Stat110s1 Exam 2

Question 1

The number of students taking STAT115 who are majoring in Neuroscience for the five years between 2018 and 2022 are shown in the table below:

29	31	21	23	20

a) The mean number of students majoring in Neuroscience and taking STAT115 is closest to:

1

- 86.75
- 108
- 100.8
- 24.8
- 124

Solution: The mean is found by adding up all the observations and dividing by the number of observations.

$$\text{Mean} = \frac{29+31+21+23+20}{5} = 24.8$$

b) The sample standard deviation for the number of students majoring in Neuroscience and taking STAT115 over the five years is closest to:

1

- 4.92
- 24.20
- 4.40
- 6.12
- 19.36

Solution: You may be able to do this using a sample standard deviation function on your approved calculator.
Or do this by hand using the formula for the sample standard deviation $\sqrt{\sum(x_i-\bar{x})^2/(n-1)}$, where x_i is each observation, $\bar{x}$ is the sample mean, and n is the sample size.
In this question this is

$$\sqrt{\frac{(29-24.8)^2+(31-24.8)^2+(21-24.8)^2+(23-24.8)^2+(20-24.8)^2}{5-1}}$$

c) Choose one of the following which best describes the type of data in this question:

1

- Nominal categorical
- Binomial
- Continuous
- Binary
- Discrete

Solution: The number of students taking the paper must be a whole number, so the the number of students taking the paper is a discrete random variable.

Question 2

Policy makers are interested in whether New Zealand motor vehicle owners are open to using more public transport. They randomly selected motor vehicle owners from the NZTA (New Zealand Transport Agency) database of the currently registered vehicle owners, and sent them a survey about their attitidues to public transport.

a) Select the type of sampling procedure used from the options below.

1

- (•) Simple random sampling.
- ( ) Stratified random sampling.
- ( ) Cluster sampling.
- ( ) A census.
- ( ) A random variable.

Solution: You'll need to know the definitions of these terms. This example here is an example of simple random sampling.

b) In this study, the NZTA database of registered vehicle owners is an example of:

1

- ( ) A cluster.
- ( ) A random sample.
- (•) A sampling frame.
- ( ) A stratum.
- ( ) A sample.

Solution: A sampling frame is where the participants in the sample are taken from. The participants in the study were taken from the NZTA database of registered vehicle owners, so this is the sampling frame.

c) What is the population of interest in the study?

1

- ( ) The vehicle owners in the sample.
- ( ) The vehicles in the sample.
- ( ) Vehicle owners who were open to using public transport.
- ( ) All New Zealand vehicles.
- (•) All New Zealand vehicle owners.

Solution: From the question brief, "policy makers are interested in whether *New Zealand motor vehicle owners* are open to using more public transport."

Question 3

Out of students taking 4 papers per semester at The University of Otago, the table below shows the probability distribution for the number of papers a student passes.

Papers Passed	Probability
0	0.02
1	0.05
2	0.04
3	?
4	0.49

a) What is the probability a student passes exactly three papers?

1

- We don't know as it isn't given
- 0.4
- 0.04
- 0.05
- 0.49

Solution: Use the fact that the sum of all the probabilities must add to 1. This means the probability of passing 3 papers is one minus the probability of passing the other number of papers:

$$1 - 0.02 - 0.05 - 0.04 - 0.49 = 0.4$$

b) What is the mean number of papers a student taking 4 papers in a semester at The University of Otago passes?

1

- 0.98
- 1
- 3.29
- 1.96
- 3.0

Solution: For a discrete probability distribution, X, the formula for the mean of that distribution is

$$\sum_{i}^{k} x_i \times \Pr(X = x_i)$$

This is on the formula sheet. What this means is that we find the mean by adding up, for each possible outcome of the variable (0,1,2,3, or 4 in this case, since this is the number of papers a student taking 4 papers in a semester can pass), that outcome times the probability of observing that outcome.
In our case here this is

$$\sum_i^k x_i \times \Pr(X = x_i) =$$
$$0 \times \Pr(X = 0) + 1 \times \Pr(X = 1) + 2 \times \Pr(X = 2) + 3 \times \Pr(X = 3) + 4 \times \Pr(X = 4)$$
$$= (0 \times 0.02) + (1 \times 0.05) + (2 \times 0.04) + (3 \times 0.4) + (4 \times 0.49)$$
$$= 3.29$$

c) The standard deviation of the number of papers a student taking 4 papers in a semester at The Univeristy of Otago passes is closest to:

1

- 18.32
- 0.00
- 0.91
- 0.83
- 1.65

Solution: For a discrete probability distribution, X, the formula for the variance of that distribution is

$$\sum_i^k (x_i - \mu_X)^2 \times \Pr(X = x_i)$$

This is on the formula sheet. The mean (μ_X) is 3.29 from above. So

$$\sum_i^k (x_i - \mu_X)^2 \times \Pr(X = x_i)$$
$$= \Pr(X = 0) \times (0 - 3.29)^2 + \Pr(X = 1)(1 - 3.29)^2 + \Pr(X = 2)(2 - 3.29)^2$$
$$+ \Pr(X = 3) \times (3 - 3.29)^2 + \Pr(X = 4) \times (4 - 3.29)^2$$
$$= 0.02 \times (0 - 3.29)^2 + 0.05 \times (1 - 3.29)^2 + 0.04 \times (2 - 3.29)^2 + 0.4 \times (3 - 3.29)^2$$
$$+ 0.49 \times (4 - 3.29)^2$$
$$= 0.83$$

This gives the variance of the number of papers passed. To get the standard deviation, we take the square root of the variance (so the standard deviation is $\sqrt{0.83} = 0.91$)

Question 4

The number of children being clinically diagnosed with ADHD by psychologists and psychiatrists has been rising rapidly over the last 25 years.
ADHD can be divided into the "inattentive ADHD", "hyperactive ADHD", and "combined ADHD" subtypes (the combined type is where the criteria for being diagnosable with inattentive and hyperactive ADHD are both met.).
The table below shows the number of children at a general practice medical centre meeting the criteria for each subtype of ADHD.
In the options below, let I be the event that a person had inattentive ADHD, and H be the event that a person has hyperactive ADHD.

		Inattentive ADHD		
		Yes (I)	No ($\bar{I}$)	Total
Hyperactive ADHD	Yes (H)	48	34	82
	No ($\bar{H}$)	30	905	935
	Total	78	939	1017

a) What is the probability a child at the practice has inattentive ADHD? (i.e. What is $\Pr(I)$?)

1

- $82/1017$
- $48/78$
- $48/82$
- $78/1017$
- $48/1017$

Solution: Of the 1017 children, 78 have inattentive ADHD. So the probability is just $78/1017$.

b) What is the probability a child at the practice has both inattentive and hyperactive ADHD? (i.e. What is $\Pr(I \cap H)$?)

1

- $82/1017$
- $48/78$
- $48/82$
- $78/1017$
- $48/1017$

Solution: Of the 1017 children, 48 have inattentive and hyperactive ADHD. So the probability is just $48/1017$.

c) Given that a child has inattentive ADHD, what is the probability that they also have hyperactive ADHD? (i.e. What is $\Pr(H \mid I)$?)

1

- $48/78$
- $48/82$
- $34/78$

- $48/1017$
- $34/1017$

Solution: Simplest way: Out of the 78 children with inattentive ADHD, 48 also have the hyperactive type. So the probability is just $48/78$.

Another way: You could use the conditional probability formula

$$\Pr(H \mid I) = \frac{\Pr(H \cap I)}{\Pr(I)}$$

$\Pr(H \cap I) = 48/1017$ because this is the total with hyperactive and inattentive ADHD (48) divided by the total number of children (1017).
$\Pr(I) = 78/1017$ by identical reasoning.
So

$$\Pr(H \mid I) = \frac{\Pr(H \cap I)}{\Pr(I)} = \frac{48/1017}{78/1017} = \frac{48}{78}$$

d) Given that a child has hyperactive ADHD, what is the probability they do not have inattentive ADHD? (i.e. What is $\Pr(\bar{I} \mid H)$)

1

- $48/82$
- $48/78$
- $34/82$
- $34/1017$
- $34/78$

Solution: Simplest way: Out of the 82 children with hyperactive ADHD, 34 didn't have the inattentive type. So the probability is just $34/82$.

Another way: You could use the conditional probability formula

$$\Pr(\bar{I} \mid H) = \frac{\Pr(\bar{I} \cap H)}{\Pr(H)}$$

$\Pr(\bar{I} \cap H) = 34/1017$ because this is the total with hyperactive but not inattentive ADHD (34) divided by the total number of children (1017).
$\Pr(H) = 82/1017$ by identical reasoning.
So

$$\Pr(\bar{I} \mid H) = \frac{\Pr(\bar{I} \cap H)}{\Pr(H)} = \frac{34/1017}{82/1017} = \frac{34}{82}$$

Question 5

Avian bird flu is a disease that affects over 100 species of wild and domestic birds. The poultry farming industry reguarly monitors for outbreaks.

At one point in time on a poultry farm, 12% of chickens had the flu.
A test for detecting the flu tests positive for 91% of chickens with the flu, and tests positive for 4% of chickens without the flu.

Let T be the event "tests positive" and F be the event "has the avian flu".

a) What is the sensitivity of the test? (i.e. what is $\Pr(T \mid F)$?)

1

- 0.09
- 0.91
- 0.96
- 0.04
- 0.12

Solution: The sensitivity of the test is the probability it tests positive in birds with the flu (how sensitive the test is to picking up that a bird with the flu has the flu). This is given in the question brief as 91%, or $91/100 = 0.91$ when expressed as a probability.

b) What is the specificity of the test? (i.e. what is $\Pr(\bar{T} \mid \bar{F})$?)

1

- 0.09
- 0.91
- 0.96
- 0.04
- 0.12

Solution: The specificity of the test is the probability it tests negative in birds that don't have the flu. In the question brief it is given that the test tests positive for 4% of birds without the flu, so the probability it tests positive in a bird without the flu is $4/100 = 0.04$, and the probability it tests negative in a bird without the flu is $1 - 0.04 = 0.96$.

c) The probability that a chicken does not have the flu and has a negative test result (i.e. $\Pr(\bar{F} \cap \bar{T})$) is:

1

- 0.1152
- 0.8448
- 0.1092
- 0.8008
- 0.0048

Solution: Use the rule on the formula sheet $\Pr(A \cap B) = \Pr(A) \times \Pr(B \mid A)$. Let A be the event $\bar{F}$ and B be the event $\bar{T}$, and then we have $\Pr(\bar{F} \cap \bar{T}) = \Pr(\bar{F}) \times \Pr(\bar{T} \mid \bar{F})$. Then we can calculate $\Pr(\bar{F} \cap \bar{T})$ by substituting in the values for $\Pr(\bar{F})$ and $\Pr(\bar{T} \mid \bar{F})$:

$$\Pr(\bar{F} \cap \bar{T}) = \Pr(\bar{F}) \times \Pr(\bar{T} \mid \bar{F}) = (1 - 0.12) \times 0.96 = 0.8448$$

d) The probability that the test produces a negative result (i.e $\Pr(\bar{T})$) is closest to:

1

- (•) 0.8556
- ( ) 0.1444
- ( ) 0.8556
- ( ) 0.954
- ( ) 0.88

Solution: The easiest way to solve this type of problem is to draw a tree diagram, and to look at the branches of the tree where there is a negative test result. See the lectures on probabilities, tree diagrams, and the sensitivity and specificity of tests for how to do this.
From the tree diagram we see that the branches of the tree where a negative test result occurs are the $(F \cap \bar{T})$ branch and the $(\bar{F} \cap \bar{T})$ branch. So $\Pr(\bar{T})$ is the probability of both those branches occurring ($\Pr(\bar{T}) = \Pr(F \cap \bar{T}) + \Pr(\bar{F} \cap \bar{T})$). Each of these can be calculated using the same technique as in part (c), of using the rule $\Pr(A \cap B) = \Pr(A) \times \Pr(B \mid A)$:

$$\Pr(F \cap \bar{T}) = \Pr(F) \times \Pr(\bar{T}|F) = 0.12 \times (1 - 0.91) = 0.0108$$

$$\Pr(\bar{F} \cap \bar{T}) = \Pr(\bar{F}) \times \Pr(\bar{T} \mid \bar{F}) = (1 - 0.12) \times 0.96 = 0.8448$$

So

$$\Pr(\bar{T}) = \Pr(F \cap \bar{T}) + \Pr(\bar{F} \cap \bar{T}) = 0.0108 + 0.8448 = 0.8556$$

e) What is the negative predictive value of this test closest to? (i.e. what is $\Pr(\bar{F} \mid \bar{T})$ closest to?)

1

- ( ) 0.88
- ( ) 0.8556
- ( ) 0.0056
- ( ) 0.9360
- (•) 0.9874

Solution: For any events A and B, the conditional probability rule

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

applies. When letting $\bar{F}$ be the event A and $\bar{T}$ be the event B this, this becomes

$$\Pr(\bar{F} \mid \bar{T}) = \frac{\Pr(\bar{F} \cap \bar{T})}{\Pr(\bar{T})}$$

which we can use to calculate $\Pr(\bar{F} \mid \bar{T})$.

You will need to know the formula $\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$ for the exam for questions like these. Either simply remember this formula, or if you forget you can derive it by taking the formula

$$\Pr(A \cap B) = \Pr(B) \times \Pr(A \mid B)$$

that's on the formula sheet, and then dividing both sides of this formula by $\Pr(B)$ to get

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

From part (c), $\Pr(\bar{F} \cap \bar{T}) = 0.8448,$ and from part (d), $\Pr(\bar{T}) = 0.8556,$ so

$$\Pr(\bar{F} \mid \bar{T}) = \frac{\Pr(\bar{F} \cap \bar{T})}{\Pr(\bar{T})} = \frac{0.8448}{0.8556} = 0.9874$$

Question 6

A dog breeder has a pregnant golden retriever and a pregnant black labrador. She knows that the pups that will be born from each dog will be pure breds (i.e. the golden retriever will give birth to golden retriever pups, and the black labrador will give birth to black labrador pups).
Suppose she will sell all the golden retriever pups for $\$2000$ and all the black labrador pups for $\$1000$.

a) What type of random variables are the total number of pups that are born of each dog breed?

1

- Continuous.
- (•) Discrete.
- Categorical ordinal.
- Categorical nominal.
- They are not random variables.

Solution: The total number of pups that are born are whole numbers, so they are discrete random variables. They are random variables, because there is uncertainty in how many will be born before we observe how many are born.

b) The average number of golden retriever pups born in a litter is 6.8, and the average number of black labrador pups born in a litter is 7.3. What is mean of the distribution for the total amount the breeder collects in sales from the pups? Assume that all the pups are sold. Hint: You will want to express the total sales in the form $W = aX + bY + c$ where X and Y represent the total number of pups of each type born, then use the rule $\mu_W = a\mu_X + b\mu_Y + c$ for the mean of combined random variables (this rule is on your formula sheet).

1

- $\$22900$
- $\$21900$
- (•) $\$20900$
- $\$21400$
- $\$28200$

Solution: Letting W be the total sales from all pups, and X and Y respectively be the numbers of golden retrievers and black labradors that are born, the total sales

$$W = 2000X + 1000Y$$

since the golden retrievers sell for 2000 and the labs sell for 1000 (and we assumed in the question brief that all the pups were sold).
Therefore

$$\mu_W = a \times \mu_X + b \times \mu_Y = 2000 \times 6.8 + 1000 \times 7.3 = 20900$$

c) Suppose the standard deviations for the number of golden retriever and black labrador pups born are 0.9 and 1.1, respectively. The standard deviation of the total sales the breeder can expect to collect for all pups that are born, assuming all the pups are sold, is closest to:

1

- (•) $2109.50
- ( ) $4450000
- ( ) $2830
- ( ) $2900
- ( ) $53.20

Solution: Use the rule that's on the formula sheet that when $W = aX + bY + c$, that $\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$. From the previous question, the total sales $W = 2000X + 1000Y$ where X is the number of golden retrievers born and Y is the number of black labradors born, so in the format $W = aX + bY + c$, $a = 2000$, $b = 1000$, and $c = 0$. Using $\sigma_X = 0.9$, and $\sigma_Y = 1.1$ as given above,

$$\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 = 2000^2 \times 0.9^2 + 1000^2 \times 1.1^2 = 4450000$$

This gives the variance of the total sales σ_W^2. But the question wanted the standard deviation of the total sales, so the standard deviation of the total sales is

$$\sqrt{\sigma_W^2} = \sqrt{4450000} = 2109.50$$

Question 7

Hippos are some of the largest animals in the world. Suppose it is known that the weights of fully grown male hippos are normally distributed with a mean of 3530kg and a standard deviation of 190kg.

a) Suppose you were asked to calculate the probability that a randomly selected male hippo weighs more than 3930kg. Which of the following statements could you make without needing any software?

1

- ○ Because the standard error depends on the sample size, we couldn't say anything about the probability without knowing the sample size used to estimate the mean and standard deviation.
- ○ We cannot say anything about the probability without the use of software.
- ○ The probability will be zero because a hippo cannot weigh 3930kg if hippo weights are distributed as given above.
- ○ The probability will be greater than 0.5.
- ◉ The probability will be less than 0.5.

Solution: For a normally distributed variable, the probability of observing an observation below (and above) the mean is exactly 0.5. So the probability of observing a weight greater than any value that's more than the mean (such as 3930kg if the mean hippo weight is 3530kg) will be less than 0.5.

b) Select the appropriate R command to calculate the probability a randomly selected male hippo weighs more than 3930kg.

1

- ○ **pnorm(q=3930,mean=3530,sd=190)**
- ○ **pnorm(q=3930,mean=3530,sd=190,lower.tail=TRUE)**
- ◉ **1-pnorm(q=3930,mean=3530,sd=190)**
- ○ **pnorm(q=3930,mean=3530,sd= $190/\sqrt{n}$)**
- ○ **1-pnorm(q=3930,mean=3530,sd= $190/\sqrt{n}$)**

Solution: **pnorm(q=3930,mean=3530,sd=190)** finds the probability that a normally distributed variable with mean 3530 and standard deviation 190 (i.e. a variable with the same distribution as the male hippo weights) takes a value *less than* 3930 (with lower.tail not entered, R defaults to looking in the lower tail (i.e. it finds the "less than" probability)). So **pnorm(q=3930,mean=3530,sd=190)** gives the probability a randomly selected male hippo weighs *less than* 3930kg, and **1-pnorm(q=3930,mean=3530,sd=190)** gives the probability a randomly selected male hippo weighs *more than* 3930kg.

c) What is the Z-score for a male hippo that weighs 3130kg ?

