For questions 1 – 30, circle one answer only. If making an error, write your final answer clearly in the margin. Ambiguous responses will be considered incorrect.

1. The mean age of five people in a room is 30 years. One of the people whose age is 50, leaves the room. The mean age of the remaining people is 25 years.

   T(correct) F

2. For a Normal distribution, the first and third quartiles are respectively one standard deviation below and above the mean.

   T F(correct)

3. A correlation coefficient based on a scatter plot measures the proportion of data lying on the regression line.

   T F(correct)

4. A phone–in poll at a radio station concluded that 70% of Canadians approve of Stephen Harper, based on the responses of 8,000 callers. The conclusion is valid since the sample size is large.

   T F(correct)

5. A random sample of 100 students are asked if they are vegetarians. Five percent respond yes, and a 99% confidence interval for the true proportion of students who are vegetarians is found to be $(0.03, 0.07)$. This implies 99% of all random samples of 100 students will have a sample proportion that falls between 0.03 and 0.07.

   T F(correct)

6. Confidence intervals are constructed for both parameters and statistics.

   T F(correct)

7. You carry out a hypothesis test that compares the means of two populations. Suppose a P-value of 0.12 is obtained. This implies that there is a 12% chance that the two population means are equal.

   T F(correct)

8. At the start of this course, I hypothesized that the students in the pharmacy program might have a higher mean grade than the students not in the pharmacy program. After this final exam is completed and marked, I will have final grades for every student. To test my hypothesis, I should use... (choose the best answer)

   **(a)** A one-sample t-test for the mean

   **(b) A two sample t-test for the mean**

   **(c)** A confidence interval for the proportion

   **(d)** The correlation coefficient

   **(e)** Linear regression

   UBC is interested in finding out if students wish to continue the U-pass program. The researchers decide that undergraduate and graduate students may have different opinions on such an issue. Undergraduate students comprise 84% of the student population and graduate students make up the other 16%. The researcher will survey 1000 students, and so she obtains a list of all UBC students. She divides them into two lists, one of undergraduate students and one of graduate students. She then randomly samples 840 undergraduate and 160 graduate students to survey. She finds that 83% of those surveyed wish to continue with the U-pass program. Use this information to answer questions 9 and 10.

9. What type of sampling technique was used?

   **(a)** Simple random sampling

   **(b) Stratified sampling**

   **(c)** Cluster sampling

   **(d)** Systematic sampling

   **(e)** Convenience sampling

10. What is the population parameter?

    **(a)** The distribution of students in an undergraduate of graduate program

    **(b)** 83%

    **(c)** The 1000 students surveyed

    **(d) The proportion of UBC students who wish to continue the U-pass program**

    **(e)** UBC students

11. Researchers at a university are interested in examining the relationship between one's smoking habits and whether or not they have lung cancer. They randomly sample 312 people with lung cancer and 427 people without lung cancer. After interviewing the patients and analyzing the data, the conclude that smoking is associated with having lung cancer. Which of the following is correct?

    **(a)** This is a prospective study

    **(b) This is a retrospective study**

    **(c)** The sample sizes must be equal to have a valid study

    **(d)** There is no control group

    **(e)** Both (b) and (c)

12. A simple random sample of 150 cars was taken to estimate the mean speed of cars driving on the sea-to-sky highway, and resulted in the following 95% confidence interval: (107,116). Circle the only correct statement.

    **(a)** About 95% of cars in this sample were driving between 107 to 116km/hr

    **(b)** A car found driving 120km/hr would be considered unusual

    **(c)** We have violated our model assumptions by taking a simple random sample

    **(d)** There is a 5% chance of making a Type II error

    **(e) None of the above**

13. The national farming association has data on the amount of pesticide used per acre and the percentage of fruit that has been contaminated by insects per acre, for 117 farms. These two variables have a correlation coefficient $r = -0.8$. For a farm that uses an amount of pesticide that is 2 standard deviations above the mean amount of pesticide used, we would predict the percentage of contaminated fruit will be

    **(a)** 1.6 standard deviations above the mean

    **(b) 1.6 standard deviations below the mean**

    **(c)** 2 standard deviations above the mean

    **(d)** 2 standard deviations below the mean

    **(e)** 1.28 standard deviations below the mean

14. The Vancouver police are interested in estimating the proportion of cars that are uninsured within the city. They assume that uninsured cars are spread randomly and uniformly throughout the city. They create a list of all major intersections in the city, and randomly select 10 to inspect. They set up roadblocks at each of the selected intersections and stop all cars passing to check for insurance. What type of sampling technique was used?

