STATS 415, Homework 1

Due Thursday Jan 18, 2018

Turn in a printout of your homework in the lecture or in your GSI's mailbox across room 305A West Hall, no later than 5pm same day,

1. Consider the students of Stats 415 in W'18 as your sample of interest.

   (a) Name one variable **related to age** you could collect or measure on this sample in each of the following categories: a categorical(nominal) variable, an ordinal variable, an interval variable, and a ratio variable.

   (b) Name a population about which we could plausibly make inferences on the variables you listed, based on the data collected from this sample.

   (c) Name a population about which we could not make valid inferences on the variables based on the data collected from this sample.

2. Consider a document-term matrix, where f_{ij} is the frequency of the jth word (term) in the ith document and n is the number of documents. Consider the variable transformation that is defined by

$$f_{ij}^{*} = f_{ij} \cdot \log \frac{n}{g_j},$$

   where g_j is the number of documents in which the jth term appears and is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation.

   (a) What is the effect of this transformation if a term occurs in one document? In every document?

   (b) What might be the purpose of this transformation?

3. This exercise relates to the `College` data set, which can be found in the file `College.csv`on the book's website. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- `Private` : Public/private indicator
- `Apps` : Number of applications received
- `Accept` : Number of applicants accepted
- `Enroll` : Number of new students enrolled
- `Top10perc` : New students from top 10% of high school class
- `Top25perc` : New students from top 25% of high school class
- `F.Undergrad` : Number of full-time undergraduates
- `P.Undergrad` : Number of part-time undergraduates
- `Outstate` : Out-of-state tuition
- `Room.Board` : Room and board costs
- `Books` : Estimated book costs
- `Personal` : Estimated personal spending
- `PhD` : Percent of faculty with Ph.D.'s
- `Terminal` : Percent of faculty with terminal degree
- `S.F.Ratio` : Student/faculty ratio
- `perc.alumni` : Percent of alumni who donate
- `Expend` : Instructional expenditure per student
- `Grad.Rate` : Graduation rate

Perform exploratory data analysis of this dataset and report your results. Comment on any interesting or significant features. You can do any exploration you like as long as your final report includes

- some numerical summaries for each variable
- some multivariate numerical summaries (e.g. pairwise correlations)
- some graphical summaries for each variable (at least one with boxplots and at least one with histograms)
- some multivariate graphical summaries (at least one with pairwise scatter plots, and at least one with side-by-side boxplots of different groups).

Pay special attention to make sure you are using appropriate summaries for different types of variables; for example, do not compute

the mean of a categorical variable, even if its values are coded with numbers

Please limit the data exploration part of your report to 6 pages. Since the first two questions in this homework do not include much math, please included typed answers to them in the same submission.

Some useful functions and exploration ideas to start from:

(a) Use the `read.csv()` function to read the data into R. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data.

(b) Look at the data using the `fix()` function. The first column is just the name of each university. Try the following commands:

```
> rownames(college) = college[,1]
> fix(college)
```

You should see that there is now a `row.names` column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column of names in the data matrix before performing numerical operations on it. Try

```
> college = college[,-1]
> fix(college)
```

Now you should see that the first data column is `Private`, and another column labeled `row.names` now appears before the `Private` column. However, this is not a data matrix column, but the name that R is giving to each row.

(c) Numerical summaries for all variables can be easily obtained using the `summary()` function. However, you may want to use `as.factor` command first to tell R which variables are categorical.

(d) A scatter plots matrix can be produced with the `pairs()` function. Recall you can select a subset of variables to plot; for instance, to plot the first 10 columns only of a data matrix A you

can apply the `pairs()` function to the matrix `A[,1:10]`. You can use options to plot different things on the diagonal of the scatter plot matrix, such as a histogram or simply the variable name.

(e) The `plot()` function will produce side-by-side boxplots if one of the variables is categorical. For example, you can make a side-by-side boxplot by plotting `Outstate` versus `Private`.

(f) You can create new variables by transforming original variables. For example, consider a new categorical variable, `Elite`, which divides universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

```
> Elite = rep(``No'', nrow(college))
> Elite[college$Top10perc >50] = ``Yes''
> Elite = as.factor(Elite)
> college = data.frame(college, Elite)
```

You can now use the `summary()` function to see how many elite universities there are and the `plot()` function to produce side-by-side boxplots of any variable of interest versus `Elite`.

(g) The `hist()` function produces histograms, and its options can be used to specify number of bins, etc. You may find the command `par(mfrow=c(2,3))` useful: it divides the print window into 6 regions (2 rows, 3 columns) so that multiple plots can be made in one figure window. You can modify the arguments to divide the screen in other ways.