Solutions

STAT 651 Final Test (2 hours) December 12th, 2012

NAME: Total number of Marks: /35

Answer all the questions in the exam (8 questions are on both sides of the paper). There are questions in this paper. Unless stated otherwise, do all tests at the 5% level.

When conducting a test, always state your null and alternative. Also state the distribution and/or the test that you use.

Write your solutions in the question paper. Good luck.

(1) (i) In an independent two sample statistical test, where we test $H_0 : \mu_1 - \mu_2 \leq 0$ against $H_A : \mu_1 - \mu_2 > 0$, the null hypothesis was not rejected. This means a type II error could have been committed. TRUE or FALSE (no justification is required). [1]

TRUE

(ii) Suppose the population variance is 1. I want to test $H_0 : \mu \leq 0$ against $H_A : \mu > 0$. Suppose the alternative is true. Consider the following two scenarios.

(A) I draw a sample of size 40, the true mean is $\mu = 2$.

(B) I draw a sample of size 60, the true mean is $\mu = 4$.

For which scenario (A or B) am I most likely to reject the null (no justification required)?

(B) Larger sample size and alternative further away from null

[1]

(2) You want to decide whether a 6 sided die (the plural is dice, it's a cube with dots on the sides) is fair or not. So you roll the die 300 times. The outcomes are given below DO NOT DO THE TEST.

1	2	3	4	5	6
55	45	57	43	48	52
50	50	50	50	50	50

(i) Looking at the above table explain whether you believe the die is fair. There is no correct answer to this question, I just require a logical explanation. [2]

On average with a fair die we would expect about 50 for each number. Just comparing this to what is observed there does don't appear to be much difference. Thus it is plausible that the die is fair.

(ii) State the null and alternative and the test you need to do to decide if the dice is fair or not (don't do the test). [2]

H_0: Probability of any one number is $1/6$

H_A: Probability of any one number is not $1/6$

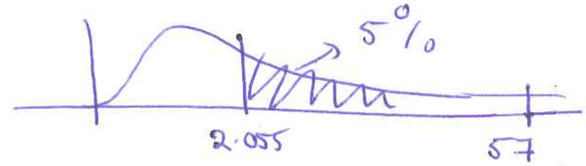
χ^2 goodness of fit test

(3) Suppose the average height of oak trees from several states in the US are being compared. Independent samples are being drawn from eight states, and an ANOVA is done.

	Sum of Squares	df	Mean square	F	Sig.
Between Groups	40	7	40/7 = 5.7	57	0.0..
Within Groups	20	200	20/200 = 0.1		
Total	60	1			

(i) Test the hypothesis H_0 : All the means are the same against the alternative H_A : At least one of the means are different. Do the test at the 5% level. [2]
Useful information: $F_{200,7}(0.05) = 3.252$, $\boxed{F_{7,200}(0.05) = 2.055}$ and $F_{7,140}(0.05) = 2.0755$.

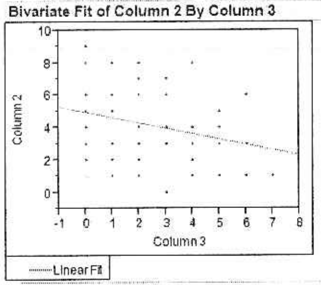
Since 57 > 2.055



(ii) What are the main assumptions to do an ANOVA and how would you check for them? [1]

- Everything is independent of each other
- Normality
- Variance for each group is about the same.

Bivariate Fit of Column 2 By Column 3



Linear Fit

Column 2 = 4.9639089 - 0.3287063*Column 3

Summary of Fit

RSquare	0.075947
RSquare Adj	0.062155
Root Mean Square Error	2.024148
Mean of Response	4.115942
Observations (or Sum Wgts)	69

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	22.56164	22.5616	5.5066
Error	67	274.51083	4.0972	Prob > F
C. Total	68	297.07246		0.0219*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.9639089	0.435842	11.39	<.0001*
Column 3	-0.328706	0.140076	-2.35	0.0219*

Figure 1: JMP output for Question 4 on M&Ms

(4) It is known that the mean number of M&Ms in a fun size bag is 18. The M&M quality control want to investigate whether there is a relationship between the number of blue and yellow M&Ms in a bag. A random sample of 69 M&M mini bags was taken. The number of blue M&Ms was plotted against the number of yellow M&Ms (see Figure 1, the x-axis corresponds to yellow, the y-axis the blues).

