



Western University
Biology 2244B/Statistics 2244B (FW2019)
Take-Home Final Exam (20 or 30%)

Available for 24 h starting at 5 pm April 23, 2020 (London, ON)
Due date: 5 pm April 24, 2020 (London, ON)

**if you are in a different time zone, it is your responsibility to determine what the equivalent due date is local to your region.*

Instructions for this take-home exam:

It is IMPERATIVE that you read through these instructions prior to beginning this exam, to ensure that you format/submit your exam correctly and are familiar with the expectations for the exam.

1. This file contains **25** pages. When you complete the exam and are ready to submit the file, the file you submit must also have **25** pages. This cover page plus instructions are pages **1** and **2**, for reference.
2. Each question on this exam is on a specific page(s), and provides *ample* space for you to submit your answer. When you complete the exam and are ready to submit the file, the locations and spaces provided for your answer must be the same as was originally provided. The question locations are listed here for handy reference/double checking that you haven't changed the spacing of the exam:

Here are many references, students checking that you haven't changed the spelling of

Section A: Multiple Choice Questions (aim for about 25 minutes dedicated to this section)		
Question 1	Page 3	You will indicate your answer by placing an X in the box corresponding to the answer option you choose.
Question 2	Page 4	
Question 3	Page 5	
Question 4	Page 6-7	
Question 5	Page 8	
Section B: Short Answer Questions (aim for about 45-60 minutes dedicated to this section)		
Question 6	Page 9-10	You will type your answer into the space provided; if you use pictures/images, these should also be placed in the space allocated for each question
Question 7	Page 11-12	
Question 8	Page 13-14	
Question 9	Page 15-16	
Section C: Data Analysis Questions (aim for about 120 minutes dedicated to this section)		
Context for Q10-13	Page 17	You will type your answer into the space provided; if you use pictures/images, these should also be placed in the space allocated for each question.
Question 10	Page 18-19	
Question 11	Page 20-21	
Question 12	Page 22-23	
Question 13	Page 24-25	

3. Do NOT edit or delete any *question text* originally present in this file. You should only ADD your answer text/graphics, etc. in the space provided for each question.
4. When you are ready to submit your completed exam (following the rules above), save your file as a single PDF file. You will upload the PDF file in TWO (2) locations by the deadline of 5 pm on Friday April 24 (time local to London ON):
 - a. On the 2244B OWL course site, through the tool, "Assignments (OWL submissions)". Upload to the 'assignment' from which you obtained the take home exam files.
 - b. On Gradescope, under the 'assignment' Take Home Exam – due April 24 at 5 pm.

Expectations for this take-home exam

1. You are permitted to use your course notes, textbook, and other course-related resources as you see fit. The take home exam can be considered an 'open book' exam. However, be leery of using non-course related internet resources or textbooks, etc.; you will be marked based on what we have covered in Biol/Stat 2244 (video) lectures and required reading.
2. You are expected to complete this exam individually, i.e. without support from your peers or other people (e.g. personal tutors, paid services, etc.). You should be aware that collaborating with other people would be considered an academic offense under Western University policy.

Honour code

Western University is a community of students, faculty, and staff involved in learning, teaching, research, and other activities. The University seeks to provide an environment of free and creative inquiry within which critical thinking, human values, and practice skills are cultivated and sustained. It is committed to a mission and to principles that will foster excellence and create an environment where its students, faculty, and staff can grow and flourish.¹ "Members of the University accept a commitment to maintain and uphold the purposes of the University and, in particular, its standards of scholarship."²

By submitting answers to this take home exam, you unequivocally state that all work is entirely your own, and is submitted with an understanding of Western University's policy on "Scholastic Discipline for Undergraduate Students" (available at:

https://www.westerncalendar.uwo.ca/PolicyPages.cfm?Command=showCategory&PolicyCategoryID=1&SelectedCalendar=Live&ArchiveID=#Page_20).

Enter your name here: Yiran Shao

Enter your student number here: 251015795

Remember!

- You have been given WAY more space than necessary for each question. "More" is not better for written answers. Take a minute to think through a question, and then respond.
- Grammar/spelling don't count. Just do your best to make your answers understandable.
- We know you only have a limited amount of time and energy to dedicate to this exam. Just focus on demonstrating your knowledge/understanding to the best of your ability. Part marks are always possible.
- If you can't completely answer a question (especially in the Data Analysis section with R), at least answer the parts you can and/or briefly describe what you would or are trying to do.

You can do this!!

