

Statistics Qualifying Examination

Answer all questions and show all work.

This exam is closed-note/book. You need to use a calculator.

1. Let X_1 and X_2 have the joint probability density function (pdf)

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} 15x_1^2x_2 & 0 < x_1 < x_2 < 1, \\ 0 & \text{elsewhere,} \end{cases}$$

- (a) Calculate the probability $\Pr(X_1 + X_2 \leq 1)$.
 - (b) Find the marginal pdf of X_2 , $f_{X_2}(x_2)$.
 - (c) Show that $f_{X_2}(x_2)$ is a (legitimate) probability density function.
 - (d) Calculate $E\left(\frac{1}{X_2}\right)$.
2. Suppose that the random variable X has a Poisson distribution such that $50\Pr(X = 1) = \Pr(X = 2)$.
- (a) Find the mean and the variance of X .
 - (b) Calculate the probability $\Pr(X \geq 1)$.
 - (c) Obtain a lower bound for the probability $\Pr(75 < X < 125)$ using the Chebyshev's inequality.

3. Let $X_1, \dots, X_n$ be independent and identically distributed with density

$$f_\theta(x) = \frac{2x}{\theta} \exp\left(-\frac{x^2}{\theta}\right) \mathbf{1}\{x > 0\}, \quad \theta > 0.$$

Set $T = \sum_{i=1}^n X_i^2$.

- (a) Find the distribution of $Y_i = X_i^2$ and hence derive the probability density function of T .
- (b) Construct an exact $100(1 - \alpha)\%$ confidence interval for θ , where $0 < \alpha < 1$.
- (c) Find the maximum likelihood estimator of θ . Show that it is unbiased, and then prove that it is the UMVUE of θ .
- (d) Find the UMVUE of θ^2 . Clearly justify your answer.

4. For each $n \geq 1$, let X_n have density

$$f_n(x) = nx^{n-1} \mathbf{1}\{0 < x < 1\}.$$

- (a) Derive the cumulative distribution function of X_n and prove that $X_n \xrightarrow{p} 1$.
- (b) Let $Y_n = n(1 - X_n)$. Derive the cumulative distribution function of Y_n . Find its limiting distribution and prove that Y_n converges to it in distribution.
- (c) Determine whether Y_n converges in probability to any constant. Explain why your answer does not contradict part (a).

5. A simple linear regression model is specified and fit:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \epsilon_i \text{ iid } N(0, \sigma^2),$$

for $i = 1, \dots, n$. Use the partially complete output below to answer questions.

Use t-values: $t_{0.025,40} = 2.021$, $t_{0.05,40} = 1.684$, $t_{0.025,5} = 2.7764$, $t_{0.05,5} = 2.1318$,
 $t_{0.025,43} = 2.016692$, $t_{0.05,43} = 1.681071$, $t_{0.025,6} = 2.44691$, $t_{0.05,6} = 1.9431$,
 $t_{0.025,44} = 2.015368$, $t_{0.05,44} = 1.68023$, $t_{0.025,3} = 3.182446$, $t_{0.05,3} = 2.353363$,

Analysis of Variance Table

Response: Y

Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	152044	****	***
Residuals	43	*****	6035	

Coefficients:

Estimate	Std. Error	t value	Pr(> t)
(Intercept)	99.595	****	****
X1	5.433	*****	****

- (a) Compute the coefficient of determination R^2 .
- (b) Conduct an F test to decide whether or not there is a linear association between X and Y . Use $\alpha = 0.05$. To get full credit, you need to give the hypotheses, test statistic, degrees of freedom, (range of) p-value, and your conclusion.
- (c) Construct a 95% confidence interval for β_1 .
- (d) What additional information is needed to calculate the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$? Explain your answer.

6. Some researchers observed the data on 20 individuals with high blood pressure. The researchers were interested in determining if a relationship exists between blood pressure and age, weight, body surface area, duration, pulse rate and/or stress level. The following variables are recorded:

- blood pressure ($Y = \text{BP}$, in mm Hg)
- age ($X1 = \text{Age}$, in years)
- weight ($X2 = \text{Weight}$, in kg)
- body surface area ($X3 = \text{BSA}$, in sq m)
- duration of hypertension ($X4 = \text{Dur}$, in years)
- basal pulse ($X5 = \text{Pulse}$, in beats per minute)
- stress index ($X6 = \text{Stress}$)

The summary of the multiple regression analysis is given below:

