STAT 740 – Exam 3 – Spring 2019

*Directions:*

* *This exam consists of five (5) questions that are not equally weighted; the number of points a problem is worth is indicated next to the item stem.*
* *Sub-parts may also be worth different point values, but these will be determined during grading.*
* *Each problem is on a different page for convenience.*
* *You should provide your answers in this file, in the space below each sub-part. It would be helpful if you would change the font color of the questions or answers to something else (e.g. red) to help with scoring, but this is optional.*
* *You may submit your answers in LaTeX form instead of using this file if you so choose.*
* *You should submit your answer file (e.g. this file) as well as a separate file containing your R code and any other necessary files (e.g. data files you have modified).*
* *If necessary, you may insert photographs of handwritten work in lieu of a typed answer. This should be rare.*
* *You should submit this exam by the deadline given in the email.*
* *Submissions should be made to the Dropbox on D2L.*
* *You may reference our course materials (e.g. the book, slides, notes, etc.), but you may not use outside resources (e.g. Google).*
* *If you have any questions, email me (*[whitakerdo@uwstout.edu](mailto:whitakerdo@uwstout.edu)*).*
* *The total exam is graded out of* ***27 points****.*

1. [5 points] (6.17a-c) The date in T6-8.dat were collected to test two psychological models of numerical cognition. Does the processing of numbers depend on the way the numbers are presented (words vs. Arabic digits)?

Thirty-two (32) subjects were required to make a series of quick numerical judgements about two numbers presented as either two number words (“two”, “four”) or two Arabic digits (“2”, “4”).

The subjects were asked to respond “same” is the two numbers had the same numerical parity (i.e. both numbers were even or both numbers were odd) and “different” if the two numbers had different parity (i.e. one number was even and the other was odd).

Half of the subjects were assigned a block of Arabic digit trials, followed by a block of number word trials. The other half received the blocks in the reverse order. Within each block, the order of “same” or “different” parity trials was randomized for each subject.

For each of the four combinations of parity and format, the median reaction times for correct responses were recorded for each subject:

* = the median reaction time for word format-different parity combination
* = the median reaction time for word format-same parity combination
* = the median reaction time for Arabic format-different parity combination
* = the median reaction time for Arabic format-same parity combination.

1. Test for treatment effects using a repeated measures design at the level.
2. Construct 95% simultaneous confidence intervals for the contrasts representing the number format effect, the parity type effect, and the interaction effect. Interpret the resulting intervals.
3. The absence of interaction supports the “M” model of numerical cognition, while the presence of interaction supports the “C and C” model of numerical cognition. Which model is supported in this experiment? [This requires no further calculations.]

Hint: The design of the study could be represented like this.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Parity | |
|  |  | Different | Same |
| Format | Words |  |  |
| Arabic Digits |  |  |

This may help with determining which means should be compared when establishing **,** the contrast matrix**.**

1. [5 points] (8.28) Survey data were collected as part of a study to assess options for enhancing food security through the sustainable use of natural resources in the Sikasso region of Mali (West Africa). A total of farmers were surveyed and observations on nine variables were recorded:

* = Family (total number of individuals in household)
* = DistRd (distance in kilometers to nearest passable road)
* = Cotton (hectares of cotton planted in year 2000)
* = Maize (hectares of maize planted in year 2000)
* = Sorg (hectares of sorghum planted in year 2000)
* = Millet (hectares of millet planted in year 2000)
* = Bull (total number of bullocks or draft animals)
* = Cattle (total)
* = Goats (total)

The data are in file T8-7.DAT.

1. Construct two-dimension scatterplots of Family versus DistRd, and DistRd versus Cattle. Remove any obvious outliers from the dataset.
2. Perform a principal component analysis using the correlation matrix. Determine the number of components to effectively summarize the variability. Use the proportion of variation explained and a scree plot to aid in your determination.
3. Interpret the first five principal components. Can you identify, for example, a “farm size” component? A, perhaps, “goats and distance to the road” component? (Note that the components may be difficult to interpret without deep context knowledge: that is okay.)
4. [5 points] In this problem you are asked to compare the rates of employment for creative jobs in several midwestern states. A description of this dataset follows (<https://www.ers.usda.gov/data-products/creative-class-county-codes.aspx>):

*The creative class thesis—that towns need to attract engineers, architects, artists, and people in other creative occupations to compete in today's economy—may be particularly relevant to rural communities, which tend to lose much of their talent when young adults leave. The ERS creative class codes indicate a county's share of population employed in occupations that require "thinking creatively." Variables used to construct the ERS creative class measure include number and percent employed in creative class occupations and a metro/nonmetro indicator for all counties, 1990, 2000, and 2007-11. A break-out of employment in the arts is included.*

The raw data is available at [also included with the exam for convenience]: <https://www.ers.usda.gov/webdocs/DataFiles/48675/creativeclass200711.csv?v=41775>

Key variables of interest are CCShare, the proportion of jobs within a county that are classified as requiring creative thinking, and BohShare, the proportion of jobs within a county that are classified as being “Bohemian”, i.e. employment in arts occupations. These values are available for each county within each state. Counties are classified as being either metro (metro03=1) or nonmetro, e.g. rural (metro03=0).

Considering only nonmetro counties in states in the Upper Midwest (WI, MN, ND, SD, and MI), determine if there is evidence that the average proportion of jobs that are classified as requiring creative thinking or are Bohemian differs among states. If you determine that there are differences, report which states differ.

