Solutions

STAT 651 Midterm Test (75 minutes) - 2:20pm - 3:35pm March, 2014

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

There are 8 questions in this paper, do not be deterred, they are all straightforward. Read each question carefully. There are questions on both 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 3 JMP outputs in this paper.

Rubric: This exam is a closed book exam, but you can use a 2-sided cheat sheet, normal and t-tables and a calculator.

Write your solutions in the question paper.

GOOD LUCK!

(0) [0.5]

(1) What type of variable is:

(i) The ethnicity of a person. Categorical [1]

(ii) The number of pets in a household. Numerical discrete [1]

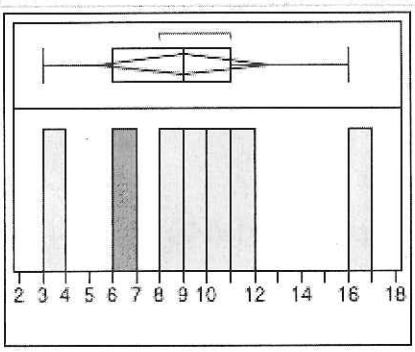
(2) You are given the data set

3, 6, 8, 9, 10, 11, 16.

It's summary statistics are summarized below:

Distributions

data



Quantiles		
100.0%	maximum	16
99.5%		16
97.5%		16
90.0%		16
75.0%	quartile	11
50.0%	median	9
25.0%	quartile	6
10.0%		3
2.5%		3
0.5%		3
0.0%	minimum	3

Summary Statistics	
Mean	9
Std Dev	4.0824829
Std Err Mean	1.5430335
Upper 95% Mean	12.775667
Lower 95% Mean	5.224333
N	7

The numbers 11 and 16 are increased. What will happen to the following (answer stay the same, increase or decrease)? [2.5]

(i) Mean Increase

(ii) Median Stay same

(iii) First Quartile Stay same

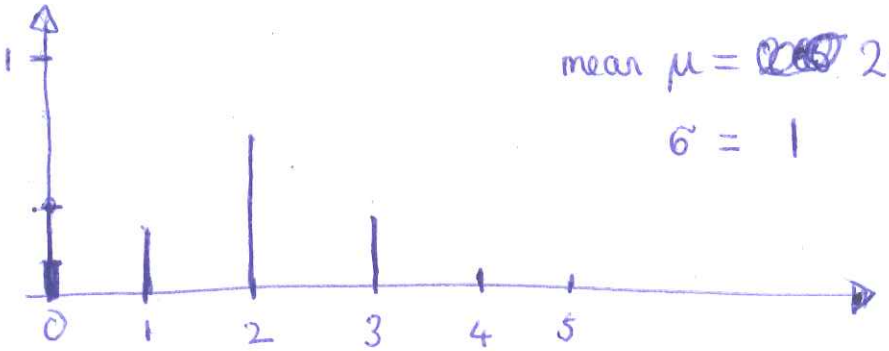
(iv) Third Quartile Increase

(v) Interquartile range. Increase

(3) The objective of a survey is to find the mean number of cars per household. A survey samples 200 households, and evaluates the average number of cars (sample mean) for this sample.

(a) What will the distribution of the number of cars look like (give a rough sketch of the likely histogram, including the mean and standard deviation - use your intuition for this part)? [2]

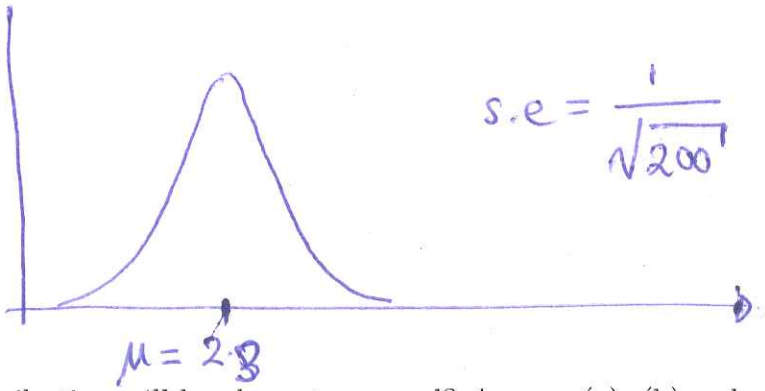
Distribution of Numerical Discrete.



