1 Stat-2410

1. In a certain factory, machines I, II and III are all producing springs of the same length. Machines I, II, III produce $3\%, 2\%$ and 4% defective springs, respectively. Of the total production of springs in factory, Machine I produces 40%, Machine II produces 50% and Machine III produces 10%.

   a) If one spring is selected at random from the total springs produced in a given day, determine the probability that it is defective. **[2 points]**

   b) Given that the selected spring is defective, find the conditional probability that it was produced by Machine II. **[2 points]**

2. The random variable X follows the distribution

$$p_X(n) = \mathbb{P}(X = n) = \frac{k}{n(n+2)(n+3)}, n \geq 1$$

   a) Find constant k that makes the p_X a probability mass function. **[3 points]**

   b) Find the $\mathbb{E}(X)$. **[2 points]**

3. Let X_1, X_2 and X_3 be random variables with equal variances but with correlation coefficients $\rho_{12} = 0.3$, $\rho_{13} = 0.5$ and $\rho_{23} = 0.1$. Find the correlation coefficient of the linear functions $Y = X_1 + X_2$ and $Z = X_2 + X_3$. **[5 points]**

4. A customer in an appliance store will purchase a washer with probability 0.38, a dryer with probability 0.0.47 and an iron with probability 0.48. She will purchase both a washer and a dyer with probability 0.11, both a washer and an iron with probability 0.12, both a dryer and an iron with probability with 0.11 and all the three items with probability 0.05. Determine the probability that the customer will purchase none of them. **[5 points]**

5. Let the independent random variables X_1 and X_2 have binomial distributions with parameters $n_1, p_1 = \frac{1}{2}$ and $n_2, p_2 = \frac{1}{2}$, respectively. Show that $Y = X_1 - X_2 + n_2$ has a binomial distribution with parameters $n = n_1 + n_2, p = \frac{1}{2}$. **[5 points]**

6. Let X have the following probability density function

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1$$

where $\alpha > 0$ and $\beta > 0$. Find the mean and variance of X. **[6 points]**

7. Assume that X and Y are independent. The probability density function of X is

$$f_X(x) = \frac{1}{6}x^3 e^{-x}, \quad x > 0,$$

and Y has the following probability density function.

$$f_Y(y) = ye^{-y}, \quad y > 0.$$

Find the probability density function of $V = \frac{X}{X+Y}$. **[5 points]**

8. Let Y_1 denote the minimum of a random sample of size n from a distribution that has pdf $f(x) = e^{-(x-\theta)} \quad \theta < x < \infty$, zero elsewhere. Let $Z_n = n(Y_1 - \theta)$. Investigate the limiting distribution of Z_n. **[5 points]**

1. We draw at random 13 cards from a full deck of cards. What is the probability that we draw 4 Hearts and 3 Diamonds (keeping 4 digits after the decimal point in your answer)?

2. Two fair dice are thrown and the smallest of the face values, Z say, is noted.

   (a) Calculate the expectation $\mathbb{E}[Z]$ of Z.

   (b) Calculate the variance $\mathrm{Var}(Z)$ of Z.

3. A continuous random variable X has CDF F given by,

$$F(x) = \begin{cases} 0, & x < 0 \\ x^3 & 0 \le x \le 1 \\ 1 & x > 1. \end{cases}$$

   (a) Calculate the expectation $\mathbb{E}(X)$ of X.

   (b) Calculate the variance $\mathrm{Var}(X)$ of X.

4. A company has three factories (1, 2 and 3) that produce the same chip, each producing 15%, 35% and 50% of the total production. The probability of a defective chip at 1, 2, 3 is 0.01, 0.05, 0.02, respectively. Suppose someone shows us a defective chip. What is the probability that this chip comes from factory 1 (keeping 4 digits after the decimal point in your answer)?

5. Let X and Y be independent standard normal random variables. Find the PDF of the quotient $U = X/Y$.

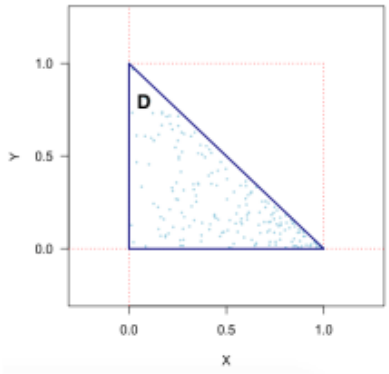
6. Consider a sequence of iid random variables $X_1, X_2, \ldots$ from a common distribution with probability density function (PDF):

$$f_X(x) = \frac{5}{2}x^4, \quad \text{for} \quad -1 \le x \le 1.$$

   (a) Let $Z_n = \max(X_1, X_2, \ldots, X_n)$. Find the limiting distribution of Z_n as $n \to \infty$.

