this probability depends on the values μ_H, μ_L, σ^2, p, and q.

Solution.

(a) Each group member is selected independently from either the high-risk group or the low-risk group. So the cholesterol level for each person in each group is a random variable independent of that for any other person. Also, since for both the treatment and control groups, with probability 1/2, a high-risk individual is chosen, and with probability 1/2, a low-risk individual is chosen, they must have the same distribution. So, the variables $X_1, \cdots, X_n, Y_1, \cdots, Y_n$ are IID from a common distribution. To compute the mean and variance, we may write X_i as

$$X_i = Z_i H_i + (1 - Z_i) L_i$$

where $H_i \sim \mathcal{N}\left(\mu_H, \sigma^2\right), L_i \sim \mathcal{N}\left(\mu_L, \sigma^2\right), Z_i \sim$ Bernoulli(1/2), and these are independent. Then

$$\begin{aligned} \mathbb{E}[X_i] &= \mathbb{E}[Z_i]\,\mathbb{E}[H_i] + \mathbb{E}[1 - Z_i]\,\mathbb{E}[L_i] \ \text{ (due to independence)} \\ &= \frac{1}{2}\mu_H + \frac{1}{2}\mu_L. \end{aligned}$$

To compute the variance, we have

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}\left[Z_i^2 H_i^2 + 2 Z_i (1 - Z_i) H_i L_i + (1 - Z_i)^2 L_i^2\right].$$

Note that since $Z_i \in \{0, 1\}$, $Z_i(1 - Z_i) = 0$, $Z_i^2 = Z_i$, and $(1 - Z_i)^2 = (1 - Z_i)$. Then

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}[Z_i]\,\mathbb{E}\left[H_i^2\right] + \mathbb{E}[1 - Z_i]\,\mathbb{E}\left[L_i^2\right] = \frac{1}{2}\mathbb{E}\left[H_i^2\right] + \frac{1}{2}\mathbb{E}\left[L_i^2\right]$$

We have $\mathbb{E}\left[H_i^2\right] = \text{Var}[H_i] + (\mathbb{E}[H_i])^2 = \mu_H^2 + \sigma^2$, and similarly $\mathbb{E}\left[L_i^2\right] = \mu_L^2 + \sigma^2$. So

$$\mathbb{E}\left[X_i^2\right] = \frac{1}{2}\left(\mu_L^2 + \mu_H^2\right) + \sigma^2,$$

and

$$\text{Var}[X_i] = \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X_i])^2 = \frac{1}{2}\left(\mu_L^2 + \mu_H^2\right) + \sigma^2 - \frac{1}{4}\left(\mu_L^2 - 2\mu_L\mu_H + \mu_H^2\right) = \sigma^2 + \frac{1}{4}(\mu_H - \mu_L)^2.$$

(b) As the X_i 's and Y_i 's are all IID, by the Central Limit Theorem, $\sqrt{n}\left(\bar{X} - \mathbb{E}[X_i]\right) \to \mathcal{N}(0, \text{Var}[X_i])$ and $\sqrt{n}\left(\bar{Y} - \mathbb{E}[X_i]\right) \to \mathcal{N}(0, \text{Var}[X_i])$ in distribution, so their difference $\sqrt{n}(\bar{X} - \bar{Y}) \to \mathcal{N}(0, 2\,\text{Var}[X_i])$. The pooled variance is

$$S_p^2 = \frac{1}{2n - 2}\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2\right) = \frac{1}{2}S_X^2 + \frac{1}{2}S_Y^2,$$

where $S_X^2 = \frac{1}{n-1}\sum_i\left(X_i - \bar{X}\right)^2$ and $S_Y^2 = \frac{1}{n-1}\sum_i\left(Y_i - \bar{Y}\right)^2$ are the individual sample variances. Using the result, $S_X^2 \to \text{Var}[X_i]$ and $S_Y^2 \to \text{Var}[Y_i] = \text{Var}[X_i]$ in probability, so the Continuous Mapping Theorem implies $S_p^2 \to \text{Var}[X_i]$. Then

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{n}}} = \frac{\sqrt{\text{Var}[X_i]}}{S_p}\frac{\sqrt{n}(X - \bar{Y})}{\sqrt{2\,\text{Var}[X_i]}} \to \mathcal{N}(0, 1)$$

in distribution by Slutsky's lemma. Hence, a test that rejects for $T > Z_\alpha$ is approximately level α for large n.

(c) The difference in this part from Part (a) is that here

$$X_i = Z_i H_i + (1 - Z_i) L_i,$$

where H_i, L_i are defined as before but $Z_i \sim$ Bernoulli(p). Then

$$\mathbb{E}[X_i] = p\mu_H + (1 - p)\mu_L.$$

Similarly, $\mathbb{E}[Y_i] = q\mu_H + (1-q)\mu_L$.
For the variances, we compute as in Part (a)

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}[Z_i]\mathbb{E}\left[H_i^2\right] + \mathbb{E}[1-Z_i]\mathbb{E}\left[L_i^2\right] = p\left(\mu_H^2+\sigma^2\right) + (1-p)\left(\mu_L^2+\sigma^2\right) = p\mu_H^2 + (1-p)\mu_L^2 + \sigma^2,$$

so

$$\operatorname{Var}[X_i] = \mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}[X_i]\right)^2 = p\mu_H^2 + (1-p)\mu_L^2 + \sigma^2 - \left(p\mu_H + (1-p)\mu_L\right)^2 = \sigma^2 + \left(\mu_H - \mu_L\right)^2 p(1-p).$$

Similarly, $\operatorname{Var}[Y_i] = \sigma^2 + \left(\mu_H-\mu_L\right)^2 q(1-q)$.

(d) In this case $S_X^2 \to \operatorname{Var}[X_i]$ and $S_Y^2 \to \operatorname{Var}[Y_i]$ in probability, so

$$S_p^2 \to \frac{1}{2}\left(\operatorname{Var}[X_i] + \operatorname{Var}[Y_i]\right) = \sigma^2 + \frac{1}{2}\left(p(1-p)+q(1-q)\right)\left(\mu_H-\mu_L\right)^2 =: c.$$

By the CLT, $\sqrt{n}\left(\bar{X} - \mathbb{E}[X_i]\right) \to \mathcal{N}\left(0, \operatorname{Var}[X_i]\right)$ and $\sqrt{n}\left(\bar{Y} - \mathbb{E}[Y_i]\right) \to \mathcal{N}\left(0, \operatorname{Var}[Y_i]\right)$. The X_i 's and Y_i 's are independent, so the difference $\sqrt{n}\left(\bar{X} - \bar{Y} - \mathbb{E}[X_i] + \mathbb{E}[Y_i]\right) \to \mathcal{N}\left(0, \operatorname{Var}[X_i] + \operatorname{Var}[Y_i]\right)$.
Then

$$T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{2/n}} = \frac{1}{\sqrt{2S_p^2}}\left(\sqrt{n}(\bar{X}-\bar{Y})\right)$$

is approximately distributed as

$$\frac{1}{\sqrt{2c}}\mathcal{N}\left(\sqrt{n}\left(\mathbb{E}[X_i] - \mathbb{E}[Y_i]\right), \operatorname{Var}[X_i] + \operatorname{Var}[Y_i]\right) = \mathcal{N}\left(\frac{\sqrt{n}\left(\mathbb{E}[X_i] - \mathbb{E}[Y_i]\right)}{\sqrt{2c}}, 1\right).$$

Let

$$m := \frac{\sqrt{n}\left(\mathbb{E}[X_i] - \mathbb{E}[Y_i]\right)}{\sqrt{2c}} = \frac{\sqrt{n}(p-q)\left(\mu_H-\mu_L\right)}{\sqrt{2c}} = \frac{\sqrt{n}(p-q)\left(\mu_H-\mu_L\right)}{\sqrt{2\sigma^2 + \left(p(1-p)+q(1-q)\right)\left(\mu_H-\mu_L\right)^2}},$$

so T is approximately $\mathcal{N}(m,1)$. Then the rejection probability is

$$\mathbb{P}[T > Z_\alpha] = \mathbb{P}[T - m > Z_\alpha - m] \approx 1 - \Phi(Z_\alpha - m) = \Phi(m - Z_\alpha).$$

This probability increases in m and only equals α when $m = 0$. If $(p-q)\left(\mu_H-\mu_L\right) > 0$, then $m \to \infty$ as $n \to \infty$, and we expect to falsely reject H_0 with probability close to 1 for large n. If $(p-q)\left(\mu_H-\mu_L\right) < 0$, then $m \to -\infty$ as $n \to \infty$, and we expect the significance level of the test to in fact be close to 0 for large n.

Problem 25. (Improving upon Bonferroni for independent tests)

(a) Let $P_1, \ldots, P_n$ be the p-values from n different hypothesis tests. Suppose that the tests are performed using independent sets of data, and in fact all of the null hypotheses are true, so $P_1, \ldots, P_n \overset{\text{IID}}{\sim} \text{Uniform}(0,1)$. Show that for any $t \in (0,1)$,

$$\mathbb{P}\left[\min_{i=1}^{n} P_i \leq t\right] = 1 - (1-t)^n.$$

(b) Under the setting of part (a), if we perform all tests at significance level $1-(1-\alpha)^{1/n}$ (that is, we reject a null hypothesis if its p-value is less than this level), show that the probability of (falsely) rejecting any of the n null hypotheses is exactly α. Is this procedure more or less powerful than the Bonferroni procedure (of performing all tests at level α/n)?

(c) Suppose, now, that all of the p-values $P_1, \ldots, P_n$ are still independent, but not necessarily all of the null hypotheses are true. (So the p-values corresponding to the true null hypotheses are still IID and distributed as Uniform $(0,1)$.) If we perform all tests at significance level $1-(1-\alpha)^{1/n}$, does this procedure control the family-wise error rate (FWER) at level α? (Explain why, or show a counterexample.)

Solution.

(a) We know that $P_1, \dots, P_n \overset{\text{IID}}{\sim} \mathrm{U}(0,1)$. So for any $t \in (0,1)$,

$$\begin{aligned}\mathbb{P}\left[\min_{i=1}^{n} P_i \le t\right] &= 1 - \mathbb{P}\left[\min_{i=1}^{n} P_i > t\right] \\ &= 1 - \mathbb{P}\left[P_i > t \quad \forall\, i = 1, \dots, n\right] \\ &= 1 - \prod_{i=1}^{n} \mathbb{P}\left[P_i > t\right] = 1 - (1-t)^n.\end{aligned}$$

(b) If all the tests are performed at significance level $1 - (1-\alpha)^{1/n}$,

$$\begin{aligned}\mathbb{P}(\text{rejecting any of the } n \text{ null hypotheses}) &= \mathbb{P}\left(P_i < 1 - (1-\alpha)^{1/n} \text{ for any } i\right) \\ &= \mathbb{P}\left(\min_{i=1}^{n} P_i < 1 - (1-\alpha)^{1/n}\right) = 1 - \left(1 - 1 + (1-\alpha)^{1/n}\right)^n = \alpha.\end{aligned}$$

Hence, the probability of (falsely) rejecting any of the n null hypotheses is exactly α.
The Bonferroni procedure rejects when $P_i \le \alpha/n$ and the above procedure rejects when $P_i \le 1 - (1-\alpha)^{1/n}$.
Note that

$$\left(1 - \frac{\alpha}{n}\right)^n > 1 - \alpha,$$

so $1 - (1-\alpha)^{1/n} > \alpha/n$. Hence, whenever the Bonferroni test rejects, this procedure also rejects, so this procedure is more powerful than the Bonferroni test.

(c) Suppose there are k true null hypotheses, and without loss of generality, let us assume that these are the first k. If all the tests are performed at significance level $1 - (1-\alpha)^{1/n}$, and V is the number of true null hypotheses that are rejected, then the FWER is

$$\begin{aligned}\mathbb{P}(V \ge 1) &= \mathbb{P}\left(\min_{i=1}^{k} P_i \le 1 - (1-\alpha)^{1/n}\right) \\ &= 1 - \mathbb{P}\left(\min_{i=1}^{k} P_i > 1 - (1-\alpha)^{1/n}\right) \\ &= 1 - \mathbb{P}\left(P_i > 1 - (1-\alpha)^{1/n} \ \forall\, i = 1, \cdots, k\right) \\ &= 1 - \left(1 - 1 + (1-\alpha)^{1/n}\right)^k = 1 - (1-\alpha)^{k/n}.\end{aligned}$$

Since $k \le n$ and $\alpha < 1, (1-\alpha)^{k/n} > (1-\alpha)$ and hence $1 - (1-\alpha)^{k/n} < \alpha$, so the FWER is controlled.

Problem 26. (Computing the MME and MLE using Newton-Raphson Method)

(a) Let $\mathbf{X} = (X_1, \dots, X_n)$ be i.i.d. random variables with Kumaraswamy distribution with parameters $a > 0$ and $b > 0$. The probability density function of the Kumaraswamy distribution is

$$f(x; a, b) = abx^{a-1}\left(1 - x^a\right)^{b-1}, \qquad \text{where } x \in [0,1].$$

Find the method of moments for the Kumaraswamy distribution using numerical methods.

(b) Implement a function that takes as input a vector of data values X, performs the Newton-Raphson iterations to compute the MLEs $\hat{\alpha}$ and $\hat{\beta}$ in the Gamma (α, β) model, and outputs $\hat{\alpha}$ and $\hat{\beta}$. (You may use the form of the Newton-Raphson update equation derived in class. You may terminate the Newton-Raphson iterations when $|\alpha^{(t+1)} - \alpha^{(t)}|$ is sufficiently small.)

(c) For $n = 500$, use your function from part (a) to simulate the sampling distributions of $\hat{\alpha}$ and $\hat{\beta}$ computed from $X_1, \dots, X_n \overset{\text{IID}}{\sim}$ Gamma(1,2). Plot histograms of the values of $\hat{\alpha}$ and $\hat{\beta}$ across 5000 simulations, and report the simulated mean and variance of $\hat{\alpha}$ and $\hat{\beta}$ as well as the simulated covariance between $\hat{\alpha}$ and $\hat{\beta}$. Compute the inverse of the Fisher Information matrix $I(\alpha, \beta)$ at $\alpha = 1$ and $\beta = 2$. Do your simulations support that $(\hat{\alpha}, \hat{\beta})$ is approximately distributed as $\mathcal{N}\left((1,2), \frac{1}{n} I(1,2)^{-1}\right)$? (You may use the formula for the Fisher information matrix $I(\alpha, \beta)$ and/or its inverse derived in class.)

Solution.

