STATS 2244 Final Exam Workbook Solutions

Activities

Summer 2024

Activity 1

Q1. Causative. The answer is A).

Q2. Young Children. The answer is A).

Q3. The interest in soccer is a response variable. The answer is B).

Q4. **Student answer 1 (full marks):**

The sampling frame the Kidz soccer club planned to use is young children aged 4-9 who are part of their own club. The population of interest for the Kidz soccer club research question is young children. One of the negative consequences of the sampling frame used is undercoverage bias. Undercoverage bias occurs when some groups of the population (young children) are left out of the sampling frame. This undercoverage bias occurs in this sampling procedure immediately when the Kidz soccer club picks its own young children to be a part of the sampling frame. This could influence how representative the sample is of the population because the sampling frame chosen could miss many young children who play in different leagues and clubs and young children who do not live in Waterloo. In my experience with being a tennis instructor, I see many kids that go to all types of different clubs for reasons such as convenience, finances, racial inclusivity, gender inclusivity, and socioeconomic status which proves that not all types of young children will have a fair representation using this sampling frame. This could affect the conclusion drawn for the research question because different kids that come from different households, economic statuses or with different experiences prefer certain learning methods over others. This would present a problem because the study does not include young children who can't afford to play in the Kidz soccer club (a group of young children that are not represented). These specific young children could have a significant impact on the results of the study if they could be included. This shows the study's statistics about PTM encouraging greater interest and retention of young children than more traditional soccer practices will be biased towards people who can afford to play and choose to play at the Kidz soccer club. Due to these discrepancies, the young children are not fairly accounted for and the study's response variable will not be accurate compared to the population parameter.

Student answer 2 (full marks):

One concern of Kidz Soccer Club's plan that may limit how well their data will reflect the population of interest being the young children given the research question, is their sampling frame and how it produces undercoverage and non-responsive bias. The sampling frame for this study are the children aged 4 to 9 from 20 teams that were selected from the four leagues (North, South, East, and West) of the Kidz Soccer Club in Waterloo, Ontario in 2024. Unfortunately, the sampling frame chosen does not accurately represent the population of interest and has undercoverage bias. Undercoverage bias occurs when parts of the population are less represented or excluded in the sample. The chosen sampling frame does not include roughly 50% of young children from the study. Many young children have not been included in the data collection as the study does not consider these children playing in other leagues, other teams which were not selected due to the sampling choice, and the bias towards children, socioeconomic status, different training, and their parents who did not respond to the questionnaire. The responses from those who chose to respond to the questionnaire will be taken highly, however, the study does not include the families who have chosen not to respond which will create a lack of accuracy and result in non-responsive bias. Based on my experiences as a basketball player, I understand the situation as I have also joined different leagues due to the high cost and long commuting time. Consequently, I have other team mates who have left our league to join another team further away due to their preference in the coaching style and structure, they are also able to afford the higher fees and demanding commuting time. Additionally, the undercoverage bias can also be due to the young children's families who could not afford the Kidz Soccer Club fees and they might have been excluded from the research study as well. Given these discrepancies, many young children are not being adequately accounted for, which means the response variable and statistics will be inaccurate when measured

Question 4

Critique **1 / 1 pt**

✔ **– 0 pts** Consistent with criteria for full credit

Short answer question addresses all parts of question AND content of the submitted answer demonstrates an understanding of the concepts that apply to the question, even if their application isn't completely correct.

Your answer achieved the following:

- identified the sampling frame clearly (but not necessarily correctly); the sampling frame was the "teams and players in the Kidz Soccer Club in 2024".
- identified the population of interest clearly (but not necessarily correctly); the population of interest was "young children"
- identified a negative consequence of the stated sampling frame and discussed that issue with respect to generalizability
- attempted to use relevant vocabulary taught in 2244 for that concern; undercoverage bias is a logical vocabulary term when discussing the alignment between sampling frame and population of interest. Other terms related to *sampling* in general that might apply are selection/sampling bias, voluntary response bias, cluster sampling, random sampling, etc.)
- referenced information about the scenario (as opposed to generically speaking about populations, sampling frame, bias in the discussion)

Activity 2

Q1. Experimental, completely randomized

Since the explanatory variable is being assigned by the researcher, it is experimental and therefore not an observation study, so right away "case control", "cohort" and "observational" are not included. A survey is also not correct as it is an observational study design, however, don't think that "questionnaire" is equivalent to a survey. A questionnaire is a means to collect data that can be used both in experimental or observational studies. A survey is a study design specific to observational studies. This study is completely randomized because the entire sample of 350 individuals were randomly assigned directly to the treatment groups, without dividing them into blocks. Remember that blocking has to occur AFTER you sample so the use of home departments is just part of the stratified sampling method used.

Q2. Version 1: Ordinal. The answer is A).

The variable collected to answer the question, "How many times did you meet with your mentor during the past term" is ordinal because "no times", "one time", "two to four times", etc implies an ordering. It is a categorical variable, as there is no measurements going on, so it can't be ratio or interval.

Version 2: Nominal. The answer is D).

This variable is categorical since the answers are just "yes" or "no", so there is no measurement going on. As a result, it can't be B) or D). Since there is no "order" involved, just labelling, it is nominal and not ordinal.

Q3. Version 1:

The correct answers are:

Average grade of first year courses is a response variable measurement.

There are four comparison groups.

There were four comparison groups in this study. There were two explanatory variables; the identity of the mentor and the type of mentoring. Type of mentoring is not a treatment, as a treatment must be a combination of these two explanatory variables, for example, "staff mentor/hidden curriculum". In this study, there is no "matching/pairing". There are four treatments and they are all independent as there is NO explicit pairing or matching of responses across all of the treatment groups. The identity of the mentor is nominal and therefore, a categorical variable. Home department is not a blocking variable, it is a stratifying variable. Remember, home department is used during the sampling process, i.e. how the individuals who will be in the study are selected. Blocking must occur after our sampling when we subdivide the individuals in order to assign them to various treatments.

Version 2:

The correct answers are:

Type of mentoring is a categorical variable.

Home department is a stratifying variable.

There were four comparison groups in this study, not only two. There were two explanatory variables; the identity of the mentor and the type of mentoring. Type of mentoring is not a treatment, as a treatment must be a combination of these two explanatory variables, for example, "staff mentor/hidden curriculum". Average grade of first year courses is a response variable not an explanatory variable. In this study, there is no "matching/pairing". There are four treatments and they are all independent as there is NO explicit pairing or matching of responses across all of the treatment groups.

Q4. **Student answer 1: (Full marks)**

First-generation status is a confounding variable since being a “first generation” student could have an impact on academic achievement (the response variable) which would make it impossible to separate its effects compared to the impact that the mentor type and mentor identity (factors of interest) would have on the response variable. One way the study design could be improved is the dean could incorporate the generation of students as a new variable into the study as cofactors by using a study design called randomized blocking. Another way the study design could be improved is by using the repeated measures design which is also a form of blocking. In the repeated measures design, blocking occurs at the most detailed level, each student of the sample of 350 would experience all 4 treatments (combinations of levels of factors of interest) in a randomized order (whether they are first generation or not) and we will collect data for academic performance (response variable) after each of the 4 treatments. Then, we will compare the difference in data for academic performance for each individual after each of the 4 treatments separately. In contrast, a randomized block design would subdivide the individuals in the sample of 350 students based on whether they are first-generation or not. Then, we would randomly assign individuals from these blocks to each treatment. The results of the survey are all compared together in the end. The repeated measures design is a better choice because it is a form of blocked design but has an extra layer of structure since there is explicit pairing in the data collected after each treatment. This study design prevents any variables such as “first generation” and variation in individuals (such as commuting or staying on campus and socioeconomic status) from affecting the response variable (academic achievement) in a confounding way across the 4 treatments since each treatment group is composed of the same individuals. This allows us to compare differences in academic achievement data (response variable) after all 4 treatments per student to conclude which type of mentoring and identity of mentor combination worked best for the majority of students since all students in the sample tried all the treatments. This shows that the "generation" of students is no longer confounding since we have equal representation in all treatment groups. Although the blocked design ensures that the variation among blocks concerning the confounding variable is distributed in all 4 treatment groups with randomization, the treatment groups will never be identical like they are in the repeated measures design. When the treatment groups are not identical, there will always be other confounding variables that could affect the response variables. We would have equal representation in all treatment groups in terms of the confounding variable but not any other variables since we are leaving that to randomization. The repeated measure ensures that all treatment groups are composed of the same individuals improving the quality of the study.

Student answer 2: (Full Marks)

To address the concern regarding first-generation status, the Dean could implement a randomized block or repeated measures design. In the randomized block approach, the sample of 350 students would be divided

into two blocks based on prior characteristics: first-generation students and non-first-generation students. Next, the students from the two blocks are randomly assigned to the four different treatment groups. By utilizing a randomized block design, researchers can assess whether there is a correlation between academic achievement and the blocking variable of first-generation status. This design enables multiple levels of comparison and the results (response variable) can be understood efficiently. Additionally, the Dean could consider incorporating a repeated measures design to further account for generational status. In a repeated measures design, blocking occurs at the individual level which offers the highest level of control over other explanatory variables that might affect the result (response variable).Moreover, each student (first generation or not) undergoes all four treatments over the course of the four terms of the

study, with the order of treatments randomized for each student. The researcher will repeat the measurement of the variable which comes from the result of the survey for the treatment of each individual. This method helps account for factors such as the potential effectiveness of a treatment depending on the term it was administered. In a repeated measures design, blocking occurs at the individual level, offering the highest level of control over other explanatory variables that might affect the response variable. For the response variables, the comparison involves examining differences in survey results across the four treatments. Each treatment group provides a distinct set of survey outcomes. The advantage here is that variation among units cannot account for the variation in response variables across treatments, as the composition of individuals does not differ between treatments. Furthermore, I propose that a randomized block design would be preferable. This design allows the Dean to conduct focused comparisons across various levels. For instance, comparisons can be made within blocks, such as comparing survey results among first-generation students across treatments. Additionally, comparisons can be made across blocks, such as evaluating the same treatment while considering whether a student is first-generation or not. This approach enables the Dean to control confounding variables by blocking on first-generation status, ensuring that differences in survey results are not influenced by whether a student is first-generation. Analyzing within and across blocks clarifies how first-generation status interacts with treatment, providing actionable insights for the Dean. In contrast, a repeated measures design over two years risks carryover effects and doesn't isolate first-generation status as well as randomized block designs.

Question 4

improvement 1 / 1 pt

✔ – 0 pts Consistent with criteria for full credit

Short answer question addresses all parts of question AND content demonstrates an understanding of the concepts that apply to the question, even if their application isn't completely correct.

Your answer achieved all of the following:

- identified TWO changes that could be made to the study design (e.g. implement blocking, collect data on the cofactor of first gen status, use a repeated measures design)
- explicitly identified how the changes would apply to the scenario (i.e. block by what variable, what would you do to get the cofactor information? what would the repeated measures look like? etc.)
- the changes were to the study design, not to aspects related to sampling (sampling would be narrowing the sampling frame, or choosing a stratified sample, etc.)
- used the vocabulary to support the answer (e.g. 'blocking', randomization, confounding, etc)
- gave a logical reason for why the selected approach is better for quality/preventing the confounding

Activity 3

Q1. Sampling error

Your answer has correctly identified that the difference between the value of 11.3 degrees C—the parameter--and the observed value of 7.6 degrees C (a statistic) is sampling error.

Q2. The correct options were:

1. The distribution of standard deviations for annual snowfall (cm) based on different samples of 10 participants
2. The distribution of mean January rainfalls (mm) that could be collected from stratified samples of 30 London ON residents.

Note that the "distribution of daily rainfall (mm) in London, ON" would be considered a population distribution, while the "distribution of responses about human experience of daily temperature recorded by the meteorologist-to-be" would be considered a distribution for a sample.

Q3. The standard deviation of the distribution is 1.3°C

Q4. Student Answer 1 (Full Marks):

There are 4 conditions for a binomial distribution:

1. The first condition is that there are a fixed number “n” of observations or trials. The observations are the participants/individuals. This condition is met since there are 10 participants in the sample (n=10). This number is fixed and it is known in advance.
2. The second condition is that the "n" observations are independent which means knowing the result of 1 observation does not change the probabilities we assign to other observations. The researcher used simple random sampling without replacement and stated that they chose 10 different people on a singular day in the month for 4 days per month. Although in this design there was no replacement, the trials are reasonably considered to be independent because the sampling process was taking simple random samples to obtain a sample of 10 individuals which is a relatively small sample compared to the population of interest which is London Ontario. Knowing the result of 1 observation does not affect the next observation. For example, if one individual is recorded as a success; their response is “less than expected”, it should not affect what another individual thinks or feels. This is because the sample (n=10) is

very small relative to the population of interest which is the entire city of London Ontario, meaning even if the research did not sample with replacement, the individuals removed will have a negligible impact on the probability of success that we assign to the next observations we take. Another reason is although we might get individuals from the same family who may share similar thoughts, this form of non-independence is broken down by simple random sampling where all combinations of samples are possible from the sampling frame. Therefore, this condition is met since all observations are reasonably independent.

3. The third condition is that each observation falls into one of just two categories ("success" or "failure"). This condition is met since each participant's experience / each observation (individual) falls into one of 2 categories; success or failure. The success is the condition we are interested in, which is the human experience of snowfall being less than the actual amount of snowfall recorded. Not perceiving the snowfall as less than the actual amount of snowfall recorded is considered a failure.
4. The fourth condition is the probability of a success "p" is the same for each observation. This condition is met since the probability of success "p" is the same for all observations. While we may not know the exact probability of success in a population, we know there is a certain percentage of people who think the snowfall is "less than expected" (success) in the population so when we randomly select an individual from this population, the probability of success will be equal to the proportion of the population with that characteristic. This means the probability of success (less than expected) will be the same for each observation and will be equal to the probability of success (less than expected) in the population of London Ontario. This is all in consideration of the observations being reasonably independent since the probability change caused by the individuals removed from the population is negligible.

Conclusion:

Therefore, using a binomial distribution is valid for this variable when considering that the observations are reasonably independent.

Student Answer 2 (Full Marks):

The Binomial distribution is not a suitable probability model for this scenario. Four conditions must be met for a Binomial distribution to be applicable: there must be a fixed number (n) observations or trials, the observations (n) are independent (the probability of one doesn't impact the subsequent observations), each observation must fall into a success or failure category, and the probability of success (p) is the same for each observation. The first condition of having a fixed number of observations has been met as there are a fixed number of 10 participants selected.

Regardless, the observations are not independent as the probability of one impact the other due to the use of sampling without replacement. This is because if we knew if one individual is a success whose response is "less than expected", then the subsequent individual is more likely to be a failure or have a low probability of being a success. Furthermore, individuals are not being directly placed or removed from the population after each trial and only taking them from the sample for a selected day. As a result, it is independent because each trial's probability of success is guaranteed by directly adding or removing individuals from the population, and because trials are chosen for a specific day, the probability of success and failure would vary. Therefore, the second condition regarding independence is not met. Thirdly, there are two possible outcomes, success or failure. In this scenario, individuals who have perceived the snowfall as less than the actual snowfall are recorded in the success category, however, those who do not are in the failure category. The third condition is met. Lastly, the condition of constant probability for each observation has not been met because the probability of success changes with each observation as the observations are not independent. Stated differently, because the trials are independent, learning whether a trial is successful influences the likelihood of both failure and success in subsequent trials, increasing the likelihood of a failure. Furthermore, because we are working with humans, their perceptions of what constitutes "less than expected" snowfall, the likelihood of success and failure can also vary. Individuals who have lived in a warmer or tropical geographic who may claim that they have experienced more snowfall than the actual amount recorded for a particular day because they have never experienced snowfall, then that person is not comparable to someone who lives in colder climate country like Canada who claims that they have experienced less snowfall than what has actually been recorded for a particular day. Given that the second and fourth conditions regarding independence and probability have not been met, thus, we can state that this situation does not follow a binomial distribution model.

Feedback

Short answer question addresses all parts of question AND content demonstrates an understanding of the concepts that apply to the question, even if their application isn't completely correct.

Your answer achieved the following:

- clearly identified conditions for Binomial (even if not correctly represented)
- clearly identified whether each condition was met or not
- answer is written in the context rather than generically (e.g. writing about 'participants' rather than
- always referring to trials/observations, about "less snowfall" rather than always success/failure)
- attempts to connect justification/discussion to content from the scenario (even if justification isn't accurate)
- attempts to incorporate relevant vocabulary (e.g. trials, success, independence, etc.)

Model answer:

Condition 1: there is a fixed number of trials. In this case, each trial is a participant from the sample. There are a fixed number: the sample size is 10 participants total (that's what the meteorologist used). Condition is met.

Condition 2: the trials are independent. I think this condition is reasonably met. The meteorologist took a SRS of participants, and it seemed like it was with replacement (they made sure 10 different participants for a given day). There's no reason to expect one participant's experience of the snowfall would impact others.

Condition 3: Each trial has two possible outcomes. A success would be experiencing less than the actual snowfall, making the failure something else (e.g. more, or the expected amount). This seems to fit the idea of two possible outcomes, so the condition is met.

Condition 4. The probability of success is constant. While I may not know the actual probability of experiencing less than actual snowfall, each person will either experience less, or not less (i.e. success or failure). So, there is some proportion of all people who experience less; because we are taking a SRS from those people, it's reasonable to assume the probability of experiencing less than actual snowfall is constant across selected participants. It's just the proportion for the population as a whole.

Activity 4

Q1. Value 0.0045 is an estimate of how far the statistic is from the parameter, typically.

That value is the standard deviation (or standard error) for the sampling distribution of sample means (as this is a confidence interval for the mean).

Q2. The mean second year grade of undergraduate students who receive academic mentoring is higher than that of undergraduate students who receive hidden curriculum mentoring

Q3. ALL three of the statements were correct.

A probability model--like a Normal distribution--can be used to describe a sampling distribution, and provides predictability to values of the statistic.

While variation exists in the values of a statistic, that variation is described by the standard deviation of the sampling distribution.

When dealing with an unbiased statistic, how close our statistic is to the mean of the sampling distribution is also how close our statistic is to the parameter being estimated.

Q4. Student Answer 1 (Full Marks):

For this specific research scenario, the null hypothesis is “There is no relation between the academic achievement of students and the identity of mentors as either upper years or staff/faculty members”. This was the null hypothesis since it is a hypothesis of no difference or no relationship between the variables of academic achievement and identity of mentors. The alternative hypothesis is “The identity of a mentor as a staff/faculty member improves the academic performance of students significantly more than upper-year students as mentors”. This was my alternative hypothesis since it is a hypothesis associated with a claim. After looking at the hypothesis tests, the dean wanted to keep the mentoring program filled with only staff and faculty. I would not have come to the same conclusion as the dean since the p-value is 0.04 which means there is a 4% likelihood of observing data under the null hypothesis. I believe p-value is not small enough to reject the null hypothesis and only select faculty members and staff as mentors in the future which would be supporting the alternative hypothesis. The dean's conclusion could have been swayed by confounding variables; thus, I think looking over the study and sampling design is important.

