**Section A: Multiple Choice Questions (5 questions, 1 point each)**

**Question 1.**

The mechanism by which people view TV shows and movies has changed dramatically over the last few years. Rather than purchasing a cable package with a series of channels, viewers purchase subscriptions to ‘video on demand’ companies like “Crave TV”, or “Netflix”, etc. Each such company offers a different selection of shows and movies for viewing. It’s unclear whether the demographics of viewers who subscribe to such companies differ in some systematic way; one obvious characteristic that may vary across company subscribers is age.

A polling company conducted a representative online survey of **890** North American TV viewers to obtain data on the association between age and company. The data from their survey are summarized in the table below. Based on these data, what percentage of Crave TV viewers are 26 to 65 years old?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Company** | | | |
|  |  | *Amazon Prime* | *Crave TV* | *Disney+* | *Netflix* |
| **Age group** | 25 y and younger | 70 | 85 | 40 | 35 |
| 26 to 40 y | 55 | 95 | 75 | 80 |
| 41 to 65 y | 50 | 25 | 60 | 110 |
| 66 y and older | 35 | 10 | 25 | 40 |

|  |  |  |
| --- | --- | --- |
|  | *Input an uppercase/capital letter* ***X*** *into the box next to the option you wish to select as your answer to this question. Leave all other boxes empty for this question.* | |
|  | A. | 21.8% |
|  | B. | 23.4% |
| x | C. | 55.8% |
|  | D. | 59.5% |

**Question 2.**

A consumers’ report service is evaluating the absorbance efficiency of paper towel brands (e.g. Brawny, Bounty, etc.) for a special report to guide consumer purchasing. They know that brands vary in the shape of each sheet (i.e. rectangular vs. square) and there are far too many brands than can be included in their study. So, they select a package of paper towels from three (**3**) different brands that have square sheets, and from three (**3**) different brands that have rectangular sheets.

For a single sheet of paper towel from each of the selected brands, they record how much of a standardized amount of water is absorbed by the sheet. The researchers notice, however, that the various brands differ in area (in cm2) of each sheet. So, the researchers also calculate the area of a sheet from each of their selected brands. They observe that—once they have taken into account differences in area of the sheets—there are no real differences in absorbency across the different brands; consequently, they encourage consumers to use price and personal preference as a guide to purchasing paper towel.

What type of variable is area of the sheets?

|  |  |  |
| --- | --- | --- |
|  | *Input an uppercase/capital letter* ***X*** *into the box next to the option you wish to select as your answer to this question. Leave all other boxes empty for this question.* | |
| x | A. | Blocking variable |
|  | B. | Stratifying variable |
|  | C. | Explanatory variable (specifically, the factor of interest) |
|  | D. | Cofactor |
|  | E. | Response variable |

**Question 3.**

A major contracting company is trying to get a sense of the preferred brick colour for houses in London, Ontario, to inform their future housing development projects. They have already surveyed people buying houses to get a sense of what buyers might prefer, but they suspect that some of the observed preferences may reflect how common brick colours already are (i.e. buyers may prefer colours that seem unique). Consequently, the company would like to survey the distribution of brick colours in houses that currently exist in London.

The company obtained a map of London and subdivided the city into five (**5**) geographical quadrants (i.e. Northwest, Northeast, Central, Southwest, Southeast). For each quadrant, they identify all the houses by address. They then selected fifty (**50**) houses from each quadrant to visit and record the colour of the bricks on the house.

What type of sample was obtained as a result of this design?

|  |  |  |
| --- | --- | --- |
|  | *Input an uppercase/capital letter* ***X*** *into the box next to the option you wish to select as your answer to this question. Leave all other boxes empty for this question.* | |
|  | A. | A stratified sample |
| x | B. | A cluster sample |
|  | C. | A multistage sample |
|  | D. | A systematic sample |

**Question 4.**

The Women’s Network plays a lot of made-for-TV romance movies; many of these movies get replayed from week to week. The Network has noticed that the level of viewership for individual movies changes from week to week in an apparently unpredictable manner. They wonder whether the variation in viewership can be predicted in some manner. They have designed one study to evaluate whether the ‘quality’ of the movie (i.e. measured by whether it would be watched again by a viewer) is a determining factor.

