UNIVERSITY OF OTAGO EXAMINATIONS 2023

STATISTICS

STAT110

Statistical Methods

Practice Exam

(TIME ALLOWED: 3 HOURS)

This examination paper comprises 35 pages

Candidates should answer questions as follows:

This is a multiple-choice examination with 93 questions.

Select ONE answer only from the 5 options provided for each question.

Record your answers on the machine-readable sheet provided.

All questions should be attempted.

Use the blank reverse side of the examination paper pages for rough work.

The following material is provided:

A detachable appendix giving a summary of formulae.

Use of calculators:

Any model of calculator can be used for this exam. Provided that it is battery powered, truly portable, silent and free of communication capabilities.

Candidates are permitted copies of: NIL

Other Instructions:

Hand in the entire paper and appendix attached to the MCQ answer sheet and green attendance receipt.

TURN OVER

Information for Questions 1 to 3
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

1. The mean number of students majoring in Neuroscience and taking STAT115 is closest to:

   A) 86.75 B) 108 C) 100.8 D) 24.8 E) 124

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

   A) 4.92 B) 24.20 C) 4.40 D) 6.12 E) 19.36

3. Choose one of the following which best describes the type of data in this question:

   A) Nominal categorical
   B) Binomial
   C) Continuous
   D) Binary
   E) Discrete

Information for Questions 4 to 6
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.

4. Select the type of sampling procedure used from the options below.

   A) Simple random sampling.
   B) Stratified random sampling.
   C) Cluster sampling.
   D) A census.
   E) A random variable.

5. In this study, the NZTA database of registered vehicle owners is an example of:

   A) A cluster.
   B) A random sample.
   C) A sampling frame.
   D) A stratum.
   E) A sample.

6. What is the population of interest in the study?

   A) The vehicle owners in the sample.
   B) The vehicles in the sample.
   C) Vehicle owners who were open to using public transport.
   D) All New Zealand vehicles.
   E) All New Zealand vehicle owners.

Information for Questions 7 to 9
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

7. What is the probability a student passes exactly three papers?

A) We don't know as it isn't given
B) 0.4
C) 0.04
D) 0.05
E) 0.49

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

A) 0.98 B) 1 C) 3.29 D) 1.96 E) 3.0

9. 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:

A) 18.32 B) 0.00 C) 0.91 D) 0.83 E) 1.65

Information for Questions 10 to 13
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

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

A) 82/1017 B) 48/78 C) 48/82 D) 78/1017 E) 48/1017

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

A) 82/1017 B) 48/78 C) 48/82 D) 78/1017 E) 48/1017

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

A) 48/78 B) 48/82 C) 34/78 D) 48/1017 E) 34/1017

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

A) 48/82 B) 48/78 C) 34/82 D) 34/1017 E) 34/78

Information for Questions 14 to 18

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".

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

A) 0.09 B) 0.91 C) 0.96 D) 0.04 E) 0.12

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

A) 0.09 B) 0.91 C) 0.96 D) 0.04 E) 0.12

16. 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:

A) 0.1152 B) 0.8448 C) 0.1092 D) 0.8008 E) 0.0048

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

A) 0.8556 B) 0.1444 C) 0.8556 D) 0.954 E) 0.88

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

A) 0.88 B) 0.8556 C) 0.0056 D) 0.9360 E) 0.9874

Information for Questions 19 to 21

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 $\$1000$.

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

A) Continuous.
B) Discrete.
C) Categorical ordinal.
D) Categorical nominal.
E) They are not random variables.

20. 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).

A) $22900 B) $21900 C) $20900 D) $21400 E) $28200

21. 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:

A) $2109.50
B) $4450000
C) $2830
D) $2900
E) $53.20

Information for Questions 22 to 24
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.

22. 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?

A) 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.
B) We cannot say anything about the probability without the use of software.
C) The probability will be zero because a hippo cannot weigh 3930kg if hippo weights are distributed as given above.
D) The probability will be greater than 0.5.
E) The probability will be less than 0.5.

