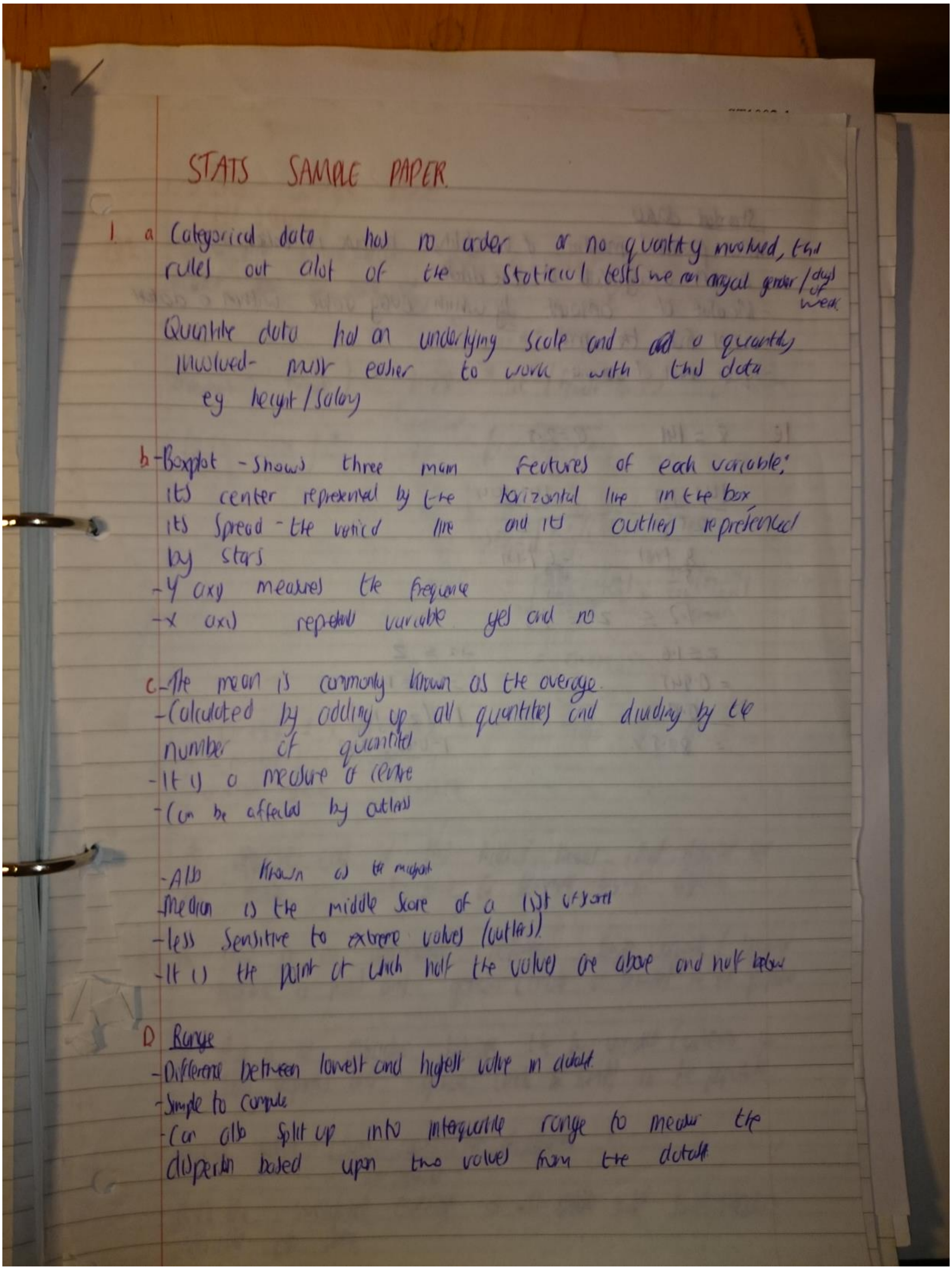


STATS SAMPLE PAPER

1 a Categorical data has no order or no quantity involved, this rules out alot of the statistical tests we can anyal gender / days of week