Call:

```
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.93213	-0.11314	0.03064	0.21834	0.48454

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-12.870476	2.556650	-5.034	0.000229 ***
X1	0.703259	0.049606	14.177	2.76e-09 ***
X2	0.969920	0.063108	15.369	1.02e-09 ***
X3	3.776491	*****	*****	***** *
X4	0.068383	0.048441	1.412	0.181534
X5	-0.084485	0.051609	-1.637	0.125594
X6	0.005572	0.003412	1.633	0.126491

Residual standard error: 0.4072 on 13 degrees of freedom

Multiple R-squared: 0.9962, Adjusted R-squared: 0.9944

F-statistic: 560.6 on 6 and 13 DF, p-value: 6.395e-15

The researchers are also given the F-tables below as well as the best subset regression using the R function *regsubsets* with *nbest=4* to perform best-subset based model selection and *summaryHH* (in the next page). Output is summarized below. *rss* denotes the sum of squares of errors/residuals (SSE). Also *rsq* denotes the R^2 .

(df1, df2)	(1,13)	(1,14)	(1,15)	(1,16)	(1,17)	(1,18)	(1,19)
$F_{0.025;df1,df2}$	6.414	6.298	6.199	6.115	6.042	5.978	5.922
$F_{0.05;df1,df2}$	4.667	4.600	4.543	4.493	4.451	4.414	4.381

(df1, df2)	(2,13)	(2,14)	(2,15)	(2,16)	(2,17)	(2,18)	(2,19)
$F_{0.025;df1,df2}$	4.965266	4.856698	4.765048	4.686665	4.618874	4.559672	4.507528
$F_{0.05;df1,df2}$	3.805565	3.738892	3.682320	3.633723	3.591531	3.554557	3.521893