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

A) **pnorm(q=3930,mean=3530,sd=190)**
B) **pnorm(q=3930,mean=3530,sd=190,lower.tail=TRUE)**
C) **1-pnorm(q=3930,mean=3530,sd=190)**
D) **pnorm(q=3930,mean=3530,sd= 190/$\sqrt{n}$)**
E) **1-pnorm(q=3930,mean=3530,sd= 190/$\sqrt{n}$)**

24. What is the Z-score for a male hippo that weighs 3130kg ?

A) -2.11 B) 2.11 C) 18.58 D) -18.58 E) 0.48

Information for Questions 25 to 26

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

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

25. Which of the following does Interval A represent?

A) A 95% confidence interval. We estimate the mean daily bamboo consumption for 95% of all pandas is in this range.

B) A 95% confidence interval. We can be 95% confident the mean daily bamboo consumption across all pandas is in this range.

C) A 95% reference range. We estimate the mean average daily bamboo consumption for 95% of all pandas is in this range.

D) A 95% reference range. We can be 95% confident the mean daily bamboo consumption across all pandas is in this range.

E) The interval has no discernable meaning.

26. Which of the following does Interval B represent?

A) A 95% confidence interval. We estimate the mean daily bamboo consumption for 95% of all pandas is in this range.

B) A 95% confidence interval. We can be 95% confident the mean daily bamboo consumption across all pandas is in this range.

C) A 95% reference range. We estimate the mean daily bamboo consumption for 95% of all pandas is in this range.

D) A 95% reference range. We can be 95% confident the mean daily bamboo consumption across all pandas is in this range.

E) The interval has no discernable meaning.

Information for Questions 27 to 29

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.

27. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

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

28. The probability that exactly 3 of the mountains have an avalanche is:

A) 0.32805

B) 0.0729

C) 0.0081

D) 0.00045

E) 0.59049

29. The probability that more than 3 of the mountains have an avalance is:

A) 0.00856 B) 0.00001 C) 0.0081 D) 0.00046 E) 0.99144

Information for Questions 30 to 34

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.

30. 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 qnorm(0.975) = 1.96.

A) At least 0.094 to 3 dp.
B) Exactly 0.094 to 3 dp.
C) No more than 0.094 to 3 dp.
D) At least 0.189 to 3 dp.
E) No more than 0.188 to 3 dp.

31. 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**.

A) 314 B) 213.95 C) 214 D) 264.08 E) 264

32. 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:

A) 62 people
B) 23 people
C) 0.629
D) 0.371
E) 2.696

33. 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 (Lower Bound, Upper Bound).

A) $(0.251, 0.491)$
B) $(0.310, 0.432)$
C) $(-1.589, 2.331)$
D) $(0.364, 0.378)$
E) $(0.251, 2.331)$

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

A) Using a larger sample size to allow for non-response.
B) Selecting the participants at random from a workers' union database, and then arranging a time that suits the participant to survey them.
C) 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.
D) Selecting the participants at random from the tax department database, and then arranging a time that suits the participant to survey them.
E) 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.

Information for Questions 35 to 39

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

35. What type of study did the researchers use?

A) Descriptive, observational, cohort study.
B) Analytic, observational, cohort study.
C) Analytic, observational, randomised control trial.
D) Analytic, experimental, randomised control trial.
E) Descriptive, experimental, randomised control trial.

36. Select the correct statement below regarding the design of the study.

A) Random sampling creates two comparable groups.
B) Random sampling eliminates confounding in the study.
C) Randomisation reduces the risk of confounding in the study.
D) Random sampling eliminates random error in the study.
E) Randomisation eliminates random error in the study.

37. 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.

A) **1-qt(0.975, ν)**
B) **qt(0.975, ν)**
C) **qt(0.995, ν)**
D) **qt(0.995, $n-1$)**
E) **1-qt(0.975, $n-1$)**

38. 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.

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

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

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

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

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

39. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

Information for Questions 40 to 42

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.

