The following is a pre-test that was used at the start of STAT 335 for four years. The course is a second course in Statistics, with most students having previously taken either STAT 200 or FRST 231. The purpose of the test was to gauge students’ retention of concepts from their first course in Statistics.

Questions with only one correct response are worth one mark each, the others are worth two marks. Overall the students performed poorly on the test, with mean scores of only around 50% typical. Last year Andy Leung investigated the results from the first sitting of (and in fact a slightly different version of) the test, and found that students who had taken more Statistics courses did not tend to do better on the pre-test, although those students did tend to perform better on the course overall.

Questions 5, 7, 10, 11, 12, 13, 15, 16, 18 and 20 in the following tend to be answered the worst. For each of these, I attempt to indicate which concepts were being targeted, common student misconceptions (and how those link to the distractors), plus any perceived problems with the questions.

1. The variance of a sample is zero. Which of the following must be true? (Choose all that apply.)
   (a) The sample mean is zero.
   (b) The sample median is nonzero.
   (c) The sample standard deviation is zero.
   (d) The sample IQR is zero.
   (e) The sample total is zero.

2. Which of the following data sets is/are likely to have a mean that is smaller than the median? (Choose all that apply.)
   (a) The salaries of the employees of a large company.
   (b) The assessed value of homes in Vancouver.
   (c) The percentage scores of students on an easy exam where most scored over 90%, but a few scored very poorly.
   (d) The percentage scores of students on a difficult exam where most scored less than 60%, but a few scored very highly.
(e) The percentage scores of students on an exam where the scores were approximately Normally distributed.

3. To get the average number of children per family in a small town, a teacher counted the total number of children in the town. She then divided by 50, the total number of families. The average number of children per family was 2.2. Which of the following is certain to be true?

(a) Half of the families in the town have more than 2 children.
(b) More families in the town have 3 children than have 2 children.
(c) **There is a total of 110 children in the town.**
(d) The most common number of children in a family is 2.
(e) None of the above.

4. The UBC Alumnae Association collects data on each UBC graduate in the last 10 years: 4th year grade point average and how much he/she contributed to UBC within a year of graduation. The Association wants to see if a student’s fourth year grade point average can be used to predict the amount of money he or she will contribute to UBC within a year of graduation.

What sort of graph will be the most help to the Alumnae Association?

(a) **A scatterplot with** \( x = 4\text{th year grade point average} \) and \( y = \text{amount of money contributed} \).
(b) A scatterplot with \( x = \text{amount of money contributed} \) and \( y = 4\text{th year grade point average} \).
(c) A histogram of the total amount of money contributed.
(d) A histogram of the 4th year grade point averages.
(e) A pie chart showing the proportion of alumnae who contributed.

5. The (sample) correlation coefficient between two variables \( x \) and \( y \) is zero. You can deduce that (Choose all that apply.):

(a) \( x \) and \( y \) are not related at all.
(b) the regression line $\hat{y} = b_0 + b_1x$ has a slope of 0.
(c) all the residuals of the regression line $\hat{y} = b_0 + b_1x$ are 0.
(d) all the points $(x, y)$ fall on a horizontal line.
(e) either the variance of $x$ or the variance of $y$ is zero.

- Concept: Two variables could have a strong non-linear relationship while having zero correlation.
- Misconceptions: (i) $r = 0$ implies that there no relationship (hence (a)), (ii) $r = 0$ implies one or other s.d. must be zero (e).
- Problem: The question contains some notation and jargon (correlation, regression, residuals).

6. A police study into robberies gathered detailed information about each burglary at properties that occurred in a Canadian city over the period of a year. Which of the following variables is/are categorical? (Choose all that apply.)

(a) the time of the robbery
(b) the number of witnesses to the crime
(c) whether or not anyone was injured in the incident
(d) the proportion of crimes in which the police detained a suspect
(e) the type of property in which the robbery occurred

7. An engineer has created a new type of light bulb, designed to be long-lasting. A company is interested in distributing the bulb, but decide firstly to test a sample of thirty bulbs. The thirty bulbs are left on until each one expires, and the lifetime of each bulb is recorded. The company tests a hypothesis using the mean lifetimes of these thirty bulbs. A parameter of interest here would be (Choose all that apply.)