¹ Text quoted from the Western University code of Student Conduct (effective April 25, 2019) available at: <https://www.uwo.ca/univsec/pdf/board/code.pdf>

² Text quoted from Western University's Academic Calendar under Academic Rights and Responsibilities, from the section on Scholastic Discipline for Undergraduate Students, available at: https://www.westerncalendar.uwo.ca/PolicyPages.cfm?Command=showCategory&PolicyCategoryID=1&SelectedCalendar=Live&ArchiveID=#Page_20

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 viewers who subscribe to Amazon Prime are 40 y or younger?

		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% |
| | C. | 55.8% |
| X | 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 cm^2) 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 shape 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.


	A. Blocking variable
X	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). They randomly selected two of the five quadrants. From each of those two selected quadrants, they randomly selected fifty (50) houses to visit and recorded 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
	B. A cluster sample
X	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 probabilities for Love in Rome, and a column called *Sunset_prob* for the 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. `> t.test(x=Rome_prob, y=Sunset_prob)`
- ```
welch Two Sample t-test

data: Rome_prob and Sunset_prob
t = -2.4796, df = 82, p-value = 0.0152
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.089907257 -0.009863535
sample estimates:
mean of x mean of y
0.5007270 0.5506124
```
- ☐ B. `> t.test(x=(Rome_prob-Sunset_prob))`
- ```
One Sample t-test

data: (Rome_prob - Sunset_prob)
t = -2.2749, df = 41, p-value = 0.02821
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.094171866 -0.005598926
sample estimates:
mean of x
-0.0498854
```
- ☐ C. `> t.test(x=Rome_prob, y=Sunset_prob, alternative="greater")`
- ```
welch Two Sample t-test

data: Rome_prob and Sunset_prob
t = -2.4796, df = 82, p-value = 0.9924
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.08335532 Inf
sample estimates:
mean of x mean of y
0.5007270 0.5506124
```
- ☐ D. `> t.test(x=Rome_prob, y=Sunset_prob, mu=-0.05)`
- ```
welch Two Sample t-test


data: Rome_prob and Sunset_prob
t = 0.0056965, df = 82, p-value = 0.9955
alternative hypothesis: true difference in means is not equal to -0.05
95 percent confidence interval:
 -0.089907257 -0.009863535
sample estimates:
mean of x mean of y
0.5007270 0.5506124
```
-

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!

Because of ethical concerns, a prospective cohort study will be conducted with the explanatory variable being the level of screentime and response variable being the attention span. A large-enough sample of young children sharing a common demographic characteristic (e.g. resemble a Canadian population based on a set of variables including gender, race, household income, etc.) will be selected by simple random sampling to participate in the study. At regular intervals over a relatively long period of time, researchers will record the average amount of time per day they spent watching TV/movies, their main content of choice and their attention span.

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:

- i. their employment status (specifically, whether they are employed or not during the most recent school term)
- ii. their course grades for the most recent school term (specifically, their mean grade (%) across all their courses during the school term).

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.

- (a) **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).
- (b) 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)
- (c) 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).

a).

1. What is your analysis goal?	My analysis goal is to assess evidence for the claim that student employment does not influence academic success.
2. How many “samples” or groups are you comparing?	I’m comparing 2 samples, which are 1) students that are employed during the most recent school term and 2) students that are not.
3. Are your “samples” independent or paired?	They are independent.
4. What type of data are you collecting? quantitative or categorical (qualitative)?	The data that I’m collecting, the course grades for the most recent school term, are interval data which is quantitative and continuous.
5. What parameter(s) are of interest? mean(s)? proportion(s)? other?	The parameter of interest is the average (mean) course grades of students that are employed during the most recent school term and the average of those who are not.

b). t procedures for difference between means.

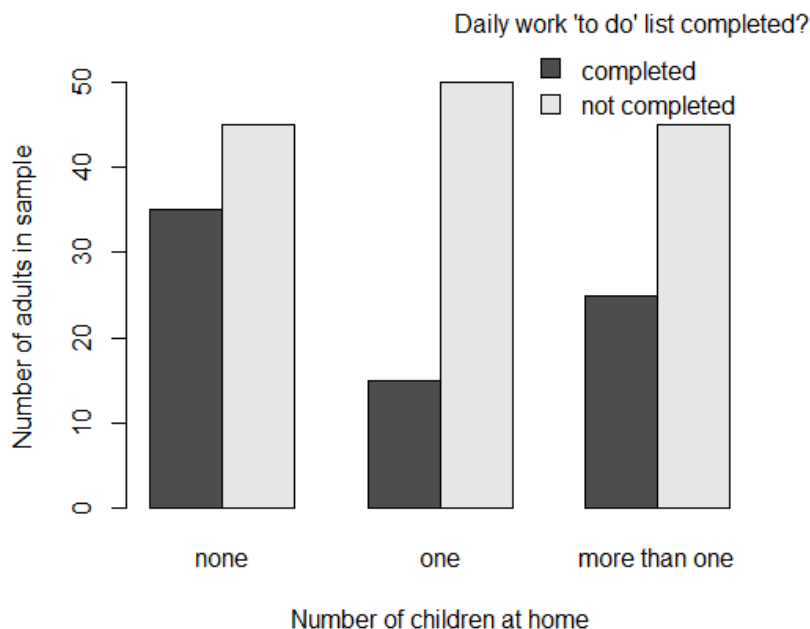
c). The sampling distribution of difference between means is the normal distribution.

