Question 1 Multiple choice: circle the best answer to each question:

(1)[2 points] A quality control manager designs a procedure to monitor lead levels in toys manufactured in factories abroad and imported by his company. One out every ten shipping containers of toys is selected, and from that container ten toys are randomly selected for testing. The lead level in each of the ten toys is then measured and recorded in units of ppm (parts per million). The population, sample, and variable in this example are, respectively

(a) consumers; the shipping container selected; the amount of lead.

(b) factories; the ten toys selected; the amount of lead.

(c) the ten toys; the shipping container; factory of origin.

(d) toys; the ten toys tested; the amount of lead.

(e) all lead based toys; the factories; the amount of lead.

(2)[2 points] Which of the following is true?

(a) All sample surveys are observational studies.

(b) All observational studies are sample surveys.

(c) A census is an observational study.

(d) (a) and (b).

(e) (a) and (c).

(3)[2 points] Which of the following sampling methods is best for avoiding bias?

(a) A convenience sample.

(b) A voluntary response sample.

(c) A simple random sample.

(d) An observational sample.

(e) A double-blind sample.

(4)[2 points] If using the table of random digits (Table A) attached to select a simple random sample of size 100 from a population of 100,000,

(a) individuals should be labelled $1, 2, \ldots, 100000$.

(b) individuals should be labelled $0, 1, \ldots, 99999$.

(c) individuals should be labelled $00000, 00001, \ldots, 99999$.

(d) individuals should be labelled $00000, 11111, \ldots, 99999$.

(e) Table A cannot be used since some labelled individuals are not represented.

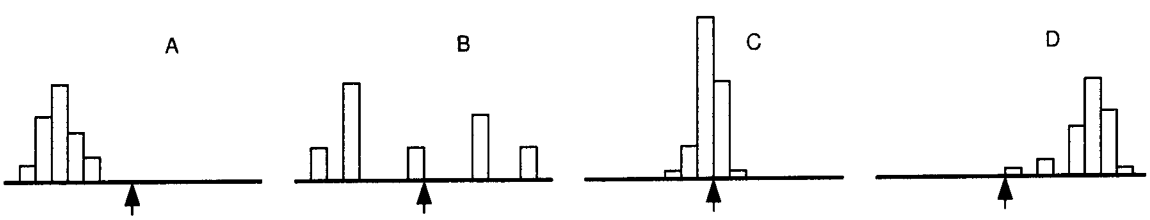
(5)[2 points] A study of student success after extra tutoring concludes "We are 95% confident that the difference between the average score for tutored students and that for students without tutoring is between 17 and 36 points." To be 99% confident, the range of points would be

(a) wider, since higher confidence requires a larger margin of error.
(b) narrower, since higher confidence requires a smaller margin of error.
(c) wider, since higher confidence requires a smaller margin of error.
(d) narrower, since higher confidence requires a larger margin of error.
(e) wider, since lower confidence requires a larger margin of error.

(6)[2 points] The confidence statement "We are 95% confident that the difference between the average score for tutored students and that for students without tutoring is between 17 and 36 points" means

(a) 95% of students will increase their score by between 17 and 36 points.
(b) We are certain that the average increase is between 17 and 36 points.
(c) The range of 17 to 36 points was determined using a method which would capture the true average in 95% of samples.
(d) 95% of averages fall between 17 and 36.
(e) 95% of adults believe this statement.

(7)[2 points] To estimate the value of some variable of interest, one hundred samples were taken from a population and the sample mean calculated for each sample. A histogram of the distribution of the sample means was then made; an arrow indicates the true value of the population mean. Which of the following histograms shows small bias and large variability in the sample means?



(a) A
(b) B
(c) C
(d) D
(e) bias and variability cannot be determined from a histogram.

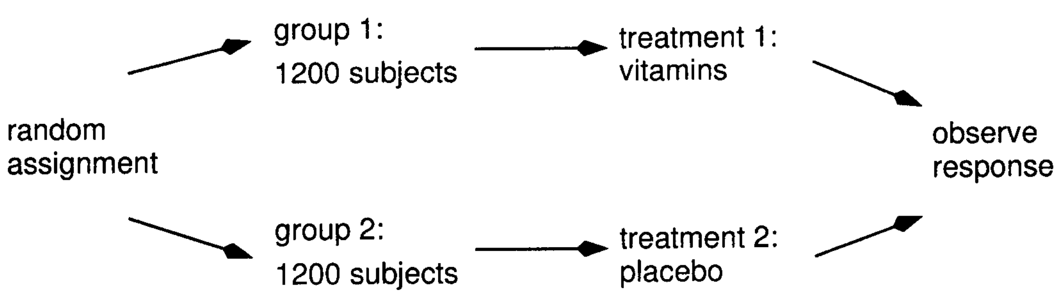
(8)[2 points] According to the 2006 Census, the city of Parksville has 11,545 residents, of whom only 1665 are in the 20-34 age bracket, while 7815 are in the 35+ age bracket. To survey individuals aged 20 and older, a survey firm establishes two strata of equal size and randomly selects individuals for each. The first stratum gives 20-34 year-olds a 1/3 chance of being selected for the survey. The second stratum is for those in the 35+ age bracket. What is the approximate chance (probability) of a Parksville resident in the 35+ age bracket being selected for the survey?

$(1665)(\frac{1}{3}) = 555$

$\therefore \frac{555}{7815} \doteq 0.07$

(a) 0.03
(b) 0.05
(c) 0.07
(d) 0.09
(e) 0.10

(9)[2 points] The study outlined in the following figure is best described as



(a) a simple random sample.
(b) a stratified random sample.
(c) randomized comparative experiment.
(d) a multistage experiment.
(e) an observational study.

(10)[2 points] In a class demonstration a pencil is measured by five different people using a poor quality ruler. A much more accurate value for the length of the pencil is then determined; call this the true value of the length. The differences between the measured values and true value are found to have an average of 0.2 cm. What is the bias in measurement using the poor quality ruler?

(a) 0.2 cm.
(b) −0.2 cm.
(c) ±0.2 cm.
(d) 1 cm.
(e) cannot be determined since the differences between measured and true values also depends on random measurement error.

(11)[2 points] The annual percentage change in value of a portfolio of the 500 most popular stocks is given in the following table:

year	1999	2000	2001	2002	2003	2004	2005	2006
change (%)	21.0	-9.1	-11.9	-22.1	28.7	10.9	4.9	15.8

sorted −22.1 −11.9 −9.1 4.9 10.9 15.8 21.0 28.7

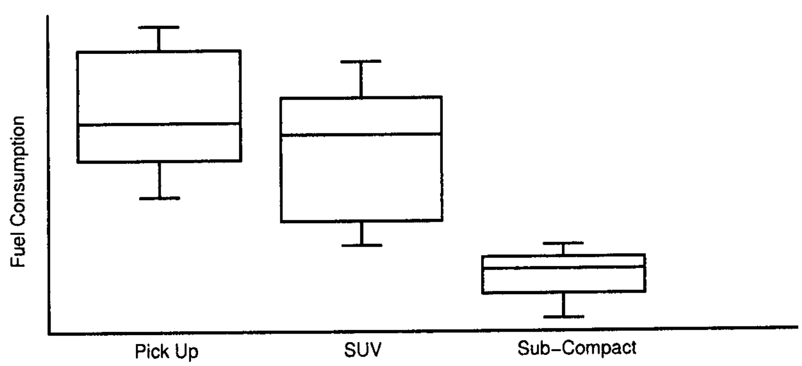
Letting Q_1 and Q_3 represent the first and third quartiles of the percentage change values, respectively, what is the approximate value of $Q_1 + Q_3$?

(a) 1.8
(b) 6.0
(c) 3.3
(d) 7.9
(e) -2.6

(12)[2 points] Using the data table from Question (11) on this page, Let M represent the median and $\bar{x}$ the mean of the percent changes. What is the approximate value of $M - \bar{x}$?

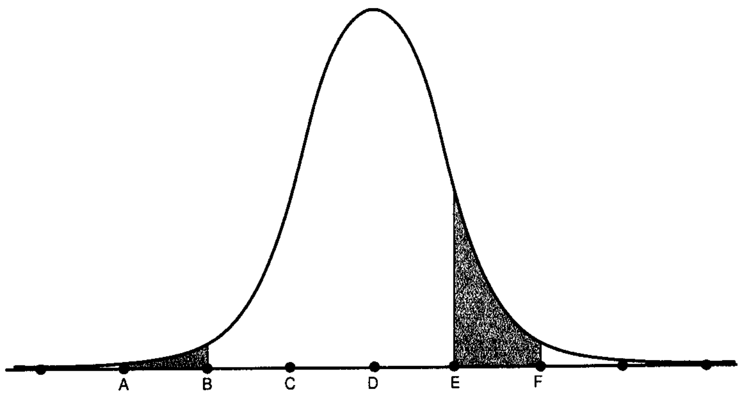
(a) 7.9
(b) 3.1
(c) -3.1
(d) 4.8
(e) 0.0

(13)[2 points] The boxplot below summarizes fuel consumption data (in litres per 100 km) for three classes of automobiles: pick up trucks, sport utility vehicles (SUV), and sub-compact (very small) cars. Which of the following is true?



(a) The third quartile of SUVs have lower fuel consumption than the third quartile of pick ups.
(b) All sub-compacts have lower fuel consumption than pick ups.
(c) The minimum fuel consumption among SUVs is greater than the third quartile of sub-compacts, yet less than the minimum among pick-ups.
(d) The bottom 50% of SUVs consume less fuel than the top 50% of pick-ups.
(e) all of the above.

(14)[2 points] The graph below represents a normal distribution and the distance between the points shown on the horizontal axis is one standard deviation. Approximately what percentage of the total area is represented by the shaded regions (the region between A and B and that between E and F)?



(a) 16%
(b) 32%
(c) 24%
(d) 30%
(e) 18%

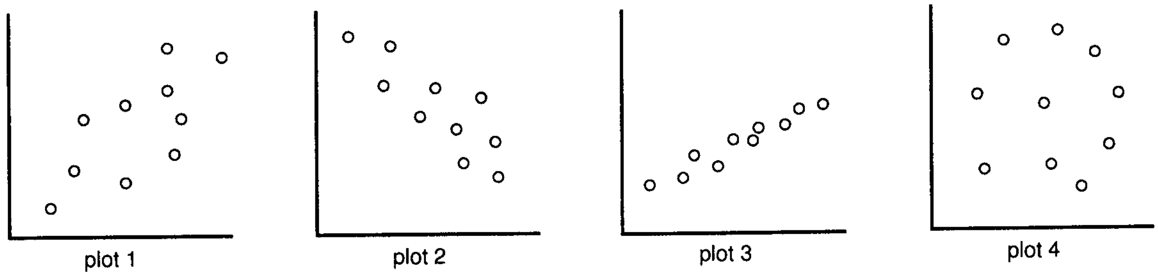
(15)[2 points] With reference to the distribution from Question (14) on this page, which point has a z-score (or standard score) closest to -2.2?

(a) A
(b) B
(c) C
(d) D
(e) E

(16)[2 points] With reference to the distribution from Question (14) on this page, 84.13% of all values are greater than what value?

(a) A
(b) B
(c) C
(d) D
(e) E

(17)[2 points] The correlations for the four scatterplots below are, respectively,



(a) $r = 1$, $r = -1$, $r = 1$, $r = 0$

(b) $r = -0.3$, $r = 0.7$, $r = -0.9$, $r = 0$

(c) $r = 0.9$, $r = -0.7$, $r = 0.3$, $r = 0$

(d) $r = 0.3$, $r = 0.9$, $r = -0.7$, $r = 0$

(e) $r = 0.3$, $r = -0.7$, $r = 0.9$, $r = 0$

(18)[2 points] With reference to the four scatterplots in Question (17) on this page, which is best described as showing an association which is weak, positive and linear?

(a) plot 1

(b) plot 2

(c) plot 3

(d) plot 4

(e) none matches this description

(19)[2 points] The least squares regression line for predicting heart disease deaths by country based on wine consumption data is $y = mx + 260.6$ for some value m. Here x represents average number of litres of wine consumed per person in the country, and y is heart disease deaths per 100,000 people in the population. For a country where, on average, residents consume 6.8 litres of wine per year, the least squares regression line predicts 104.4 deaths per 100,000 people. What must be the value of m.