The researchers obtained a sample of typical viewers of the Women’s Network. They have asked the viewers to come to a theatre where they will watch two movies: (i) “Love in Rome”, and (ii) “Love at Sunset Terrace”. The researchers have randomly assigned each viewer to one of two groups; group **A** will watch Love in Rome first, then will watch Love at Sunset Terrace. Group **B** will watch Love at Sunset Terrace first, then Love in Rome. After each movie, the viewers are asked to indicate the probability (from 0 to 1) that they would watch the movie again if it were available. Interestingly, they observed that the mean probability for Love in Rome was **0.50**, while that for Love at Sunset Terrace was **0.55**.

The researchers conducted a hypothesis test to evaluate whether the two movies differed in the probabilities that viewers would watch them again (assume that the relevant model conditions have been met). Which of the following R outputs would be appropriate for this hypothesis test? The data collected are stored in a single datafile that has a column for the viewer identification (i.e. the participant ID), a column called *Rome\_prob* that has the viewers’ probabilities for Love in Rome, and a column called *Sunset\_prob* for the viewers’ probabilities for Love at Sunset Terrace*.*

**The answer options are on the next page.**

***Question 4, continued.***

|  |  |  |
| --- | --- | --- |
|  | *Input an uppercase/capital letter* ***X*** *into the box next to the option you wish to select as your answer to this question. Leave all other boxes empty for this question.* | |
|  | A. |  |
|  |
|  | B. |  |
|  |
|  | C. |  |
|  |
|  | D. |  |
|  |

**Question 5.**

On Twitter (a social media platform), there is a friendly debate over what the ‘best’ programming language is for data analysis. Two of the dominant languages that battle back and forth for ease of use are Python and R. To get to the bottom of which of these two languages have a higher ranking for ease of use, a Twitter user created a survey that collected data from post-secondary school students who had taken one course in either Python, or, R (but not both). The survey asked the students to rank the programming language they learned in terms of ease of use on a Likert-style scale from **1** (extremely difficult) to **5** (extremely easy); a value of **3** was interpreted as neutral.

The survey creator separated the data into a sample from students who learned Python, and a sample who learned R. She computed the proportion from each sample who ranked the programming language as 4 or 5 (i.e. who found the program easy to some degree), and conducted a hypothesis test for the claim that there is no difference in proportion of students who find the programs they learned easy to learn when comparing Python students to R students. She obtained a P-value for the hypothesis test of **0.073** (assume that the model for the hypothesis test used was valid). On Twitter, she made the following statement about her results, “*The probability that there is no difference between the proportion of students who think Python vs. R are easy to use is 0.073.*”

As a knowledgeable student of statistics, which of the following statements about the Twitter user’s statement of her results is correct?

|  |  |  |
| --- | --- | --- |
|  | *Input an uppercase/capital letter* ***X*** *into the box next to the option you wish to select as your answer to this question. Leave all other boxes empty for this question.* | |
| x | A. | The Twitter user’s statement is incorrect: the hypothesis test results provide a probability of getting an unusual sample statistic if there is no difference in the populations, not that there is no difference. |
|  | B. | The Twitter user’s statement is incorrect: the hypothesis test results provide a probability that the null hypothesis is incorrect, not that the null hypothesis is correct. |
|  | C. | The Twitter user’s statement is incorrect: the hypothesis test results provide a probability that the sample statistic is equal to the population parameter, not that there is no difference. |
|  | D. | The Twitter user’s statement is incorrect: the hypothesis test results provide a probability that the alternative hypothesis is correct, not that the null hypothesis is correct. |

