STAT 210: Exam #3 Take-home portion Name: _____
Fall 2017
Points: 50

CitiBike Bike Share

Consider again the Citibike bike share data. The most recent data provided on this website is fro September 2017. To advoid any confusion from Exam #2, you should download this data again – the JC- datasets should *not* be used for this exam.

Source: https://www.citibikenyc.com/system-data
Data Download: https://s3.amazonaws.com/tripdata/index.html

Download the most recent data – appears to be 201709-citibike-tripdata.csv.zip, which is about three-fourths of the way down the page and just before the start of the JC- data files. The file being downloaded is a *.zip file. WSU laptops come with software to unzip files of this type. Unzip this file so that the 201709-citibike-tripdata.csv can be opened into JMP.

201709-citibike-tripdata.csv.zip	Oct 3rd 2017, 10:52:58 am	65.88 MB	ZIP file

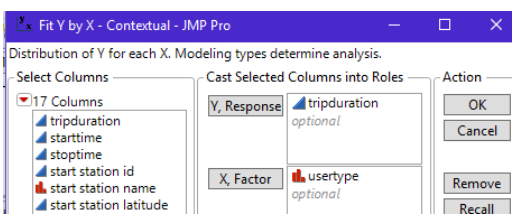
Consider the following columns in this dataset.

- tripduration (duration of trip in seconds)
- usertype (Customer = 24 hour pass or 3-day pass -- temporary type member; Subscriber = Annual pass – more permanent type member)
- Year of Birth
- Gender (Zero = unknown, 1 = male, 2=female)

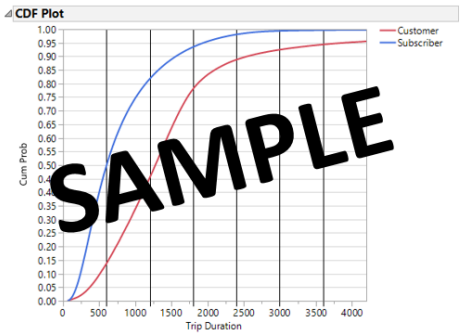
Consider a summary of tripduration across the usertype. In JMP, place tripduration in the Y, Response box and usertype in the X, Factor box. Recreate the CDF plot below with the following criteria.

- Y axis should go from 0 to 1 with grid lines every 0.05.
- X axis should range from 0 to 4200 or so. Vertical lines placed at 600, 1200, 1800, 2400, 3000 and 3600.

Fit Y by X Dialog box



Recreate this plot



1. Answer the following using your CDF plot.

    a. Why were vertical lines placed at every 600 units along the x-axis? Briefly discuss. (2 pts)

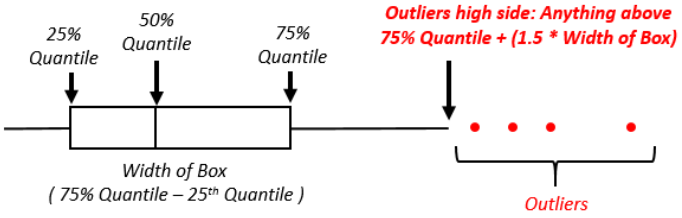
    b. What happens to this plot when the x-axis is allowed to go span the entire range of tripduration? Briefly discuss. (2 pts)

    c. What information is gained from a careful review of this plot? Discuss. (3 pts)

    d. The tripduration variables has many outliers. Do the outliers mess up this plot? Briefly discuss. (2 pts)

Tripduration values that are outliers (on the high-end) are known to cause problems for a bike share program – all stations need bikes available. A compounding problem is that most often bikes are not returned to the same station they were rented from – that is, the start and stop stations are different. Due to the extreme skewness in this data, the boxplot approach will be used to determine outliers – the z-score approach uses the mean and standard deviation which are adversely affected by the extreme outliers present in this data.

