STAT 651 Midterm Test (75 minutes) - 12:45pm - 2:00pm October, 2012

NAME: **Total number of Marks: /25**

There are 5 questions in this paper, do not be deterred, they are all straightforward. Read each question carefully. There are questions on <u>both</u> side of the page. The number of marks for each question are given in brackets. Be smart about how you answer. If you can't answer one question move on the to next and return to the questions you could not do after answering all the other questions! There are three JMP outputs in this question.

Rubric: This exam is an open book exam you can use all written materials that you want, normal tables and a calculator.

Write your solutions in the question paper.

GOOD LUCK!

(1a) Suppose you want to construct a 99% confidence interval for the mean height of the human population. You do not observe the whole population, but you have the choice of the following samples.

(A) The standard deviation of a random selected male height is 3 inches and you observe the heights of 100 randomly selected males.

(B) The standard deviation of a randomly selected female height is about 2.6 inches and you observe the heights of 100 randomly selected females.

(C) The standard deviation of a randomly selected human height is about 3.1 inches and you observe the heights of 50 randomly selected individuals (a mixture of males and females).

(D) None of the above.

Choose one option and give a reason for your answer. [1]

(b) Ebony draws 500 samples, each sample is of size 30. For each sample, Ebony constructs a 99% CI, on average how many of these confidence intervals will contain the true mean? [1]

(2) (i) A certain medical test is based on counting the number of abnormal cells from a patient's blood sample. The probability of the observed number of abnormal cells given that the patient is healthy is evaluated. If this probability turns out to be below 5%, the person is requested to take further medical check-up. Pose this problem as a statistical test, giving the null and alternative hypotheses, and the type I or type II errors if they are known. [2]

(ii) Let μ denote the mean number of abnormal cells in a blood sample. If a person is normal, then their mean number of abnormal cells is $\mu = 10$. A statistical test, to test the hypothesis $H_0 : \mu \leq 10$ against $H_A : \mu > 10$, is done at the 5%. It is known if the mean number of abnormal cells is above 50 the patient has a serious condition. The power in the test for the alternative $H_A : \mu \geq 50$ is 98%.

It is found that for a certain patient the sample mean number of abnormal cells is $\bar{X} = 20$. The above test was done and we were unable to reject the null. What can you say about the likelihood of the patient having this severe condition. Give a reason for your answer. [2]

(3) A winery is trying to determine whether the gender of an individual has an influence of the wine preference. The winery was looking at wine A and wine B. To see whether there was any influence, they asked 1000 randomly selected volunteers to taste the wine. A scarf was put round the volunteers eyes and they were each given a taste of both wines (not knowing which was which) and the ordering for each volunteer was randomly assigned. To the question which wine did they prefer, they could only give one of two answers, wine A or wine B.

The summary of the result is given below:

- Out of the 1000, 600 prefered wine A and 400 prefered wine B.
- There were 600 males in the group.
- 200 females prefered wine A.

(i) Explain why the volunteers were not told which wine was which and the ordering the wines were given was randomly changed. [1]

(ii) Based on this group of volunteers, is there an influence of gender on wine preference, explain your answer. [2]

(4) The difference in weights before and after a diet of 25 volunteers is measured (evaluated as After diet weight - Before diet weight). The JMP output is summarised in Figure 1 (next page).

(i) Using the information from the output and treating the standard deviation (for now) as known (use a normal distribution) calculate a 99% CI for the mean weight difference. [1]

(ii) Comment on the accuracy of the 99% CI based on plots of the data in Figure 1. [1]

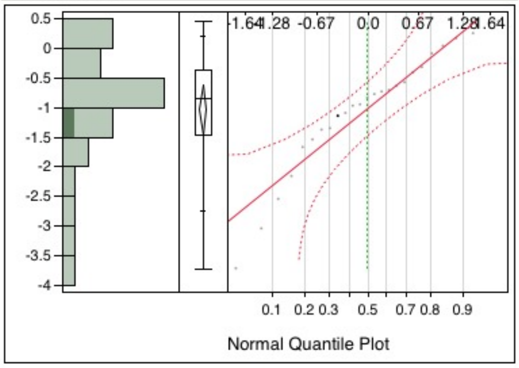
(iii) Approximately how large a sample size should we use in order that the 99% confidence interval for mean weight difference has length 0.2. [1]

(iv) Use the information from Figure 1 and treat the standard deviation (for now) as if it were known (use a normal distribution) calculate the probability that the sample mean will be less than -1.03, when true mean difference is zero. [2]

(v) Suppose we want to test that the diet was effective for weight loss. State the null and alternative hypothesis and do the test at the 5% level (again assume that the standard deviation is known). [2]

Distributions

Column 2



Quantiles

100.0%	maximum	0.43985
99.5%		0.43985
97.5%		0.43985
90.0%		0.18878
75.0%	quartile	-0.3758
50.0%	median	-0.8465
25.0%	quartile	-1.4655
10.0%		-2.7475
2.5%		-3.7156
0.5%		-3.7156
0.0%	minimum	-3.7156

Summary Statistics

Mean	-1.037607
Std Dev	1.0093195
Std Err Mean	0.2018639
Upper 95% Mean	-0.620981
Lower 95% Mean	-1.454234
N	25

Figure 1: JMP output for Question 4.

