DEPARTMENT OF MATHEMATICAL SCIENCES
MONTCLAIR STATE UNIVERSITY

MASTER OF SCIENCE COMPREHENSIVE EXAMINATION

STATISTICS — 3 Hour — STUDY GUIDE

The following questions are representative of the type of content that can be expected in this THREE (3) hour examination. Note that some questions may combine concepts from more than one course.

Instructions:
In this comprehensive examination you must answer all SIX (6) questions.
Probability Tables, Graphs and SAS / JMP output will be provided as needed.

Methodology: Significance testing, confidence intervals, power, analysis of variance, contrasts, multiple comparisons, assumptions, principles of experimental designs, blocking designs, latin squares, factorial and nested designs, fixed, random and mixed effects, expected mean squres, model evaluation, residual diagnostics, correlation, simple and multiple linear regression, polynomial regression, sequential and partial sums-of-squares (Type I and III analyses), collinearity, interpretation of output from statistical software such as SAS, JMP and graphing calculators.

Theory: Random variables, univariate and multivariate probability models, distribution functions, transformations, expectation, moment generating functions, Chebychev's inequality, sampling distributions, limit theorems, properties of point estimators, maximum likelihood estimation, sufficiency, exponential class of distributions, Fisher information and Cramer-Rao inequality, simple and composite hypotheses, most powerful tests, Neyman-Pearson lemma, uniformly most powerful tests, likelihood ratio tests.

Courses: STAT 541, 542, 543, 544, 547, 548

References:

The **candidate's** lecture notes and **textbook** from each of the required courses

Moore and McCabe, *Introduction to the Practice of Statistics*
Ott, *An Introduction to Statistical Methods and Data Analysis*
Kutner, Nachtsheim and Neter, *Applied Linear Regression Models*
Draper and Smith, *Applied Regression Analysis*
Montgomery, *Design and Analysis of Experiments*
Kuehl, *Design of Experiments: Statistical Principles of Research Design and Analysis*
Delwiche and Slaughter, *The Little SAS Book: A Primer*
SAS Institute Inc., *PC-SAS Software*
Hogg, McKean and Craig, *Introduction to Mathematical Statistics*
Hogg and Tanis, *Probability and Statistical Inference*

Question 1

The following ANOVA table was obtained from a *balanced* completely randomized design:

Source	df	SS	MS	F
Treatment		126		
Error	20		16	
Total	23			

(a) By completing the ANOVA table above determine:

(i) The number of treatments.

(ii) The number of replications per treatment.

(b) Hence state the model and the assumptions needed for inference purposes.

(c) Test whether there is a difference between the true mean responses to the treatment.

(d) Determine the $P-$value of this test and explain what it represents.

Question 2

In a study of fitness, researchers were interested in the effects of weight and oxygen capacity on the time to run 1.5 miles in working adults. Data were collected on 31 individuals. Weight was measured in kg and Runtime in minutes. Higher values of oxygen capacity (Oxy) are associated with better aerobic fitness.

(a) Figure 1 and Table 1 give the results from fitting different Simple Linear Regressions (SLR's). What conclusions are suggested by each SLR ?

(b) Based on your conclusions from Figure 1 and Table 1, find a 95% confidence interval for the mean time for an individual whose oxygen capacity is 51 to run 1.5 miles.

(c) Can the correlation between `Runtime` and `Oxy` be determined from Table 1 ?
If so, compute it ; if not, explain why.

(d) If both weight and oxygen capacity are used as explanatory variables, would you expect your conclusions as to the usefulness of weight and oxygen capacity to explain average running time to be the same as those in part (a) ? Explain.