   (b) Let $S_n = X_1 + X_2 + \cdots + X_n$. Using the Central Limit Theorem to find the approximated probability of $\mathbb{P}(S_{100} \le 1.50)$.

7. We select a point (X, Y) in the triangle $D = \{(x, y) \in (0.1)^2 \subset \mathbb{R}^2,\ x + y < 1\}$, as shown in the figure below:



Let the joint PDF of random variables X and Y be

$$f_{(X,Y)}(x, y) = \begin{cases} \frac{1}{1-x} & (x, y) \in D, \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Find the marginal PDFs of X and Y.

(b) Find the joint probability $\mathbb{P}\left(X \leq \frac{1}{2},\ Y > \frac{1}{2}\right)$.

(c) Find the conditional PDFs of X and Y.

(d) Given $X = \frac{1}{2}$, find the conditional probability that $Y \leq \frac{1}{4}$.

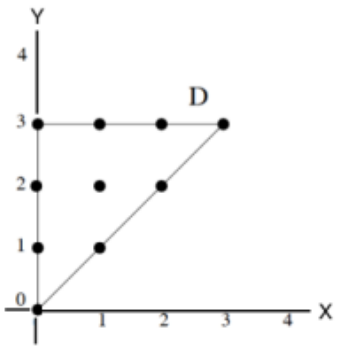
(e) Are X and Y independent?

(f) Let $V = X + Y$, find the probability distribution of V.

(g) Find the probability that $V \geq \frac{1}{2}$.

(h) Find the expectation of V.

8. We draw a point (X, Y) from the 10 points on the triangle D and assume that each point is equally likely to be dawn.



(a) Find the joint PMF of X and Y.
(b) Find the marginal PMFs of X and Y.
(c) Find the joint probability $\mathbb{P}(X \geq 1 \text{ and } Y \leq 2)$.
(d) Find the conditional probability that $Y < 3$ given that $X = 1$.
(e) Are X and Y independent?
(f) Let $U = XY$, find the distribution of U.
(g) Find the probability that $U \geq 5$.
(h) Find the expectation of U.

2 Stat-3411

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with probability density function $f(x;\theta), \forall \theta \in \Omega$, where Ω is an interval. Suppose that Y is a complete sufficient statistic for θ. Prove that Y is independent of any ancillary statistic Z.

[5 points]

2. Let $X_1, \ldots, X_n$ be a random sample of size n the uniform distribution $U(0, \theta)$. Let Y_n denote the maximum of the sample and let $Z_n = n(\theta - Y_n)$. Show $Z_n \xrightarrow{D} Exp(\theta)$. **[5 points]**

3. Let $X_1, \ldots, X_n$ be a random sample of size n from $U(0, \theta)$. At level $(1 - \alpha)$, find a confidence interval with equal tails for θ. **[5 points]**

4. Let $X_1, \ldots, X_n$ be a random sample of size n from Binomial Distribution $B(2, \theta)$.

   a) Use statistic $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ and find unbiased estimators for parameters θ and θ^2, respectively. **[5 points]**

   b) Use part (a) and find the Rao–Cramer Lower Bound for the estimators. **[5 points]**

5. Let $X_1, \ldots, X_n$ be a random sample of size n following the distribution function

$$F(x; \theta_1, \theta_2) = \begin{cases} 1 - (\theta_1/x)^{\theta_2} & \theta_1 \leq x \\ 0 & \text{elsewhere} \end{cases}$$

Find the maximum likelihood estimators of θ_1 and θ_2. **[10 points]**

6. Let $X_1, \ldots, X_n$ be a random sample of size n the distribution $N(\theta_1, \theta_2)$. Show taht the likelihood ratio principle for testing $H_0 : \theta_2 = \theta_2^{(0)}$ specified, and θ_1 unspecified against $H_1 : \theta_2 \neq \theta_2^{(0)}$, θ_1 unspecified, leads to a test that rejects when $\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq c_1$ or $\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq c_2$ where $c_1 < c_2$ are selected appropriately. **[7 points]**

