on - Stat-403/ 650/ 890 (BPK)
Final Exam

Answer all questions in the exam booklets. All parts have equal worth. Be succinct - if you are writing a thesis, you are doing far too much!

Cheese Please

This question is adopted from the `StatSci.org` story pages about this topic.

As cheddar cheese matures, a variety of chemical processes take place. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. Overall taste scores were obtained by combining the scores from several tasters.

A portion of the raw data is presented at the end of the examination. Each line in the data table represents one batch of cheese manufactured using the specified predictor variables. The relevant variables are

- *Taste*: Subjective taste test score, obtained by combining the scores of several tasters. Larger values indicate a more pleasant taste.
- *Acetic*: Natural logarithm of concentration of acetic acid.
- *H2S*: Natural logarithm of concentration of hydrogen sulfide.
- *Lactic*: Concentration of lactic acid.

1. Why was the average score from the multiple tastes used as the response variable as opposed to the individual scores?
   Solution: The multiple score on the SAME BATCH of cheese are all pseudo-replicates and not independent of each other.

   Several students said that the average would be used to reduce the variation caused by individual tasters. While that is true, that is NOT the core issue. If you prepared several batches of cheese (all using the same concentrations of the chemicals) and each judge only tasted one batch, it would

NOT be necessary to average the readings and the individual readings could have been used. The key issue is the SAME batch (experimental unit) was sampled by multiple tasters (observational unit).

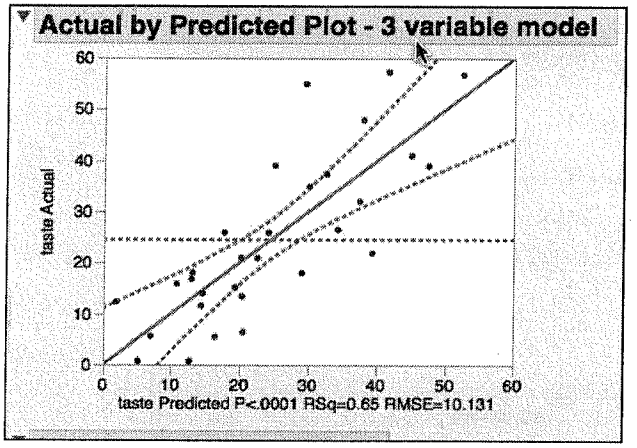
Also note that we implicitly assumed that each judge only tasted one batch of these. If judges also sampled multiple batches, this makes the analysis more difficulty because now a judge is a block, may not have sampled all batches (an incomplete block), with pseudo-replication (each batch measured multiple times). Give me a call.

Several students also said that the averages will be "less biased". Averaging does not remove bias. Bias is a 'hidden' feature of the PROCEDURE (e.g. the measuring instrument was not calibrated properly) and averaging does not remove bias. See the relevant section of my course notes for a discussion of bias, precision, and accuracy.

2. Interpret the estimated coefficient for the *H2S* variable.
Solution: The estimated mean taste increases by 3.9 units for every unit change in log *H2S* assuming all other variables are kept fixed, i.e. a marginal increase by 3.9 taste units.

3. The analyst also ran a simple regression of *taste* vs. only *acetic acid*, i.e. a single variable regression. This gave a p-value for the slope associated with *acetic acid* of 0.0017. Yet in the three variable model, the p-value associated with the slope for *acetic acid* was 0.9420. Why the dramatic change in the p-value between the two models?
Solution: Acetic acid is highly correlated with the other two variables and so must be redundant. This is an example of the effects of multi-collinearity where other (correlated) variables can suppress the marginal effect of another variable.