(a) The Kumaraswamy distribution is an alternative to the Beta distribution, which offers similar flexibility in terms of the shape of the density. An appealing feature of the Kumaraswamy distribution is that it only involves algebraic terms, in contrast to the Beta distribution, which requires the evaluation of the Beta function. The mean and variance of this distribution are

$$\mu_1 = E[X] = \frac{b\Gamma\left(1+\frac{1}{a}\right)\Gamma(b)}{\Gamma\left(1+\frac{1}{a}+b\right)},$$

$$\text{Var}[X] = E\left[X^2\right] - E[X]^2 = \mu_2 - \mu_1^2 = \frac{b\Gamma(1+2/a)\Gamma(b)}{\Gamma(1+b+2/a)} - \mu_1^2.$$

The MME of a and $b, \tilde{a}$ and $\tilde{b}$, are the solution to the system of non-linear equations,

$$\overline{\mathbf{X}} = \frac{b\Gamma\left(1+\frac{1}{a}\right)\Gamma(b)}{\Gamma\left(1+\frac{1}{a}+b\right)},$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{\mathbf{X}}^2 = \frac{b\Gamma(1+2/a)\Gamma(b)}{\Gamma(1+b+2/a)} - \mu_1^2.$$

which cannot be found in closed form. Thus, we can only obtain the MME using numerical methods. The following R code shows how to calculate the MME numerically.

```
rm(list = ls()) # Delete memory
# Moments of order n of the Kumaraswamy(a,b) distribution
# https://en.wikipedia.org/wiki/Kumaraswamy_distribution

mn <- function(a,b,n){
  log.num <- log(b) + lgamma(1 + n/a) + lgamma(b)
  log.den <- lgamma(1 + b + n/a)
  return(exp(log.num-log.den))
}

#--------------------------------
# Method of Moments Estimates
#--------------------------------
library(nleqslv) # Required packages

# Function containing the implementation of the MME
# It solves a system of nonlinear equations

MME <- function(data){
  x.bar <- mean(data)
  x2.bar <- mean(data^2)

  fn <- function(par){
    mom1 <- mn(exp(par[1]),exp(par[2]),1) - x.bar
    mom2 <- mn(exp(par[1]),exp(par[2]),2) - mn(exp(par[1]),exp(par[2]),1)^2 - x2.bar + x.bar^2
    return(c(mom1,mom2))
  }

  est <- as.vector(exp(nleqslv(c(0,0), fn)$x))

  return(est)

}

# Examples
```

```
#--------------------------------
# Example 1
#--------------------------------
a0 <- 1
b0 <- 1
ns <- 1000

# Simulation of Kumaraswamy distributed data
set.seed(123)
dat <- (rbeta(ns, shape1 = 1, shape2 = b0))^(1/a0)

# MME
MME(dat)
## [1] 1.018995 1.032939
#--------------------------------
# Example 2
#--------------------------------
a0 <- 0.5
b0 <- 2
ns <- 1000

# Simulation of Kumaraswamy distributed data
set.seed(123)
dat <- (rbeta(ns, shape1 = 1, shape2 = b0))^(1/a0)

# MME
MME(dat)
## [1] 0.5059483 2.1418411
#--------------------------------
# Example 3
#--------------------------------
a0 <- 5
b0 <- 3
ns <- 1000

# Simulation of Kumaraswamy distributed data
set.seed(123)
dat <- (rbeta(ns, shape1 = 1, shape2 = b0))^(1/a0)

# MME
MME(dat)
## [1] 4.796745 2.915643
```

(b) We denote the function

$$f(\alpha) = \log\alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \log\alpha - \psi(\alpha),$$

where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function. Its derivative is

$$f'(\alpha) = \frac{1}{\alpha} - \psi'(\alpha),$$

where $\psi'(\alpha)$ is the trigamma function.

The Newton-Raphson update rule is

$$\alpha^{(t+1)} = \alpha^{(t)} + \frac{-f\left(\alpha^{(t)}\right) + \log\bar{X} - \frac{1}{n}\sum_{i=1}^{n}\log X_i}{f'\left(\alpha^{(t)}\right)}.$$

We can implement this as follows:

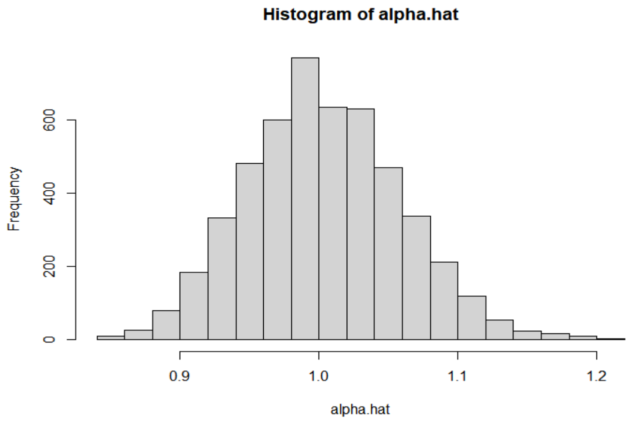
```
gamma.MLE = function(X) {
  ahat = compute.ahat(X)
  bhat = ahat/mean(X)

  return(c(ahat,bhat))
}

# estimate ahat by Newton-Raphson
compute.ahat = function(X) {
  a.prev = -Inf
  a = mean(X)^2/var(X)#initialguess

  #while not converged, do Newton-Raphson update
  while(abs(a-a.prev)>1e-12) {
    a.prev = a
    numerator = -f(a.prev)+log(mean(X))-mean(log(X))
    denominator = f.prime(a.prev)
    a = a.prev+numerator/denominator
  }

  return(a)
}

#define some helper functions
f=function(alpha) {

  return(log(alpha)-digamma(alpha))
}
f.prime=function(alpha) {
  return(1/alpha-trigamma(alpha))
}
```

(c) We run the simulation:

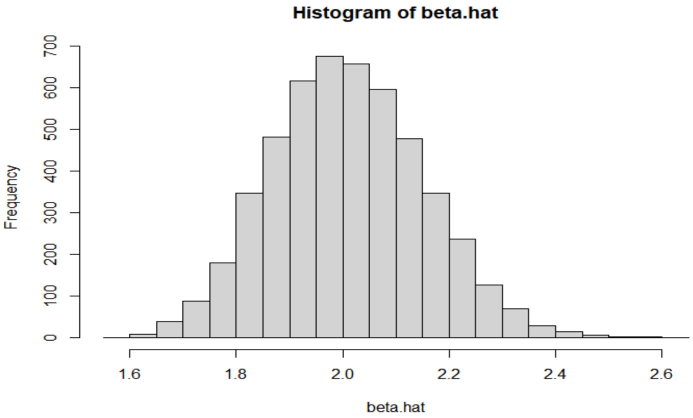
```
# In R you may simulate samples from Gamma (alpha, beta) using

X = rgamma(n , alpha, rate = beta )

# The sample variance  of a vector of values of X is given by var(X), and the sample covariance
# between two vectors of values X and Y (of the same length) is given by cov(X, Y).

n=500
n.reps=5000
alpha=1
beta=2

alpha.hat=numeric(n.reps)
beta.hat=numeric(n.reps)

for(i in 1:n.reps) {
  X=rgamma(n,shape=alpha,rate=beta)
  estimates=gamma.MLE(X)
  alpha.hat[i] = estimates[1]
  beta.hat[i] = estimates[2]
}
```

Here are the resulting histograms:

```r
hist(alpha.hat, breaks=20)
```



```r
hist(beta.hat, breaks=20)
```



The inverse of the Fisher information matrix was computed to be

$$I(\alpha,\beta)^{-1} = \frac{1}{\psi'(\alpha)\frac{\alpha}{\beta^2} - \frac{1}{\beta^2}} \begin{pmatrix} \frac{\alpha}{\beta^2} & \frac{1}{\beta} \\ \frac{1}{\beta} & \psi'(\alpha) \end{pmatrix}.$$

Plugging in $\alpha = 1, \beta = 2$ gives

$$I(1,2)^{-1} = \frac{4}{\psi'(1) - 1} \begin{pmatrix} 1/4 & 1/2 \\ 1/2 & \psi'(1) \end{pmatrix} \approx \begin{pmatrix} 1.551 & 3.101 \\ 3.101 & 10.202 \end{pmatrix}$$

using $\psi'(1) = \pi^2/6 \approx 1.645$.

Now, let's take a look at the empirical moments produced by the simulation. Here are the means:

```r
mean(alpha.hat) # should be close to 1
## [1] 1.003308
```

```r
mean(beta.hat) # should be close to 2
## [1] 2.010126
```

They are close to the true values of $\alpha = 1$ and $\beta = 2$, as expected. Now here are the variance and covariance terms:

```
var(alpha.hat)
## [1] 0.003079451

var(beta.hat)
## [1] 0.02048946

cov(alpha.hat,beta.hat)
## [1] 0.006190339
```

In order to compare to the Fisher information, we need to scale by n.

```
n*var(alpha.hat) #should be close to 1.551
## [1] 1.539726

n*var(beta.hat) #should be close to 10.202
## [1] 10.24473

n*cov(alpha.hat,beta.hat) #should be close to 3.101
## [1] 3.09517
```

Problem 27. (MLE in a misspecified model)

Suppose you fit the model Exponential(λ) to data $X_1, \ldots, X_n$ by computing the MLE $\hat{\lambda} = 1/\bar{X}$, but the true distribution of the data is $X_1, \ldots, X_n \overset{IID}{\sim}$ Gamma$(2, 1)$.

(a) Let $f(x \mid \lambda)$ be the PDF of the Exponential(λ) distribution, and let $g(x)$ be the PDF of the Gamma $(2, 1)$ distribution. Compute an explicit formula for the KL-divergence $D_{\mathrm{KL}}(g(x) \| f(x \mid \lambda))$ in terms of λ, and find the value λ^* that minimizes this KL-divergence.

(You may use the fact that if $X \sim$ Gamma(α, β), then $\mathbb{E}[X] = \alpha/\beta$ and $\mathbb{E}[\log X] = \psi(\alpha) - \log \beta$ where ψ is the digamma function.)

(b) Show directly, using the Law of Large Numbers and the Continuous Mapping Theorem, that the MLE $\hat{\lambda}$ converges in probability to λ^* as $n \to \infty$.

(c) Perform a simulation study for the standard error of $\hat{\lambda}$ with sample size $n = 500$, as follows: In each of $B = 10000$ simulations, sample $X_1, \ldots, X_n \overset{\text{IID}}{\sim}$ Gamma$(2, 1)$, compute the MLE $\hat{\lambda} = 1/\bar{X}$ for the exponential model, compute an estimate of the standard error of $\hat{\lambda}$ using the Fisher information $I(\hat{\lambda})$, and also compute the sandwich estimate of the standard error, $S_X / \left(\bar{X}^2 \sqrt{n}\right)$.

Report the true mean and standard deviation of $\hat{\lambda}$ that you observe across your 10000 simulations. Is the mean close to λ^* from part (a)? Plot a histogram of the 10000 estimated standard errors using the Fisher information, and also plot a histogram of the 10000 estimated standard errors using the sandwich estimate. Do either of these methods for estimating the standard error of $\hat{\lambda}$ seem accurate in this setting?

Solution.

(a) The KL-divergence is given by

$$\begin{aligned}
D_{\mathrm{KL}}(g(x) \| f(x \mid \lambda)) &= \mathbb{E}_g\left[\log \frac{g(X)}{f(X \mid \lambda)}\right] \\
&= \mathbb{E}_g\left[\log \frac{\frac{1}{\Gamma(2)} X e^{-X}}{\lambda e^{-\lambda X}}\right] \\
&= \mathbb{E}_g[-\log \Gamma(2) + \log X - X - \log \lambda + \lambda X] \\
&= -\log \Gamma(2) - \log \lambda + \mathbb{E}_g[\log X] + (\lambda - 1)\mathbb{E}_g[X] \\
&= -\log \Gamma(2) - \log \lambda + \psi(2) + 2(\lambda - 1).
\end{aligned}$$

Setting the derivative with respect to λ equal to 0, is minimized at $\lambda^* = 1/2$.

(b) By the Law of Large Numbers, $\bar{X} \to \mathbb{E}_g[X] = 2$ in probability, so $\hat{\lambda} = 1/\bar{X} \to 1/2$ in probability by the Continuous Mapping Theorem.

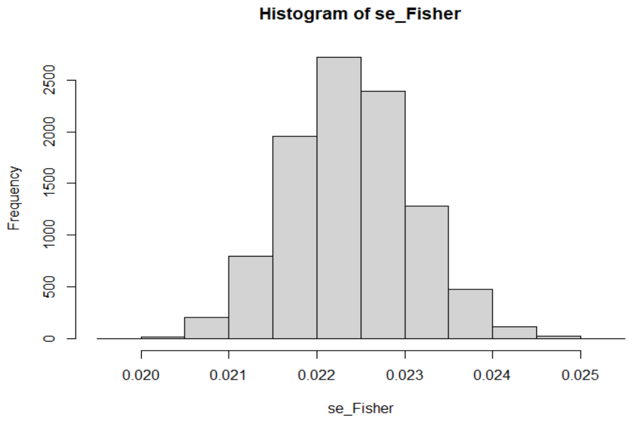
(c) The Fisher information in the exponential model is given by

$$I(\lambda) = -\mathbb{E}_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log f(X \mid \lambda)\right] = -\mathbb{E}_\lambda\left[\frac{\partial^2}{\partial\lambda^2}(\log\lambda - \lambda X)\right] = 1/\lambda^2$$

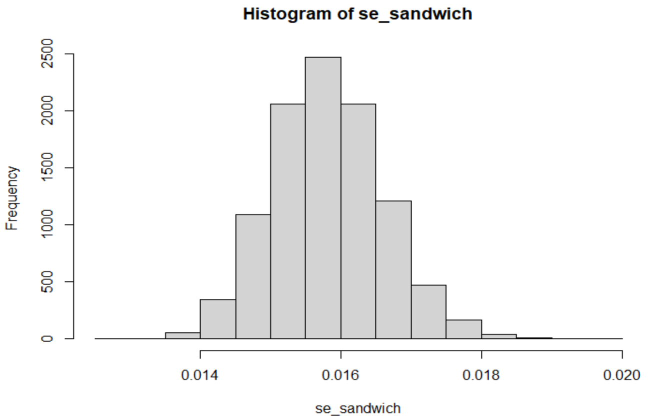
The corresponding plug-in estimate of the standard error is $\sqrt{\frac{1}{nI(\hat{\lambda})}} = \frac{1}{\bar{X}\sqrt{n}}$.

```
n = 500
B = 10000
lamb_hat = numeric(B)
se_Fisher = numeric(B)
se_sandwich = numeric(B)
for(i in 1:B){
  X = rgamma(n,2,rate=1)
  lamb_hat[i] = 1/mean(X)
  se_Fisher[i] = 1/(mean(X)*sqrt(n))
  se_sandwich[i] = sd(X)/(mean(X)^2*sqrt(n))
  }
print(mean(lamb_hat))
## [1] 0.5006192
print(sd(lamb_hat))
## [1] 0.01595704
```

```
hist(se_Fisher)
```



```
hist(se_sandwich)
```



The empirical mean and standard error of $\hat{\lambda}$ are 0.501 and 0.016; the mean is close to $\lambda^* = 0.5$ from part (a). The Fisher-information-based estimate of the standard error is incorrect – it estimates the standard error as approximately 0.022. The sandwich estimate of the standard error seems correct – it estimates the standard error as 0.016, with some variability in the third decimal place.