Figure 1: Plots for Question 2

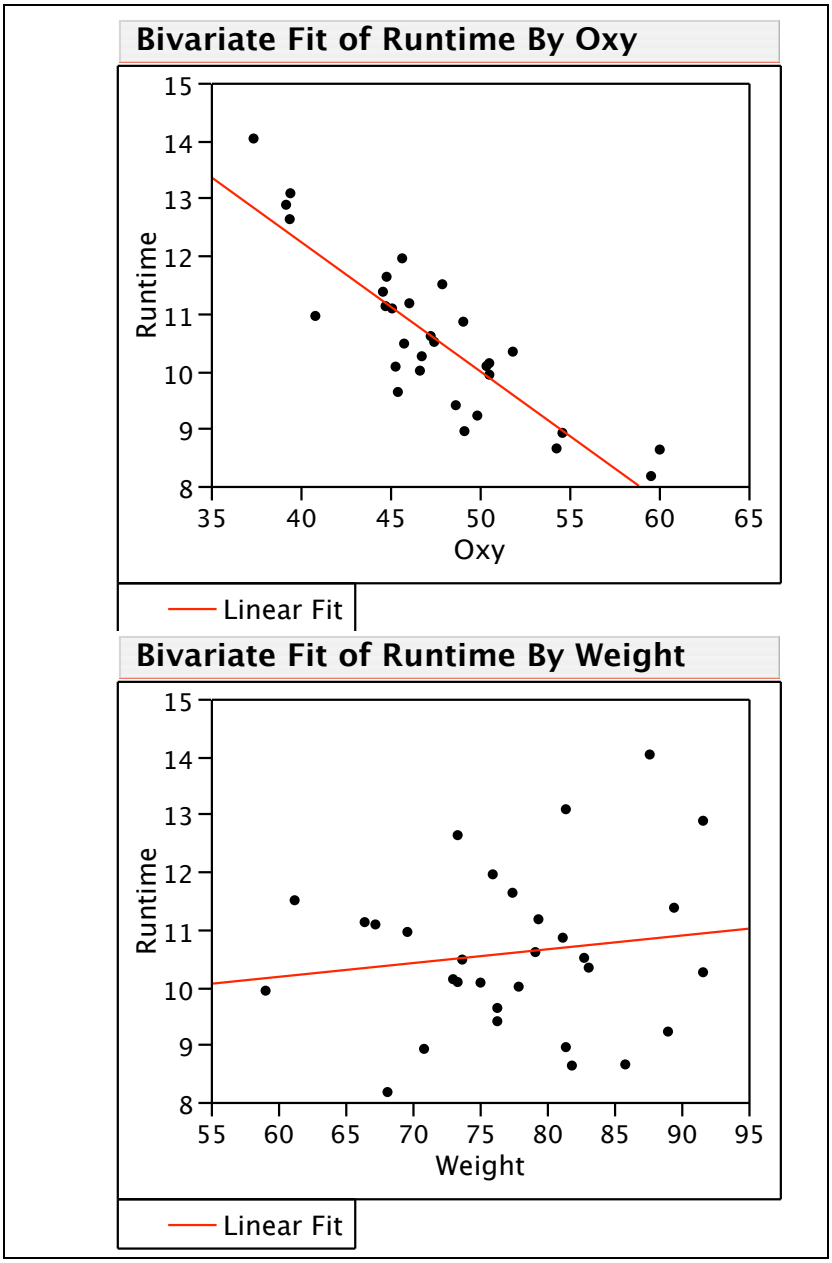

Table 1: SAS Output for Question 2

Dependent Variable: Runtime

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	42.92405	42.92405	83.97	<.0001
Error	29	14.82349	0.51115		
Corrected Total	30	57.74754			

Root MSE	0.71495	R-Square	0.7433
Dependent Mean	10.58613	Adj R-Sq	0.7345
Coeff Var	6.75366		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21.22219	1.16775	18.17	<.0001
Oxy	1	-0.22450	0.02450	-9.16	<.0001

Dependent Variable: Runtime

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.18928	1.18928	0.61	0.4412
Error	29	56.55826	1.95028		
Corrected Total	30	57.74754			

Root MSE	1.39653	R-Square	0.0206
Dependent Mean	10.58613	Adj R-Sq	-0.0132
Coeff Var	13.19204		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.73472	2.38411	3.66	0.0010
Weight	1	0.02391	0.03061	0.78	0.4412

Question 3

A randomized complete block design was employed to study the effects of three diets on the *reduction* of total lipid (fat) level in blood plasma. Data and SAS output from fifteen male subjects who were within 20% of their ideal body weight is given in **Table 2** below.

Table 2: Lipid Readings

Age Group	Fat Content of Diet		
	Extremely Low	Fairly Low	Moderately Low
15 - 24	0.73	0.67	0.15
25 - 34	0.86	0.75	0.21
35 - 44	0.94	0.81	0.26
45 - 54	1.40	1.32	0.75
55 - 64	1.62	1.41	0.78

Source	DF	Sum of Squares	Mean Square
AGE	4	1.41896	0.354740
FAT	2	1.32028	0.660140
Error	8	0.01932	0.002415
Corrected Total	14	2.75856	

Level of AGE	N	Y Mean	Y SD
15-24	3	0.51666667	0.31895663
25-34	3	0.60666667	0.34789845
35-44	3	0.67000000	0.36097091
45-54	3	1.15666667	0.35444793
55-64	3	1.27000000	0.43714986

Level of FAT	N	Y Mean	Y SD
ext_low	5	1.11000000	0.38078866
fair_low	5	0.99200000	0.34557199
mod_low	5	0.43000000	0.30846394

(a) Write the model and assumptions for this experiment.

