

MATH 204: Principles of Statistics 2

Final Examination: Solutions

Winter 2026

April 22nd 2026: 9.00am – 12.00pm

MATH 204 Specific Instructions.

Please write your answers on the exam in the boxes provided for each question; only answers written inside the correct boxes will be marked. Extra space for rough work can be found at the end of the exam paper. Please write clearly. Please do not use highlighter.

This paper contains five questions. Each question carries 20 marks. Credit will be given for all questions attempted. The total mark available is 100 but rescaling of the final mark may occur.

Calculators may be used. Dictionaries and Translation dictionaries are permitted. A formula sheet is provided. Standard notation from the course will be used, and you may use standard notation in your answers without definition. For full marks, you must give justification for your answers.

1. In an experiment to discover preference for different ice cream brands, $n = 27$ subjects were recruited and randomly allocated to test one of four ice cream brands (labelled A, B, C and D). Each subject reported a satisfaction score on a range from 0 to 100; this score will be treated as being a measure on a continuous scale.

The data were analyzed as a Completely Randomized Design (CRD). In the usual notation of the course, it was discovered that two of the sums of squares quantities were $SS = 5186.018$ and $SSR = 4224.309$.

(a) Construct the ANOVA table from the information given.

8 MARKS

Q1a

As the CRD can be treated as a Linear Model, we have of course that $SSR = SST$.

Then, using the usual computation rules, we find the table to be

Source	DF	SS	MS	F
Brand	3	4224.309	1408.103	33.676
ERROR	23	961.709	41.813	
Total	26	5186.018		

(b) Carry out the ANOVA F-test and report the conclusion.

4 MARKS

Q1b

We compare the F-statistic with the $F_{3,23}$ distribution upper $\alpha = 0.05$ point, which is from tables found to be 3.03. The test statistic lies in the rejection region, and we therefore reject H_0 and conclude that the ice cream brands do have different preference scores.

- (c) Taking Brand A as the baseline group, the following table reports the parameter summaries using the usual contrast parameterization.

	Estimate	Std. Error	t value	p-value
(Intercept)	70.1661	2.2862	30.6912	0.0000
BrandB	-28.6555	3.4922	-8.2055	0.0000
BrandC	-2.6406	3.4922	-0.7561	0.4572
BrandD	-23.4050	3.3466	-6.9936	0.0000

- (i) Report the mean of the satisfaction scores from the subjects who sampled Brand B ice cream. 2 MARKS
- (ii) Is the mean of the satisfaction scores from the subjects who sampled Brand C ice cream significantly different from zero? Justify your answer. 2 MARKS

Q1c

- (i) The estimate is the sum of the intercept and the Brand B contrast, that is

$$70.1661 + (-28.6555) = 41.5106.$$

- (ii) Yes it is: the p-value is for the contrast, not the mean itself.
(this answer sufficient for full marks)

The estimate is therefore $70.1661 + (-2.6406) = 67.5255$, We know from the ANOVA table that the estimate of σ is

$$s = \sqrt{41.8130} = 6.4663$$

so the 95% interval clearly excludes zero.

- (d) Report the usual estimate of the residual error variance, σ^2 . 2 MARKS
Identify a key assumption that allows this estimate to be used. 2 MARKS

Q1d

This is the MSE, and we have that $\hat{\sigma}^2 = 41.813$.

The key assumption is the equal variance assumption; that is, the variance in the different treatment group populations must be the same.

The other assumptions are not so important: independence is necessary, but guaranteed by the design. Normality is not important – this estimate of the variance is ok even if we do not have Normality.