4. Two diagnostic plots were produced for the 3 variable models. What do you conclude from these plots? For each plot, give one example of a what a bad plot might look like and how to modify the regression model to "fix the problem".
Solution: This first diagnostic plot examines the relationship between the observed values (Y-axis) and the predicted values (X-axis):

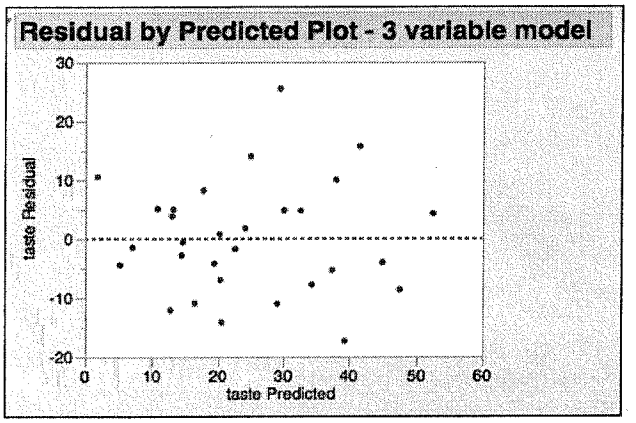


If the model fits well, then the points should fall above or below the 45° ($Y = X$) reference line. The dashed lines are NOT confidence limits on the curve and have no easy interpretation.

We see that the scatter above/below the reference line is approximately equal which indicates that the standard deviation around the fitted curve is roughly equal (although the residual plots makes it easier to see this). There are no obvious outliers (points well above or below the reference line). Note that points outside the dashed line are NOT outliers. The dashed line has nothing to do with individual points. There is no evidence of leverage (a single point in the upper right part of the plot). The relationship doesn't appear to be missing any explanatory variables because the scatter doesn't have any obvious pattern.

A bad plot could have outliers (run the model with the outliers removed), leverage points (run the model with the leverage points removed), or show a curvilinear relationship (add more explanatory variables such as the square of a variable or a new explanatory variable). See my course notes for details.

The second diagnostic plot is the residual plot. The residuals (difference between observed and predicted values) plotted against the predicted values. The residual plot is actually identical to the previous plot if you rotate the reference line (and points) by 45°in a clockwise fashion.



If the model fits well, then the points should fall above and below the reference line at 0 in a random scatter.

The scatter above/below the reference line is approximately equal which again indicates that the standard deviation around the fitted curve is roughly equal. There are no obvious outliers or pattern to the residuals.

A bad residual plot could show outliers (points well above or below 0 line), leverage (a solitary point to the right of the plot), or non-linearity (a non-random scatter), or increasing variation with predicted values (a funnel shape to the residuals). For outliers and/or leverage points, re-run the regression excluding these points; for non-linearity, add new explanatory variables (the square of some variables or new variables); for unequal variance, a weighted least squares, a transformation (such as $\log Y$) are often used. See my course notes for details.

Some students indicated that the plots showed that the data had a normal distribution. You can't tell that from the plots. As well, the RESIDUALS need to have a normal distribution and not the raw data – see my course notes for details.

5. The model was used to predict the taste score at new values of the three variables. Refer to the output at the end of the exam.

   Show how the predicted value was computed, i.e. give the equation.

   Compare and contrast the two intervals given on the output. **Hint: Carefully think about what is the actual response variable.**
   Solution: The predicted value is found by substituting in the individual values of X into the regression equation:

$$\widehat{Y} = -28.9 + 0.328(5) + 3.91(6) + 19.67(1.5)$$

   The confidence interval for the mean response indicates the uncertainty about the estimated average taste (over the entire population of people and not just those tasters in the experiment) at those X values.

The prediction interval measures the uncertainty in the measured response for a batch of cheese with those attributes. Here is where you have to be careful. **Many students just parroted the phrase "variation in individual responses", but the response in this experiment is sample AVERAGE of several tastes.** Hence the prediction interval shows the uncertainty in the sample average of several tasters. That was the purpose of the hint given above.

If you want the variation in the scores for individual tasters, you would have to add an additional source of variation (variation of taster values around the sample mean) which would make the interval even wider.