(a) the mean lifetime of the sample of thirty bulbs.
(b) the sample size.
(c) the mean lifetime of the new type of bulb.
(d) whether the new bulb seems to last longer than another type of bulb.
(e) the distribution of the lifetimes of the new bulb.

- **Concept:** The meaning of *parameter* in a statistical context.
- **Misconceptions:** The sample statistics (a and b), hypotheses (d) or distributions (e) can be parameters.
- **Problem:** This is a vocabulary check: do they understand what “parameter” means in a Statistics course?

8. Consider two events A and B defined on the same sample space. It is given that \( P(A \text{ and } B) = 0.16 \). Then one can deduce that (choose all that apply)

(a) the two events A and B are not independent.
(b) **the two events A and B are not disjoint.**
(c) \( P(A) \geq 0.4 \).
(d) \( P(B^c) \geq 0.6 \).
(e) event A is not the complement of event B.

9. A Canadian male is rushed to hospital in an ambulance. Which of the following is least likely?

(a) The man is over 55 years old.
(b) The man is over 56 years old.
(c) The man has had a heart attack.
(d) The man is over 55 years old and has had a heart attack.
(e) **The man is over 56 years old and has had a heart attack.**

10. Someone tosses a fair coin three times and announces that he flipped at least one tail. What is the probability that he flipped three tails?

(a) \( 1/3 \)
(b) \( 1/4 \)
(c) \( 1/6 \)
(d) **1/7**
11. Twenty fair dice are rolled independently. The number of dice that produce an odd score is a random variable that (circle all that apply):

(a) has a mean of 5.
(b) has a mean of 10.
(c) has a variance of 5.
(d) has a variance of 25/9.
(e) has a variance of 10.

- **Concept:** Mean and variance of a Binomial distribution.
- **Misconceptions:** Confusing formulae for mean and variance (hence (a) and (e)).
- **Problem:** As long as a student recognizes it is a Binomial set-up (though one suspects not many do!), recollection of formulae is all that is required.
12. Recall that the variance of a sample is usually defined as

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.
\]

We take a random sample of size 15 from a very large population, and compute the sample variance \( s^2 \). You then take a random sample of size 150 from the same population, and again compute the sample variance. Which of the following best describes how we should expect the two sample variances to compare?

(a) The two variances would be identical.
(b) The second variance would be expected to be larger than the first.
(c) The second variance would be expected to be smaller than the first.
(d) We would not expect the second variance to be much bigger or much smaller than the first.
(e) It is impossible to say anything about the relative values of the two variances.

- **Concept:** The sample variance is a consistent estimator of the population variance.
- **Misconceptions:** Most students (around 90%) believe the sample variance shrinks as the sample size increases, apparently seduced by the \((n - 1)\) in the denominator (c).
- **Problem:** Is providing the formula for \( s^2 \) distracting? Is the wording clear?

13. Which of the following statements is/are correct about the sampling distribution of the sample mean, and hold for any sample size? Choose all that apply.

(a) The standard deviation of the sample mean is a measure of the variability of the sample mean among repeated samples.
(b) The standard deviation of the sample mean will decrease as the sample size increases.
(c) The sampling distribution of the sample mean shows how the sample data were distributed around the sample mean.

(d) **The sampling distribution of the sample mean shows how the sample mean will vary among repeated samples.**

(e) The sampling distribution of the sample mean is exactly Normal.

- **Concept:** Properties of the sampling distribution of the sample mean.
- **Misconceptions:** The sampling distribution of $\bar{x}$ indicates how the data are distributed (hence (c)), the sampling distribution will be Normal whatever the sample size (e).
- **Problem:** The term “sampling distribution” is meaningless jargon to a student prior to taking a Statistics course.

14. In 1997, census results indicated that the age at which Canadian men retired had a mean of 58.5 years. In 2007, a random sample was taken of 500 men who have retired. Their ages at retirement had a skewed distribution with mean 60.3 years and standard deviation 6 years.

Some sociologists believe that the mean age at retirement now, $\mu$, is greater than that from 10 years ago. To test this, which of the following hypotheses should you compare?

