STAT 530/J530
October 20th, 2005

Instructor: Brian Habing
Department of Statistics
LeConte 203
Telephone: 803-777-3578
E-mail: habing@stat.sc.edu

Exam

The raw data set used for all of the following questions can be found at:

`http://www.stat.sc.edu/~habing/courses/data/scdata.txt`

```
scdata<-as.matrix(sctable[,3:27])
rownames(scdata)<-sctable[,1]
regions<-sctable[,2]
```


Question 1

Choose one of the twenty-four variables as an example for each of the following: approximately normally distributed, approximately normally distributed except for an outlier, skewed right, skewed left, heavy tailed, light tailed. Provide an appropriate statistic or graph to justify each choice.

Question 2

South Carolina is often broken into distinct regions. One way of breaks it into four regions: the Low Country, the Midlands, the PeeDee, and the Upstate. Construct an appropriate graphical display that compares the four regions to each other in terms of the percent population change from 1990-2000. Briefly compare the four regions' distributions in terms of the center, spread, and skew.

Question 3

Consider the set of the three variables Coll4, HSGrad, and Income. Identify any counties that appear to be potential outliers for this set, justifying your method of locating them.

Question 4

Construct a plot that seems to show an apparently strong relationship between high school graduation rates and the percentage of minorities in a county. Construct a follow up plot that demonstrates this apparent relationship may be due to the average income level.

Question 5

Check whether the variables Urban and Income seem to satisfy the properties of a multivariate normal distribution.

Question 6

Using the *South Carolina Statistical Abstract 2005* we could convert most of the above variables into county totals instead of county rates (e.g. 2=total deaths in the county, 9=total income for all county residents). Briefly explain why it is probably more useful to use the rates instead of the totals. Also briefly explain why simply taking the standardized values (z-scores) of the totals would also not work as well as using the rates.

Part II – Question 1

It is desired to reduce the set of eleven financial variables (9=Income to 19=Unemp) to a more manageable size by using principal components.

a) Choose to use either the correlation or covariance matrix and justify your choice.

Part II – Question 1

b) Indicate the minimum number of principal components required to explain 95% of the variation in the financial variable portion of the data

c) The number of components you chose in (b) explain at least 95% of the total variation in the financial variables. Check if any of the financial variables has significantly less of its variation explained than the other variables.

Part II – Question 1

d) Describe the characteristics a county with a large positive first principal component is likely to have.

e) Construct a graph of the first two principal components that identifies the counties.

f) Why would multicollinearity not be a problem if we replaced the original predictor variables with the principal components instead?

Part II – Question 2

It is desired to conduct a factor analysis on 1=Birth to 8=Urban, 9=Income, 15=MoblHms, 19=Unemp, 20=Coll4, and 22=HSGrad.

a) Identify the appropriate number of factors needed in order to get a model that is parsimonious, fits, and has enough large factor loadings when using the varimax rotation. Justify your answer.

Part II – Question 2

b) Identify the practically significant loadings for each factor of the model you chose in (a). Briefly give a feeling for what each factor seems to be representing.

c) Identify which variable has the most variance explained by the common factors in the model and which has the least.

Part II – Question 2

d) Indicate the correlation coefficient between Income and your factor 1. What percentage of the variation in income is explained by your factor 1?

e) If you had to give summary values to the counties based on the underlying factors you discovered, which do you think would work best for this particular data set: factor scores, single representative question, or sum score. Justify your answer.

Part II – Question 2

f) There is something (a statistical reason) that should cause you to have little faith in the results of the factor analysis you conducted in 2. What is it?