(vi) Briefly comment on whether the p-value for the above test calculated with the standard deviation estimated from the data will be smaller or larger than the p-value calculated when the standard deviation is known. Give a reason for your answer.

[1]

(vii) Suppose we want to test whether the diet lead to an <u>increase</u> in weight (assume the standard deviation is unknown and has to be estimated from the data). State the null and alternative hypothesis and do the test at the 5% level. [1]

(5) University Police department are trying to determine whether policing Ireland street (in particular giving citations to cyclists) is having an impact on the number of accidents that happen on Ireland. To do the analysis, for 15 months they refrain (this was very hard for them) from policing the street and for each month they count the number of accidents that happen. The average number of accidents over these 15 months is 2.26.

For another 17 months they police the street (they were more comfortable doing this) and each month they count the number of accidents which happen. The average number of accidents over these 17 months is 1.58.

A summary of the data is given in Figure 2, where $Column2 = 0$ refers to the case when no policing took place and $Column2 = 1$ when policing took place.

An independent t-test was done to test whether policing reduced the number of accidents, these results are reported in Figure 3.

(i) From how the data is collected and the plots in Figure 2 what sort of variable (ie. numerical continuous, categorical, numerical discrete, binary etc.) are the number of accidents per month? [1]

(ii) Do you believe this data is close to normally distributed, give a reason for your answer. [1]

(iii) Test the research hypothesis that policing Ireland reduced the number of accidents. State precisely the null and alternative, and use the JMP output do the test at the 5% level. State the standard error and also the distribution the test uses. [2]

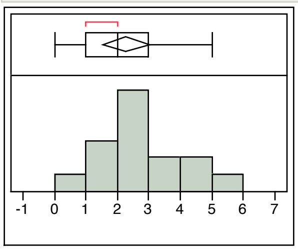
(iv) Based on Figures 2 and 3, have the assumptions to do the test been satisfied? [1]

(v) An adminstrator queries why the need to do the test, he says 'clearly the average number of accidents after policing has gone down'. Explain why a statistical test was necessary. [2]

two_sample3: Distribution

Distributions Column 2=0

Column 1



Quantiles

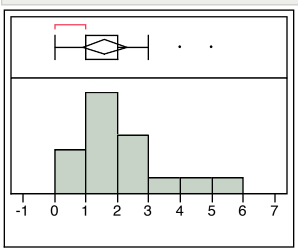
100.0%	maximum	5
99.5%		5
97.5%		5
90.0%		4.4
75.0%	quartile	3
50.0%	median	2
25.0%	quartile	1
10.0%		0.6
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statis

Mean	2.26
Std Dev	1.33
Std Err Mean	0.34
Upper 95% Mean	3.0
Lower 95% Mean	1.52
N	

Distributions Column 2=1

Column 1



Quantiles

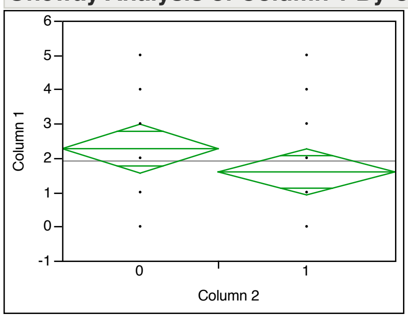
100.0%	maximum	5
99.5%		5
97.5%		5
90.0%		4.2
75.0%	quartile	2
50.0%	median	1
25.0%	quartile	1
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statis

Mean	1.58
Std Dev	1.37
Std Err Mean	0.33
Upper 95% Mean	2.29
Lower 95% Mean	0.88
N	

Figure 2:

Oneway Analysis of Column 1 By Column 2



Oneway Anova

Summary of Fit

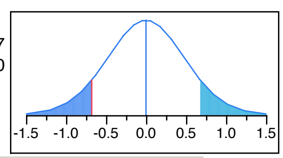
Rsquare	0.062463
Adj Rsquare	0.031212
Root Mean Square Error	1.354634
Mean of Response	1.90625
Observations (or Sum Wgts)	32

t Test

1-0

Assuming equal variances

Difference	-0.6784	t Ratio	-1.41377
Std Err Dif	0.4799	DF	30
Upper CL Dif	0.3016	Prob > |t|	0.1677
Lower CL Dif	-1.6585	Prob > t	0.9161
Confidence	0.95	Prob < t	0.0839



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Column 2	1	3.667770	3.66777	1.9987	0.1677
Error	30	55.050980	1.83503		
C. Total	31	58.718750			

Means for Oneway Anova

Figure 3: