STAT 651 Midterm Test (60 minutes) - 1:50pm - 2:50pm April, 2018

NAME: Total number of Marks: /25

Solutions

There are 6 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!

Rubric: This exam is a closed book exam, but you can use a 4-sided cheat sheet, normal and t-tables and a calculator.

Write your solutions in the question paper.

GOOD LUCK!

(1) [Quick questions]

(b) Consider the population standard deviation and sample standard deviation. Which changes as the sample size grows and why? [1]

The population standard deviation is fixed and depends only on the population. On the other hand, the sample standard deviation, like the sample mean, depends on the sample. As the sample size grows the sample st. dev. will tend to be closer to the population standard dev.

(c) Explain the difference between the population standard deviation and (population) standard error. Which changes as the sample size grows and how does it change? [1]

The standard error of the sample mean gets smaller as sample size grows. Since $s.e. = \frac{\sigma}{\sqrt{n}}$ $\rightarrow$ gets larger $\Big\}$ gets small.

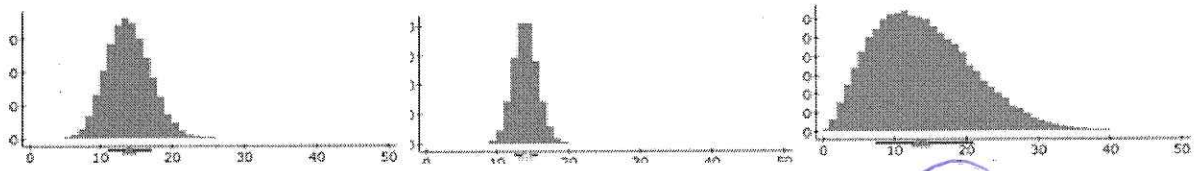
(d) Explain when a t-distribution is used instead of a normal distribution when constructing a 95% confidence interval for the mean. [1]

If we use the sample standard deviation instead of the population standard deviation then we use the t-distribution.

(e) Given the 95% confidence interval [20, 30] for the mean. What is the sample mean and what is the margin of error? [1]

$\bar{x} = 25$ $\quad$ MoE $= 5$

(e) The distribution of the **sample means** (data drawn from the same population) for various sample sizes ($n = 5, 15$ and 50) are plotted below. Match the sample size with the plot. [1]



15 $\quad$ 50 $\quad$ 5

(2) The proportion of fathers greater than 6 feet is 0.1. However, the probability a son will grow to over 6 feet **if (or given)** their father is over 6 feet is 0.8.

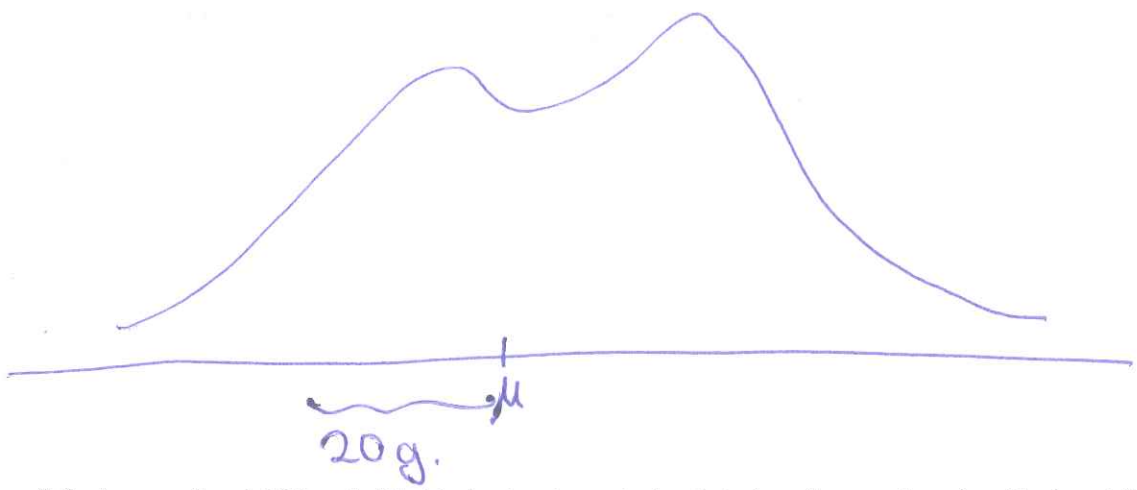
Using the information above, calculate the chance that an adult son **and** their father are both over 6 feet. [2]

$P(\text{Father} > 6 \text{ and } \text{Son} > 6)$

$= P(\text{Son} > 6 \mid \text{Father} > 6) \times P[\text{Father} > 6] = 0.8 \times 0.1 = 0.08$

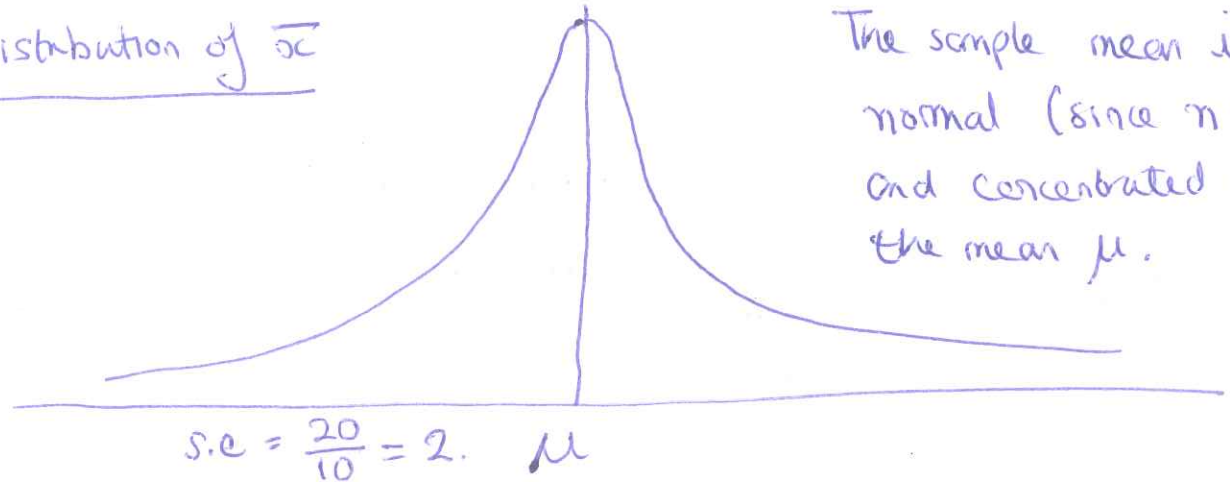
(3) Golaith birdeater tarantulas are the largest species of spiders. It is known that females tend to be **heavier** than males and the standard deviation for their weight is $\sigma = \mathbf{20g}$. The mean weight of a golaith birdeater tarantula is unknown we denote it as μ

(a) Using the information above, sketch the distribution of weight of a golaith birdeater tarantula (regardless of gender, males and females are mixed together). Making sure to indicate μ on the sketch, the standard deviation $\sigma = 20$g and any other major feature. [1.5]



(b) A sample of **100** golaith birdeater tarantulas is taken from a jungle. Each spider is weighed and the sample mean of the 100 spiders evaluated. Sketch the distribution of the **sample mean** based on $n = 100$. Use the same scale as in part (a). As in part (a) include the mean and standard error. [1.5]