(i) Looking at the plot, what you know about M&Ms and the normal distribution, do you believe the distribution of blue and yellow M&Ms are normally distributed (give a reason for your answer)? [1]

No, M&Ms take numerical discrete values, 1, 2, 3, ...
Whereas normally dist. r.vs are continuous.

(ii) Given the JMP output do you believe that the number of yellow M&Ms has an influence on the number of blue M&Ms (pose this in a statistical sense and give a reason for your answer). [2]

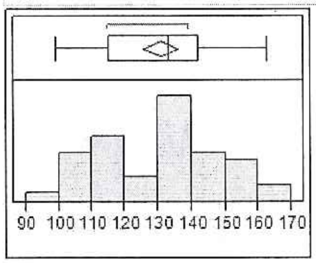
Yes. The slope is -0.3 and its p-value is $0.0219 < 0.05$.
Thus it appears statistically significant at the 5% level.

(iii) Give a reason as to why there may be negative correlation between the number of blue and yellow M&Ms? [2]

On average there are 18 M&Ms in a bag. If there are
more blues it is likely there are less yellows.

Distributions

Wt 7



Quantiles

100.0%	maximum	163
99.5%		163
97.5%		163
90.0%		155
75.0%	quartile	142.25
50.0%	median	133
25.0%	quartile	115.25
10.0%		108
2.5%		100
0.5%		99
0.0%	minimum	99

Summary Statistics

Mean	130.59091
Std Dev	16.761415
Std Err Mean	2.5268784
Upper 95% Mean	135.68685
Lower 95% Mean	125.49497
N	44

Figure 2: JMP output for Question 5 on the weight of 7 week calves

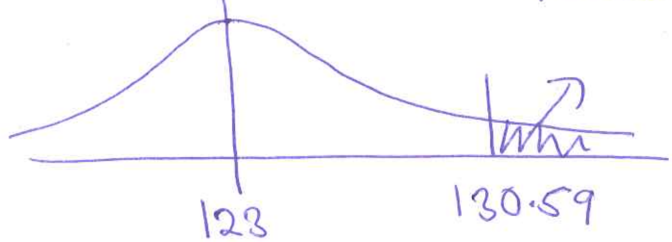
(5) A farmer is interested in the mean weight of 7 week old calves. Previously he was feeding his cows corn feed. It is known that the mean weight of a 7 week calf that is fed corn feed is 123 pounds. In the past year the Farmer has changed the feed to grass, and wants to know whether this has led to an increase in the weight of his calves. To test this hypothesis he takes the weight of 44 7 week old calves. The data is summarised in Figure 2. To simplify the calculations in this question you can assume, without any penalisation, that the standard deviation given in the output is the true population standard deviation (and was not estimated from the data) - if you don't that also does not matter.

(i) Using the output in Figure 2, state precisely the null and alternative (research) hypothesis and do the test at the 5% level. [2]

$H_0: \mu \leq 123$ $\quad$ $H_A: \mu > 123$

the s.e $= 2.52$

$$\text{p-value} = P\left(Z > \frac{130.59-123}{2.52}\right) = P(Z > 3.012) = 0.0012.$$



Since $0.0012 < 0.05$ there is enough evidence to reject the null and suppose the mean weigh is greater than 123 ponds.

