For questions $1 - 27$, there is *only one correct answer* from those given. Mark your answer to each question with a pencil on the sheet provided. Ambiguous responses will be considered incorrect.

1. If two lists of numbers have exactly the same mean of 30 and the same standard deviation of 5, then the percentage of numbers between 25 and 35 must be exactly the same for both lists.

   (a) True

   (b) **False**

2. If the slope of a regression line is zero, then the correlation coefficient is zero. Likewise, if the slope of the line is one, then the correlation coefficient is one.

   (a) True

   (b) **False**

3. I calculate the residuals from a regression line predicting weight from height. I find that all the residuals are negative. This is impossible; I must have made a mistake in my calculations.

   (a) **True**

   (b) False

4. In drawing a sample from a population, the sampling variability increases with sample size.

   (a) True

   (b) **False**

5. A 90% confidence interval for a population proportion p is constructed from a random sample of size 500 and is found to be 0.48 ± 0.04. If another random sample of the same size is drawn, the 90% confidence interval for p constructed based on this new sample will have a 90% chance of including the value 0.48.

   (a) True

   (b) **False**

6. Farmers concerned with the affect of snowfall levels on their crops found that there wasn't sufficient evidence of a decrease in crop growth. They based this conclusion on a hypothesis test using $\alpha = 0.05$. They would have made the same decision at $\alpha = 0.01$.

   (a) **True**

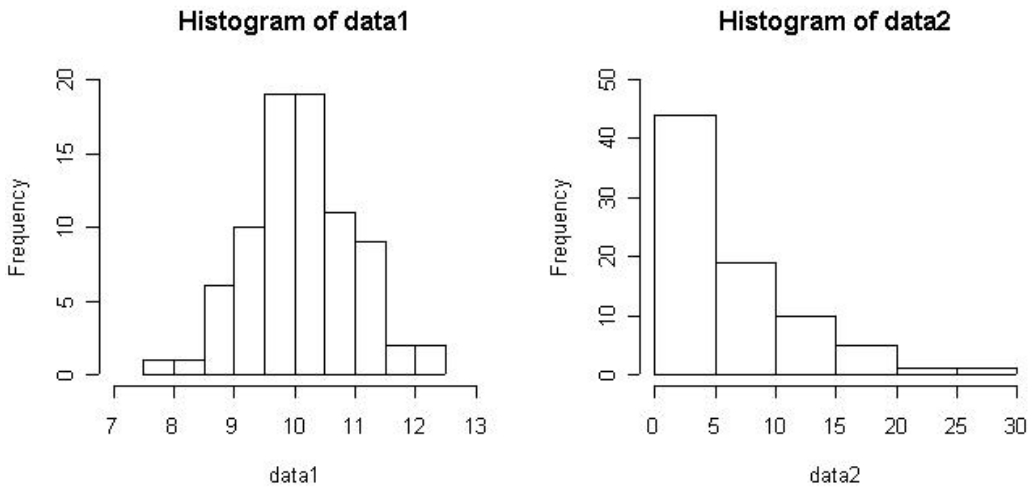
   (b) False

7. One-way ANOVA provides evidence against the null hypothesis that the population means are all equal when the between-group variation is large compared to the within-group variation.

   (a) **True**

   (b) False

8. Consider the following histograms of two data sets (both contains 80 observations):



Which of the following statements is (are) true about the two data sets?

(1) Data 1 has a smaller median that data 2.

(2) Data 1 has a larger standard deviation than data 2.

(3) Data 1 has a smaller IQR than data 2.

(4) The range of the two data sets is approximately the same.