Distribution of $\bar{x}$



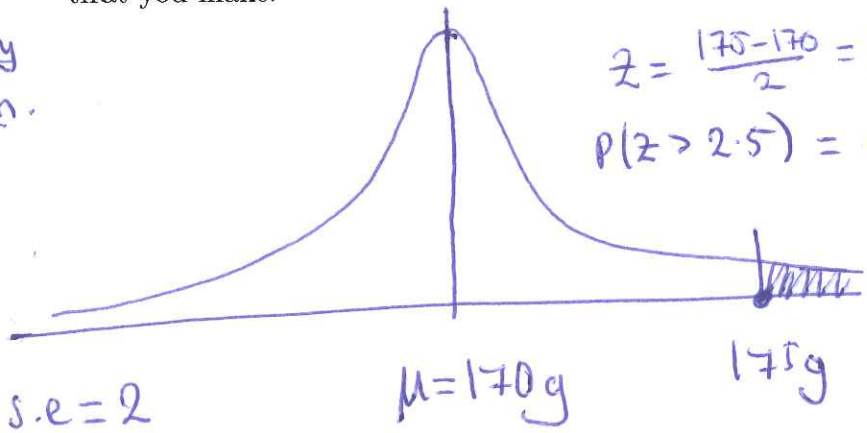
The sample mean is normal (since $n = 100$) and concentrated about the mean μ.

(c) Calculate the probability that the **sample mean** of **100** golaith birdeater tarantulas is **over** 175g given that the population mean is $\mu = 170$g. State all assumptions that you make. [3]

Assume normality of sample mean.

$z = \frac{175-170}{2} = 2.5$

$P(z > 2.5) = 0.0062$



(d) Using (c) and that $n = 100$ and $\bar{X} = 175$, test the hypothesis $H_0 : \mu = 170$g vs $H_A : \mu \neq 170$g. Conduct the test at the 5% significance level. [1]

Using the result from (c) since it is a two-sided test p-value = $2 \times 0.62\% = 1.24\%$. Reject null at the 5% level; since $1.24\% < 5\%$

There is some evidence to suggest the ~~weig~~ mean weight $\neq 170$g.

(4) Researchers want to investigate the influence hens living in white or red light has on the quality of the eggs they produce. The conjecture/idea is that red light is **less** stressful than white light, thus resulting in eggs with better protein quality i.e. hens exposed to **red light** will have a **larger haugh unit** than hens exposed to **white light**. They want to see if there is any evidence of this in the data they collect.

Two groups of chickens were put under red or white light and the protein quality of their eggs measured. The sample mean for eggs of hens under red light is **103.09** whereas the sample mean for eggs of hens under white light is **101.7**. The sample size in white light treatment group is $n = \mathbf{88}$. The sample size in the red light treatment group is $m = \mathbf{114}$. The data is summarized below.

(a) State the hypothesis of interest and the conclusion of the test (at the 5% significance level). Remember to give the p-value. [2]

$H_0: \mu_{white} - \mu_{red} \geq 0$ $H_A: \mu_{white} - \mu_{red} < 0$.

p-value = $0.63\% < 5\%$ reject null evidence to suggest that red light increases strength of egg shells.