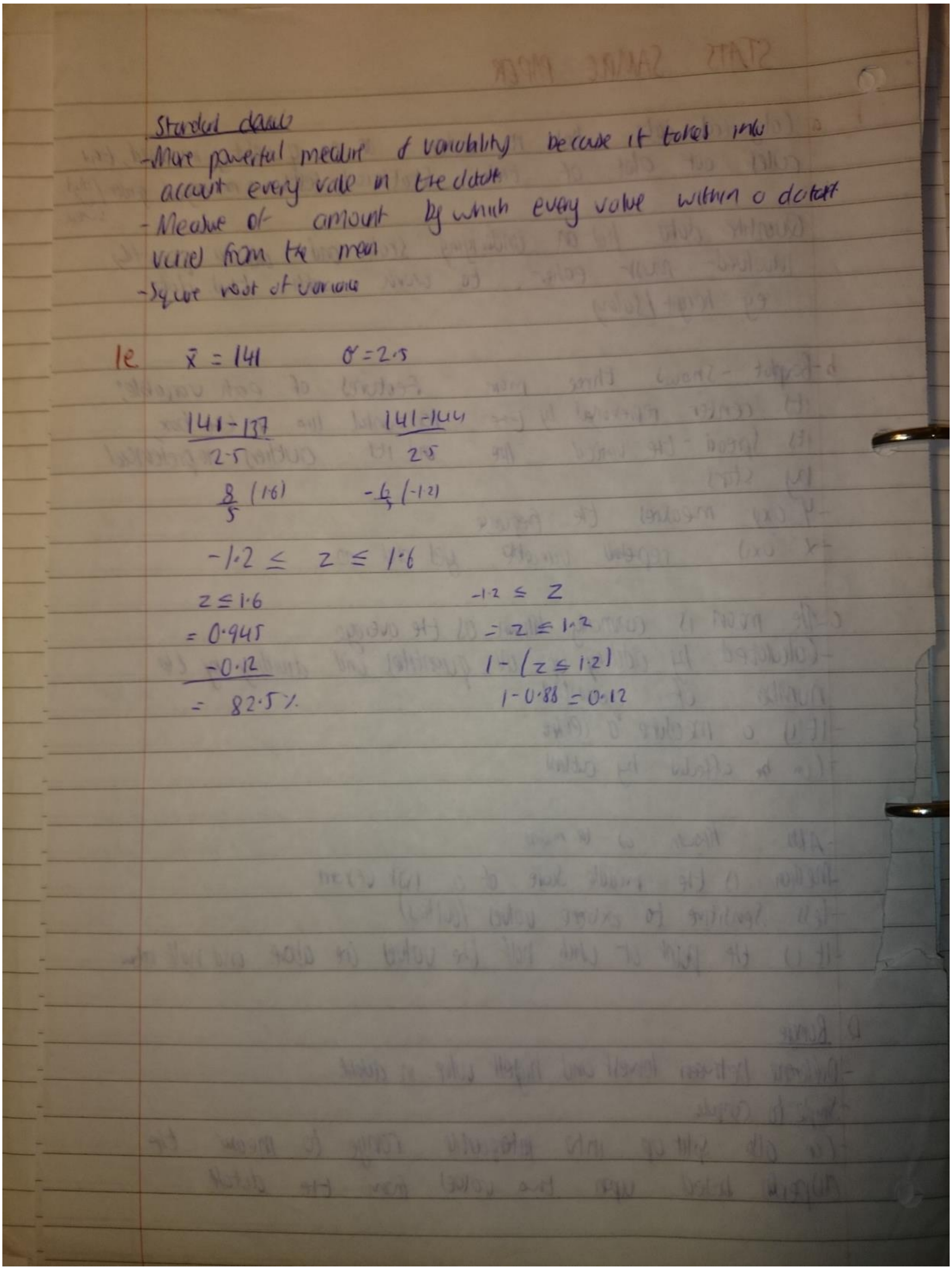
Quantitive data has an underlying scale and ~~ad~~ a quantity involved - much easier to work with this data
eg height / salary

b -Boxplot - shows three main features of each variable;
its center represented by the horizontal line in the box
its spread - the vertical line and its outliers represented by stars
-y axy measures the frequence
-x axis repreated variable yes and no

c -The mean is commonly known as the average.
-Calculated by adding up all quantities and dividing by the number of quantites
-It is a measure of centre
-Can be affected by outliers

-Also known as the midpoint
Median is the middle score of a list of sorted
-less sensitive to extreme values (outliers)
-It is the point at which half the values are above and half below

D Range
-Difference between lowest and highest value in dataset
-Simple to compute
-Can also split up into interquartile range to measure the dispersion based upon two values from the dataset



Standard deviation

- More powerful measure of variability because it takes into account every value in the dataset
- Measure of amount by which every value within a dataset varied from the mean
- Square root of variance