(a) $m = 22.97$

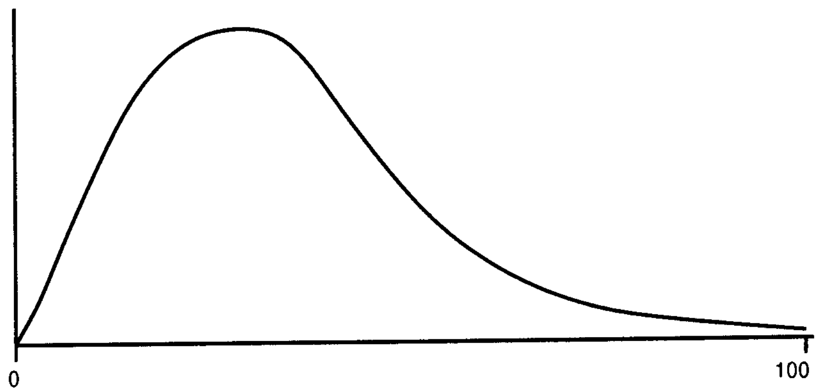
(b) $m = -22.97$

(c) 2.43

(d) -2.43

(e) 383.73

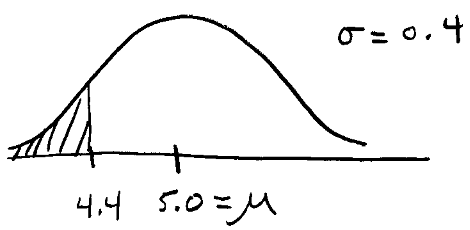
(20)[**2 points**] Three students are comparing grades after the final exam. The first student excitedly proclaims "I scored 15 marks above the class mean." The second student responds with "My score was 15 above the class median." The third observes that her grade is the average of the first two. If the grades were distributed according to the following distribution, who did better, the first student, second or third student?



(a) the first.
(b) the second.
(c) the third.
(d) they are all about the same.
(e) there is not enough information to decide.

Question 2 Gasoline consumption sensors are installed in 1000 new fuel hybrid vehicles as they leave the factory so that fuel consumption data can be gathered for the new vehicles. After one year of driving the data is downloaded and processed: for each vehicle the average fuel consumption over the year is calculated. These 1000 average fuel consumption values are then plotted and found to have a distribution which is approximately normal with a mean of $\mu = 5.0$ litres per 100 km and a standard deviation of $\sigma = 0.4$ litres per 100 km.

(a)[3 points] Approximately how many of the 1000 vehicles have fuel consumption values of 4.4 litres per 100 km or less? Round to the nearest vehicle.

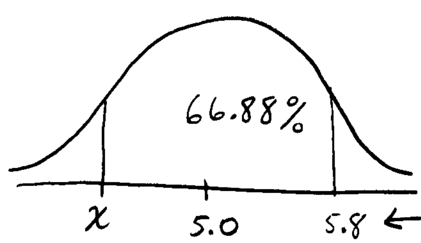


$$z = \frac{4.4 - 5.0}{0.4} = -1.5$$

$\therefore$ area represents 6.68% of vehicles, i.e.

$(0.0668)(1000) \doteq \boxed{67 \text{ vehicles.}}$

(b)[4 points] 66.88% of vehicles have fuel consumption values below 5.8 but above what value?