mean $\mu = 2.8$

$\sigma = 1$

(b) What will the distribution of the sample mean look like (give a rough sketch of the histogram, remember it should be based on the numbers you used in plot (a))? [2]



$s.e = \frac{1}{\sqrt{200}}$

(c) Which distribution will be closer to normal? Answer (a), (b) or both. [1]

(b) [by the central limit theorem]

(4) A population has a mean μ and standard deviation σ. From this population 1000 samples are drawn each sample consists of between 100-150 observations ($n = 100 - 150$). For each sample a 95% confidence interval for the mean is constructed.

On average how many of these samples will contain the population mean μ? [2]

On average $0.95 \times 1000 = 950$ intervals will contain the mean μ.

(5) A biologist wants to estimate the yield of cotton plants. Before she starts the experiment, she needs to decide how many plants to grow such that the margin of error of the 95% confidence interval for the mean is **0.5**. The standard deviation (σ) for the amount of cotton grown in a plant is known to be between $\mathbf{3-7}$ (σ is between 3-7). Calculate the minimum number of plants she needs to grow to be sure her Margin of Error is **less** than 0.5. [2]

Use $\sigma = 7$ as this will give an upper bound for n.

$$1.96 \times \frac{7}{\sqrt{n}} = 0.5$$

$$n = \left(\frac{1.96 \times 7}{0.5}\right)^2 \simeq 753$$

(6) Recently there has been an interesting convergence between the three disciplines, Anthropology, Genetics and Medicine.

Researchers want to know whether Neanderthals (a possible subspecies of Homo sapiens) are one of the contributing factors as why humans have diabetes. To find out if there is a possible link, they sampled 450 people who had Neanderthal DNA and 700 people who did not have Neanderthal DNA and tested them for diabetes. The data us summarized below

	With Neanderthal DNA	Without Neanderthal DNA	
Diabetes	70	90	160
No Diabetes	380	610	990
	450	700	

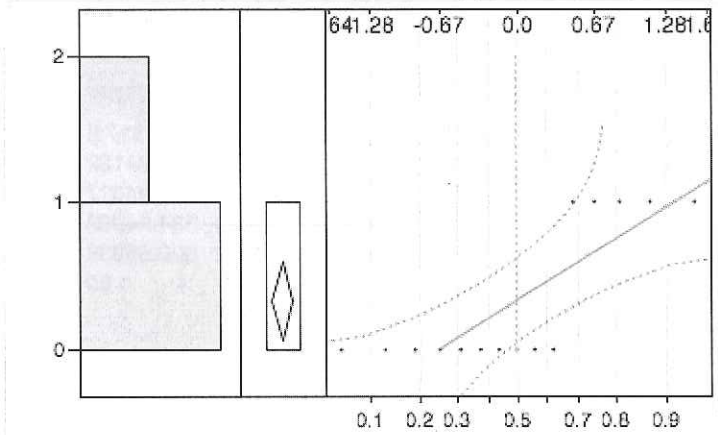
By using probabilities as the basis of your argument, is there an association (dependence) between whether the person has Neanderthal DNA and their diabetes status? [3]

The proportion of people with Neanderthal DNA who had diabetes $= 70/450 \simeq 15.5\%$.

The proportion of people without Neanderthal DNA who had diabetes $= 90/700 \simeq 12.8\%$.

The proportion of people with Neanderthal DNA who has diabetes is 15.5% is slightly leyes than those who did not have Neanderthal DNA. This suggests there is an association with having the DNA and their Diabetes status. Though we need to bear in mind this is only a sample.

Like?