(b) Use the SAS output to explain why *two* of the assumptions appear to be satisfied.

(c) Use the ANOVA results to state "appropriate" conclusions about the two factors.

(d) Give two orthogonal contrasts that could be used to compare the `FAT` levels.

Question 4

Let $X_1, \ldots, X_n$ be a random sample obtained from the *shifted* Exponential distribution:

$$f(x) = \lambda \exp\{-\lambda(\, x - \theta \,)\} \quad , \qquad x \geq \theta, \quad \lambda > 0$$

(a) Find the maximum likelihood estimator for λ assuming θ is known.

(b) Derive the cumulative distribution function for $\quad Y_1 = \min\{\, X_1, \ldots, X_n \,\}$
[NOTE: Remember that $\quad X_i \geq \theta \quad$ for all $i = 1, 2, \ldots, n$]

(c) Hence, show Y_1 is a *consistent* estimator of θ.

Question 5

SAS output from fitting three regression models to the data below is given in **Table 3**.

Y	X_1	X_2
28	3	49
33	4	41
42	7	33
55	8	31
75	16	7
91	21	5

(a) For each regression, what are your conclusions ?

That is, briefly discuss the overall significance, quality-of-fit of the regression, and the significance of the explanatory variable(s) in the regression model.

(b) What "conflict" appears to arise between the three regression results ?

(c) (i) Do a rough sketch of X_1 versus X_2

(ii) Explain why this resolves the apparent conflict in part (b) above.

(d) Assume the dataset above exists as a permanent SAS dataset `"aperm"` with variable names `Y X1 X2` in the directory `C:\Desktop\MyLib`

Write SAS code that would produce the output below.

Table 3: SAS Output for Question 5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3004.41465	3004.41465	177.81	0.0002
Error	4	67.58535	16.89634		
Corrected Total	5	3072.00000			

Root MSE	4.11052	R-Square	0.9780
Dependent Mean	54.00000	Adj R-Sq	0.9725

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20.23610	3.03764	6.66	0.0026
X1	1	3.43362	0.25749	13.33	0.0002

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2913.36198	2913.36198	73.46	0.0010
Error	4	158.63802	39.65950		
Corrected Total	5	3072.00000			

Root MSE	6.29758	R-Square	0.9484
Dependent Mean	54.00000	Adj R-Sq	0.9355

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	91.17851	5.04245	18.08	<.0001
X2	1	-1.34380	0.15679	-8.57	0.0010

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3009.92646	1504.96323	72.73	0.0029
Error	3	62.07354	20.69118		
Corrected Total	5	3072.00000			

Root MSE	4.54876	R-Square	0.9798
Dependent Mean	54.00000	Adj R-Sq	0.9663

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.93210	26.74829	1.27	0.2941
X1	1	2.78476	1.28906	2.16	0.1195
X2	1	-0.26442	0.51232	-0.52	0.6414

Question 6

A study to investigate possible causes for high serum phosphate levels in adults with sickle cell anemia included three groups: five sickle cell patients with high serum phosphate levels ($>$ 4.5 mg/dliter), four sickle cell patients with normal serum phosphate levels (3 - 4.5 mg/dliter), and 7 controls. The response variable was tubular reabsorptive capacity for phosphate (TRCP). [TRCP is a measure of the phosphate reabsorption rate in the kidney.]

(a) In planning your pre-ANOVA analysis determine appropriate orthogonal contrasts that would help experimenters analyze the experiment. State the contrasts in words and give the coefficients for the contrast.

(b) State the ANOVA model and state the necessary assumptions.

(c) What methods would you use to test the assumptions of part b).
[**Note:** Do not perform any tests, just discuss the methods.]

(d) What options would you consider if these assumptions are violated ?

(e) Use the SAS Output given in **Table 4** to analyze the data. Summarize the results and conclusions of your analysis.