1

- ◉ -2.11
- ○ 2.11
- ○ 18.58
- ○ -18.58
- ○ 0.48

Solution: The Z-value or Z-score for an observation or value tells us how many standard deviations the observation or value is above or below the mean. The formula for calculating the

Z-score for a value is

$$\frac{\text{Value} - \text{Mean}}{\text{Standard Deviation}}$$

In this example

$$\frac{\text{Value} - \text{Mean}}{\text{Standard Deviation}} = \frac{3130 - 3530}{190} = \frac{-400}{190} = -2.11$$

Question 8

The diet of pandas consists almost entirely of bamboo.
A study of 50 pandas from around the world attempted to measure the average daily bamboo consumption of pandas. In the sample, presume that they found the mean daily bamboo consumption for each panda exactly. Across the 50 pandas in the sample, the mean of the pandas' mean daily bamboo consumptions was 25kg, while the standard deviation of the 50 pandas' mean daily bamboo consumptions was 6.5kg.
Assume that the distribution for the mean daily bamboo consumption across all individual pandas is normally distributed.

The researchers constructed the below two intervals from their sample data using $\alpha = 0.05$.

Interval A : $25 \pm t_{\left(1-\frac{\alpha}{2},49\right)} \times \dfrac{6.5}{\sqrt{50}}$

Interval B : $25 \pm z_{\left(1-\frac{\alpha}{2}\right)} \times 6.5$

a) Which of the following does Interval A represent?

1

- A 95% confidence interval. We estimate the mean daily bamboo consumption for 95% of all pandas is in this range.
- A 95% confidence interval. We can be 95% confident the mean daily bamboo consumption across all pandas is in this range.
- A 95% reference range. We estimate the mean average daily bamboo consumption for 95% of all pandas is in this range.
- A 95% reference range. We can be 95% confident the mean daily bamboo consumption across all pandas is in this range.
- The interval has no discernable meaning.

Solution: A 95% confidence interval for a population mean is a range of values we can be 95% confident contains the population mean. So a 95% CI for the mean daily bamboo consumption of all pandas is a range we can be 95% confident contains the mean daily bamboo consumption of all pandas.
From the formula sheet, if we were to construct a $100(1-\alpha)\%$ confidence interval for the population mean daily bamboo consumption across all pandas, presuming we had a sample of size larger than 20 or our underlying population was normal (in this case, both of these assumptions are met), we would construct a $100(1-\alpha)\%$ CI using $\text{Estimate} \pm \text{Multiplier} \times \text{Standard Error}$ where the estimate is the sample mean bamboo consumption (25, in this case), the multiplier is $t_{(1-\alpha/2,n-1)}$ ($t_{(1-\alpha/2,49)}$, in this case. See note at the bottom of this solution for what $t_{(1-\alpha/2,n-1)}$ represents.), and the standard error is $s/\sqrt{n}$ ($6.5/\sqrt{50}$, in this case). So

$$\text{Estimate} \pm \text{Multiplier} \times \text{Standard Error} = 25 \pm t_{(1-\alpha/2,49)} \frac{6.5}{\sqrt{n}}$$

is a $100(1-\alpha)\%$ confidence interval for the mean daily bamboo consumption across all pandas. With $\alpha = 0.05$, this is a $100(1-0.05)\% = 95\%$ CI.

A 95% CI for a population mean is an interval we can be 95% confident contains the population mean. So we can be 95% confident this range contains the mean daily bamboo consumption across all pandas. It is NOT a range which the mean daily consumption for 95% of

individual pandas lies between. This would be a 95% reference range.

Comment on what $t_{1-\alpha/2,n-1}$ refers to: $t_{p,n-1}$ refers to the t-value, or the value on the horizontal axis of the t-distribution curve with $n-1$ degrees of freedom, for which the area under the t-distribution curve spanning over all horizontal-axis values below that value is p. It is found in R with the command $\mathrm{qt}(1-\alpha/2, n-1)$.

b) Which of the following does Interval B represent?

1

- A 95% confidence interval. We estimate the mean daily bamboo consumption for 95% of all pandas is in this range.
- A 95% confidence interval. We can be 95% confident the mean daily bamboo consumption across all pandas is in this range.
- A 95% reference range. We estimate the mean daily bamboo consumption for 95% of all pandas is in this range.
- A 95% reference range. We can be 95% confident the mean daily bamboo consumption across all pandas is in this range.
- The interval has no discernable meaning.

Solution: A 95% reference range is a range of values which 95% of the individuals in a population sits between. So a 95% reference range for the mean daily bamboo consumption of individual pandas is a range which the mean daily bamboo consumption for 95% of individual pandas sits between.
We are told in the question to assume the distribution of individual pandas' mean daily bamboo consumptions is normally distributed. Our best estimate for the mean and standard deviation of this distribution is the sample mean (25) and the sample standard deviation (6.5), respectively. From the lecture notes, 95% of individuals from a normally distributed population are within $z_{0.975} \approx 1.96$ standard deviations of its mean (Note: see the comment at the bottom of this solution for what $z_{0.975}$ represents). Therefore, we estimate that the mean daily bamboo consumption for 95% of individual pandas is in the range

$$25 \pm z_{(0.975)} \times 6.5 = 25 \pm z_{\left(1-\frac{\alpha}{2}\right)} \times 6.5$$

with $\alpha = 0.05,$ and the interval above is an estimate for a 95% reference range for the mean daily bamboo consumption for individual pandas. (Note: It is an estimate for the reference range rather than an exact reference range since we are estimating the true population mean and standard deviation using the sample mean and sample standard deviation.)

Comment on what $z_{0.975}$ refers to: z_p refers to the z-value, or the value on the horizontal axis of the Z (standard normal) distribution, for which the area under the standard normal distribution curve spanning all horizontal-axis values below that value is p (i.e. it is the horizontal-axis value on the Z-distribution graph for which the probability that the Z-distributed variable takes a value less than that value is p). This can be found using the R command $\mathrm{qnorm}(0.975)$.

Question 9

A district with several mountains is renowned for being a high-risk zone for avalanches. A geologist is trying to predict the number of mountains in the district that will have an avalanche in a given year.
She decides to use a binomial model to estimate the number of mountains in the district that will have one or more avalanches in a given year.

a) Letting X be the number of mountains in the district that have an avalanche in a given year, select the conditions that must apply for X to be a binomial random variable.

1

- ○ Fixed number of mountains in the district; the probability each mountain has an avalanche differs across all mountains; the probability that each mountain has an avalanche depends on whether some of the other mountains have had an avalanche; exactly two possible outcomes for each mountain.
- ◉ Fixed number of mountains in the district; the probability each mountain has an avalanche is the same across all mountains; the probability that each mountain has an avalanche does not depend on whether any of the other mountains have had an avalanche; exactly two possible outcomes for each mountain.
- ○ Two mountains in the district; the probability each mountain has an avalanche is between 0 and 1; the probability that each mountain has an avalanche does not depend on whether any of the other mountains have had an avalanche; multiple possible outcomes for each mountain.
- ○ Two mountains in the district; the probability each mountain has an avalanche is the same across all mountains; the probability that each mountain has an avalanche depends on whether some of the other mountains have had an avalanche; multiple possible outcomes for each mountain.
- ○ Fixed number of mountains in the district; the probability each mountain has an avalanche is between 0 and 1; the probability that each mountain has an avalanche does not depend on whether any of the other mountains have had an avalanche; One possible outcome for each mountain.

Solution: The four conditions that must be met for a variable to be binomial, using the generic terminology introduced in lectures for binomial variables, are:
- Fixed number of trials.
- Constant probability of "success" in each trial.
- Two possible outcomes for each trial.
- Each trial independent of all the other trials.

In this example,
- The "trials" are whether or not each individual mountain has an avalanche.
- "Success" in each trial is when the mountain has an avalanche.
- The two "outcomes" for each mountain are "has an avalanche" and "does not have an avalanche"

Translating the meaning of the generic binomial terms (e.g. "trial", "success", "outcome") for our specific example into the generic 4 conditions that must apply for a variable to be binomial, we get the 4 conditions that must apply for the number of avalanches on the mountains in the district to be binomial.

Suppose the district has 5 mountains, and the geologist assumes the probability each mountain has an avalanche is $\pi = 0.1$. The table below displays the probability distribution for the number of mountains in the district that have an avalanche with $n = 6$ and $\pi = 0.1$:

x_i	0	1	2	3	4	5
$\Pr(X = x_i)$	0.59049	0.32805	0.0729	0.0081	0.00045	0.00001

b) The probability that exactly 3 of the mountains have an avalanche is:

1

- 0.32805
- 0.0729
- 0.0081
- 0.00045
- 0.59049

Solution: This is can be read directly from the table.

c) The probability that more than 3 of the mountains have an avalance is:

1

- 0.00856
- 0.00001
- 0.0081
- 0.00046
- 0.99144

Solution: If more than 3 of the 5 mountains have an avalanche then either 4 or 5 of the mountains have an avalanche. So the probability that more than 3 of the mountains have an avalance is the probability that 4 or 5 of the mountains have an avalanche. So

$$\Pr(X > 3) = \Pr(X = 4) + \Pr(X = 5) = 0.00045 + 0.00001 = 0.00046$$

Question 10

A poll by YouGov, and international research and data analytics organisation, sought to explore levels of job satisfaction across workers in the UK. Within the poll, one of the questions asked was whether the respondents believed their job made no meaningful contribution to the world.

a) If 108 UK workers were polled, without knowing the proportion that believe their job has no meaningful impact, select the correct statement from the options below regarding the margin of error of the 95% confidence interval for the proportion of UK workers that believe their job has no meaningful contribution to the world. Use $\text{qnorm}(0.975) = 1.96$.

1

- At least 0.094 to 3 dp.
- Exactly 0.094 to 3 dp.
- **No more than 0.094 to 3 dp.**
- At least 0.189 to 3 dp.
- No more than 0.188 to 3 dp.

Solution: The margin of error for a confidence interval is $\text{Multiplier} \times \text{Standard Error}$. This is how far our interval spans each side of our estimate.
The multiplier and standard error for the CI for a single proportion is on the formula sheet as $\text{Multiplier} = \text{qnorm}(1-\alpha/2) = \text{qnorm}(0.975) = 1.96$ for a 95% CI, and

$$\text{Standard Error} = \sqrt{\frac{p(1-p)}{n}}$$

where p is the sample proportion.
As a result of this, when we don't know the sample proportion p, we won't know the exact standard error. However, as demonstrated in the lectures, this standard error will be largest when $p = 0.5$, for a fixed sample size n. So we know that

$$\text{Standard Error} = \sqrt{\frac{p(1-p)}{n}} \leq \sqrt{\frac{0.5 \times (1-0.5)}{108}} = 0.048$$

The margin of error for the 95% CI will therefore be no more than

$$\text{Multiplier} \times \text{Standard Error} = 1.96 \times 0.048 = 0.094$$

b) Suppose we wanted the margin of error to be no more than 0.067. What is the minimum sample size required to ensure the margin of error will be no more than this? Use **qnorm(0.975)=1.96**.

1

- 314
- 213.95
- **214**
- 264.08
- 264

Solution: The margin of error for a 95% CI is $\text{Multiplier} \times \text{Standard Error}$. If this needs to be less than 0.067 then

$$\text{Multiplier} \times \text{Standard Error} <= 0.067$$
$$\text{qnorm}(1-\alpha/2) \times \sqrt{\frac{p(1-p)}{n}} <= 0.067$$

For a 95% CI, $\alpha = 1 - 0.95 = 0.05$, so
$\text{qnorm}(1-\alpha/2) = \text{qnorm}(1-0.05/2) = \text{qnorm}(0.975)$. The largest possible value of $\sqrt{p(1-p)/n}$ for a fixed n is when $p = 0.5$, so if
$\text{qnorm}(1-\alpha/2)\sqrt{p(1-p)/n} <= 0.067$ then

$$\begin{aligned}
\text{qnorm}(0.975) \times \sqrt{\frac{0.5(1-0.5)}{n}} &<= 0.067 \\
1.96 \times \sqrt{\frac{0.25}{n}} &<= 0.067 \\
\sqrt{\frac{0.25}{n}} &<= \frac{0.067}{1.96} \\
\frac{0.25}{n} &<= \left(\frac{0.067}{1.96}\right)^2 \\
\frac{0.25}{\left(\frac{0.067}{1.96}\right)^2} &<= n \\
213.945 &<= n
\end{aligned}$$

Since the sample size must be a whole number, $n >= 213.945$ means that n needs to be at least 214.

c) Suppose that the poll actually surveyed 62 UK workers, and 23 of them reported that they believed their job made no meaningful contribution to the world. The sample proportion that believed their job made no meaningful contrbution to the world is closest to:

1

- 62 people
- 23 people
- 0.629
- 0.371
- 2.696

Solution: The sample proportion is the total with that outcome divided by the total number surveyed. So it is $23/62 = 0.371$

d) Select the 95% confidence interval for the proportion of UK workers that believe their job makes no meaningful contribution to the world from the options below. The answers are in the format $(\text{Lower Bound}, \text{Upper Bound})$.

1

- (0.251, 0.491)
- (0.310, 0.432)
- (−1.589, 2.331)
- (0.364, 0.378)
- (0.251, 2.331)

Solution: The 95% confidence interval is constructed using

$$\text{Estimate} \pm \text{Multiplier} \times \text{Estimated Standard Error}$$

The estimate is our sample proportion, $p = 23/62 = 0.370968$.
The multiplier is $\text{qnorm}(1 - \alpha/2) = \text{qnorm}(0.975) = 1.96$.
The standard error is

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.370968(1-0.370968)}{62}} = 0.061349$$

So the 95% confidence interval for the true proportion is

$$0.370968 \pm 1.96 \times 0.061349 = (0.251, 0.491)$$

Note: More decimal places were used for the estimate and the standard error in this calculation to get the answer more accurate. This is advised to reduce the risk of error due to rounding.

e) Which of the below approaches is likely to help the researchers minimise bias in the survey the most?

1

- Using a larger sample size to allow for non-response.
- Selecting the participants at random from a workers' union database, and then arranging a time that suits the participant to survey them.
- Selecting the participants at random from a workers' union database, and then calling them at 5:30pm when they're more likely to have time to talk.
- Selecting the participants at random from the tax department database, and then arranging a time that suits the participant to survey them.
- Selecting the participants at random from the tax department database, and then calling them at 5:30pm when they're more likely to have time to talk.

Solution: Using a larger sample size does not minimise bias from non-response (or any type of bias). Increasing the sample size only reduces the uncertainty in our estimates purely due to random variation across the population.
The options involving contacting participants through the tax department database will be better for minimising bias than contacting participants through a workers' union, since workers in unions are likely to systematically differ from non-union workers.
Arranging a time that suits participants to talk will also reduce bias more than calling at a time that more participants are likely to be available, since people who are available at a particular time to talk (such as 5:30pm) may systematically differ from people who aren't available at that time.

Question 11

The direct and indirect consequences of insomnia are estimated to cost the United States $100 billion annually. A study was conducted to explore whether exercise could help to improve sleep.

In the study, 50 middle-aged men who currently did not exercise were randomly selected from across the population. The men were randomly allocated to a control or intervention group. The control group did not change their exercise patterns (so they continued to not exercise), while the intervention group undertook 30 minutes per day of moderate-intensity exercise.
One metric for how well a person sleeps is their sleep latency, or how long it takes them to fall asleep at the start of the night. For each person in the sample, they measured their sleep latency on one night before the start of the study, and one night after 8 weeks following the study protocols.
The table below summarises the change in sleep latencies between the two readings for both the intervention and control groups. The change in sleep latency is calculated **using the order of differencing "sleep latency after study" minus "sleep latency before study"**.

	Group	
	Control	Intervention
Participants (n_i)	25	25
Mean change in minutes ($\bar{x}_i$)	-0.1	-5.5
Standard Deviation in minutes (s_i)	0.02	0.79

a) What type of study did the researchers use?

1

- ○ Descriptive, observational, cohort study.
- ○ Analytic, observational, cohort study.
- ○ Analytic, observational, randomised control trial.
- ◉ Analytic, experimental, randomised control trial.
- ○ Descriptive, experimental, randomised control trial.

Solution: The study is an analytic and not a descriptive study, because it is exploring the effect one variable (exercise) has on another (sleep), rather than simply exploring a single variable.
It is experimental rather than observational, because the researchers are intervening in the study participants' behaviours, rather than simply observing their natural behaviour.
It is a randomised control trial, because participants are randomly allocated to the study groups (control group and intervention group).

b) Select the correct statement below regarding the design of the study.

1

- ○ Random sampling creates two comparable groups.
- ○ Random sampling eliminates confounding in the study.
- ◉ Randomisation reduces the risk of confounding in the study.
- ○ Random sampling eliminates random error in the study.
- ○ Randomisation eliminates random error in the study.

Solution: Random sampling refers to the process of bringing people into the study, whereby participants are randomly selected from some sampling frame to join the study.
Randomisation differs from random sampling in that it refers to the process of randomly allocating people to one of the study sub-groups (control group, or intervention group, in this randomised control study) once they're already in the study.

Randomisation seeks to create two groups who are as similar as possible in all ways, with the exception of the exposure variable being tested (e.g. in this study, it seeks to create two groups who are as similar as possible, except that one group exercises for 30-minutes a day, while the other group doesn't). As a result, this reduces confounding, because it seeks to eliminate all other differences between the study sub-groups besides the exposure variable being tested. When confounding exists, the confounding variable will be associated with the exposure variable, so the confounding variable will differ between the study sub-groups given that the study sub-groups will differ with respect to the exposure variable being tested. Because of this, the confounding variable can distort the results between the groups from the true effect that exists purely due to differences with respect to the exposure variable.