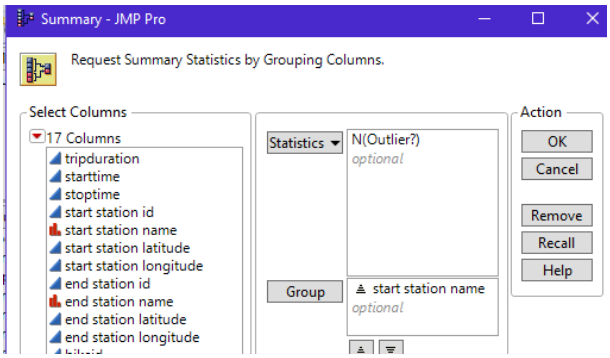
Outlier rule via boxplots: Anything above: 75% Quantile + (1.5 * Width of Box)
Anything below: 25% Quantile – (1.5* Width of Box)



2. Consider only outliers for which trip duration is atypically large. Determine the top 5 stations that have the most outliers on the high-side using the outlier rule provided above. When making this determination, you should only consider rentals that were returned to a station different from the originating station. (5 pts)

Rank	Station Name
1 (Most)	
2	
3	
4	
5	

Counting Outliers by Station can be accomplished using Tables > Summary with the following setup



WSU Student Survey

In some semesters, the extra credit assignment for STAT 110 and STAT 210 involves taking a survey to gather data on WSU students. The data from a past semester can be downloaded from our course website. The data collected can be used to answer a wide-variety of questions – such as the following.

A. I've heard some people over the years say, "At Winona State, women are smarter than men." Some believe this to be the case because women are often nursing or elementary education majors and these programs have strict entrance requirements. Run an appropriate statistical test to determine whether or not there is any truth to the statement, "At Winona State, women's WSU GPA is higher than men's WSU GPA."

B. The demands of athletes at Winona State are often higher than a typical student. As a result, academic advisors / coaches of athletes tend to be more diligent in keeping track of the progress of their students and in making sure student athletes take advantage of available resources to ensure success. Run an appropriate statistical test to verify whether or not it is true that athletes at WSU tend to have a higher WSU GPA than non-athletes?

C. My grandma often said, "Don't do any of those risky behaviors (like smoking, drinking, or having a significant other) as these things tend to distractive you from doing well in school." I think Grandma was just trying to scare me away from doing these things and do not believe statements like this are true. However, go ahead and run an analysis for, say drinking, to test whether or this has any impact on WSU GPA.

D. Our car insurance company offers a discount when our children get above a certain GPA level. Thus, our insurance company must believe a relationship exists on driving habits and GPA. Run an appropriate statistical test to verify the following statement, "Those that have been in car crash have a lower WSU GPA than those that have not been in car crash."

1. Fill in the following table for your analyses. (8 pts)

Research Question	P-Value?	*Brief discussion regarding which test you ran* *Choices: 1) When variances are similar use pooled t-test, 2) When one variance is more than double the other, use t Test, 3) Use Wilcoxon or Median test when data fails normality or when outliers are persent*
A		
B		
C		
D		

2. Consider the following phrases that captures the essence of each statement. Write a response that is void of statistical language, but is based on what you learned from your statistical analysis. (8 points)

Research Question	Phrase	Response?
A	*Women tend to have a higher GPA*	
B	*The diligence by advisors/coaches tends to improve an athletes GPA*	
C	*Grandma was right – drinking tends to affect GPA*	
D	*A legitimate rationale exists for insurance companies to give discounts to those with higher GPAs*	

Discriminant Analysis

A thorough discriminant analysis is beyond the scope of this class; however, some of the concepts that you've learned in class can be used here to perform an ad hoc discriminant analysis. See the Discriminant Analysis Wikipage for more information -- https://en.wikipedia.org/wiki/Discriminant_function_analysis

The "big-picture" view of a discriminant analysis is to find variables that yield good separation in a set of measurements across two or more groups. For the sake of discussion, suppose I have three different traits that are being measured with the goal of separating Group 1 from Group 2. In the graphs below, Trait 1 would be the best discriminator because for Trait 1 there is the least amount of overlap between the distributions. Trait 2 has more separation between the averages, i.e. 20 vs 7; however, the variation in each group is larger causing greater overlap in the distributions. Likewise, Trait 3 has a difference of 7 – like Trait 1, but again an increased level of variation hurts our ability to discriminant between groups.