In the sampling design, students were randomly selected by stratified sampling since he chose 50 students randomly out of each department, this is a good sampling design to use. The study design used was a “completely randomized design” where students from the sample were

randomly put into a total of 4 treatment groups. Although this process is randomized, there is no mention of control or blocking of variables that may cause differences in treatment responses and may influence the response variable as these variables could be confounding. The bar graph attached to the scenario showed that the upper-year students had fewer mentoring hours with the students than the faculty/staff. This could be because upper-year students may have exams and studying to do for themselves leaving them with a tighter schedule to do mentoring sessions. On the other hand, faculty and staff members may have more time and availability to do mentoring hours since they are not full-time students. This is shown in the graph since there was a smaller proportion of "no times met" in faculty/staff than in upper-year students and also there was a larger proportion of "more than four times met" with staff/faculty than with upper-year students. These factors may have been the factors that swayed the dean's decision even though I believe the p-value is too small to reject the null hypothesis in this research study.

I do not think the dean's decision to use only staff/faculty as mentors is warranted since I believe that upper years could provide an insight that is a lot more relatable than a faculty member such as balancing different assignments/full course loads and social times like going to social events or emotional relationships at an age of first year and second year. The relationships built may be stronger when with an upper year due to close age and experiences. The staff and faculty on the other hand are a lot better at providing academic support as they have a lot of years of experience. Both mentor types have their own strengths. These forms of confounding could be taken into account when designing the study design by applying some blocking. The blocking could help us understand and evaluate how the confounding variables affect the study, preventing them from being confounding and allowing us to better investigate the relationship between the identity of mentors and academic success. This blocking could be done on a small scale where each sample is blocked by using the repeated measures design or even the paired measures design. The dean could also keep the number of meeting times constant between both types of mentors so we could see a better relationship between the identity of the mentor and academic achievement.

Student Answer 2 (Full Marks):

In this scenario, the null hypothesis posits that the identity of the mentor does not affect students' academic achievement. Conversely, the alternative hypothesis suggests that faculty or staff mentors significantly influence students' academic performance more than upper-year student mentors. Based on the results of the hypothesis test, I would not have drawn the same conclusion as the Dean, who intends to exclusively use staff or faculty as mentors in the future mentoring program. I have decided this because the p-value of 0.04 (indicating a 4% likelihood of observing the data under the null hypothesis) provides weak evidence supporting the claim that staff or faculty members should serve as mentors in the future, favoring the alternative hypothesis. There are several potential confounding variables that may have influenced the

Dean's conclusion. In the study design, students were randomly assigned to the four treatment groups (a completely randomized design) without any control or blocking for confounding variables. Additionally, the bar graph shows that faculty mentors generally provided more mentoring hours and had more frequent meetings with students compared to upper-year student mentors. Upper year students have their own academic commitments, leading to less stable availability for mentoring sessions. Faculty and staff members, on the other hand, have more flexible schedules and more consistent availability, with only a small proportion indicating no availability for mentoring sessions. These factors likely contributed to the Dean's conclusion, despite the weak evidence suggested by the p-value. Therefore, I do not agree with the Dean's decision to restrict the mentoring program exclusively to staff or faculty mentors. Furthermore, I believe the Dean's decision to use only staff and faculty as mentors in the future is unwarranted. Upper-year students, having recently faced similar experiences and challenges, can relate better to undergraduate students. Although staff and faculty mentors offer more experience, a better understanding of academic requirements, and more stable availability, I recommend implementing a blocking method to control for any confounding variables. This approach could provide a clearer understanding of the interaction between different types of mentors and student outcomes. I recommend that the Dean consider an alternative study design, such as repeated measures, to determine which mentoring approach is more effective. Additionally, a mixed approach combining the perspectives of upper-year students and faculty/staff could provide comprehensive benefits. If upper-year students were able to maintain frequent and consistent meetings, the conclusions might have shown a stronger correlation between the type of mentor and improved academic achievements.

Feedback

Short answer question addresses all parts of question AND content demonstrates an understanding of the concepts that apply to the question, even if their application isn't completely correct.

Your answer achieved the following:

- clearly stated whether you would draw the same conclusion as the dean
- clearly stated whether you think the Dean's decision to use only staff/faculty as mentors is warranted
- provided a discussion to support both points
- attempted to ground your discussion in concepts taught in 2244 related to drawing conclusions in hypothesis testing and understanding of P-values

Model Answer:

Note, there are many possible ways to answer this question; this is just an example

I would probably draw the same conclusion as the Dean with one caveat: I'd want to know whether the relevant conditions/assumptions for the hypothesis test conducted were value. I say this because the study used by the Dean was quite good (a diverse sample, strong completely randomized experiment), so I trust the data collected, and hence, interpret a P-value of 0.04 as indicating an unlikely outcome. I just have the 'caveat' because I know the 'accuracy' of that P-value is only as good as the fit of the probability model used to calculate it. In terms of whether I would focus on only staff/faculty for the future, I probably would not. Looking at Figure 4, the distribution of number of meeting times does obviously differ (hence, the small P-value), but the main differences are in the frequency of "no times" for meeting and "two to four" times across the mentor identity types. Looking at the data for the two types of mentors, the frequency of "no times" is still low and consistent for both types of mentors. If mentoring really is a benefit, then my big concern would be supporting the mentor identity that is most likely to get students to engage with the mentoring at all; that doesn't seem to differ based on the two mentor types. So, I'd simply keep supporting both types of mentors (maybe encouraging more faculty/staff mentorships but not exclusively), so students have options for what they are comfortable with.

Activity 5

Q1.

Version 1: none of the procedures taught in Biol/Stat 2244 would be appropriate

Remember from Topic 5 (Data Structure and Planning Analysis), we think about:

- the type of variables involved, which can lead to what parameter is of interest
- the number of comparison groups
- the structure of the comparison groups (matched/paired vs. independent)
- the analysis goal (estimating a parameter or testing a claim about a parameter).

When we review the research question, it involves comparing multiple comparison groups (the different household salary levels)—a categorical explanatory variable—for a categorical response variable, the value of the Western Fair entrance fee. Remember, that 'value' variable was categorical ordinal. All of the confidence intervals and hypothesis tests that we have seen in 2244 involve a quantitative response variable. Hence, none of these methods would be appropriate.

Version 2: t test for difference between means

Q2. The estimated mean number of tickets purchased if a household has no residents is 10.55.

Q3. Mosaic plot, Bar Graph

The research question is focused on the age—which was measured as a categorical ordinal variable (i.e. age groups)—and whether carnival games are played (which was measured as a categorical, nominal variable). If we are focused on graphing the relationship or association between two categorical variables, the only graph types listed that are appropriate are: bar graph and mosaic plot.

For reference,

- bar graphs and mosaic plots handle univariate, bivariate, and multivariate categorical variables
- histograms and dotplots handle a univariate quantitative variable
- scatterplots work for the relationship between two quantitative variables, possibly with additional categorical explanatory variables (as colours/symbols)
- means plots, boxplots, and stripcharts are appropriate for: (i) univariate quantitative, or, (ii) quantitative response and one or more categorical explanatory variables

Q4.

Version 1:

Student Answer (Full Marks):

Mean number of ride tickets does not differ based on the types of tickets the participant purchased online (F = 0.436, df= 3,196, P=0.727).

Version 2:

Student Answer (Full Marks):

The mean number of ride tickets purchased by the Western Fair participants does not differ based on the perceived value of the entrance fee (df=4, 195, F=0.372, P=0.828).

Feedback:

Short answer question addresses all parts of question AND content demonstrates an understanding of the concepts that apply to the question, even if their application isn't completely correct.

Your answer achieved the following:

> Showed a clear attempt at proper conventional format for hypothesis test conclusions; that would reflect:
> - Writing in the context of the scenario
> - Having a sentence that states the conclusion
> - Placing 'evidence' for the conclusion in parentheses (i.e. P-value, df / n, sample statistics)
> Logically related to the research question

Model Answer:

The mean number of ride tickets purchased by Western Fair participants does not differ based on perceived value of the entrance fee (F = 0.327, df = 4,195, P=0.828).

Comments on that model answer

- it doesn't mention hypotheses, statistical significance, rejecting/accepting hypotheses, significance levels, etc.

- it makes a clear statement that addresses the research question, thereby communicating what the conclusion was for the hypothesis test (in this case, the null hypothesis was of no difference, the conclusion sentence states clearly whether that was the conclusion or not)
- any reference to test statistics, P-values, sample size or degrees of freedom, and sample statistics (like sample means, standard deviations, etc.) are in parentheses

Note as well that the parameter in this case is the "difference between means", i.e. we compare the mean for those who had a value of 1, to the mean of those with value 2, to the mean of those with value 3, and so on. We are not computing a variable that are the "differences" (or changes), and then taking the mean of those 'differences'. Consequently, it's not accurate to phrase the parameter as the "mean difference". That's like saying "mean weight" when what you really are focused on is the "mean height"; it's the wrong variable that the mean summarizes.

Winter 2024

Activity 1

Q1. Causative. The answer is B). The research question is "Do instructor gender and presentation mode for swimming lessons impact the speed at which preschool children learn fundamental water safety skills?" So, we are trying to find a relationship between the variables, gender/presentation mode and the speed at which children learn water safety skills.

Q2.

Version 1: Preschool children is the population of interest The answer is B). The question to be researched is "Do instruction gender and presentation mode for swimming lessons impact the speed at which preschool children learn fundamental water safety skills?"

The YMCA-SWO didn't take their sample from this population of interest, rather, their sample was the 50 preschool children who signed into the 10 classes. Remember, the sample can only be made up of YMCA-SWO members as these classes are only available to these members. The sampling frame is best described as preschool children who are members of the YMCA-SWO.

Version 2: The sampling frame is "the preschool children who are members of the YMCA-SWO". The answer is C).

Q3.

Version 1: "Speed of learning fundamental water safety skills is a response variable" correctly describes the variable in this study. The answer is C).

Version 2: "Instructor gender is an explanatory variable" correctly describes the variable in this study. The answer is B).

NOTE: Play-based presentation mode is not a variable at all. It is a level or value of the variable presentation mode.

Q4. Student answer 1 (full marks):

A concern I have is that voluntary response sampling is applied here, which results in self-selection bias. Because the new classes are an open invitation to preschool children who are members, this is voluntary response sampling, which is a form of non-probability sampling. From my experience, boys are more encouraged into sports by their parents, especially at a young age. Because the registration spots are limited, and it is the parents who would be registering them, boys are more likely to be registered into these classes (i.e. boys will have a higher probability of being chosen into the sample). This will impact the data because if the sample is comprised mostly of one gender, this will skew the data collected, especially in relation to instructor gender. From personal experience, boys prefer to be taught, especially in an athletic setting, by male instructors which would result in improper data collection of the relationship between instructor gender and speed of learning fundamental water safety skills.

Student answer 2 (full marks):

One concern of YMCA-SWO's plan that may limit how well their data will reflect the population of interest being of preschool children of Southwestern Ontario, is their sampling frame and how it produces undercoverage bias. The sampling frame for this study are preschool children aged 3-5 years at a preschool 1 skill level during winter 2024, who have an active account with YMCA-SWO in either their Stoney Creek (London) or Woodstock location. Unfortunately, this Sampling frame chosen does not readily represent the population of interest for the study, and shows undercoverage bias. YMCA-SWO chose to conduct the study at their two London and Woodstock locations during the winter 2024 session, due to this, many preschool children may have been missed who cannot make it to either of these locations (do not live in either London or Woodstock) or only take swimming lessons in the summertime. Based on my experience living in London Ontario, the summer would be the most popular time to take swimming lessons, due to an increase of outdoor swimming activities in the warmer temperatures. Another example of undercoverage bias occuring is due to preschoolers whos families could not afford the YMCA-SWO membership fees (although being fairly inexpensive), they may have been missed from the study as well. Due to these discrepancies, the preschoolers in the large area of Southwestern Ontario are not fairly accounted for, and they study's response variable will not be accurate when measured.

NOTE: Comments included that it wasn't clear how the difference stated is relevant to a research question on the effect of gender and presentation mode on speed of learning. Be sure when you identify such differences that you clearly link that difference or variable to being relevant based on the research question (for future assignments or activities).

Activity 2

Q1. The answers are A) and B).

Undercoverage and Non-response are bias that are present in the design. (**Self-selection** is NOT chosen). NOTE: Undercoverage occurs when our chosen sampling frame leaves out some part/group of the population. It only used London and they said it tends to be "older individuals or small business owners" who come into the bank in person. **Non-response** occurs when researchers select individuals to participate and then they don't participate and this clearly happened since they only got 257/1000 questionnaires back. **Self-selection bias** occurs from voluntary response samples which we don't have in this case.

Q2.

Version 1: The second stage consists of simple random sampling. The answer is D).

Version 2: The first stage consists of cluster sampling. The answer is C).

The population of interest was the Bank of Montreal customers. They chose to take their sample from customers visiting the in-person branches located in London (14 such branches), so this defines their sampling frame. The first stage in this multi stage sample involves cluster sampling as they select a subset of people from that sampling frame since they randomly select 5 branches.

After that, from the customers at those five branches, the bank assigns each person a unique number and randomly chooses 1000 numbers, so this step is a SRS (simple random sampling).

There is an obvious non-response bias, which is unavoidable as the researcher has to obtain consent to participate.

Q3.

Version 1: age (years)

The issues are that it includes spaces and it uses special symbols ie. parentheses.

The answer is A) and B).

Version 2: 1app based

The issues are that it includes spaces and it starts with a number.

The answer is A) and B).

Q4.

Student answer 1 (full marks):

i) The piece or type of information I think is the most important to include in the metafile to maintain data sharing and reproducibility would be the methodology of the study's data collection. ii) I would include the multistage sampling used to gather the data and incorporate information about the randomness used by assigning numbers to ensure no constraints or limitations were put on the sample. The methodology of gathering the data would be described, so that the data in the datafile makes sense and can be used by other interested researchers. iii) For other researchers to be able to interpret the data in the datafile, it is most important to understand how the data was taken and if it holds validity. This piece of information is essential for the reproducibility and data sharing of the datafile, and can provide researchers with the research methods used to also determine any bias.

Student answer 2 (full marks):

(i) It is most important to include the sampling strategy in the metadata file, especially as the sampling frame consisted of only customers who visit in-person branches. As stated in the research scenario, older individuals tend to prefer in-person banking, which means an undercoverage bias is present in this study.

(ii) The datafile shows the age of different customers. Of the customers in the screenshot, the range of age is 55 years old to 82 years old, an example of the undercoverage bias which results in most of the sample being comprised of older individuals which may skew the data. For example, this can be seen as most of the individuals in the screenshot said they would not use app-based banking.

(iii) This information is critical in the context of data sharing because this sampling methodology results in undercoverage bias that leads to the sample comprising of mostly older individuals. If other researchers attempt to replicate the study, they must understand how the sample is chosen and how this introduces bias which leads to differences between their results and this studies results.

Activity 3

Q1. Version 1: Nominal and categorical. The answer is C) and D).

Version 2: Categorical and Ordinal. The answer is A) and D).

Q2. Survey is the study design. The answer is F).

Q3. The answer is B), E, and F).

Comparative design, replication and collecting data on cofactors

Comparative design: The researchers were comparing two different migration approaches which were "all at once" and "some get both".

Replication: If you look at the comparison groups, they used more than one student in each group. You can tell this as they used stratified sampling, or you can notice that in figure 1 it talks about 1000 + students being used.

Collecting data on cofactors: The questionnaire they used gave us information regarding the student's opinion and the amount of time using the LMSs, it also gave other information, such as age, faculty, year of study, etc.

Why are the other answers incorrect?

Blocking: Since the study design was an observational study and not an experimental approach involving a control group, blocking couldn't occur. Don't mix up stratified sampling with blocking as they are completely different. The selection of students from the years 2012 and 2024 were not blocks, but rather, samples from two populations beings used, i.e. a migration approach of "all at once" vs. a migration approach of "some get both".

Randomization: This is another experimental principle, so like blocking, it didn't not occur. When we randomly assign individuals from ***the sample into treatments***, this is called randomization. Do NOT mix up random sampling with randomization. Remember, random sampling is a sampling method or a way to obtain individuals in your study. On the other hand, ***randomization*** is a process used in an experiment on the individuals who are ALREADY in our sample to assign them to treatments.

Hold other variables constant: There wasn't any indication of holding any variables constant, such as age, faculty, year of study. The data for the sample came from a variety of ages, faculties, etc.

Q4. Model answer: One concern I have about this study design is that there is a confounding variable, being the LMS switch. There are two comparison groups “full group” migration approach and “gradual migration” approach. The full group switched from WebCT to Sakai, whereas the gradual group switches from Sakai to Brightspace. As a result, if a students’ response is in favour of “gradual migration”, we don’t know whether it is because of gradual migration or is it because they favour the switch from Sakai to Brightspace? I would redesign the study so that both migration approaches occur in 2024. I would then have some students in the gradual approach and others who are fully in Sakai switch over in September to Brightspace. As a result, I would do the two different approaches simultaneously right now and the variable “LMS switch” would be held constant.

Student answer 1 (full marks):

This study is a survey which is an observational study. Thus, the explantory variables are not directly manipulated and so exhibiting "control" on this study design is difficult. An important confounding variable to consider is student's previous experience with a Learning Management System (LMS). I would collect this data through the questionnaire by asking: "Have you used (new LMS) before?" The possible responses would be "Yes" or "No". (new LMS) would be replaced in the questionnaire dependent on which migration the student participated in.

Student's with experience in the new LMS may report that they find using it to be easier, or recognize different features or aspects they enjoy. For example, from my experience with brightspace in highschool, I find its calendar feature to be more intuitive, so the cofactor previous experience would affect the response variable, which is Western University undergraduate student opinions of a new LMS. Although we cannot exhibit "control" to effectively prevent the effect of this confounding variable as this is an observational study, we can collect data on if students have previous experience with an LMS so its impact on the response variable can be accounted for, as a cofactor.

Student answer 2 (full marks):

In this observational survey study design, a questionnaire is presented to students selected through stratified sampling and is used to collect categorical data. One issue survey study designs (observational studies) pose are the potential of confounding factors which can interfere with the response variable data. A way to combat these potential confounding factors would be to introduce some type of randomization in the stratified sampling aspect of the study. Stratified sampling can be made random sampling through having proportionate groups of each strata of faculty and year of study. Although the stratified sampling used attempts to rid of confounding factors through known variables, introducing randomization into the study design would be more representative of the study, rid of bias and further aid in reducing confounding factors. Establishing randomness in the study design will reduce potential confounding variables introduced through observational study designs by allowing for them to be balanced across each strata through randomization.

Activity 4

Q1. Version 1: Bivariate. The answer is B).

The two variables involved are the number of COVID-19 vaccine doses and the number of children that live in a residence, so this is called bivariate data.

Version 2: Univariate. The answer is B).

Here, there is only one variable involved which is the source of information about vaccine safety, so this is called univariate data.

Q2. Approximately symmetric and unimodal. The answer is A) and E).

Here, we have only ONE peak, so it is unimodal. Yes, it is very spread out, but that is fine!

Q3.

Version 1: Stripchart, Means plot and Boxplot. i.e. answer is D, E, and F

We have two variables here, or bivariate data. The two variables are age and frequency of a vaccine. Ask yourself what type of variables each of these is. Age is entered in years and so it is a quantitative variable. Frequence of vaccine was collected as data: 0 (never), 1 (once or twice), 2 (almost every year) or 3 (every year). Since these are words and not numerical data, like 1,2,3,4, they represent categorical data and not quantitative. They are "ordered", so they are ordinal, not nominal, but categorical nonetheless.

If we have a categorical explanatory variable and a quantitative response variable, we can us boxplot, means plot and stripchart only.