6. Suppose you had complete freedom to DESIGN this study and were able to "choose" values for the three predictor variables. Describe how you would design and run this study. The potential range of predictor variables is close to that seen in the raw data presented.
Solution: There are many ways to design this experiment. You could do an ANOVA type of design where the X variables are classified into categories; you could also continue with a regression type of design. The key considerations are:

   - Contrast. Choose levels of each X variable that has a wide range as possible and practical. You don't want to chose levels of a variable that are so similar that it is impossible to detect and effect.
   - Avoid collinearity. Choose combinations of levels of X so that low value of each X variable are matched against high values of other X variables and vice versa. A factorial ANOVA type model would automatically do this. With regression types of experiments, you need more care in choosing the X variables.
   - Avoid pseudo-replication. While having multiple tasters test each batch is fine (and correctable), a better experiment would make multiple batches at the same combination of X values to assess batch-to-batch variation rather than taster-to-taster variation.
   - Randomize, randomize, randomize.
   - Blinding.

   Many students ignored the question asking about DESIGN considerations and gave alternate ways to analyze the data. This missed the entire point of the question.

And they'll have fun, fun, fun, ...

This question is adopted from the `StatSci.org` story pages about this topic.

Groups of dolphins were observed off the coast of Iceland near Keflavik in 1998. The data here give the time of the day and the main activity of the group, whether travelling quickly, feeding or socializing. The dolphin groups varied in size - usually feeding or socializing groups were larger than travelling groups.

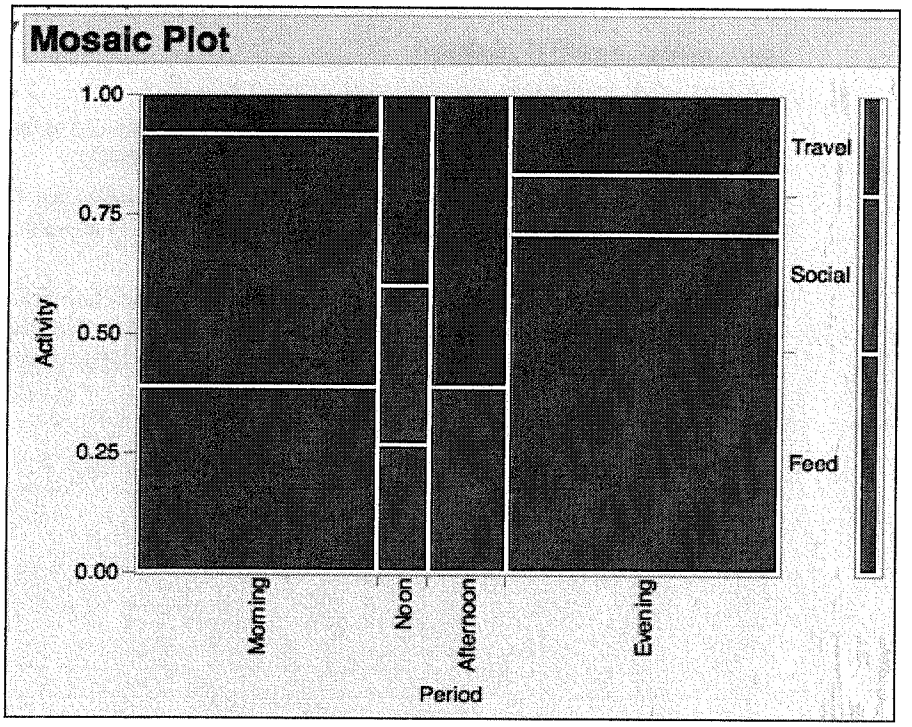
Parts of the raw data are shown at the end of the exam. The data table has the following three variables:

- *Activity.* Main activity of group: travelling (Travel), feeding (Feed) or socializing (Social)
- *Period.* Time of the day: Morning, Noon, Afternoon or Evening
- *Groups.* Number of groups observed

7. Identify the factor(s), their levels, the treatments, and the response variable.
   Solution: Factor is *time* of the day with 4 levels. Response variable is the *activity*. The count is NOT the response variable, but simply indicates how dolphin groups were observed in that combination of time and activity.

   Some students indicated that the factor was the activity and the response was the time of day. While technically the chi-square test is symmetric in the treatment of the factor and response, the way the experiment had to be conducted, i.e. you go out on the water at various times of day and observe activities implies the time of the day is the factor. We didn't randomly select a dolphin that was feeding and then see what the time of day was.