**(a)** Simple random sampling

**(b)** Stratified sample

**(c) Cluster sampling**

**(d)** Convenience sampling

**(e)** Systematic sampling

15. The slope of a regression line and the correlation are similar in the sense that

**(a) they both have the same sign.**

**(b)** they do not depend on the units of measurement of the data.

**(c)** they both fall between –1 and 1 inclusive.

**(d)** neither of them can be affected by outliers.

**(e)** both can be used for prediction.

16. Which of the following is/are *incorrect* statement(s) about the correlation between two quantitative variables $X$ and $Y$?

I. A correlation of –0.8 indicates a stronger linear association between X and Y than a correlation of 0.5.
II. A correlation of 0 implies X and Y are not related at all.
III. A correlation of –1 indicates that $Y = -X$.

**(a)** I only

**(b)** II only

**(c)** I and II only

**(d) II and III only**

**(e)** I, II and III

17. A certified fitness coach wanted to test the effectiveness of a new fitness program in reducing weight among obese patients. Fifty female patients and fifty male patients participated in the experiment. Within each gender group, the patients were randomly assigned to one of the two fitness programs – the new and the existing fitness programs. Upon completion of the program, reduction in weight was measured for each patient. Which of the following statements is incorrect about this experiment?

**(a) There are four treatments in the study.**

**(b)** Gender is a blocking variable.

**(c)** The patients were not guaranteed to lose weight due to the experiment.

**(d)** Reduction in weight is the response variable.

**(e)** Type of fitness program is a factor.

18. In testing a two–sided hypothesis test for a mean, the test statistic was –2.12 which is expected to be a value from the standard Normal distribution under the null hypothesis. The P–value of the hypothesis test is (choose the most appropriate answer)

    **(a)** Between 0 and 0.025

    **(b) Between 0 and 0.05**

    **(c)** Between 0 and 0.003

    **(d)** Between 0.95 and 0.997

    **(e)** Between 0.05 and 1

19. If examining the plot of the residuals against the explanatory variable $x$ for a linear model, when the model fits well one would expect

    **(a)** the residuals to lie on a line of positive slope.

    **(b)** the residuals to lie on a line of negative slope.

    **(c)** the residuals to scatter about a line of positive slope.

    **(d)** there to be no variation in the residuals.

    **(e) there to be no obvious pattern in the residuals.**

20. In a large city, 37% of all restaurants accept both master and visa credit cards, and 50% accept master cards and 60% accept visa cards. A tourist visiting the city picks at random a restaurant at which to have lunch. Define the following events:

$$\begin{aligned} M &= \{\text{the randomly chosen restaurant accepts master credit cards}\}, \\ V &= \{\text{the randomly chosen restaurant accepts visa credit cards}\}. \end{aligned}$$

    Are M and V independent?.

    **(a)** Yes.

    **(b) No.**

    **(c)** Insufficient information to tell.

21. A type of thread is being studied for its tensile strength. Fifty-one pieces were tested under similar conditions, the mean tensile strength being 78.30kg and the standard deviation being 5.60kg.

**(1)** Give an approximate 95% confidence interval for the mean tensile strength of the thread.

**(a)** $78.30 \pm 2 \times 5.60$
**(b)** $78.30 \pm 2 \times \frac{5.60}{\sqrt{51}}$
**(c)** $78.30 \pm 2.009 \times 5.60$
**(d)** $78.30 \pm 2.009 \times \frac{5.60}{\sqrt{51}}$ (correct)

**(2)** Assuming the strength of the thread follows the normal model with mean and SD having the values given in the above, estimate the tensile strength that would be exceeded by 97.5% of such threads.