2. The following ANOVA tables were constructed from analyses of a data set from a randomized block design (RBD) for treatment factor A using blocking factor B:

SOURCE	DF	SS	MS	F	p
Treatments	7	20.676	2.954	2.665	0.0175
Blocks	3	99.293	33.098	29.861	0.0000
Interaction	21	21.229	1.011	0.912	0.5773
ERROR	64	70.938	1.108		
TOTAL	95	212.137			

First Analysis

SOURCE	DF	SS	MS	F	p
Treatments	7	20.676	2.954	2.724	0.0134
Blocks	3	99.293	33.098	30.524	0.0000
ERROR	85	92.168	1.084		
TOTAL	95	212.137			

Second Analysis

SOURCE	DF	SS	MS	F	p
Treatments	7	20.676	2.954	1.358	0.2334
ERROR	88	191.461	2.176		
TOTAL	95	212.137			

Third Analysis

The sums of squares quantities are reported to 3 decimal places, and the p-values are reported to 4 decimal places.

- (a) Identify the numbers of levels of factor A and factor B, the number of replicates, and the total sample size. 4 MARKS

Q2a

The numbers are

$$k = 8 \quad b = 4 \quad r = 3 \quad n = 96$$

- (b) Explain briefly **why** we would use a randomized block design in the investigation of the effect of a treatment factor. You do not need to explain the precise details of a randomized block design. 4 MARKS

Q2b

We would use a blocking factor if we believed that the population was comprised of different subpopulations that we know or strongly suspect will respond differently to the treatment factor. We will randomly allocate experimental treatments to individuals within a sample from each subpopulation.

- (c) Explain why the SSE quantity is different in each table. 2 MARKS

Write down the sample variance for the response variable. 2 MARKS

Q2c

Because in the second analysis, the SSE is increased by the SSI quantity, and in the third analysis, it gets further increased by the SSB quantity. In each case, when we remove a term from the model, the sum of squares associated with that term gets added to the SSE.

The sample variance $SS/(n - 1) = 2.233$.

- (d) Based on the analyses, is the effect on the outcome of changing treatment level the same in each block ? Briefly justify your answer. 2 MARKS.

Q2d

Yes: the interaction in the first analysis is not significant.

- (e) Based on all of the information given, and **not for each analysis separately**, state whether it should be concluded from these results that there is a significant effect of the treatment factor A, and justify your conclusion. 6 MARKS

Q2e

There does appear to be a significant effect of factor A. The blocking factor is clearly significant, as evidenced by the p-values for the blocks in the first and second analyses, and the one factor Third Analysis is ignoring the variation due to the blocking factor.

3. In a study of blood viscosity (y) in samples taken from 32 hospital patients, it is desired to assess the influence of the continuous covariates PCV (x_1), plasma fibrinogen (x_2) and plasma protein (x_3) on the outcome.

(a) In an analysis using only x_1 as a predictor, the following ANOVA table was produced:

	DF	SS	MS	F
X1	1	9.2295	9.2295	101.7645
ERROR	30	2.7209	0.0907	
TOTAL	31	11.9504		

Compute the R-squared value from this analysis.

4 MARKS

Q3a

This is simply

$$R^2 = \frac{SSR}{SS} = \frac{9.2295}{11.9504} = 0.772$$

- (b) In an analysis using only x_2 as a predictor, it was discovered that $SSR = 2.4990$. Construct the ANOVA table for the analysis based on x_2 , and report whether x_2 should be regarded as an influential predictor in that model.

8 MARKS

Q3b

The table is this

	DF	SS	MS	F
X2	1	2.4990	2.4990	7.9323
ERROR	30	9.4513	0.3150	
TOTAL	31	11.9504		

The F-statistic exceeds the critical value which is $F_{0.05,1,30} = 4.17$, and so should be regarded as an influential predictor.

- (c) Several model fits were implemented, and the key summaries are included in the table below. Each model is denoted in standard model notation. In the table, p records the **total** number of parameters, including the intercept.

	Model	p	SSE
M_1	X_1	2	2.721
M_2	X_2	2	9.451
M_3	X_3	2	11.828
M_4	$X_1 + X_2 + X_3$	4	2.582
M_5	$X_1 + X_2 + X_1 : X_2$	4	2.598
M_6	$X_1 + X_2 + X_3 + X_1 : X_2$	5	2.579

On the basis of this information, identify the most suitable model for explaining the variation in the response among the models fitted. Any appropriate model search strategy may be used. 8 MARKS

Use this space for rough work, and give your answer on the next page, summarizing your selection strategy.