8. Create a suitable plot.
   Solution: Create a mosaic plot similar to the following:



It isn't necessary to have the bars unequal width to reflect the different sample sizes. Nor is it necessary to color the bars.

A side-by-side bar chart could also be done.

If you use a pie chart, it is instant failure for the course. See `https://www.stevefenton.co.uk/Content/Pie-Charts-Are-Bad/` and other similar sites.

Some student reversed the roles of time of day and activity. Again, while technically not incorrect, it is not preferred (see solution to previous question).

9. What hypothesis is being "tested"?
Solution: We are interested in examining if the PROPORTION of dolphin groups engaged in different activities is the same across the times of day. So the hypothesis being tested is that the PROPORTION dolphin groups engaged in various activities is the same across the various time periods.

Some students said that the proportions must all be equal to 0.33 (because there are 3 activities). This is incorrect. Just like in ANOVA, we don't actually care what the actual proportions are – we are just interested if there is evidence that they differ across the times of day.

Some students talked about the mean time in each activity. The time in

each activity was NOT measured.

10. What do you conclude?
Solution: There was strong evidence that the proportion of groups in each activity differed among the time of day levels ($p < .0001$).

Some students tried to state their conclusions in terms of mean, e.g. "there was evidence that the mean proportion differed . . .". There is no such thing as a mean proportion.

Some students concluded that the "there was evidence that the activities varies among the times of day". It makes no sense to talk about individual activities – it is the PROPORTION of groups in each activity that is of interest.

Misc topics

11. Many student use a yellow highlighter to annotate their notes to make it easier to study. Does this have any effect for short term memory?

    Consider an experiment where subjects are presented with a word list of 25 words for one minute to memorize. One factor is that the subjects can highlight words or not highlight the words. Another factor is the level of distraction where the levels are silence or classical music playing in the back ground. Explain how you could run this experiment as a CRD, RCB, or split-plot design (only one variant of a split-plot is necessary).
    Solution: There are 2 factors (highlighting and music) each with two levels. There are 4 treatments.

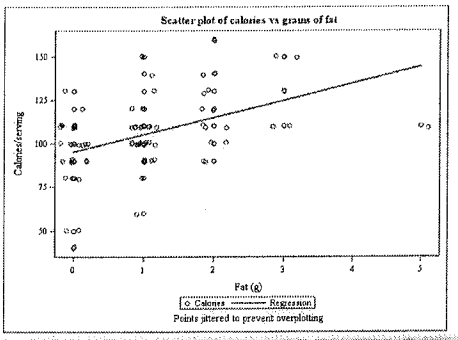
    CRD: Students are randomly assigned to one of the four treatments. Each student does only one treatment. Each treatment is replicated with multiple students.

    RCB. There are many ways to run this experiment as a RCB. For example, you could use subjects as blocks, and then have each student do all 4 treatments in random order. Or you could block by day. On each day, you have 4 separate students and each student does one of the 4 treatments so that all 4 treatments are run in each day.

    Split-plot. The key is now that students do NOT do all four treatments, but do two of the four treatments. For example, students could be assigned to the music or silence group. Then each student does both highlighting levels (in random order). You have to be careful here. Suppose you had two rooms and in each room either silence or music was playing. Then if students within a room did both the highlighting levels, this looks like a split-plot design, but there is a subtle problem – there is NOT replicates at the music level. The experimental unit for music is the ROOM, not the individual student, i.e. you've committed pseudo-replication. For example, suppose that the room where music was playing also had a bad smell. You would be unable to distinguish the music effect from the smell effect.

    Some students claimed that only 4 students are needed for the CRD. If you only use 4 students, you would have only a single replicate for each treatment and analysis is not possible.

12. A sample of breakfast cereals was analyzed measuring the number of calories and the grams of fat per serving. The following is some output from analysis. Write a short results paragraph.



Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	95.13158	3.14122	30.28	<.0001	88.87394	101.38922
fat		1	9.80601	2.20690	4.44	<.0001	5.40964	14.20237

Solution:
The relationship between the calories/serving and the grams of fat/serving was investigated using linear regression (Figure 1). The fitted equation is

$$Calories = 95 + 9.8(Fat)$$

There was strong evidence that the slope is different from 0 ($p < .0001$). For every gram of fat, the calories/serving is expected to increase by 9.8 (SE 2.2) calories/gram of fat.

Common problems in solutions from students include:

- Reporting too many decimal places. Seldom do you need to report more than two significant digits.
- The intercept is usually not of interest and so you don't usually spend anytime discussing it.
- Many student reported the t-values. These are not of interest and should not be reported.
- The whole point of regression is to estimate the slope. So the discussion needs to be about the slope. Many students discussed "differences in means" (which is not sensible), or "differences in the mean among groups" which is again not sensible. These students were likely confusing regression with ANOVA.
- Don't just give the table values as "facts' – add some interpretation to the information in the table. For example, many student had sentences such as "The parameter estimate for Fat was 9.8. The standard error was 2.21. The t-value was 4.44 and the p-value was

< .0001 so we rejected the null hypothesis". These types of sentence provide no useful information to the reader over and above the table.

13. In the following table are some observations from an experiment on the length of time needed to stack boxes. Estimate the main effect of sex.

	method (a)	method (b)	method (c)
m	20, 25, 30	32	40, 45, 50
f	35	35, 40, 45	51

Solution: We first compute the average response in each of the 6 cells of the study. The sex effect for each method is the difference in averages within a method. These differences are $10 = 35 - 25$; $8 = 40 - 32$; and $6 = 51 - 45$. The main effect of sex is the average of the sex effects or $(10 + 8 + 6)/3 = 8$.

14. Give two examples of pseudo-replication - one of which is "fixable" and one of which is "not fixable." You will make your life easier if the same experiment is used for both examples.
Solution:
Non-fixable.. Two tanks each with 5 fish. One tank has a chemical added to the water; the other tank is control. The growth of individual fish is measured. This is pseudo-replicated because the experimental unit is the tank, but the observational unit is the fish. There are NO real replicates as there is only 1 tank that is treatment and 1 tank that is control. When you find the average growth of fish in each tank, you get two numbers that cannot be analyzed.

Fixable. Four tanks, each with 5 fish. Two tanks have chemical added to the water; two tanks are control. The growth of individual fish is measured. This is still pseudo-replicated because the experimental unit is the tank, but the observational unit is the fish. But, now after averaging over the fish in a tank, you end up with 4 numbers, 2 for treatment and 2 for control and you can now analyze the averages.

15. **Bonus**. It is well known that if your parents didn't have children, you also won't have children. Why type of inheritance is this, e.g. autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive, co-dominant, or mitochondrial.

Part of the raw output for the Cheese question.

Part of the raw data:

	taste	Acetic	H2S	Lactic
1	12.3	4.543	3.135	0.86
2	20.9	5.159	5.043	1.53
3	39	5.366	5.438	1.57
4	47.9	5.759	7.496	1.81
5	5.6	4.663	3.807	0.99
6	25.9	5.697	7.601	1.09
7	37.3	5.892	8.726	1.29
8	21.9	6.078	7.966	1.78
9	18.1	4.898	3.85	1.29
10	21	5.242	4.174	1.58

Results from a three-variable model for the cheese experiment:

Summary of Fit

RSquare	0.651775
RSquare Adj	0.611595
Root Mean Square Error	10.13071
Mean of Response	24.53333
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	4994.4756	1664.83	16.2214
Error	26	2668.4111	102.63	Prob > F
C. Total	29	7662.8867		<.0001*