$$z = \frac{5.8 - 5.0}{0.4} = 2$$

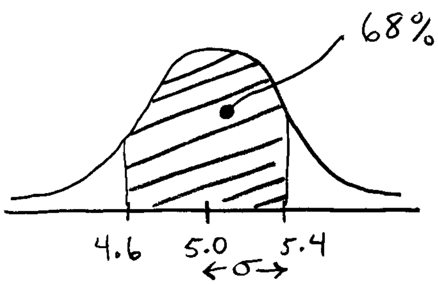
$\therefore$ 97.73%

$\therefore 97.73 - 66.88 = 30.85\%$ which corresponds to a z-score of -0.5, so

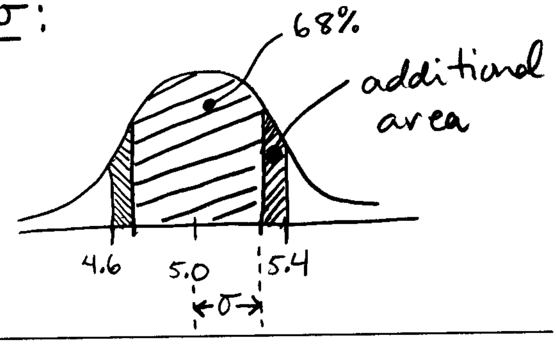
$x = 5 - (0.5)(0.4) = \boxed{4.8 \text{ L}/100 \text{ km}}$

(c)[3 points] A mistake is discovered and σ is smaller than first thought. As a result, will the percentage of vehicles with consumption values between 4.6 and 5.4 be larger or smaller than first thought? Explain briefly. $\boxed{\text{Larger}}$

original σ:



revised σ:



Question 3 This is a probability question which involves rolling a pair of dice. The first die has sides numbered 1 and 2 coloured red while the other four sides are green. The second die has sides numbered 1, 2 and 3 coloured red while the other three sides are green.

(a)[2 points] Both dice are rolled. What is the probability that both come up red?

$$P(\text{both red}) = P(\text{first red})\,P(\text{second red}) = \left(\tfrac{2}{6}\right)\left(\tfrac{3}{6}\right) = \frac{6}{36} = \boxed{\tfrac{1}{6} \doteq 0.17}$$

(b)[2 points] Both dice are rolled. What is the probability that at least one comes up red?

$$P(\text{at least one red}) = 1 - P(\text{both green}) = 1 - P(\text{first green})\,P(\text{second green}) = 1 - \left(\tfrac{4}{6}\right)\left(\tfrac{3}{6}\right) = \boxed{\tfrac{2}{3} \doteq 0.67}$$

(c)[3 points] Both dice are rolled. What is the probability of both coming up red or getting a total of 6?

Event "both red or total of 6" consists of (R1,R1) (R1,R2) (R1,R3) (R2,R1) (R2,R2) (R2,R3) (R1,G5) (R2,G4) (G3,R3) (G4,R2) (G5,R1)

$\therefore$ P(both red or total 6) $= \boxed{\tfrac{11}{36} \doteq 0.31}$

(d)[3 points] The first die only is rolled and you receive \$3 if a 1 or 2 comes up, \$12 if a 3, 4 or 5 comes up, and \$18 if a 6 comes up. What are your expected winnings?

$$\text{Expected winnings} = (3)\left(\tfrac{2}{6}\right) + (12)\left(\tfrac{3}{6}\right) + (18)\left(\tfrac{1}{6}\right) = \frac{6+36+18}{6} = \boxed{\$10}$$

Question 4

(a)[5 points] Of the total number of characters (individual letters) in a book written in English, a student wishes to estimate what proportion are vowels (the letters a, e, i, o, u, y). To do this, the student walks into the library, pulls a book at random off the shelf, opens up the book to a random full page of text and counts the characters. The student finds that it contains 2100 characters, of which 1176 are vowels. Assuming that the page of text represents a random sample, give a 90% confidence interval for the true proportion of vowels which appear in English writing. Round your answer to two decimals.

$$\hat{p} = \frac{1176}{2100} = 0.56$$

$$z^* = 1.64$$

$\therefore$ confidence interval is $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$$= 0.56 \pm (1.64)\sqrt{\frac{0.56(1-0.56)}{2100}}$$

$$\doteq \boxed{0.56 \pm 0.02}$$

(b)[5 points] The page of text which formed the random sample in part (a) can also be used to estimate the mean word length of English words used in writing (the length of a word is the number of characters in the word). Suppose the 2100 characters on the random page of text form 467 words on the page. The standard deviation of the word lengths is 1.1 letters. Use this information to construct a 95% confidence interval for the mean word length of English words used in writing. Again round your answer to two decimals.