**Section B: Short Answer (4 questions, 27 points total for this section)**

**Question 6.**

|  |
| --- |
| Recommendations that young children limit the amount of “screentime” (i.e. time spent watching TV/movies or playing on tablets) are plentiful on the internet; however, there is little solid research on what impact screentime has on children. It is hypothesized that high levels of screentime in children has a negative impact on attention span, but again, there is little research to support this claim. A research team has decided to investigate this hypothesis (they already have ethics approval for their research and willing participants). As they develop their research plan, they recognize that the *content* of what children are viewing may be an important factor; specifically, they are concerned that educational shows (i.e. animal documentaries, etc.) might influence children differently than purely recreational shows (i.e. cartoons, etc.). |

Briefly describe a **study design** that can be used to address the research goal and concerns described above. In your description, be sure to properly identify/label aspects of the study design with study design vocabulary taught in the course (this could be achieved by using vocabulary directly in sentences, or by identifying vocabulary terms by putting them in parentheses where they apply). *(4 marks, suggested length is a paragraph, e.g. 5-7 typical length sentences).*

Note: the purpose of this question is for you to demonstrate that you can select and accurately describe a study design to meet a particular purpose AND show accurate application of relevant statistical vocabulary.

You have the rest of this page plus the entirety of the next page for your answer (mainly to accommodate handwritten answers). Keep in mind the suggested length above!

**Question 7.**

|  |
| --- |
| Over the last 20 years, the number of students who hold a job while attending university fulltime has increased. Work responsibilities may ‘compete’ for time and energy with course responsibilities, and consequently, may affect student academic success. An educational researcher is interested in determining whether student employment influences academic success. The research has obtained a relevant sample of university students, and has determined the following information for *each* student:   1. their employment status (specifically, whether they are employed or not during the most recent school term) 2. their course grades for the most recent school term (specifically, whether they obtained a mean grade of 80% or higher, or not). |

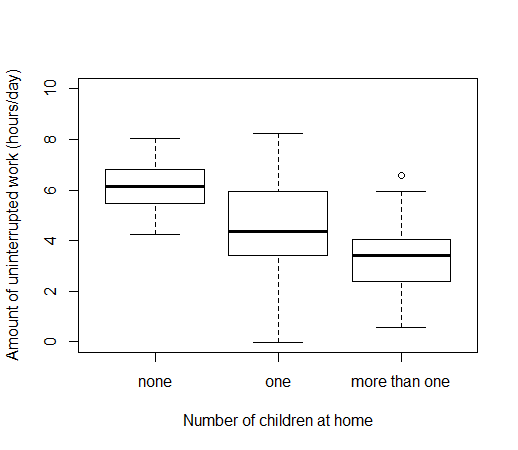
In our lecture on *Planning Ahead: Sampling variability*, you were introduced to a set of five (5) questions that can be used to help decide upon a relevant statistical inference procedure.

* + - 1. **Identify** each of the five (5) questions and **answer** each question based on the scenario described above. *(5 marks, suggested length is 10 short sentences, i.e. 5 questions plus 5 answers).*
      2. Based on your answers to *part a*, identify which of the inference procedures covered in 2244 is most appropriate to analyse the researcher’s data? *(1 mark, suggested length is 1 sentence)*
      3. Assume the researchers for the above scenario have obtained relevant simple random sample(s). **Identify** (by name) the sampling distribution that underlies the statistical inference procedure named in part b. Then, **give the formulae** for the mean and standard deviation for the sampling distribution you have identified. *(5 marks, suggested length is 2-4 sentences [including formulae])*.

Note: you have the rest of this page plus the entirety of the next page for your answer (mainly to accommodate space for handwritten formula, if necessary).

**Question 8.**

Consider the following graph, which summarizes a researcher’s data from a sample of adults who are work from home.