The other options are incorrect.
Random sampling creates two comparable groups: This is incorrect, because random sampling just refers to the process of bringing people into the study, and not of creating comparable sub-groups once the study participants are selected. For example, if we had used random sampling to bring people into the study but participants had chosen whether to join the exercising or non-exercising group themselves, those that chose to exercise may differ from those that didn't in other ways besides exercise that affect sleep (e.g. the exercisers diet may be better which may improve sleep by more than the effect of the exercise alone). So the groups may not be comparable in ways other than the variable being examined in the study (exercise).
Random sampling eliminates confounding in the study: Random sampling doesn't eliminate confounding. For example, if random sampling is used to bring people into the study but we didn't intervene in the participants' behaviours (e.g. tell one group to exercise, and the other group to not), the natural associations that would exist between people who chose to change their habits and exercise, and people who didn't change their habits, would still exist, doing nothing to reduce confounding. (For example, people who chose to begin exercising may be more motivated to improve their health in general, so may also choose to eat better. So if diet influences sleep, then diet could confound the relationship between exercise and sleep).
Random sampling and randomisation eliminates random error in the study: Random error refers to how random variation across individuals leads to study results differing from the truth across the whole population. For example, in this study, there will be random variation across individuals in relation to the effect that exercise has in improving sleep (exercise won't improve or harm sleep by the same amount for each individual), meaning the mean effect of exercise on sleep in the study will differ from the true mean effect across the whole population. Neither random sampling or randomisation can eliminate this, because random differences will always exist between individuals. (Note: For the same reason, randomisation won't eliminate confounding. It'll just reduce it).

c) Select the appropriate R code to calculate the multiplier for the 99% confidence interval for the difference between the mean sleep latency change for exercising men and the mean sleep latency change for non-exercising men. Assume ν is the appropriate degrees of freedom.

1

- **1-qt(0.975, ν)**
- **qt(0.975, ν)**
- **qt(0.995, ν)**
- **qt(0.995, $n-1$)**
- **1-qt(0.975, $n-1$)**

Solution: From the formula sheet, the multiplier for a $100(1-\alpha)\%$ CI is $\text{qt}(1-\alpha/2, \nu)$, where ν will be given (the formula for ν was given in lectures, and is too complex to be expected to be calculated by hand).
For a $100(1-\alpha)\% = 99\%$ CI, $(1-\alpha) = 99/100$, so $\alpha = 0.01$. Therefore, the multiplier is $\text{qt}(1-\alpha/2, \nu) = \text{qt}(1-0.01/2, \nu) = \text{qt}(995, \nu)$.
Note that ν is not $n-1$ for CI for a difference between 2 means; $n-1$ is the degrees of

freedom for a CI for a single mean and a paired difference between means.

d) Select the appropriate calculation of the 99% confidence interval for the difference between the mean sleep latency change for exercising men and the mean sleep latency change for non-exercising men.

1

- $\big(-5.5-(-0.1)\big) \pm \text{Multiplier} \times \sqrt{\frac{-5.5^2}{25}+\frac{-0.1^2}{25}}$
- $\big(-0.1-(-5.5)\big) \pm \text{Multiplier} \times \sqrt{\frac{0.79^2}{25}+\frac{0.02^2}{25}}$
- $\big(-5.5-(-0.1)\big) \pm \text{Multiplier} \times \sqrt{\frac{0.79^2}{25}+\frac{0.02^2}{25}}$
- $\big(-0.1-(-5.5)\big) \pm \text{Multiplier} \times \big(\frac{0.79^2}{25}+\frac{0.02^2}{25}\big)$
- $\big(-5.5-(-0.1)\big) \pm \text{Multiplier} \times \big(\sqrt{\frac{0.79^2}{25}}+\sqrt{\frac{0.02^2}{25}}\big)$

Solution: Our confidence intervals are calculated using the format

$$\text{Estimate} \pm \text{Multiplier} \times \text{Estimated Standard Error}$$

Estimate: The CI in this question is for the difference between the mean change in sleep latency for exercising men and the mean change in sleep latency for non-exercising men. So the estimate is the difference between the mean change in sleep latencies observed in the sample $\bar{x}_I - \bar{x}_C$ (where $\bar{x}_I$ and $\bar{x}_C$ are the sample means in the intervention (exercising) and control (non-exercising) groups). Note that the question necessitates that we take the order of differencing $\bar{x}_I - \bar{x}_C$ rather than $\bar{x}_C - \bar{x}_I$, because it asks for the difference between the mean change in sleep latencies for exercising men and the mean change in sleep latencies for non-exercising men. If the question had asked for the CI for the difference between the mean change in sleep latency between non-exercising and exercising men, our estimate would be the other way around ($\bar{x}_C - \bar{x}_I$, rather than $\bar{x}_I - \bar{x}_C$).

Standard Error: The standard error is (from the formula sheet)

$$\text{Estimated Standard Error} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

where s_1 and s_2 are the sample standard deviations for the two groups (in this example, the intervention and the control group), and n_1 and n_2 are the sample sizes for the two groups. If we let s_1 and n_1 refer to the intervention group, and s_2 and n_2 refer to the control group (note: it would have been fine to order the groups the other way around, so let s_1 and n_1 refer to the control group and s_2 and n_2 refer to the intervention group), we get

$$\text{Estimated Standard Error} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{0.79^2}{25}+\frac{0.02^2}{25}}$$

e) If the 99% confidence interval for the difference between the mean sleep latency change for exercising men and the mean sleep latency change for non-exercising men is $(-5.81, -4.99)$, select the appropriate conclusion.

1

- We are 99% confident that the difference between the mean sleep latency change for exercising men and the mean sleep latency change for non-exercising men is in the range $(-5.81, -4.99)$. Since the interval is fully below zero, we have evidence that the sleep latency for exercising men reduces by more than the sleep latency for non-exercising men.
- We are 99% confident that the difference between the mean sleep latency change for exercising men and the mean sleep latency change for non-exercising men is in the range $(-5.81, -4.99)$. Since the interval is fully below zero, we have evidence that the sleep latency for non-exercising men reduces by more than the sleep latency for exercising men.
- We are 99% confident that the difference between the mean sleep latency change for exercising men and the mean sleep latency change for non-exercising men is in the range $(-5.81, -4.99)$. Since the interval contains zero, we have no evidence that the mean sleep latency change differs between exercising and non-exercising men.
- We are 99% confident that the difference between the mean sleep latency change for exercising men and the mean sleep latency change for non-exercising men is in the range $(-5.81, -4.99)$. Since the interval is largely below zero, there is evidence that the mean sleep latency for exercising men reduces by more than the sleep latency for non-exercising men.
- We are 99% confident that the difference between the mean sleep latency change for exercising men and the mean sleep latency change for non-exercising men is in the range $(-5.81, -4.99)$. Since the interval is largely below zero, there is evidence that the mean sleep latency for non-exercising men reduces by more than the sleep latency for exercising men.

Solution: Since the interval $(-5.81, -4.99)$ is a 99% CI for the mean sleep latency change for exercising men and the mean sleep latency change for non-exercising men, we can be 99% confident that the difference between the mean change in sleep latency between exercising and non-exercising men is between -5.81 and -4.99 minutes. The interval is a plausible range for the difference between the mean change in sleep latencies.

If the interval is fully negative, the plausible range of values for the difference between the mean change are all negative. Since the interval is for the difference between the mean sleep latency change in exercising men and the mean sleep latency change in non-exercising men, the interval is for "the mean sleep latency change in exercising men" minus "the mean sleep latency change in non-exercising men". This being negative means a lower value for the sleep latency change in the exercising men. A lower value for the sleep latency change in the exercising group means a greater reduction in sleep latency. (To be clearer on this, it stated in the question brief that the values observed for the sleep latency change are calculated using "sleep latency after" minus "sleep latency before", meaning negative values imply a reduction in sleep latency between before and after the study.) The interval being fully negative therefore means we have evidence that the sleep latency reduces by more in exercising men than non-exercising men.

If the interval contains zero, the value indicating no difference between the mean change in sleep latencies between exercising and non-exercising men (i.e. zero) is in the interval, so we have no evidence (at the $1 - 0.99 = 0.01$ significance level) that the mean sleep latency reduces by a different amount in exercising and non-exercising men.

Question 12

Loneliness is a major mental health issue among elderly people.
Loneliness causes stress, among other things. Blood cortisol levels measured first thing in the morning are one of the most reliable methods of measuring a person's stress. Some researchers did a study to examine whether the companionship provided by a pet could help reduce stress (gauged by their morning blood cortisol levels) among elderly people.
They found a group of 8 elderly people who previously did not have pets and suffered from loneliness, and gave them a pet of their choosing. The researchers measured their morning blood cortisol levels on one morning before they got the pet, and on one morning three months after they got the pet.
The cortisol levels (expressed in mcg/dL) before and after getting the pet for the 8 participants are displayed in the table below, along with the difference between the before and after readings.

Person	Before	After	Difference
1	23	20.2	2.8
2	19.4	19.5	-0.1
3	27.9	19.4	8.5
4	37.2	28.8	8.4
5	12.2	13	-0.8
6	21.2	17.6	3.6
7	16.8	11.9	4.9
8	18.4	18.3	0.1

You may assume that the differences are Normally distributed.

a) The estimate for the mean difference between elderly people's cortisol levels before and after 3 months with a pet is closest to:

1

- 27.31
- 160.00
- 3.91
- 3.43
- 18.59

Solution: The estimate for the mean of the paired difference is just the mean of the paired differences in the sample.
We get the mean of the differences by adding all the differences and dividing by the number of observations (8):

$$\mu_d = \frac{(2.8 + -0.1 + 8.5 + 8.4 + -0.8 + 3.6 + 4.9 + 0.1)}{8} = 3.43$$

b) Based off this study, given that the standard deviation of the paired differences in the sample is 3.67, the estimated standard error for the mean difference between elderly people's cortisol levels before and after 3 months with a pet is closest to:

1

- 1.30
- 1.39
- 0.46

- 3.26
- 5.49

Solution: From the formula sheet, the standard error for a mean of a paired difference is $s_d/\sqrt{n}$, where s_d is the standard deviation of the paired differences in the sample. So the standard error is

$$\frac{s_d}{\sqrt{8}} = \frac{3.67}{\sqrt{8}} = 1.30$$

c) The 95% confidence interval for the mean difference between elderly people's cortisol levels before getting a pet and 3 months after getting a pet is $(0.35, 6.50)$. Select the correct interpretation of this confidence interval.

1

- We can be 95% confident that the mean difference between elderly people's cortisol levels before and 3 months after getting a pet is between 0.35 and 6.50 mcg/dL. As the interval contains zero, there is no evidence that elderly people's cortisol levels change, on average, after getting a pet.
- We can be 95% confident that the mean difference between elderly people's cortisol levels before and 3 months after getting a pet is between 0.35 and 6.50 mcg/dL. As the interval is entirely above zero, there is evidence that elderly people's cortisol levels $reduce$, on average, after getting a pet.
- We can be 95% confident that the mean difference between elderly people's cortisol levels before and 3 months after getting a pet is between 0.35 and 6.50 mcg/dL. As the interval is entirely above zero, there is evidence that elderly people's cortisol levels $increase$, on average, after getting a pet.
- We can be 95% confident that the difference between 95% of the elderly people's cortisol levels before and 3 months after getting a pet is between 0.35 and 6.50 mcg/dL. As the interval is entirely above zero, there is evidence that elderly people's cortisol levels $reduce$, on average, after getting a pet.
- We can be 95% confident that the difference between 95% of the elderly people's cortisol levels before and 3 months after getting a pet is between 0.35 and 6.50 mcg/dL. As the interval contains zero, there is no evidence that elderly people's cortisol levels change, on average, after getting a pet.

Solution: Since the interval $(0.35, 6.50)$ is a 95% CI for the mean difference between elderly people's cortisol levels before getting a pet and 3 months after getting a pet, we can be 95% confident that the mean difference is between 0.35 and 6.50 mcg/dL. The interval is a plausible range for the mean difference between elderly people's cortisol levels before and after getting a pet. (Note: The interval is therefore not a range for the cortisol difference for 95% of the elderly individuals (This would be a reference range for the difference.). So the options mentioning this are incorrect.)
If zero is in the interval, the value indicating no mean difference between the before and after cortisol levels (zero) is in the plausible range of values. We therefore have no evidence (at the $\alpha = 100 - 95 = 5\%$ significance level) that elderly people's cortisol levels change, on average, after getting a pet.
If zero is outside the interval, since the interval is for the mean difference between cortisol levels before and after getting the pet, the interval takes the form "cortisol levels before" minus "cortisol levels after" (rather than "cortisol levels after" minus "cortisol levels before"). Positive values for this indicate that the cortisol levels are higher before the pet than they are after the pet. If the interval is entirely above zero, the plausible range of values for the mean difference indicate the cortisol levels are higher (on average) before than they are after, so we have evidence that the cortisol levels $decrease$, on average, after getting a pet.

On the other hand, if the interval is entirely below zero, the plausible range of values for the mean difference indicate the cortisol levels are lower (on average) before than they are after, so we have evidence that the cortisol levels *increase*, on average, after getting a pet.

Question 13

A group of psychology reseachers are interested in whether people who attended single sex schools have differing rates of social anxiety to those that didn't (those that attended a `co-ed' school).
They surveyed a group of 20-year-olds who had left school, and found the following numbers were clinically diagnosable with social anxiety:

		Had Social Anxiety		
		Yes	No	Total
School	Single Sex	22	186	208
	Co-Ed	16	202	218
	Total	38	388	426

a) Based off this study, the estimated risk of a person who attended a single-sex school having social anxiety is closest to:

1

- 0.1183
- 0.1058
- 0.0792
- 0.0734
- 1.44

Solution: The risk in a group is the number with the outcome (social anxiety) divided by the total number in that group. For the single sex group this is $22/208 = 0.1058$

b) Based off this study, the estimated risk of a person who attended a co-ed school having social anxiety is closest to:

1

- 0.1183
- 0.1058
- 0.0792
- 0.0734
- 1.44

Solution: The risk in a group is the number with the outcome (social anxiety) divided by the total number in that group. For the co-ed group this is $16/218 = 0.0734$

c) Based off this study, the estimated relative risk of people who attended single sex schools having social anxiety compared to people who attended co-ed schools is closest to:

1

- 1.44
- 0.69
- 0.1058
- 0.0734
- 1.49

Solution: The RR of social anxiety in the single sex group relative to the co-ed group is the

$$\frac{\text{Risk of social anxiety in single sex group}}{\text{Risk of social anxiety in co-ed group}} = \frac{0.1058}{0.0734} = 1.44$$

Be sure to get the order right here. The question is asking for the relative risk of people who attended single sex schools having social anxiety compared to people who attended co-ed schools, so it should be the risk among those who attended single sex schools divided by the risk among those who attended co-ed schools (and not the other way around).

d) The interpretation of the relative risk above is:

1

- Among the people in the study, the risk of people who attended single-sex schools having social anxiety is greater than 0.
- Among the people in the study, the risk of having social anxiety is greater among people who attended single-sex schools than it is among people who attended co-ed schools.
- The risk of having social anxiety is greater among people who attended co-ed schools than it is among people who attended single-sex schools.
- Among the people in the study, the risk of having social anxiety is the same among people who attended single-sex schools as it is among people who attended co-ed schools.
- Across the whole population, the risk of having social anxiety is greater among people who attended single-sex schools than it is among people who attended co-ed schools.

Solution: The solution depends on whether your relative risk was greater or less than 1:
If the RR is greater than 1 then the risk is greater among people who attended single-sex schools in the study, because it is the relative risk of people who attended single sex schools having social anxiety compared to people who attended co-ed schools that was calculated. So the RR is

$$\frac{\text{Risk of social anxiety in single sex group}}{\text{Risk of social anxiety in co-ed group}}$$

so the risk is greater among single-sex students if this ratio is greater than 1.
For the same reason, if the RR is less than 1, the risk is greater among students who attended co-ed schools.
Note the RR is just the RR among people in the study. The RR could be different across the whole population. The confidence interval in the question below is a range we are confident the true RR across people across the whole population sits between.

e) Based off this study, the estimated standard error for the log of the risk ratio is closest to:

1

- 0.3139
- 0.3439
- 0.3426
- 0.3125
- 0.0986

Solution: You can use the formula given on the formula sheet of

$$\text{SE} = \sqrt{\frac{1}{a} - \frac{1}{r_1} + \frac{1}{b} - \frac{1}{r_2}} = \sqrt{\frac{1}{22} - \frac{1}{208} + \frac{1}{16} - \frac{1}{218}}$$

Note that the a, b, r_1 and r_2 as given in the standard error formula on the formula sheet are given in the contingency table on the first page of the formula sheet.

f) The 95% confidence interval for the risk of people who attended single-sex schools having social anxiety compared to people who attended co-ed schools is (0.78, 2.67). Select the appropriate conclusion.

1

- As the interval is fully above zero, we can be 95% confident that the risk of having social anxiety is greater among people who attended single-sex schools than it is among people who attended co-ed schools.
- As the interval is fully above zero, we can be 95% confident that the risk of having social anxiety is greater among people who attended co-ed schools than it is among people who attended single-sex schools.
- **As the interval contains 1, we have no evidence that there is a difference in the risk of having social anxiety between people who attend single-sex and co-ed schools.**
- As the interval is fully above 1, we can be 95% confident that the risk of having social anxiety is greater among people who attended single-sex schools than it is among people who attended single sex schools.
- As the interval is fully above 1, we can be 95% confident that the risk of having social anxiety is greater among people who attended co-ed schools than it is among people who attended single-sex schools.