$\bar{x}$ = sample mean length of word

$\bar{x} = \frac{2100}{467} \doteq 4.4968$ characters

$s = 1.1$

$n = 467$, $z^* = 1.96$

$\therefore$ confidence interval is $\bar{x} \pm z^* \frac{s}{\sqrt{n}}$

$$= 4.4968 \pm (1.96)\frac{1.1}{\sqrt{467}}$$

$$\doteq \boxed{4.50 \pm 0.10}$$

Question 5 [10 points] Bags of top grade frozen peas are supposed to have no more than 2% poor quality peas in the mix. Poor quality means discoloured, shrivelled, broken, etc. A consumer buys a 1 kg bag of frozen peas and decides to test the producer's claim that the peas are top grade. Out of 1134 peas in the bag, 29 are of poor quality. Is this sufficient evidence to refute the producer's claim that the peas are top grade? Test at the $\alpha = 0.05$ level of significance, and clearly state the parameter you are testing as well as the hypotheses.

Let p = true proportion of poor quality peas produced.

H_0 : $p = 0.02$

H_a : $p > 0.02$

$$\hat{p} = \frac{29}{1134} \doteq 0.02557.$$

All such $\hat{p}$ have a distribution which is approximately normal, $\mu_{\hat{p}} = 0.02$, $\sigma_{\hat{p}} = \sqrt{\frac{(0.02)(1-0.02)}{1134}}$

$\doteq 0.004157$

$$\therefore z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.02557 - 0.02}{0.004157} \doteq 1.3$$

$$P(z \geq 1.3) \doteq \frac{100 - 90.32}{100} = 0.0968.$$

Since $0.0968 > \alpha = 0.05$, we do <u>not</u> have enough evidence to reject H_0.

Question 6 [10 points] A car manufacturer lists the fuel consumption data for their vehicles alongside the sale prices. The fuel consumption value for a particular model is actually a mean for all vehicles of that model. Suppose a consumer reporting agency tests the manufacturer's claim that a particular model of vehicle gets 23 miles per gallon. To test this, fuel consumption data for 50 new vehicles is gathered and found to have a mean of 21.1 miles per gallon and a standard deviation of 3.9 miles per gallon. Is this sufficient evidence to conclude that the mean miles per gallon rating of the vehicles is less than what is stated by the manufacturer? Test at the $\alpha = 0.01$ level of significance, and clearly state the parameter you are testing as well as the hypotheses.

Let μ = mean miles per gallon rating of vehicles.

$H_0: \mu = 23$
$H_a: \mu < 23$

$\bar{x} = 21.1$
$s = 3.9$
$n = 50$

All such $\bar{x}$ have a distribution which is approximately normal with $\mu_{\bar{x}} = 23$, $\sigma_{\bar{x}} = \frac{3.9}{\sqrt{50}} \doteq 0.5515$

$\therefore \quad z = \frac{21.1 - 23}{0.5515} \doteq -3.4$

$P(z \leq -3.4) = \frac{0.03}{100} = 0.0003$

Since $0.0003 < \alpha = 0.01$, we reject H_0 in favour of H_a.

Question 7 Is eye colour and handedness (preference for using left or right hand) related? Data on 1166 people with either blue or brown eyes was distributed as follows:

	left	right	total
blue	15 (23.7)	238 (229.3)	253
brown	94 (85.3)	819 (827.7)	913
total	109	1057	1166

(a)[8 points] Test at the $\alpha = 0.01$ level of significance to determine if there is significant evidence of an association between eye colour and handedness. Clearly state the null and alternative hypotheses as well as your conclusion.

H_0: There is no association between eye colour and handedness

H_a: There is an association.

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \doteq 4.58$$

$df = (2-1)(2-1) = 1$

Critical number corresponding to $df=1$, $\alpha = 0.01$
$= 6.63$.

Since $\chi^2 = 4.58 < 6.63$, we do not have sufficient evidence to reject H_0.

(b)[2 points] What is the greatest value of α in Table 24.1 at which the evidence is statistically significant?

$\alpha = 0.25$

[Note: the intended question was "What is the smallest value of α in Table 24.1 at which the evidence is statistically significant?", the answer to which is $\alpha = 0.05$.]