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

A) 27.31 B) 160.00 C) 3.91 D) 3.43 E) 18.59

41. 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:

A) 1.30 B) 1.39 C) 0.46 D) 3.26 E) 5.49

42. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

Information for Questions 43 to 49

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

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

A) 0.1183 B) 0.1058 C) 0.0792 D) 0.0734 E) 1.44

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

A) 0.1183 B) 0.1058 C) 0.0792 D) 0.0734 E) 1.44

45. 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:

A) 1.44 B) 0.69 C) 0.1058 D) 0.0734 E) 1.49

46. The interpretation of the relative risk above is:

A) Among the people in the study, the risk of people who attended single-sex schools having social anxiety is greater than 0.

B) 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.

C) 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.

D) 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.

E) 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.

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

A) 0.3139 B) 0.3439 C) 0.3426 D) 0.3125 E) 0.0986

48. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

49. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

Information for Questions 50 to 53

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

50. 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.

A) $H_0 : \mu_E = \mu_R; \quad H_A : \mu_E \neq \mu_R$

B) $H_0 : \mu_E \neq \mu_R; \quad H_A : \mu_E = \mu_R$

C) $H_0 : \mu_E - \mu_R \neq 0; \quad H_A : \mu_E - \mu_R = 0$

D) $H_0 : \bar{x}_E = \bar{x}_R; \quad H_A : \bar{x}_E \neq \bar{x}_R$

E) $H_0 : \bar{x}_E - \bar{x}_R = 0; \quad H_A : \bar{x}_E - \bar{x}_C \neq 0$

51. 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:

A) 12.28
B) 544.01
C) 150.86
D) 295951.92
E) 27.74;

52. 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.

A) -11.87 B) 11.87 C) 750.22 D) 2.33 E) 7.91

53. 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.

A) 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.
B) 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.
C) 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.
D) 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.
E) 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.

Information for Questions 54 to 58

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.

54. 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.

A) $H_0 : p = 0.117; \quad H_A : p \neq 0.117$

B) $H_0 : \pi \neq 0.117; \quad H_A : \pi = 0.117$

C) $H_0 : \pi = 0.117; \quad H_A : \pi \neq 0.117$

D) $H_0 : \pi = 0.206; \quad H_A : \pi = 0.206$

E) $H_0 : \mu = 0.206; \quad H_A : \mu \neq 0.206$

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

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

56. 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?

A) Using a sample size smaller than 97. Increasing α to 0.05.

B) Using a sample size smaller than 97. Decreasing α to 0.001.

C) Using a sample size larger than 97. Increasing α to 0.05.

D) Increasing the power of the study. Decreasing α to 0.001.

E) Increasing the power of the study. Increasing α to 0.05.

57. Choose the appropriate calculation of the test statistic for this hypothesis test.

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

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

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

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

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

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

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

Information for Questions 59 to 63

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.

59. 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?

A) 0.5 B) 0.29 C) -0.29 D) 0.64 E) 0.31

60. 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:

A) 0.0649 B) 0.0665 C) 0.0670 D) 0.0046 E) 0.0681

61. 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.

A) 60.237 B) 5.259 C) -4.259 D) 4.259 E) 3.759

62. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

63. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

Information for Questions 64 to 67

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.

64. What type of study did The Ministry of Education use?

A) Descriptive, observational, cohort study.

B) Analytic, observational, cohort study.

C) Analytic, observational, case-control study.

D) Analytic, experimental, cohort study.

E) Descriptive, experimental, case-control.

65. 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:

A) the probability of making a type I error.

B) the probability of having a false negative result.

C) the probability of making a type II error.

D) the positive predictive value of the test.

E) the power of the test.

66. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

67. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) Having other chronic health conditions neither acts as a confounder or a bias in the relationship between CBT treatment and depression.