**(a) 67.10 kg**
**(b)** 89.50 kg
**(c)** 76.73 kg
**(d)** 79.87 kg
**(e)** 61.50 kg

22. Eight marksmen, labeled A, B,. . . ,H, shot at targets with two types of rifle. Their scores were as in the table below:

| | Marksman | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | sample mean | sample SD |
| Rifle Type 1 | 93 | 99 | 90 | 87 | 85 | 94 | 88 | 91 | 90.875 | 4.45 |
| Rifle Type 2 | 89 | 93 | 86 | 92 | 78 | 90 | 91 | 87 | 88.25 | 4.77 |
| Difference(Type1-Type2) | 4 | 6 | 4 | -5 | 7 | 4 | -3 | 4 | 2.625 | 4.27 |

**(1)** To perform the hypothesis testing, is there any assumption needed?

**(a)** No assumption is needed.
**(b)** Differences are assumed to follow the t distribution.
**(c) Differences are assumed to follow the Normal distribution.**
**(d)** Differences are assumed to follow the Binomial distribution.

**(2)** When testing the hypothesis that the rifles are of equal quality, what is the test statistic?

**(a)** 1.1381
**(b) 1.7388**
**(c)** 2.4452
**(d)** 2.4590

**(3)** What is the P-value for this hypothesis test?

**(a)** Greater than 0.2
**(b) Between 0.1 and 0.2**
**(c)** Between 0.05 and 0.1
**(d)** Between 0.02 and 0.05

**(4)** Is there a significant difference between the two types of rifles at the 10% significant level?

**(a)** Yes.
**(b) No.**
**(c)** There is insufficient information to tell.

23. How well do the size and age of a house determine the annual tax house owners are paying? Nineteen houses are randomly selected from a city. The age (# years since the house was built), house size (measured in square feet of living space) and the amount of annual tax (in dollars) are recorded for each of the 19 houses. Here are the summary statistics for two of the three variables:

house size : mean = 1456 sqft, standard deviation = 374 sqft
annual tax : mean = \$1707, standard deviation = \$323

**(1)** The linear regression line that predicts the amount of annual tax from the house size has a slope of \$0.81 per square foot. Find the value of the correlation between the size of a house and the amount of annual tax charged.

**(a) 0.94**
**(b)** 0.69
**(c)** 0.70
**(d)** 0.95
**(e)** 0.91

**(2)** Another linear regression line is fitted to predict the amount of annual tax from the age of a house. This regression line has a slope of -\$92.4 per year. Based on the information that is available to you in this question, which of the following is a correct statement?

**(a)** The correlation between house size and amount of annual tax is stronger than that between age of a house and amount of annual tax.
**(b)** The correlation between house size and amount of annual tax is weaker than that between age of a house and amount of annual tax.
**(c)** The correlation between house size and amount of annual tax is the same as that between age of a house and amount of annual tax.
**(d) There is insufficient information to tell.**

**(3)** Predict the annual tax paid by owners owning a 1500-square foot house.

**(a)** \$1215.
**(b)** \$2915.
**(c)** \$1466.
**(d) \$1743.**

24. To compare four treatments via ANOVA, 64 subjects were split at random into four groups each containing sixteen subjects. The test statistic was found to be 21.02. To compute the P-value here we would find the area under the density curve

(a) of $F_{3,16}$ to the left of 21.02.
(b) of $F_{3,64}$ to the right of 21.02.
(c) **of $F_{3,60}$ to the right of 21.02.**
(d) of $F_{3,64}$ to the left of 21.02 and the right of 0.0476.
(e) of $F_{3,60}$ to the right of 21.02 and the left of 0.0476.

25. A study investigated whether month of birth impacts on the time a baby learns to crawl. Parents with children born in January, May or October were asked the age, in weeks, at which their child could crawl one metre within a minute. The data are summarized below:

| | | Crawling age | | |
|---|---|---|---|---|
| | | Mean | SD | size |
| Birth month | January | 29.84 | 7.08 | 34 |
| | May | 28.58 | 8.06 | 29 |
| | October | 33.83 | 6.93 | 40 |

The data from each birth month are assumed to follow a Normal distribution. The analysis is via ANOVA, with an incomplete ANOVA table given below:

| Source | Sum of squares | df. | MS | F |
|---|---|---|---|---|
| Between groups | 505.26 | | | |
| Error | | | 53.45 | |
| Total | | | | |

(a) Which of the following statements you consider to be correct? CHECK ALL THAT APPLY.