1. Write a research question that could be answered based on this graph. *(~3 marks, suggested length is one sentence).*
2. Use this graph—as you might during the Data phase of the PPDAC framework—to describe two (**2**) distinct patterns, or interesting characteristics about these sample data/or and distributions. *(4 marks, suggested length of 2-4 sentences total).*

Note: The purpose of this question is to demonstrate that you understand and can interpret this graph.

Note: you have the rest of this page plus the entirety of the next page for your answer (mainly to accommodate space for handwritten answers).

**Question 9.**

|  |
| --- |
| A gardener thinks that his plants are being affected by how closely together he places the plants in his garden. Specifically, he thinks that the height that his plants achieve by the end of the growing season differs according to how much space he allocates between each plant. Consequently, he wishes to estimate how different the plant heights (in centimetres) are between rose plants that are planted either **6** inches apart, or **12** inches apart. He has a large garden, so has subsequently planted ten (**10**) rose plants **6** inches apart, and another ten (**10**) rose plants **12** inches apart. |

1. **Identify** (i.e. name) which of the inference procedures covered in 2244 is most appropriate to analyse this researcher’s data? *(1 mark, suggested length is 1 sentence)*
2. Each inference procedure has a set of model conditions that must be evaluated prior to using the procedure. For the inference procedure you have identified in part a, **list** each required condition using the context of the research scenario described above (that is, the variables/language of this specific scenario). *(~4 marks, suggested length is 1 sentence per condition)*

Note: This question is NOT asking you to *evaluate/justify* the validity of model conditions for an inference procedure. You are simply being asked to demonstrate that you can correctly “translate” general conditions into the context of a specific example.

Note: you have the rest of this page plus the entirety of the next page for your answer (mainly to accommodate space for handwritten answers).

**Section C: Data Analysis Questions (4 questions, 29 points total for this section)**

All questions in this section are based on the information provided below (a brief description of the research data) and the dataset.csv file that was provided to you with this take-home exam file. The ***Research Question*** you will be addressing for this data analysis is also given below. For some questions, you will need to use R (and/or R Studio)—and illustrate its use—by reporting your code plus output to answer questions that indicate to do so. Use the same Reminders/Tips for Success for reporting R code/output and answering these questions as you were provided for Lab Assignments.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The Olympic Summer Games is an international multi-sport event that has been held every four years since 1896. Competitors from countries across the world compete against one another, representing their countries, in one or more sports events such as wrestling, rowing, high jump, running races, etc. Events may be separated by the sex of the competitors (male or female) or may be mixed (i.e. males and females competing with each other in the same event). At the end of each event, the top three competitors are awarded a medal: bronze for third place, silver for second place, and gold for first place.  The dataset.csv file you will be working with contains results from the Olympic Summer Games for a subset of the years that these Games have run. There are 9 columns (i.e. vectors) in the dataset, and 1131 rows (i.e. observations). There are NO missing values in the dataset.csv file. A description of the columns is as follows:   |  |  |  | | --- | --- | --- | | ***Name of vector*** | ***Description of data*** | ***Example values*** | | sex | Character vector that gives the sex (M=male or F=female) of the athlete | M  F | | age | Integer vector the gives the age (in years) of the athlete | 28 | | height | Integer vector that gives the height (in centimetres) of the athlete | 175 | | weight | Numeric vector that gives the weight (in kilograms) of the athlete | 64 | | event | Character vector that identifies the event in which the athlete participated | Athletics Men’s Long Jump | | year | Integer vector that identifies the year in which the athlete participated in the Games | 1992 | | sport | Character vector the identifies the generic category of sport the athlete participated in | Basketball  Athletics | | season | Character vector that identifies the season during which the Games occurred; all values are Summer | Summer | | medal | Character vector that identifies the type of medal (bronze, silver, or gold) won by the athlete | Gold  Silver  Bronze |     **The Research Question:** Does age differ across medal type for male competitors of the Summer Olympic Games? |