Parameter Estimates

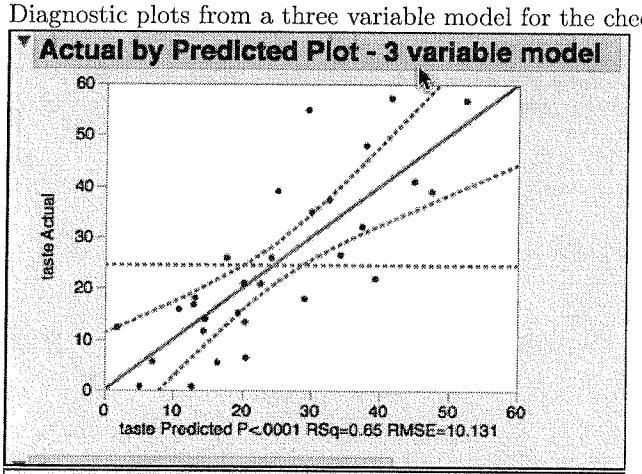
Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	-28.87677	19.73542	-1.46	0.1554	-69.4435	11.689964	.
Acetic	0.3277413	4.459757	0.07	0.9420	-8.83942	9.4949022	1.8315892
H2S	3.9118411	1.24843	3.13	0.0042*	1.3456559	6.4780263	1.9921996
Lactic	19.670543	8.629055	2.28	0.0311*	1.9332671	37.40782	1.9379116

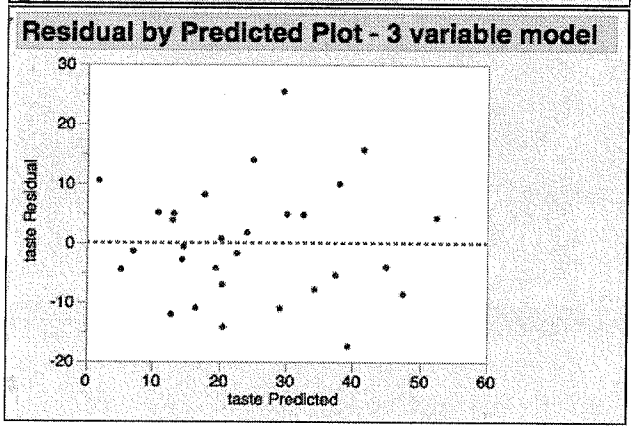
Results from a model using acetic acid only for the cheese experiment:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	-61.49861	24.84638	-2.48	0.0196*	-112.3941	-10.60311	.
Acetic	15.647767	4.495773	3.48	0.0017*	6.4385941	24.85694	1

Diagnostic plots from a three variable model for the cheese experiment:





Predictions and intervals from the three variable model for the cheese experiment:

	taste	Acetic	H2S	Lactic	Predicted taste
1	•	5	6	1.5	25.74
2					

icted taste	Lower 95% Mean taste	Upper 95% Mean taste	Lower 95% Indiv taste	Upper 95% Indiv taste
25.74	19.42	32.06	3.98	47.50

Part of the raw output for the Dolphin question.

Part of the raw data for the dolphin survey:

	Activity	Period	Groups
1	Travel	Morning	6
2	Feed	Morning	28
3	Social	Morning	38
4	Travel	Noon	6
5	Feed	Noon	4
6	Social	Noon	5
7	Travel	Afternoon	14
8	Feed	Afternoon	0
9	Social	Afternoon	9
10	Travel	Evening	13
11	Feed	Evening	56
12	Social	Evening	10

Part of the output from the dolphin analysis:

Contingency Table

Period	Count / Total % / Col % / Row %	Feed	Social	Travel	
		Activity			
Morning		28	38	6	72
		14.81	20.11	3.17	38.10
		31.82	61.29	15.38	
		38.89	52.78	8.33	
Noon		4	5	6	15
		2.12	2.65	3.17	7.94
		4.55	8.06	15.38	
		26.67	33.33	40.00	
Afternoon		0	9	14	23
		0.00	4.76	7.41	12.17
		0.00	14.52	35.90	
		0.00	39.13	60.87	
Evening		56	10	13	79
		29.63	5.29	6.88	41.80
		63.64	16.13	33.33	
		70.89	12.66	16.46	
		88	62	39	189
		46.56	32.80	20.63	