Version 2: Bar Graph and Mosaic Plot i.e. answer is A and F

The population here is just "people" and the two variables are household annual salary and frequence of a seasonal vaccine. The household income is collected as under 25,000, 25000-40000, 40 0001-80000, etc. These are groups are labels or categories and not a quantitative variable. The second variable, frequency of the vaccine is the same as version 1 above. So, we have a categorical response variable and also a categorical explanatory variable. Thus, we choose only the bar graph and mosaic plot options.

Q4. A sample of the act4.csv is below for the first 20 people:

	A	B	C	D	E	F	G	H	I	J	K	co
1	age	salary	flu	covid_doses	trust	mRNA	covid_none	covid_science	covid_friends	covid_health	covid_social	
2	82	$40001-$80000	1	2	4	Yes	1	0	0	0	1	
3	33	$80001 - $120000	1	3	4	I don't know	1	0	1	0	1	
4	82	$80001 - $120000	1	3	3	Yes	0	0	0	0	1	
5	66	$40001-$80000	2	2	3	Yes	1	0	1	0	1	
6	78	$40001-$80000	1	2	3	I don't know	0	0	0	1	1	
7	86	$40001-$80000	3	3	3	Yes	0	0	1	0	0	
8	74	$25000-$40000	1	1	1	I don't know	0	0	0	0	0	
9	27	$80001 - $120000	0	3	4	I don't know	0	0	0	0	0	
10	39	$80001 - $120000	2	2	3	No	0	0	0	0	0	
11	30	$80001 - $120000	1	3	4	I don't know	1	0	0	0	1	
12	51	$80001 - $120000	2	3	4	Yes	0	0	0	0	1	
13	21	$25000-$40000	2	2	5	Yes	0	0	0	0	0	
14	76	$40001-$80000	0	2	3	I don't know	1	1	0	1	1	
15	49	$40001-$80000	1	2	4	I don't know	0	0	0	0	1	
16	42	$80001 - $120000	2	2	2	Yes	1	0	0	0	1	
17	81	$40001-$80000	1	3	4	I don't know	0	0	1	0	1	
18	30	$80001 - $120000	1	2	1	I don't know	0	0	0	0	1	
19	21	$25000-$40000	1	3	4	I don't know	0	0	1	0	1	
20	28	$40001-$80000	1	3	4	Yes	0	0	0	0	1	

L	M	N	O	P	Q	R	S	T	U	V
covid_doctor	num_adults	num_children	full_vaccine	child_consult	vaccine_doctor	vaccine_science	vaccine_friends	vaccine_health	vaccine_social	vaccine_other
0	2	2	Yes	Yes	0	0	0	0	1	0
0	1	0	NA	NA	NA	NA	NA	NA	NA	NA
1	1	3	Yes	Yes	0	1	0	0	1	0
0	1	2	Yes	Yes	0	0	1	0	1	1
0	4	2	Yes	No	NA	NA	NA	NA	NA	NA
1	2	5	Yes	No	NA	NA	NA	NA	NA	NA
0	2	0	NA	NA	NA	NA	NA	NA	NA	NA
0	2	2	I don't know	No	NA	NA	NA	NA	NA	NA
1	3	3	Yes	Yes	1	0	0	0	1	0
1	1	0	NA	NA	NA	NA	NA	NA	NA	NA
0	2	0	NA	NA	NA	NA	NA	NA	NA	NA
0	3	1	I don't know	No	NA	NA	NA	NA	NA	NA
0	1	2	No	Yes	0	1	1	0	1	0
0	1	2	I don't know	No	NA	NA	NA	NA	NA	NA
0	1	0	NA	NA	NA	NA	NA	NA	NA	NA
0	2	0	NA	NA	NA	NA	NA	NA	NA	NA
0	2	2	Yes	No	NA	NA	NA	NA	NA	NA
0	2	2	No	No	NA	NA	NA	NA	NA	NA
0	2	0	NA	NA	NA	NA	NA	NA	NA	NA

Student answer 1 (full marks):

> Figure 2. Frequency of number of respondents to question 6 in 2020 survey of public health in Ontario (n = 2844 for "Yes", n = 1378 for "No", n = 4125 for "I don't know"). Distribution of responses to question 5 for each response in question 6 is shown by stacked bar graph (n = 835 for strongly diagree [1], n = 822 for moderately disagree [2], n = 2483 for neither agree or disagree [3], n = 3343 for moderately agree [4], n = 864 for strongly agree [5]).

Student answer 2 (full marks):

Figure 2. Stacked bar graph representing the distribution of respondent (n= 20,000) answers for two questions within a questionnaire presented by the Ontario Ministry of Health and Long-term Care. Individuals responded following a 5-point scale for opinion question: 1 = strongly disagree, 2 = moderately disagree, 3 = neither agree nor disagree, 4 = moderately agree, 5 = strongly agree. Participants also answered a knowledge check question with either yes, no or I don't know. All citizens who participated where eligible to vote in Ontario and Municipal elections.

Comments about any model answer:

- There are values or acronyms on the graph that must be defined. These are the numbers to measure how strongly they agree i.e. 1,2,3,4, 5. Since the legend doesn't tell us that green is "strongly disagree", it must be defined as 1=strongly disagree, etc.
- Trust in the Ministry is the second variable and it isn't shown on the graph at all by name, but is only represent using numbers. It would have been clearer if the actual legend provided had the variable name "trust in Ministry" above the numbers/colours, it would have been better. However, since it wasn't, it is helpful to include "reflected by colour" in the caption
- There was no information given on the "what, when or where". Only the "who", ie. Ontario citizens was included. Probably most of the others don't need to be included, however, if the when was during the years 2020, 2021, when the vaccination topic was very "HOT" due to COVID-19, it might have been important to include it
- There is no conclusion or interpretation in the figure caption, ie. nothing to tell readers what they "should" be seeing and there is also no discussion on how to interpret the graph. Here, we assume readers know what a bar graph is and how to interpret it, it Is not explained what the bar heights represent, etc.

Activity 5

Q1. Estimator

Recall that a parameter is a number that summarizes a variable (or some relationship) for a population, whereas a statistic is a number that summarizes a variable (or some relationship) for a sample. The estimator is some function of the sample data that is used to estimate a parameter. It is the equation or function you use by entering the sample data you observe in order to calculate the statistic. So, the equation here % of grade 12's $=\frac{x}{n}$ is the estimator. If you put some sample data into this equation, you would have the statistic.

Q2. 2 and 38% involve sampling error

Sampling error is the difference between the observed value of the statistic and the value of the parameter that it is estimating, i.e. sampling error = statistic – parameter. As a result, any numbers you calculate based on your sample data will involve some sampling error, some more than others!

Here, 2 is the median and this is based on the registration information which is a sample so it is a statistic and does involve sampling error.

38% is the percentage interested in the Faculty of Health sciences of those students participating, so again, this is from the sample and it is a statistic and therefore also involves sampling error.

40.5% is the is a parameter and it does not have sampling error. It is the percentage of Western's population of undergrads who are from Toronto and the GTA region.

Grade 11 is the grade they list that most students indicated on the sample. It is a value for the variable and it isn't a statistic or a parameter and consequently it has no sampling error.

Q3.

The answers are:

1. A trial would be an individual respondent's ranking for the statement and

2. The mean of the distribution will depend on the number of respondents

Recall, the shape of the binomial distribution depends on the probability of success, p. The closer the p is to 0 or 1, the more skewed the distribution will be. Binomial distributions are always unimodal. So, the shape is NOT dependent on the number of respondents, n.

Also, a binomial experiment counts how many successes in a fixed number of trials, n. Each individual can either be a success or a failure and here success refers to agreeing and failure is not agreeing. The count or total number of successes is the value the Binomial takes on.

A trial here is an individual's response as either success or failure, i.e. agree or not agree, so 1 above is true. Lastly, since the mean of a binomial is $\mu=np$ (the number of trials, n times the probability of success, p) so the mean DOES depend on n, the number of trials which in this question is the number of respondents.

Q4.

Student answer 1 (full marks):

A confidence interval accounts for sampling variability to offer a better estimate of a the true parameter mean compared to the sample mean. Rather than just stating the mean high school grade, a confidence interval can be generated from a point estimate (the mean high school grade), a critical value, and the standard deviation of our estimator. We can then determine an interval within which we are confident that the true parameter lies. For example, when referring to a 95% confidence interval, if many samples are drawn from a population and the interval is computed for each, we can expect that 95% of those intervals would contain the mean high school grade of the true population. Different samples, such as students registered for the open house, will have different mean high school grades, which results in sampling variability. The confidence interval portrays the variability found in the sampling distribution, to more accurately represent the sample mean. Thus, the mean high school grade may be 82% but the 95% confidence interval could be between 67% and 97%. This would indicate the point estimate could considerably deviate from the true parameter of the population, highlighting the importance of the considering confidence intervals in our analysis.

Student answer 2 (full marks):

The confidence interval is a range of plausible values for the parameter (population mean high school grade), with a confidence level that the true population will be within the interval; this can address issues of uncertainty and variability that can arise given a sample population of students. From the students attending the spring open house, the collected data is only taken from only a small portion of the population. Due to random sampling variation, each sample will have inconsistencies in the mean grade; this variation is innate in sampling and needs to be accounted for when extrapolating data for a true population. By taking a single value for the mean high school grade will not account for these factors on its own. The sampling distribution from solely collecting sample means of various samples can give insight on population variability and aid in quantifying the uncertainty of our estimate of the parameter at hand. The confidence interval calculated would allow confidence in the estimate, in which account for factors of uncertainty and variation in a population, rather than just stating the mean grade based on the data that will account for none.

Feedback:

Consistent with criteria for full credit

Short answer question addresses all parts of question AND content demonstrates an understanding of the concepts that apply to the question, even if their application isn't completely correct.

Your answer achieved the following:

- Attempted to explain why confidence intervals are used instead of just the estimate
- Used vocabulary from 2244 related to confidence intervals (e.g. point estimate, parameter, statistic, estimator, sampling error, sampling variability)
- Made reference to sampling distributions

Model answer:

Confidence intervals are a range of values that we think contains the value of a population parameter we are estimating. We know that when we take a sample and calculate a statistic (which is our point estimate of the parameter), that statistic will differ from the parameter by sampling error. Unfortunately, we really don’t know how big that sampling error is, so we don’t know how close our estimate might be to the parameter. The confidence interval uses knowledge of the variability of the chosen estimator to get an idea of the average sampling error—by using the standard deviation of the sampling distribution of the estimator. That average sampling error is then used to build an interval around our statistic, which helps communicate the lack of precision in our estimate (so…big intervals would suggest our statistic could be quite far from the real value, mall intervals would suggest the statistic is probably close).

Practice Exam Questions

1. PPDAC

1. **Solution: B.**

See Section A of the booklet. “Having agreed on the problem definition the next stage is to formulate an approach that has the best possible chance of addressing the problem and achieving answers (outcomes) that meet expectations”.

2. **Solution: D.**

A ***predictive research goal*** involves us taking a certain individual in the population of interest and assigning a certain value of the ***response variable***. So, the height of the bean plant reaches based on a 10mg size seed would be predictive. The other options focus on the entire group of Western students (not on whether a particular student is in support), the entire group of 2244 students (rather than if a certain student will score above 75%) and grocery bags in general. (rather than the lifetime of a single plastic bag)

3. **Solution: B.**

A ***descriptive goal*** is one where we are trying to characterize some attribute(s) of a population and we are describing the distribution of the variables for this population. A ***causative goal*** is one in which we want to establish a cause-and-effect relationship between the explanatory and response variables. We want to see that some change in the x-variable (explanatory) causes a reliable change in the y-variable (response). A ***predictive goal*** is one in which we want to predict the value of a variable for a new or future individual (unit) of a population. In this case she wants to determine what grade to expect based on how much time is spent working, so this is predictive. There is no focus on obtaining a general relationship between variables in a population, which would be causative i.e. determining if more hours worked results in lower grades in general. The focus here is on the value of a grade (***variable***) for a particular individual.

4. **Solution: C.**

The ***population of interest*** is the total group of individuals or units that we want to obtain information about. Thus, chocolate milk consumers is the population since Neilsen wants to see if a new boiling process causes a difference in flavour.

5. **Solution: A.**

There is only one ***variable*** which is a response variable or outcome which is the opinion on gun control legislation.

6. **Solution: A.**

The ***population of interest*** is US citizens as they want to know "US citizen opinions' on gun control".

7. **Solution: A.**

Researchers want to determine if the filtration process influences the flavour of their chocolate milk, so the ***explanatory variable*** is the filtration process and the ***response variable*** is the flavour of the milk. Note: If asked, the ***explanatory variable*** has to different values or levels: boiled (new) and filtered (old).

8. **Solution: C.**

The ***population of interest*** is the group of individuals or units we want to obtain information about and it says "interested in the level of public support from the suburb's adult residents…".

2. Sampling Design and Strategies

1. Solution: B.

Since they are **randomly** selecting 12 boxes from each production line each hour, this is classified as ***stratified sampling***. This is not an answer choice. It is not **cluster** as they are not randomly selecting some of the production lines and performing a census of them. It is not **SRS** as they are taking a random sample of **EACH** line and random sampling requires **EACH** unit has an equal chance of being selected. This is not the case as we are selecting 12 from each line and that is not the same as selecting 120 from the total. It is ***probability sampling*** because there is "***random sampling*** from each line". Is it random sampling? A ***stratified sampling*** method is considered random if the number of units from each stratum is chosen at a constant proportion. However, they tell us that the lines operate at different rates, so the selection of 12 is not going to be a constant proportion.

2. **Solution**:

There are many appropriate ways to answer this 3-point question. 3 points are allocated if the answer demonstrates:

- all use of sampling vocabulary in the answer is accurate
- the suggested sampling design would reasonably result in the pattern/composition of the sample that was provided
- the answer clearly justifies/explains why the approach would result in the pattern

Example: This ***sampling strategy*** involves a ***multistage*** involving first ***cluster sampling*** (sampling of players by province) and second a ***stratified sampling*** method (players sampled by the league they play in). The players can be found throughout Canada in all 10 provinces and all 3 territories but the sample only includes players from Ontario, BC and Quebec. This would likely be due to ***cluster sampling*** where we randomly select a few provinces. Normally, with ***cluster sampling***, we would then perform a census of players in those 3 randomly selected areas, however, there are such small numbers of players that it is more likely that a stratified sampling method was applied next. This way, there is a random selection from each of the three types of organized leagues in each of the randomly selected provinces.

3. Solution: B.

Here the ***population of interest*** is the organizations that the credit union supports and so is the sampling frame! (they are the same here). Since they use a computer to randomly select 2 of the classes of the organization, the groups are subdivided based on a pre-existing condition or variable, so it could be ***stratified*** or ***cluster***. In this case, they try to use all the organizations (units) in the selected groups, so it is **cluster *sampling***. It is random sampling since the clusters were chosen at random, so ***random sampling*** is correct. Since they email each of the organizations that are selected, it is not a voluntary response sample, so it is B.

4. Solution: A,C.

The coach is looking for the relationship between race time (***response***) and training regime (***explanatory***). The coach separates the athletes by sex and experience level, so it is ***stratified*** by both of these. Once some of the athletes are chosen from each sex/experiment level, they are subdivided by age and assigned to training regimes, but there is no sampling at this point, so there is no other stratum. Note: Under 25 years old is a ***block*** since it is a grouping that was used AFTER we conducted the sample, a way to randomly assign individuals within each block (age grouping) to the treatment (training regime) in this experiment.

5. Solution:

Here's an example answer:

The ***population of interest*** is grocery stores in London Ontario. I will use a ***sampling frame*** of all grocery stores open 7 days a week for simplicity of conducting my study. I will use a ***multistage sampling*** method in which I divide the grocery stores up based on square footage: small, medium and large stores. I will take a ***stratified sample*** of 15% of each size group and this is my first stage of sampling. I will then subdivide those stores into four quadrants: North, South, East and West. I will then use ***cluster sampling*** as my second stage of sampling and randomly select two of these four quadrants and the stores in these regions will make up my sample. Note: you don't need to worry about why you would/would not want to subdivide into quadrants. The point of the question is merely to see if you understand obtaining:

- a sampling design consistent with multi-stage sampling. That would mean that the sampling frame is reduced in at least 2 separate steps.
- your answer correctly identified the (1) population, (2) sampling frame, and (3) sample.
- your answer was related to the research question (i.e. written in the context of grocery stores, etc.).

6. Solution: A, C.

This research study involves a survey using a questionnaire to see whether peers in the program are interested in medical research. The ***population of interest*** is Queen's university, undergrad medical program individuals. Since she ***randomly*** selects 3 courses and the selection of courses is a way to take all the peers from some of the courses, this is ***cluster sampling***. It is also probability sampling because there is an element of chance.

7. Solution: B, D.

This research problem involves the study of the effect on population size (***response***) due to the impact of predators (***explanatory***). The ***sampling frame*** is the phytoplankton in the tank the researcher has. First, they divide the phytoplankton based on species and sex and then 100 individuals are randomly selected from each of these subgroups. The ***strata*** are, therefore, species and sex. When the phytoplankton are categorized by reproductive maturity, the question is: is there are second stage after the ***stratified sampling*** and only SOME of the phytoplankton are selected or is the further subdivision just part of our ***study design***. In this case, the phytoplankton from each maturity group are randomly assigned to tanks which each contain a different predator (explanatory variable), so this is just assigning treatments in the ***experiment*** and this is not a form of sampling.

8. Solution: C.

The ***individuals/units*** in this question are the books. Since the library already has the books subdivided into 20 piles and then they select 4 piles randomly, this seems to be ***cluster sampling***. However, after that, they choose 15 books from each of the 4 piles, so this is an example of ***multistage sampling***.

9. Solution: B.

The ***population of interest*** is homeless people in London since they want to determine what type of supports are needed for homeless in London. Since the ***sample*** is selected only from the homeless living in tents along the Thames River bank, this is the ***sampling frame***. The ***sample*** consists of the 15 people who are willing to speak to them.

10. Solution: D.

This is an example of ***convenience sampling*** as once they ***randomly*** selected the 4 busy intersections; they simply choose the first 10 cars! It is in fact a ***multistage sample*** as they first use a ***probability sampling method*** of intersections and the second stage is a convenience sample of cars.

11. Solution: D.

Cluster sampling involves dividing the ***population*** into clusters based on some pre-existing variables and then randomly selecting a subset of the clusters and collecting data from each of these clusters. Here, the ***population of interest*** is the bottles of nail polish and they are being subdivided by brand and finish (clusters). However, it is ***not cluster sampling*** because we only take 10 bottles across all combinations of brand/finish. It isn't ***stratified sampling*** as for stratified we would need to select a ***random sample*** from EACH brand/finish combination and this is not done, so it isn't just a ***stratified sample***. There is no convenience sampling as we aren't taking bottles closer to the end of the shelf, etc. This is a ***multistage sample*** involving taking samples within samples. We could say the first stage is ***cluster sampling*** of bottles by brands and the second stage is a ***cluster sample*** of bottles by the brand/finish and the final or third stage involves a ***stratified sample*** of bottles by brand/finish. We could also say we take an ***SRS*** of brands, then a SRS within that SRS of brand/finish combinations from the chosen brands and then a SRS of bottles from the randomly selected brand/finish combinations.