Problem 28. (Confidence intervals for a binomial proportion)

Let $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(p) be n tosses of a biased coin, and let $\hat{p} = \bar{X}$. In this problem, we will explore two different ways to construct a 95% confidence interval for p, both based on the Central Limit Theorem result

$$\sqrt{n}(\hat{p} - p) \to \mathcal{N}(0, p(1-p)). \tag{4}$$

(a) Use the plugin estimate $\hat{p}(1-\hat{p})$ for the variance $p(1-p)$ to obtain a 95% confidence interval for p. (This is the procedure discussed in class, yielding the Wald interval for p.)

(b) Instead of using the plugin estimate $\hat{p}(1-\hat{p})$, note that Eqn. (4) implies, for large n,

$$\mathbb{P}_p\left[-\sqrt{p(1-p)}Z_{\alpha/2} \le \sqrt{n}(\hat{p}-p) \le \sqrt{p(1-p)}Z_{\alpha/2}\right] \approx 1-\alpha.$$

Solve the equation $\sqrt{n}(\hat{p}-p) = \sqrt{p(1-p)}Z_{\alpha/2}$ for p in terms of $\hat{p}$, and solve the equation $\sqrt{n}(\hat{p}-p) = -\sqrt{p(1-p)}Z_{\alpha/2}$ for p in terms of $\hat{p}$, to obtain a different 95% confidence interval for p.

(c) Perform a simulation study to determine the true coverage of the confidence intervals in parts (a) and (b), for the 9 combinations of sample size $n = 10, 40, 100$ and true parameter $p = 0.1, 0.3, 0.5$. (For each combination, perform at least $B = 100,000$ simulations. In each simulation, you may simulate $\hat{p}$ directly from $\frac{1}{n}$ Binomial(n,p) instead of simulating $X_1, \ldots, X_n$.) Report the simulated coverage levels in two tables. Which interval yields true coverage closer to 95% for small values of n?

Solution.

(a) $\hat{p} \to p$ in probability, hence $\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \to \mathcal{N}(0,1)$ in distribution by Slutsky's Lemma. So for large n

$$\mathbb{P}\left[-Z_{0.025} \le \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \le Z_{0.025}\right] \approx 0.95.$$

We may rewrite the above as

$$\mathbb{P}\left[\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}Z_{0.025} \le p \le \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}Z_{0.025}\right] \approx 0.95,$$

so an approximate 95% confidence interval is $\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}Z_{0.025}$.

(b) The following conditions are equivalent:

$$-\sqrt{p(1-p)}Z_{\alpha/2} \le \sqrt{n}(\hat{p}-p) \le \sqrt{p(1-p)}Z_{\alpha/2}$$

$$\Updownarrow$$

$$n(\hat{p}-p)^2 \le p(1-p)Z_{\alpha/2}^2$$

$$\Updownarrow$$

$$\left(n+Z_{\alpha/2}^2\right)p^2 - \left(2n\hat{p}+Z_{\alpha/2}^2\right)p + n\hat{p}^2 \le 0.$$

This occurs when p is between the two real roots of the above quadratic equation, which are given by

$$\frac{2n\hat{p}+Z_{\alpha/2}^2 \pm \sqrt{\left(2n\hat{p}+Z_{\alpha/2}^2\right)^2 - 4\left(n+Z_{\alpha/2}^2\right)n\hat{p}^2}}{2\left(n+Z_{\alpha/2}^2\right)}.$$

Taking $\alpha = 0.05$ and simplifying the above, we obtain an approximate 95% confidence interval of

$$\frac{\hat{p}+\frac{Z_{0.025}^2}{2n} \pm Z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{Z_{0.025}^2}{4n^2}}}{1+\frac{Z_{0.025}^2}{n}}.$$

(c)
```
ns = c(10,40,100)
ps = c(0.1,0.3,0.5)
B=100000
z = qnorm(0.975)
for (n in ns) {
  for (p in ps) {
    cover_A = numeric(B)
    cover_B = numeric(B)
    for (i in 1:B) {
      phat = rbinom(1,n,p)/n
      U = phat+z*sqrt(phat*(1-phat)/n)
      L = phat-z*sqrt(phat*(1-phat)/n)
      if (p <= U && p >= L) {
        cover_A[i] = 1
      } else {
          cover_A[i] = 0
      }
      U = (phat+z^2/(2*n)+z*sqrt(phat*(1-phat)/n+z^2/(4*n^2)))/(1+z^2/n)
      L = (phat+z^2/(2*n)-z*sqrt(phat*(1-phat)/n+z^2/(4*n^2)))/(1+z^2/n)
      if (p <= U && p >= L) {
        cover_B[i] = 1
      } else {
          cover_B[i] = 0
      }
    }
    print(c(n,p,mean(cover_A),mean(cover_B)))
  }
}
```

For the interval from part (a), we obtain the following coverage probabilities:

	$p = 0.1$	$p = 0.3$	$p = 0.5$
$n = 10$	0.65	0.84	0.89
$n = 40$	0.91	0.93	0.92
$n = 100$	0.93	0.95	0.94

For the interval from part (b), we obtain the following coverage probabilities:

	$p = 0.1$	$p = 0.3$	$p = 0.5$
$n = 10$	0.93	0.92	0.98
$n = 40$	0.94	0.94	0.96
$n = 100$	0.94	0.94	0.94

The intervals from part (b) are more accurate when n is small.

Problem 29. (Power Comparisons)

Consider the problem of testing

$$H_0 : X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}(0,1)$$
$$H_1 : X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}(\mu,1)$$

at significance level $\alpha = 0.05$, where $\mu > 0$. We've seen four tests that may be applied to this problem, summarized below:

- Likelihood ratio test: Reject H_0 when $\bar{X} > \frac{1}{\sqrt{n}} Z_{0.05}$.
- t-test: Reject H_0 when $T := \sqrt{n}\bar{X}/S > t_{n-1;0.05}$, where $S^2 = \frac{1}{n-1}\sum_i \left(X_i - \bar{X}\right)^2$.
- Wilcoxon signed rank test: Reject H_0 when $W_+ > \frac{n(n+1)}{4} + \sqrt{\frac{n(n+1)(2n+1)}{24}} Z_{0.05}$, where W_+ is the Wilcoxon signed rank statistic.
- Sign test (from Problem 4 of TD-3): Reject H_0 when $S > \frac{n}{2} + \sqrt{\frac{n}{4}} Z_{0.05}$, where S is the number of positive values in $X_1, \ldots, X_n$.

(For the Wilcoxon and sign test statistics, we are using the normal approximations for their null distributions.) These tests are successively more robust to violations of the $\mathcal{N}(0,1)$ distributional assumption imposed by H_0.

(a) For $n = 100$, verify numerically that these tests have significance level close to α, in the following way: Perform 10,000 simulations. In each simulation, draw a sample of 100 observations from $\mathcal{N}(0,1)$, compute the above four test statistics $\bar{X}, T$, W_+, and S on this sample, and record whether each test accepts or rejects H_0. Report the fraction of simulations for which each test rejected H_0, and confirm that these fractions are close to 0.05.

(b) For $n = 100$, numerically compute the powers of these tests against the alternative H_1, for the values $\mu = 0.1, 0.2, 0.3,$ and 0.4. Do this by performing 10,000 simulations as in part (a), except now drawing each sample of 100 observations from $\mathcal{N}(\mu,1)$ instead of $\mathcal{N}(0,1)$. (You should be able to re-use most of your code from part (a).) Report your computed powers either in a table or visually using a graph.

(c) How do the powers of the four tests compare when testing against a normal alternative? Your friend says, "We should always use the testing procedure that makes the fewest distributional assumptions because we never know in practice, for example, whether the variance is truly 1 or whether data is truly normal." Comment on this statement. Rice says, "It has been shown that even when the assumption of normality holds, the [Wilcoxon] signed rank test is nearly as powerful as the t test. The [signed rank test] is thus generally preferable, especially for small sample sizes." Do your simulated results support this conclusion?

Solution.

(a) The code below runs the simulations for the null case ($\mu = 0$) as well as for $\mu = 0.1, 0.2, 0.3, 0.4$:

```
set.seed(1)
n = 100
B = 10000
for (mu in c(0,0.1,0.2,0.3,0.4)) {
  output.Z = numeric(B)
  output.T = numeric(B)
  output.W = numeric(B)
  output.S = numeric(B)
  for (i in 1:B) {
    X = rnorm(n, mean=mu, sd=1)
    if (mean(X) > 1/sqrt(n)*qnorm(0.95)) {
      output.Z[i] = 1
    } else {
        output.Z[i] = 0
    }
    T = t.test(X)$statistic
    if (T > qt(0.95,df=n-1)) {
      output.T[i] = 1
    } else {
        output.T[i] = 0
    }
    W = wilcox.test(X)$statistic
    if (W > n*(n+1)/4+sqrt(n*(n+1)*(2*n+1)/24)*qnorm(0.95)) {
      output.W[i] = 1
    } else {
        output.W[i] = 0
    }
    S = length(which(X>0))
    if (S > n/2+sqrt(n/4)*qnorm(0.95)) {
      output.S[i] = 1
    } else {
        output.S[i] = 0
    }
  }
  print(paste("mu = ", mu))
  print(paste("Z: ", mean(output.Z)))
  print(paste("T: ", mean(output.T)))
  print(paste("W: ", mean(output.W)))
  print(paste("S: ", mean(output.S)))
}
```

Under H_0 (case $\mu = 0$), we obtained the results:

Test stat	Type-I Error
Likelihood ratio test	0.0507
t-test	0.0505
Wilcoxon signed-rank test	0.053
Sign test	0.0441

(b) Under these alternatives, we obtained the results:

Test stat	Power over $\mathcal{N}(\mu, 1)$			
	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.4$
Likelihood ratio test	0.2631	0.6356	0.913	0.9895
t-test	0.261	0.6306	0.9085	0.9885
Wilcoxon signed-rank test	0.252	0.6164	0.8978	0.9847
Sign test	0.1805	0.4617	0.7521	0.9318

(c) The powers of the tests against $\mathcal{N}(\mu, 1)$ decrease as we increasingly relax the distributional assumptions (from $\mathcal{N}(0, 1)$ to $\mathcal{N}\left(0, \sigma^2\right)$ to any symmetric PDF f about 0 to any PDF f with median 0). The sign test makes the fewest distributional assumptions under H_0, but its power is substantially lower than the other three tests. Hence if we have good reason to believe that the data distribution under H_0 is symmetric (for example, if each data value is the difference of paired samples (X_i, Y_i), and (X_i, Y_i) should have the same distribution as (Y_i, X_i) under H_0), then we should at least opt for using the Wilcoxon test. The difference in powers between the Wilcoxon test, t-test, and the most-powerful likelihood ratio test is indeed very small, which supports Rice's claim (at least for the tested sample size $n = 100$).

Problem 30. (Testing gender ratios)

In a classical genetics study, Geissler (1889) studied hospital records in Saxony and compiled data on the gender ratio. The following table shows the number of male children in 6115 families having 12 children:

Number of male children	Number of families
0	7
1	45
2	181
3	478
4	829
5	1112
6	1343
7	1033
8	670
9	286
10	104
11	24
12	3

Let $X_1, \ldots, X_{6115}$ denote the number of male children in these 6115 families.

(a) Suggest two reasonable test statistics T_1 and T_2 for testing the null hypothesis

$$H_0 : X_1, \ldots, X_{6115} \overset{IID}{\sim} \text{Binomial}(12, 0.5).$$

(This is intentionally open-ended; try to pick T_1 and T_2 to "target" different possible alternatives to the above null.) Compute the values of T_1 and T_2 for the above data.

(b) Perform a simulation to simulate the null distributions of T_1 and T_2. (For example: Simulate 6115 independent samples $X_1, \ldots, X_{6115}$ from Binomial$(12, 0.5)$, and compute T_1 on this sample. Do this 1000 times to obtain 1000 simulated values of T_1. Do the same for T_2.) Plot the histograms of the simulated null distributions of T_1 and T_2. Using your simulated values, compute approximate p-values of the hypothesis tests based on T_1 and T_2 for the above data. For either of your tests, can you reject H_0 at significance level $\alpha = 0.05$?

(c) In this example, why might the null hypothesis H_0 not hold? (Please answer this question regardless of your findings in part (b).)

Solution.

(a) There are many possible answers. We may take T_1 to be the average number of male children per family,

$$T_1 = \bar{X},$$

and perform a two-sided test based on T_1 to check whether roughly half of the children are male. We may take T_2 to be Pearson's chi-squared statistic

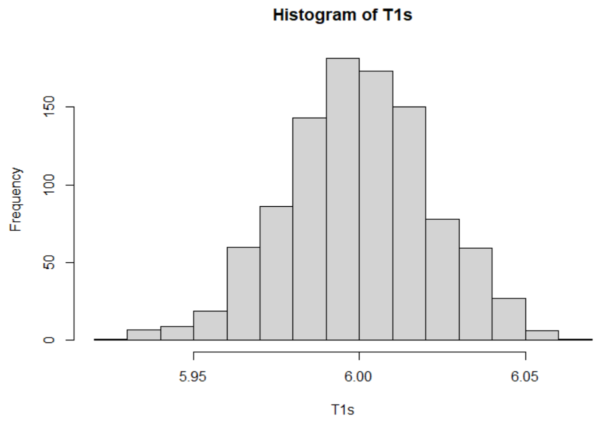
$$T_2 = \sum_{k=0}^{12} (O_k - E_k)^2 / E_k$$

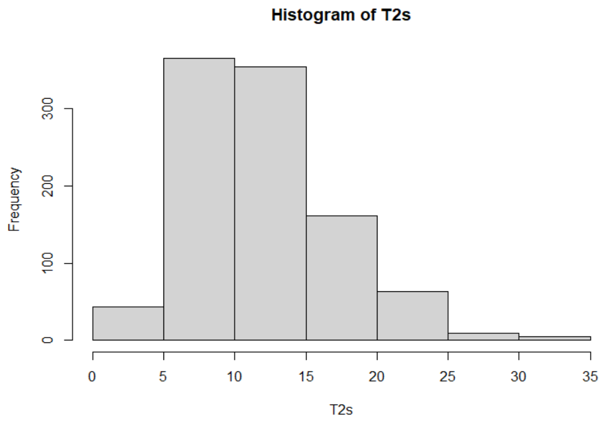
where O_k is the number of families with k male children and E_k is the expected number under the hypothesized binomial distribution, i.e. $E_k = 6115 \times \begin{pmatrix} 12 \\ k \end{pmatrix} (0.5)^{12}$, and perform a one-sided test that rejects for large T_2 to check whether the shape of the observed distribution of $X_1, \ldots, X_{6115}$ matches the shape of the binomial PDF.

