This is a take-home exam that intends to assess your learning abilities after the first five weeks of the STAT 530 course (Exploring Multivariate Data). Being a take-home exam implies you may use class notes, textbooks, calculators etc. to help you answer questions but absolutely no other human sources. The test is due Tuesday, October the 25th by 4:00 p.m. The test may be turned to Dr. Habing personally, left in his mail-box in 216 LeConte, faxed to (803) 777-4048, or sent by e-mail in a pre-approved format.

There are 20 multiple-choice items, and for each item you have to select the BEST answer for that question by circling the letter. The final test score is based on the number of correct answers, so be sure you attend and answer each item. The point value of each item is 0.5 for a 10 points total. This counts as an optional 10 points homework.
Multivariate data come from measuring several variables on each “unit” of the sample. In order to be able to understand the data we need to use methods that are descriptive as well as inferential. For a data set with a large number of variables someone can use a variety of statistic methods to describe it.

The following first 7 questions pertain to general statistical knowledge about data.

1. The $P(\text{rejecting } H_0 | H_0 \text{ is true})$ it is called______________.
   
   a. significance level  
   b. power  
   c. confidence interval

2. A t-test is used to compare__________.
   
   a. means  
   b. standard deviations  
   c. variances

3. Below are some Multivariate Graphical Displays. Column 1 contains names of 3 display types. Please match each graph with the BEST descriptive name by placing the letter assigned to each graph on the line to the left of the descriptive name. Each graph matches only one display type.
   Column 1:
   _____ a. scatterplot
   _____ b. contour plot
   _____ c. 3D density plot

   A.  
   B.  
   C.  

4. Normal data should result in boxplots that are ________.
   
   a. asymmetric  
   b. skewed  
   c. symmetric
5. Normal data should result in qq plots that are __________.
   a. like an s-shaped curve
   b. like a side-ways s-shaped curve
   c. a straight line

6. Imputation is when missing values are dealt with by__________.
   a. removing these observations
   b. replacing the values with 0’s
   c. replacing the observations with estimated values

7. For a multivariate normal random sample \((X-q)^T \Sigma^{-1} (X-q)\) is distributed as __________;
   where \(q\) is the number of variable.
   a. a t with df = \(n – q\)
   b. an F with df=(\(q, n-q\))
   c. Chi-squared with df =q

The next 6 questions regard Principal Component Analysis (PCA).

8. Principal Component Analysis (PCA) can be applied to?
   a. nominal variables
   b. ordinal variables
   c. interval variables

9. Why are correlations preferred to covariance?
   a. Correlations retain the scales of the variables.
   b. Correlations are all between –1 and 1.
   c. Correlations are scale free.

10. Which of the following is a true statement about principal components?
    1. T or F The coefficient vector should be equal to the eigen values.
    2. T or F The coefficient vector should be orthogonal to the previous one(s).
    3. T or F The coefficient vector should explain as little variance as possible.

11. Eigen vectors of the covariance matrix for a PC analysis are __________.
    a. the covariances of the PC
    b. the coefficients for the transformation
    c. the correlations of the transformation
12. Which of the following statement about the eigen values of the covariance matrix is true?

a. The eigen values are the variances of the new variables.
b. The eigen values are the variances of the original variables.
c. The eigen values are the standard deviation of the original variables.

13. The eigen value rule is to choose factors with eigen values greater than__________.

a. 0.05  
b. 0.95  
c. 1.00

The below figures depicts two scree plots. Please take a closer look and answer questions 14 and 15.

14. Just by looking at the A scree plot how many components would you take into consideration?

a. 1  
b. 2  
c. 3

15. Just by looking at the B scree plot how many components would you take into consideration?

a. 1  
b. 4  
c. 5
Answer questions 16, 17, and 18 given the below loadings by using the method "iterated principal factor - no rotation"

<table>
<thead>
<tr>
<th>loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor1</td>
</tr>
<tr>
<td>m0</td>
</tr>
<tr>
<td>m25</td>
</tr>
<tr>
<td>m50</td>
</tr>
<tr>
<td>m75</td>
</tr>
<tr>
<td>w0</td>
</tr>
<tr>
<td>w25</td>
</tr>
<tr>
<td>w50</td>
</tr>
<tr>
<td>w75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>communalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>m0  m25  m50  m75  w0  w25  w50  w75</td>
</tr>
<tr>
<td>1.001 0.604 1.043 0.715 0.977 0.984 0.960 0.905</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>variance explained</td>
</tr>
<tr>
<td>5.523 1.243 0.423</td>
</tr>
<tr>
<td>percent explained</td>
</tr>
<tr>
<td>0.690 0.155 0.053</td>
</tr>
</tbody>
</table>

16. Specify the number of variables that significantly load on Factor 2.

   a. 2
   b. 4
   c. 8

17. In general we want the communalities to be __________.

   a. close to 0
   b. close to 0.5
   c. close to 1

18. Using the output at the top of page 5, if only Factor 3 would be used to summarize the data, ____ percent of information it will be lost.

   a. 10.2%
   b. 69%
   c. 94.7%
Below is an illustration of an R summary. Use it to answer questions 19 and 20.

Importance of components:

|       | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 | Comp.8 | Comp.9 | Comp.10 | Comp.11 | Comp.12 | Comp.13 | Comp.14 | Comp.15 | Comp.16 | Comp.17 | Comp.18 | Comp.19 | Comp.20 |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| SD    | 1.766  | 1.567  | 1.361  | 1.307  | 1.047  | 0.991  | 0.977  | 0.977  | 0.977  | 0.947   | 0.849  | 0.817   | 0.773   | 0.748   | 0.677   | 0.668   | 0.637   | 0.575   | 0.544   |
| Prop  | 0.159  | 0.125  | 0.094  | 0.087  | 0.056  | 0.050  | 0.049  | 0.049  | 0.049  | 0.045   | 0.037  | 0.034   | 0.031   | 0.028   | 0.023   | 0.023   | 0.021   | 0.015   | 0.013   |
| cum   | 0.159  | 0.284  | 0.378  | 0.465  | 0.521  | 0.571  | 0.619  | 0.665  | 0.708  | 0.747   | 0.784  | 0.818   | 0.848   | 0.876   | 0.925   | 0.947   | 0.968   | 0.984   | 1.000   |

19. From the above R output, using Kaiser’s criterion we will keep at least ___ factors.
   a. 1
   b. 5
   c. 20

20. Using the percentages of the variation _____ factors will explain at least half of the variance?
   a. 1
   b. 5
   c. 20