1e $\bar{x} = 141$ $\sigma = 2.5$

$\frac{141-137}{2.5}$ $\frac{141-144}{2.5}$

$\frac{8}{5}$ (1.6) $-\frac{6}{5}$ (-1.2)

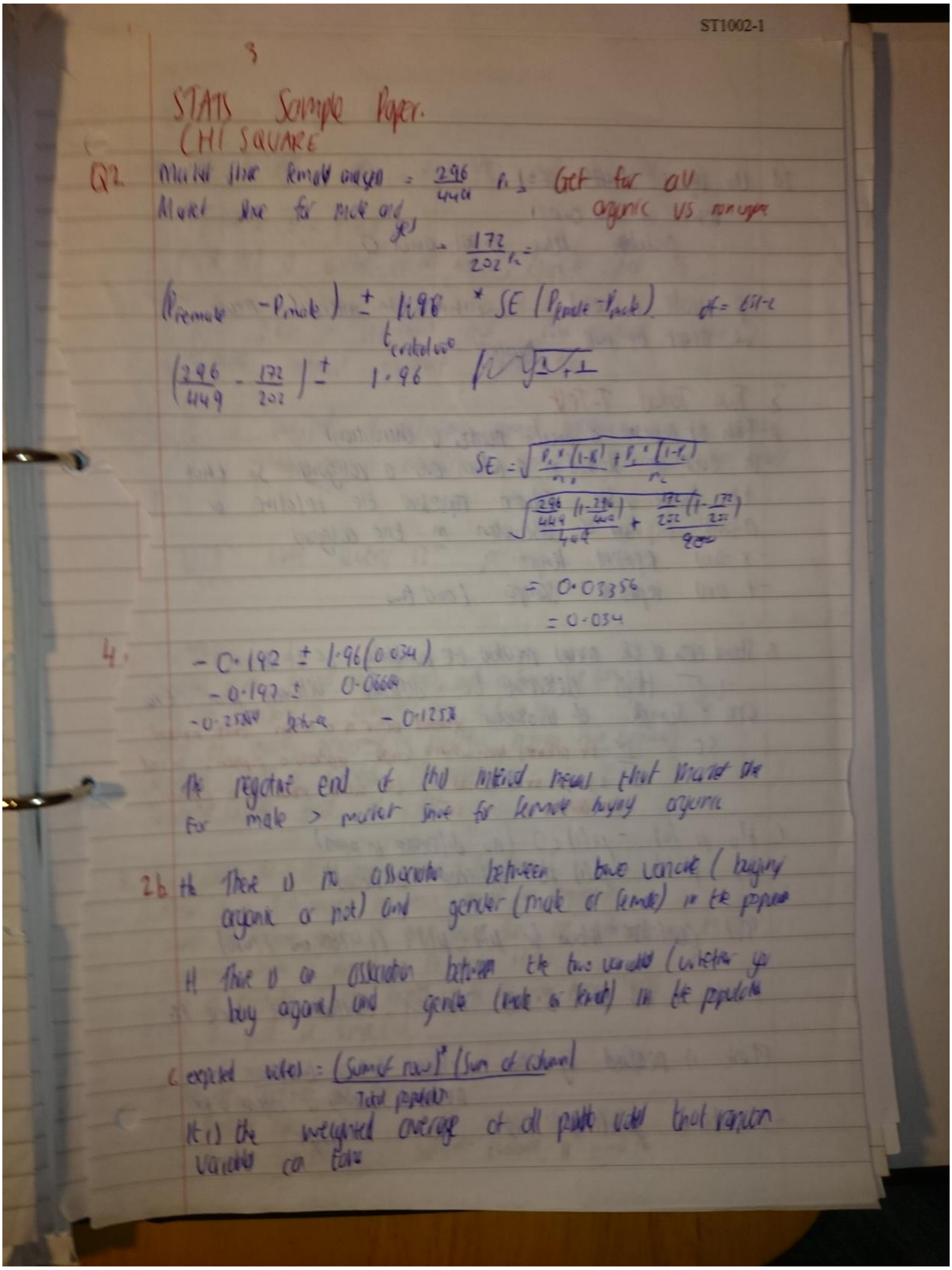
$-1.2 \leq z \leq 1.6$

$z \leq 1.6$ $-1.2 \leq z$

$= 0.945$ $= z \leq 1.2$

-0.12 $1-(z \leq 1.2)$

$= 82.5\%$ $1-0.88 = 0.12$



ST1002-1

3

STATS Sample Paper.