(b) $\mathbb{R}$ code corresponding to the above T_1 and T_2 is as follows:

```
ks = seq(0,12)
counts = c(7,45,181,478,829,1112,1343,1033,670,286,104,24,3)
expected = 6115*choose(12,ks)*(0.5^12)

T1_obs = sum(ks*counts)/6115
T2_obs = sum((counts-expected)^2/expected)

T1s = numeric(1000)
T2s = numeric(1000)
for (i in 1:1000) {
  X = rbinom(6115, 12, 0.5)
  T1s[i] = mean(X)
  counts = numeric(13)
  for (k in 0:12) {
    counts[k+1] = length(which(X==k))
    }
  T2s[i] = sum((counts-expected)^2/expected)
  }
hist(T1s)
hist(T2s)
T1_pvalue = length(which(T1s<T1_obs))/1000 * 2
T2_pvalue = length(which(T2s>T2_obs))/1000
```

Histograms of the null distributions of T_1 and T_2 are below:





The values of the test statistics for the observed data are $T_1 = 5.77$ and $T_2 = 249$, which are both far outside the range of the simulated null distributions above. The simulated p-values for the two tests are both < 0.001, and there is strong evidence that H_0 is not correct.

(c) There may be biological and sociological reasons why H_0 is false. Biologically, the human male-to-female sex ratio at birth is not exactly $1 : 1$. The probability p that a child is male might also vary from family to family. The sexes of children within a family might be dependent; in particular, one source of dependence is the presence of identical twins.

Sociologically, there may be a relationship between family size and the sex ratio of children because the current sex ratio influences parents' decision of whether to have another child. Note that the given data is only for families with 12 children, which is quite large even for that time. There is a noticeable bias towards families with more girls than boys, which may be explained if parents tended to continue having children when their current children were predominantly female.

Problem 31. (Monte Carlo Simulations)

(a) Consider the problem of approximating tail probabilities of a Standard Normal distribution. Find $\mathcal{P}(X > 2.04)$ for $X \sim N(0,1)$ using Monte Carlo methods. While we can easily calculate this in R using the pnorm function, check how a Monte Carlo method would perform. How does the estimate change with the number of Monte Carlo samples?

(b) Consider the problem of calculating the expected present value of the payoff of a call option on a stock. Let $S(t)$ denote the price of the stock at time t. Consider a European call option, that is, the holder has the right to buy the stock at a fixed price K at a fixed time T in the future, $t = 0$ being the current time. If $S(T) > K$, the holder exercises the option for a profit of $S(T) - K$. If $S(T) \leq K$, the option expires worthless. The payoff at time T is thus

$$\max\{S(T) - K, 0\}.$$

The expected present value of this payoff is

$$E\left[e^{-rT}\max\{S(T) - K, 0\}\right],$$

where r is the compound interest rate. The evolution of the stock price $S(t)$ can be modeled via the Black-Scholes model expressed as

$$\frac{dS(t)}{S(t)} = rdt + \sigma dW(t),$$

where W is a standard Brownian motion. The solution to the above Stochastic Differential Equation is given by

$$S(T) = S(0)\exp\left\{\left(r - \frac{1}{2}\sigma^2\right)T + \sigma W(T)\right\},$$

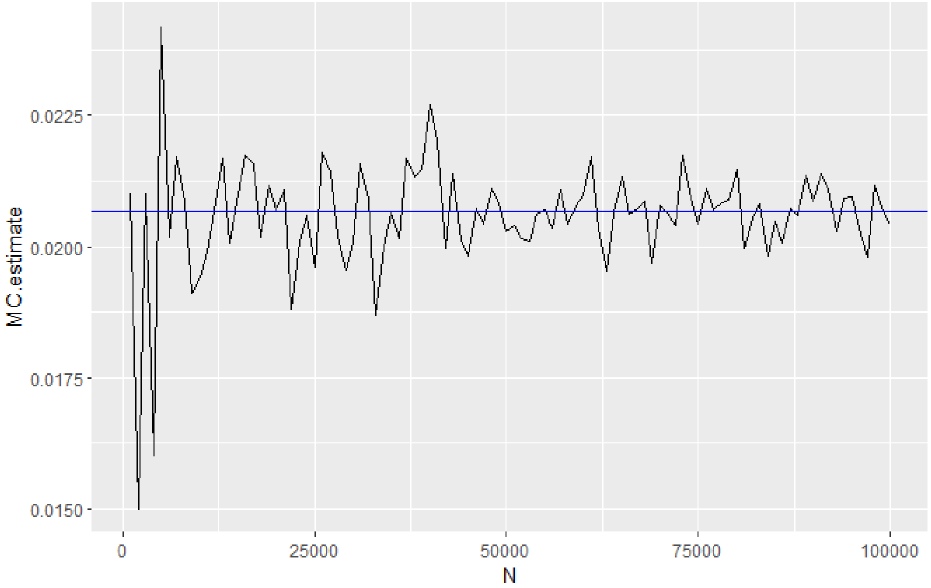
where $S(0)$ is the current price of the stock and $W(T) \sim N(0,T)$, that is, $W(T) = \sqrt{T}Z, Z \sim N(0,1)$. Thus, the logarithm of $S(T)$ is Normally distributed, or, in other words, the distribution of $S(T)$ is log-normal. The expected payoff $E\left[e^{-rT}\max\{S(T) - K, 0\}\right]$ is thus an integral w.r.t. to the lognormal density.

Can you perform Monte Carlo sampling to estimate this integral (that is, find the expected payoff)? This will require generating samples from the standard Normal distribution, followed by computing the value of the function inside the expectation and then taking the average. Can you also compute the estimated standard error of the estimator?

Solution.

(a)

```
N <- 1e06 # number of samples
t <- 2.04
Xsamp <- rnorm(N) #simulating from N(0,1)
tail.prob <- mean(Xsamp > t)
exact.tail.prob <- pnorm(t, lower.tail = FALSE)
c(MC.estimate = tail.prob, Exact.value = exact.tail.prob)
# Standard error of the estimate
MC.se <- sqrt(var(Xsamp > t)/N)
MC.se

library(tibble)
library(ggplot2)

MC_func <- function(N){
  Xsamp <- rnorm(N) # simulate from N(0,1)
  tail.prob <- mean(Xsamp > t)
  MC.se <- sqrt(var(Xsamp > t)/N)
  c(N = N, MC.estimate = tail.prob, MC.se = MC.se, Exact.value = exact.tail.prob)
}
N <- seq(1000, 100000, by = 1000)
out <- NULL
for (i in N) {
  out <- rbind(out, MC_func(i))
}
out <- as_tibble(out)
ggplot(out, mapping = aes(N,MC.estimate))
+ geom_line() + geom_hline(yintercept = exact.tail.prob, color = "blue")
```



(b)
```
payoff_func <- function(S_0, r, s, T, K, N=1e06){
  Z <- rnorm(N)
  S_T <- S_0*exp((r - s^2/2)*T + s*sqrt(T)*Z)
  payoff <- exp(-r*T)*pmax((S_T - K), 0) # pmax function in the formula
  ex_payoff <- mean(payoff)
  se_ex_payoff <- sqrt(var(payoff)/N)
  out <- c(expected.payoff = ex_payoff, SE = se_ex_payoff)
  return(out)
}
payoff_func(S_0 = 100, r = 0.07, s = 0.3, T = 5, K = 120)
```

Problem 32. (Simple Linear Regression)

A study was made on the effect of temperature on the yield of a chemical process. The following data were collected:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

(a) Assuming a model, $Y = \beta_0 + \beta_1 X + \epsilon$, what are the least square estimates of β_0 and β_1? What is the fitted equation?

(b) Construct the ANOVA table and test the hypothesis $H_0 : \beta_1 = 0$ with $\alpha = 0.05$.

(c) What are the confidence limits for β_1 at $\alpha = 0.05$?

(d) What are the confidence limits for the true mean value of Y when $X = 3$ at $\alpha = 0.05$?

(e) What are the confidence limits at $\alpha = 0.05$ level of significance for the difference between the true mean value of Y when $X_1 = 3$ and the true mean value of Y when $X_2 = -2$?

[Given, $F_{0.05,1,9} = 5.12,\ t_{0.05,9} = 1.833,\ t_{0.025,9} = 2.263,$]

Solution:

(a) The given data is (x_i, y_i); $i = 1, 2, \ldots, 11$. The linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ \text{ thus, } S = \sum_{i=1}^{11} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{11} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{11} x_i^2 - n\bar{x}^2} = \frac{158}{110} = 1.44; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 9.27$$

Thus, $\hat{y_i} = 9.27 + 1.44x_i$ is the fitted model.

(b) The sum of squares can be calculated as follows:

$$SS_T = \sum_{i=1}^{11} (y_i - \bar{y})^2 = 248.18, \ SS_{Reg} = \hat{\beta}_1^2 S_{xx} = 226.94, \ SS_{Res} = \sum_{i=1}^{11} (y_i - \hat{y_i})^2 = SS_T - SS_{Reg} = 22.23$$

Thus, the ANOVA table is

Source	DF	SS	MS = SS/DF	F-obs=MS$_{Reg}$/MS$_{Res}$	F-tab
Regression	1	226.94	226.94	96.17	$F_{0.05,1,9} = 5.12$
Residual	9	22.23	2.36		
Total	10	248.18			

To test the hypothesis $H_0 : \beta_1 = 0$ **vs** $H_1 : \beta_1 \neq 0$.

Here, we reject H_0, since $F_{obs} > F_{tab}$. Y and X are linearly dependent.

Alternative method for testing:

To test the hypothesis $H_0 : \beta_1 = 0$ **vs** $H_1 : \beta_1 \neq 0$.

Under $H_0 : \beta_1 = 0$, $\quad t = \frac{\hat{\beta}_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$. Hence, $t = \frac{1.44}{\sqrt{\frac{2.36}{110}}} = 9.83 > t_{0.05,9} = 1.833$. Hence, we reject $H_0 : \beta_1 = 0$.

Hence, $F \equiv t^2$, under H_0.

(c) The confidence interval of β_1 can be computed as follows:

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim \mathrm{N}(0,1) \quad [\sigma^2 \text{ is unknown and estimated by } MS_{Res}]$$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim \mathrm{t}_{n-2}$$

$$\Rightarrow P\left[-\mathrm{t}_{\alpha/2,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \leq \mathrm{t}_{\alpha/2,n-2}\right] = 1 - \alpha$$

$$\Rightarrow \hat{\beta}_1 - \mathrm{t}_{\alpha/2,n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + \mathrm{t}_{\alpha/2,n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}}$$

$$\Rightarrow 1.44 - 2.263 * 0.146 \leq \beta_1 \leq 1.44 + 2.263 * 0.146 \ \Rightarrow 1.11 \leq \beta_1 \leq 1.77.$$

(d) The confidence limits for $E(Y|X = 3)$ can be computed as follows:

95% CI for $E(Y$ at $X = x_0)$ is $\beta_0 + \beta_1 x_0$. An unbiased estimator $E(Y$ at $X = x_0)$ is $\hat{\beta}_0 + \hat{\beta}_1 x_0$.

Therefore, $\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim \mathrm{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$.

$$A = \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - (\beta_0 + \beta_1 x_0)}{\sqrt{MS_{Res}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}, \ \therefore P\left\{-t_{\alpha/2,n-2} \le A \le t_{\alpha/2,n-2}\right\} = 1-\alpha$$

$$\therefore \left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) \pm t_{\alpha/2,n-2} \times \sqrt{MS_{Res}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \text{ is the confidence interval for } \beta_0 + \beta_1 x_0.$$

Thus, replacing the values we get, the required confidence interval is $12.15 \le \beta_0 + \beta_1 x_0 \le 15.03$.

(e) The confidence limits for the difference between the true mean value of Y when $X_1 = 3$ and the true mean value of Y when $X_2 = -2$ can be computed as follows:

$$E(Y \text{ at } X_1 = 3) - E(Y \text{ at } X_2 = -2) \equiv Z_1 - Z_2.$$

Unbiased estimators of Z_1 and Z_2 are: $\hat{Z}_1 = \hat{\beta}_0 + \hat{\beta}_1 3$ and $\hat{Z}_2 = \hat{\beta}_0 + \hat{\beta}_1(-2)$.
Thus, $\hat{Z}_1 - \hat{Z}_2 = (\hat{\beta}_0 + \hat{\beta}_1 3) - (\hat{\beta}_0 + \hat{\beta}_1(-2)) = 5\hat{\beta}_1 = 7.20$ (point estimation).
Confidence interval for $Z_1 - Z_2$:
Compute $Var(\hat{Z}_1 - \hat{Z}_2) = Var(5\hat{\beta}_1) = \frac{25\sigma^2}{S_{xx}} = \frac{25\sigma^2}{110}$; thus $\hat{Z}_1 - \hat{Z}_2 \sim \mathrm{N}(Z_1 - Z_2, \frac{25\sigma^2}{110}) \Rightarrow \frac{(\hat{Z}_1-\hat{Z}_2)-(Z_1-Z_2)}{\sqrt{\frac{25MS_{Res}}{110}}} \sim \mathrm{t}_{n-2}$.
So, C.I. of $Z_1 - Z_2$ is:

$$(\hat{Z}_1 - \hat{Z}_2) - \mathrm{t}_{\alpha/2,9}\sqrt{\frac{25 \times 2.36}{110}} \le Z_1 - Z_2 \le (\hat{Z}_1 - \hat{Z}_2) + \mathrm{t}_{\alpha/2,9}\sqrt{\frac{25 \times 2.36}{110}} \Rightarrow 5.54 \le Z_1 - Z_2 \le 8.86.$$

Problem 33. (Fake Data in Linear Regression)

There are a few occasions where it makes sense to fit a model without an intercept β_0. If there were an occasion to fit the model $y = \beta x + \epsilon$ to a set of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the least square estimate of β would be

$$\hat{\beta} = b = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

Suppose you have a programmed calculator that will fit only the intercept model $y = \beta_0 + \beta_1 x + \epsilon$, but you want to fit a non-intercept model. By adding one more fake data point $(m\bar{x}, m\bar{y})$ to the data above, where $m = \frac{n}{(n+1)^{1/2}-1} = \frac{n}{a}$, say, and letting the calculation fit $y = \beta_0 + \beta_1 x + \epsilon$, can you estimate β by using b?