Table 4: SAS Output for Question 6

Group	N	Mean	Std Dev	Minimum	Maximum
Control NonS	7	3.3071429	0.5498398	2.6000000	4.0800000
High Serum	5	4.6420000	0.4506884	4.1200000	5.3600000
Normal Serum	4	3.5750000	0.2998333	3.1800000	3.8900000

Dependent Variable: TRCP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.44645214	2.72322607	12.22	0.0010
Error	13	2.89612286	0.22277868		
Corrected Total	15	8.34257500			

Table 4 Continued: SAS Output for Question 6

R-Square	C.V.	Root MSE	TRCP Mean
0.652850	12.44957	0.471994	3.791250

Source	DF	Anova SS	Mean Square	F Value	Pr > F
GROUP	2	5.44645214	2.72322607	12.22	0.0010

Analysis of Variance Procedure

Level of GROUP	N	TRCP Mean	TRCP SD
Control NonS	7	3.30714286	0.54983980
High Serum	5	4.64200000	0.45068836
Normal Serum	4	3.57500000	0.29983329

Duncan's Multiple Range Test for variable: TRCP

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 13 MSE= 0.222779
WARNING: Cell sizes are not equal.
Harmonic Mean of cell sizes= 5.060241

Number of Means	2	3
Critical Range	.6411	.6714

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	GROUP
A	4.6420	5	High Serum
B	3.5750	4	Normal Serum
B			
B	3.3071	7	Control NonS

Question 7

Researchers conducted an experiment to determine the content uniformity of tablets used to lower blood pressure which were produced at two blending sites. A random sample of three batches at each site was collected and from each batch, a random sample of five tablets were analyzed for content uniformity. Partial SAS output from this analysis is presented below.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Site	1	0.1255	0.1255	10.37	0.0036
Batch(Site)	4	0.4540	0.1135	9.38	0.0001
Error	24	0.2904	0.0121		
Corrected Total	29	0.8699			

(a) Write the model and assumptions for this experiment.

(b) Assume the data were entered correctly and the model assumptions satisfied.

What conclusions would the researchers make "if" they based their interpretation on the SAS output above, **as given**. [NOTE: Answer this *in the context of the study.*]

(c) (i) Although the numbers in the SAS output are correct, the researchers based their conclusions on the wrong analysis. Explain why.

(ii) Hence, perform the correct analysis and state your conclusions.

(d) What is the **specific** problem that the researchers now need to resolve and what additional piece of useful information can you give them based on the SAS output?

Question 8

This question refers to the study and data given in Question 3.

(a) Explain exactly how the randomization for this design would have been done.

(b) What are the advantages and disadvantages of using a randomized block design versus a completely randomized design?

(c) Was blocking by `AGE` group effective ? Explain your answer.

(d) Give the standard error of the difference between two means for any pair of diets.

Question 9

Let X_1, X_2 ..., X_{10} denote a random sample of size 10 from a distribution having probability density function $f(x;p) = p^x(1-p)^{1-x}$ for $x = 0, 1$ and zero elsewhere.

(a) Show that the critical region defined by $\sum_{i=1}^{10} X_i \leq 2$ is a best critical region for testing $H_o : p = \frac{1}{2}$ against $H_1 : p = \frac{1}{4}$

(b) Determine, for this test, the significance level α.

(c) Determine, for this test, the power at $p = \frac{1}{4}$.

(d) Is this also the "best" test for $H_o : p = \frac{1}{2}$ against $H_2 : p < \frac{1}{2}$? Show your reasoning.

Question 10

Data were collected at a large university on all first-year computer science majors in a particular year. The purpose of the study was to attempt to predict success in the early university years. One measure of success was the cumulative grade point average (GPA) after three semesters. Explanatory variables under study were average high school grades in mathematics (HSM), science (HSS), and English (HSE). We also include SAT mathematics (SATM) and SAT verbal (SATV) scores as explanatory variables. The SAS output below relates to this problem.

(a) Write down the model under consideration, including all assumptions.

(b) Describe how each assumption would be investigated and what course of action might be taken if an assumption were not met.

(c) Based on the SAS output, indicate which explanatory variables seem to be significant predictors of GPA for this population of students.

(d) (i) Briefly discuss the difference between the "`Type I SS`" and "`Type III SS`" (Sums of Squares) presented in the SAS output below.

(ii) Hence explain why SATM (`satm` in the SAS output below) is significant under Type I Sums of Squares and *not* significant under Type III Sums of Squares.

Dependent Variable: gpa

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	28.6436449	5.7287290	11.69	<.0001
Error	218	106.8191439	0.4899961		
C Total	223	135.4627888			

R-Square	Coeff Var	Root MSE	gpa Mean
0.211450	26.56311	0.699997	2.635223

Source	DF	Type I SS	Mean Square	F Value	Pr > F
satm	1	8.58293360	8.58293360	17.52	<.0001
satv	1	0.00090549	0.00090549	0.00	0.9658
hsm	1	17.72646964	17.72646964	36.18	<.0001
hss	1	1.37653218	1.37653218	2.81	0.0952
hse	1	0.95680397	0.95680397	1.95	0.1637