(ii) Construct a 95% CI for the weight of a randomly selected 7 week old calf fed on grass. [2]

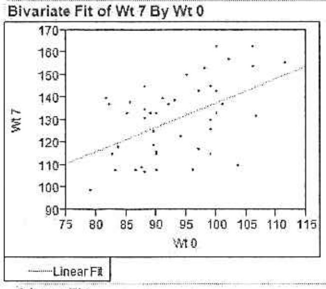
$$\left[130.5 - 1.96\sqrt{16.76^2+2.52^2},\ 130.5+1.96\sqrt{16.76^2+2.52^2}\right]$$

$$= [97, 163]$$

(iii) Nice, old, vegetarian Granny Nandi wants to buy a 7 week old calf, but she would like that the calf is over 92 pounds. Using your answer in (ii), would you recommend she buys a grass fed calf (give a reason for your answer)? [1]

Since 92 does not lie in the CI above and the CI above is the interval where the mean weight of a calf is most likely to lie, it seems reasonable to suppose that most calves fed on grass fed will weigh more than 92 pounds.

Bivariate Fit of Wt 7 By Wt 0



Linear Fit

Wt 7 = 29.826307 + 1.0809807*Wt 0

Summary of Fit

RSquare	0.251654
RSquare Adj	0.233836
Root Mean Square Error	14.6714
Mean of Response	130.5909
Observations (or Sum Wgts)	44

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	3040.135	3040.14	14.1237
Error	42	9040.501	215.25	Prob > F
C. Total	43	12080.636		0.0005*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	29.826307	26.90331	1.11	0.2739
Wt 0	1.0809807	0.287636	3.76	0.0005*

Figure 3: JMP output for Question 6 on the relationship between weights at week 0 and week 7

(6) A farmer wants to investigate whether the birth weight of a calf can be used to predict the weight of a calf at week 7. To investigate this 44 calves were weighed at birth and seven weeks. The JMP output is given in Figure 3 (Figure 4 may also be mildly useful too).

(i) Using the output what is the prediction equation for predicting the average weight of 7 week old based on its week 0 weight? [1]

$\hat{Y} = 1.08 \times \text{week0} + 29.82$

(ii) Predict the weight of a calf whose birth weight is 60 pounds, and explain why we need to be cautious about this prediction. [1]

$\hat{Y} = 1.08 \times 60 + 29.82 = 94.78$ pounds.

60 pond lie outside the range of birth weights - extrapolation.

(iii) Using the output what is the estimated correlation coefficient between the week 0 (birth) and week 7 weights? [1]

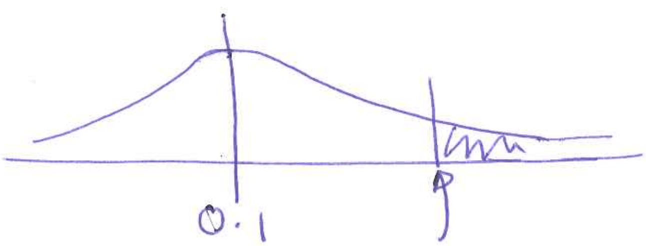
$\hat{\rho}_{xy} = 0.5009$

(iii) Test whether the correlation between the weight at week zero (birth) and weight at week 7 is greater than 0.1 (this is the research hypothesis). State precisely your null and alternative. [3]

s.e for correlation coefficient is $\sqrt{\frac{1-0.251}{42}} = 0.133$

$H_0: \rho \leq 0.1$

$H_A: \rho > 0.1$



$0.1 + 0.133 \times 1.68$

$= 0.32$

since $0.5001 > 0.32$ there is enough evidence to reject the null and suppose the true correlation is greater than 0.1.

Matched Pairs

Difference: Wt 7-Wt 0

Wt 7	130.591	t-Ratio	17.0819
Wt 0	93.2159	DF	43
Mean Difference	37.375	Prob > |t|	<.0001*
Std Error	2.18799	Prob > t	<.0001*
Upper 95%	41.7875	Prob < t	1.0000
Lower 95%	32.9625		
N	44		
Correlation	0.50165		

Figure 4: JMP output for Question 7, the difference between birth and week 7 weight

(7) The same farmer wants to investigate whether the mean weight of a calf at week 7 is greater than the mean weight at birth. He uses the same data that he collected above to do the test. The result of the test is given in Figure 4.