Solution:

Given the data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we fit: $y = \beta x + \epsilon$, where the least square estimate of β is calculated as follows:

$$\mathrm{SSE} = \mathrm{S} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta} x_i)^2; \quad \frac{\partial S}{\partial \beta} = 0 \Rightarrow \sum (y_i - \hat{\beta} x_i) x_i = 0 \Rightarrow \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}.$$

For the computer-fitted model with intercept, we have:

$$y = \beta_0 + \beta_1 x + \epsilon \text{ where, the LSEs are } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \text{ and } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Adding one more data point: $(m\bar{x}, m\bar{y})$; $m = \frac{n}{(n+1)^{1/2}-1} = \frac{n}{a} \Rightarrow (a+1)^2 = (n+1)$.
Thus the new dataset is given by $(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n), (u_{n+1}, v_{n+1}) = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), (m\bar{x}, m\bar{y})$.

Hence we have

$$\bar{u} = \frac{n\bar{x} + (n/a)\bar{x}}{n+1} = \frac{n\bar{x} + m\bar{x}}{n+1} = \frac{n\bar{x}(\frac{a+1}{a})}{n+1} = \frac{n\bar{x}}{(a+1)^2} \times \frac{a+1}{a} = \frac{n\bar{x}}{a(a+1)} \text{ similarly } \bar{v} = \frac{n\bar{y}}{a(a+1)}$$

$$S_{uu} = \sum_{i=1}^{n} x_i^2 + (m\bar{x})^2 - (n+1)\bar{u}^2 = \sum_{i=1}^{n+1} u_i^2 - (n+1)\bar{u}^2 = \sum_{i=1}^{n} x_i^2 + \frac{n^2}{a^2}\bar{x}^2 - (n+1)\frac{n^2\bar{x}^2}{a^2(a+1)^2} = \sum_{i=1}^{n} x_i^2$$

$$S_{uv} = \left(\sum_{i=1}^{n} x_i y_i (m\bar{x}m\bar{y})\right) - (n+1)\left[\frac{n\bar{x}}{a(a+1)}\right]\left[\frac{n\bar{y}}{a(a+1)}\right] = \sum_{i=1}^{n} x_i y_i + m^2\bar{x}\bar{y} - \frac{(n+1)m^2\bar{x}\bar{y}}{(a+1)^2}$$

$$= \sum_{i=1}^{n} x_i y_i + m^2\bar{x}\bar{y} - m^2\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i$$

For the programmed calculation: $\hat{\beta}_1 = \frac{S_{uv}}{S_{uu}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

Problem 34. (Multiple Linear Regression)

Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ for the data given below. Provide an ANOVA table and perform the partial F-tests to test $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ for $i = 1, 2$; given the other variable is already in the model. Comment on the relative contributions of the variables X_1 and X_2, depending on whether they enter the model first or second. Find the regression equation.

X_1	-5	-4	-1	2	2	3	3
X_2	5	4	1	-3	-2	-2	-3
Y	11	11	8	2	5	5	4

Solution:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Fit the model with both X_1, X_2. OLS estimates can be obtained using $\hat{\beta} = (X^T X)^{-1} X^T Y$

$$X = \begin{pmatrix} 1 & -5 & 5 \\ 1 & -4 & 4 \\ 1 & -1 & 1 \\ 1 & 2 & -3 \\ 1 & 2 & -2 \\ 1 & 3 & -2 \\ 1 & 3 & -3 \end{pmatrix}; \quad \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 68 & -67 \\ 0 & -67 & 63 \end{pmatrix}^{-1} \begin{pmatrix} 46 \\ -66 \\ 69 \end{pmatrix} = \begin{pmatrix} \frac{46}{7} \\ 1 \\ 2 \end{pmatrix} \therefore \hat{Y} \quad = \frac{46}{7} + X_1 + 2X_2$$

Source	DF	SS	MS	F-obs	F-tab
Regression	2	72.00	36.00	83.72	$F_{0.05,2,4} = 6.94$
Residual	4	1.71	0.43		
Total	6	73.71			

$SS_T = \sum \left(Y_i - \bar{Y}\right)^2 = 73.71, \quad e_i = y_i - \hat{y}_i, \quad SS_{Res} = \sum_{i=1}^{n} e_i^2, \quad H_0 : \beta_1 = \beta_2 = 0$ vs. $H_1 : H_0$ is not true. To test this hypothesis, we have $F = 83.72 > F_{0.05,2,4} = 6.94 \Rightarrow H_0$ is rejected.

Partial F-tast:- Fit the model with X_1 as $Y = \beta_0 + \beta_1 X_1 + \epsilon$, $\hat{Y} = \frac{46}{7} - \frac{66}{68} X_1$ and test the hypothesis $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$

$$F = \frac{\left\{SS_{\text{Reg}}\,(\text{Full}) - SS_{\text{Reg}}\,(\text{Restricted Model})\right\}/1}{MS_{\text{Res}}} = \frac{72 - 64.06}{0.43} = 18.53 > F_{0.05,1,4} = 7.71$$

$\therefore H_0$ is rejected at 5% level of significance, i.e., X_2 is significant in the presence of X_1.

Partial F-tast:- Fit the model with X_2 as $Y = \beta_0 + \beta_2 X_2 + \epsilon$, $\hat{Y} = \frac{46}{7} - \frac{69}{68} X_2$ and test the hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

$$F = \frac{\{SS_{\text{Reg}}\,(\text{Full}) - SS_{\text{Reg}}\,(\text{Restricted Model})\}\,/1}{MS_{\text{Res}}} = \frac{72 - 70.01}{0.43} = 4.64 < F_{0.05,1,4} = 7.71$$

$\therefore$ H_0 is not rejected at 5% level of significance, i.e., X_1 is not significant in the presence of X_2.

Implication:- If X_2 is in the model, we donot need X_1. If X_1 is in the model, X_2 helps out significantly. Then X_2 is clearly more useful variable and it explains $R^2 = \frac{70.01}{73.71} = 95\%$ of the total variability in Y about mean, where as X_1 alone explains $R^2 = \frac{64.06}{79.71} = 86\%$ of total variability in Y about mean. And X_1 and X_2 together explain $\frac{72.00}{73.71} = 97\%$ of total variability.

Note: In this problem, $X_1 + X_2 \approx 0$ and the presence of multicollinearity (X_1 and X_2 are not independent), and that is why Partial F-test suggests $\beta_0 = 0$ is accepted.

Problem 35. (More on Multiple Linear Regression)

Given a 2-variables linear regression problem $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \epsilon$, yield the following

$$X^T X = \begin{bmatrix} 33 & 0 & 0 \\ 0 & 40 & 20 \\ 0 & 20 & 60 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 132 \\ 24 \\ 92 \end{bmatrix}, \quad \text{and} \quad \sum (Y - \bar{Y})^2 = 150.$$

(a) What is the sample size?

(b) Write the normal equations and solve for the regression coefficients.

(c) Estimate the standard error of β_2 and test the hypothesis that $\beta_2 = 0$

(d) Compute R^2 and interpret it. Also, interpret the value of regression coefficients.

(e) Predict the value of y given $x_1 = -4$ and $x_2 = 2$

(f) Comment on the possibilities of any regressors being a dummy variable.

Solution:

(a) The variance-covariance matrix of the design matrix $X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$ is given by

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1} x_{i2} \\ \sum x_{i2} & \sum x_{i1} x_{i2} & \sum x_{i2}^2 \end{bmatrix} = \begin{bmatrix} 33 & 0 & 0 \\ 0 & 90 & 20 \\ 0 & 20 & 60 \end{bmatrix}$$

$\therefore$ The sample size, $n = 33$.

$$X^T Y = \begin{bmatrix} 1 & \ldots & 1 \\ x_{21} & \ldots & x_{n1} \\ x_{12} & \ldots & x_{n2} \end{bmatrix}_{3\times n} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \sum x_{i2} y_i \end{bmatrix} = \begin{bmatrix} 132 \\ 24 \\ 92 \end{bmatrix}$$

(b) We know the normal equation is given by $\hat{\beta} = (X^T X)^{-1} X^T Y \Rightarrow (X^T X)\hat{\beta} = X^T Y$. Thus we have

$$(X^T X)\hat{\beta} = X^T Y \Rightarrow \begin{bmatrix} 33 & 0 & 0 \\ 0 & 40 & 20 \\ 0 & 20 & 60 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 132 \\ 24 \\ 92 \end{bmatrix} \Rightarrow \begin{bmatrix} 33 & 0 & 0 \\ 0 & 40 & 20 \\ 0 & 0 & 50 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 832 \\ 24 \\ 80 \end{bmatrix}$$

$$\left[\text{using Gaussian elimination:} R_3' = R_3 - \frac{1}{2} R_2\right]$$

Thus we obtain $b_3 = 1.60, 40b_2 + 20b_3 = 24 \Rightarrow 40b_2 = -8 \Rightarrow b_2 = -0.20, 33b_1 = 132 \Rightarrow b_1 = 4$. Hence, the regression equation is $\hat{y} = 4 - 0.20x_1 + 1.60x_2$

(c)

$$\left|X^T X\right| = 33(2400 - 400) = 66000,$$

$$\text{S.E. of } b_2 = S\sqrt{c_{22}} \;\text{ where } S = \sqrt{\frac{RSS}{n-k}} = \sqrt{\frac{7.6}{2}} = \sqrt{3.8} \text{ and } C_{22} = \frac{\det(\text{co-factor})}{|X^T X|} = \frac{1980}{66000} = 0.003$$

$$\text{SE}(b_2) = \sqrt{0.003 \times 3.8} = 0.106$$

To test the hypothesis $H_0 : b_2 = 0$ vs. $H_1 : b_2 \neq 0$, we use the following test statistic $t = \frac{b_2}{\text{SE}(b_2)} = \frac{-0.20}{0.106} = 1.88 < 4.303$. We reject H_0.

(d) $\text{TSS} = \text{RSS} + \text{SS}_{\text{Reg}}$

Given that $TSS = 150$, we can calculate $\text{SS}_{\text{Reg}} = \beta^{*T}(X^T Y)^* = \begin{bmatrix} -0.20 & 1.60 \end{bmatrix} \begin{bmatrix} 24 \\ 92 \end{bmatrix} = 142.4$ and thus we have $RSS = 150 - 142.4 = 7.6$

Thus, the coefficient of determination R^2 is given by $R^2 = 142.4/150 = 0.9493$ explains 94.93% of the total variability in the response variable.

(e) The predicted value of y for the given values of $x_1 = -4$ and $x_2 = 2$ is given by $\hat{y} = 4 - 0.20(-4) + 1.60(2) = 8$.

Problem 36. (Presence of Multicollinearity)

Can we use the data given below to get a unique fit to the model $Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

X_1	-4	3	1	4	-3	-1
X_2	1	2	3	4	5	6
X_3	3	-5	-4	-8	-2	-5
Y	7.4	14.7	13.9	18.2	12.1	14.8

Solution:

The LSE of the regression model $Y = X\beta + \epsilon$ is given by $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$

$$X = \left[\begin{array}{cccc} x_0 & x_1 & x_2 & x_3 \\ 1 & -4 & 1 & 3 \\ 1 & 3 & 2 & -5 \\ 1 & 1 & 3 & -4 \\ 1 & 4 & 4 & -8 \\ 1 & -3 & 5 & -2 \\ 1 & -1 & 6 & -5 \end{array}\right]$$

Note that, $X_1 + X_2 + X_3 = 0$ which implies that $\left(X^T X\right)$ is a singular matrix, i.e., $|\left(X^T X\right)| = 0$. No, we can't compute $\hat{\beta}$ uniquely here.

Problem 37. (Residual Vs. Response Variable)

Show that in linear regression with a β_0 term in the model:

(a) The correlation between the vector e and Y is $(1 - R^2)^{1/2}$. This result implies that it is a mistake to find defective regressions by a plot of residuals e_i versus observations Y_i as this always shows a slope.

(b) Show further that the correlation between e and $\hat{Y}$ is zero.

Solution:

(a)

$$\text{Cor}(e, Y) = \frac{\sum (e_i - \bar{e})\left(Y_i - \bar{Y}\right)}{\sqrt{\sum (e_i - \bar{e})^2 \sum \left(Y_i - \bar{Y}\right)^2}}$$

$$\sum (e_i - \bar{e})\left(Y_i - \bar{Y}\right) = \sum e_i \left(Y_i - \bar{Y}\right) = \sum e_i Y_i = Y^T e \text{ [since, } \bar{e} = 0 \text{ if } \beta_0 \text{ is in the model]}$$

Note that,

$$Y = X\beta + \epsilon,\ \hat{\beta} = \left(X^T X\right)^{-1} X^T Y \ \Rightarrow \hat{Y} = X\left(X^T X\right)^{-1} X^T Y = X\hat{\beta} = HY, \text{ where } H = X\left(X^T X\right)^{-1} X^T$$

$$e = Y - \hat{Y} = (I - H)Y \Rightarrow e^T e = Y^T (I-H)^T (I-H) Y = Y^T (I-H) Y = Y^T e \text{ [since, } H^2 = H]$$

Thus we have

$$\text{Cor}(e, Y) = \frac{Y^T e}{\sqrt{(e^T e) SS_T}} = \sqrt{\frac{e^T e}{SS_T}} = \sqrt{\frac{SS_{Res}}{SS_T}} = \sqrt{1 - \frac{SS_{Reg}}{SS_T}} = \sqrt{1 - R^2}.$$

Implication: This is why we plot $\hat{Y}_i$ and e_i but not Y_i and e_i since they are correlated.