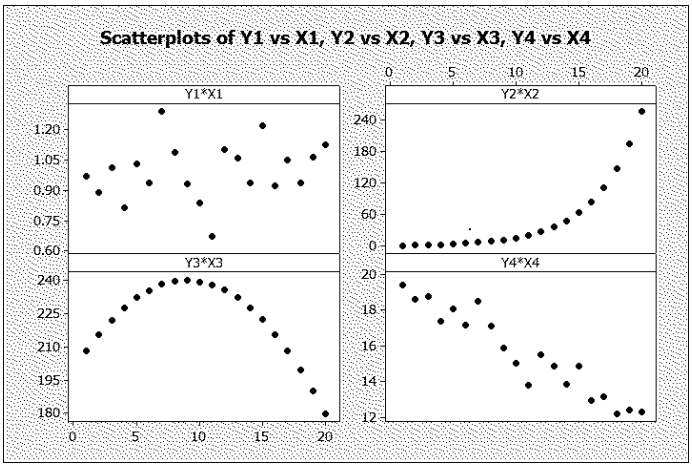
(a) **(3) only**

(b) (1) and (2) only

(c) (1) and (3) only

(d) (1) and (4) only

9. For which of the 4 plots below will you expect the corresponding residual plot to be patternless?



(a) "Y1 * X1" only

(b) "Y4 * X4" only

(c) "Y1 * X1" and "Y2 * X2"

(d) "Y3 * X3" only

(e) **"Y1 * X1" and "Y4 * X4"**

10. Data were collected within a group of males in an athletic association in BC. Based on this dataset, a regression model was computed to predict *weight* Y (in kg) from *height* X (in cm). The model fitted was

$$Y = 0.28X + 27.00.$$

If intending to predict *height* from *weight* using the same dataset, which of the following statements most precisely describes what you can say about the appropriate regression line?

(a) The slope of the regression line would be 0.28.

(b) The slope of the regression line would be 27.00

(c) **The slope of the regression line would be positive.**

(d) The slope of the regression line would be negative.

(e) The slope of the regression line would be 3.57.

11. A study on the accreditation rate of food-and-drink businesses by the Better Business Bureau obtained a sample from the telephone directory's Yellow Page listings of food-and-drink businesses in Greater Vancouver. The investigator of this study first drew a simple random sample of 4 cities in Greater Vancouver. Then within each selected city, he randomly sampled 50 businesses. For various reasons, the study got no response from 39.5% of the 200 businesses chosen. Interviews were completed with 121 businesses that responded.

    (1) The population of interest to the investigator is

        (a) all food-and-drink businesses in Greater Vancouver that are listed under the telephone directory's Yellow Page.

        (b) **all food-and-drink businesses in Greater Vancouver.**

        (c) the 200 businesses that were chosen by the investigator.

        (d) the 121 businesses that responded.

    (2) What is the statistic?

        (a) The proportion of all food-and-drink businesses in Greater Vancouver listed under the telephone directory's Yellow Pages that are accredited by the Better Business Bureau.

        (b) The proportion of all food-and-drink businesses in Greater Vancouver that are accredited by the Better Business Bureau.

        (c) The proportion of the 200 businesses sampled by the investigator that are accredited by the Better Business Bureau.

        (d) **The proportion of the 121 businesses that responded which are accredited by the Better Business Bureau.**

12. An education researcher was interested in examining the effect of teaching method and the effect of the particular teacher on students' scores on a reading test. In a study, there were two different teachers (Juliana and John) and three different teaching methods (method A, method B, and method C). Two hundred and fifty students were randomly assigned to a teaching method and teacher. The researcher wants to compare the scores on the reading test between different treatment (teaching method-teacher combination) groups. Which of the following four display(s) is (are) appropriate for the comparison?

    (1) contingency table

    (2) side-by-side boxplots

    (3) scatterplot

    (4) bar graph

    (a) **(2) only**

    (b) (2) and (4) only

    (c) (1) and (3) only

    (d) (2), (3) and (4) only

13. Consider sampling with replacement from a large population of people. Within the population, the variance of IQ is denoted σ^2. The sample size is n, and the mean of the sample IQs is found. The variance of this sample mean is

    (a) σ^2 provided n is large.

    (b) σ^2 for any value of n.

    (c) σ^2/n provided n is large.

    (d) σ^2/n **for any value of** n.

    (e) $\sigma^2/\sqrt{n}$ for any value of n.