(i) By using the results in Question 6, what test must be used for this data set (give a reason for your answer). [1]

Since there is correlation between weight 0 and weight 7 of a calf we need to use a paired t-test.

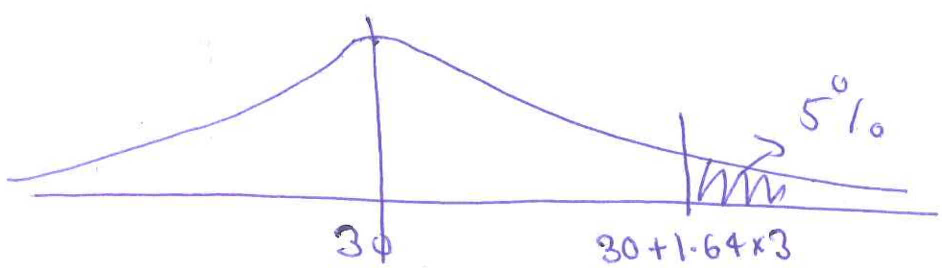
(ii) Test the research hypothesis that the mean weight at week 7 has increased since birth (state precisely the null and alternative). [2]

$H_0: \mu_7 - \mu_0 \leq 0$ $\quad H_A: \mu_7 - \mu_0 > 0$.

the p-value < 0.0001 hence enough evidence to reject null and suppose mean weight has increased.

(8) Suppose X is a normally distributed random variable with standard deviation $\sigma = 3$. To answer this question I strongly suggest you make a sketch.

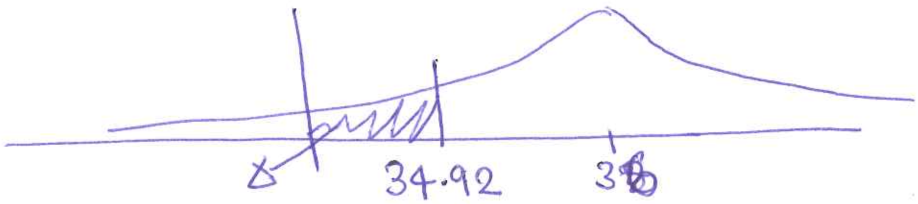
(i) Suppose the mean of X is $\mu = 30$, calculate the value x such that the probability of drawing a value greater than ~~above~~ x is 5% (in other words calculate the value on the x-axis of the normal distribution, such that the area to the right of this value is 5%, or equivalently the rejection region when the null has mean 30 and the standard deviation is 3). [2]



$30 + 1.64 \times 3$

$= 34.92$

(ii) Suppose the mean of the random variable is $\mu = 38$, calculate the probability of drawing a value less than the number x you calculated in (i) when the true mean is 38 (if you did not calculate x use the value $x = 35$ instead). [2]



$P\left(Z < \frac{34.92 - 36}{3}\right) = 0.36$. Probability of drawing less than 34.92 when the mean is 36 is 36%.

(iii) It is believed that blood glucose levels follow a normal distribution with mean μ and standard deviation 3. A persons glucose level is considered good if their true mean is $\mu = 30$. Their blood level is determined to be very high if their true mean is $\mu = 38$. A patient has her glucose level taken and it is $X = 34$, using your answers in (i) and (ii) can we say anything about the patients mean glucose level (is it good or high). [Hint: You can articulate this as a test where the null is the good level and the alternative is the high level.] [2]

Suppose we do a test $H_0: \mu \leq 30$ $H_A: \mu > 30$

The rejection region is a blood level greater than 34.92. In this case we cannot reject the null. Hence there is not enough evidence it is greater than 30. However, the chance of not rejecting the null when the level is high (ie 36) is 36%, which is quite large. Hence we are unable to say whether the patient has high ~~blood~~ level or not.

(iv) State one very simple measure that the doctors office could implement to improve the sensitivity of the above test [hint: sample size] (give a reason for your answer)? [1]

If we take a few blood samples and take the average this will be a better estimator of the mean. And it will allow us to more easily determine whether some one has a high sugar level or not.