CHI SQUARE

Q2 Market share female organic = $\frac{296}{449}$ p_1 → Get for all organic VS non organic

Market share for male organic = $\frac{172}{202}$ p_2 =

$(p_{female} - p_{male}) \pm 1.98 * SE(p_{female} - p_{male})$ df = 651-2

$t_{critical value}$

$\left(\frac{296}{449} - \frac{172}{202}\right) \pm 1.96 \; p\sqrt{\frac{1}{\ } + \frac{1}{\ }}$

$$SE = \sqrt{\frac{p_1 * (1-p_1)}{n_1} + \frac{p_2 * (1-p_2)}{n_2}}$$

$$\sqrt{\frac{\frac{296}{449}\left(1-\frac{296}{449}\right)}{449} + \frac{\frac{172}{202}\left(1-\frac{172}{202}\right)}{202}}$$

$= 0.03356$

$= 0.034$

4. $-0.192 \pm 1.96(0.034)$

-0.192 ± 0.0666

-0.2586 between -0.1258

The negative end of this interval means that market share for male > market share for female buying organic

2b H_0 There is no association between two variables (buying organic or not) and gender (male or female) in the population

H_1 There is an association between the two variables (whether you buy organic and gender (male or female) in the population

c expected values = $\frac{(\text{Sum of row}) * (\text{Sum of column})}{\text{Total population}}$

It is the weighted average of all possible values that random variable can take

2d chi squari = 25.479 df = 1
result is < 0.001
probability is the 0.001 and 0

P vale (0.000) is within the interval - mean we acept the null hypothesis

3 Two Tailed T-Test

A - Each dot represents a Specific number of observations
- The dots are Stacked in a column over a category the height of the column represents the absolute frequency of observation in the category
- x - axis represent length
- y axis represent sample 1 and two

B Stand error of the mean measures the variability of the sampled = [illegible] Helps determine the difference between more than one sample of information standard error is a term that measures the accuracy with which a sample represents a population. In stat a sample mean deviates from the actual mean of a population, this deviation is the SE.

$$SE = \frac{\sigma}{\sqrt{n}}$$

C H_0 $\mu(0) - \mu(1) = 0$ (no difference in means)
H_1 $\mu(0) - \mu(1) \neq 0$ (difference in means)

95% confidence interval for $\mu(0) - \mu(1)$ = (1.4260 and 2.2050)

P value is less than 0.055 we can conclude we reject H_0

T-test is measured by ratio $\frac{\text{difference between two mean}}{\text{measure of variability or dispersion of scores}}$

- T vale is positive, first mean is larger

3c look up t-table with $\alpha = 0.05$ and $df = 198$

$1.97 < t < 1.97$

T (28.41) is outside confidence interval, evidence against H_0

~~Pvalue of~~

p of 0.000 is less than 0.05 and outside the confidence interval, evidence against H_0

3d It is a method of selecting sample members from a large population according to a random starting point and a fixed periodic interval

$\frac{1136}{35}$ take every 32nd person

4 a Scatterplot

- Consists of an x-axis and Y axis, and a Series of dots
- Each dot represents one observation from a data set.
- Useful for visually determining the correlation between two variables.
- Independent variable on x-axis (temperature).
- Dependent variable yield on y-axis

b The equation for yield in grams is $17.0 + 2.0$ (temp c)

- There is a 2.0 increase for each increase in temperature
- coef of constant gives the exact figure for the coefficient in the equation which is rounded to 17 from 17.002
- coef of temperature is also rounded up to ~~1~~ 2.00 from 1.99517.
- 17.0 is yield at temp (0).

- Pvalue of constant test hypotheses:

H_0 Population slope $= 0$ $\quad$ $p = 0.000$ which is < 0.001

H_1 Population slope $\neq 0$ $\quad$ Sufficient evidence to reject H_0

- p value in the constrine test:

H_0 population intercept $= 0$ $p < 0.001$

H_1 pop intercept $\neq 0$ reject H_0

- $S = 4.01967$
- S is estimate of standard deviation of y for fixed x. (SD of residuals)

- Residual should be normally distributed
- There should be no relationship between residual and predicted value

- Residual is value of observed cost (actual cost) - predicted value from eqn.
- R^2 measures the fit of the model to the data

98.4% of variance of yield is accounted for

C. $T = 75$ $17 + 2(75)$

$17 + 150 = 167$ grams