_____ It is inappropriate to use ANOVA here since there is evidence that the Normal distributions underlying each sample have different variances.
_____ It is inappropriate to use ANOVA here since the sample sizes are unequal.
_____ The table shown above is a contingency table.
_____ It would be inappropriate to calculate correlations for these data. **(correct)**
_____ This is a randomized block design experiment.
_____ None of the above.

(b) State clearly the null hypothesis for the AVOVA test.
$H_0$ : *The true mean time to crawling is the same for each birth month.*

(c) Compute the test statistic for the test in (b).
*The complete table is*

| Source | Sum of squares | df | MS | F |
|---|---|---|---|---|
| Treatments | 505.26 | 2 | 252.63 | 4.73 |
| Error | 5345.00 | 100 | 53.45 | |
| Total | 5850.26 | | | |

*So the test statistic is 4.73.*

(d) Test the hypothesis at the 5% significance level.
*We compare 4.73 with the* $F_{2,100}$ *distribution. The critical value* $F_{\alpha=0.05,2,100} = 3.09$. *Since the F-ratio 4.73 is larger than the critical value, we would reject the null hypothesis at the 5% level. We conclude that among the three groups of children born in January, May and October, at least one group has a significantly different mean crawling age than the other two.*

26. A multiple choice exam consists of 10 questions, each question having 4 possible answers to choose from. Suppose a student has not studied for the exam, and will make completely random guesses at the answer for each of the questions.

**(a)** What is the probability that this student gets at least 3 answers in the exam correct?
*Let $X = \#$ of correct answers, $n = 10$, $p = \frac{1}{4}$, $X \sim Bin(10, \frac{1}{4})$*

$$\begin{aligned} P(X \geq 3) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - {}_{10}C_0(\frac{1}{4})^0(\frac{3}{4})^{10} - {}_{10}C_1(\frac{1}{4})^1(\frac{3}{4})^9 - {}_{10}C_2(\frac{1}{4})^2(\frac{3}{4})^8 \\ &= 1 - 0.0563 - 0.1877 - 0.2816 \\ &= 0.4744 \end{aligned}$$

**(b)** Now consider an exam of 100 multiple choice questions with each question having 4 possible answers to choose from. If the student will make completely random guesses at all of the answers, is it usual that he gets at least 30 answers correct on the exam? Justify your answer probabilistically.
*$\hat{p} =$ proportion of correct answers among the 100 questions, $n = 100$, and $p = \frac{1}{4}$*
*Since $np = 100 \times \frac{1}{4} = 25 > 10$, $n(1-p) = 100 \times (1 - \frac{1}{4}) = 75 > 10$, we can use normal approximation to sample proportion.*
$p = \frac{1}{4}$, $\sigma_{(\hat{p})} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{1/4(1-1/4)}{100}} = 0.0433$, and $\hat{p} \sim_{approx} N(\frac{1}{4}, 0.0433)$
$P(\hat{p} > \frac{30}{100}) = P(Z > \frac{0.3-p}{\sigma_{(\hat{p})}}) = P(Z > \frac{0.3-0.25}{0.0433}) = P(Z > 1.15) = 0.1251$
*The probability isn't too low, so I think it is not too unusual.*

27. iPods have been criticized for having a battery that doesn't last very long. I am interested in studying a few things about the mean lifetime of a fully charged battery. I randomly sample and test 32 iPods of the same model, and find a sample mean lifetime of 6.15 hours with a sample standard deviation of 45 minutes.

**(a)** Suppose we want to re-estimate the true mean lifetime of the battery using a 95% confidence interval with a margin of error no larger than 10 minutes. How large a sample should we take?