(b) The correlation between e and $\hat{Y}$ can be computed as follows:

$$\text{Cov}(e, \hat{Y}) = \sum (e_i - \bar{e})\left(\hat{Y}_i - \bar{\hat{Y}}\right) = e^T \hat{Y} = Y^T (I - H) H Y = Y^T (H - H^2) Y = 0 \text{ [since, } \hat{Y} = HY, e = (I-H)Y]$$

$$\text{Cor}(e, \hat{Y}) = 0.$$

Problem 38. (Multiple Coefficient of Determination)

Prove that the multiple coefficient of determination R^2 equals the square of the correlation between Y and $\hat{Y}$.

Solution: We have to prove that $[\text{Cor}(Y, \hat{Y})]^2 = R^2 = \frac{SS_{Reg}}{SS_T}$

Note that, $\sum e_i = 0 \Rightarrow \sum \left(Y_i - \hat{Y}_i\right) = 0 \Rightarrow \sum Y_i = \sum \hat{Y}_i \Rightarrow \bar{Y} = \bar{\hat{Y}}$ [since, $Y_i = \hat{Y}_i + e_i$]

$$\text{Cor}\left(e, \hat{Y}\right) = 0 \Rightarrow \text{Cov}\left(e, \hat{Y}\right) = 0 \Rightarrow \sum \left(\hat{Y}_i - \bar{Y}\right) e_i = 0$$

$$\text{r}_{Y\hat{Y}} = \frac{\sum \left(\hat{Y}_i - \bar{\hat{Y}}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2 \sum \left(Y_i - \bar{Y}\right)^2}} = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)\left(\hat{Y}_i - \bar{Y}\right) + \sum \left(\hat{Y}_i - \bar{Y}\right) e_i}{\sum \left(Y_i - \bar{Y}\right)\left(\hat{Y}_i - \bar{Y}\right)^2} = \sqrt{\frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2}} = \frac{SS_{Reg}}{SS_T} = \sqrt{R^2}$$

Problem 39. (Polynomial Regression)

A new born baby was weighted weekly. Twenty such weights are shown below, recorded in ounces. Fit to the data, using orthogonal polynomials, a polynomial model of degree justified by the accuracy of the figures, that is, test as you go along for the significance of the linear, quadratic and so fourth, terms.

No. of weeks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Weights	141	144	148	150	158	161	166	170	175	181	189	194	196	206	218	229	234	242	247	257

Solution:

We wish to fit the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_k X^k + \epsilon$$

Polynomial fitting using orthogonal polynomial:

$$Y = \alpha_0 + \alpha_1 P_1(X) + \alpha_2 P_2(X) + \ldots + \alpha_k P_k(X) + \epsilon; \quad \left[P_i(x) \text{ are } i^{th} \text{ orthogonal polynomial}\right].$$

The parameters can be estimated as $\hat{\alpha_0} = \bar{y}$ and $\hat{\alpha_j} = \frac{\sum P_j(x_i) y_i}{\sum P_j^2(x_i)}$. Given the data (x_i, y_i), we compute total variability (SS_T), then build a model to explain this variability. Suppose that $SS_{Reg}(\alpha_1)$ represents how much of the total variability in Y is explained by the linear term, similarly for cubic and quadratic.

$$SS_{Reg}(\alpha_1) = \hat{\alpha}_1 \sum_{i=1}^{n} y_i P_1(x_i) = 25438.75 \text{ (linear term)}, \; SS_{Reg}(\alpha_2) = \hat{\alpha}_2 \sum_{i=1}^{n} y_i P_2(x_i) = 489 \text{ (quadratic term)}$$

$$SS_{Reg}(\alpha_3) = \hat{\alpha}_3 \sum_{i=1}^{n} y_i P_3(x_i) = 1.15 \text{ (cubic term)}, \; SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 26018$$

Source	DF	SS	MS	F-obs	F-tab
Regression (α_1)	1	25438.75	25438.75	4558.98	$F_{0.05,1,6} = 4.49$
Regression (α_2)	1	489	489	84.63	
Regression (α_3)	1	1.15	1.15	0.21	
Residual	16	89.30	5.58		
Total	19	26018			

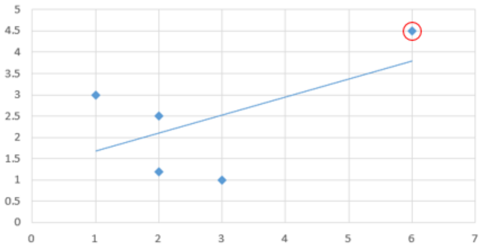
As we can see, the linear term explains the major part of variability, and α_2 is also significant (F-value). Thus the final model is $\hat{y} = 136.227 + 2.68x + 0.167x^2$, (quadratic fit). If residual SS_{Res} is large, you may check for fourth-degree polynomial fitting.

Problem 40. (Outliers)

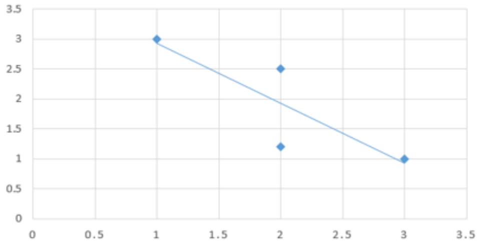
If you are asked to fit a straight line to the data $(X, Y) = (1, 3), (2, 2.5), (2, 1.2), (3, 1),$ and $(6, 4.5)$. What would you do about it?

Solution:

From the figure below, we can observe that $(6, 4.5)$ is an influential observation. The regression model has a positive slope when the influential observation is included and a negative slope when it is not.



With influential observation



Without influential observation

Recommendation: You can ignore influential observation if it's small in number. Some observations between $X = 3$ and $X = 6$ would be useful here.

Problem 41. (Calculating R^2 but Friends Forever)

Your friend says he/she has fitted a plane to $n = 33$ observations on (X_1, X_2, Y) and his/ her overall regression (given β_0) is just significant at the $\alpha = 0.05$ level. You ask him/ her for R^2 value, but s/he does not know. You work it out for him/ her based on what s/he has told you.

Solution:

The equation of the plane is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Here we have n = 33 observations

Source	DF	SS	MS	F-obs	F-tab
Regression	2	SS_{Reg}	MS_{Reg}	F	$F_{0.05,2,30} = 3.32$
Residual	30	SS_{Res}	MS_{Res}		
Total	32	SS_T			

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{SS_{Reg}}{SS_{Reg} + SS_{Res}} = \frac{SS_{Reg}/MS_{Res}}{\frac{SS_{Reg}}{MS_{Res}} + \frac{SS_{Res}}{MS_{Res}}} = \frac{2MS_{Reg}/MS_{Res}}{\frac{2MS_{Reg}}{MS_{Res}} + \frac{30MS_{Res}}{MS_{Res}}} = \frac{2F}{2F + 30} = \frac{2 \times 3.32}{(2 \times 3.32) + 30} = 0.1812$$

$\therefore$ 18% of the total variability is explained by the model.

Implication: Thus, R^2 is a "good" measure to measure the "goodness of fit" even when the statistical test suggests that the regression is significant.

Problem 42. (Understanding Regression Output)

You are given a regression printout that shows a planar to fit X_1, X_2, X_3, X_4, X_5 plus an intercept term obtained from a set of 50 observations. The overall F for regression is ten times as high as the 5% upper-tail F percentage point. How big is R^2?

Solution:

The regression equation is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

The ANOVA table is We know,

Source	DF	SS	MS	F-obs	F-tab
Regression	5	SS_{Reg}	MS_{Reg}	F	$F_{0.05,5,44} = 2.43$
Residual	44	SS_{Res}	MS_{Res}		$10 \times F_{0.05,5,44} = 24.3$
Total	49	SS_T			

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{SS_{Reg}}{SS_{Reg} + SS_{Res}} = \frac{SS_{Reg}/MS_{Res}}{\frac{SS_{Reg}}{MS_{Res}} + \frac{SS_{Res}}{MS_{Res}}} = \frac{5MS_{Reg}/MS_{Res}}{\frac{5MS_{Reg}}{MS_{Res}} + \frac{44MS_{Res}}{MS_{Res}}} = \frac{5F}{5F + 44} = \frac{5 \times 24.3}{(5 \times 24.3) + 44} = 0.7343.$$

Conclusion: 73.43% of the total variability in the response variable is explained by the fitted model.

Problem 43. (Weighted Least Square)

Consider the simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where the variance of ϵ_i is proportional to x_i^2, i.e., $\mathrm{V}(\epsilon_i) = \sigma^2 x_i^2$ (assumption of constant variance is NOT satisfied).

(a) Suppose that we use these transformation $y' = \frac{y}{x}$ and $x' = \frac{1}{x}$. Is this a variance-stabilizing transformation?

(b) What are the relationships between the parameters in the original and the transformed model?

(c) Suppose we use the method of weighted least squares with $w_i = \frac{1}{x_i^2}$. Is this equivalent to the transformation introduced in part (a)?

Solution:

(a)

$$
\begin{aligned}
&y_i = \beta_0 + \beta_1 x_i + \epsilon_i \longrightarrow \text{ original model}\\
&\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\epsilon_i}{x_i},\ y_i' = \beta_0 x_i' + \beta_1 + \epsilon_i' \longrightarrow \text{ Transformed model}
\end{aligned}
$$

Now, $\text{Var}(y_i') = \text{Var}\left(\frac{\epsilon_i}{x_i}\right) = \frac{\sigma^2 x_i^2}{x_i^2} = \sigma^2$. Yes, it's a variance-stabilizing transformation.

(b) Slope in the transformed model became an intercept in the transformed model and vice-versa.

(c) **Weighted LS function:**

$$
\begin{aligned}
S_{OLS}(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \implies \text{ Oridinary Least Squares}\\
S_{WLS}(\beta_0, \beta_1) &= \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2 \implies \text{ Weighted Least Squares}\\
&= \sum_{i=1}^{n} \frac{1}{x_i^2} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \left(\frac{y_i}{x_i} - \frac{\beta_0}{x_i} - \beta_1\right)^2
\end{aligned}
$$

Calculate β_0 and β_1 by minimizing $S_{WLS}(\beta_0, \beta_1)$ in Weighted Least Squares method.

For the transformed model, OLS estimates are:

$$
S^*(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i' - \beta_0 x_i' - \beta_1)^2 = \sum_{i=1}^{n} \left(\frac{y_i}{x_i} - \frac{\beta_0}{x_i} - \beta_1\right)^2
$$

Calculate β_0 and β_1 by minimizing $S^*(\beta_0, \beta_1)$ for the transformed model.

Here, $S^*(\beta_0, \beta_1) \equiv S_{WLS}(\beta_0, \beta_1)$.

Problem 44. (Autoregression)

Consider the simple linear regression model $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ where the errors are generated by second-order autoregressive process

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + z_t,$$

where z_t is an NID $(0, \sigma_z^2)$ random variable, and ρ_1 and ρ_2 are autocorrelation parameters. Discuss how the Cochrane-Orcutt iterative procedure could be used in this situation. What transformations would be appropriate for the variables y_t and x_t? How would you estimate the parameters ρ_1 and ρ_2?

Solution:

$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ where $\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + z_t$; $z_t \overset{\text{ind}}{\sim} N(0, \sigma_z^2)$ [In OLS-based SLR, we assume $\epsilon_t \overset{\text{ind}}{\sim} \text{N}(0, \sigma^2)$ which is not true here.]

$$
\begin{aligned}
y_t \to y_t' = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} &= (\beta_0 + \beta_1 x_t + \epsilon_t) - \rho_1(\beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1}) - \rho_2(\beta_0 + \beta_1 x_{t-2} + \epsilon_{t-2})\\
&= (\beta_0 - \rho_1\beta_0 - \rho_2\beta_0) + \beta_1(x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}) + (\epsilon_t - \rho_1\epsilon_{t-1} - \rho_2\epsilon_{t-2}) = \beta_0' + \beta_1 x_t' + z_t
\end{aligned}
$$

Now z_t's are independent and $z_t \sim N(0, \sigma_z^2)$. But, (y_t', x_t') cannot be directly used as

$$y_t' = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} \text{ and } x_t' = x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}$$

are functions of unknown parameters ρ_1 and ρ_2.

We know $\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + z_t$; (How to estimate ρ_1 and ρ_2?).

- Fit $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ using OLS and obtain e_i (ignoring autocorrelation)

- Regress e_i on e_{i-1} and e_{i-2}, i,e., $e_i = \rho_1 e_{i-1} + \rho_2 e_{i-2} + z_t$ (MLR with two regressors)
- Compute $S(\rho_1, \rho_2) = \sum (e_i - \rho_1 e_{i-1} - \rho_2 e_{i-2})^2$ and minimize $S(\rho_1, \rho_2)$ and obtain $\hat{\rho}_1$ and $\hat{\rho}_2$.

  $$\frac{\partial S}{\partial \rho_1} = 0 \Rightarrow \sum (e_i - \rho_1 e_{i-1} - \rho_2 e_{i-2}) e_{i-1} = 0, \ \frac{\partial S}{\partial \rho_2} = 0 \Rightarrow (e_i - \rho_1 e_{i-1} - \rho_2 e_{i-2}) e_{i-2} = 0$$

  These will generate the LSE of $\hat{\rho}_1$ and $\hat{\rho}_2$.
- $y'_t = y_t - \hat{\rho}_1 y_{t-1} - \hat{\rho}_2 y_{t-2}$ and $x'_t = x_t - \hat{\rho}_1 x_{t-1} - \hat{\rho}_2 x_{t-2}$ and apply OLS to the transformed data $y'_t = \beta'_0 + \beta_1 x'_t + z_t$, where $z_t \overset{\text{ind}}{\sim} N(0, \sigma_z^2)$.
- Final fitted model is: $\hat{y}'_t = \hat{\beta}'_0 + \hat{\beta}_1 x_t$.

Problem 45. (Durbin-Watson Test)

The following 24 residuals from a straight-line fit are equally spaced and given in time sequential order. Is there any evidence of lag-1 serial correlation?

$$8, -5, 7, 1, -3, -6, 1, -2, 10, 1, -1, 8, -6, 1, -6, -8, 10, -6, 9, -3, 3, -5, 1, -9$$

Use a two-sided test at level $\alpha = 0.05$.

Solution:

The correlation between the residuals, $e_i, \quad i = 1(1)24$ is $\text{Cor}(e_u, e_{u+1}) = \rho$. (if $\rho \neq 0$, there is autocorrelation).