7. Let $X_1, \ldots, X_n$ be independent random variables from

$$f(x_i; \theta, \lambda) = \lambda e^{-\lambda(x_i - i\theta)}, \ \lambda > 0, \ x_i \geq i\theta, i = 1, \ldots, n$$

   a) Find a sufficient statistic for parameter $\Theta = (\theta, \lambda)$. **[5 points]**

   b) Find a minimal sufficient statistic for parameter $\Theta = (\theta, \lambda)$. **[5 points]**

8. Let X be a random variable from the truncated binomial distribution at zero with

$$\mathbb{P}(X = x) = c\binom{n}{x} p^x q^{n-x}, \quad c > 0, \quad 0 < p < 1, \quad x = 1, \ldots, n.$$

where $q = 1 - p$. Find the UMVUE for parameter $\frac{p}{1-q^n}$. **[8 points]**

1. (14 points) Let $X = (X_1, \ldots, X_n)$ be a random sample of size n. We say $T_n = t_n(X)$ is consistent in mean square if $\lim_{n\to\infty} E_\theta\{(T_n - \theta)^2\} = 0$ for all $\theta \in \Omega$. The quantity $E_\theta\{(T_n - \theta)^2\}$ is called the mean squared error (MSE) of the estimator T_n to estimate θ.
   (a) (6 points) Show that T_n is consistent in mean square if (1) $Var_\theta(T_n)$ converges to 0 as $n \to \infty$, and (2) T_n is unbiased in the limit as $n \to \infty$.
   (b) (6 points) Show that if $\lim_{n\to\infty} E_\theta\{(T_n - \theta)^2\} = 0$ for all $\theta \in \Omega$, then T_n converges in probability to θ for all $\theta \in \Omega$; that is, $T_n \xrightarrow{p} \theta$.
   (c) (2 points) For a given sample size n, a desired property of an estimator is to have a small MSE comparing with other estimators. Briefly discuss why this is an important property.

2. (22 points) Let $(X_1, \ldots, X_n)$ be a random sample from the distribution with p.d.f.
$$f(x;\theta) = \frac{2\theta^2}{x^3}, \qquad 0 < \theta \le x,$$
and c.d.f.
$$F(x;\theta) = 1 - \left(\frac{\theta}{x}\right)^2, \qquad 0 < \theta \le x.$$
   (a) (8 points) Find the probability density function (p.d.f.) of $T = X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $E_\theta(T)$.
   (b) (8 points) Show that T is a complete sufficient statistic.
   (c) (4 points) Find the minimum variance unbiased estimator (M.V.U.E.) of θ.
   (d) (2 points) Let $\hat{\theta}_n$ denote the maximum likelihood (M.L.) estimator of θ. Find $\hat{\theta}$.

3. (22 points) Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n > 2$ from a distribution with p.d.f.
$$f(x;\theta) = \theta\, x^{\theta-1}, \qquad 0 < x < 1, \quad \theta > 0,$$
and 0 elsewhere.
   (a) (6 points) Let $\hat{\theta}$ denote the maximum likelihood estimator (M.L.E.) of θ. Show that $\hat{\theta} = \frac{n}{T}$, where $T = -\sum_{i=1}^n \log X_i$.
   (b) (4 points) Let $Y = \left(\frac{n-1}{n}\right)\hat{\theta}$. Show that Y is an unbiased estimator of θ. [*Hint*: $T = -\sum_{i=1}^n \log X_i \sim \text{Gamma}(n, \frac{1}{\theta})$].
   (c) (4 points) Find Rao-Cramér Lower Bound (R.C.L.B.) for estimating θ.
   (d) (5 points) Show whether Y is an efficient estimator of θ. [*Hint*: $E_\theta(Y^2) = \frac{(n-1)\theta^2}{n-2}$]
   (e) (3 points) Find the likelihood ratio test (L.R.T.) statistic Λ for testing $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$.

4. (22 points) Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size $n > 2$ from a distribution with p.d.f.

$$f(x;\theta) = \theta\, x^{\theta-1}, \qquad 0 < x < 1, \quad \theta > 0,$$

and 0 elsewhere.

   (a) (6 points) Show that $X_1, \ldots, X_n$ has an exponential family (E.F.) distribution.

   (b) (8 points) Show that $X_1, \ldots, X_n$ has a regular exponential family (R.E.F.) distribution. Find a complete sufficient statistic for this model.