Source	DF	Type III SS	Mean Square	F Value	Pr > F
satm	1	0.92799885	0.92799885	1.89	0.1702
satv	1	0.23265187	0.23265187	0.47	0.4915
hsm	1	6.77243120	6.77243120	13.82	0.0003
hss	1	0.44214334	0.44214334	0.90	0.3432
hse	1	0.95680397	0.95680397	1.95	0.1637

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.3267187390	0.39999643	0.82	0.4149
satm	0.0009435925	0.00068566	1.38	0.1702
satv	-.0004078495	0.00059189	-0.69	0.4915
hsm	0.1459610795	0.03926097	3.72	0.0003
hss	0.0359053199	0.03779841	0.95	0.3432
hse	0.0552925813	0.03956869	1.40	0.1637

Question 11

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from $N(\mu_x, \sigma_x^2)$ and $Y_1, Y_2, \ldots, Y_m$ a random sample of size m from $N(\mu_y, \sigma_y^2)$ where the normal populations are independent.

(a) Let $\theta = \sigma_x^2$ and assume μ_x is known. Show that the maximum likelihood estimator (MLE) of θ is unbiased.

(b) Now assume all parameters unknown. Derive an *exact* 95% confidence interval for $\mu_x - \mu_y$ under the assumption that $\sigma_x^2 = k\sigma_y^2$ where $k > 0$ is a known constant.

(c) Explain what is done in practice to obtain an approximate 95% confidence interval for $\mu_x - \mu_y$ when σ_x^2 and σ_y^2 are unknown and no additional assumptions are imposed.

Question 12

The following questions relate to the problem and SAS output presented in Question 10. Discuss the *main* issue(s) that arise in each of the scenarios that are presented.

(a) The correlation between `satm` and `hsm` is 0.99

(b) The use of R-Square versus Adjusted R-Square to select the best subset model.

(c) Selecting the "best" subset model by eliminating *all* nonsignificant explanantory variables (based on `Pr > |t|`) listed in the `Parameter` part of the SAS output.

(d) Fitting the polynomial model: $\texttt{gpa} = \beta_0 + \beta_1 \texttt{hsm} + \ldots + \beta_{100}(\texttt{hsm})^{100} + \epsilon$

(e) Including all possible interactions between the explanatory variables in the regression model. That is, `satm*hsm` (etc., ... up to) `satm*satv*hsm*hss*hse`

Question 13

A process for refining sugar yields up to 1 ton of pure sugar per day, but the actual amount produced, Y, is a random variable because of machine breakdowns and other slowdowns. Suppose that Y has density function given by

$$f(y) = \begin{cases} 2y & 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

The company is paid at the rate of \$300 per ton for the refined sugar, but it also has a fixed overhead cost of \$100 per day.

(a) Find the mean and variance of Y.

(b) Express the daily profit, P say, as a function of Y.

(c) Hence, find the probability density function of P.

(d) Find $E(P)$, the average daily profit (in hundreds of dollars).

Question 14

Let X_1, X_2, X_3, X_4 denote a random sample of size $n = 4$ from a normal distribution with mean $\mu = 7$ and variance $\sigma^2 = 4$. Let Y_1, Y_2, ..., Y_9 denote a random sample of size $n = 9$ from a normal distribution with mean $\mu = 10$ and variance $\sigma^2 = 9$.

(a) State the distribution of $\overline{X}$.

(b) Hence, derive the distribution of $\overline{Y} - \overline{X}$.

(c) Find $P[\overline{X} < \overline{Y}]$.

Question 15

Let X_1, $X_2, \ldots$, X_5 denote a random sample of size 5 from a normal distribution $N(0, \theta)$

(a) By initially setting $\theta' = 2$, derive the uniformly most powerful test of size = 0.05 for testing $H_o : \theta = 1$ against $H_1 : \theta > 1$

(b) What is meant by the term"uniformly" in the phrase "uniformly most powerful"?

(c) Suppose we observe the following data: -1, 0, 1, 1, 2

(i) Use your test from part (a) to test $H_o : \theta = 1$ against $H_1 : \theta > 1$

(ii) What is your conclusion? Interpret this result in plain English.

Question 16

Let X_1, $X_2, \ldots$, X_{25} denote a random sample of size $n = 25$ from the distribution that has the probability density function

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right)$$

(a) Give the best critical region of a size $\alpha = 0.05$ test of $H_o : \theta = 0$ vs. $H_a : \theta = 1$.

(b) Give the uniformly most powerful (UMP) test of $H_o : \theta = 0$ vs. $H_a : \theta > 0$

(c) Explain why there does **not** exist a UMP test of $H_o : \theta = 0$ vs. $H_a : \theta \neq 0$