Mean: $\mu_1 - \mu_2$

Standard deviation: $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Question 8.

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



- (a) Write a research question that could be answered based on this graph. (*~3 marks, suggested length is one sentence*).
- (b) 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).

- a). Does the number of children at home influence the ability of adults to complete their daily work 'to do' list?
- b).
1. The proportions of adults that did not complete their daily work 'to do' lists are always greater than that of those who completed them, regardless of how many children at home they have.
 2. Adults with one child at home are most unlikely to complete their daily work 'to do' list among the three groups, following by adults with more than one children at home. Adults with no children at home are most likely to complete their daily work 'to do' list.

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 evaluate whether there is a relationship between plant height and distance between plants. He has a large garden, so has subsequently planted ten (**20**) rose plants at a variety of distances from one another (anywhere ranging from **6** inches to **12** inches apart).

- (a) **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)*
- (b) 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).

a). Correlation and regression

b).

1. There is a linear relationship between distance between plants and plant height.
2. The observations of plant height are independent.
3. The population distribution of plant height is Normally distributed for each distances used in the experiment.
4. The standard deviation of population distribution of plant height is the same for all of the heights recorded at one distance used.

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 7974 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	Integer 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 weight differ across medal type for athletes competing in the Athletics sports 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
X	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!

b).

Null hypothesis (H_0): All mean weights of the athletics winning three types of medal, gold, silver and bronze, competing in the Athletics sports of the Summer Olympic Games are equal.

Alternative hypothesis (H_A): At least the mean weight of athletics winning one type of the medals competing in the Athletics sports of the Summer Olympic Games is different.

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.

- a) one-factor ANOVA
- b)
- i) Conditions
 - 1) Observations of weight from populations of athletics are independent samples.
 - 2) Each sample of weight is a simple random sample from its type of medal group
 - 3) Each sample's observations come from a population of medal group that are Normally distributed
 - 4) The standard deviation or variance of the population distribution is the same across the population.
 - ii) Evaluations of validity of the conditions (in the same order as b(i))
 - 1) The sampling of the study design procedure does not have pairing structure so they are independent.
 - 2) We can check for the study design and sampling method to see if each of the sample is SRS. If the subset of the years are selected randomly, this condition is valid.
 - 3) Firstly, we can make a graph showing the population distribution for all of the three medal groups on an x-axis of weight to see if they all follow a Normal model. Secondly, we can make a QQ plot of 'residuals' to check the normality.

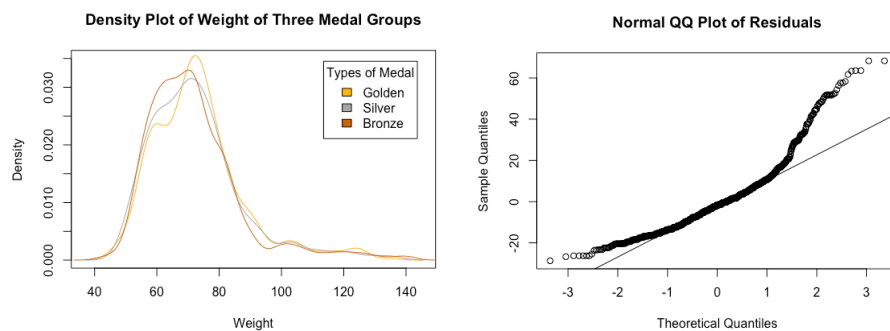
```
# Make a graph to describe the population distribution of the three medal groups
weight_medal <- subset(dataset, select=c(weight, medal, sport))
# Make three vectors containing the density of the weight of athletics winning three types of medal
Gdensity <-
density(weight_medal$weight[weight_medal$medal=="Gold"&weight_medal$sport=="Athletics"])
Sdensity <-
density(weight_medal$weight[weight_medal$medal=="Silver"&weight_medal$sport=="Athletics"])
)
```

```

Bdensity <-
density(weight_medal$weight[weight_medal$medal=="Bronze"&weight_medal$sport=="Athletics
"])
# Make the density curve of the weight of the three medal groups
plot(Gdensity, main = "Density Plot of Weight of Three Medal Groups", ylim = c(0,0.035), xlab =
"Weight", col = "darkgoldenrod1")
lines(Sdensity, col = "darkgray")
lines(Bdensity, col = "darkorange3")
legend("topright", inset=0.05, title = "Types of Medal", c("Golden", "Silver", "Bronze"),
fill=c("darkgoldenrod1","darkgray","darkorange3"))

# Make a QQ plot of residuals
Gweight <-
weight_medal$weight[weight_medal$medal=="Gold"&weight_medal$sport=="Athletics"]
Sweight <-
weight_medal$weight[weight_medal$medal=="Silver"&weight_medal$sport=="Athletics"]
Bweight <-
weight_medal$weight[weight_medal$medal=="Bronze"&weight_medal$sport=="Athletics"]
weight.residual <- c((Gweight-mean(Gweight)), (Sweight-mean(Sweight)), (Bweight-
mean(Bweight)))
qqnorm(weight.residual, main="Normal QQ Plot of Residuals")
qqline(weight.residual)