12. Solution: B, C, D.

First, the ***population of interest*** is injection drug users and the ***sampling frame*** are the injection drug users in the methadone clinic. The ***sample*** will be taken from the injection drug users in the methadone clinic. The ***sampling design*** involves first ***randomly selecting*** a day of the week and then ***randomly selecting*** 20 of the 100 individuals on that particular day. This involves two stages, so it is definitely ***multistage***, but that is not an option. It also involves the element of chance i.e. some randomness, so it is definitely ***probability sampling***. There is no ***systematic sampling*** as we aren't selecting every 10^{th} person, for example. We are using both an ***SRS*** and ***cluster sampling***. In the first stage, cluster sampling is used to subdivide the individuals by day and then we take all the individuals from that day. In the second step or stage, we randomly select 20 of 100 individuals and this is an SRS.

13. Solution: D.

By definition, a ***sampling frame*** is the subset of the population from which the actual sample is drawn. In this case, the population of interest is Dr. Lee's patients diagnoses, but the review board is only selecting diagnoses from televised patients. Thus, the sampling frame is only the patients' diagnoses who were televised.

14. Solution: C.

The ***population*** is the US. If the sample is representative, then it is a random sample in which all individuals in the population have an equal chance of being chosen. Because the invitations were only randomly sent to America Online subscribers, it is not representative. The subset of the population including non-subscribers were left out of the sampling process. This is ***undercoverage*** and as a result, even though it was done using a ***random sample***, it is coming from a ***biased sampling frame***, or one that is unlikely to be representative of the US as a whole. There is no self selection/voluntary response bias because the participants in the survey were sent an email invitation directly, and they weren't just responding to an open invitation where they "self-select". We could have ***non response bias*** though as some individuals may choose not to respond to their invitation.

15. Solution: C.

The ***sampling frame*** is the subset of the population of interest from which the actual sample is drawn and here it says they are only willing to hire individuals who live in the US and are adults with an AMT account, so the answer is C.

16. Solution: D.

Stratified sampling is the best sampling method so that you test EACH region. ***Stratified sampling*** involves splitting the population into ***strata*** and taking an ***SRS*** of each strata. This ensures you are testing the water quality in each region whereas just doing an SRS won't ensure this, especially since the water is not flowing well between all regions.

17. Solution: C.

In order to be ***stratified***, you need to choose an ***SRS*** from EACH group or ***strata*** and in this case, he only takes a ***random sample*** from 10 of his trees. If you had collected all of the apples from each of the 10 randomly selected trees you would have a cluster sample. To be a random stratified sample, each tree would have to have a proportional number of apples chosen, for example, 10% of the apples from each tree.

18. Solution: A, C.

Both are correct answers. ***Undercoverage*** occurs when you leave out certain groups from the population of interest when choosing your sampling frame. In this example, they put up posters in municipal offices and therefore, only those who can read could possibly participate in the study. ***Selection bias*** or bias resulting from voluntary response is another serious flaw. Since these posters are an open invitation to those to are in the right place at the right time, possibly individuals who are more interested in this top or have stronger opinions, are more likely to self-select themselves are part of the sample. This example doesn't involve ***non-response bias*** as that occurs when you invite people to participate and some choose not to respond or participate in your study. An open invitation doesn't allow for any non-response to occur.

19. Solution: C

20. Solution: C.

Recall that ***statistics*** are characteristics of a ***sample***. ***Parameters*** are characteristics of the ***population of interest***. Statistics are used to estimate the value of a population parameter.

21. Solution: C.

Most of the time random samples, i.e. those chosen using a ***random sampling strategy*** will be representative of the ***population of interest*** they are chosen from.

Despite the fact that opinions are based on personal preference, this doesn't automatically mean bias. Because the sample was selected randomly, it will be representative of the population. It is very important that all of the telephoned individuals responded with their opinion on the question. If some didn't respond, then it would be only the data from those who "volunteered" and the data would NOT necessarily be representative of all employees. If they had chosen the sample of 300 people non-randomly, such as only selecting the category "staff", the sample would not be random and not representative.

22. Solution: False.

An ***SRS*** requires by definition that **EVERY** unit or individual has an equal chance to be chosen. In this example, there are three groups presumably of different sizes and we are selecting 5 from each group. As a result, it may be more likely to be selected from the group of females if it contains less individuals than the male group, etc. This scenario is actually a ***stratified sample***, since they are ***randomly selecting*** 5 from each of the 3 groups. We don't know if the gender groups are equal in size though, so it isn't a ***random sample***. This would be an example of ***probability sampling*** though as that just means it involves an element of chance in choosing which individuals will be in the sample.

23. Solution: C.

It is performing a census of 2 randomly selected sections.

24. Solution: A and C.

Random sampling is NOT correct because each sausage doesn't have the same probability to be chosen. They are chosen from each group and for example, there are many less weisswurst.

Cluster is **NOT** correct because it would involve a census of a few randomly selected groups.

Stratified involves a ***random sample*** of each group and ***probability sampling*** is correct too. ***Probability sampling*** simply requires the element of chance to be involved in the ***sampling strategy***.

25. Solution: C.

Statistic- characteristic of a sample e.g. $\bar{x}, s$
Sample – subset of the units of the population that we actually collect data from
Parameter- characteristic that describes the population e.g. μ, σ
Sampling error – difference between the statistic and the parameter it estimates. It is attached only to statistics.

A. False, number of games won, 11, is the highest number of games run across the league ∴ it is a parameter
→ when would it be a statistic? If we first choose some friends and then found the maximum value for that subset of friends
B. This is False, since it is a statement about the total individuals in the league. It is describing the population, so it can't be a sample.
$\boxed{C}$ is True. 18 is a statistic (based on some of the games) so it has sampling error.
D. False, the standard deviation is based on a sample of 11 of 66 games. It is describing variation in a sample

26. **Solution: B.**

The **units** in the sample are the objects being studied and they studied the 600 schools from 10 selected cities.

NOTE: all Canadian schools would be the population and the 40% of schools that received the intervention would be the replicates in one of the treatment groups.

27. Solution: D.

Sampling error is the difference between the statistic and the parameter it estimates.

28. Solution: A.

We don't need any information about subjects to do an ***SRS***.

NOTE: ***Stratified*** and ***cluster*** require knowing information about each individual with respect to the variable by which you intend to stratify or cluster.

Case control sampling isn't a ***sampling method***, but a type of ***observational study***. So, this answer isn't relevant.

29. **Solution: A**. 90% is a statistic and therefore has sampling error.

B. False – the ***sample*** is the 50 who actually responded
C. A ***statistic*** is a characteristic of a sample. The 12.5% tells us what fraction of the population responded to the questionnaire so it isn't describing the population or a statistic.
D. False – the ***population of interest*** is parents with children in certified childcare.

30. Solution: D.

A. False – we don't take ***random samples*** of each colour ∴ not stratified.
B. False – each Smartie doesn't have an equal chance; he scoops them out.
C. Convenience is a ***non-probability sampling*** in which we take the most accessible, but here there was some probability involved when selecting the colours and mixing containers.

31. Solution: D.

It was not a ***random selection*** because all cookie types were not equally likely to be chosen. He just chose the 3 most abundant types.

∴ $\boxed{D}$ Non-random sampling is correct because it involved chance that is not equal across all units/individuals.

32. **Solution: B**. A ***statistic*** summarizes a variable in the ***sample***.

A. 6 doctors – ***population*** ∴ $parameter$ ∴ False
$\boxed{B}$ 15% is true based on observing 250 patients which is a ***sample***
C. 45 minutes is not connected to a particular sample. It is a ***parameter*** because it refers generally to patients of the hospital.
D. 250 patients is the size of the ***sample***. It doesn't summarize a ***variable*** associated with a sample ∴ not a statistic.

3. Study Design

1. This ***research question*** is about whether conditioning chemicals are necessary to prevent premature death of fish in a fish tank. For a ***randomized block design***, I would begin by blocking my sample of fish by their species, i.e. a block of guppies, a block of danios, etc. Then I will use ***randomization*** to randomly assign some of each species to different treatments. This means, for the guppy group, treatment 1 would be with chemical, and group 2 would be treatment 2, without the chemical.

The ***explanatory variable*** is the variation in chemical and the ***response variable*** is whether or not the fish survive.

2. **Solution**:

homeowner_ID	age	salary	device_type	room	type_of_command	number_of_commands
1	45	$80 000 to $110 000	Amazon Echo	kitchen	shopping list	33
1	45	$80 000 to $110 000	Amazon Echo	kitchen	look up info	54
1	45	$80 000 to $110 000	Amazon Echo	kitchen	play music	15
1	45	$80 000 to $110 000	Amazon Echo	kitchen	listen to news	17
2	33	<$50 000	Google Home	living room	shopping list	2
2	33	<$50 000	Google Home	living room	listen to news	24
2	33	<$50 000	Google Home	living room	play music	40
2	33	<$50 000	Google Home	living room	look up info	21
2	33	<$50 000	Google Home	living room	set alarm	4
3	56	$50 000 to $80 000	Google Home	basement	play music	54
3	56	$50 000 to $80 000	Google Home	basement	listen to news	23

I think it is pretty easy to tell the best way to label each column, i.e. age, salary, device type, etc., so I don't believe you need a key to explain any of the titles. The salaries were given as ranges of values, not exact ratio data. This is a ***categorical variable***, not a quantitative one.

3.

1) ***Random sampling*** is a method of sampling used to obtain the individuals or units we will use in a research study. ***Randomization***, on the other hand, is a method use to assign individuals from our sample to various treatments groups. If ***randomization*** is used, individuals would be randomly assigned to different ***treatment*** groups. The difference is when each of them is involved in our process. ***Random sampling*** is used in the sampling as a ***sampling strategy***, but ***randomization*** is a component of the ***study design***. For example, see question 1 in this section, randomization was used to randomly assign fish to each of the treatments, "with chemical" and "without". A simple example of random sampling is to put 100 names into a hat and mix them up and select 20 names. The twenty names would be a random sample.

2). ***Random sampling*** helps to ensure that our ***sample*** is representative of the ***sampling frame***, and the ***population of interest*** (hopefully). It allows every unit to have the same chance of being chosen. ***Randomization*** aids to distribute the ***variation*** in the individuals in your sample across all of the treatment groups. This is done to prevent the negative effects of confounding variables by making our ***treatment*** groups fairly similar in composition to the individuals in the sample.

4. To obtain full marks, make sure to have:

1. **Study Design Consistency**: The study employed a ***matched pairs design*** with ***randomization***, ensuring that participants with similar characteristics were paired **before *random assignment*** to ***treatment*** groups. This method enhances the **reliability** of the results by controlling for variables that might confound the relationship between the treatment and the outcome.

2. **Relevance to Research Question**: The chosen design is relevant as it directly investigates the impact of tattoos on the participants, comparing outcomes between those with previous tattoos and those without. By focusing on this specific comparison, the study is well-equipped to provide data that addresses the research question effectively.

3. **Addressing Client History**: The study design incorporates participants' histories by classifying them into matched pairs based on whether they had previous tattoos. This concrete approach ensures that any effects observed can be attributed to the treatment rather than pre-existing differences, thereby enhancing the validity of the findings.

Sample Example: I will take a sample of 20 pairs of relatives who want to get a tattoo, as it is likely that relatives have similar pain tolerances. I will use ***randomization*** i.e. randomly assign the relatives to ***treatments*** based on the brand of numbing cream; Dr. Numb, Linopain, etc. These brands are the levels of the ***explanatory variable***. The ***response variable*** I would measure if the level of pain individual's experiences during the tattoo process. One relative will receive Dr. Numb and the other Linopain. Since previous tattoos may influence a subject's perception of pain level, I will record whether or not individuals have a previous tattoo and make that a cofactor in my study.

5.
a) Mortgage rates, for example, 4.39%, are ***quantitative***, ***continuous*** and ***ratio data***.

b) The analysis goal here is assessing evidence for a claim. We want to determine if mortgage rates given to clients are influenced by their level of education.

c) The ***explanatory variable*** is the type of post-secondary degree; there are two levels, "college" or "university". Therefore, there are only 2 comparison groups: college and university.

d) The data should be ***independent samples***, **not** ***matched/paired samples***.

Each of the ***comparison groups*** contains different individuals, who either went to college or university, rather than a ***repeated measures*** study where we have repeated measures for each individual or related individuals split into two groups.

6.
a) The researcher likely wanted to use ***stratified sampling*** so that they were taking a ***random sample*** of EACH field in order to control differences that may occur due to different types of plants in the various fields, as well as the fact that different famers vary in farming practice, such as watering. By using ***stratified sample***, they could obtain data from wheat/Farmer A, wheat/Farmer B, corn/Farmer A, corn/Farmer B and determine if the fertilizer is effective. This would allow for all for all of the combinations of type of plant and farmer to be studied.

b) The ***strategy*** outlined was ***stratified sampling***, but not a ***multi-stage sample*** and it was ***random***. They took 2% of each type/farmer combination and applied fertilizer to half of them. In order to be multi-stage, they would have to add in another step in the sampling process. For example, they could take a ***random sample*** of farms in Middlesex County first, if that is the region they want to study. Then, the second step could be ***stratified random sampling***. We could make the second step ***non-random stratified*** by taking a constant number of crops (4, for example) for each plant/farmer combination rather than 2% from each stratum.

7.
a) The ***sampling frame*** consists of owners of Dell laptops who registered their laptop after purchase.

b) The ***population*** is Dell laptops and the sample is 80 owners; 40 who made a claim and 40 who hadn't made a claim. The ***explanatory variable*** or factor of interest was the nature of the use of the laptop, i.e. uses related to school, business, or personal/recreational use. The ***response variable*** is the "status of the warranty claim", i.e. made a claim or not

c) This ***study design*** is a ***Case-control study*** which is observations, i.e. not an experiment. Recall that Case-control studies involve researchers selecting individuals based on their value of a particular response variable, in this case "warranty status" and then collecting data about the explanatory variable, type of use of the laptop. Here, the cases would be the 40 (registered) laptop owners who had made a warranty claim and the controls are the other 40 who didn't make a claim.

8.
Sample answer: I would take a sample of 20 individuals who are willing to participate in my study and I would ***randomize*** them to 1 of 2 different ***treatment groups***; tasting a sample of "Cocsi" made from sugar and tasting a sample of "Cocsi" made from artificial sweetener. These two groups illustrate two levels of the ***explanatory variable***, type of "sweetener". I have used ***replication*** by using 20 individuals, 10 in each ***treatment*** group. I would made the study ***blind*** so that participants are unaware of which treatment they are getting. After tasting, I would ask them whether they believe they tasted the sample of "Cocsi" sweetened with sugar or artificial sweetener and record this data. Since the researchers are concerned about familiarity with "Cocsi", I would ask each individual if they have previously drunk "Cocsi" and use that as data on a ***cofactor***.

9.
a) Frequency of visits is measured as the number of times a bee visits a flower, so this is a ***quantitative variable*** and it is ***discrete***, i.e. countable, like 3,4,5 times. It is also ***ratio data*** since 0 is meaningful and 0 means the bee doesn't visit the flower at all.

b) There are 4 ***comparison groups*** being studied that all vary in number of ppm of caffeine.

c) The ***comparison groups*** are ***matched/paired*** and **not** ***independent***. It is a ***repeated measures design*** as they watch EACH bee as it visits EACH "station" of flowers with different levels of caffeine. They can then "match" a particular frequency in one group with that in the other groups.

d) They want to know if the frequency of visits is influenced by the amount of caffeine, so this analysis goal is to assess evidence for a claim and NOT to estimate the value of a ***population characteristic***. We are not estimating the value as so many visits, but rather looking for whether more caffeine results in more visits.

10.
a) ***Blocking*** allows for better control by ensuring that the ***treatment groups*** in the experiment have a similar composition based on the blocking variable. This is achieved by ***randomly assigning*** individuals from each block to the different ***treatment groups***.

b) I would use ***blocking*** to control for "age of toddler" in this study. I can use various age groups as my blocks, such as 2-2.5, 2.5-3, etc. I can't hold the ***variable constant*** for age, because we want to study if the amount of time toddlers watch TV influences sleep and the amount of sleep of toddlers changes with different age groups. If I separate the toddlers by age, I can assign the amount of time spent watching TV as the treatments.

Other answers are possible and correct if you explain and pick a variable that can be blocked and explain why the variable can't be held constant.

Note: You can't choose the amount of time spent watching TV as a ***blocking factor*** as that is already the ***explanatory variable***. You also can't block the amount of sleep as that is the ***response variable*** in the study.

11.
a) The likelihood of angry outburst was measured as "high", "moderate/average" and "low", so this is ***categorical data***, not ***quantitative***. It is also ***ordinal*** as there is an ordering to the level of anger.

b) There are 3 ***comparison groups*** or "samples" that are "high likelihood", "moderate/average likelihood" and "low likelihood".

c) The ***comparison groups*** are ***independent***. Each individual in the study is characterized by either "high", "moderate/avg" or "low" in terms of anger level. There are not family members being "***matched***" into different anger groups. There is no connection of any individual in the "high group" to a particular person in another anger level group.

d) They want to see if personality traits influence blood pressure, so this is assessing evidence about a claim. For example, they might be seeing if anger likelihood causes higher blood pressure.

12.
a) The ***population of interest*** was Facebook platform users. The ***sampling frame*** was all English-language Facebook platform users who had their account for at least 5 years. The ***sampling frame*** is the sub-group of the population from which you obtain your sample. In this case, since the questionnaire if presented over the course of one week, it is likely that the ***sampling frame*** would be those English-language users who logged on during that time period.

b) The ***explanatory variable*** is the number of Facebook social media platforms used, e.g. Instagram, Messenger, etc. The response variable is the reaction to the name change.

c) The ***explanatory variable*** is measured as a number, like 0,1,2, etc. so this is a ***quantitative*** and ***ratio variable*** as 0 is meaningful and would mean using 0 social media platforms. This is also a ***discrete variable*** as it is **countable**. The ***response variable*** is given as 1=don't like, 2 = like it,3 = didn't know, 4 = prefer not to answer so this is a ***categorical variable***. It is also ***nominal***, since all of these responses are not "ordered".

13.
In this research scenario, the ***explanatory variable*** is hat-wearing and the r***esponse variable*** is balding. For a ***case-control study***, I will obtain a sample of cases of men who have balding hair. For my ***control***, I will select a sample of men who are not balding. Because age is a possible ***confounding variable*** that could affect my data, I will determine the age of each of my "***cases***" and then use a "***control***" of a man of the same age, thereby "***matching***" up each case/control by age of the man. I would give each man a questionnaire to report how often they wear hats and record the data as an ordinal variable, 1=never, 2=sometimes, 3=often, 4=…etc.

14.
a) The ***population of interest*** is abandoned dogs and cats in the city of Freetown. They are trying to determine how frequence common medical supplies are needed to treat abandoned dogs and cats in Freetown.
The ***sampling frame*** were those animals in the Freetown Pound when the study was conducted and the sample was the 10 animals who were most recently caught.

b) **Possible Answer 1**: I believe the bigger problem with the ***sampling plan*** is ***undercoverage***. The ***sampling frame*** used was the 10 most recently caught animals in the Freetown pound. Since these animals were just caught, they would likely be in pretty good condition. Many could be house pets who "escaped" their owners/homes and therefore, they might not require much in terms of medical supplies. Since the pound is called when these "problem animals" are disrupting the neighbourhood, it may also be more likely that these animals are runaways, rather than animals without any home. It would be better to obtain a sample of abandoned animals from various parts of

town. It is also likely that the most recently caught animals in this case would be in better shape than animals caught some time ago, i.e. have less fleas, etc.
Possible Answer 2: I think ***sampling bias*** is a bigger problem than ***undercoverage*** in this scenario. Despite the fact that there may be some ***undercoverage*** as a result from only choosing animals caught in the neighbourhoods surrounding Freetown pound, I don't think it would be misrepresentative of Freetown's abandoned animal population. The fact that the pound used a convenience sample to test the 10 most recently caught animals is much more problematic. The most recently caught animals have not been exposed to tons of animals already in the pound, who have lots of fleas and other medical issues. As a result of only looking at these 10 animals that were last caught, we may be badly underestimating the number of animals who require medical supplies for various ailments.