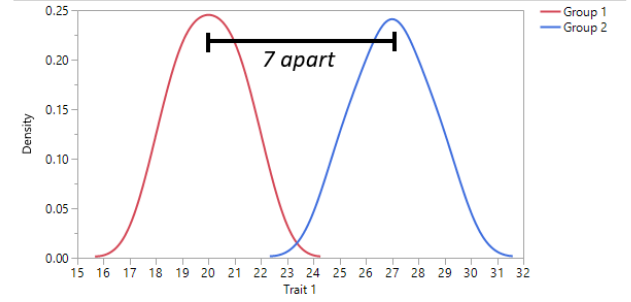
Compare Densities for Trait 1

Means and Std Deviations

Level	Number	Mean	Std Dev
Group 1	8	20.0000	1.30931
Group 2	6	27.0000	1.41421

Differences in averages for Trait 1 = 7
Trait 1 is a *good* discriminator -- not much overlap between Group 1 and 2

Compare Densities



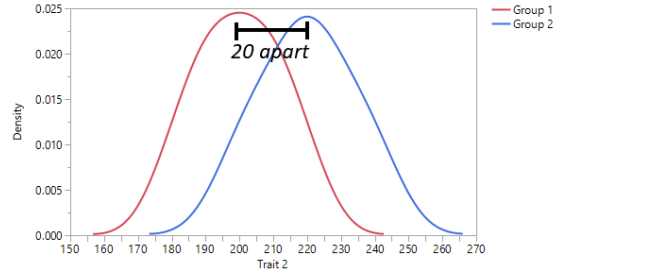
Compare Densities Trait 2

Means and Std Deviations

Level	Number	Mean	Std Dev
Group 1	8	200.000	13.0931
Group 2	6	220.000	14.1421

Differences in averages for Trait 2 = 20
Trait 2 is a *not* a good discriminator – considerable overlap between Group 1 and 2

Compare Densities



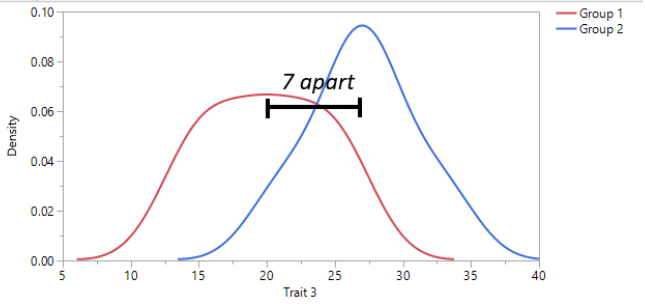
Compare Densities for Trait 3

Means and Std Deviations

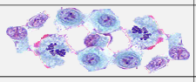
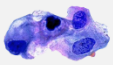
Level	Number	Mean	Std Dev
Group 1	8	20.0000	4.47214
Group 2	6	27.0000	4.00000

Differences in averages for Trait 3 = 7
Trait 3 is a *not* a good discriminator – considerable overlap between Group 1 and 2

Compare Densities



Here, we will investigate what cell features or characteristics best discriminate between benign (normal) cells and malignant (cancerous) cells.

Normal	Cancer	
		Large, variably shaped nuclei
		Many dividing cells; Disorganized arrangement
		Variation in size and shape
		Loss of normal features

Source: http://sphweb.bumc.bu.edu/otlt/mph-modules/ph/ph709_cancer/ph709_cancer7.html

We will use the well-known Wisconsin Diagnostic Breast Cancer (WDBC) dataset to preform our discriminant analysis. This dataset is available on our course website.

	Diagnosis	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity	Concave Points	Symmetry	Fractal Dimension
1	Malignant	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
2	Malignant	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
3	Malignant	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999
4	Malignant	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744
5	Malignant	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883

Breast Cancer – Data: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data
Breast Cancer – Variable Descriptions: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names

The cell characteristics under consideration are provided here – from the variable description link above.