```

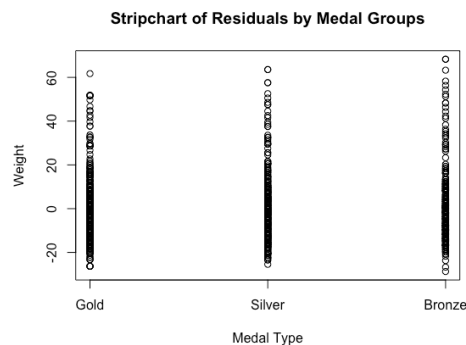


- 4) We can make a stripchart of residuals by medal groups to check the constant standard deviation.

```

weight_medal.wide <- list("Gold"=Gweight, "Silver"=Sweight, "Bronze"=Bweight)
stripchart(weight_medal.wide, main="Stripchart of Residuals by Medal Groups", xlab="Medal
Type", ylab="Weight", vertical=TRUE, pch=1)

```



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.

- a) one-factor ANOVA
- b) Null hypothesis (H_0): All mean weights of the athletics winning three types of medal, gold, silver and bronze, are equal.
Alternative hypothesis (H_A): At least the mean weight of athletics winning one type of the medals is different.

c)

Code:
`Athletics_data <- subset(dataset, sport=="Athletics")`
`weight_medal.aov <- aov(Athletics_data$weight~Athletics_data$medal, data=Athletics_data)`
`summary(weight_medal.aov)`

Output:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Athletics_data\$medal	2	535	267.4	1.176	0.309
Residuals	1283	291608	227.3		

- d) Mean weights are the same for the athletics winning all three types of the medals ($F_{2, 1283}=1.176$, $P=0.309$).

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.

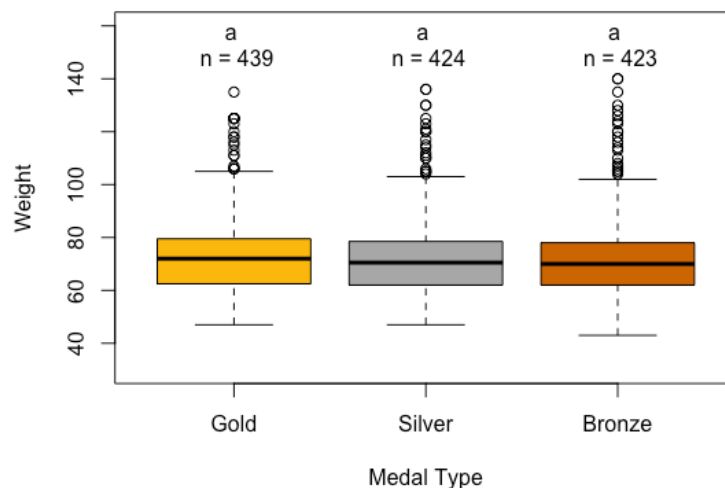


Fig 1. Comparison of mean weight of athletes competing in the Athletics sports of the Summer Olympic Games with three medals from Olympic Summer Games for a subset of the years that these Games have run. Medal types that do not share the same letter have significantly different means at the $p < .05$ level. Sample sizes of each group are shown above the boxplots.

Code:

```
# Re-order the levels of the medals
Athletics_data$medal <- ordered(Athletics_data$medal, levels=c("Gold", "Silver", "Bronze"))
# Make a boxplot of the athletics' weights to their medal type
boxplot(Athletics_data$weight~Athletics_data$medal, data=Athletics_data, xlab="Medal Type", ylab="Weight",
        ylim=c(30,160), col=c("darkgoldenrod1","darkgray","darkorange3"))
text(x=1:3, y=153, labels=c("a \n n = 439","a \n n = 424","a \n n = 423"))
```