   (c) (2 points) Find a minimal sufficient statistic for this model.

   (d) (6 points) Find the minimum variance unbiased estimator (M.V.U.E.) of θ. [*Hint*: $-\sum_{i=1}^{n} \log X_i \sim \text{Gamma}(n, \frac{1}{\theta})$]

5. (20 points) Let $X = (X_1, \ldots, X_n)$ be a random sample from the Exp(θ) distribution.

   (a) (2 points) State Basu's theeorem.

   (b) (15 points) Show that $T(X) = \sum_{i=1}^{n} X_i$ and $U(X) = \left(\frac{X_1}{T}, \ldots, \frac{X_n}{T}\right)$ are independent random variables.

   (c) (3 points) Find $E\left[\frac{X_1}{T}\right]$.

3 Stat-3420

1. (5 points) Briefly answer the following question:

   (a) (2 points) What is the main difference between an experiment and an observational study?

   (b) (2 points) Name and define the basic principles of a randomized comparative experiment.

   (c) (2 points) What is a blocking factor? Give an advantage of the blocking technique.

2. (70 points) Four different designs for a digital computer circuit are being studied to compare the amount of noise present. A completely randomized design was conducted. The results are shown in the following table:

Circuit Design	Noise Observed					Totals	Averages
1	19	20	19	30	8	96	19.2
2	80	61	73	56	80	350	70.0
3	47	26	25	35	50	183	36.6
4	95	46	83	78	97	399	79.8

   (a) (3 points) Let τ_i denote the effect of the ith treatment, $i = 1, \ldots, a$, where $a = 4$. Write down the effects model for this experiment and define all elements of the model. Make it sure that assumptions are clearly stated.

   (b) (5 points) Let $\sum_{i=1}^{4} \tau_i = 0$. Derive the least square estimators of the model parameters.

   (c) (4 points) Give a reasonable point estimate for each parameter in the model defined in part (a).

   (d) (6 points) What is a reasonable estimator of μ_2? Under the model assumptions stated in part (a), find the expectation and variance of it. What is the distribution of it?

   (e) (2 points) It is known that under the constraint $\mu = 0$, the least squares estimator of τ_i is $\bar{y}_{i.}$, where $\bar{y}_{i.} = \frac{\sum_{j=1}^{n} y_{ij}}{n}$, $i = 1, \ldots, a$ (You do not need to show this result). Is τ_i estimable? Why?

(f) (15 points) An incomplete ANOVA table is given below:

Source	SumSq	DoF	MeanSq	F-value
Design	12042			
Residuals				
Total				

For the given data set, $\sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij}^2 = 67830$ and $y.. = 1028$. Test the hypothesis that the same amount of noise is present in all four circuit designs. Define the null and alternative hypotheses. Write down your conclusion. Use $\alpha = 0.05$.

(g) (5 points) Calculate $2 \times \sqrt{\frac{MSE}{n}}$. Informally compare the treatment means. What conclusions can you draw?

(h) (1 point) If you were a design engineer and you wished to minimize the noise, which circuit design would you select?

(i) (3 points) Calculate the residuals for observations in Treatment 4.

(j) (2 points) Let Γ be a contrast of treatment means, where $\Gamma = \mu_1 + \mu_2 - \mu_3 - \mu_4$. What are the contrast coefficients in Γ?

(k) (2 points) Estimate the contrast Γ defined in part (n).

(l) (5 points) Construct a 99% confidence interval for Γ defined in part (n).

(m) (5 points) Use Tukey's test to test the hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ at 5% level of significance.

(n) (5 points) You have been reported that the "Noise Observed" data have been misrecorded. Each observation should be 7 measurements less than what was reported in the above data table; that is, $y_{ij}^* = y_{ij} - 7$, (for $i = 1, \ldots, 4$, and $j = 1, \ldots, 5$), where y_{ij}^* is the value of the correct observation and y_{ij} is the value reported in the data table. Test the hypothesis that the same amount of noise is present in all four circuit designs. Write down your conclusion. Use $\alpha = 0.05$.

3. (15 points) An engineer is studying the mileage performance characteristics of five types of gasoline additives. In the road test he wishes to use cars as blocks; however, because of a time constraint, he must use an incomplete block design. He runs the balanced design with the five blocks that follow.