15.
a) Assessing evidence for a ***claim***. They are trying to determine which toy sound is most attractive to babies, i.e. toys that rattle, or squeak, or crinkle, etc. It is not focused on estimating a value for a population characteristic.

b) There are 4 ***comparison groups*** that are comprised of toys that make various sounds: rattle, crinkle, squeak and bells.

c) The ***comparison groups*** should be ***matched/paired***. This way, we can measure every baby's response to all of the possible toy noises and compare the data for all 30 babies. It is a repeated measures design.

16. a) Means plot

b) There are 3 ***variables***: delay time, destination, and airline brand

c) International flights have longer delay times than domestic, regardless of the airline brand chosen. We can see this based on the triangles being higher vertically than the circles. Also, international flights run by Air Canada have the highest mean delay time, since the triangle representing the mean is higher than all the other means. It is 45 minutes.

17. a) This is a side-by-side boxplot.

b) There are 2 ***variables*** being represented: first year salary ($) and Undergraduate module

c) There are lots of different answers, but make sure you discuss both centre and spread and support your advice. I would recommend the Statistics module. Even though the median is very slightly lower than that for Actuarial Science ($85,000 instead of $90,000), the variability or spread of the first-year salaries is much lower than that of Actuarial Science. The IQR for Actuarial Science is $15,000 and it is only about $6000.

The ***range*** of Actuarial Science is $120,000 - $725,000=$48,000 whereas the ***range*** for Statistics is only $95,000 - $75,000=$20,000. Because of the lower ***variability***, in Statistics your first-year salary is likely to be around $85,000 but with Actuarial science you could get a much lower salary due to the larger spread.

NOTE: You could just as easily argue you would recommend Actuarial science because it has a much higher maximum salary, so there is greater potential for an extremely high salary!!

18. **Solution: A, C.**

A ***confounding variable*** is any variable other than the ones we are studying that varies/differs across treatment groups. The ***treatment*** in the question is pill or no pill. Because the pill group puts something in their mouth and the "no pill" group doesn't, these groups differ in whether they are actually putting something in their mouths. So, the act of taking a pill differs and is a ***confounding variable***. This is why we typically use *a placebo* pill when the active ***treatment*** is not a pill. As far as activity levels goes, we used ***randomization*** so we can trust that our ***treatment*** groups are similar in composition and ***confounding*** would not occur as a result.

NOTE: We are not saying all individuals have the same activity level but ***randomization*** ensures when we compare ***treatment groups*** that the groups are similar in terms of activity level. The rate of weight loss is the variable we are measuring, so it can't be a confounding variable. Sex is a ***confounding variable*** as it does differ and it is not the ***explanatory*** or ***response variable***.

19. **Solution: A.**

Replication refers to using more than one individual/unit in a ***treatment group***. We are investigating weight loss in males, so the ***replicates*** are individual males. The ***explanatory variable*** is the type of pill and there are two levels: new pill and currently available pill. The researchers are also ***blocking*** by daily activity level (not very, moderately, very). A ***treatment*** would be a combination of type of pill and activity level, example, moderate activity/new pill. The individuals in this combination would be replicates that could be compared to other activity level/pill type combinations.

Randomization refers to the random assignment of individuals from the sample to the treatment groups. The random assignment to the pill type is "they split the 50 males into three groups based on their activity level separately, they put the men's names into a hat and randomly pull-out half of the names. These pulled names are assigned to one group who will receive the new weight loss pill. The leftover names are assigned to a group to receive the currently available weight loss pill.

The current weight loss pill is **NOT** a placebo as a placebo is a non-active treatment and this pill is a real medication.

20. **Solution: D.**

A ***confounding variable*** is any variable other than the ones we are studying that varies/differs across ***treatment groups***. They are recruiting men of the same age, so that is not confounded. They are studying those who have and have not had previous heart attack, so that is not confounded. Also, they are studying whether exercise reduces heart attack so past exercise habits is not confounded either. The family history of heart attacks would be a ***confounding variable***.

21. **Solution: C.**

She doesn't randomly assign individuals to treatments, so ***randomization i***s not done. She does use random sampling as she uses a random number table. She is repeating the experiment with many students, so replication is not missing.

22. **Solution: D.**

We should use ***randomized block design*** so we can first block the sausages based on the four types and then for each of the four types, apply each treatment or preservative.

23. **Solution: B.**

The students are sent personalized emails inviting them to participate, so many will not participate and therefore ***non response*** is the most problematic.

24. **Solution: A.**

We are trying to see the effect of a new allergy medication on seasonal allergy symptoms. The allergy symptoms are the *response* ***variable*** and the medication type is the ***explanatory variable***. This is an experiment and not an observational study, so C is not correct. The population is allergy sufferers and those interested get separated into groups based on the severity of their allergies. This is a new variable introduced. Next, individuals from each of the groups are selected to make up the sample that will actually participate in the study. So, they asked people to volunteer based on an open invitation and then used a stratified sampling process. However, the question doesn't ask about sampling method, but rather study design so the answer is not B. Because we haven't subdivided this sample in any way before assigning to the treatments (which are based on the ***explanatory variable***), this design doesn't involve any ***blocking***, so the answer is not D. We have assigned individuals from the entire sample of 75 to their treatment groups in a random manner so it is ***completely randomized***.

25. **Solution: A.**

Here, no ***treatment*** is imposed, they are just giving questionnaires to 32 women who had babies with birth defects and 32 who didn't have babies with birth defects. So, it isn't an experiment, so therefore it is an observational study and must not be B, C or D. Recall, a *Case* ***control study***= type of observational study in which we select a group of cases who have a particular value for the ***response variable*** and a group of controls that don't have that value of the ***response variable.***

26. **Answer: B.**

Here, the grade categories are the *strata* and the 24 students are the ***stratified sample*****.** She randomly assigns the students to different values of the explanatory variable (different programming languages) and she is investigating the speed (response variable). This is **a *completely randomized* design**.

27. **Answer: A.**

Since John is only interested in the overall proportion of residents in the municipality who plan to vote for him as well as the proportions of people within each area who plan to vote for him, the sampling design needs to ensure representation of residents in each area. So, we need to use ***stratified random sampling. Cluster*** would involve randomly choosing a few areas and surveying all residents in those areas, so he wouldn't end up with information from all areas this way. ***SRS*** wouldn't guarantee that he ends up with a sample of residents from all areas. ***Randomized block design*** is a type of experiment. The nature of John's question is descriptive and he just wants to find out proportions, not evaluate the impact of a treatment so this isn't applicable. ***Multi-stage sampling*** would involve randomly choosing some areas, then randomly

selecting a sample resident from each of those areas to survey so he wouldn't end up with information from all areas with this method.

28. **Answer: C.**

This is a ***matched pairs design***. The researchers are giving the drug valproate and a different drug and they are randomly assigned to groups and then after two months, they are switched from what they are on to the other drug. This is an experiment, so it isn't A or B as they are ***observational studies***. It is matched pairs and not completely randomized as the individuals are paired since in the second part of the study, they are switched to the other drug.

29. **Answer: B, D**.

A is false, as the ***explanatory variable*** is the number of legs or whether they have any removed or not. This study is experimental, so B is correct. The researchers are comparing the survival in the two groups of spiders, based on whether the spiders have all their legs or not. C is incorrect as this is not an ***observational study***, so it can't be case- control. A ***case-control study*** is a type of observational study in which we select a group of cases who have a particular value for the ***response variable*** and a group of controls that don't have that value of the response variable and then look for differences between the groups. D is correct as the response variable is the survival of the males. The researchers are recording survival between the two groups.

30. **Answer: A**

The population of interest is the Carolinian Population which consists of woody debris and rocks in natural woodlands as well as construction sites***. Undercoverage*** occurs when some group(s) in the population are left out in the process of choosing the sample and here it occurs because the sample is only taken from woody debris and rocks.

Non-response bias occurs when a selected individual can't be contacted or refuses to cooperate. In this context, it is impossible as a lizard can't refuse to cooperate or provide a blood sample. So, there isn't any non-response occurring in this scenario.

Sampling error is the difference between a sample statistic and the population parameter and it is inevitable and always occurs when taking a sample, so it is not a statistical problem.

Confounding occurs when a variable other than the explanatory and response variables are varying with the factor(s) of interest, such that we can't tell their individual effects. In this situation, the study is interested in genetic diversity of the population, but there is no response variable and it looks like an exploration. Therefore, there are no confounding variables present.
The answer is A.

31. **Answer: C.**

The mail carriers are divided into 4 groups according to sex and mode of transport and then a random selection is made from each group, so this is ***stratified sampling***.

32. **Answer: A.**

You are randomly selecting half to go to a smiling teller and half to go to a non-smiling teller, so this is ***randomization.***

33. **Answer: B.**

Completely randomized design means participants are randomly assigned to treatments. In an observational study, an experiment is not being performed and there is no manipulation of variables, placebo, etc. In randomized block design units are grouped into blocks according to known or suspected variation between blocks (similarities within blocks). The variability between the blocks is less than the variability within blocks.

In this question, they don't subdivide the plants into groups, so it is not a block design. They do randomly assign groups of plants to each area, so it is a completely randomized design.

34. **Answer: B.**

We have a sample of 15 chimpanzees and each chimp is put in a situation where they are confronted with either a friend or a "non-friend". The researchers expect the chips in the friend situation to pull the trust rope more often that delivers food to both of them. The treatment is "friend vs not friend", related to the explanatory variable. The nature of the rope pulled is what the researchers are observing as a consequence of the treatment, which is the ***response variable***.

35. **Answer: B.**

They are taking the population of rats and subdividing it by size class, so this is the ***blocking variable***. It is a blocking variable as it is controlled by the experimenter.

36. **Answer: A.**

The ***explanatory variable*** is the species and the response variable is the wavelength of light to which the cells in the dissected eyes are most sensitive. There is no blocking or stratifying variable.

37. **Answer: B.**

They are separating the rats into immune-compromised or not immune-compromised, so that is blocking. Then, the researchers assign half to get newsprint on them and half to get water, for each group. The answer to the type of study design is ***randomized block design***.

38. **Answer: A and B.**

They chose Southwestern Ontario communities, so this is not random sampling since they only chose Southwestern Ontario ones. So, C is false. They chose 4 of the 8 to see ad 1 and the other 4 to see ad 2 so this is stratified sampling, so B is true. They did use ***randomization*** by taking an SRS of 4 of the communities to assign to one of the treatments, i.e. randomly assigning individuals to treatments, so A is correct. We did not use ***blocking*** because we did not first split up the sample of 8 communities based on some pre-existing characteristic before assignment them to treatments. So, D is false.

39. **Answer: D.**

Since only 65 of the 90 were handed back in the biggest problem is ***non-response bias.***

40. **Answer: A, B, C, D.**

A is correct. ***Replication*** is used because there are 30 participants so the experiment is repeated. B is correct. There is randomization because the order of the pills were randomly assigned to each participant. ***Randomization*** is necessary to avoid confounding factors, such as time of year or carryover effects of the Sleepeze. More people are likely to experience insomnia closer to the holidays, so by having some on a placebo and some on the pill and having a week off in between, this will balance out the effects of confounding variables. This experiment is ***double-blinded*** because the participants and researchers don't know which pill is Sleepeze. ***Blocking*** is also correct as a block is an individual patient. A single individual is identical in every way to themselves, so in a repeated measures design like this one, the "block" is each person. Recall that a block is a group of individuals similar to each other in a meaningful way e.g. Same gender, same age group and here, the same person. The blocking here avoids confounding between the person's underlying health and genetics and the two pills, the placebo and Sleepeze.

41. **Answer: A.**

We want the design that would enable the researcher to account for genetic background while also answering the question "what is the impact of water-soluble fertilizer on the aquatic animal?" The different amounts of fertilizer would be the treatments (fertilizer is the explanatory variable). In order to account for genetic background, the researcher would need to have variation in genetic background and be able to make comparisons among all the different backgrounds. The researcher must use a study design that allows the distribution of units from different genetic backgrounds into each treatment. This would involve a ***randomized block design***. The researcher would subdivide the 1000 eggs by genetic background (mother) and then randomly assign eggs from each subdivision or ***block*** to the treatments. This would ensure that each treatment level has representation from each mother.

42. **Answer: A.**

Customers are using ***non probability sampling*** techniques, so this is ***selection bias***.

43. **Answer: D.**

A ***treatment*** is a specific experimental condition applied to the subjects. A treatment is a combination of factors, where each factor is an explanatory variable. In this experiment, they are interested in whether the leaves of gingko trees can influence the post-lunch dip AND whether the timing of taking the leaves influences the effect. So, we have two ***explanatory variables or factors***, i.e. what they take and what time they take their pill. So, a treatment would be a combination of these two things, and that would mean D is the answer as it addresses the pill and the timing.
A, the number of times a participant missed an "e" is the response variable.
B is not a complete treatment as it doesn't discuss timing of the pill.
C, is a ***unit*** in the sample of this study

44. **Answer: D.**

It says the quality control group recognizes that employees vary in experience, so they subdivide their 200 selected employees into groups based on experience. This is the ***blocking variable,*** so the answer is D. This is a ***randomized block design experiment***.

45. **Answer: C.**

A ***blocking*** variable refers to a variable that the experimenter controls, while ***stratifying*** variables are those the experimenter doesn't control, but the subjects bring with them to the experiment, such as male/female, age, past medical events. A blocking variable might involve drug vs placebo or different dosages.

The stratifying variable is the degree program. The response variable is their marks and the explanatory variable is the student status i.e. new student or repeating student.

46. **Answer: D.**

The ***explanatory variable*** is the vaccine and participants are either given a placebo or a vaccine, so the variable is manipulated and it is an experiment, ***not an observational study***, so it can't be Case Control. For ***matched pairs and randomized block designs***, we would first have to subdivide our sample by a pre-existing characteristic before assigning them to treatment. In this study design, there is no subdivision of the sample before the treatments are assigned. The sample is immediately subdivided into treatments in a random manner, so this ***is completely randomized.***

4. Summarizing and Exploring Data

1. **Solution: A,B,C.**

This data is ***categorical***, so proportions or percentages summarize it well. We can also use bar graphs and pie charts to summarize ***categorical data***. A ***box plot*** is not correct, since it is used to summarize one ***quantitative variable***.

2. **Solution: D**

This data is ***categorical***, since it is "shopping list, reuseable containers, etc.". A mean would be used for ***quantitative*** data and a histogram would display one ***quantitative variable***. A ***boxplot*** and a mean plot also summarized one ***quantitative variable***. A ***dot plot*** and ***scatterplot*** are to summarize 2 ***quantitative variables***.

3. **Solution: D.**

A) is false since the median does not increase in the graphs going from left to right, i.e. low to moderate to high.
B) is also false as the maximum is the largest value and the maximum is way above 100 for moderate, but exactly 100 for high. **Note** that ***outliers*** still count as the highest value, rather than just looking at the upper "line" of the boxplot.
C) is false because the Q3 has a value of 60, but that does NOT mean that there is a data point at 60, it could be the average of two values, for example.

4. **Solution: A, B, C are all correct.**

5. **Solution: B.**

This data has a mean that is greater than the median, which occurs in a right skewed graph.

6. **Solution: C.**

This data is ***categorical*** or ***qualitative*** so we can use percentages or proportions, but not a mean, median or range.

7. **Solution: B, D.**

Recall that standard deviations are affected by ***skew*** and ***outliers*** just like the mean, and **MUCH** more so than the IQR, for example. Since the female heart rates have a higher spread based on standard deviations than males, but a lower spread than males for IQR's, the female standard deviation must be affected by outliers and/or skew.

8. **Solution: A.**

There are MANY more students to the left or below the high "peak" of 70-80% than above it, so it is left-skewed.

9. **Solution: C.**

A) is false because there are about 19/64 students in the 62 to 65 range and this is NOT half. This is not ***symmetric*** as if we split it into two equal parts down the middle, they are not mirror images, so B) is false.
C) is true since the shortest person lies between 59 and 62 inches, so they are definitely less than 63 inches. The tallest person lies between 77 and 80, but we can't say they are at least 79 inches as they could be 77.1.

10. a) means plot

b) There are three ***variables*** being summarized: course grade in percentage which is quantitative, gender (***categorical***) and type of final exam (multiple-choice/short answer, also ***categorical***)

Make sure you don't write the ***variable*** is **MEAN** course grade as this will **NOT** get you the mark. The variable is course grade only.

c) There are several possible correct answers.
For example,

1. The greatest ***standard deviation*** occurs for females writing multiple-choice exams. This can be seen by the longest length in terms of the vertical lines. For this group, the standard deviation is approximately 69 to 71.5, or about 2.5%.

The smallest ***standard deviation*** occurs in the male short answer group, with the standard deviation being from 72 to 73, or 1 %.

2. The type of final exam, multiple-choice vs. short answer doesn't seem to influence the mean grade for males as the mean for male taking short answer is about 73 and for males taking a multiple-choice exam about 73.5%.

11. **Solution**:

You can get full marks if you pick one of these graphs and explain why you think that graph is best suited. There is no "one correct answer" here.

Example. I think the ***mosaic plot*** provides a clearer picture of the interaction between type of wine and type of music because being able to see the relative frequency of "level of enjoyment" for different combinations of wine and music is easier to interpret. For red wine, the type of music doesn't seem to affect the level of enjoyment, i.e. the same relative areas of the segments, whereas with white wine, the type of music does seem to affect enjoyment, with jazz having higher levels most often. With the bar graph, it is more difficult to see the relative impact of the wine/music because the interpretation is based on different numbers of observations for each treatment in the study.

12. **Solution: A.**

They are reporting on which agricultural sector they belong to, which is ***categorical***, like beef, poultry, etc. So, this is ***nominal data***. **Note** that ***discrete*** and ***continuous data*** refers only to ***quantitative variables***.

13. **Solution: B.**

The ***response variable*** here is the type of protists. He is trying to determine that proximity to human habitation influences the type of protists in the water. The type of protists is ***nominal***, or just names.

14. **Solution: D.**

There are 6 ***treatments*** in this study. There are two ***explanatory variables***, counselling and anti-anxiety medication. For counselling, there are 2 levels, i.e. no counselling or counselling and for medications, there are 3 different types, so altogether that makes 2(3)=6 different treatments.

15. **Solution: D.**

16. **Solution: A, B, C.**

Female Dispersion index is a ***quantitative variable*** and it is ***ratio data***, since 0 would mean 0 out of 4 women. The data is ***discrete*** and **not *continuous*** because discrete data are countable and here, we can only get 0, 0.25, 0.5 ,etc. not every decimal, which would be continuous.

17. **Solution: A.**

Here, the points you can win are a ***quantitative variable***, so it is either ***interval*** or ***ratio***. Since $0 would mean winning nothing, it is ratio and NOT interval. The number of points is countable, or ***discrete***, ***not continuous*** which would represent a range of values, and a ***variable*** we could make more accurate by increasing our precision.