Solution: The confidence interval is a plausible range for the risk ratio of people who attended single-sex schools having social anxiety compared to people who attended co-ed schools.

$$\frac{\text{Risk of a person who attended a single-sex school having social anxiety}}{\text{Risk of a person who attended a co-ed school having social anxiety}}$$

If this is 1, then the risk of having social anxiety across the two populations (those who attended single-sex schools, and those who attended co-ed schools) is the same. If this is greater than 1, the risk is higher among those who attended single-sex schools. While if this is less than 1, the risk is higher among those who attended co-ed schools.
So if the interval contains 1: 1 (the value indicating the risk is the same) is in the plausible range for the true risk ratio, and we don't have evidence (at the 5% significance level) that the risk of social anxiety differs between those who attended single-sex and co-ed schools.
If the interval is fully above 1: The plausible range for the risk ratio only contains values where the risk is greater for those who attended single-sex schools. So we have evidence the risk of social anxiety is greater for those who attended single-sex schools.
If the interval is fully below 1: The plausible range for the risk ratio only contains values where the risk is greater for those who attended co-ed schools. So we have evidence the risk of social anxiety is greater for those who attended co-ed schools.

g) Suppose that among people who attended co-ed schools, that people with social anxiety were less likely to be willing to participate in the study, while this did not occur among people who attended single-sex schools (i.e. those who attended single-sex schools were equally likely to participate whether they had social anxiety or not). Select the correct option below regarding how this could affect the study results.

1

- This is an example of information bias. It could bias the study results, explaining some of any increased risk of social anxiety among people who attend single-sex schools.
- This is an example of information bias. It could bias the study results, masking some of any increased risk of social anxiety among people who attend single-sex schools.
- **This is an example of selection bias. It could bias the study results, explaining some of any increased risk of social anxiety among people who attend single-sex schools.**
- This is an example of selection bias. It could bias the study results, masking some of any increased risk of social anxiety among people who attend single-sex schools.

- This is an example of random error. It could distort the study results, masking some of any increased risk of social anxiety among people who attend single-sex schools.

Solution: If people from co-ed schools with social anxiety were less likely to be willing to participate in the study than those without social anxiety, one would expect the risk of social anxiety in co-ed attendees to be lower in the study than it was in reality across the whole populationpf co-ed attendees. One would therefore expect the relative risk of social anxiety in single-sex attendees relative to co-ed attendees to be higher in the study than it was in reality. This could therefore explain some of any increased risk of social anxiety among single-sex attendees that was observed in the study.
This is an example of bias, because it systematically distorts (e.g. due to flaws in the study design, or human behaviour) the results observed in the study from the true results.
It is an example of selection bias, because it is to do with the process of bringing people into the study (in contrast to information bias, which is to do with the information obtained from people once they've joined the study).

Question 14

There is debate in philosophy and psychology as to whether humans are more motivated by emotion or reason.
To test whether people are more motivated to donate to charity by being logically convinced to through reason, or by being moved to donate emotionally, a charity sends out two groups of 26 campaigners to collect money from people in the street towards starving children in Africa.
One group, the "reason" group, tries to convince people to donate by using facts about child malnutrition and logical arguments, while the other, the "emotion" group, tries to convince people by displaying graphic images of children suffering from starvation and appealing to emotions like empathy.
They seek to explore whether people are more moved to donate through emotion or reason by examining the difference between the mean amount collected per "emotion" or "reason" campaigner.
The mean and variances of the amounts collected across the two groups of campaigners were:

	Advertising Campaign	
	Emotion	Reason
Campaigners	26	26
Mean ($\bar{x}_i$)	\$803.79	\$657.97
Variance	\$1931.03	\$1991.45

a) Letting μ_E and μ_R refer to the population-level mean amounts collected by campaigners using the "emotion" and "reason" strategies, respectively. Select the appropriate null and alternate hypotheses for the question the charity is interesting in answering.

1

- (●) $H_0 : \mu_E = \mu_R;\ \ H_A : \mu_E \neq \mu_R$
- ( ) $H_0 : \mu_E \neq \mu_R;\ \ H_A : \mu_E = \mu_R$
- ( ) $H_0 : \mu_E - \mu_R \neq 0;\ H_A : \mu_E - \mu_R = 0$
- ( ) $H_0 : \bar{x}_E = \bar{x}_R;\ \ H_A : \bar{x}_E \neq \bar{x}_R$
- ( ) $H_0 : \bar{x}_E - \bar{x}_R = 0;\ H_A : \bar{x}_E - \bar{x}_C \neq 0$

Solution: In the question brief it said that the charity seeks to explore whether people are more moved to donate through emotion or reason by examining the difference between the mean amount collected per "emotion" or "reason" campaigner.
The null hypothesis, being the "no difference" hypothesis, will be that there's no difference between the means, while the alternate hypothesis will be that there is a difference between the means.
The hypotheses concern the population-level means μ_E and μ_R, rather than the sample means $\bar{x}_E$ and $\bar{x}_R$. In statistics we are trying to find out about the population-level parameters using sub-samples of that population.

b) Based off this sample, the estimated standard error (in \$) for the difference between the mean amount collected per campaigner using emotion and reason is closest to:

1

- (●) 12.28
- ( ) 544.01
- ( ) 150.86
- ( ) 295951.92
- ( ) 27.74;

Solution: The formula for the standard error for a hypothesis test for a difference between means is the same as the formula for the standard error for confidence intervals for a difference between means.

From the formula sheet, this is

$$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

We're given the sample variances s_1^2 and s_2^2 in the table in the question (1931.03 and 1991.45), while the sample sizes within the two groups are both 26. So

$$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}=\sqrt{\frac{1931.03}{26}+\frac{1991.45}{26}}=12.28$$

c) Select which option the test statistic for their hypothesis test is closest to. Use the order of differencing $\bar{x}_E-\bar{x}_R$ for the observed sample value.

1

- -11.87
- 11.87
- 750.22
- 2.33
- 7.91

Solution:

$$\text{Test Statistic}=\frac{(\text{Observed Sample Value}-\text{Null Hypothesis Value})}{\text{Estimated Standard Error}}$$

Observed sample value: The observed sample value is the difference between sample means observed. The question said to use the order $\bar{x}_E-\bar{x}_R$ (in contrast to $\bar{x}_R-\bar{x}_E$), so the observed sample value is $\bar{x}_E-\bar{x}_R=803.79-657.97=145.82$.
Null value: The difference between means assumed under H_0 is zero, so the null value is zero.
Estimated Standard Error: Is as calculated in the previous question.
So

$$\text{Test Statistic}=\frac{(\text{Observed Sample Value}-\text{Null Hypothesis Value})}{\text{Estimated Standard Error}}=\frac{(145.82-0)}{12.28270}=11.87$$

Note more decimal places were used for the standard error in this calculation. This is advised so that the answer calculated is more accurate (although, there's less need to do this in a multi-choice exam, unless some of the multi-choice options given are very close to each other).

d) If the *p*-value calculated from the test statistic above is 0.0000 (when rounded to 4DP), select the approprate conclusion to the hypothesis test using $\alpha=0.01$ as the significance level.

1

- Reject the null hypothesis that the "emotion" and "reason" strategies are just as effective at collecting donations. Since the mean amount collected using the "emotion" strategy is greater in the sample, there is evidence that the "emotion" strategy is more effective.
- Fail to reject the null hypothesis that the "emotion" and "reason" strategies are just as effective at collecting donations. Since the mean amount collected using the "emotion" strategy is greater in the sample, there is evidence that the "emotion" strategy is more effective.
- Fail to reject the null hypothesis that the "emotion" and "reason" strategies are just as effective at collecting donations. There is no evidence (at the $\alpha=0.01$ significance level) of a difference between the effectiveness of the "emotion" and "reason" strategies.

- Reject the null hypothesis that the "emotion" strategy is more effective than the "reason" strategy at collecting donations. There is no evidence (at the $\alpha = 0.01$ significance level) of a difference between the effectiveness of the "emotion" and "reason" strategies.
- Fail to reject the null hypothesis that the "emotion" strategy is more effective than the "reason" strategy at collecting donations. Since the mean amount collected using the "emotion" strategy is greater in the sample, there is evidence that the "emotion" strategy is more effective.

Solution: In a hypothesis test, we reject the null hypothesis if our p-value is less than the significance level α, and fail to reject the null if the p-value is greater than or equal to α. The lower the p-value, the more evidence against the null hypothesis. The null hypothesis in this example is that the "emotion" and "reason" strategies are just as effective at collecting donations, or that there is no difference between the mean amounts collected per campaigner using each of the "emotion" and "reason" methods.
So **if the p-value is less than $\alpha = 0.01$:** we reject the null hypothesis that the strategies are just as effective (that the mean amount collected is the same whether potential donors are convinced using emotion or reason). We have evidence (at the $\alpha = 0.01$ significance level) against the strategies being just as effective. Since the mean amount collected using the emotion strategy is greater in the sample, there is evidence that the emotion strategy is more effective.
While **if the p-value is greater than or equal to $\alpha = 0.01$:**, we fail to reject H_0 at the 0.01 significance level. We have no evidence at the 0.01 significance level to reject the hypothesis that the strategies are just as effective (note: this doesn't mean we have evidence that the strategies are just as effective - no evidence of difference is NOT evidence of no difference).

Question 15

In psychology, "The Boomerang Effect" refers to the phenomena where an attempt to persuade has the opposite to the intended effect, resulting in people being more likely to believe or behave in manners opposing the intended effect of the persuation.
An example of this would be if an advertising campaign to reduce the proportion of teenagers vaping actually increased the proportion of teenagers vaping.
Some social commentators believe that exactly this occurred in the United States in 2018.
Suppose that in December 2017, it was known that 11.7% of U.S. high school students had vaped within the last month.
An anti-teen-vaping advertising campaign was ran by the U.S. Food and Drug Administration (FDA) across 2018, and in December 2018, a survey of 97 U.S. high school students found that 20 had vaped in the last month.

a) Set up the appropriate null and alternative hypothesis for addressing the question of whether the proportion of U.S. high school students who had vaped in the last month differed between December 2017 and December 2018. In the options below, π and p refer to the population and sample proportions who had vaped within the last month in December 2018.

1

- $H_0 : p = 0.117; \;\; H_A : p \neq 0.117$
- $H_0 : \pi \neq 0.117; \;\; H_A : \pi = 0.117$
- **$H_0 : \pi = 0.117; \;\; H_A : \pi \neq 0.117$**
- $H_0 : \pi = 0.206; \;\; H_A : \pi = 0.206$
- $H_0 : \mu = 0.206; \;\; H_A : \mu \neq 0.206$

Solution: The null hypothesis is the hypothesis of no difference, so the null hypothesis is that the proportions are the same between 2017 and 2018.
In the question we are given that the proportion in 2017 was 0.117, so therefore the null hypothesis is that the proportion in 2018 was 0.117 too.
The null hypothesis concerns π, the population proportion, rather than p (we are always testing for the population level parameters in hypothesis tests). So H_0 is that $\pi = 0.117$.
The alternate hypothesis is the opposite statement to the null hypothesis, so if H_0 is that $\pi = 0.117$, H_A is that $\pi \neq 0.117$.

b) Select the correct statement below about type II errors and this study.

1

- A type II error in this study would be rejecting the hypothesis that the proportion vaping is the same between 2017 and 2018, when the proportions differ.
- A type II error in this study would be rejecting the hypothesis that the proportion vaping is the same between 2017 and 2018, when the proportions are the same.
- A type II error in this study would be failing to reject the hypothesis that the proportion vaping is the same between 2017 and 2018, when the proportions are the same.
- **A type II error in this study would be failing to reject the hypothesis that the proportion vaping is the same between 2017 and 2018, when the proportions differ.**
- A type II error in this study would be failing to reject the hypothesis that the proportion vaping differ between 2017 and 2018, when the proportions differ.

Solution: A type II error is failing to reject the null hypothesis when it is false. It is also called a "false negative" error. Our null hypothesis in this example is that the proportions vaping are the same between 2017 and 2018, so a type II error is failing to reject the hypothesis that the

proportions are the same, when they actually differ.

c) Presuming the researchers were planning to condust a hypothesis test with $\alpha = 0.01$, tweaking which of the below study design features would definitely increase the chance of the researchers making a type II error?

1

- Using a sample size smaller than 97. Increasing α to 0.05.
- Using a sample size smaller than 97. Decreasing α to 0.001.
- Using a sample size larger than 97. Increasing α to 0.05.
- Increasing the power of the study. Decreasing α to 0.001.
- Increasing the power of the study. Increasing α to 0.05.

Solution: Sample size and the significance level are the two factors study designers can alter that influence the type II error rate. Using diagrams is very helpful in understanding how these influence the type II error rate, so reviewing the hypothesis test lecture on power where this is explained using diagrams may be a good idea to understand this. But here is the explanation using words alone:
Sample Size: Using a larger sample size reduces the chance of making a type II error (the type II error rate), while using a smaller sample size will increase the chance of making a type II error. The reason is because the smaller the sample size used, the wider the sampling distributions are, and the larger the overlap between what the sampling distribution would be if H_0 was true and the actual sampling distribution. The probability of observing a sample result (a sample proportion, in this case) within each range is determined by the actual sampling distribution. If this has more overlap with the sampling distribution under H_0 as it would when the sample size is smaller, then the chance of observing a sample result within the ranges within the centre of the distribution under H_0 where we'd fail to reject H_0 increases, so the probability of failing to reject H_0 when it is false (of making a type II error) increases.
Significance Level: Reducing α increases the type II error rate, while increasing α reduces the type II error rate. Reducing α means we require lower *p*-values before we reject H_0. It means we are failing to reject H_0 over a larger range of values, so increases the chance of us failing to reject H_0, and increases the chance of failing to reject H_0 when it is false (of making a type II error).

d) Choose the appropriate calculation of the test statistic for this hypothesis test.

1

- $\sqrt{\frac{0.206(1-0.206)}{97}}$
- $\sqrt{\frac{0.117(1-0.117)}{97}}$
- $\frac{(0.117-0.206)}{\sqrt{\frac{0.117(1-0.117)}{97}}}$
- $\frac{(0.206-0.117)}{\sqrt{\frac{0.117(1-0.117)}{97}}}$
- $\frac{(0.206-0.117)}{\sqrt{\frac{0.206(1-0.206)}{97}}}$

Solution: Test statistics for a hypothesis test for a proportion are calculated using

$$\text{Test Statistic} = \frac{(\text{Observed Sample Value} - \text{Null Hypothesis Value})}{\text{Standard Error}}$$

Observed sample value: For a HT for a proportion, this is the proportion observed in the sample (0.206).
Null Value: Under the null hypothesis, the proportions vaping are the same between 2017 and 2018. Since we're assuming the proportion in 2017 is 0.117, the proportion we're assuming in 2018 is also 0.117, so the null proportion is 0.117.
Standard Error: This is what the standard error for the proportion (the standard deviation of the distribution of sample proportions) would be if the null hypothesis was true

$$\sqrt{\frac{\pi_0(1-\pi_0)}{n}} = \sqrt{\frac{0.117(1-0.117)}{97}} = 0.033$$

where π_0 is the proportion vaping under the null hypothesis. Note that this differs from the standard error for a confidence interval for a single proportion

$$\sqrt{\frac{p(1-p)}{n}}$$

which uses the sample proportion *p* instead of the proportion assumed under H_0. The reason we use the null hypothesis proportion for the standard error in a HT is because the *p*-value is the probability of observing the sample data we observed, or data further away from the data we observed, *assuming the null hypothesis is true.*

Therefore, the test statistic is

$$\begin{aligned}\text{Test Statistic} &= \frac{(\text{Observed Sample Value} - \text{Null Hypothesis Value})}{\text{Standard Error}} \\ &= \frac{(0.206 - 0.117)}{\sqrt{\frac{0.117(1-0.117)}{97}}}\end{aligned}$$

e) The *p*-value calculated from the test statistic is 0.006 (rounded to 3DP). Select the appropriate conclusion to the hypothesis test.

1

- As the *p*-value is greater than or equal to 0.01, we fail to reject H_0 at the $\alpha = 0.01$ significance level. There is no evidence at the 0.01 significance level that the proportion vaping across the population differed between 2017 and 2018.
- As the *p*-value is less than or equal to 0.01, we reject H_0 at the $\alpha = 0.01$ significance level. Since the sample proportion vaping in 2018 is less than the sample proportion vaping in 2017, there is evidence at the $\alpha = 0.01$ significance level that the proportion vaping across the population reduced between 2017 and 2018.
- (selected) As the *p*-value is less than or equal to 0.01, we reject H_0 at the $\alpha = 0.01$ significance level. Since the sample proportion vaping in 2018 is greater than the sample proportion vaping in 2017, there is evidence at the $\alpha = 0.01$ significance level that the proportion vaping across the population increased between 2017 and 2018.
- As the *p*-value is greater than or equal to 0.01, we reject H_0 at the $\alpha = 0.01$ significance level. Since the sample proportion vaping in 2018 is greater than the sample proportion vaping in 2017, there is evidence at the $\alpha = 0.01$ significance level that the proportion vaping across the population increased between 2017 and 2018.

- As the p-value is less than 0.01, we fail to reject H_0 at the $\alpha = 0.01$ significance level. There is no evidence at the 0.01 significance level that the proportion vaping across the population differed between 2017 and 2018.

Solution: If the p-value is less than 0.01, we reject H_0 at the 0.01 significance level, while if the p-value is greater than or equal to 0.01, we fail to reject H_0 at the 0.01 significance level. We have no evidence at the 0.01 significance level that the proportions vaping differed between 2017 and 2018.
If we fail to reject H_0 at the 0.01 significance level (i.e. if the p-value is greater than or equal to 0.01), we have no evidence at the 0.01 significance level to reject the null hypothesis that the proportions vaping across the whole population are the same between 2017 and 2018.
If we reject H_0 at the 0.01 significance level (i.e. the p-value is less than 0.01), we have evidence against the proportions vaping in 2018 and 2017 are the same, and that the proportions vaping differ between 2018 and 2017. If the sample proportion vaping is greater in 2018, we have evidence that the proportion across the whole population increased between 2017 and 2018. If the sample proportion vaping is less in 2018 than 2017, there is evidence that the proportion vaping reduced between 2017 and 2018.