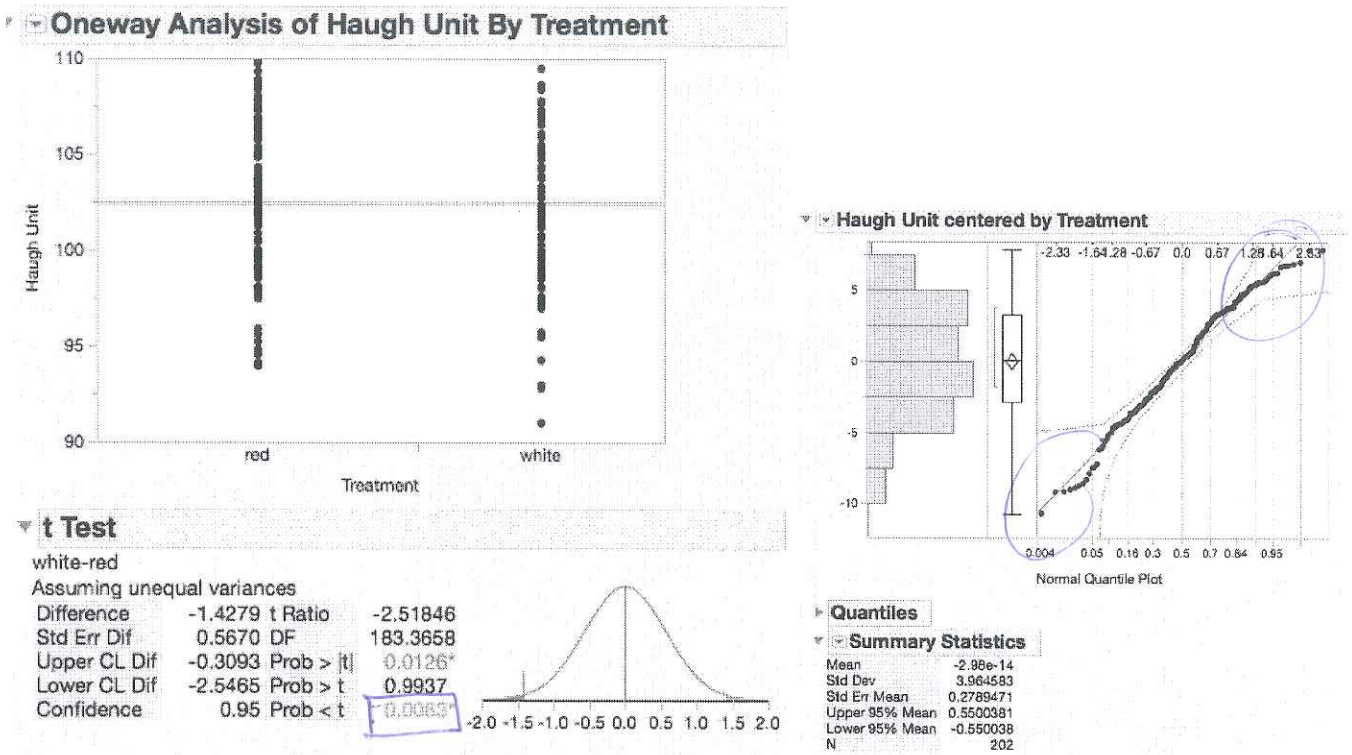


Figure 1: Left: Output for t-test. Right: QQplot of the residuals for question 4.

(b) A QQplot of the residuals is given in Figure 1. Using this QQplot discuss if (a) the data is normal (b) if the sample means will be close to normal and (c) the impact this has on p-values calculated in part (i). [2]

There is some deviation from normality in the tails of the distribution of the residuals.

However the sample sizes of 88 and 114 are sufficiently large to assume normality of $\bar{X} - \bar{Y}$. Thus the p-values calculated in the output are close to the truth.

(c) Explain why the sample mean for the residuals is zero? [1]

Residuals are defined as $X_i - \bar{X}$ and $Y_i - \bar{Y}$.

Since it centralises the data, the average of residuals is always zero.

(d) Using the data in Figure 1 test the hypthothesis that $H_0 : \mu_R - \mu_W \leq 2$ against $H_A : \mu_R - \mu_W > 2$ (where μ_R and μ_W denotes the mean haught strength under red and white light respectively).

Hint: Read the hypothesis and the output carefully (in particular how **Difference** is defined) and match them correctly. [2]

Below are the critical values for a t-distribution with 183.3df.

probability	0.3	0.15	0.10	0.05	0.025	0.01	0.005
t^*	0.52	1.09	1.28	1.66	1.99	2.37	2.64

The above hypothesis is the same as

$H_0: \mu_w - \mu_R \geq -2$ against $H_A: \mu_w - \mu_R < -2$

$\bar{w} - \bar{R} = \underbrace{-1.429}$

This is completely consistent with the null being true that we cannot reject the null.