```

> library(leaps)
> BP.bestreg<-regsubsets(Y~., data=data1, method="exhaustive",nbest=4)
> library(HH)
> # S3 method for regsubsets
> summaryHH(BP.bestreg)

```

	model	p	rsq	rss	adjr2	cp	bic	stderr
1	X2	2	0.903	54.53	0.897	312.81	-40.59	1.740
2	X3	2	0.750	140.14	0.736	829.07	-21.71	2.790
3	X5	2	0.520	268.56	0.494	1603.41	-8.71	3.863
4	X1	2	0.434	316.73	0.403	1893.93	-5.41	4.195
5	X1-X2	3	0.991	4.82	0.990	15.09	-86.10	0.533
6	X2-X6	3	0.920	44.87	0.910	256.56	-41.50	1.625
7	X2-X5	3	0.919	45.59	0.909	260.90	-41.18	1.638
8	X2-X4	3	0.914	48.43	0.903	278.06	-39.97	1.688
9	X1-X2-X3	4	0.995	3.06	0.994	6.43	-92.23	0.437
10	X1-X2-X5	4	0.992	4.33	0.991	14.10	-85.27	0.520
11	X1-X2-X4	4	0.992	4.65	0.990	16.01	-83.86	0.539
12	X1-X2-X6	4	0.992	4.66	0.990	16.11	-83.79	0.540
13	X1-X2-X3-X4	5	0.995	2.72	0.994	6.41	-91.56	0.426
14	X1-X2-X3-X6	5	0.995	2.84	0.994	7.12	-90.72	0.435
15	X1-X2-X3-X5	5	0.995	3.02	0.993	8.18	-89.51	0.448
16	X1-X2-X5-X6	5	0.994	3.39	0.992	10.47	-87.14	0.476
17	X1-X2-X3-X5-X6	6	0.996	2.49	0.994	6.99	-90.37	0.421
18	X1-X2-X3-X4-X5	6	0.995	2.60	0.994	7.67	-89.49	0.431
19	X1-X2-X3-X4-X6	6	0.995	2.60	0.994	7.68	-89.47	0.431
20	X1-X2-X4-X5-X6	6	0.994	3.10	0.992	10.71	-85.94	0.471
21	X1-X2-X3-X4-X5-X6	7	0.996	2.16	0.994	7.00	-90.22	0.407

- (a) Interpret estimates $\hat{\beta}_4$ in the full model (the model with all 7 covariates). Please use word only no Y 's and X 's.
- (b) Is it worth including $X3$ in the model that already contains $X1$, $X2$, $X4$, $X5$, and $X6$? To get full credit, give the hypotheses, test statistic, degrees of freedom, (range of) p-values, and your conclusion. Use $\alpha = 0.05$.
- (c) Assume that the modeling assumptions hold. A researcher wants to determine whether an additional 1 kg increase in a patient's weight is associated with an increase in blood pressure of at least 0.9 mm Hg. To receive full credit, state the hypotheses, the test statistic, the degrees of freedom, the (range of) p-value, and your conclusion. Use a significance level of $\alpha = 0.05$.
- (d) Starting with the model that contains no predictor variables, perform forward variable selection using a significance level of $\alpha = 0.05$. To receive full credit, present the first two steps of the forward selection procedure, including the variable considered for entry at each step, the corresponding test statistic (or p-value), and the decision made.

7. A materials engineer studies the tensile strength (in MPa) of 3D-printed test specimens. Three process factors are investigated: the *infill density* A at three levels (20%, 50%, 80%), the *layer height* B at two levels (0.10 and 0.20 mm), and the *print speed* C at two levels (40 and 80 mm/s). All three factors are fixed. Three specimens are printed at each of the $3 \times 2 \times 2 = 12$ factor-level combinations, and the 36 specimens are produced and tested in random order. The measured tensile strengths are:

Layer	Speed=40 mm/s	Speed=80 mm/s					
Infill	(mm)	(three specimens)			(three specimens)		
20%	0.10	33.8	39.4	39.5	38.9	35.4	36.8
20%	0.20	31.6	29.2	32.5	32.1	27.1	27.4
50%	0.10	45.6	41.9	44.9	38.1	44.2	40.0
50%	0.20	42.9	40.1	37.5	32.9	38.4	39.8
80%	0.10	44.7	49.3	51.1	47.6	48.6	47.2
80%	0.20	46.2	42.6	44.6	44.0	43.8	44.2

A three-factor ANOVA was fit with SAS PROC GLM; part of the output is below. (Note that here the design is balanced, and the Type I and Type III sums of squares coincide.)

The GLM Procedure

Dependent Variable: strength (tensile strength, MPa)

Sum of

Source	DF	Squares	Mean Square	F Value	Pr > F
Model	(1)	1233.1075	(2)	(3)	<.0001
Error	24	146.8600	(4)		
Corrected Total	35	1379.9675			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
infill	2	943.1267	471.5633	(5)	<.0001
layer	1	225.5003	225.5003	36.85	<.0001
speed	1	26.5225	26.5225	4.33	0.0482
infill*layer	2	24.0356	12.0178	1.96	0.1622
infill*speed	2	11.7067	5.8533	0.96	0.3984
layer*speed	1	0.4669	0.4669	0.08	0.7847
infill*layer*speed	2	1.7489	0.8744	0.14	0.8676

The infill-density level means are $\bar{y}_{20\%..} = 33.642$, $\bar{y}_{50\%..} = 40.525$, $\bar{y}_{80\%..} = 46.158$ (each averaging 12 observations). You may use $t_{0.025, 24} = 2.064$ and the Tukey value $q_{0.05, 3, 24} = 3.53$.

- (a) Write down an appropriate statistical model for this experiment, defining every term and stating the model assumptions and the side conditions (constraints) on the fixed effects.
- (b) Fill in the five missing values labeled (1)–(5) in the ANOVA table. Show your steps.
- (c) At the $\alpha = 0.05$ level, state which main effects and interactions are statistically significant, and interpret your findings in the context of the experiment.
- (d) The interactions are negligible. Estimate the difference in mean tensile strength between the 80% and 20% infill levels, and construct a 95% confidence interval for this difference.
- (e) Use Tukey's method to compare the three infill-density levels at a family significance level of $\alpha = 0.05$. State which levels differ significantly.

8. A food scientist compares four protein-bar recipes (1, 2, 3, 4) with respect to firmness (in newtons, N). Firmness is rated by trained sensory judges, but a judge suffers palate fatigue and can reliably evaluate only *three* of the four recipes in one tasting session. Four judges are used, and the recipes are assigned so that each judge evaluates a different set of three recipes and every pair of recipes is evaluated together the same number of times. The design used and the measured firmness values are shown below.

Judge (block)					
Recipe	1	2	3	4	Y _{i.}
1	28.1	22.7	26.4		77.2
2	28.3	23.8		26.8	78.9
3	32.8		30.3	30.8	93.9
4		22.8	27.1	25.0	74.9
Y _{.j}	89.2	69.3	83.8	82.6	324.9

A SAS PROC GLM analysis (MODEL firm = recipe judge;) gave the output below.

The GLM Procedure

Dependent Variable: firm (firmness, N)

Sum of

Source	DF	Squares	Mean Square	F Value	Pr > F
Model	6	110.4450	18.4075	43.06	<.0001
Error	5	2.1375	0.4275		
Corrected Total	11	112.5825			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
recipe	3	74.0892	24.6964	57.77	0.0003
judge	3	36.3558	12.1186	28.35	0.0015

Source	DF	Type III SS	Mean Square	F Value	Pr > F
recipe	3	39.0025	13.0008	30.41	0.0012
judge	3	36.3558	12.1186	28.35	0.0015

The adjusted (least-squares) recipe means are $\widehat{\mu + \tau_1} = 25.74$, $\widehat{\mu + \tau_2} = 26.53$, $\widehat{\mu + \tau_3} = 30.34$, $\widehat{\mu + \tau_4} = 25.70$. Selected critical values:

(df_1, df_2)	(4, 4)	(4, 5)	(4, 6)	df	4	5	6
$q_{0.05; df_1, df_2}$	5.7571	5.2183	4.8956	$t_{0.025; df}$	2.7764	2.5706	2.4469
$q_{0.025; df_1, df_2}$	7.0882	6.2567	5.7717	$t_{0.05; df}$	2.1318	2.0150	1.9432

- (a) What kind of design is used in this experiment? Determine the design parameters and give a model to analyze this data set together with the model assumptions and constraints.
- (b) Using the SAS output, test whether the four recipes differ in mean firmness (use $\alpha = 0.05$). To get full credit, give the test statistic, its degrees of freedom, the p -value, and your conclusion.
- (c) Suppose Tukey's method is used to compare the four recipes pairwise. Calculate the critical difference (CD) at $\alpha = 5\%$. (No comparisons or intervals are required—only the CD.)
- (d) Construct a 95% confidence interval for $\tau_3 - \tau_2$ (the recipe-3 main effect minus the recipe-2 main effect), and state whether recipes 3 and 2 differ significantly.