We test the hypothesis $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$, thus H_0 indicates that there is no lag-1 autocorrelation. To perform this test, we compute the Durbin - Watson test statistic:

$$d = \frac{\sum_{u=2}^{24} (e_u - e_{u-1})^2}{\sum_{u=1}^{24} e_u^2} = \frac{2225}{834} = 2.67 \Rightarrow 4 - d = 1.33$$

Now, compare with d_L and d_U values from d table. For $\alpha = 0.025$. (two-sided test) $n = 24, k = 1$ (since straight line fit with one regressor variable) $d_L = 1.16, d_U = 1.33$

- If $d < d_L$ and $4 - d < d_L$: reject H_0. Here, we accept H_0 as $d = 2.67 > 1.16$ (then is no lag-1 autocorrelation).
- If $d > d_U$ and $4 - d > d_U$: accept H_0. Here, we accept H_0 as $d = 2.67 > 1.33$ (then is no lag-1 autocorrelation).

Thus, there is no lag-1 autocorrelation/serial correlation in the data.

Problem 46. (Nonlinear Regression)

Estimate the parameters α & β in the non-linear model $Y = \alpha + (0.49 - \alpha)e^{-\beta(X-8)} + \epsilon$ from the following observations:

X	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42
Y	0.490	0.475	0.450	0.433	0.458	0.423	0.407	0.407	0.407	0.405	0.393	0.405	0.400	0.395	0.400	0.390	0.407	0.390

Solution:

The problem is to estimate α, β of the non-linear model using the data. The residual sum of square can be written as

$$S(\alpha, \beta) = \sum_u (y_u - f(x_u, \alpha, \beta))^2 = \sum_u (y_u - \alpha - (0.49 - \alpha)e^{-\beta(x_u - 8)}))^2$$

$$f(x_u, \alpha, \beta) = \alpha + (0.49 - \alpha)e^{-\beta(x_u - 8)}$$

[Since, f is nonlinear, we solve the system of nonlinear eqns by via Taylor series approximation of nonlinear into linear one]

$$\frac{\partial f}{\partial \alpha} = 1 - e^{-\beta(x_u - 8)}, \quad \frac{\partial f}{\partial \beta} = -(0.49 - \alpha)e^{-\beta(x_u - 8)}(x_u - 8)$$

Linearization: Taylor series expansion of $f\,(x_u, \alpha, \beta)$ about the point (α_0, β_0) is

$$\begin{aligned}
f(x_u;\alpha,\beta) &= f(x_u,\alpha_0,\beta_0) + (1-e^{-\beta_0(x_u-8)})(\alpha-\alpha_0) + [-(0.49-\alpha_0)e^{-\beta_0(x_u-8)}\,(x_u-8)](\beta-\beta_0)\\
&= f_u^0 + z_{1u}^0(\alpha-\alpha_0) + z_{2u}^0(\beta-\beta_0) \text{ [linear function from nonlinear function using Taylor approximmation]}\\
Y_u &= f_u^0 + z_{1u}^0(\alpha-\alpha_0) + z_{2u}^0(\beta-\beta_0) + \epsilon_u\\
\Rightarrow Y_u - f_u^0 &= z_{1u}^0(\alpha-\alpha_0) + z_{2u}^0(\beta-\beta_0) + \epsilon_u \longrightarrow \text{[MLR model]}\\
\Rightarrow Y_0 &= z_0\theta_0 + \epsilon \longrightarrow \text{[In matrix form]}\\
\Rightarrow \hat{\theta}_0 &= (z_0^T z_0)^{-1} z_0^T Y_0 \text{ is the least square estimate}
\end{aligned}$$

where

$$Y_0 = \begin{bmatrix} Y_1 - f_1^0 \\ \vdots \\ Y_n - f_n^0 \end{bmatrix},\ z_0 = \begin{bmatrix} z_{11}^0 & z_{21}^0 \\ \vdots & \vdots \\ z_{1n}^0 & z_{2n}^0 \end{bmatrix},\ \theta_0 = \begin{bmatrix} \alpha-\alpha_0 \\ \beta-\beta_0 \end{bmatrix},\ \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- If we begin the iteration with initial guess $\alpha_0 = 0.30, \beta_0 = 0.02$

$$z_0 = \begin{bmatrix} 1-e^{-\beta_0(x_1-8)} & -(0.49-\alpha_0)(x_1-8)e^{-\beta_0(x_1-8)} \\ \vdots & \vdots \\ 1-e^{-\beta_0(x_n-8)} & -(0.49-\alpha_0)(x_n-8)e^{-\beta_0(x_n-8)} \end{bmatrix}$$

iteration	α_j	β_j
0	0.30	0.02
1	0.84	0.10

- Iteration continues and obtain α_{j+1} and β_{j+1}.
- This process continue until $|\alpha_{j+1} - \alpha_j| < \delta$ and $|\beta_{j+1} - \beta_j| < \delta = 0.0001$. So, we stop here.

iteration	α_j	β_j
2	0.3901	0.1004
3	0.3901	0.1016
4	0.3901	0.1016

Problem 47. (Statistician's Dilemma) Look at these data. I don't know whether to fit two straight lines, one straight line, or something else. How can I solve this dilemma?

Solution:

If we attach a dummy variable Z to distinguish the two groups (such that $Z = 0$ for set A and $Z = 1$ for set B), we can look at all 4 possibilities.

$$Y = (\beta_0 + \beta_1 X) + Z(\alpha_0 + \alpha_1 X) + \epsilon = \beta_0 + \beta_1 X + \alpha_0 Z + \alpha_1 XZ + \epsilon$$

Thus the $\underset{\sim}{X}$ matrix becomes

$$\underset{\sim}{X} = \left[\begin{array}{cccc} \mathbf{1} & \mathbf{X} & \mathbf{Z} & \mathbf{XZ} \\ 1 & 8 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 12 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 9 & 1 & 9 \\ 1 & 7 & 1 & 7 \\ 1 & 8 & 1 & 8 \\ 1 & 6 & 1 & 6 \end{array}\right]$$

Thus we have $Y = X\beta + \epsilon \Rightarrow \beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ and $\hat{\beta} = (X^TX)^{-1}X^TY \Rightarrow \hat{Y} = 1.142 + 0.506X - 0.0418Z - 0.036XZ$.

Case: Test if a single line is sufficient i.e., $H_0 : \alpha_0 = \alpha_1 = 0$

$$\begin{aligned} F &= \frac{\{SS_{Reg}(\textbf{Full}) - SS_{Reg}(\textbf{Restricted Model})\} / \{df(\textbf{Full}) - df(\textbf{Restricted Model})\}}{MS_{Res}} \\ &= \frac{0.1818/(3-1)}{0.3272/4} = 1.11 < F_{0.05,2,4} \end{aligned}$$

Hence, we fail to reject H_0 and can go for a single straight-line fit.

Problem 48. (Principal component Regression)

Let $\underset{\sim}{x}$ be a vector of p random variables and α_k is a vector of p constants and we write $\alpha_k^T\underset{\sim}{x} = \sum_{j=1}^{p} \alpha_{kj}x_j$. Also, let S be the (known) sample covariance matrix for the random variable $\underset{\sim}{x}$. For $k = 1, 2$, show that the k^{th} principal component is given by $z_k = \alpha_k^T\underset{\sim}{x}$ where α_k is an eigenvector of S corresponding to its k^{th} largest eigenvalue λ_k.

Solution:

To calculate the k^{th} principal component $z_k = \alpha_k^T\underset{\sim}{x}$ we apply the constrained maximuization method. By using the Lagrange multiplier we maximize var $\left(\alpha_k^T\underset{\sim}{x}\right) = \alpha_k^T\Sigma\alpha_k$ subject to $\alpha_k^T\alpha_k = 1$. We maximize the function

$$\alpha_k^T\Sigma\alpha_k - \lambda\left(\alpha_k^T\alpha_k - 1\right)$$

w.r.t. to α_k by differentiating w.r.t. to α_k. This results in

$$\begin{aligned} \frac{d}{d\alpha_k}\left(\alpha_k^T\Sigma\alpha_k - \lambda_k\left(\alpha_k^T\alpha_k - 1\right)\right) &= 0 \\ \Sigma\alpha_k - \lambda_k\alpha_k &= 0 \\ \Sigma\alpha_k &= \lambda_k\alpha_k \end{aligned}$$

This is recognizable as an eigenvector equation where α_k is an eigenvector of Σ and λ_k is the associated eigenvalue.

If we recognize that the quantity to be maximized

$$\alpha_k^T\Sigma\alpha_k = \alpha_k^T\lambda_k\alpha_k = \lambda_k\alpha_k^T\alpha_k = \lambda_k$$

then we should choose λ_k to be as big as possible. So, calling λ_1 the largest eigenvector of Σ and α_1 the corresponding eigenvector then the solution to

$$\Sigma\alpha_1 = \lambda_1\alpha_1$$

is the 1^{st} principal component of $\underset{\sim}{x}$.

The second PC, $\alpha_2^T\underset{\sim}{x}$ maximizes $\alpha_2^T\Sigma\alpha_2$ subject to being uncorrelated with $\alpha_1^T\underset{\sim}{x}$. The uncorrelation constraint can be expressed using any of these equations

$$\begin{aligned} \text{cov}\left(\alpha_1^T\underset{\sim}{x}, \alpha_2^T\underset{\sim}{x}\right) &= \alpha_1^T\Sigma\alpha_2 = \alpha_2^T\Sigma\alpha_1 = \alpha_2^T\lambda_1\alpha_1^T \\ &= \lambda_1\alpha_2^T\alpha = \lambda_1\alpha_1^T\alpha_2 = 0 \end{aligned}$$

Of these, if we choose the last we can write a Lagrangian to maximize α_2

$$\alpha_2^T\Sigma\alpha_2 - \lambda_2\left(\alpha_2^T\alpha_2 - 1\right) - \phi\alpha_2^T\alpha_1$$

Differentiation of this quantity w.r.t. α_2 (and setting the result equal to zero) yields

$$\begin{aligned} \tfrac{d}{d\alpha_2}\left(\alpha_2^T\Sigma\alpha_2 - \lambda_2\left(\alpha_2^T\alpha_2 - 1\right) - \phi\alpha_2^T\alpha_1\right) &= 0 \\ \Rightarrow \Sigma\alpha_2 - \lambda_2\alpha_2 - \phi\alpha_1 &= 0 \end{aligned}$$

If we left multiply α_1 into this expression

$$\begin{aligned} \alpha_1^T\Sigma\alpha_2 - \lambda_2\alpha_1^T\alpha_2 - \phi\alpha_1^T\alpha_1 = 0 \\ \Rightarrow 0 - 0 - \phi 1 = 0 \end{aligned}$$

then we can see that ϕ must be zero and that when this is true we are left with

$$\Sigma\alpha_2 - \lambda_2\alpha_2 = 0$$

Clearly

$$\Sigma\alpha_2 - \lambda_2\alpha_2 = 0$$

is another eigenvalue equation and the same strategy of choosing α_2 to be the eigenvector associated with the second largest eigenvalue yields the second PC of $\underset{\sim}{x}$, namely $\alpha_2^T\underset{\sim}{x}$.

Problem 49. (Shrinkage Methods in Regression)

Show that $\|\hat{\beta}_{\text{Ridge}}\|$ increases as its tuning parameter $\lambda \to 0$. Does the same property hold for the LASSO regression?

Solution

SVD Decomposition of $X = U_{n\times p}D_{p\times p}V_{p\times p}^T$

$$\hat{\beta}_{\text{Ridge}} = \left(X^TX + \lambda I\right)^{-1}X^TY = \left(VD^2V^T + \lambda I\right)^{-1}VDU^TY = \left(V\left(D^2 + \lambda I\right)V^T\right)^{-1}VDU^TY = V^T\left(D^2 + \lambda I\right)^{-1}DU^TY.$$

$$\|\hat{\beta}_{\text{Ridge}}\|_2^2 = Y^TUD(D^2 + \lambda I)^{-1}(D^2 + \lambda I)^{-1}DU^TY = (U^TY)^T[D(D^2 + \lambda I)^{-2}D](U^TY) = \sum_{j=1}^{p}\frac{d_j^2(U^TY)_j^2}{(d_j^2 + \lambda)^2};$$

where $D\left(D^2 + \lambda I\right)^{-1}D$ represents a diagonal matrix with elements $\frac{d_j^2}{\left(d_j^2+\lambda\right)^2}$.
Therefore, we see that $\|\hat{\beta}_{\text{Ridge}}\|$ increases as its tuning parameter $\lambda \to 0$. Recall the dual form of LASSO as defined below:

$$\begin{aligned} \hat{\beta}_{\text{Lasso}} &= \arg\min_{\beta}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 \\ &\text{s.t. } \sum_{j=1}^{p}|\beta_j| \leq t \\ \hat{\beta}_{\text{Lasso}} &= \arg\min_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \end{aligned}$$

It is easy to see that t and λ have an inverse relationship; therefore, as $\lambda \to 0, t$ increases and so does the norm of optimal solutions.

Problem 50. (Logistic Regression)

Consider a two-class logistic regression problem with $x \in \mathbb{R}$. Characterize the maximum-likelihood estimates of the slope and intercept parameter if the sample x_i for the two classes are separated by a point $x_0 \in \mathbb{R}$. Generalize this result to (a) $x \in \mathbb{R}^p$ and (b) more than two classes.