The next four questions (Q14 - Q17) refer to the following situation.

A politician must decide whether or not to run the next local election. He would be inclined to do so if more than 30% of the voters would favour his candidacy. The results of a poll of 225 local citizens showed that 81 favour the politician. Should the politician decide to run the election based on the results of this survey? Carry out an appropriate hypothesis test at $\alpha = 0.01$ to answer this question.

14. What is the parameter of interest?

    (a) 30%

    (b) The proportion of citizens who favor the politician among the poll of 225 local citizens.

    (c) 36%

    (d) **The proportion of all citizens who favor the politician.**

15. Denote p as the parameter of interest. What are the correct hypotheses?

    (a) $H_0 : p = 0.36$ v.s $H_A : p > 0.36$

    (b) $H_0 : p = 0.30$ v.s $H_A : p \neq 0.30$

    (c) $H_0 : p = 0.30$ **v.s** $H_A : p > 0.30$ **(Correct)**

    (d) $H_0 : p = 0.36$ v.s $H_A : p < 0.36$

16. The test statistic is

    (a) 1.88

    (b) -1.88

    (c) **1.96**

    (d) -1.96

17. Which of the following is correct?

    (a) We reject the null hypothesis and advise the politician to run the election.

    (b) **We do not reject the null hypothesis and advise the politician not to run the election.**

    (c) We accept the alternative hypothesis and advise the politician to run the election.

    (d) We do not accept the alternative hypothesis and advise the politician not to run the election.

Use the following information for questions 18 and 19:

Researchers are interested in determining if, during an exam period, SFU undergraduates tend to sleep more than UBC undergraduates. Ten SFU undergraduates were chosen at random and, independently, ten UBC undergraduates were chosen at random. A data file was constructed consisting of a line for each student containing:

ID: student ID,

SCH: School attended (SFU/UBC),

APR16: number of hours slept during the period on April 16 from 12:01 am to 11:59 pm,

APR17: number of hours slept during the period on April 17 from 12:01 am to 11:59 pm.

18. True or false? A good way to study the primary research question is to make a scatterplot of UBC students' numbers of hours slept Apr 16 on the x axis and SFU students' numbers of hours slept Apr 16 on the y axis.

    (a) True

    (b) **False**

19. To test the null hypothesis that SFU undergraduates and UBC undergraduates tend to sleep the same, on average, during exam period, we would need which one of the following?

    (a) the t model with 8 degrees of freedom.

    (b) the t model with 19 degrees of freedom.

    (c) **the t model with 9 degrees of freedom.**

    (d) the binomial model.

Use the following information for questions 20-22:

The owner of a small clothing store is concerned that her average sales each day are only \$149, not enough to cover rent and salary. She decides to try out some new window displays, to see if these will increase her average sales. She buys the new window displays on trial. To decide if she should keep the new displays, she collects sales data for 20 days to test the null hypothesis that the daily expected sales are unchanged (equal to \$149) versus the alternative hypothesis that expected daily sales are greater than \$149.

20. Suppose that the displays really do work. If the store owner extends her trial period from 20 days to 30 days, which statement most precisely describes what can be said about the chance of committing a type II error?

    (a) The chance of committing a type II error would increase.

    (b) The chance of committing a type II error would stay the same.

    (c) **The chance of committing a type II error would decrease.**

    (d) The chance of committing a type II error would remain zero.

    (e) The chance of committing a type II error could be chosen to be 5%.

21. Suppose that, based on the data collected in the trial, the owner calculates a P-value of 0.04. This means

    (a) there is a 4% chance that sales increased during the trial period.

    (b) there is a 4% chance that sales decreased during the trial period.

    (c) during the trial period, sales increased by 4%.

    (d) during the trial period, sales decreased by 4%.

    (e) **during the trial period, her sales figures were pretty high, if indeed the new displays typically would have no effect.**

22. Suppose that, based on the data collected in the trial, the owner of the store decides to keep the new displays. Then

    (a) **she is in danger of making a Type I error.**

    (b) she is in danger of making a Type II error.