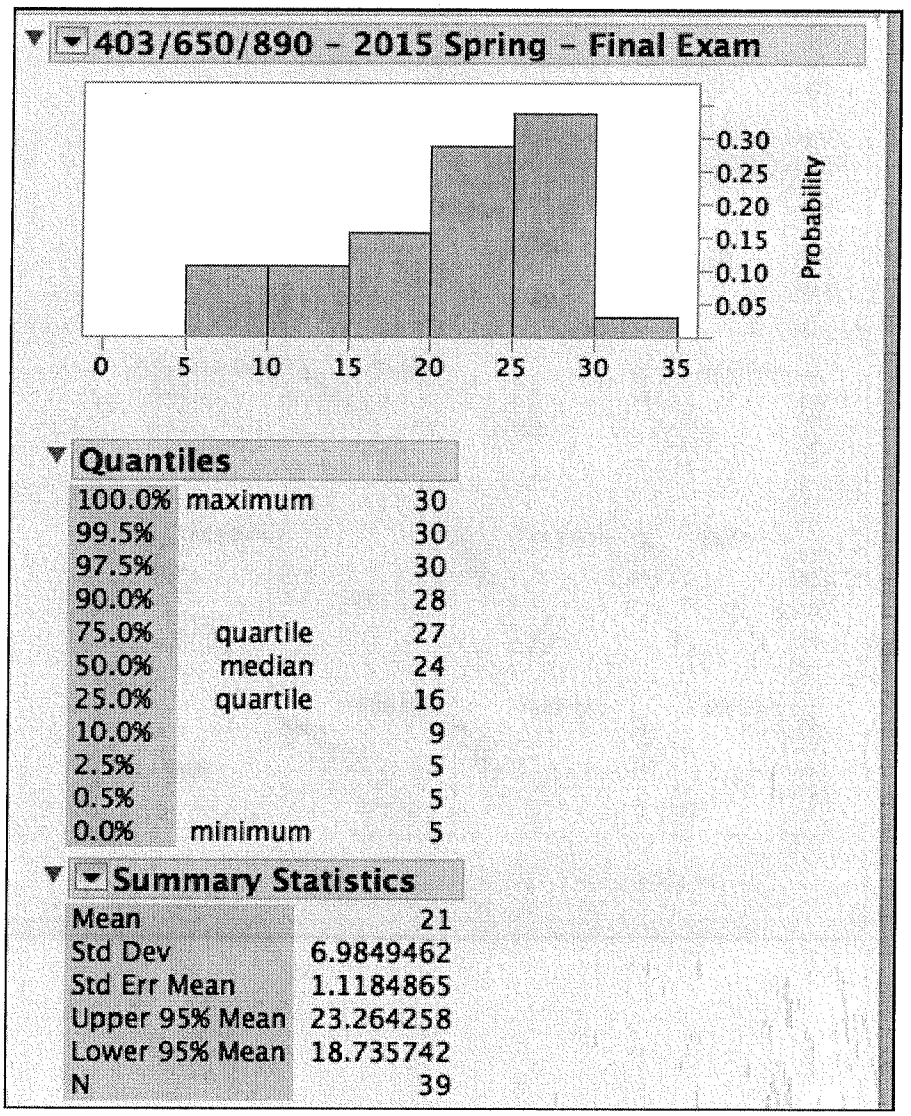
Tests

N	DF	-LogLike	RSquare (U)
189	6	37.215041	0.1880

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	74.430	<.0001*
Pearson	68.465	<.0001*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Statistics about student performance



Quantiles

100.0%	maximum	30
99.5%		30
97.5%		30
90.0%		28
75.0%	quartile	27
50.0%	median	24
25.0%	quartile	16
10.0%		9
2.5%		5
0.5%		5
0.0%	minimum	5

Summary Statistics

Mean	21
Std Dev	6.9849462
Std Err Mean	1.1184865
Upper 95% Mean	23.264258
Lower 95% Mean	18.735742
N	39

Stat 403/ 650/ 890 – Spring 2015 – Correlations among components

	Project	T01	T02	Final
Project	1.0000	0.7310	0.6250	0.6056
T01	0.7310	1.0000	0.7421	0.7544
T02	0.6250	0.7421	1.0000	0.8837
Final	0.6056	0.7544	0.8837	1.0000

There are 7 missing values. The correlations are estimated by REML method.

Scatterplot Matrix