18. **Solution: B.**

The ***response variable*** here is ***quantitative***, the amount of time spent fluttering, so the answer cannot be A) which is reserved for ***categorical data***. It can't be a ***histogram*** or ***boxplot*** as that is for ***1 quantitative variable***, and here we are determining if there is a relationship between ***two quantitative variables***, i.e. (x,y) pairs for the chickadees and their corresponding mother's data. For bivariate quantitative variables, we can use both a ***scatterplot*** or a ***dot plot***.

19. **Solution: B,C,D**.

Here, we have only one ***variable***, so ***univariate*** data, and it is ***quantitative*** since it is volume measured in gigabytes. So, we can't use a ***bar graph*** as that is for 1 or more ***categorical variables***. So, we could use a ***boxplot***, ***dot plot*** or a ***histogram***.

20. **Solution: A.**

If you look at the box for right female, it shows a much larger distance from the median down to the minimum than from the median up to the maximum value.
B) is false as we have no idea from a ***boxplot*** how many individuals were sampled in each group.
C) is incorrect as the range for right males is from 10 up to about 65, so range = max – min= 65 – 10 = 55. Finally, D) is false as the median is NOT consistent for females right/left or for males right/left.

21. **Solution: A,C.**

The ***explanatory variable*** is the number of kids and the ***response variable*** is 0=not at all, 1 =slightly accurate…etc. so that response variable is ***categorical*** and ***ordinal***, i.e. it does imply order. As a result, it is **not *quantitative*** so the median is not appropriate and neither is a histogram which is used for one ***quantitative variable***. This question represents ***univariate*** data as well since they are only talking about summarizing the response of the mothers, i.e. one variable. A) is correct, since relative frequencies, proportions or percentages can be used for ALL types of data and a bar graph C) is also correct since we do have categorical data.

22. **Solution: B.**

A) is false since the second column is relative frequencies, it should add up to 100% and it does. For non-life threating, the 3rd number, we have 18+3+27 = 48% which is approximately half, so B) is correct.
C) is incorrect as the data for AIC is ***categorical***, so there is no discussion of median, which is a measurement of the middle value of quantitative variables. The chart is not for 100 people, we only know the % has to add to 100%.

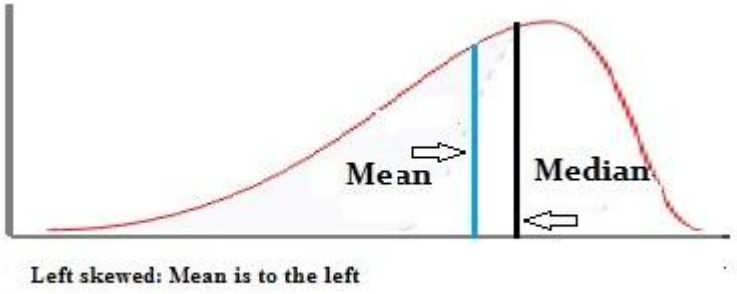
23. **Solution: A, F.**

If you want ONE graph to summarize all of this data, you need to realize which type(s) of variables are involved. Here, we have % of mothers employed in 1975 and 1998 (categorical) and we have ages of children as 0-2, 2-6, >6 years old as well as whether the women. Because the women are categorized by year as well as the percentage employed, we can't use a pie chart. The age of children is categorical and ordinal, not quantitative. We can't use a histogram as it is only for 1 quantitative variable. Here, since we really have 3 variables, all categorical, the only graph that would work is a bar graph, side by side or stacked or a mosaic plot.

24. **Solution: B.**

The mean is less than the median, so left skewed

Q1 medianQ3
|____|____|____|____|
0 3 1.25
$\overbrace{larger}$ $\overbrace{smaller}$



Left skewed: Mean is to the left

25. **Solution: C.**

It is performing a census ∴ whole population ∴ called a parameter

26. **Solution: D.**

A. false – it is skewed right, longer tail from the median to the maximum
B. We can't tell mean from box plots
C. False, the range=max-min= 90 000 – 40 000 = 50 000
$\boxed{D}$ true – above Q3 is $\frac{1}{4}\ data = \frac{25}{4} = 6.25$

27. **Solution: E.**

A. false – continuous
B. false, it is a bar graph.
C. false
D. false – the opposite is true
$\boxed{E}$ true

28. **Solution: D.**

Rankings of players is ***ordinal***. Salaries is ***ratio***, marital status is ***nominal*** and lifespan is ***ratio*** as well.

29. **Solution: D, G.**

The levels are being ranked, but they are ***categorical***, so ***ordinal***. Hated it…loved it are "rankings". This data is ***categorical*** as there is no measurement going on.

30. **Solution: C, D.**

It is ***ordinal***, since there is a ranking to this ***qualitative*** or ***categorical*** data. The "gradings" are 0=low severity, 1=medium severity and 2=high severity, so this involves levelling or ranking of the options.

31. **Solution: D, F.**

Discrete (countable) is true, for example, 18 -20, could be 18.25, 18.5, 18.75 or 20. It is also ratio data since 0 is meaningful in that a score of 0 means a complete absence of the variable.

32. **Solution: A., E.**

2 ***quantitative variables*** is best displayed as a ***scatter plot*** or a ***dot plot***.
C), F) and H) are for ***categorical variables***.
B) is for only ***1 quantitative variable*** and a single box plot can't be used to display 2 quantitative variables.

33. **Solution: D.**

NOTE: ***selection bias*** is bias resulting from an error in getting participation of study subjects.

34. **Solution: D, E.**

The data is percentages or relative frequences, so it is ***categorical data*** and therefore, we can't calculate the mean, or use a ***histogram*** which is for ***1 quantitative variable***. A ***box plot*** is only for ***quantitative variable(s)*** as well. We can use a ***frequency table*** as well. A ***means plot*** is for 1 quantitative and 1 or more categorical variables.

5. Probability Models, Sampling Distributions, & Modelling Relationships

1.Solution: D

Option A, "the number of players on the basketball team who scored more than 10 points in the most recent game," is not a random variable. Although the outcome is numeric (i.e., the count of players), it is neither continuous nor random. The "number of players" is a count (e.g., 0, 1, 2, 3, ...), making it a discrete variable. Since the game is fixed ("most recent"), there is no uncertainty regarding the number of players who meet the criteria. A specific number will be determined, so Option A is incorrect.

Option B, "the number of fans who will purchase tickets for the basketball game tomorrow," can be considered a random variable. The outcome is numeric (i.e., the count of fans) and uncertain (the number will not be known until after the game). However, it is not a continuous random variable, as the "number of fans" is a count and thus a discrete variable.

Option C, "the city in which a game will be played on a randomly selected day," has an uncertain outcome (the result of selecting a day randomly). However, this is not a random variable because the outcome is not numeric; random variables must have numeric outcomes. The "city" is a categorical variable (e.g., London, Toronto, Chicago, etc.), making Option C incorrect.

Option D, "the weight (in kilograms) of a randomly selected player," is a continuous random variable. Weight in kilograms is a quantitative, ratio-level, continuous variable, fulfilling the requirement for a numeric and continuous outcome. The value is uncertain and results from a random process (i.e., selecting a player randomly). This makes

Option D the correct answer.

2.**Solution: D**

An outcome refers to one possible result of a random event. In this case, we are focusing on the number of days the person exercised (not the duration of exercise). A possible outcome could be 4 days, while multiple occurrences of this would form a subset of possible outcomes (i.e., an event). The sample space, which includes all possible outcomes, consists of the values: 0 days, 1 day, 2 days, 3 days, 4 days, 5 days, 6 days, and 7 days.

3. **Solution: B, E**
In this example, the means of the distributions differ, with Distribution 1 having a smaller mean than Distribution 2. This indicates that Distribution 1 is centered at a lower value compared to Distribution 2, so it will be positioned farther to the left on the horizontal axis, assuming the axes are scaled identically for both graphs. However, the standard deviation of Distribution 1 is smaller than that of Distribution 2, meaning that Distribution 1 has less spread and will cover a narrower range of values.

4. **The correct answer is B. The distribution of the population.**

Explanation:

In this case, the Canadian Census is a survey of **all** Canadians, meaning it covers the entire population. The histogram created by Statistics Canada would represent the total household salary for every Canadian household that completed the survey. Since the data comes from the entire population, it represents the **distribution of the population**, not a sample.

Option A, "The distribution of a sample," would apply if the survey was conducted on a subset (sample) of Canadians, not the entire population.

Option C, "The sampling distribution of a statistic," refers to the distribution of a statistic (such as the mean or median) across multiple samples, not the entire population. This doesn't apply to this case, as the histogram is based on data from the whole population.

5. **Solution: True**

Sampling distributions are the distributions of a statistic, derived from all possible samples of a specific size. Each sample size has its own unique sampling distribution, and different statistics will have distinct sampling distributions.

6. **Solution: A**

2.092 represents the standard deviation of the number of migraines over the three-week period for the 121 people in the acupuncture group. Since it is calculated for the sample (i.e., a characteristic of the sample), it is considered a statistic by definition.

7. **The correct answer is A. sampling distribution**.

Explanation:

A **sampling distribution** refers to the distribution of a statistic (in this case, the median) calculated from all possible samples of a given size (25 physicians). In this scenario, if the government were to take multiple simple random samples of 25 physicians and compute the median earnings for each sample, the resulting distribution of medians would represent the **sampling distribution** of the median earnings.

Option B, **probability distribution**, is too general and does not specifically refer to a statistic derived from repeated sampling.

Option C, **Normal distribution**, and Option D, **Binomial distribution**, refer to specific types of distributions and do not directly apply to the context of median earnings in this example.

8. **Answer: A, B, and C**

Let's evaluate each statement:

A. The value, 0.70, is a parameter.
This statement is **correct**. The value 0.70 refers to the HPV infection rate in the population of Canadians, which is a fixed characteristic of the entire population. Since it describes a population, it is a **parameter**.

B. The sampling frame in the epidemiologic study is hospitalized patients.
This statement is **correct**. The sampling frame refers to the group of people from which the sample is drawn. In this case, the study recruited hospitalized patients, so the sampling frame consists of **hospitalized patients**.

C. The value, 0.57 (i.e. 57 /100) should be considered a statistic.
This statement is **correct**. The value 0.57 represents the proportion of cancer patients in the sample who were infected with HPV. Since this value is calculated from the sample of 100 cancer patients, it is a **statistic**.

D. The epidemiological study used a cluster sampling procedure.
This statement is **incorrect**. Cluster sampling involves selecting groups (clusters) randomly and then sampling from within those groups. In this study, the researchers selected patients with or without cancer, not groups or clusters, so this is not an example of **cluster sampling**.

Correct answers: A, B, C.

9.**Answer: B**

Let's evaluate each option:

A. A teacher collects all test scores for his class and creates a histogram of the scores.
This is not an example of a **sample** distribution. It involves the entire class (the whole population in this case), so it represents the distribution of the population, not a sample.

B. A mother selects 15 fries from a large bag of fries in the freezer, measures the length of each fry, and creates a histogram of the fry lengths.
This is an example of a **sample** distribution. The mother is selecting 15 fries from a larger population (the entire bag of fries), and the histogram shows the distribution of fry lengths in that specific sample.

C. An administrator collects 10 different subsets of 4 employees, measures each employee's height, and creates a histogram of the maximum heights from each subset.
This is an example of a **distribution of a sample statistic** (the maximum height from each subset of 4 employees). It describes the variation of a statistic (maximum height) across multiple samples, not the distribution of a sample itself.

Correct answer: B.

10.Answer: D

A is false because a histogram of all first-year philosophy students is a distribution of ages (a variable) for a population (all 1st year philosophy students who live at home)

B is false since it is a distribution of body temperatures for a population

C is false since it only contains information from one sample, while a sampling distribution of a statistic summarizes information from multiple samples

11. **Answer: D**

Let's evaluate each option:

A. The ten-sample means should be the same.
This statement is **incorrect**. The sample means are likely to vary because each sample of 20 students is randomly selected, and there will naturally be some variation between the samples. The means will not be identical.

B. There will be variation among the sample means, and the sample means should follow a Normal distribution.
This statement is **partially correct**, but it's more accurate to say that the sample means will **tend to follow a normal distribution** (according to the Central Limit Theorem) if the sample size is sufficiently large. In this case, since each sample has 20 students, the sample means may approximate a Normal distribution, but this depends on the population's distribution. So, this answer is not fully precise.

C. There will be variation among the sample means, and this variation should be close to the age variation among the UWO undergraduates.
This statement is **incorrect**. The variation among the sample means (i.e., the standard deviation of the sample means) will be less than the variation of the entire population. The sample means will have less variability due to the Central Limit Theorem, which suggests that the distribution of sample means will have a smaller spread than the population itself.

D. There will be variation among the sample means, and this variation should be less than the age variation among UWO undergraduates.
This statement is **correct**. According to the Central Limit Theorem, the variability (standard deviation) of the sample means will be smaller than the variability of the population. The standard deviation of the sample means is given by the population standard deviation divided by the square root of the sample size. Since the sample size is 20, the variation among the sample means will be less than the variation in the population.

E. There will be variation among the sample means, but this variation will be unpredictable.
This statement is **incorrect**. The variation among the sample means is predictable and can be estimated using the population's standard deviation and the sample size. The variation is not entirely unpredictable.

Correct answer: D.

12. The correct answer is **C. A numerical variable whose value depends on chance**.

Explanation:

A **random variable** is a variable whose value is determined by the outcome of a random process or experiment. It is typically numerical, and its value depends on chance. For example, the number of heads in a series of coin flips or the height of a randomly chosen individual are random variables because their outcomes depend on random processes.

Let's go through the other options:

A. An arbitrarily selected variable - This is incorrect because a random variable is not chosen arbitrarily but depends on random outcomes.

B. A random number generated by a computer - While a computer can generate random numbers, a random variable is not simply a generated number; it refers to a variable whose value depends on random events.

D. A quantity with an unknown value - This is too vague. While the value of a random variable is unknown before the random process, the defining feature of a random variable is that its value depends on chance and is numerical.

Correct answer: C. A numerical variable whose value depends on chance.

13. **A is correct**

Let's evaluate each option to determine which variables are discrete random variables:

A. Number of patients who will visit a doctor on Thursday next week
This is a **discrete random variable**. The number of patients is countable (e.g., 0, 1, 2, 3, ...) and can take on only integer values. The number of patients depends on random factors, so it is a discrete random variable.

B. Trunk diameter of a randomly selected Redwood in a National Park
This is **not a discrete random variable**. The trunk diameter is a continuous variable because it can take any value within a range (e.g., 15.2 cm, 15.23 cm, 15.232 cm, etc.), not just specific, countable values.

C. Shirt colour of a randomly selected student in a 2244 class
This is **not a discrete random variable**. Shirt colour is a **categorical variable** (e.g., red, blue, green) rather than a numerical one, so it does not qualify as a random variable in the typical statistical sense. Discrete random variables are typically numeric.

D. The size of the Biol/Stat 2244 class from last term, including all sections
This is Not a random variable because it is already known i.e. from last term, so there is no element of random chance involved.

Correct answer: A.

14. Answer: B.

Let's evaluate each option to determine which can be described as a random variable:

A. The number of innings the Blue Jays had to play before winning the playoff game this past Thanksgiving weekend.
This is not a random variable because there is no element of chance involved as it has already occurred last weekend.

B. The number of stadium seats sold by eBay auction for a randomly selected professional football game.
This can also be described as a **random variable**. The number of seats sold is a countable, numeric value that depends on random factors (e.g., demand, time of the game, ticket prices). Therefore, it qualifies as a random variable.

C. The degree of stress (e.g., low, medium) that a randomly selected student experiences in the days leading up to the 2244 test.
This is **not a random variable**. The degree of stress is a **categorical variable** (low, medium, high), not a numeric one. While it involves randomness in the selection process, it doesn't meet the typical criteria for a random variable, which usually refers to numerical outcomes.

Correct answer: B.

15.A. is false because probability is quantitative

C. is false because probability is ratio data.

$\therefore$ $\boxed{B}$ & $\boxed{D}$ are correct

The probabilities are discrete because there is a finite number of possible values assigned to each outcome.

16. **Solution: D**

The mean of the sampling distribution is equal to the mean of the population, since n is large**.**

When working with simple random samples (with replacement) of size n from a population, the mean of the sampling distribution of sample means will be equal to the population mean

(μ of x bars=μ), and the standard deviation of the sampling distribution of sample means will be the population standard deviation divided by the square root of the sample size

(σ of x-bars = sigma\root n). This holds true regardless of whether the population distribution is normal distributed or not. In this case, we are given that μ=18.48 and σ=0.573. So, the sampling distribution of sample means for n=9 will have a mean of 18.48 and a standard deviation of 0.5739/sqrt{9}}=0.191.

17.**A, B and C are all true:**

Let's evaluate each statement based on the given information about the population and the sampling distribution:

Given:

- The population mean (μ\muμ) = 23.4 cm
- The population standard deviation (σ\sigmaσ) = 1.7 cm
- The sample size (nnn) = 25
- The sampling distribution of the sample means is Normally distributed (due to the Central Limit Theorem, especially since the population is already Normally distributed).

A. The mean of the sampling distribution is 23.4 cm.

This statement is **true**. The mean of the sampling distribution of the sample means is equal to the population mean, so the mean of the sampling distribution is 23.4 cm.

B. The standard deviation of the sampling distribution is 0.34 cm.

This statement is **true**. The standard deviation of the sampling distribution (also called the **standard error**) is calculated as:1.7/root 5 = 0.34

C. The standardized scores for the sample means will be Normally distributed.

This statement is **true**. Since the sampling distribution of the sample means is Normally distributed (as per the Central Limit Theorem), the standardized scores (also known as **z-scores**) for the sample means will also follow a Normal distribution. Standardizing the sample means (subtracting the mean and dividing by the standard error) will yield a standard Normal distribution.

Conclusion:

The correct answers are: **A, B, and C**.

6. Confidence Intervals and Hypothesis Testing

1. **Solution: C**

The focus here is on the **mean difference** in calorie intake. The data collected involves paired observations: each child was given both types of breakfast (high GI and low GI). For each child, we can compute the difference in calorie intake between the high GI and low GI breakfasts, and then calculate the mean of these differences (referred to as the "mean difference"). This makes the situation a one-sample t-test for the mean difference, where the hypotheses are as follows:

H0: mean difference=0

Ha: mean difference<0

The alternative hypothesis is left-tailed because we are calculating the difference as high GI minus low GI (H - L), and the claim suggests that a high GI breakfast leads to a lower calorie intake at the next meal (lunch) compared to a low GI breakfast. According to this claim, we would expect the caloric intake after the high GI breakfast (H) to be lower than that after the low GI breakfast (L), making the difference (H - L) negative.

This question is asking about the **P-value**, which is the probability of obtaining a sample result as extreme or more extreme (in the direction specified by the alternative hypothesis) if the null hypothesis is true. In the context of this study, the P-value represents:

The probability of obtaining a sample mean difference (H - L) of -0.5 kcal or a lower value, assuming there is no difference in mean calorie intake between the two breakfasts (i.e., the mean difference is zero).

Among the answer options, the one that best matches this definition is: *"It is the probability of taking a sample that has a mean difference in calorie intake of -0.5 kcal or lower, assuming there is really no difference in mean calorie intake."*

2. The answer is C.

The most appropriate statistical test for this situation would be **C. t-test for beta**.

Here's why:

- The insurance company is investigating the relationship between two continuous variables: the **number of firefighters** and the **amount of property damage**.
- This suggests that the goal is to assess how changes in one variable (number of firefighters) affect the other variable (property damage). This involves 2 QUANTITATIVE variables!
- A **t-test for beta** tests the significance of the regression coefficient (beta) in a linear regression model. This type of analysis is used to determine if there is a significant relationship between the two variables (i.e., whether the number of firefighters affects the amount of property damage).