    (c) she is in danger of making a Type III error.

    (d) she will get a bigger α.

    (e) she will get a smaller α.

23. Three different labs tested two types of cream, A and B, recording the percentage of solubility in some liquid. Each lab repeated each experiment, and the data are given below:

	Cream type	
Lab	A	B
1	6.8, 6.6	5.3, 6.1
2	7.5, 7.4	7.2, 6.5
3	7.8, 9.1	8.8, 9.1

Differences in the measurements may be due to differences in solubility in the cream types, differences between the labs or both of these possible sources of variation. To investigate this, you could use

(a) Binomial model.

(b) matched pairs t test.

(c) the ANOVA F-test.

(d) linear regression model.

(e) **No method that has been encountered in STAT 200.**

For questions 24–26, consider studying if gender and the highest academic qualification obtained (none, high school diploma, bachelor's degree, post-graduate degree) are independent.

24. True or false? To study independence of gender and the highest academic qualification obtained, it would be useful to construct side-by-side boxplots.

(a) True (b) False(correct answer)

25. True or false? To study independence of gender and the highest academic qualification obtained, it would be useful to calculate a correlation coefficient.

(a) True (b) False(correct answer)

26. True or false? To study independence of gender and the highest academic qualification obtained, it would be useful to calculate a chi-square statistic.

(a) True(correct answer) (b) False

27. Every day Lucky Louie plays a die roll game. He rolls a die five times and counts the number of ones. If he rolls exactly two ones, then he treats himself and buys a Barstucks Macchiato. That is the only way he treats himself. Let X be the number of Macchiatos Lucky Louie buys in the month of June, a month with thirty days. Then X has a Binomial model defined by two parameters, denoted as usual n and p.

(a) What is the value of n here? (circle the correct one)

2 5 10 30(correct answer)

(b) What is the value of p? (circle the correct one)

$\frac{1}{6}$ $(\frac{1}{6})^2$ ${}_5C_2(\frac{1}{6})^2(\frac{5}{6})^3 (correct\ answer)$ ${}_5C_2(\frac{1}{6})^2$

28. In this class last year, there were 211 students who wrote both the midterm exam and final exam. Below you can find summary statistics for the 211 students who wrote both exams.

- The mean grade for the midterm was 80%
- The standard deviation for the midterm grade was 16%
- The mean grade for the final exam was 73%
- The standard deviation for the final exam grade was 12%
- The correlation between the two exam grades was 0.63

(a) What final exam grade would you predict for a student who scored 65% on their midterm?
Let X = midterm grade, Y = final exam grade
Given, $\bar{x} = 80\%, \quad s_x = 16\%, \quad \bar{y} = 73\%, \quad s_y = 12\%$
$b_1 = r \times (\frac{s_y}{s_x}) = 0.63 \times \frac{12}{16} = 0.4725$ *, and* $b_0 = \bar{y} - b_1\bar{x} = 73 - 0.4725 \times 80 = 35.2\%$
So, $\hat{y} = b_0 + b_1 x = 35.2 + 0.4725x$
when $x = 65\%, \quad \hat{y} = 35.2 + 0.4725 \times 65 = 65.9125\%$
That is, we predict a final grade of 65.9125% for a student who scored 65% on the midterm.

(b) Write an interpretation of the slope in the context of this example.
For every 1% increase in the midterm grade, we expect a 0.4725% increase in the student's final exam grade.

(c) Suppose that another students' grades were added to the dataset. This student scored 10% on the midterm and 92% on the final exam. Would this new observation affect the correlation, and if so, how?
This new observation would decrease the correlation. This is because it deviates largely from the regression line and is likely to increase scatter.