*Note that the t-score depends on the sample size, but the sample size is what we need to solve for here. We will assume that the sample size is large such that the z-score and the t-score have similar values. We will hence use the z-score which is constant (a value of 2 for 95% confidence) for our calculation.*

*Margin of error $ME = 10$ min, $s = 45$ min, $z^* = 2$*

$$n = \left(\frac{z^* s}{ME}\right)^2 = \left(\frac{2 \times 45}{10}\right) = 81 \text{ ipods}$$

*So the sample size should be $n = 81$ iPods.*

**(b)** The company claims that the true mean lifetime of a fully charged battery is significantly greater than 6 hours. Test this claim using a significance level of 5%.
item $H_0 : \mu = 6$ *hours, vs.* $H_A : \mu > 6$ *hours*
*Test statistic is* $t_0 = \frac{\bar{y}-\mu_0}{s/\sqrt{n}} = \frac{6.15-6}{0.75/\sqrt{32}} = 1.131$ *df* $= n - 1 = 32 - 1 = 31$, *P-value*$= P(t_{31} \geq 1.131) \approx P(t_{30} \geq 1.131)$, and $0.10 <$ *P-value*.
*Since P-value $> \alpha = 0.05$, we fail to reject $H_0$ and conclude that there is not enough evidence to say the true mean lifetime of a fully charged battery is greater than 6 hours at a significance level of 5%.*

28. A study was conducted to determine whether an expectant mother's cigarette smoking has any effect on the bone mineral content of her otherwise healthy child. A sample of 30 newborns whose mothers smoked during pregnancy has a mean bone mineral content of 0.092 g/cm and a standard deviation of 0.026 g/cm; a sample of 72 infants whose mothers did not smoke has a mean of 0.105 g/cm and a standard deviation of 0.025 g/cm.

**(a)** Do the data suggest that the population mean bone mineral content of newborns differ between mothers who smoked and those who did not smoke during pregnancy? Use a significance level of $\alpha = 0.05$. Define clearly the parameter(s) and variable(s) that relate to your test.
*Let $y_1$ be the bone mineral content of a newborn whose mother smoked, $y_2$ the corresponding variable for a baby whose mother did not smoke. Let $\mu_1$ and $\mu_1$ be the respective means, $\sigma_1$ and $\sigma_2$ their respective standard deviations. $\sigma_1$ and $\sigma_2$ are unknown. We test*

$$H_0 : \mu_1 = \mu_2$$

*against*

$$H_A : \mu_1 \neq \mu_2.$$

*the test statistic is*

$$t = \frac{\bar{y_1} - \bar{y_2}}{SE(\bar{y_1} - \bar{y_2})} = \frac{\bar{y_1} - \bar{y_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{0.092 - 0.105}{\sqrt{\frac{0.026^2}{30} + \frac{0.025^2}{72}}} = -2.33.$$

*Under $H_0$ this should be from the $t_{29}$ distribution, since* $\min(30 - 1, 72 - 1) = 29$.
P-value $= 2 \times P(t_{29} > |-2.33|)$, *and* $0.01 < P(t_{29} > |-2.33|) < 0.025$, *so* $0.02 <$ *P-value* $< 0.05$.
*Since P-value $< \alpha = 0.05$, we reject $H_0$ and conclude there is a difference between the underlying means.*

**(b)** Based on the results obtained from part (a), you can confidently say that (circle all that apply):

**(i)** smoking causes a decrease in the bone mineral content in the newborns.

**(ii) smoking is associated with the bone mineral content in the newborns.**

**(iii)** smoking has no effect on the bone mineral content in the newborns.

**(iv)** smoking is independent of the bone mineral content in the newborns.

**(c)** Would you expect a 95% confidence interval for the true difference in the population means to contain the value 0? (Circle one)

Yes No(correct)

Briefly justify your answer.
*The two–sided test above is at the 5% significance level and rejects the hypothesis that $\mu_1 - \mu_2 = 0$.*

29. In a certain city, 25% of residents are European. Suppose 120 people are called for jury duty, and only 24 of them are European. Does this indicate that Europeans are under-represented in the jury selection system? Carry out an appropriate hypothesis test at the 1% significance level. Remember to define the parameter(s) that relates to your test.

*Let $p$ be the true proportion called for jury service that are Europeans. We test*

$$H_0 : p = 0.25$$

*against*

$$H_A : p < 0.25.$$

*With $n = 120$, and*

$$np_0 = 120 \times 0.25 = 30 > 10$$

$$n(1 - p_0) = 120 \times (1 - 0.25) = 90 > 10,$$

*then approximately*

$$\hat{p} \sim N\left(0.25, \sqrt{\frac{0.25 \times 0.75}{120}}\right)$$

*under $H_0$. The test statistic is*

$$z = \frac{\frac{24}{120} - 0.25}{\sqrt{\frac{0.25 \times 0.75}{120}}} = -1.\,265.$$

*The P-value is* $P\,(Z < -1.265) = 0.103$ so P-value$> \alpha = 0.01$.