Solution

Without loss of generality, suppose that $x_0 = 0$ and that the coding is $y = 1$ for $x_i > 0$ and $y = 0$ for $x_i < 0$. Now, suppose that

$$p(x;\beta) = \frac{\exp\left\{\beta x + \beta_0\right\}}{1 + \exp\left\{\beta x + \beta_0\right\}}$$

so that

$$1 - p(x;\beta) = \frac{1}{1 + \exp\left\{\beta x + \beta_0\right\}}$$

Since $x_0 = 0$ is the boundary then $p\left(x_0\right) = 1 - p\left(x_0\right)$ then $\beta_0 = 0$. Therefore,

$$p(x;\beta) = \frac{\exp\{\beta x\}}{1 + \exp\{\beta x\}}$$

so that

$$1 - p(x;\beta) = \frac{1}{1 + \exp\{\beta x\}}.$$

Therefore, the likelihood function

$$L(\beta; y, x) = \prod_{i=1}^{N} p\left(x_i;\beta\right)^{y_i}\left[1 - p\left(x_i;\beta\right)\right]^{1-y_i} = \prod_{i=1}^{N}\left[\frac{p\left(x_i;\beta\right)}{1 - p\left(x_i;\beta\right)}\right]^{y_i}\left[1 - p\left(x_i;\beta\right)\right] = \prod_{i=1}^{N}\left[\exp\left\{\beta x_i\right\}\right]^{y_i}\left[1 + \exp\left\{\beta x_i\right\}\right]$$

so that the log-likelihood function

$$l(\beta; y, x) = \sum_{i=1}^{N} y_i\left[\beta x_i\right] - \log\left[1 + \exp\left\{\beta x_i\right\}\right]$$

Taking the derivative with respect to β and substituting in the proper coding of y_i gives

$$\begin{aligned}\frac{dl(\beta; x, y)}{d\beta} &= \sum_{i=1}^{N} x_i\left(y_i - \frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right)\\ &= \sum_{x_i > 0} x_i\left(1 - \frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right) - \sum_{x_i < 0} x_i\left(\frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right)\\ &= \sum_{x_i > 0} x_i - \sum_{x_i > 0} x_i\left(\frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right) - \sum_{x_i < 0} x_i\left(\frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right).\end{aligned}$$

Setting the above equal to zero gives

$$\sum_{x_i > 0} x_i = \sum_{i=1}^{N} x_i\left(\frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right).$$

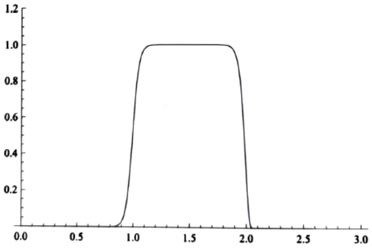
Clearly, for any data set $\{x_i\}_{i=1}^{N}$ we must have that $\beta \to \infty$ for the above equality to hold.

(b) Now, suppose that there are K classes such that x_1 seperates classes one and two, x_2 seperates classes two and three, and so on to x_{K-1} that seperates classes $K-1$ and K with $-\infty = x_0 < x_1 < x_2 < \ldots < x_{K-1} < x_K = \infty$. Now, define probabilities

$$\begin{aligned}p_1(x;\beta) &= \frac{\exp\left\{\beta_1 x + \beta_{01}\right\}}{1 + \sum_{j=1}^{K-1}\exp\left\{\beta_j x + \beta_{0j}\right\}}\\ p_2(x;\beta) &= \frac{\exp\left\{\beta_2 x + \beta_{02}\right\}}{1 + \sum_{j=1}^{K-1}\exp\left\{\beta_j x + \beta_{0j}\right\}}\\ &\vdots\\ p_{K-1}(x;\beta) &= \frac{1}{1 + \sum_{j=1}^{K-1}\exp\left\{\beta_j x + \beta_{0j}\right\}}.\end{aligned}$$

Now, suppose that the coding is $y_i = 1$ if $x_{j-1} < x_i < x_j$ and $y_i = 0$ otherwise for observation $i = 1, \ldots, N$ and class $j = 1, \ldots, K$. Therefore, the likelihood function

$$L(\beta; y, x) = \prod_{j=1}^{K}\prod_{i=1}^{N_j}\left[p_j\left(x_i;\beta\right)\right]^{y_i}$$



where N_j is the number of observations in class j, so that the log-likelihood function

$$\begin{aligned} l(\beta; y, x) &= \sum_{j=1}^{K-1}\sum_{i=1}^{N_j} y_i \log\left[\frac{\exp\{\beta_j x_i+\beta_{0j}\}}{1+\sum_{j=1}^{K-1}\exp\{\beta_j x+\beta_{0j}\}}\right] + \sum_{i=1}^{N_k} y_i \log\left[\frac{1}{1+\sum_{j=1}^{K-1}\exp\{\beta_j x_i+\beta_{0j}\}}\right] \\ &= \sum_{j=1}^{K-1}\sum_{i=1}^{N_j} y_i\left[\beta_j x_i+\beta_{0j}\right] - \sum_{j=1}^{K}\sum_{i=1}^{N_j} y_i \log\left[1+\sum_{j=1}^{K-1}\exp\{\beta_j x_i+\beta_{0j}\}\right] \end{aligned}$$

Now, we determine the values of β_{0j}. First, note that β_{0j} is a function of β_j, x_{j-1}, and x_j. So that the expression $p(x;\beta)$ maintains proper form, for $x_{j-1} < x < x_j$ we define

$$\begin{aligned} p\left(x;\beta_j\right) &= \frac{\exp\{\beta_j\left(x-x_{j-1}\right)\}-\exp\{\beta_j\left(x-x_j\right)\}}{1+\sum_{j=1}^{K-1}\exp\{\beta_j x_i+\beta_{0j}\}} \\ &= \frac{\exp\{\beta_j x\}\left[\exp\{\beta_j x_{j-1}\}-\exp\{\beta_j x_j\}\right]}{1+\sum_{j=1}^{K-1}\exp\{\beta_j x_i+\beta_{0j}\}} \\ &= \frac{\exp\{\beta_j x+\beta_{0j}\}}{1+\sum_{j=1}^{K-1}\exp\{\beta_j x_i+\beta_{0j}\}} \end{aligned}$$

where $\beta_{0j} = \log\left[\exp\{\beta_j x_{j-1}\}-\exp\{\beta_j x_j\}\right]$. The reason for the beginning step of the formulation above is because, for example, when $x \in (x_1, x_2)$ so that x classifies to class two, the probability function appears as in the following figure, where it was assumed that $x_1 = 1$ and $x_2 = 2$.

Now, taking the derivative with respect to $\beta = (\beta_1, \ldots, \beta_{K-1})$ and substituting in the proper coding of y_i gives

$$\begin{aligned} \frac{dl(\beta; x, y)}{d\beta_j} &= \sum_{i=1}^{N_j} x_i + \sum_{i=1}^{N_j} \frac{\exp\{\beta_j x_{j-1}\}x_{j-1}-\exp\{\beta_j x_j\}x_j}{\exp\{\beta_j x_{j-1}\}-\exp\{\beta_j x_j\}} \\ &\quad - \sum_{i=1}^{N_j}\left[x_i + \frac{\exp\{\beta_j x_{j-1}\}x_{j-1}-\exp\{\beta_j x_j\}x_j]}{\exp\{\beta_j x_{j-1}\}-\exp\{\beta_j x_j\}}\right]\left(\frac{\exp\{\beta_j x_i-\beta_{0j}\}}{1+\sum_{j=1}^{K-1}\exp\{\beta_j x_i+\beta_{0j}\}}\right) \end{aligned}$$

Note that the $\frac{\exp\{\beta_j x_{j-1}\}x_{j-1}-\exp\{\beta_j x_j\}x_j}{\exp\{\beta_j x_{j-1}\}-\exp\{\beta_j x_j\}}$ term in the above is a constant in the sum over $i = 1, \ldots, N_j$. Therefore, setting the above equal to zero for each $j = 1, \ldots, K-1$ and solving for β_j similarly gives the maximum likelihood estimators to the two-class case that $\beta_j \to \infty$.

Some Useful Formulae

- The volume of a sphere of radius r is $\frac{4}{3}\pi r^3$.
- A Maclaurin series is a Taylor series expansion of a function about 0,

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \ldots + \frac{f^{(n)}(0)}{n!}x^n + \ldots.$$

- The generic Chernoff bound for a random variable X is attained by applying Markov's inequality to e^{tX}. This gives a bound in terms of the moment-generating function of X. For every $t \geq 0$:

$$\Pr(X \geq a) = \Pr\left(e^{t\cdot X} \geq e^{t\cdot a}\right) \leq \frac{\mathrm{E}\left[e^{t\cdot X}\right]}{e^{t\cdot a}}.$$

- Uniform distribution on the interval $(0,1)$ has pdf

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

- The standard normal distribution has a pdf

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \quad \text{for } x \in \mathbb{R}$$

- Beta density with parameters α and β :

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$E[X] = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$\alpha = 1$ and $\beta = 1$ gives the Uniform distribution on $(0,1)$.

- Gamma distribution with parameters $r > 0$ and $\lambda > 0$:

$$f(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \quad E[X] = \frac{r}{\lambda}, \mathrm{Var}(X) = \frac{r}{\lambda^2}$$

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^r \quad \text{for } t < \lambda.$$

$r = 1$ gives the Exponential distribution with mean $1/\lambda$.

- Exponential distribution with parameter $\lambda > 0$:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \quad E[X] = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Normal distribution with mean μ and variance $\sigma^2 > 0$:

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2} & \text{if } x \in \mathbb{R} \\ 0 & \text{otherwise.} \end{cases} \quad E[X] = \mu, \quad \mathrm{Var}(X) = \sigma^2$$

$M_X(t) = e^{\mu t + t^2\sigma^2/2}$ for $t \in \mathbb{R}$. $\mu = 0$ and $\sigma^2 = 1$ gives the standard normal distribution.

- Poisson distribution with mean $\lambda > 0$:

$$f(k) = \begin{cases} \frac{\lambda^k}{k!}e^{-\lambda} & \text{if } k = 0, 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases} \quad E[X] = \lambda, \quad \mathrm{Var}(X) = \lambda$$

$$M_X(t) = \exp\left\{\lambda\left(e^t - 1\right)\right\} \quad \text{for } t \in \mathbb{R}.$$

Some Useful Distributional Properties

1. If $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then
$$Z = \frac{Y-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

2. $Y \sim \mathcal{N}(0,1) \Longrightarrow Y^2 \sim \chi^2(1)$

3. $Y \sim \mathcal{N}\left(\mu, \sigma^2\right) \Longrightarrow aY+b \sim \mathcal{N}\left(a\mu+b, a^2\sigma^2\right)$

4. $Y \sim \mathcal{U}(0,1) \Longrightarrow -\ln Y \sim$ exponential(1)

   Generalization: $Y \sim \mathcal{U}(0,1) \Longrightarrow -\beta \ln Y \sim$ exponential(β)

   Related: $Y \sim$ beta$(\alpha, 1) \Longrightarrow -\ln Y \sim$ exponential$(1/\alpha)$

   Related: $Y \sim$ beta$(1, \beta) \Longrightarrow -\ln(1-Y) \sim$ exponential$(1/\beta)$

5. $Y \sim$ exponential$(\alpha) \Longrightarrow Y^{1/m} \sim$ Weibull(m, α)

   Related: $Y \sim$ Weibull$(m, \alpha) \Longrightarrow Y^m \sim$ exponential(α)

6. $Y \sim \mathcal{N}\left(\mu, \sigma^2\right) \Longrightarrow e^Y \sim$ lognormal $\left(\mu, \sigma^2\right)$ or equivalently if $U \sim$ lognormal $\left(\mu, \sigma^2\right) \Longrightarrow \ln U \sim \mathcal{N}\left(\mu, \sigma^2\right)$

7. $Y \sim$ beta$(\alpha, \beta) \Longrightarrow 1-Y \sim$ beta(β, α)

8. $Y \sim \mathcal{U}(-\pi/2, \pi/2) \Longrightarrow \tan Y \sim$ Cauchy

9. $Y \sim$ gamma$(\alpha, \beta) \Longrightarrow cY \sim$ gamma$(\alpha, \beta c)$, where $c > 0$

   Special case: $2Y/\beta \sim \chi^2(2\alpha)$

10. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim}$ Bernoulli $(p) \Longrightarrow \sum Y_i \sim b(n,p)$

11. $Y_i \sim$ gamma $(\alpha_i, \beta), i = 1, 2, \ldots, n$ (mutually independent)
$$\Longrightarrow \sum Y_i \sim \text{gamma}\left(\sum \alpha_i, \beta\right)$$

    Special case: $\alpha_i = 1$, for $i = 1, 2, \ldots, n$. Then $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim}$ exponential$(\beta) \Longrightarrow \sum Y_i \sim$ gamma(n, β)

    Special case: $\alpha_i = \nu_i/2, \beta = 2$. Then $Y_i \sim \chi^2(\nu_i), i = 1, 2, \ldots, n$ (mutually independent) $\Longrightarrow \sum Y_i \sim \chi^2\left(\sum \nu_i\right)$

    Combination: If $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim}$ exponential(β), then
$$\frac{2\sum Y_i}{\beta} \sim \chi^2(2n)$$

12. $Y_i \sim$ Poisson $(\lambda_i), i = 1, 2, \ldots, n$ (mutually independent)
$$\Longrightarrow \sum Y_i \sim \text{Poisson}\left(\sum \lambda_i\right)$$

13. $Y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right), i = 1, 2, \ldots, n$ (mutually independent)
$$\Longrightarrow \sum a_i Y_i \sim \mathcal{N}\left(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2\right)$$

    Special case: $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$, for $i = 1, 2, \ldots, n$. Then $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
$$\Longrightarrow \sum a_i Y_i \sim \mathcal{N}\left(\mu \sum a_i, \sigma^2 \sum a_i^2\right)$$

Special case of iid result: If $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, then $\bar{Y} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$

Special case of iid result: If $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, then $\sum Y_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right)$

14. If $Y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right), i = 1, 2, \ldots, n$ (mutually independent), then

$$Z_i = \frac{Y_i - \mu_i}{\sigma_i} \sim \mathcal{N}(0,1),$$

for $i = 1, 2, \ldots, n$. Therefore, $U = \sum Z_i^2 \sim \chi^2(n)$ because $Z_1^2, Z_2^2, \ldots, Z_n^2$ are iid $\chi^2(1)$

15. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim}$ geometric$(p) \Longrightarrow U = \sum Y_i \sim$ nib(n, p)

16. $Y_1, Y_2 \overset{iid}{\sim} \mathcal{N}(0,1) \Longrightarrow U = Y_1/Y_2 \sim$ Cauchy

17. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim}$ exponential$(\beta) \Longrightarrow Y_{(1)} \sim$ exponential (β/n)

18. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim}$ Weibull$(m, \alpha) \Longrightarrow Y_{(1)} \sim$ Weibull$(m, \alpha/n)$

19. If $Z \sim \mathcal{N}(0,1), W \sim \chi^2(\nu)$, and $Z \perp\!\!\!\perp W$, then

$$T = \frac{Z}{\sqrt{W/\nu}} \sim t(\nu)$$

20. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$

$$\Longrightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

21. If $W_1 \sim \chi^2\left(\nu_1\right), W_2 \sim \chi^2\left(\nu_2\right)$, and $W_1 \perp\!\!\!\perp W_2$, then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F\left(\nu_1, \nu_2\right)$$

22. If $F \sim F\left(\nu_1, \nu_2\right)$, then $1/F \sim F\left(\nu_2, \nu_1\right)$

23. If $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$

24. If $W \sim F\left(\nu_1, \nu_2\right)$, then

$$\frac{\left(\nu_1/\nu_2\right) W}{1 + \left(\nu_1/\nu_2\right) W} \sim \text{beta}\left(\nu_1/2, \nu_2/2\right)$$