$t = \frac{-1.429 - (-2)}{0.56} =$

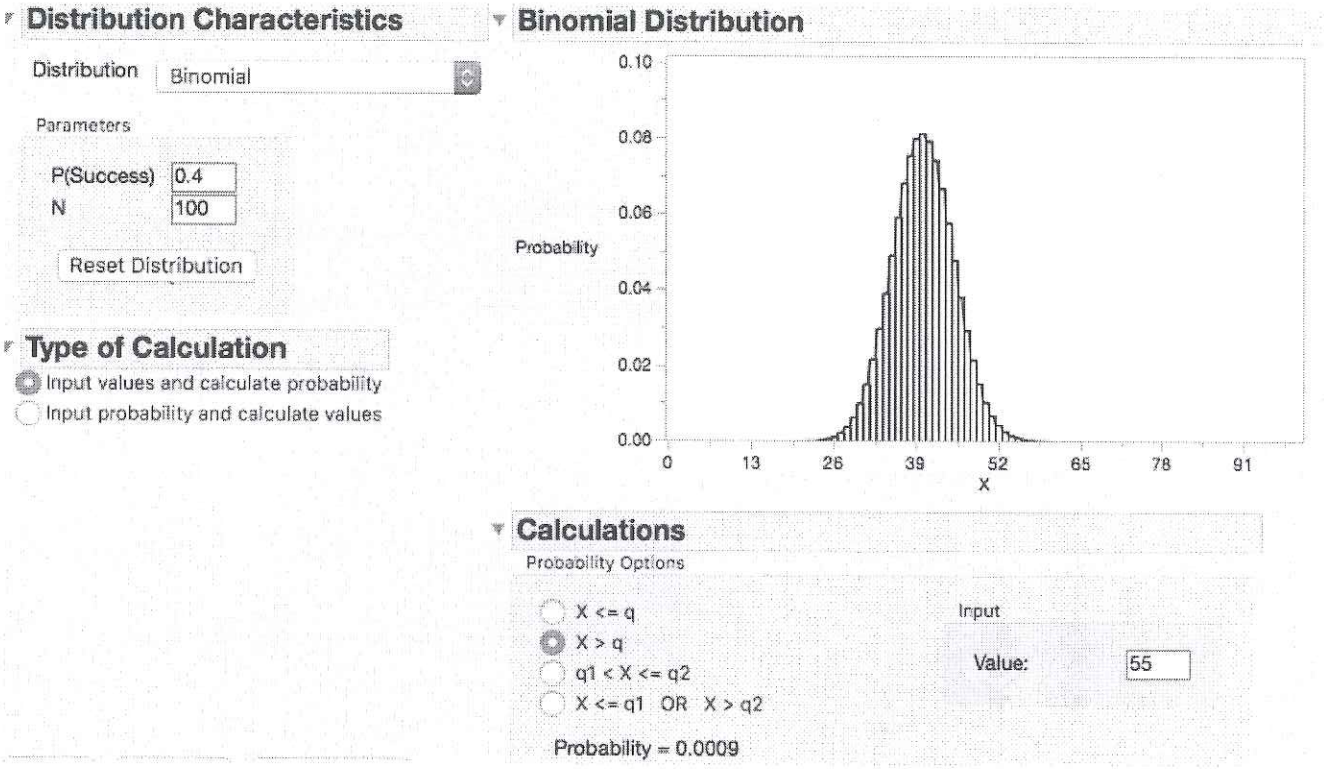
(5) [Interpretation of probabilities]

For the past 30 years, it has been well known to researchers in public health that an increase in the number of privately owned guns is associated with an increased risk of a gun related death. However, until recently, the proportion of the American public who supported tighter gun control measures was only 40% of the population.

After the recent tragic events public health researchers and sociologists are asking if the proportion of the American public who support gun control has **increased**?

A survey was recently conducted, where a simple random sample of 100 American adults were asked if they supported tighter gun controls. 56 in that sample said they did. Using the probability below, what can one say about the proportion of Americans who support gun control?

Carefully explain your answer, if you can in terms of a statistical test using the data and the calculation below. If not, in terms of the data collected and the calculation below. [2]



The ~~chance~~ probability that in a sample 55 or more out 100 will support gun control, when the proportion in the population support gun control is 40%, is 0.09%. This is very small, which strongly suggest that a better explanation for the data is that population proportion is greater than 40%.