Q3c

By inspection of the SSE quantities, it is clear that the SSEs for many of the models are similar provided x_1 is included, and that there does not really seem to be much improvement from the complex models compared to the simplest. Therefore we jump straight to comparing Model M_6 as the complete model with M_1 as the reduced model. In this case

$$F_{obs} = \frac{(SSE_1 - SSE_6)/(k - g)}{SSE_6/(n - k - 1)}$$

where $k = 4, g = 1$ so that

$$k - g = (n - g - 1) - (n - k - 1) = 5 - 2 = 3.$$

Therefore

$$F_{obs} = \frac{(2.721 - 2.579)/(5 - 2)}{2.579/(32 - 5)} = 0.494$$

We compare this with the 0.05 upper α point of the $F_{k-g, n-k-1}$ distribution, which in this case is

$$F_{0.05, 3, 27} = 2.96$$

and we do not reject this as an adequate simplification. Therefore, we select the model with only x_1 included as the simplest adequate model.

4. A study was carried out to assess predictors that were associated with patient pain levels one month after knee surgery. In the study, three different surgical approaches were compared; each patient was also randomly assigned to one of two physiotherapy regimens to aid recovery after the surgery. At the end of one month, a detailed pain questionnaire was completed by the patients. The outcome in the study was a measured pain score, a value on a continuous scale from 0 to 100.

In the data set, there are four recorded variables:

- Surgery type (factor predictor with three levels labelled I, II and III)
- Physiotherapy regimen (factor predictor with two levels labelled P1 and P2)
- Age of patient (continuous predictor)
- Pain Score (continuous outcome)

The numbers of data points in the sample falling into each cross-classification category given by the levels of the factor predictors are given in the following table.

		Surgery		
		I	II	III
Physio	P1	35	66	35
	P2	29	65	46

- (a) A model was fitted and used to produce the following ANOVA table:

SOURCE	DF	SS	MS	F	p
Age	1	4318.342	4318.342	236.249	0.0000
Surgery	2	72.116	36.058	1.973	0.1411
Physio	1	1655.437	1655.437	90.566	0.0000
ERROR	271	4953.551	18.279		
TOTAL	275	10999.445			

On the basis of these results alone, is it possible to conclude that Surgery type has no effect on pain score ? Justify your answer. 5 MARKS

Q4a

These results alone do not allow us to conclude that Surgery type is not an influential variable as the study is unbalanced, so the F-statistics and p-values do not represent the true effect of any of the variables.

(b) A second model was fitted and used to produce the following ANOVA table:

SOURCE	DF	SS	MS	F	p
Surgery	2	113.867	56.933	1.428	0.2416
ERROR	273	10885.578	39.874		
TOTAL	275	10999.445			

On the basis of these results alone, is it possible to conclude that Surgery type has no effect on pain score ? Justify your answer. 5 MARKS

Q4a

Taking the results in isolation, we can conclude that there is no effect. Balance does not matter, as we are treating this as a CRD-type analysis. We might question whether the right model is being fitted.

In reality, these results alone do not allow us to conclude that Surgery type is not an influential variable. Here the lack of balance does not matter, but we also know from the previous analysis that it seems likely that one of the other variables is influential. Therefore we need to condition on those other variables to get a full picture.

(c) A third model was fitted and used to produce the following ANOVA table:

SOURCE	DF	SS	MS	F	p
Age	1	4318.342	4318.342	235.983	0.0000
Physio	1	1685.364	1685.364	92.099	0.0000
ERROR	273	4995.740	18.299		
TOTAL	275	10999.445			

Given all the information presented in the three analyses in (a), (b) and (c), is it possible to conclude that Surgery type has no effect on pain score ? Justify your answer. 6 MARKS

Answer this question on the next page

Q4c

Now we can do a model comparison between the model in (a) and the model in (c):

$$F = \frac{(4995.740 - 4953.551)/(4 - 2)}{4953.551/271} = 1.144$$

which is not significant when we compare against the $F_{2,271}$ upper 0.05 point. Now we can conclude that Surgery type is not an influential variable.