*Hence, we do not reject $H_0$ and conclude there is no evidence to suggest the under–representation of Europeans in the jury selection system.*

30. A medical research is interested in examining the relationship between the duration of catheterization and whether or not an infection occurred. The thought is that whether or not an infection occurs may be related to the duration of catheterization. She collects data on a random sample of 266 patients and the data is presented in the contingency table below.

| | Duration(days) | | | | Total |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | $\geq 4$ | |
| Infection | 5 | 10 | 8 | 18 | 41 |
| No Infection | 46 | 64 | 39 | 76 | 225 |
| Total | 51 | 74 | 47 | 94 | 266 |

**(a)** For this set of data, what is the probability of a person getting an infection given that their duration was between 1 and 2 days?
*Since* $51 + 74 = 125$ *patients had duration between 1 and 2 days, and among them* $5 + 10 = 15$ *patients got infections,*
*P(getting an infection given that duration is between 1 and 2 days)*= $\frac{15}{125} = 0.12$

**(b)** Complete a hypothesis test to decide if whether or not an infection occurred is independent of the duration. Use a significance level of 1% and make sure to state your hypotheses and a conclusion.
$H_0$*: infection and duration are independent vs.* $H_A$*: there is an association between infection and duration*
Table of expected counts:

| Duration (days) | 1 | 2 | 3 | $\geq 4$ | Total |
|---|---|---|---|---|---|
| Infection | $\frac{51\times41}{266} = 7.96$ | $\frac{71\times41}{266} = 11.41$ | $\frac{47\times41}{266} = 7.24$ | $\frac{94\times41}{266} = 14.49$ | 41 |
| No infection | $\frac{51\times225}{266} = 43.14$ | $\frac{74\times225}{266} = 62.59$ | $\frac{47\times225}{266} = 39.76$ | $\frac{94\times225}{266} = 79.51$ | 225 |
| Total | 51 | 74 | 47 | 94 | 266 |

We have counts of categorical variables, random sample of patients. All expected counts are greater than 5.
Under these conditions, the sampling distribution of the test statistic is $\chi^2$ with $(R-1)(C-1) = (2-1)(3-1) = 3$ degrees of freedom, and we will perform a chi-square test of independence.

$\chi^2 = \frac{(5-7.86)^2}{7.86} + \frac{(10-11.41)^2}{11.41} + \frac{(8-7.24)^2}{7.24} + \frac{(18-14.49)^2}{14.49} + \frac{(46-43.14)^2}{43.14} + \frac{(64-62.59)^2}{62.59} + \frac{(39-39.76)^2}{39.76} + \frac{(76-79.51)^2}{79.51} = 2.536$
P-value $= P(\chi^2_3 > 2.536) > 0.1$

P-value $> 0.01$, we fail to reject the null hypothesis. There is insufficient evidence to say infection and duration of catheterization are associated.

**(c)** What is the distribution of the duration, conditioned on them having an infection?
*The distribution of duration conditioned on having an infection is:*

| Duration (days) | 1 | 2 | 3 | $\geq 4$ | Total |
|---|---|---|---|---|---|
| # Infection | 5 | 10 | 8 | 18 | 41 |
| Proportion | $\frac{5}{41} = 0.122$ | $\frac{10}{41} = 0.244$ | $\frac{8}{41} = 0.195$ | $\frac{18}{41} = 0.439$ | 1 |

**(d)** What is the marginal distribution of the duration of catheterization?
*The marginal distribution of duration of catheterization is:*

| Duration (days) | 1 | 2 | 3 | $\geq 4$ | Total |
|---|---|---|---|---|---|
| Total # | 51 | 74 | 47 | 94 | 266 |
| Proportion | $\frac{51}{266} = 0.192$ | $\frac{74}{266} = 0.278$ | $\frac{47}{266} = 0.177$ | $\frac{94}{266} = 0.353$ | 1 |