Question 16

A group of educational psychologists are interested in whether attending early childhood education (ECE) improves educational attainment across children's lives. They select a group of 18-year-olds, 100 of whom attended ECE, and 100 of whom didn't.
They find that 78 of the 18-year-olds who attended ECE had attained University Entrance (UE), while 49 of those who hadn't attended ECE had attained UE.
They conduct a hypothesis test to explore the relationship between ECE attendance and UE attainment rates. Their null hypothesis is that there is no difference between the UE attainment rates among those who had and hadn't attended ECE.

a) Based off this sample, what is the estimate of the difference between the proportion of ECE attendees that attained UE, and the proportion of non-ECE attendees that attained UE?

1

- 0.5
- 0.29
- -0.29
- 0.64
- 0.31

Solution: Let p_a be the sample proportion of ECE attendees that attained UE, and p_{na} be the sample proportion of non-ECE attendees that attained UE.
The estimate of the difference between the true proportion of ECE attendees that attained UE (π_a) and the true proportion of non-ECE attendees that attained UE (π_{na}) is the difference between the sample proportions

$$p_a - p_{na} = 78/100 - 49/100 = 0.29$$

Note: The order of the difference is important. If the question had asked for the difference between the proportion of non-ECE attendees that attained UE, and the proportion of ECE attendees that attained UE, we would have subtracted in the other order (i.e. done $p_{na} - p_a$ rather than $p_a - p_{na}$).

b) The estimated standard error for the difference between the proportion of ECE attendees that attained UE, and the proportion of non-ECE attendees that attained UE is closest to:

1

- 0.0649
- 0.0665
- 0.0670
- 0.0046
- 0.0681

Solution: From the formula sheet, the standard error in a hypothesis test for the difference between 2 proportions is

$$\sqrt{\frac{p^*(1-p^*)}{n_1} + \frac{p^*(1-p^*)}{n_2}}$$

where p^* is the pooled proportion across both groups in the sample

$$p^* = \frac{x_1 + x_2}{n_1 + n_2}$$

where x_1, x_2, n_1 and n_2 respectively are the number with the outcome we're interesting in (attaining UE) in the first sample, the number with the outcome we're interested in in our second sample, the total number in the first sample, and the total number in the second sample (Note: it doesn't matter which sample we label as the first and second sample when calculating the standard error). Letting x_1 and n_1 refer to the ECE attendees sample, and x_2 and n_2 refer to the non-ECE attendees sample, the pooled proportion is

$$p^* = \frac{x_1 + x_2}{n_1 + n_2} = \frac{(78 + 49)}{(100 + 100)} = 0.64$$

Then the standard error is

$$\sqrt{\frac{p^*(1-p^*)}{n_1} + \frac{p^*(1-p^*)}{n_2}} = \sqrt{\frac{0.64(1-0.64)}{100} + \frac{0.64(1-0.64)}{100}} = 0.0681$$

Note: A common mistake is to compute the standard error for a confidence interval for a difference between two proportions

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

instead of the standard error for a hypothesis test for a difference between 2 proportions.

c) The test statistic for their hypothesis test is closest to which option below? Use the order of differencing "proportion of ECE attendees that attained UE" minus "proportion of non-ECE attendees that attained UE" for the observed sample value when calculating the test statistic.

1

- 60.237
- 5.259
- -4.259
- 4.259
- 3.759

Solution: The test statistic is calculated using

$$\text{Test Statistic} = \frac{\text{Observed Sample Value} - \text{Null Value}}{\text{Estimated Standard Error}}$$

The observed sample value is the difference between the sample proportions. The question stated to use the order of differencing "proportion of ECE attendees that attained UE" minus "proportion of non-ECE attendees that attained UE", so using this order of differencing the difference between the sample proportions is $78/100 - 49/100 = 0.29$
The null hypothesis, being the "no difference" hypothesis, is the hypothesis that there is no difference between the proportion attaining UE who attend and don't attend ECE. So the null value is 0.
And the standard error is calculated in the previous question. So

$$\text{Test Statistic} = \frac{\text{Observed Sample Value} - \text{Null Value}}{\text{Estimated Standard Error}} = \frac{(0.29 - 0)}{0.068085} = 4.259$$

Note: Even though the previous question gave the answer for the standard error to 4 DP, it is advised to use as many decimal places as possible when using previously calculated results to calculate other results to get as accurate and answer as possible. This is why the standard error to 6 DP was used in this calculation.

d) The *p*-value for their hypothesis test for the difference between the proportion of ECE attendees that attained UE, and the proportion of non-ECE attendees that attained UE is 0.0000 (when rounded to 4DP). Select the appropriate conclusion to their hypothesis test.

1

- (●) Since the *p*-value is less than 0.05, there is evidence at the $\alpha = 5\%$ significance level to reject the null hypothesis that the probability a person attains UE does not depend on whether they attended ECE or not. Since the sample proportion attaining UE is higher among ECE attendees than non-ECE attendees, there is evidence at the 5% level that the proportion attaining UE is higher among ECE attendees than non-ECE attendees.
- ( ) Since the *p*-value is less than 0.05, there is evidence at the $\alpha = 5\%$ significance level to reject the null hypothesis that the probability a person attains UE does not depend on whether they attended ECE or not. Since the sample proportion attaining UE is higher among non-ECE attendees than ECE attendees, there is evidence at the 5% level that the proportion attaining UE is higher among non-ECE attendees than ECE attendees.
- ( ) Since the *p*-value is less than 0.05, there is no evidence at the 5% significance level that the probability a person attains UE depends on whether they attended ECE or not.
- ( ) Since the *p*-value is greater than or equal to 0.05, there is no evidence at the 5% significance level that the probability a person attains UE depends on whether they attended ECE or not.
- ( ) Since the *p*-value is greater than or equal to 0.05, there is evidence at the $\alpha = 5\%$ significance level to reject the null hypothesis that the probability a person attains UE does not depend on whether they attended ECE or not. Since the sample proportion attaining UE is higher among ECE attendees than non-ECE attendees, there is evidence at the 5% level that the proportion attaining UE is higher among ECE attendees than non-ECE attendees.

Solution: **If the *p*-value is less than 0.05 :** The *p*-value is lower than the 0.05, so there is less than a 5% chance of observing a difference between sample proportions as great or greater than the one observed, if the null hypothesis (that there is no difference between the proportion attaining UE between ECE and non-ECE attendees across the population) is true. Our sample result is therefore unlikely enough at the $0.05 = 5\%$ significance level that we have evidence to reject H_0 at the 5% significance level.

Since the sample proportion that attained UE is higher among ECE attendees ($73/100 = 0.73$) than non-ECE attendees ($49/100 = 0.49$, there is evidence that the population-level proportion attaining UE is greater among ECE attendees than non-ECE attendees.
If the *p*-value is greater than or equal to 0.05 : The *p*-value is greater than or equal to 0.05, so there is more than a 5% chance of observing a difference between sample proportions as great or greater than the one observed, if the null hypothesis (that there is no difference between the proportion attaining UE between ECE and non-ECE attendees across the population) is true. This is not an unlikely enough result to reject H_0 at the 5% significance level.

e) In determining whether attending ECE improves the chance of a person attaining UE, the educational psychologists conducting the study considered whether socioeconomic background could be confounding the relationship between attending ECE and attaining UE. Select the correct statement below.

1

- If coming from a higher socioeconomic background is positively associated with attending ECE, and coming from a higher socioeconomic background improves educational attainement, socioeconomic background could confound the relationship between attending ECE and attaining UE, making it appear like attending ECE improves UE attainement *less* than it actually does.
- **If coming from a higher socioeconomic background is positively associated with attending ECE, and coming from a higher socioeconomic background improves educational attainement, socioeconomic background could confound the relationship between attending ECE and attaining UE, making it appear like attending ECE improves UE attainement *more* than it actually does.**
- If coming from a higher socioeconomic background is negatively associated with attending ECE, and coming from a higher socioeconomic background improves educational attainement, socioeconomic background could confound the relationship between attending ECE and attaining UE, making it appear like attending ECE improves UE attainement *more* than it actually does.
- If coming from a higher socioeconomic background is not associated with attending ECE, and coming from a higher socioeconomic background improves educational attainement, socioeconomic background could confound the relationship between attending ECE and attaining UE, making it appear like attending ECE improves UE attainement *less* than it actually does.
- If coming from a higher socioeconomic background is associated with attending ECE, and coming from a higher socioeconomic background is not associated with educational attainement, socioeconomic background could confound the relationship between attending ECE and attaining UE, making it appear like attending ECE improves UE attainement *less* than it actually does.

Solution: For a variable to confound in the relationship between an exposure and an outcome, at a minimum, it needs to be associated (whether it be positively or negatively associated) with both the exposure and the outcome. So socioeconomic background needs to at least be either positively or negatively associated with both attending ECE and attaining UE to be a confounder in the relationship between attending ECE and attaining UE.
If higher socioeconomic background is positively associated with ECE attendance, then those attending ECE are more likely to from a higher socioeconomic background. If higher socioeconomic background improves educational attainement (e.g. the chance of attaining UE), then those attending ECE are likely to have higher UE attainement rates not because of the effect of attending ECE, but because of their socioeconomic background. So socioeconomic background could make attending ECE appear to be more beneficial in improving UE attainement than it actually is.

Question 17

In an attempt to reduce depression rates in primary school age children, The Ministry of Education are considering whether to fund cognitive behavioural therapy (CBT) sessions with a counsellor for children clinically diagnosed with depression.
The Childrens' Depression Rating Scale (CDRS) is used to assess whether children between 6-12 years old are depressed. A CDRS score of higher than 40 indicates a child is depressed.
The Ministry of Education conducted the following study: They found a group of depressed 6-12 year old children, and assessed their CDRS score. They followed up with these children one year later, reassessing their CDRS score. They compared the mean reduction in the CDRS scores for children who had and hadn't gone through a program of CBT.

a) What type of study did The Ministry of Education use?

1

- ○ Descriptive, observational, cohort study.
- ◉ Analytic, observational, cohort study.
- ○ Analytic, observational, case-control study.
- ○ Analytic, experimental, cohort study.
- ○ Descriptive, experimental, case-control.

Solution: The study is an analytic and not a descriptive study, because it is exploring the effect one variable (CBT sessions) has on another (depression), rather than simply exploring a single variable.
It is observational rather than experimental, because the researchers are simply observing results of the participants' natural behaviours, rather than intervening in their behaviours.
It is a cohort study, because it tracks the study participants over time, rather than looking back at their past behaviour (as a case-control study would).

b) The Ministry of Education conducted a hypothesis test for the difference in the mean CDRS score reductions between children who had CBT sessions and those who didn't. The probability that they conclude there is a difference in the mean CDRS score reduction between children who went through the program and children who didn't, when there is no difference between the mean CDRS score reductions, is:

1

- ◉ the probability of making a type I error.
- ○ the probability of having a false negative result.
- ○ the probability of making a type II error.
- ○ the positive predictive value of the test.
- ○ the power of the test.

Solution: When one rejects the null hypothesis when it is actually true, they make a type I error. The null hypothesis in the test The Ministry are conducting would be that there is no difference between the mean CDRS score reductions. So concluding there is a difference between the mean CDRS score reductions (rejecting the hypothesis that there is no difference between the mean CDRS score reductions) when there is in fact no difference is a case of rejecting the null hypothesis when it is true, and is therefore an example of making a type I error.

c) Suppose The Ministry of Education deems CBT to have a clinically significant benefit at treating children's depression if it reduces the mean CDRS score by at least 10 points more than not going through CBT does. This is the minimum reduction in CDRS score the ministry requires for them to fund CBT sessions for depressed

primary school children.
The study's 95% confidence interval for the difference in the mean CDRS score **reduction** between children who went through CBT and those who didn't is $(2, 13)$ (so, positive numbers here indicate a greater CDRS score reduction in the study's CBT group). Select the appropriate conclusions regarding whether the study provides statistical and clinical significance at the $\alpha = 0.05$ significance level.

1

- The result is not statistically significant, in that it provides evidence that the CBT treatment provides some benefit. And the result provides evidence that the CBT treatment provides clinically significant benefit.
- The result is not statistically significant, in that it does not provide evidence as to whether or not the CBT treatment provides some benefit. And the result does not provide evidence as to whether or not the CBT treatment provides clinically significant benefit.
- The result is statistically significant, in that it provides evidence that the CBT treatment provides some benefit. And the result provides evidence that the CBT treatment provides clinically significant benefit.
- The result is statistically significant, in that it provides evidence that the CBT treatment provides some benefit. And the result provides evidence that the CBT treatment does not provide clinically significant benefit.
- **(selected)** The result is statistically significant, in that it provides evidence that the CBT treatment provides some benefit. And the result does not provide evidence as to whether or not the CBT treatment provides clinically significant benefit.

Solution: To answer this question, we need to consider whether or not 0 and the clinically significant value (10) is in the confidence interval, and if any of these are outside the interval, whether or not the interval is fully above or below the value(s) outside the interval.
Statistical significance concerns whether or not we have evidence to reject the null hypothesis. The null hypothesis, being the "no difference" hypothesis, is that there is no difference in the mean depression reduction between going through the CBT treatment and not going through the CBT treatment. So statistical significance concerns whether we have evidence against the difference in mean depression score reduction being 0. If 0 is outside the 95% CI, we have evidence against the difference being 0 at the $1 - 0.95 = 0.05 = 5\%$ significance level, so we have statistical significance. 0 is outside the interval $(2, 13)$ in this example. Since the interval is fully above 0, we have evidence that the CBT treatment reduces depression score by more than not having the treatment (i.e. that it provides some benefit).
Clinical significance concerns whether or not the treatment provides a noteworthy benefit. In this example, it is the minimum benefit for the ministry to fund the treatment. The confidence interval $(2, 13)$ contains the clinically significant value 10, so it is inconclusive with regard to clinical significance: it does not provide evidence as to whether the treatment meets, or does not meet the clinically significant benefit. If 10 was outside the interval, we would have evidence as to whether clinical significance was met: if the interval was fully above 10, we would have evidence clinical significance was met; if the interval was fully below 10, we would have evidence that clinical significance was not met.

d) Suppose that children whose parents provide them with CBT treatment for their depression are more likely to have other chronic health conditions, that having other chronic health conditions perpetuates depression, and other chronic health conditions are not part of any mechanism by which CBT influences depression. Select the correct statement below.

1

- Having other chronic health conditions could be biasing the relationship between undergoing CBT treatment and depression. It could be making CBT treatment appear less beneficial than it actually is at reducing depression.

- Having other chronic health conditions could be biasing the relationship between undergoing CBT treatment and depression. It could be making CBT treatment appear more beneficial than it actually is at reducing depression.
- Having other chronic health conditions could be confounding the relationship between undergoing CBT treatment and depression. It could be making CBT treatment appear less beneficial than it actually is at reducing depression.
- Having other chronic health conditions could be confounding the relationship between undergoing CBT treatment and depression. It could be making CBT treatment appear more beneficial than it actually is at reducing depression.
- Having other chronic health conditions neither acts as a confounder or a bias in the relationship between CBT treatment and depression.

Solution: If children whose parents provide them with CBT treatment for their depression are more likely to have other chronic health conditions, and having other chronic health conditions perpetuates depression, having other chronic health conditions would be associated with the study's exposure (undergoing CBT treatment), and it would also effect the study's outcome (change in depression score). If we assume other chronic health conditions are not part of any mechanism by which CBT treatment influences depression, then it is not on the causal pathway between the exposure (CBT treatment) and the outcome (change in depression). It therefore satisfies the definition of being a confounder in the relationship between CBT treatment and change in depression.
If children whose parents provide them with CBT treatment for their depression are more likely to have other chronic health conditions, and having other chronic health conditions perpetuates depression, having other chronic health conditions would be working against any benefit the CBT provides in reducing depression in the study participants, making the CBT appear less beneficial than it actually is in reducing depression.

Question 18

A group of criminologists are interested in exploring the correlation between educational attainment and crime.
They utilised a large dataset featuring many socioeconomic variables, including ones measuring educational attainment and crime, across a sub-sample of 2118 communities in the U.S.. The dataset was built by linking data from the 1990 U.S. Census with FBI data.
The criminologists examined the extent to which educational attainment - as measured by the percentage of people over the age of 25 in each community who had graduated high school (**HSGrad**), predicted crime (**CrimeRate**) - as measured by the number of non-violent criminal offences recorded per $100,000$ people per year in each community, by running a simple linear regression model on the dataset. The output of their analysis in R is below.

```
> summary(mymodel)

Call:
lm(formula = CrimeRate ~ HSGrad)

Residuals:
    Min      1Q  Median      3Q     Max
-7044.3 -1626.6  -440.3  1174.9 22881.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12412.44     391.83   31.68   <2e-16 ***
HSGrad        -96.52       4.99  -19.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2526 on 2116 degrees of freedom
  (97 observations deleted due to missingness)
Multiple R-squared:  0.1502,    Adjusted R-squared:  0.1498
F-statistic: 374.1 on 1 and 2116 DF,  p-value: < 2.2e-16
```

a) What roles do the variables play in the regression model? In the model, **HSGrad** and **CrimeRate** refer to the percentage of people who had graduated high school by age 25 and the non-violent crime rates per $100,000$ people per year, respectively.

1

- (●) Explanatory variable: **HSGrad**; Response variable: **CrimeRate**.
- ( ) Explanatory variable: **CrimeRate**; Response variable: **HSGrad**.
- ( ) Explanatory variable: **Community**; Response variable: **CrimeRate**.
- ( ) Dependent variable: **HSGrad**; Independent variable: **CrimeRate**.
- ( ) Dependent variable: **Community**; Independent variable: **HSGrad**.

Solution: From the question brief, the criminologists are examining the extent to which the percentage who have graduated high school in a community (**HSGrad**) predicts its crime rate (**CrimeRate**). So **HSGrad** is the explanatory, or independent, or predictor variable, while **CrimeRate** is the response, dependent, or outcome variable.

b) What is the equation for the fitted model regression line in the simple linear regression analysis that the researchers ran?