Normal Quantile Plot

Quantiles

Summary Statistics

Mean	0.3333333
Std Dev	0.48795
Std Err Mean	0.1259882

(7) **15** people were interviewed to see whether they liked or disliked Marmite (a rather strange tasting British food spread). The could either answer like (coded as 1) or dislike (coded as 0). The data is summarized above.

(i) Construct a 99% confidence interval for the proportion of the population who like Marmite. [2]

$0.33 \pm 2.977 \times 0.126$

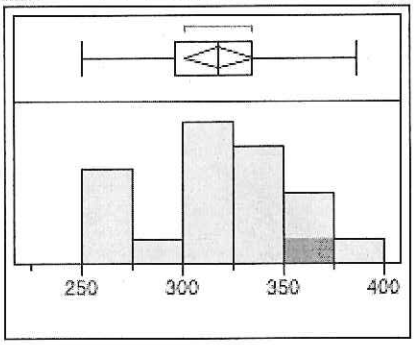
(ii) From the data, do you believe there really is **99%** confidence in the interval you have constructed above? [2]

(explain your answer)

The data is clearly not normal (infact it is binary data only 0 or 1). This means that it will take a large sample size for the sample mean to be close to normal. A sample size of 15 ~~does not~~ is not large enough. This means the 99% CI for the mean we constructed ~~for the~~ in (d) is unlikely to be have 99% confidence. The interval is not reliably a 99% confidence interval.

Distributions

Blood



Summary Statistics

Mean	317.8
Std Dev	36.624122
Std Err Mean	8.1894027
Upper 95% Mean	334.94062
Lower 95% Mean	300.65938
N	20

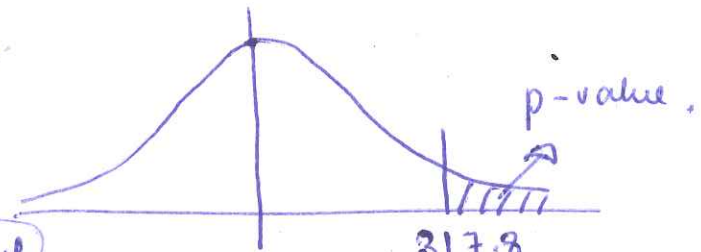
300

(8) Recently there has been a huge advertising campaign (blood drive) to increase the number of people who donate blood. Before the 'blood drive' on average the number of people who donated blood in a day was 300. After the 'blood drive', the number of people who donated blood was monitored over 20 successive days (the number of people who gave blood was counted each day). The JMP output is summarized above (note that the standard deviation has been calculated from the data).

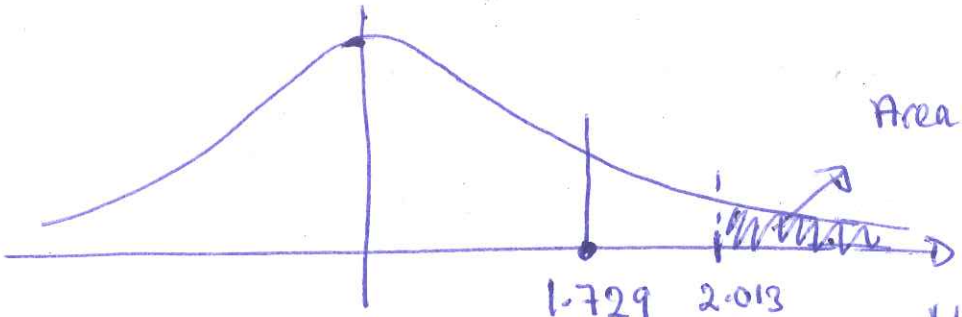
Is there any evidence that the blood drive has worked (do the test at the 5% level)? Remember to state the hypotheses of interest and all the assumptions that you make to do the test. [4]

$H_0: \mu \leq 300$ $H_A: \mu > 300$

$$t = \frac{317.8 - 300}{8.8189} = 2.017$$



Looky up the t-tables with 19 df.



5% level.

Area < 5%. The p-value so less than 5%.. Thus there is evidence at the 5% to reject the null.