	Car				
Additive	1	2	3	4	5
1		17	14	13	12
2	14	14		13	10
3	12		13	12	9
4	13	11	11	12	
5	11	12	10		8

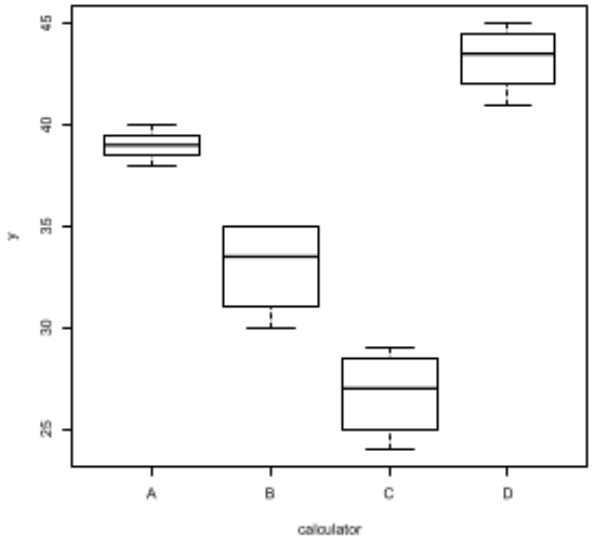
(a) (3 points) What are the parameters of this incomplete block design? Verify that this a balanced incomplete block design (BIBD).

(b) (10 points) Test whether the treatment means are equal or not (use $\alpha = 0.05$) and draw conclusions. You can use the following results:
The row totals are $y_{1.} = 56$, $y_{2.} = 51$, $y_{3.} = 46$, $y_{4.} = 47$ and $y_{5.} = 41$.
The column totals are $y_{.1} = 50$, $y_{.2} = 54$, $y_{.3} = 48$, $y_{.4} = 50$ and $y_{.5} = 39$.
The correction factor is $CF = \frac{y^2}{N} = \frac{241^2}{20} = 2904.05$ and $SS_T = 76.95$.
Also, $Q_1 = 33/4 = 8.25$, $Q_2 = 11/4 = 2.75$, $Q_3 = -(3/4) = -0.75$, and $Q_4 = -(14/4) = -3.5$.

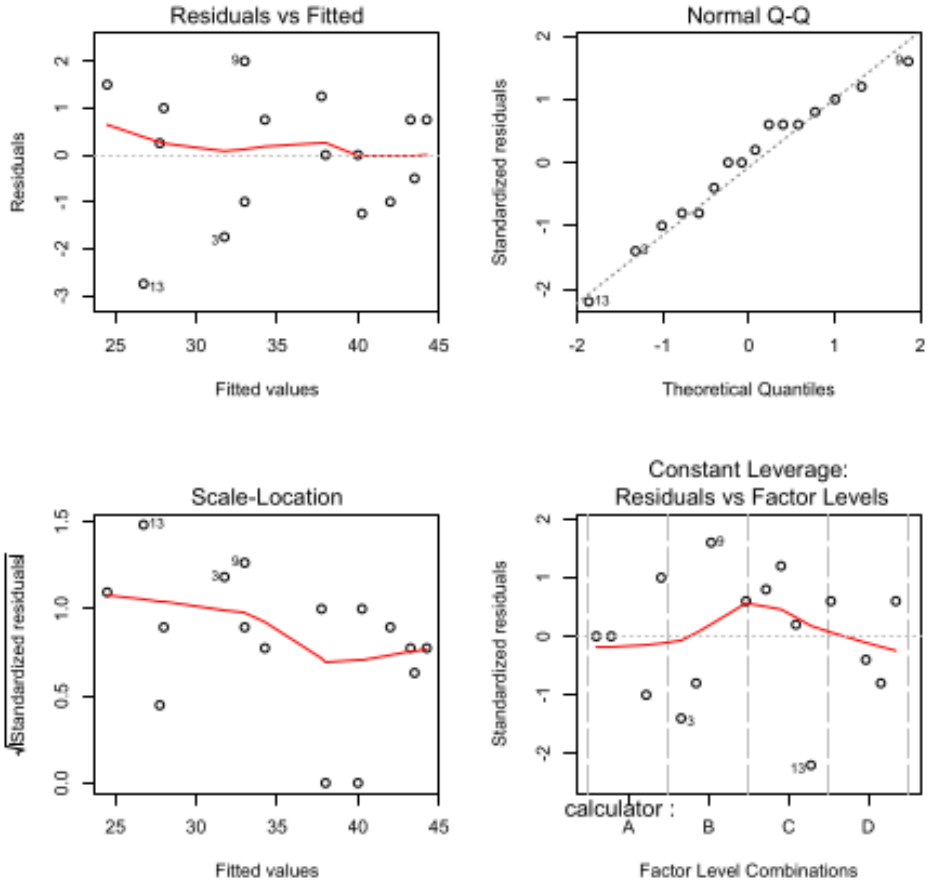
(c) (2 points) The engineer would like to assess the block effects. Write down the symbolic notation of the partition of the total sum of squares to assess the block effects.