Form $H_0: p \leq 0.4$ vs. $H_A: p > 0.4$ p-value = 0.09%

reject null, $p > 0.4$ is a better explanation for the data

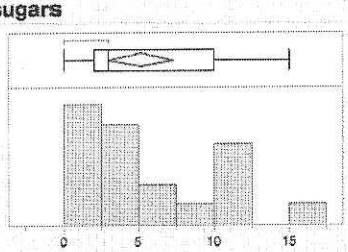
(6) [Interpretation of statistical output]

A healthy eating consumer group wants to understand if there is an association between the amount of sugar in a cereal box and its location on supermarket shelves. To see if there is, they surveyed the cereal section in a supermarket. They recorded the amount of sugar per 30 grams for each brand of cereal and the location of the cereal (**top shelf**, **middle shelf** and **bottom shelf**).

Using the data that is summarized below, write a few sentences on what you infer about the population based on the data. Back all statements with data, but you do not need to do any formal tests. [2]

Distributions shelf=Bottom

sugars

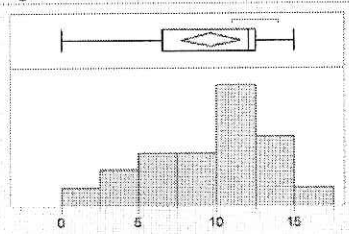


Summary Statistics

Mean	5.1052632
Std Dev	4.4832371
Std Err Mean	1.0285251
Upper 90% Mean	6.8887912
Lower 90% Mean	3.3217351
N	19

Distributions shelf=Middle

sugars

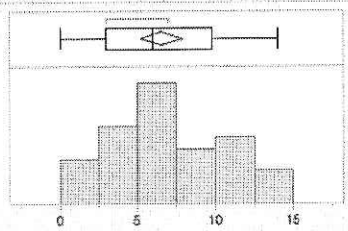


Summary Statistics

Mean	9.6190476
Std Dev	4.1288762
Std Err Mean	0.9009947
Upper 90% Mean	11.17301
Lower 90% Mean	8.0650857
N	21

Distributions shelf=Top

sugars



Summary Statistics

Mean	6.5277778
Std Dev	3.835817
Std Err Mean	0.6393028
Upper 90% Mean	7.6079262
Lower 90% Mean	5.4476293
N	36

The sample mean $\bar{x}_B = 0.51$

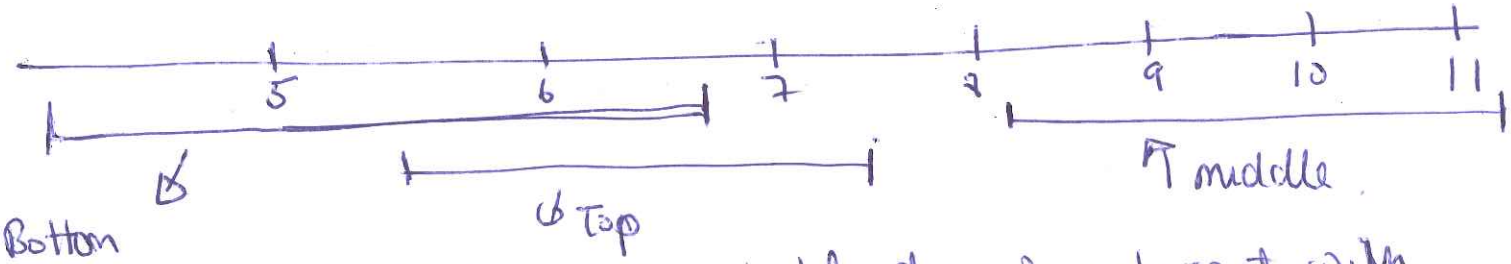
$\bar{x}_m = 9.61$

$\bar{x}_s = 6.52$

The 90% confidence intervals for ~~see bottom, middle~~ the mean amount of sugar in bottom, middle and top are

[3.32, 6.89], [8.06, 11.2]

and [~~6~~, [5.4, 7.6]



Since the ~~mm~~ CI of the middle shelf does not intersect with the other intervals, this "suggests" that the sugar content of cereals in the middle shelf is greater than the others.