Perform all steps of the analysis and justify your decisions. Be sure to account for the multivariate structure of the response vector (CCShare, BohShare).

1. [5 points] The description of the dataset for this problem follows (<https://catalog.data.gov/dataset/2009-vha-facility-quality-and-safety-report-hospital-settings>):

*The 2008 Hospital Report Card was mandated by the FY08 Appropriations Act, and focused on Congressionally-mandated metrics applicable to general patient populations. The 2009 VHA Facility Quality and Safety Report report, not required by Congress, shifts to Veteran-centered metrics, and includes information related to infrastructure, care provided in outpatient and hospital settings, quality of care within given patient populations, accreditation status, patient satisfaction and patient outcomes for FY2008. The data in this report have been compiled from multiple sources throughout VHA. This dataset includes adjusted mortality rate for three defined populations: Pneumonia, Congestive Heart Failure, and Acute Myocardial Infarction, Nosocomisal Infections, Percent of patients on prophalaxis for deep vein thrombosis and Observed minus expected length of stay.*

The dataset is also included with the exam: VHA\_Facility\_Quality\_and\_Safety\_Report\_Hospital\_Settings.csv

Additional (*optional*) information about this report is available here: <https://www.va.gov/health/docs/HospitalReportCard2009.pdf>

Your task is to determine if a reduction in the number of variables is advisable. Perform all steps of the analysis and justify your decisions. Note the differences between multiple approaches you use, if applicable.

1. [7 points] **Note: No data analysis is needed for this problem.**   
     
   The description of the dataset for this problem follows (<https://catalog.data.gov/dataset/age-adjusted-death-rates-for-the-top-10-leading-causes-of-death-united-states-2013>):

*This dataset presents the age-adjusted death rates for the 10 leading causes of death in the United States beginning in 1999. Data are based on information from all resident death certificates filed in the 50 states and the District of Columbia using demographic and medical characteristics. Age-adjusted death rates (per 100,000 population) are based on the 2000 U.S. standard population. Populations used for computing death rates after 2010 are postcensal estimates based on the 2010 census, estimated as of July 1, 2010. Rates for census years are based on populations enumerated in the corresponding censuses. Rates for non-census years before 2010 are revised using updated intercensal population estimates and may differ from rates previously published. Causes of death classified by the International Classification of Diseases, Tenth Revision (ICD-10) are ranked according to the number of deaths assigned to rankable causes. Cause of death statistics are based on the underlying cause of death.*

The data are available in the file NCHS\_-\_Leading\_Causes\_of\_Death\_\_United\_States.csv, though you do not necessarily need to import this dataset into R. A random sample of 10 rows is reproduced below for your convenience. **This may be enough context to answer the questions**.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Row** | **Year** | **113 Cause Name** | **Cause Name** | **State** | **Deaths** | **Age-adjusted Death Rate** |
| 3737 | 2002 | Cerebrovascular diseases (I60-I69) | Stroke | United States | 162672 | 57.2 |
| 4920 | 2008 | Chronic liver disease and cirrhosis (K70,K73-K74) | Chronic liver disease and cirrhosis | New Mexico | 392 | 18.8 |
| 5793 | 2008 | Chronic lower respiratory diseases (J40-J47) | CLRD | Maryland | 1982 | 35.8 |
| 5873 | 2009 | Chronic lower respiratory diseases (J40-J47) | CLRD | Washington | 2933 | 44.3 |
| 6287 | 2000 | Diabetes mellitus (E10-E14) | Diabetes | Vermont | 163 | 26.1 |
| 6825 | 2011 | Diabetes mellitus (E10-E14) | Diabetes | Idaho | 393 | 24.2 |
| 8311 | 2005 | Essential hypertension and hypertensive renal disease (I10,I12,I15) | Essential hypertension and hypertensive renal disease | Tennessee | 589 | 10.1 |
| 9520 | 2011 | Influenza and pneumonia (J09-J18) | Influenza and pneumonia | Wisconsin | 986 | 14.2 |
| 10677 | 2000 | Malignant neoplasms (C00-C97) | Cancer | Kansas | 5223 | 187 |
| 12263 | 2013 | Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) | Kidney Disease | Tennessee | 1068 | 14.8 |

You may assume that you can reasonably clean the data (e.g. excluding row 3737 for a state-level analysis) and change the way it is organized (e.g. creating a column for each Cause Name using the reshape2 package).

1. Would it be appropriate to use this dataset to compare states using MANOVA (or ANOVA) procedures? Why or why not? (Describe any assumptions you would have to make to analyze the data or assumptions that would be violated.)
2. Would it be appropriate to use this dataset to conduct a principal components analysis to determine if a few linear combinations of causes could explain a large proportion of the variation? Why or why not? (Describe any assumptions you would have to make to analyze the data or assumptions that would be violated.)
3. Would it be appropriate to use this dataset to conduct a factor analysis to determine if a few underlying factors explain the variability in the deaths or death rates by underlying cause of death? That is, would it be appropriate to attempt to construct factors based on sets of causes of death? Why or why not? (Describe any assumptions you would have to make to analyze the data or assumptions that would be violated.)
4. What information available in this dataset is not accounted for directly by the analysis types above? What is the effect of ignoring this information on analyses?

Hint: For the above questions, you may not feel that you can give a definitive answer – that’s okay. Think about the assumptions/conditions and articulate your thinking about what you do know and what you would want to know more about. This question is *not* about trying to anticipate what the results of an analysis might be but instead to get you to think about what types of analyses are more appropriate than others for a new, unfamiliar dataset.