Information for Questions 68 to 77

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
```

68. 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.

A) Explanatory variable: **HSGrad**; Response variable: **CrimeRate**.

B) Explanatory variable: **CrimeRate**; Response variable: **HSGrad**.

C) Explanatory variable: **Community**; Response variable: **CrimeRate**.

D) Dependent variable: **HSGrad**; Independent variable: **CrimeRate**.

E) Dependent variable: **Community**; Independent variable: **HSGrad**.

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

A) $391.83 + 4.99\hat{\beta}_1$

B) $-96.52 + 12412.44x$

C) $12412.44 - 96.52x$

D) $12412.44 - 96.52\hat{\beta}_1$

E) $12412.44 + 391.83x$

70. 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?

A) -91.2 B) -20.0 C) 91.2 D) 20.0 E) 6530

71. 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.

A) 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.

B) 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.

C) 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.

D) 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%.

E) 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.

72. The estimated standard error for the slope of the regression line is:

A) 12412.44

B) −96.52

C) 4.99

D) 2526

E) 0.1502

73. 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.

A) $2 \times$ pt(−96.52/4.99, 2117)

B) $2 \times$ pt(−96.52/4.99, 2116)

C) $2 \times$ pt(−96.52/4.99, 2116, lower.tail=False)

D) $2 \times$ pt(12412.44/391.83, 2116, lower.tail=False)

E) $2 \times$ pt(12412.44, 2117, lower.tail=False)

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

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

75. 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.

A) $\hat{\beta}_0$ 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.

B) $\hat{\beta}_0$ 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.

C) $\hat{\beta}_0$ 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.

D) $\hat{\beta}_0$ 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.

E) $\hat{\beta}_0$ 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.

76. 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.

A) 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.

B) Their recommendations should be ignored. They must have made a mistake in their calculation because the correlation coefficient cannot be negative.

C) 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.

D) 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.

E) 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.

77. 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?

A) Simple linear regression with "had ram raid" (yes=1/no=0) as the response variable and high school graduation rate as the explanatory variable.

B) Logistic regression with 'had ram raid" (yes=1/no=0) as the response variable and high school graduation rate as the explanatory variable.

C) Multiple linear regression with "had ram raid" (yes=1/no=0) as the response variable and high school graduation rate as the explanatory variable.

D) ANOVA with high school graduation rate as the response variable and "had ram raid" (yes=1/no=0) as the predictor variable.

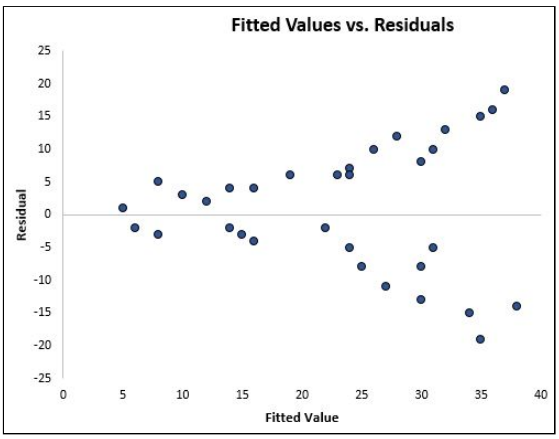
E) Logistic regression with high school graduation rate as the response variable and "had ram raid" (yes=1/no=0) as the explanatory variable.

Information for Question 78

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.



78. 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?

A) Normality of residuals.

B) Equality of variances.

C) Linearity.

D) Both equality of variances and linearity appear to have been violated.

E) None of the assumptions appear to have been violated.

Information for Questions 79 to 85

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
```

79. Which is the response variable in the model?

A) The perceived wine quality.
B) The intercept.
C) The citric acid content.
D) Red wine.
E) The sulphate content.

80. 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.

A) 6.485
B) −18.308
C) 7.136
D) 6.043
E) 6.708

81. Letting β_2 be the slope parameter for the chlorides variable, what does β_2 represent?

A) β_2 is the true mean change in chloride levels for every unit change in perceived wine quality.
B) β_2 is the estimated mean change in perceived wine quality for every unit change in chloride levels, based off this sample.
C) 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.
D) β_2 is the true mean change in perceived wine quality for every unit change in chloride levels.
E) 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.

82. 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.

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

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

C) $-21.33445 \pm t_{0.975,1592} \times 11.81275$

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

E) $-21.33445 \pm t_{0.975,1598} \times 11.81275$

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

A) There is no statistically significant evidence at the 5% significance level that the perceived quality of wine is associated with its density.

B) 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.

C) We have statistically significant evidence at the 5% significance level that the perceived quality of wine tends to decrease as its density increases.

D) 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.