1

- ( ) $391.83 + 4.99\hat{\beta}_1$
- ( ) $-96.52 + 12412.44x$
- (●) $12412.44 - 96.52x$
- ( ) $12412.44 - 96.52\hat{\beta}_1$
- ( ) $12412.44 + 391.83x$

Solution: In running a simple linear regression model to predict the crime rates of communities based off their high school graduation rates, we are assuming the relationship between the mean crime rates (μ_Y) across communities where $x\%$ graduated high school follows the relationship

$$\mu_Y = \beta_0 + \beta_1 \times x$$

The fitted model for a simple linear regression line is our estimate of the true regression line. This takes the form $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. $\hat{y}$ is both the predicted mean crime rate in communities where $x\%$ graduated high school, and the estimated crime rate for a community where $x\%$ graduated high school.
To find the regression line equation, we need to find $\hat{\beta}_0$ and $\hat{\beta}_1$. Both of these can be read from the regression summary output table, with $\hat{\beta}_0$ being the value in the "Estimate" column and "(Intercept)" row (12412.44), and $\hat{\beta}_1$ being the value in the "Estimate" column and "HSGrad" row of the table (-96.52). So the fitted line equation is

$$\hat{\beta}_0 + \hat{\beta}_1 \times x = 12412.44 + (-96.52)x = 12412.44 - 96.52x$$

Below is a general guide to the outputs from R regression summaries that you'll need to be able to read in the final exam, and how to read these outputs.
The rows of the regression table: Each row corresponds to the independent variables in the model, with the exception of the top "(Intercept)" row, which gives us information on the intercept ($\hat{\beta}_0$) of the fitted model. In this example, the percentage that graduated high school (**HSGrad**) is the only independent variable, so this is the only row in the table besides the intercept row.
The estimates for the intercept and slope parameters: These can be read from the "Estimate" column of the output table. The value in the "(Intercept)" row is the value for $\hat{\beta}_0$, while the estimated slope parameters for each independent variable is in the row corresponding to that variable. In this example, the slope parameter for the **HSGrad** variable is -96.52.
The estimated standard errors: The estimated standard errors for the intercept and slope parameters can be found in the "Std. Error" column. For example, the estimated standard error for the slope parameter β_1 for the **HSGrad** variable is 4.99 in this example. This describes how much we can expect the values for $\hat{\beta}_1$ to vary between different samples.
t value column: For the intercept and each independent variable in the model, this column gives the value of the test statistic for the hypothesis test that the parameter corresponding to that variable is zero. For example, the value in the "t value" column and "HSGrad" row gives the test statistic in the hypothesis test that the slope parameter $\hat{\beta}_1$ for the **HSGrad** variable is zero.
Pr(>|t|) column: For the intercept and each independent variable in the model, this column gives the *p*-value for the hypothesis test that the parameter corresponding to the intercept or that variable is zero. For example, the value in the "Pr(>|t|)" column and "HSGrad" row gives the *p*-value in the hypothesis test that the slope parameter $\hat{\beta}_1$ for the **HSGrad** variable is zero.
Residual Standard Error: As the name suggests, this gives the residual standard error of the sample data around the fitted model.
Multiple R-squared: This is the R^2 value, or the proportion of the variation in the dependent variable (the crime rates, in our example) that can be explained by the independent variable (the percentage who had graduated high school, in this example).

c) In one community in the sample, 60% of people graduated high school by age 25, and the crime rate was 6530 offences per $100,000$ people per year. Which of the below options is the raw residual for this data point closest to?

1

- (●) -91.2
- ( ) -20.0
- ( ) 91.2
- ( ) 20.0
- ( ) 6530

Solution: The raw residual is the difference between the actual value of a response variable for an individual and the value predicted by the model based off the value(s) of its predictor variable(s). So, in this case, it is the crime rate observed in this community (y_i) minus the crime rate predicted by the model based off the community's high school graduation rate being 60%. The observed crime rate in the community is $y_i = 6530$, while the predicted crime rate is $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 \times x = 12412.44 - (96.52 \times 60)$, so the raw residual is

$$\hat{e_i} = y_i - \hat{y_i} = y_i - (\hat{\beta}_0 + \hat{\beta}_1 \times x) = 6530 - (12412.44 - 96.52 \times 60) = -91.2$$

d) The 95% prediction interval for the crime rate (offences per $100,000$ people per year) in a community where 60% of people graduated high school is $(1670.3, 11572.2)$. Select the appropriate conclusion that can be drawn from this interval.

1

- ( ) There is a 95% chance that any given community's crime rate is between 1670.3 and 11572.2 offences per $100,000$ people per year.
- ( ) There is a probability of 0.95 the mean crime rate across communities where 60% of people have graduated high school by the age of 25 is between 1670.3 and 11572.2 offences per $100,000$ people per year.
- (●) If we take a randomly selected community where 60% of people have graduated high school by the age of 25, there is a probability of 0.95 that their crime rate is between 1670.3 and 11572.2 offences per $100,000$ people per year.
- ( ) There is a probability of 0.95 that a community where the crime rate is between 1670.3 and 11572.2 offences per $100,000$ people per year has a high school graduation rate of 60%.
- ( ) For every 1% increase in high school graduation rates, there is a probability of 0.95 that the crime rate increases by between 1670.3 and 11572.2 offences per $100,000$ people per year.

Solution: A 95% prediction interval in a regression analysis is an interval for which there is a probability of 0.95 that the response variable for a randomly selected *individual* (e.g. a community) is between, *for a specific value* of the predictor variable (e.g. for a specific high school graduation rate). So in this example, if $(1670.3, 11572.2)$ is a 95% prediction interval for the crime rate in a community with a high school graduation rate of 60%, then the probability is 0.95 that any randomly selected community with a graduation rate is 60% has a crime rate between 1670.3 and 11572.2.

A common mistake is that the prediction interval does not give a range for the mean response across all individuals with some specific value of the predictor variable. For example, the 95% prediction interval does not give a range for which the probability is 0.95 that the mean crime rate across communities where the high school graduation rate is 60%. This is closer to what a confidence interval for the mean response provides. Prediction intervals relate to single individuals (e.g. individual communities).

Prediction intervals also don't give a range we can be confident the response variable is between across all values of the predictor variable: they only give a range we can be confident the

response variable is between for a specific value of the predictor variable (e.g. a 80% high school graduation rate). We might expect crime rates across communities where 99% graduated high school to differ significantly from communities where 60% graduated high school, for instance.

e) The estimated standard error for the slope of the regression line is:

1

- 12412.44
- -96.52
- 4.99
- 2526
- 0.1502

Solution: Following the guide above for interpreting regression summary tables in the earlier answer, it is the entry in the "Std. Error" column and "HSGrad" row of the output table.

f) Select the appropriate R command for how the values for $\hat{\beta}_1$ and s_{β_1} that are given in the R summary table within the question brief can be used to compute the *p*-value in a hypothesis test for whether the slope parameter β_1 is zero. Recall that the sample featured 2118 communities.

1

- $2 \times \text{pt}(-96.52/4.99, 2117)$
- $2 \times \text{pt}(-96.52/4.99, 2116)$
- $2 \times \text{pt}(-96.52/4.99, 2116, \text{lower.tail=False})$
- $2 \times \text{pt}(12412.44/391.83, 2116, \text{lower.tail=False})$
- $2 \times \text{pt}(12412.44, 2117, \text{lower.tail=False})$

Solution: To compute the *p*-value for a hypothesis test, we need to know the test statistic. The test statistic for the slope of a regression line takes the form

$$\frac{\text{Estimate} - \text{Null Value}}{\text{Estimated Standard Error}}$$

Estimate: Since we're doing a hypothesis test for the slope parameter, the estimate is the estimate for the slope parameter $\hat{\beta}_1$. This is in the "Estimate" column of the "HSGrad" row of the table as -96.52 (It is stated in the question brief that "HSGrad" refers to the high school graduation rates in the model).
Null Value: The null value in a hypothesis test for a regression slope parameter is always zero (at least in this course), since we're interested in whether there is an association between the variables, and this is based off whether the true slope parameter is zero (meaning no association) or is not zero (meaning some association).
Standard Error: The standard error for the slope of the regression line can be read from the entry in the "Std. Error" column and "HSGrad" row of the regression line.

So the test statistic for the hypothesis test is

$$\frac{\text{Estimate} - \text{Null Value}}{\text{Estimated Standard Error}} = \frac{-96.52 - 0}{4.99} \frac{-96.52}{4.99}$$

To get the *p*-value from the test statistic, recall that the *p*-value is the probability of observing a sample result as far or further from the null value than the one observed, if the null hypothesis is true. This corresponds to the probability of observing test statistics further from zero than the one observed if H_0 is true. The distribution the test statistics follows under the null hypothesis in a hypothesis test for the slope of the regression line is the *t* distribution with $n - k - 1$ degrees

of freedom, where n is the sample size, and k is the number of independent variables in the model. $n = 2118$ and $k = 1$ here, so the degrees of freedom is $2118 - 1 - 1 = 2116$. So the p-value, or probability of observing a test statistic further from zero than the one observed (of $-96.52/4.99$), is then the probability of observing a value further from zero than $-96.52/4.99$ on the t-distribution with 2116 degrees of freedom. Values further from zero than $-96.52/4.99$ are ones less than $-96.52/4.99$, and ones more than $96.52/4.99$. $\text{pt}(-96.52/4.99, 2116)$ gives the lower tail probability of observing a value less than $-96.52/4.99$, while the upper tail probability featuring values more than $96.52/4.99$ will be the same by the symmetry of the t-distribution, so the total probability or the p-value is

$$\text{pt}(-96.52/4.99, 2116) + \text{pt}(-96.52/4.99, 2116) = 2 \times \text{pt}(-96.52/4.99, 2116)$$

g) If the 95% confidence interval for the slope parameter is $(-106.5, -86.54)$, select the appropriate conclusion.

1

- (●) We can be 95% confident that the true slope parameter β_1 is between -106.5 and -86.54. The interval provides evidence that there is an association between crime rates and high school graduation rates, with crime rates tending to decrease as high school graduation rates increase.
- ( ) We can be 95% confident that the true slope parameter β_1 is between -106.5 and -86.54. The interval provides evidence that there is an association between crime rates and high school graduation rates, with crime rates tending to increase as high school graduation rates increase.
- ( ) We can be 95% confident that the estimated slope parameter $\hat{\beta}_1$ is between -106.5 and -86.54. The interval provides evidence that there is an association between crime rates and high school graduation rates, with crime rates tending to decrease as high school graduation rates increase.
- ( ) We can be 95% confident that the estimated slope parameter $\hat{\beta}_1$ is between -106.5 and -86.54. The interval provides no evidence that there is an association between crime rates and high school graduation rates.
- ( ) We can be 95% confident that the true slope parameter β_0 is between -106.5 and -86.54. The interval provides evidence that there is an association between crime rates and high school graduation rates, with crime rates tending to increase as graduation rates increase.

Solution: The true slope parameter is β_1, while the estimated slope parameter is $\hat{\beta}_1$. A 95% confidence interval for the slope parameter is always for the true slope parameter β_1, and it is a range we can be 95% confident the true slope parameter sits between. We don't need a confidence interval for the estimated slope parameter $\hat{\beta}_1$: we already have this from our sample.
Since the confidence interval is fully below zero, the interval provides evidence that as the predictor variable x (high school graduation rates) increases, the y variable (crime rates) decreases. Since the mean crime rates for communities with a high school graduation rate of x is $\mu_Y = \beta_0 + \beta_1 \times x$, if β_1 is negative, then $\beta_1 \times x$ decreases as x increases, so $\mu_Y = \beta_0 + \beta_1 \times x$ decreases as x increases, and the crime rates tend to decrease as high school graduation rates increase.
If the interval contained zero, the value for β_1 indicating that there was no association between high school graduation and crime rates (zero) would be in the interval, so we would have no statistically significant evidence that crime rates change as high school graduation rates change.

h) Select the most suitable interpretation of what $\hat{\beta}_0$ and $\hat{\beta}_1$ represent in the simple linear regression model. Recall that educational attainment is expressed as the percentage of people who graduated high school by age 25 in the model.

- $\hat{\beta}_0$ is the true crime rate in a community where no one had graduated high school; $\hat{\beta}_1$ is the true change in crime rate for every 1% increase in high school graduation rates.
- $\hat{\beta}_0$ is the true high school graduation rate in a community where no crimes were committed; $\hat{\beta}_1$ is the true change in high school graduation rate for every 1% increase in crime rate.
- $\hat{\beta}_0$ is the crime rate predicted by the regression model in a community where no one had graduated high school; $\hat{\beta}_1$ is the change in crime rate predicted by the model for every 1% increase in high school graduation rates.
- $\hat{\beta}_0$ is the high school graduation rate predicted by the regression model in a community where no crimes were committed; $\hat{\beta}_1$ is the change in high school graduation rate predicted by the model for every 1% increase in crime.
- $\hat{\beta}_0$ is the change in crime rate predicted by the regression model for every 1% increase in high school graduation rates; $\hat{\beta}_1$ is the crime rate predicted by the model in a community where no one had graduated high school.

Solution: Any simple linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times x$ is an estimate of the true relationship $\mu_Y = \beta_0 + \beta_1 \times x$ between a response variable y and the predictor variable x. So $\hat{\beta}_0$ and $\hat{\beta}_1$ are predicted values, rather than true values. In our example, high school graduation rates (x) is the predictor variable, while crime rates (y) is the response variable. For a community where no one has graduated high school by age 25, $x = 0$, so $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times 0 = \hat{\beta}_0$. Therefore, $\hat{\beta}_0$ is the crime rate in a community where no one graduated high school that's predicted by the model. (Note: in practice we shouldn't use a model to predict values of the repsonse variable outside the range of values of the predictor variables that featured in the sample. So for example, unless we really did have a good number of communities in the sample where around 0% graduate high school, we wouldn't predict the crime rate in a community where no one graduated high school with the model.)

In terms of what $\hat{\beta}_1$ represents, the crime rate predicted by the model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times x$, so every time x increases by 1, the predicted response $\hat{y}$ changes by

$$\big(\hat{\beta}_0 + \hat{\beta}_1(x+1)\big) - \big(\hat{\beta}_0 + \hat{\beta}_1 \times x\big) = (\hat{\beta}_1(x+1)) - (\hat{\beta}_1 \times x) = \hat{\beta}_1$$

so $\hat{\beta}_1$ gives the predicted change in crime rates for every unit increase in high school graduation rates.

Note that $\hat{\beta}_0$ is the predicted value of the response variable (in this example, crime rates) when the predictor variable (high school graduation rates) is zero, and not the predicted value of the predictor variable (high school graduation rates) when the response variable (crime rates) is zero. So the options mentioning this are incorrect.
Also note that $\hat{\beta}_1$ is the change in the response variable (crime rates) per unit increase in the predictor variable (high school graduation rates), and not the change in predictor variable (high school graduation rates) per unit increase in the response variable (crime rates). So the options mentioning this are wrong, incorrect.

i) The criminologists computed the correlation coefficient between high school graduation rates and crime to be -0.388. Based off this, they recommended that investing in improving educational attainment would be a smart strategy to reduce crime. Select the most appropriate comment on their recommendation.

1

- Their analysis is valid. The correlation coefficient implies that crime rates tend to decrease as educational attainment increases, which means that improving educational attainment is likely to reduce crime rates.
- Their recommendations should be ignored. They must have made a mistake in their calculation because the correlation coefficient cannot be negative.
- The correlation coefficient of -0.388 indicates that crime rates may tend to decrease as educational attainment increases, but this does not imply that improving educational attainment will cause a reduction in crime.
- Their analysis is invalid. Since the correlation coefficient is less than 0.5, they don't have statistically significant evidence that there is a correlation between educational attainment and crime rates.
- Their analysis is invalid. The coefficient -0.388 describes the proportion of variation in crime rates that is described by educational attainment, and not the degree to which there is an association between educational attainment and crime rates.

Solution: The correlation coefficient describes the extent to which there is a linear relationship between the response (crime rates) and predictor variable (high school graduation rates), and whether both increase together or one decreases as the other increases. The strength of linear relationship between the variables is determined by how close the correlation coefficient is to 1 or -1 (or how far the coefficient is from zero), while whether both variables increase together is indicated by whether the coefficient is positive or negative. If it is negative as the coefficient of -0.388 is here, this indicates that one decreases as the other increases, or crime rates tend to decrease as high school graduation rates increase.
This does not imply that higher high school graduation rates **causes** lower crime, however. Correlation does not imply causation. The correlation coefficient only tells us how variables are correlated with each other, rather than whether one has a causal influence on the other. So, the most appropriate comment on their recommendation is that, even though the correlation coefficient seems to imply that crime tends to decrease as educational attainment improves, it does not follow that improving educational attainment will cause a change in crime rates. For example, ice cream consumption is nicely correlated with number of drownings because people eat more ice cream and swim more on hot days, but this does not mean increasing ice cream consumption will cause more drownings.

j) Suppose that another group of criminologists were also interested in the extent to which high school graduation rates predicted crime rates within communities, but they sought to measure the crime rate by whether the community had a ram raid within the last year. What type of regression analysis should they have performed to explore this?

1

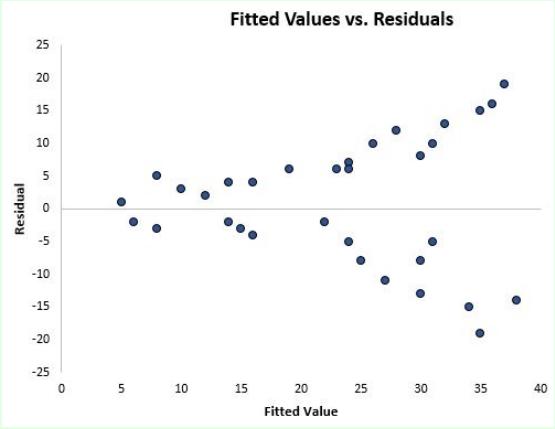
- Simple linear regression with "had ram raid" (yes=1/no=0) as the response variable and high school graduation rate as the explanatory variable.
- Logistic regression with 'had ram raid" (yes=1/no=0) as the response variable and high school graduation rate as the explanatory variable.
- Multiple linear regression with "had ram raid" (yes=1/no=0) as the response variable and high school graduation rate as the explanatory variable.
- ANOVA with high school graduation rate as the response variable and "had ram raid" (yes=1/no=0) as the predictor variable.
- Logistic regression with high school graduation rate as the response variable and "had ram raid" (yes=1/no=0) as the explanatory variable.