29. Suppose that Math Proficiency scores of 12th graders are Normally distributed with mean 80 and standard deviation 12. What is the probability that the average Math Proficiency score of a random sample of 400 students is between 80 and 81.2?
Let Y be the Math Proficiency score of 12th graders. We have $Y \sim N(80, 12)$
$\bar{y}$ is the average Math Proficiency score of a random sample of 400 students
So,

$$\bar{y} \sim N(80, \frac{12}{\sqrt{400}}) \iff N(80, 0.6)$$

$P(80 < \bar{y} < 81.2) = P(\frac{80-80}{0.6} < \frac{\bar{y}-80}{0.6} < \frac{81.2-80}{0.6}) = P(0 < Z < 2) = 0.475$ *(By the 68-95-99.7 rule)*

30. A manufacturing company has 2 different instruments they use to measure the Rockwell hardness of an object. They believe that one of the instruments may not be working properly, and giving readings that are not completely accurate. To test this, they do the following. They take a large sheet of metal, and cut it into 60 different pieces, and randomly divide the pile in two. They believe it is safe to assume that the hardness of the metal is the same for any particular piece cut from the same sheet of metal. They measure the Rockwell hardness of 30 randomly selected pieces of metal using instrument 1 and find a sample mean hardness of 45.8 with a sample standard deviation of 1.2. They measure the hardness for the other 30 pieces of metal using instrument 2 and find a sample mean hardness of 46.2 with a sample standard deviation of 1.1.

(a) Make and interpret a 95% confidence interval for the difference in the mean hardness readings using instrument 1 and instrument 2.
Given $n_1 = 30, \quad \bar{y}_1 = 45.8, \quad s_1 = 1.2$ *, and* $n_2 = 30, \quad \bar{y}_2 = 46.2, \quad s_2 = 1.1$
$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{1.2^2}{30} + \frac{1.1^2}{30}} = 0.2972$
$df = \min(n_1 - 1, n_2 - 1) = 29$ *, and* $t^*_{29} = 2.045$

$(\bar{y}_1 - \bar{y}_2) \pm t^*_{29} SE(\bar{y}_1 - \bar{y}_2) = (45.8 - 46.2) \pm 2.045 \times 0.2972 \Longrightarrow (-1.0078, 0.2078)$
Therefore, we are 95% confident that the true mean hardness readings determined by instrument 1 is between 1.0078 lower and 0.2078 higher than the true mean hardness readings determined by instrument 2.

(b) Based on this interval alone, would you reject or fail to reject the null hypothesis that the mean hardness readings are equal for the two measurement instruments? Make sure to justify your answer with a reason.
Based on this interval, we would fail to reject the null hypothesis because zero is within the interval.

(c) Complete a hypothesis test to decide if you think that the two measurement instruments result in significantly different readings in the hardness of an object. Use a significance level of 5%.
Let μ_1 be the true mean hardness reading using instrument 1, and μ_2 be the true mean hardness reading using instrument 2. Here, we test

$$H_0: \quad \mu_1 = \mu_2$$

against

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

From part (a), $SE(\bar{y}_1 - \bar{y}_2) = 0.2972$.
Test statistic is $t_0 = \frac{\bar{y}_1 - \bar{y}_2}{SE(\bar{y}_1 - \bar{y}_2)} = \frac{45.8 - 46.2}{0.2972} = -1.346$
df $= \min(n_1 - 1, n_2 - 1) = 29$
P-value$= 2P(t_{29} > |t_0|)$, *and* $0.05 < P(t_{29} > 1.346) < 0.1$, *so* $0.1 <$P-value< 0.2.
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 hardness readings from the two instruments are significantly different at 5% significance level.*

31. A geneticist is interested in studying the function of a particular gene. As a starting point to the study she wishes to determine if the gene's expression is the same in a stem cell, a mesoderm cell and a neuronal cell. She estimates the gene's expression for each cell type a few times. A summary of the measurements can be seen in the table below.

Cell Type	n	Mean Expression	SD of Expression
Stem	11	10.2	1.8
Mesoderm	9	7.5	1.5
Neuronal	11	6.9	2.2

(a) Test if the mean gene expression is the same for each type of cell, using a significance level of 5%.