E) We have statistically significant evidence at the 5% significance level that the perceived quality of wine tends to increase as its density increases.

84. 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.

A) We have strong evidence that the perceived quality of wine tends to improve as citric acid content increases.

B) 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.

C) 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.

D) We have no evidence that the perceived quality of wine is associated with citric acid levels.

E) 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.

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

A) 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%.

B) 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%.

C) 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.

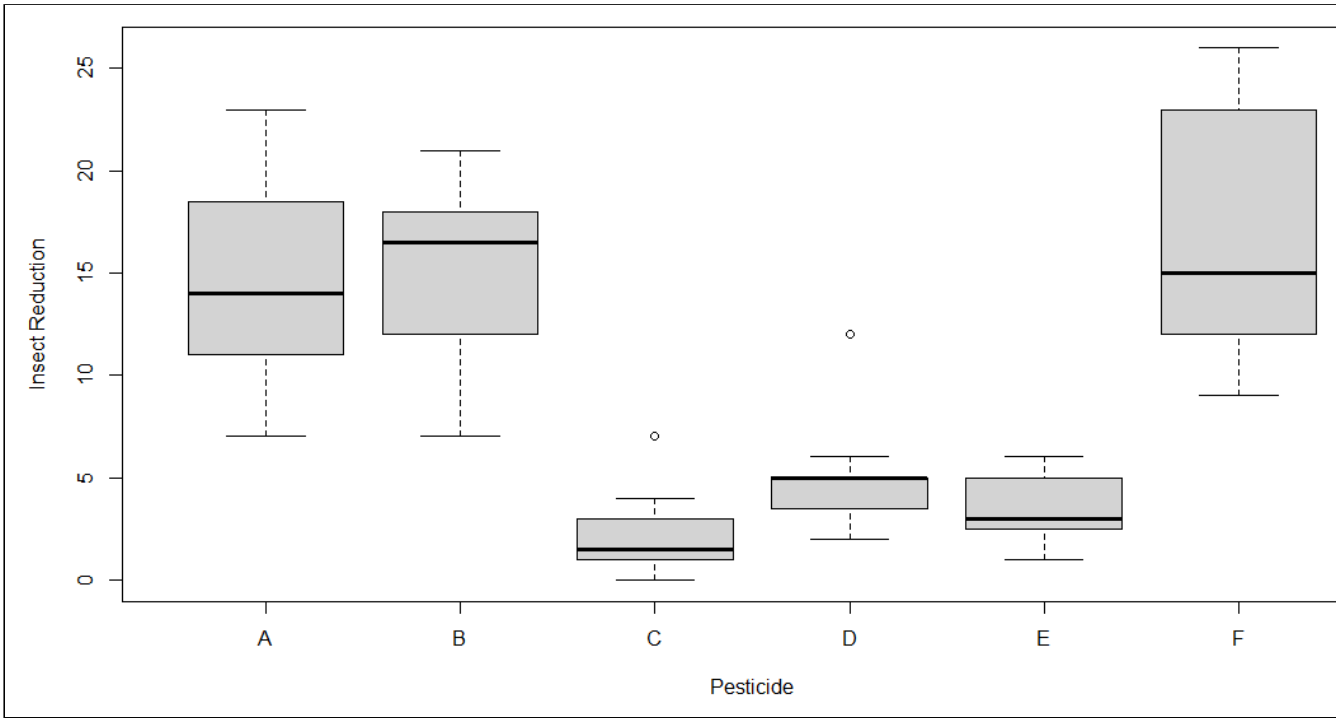
D) Since the R^2 value is greater than 0.05, the data provide no statistically significant evidence that the citric acid, chloride, pH, density, alcohol, and sulphate content predict perceived wine quality.

E) According to the model, 30.88% of the variation in perceived wine quality can be predicted. The rest is purely due to random variation.