Solution: Logistic regression is required when the response variable is categorical, rather than continuous. Whether the community had a ram raid within the last year is a binary categorical variable (a community either had one or they didn't), so logistic regression is required.

"Had rain fall" is the response rather than the explanatory variable here, because the researchers are trying to predict it based off high school graduation rates.

Question 19

Cyanobacteria, also known as "toxic algae", is a bacteria found in New Zealand waterways that can destroy marine life.
A group of ecologists are interested in exploring whether nitrogen levels could influence the growth of cyanobacteria.
They collect data of nitrogen and cyanobacteria levels from waterways around New Zealand, and run a simple linear regression model on the data with nitrogen levels as the predictor variable and cyanobacteria as the response. The residuals plot comparing their fitted values with their raw residuals is displayed below.



a) The validity of a linear regression model relies on whether four assumptions are met.
Does the residuals plot above indicate that any of these were violated? If so, which assumption(s) appear to have been violated?

1

- Normality of residuals.
- Equality of variances.
- Linearity.
- Both equality of variances and linearity appear to have been violated.
- None of the assumptions appear to have been violated.

Solution: The two model assumptions that can be tested using a residuals versus fitted values plot (such as the one above) are linearity and equality of variances. The normality of residuals assumption is best checked using a Q-Q plot (see the lectures), while the independence assumption is tested by thinking about study design (note: mathematical methods for testing independence do exist, but are well beyond the scope of the course).
Homoscedasticity, or equality of variances: This is assessed by examining whether the residuals are equally spread out across the different values for the fitted value. In the above image, it appears the residuals are more spread out from one another as the fitted values increase, indicating homoscedasticity (equality of variances) is violated.
Linearity: The linearity assumption is tested by examining whether the residuals appear to be centered around zero across all fitted values. Even though the residuals appear to be further away from zero as the fitted values increase, they are still centered around zero for each fitted value. So linearity does not appear to be violated.

Therefore, it appears from the plot that only equality of variances is violated.

Question 20

Some food scientists are interested in the variables which predict people's perception of the quality of red wine. Knowing this could help them to create higher quality wine.
For 1599 batches of wine, they measured the citric acid, chloride, pH, density, alcohol, and sulphate content. They sought to examine whether these six variables could predict the perceived quality of the wine. The perceived quality of each batch of wine, as measured on a scale from 0 (very bad) to 10 (excellent), was assessed by at least 3 taste testers.
They ran a multiple linear regression model in R to explore this, resulting in the output given below.

```
> summary(lm(quality~citric.acid+chlorides+pH++density+alcohol+sulphates))

Call:
lm(formula = quality ~ citric.acid + chlorides + pH + +density +
    alcohol + sulphates)

Residuals:
     Min       1Q   Median       3Q      Max
-2.53574 -0.36168 -0.09678  0.49107  1.95895

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.79312   11.88601   2.086 0.037146 *
citric.acid   0.45702    0.11909   3.838 0.000129 ***
chlorides    -2.79172    0.40668  -6.865 9.52e-12 ***
pH           -0.50140    0.13774  -3.640 0.000281 ***
density     -21.33445   11.81275  -1.806 0.071099 .
alcohol       0.30439    0.02089  14.570  < 2e-16 ***
sulphates     1.08759    0.11278   9.644  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6726 on 1592 degrees of freedom
Multiple R-squared:  0.3088,    Adjusted R-squared:  0.3062
F-statistic: 118.6 on 6 and 1592 DF,  p-value: < 2.2e-16
```

a) Which is the response variable in the model?

1

- (•) The perceived wine quality.
- ( ) The intercept.
- ( ) The citric acid content.
- ( ) Red wine.
- ( ) The sulphate content.

Solution: The food scientists are examining whether the citric acid, chloride, pH, density, alcohol, and sulphate content of wine could be used to predict its perceived wine quality, so perceived quality is the response variable.

b) Based off this model, predict the perceived quality of a red wine sample with a citric acid content of 0.02, a chloride content of 0.075, a pH of 3.4, a density of 0.99, an alcohol content of 13.0, and a sulphate content of 0.7 units.

1

- (•) 6.485
- ( ) −18.308
- ( ) 7.136
- ( ) 6.043
- ( ) 6.708

Solution: Let $x_1,\ x_2,\ x_3,\ x_4,\ x_5,$ and x_6 refer to the predictor variables listed in the order displayed in the regression output table (citric acid, chloride, pH, density, alcohol, sulphate), and $\hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\beta}_3,\ \hat{\beta}_4,\ \hat{\beta}_5,$ and $\hat{\beta}_6$ respectively refer to the estimated slope coefficients for the predictor variables in the fitted model listed in that order. Then the wine quality predicted by the fitted model for some specific values of the predictor variables $x_1, \ldots, x_6$ is

$$\hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 \times x_1) + (\hat{\beta}_2 \times x_2) + (\hat{\beta}_3 \times x_3) + (\hat{\beta}_4 \times x_4) + (\hat{\beta}_5 \times x_5) + (\hat{\beta}_6 \times x_6)$$

This is also the predicted mean wine quality for batches of wine for those specific values of $x_1, \ldots, x_6$.
To get the predicted wine quality for the values of the variables given in the question, we need to find the values for the slope coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_6,$ and plug in the values for the $x_1, \ldots, x_6$ variables.
The estimates for the slope coefficients is given in the "Estimates" column of the regression summary table from R as $\hat{\beta}_1 = 0.45702,\ \hat{\beta}_2 = -2.79172,\ \hat{\beta}_3 = -0.5014,$ $\hat{\beta}_4 = -21.33445,\ \hat{\beta}_5 = 0.30439,$ and $\hat{\beta}_6 = 1.08759$ (see the guide below this solution for the outputs from R regression summaries that you'll need to be able to interpret in the final exam, and how to gather these from the R regression summaries you're given). So the predicted wine quality $\hat{y}$ is

$$\begin{aligned}\hat{y} =& \hat{\beta}_0 + (\hat{\beta}_1 \times x_1) + (\hat{\beta}_2 \times x_2) + (\hat{\beta}_3 \times x_3) + (\hat{\beta}_4 \times x_4) + (\hat{\beta}_5 \times x_5) + (\hat{\beta}_6 \times x_6) \\ =& 24.79312 + (0.45702 \times 0.02) + (-2.79172 \times 0.075) + (-0.5014 \times 3.4) \\ & + (-21.33445 \times 0.99) + (0.30439 \times 13) + (1.08759 \times 0.7) \\ =& 6.485\end{aligned}$$

General guide to how to read the outputs from R regression summaries that you'll need to be able to read in the final exam:
The rows of the regression table: Each row corresponds to the independent variables in the model, with the exception of the top "(Intercept)" row, which gives us information on the intercept ($\hat{\beta}_0$) of the fitted model. In this example, there are six independent variables (citric acid, chloride, pH, density, alcohol, and sulphate content).
The estimates for the intercept and slope parameters: These can be read from the "Estimate" column of the output table. The value in the "(Intercept)" row is the value for $\hat{\beta}_0,$ while the estimated slope parameters for each independent variable is in the row corresponding to that variable. For example, the slope parameter for the citric acid variable is $0.45702.$
The estimated standard errors: The estimated standard errors for the intercept and slope parameters can be found in the "Std. Error" column. For example, the estimated standard error for the slope parameter β_1 for the citric acid variable is 0.11909 in this example. This describes how much we can expect the values for $\hat{\beta}_1$ to vary between different samples.
t value column: For the intercept and each independent variable in the model, this column gives the value of the test statistic for the hypothesis test that the parameter corresponding to that variable is zero. For example, the value in the "t value" column and "citric acid" row gives the test statistic in the hypothesis test that the slope parameter $\hat{\beta}_1$ for the citric acid variable is zero.
Pr(>|t|) column: For the intercept and each independent variable in the model, this column gives the *p*-value for the hypothesis test that the parameter corresponding to the intercept or that variable is zero. For example, the value in the "Pr(>|t|)" column and "citric acid" row gives the *p*-value in the hypothesis test that the slope parameter $\hat{\beta}_1$ for the citric acid variable is zero.
Residual Standard Error: As the name suggests, this gives the residual standard error of the sample data around the fitted model.
Multiple R-squared: This is the R^2 value, or the estimated proportion of the variation in the dependent variable (the wine quality, in our example) that can be explained by the independent variables (the citric acid, chloride, pH, density, alcohol, and sulphate content, in this example).

c) Letting β_2 be the slope parameter for the chlorides variable, what does β_2 represent?

1

- β_2 is the true mean change in chloride levels for every unit change in perceived wine quality.
- β_2 is the estimated mean change in perceived wine quality for every unit change in chloride levels, based off this sample.
- β_2 is the estimated mean change in perceived wine quality for every unit change in chloride levels based off this sample, having adjusted for citric acid, pH, density, alcohol, and sulphate content.
- β_2 is the true mean change in perceived wine quality for every unit change in chloride levels.
- (selected) β_2 is the true mean change in perceived wine quality for every unit change in chloride levels, having adjusted for citric acid, pH, density, alcohol, and sulphate content.

Solution: The assumed model for how perceived wine quality relates to the six predictor variables used in the model is

$$\mu_Y = \beta_0 + (\beta_1 \times x_1) + (\beta_2 \times x_2) + (\beta_3 \times x_3) + (\beta_4 \times x_4) + (\beta_5 \times x_5) + (\beta_6 \times x_6)$$

This gives the mean perceived wine quality (μ_Y) for batches of wine with some common values of the predictor variables $x_1, \ldots, x_6$ (citric acid, chloride, pH, density, alcohol, and sulphate content). For example, it might give the mean perceived wine quality across batches of wine where the citric acid content of the wine is 0.02, the $\text{chlorides} = 0.075$, the $\text{pH} = 3.4$, the $\text{density} = 0.99$, the $\text{alchohol} = 13.0$, and the sulphate content is 0.7.
With β_2 being the slope parameter for the chlorides variables (as assumed in the question), β_2 will give the increase in mean perceived wine quality every time the chloride levels increase by 1, **assuming all the other predictor variables remain fixed**, or **having adjusted for the other predictor variables**. It does not give the mean increase in perceived wine quality for every increase in chloride levels by 1 **in general** (i.e. not fixing the values for the other predictor variables), as if there is an association between chloride levels and the other predictor variables, every unit increase in chloride levels in general will be accompanied by some changes in the other predictor variables, resulting in the other predictor variables besides chlorides changing μ_Y too.

β_2 is the true slope parameter assumed in the model, not the predicted slope parameter based off the sample (this would be $\hat{\beta}_2$, not β_2). We assume in running the multiple regression model that the true mean perceived wine quality μ_Y follows the relationship

$$\mu_Y = \beta_0 + (\beta_1 \times x_1) + (\beta_2 \times x_2) + (\beta_3 \times x_3) + (\beta_4 \times x_4) + (\beta_5 \times x_5) + (\beta_6 \times x_6)$$

We estimate the true values for $\beta_0, \beta_1, \ldots, \beta_6$ based off our sample. These estimates we denote $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_6$, so our **predicted** mean wine quality (and also our estimate for the perceived wine quality $\hat{y}$) based off our sample is

$$\hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 \times x_1) + (\hat{\beta}_2 \times x_2) + (\hat{\beta}_3 \times x_3) + (\hat{\beta}_4 \times x_4) + (\hat{\beta}_5 \times x_5) + (\hat{\beta}_6 \times x_6)$$

d) Letting β_4 be the slope parameter for the density variable, select the appropriate option below for how to calculate the 95% confidence interval for β_4.

1

- $-21.33445 \pm t_{0.975,1597} \times 11.81275$

- $-21.33445 \pm t_{0.975,1597} \times \dfrac{11.81275}{\sqrt{1599}}$
- $-21.33445 \pm t_{0.975,1592} \times 11.81275$
- $-21.33445 \pm t_{0.975,1592} \times \dfrac{11.81275}{\sqrt{1599}}$
- $-21.33445 \pm t_{0.975,1598} \times 11.81275$

Solution: Confidence intervals for slope parameters in linear regression are calculated using

$$\text{Estimate} \pm \text{Multiplier} \times \text{Standard Error}$$

Estimate: This is the slope parameter in the fitted model, $\hat{\beta}_4$. This is the entry in the "Estimate" column and "density" row of the output table from R (21.33445).
Multiplier: For linear regression, this is $t_{1-\alpha/2,n-k-1}$, where α is the tail probability dependent on the level of confidence of the confidence interval, n is the sample size, and k is the number of independent variables in the model. For a 95% CI, $\alpha = 1 - .95 = 0.05$. There were 1599 batches of wine in the sample, so $n = 1599$. And $k = 6$, as there are 6 independent variables in the regression analysis. So

$$t_{1-\alpha/2,n-k-1} = t_{1-0.05/2,1599-6-1} = t_{0.975,1592}$$

Standard Error: This is the estimate for how much we could expect the slope parameter in the sample to vary between samples (for each sample of 1599 batches of wine, we would get different values for $\hat{\beta}_4$). This can be read from the regression summary from R as the number in the "Std. Error" column and "density" row (11.81275).

e) The 95% confidence interval for the slope parameter β_4 is $(-44.49, 1.82)$. Select the appropriate interpretation of the interval.

1

- There is no statistically significant evidence at the 5% significance level that the perceived quality of wine is associated with its density.
- There is no statistically significant evidence at the 5% significance level that the perceived quality of wine is associated with its density, having adjusted for its citric acid content, chloride levels, pH, alcohol, and its sulphate content.
- We have statistically significant evidence at the 5% significance level that the perceived quality of wine tends to decrease as its density increases.
- We have statistically significant evidence at the 5% significance level that the perceived quality of wine tends to decrease as its density increases, having adjusted for its citric acid content, chloride levels, pH, alcohol, and sulphate content.
- We have statistically significant evidence at the 5% significance level that the perceived quality of wine tends to increase as its density increases.

Solution: β_4, or the slope parameter for the density variable, gives the unit change in mean perceived wine quality for each change of density of 1, assuming all the other predictor variables remain fixed. This can be easily seen given the relationship between mean perceived wine quality μ_Y and the six predictor variables is

$$\mu_Y = \beta_0 + (\beta_1 \times x_1) + (\beta_2 \times x_2) + (\beta_3 \times x_3) + (\beta_4 \times x_4) + (\beta_5 \times x_5) + (\beta_6 \times x_6)$$

It does not give the mean perceived change in wine quality for each change of density of 1 **in general**, as if the density of wine is associated with the other predictor variables then each unit change in density in general will be accompanied by changes in the other predictor variables, meaning the other predictor variables may alter the perceived wine quality as well in conjunction

with the unit change in density.
If β_4 is not zero then the mean wine quality changes with its density having adjusted for the other predictor variables (if $\beta_4 < 0$ it decreases with density; if $\beta_4 > 0$ it increases with density), meaning there is an association between perceived wine quality and density having adjusted for the other predictor variables. If $\beta_4 = 0$ then there is no association between wine quality and density having adjusted for the other predictor variables (i.e. wine quality does not tend to change with density).
The 95% confidence interval given includes zero, so the value indicating no association is in the plausible range of values for β_4. We therefore have no evidence at the $1 - 0.95 = 0.05$ significance level that wine quality is associated with density, having adjusted for the other predictor variables. (Note: A 95% CI gives or doesn't give statistically evidence at the $1 - 0.95 = 0.05 = 5\%$ significance level.)

f) The *p*-value for the hypothesis test that the slope parameter for the citric acid variable is zero is 0.0001, when rounded to 4DP. Select the appropriate interpretation of this.

1

- We have strong evidence that the perceived quality of wine tends to improve as citric acid content increases.
- **We have strong evidence that the perceived quality of wine tends to improve as citric acid content increases, having adjusted for chloride levels, pH, alcohol, the density, and sulphate content.**
- We have strong evidence that the perceived quality of wine tends to decrease as citric acid content increases, having adjusted for chloride levels, pH, alcohol, density, and sulphate content.
- We have no evidence that the perceived quality of wine is associated with citric acid levels.
- We have no evidence that the perceived quality of wine is associated with citric acid levels, having adjusted for chloride levels, pH, alcohol, density, and sulphate content.

Solution: Let β_1 be the slope parameter for the citric acid variable. β_1 gives the mean change in wine quality for every unit change in citric acid levels when keeping the other predictor variables fixed (i.e. having adjusted for the other predictor variables).
Since the *p*-value is less than 0.01, we have strong evidence to reject H_0 (if the *p*-value was between 0.01 and 0.05, we would have **some** evidence to reject H_0. Recall the hypothesis testing lecture that discussed that, in this course, we consider *p*-values between 0.01 and 0.05 to constitute some evidence against H_0, and *p*-values less than 0.01 to constitute strong evidence against H_0). In a hypothesis test for whether $\beta_1 = 0$, H_0 is the statement that $\beta_1 = 0$ (while H_A is that $\beta_1 \neq 0$). So we have strong evidence to reject the hypothesis that $\beta_1 = 0$ based off our *p*-value. $\beta_1 \neq 0$ means there is an association between perceived wine quality and citric acid content having adjusted for the other predictor variables, with perceived wine quality tending to improve with citric acid levels if $\beta_1 > 0$, and tending to get worse with higher citric acid levels if $\beta_1 < 0$. Since the sample estimate for β_1 is greater than zero ($\hat{\beta}_1 = 0.45702$. This can be read from entry in the "Estimate" column and "citric.acid" row of the regression output), we have evidence that perceived wine quality tends to increase as citric acid levels increase, having adjusted for the other predictor variables.

g) The R^2 value in their analysis is $R^2 = 0.3088$. Select the appropriate interpretation of this.