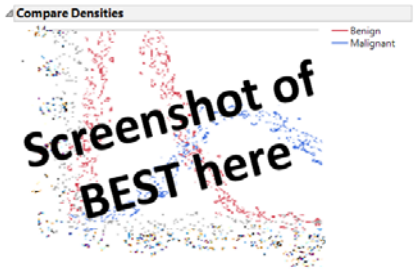
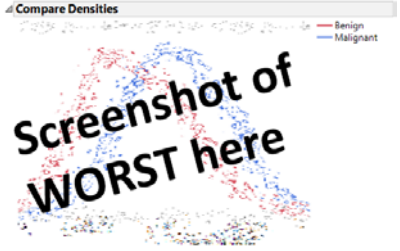
```
Ten real-valued features are computed for each cell nucleus:

        a) radius (mean of distances from center to points on the perimeter)
        b) texture (standard deviation of gray-scale values)
        c) perimeter
        d) area
        e) smoothness (local variation in radius lengths)
        f) compactness (perimeter^2 / area - 1.0)
        g) concavity (severity of concave portions of the contour)
        h) concave points (number of concave portions of the contour)
        i) symmetry
        j) fractal dimension ("coastline approximation" - 1)
```

Goal: Investigate the ability of each cell characteristic to discriminant between benign and malignant cells.

3. Using Fit Y by X in JMP, obtain the Mean and Standard Deviation for each cell characteristic. Fill-in the following table. You can leave the Rank column empty for now. (5 pts)

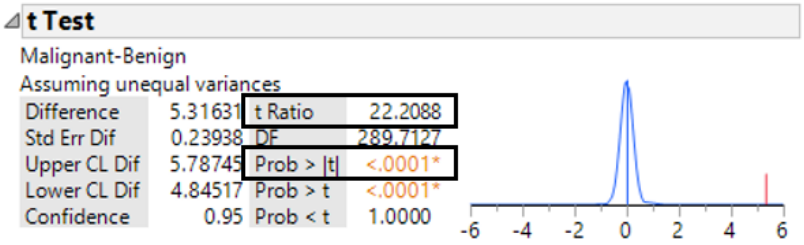
Cell Characteristic	Benign		Malignant		Rank 1: Best 10: Worst
	Mean	Standard Deviation	Mean	Standard Deviation	
Radius					
Texture					
Perimeter					
Area					
Smoothness					
Compactness					
Concavity					
Concave Points					
Symmetry					
Fractal Dimension					

4. Next, from the Fit Y bx X output, select using Densities > Compare Densities. Provide a screen-shot of the Compare Densities plots that you believe are the best and worst discriminating characteristic to separate benign cells from malignant cells. (2 pts each)

Best Discriminating Characteristic	Worst Discriminating Characteristic
	

5. On possible way to measure the degree of separation in the measurements between the benign and malignant cells is to simply conduct a t-test (akin to Handout #10). For simplicity, we will use the unequal variance version, i.e. select t Test from drop-down menu – we will not use the pooled version of the t-test.

   The cell characteristics with smallest two-tailed p-value will indicate the greatest degree of separation between the average of the benign cells and the average of the malignant cells.



   Unfortunately, the degree of separation cannot be determined because several cell characteristics have a two-tailed p-value listed as < 0.0001 and thus cannot be ranked. An alternative to using the p-values would be to use the t Ratio values. The largest t Ratio value (furthest from 0) would imply the most separation between the averages and the smallest (closest to 0) would imply the least separation between the averages.

   Use the t Ratios value from each cell characteristic t-test to rank the degree of separation between the benign and malignant cells. A rank of 1 should indicate the largest amount of separation and a rank of 10 the least. Place your ranks in the table provided above. (3 pts)

6. As discussed above, the amount of separation in the two distributions must consider the amount of separation in the averages *relative* to the amount of inherent variation in the distributions. Use whatever resources you can find to find the formula for the t-Ratio value for conducting an unequal variance t-test.

   a. Show the actual calculations for at least one t Ratio value computed by JMP. (3 pts)

   b. What is the numerator measuring, in the context of your example? Briefly discuss. (2 pts)

   c. Some argue that the formula for a t Ratio is similar to that of a Z-Score value. I'm not sure if I agree or disagree with this. What do you think? Discuss. (3 pts)

      From Handout #9, the formula for Z-Score is: $Z - Score = \frac{(Data\ Point - Mean)}{Standard\ Deviation}$