Appendix for Question 3

Box Plots



Residual Plots



4. (Bonus, 10 points) An engineer suspects that the surface finish of a metal part is influenced by the feed rate and the depth of cut. She selects three feed rates and four depths of cut. She then conducts a factorial experiment and obtains the following data:

	Depth of Cut (in)			
Feed Rate (in/min)	0.15	0.18	0.20	0.25
0.20	74	79	82	99
	64	68	88	104
	60	73	92	96
0.25	92	98	99	104
	86	104	108	110
	88	88	95	99
0.30	99	104	108	114
	98	99	110	111
	102	95	99	107

We may analyze these data using the two-way ANOVA model: $y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$, where $i = 1, 2, 3$, $j = 1, 2, 3, 4$ $k = 1, 2, 3$, and $\varepsilon_{ijk} \sim \text{NID}(0, \sigma^2)$. Some of the results from fitting this model are given in the following incomplete ANOVA table:

Source	Sum Sq	DF	Mean Sq	F-Value
A-Depth		3	708.37	
B-Feed		2	1580.25	
AB				
Residual				
Total	6532.00	35		

The value of $SS_{Subtotals}$ is 5842.67. Complete the ANOVA table. Based on the complete ANOVA table, what can you conclude?

1. In the selection of regression models:

(a) show that Mallows' C_p method containing $(p-1)$ covariates is equal to p? **[2 point]**

(b) Mallows' C_p selects the best subset of predictors based on two criteria. State those two criteria and explain what they each imply? **[2 point]**

2. Show that general linear test statistic can be expressed in terms of coefficients of multiple determination for full model and reduced model. Denoting these by R_F^2 and R_R^2, respectively, show that **[4 points]**

$$F^* = \frac{R_F^2 - R_R^2}{df_R - df_F} \div \frac{1 - R_F^2}{df_F}$$

3. In a paper published in the Proceedings of the National Academy of Sciences (PNAS), researchers look at the data from three countries including Canada, USA and Mexico. They are interested in the relationship between mean annual temperatures x and annual mortality index y from Covid-19. The regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

(a) Write down a regression model that can be used to compare the mortality index of different countries. **[2 points]**

(b) The researchers believe that mortality index for each country is affected by temperature. How the regression model (a) is changed by this new information. Expand the regression model (a) to cover this information. **[2 points]**

Use the regression model in part (b), and answer the following questions:

(c) State the response function for each country separately. **[2 points]**

(d) We wish to test if the response function for all the countries is the same or not. State clearly the null and alternative hypotheses, test statistic and decision rule. **[2 points]**

[**Hint**: Country is a qualitative variable.]

4. In a regression model, the following statistics were collected: $SSE = 210$ and $n = 7$ with

$$(X'X)^{-1} = \begin{pmatrix} 3.20 & -0.26 & -0.30 \\ & 0.06 & 0.01 \\ & & 0.04 \end{pmatrix}$$

$$(X'Y) = \begin{pmatrix} 28 \\ 126 \\ 174 \end{pmatrix}$$

(a) Estimate the coefficients and write down the estimated regression function. **[2 points]**

(b) Interpret the estimates of β_1? **[2 points]**

(c) Estimate the standard errors of each of the estimates? **[2 points]**

(d) Find 90% joint Bonferroni confidence interval for β_1 and β_2 and then interpret the joint confidence interval? **[2 points]**

(e) Find 90% prediction interval for a new observation with $x_1 = 4,\ x_2 = 5$? **[2 points]**

5. A hospital administrator wished to study the relation between patient satisfaction y and severity of illness x. The following data were collected (data have been rescaled):

x_i	12	6	12	6	8	8
y_i	4	3	8	3	5	6

(a) Complete the following ANOVA table for the collected data. **[3 points]**

Source	df	SS	MS
Regression	?	7.440	?
Error	?	?	?
Total	?	?	