- (d) Identify the checks that you would carry out to verify that the conclusion about Surgery type made in (c) is valid. 4 MARKS

Q4d

You need to check the three assumptions, in particular, that the residuals are patternless and have a constant variance. This can be done using scatterplots against the two predictors, histograms, and QQ-plots.

5. (a) In a large environmental health case control study, the association between pollution due to industrial waste disposal processes and lung cancer was investigated. In the study, 470 lung cancer cases and 478 matched controls were identified, and classified by proximity to an industrial waste disposal plant. The following counts table was recovered.

		Lung Cancer		
		Control	Case	Total
Proximity to plant	No	462	431	893
	Yes	16	39	55
Total		478	470	948

Based on these data, compute the estimate, $\hat{\psi}$, of the odds ratio for assessing the association between the two factors. 2 MARKS

Using a large sample test based on the standardized log-scale statistic, state whether there is any statistical evidence for an association. 4 MARKS

Recall that the natural logarithm scale is used in this calculation.

Q5a

Here the estimate is

$$\hat{\psi} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 2.613$$

The statistic is

$$Z = \frac{\log \hat{\psi}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

which in this case is equal to $Z = 3.156$ which we compare against the standard Normal distribution: this leads to a rejection of the null hypothesis, and confirms evidence of an association.

- (b) Students from three Faculties (labelled I, II and III) take a class, and the number of A grades assigned to students is counted. This yield the following table:

		Faculty			
		I	II	III	Total
A grade	No	56	31	46	133
	Yes	19	9	6	34
	Total	75	40	52	167

Using a chi-squared test, assess whether there is evidence to suggest that there is a significant difference between the proportion of students who achieve an A grade across Faculties.

8 MARKS

Q5b

Using the formula, the table of fitted values is

		Faculty			
		I	II	III	Total
A grade	No	59.731	31.856	41.413	133
	Yes	15.269	8.144	10.587	34
	Total	75	40	52	167

which yields $X^2 = 3.753$, which we need to compare with the χ^2_2 distribution upper 0.05 point, which is 5.991. Hence we do not reject the null hypothesis that the proportion of A grades is the same across Faculties.

- (c) The **large sample** version of the non-parametric Mann-Whitney-Wilcoxon (MWW) test proceeds as follows: suppose that samples from populations 1 and 2 are available. Let the sample sizes be n_1 and n_2 , where $n_1 > n_2$. Let R_1 and R_2 be the **sum of the ranks** for sample values from **population 1** and **population 2**, respectively. The test statistic Z is given by

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \quad \text{where} \quad U = R_2 - \frac{n_2}{2}(n_2 + 1).$$

If the null hypothesis of **equal population medians** is **true**, Z is approximately distributed as a $Normal(0, 1)$ random variable.

Use the large sample MWW procedure to test for equality of medians if the sample sizes are $n_1 = 15$ and $n_2 = 12$, with $R_2 = 233$. 4 MARKS

Based on this evidence, would you assess that the sample median for the sample from population 2 is greater than the sample median for the sample from population 1? Justify your answer. 2 MARKS

Q5c

Here we have

$$U = R_2 - \frac{n_2(n_2 + 1)}{2} = 233 - 12 \times 13/2 = 155$$

and thus $Z = 3.172$. This lies in the two-sided rejection region $\{|Z| \geq 1.96\}$, so we reject the null hypothesis of equal medians.

The sum of the ranks is always equal to $n(n + 1)/2 = 378$, so $R_1 = n(n + 1)/2 - R_2 = 145$. Hence $R_2 > R_1$, and it is therefore likely that the sample median is greater in population 2.