(a) $H_o: \mu = 60.3$ vs. $H_a: \mu = 58.5$
(b) $H_o: \mu = 60.3$ vs. $H_a: \mu > 60.3$
(c) $H_o: \mu = 58.5$ vs. $H_a: \mu > 58.5$
(d) $H_o: \mu = 60.3$ vs. $H_a: \mu \neq 58.5$
(e) $H_o: \mu = 0$ vs. $H_a: \mu > 0$

15. Refer to the scenario described in the previous question. Which distribution should you use to construct an approximate 99% confidence interval for the mean age at retirement in 2007?

(a) **The standard Normal distribution.**
(b) The Chi–squared distribution with 499 degrees of freedom.
(c) The t-distribution with 49 degrees of freedom.
(d) The t-distribution with 9 degrees of freedom.
(e) None of the above because the distribution of ages at retirement is not Normal.

- **Concept:** With a large sample size ($n = 500$ here), Normal approximations will work fine when finding confidence intervals for a mean.
- **Misconceptions:** Failure to appreciate that $t_{499} \approx N(0,1)$ (but it would be incorrect to give $t_{499}$ as a distractor, though two t distributions are given in (c) and (d)); non-Normal data makes inference impossible (e).
- **Problem:** Links to previous question, so if that was misunderstood, this will probably be done poorly.

16. In a hypothesis testing procedure, suppose the null hypothesis is rejected if a test statistic, $W$, is too large. Suppose the test with significance level $0.05$ rejects the null hypothesis if $W > 4.2$. Which of the statements below is/are true? (Choose all that apply.)

(a) **With this level 0.05 procedure, the probability of rejecting the null hypothesis when it is true is 0.05.**
(b) With this level 0.05 procedure, if the null hypothesis is true, the probability of accepting the null hypothesis is 0.05.
(c) A level 0.10 hypothesis test rejects the null hypothesis if $W > c$, where $c$ is some number larger than 4.2.
(d) **A level 0.10 hypothesis test rejects the null hypothesis if $W > c$, where $c$ is some number smaller than 4.2.**
(e) If the p-value is 0.07, then $W$ must be bigger than 4.2.

- **Concept:** Link between significance level and critical value.
- **Misconceptions:** Confusing meaning of significance level (b, c), misunderstanding p-value (e).
- **Problem:** Relies on some recollection of hypothesis testing. Lack of context maybe makes question too abstract.
17. The point estimate of the mean of a distribution (or population) which has lowest variance is the sample

(a) mean
(b) standard deviation
(c) mode
(d) interquartile range
(e) median

18. Researchers at a university planted Poplar Clone 252 on two different sites: one, a rich site by a creek, and the other, a dry, sandy site on a ridge. They measured the diameter in centimeters, height in metres, and the dry weight of wood in kilograms for a sample of three-year-old trees so that they could determine the weight of a tree from its diameter and height measurements. A plot of dry weight against height for the trees is given below, including the regression line:

Suppose that an additional tree was added to the plot above that had a height of 2.5m and a dry weight of 0.3 kg. With respect to the regression line for the new data, adding this point to the plot above would

(a) decrease the intercept, decrease the slope.
(b) decrease the intercept, increase the slope.
(c) **increase the intercept, decrease the slope.**
(d) increase the intercept, increase the slope.
(e) not affect the regression line.

- **Concept:** Leverage and influence in regression
- **Misconceptions:** Not clear, but many seem to get this wrong!
- **Problem:** Are the words “intercept” and “slope” the best to use? Is the question somehow misconstrued as to reverse the intended direction of the changes?

19. Based on a random sample of 30 from a Normal distribution, a 95% confidence interval for the mean was computed. Using the same sample, a 99% confidence interval would be

(a) the same width
(b) possibly narrower, possibly wider
(c) impossible to determine from such a sample
(d) narrower
(e) **wider**

20. Recall that in the analysis of variance (ANOVA), the data are in groups. In performing the usual hypothesis test, which of the following is/are assumed? (Choose all that apply.)

(a) The within–group sample means are all equal.
(b) The within–group sample standard deviations are all equal.
(c) **The observations within a group are from a Normal distribution.**
(d) The observations within each group are independent.
(e) Observations in different groups are independent.

- **Concept:** Assumptions in ANOVA.
• **Misconceptions:** Assumptions relate to sample statistics (a and b), failure to recall other assumptions.

• **Problem:** Students could answer this from memory despite having little understanding of ANOVA.