In contrast, the other options are not suitable for this scenario:

- **A. t-test for mu1 – mu2**: This is used for comparing the means of two independent groups, not for assessing the relationship between two continuous variables.
- **B. t-test for mu**: This is used for testing a single mean, not for examining the relationship between two variables.
- **D. One-way ANOVA**: This is used to compare means among three or more groups, which is not relevant here since the data involve two continuous variables, not categorical groups.

Thus, **C. t-test for beta** is the correct choice for analyzing the relationship between the number of firefighters and property damage.

3. **Solution: A** That alternative hypothesis here is: mu=/= $80 000

4. **Solution: A.**

The most appropriate test for this scenario is **A. t-test for difference between means**.

Here's why:

- The executive is interested in comparing the **mean length of long-distance telephone calls** between two **different departments** (shipping and human resources).
- The situation involves two independent groups (the shipping department and the human resources department), and you want to compare the **means** of these two groups.
- The **t-test for difference between means** is specifically designed to compare the means of two independent groups to determine if there is a significant difference between them.

The other options are not appropriate for this situation:

- **B. t-test for mean**: This test is used to compare the sample mean to a known population mean, not for comparing the means of two different groups.
- **C. t-test for slope**: This test is used for testing the slope of a regression line, which is not relevant here, as you're comparing two means, not modeling a relationship.
- **D. large sample test for difference between proportions**: This test is used for comparing proportions (e.g., success rates or proportions in different categories), not means.

Therefore, the correct choice is **A. t-test for difference between means**.

5. **The level of confidence** for my confidence interval would be affected. In constructing a confidence interval**, the critical value** is determined by the chosen level of confidence. This value is then derived from the selected probability model (such as the Normal model in this case). If I select the critical value assuming the Normal model is a good fit, but it actually isn't, then the critical value will be inaccurate. As a result, the actual confidence level of my interval may be higher or lower than what I initially expected, depending on how my assumption about the model is violated.

6. **Part 1:**
There are several ways to answer this question correctly, as long as your response demonstrates an understanding of probability models and their relationship to confidence intervals. A well-constructed answer might be:
Probability models, such as the Normal model, are essential in constructing confidence intervals because they help us determine the "critical value" that defines the confidence level for the interval. These models describe the sampling distribution, and by using them, we can calculate the margin of error needed to achieve the desired confidence level.

Part 2:
There are various accurate ways to answer this question, but it should reflect an understanding of sampling distributions and how they differ from probability models in general. A strong response might be:
Sampling distributions are used in computing confidence intervals to estimate how far, on average, our statistic is from the true parameter we are trying to estimate. This is done by incorporating the standard deviation (or standard error) from the sampling distribution into the margin of error. This allows us to assess the reliability of our estimate based on the variability of the statistic.

7. Answer: C.

The confidence interval tells us the range of values for the standard deviation of blood pressures for all adult males taking Viagra, not just a single individual.

Explanation:

Your friend's explanation is incorrect because they are confusing the interpretation of a confidence interval for a population parameter (standard deviation) with individual values. A 90% confidence interval provides a range of values for the population standard deviation, not for individual blood pressure values. Therefore, the correct interpretation is that the confidence interval estimates the range of values for the population's standard deviation of blood pressures for adult males taking Viagra.

8.

a) there are two comparison groups; they are (1) waxed apples, and (2) unwaxed apples
b) the comparison groups are paired since two apples come from each tree, one that was waxed and one that wasn't waxed, i.e. unwaxed

c) the response variable is the mass (in grams) of the apples and this is a quantitative variable (it is also ratio and continuous)

d) let μ_{wax} be the mass (in grams) of all waxed apples

μ_{unwax} is the mean mass (in grams) of all unwaxed apples.

H0: $\mu_{wax} - \mu_{unwax} = 0$ (this means the means are EQUAL)

Ha: $\mu_{wax} - \mu_{unwax} > 0$ (this means the mean of waxed is GREATER than that of the uwaxed)

The test is one sided or one-tailed, meaning we have a right tailed test here and the p-value would be to shade ABOVE or to the right of our test statistic. Here, it is reasonable to assume that waxy coating makes the fruit weigh more.

9. A large P-value arises when our sample statistic (the sample proportion) is significantly extreme, meaning it is far from the hypothesized null value of 0.5. However, it is also inconsistent with the alternative hypothesis. For instance, if our sample proportion were 0.1, the area under the curve of the sampling distribution to the right of that statistic (since larger values of the proportion would support the alternative hypothesis) would cover almost the entire distribution, resulting in a very large P-value.

10. In this scenario, the sampling distribution of the sample mean (for the number of USB ports in a student computer) will be the foundation for computing the P-value in the hypothesis test. Here's what we know about this sampling distribution:

    1. **Sampling Distribution**: The sampling distribution refers to the distribution of the sample mean values obtained from repeatedly sampling from the population of student computers, each sample consisting of 250 students. Since we are using a simple random sample (with replacement), the sampling distribution will show how the sample mean varies across multiple samples.

    2. **Shape of the Distribution**: If the sample size (n = 250) is sufficiently large, the sampling distribution of the sample mean will approximately follow a Normal distribution, even if the underlying population distribution of the number of USB ports is not Normal. This result comes from the **Central Limit Theorem** (CLT), which tells us that for large sample sizes, the distribution of sample means tends to be Normal, regardless of the original population's distribution.

    3. **Mean of the Sampling Distribution**: The mean of the sampling distribution of the sample means will be equal to the population mean, assuming the null hypothesis is true. In this case, the null hypothesis states that the mean number of USB ports is 3. So, the mean of the sampling distribution will be 3.

    4. **Standard Deviation (Standard Error)**: The standard deviation of the sampling distribution, also known as the **standard error**, will depend on the population standard deviation (σ) and the sample size (n). Specifically, it is calculated as: SE=sigma/root n

Since the sample size is 250, we expect the standard error to be relatively small, leading to a more precise estimate of the population mean. If the population standard deviation is not known, we would use the sample standard deviation (s) as an estimate.

    5. **P-value Calculation**: The P-value for this t-test is calculated from the sampling distribution of the sample mean, and it represents the probability of obtaining a sample mean as extreme or more extreme than the one observed, assuming that the null hypothesis is true. In this case, it’s the probability of obtaining a sample mean number of USB ports that is different from 3, either greater than 3 or less than 3, by the observed amount or more.

In summary, the sampling distribution in this context is approximately Normal due to the large sample size, centered at the hypothesized population mean of 3, with a standard error determined by the sample size. The P-value is derived from this sampling distribution to test if the observed sample mean is significantly different from 3.

11. Solution: B

The population proportion of Grade 12 students who are left-handed in the school board (i.e., p) is 0.15. The sample proportion was the 24 of the 120 students in the sample were left-handed (i.e. p-hat=24/120=0.21).

12. Solution: D

The company is interested in the percentage of its cardholders who possess a specific characteristic (i.e., paid exactly the minimum payment last month), which is a population proportion (p), rather than the mean of a quantitative variable. Consequently, the company is focusing on a population proportion (p) and will use the sample proportion (p-hat) as an estimate of this quantity, based on the data collected from their sample of 100 cardholders.

13. **Solution: B**

The primary purpose of a confidence interval is to **estimate** a population parameter, not just to predict or estimate sample statistics.

- **A. To estimate about a population parameter**: Not exactly, we want to estimate the actual value of the parameter.
- **B. To estimate the value of a population parameter**: This is also correct, though less specific than option A. It essentially reiterates the primary purpose, but option A is a better phrasing because it suggests estimating a range of values.
- **C. To predict the frequency with which a sample value will occur**: This is incorrect. Confidence intervals are not used to predict sample values but rather to estimate the population parameter based on sample data.
- **D. To estimate the value of a sample statistic**: This is incorrect. Confidence intervals are used to estimate population parameters, not sample statistics.

14. Answer: C.

A. Sample estimate

- The **sample estimate** refers to the point estimate, which is 0.150.

15 in this case. This is not the correct description of the value 0.020.

B. Population parameter

- The **population parameter** is the true proportion of cars the United States government stops, which we are trying to estimate. The margin of error does not represent the population parameter but the uncertainty around the sample estimate. Therefore, this is not the correct description.

C. Margin of error

- The **margin of error** is the value added and subtracted from the sample estimate to create the confidence interval. It indicates the range within which we expect the true population parameter to lie, given the sample data. In this case, 0.02 is the margin of error because it describes the uncertainty around the point estimate of 0.150.

D. Sampling error

- **Sampling error** refers to the difference between the sample estimate and the true population parameter. While the margin of error provides an estimate of the range of this error, the sampling error is not exactly the same thing.

Conclusion:

The best description of the value 0.020.020.02 in the confidence interval is the **margin of error**, as it represents the range of uncertainty around the sample estimate.

15. **Answer: C**

The correct interpretation of the 85% confidence interval is:

C. 85% of all samples of 25 players will result in a range of heights that include the mean of all the players.

Here's why:

- **A. There is an 85% chance that the mean height of all players is a value within the range, 76" plus/minus 8"**: This is incorrect. The population mean is a fixed value, not a random variable, so we can't say there's a probability that it falls within the interval. The interval estimates where the population mean is likely to be, but this is not a probability about the mean.
- **B. 85% of all player heights will fall within the range of 76" plus/minus 8"**: This is incorrect because the confidence interval refers to the population mean,

not individual data points. The individual heights of players will not necessarily fall within this range.

- **C. 85% of all samples of 25 players will result in a range of heights that include the mean of all the players**: This is correct. The 85% confidence interval means that if we repeatedly take samples of 25 players, 85% of the confidence intervals we calculate from these samples will contain the true population mean height.

16. Solution: D

- **Option A:** The confidence interval (CI) is designed to estimate the population mean, μ. While it is true that sample means (x-bar) are point estimates of μ and will vary around the true population mean (presumably following a normal distribution, which is a requirement for the CI), we cannot know the exact location of μ or how "extreme" the particular sample mean we used to calculate the CI is. As a result, we cannot determine what percentage of sample means will fall within the interval. We would only be able to say that 90% of sample means fall within the interval if the sample mean were exactly equal to μ, but even then, we're talking about sample means, not individual observations. Since sample means are less spread out than individual data points, even if x-bar, the percentage of individual data points that fall within the interval would be much less than 90%.

- **Option B:** The statement "The mean gestation period is between 257.2 and 274.8 days 90% of the time" is incorrect. The population mean, μ, is a constant value, not a random variable. Therefore, we cannot say that the population mean will be in the confidence interval some of the time (e.g., 90% of the time). It is either in the interval or it is not. The concept of "90% of the time" applies to the process of constructing intervals, not to the mean itself.

- **Option C:** This statement is also incorrect. The purpose of a CI is to estimate the population parameter (in this case, μ). We know that the sample statistic (e.g., x-bar) used to create the interval will always be within the interval because it is the point estimate used for the CI. If we were to repeat the sampling process, 90% of the sample means would not necessarily fall within this specific range due to variability in sample means. Furthermore, if we replaced x-bar with μ, the statement would still be incorrect for the same reason as in Option B: μ is either in the interval or it is not, and there is no probability associated with it.

- **Option D:** This statement is correct. While we can never be certain whether the interval we calculated contains μ, the confidence we have in the CI comes from

the process. Based on a 90% confidence level, we can be 90% confident that our interval contains μ. This means that 90% of the confidence intervals constructed from all possible simple random samples (SRS) of the same size from the population will contain the true population mean μ.

17. **Answer: D**

An interval of values computed from a sample estimate that is based on a method that produces intervals capturing the parameter at a rate of 0.95.

Explanation:

- A is incorrect because the margin of error is not a fixed value like ±0.95; instead, it depends on the standard error and the desired confidence level.
- B is incorrect because a 95% confidence interval does not guarantee that the interval will contain the parameter with a probability of 0.95. It only means that if you repeated the sampling process many times, 95% of the resulting intervals would capture the true parameter.
- C is incorrect because it implies that the specific interval calculated from the sample data is correct 95% of the time, but this is not true. The interval either contains the parameter or it does not.
- D is correct because a 95% confidence interval is based on a method that, when repeated over many samples, will produce intervals that contain the true population parameter 95% of the time.

18. Answer: C

We have a 90% chance of selecting a simple random sample from the population that results in a confidence interval that includes the population median duration.

Explanation:

- **A** is incorrect because the confidence interval applies to the population parameter (the median), not to individual samples. The chance is not about another sample but about the method used to create the interval.
- **B** is incorrect because the confidence interval is about the population parameter (the median) and not the individual responses.
- **D** is incorrect because the confidence interval applies to the median, not the mean. The population mean is not being estimated here; rather, it's the population median.

  C is correct because it accurately reflects the concept of confidence intervals. It tells us that if we repeated the sampling process many times, 90% of the confidence intervals created from those samples would contain the true population median.

19. **Answer: D.**

Explanation:

- **A. Parameter** is incorrect because a parameter refers to a value that describes a population, but 0.15 is based on a sample, not the entire population.
- **B. Critical value** is incorrect because the critical value is a factor used to compute the margin of error, and it is not the value in the confidence interval.
- **C. Sample error** is incorrect because the sample error refers to the difference between the sample statistic and the population parameter, not the value of the statistic itself.
- **D. Statistic** is correct because 0.15 represents the sample proportion of cars stopped for inspection, which is a statistic calculated from the sample data.

  Thus, 0.15 is the **statistic** (the sample estimate) used to calculate the confidence interval.

20. **Answer: D.**

All confidence intervals follow a general format of estimate ± margin of error. The "estimate" refers to a sample statistic derived from a sample of size n. The margin of error (the plus/minus part) reflects the precision of the estimate and depends on the chosen confidence level, which is determined by a critical value, as well as the expected variability among samples, which is influenced by the standard deviation of the estimator.

21. **Answer: A.**

$H_0\ \mu = 19500$ vs $H_a\ \mu < 19500$ (a less than, 1 sided test)

22. **Answer: A.**

23. **Solution: G**

In hypothesis testing, the hypotheses are stated in terms of population parameters, not sample statistics. This means the hypothesis should involve μ, not $\bar{x}$. In this case, the study aims to determine if there is evidence against Ontario's wait time matching BC's (i.e., whether it differs from 6 months). Therefore, the null hypothesis, H_0, is μ = 6 months.

24. **Answer: B**

The model for a t confidence interval for the mean is reasonable.

Explanation:

In this scenario, researchers are dealing with a sample of 42 healthy adults from London and are estimating the mean duration of a typical cold. Since they are estimating the population mean (not a proportion) and the sample size is relatively small (42 individuals), it is appropriate to use a **t-distribution** for the confidence interval. The t-distribution is used when the population standard deviation is unknown, which is typically the case in such situations.

Why not A? Option A refers to the model for a large sample confidence interval for proportions, which is used when you are estimating a population proportion (e.g., the proportion of people who think a cold lasts a specific amount of time). However, the researchers are estimating the mean duration of a cold, not a proportion, so the use of a proportion model is not applicable here.

Why not C? While it's true that the situation described involves a sample of 42 people, this sample size is sufficient to use a t-distribution, as long as the data are approximately normally distributed or the sample size is large enough to apply the central limit theorem. Therefore, the "one-sample" t-confidence interval model is reasonable, making option C incorrect.

25. **Answer: A.**

26. **Answer: A**.

Explanation:

The airline wants to evaluate whether the mean flight price for new customers is more than $350, which involves testing a claim or hypothesis about a population parameter (in this case, the population mean price). The appropriate statistical procedure for this situation is a **hypothesis test**, where you would:

- **Null Hypothesis (H_0)**: The mean price of flights for new customers is equal to $350 ($\mu = 350$).
- **Alternative Hypothesis (H_1)**: The mean price of flights for new customers is more than $350 ($\mu > 350$).

A **confidence interval** (option B) is used to estimate the range of plausible values for a population parameter (e.g., the population mean), but this scenario is focused on testing whether the mean exceeds a specific value, which requires a hypothesis test rather than just estimating the parameter. Therefore, the correct procedure is a hypothesis test.

27. **Answer: A.**

Explanation:

In this scenario, the cosmetics company wants to estimate the percentage of customers who prefer the new make-up over the older products. Since the goal is to **estimate** the population proportion (the percentage of customers who prefer the new make-up), the appropriate procedure is a **confidence interval** for a population proportion.

A **confidence interval** will give a range of values within which the true proportion of customers who prefer the new make-up is likely to fall, based on the sample data.

On the other hand, a **hypothesis test** would be used if the company were testing a specific claim (e.g., whether the proportion of customers who prefer the new make-up is greater than a certain threshold), which is not the case here.

Therefore, the correct answer is a **confidence interval**.

28. **Answer: A.**

Explanation:

In this situation, the City of London is interested in estimating the **average commuting time** for people who work in the city. The goal is to find an estimate for the population mean commuting time based on the sample of 100 people.

Since the objective is to **estimate** the average commuting time, the appropriate procedure is a **confidence interval** for the population mean. This will give a range of values within which the true average commuting time is likely to fall, based on the sample data.

A **hypothesis test** would be used if the goal were to test a specific claim or hypothesis about the population (e.g., whether the average commuting time is greater than a certain value), which is not the case here.

Therefore, the correct answer is a **confidence interval**.

29. Answer: D.

30.Answer: B.

We're looking at the number of cats who turned over the vase and whether or not each cat can turn over the vase ∴ proportions and only one sample

31. Answer: D.

A. is 2 quantitative variables ∴regression would be best

B. $\mu\bar{x} = \mu$

C. Use a 2- sample t-test (independent)

D. is correct, we would use a confidence interval to estimate the first quartile for the distribution of starting salaries for graduates of Ivey

32. Answer: D.

$H_o\ p = 0.05$

$H_a\ p < 0.05 \quad (one\ sided\ less\ than\ test)$

33. Solution: B

The standard error of the sample mean is equal to SE=s/square root, and this is an estimate of the standard deviation of the sampling distribution of X-bar based on results from ONE sample.

From the output, we can see that n=length=10 and s=sd (times)=8.2792.

Therefore, SE x-bar= 8.2792/square root (10) = 2.6181

34. **Solution: D**

In a hypothesis test, the hypotheses are stated in terms of population parameters, not statistics. This means the hypothesis would reference the population proportion p, not the sample proportion p^. In this case, the study is designed to determine whether there is evidence to suggest that Ontario's proportion of new drivers passing their driving tests on the first attempt differs from BC's (i.e., whether it differs from 0.20). Therefore, the alternative hypothesis is p≠0.20.

35. **Solution: A**.

We would say that we do have enough evidence to reject H0 and conclude that there is evidence against our null hypothesis and that the results are too extreme to be solely due to random chance alone.

36. Answer: B.

$H_0\ \mu = 27$ vs $H_a\ \mu > 27$ (greater than, 1 sided test)

See the second box on the right says t.test for true mean greater than 17, so the t test=2.6364 and the p-value is 0.005628

37. Answer: B.

In this scenario, the goal is to investigate whether the typical difference between shoe and foot sizes for students exceeds 0.5 UK Adult Shoe sizes. Let's break down the key points and determine which inference procedure is most appropriate:

Key Points:

- The study involves a **sample of 12 students**.
- The **mean difference** between shoe and foot sizes is 0.75 UK Adult Shoe Size units.
- The **sample size is small (n = 12)**.
- We're interested in whether the **typical difference exceeds 0.5**.