1

- The correlation coefficient between the predictor variables of the citric acid, chloride, pH, density, alcohol, and sulphate content of wine, and the response variable of the perceived wine quality, is 0.3088%.
- The correlation coefficient between the predictor variables of the citric acid, chloride, pH, density, alcohol, and sulphate content of wine, and the response variable of the perceived

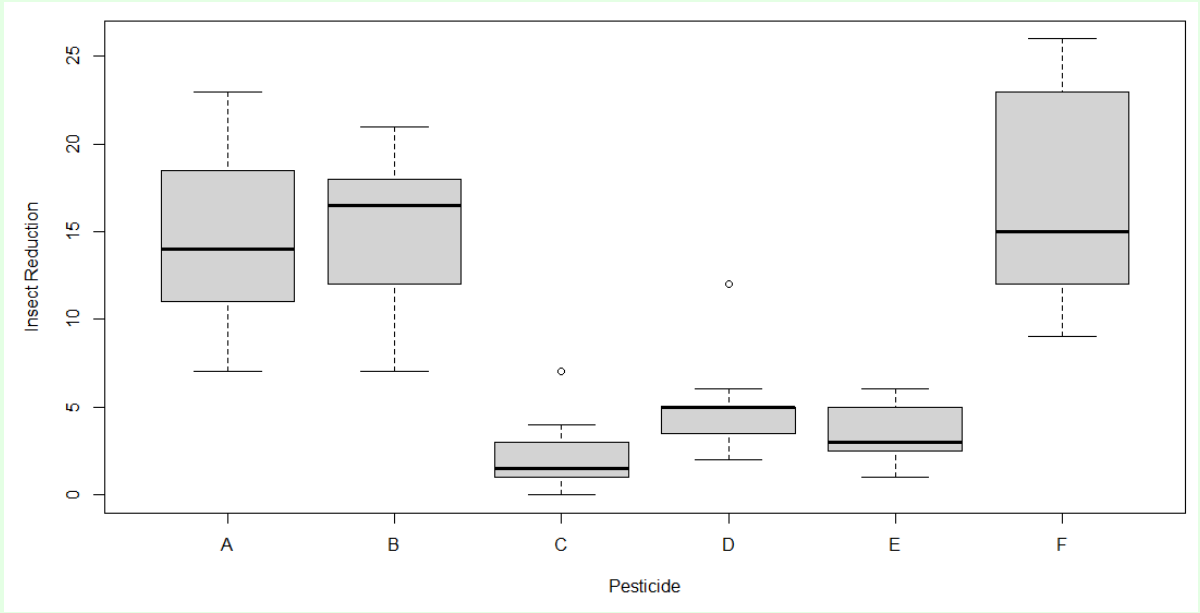
wine quality, is 30.88%.

- (●) According to the model, the citric acid, chloride, pH, density, alcohol, and sulphate content of wine describes 30.88% of the variation in perceived wine quality.
- ( ) Since the R^2 value is greater than 0.05, the data provide no statistically significant evidence that the citric acid, chloride, pH, density, alcohol, and sulphate content predict perceived wine quality.
- ( ) According to the model, 30.88% of the variation in perceived wine quality can be predicted. The rest is purely due to random variation.

Solution: The R^2 value in a regression model gives the proportion of the variation in the response variable that can be explained by the predictor variables in the model. The predictor variables in this model are citric acid, chloride, pH, density, alcohol, and sulphate content, while the response variable is perceived wine quality, so the variables citric acid, chloride, pH, density, alcohol, and sulphate content describe 30.88% of the variation in perceived wine quality (according to the model).

Question 21

A study is conducted to test the effectiveness of 6 different pesticides (labelled 'A' through 'F') at reducing insect numbers on cropped farmland. 72 small segments of various crop fields were portioned off. Each of the 72 segments were allocated one pesticide, with each of the six pesticides being allocated to 12 fields. For each segment, the number of insects on the segment was recorded before the pesticide was applied, and after one month of the pesticide being applied daily.
A box plot for the reduction in the number of insects between before and after the pesticide was applied for each of the pesticides is displayed below.



The data were analysed in R, producing the output below.

```
> anova(lm(reduction~pesticide))
Analysis of Variance Table

Response: reduction
          Df Sum Sq Mean Sq F value    Pr(>F)
pesticide  5 2668.8  533.77  34.702 < 2.2e-16 ***
Residuals 66 1015.2   15.38
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a) The ANOVA model used here can be written as

$$Y_{ij} = \mu_i + e_{ij}$$

where Y_{ij} represents the random variable for the reduction in the number of insects in the j^{th} field segment that was applied the i^{th} pesticide. Select what μ_i and e_{ij} represent in the above equation from the options below.

1

- μ_i is the mean number of insects on segments that were applied pesticide i before the pesticide was applied in our sample; e_{ij} is the mean reduction in the number of insects in field segments that were applied pesticide i in our sample.

- μ_i is the mean reduction in the number of insects in field segments that were applied pesticide i in our sample; e_{ij} is the 'residual', or difference between the reduction in the number of insects observed in the j^{th} field segment that was applied pesticide i, and μ_i.
- μ_i is the population-level mean reduction in the number of insects on field segments that are applied pesticide i; e_{ij} is the 'residual', or difference between the reduction in the number of insects observed in the j^{th} field segment that was applied pesticide i, and μ_i.
- μ_i is the mean reduction in the number of insects in field segments that were applied pesticide i in our sample; e_{ij} is the standard deviation of the j^{th} observation in group i.
- μ_i is the population-level mean reduction in number of insects on field segments that are applied pesticide i; e_{ij} is the standard deviation of the j^{th} observation in group i.

Solution: The generic form of the one-way ANOVA model (the ANOVA method taught in this course) is that $Y_{ij} = \mu_i + e_{ij}$, where Y_{ij} is the j^{th} `response' in the $i^{th}$ `group', μ_i is the *population* (not the sample) mean response across group i, and e_{ij} is the `error term' or `residual' or difference between μ_i (the mean response across all of group i) and Y_{ij} - the j^{th} response in group i.
Here, the `groups' are the pesticide applied, the group means are the mean reductions in the number of insects on field segments between before and after that pesticide is applied to them, while each of the individual responses refer to the reduction in number of insects on an individual field segment between before and after a particular pesticide was applied.
Therefore, in this example, Y_{ij} refers to the reduction in the number of insects on field segment j that was applied pesticide i; μ_i is the population-level mean reduction in the number of insects on field segments that are applied pesticide i (Notes: There could be a few reasonable ways to define our `population' here. The key thing to note is that μ_i doesn't refer to the mean response across our sample, but across whatever our population is defined to be. Also, to be technically more accurate, the mean response here would be the reduction in insect levels between before the pesticide was applied, and after one month of the pesticide being applied daily.); and e_{ij} is the 'residual', or the difference between the reduction in insect numbers observed in the j^{th} field segment that was applied pesticide i, and μ_i.

b) Which of the following statements regarding the e_{ij} term in the ANOVA model is correct?

1

- The e_{ij} values for each group i are assumed to normally distributed with a zero mean and constant variance.
- The e_{ij} values for each group i are assumed to follow a t-distribution.
- The e_{ij} values are only approximately normal if the sample sizes for each group of observations are the same.
- The e_{ij} terms do not need to follow any particular distribution, but the overall sample size must be large enough such that the e_{ij} terms follow a t-distribution.
- The e_{ij} terms do not need to follow any particular distribution, but the sample size is large enough that the e_{ij} terms follows a normal distribution.
- The e_{ij} terms do not need to follow any particular distribution, but the sample size is large enough that the sampling distribution for the mean of the e_{ij} terms follows a t-distribution.

Solution: The key assumption is that the error terms are normally distributed within each group with zero mean and have a common variance (see lecture slides).
The other ANOVA assumption is that the observations are independent.
The normality and constant variance assumptions in ANOVA are analogous to the normality and homoscedasticity (equality of variances) assumptions for the error terms in regression.

c) Select the hypothesis that is being tested by the F-statistic in the ANOVA table above.

1

- $H_0: \mu_A = \mu_B = \mu_C = \mu_D = \mu_E = \mu_F$; H_A : All the means are different to each other.
- H_0 : All the means are different to each other; H_A : All the means are the same.
- (selected) $H_0: \mu_A = \mu_B = \mu_C = \mu_D = \mu_E = \mu_F$; $H_A: \mu_A, \mu_B, \mu_C, \mu_D, \mu_E, \mu_F$ are not all equal.
- $H_0: \hat{\mu}_A = \hat{\mu}_B = \hat{\mu}_C = \hat{\mu}_D = \hat{\mu}_E = \hat{\mu}_F$; $H_A: \hat{\mu}_A, \hat{\mu}_B, \hat{\mu}_C, \hat{\mu}_D, \hat{\mu}_E, \hat{\mu}_F$ are not all equal.
- H_0 : The sample means are all equal to the population means; H_A : The sample means are not all equal to the population means.

Solution: The one-way ANOVA test is testing whether some of the population means of the groups being studied differ. The null hypothesis is that the population group means are all the same, while the alternative hypothesis is the opposite of this - that some of them differ. Note:
- Some of the means differing (i.e $H_A: \mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$ are not all equal) is a different statement to all of the means differing (i.e. $\mu_1 \neq \mu_2$, $\mu_1 \neq \mu_3$, ..., $\mu_2 \neq \mu_3$, ...).
- $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_4$, $\hat{\mu}_5$, $\hat{\mu}_6$ refer to the sample means across each group. The null and alternative hypotheses are around the population-level means for each group ($\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$).

d) The F-statistic in the ANOVA analysis is based off various sums of squares (SS). Select the correct statement below regarding the sums of squares in this analysis.

1

- The total SS is 2668.8; the residual SS is 1015.2.
- (selected) The total SS is 3684.0; the residual SS is 1015.2.
- The group SS is 1015.2. The residual SS is 2668.8.
- The group SS is 3684.0; the residual SS is 1015.2.
- The total SS is 3684.0; the group SS is 1015.2.

Solution: The group and residual SS are given in the table. The residual SS is, as the output summary would suggest, the entry in the "Residuals" row and "Sum Sq" column of the output (1015.2). The group SS is the SS for whatever the group variable is in our ANOVA. Our group variable in this ANOVA is the pesticide used, so the group SS is the entry in the "pesticide" row and "Sum Sq" column of the output (2668.2). Therefore, Total SS=Group SS+Residual SS= $2668.8 + 1015.2 = 3684.0$

e) Select the circumstance under which we would be *most* likely to reject the null hypothesis in a one-way ANOVA analysis.

1

- When the variability between the group means is small, and the variability within the study groups is small.
- When the variability between the group means is small, and the variability within the groups is large.
- When the variability between the group means is large, and the variability within the groups is large.
- (selected) When the variability between the group means is large, and the variability within the groups is small.
- When the overall variability across all of the group observations is large.

Solution: **Conceptual explanation:** The ANOVA analysis is exploring the evidence for H_A (that the population means of some of the groups differ), based off our given data. It does this by looking at the F-statistic, which compares the variability observed *between* the groups in the

data (the GSS (group sum of squares) and GMS (group mean square)) with the variability observed *within* the groups in the data (the RSS (residual sum of squares) and GMS (group mean square)).
(OPT A) Differences observed between group means in the sample could be due to there being true differences between the population-level group means, or due to random variation across individuals (fields, in this case) causing our sample group means to differ from the population group means. Understandably, larger differences between sample group means constitute more evidence against the null (that there is no difference between population group means). But greater variability observed *within* groups makes it more plausible that the population means don't actually differ (i.e. that H_0 is true), but the observed differences between sample means were just due to random variation. We are therefore most likely to reject H_0 and conclude there is evidence of difference between population group means when we observe *larger* variability *between* groups and *smaller* variability *within* groups in our sample.

Alternate explanation: The F-statistic in an ANOVA analysis is

$$F = \frac{GSS/(K-1)}{RSS/(n-K)}$$

where $GSS/(K-1)$ is a gauge of the variability between the group means, and $RSS/(n-K)$ is (kind of) the sum of the variances within the groups. Therefore, the larger the variability between group means (i.e. the larger $GSS/(K-1)$ is), and the smaller the variability within group means (i.e. the smaller $RSS/(n-K)$ is), the larger the F-statistic. By observing the F-distribution, the larger the F-statistic, the lower the p-value, and the more likely we are to reject the null hypothesis that there is no difference between the group means.

f) Letting **f** represent the F-statistic calculated in this ANOVA, select the appropriate R command to calculate the p-value in the ANOVA table printed above.

1

- **pf(f,71,lower.tail=FALSE)**
- **1-pf(f,71,lower.tail=TRUE)**
- **pf(f,5,66,lower.tail=FALSE)**
- **1-pf(f,5,66,lower.tail=FALSE)**
- **2*(pf(f,5,66,lower.tail=FALSE)**

Solution: Under the null hypothesis that there are no differences between any of the population group means, the probability distribution for observing each F-statistic

$$F = \frac{GSS/(K-1)}{RSS/(n-K)}$$

is described by the F-distribution with $K-1$ and $n-K$ degrees of freedom, where K is the number of groups whose means are being compared, and n is the total number of individual observations across all of the groups in the study sample.
$K = 6$ and $n = 72$ in this study, since the insect counts across 72 individual field segments that were applied one of 6 different pesticides were tested in the study. Therefore, under the null hypothesis, the F-statistic follows the F-distribution with $K-1 = 5$ and $n-K = 72-6 = 66$ degrees of freedom.
The p-value for an F-statisitc f is calculated by finding the area in the upper tail above f on the F-distribution curve with $K-1$ and $n-K$ degrees of freedom: similarly to z-statistics and t-statistics, the `most likely' f-statistics under the null hypothesis are ones closest to zero; however, unlike z-statistics and t-statistics, the F-distribution can only take positive values, so we therefore just find the area in the upper tail above f. The appropriate command is therefore **pf(f,5,66,lower.tail=False)**.

g) Using $\alpha = 0.01$ as the significance level, the most appropriate conclusion to draw from the *p*-value in the ANOVA table above is:

1

- Since our *p*-value is greater than $\alpha = 0.01$, we do not have statistically significant evidence to reject H_0, and we do not have statistically significant evidence that the mean reduction in the number of insects differs between some of the pesticide types.
- Since our *p*-value is less than $\alpha = 0.01$, we have statistically significant evidence at the 0.01 significance level to reject H_0. Since the observed sample means (going from smallest to largest) were for pesticide C, E, D, A, F, and B, we have evidence that B is the most effective pesticide at reducing insect numbers, followed by F, A, D, E, then C.
- **Since our *p*-value is less than $\alpha = 0.01$, we have statistically significant evidence at the 0.01 significance level to reject H_0 and conclude that the mean reduction in number of insects differs between some of the pesticide types.**
- Since our *p*-value is less than $\alpha = 0.01$, we do not have statistically significant evidence at the 0.01 significance level to reject H_0, and we do not have statistically significant evidence that the mean reduction in number of insects differs between pesticide types.
- Since our *p*-value is less than $\alpha = 0.01$, we have statistically significant evidence at the 0.01 significance level to accept H_0 and conclude that the mean reduction in number of insects is the same between the different pesticide types.

Solution: The *p*-value can be found by viewing the value under the `Pr(>F)' heading in the ANOVA table output from R. In this case, the *p*-value in the table is $2.2\text{e}-16$. This means the *p*-value is 2.2×10^{-16} (or 0.00000000000000022).
This is significantly less than $\alpha = 0.01$, meaning we have statistically significant evidence that the null hypothesis is false, and some of the group means differ (the mean reduction in number of insects is not the same across all pesticides).
Note that rejecting H_0 in an ANOVA does not give us evidence for the order which the means are greater in (e.g. it does not allow us to make a statment like "we have evidence that B is the most effective pesticide at reducing insect numbers, followed by F, A, D, E, then C".). This is in contrast to hypothesis tests for a difference between 2 means or proportions, where we have evidence that the population mean/proportion with the highest sample mean/proportion is greater. Rejecting H_0 in an ANOVA just provides evidence that some of the means differ, but not the order which they differ in.

h) It turns out that the field segments were in four different countries, with three field segments from each country being used to test each pesticide. Suppose that the insects from the four different countries have very different resiliencies to pesticides. The following statements concern the possibility of including "country" as an additional variable in the analysis. Which of them is correct?

1

- Including country as a blocking variable in the analysis is likely to soak up some of the total variation across the entire sample, helping to detect differences that exist between the effectiveness of the pesticides.
- **Including country as a blocking variable in the analysis is likely to soak up some of the residual variation, helping to detect differences that exist between the effectiveness of the pesticides.**
- Including country as a blocking variable in the analysis is likely to soak up some of the residual variation, making it harder to detect differences that exist between the effectiveness of the pesticides.
- Including country as a blocking variable in the analysis is likely to soak up some of the variation between groups, helping to detect differences that exist between the effectiveness of the pesticides.

- Including country as a blocking variable in the analysis is likely to soak up some of the variation between groups, making it harder to detect differences that exist between the effectiveness of the pesticides.

Solution: This ANOVA assesses the evidence against all the pesticides being just as effective at reducing insect numbers (i.e. against the mean response for the different treatments being the same) by comparing the amount of variation explained by differences between treatment (pesticide) groups (as measured by the group sum of squares and group mean square) with the amount of variation not explained by factors in the ANOVA model (the residual sum of squares and residual mean square). The higher the amount of variation explained by differences between treatment groups relative to the amount of unexplained variation, the more evidence there is that there are differences between the effectiveness of the pesticides.
If there is a large amount of variation caused by some factor other than the treatment (pesticide) used, such as variability in the resiliencies of the insects in the different countries to pesticides, it could be that the different treatments really could have differing effects, but this is not picked up in the ANOVA because of the excess "noise" or variation caused by the other factor (such as the country the insects are in) causing the residual sum of squares to be higher relative to the group sum of squares.
Including country as a blocking variable in the analysis could explain or soak up some of this previously unexplained residual variation, making it easier to detect the true differences in the effectiveness of the different pesticides (making the variation between groups (pesticides) more apparent relative to the unexplained residual variation).

(Supplementary Information:) The "blocking variable" is the variable added in to control for some of the unexplained variation to help us observe the true effect of the variable being examined in the study. So "country" is the blocking variable in the alternate design, because it is the variable that seeks to control for some variation not explained by the effect of the different pesticides - the variable being explored in the study. (One can think that it is called the blocking variable because the study is divided into "blocks", where the observations in any single block are similar to the other observations in that block with regard to the blocking variable. For example, the observations for the six different pesticides applied to field segments in a single country would be one block in the proposed alternate design. There would be four blocks in the alternate design - one for each country).