Let μ_1 be the true mean gene expression for a stem cell
Let μ_2 be the true mean gene expression for a mesoderm cell
Let μ_3 be the true mean gene expression for a neuronal cell

$$H_0: \mu_1 = \mu_2 = \mu_3 \text{ vs. } H_A: \mu_j \neq \mu_{j'} \text{ for some } j \neq j'$$

Since $N = 11 + 9 + 11 = 31$, $k = 3$, we have

$$\overline{\overline{y}} = \frac{11 \times 10.2 + 9 \times 7.5 + 11 \times 6.9}{31} = 8.245$$

$$SS_T = \sum_{j=1}^{k} n_j (\overline{y}_j - \overline{\overline{y}})^2 = 11 \times (10.2 - 8.245)^2 + 9 \times (7.5 - 8.245)^2 + 11 \times (6.9 - 8.245)^2 = 66.937$$

$$
\begin{aligned}
SS_E &= \sum_{j=1}^{k}(n_j-1)s_j^2 = (11-1)\times 1.8^2 + (9-1)\times 1.5^2 + (11-1)\times 2.2^2 = 98.8 \\
MS_T &= \frac{SS_T}{k-1} = \frac{66.937}{3-1} = 33.469 \\
MS_E &= \frac{SS_E}{N-k} = \frac{98.8}{31-3} = 3.529 \\
F &= \frac{MS_T}{MS_E} = \frac{33.469}{3.529} = 9.48
\end{aligned}
$$

The critical value is $F_{\alpha=0.05,2,28} = 3.34$. Since F is larger than the critical value, we reject the null hypothesis and conclude that the true mean gene expression for at least two of the cell types are significantly different at the 5% significance level.

(b) If your conclusion from part (a) were incorrect in reality, then what type of error would you have made? State what this would mean in the context of the example.

If the conclusion from (a) were incorrect in reality, we would have made a Type I error by rejecting the null when the null is indeed true. In the context of this question, this means we concluded that the true mean gene expression for at least two of the cell types are significantly different, when in reality the true mean expression for all three cell types are equal.

32. What affects how a person chooses "at random"? Each of 92 randomly sampled university students was given a slip of paper that said

"Randomly choose one of the letters S or Q".

Of these 92 students, 61 chose S. The remaining 31 students chose Q. Another 98 randomly sampled university students were given a slip of paper that said

"Randomly choose one of the letters Q or S".

Of these 98 students, 45 chose S. The remaining 53 students chose Q.
Is there an association between how the students responded and the ordering of the letters in the question? Carry out the appropriate test at level 0.05. Clearly show the calculation of your test statistic and your rejection rule (in particular, clarify which of the tables provided you have used, if any). State your conclusion, in the context of this problem.

The null and alternative hypotheses are as follows:

H_0: How students responded is independent of the order of the letters in the question.

H_A: How students responded is associated with the order of the letters in the question.

To perform the hypothesis testing, we first tabulate a table with observed and expected counts as follows:

	Choosing S	*Choosing Q*	*Total*
Order 1: S or Q	*61 (51.33)*	*31 (40.67)*	*92*
Order 2: Q or S	*45 (54.67)*	*53 (43.33)*	*98*
	106	*84*	*190*

(Expected counts are in brackets, e.g. for the top left cell, $E = \frac{92\times 106}{190} = 51.33$.)

We have counts of categorical varaibles, random sample of students. All expected counts are greater than 5. Under these conditions, we adopt a Chi-square test with degrees of freedom

$$df = (r-1)\cdot(c-1) = (2-1)\cdot(2-1) = 1.$$

The test statistic is as follows:

$$\chi^2 = \frac{(61 - 51.33)^2}{51.33} + ... + \frac{(53 - 43.33)^2}{43.33} = 7.1903.$$

At 5% significance level, the critical value is 3.84. Since $\chi^2 > 3.84$, we reject the null hypothesis and conclude that the data provide sufficient evidence (i.e. at 5% significance level) that how students responded is associated with the order of letters in the question.