(b) Expand the ANOVA table to include the decomposition of the error sum of squares into pure error sum of squares and lack of fit sum of squares. **[2 points]**

Use the result of part (b) and answer the following questions.

(c) State the hypotheses. **[2 points]**

(d) Indicate full model and reduced model. **[2 points]**

(e) Lack of fit test statistic and compute its value based on the data in the study. **[2 points]**

(f) Use $\alpha = 0.05$, indicate the decision rule and the conclusion in a plain language. **[2 points]**

6. A biologist employed linear regression model to study the relation between the concentration of a drug in plasma y and the log-dose of the drug x_1, the Amino acid function x_2, the mRNA function x_3. The following statistics have been obtained from $n = 50$ samples in this study.

$$
\begin{aligned}
&Model\ 1: \quad y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i, \qquad SSE = 36, \qquad SSTO = 51 \\
&Model\ 2: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad SSE = 10 \\
&Model\ 3: \quad y_i = \beta_0 + \beta_1 x_{i2} + \epsilon_i, \qquad SSE = 30 \\
&Model\ 4: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \qquad SSE = 3
\end{aligned}
$$

(a) Complete the following table based on above information. **[3 points]**

Source	df	SS	MS
x_1	?	?	?
$x_2 \mid x_1$	?	?	?
$x_3 \mid x_1, x_2$	?	?	?
Error	?	?	

(b) We wish to test whether X_1 and X_3 should be dropped from the model. Use $\alpha = 0.05$, state the hypotheses, test statistics, decision rule and conclusion. **[2 points]**

(c) Calculate partial coefficient of determination $R^2_{y2|1}$ and interpret that in plain language. **[2 points]**

7. Suppose that a linear regression model is given by $Y = X\beta + \epsilon$. Consider standardized regression model with transformed variables

$$
\begin{aligned}
X^*_{ik} &= \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{s_k} \right), \ k = 1, \ldots, p-1 \\
Y^*_i &= \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \bar{Y}}{s_y} \right).
\end{aligned}
$$

where

$$
s^2_k = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ik} - \bar{X}_k)^2, \quad k = 1, \ldots, p-1
$$

Show that $(X^{*'} X^*)^{-1} = r^{-1}_{xx}$ where where r_{xx} is the correlation matrix of X variables. **[4 points]**

1. Briefly describe when you would use the ratio and regression estimator instead of the sample mean to estimate the population mean.

[4 marks]

2. In many surveys, there is interest in estimating strata means or differences in strata means.

a) In general, for simple random sampling within each stratum, write down the large sample distribution for estimators $\bar{y}_h$ and $\bar{y}_h - \bar{y}_k$.

[6 marks]

b) To estimate average water quality of wells, a survey of residential wells was carried out in the rural parts of Avalon. The population of 13,345 wells was identified from assessment records. Three strata were created. The first stratum includes wells in farms with animals, the second stratum includes wells in farms without animals, and the third stratum includes wells in houses. A random sample of wells was selected from each stratum and the water quality was tested for their sodium (Na) concentration (mg/L).

Stratum	Population Size	Sample Size	Sample Mean Na Concentration	Standard Deviation Na Concentration
Farms with animals	2,365	150	237.3	41.45
Farms without animals	1,297	100	245.6	37.62
Houses	9,683	250	220.1	51.23

Find an approximate 95% confidence interval for the average Na difference between the two types of farm wells. Interpret the confidence interval.

[6 marks]

3.

a) Prove that the Horvitz-Thompson estimator of the population total

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

is an unbiased estimator of the population total under a random sampling given the inclusion probabilities π_i for $i = 1, 2, \dots, N$.

[7 marks]

b) Find the variance of the Horvitz-Thompson estimator of the population total.

[7 marks]

4. A taxi-cab company wants to estimate the proportion of unsafe tires (excluding spare tires) on their 175 cabs. Because it would imply taking more vehicles of the roads, it is impractical to select a simple random sample of tires. Thus, cluster sampling is used instead, with each cab being a cluster. A simple random sample of 25 cabs gives the following number of unsafe tires per cab:

2, 4, 0, 1, 2, 0, 4, 1, 3, 1, 2, 0, 1, 1, 2, 2, 4, 1, 0, 0, 3, 1, 2, 2, 1

Estimate the proportion of unsafe tires being used on the company's cabs, and give a 95% confidence interval for that proportion.

[6 marks]