Inference Procedure Considerations:

1. **Large Sample Confidence Interval for p** (Option A):

- This is used for estimating a population proportion, which is not applicable here since we are dealing with **differences in shoe and foot sizes** (a continuous variable, not a proportion).

2. **t-test for mu** (Option B):
   - This procedure is used for testing hypotheses about the **mean** of a population when the sample size is small and the data are approximately normally distributed.
   - The scenario is testing whether the **mean difference between shoe and foot sizes exceeds 0.5**. This suggests a one-sample t-test for the mean of the differences.
   - The **null hypothesis** would be that the mean difference is equal to 0.5, and the **alternative hypothesis** would be that the mean difference is greater than 0.5.
3. **Large Sample Test for p** (Option C):
   - This is used for testing a population proportion, similar to option A, so it is not relevant to this case.
4. **None of the Above** (Option D):
   - This option suggests that none of the procedures are appropriate. However, as explained above, a **t-test for the mean** (option B) is a valid approach for this situation.

Conclusion:

The most appropriate inference procedure for this scenario is the **t-test for the mean** (Option B), as we are testing if the mean difference between shoe size and foot size exceeds 0.5, and the sample size is small.

38. Answer: B.

- **Mean of the sampling distribution**: The mean of the sample means will be the same as the population mean, 4.75 hours.
- **Shape of the sampling distribution**: Due to the application of the CLT, the sampling distribution of the sample mean for 25 individuals will be **approximately normal**, even though the individual data are right-skewed.

- **Standard deviation of the sampling distribution**: The standard deviation of the sample means (i.e., the standard error) is 0.12 hours.

Thus, the correct answer is:

B. Approximately Normal with mean 4.75 hours and standard deviation 0.12 hours.

39. **Answer: B.**

To address this question, let's first clarify the distinction between a **one-tailed** and **two-tailed** hypothesis test.

- A **two-tailed test** is used when we want to test for any difference (either positive or negative) from the hypothesized value. The alternative hypothesis is typically that the true parameter is **not equal** to the hypothesized value.
- A **one-tailed test** is used when we are only interested in detecting a difference in one specific direction (either greater than or less than the hypothesized value). The alternative hypothesis is typically that the true parameter is **greater than** or **less than** the hypothesized value, but not both.

Impact of conducting a one-tailed test instead of a two-tailed test:

1. **Magnitude of the test statistic**:
    - The magnitude of the **test statistic** (e.g., t-statistic) would not change simply because it's a one-tailed or two-tailed test. The statistic is calculated based on the difference between the observed data and the hypothesized value, and this calculation remains the same regardless of the direction of the test.
2. **P-value**:
    - The **P-value** is the probability of observing a test statistic as extreme as, or more extreme than, the one obtained in the sample data under the null hypothesis.
    - In a **two-tailed test**, the P-value considers both tails (both the positive and negative extremes). In a **one-tailed test**, the P-value only considers one tail (either the positive or negative extreme, depending on the direction of the test).

- Since a **one-tailed test** considers only one direction, it would generally produce a **smaller P-value** than a two-tailed test, even if the difference is in the other direction.

3. **Significance level**:
    - The **significance level (α)** is predetermined and reflects the probability of rejecting the null hypothesis when it is actually true. It is usually set to a value like 0.05.
    - The significance level is not directly affected by whether the test is one-tailed or two-tailed, though the way the P-value is compared to the significance level differs between the two types of tests.
    - **For a two-tailed test**, the critical region is split between both tails (α/2 in each tail). **For a one-tailed test**, the entire α is in one tail. However, this doesn't affect the magnitude of the significance level itself; it affects how the P-value is compared to it.

Conclusion:

The most significant impact of conducting a one-tailed test instead of a two-tailed test would be on the **P-value**. Specifically, the P-value would be **smaller** than it should have been for a two-tailed test, potentially leading to a mistaken conclusion of statistical significance.

Thus, the correct answer is:

B. The P-value would be smaller than it should have been.

40. **Answer: B.**

The researchers are interested in estimating the **mean number of tattoos** among adults who reported experiencing an **adverse reaction** to their tattoos. Let's break down the options and identify the most appropriate statistical procedure for this scenario.

Key Points:

- The research question involves estimating the **mean number of tattoos** for adults who reported having an **adverse reaction** to their tattoos.
- The sample consists of **300 American adults** who have at least one tattoo, and the focus is on those who had an adverse reaction.
- The question involves **estimating** the mean, which suggests we are looking for a procedure to **estimate a population mean**.

Let's examine each option:

A. Large sample confidence interval for p

- This procedure is used to estimate the population **proportion** (p), not the mean. It is typically used for binary outcomes (yes/no, success/failure).
- Since the researchers are interested in the **mean number of tattoos**, this procedure is not appropriate.

B. t confidence interval for mu

- This procedure is used to estimate the **population mean** (mu) when the sample comes from a normally distributed population or when the sample size is large enough (like 300) for the Central Limit Theorem to apply.
- Since the researchers are looking to **estimate the mean number of tattoos** among adults who reported adverse reactions, this is the correct choice.

C. Large sample test for p

- This procedure is used for **hypothesis testing** concerning population proportions, not means. The researchers are not testing a hypothesis about proportions but rather estimating a mean.
- This option is not appropriate.

D. t-test for mu

- A **t-test for mu** is used to **test hypotheses** about a population mean (e.g., testing if the mean number of tattoos is a certain value). However, the researchers are not testing a hypothesis but are instead estimating the mean.
- Therefore, this procedure is not the best choice.

Conclusion:

The most appropriate procedure to **estimate the mean number of tattoos** for adults who reported having an adverse reaction is the **t confidence interval for mu** (option B).

Thus, the correct answer is: **B. t confidence interval for mu.**

7. ANOVA and Regression

1. Answer: D.

- The research involves **four independent groups** (dog owners, cat owners, other pet owners, and no pet owners).
- The variable of interest, **quality of life**, is continuous and measured on a numerical scale.
- The goal is to test whether there are differences in **mean quality of life scores** among the four groups.

Now, let's evaluate the options:

A. Conduct two-sample hypothesis tests to see if any of the groups of seniors are different from each other.

- Two-sample hypothesis tests are used for comparing the means of **two** independent groups. Since there are four groups, this approach would require multiple pairwise comparisons (e.g., comparing dog owners vs. cat owners, dog owners vs. no pets, etc.), which can increase the risk of Type I error (false positives).
- This method is not suitable for comparing more than two groups.

B. Combine all pet owners into one group and test whether this group has different quality of life scores from the no-pet group.

- While this simplifies the analysis by reducing the number of groups, it ignores the **potential differences between the types of pets** (dog, cat, and other pets). The researcher is interested in comparing **four distinct groups**, so combining them into one "pet owner" group would not answer the research question about whether the **type of pet** matters.
- This approach loses important information and is not ideal for the study's goals.

C. Compare the sample means of each of the four groups to see if there are any differences in quality of life scores.

- While this option correctly suggests comparing the sample means of the four groups, it does not specify the correct statistical test for comparing more than two groups.
- The correct procedure for comparing the means of **four independent groups** is **ANOVA** (Analysis of Variance), not a simple comparison of means.

D. Conduct an ANOVA analysis to test whether any of the four groups have mean quality of life scores different from the others.

- **ANOVA** (Analysis of Variance) is specifically designed to test for differences in **mean scores** across **three or more independent groups**.
- Since the study involves **four groups**, **ANOVA** is the most appropriate statistical procedure to test whether there are significant differences in the mean quality of life scores among the four groups.

Conclusion:

The most appropriate statistical procedure for comparing the means of the four independent groups is **ANOVA** (Analysis of Variance).

Thus, the correct answer is:

D. Conduct an ANOVA analysis to test whether any of the four groups have mean quality of life scores different from the others.

2.**Answer: D.**

Given the regression analysis output, the **slope** is 9.231 and the **y-intercept** is 114.716. The **p-values** are both 0, indicating that the slope and intercept are statistically significant.

In a simple linear regression equation, the general form is:

$$\text{Mortality} = \beta_0 + \beta_1 x$$

where: β_0= y-intercept of 114.716 and β_1=slope of 9.231

- β_0 is the **y-intercept** (114.716), which represents the predicted mortality when exposure is zero.
- β_1 is the **slope** (9.231), which indicates the change in mortality for a one-unit increase in exposure.

Let's analyze each statement:

A. If the exposure is decreased by one unit, the mortality is expected to increase by 9.231 deaths per 100,000.

- This statement is **incorrect**. The **slope of 9.231** means that **for each increase in exposure by one unit**, the mortality is expected to **increase** by 9.231 deaths per 100,000. A **decrease** in exposure would lead to a **decrease** in mortality, not an increase.

B. The predicted mortality is 92.31 deaths per 100,000 when the exposure score is 10 units.

- This statement is **incorrect**. To find the predicted mortality for an exposure of 10 units, we substitute into the regression equation:

$$\text{Mortality} = 114.716 + (9.231 \times 10) = 114.716 + 92.31 = 207.026$$

The predicted mortality would be **207.026 deaths per 100,000**, not 92.31.

C. If mortality increases by 1 death per 100,000, the exposure is expected to increase by 9.231 units.

- This statement is **incorrect**. The slope of the regression line describes the **change in mortality** for a given change in exposure, not the other way around. The slope tells us how **mortality** changes with **changes in exposure**, but it does not describe how exposure changes with a given change in mortality. The relationship is not reciprocal in this way.

D. The estimated mortality is about 114.7 deaths per 100,000 when there is no exposure.

- This statement is **correct**. The **y-intercept** of the regression equation represents the predicted mortality when the exposure is zero. In this case, the predicted mortality when there is no exposure (exposure = 0) is:

Mortality=114.716

So, when there is no exposure, the estimated mortality is **114.7 deaths per 100,000**, which matches the given value for the y-intercept.

3. **Answer: D.**

A variable refers to a characteristic of a unit (for example, a characteristic of a home that is for sale in this case). The response variable is the main outcome of interest in the study, which in this situation is the number of days it takes to sell a home, as this is the aspect the researcher wants to predict. This is the variable believed to be influenced by changes in the price of a home (i.e., by changes in the explanatory variable).

4. Answer: B.

The strength of the linear relationship between lab and exam scores is strong since the points are all close to a straight line.

5. Answer: D.

On this graph, the line is going up and to the right, so there is a positive relationship, but it is fairly weak, i.e. the points are NOT all very close the line, so it can't be as high as 0.8, which is close to 1, indicating a strong, positive, linear relationship. Therefore, the correlation coefficient must be closer to 0, such as 0.3.

6. Answer: B.

The correlation coefficient $r=-0.8$ indicates a strong negative relationship between the maximum daily outside temperature and the number of classroom disruptions. This means that as the temperature increases, the number of disruptions tends to decrease, or vice versa.

However, **correlation does not imply causation**. Just because there is a negative correlation between these two variables, it does not mean that an increase in temperature directly causes a decrease in disruptions. The correlation shows an association, but it doesn't establish a cause-and-effect relationship. There could be other factors at play, or the observed pattern could be due to chance.

Conclusion:

The statement that "an increase in the maximum daily outside temperature decreases the number of classroom disruptions by students" is **not conclusively supported by the correlation alone**, as causation cannot be inferred from correlation.

7. Answer: C.

In this research scenario, the goal is to explore the relationship between **obesity (measured by BMI)** and **serum estradiol levels** in premenopausal women. Specifically, the researchers want to understand whether BMI, as a continuous variable, is related to the concentration of serum estradiol.

Let's break down the options:

A. Two-sample (matched pairs) test for mean difference

- A **matched pairs test** is used when you have two measurements taken from the same subjects or units (e.g., before and after treatment, or two variables measured on the same individual). This is not applicable here, because BMI and serum estradiol levels are two separate measurements for each individual (BMI and estradiol are not paired in the sense of pre- and post-measurements).
- This option is not appropriate.

B. Two-sample (independent samples) test for difference in means

- A **two-sample independent t-test** is used to compare the means of two independent groups (e.g., comparing two different groups of people, such as smokers vs. non-smokers). In this case, however, the study does not involve comparing two groups but rather examining the relationship between **two continuous variables** (BMI and estradiol).
- This option is not appropriate.

C. Test for linearity (i.e., hypothesis test for slope of regression line)

- A **regression analysis** is appropriate when examining the relationship between two continuous variables (in this case, BMI and serum estradiol levels). Specifically, a hypothesis test for the **slope of the regression line** would assess whether there is a linear relationship between BMI and serum estradiol levels. This test would allow the researchers to determine whether an increase in BMI is associated with an increase (or decrease) in estradiol levels.
- This is the most appropriate option for examining the relationship between the two continuous variables.

D. ANOVA for difference in means

- **ANOVA** (Analysis of Variance) is typically used to compare the means of **more than two groups**. This scenario only involves two continuous variables and does not involve comparing multiple groups, so ANOVA is not appropriate in this case.
- This option is not suitable.

Conclusion:

The most appropriate hypothesis test to assess the relationship between BMI and serum estradiol levels is a **regression analysis** to test for the linearity (slope) of the relationship between the two continuous variables.

Thus, the correct answer is:

C. Test for linearity (i.e. hypothesis test for slope of regression line).

8. Answer: 6.

There are FOUR (4) treatments (i.e. environments) in this scenario:

With four treatments, there are a total of 6 pairwise comparisons possible

If we let n=6, we have n(n-1)/2 = 4(3)/2 = 6 treatments or pairwise comparisons of means

They would be 1&2, 1&3, 1&4, 2&3, 2&4, 3&4

9. Answer: A.

The Type I error rate is influenced by multiple comparisons. When performing individual two-sample t-tests at alpha = 0.05, the overall Type I error rate is not controlled. This means that even if there are no actual differences between the means, there is a greater than 5% chance of incorrectly rejecting at least one null hypothesis (H0) due to random chance. In other words, you are more likely to detect differences between means that do not actually exist more than 5% of the time.

10. **Answer: A**. True.

After rejecting the null hypothesis in ANOVA, we are saying there is enough evidence to support that there is at least one mean that is different than the others and then we would use Tukey's HSD post hoc test to see which means are different and that uses a 2-sample t test.

11. **Answer: D**.

In this scenario, the teacher is interested in testing whether the **extent of previous practice** (with four distinct levels) influences **math anxiety levels** (measured on a continuous scale). The appropriate statistical test depends on the number of groups and the type of data.

Breakdown of options:

A. t-test for difference in means

- A **t-test for difference in means** is used to compare the means of **two groups**. Since there are **four** groups in this case (the four levels of previous practice), this test is not appropriate. This test is designed for comparing two groups, not multiple groups.
- **Not suitable**.

B. Large sample test for difference in proportions

- A **test for proportions** is used when comparing proportions or percentages across different groups. Since **math anxiety levels are measured on a continuous scale**, this test is not applicable.
- **Not suitable**.

C. t-test for slope

- A **t-test for slope** is used in the context of regression analysis to test whether the slope of the regression line is significantly different from zero. However, this is not the situation here, as the teacher is comparing groups, not looking at a relationship between two continuous variables.
- **Not suitable**.

D. One-factor ANOVA

- **One-factor ANOVA** (Analysis of Variance) is used to compare the means of **three or more independent groups**. Since the teacher is comparing four groups (levels of practice), this is the correct test for comparing the means of these groups.
- **Appropriate choice**.

E. None of the above options

- Since **one-factor ANOVA** is appropriate, this option is not correct.
- **Not needed**.

12. **Answer: C**.

Although it's possible that the population distributions are not Normal or that the variances are not constant, the primary issue here— and a significant problem—is that the samples are not independent. Since the comparison is being made across family members, the data structure involves a matching element.

13. Answer: B.

In order to assess the conditions for conducting a One-factor ANOVA, it is important to check whether the data meets certain assumptions, including normality of the residuals and homogeneity of variance (constant variance across groups). Let's evaluate the listed outputs in the context of these assumptions:

A. Normal Quantile (QQ) plots

- **QQ plots** are commonly used to check the assumption of **normality**. They display how closely the data follows a normal distribution. If the data points lie along a straight line, the data is approximately normally distributed. This output would definitely be used to assess ANOVA conditions.
- **Appropriate for checking assumptions.**

B. ANOVA table

- The **ANOVA table** provides the results of the One-factor ANOVA test, including the F-statistic, p-value, and degrees of freedom. It is used to test whether there are significant differences between group means but does not help directly with checking the conditions for performing ANOVA. An ANOVA table would be done AFTER you have checked your assumptions and you decided to proceed with ANOVA.
- **Does not evaluate assumptions.**

C. Stripchart

- A **stripchart** is a graphical display that can show the distribution of data points within each group. It can help visualize whether the data is spread evenly and may also provide insight into **variance**. It can help assess the assumption of **homogeneity of variance**.
- **Appropriate for checking assumptions.**

D. Histograms

- **Histograms** display the distribution of data and can help assess the **normality** assumption. By plotting the data for each group, one can visually check whether the data is approximately normal.
- **Appropriate for checking assumptions.**

14. **Answer: A, B, and C are all correct**. Graph 4 should have residuals on the y-axis.

15. **Answer: D**.

$H_o \ \beta = 0$
$H_a \ \beta \neq 0$
$p - value \ is \ at \ bottom \ of \ table$
$p - value = 0.03349 < 0.10 = \alpha(given \ in \ question)$
$\therefore \ strong \ evidence \ against \ H_0$ and we could conclude that the slope is not equal to 0 and that there is evidence of a straight-line (linear) relationship.
$\therefore \ slope \neq 0$ There is a linear relationship.

We can never say there is "definitely" a linear relationship, as we are using samples, so no answer is 100% accurate.

16. Answer: A.

$$sd = \sqrt{MSE} = \sqrt{206} = 14.35$$

17. **Answer: B.**

We are dealing with a sample proportion of 60 out of 100 people, so our analysis must involve a proportion and not mean(s). There is only one sample, so it can't be D.

18. Answer: D

19.Solution: B

Here, the explanatory variable is temperature and the response is rate constant.

We are given the regression equation:

(rate constant) = 3.2 + 0.051 (temperature)

We can compute the predicted value based on Jill's temperature of 20:

predicted rate = 3.2 + 0.051(20) = 4.22

Then, the residual is:

residual = observed y - -predicted y

=3.51 - 4.22

= -0.71

Recall, the residual is negative any time the actual data lies below the line

20.Answer: A, B, C, D

In this linear relationship, height is the response variable and 'age' is the explanatory variable. The slope is 6.01 and the y-intercept is 50.3.

The y-intercept tells us the height when the age is 0.

The slope tells us the rate of change, i.e. the increase in height for each increase of age in years.

So, if age is increased by 1, the height would increase by 6.01 cm. So, A is true.

If y=50.3 + 6.01(5)=80.35

And if y=50.3 +6.01(18)=158.48, so yes, the height is about half at age 5 than age 18 years old, so C is true.

If age=10, we get y=50.3 + 6.01(10)=110.4 cm, so B is true.

If age=8, we get y=50.3 + 6.01(8)=98.38, so if her niece is 115cm tall, she is likely in a high percentile for her height as that is much higher than the estimate of 98.38 cm for an average 8 year old.

21.Answer Key: C

Since we want to predict fourth finger length based on second finger length, the x, or explanatory variable is 2nd finger length and the y, or response variable is the 4th finger length.

Find the equation:

Fourth finger length = 1.1328 + 0.8679 (2nd)

=1.1328 + 0.8679 (7.25)

=7.425

A residual= observed y – predicted y=7.5 – 7.425 =0.075,

and the closest answer is C.

Best of luck on the exam!!