**Question 10.**

a) Based on the information provided, which of the following inference procedures is most appropriate to address the Research Question? *(1 mark)*

|  |  |  |
| --- | --- | --- |
|  | *Type the uppercase/capital letter X into the box next to the answer option you wish to select as your answer to this question. Leave all other boxes empty for this question.* | |
|  | A. | t test for the population mean |
|  | B. | large sample test for the population proportion |
|  | C. | t test for the difference between means |
|  | D. | large sample test for the difference between proportions |
|  | E. | t test for slope |
|  | F. | one-factor ANOVA |

b) Write the pair of statistical hypotheses (i.e. null and alternative) in *sentence form* that would be relevant, based on the inference procedure you selected in *part a*. Your hypotheses should be phrased in the context/language of the specific research question/scenario. *(5 marks, suggested length is 2 sentences)*

Note: you have the rest of this page plus the entirety of the next page for your answer. The second page is totally unnecessary, but provided nevertheless!

**Question 11.**

**a)** Copy/paste the name of the inference procedure you selected from *Question 10 part a*.

**b)** Before applying an inference procedure, we must evaluate the model conditions for the procedure.

(i) write each **condition** required for the hypothesis test to be valid, and,

(ii) below each condition, **briefly** **describe/identify** the method (i.e. how?) you should use to evaluate the validity of the condition. For conditions that involve manipulating/investigating some aspect of the data, use **R** to produce the R output necessary for evaluation using the dataset.csv file you were provided. Include the R code and output in your answer *(6 marks, suggested length is 2-5 sentences, along with any R code and output)*

|  |
| --- |
| **Remember!** If you can’t get R to do what you want, at least show the code you were attempting to make work, and briefly describe what you were trying to do. |

Note: this question is NOT asking you to actually evaluate/justify whether the condition(s) are valid. You are asked to demonstrate *how/what we would use* to evaluate each condition(s). Be explicit and specific to this dataset, identifying relevant variables, subsets, etc (i.e. demonstrate you understand how the conditions apply to this particular Research Question and dataset).

*Note: you have the rest of this page plus the entirety of the next page for your answer.*

Question 12.

Use R to conduct the inference procedure you selected in *Question 10* *part a*. To answer this question:

**a)** Copy/paste the name of the **inference procedure** you selected in *Question 10 part a*.

b) Copy and paste your null and alternative hypotheses in *sentence form* from *Question 10 part b*.

c) Provide the R code and Output you used to conduct the inference procedure. It should be clear from the code you report (and any brief #comments) where any values/vectors you use as arguments come from (i.e. we should be able to figure out/recreate what you did based on your code). *(3 marks)*

d) Provide a proper scientific conclusion for the results of your analysis. *(5 marks, suggested length is 1-2 sentences)*

|  |
| --- |
| **Remember!** If you can’t get R to do what you want, at least show the code you were attempting to make work, and briefly describe what you were trying to do. |

*Note: you are provided the rest of this page, plus the entirety of the next page to input your answer to this question.*

**Question 13.**

Create a SINGLE **graph** in R, plus write an appropriate **figure legend** that would be useful to accompany the results of your analysis from *Question 12*. As your answer to this question, show your graph and figure legend. Below your graph and figure legend, provide all the **R code** you used to generate your graph (with any brief #comments necessary to help us interpret the code; be sure to include any preliminary data manipulation, variable creation, etc. that you did!). *(9 marks)*.

Note: the expectations for this question are effectively analogous to those from a similar question on Lab Assignment 2. That means the graph produced doesn’t have to be “perfect”, it just needs to address the research question and demonstrate some minimal customization.

|  |
| --- |
| **Remember!** If you can’t get R to do what you want, at least show the code you were attempting to make work, and briefly describe what you were trying to do. |

*Note: you are provided the rest of this page and the entirety of the next page for your answer to this question.*