Information for Questions 86 to 93

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
```

86. 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

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

A) The e_{ij} values for each group i are assumed to normally distributed with a zero mean and constant variance.

B) The e_{ij} values for each group i are assumed to follow a t-distribution.

C) The e_{ij} values are only approximately normal if the sample sizes for each group of observations are the same.

D) 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.

E) 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.

F) 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.

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

A) $H_0: \mu_A = \mu_B = \mu_C = \mu_D = \mu_E = \mu_F$; H_A : All the means are different to each other.

B) H_0 : All the means are different to each other; H_A : All the means are the same.

C) $H_0: \mu_A = \mu_B = \mu_C = \mu_D = \mu_E = \mu_F$; $H_A: \mu_A, \mu_B, \mu_C, \mu_D, \mu_E, \mu_F$ are not all equal.

D) $H_0: \hat{\mu}_A = \hat{\mu}_B = \hat{\mu}_C = \hat{\mu}_D = \hat{\mu}_E = \hat{\mu}_F$; $H_A: \hat{\mu}_A, \hat{\mu}_B, \hat{\mu}_C, \hat{\mu}_D, \hat{\mu}_E, \hat{\mu}_F$ are not all equal.

E) 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.

89. 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.

A) The total SS is 2668.8; the residual SS is 1015.2.
B) The total SS is 3684.0; the residual SS is 1015.2.
C) The group SS is 1015.2. The residual SS is 2668.8.
D) The group SS is 3684.0; the residual SS is 1015.2.
E) The total SS is 3684.0; the group SS is 1015.2.

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

A) When the variability between the group means is small, and the variability within the study groups is small.
B) When the variability between the group means is small, and the variability within the groups is large.
C) When the variability between the group means is large, and the variability within the groups is large.
D) When the variability between the group means is large, and the variability within the groups is small.
E) When the overall variability across all of the group observations is large.

91. 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.

A) **pf(f,71,lower.tail=FALSE)**
B) **1-pf(f,71,lower.tail=TRUE)**
C) **pf(f,5,66,lower.tail=FALSE)**
D) **1-pf(f,5,66,lower.tail=FALSE)**
E) **2*(pf(f,5,66,lower.tail=FALSE)**

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

A) 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.
B) 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.
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.
D) 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.
E) 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.

93. 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?

A) 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.

B) 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.

C) 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.

D) 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.

E) 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.

END OF QUESTIONS

Summary of Formulae

Sample mean and variance

$$\text{Mean: } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{Variance: } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Probability Rules

$\Pr(A \text{ or } B) = \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$ $\quad \Pr(A \text{ and } B) = \Pr(A \cap B) = \Pr(A)\Pr(B|A)$

$= \Pr(B)\Pr(A|B)$

Random Variables If X and Y are independent random variables, then $W = aX + bY + c$ has:

$$\text{Mean: } \mu_W = a\,\mu_X + b\,\mu_Y + c \qquad \text{Variance: } \sigma_W^2 = a^2\,\sigma_X^2 + b^2\,\sigma_Y^2$$

Discrete Distributions

$$\text{Mean: } \mu_X = \sum_{i=1}^{k} x_i \Pr\left(X = x_i\right) \qquad \text{Variance: } \sigma_X^2 = \sum_{i=1}^{k} \left(x_i - \mu_X\right)^2 \Pr\left(X = x_i\right)$$

Binomial Distribution

$$\mu_X = n\pi \qquad \sigma_X^2 = n\pi(1-\pi) \qquad \Pr\left(X = x\right) = \binom{n}{x}\pi^x\left(1-\pi\right)^{n-x} \qquad \binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

If $n\pi \pm 3\sqrt{n\pi(1-\pi)}$ are between 0 and n, then X is approximately normally distributed with mean μ_X and variance σ_X^2.

Normal Distribution A standard normal random variable, Z, has $\mu_Z = 0$ and $\sigma_Z^2 = 1$. To transform a normal random variable X into a standard normal (and vice versa):

$$Z = \tfrac{X - \mu_X}{\sigma_X} \qquad X = Z\sigma_X + \mu_X$$

Distributions of Statistics

- The mean $\bar{X}$ of a random sample of size n has mean $\mu_{\bar{X}} = \mu_X$ and standard error $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$.
- The sample proportion P computed from a binomial distribution with parameters n and π has a mean of $\mu_P = \pi$ and standard error $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$. If $n\pi \pm 3\sqrt{n\pi(1-\pi)}$ are between 0 and n, then P will be approximately normally distributed.
- The distribution of the difference between two sample means $\bar{X}_1 - \bar{X}_2$ has a mean of $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$ and a standard error of $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Contingency Tables

	Factor 2		
Factor 1	Level 1	Level 2	Total
Level 1	a	b	$r_1 = a + b$
Level 2	c	d	$r_2 = c + d$
	$c_1 = a + c$	$c_2 = b + d$	$n = a + b + c + d$

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- R and C are the number of rows and columns respectively
- $e_{ij} = \frac{r_i c_j}{n}$, where r_i is the ith row total and c_j is the jth column total
- o_{ij} is the observed value in row i column j

<u>With Factor 2 as "Outcome" variable:</u>

Odds ratio: $OR = (a/b)/(c/d) = ad/bc$

Relative risk: $RR = (a/r_1)\,/\,(c/r_2)$

Attributable risk: $AR = a/r_1 - c/r_2$

Confidence Intervals and Hypothesis Tests

All of the $100(1-\alpha)\%$ confidence intervals calculated in this course are of the form:
Estimate ± multiplier × standard error
In the table $\bar{x}$, p etc are the values calculated from the samples.

	Estimate	df (ν)	Multiplier	Standard Error
Population mean				
• Random sample, σ_X known	$\bar{x}$	NA	$z_{(1-\alpha/2)}$	$\frac{\sigma_X}{\sqrt{n}}$
• Normal population, σ_X unknown	$\bar{x}$	$n-1$	$t_{(1-\alpha/2,\nu)}$	$\frac{s}{\sqrt{n}}$
• Large random sample ($n \geq 20$), σ_X unknown	$\bar{x}$	$n-1$	$t_{(1-\alpha/2,\nu)}$	$\frac{s}{\sqrt{n}}$
Difference between population means				
• Large random samples (both ≥ 20)	$\bar{x}_1 - \bar{x}_2$	Will be provided	$t_{(1-\alpha/2,\nu)}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
• Paired difference in random samples from a normal population	$\bar{d}$	$n-1$	$t_{(1-\alpha/2,\nu)}$	$\frac{s_d}{\sqrt{n}}$
Population proportions				
• Population proportion	p	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{p(1-p)}{n}}$
• Difference between 2 population proportions	$p_1 - p_2$	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
• Difference between 2 population proportions (hypothesis test)	$p_1 - p_2$ (Use $p^* = \frac{x_1+x_2}{n_1+n_2}$ for hypothesis test)	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{p^*(1-p^*)}{n_1} + \frac{p^*(1-p^*)}{n_2}}$
Odds ratio, relative risk, risk difference (see contingency table on previous page for a, b, c and d)				
• Log odds ratio	$\ln(OR)$	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
• Log relative risk	$\ln(RR)$	NA	$z_{(1-\alpha/2)}$	$\sqrt{\frac{1}{a} - \frac{1}{r_1} + \frac{1}{c} - \frac{1}{r_2}}$
• Risk difference – as for the difference between two population proportions with $p_1 = a/r_1$ and $p_2 = c/r_2$				
After ANOVA and Regression				
• Estimate, multiplier and standard error determined from output				

Regression

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{\beta}_1 = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Standard error of the slope is $s_{\hat{\beta}_1} = \dfrac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}}$ where $s_e = \sqrt{\dfrac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\dfrac{\text{RSS}}{n-2}}$

Standard error of a prediction at $X = x_0$ is $PE(\hat{y}_0) = s_e\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$

ANOVA

$y_{ij} = \mu_i + e_{ij}$ $\quad$ $\hat{\mu}_i = \bar{y}_{i\cdot}$

$$\underbrace{\text{TSS}}_{\substack{\text{Total sum}\\ \text{of squares}}} = \underbrace{\text{GSS}}_{\substack{\text{Group sum}\\ \text{of squares}}} + \underbrace{\text{RSS}}_{\substack{\text{Residual sum}\\ \text{of squares}}}$$

Post ANOVA analysis

$s_e = \sqrt{RMS}$

END