

11.220: Quantitative Reasoning and Statistical Methods for Planning I

Test-out exam: January 2005 (Computer Lab Portion)

Name: _____ MIT E-mail address: _____@mit.edu

Academic advisor's name: _____

Did you attend the brush-up? _____ All _____ Some _____ None

Instructions:

1. Relax. Breathe. We recommend that you look through the exam and start with those problems you feel most comfortable answerin
2. You are allowed to use up to two textbooks of your choice, your own notes, plus any online 11.220 notes and online software manuals during the exam. Please note that you may not share these materials with other students.
3. For each problem, show your work wherever possible. Partial credit will be given if we have evidence that you framed the problem correctly and/or were headed in the right direction to obtain the correct answer. Points will be deducted if you are asked to show your work for a problem and fail to do so. You may use the back of any page to complete your work. In all cases, indicate clearly what your (single) final answer is for each question and make sure that your name is readable on any printouts and answer sheets that you turn in.
4. Your answers should be submitted to the staff as hard-copy at the end of the exam. You are welcome to assemble your answers in the online editor of your choice. You can also print out the PDF version of the exam and write your textual answers on the printed test. However, you will also need to submit printed graphics and maps. Make sure your name is written on each page of work that you submit. In addition to the textual answers, you are asked to submit 1 boxplot (#3A), 1 histogram (#3B), 1 scattergram (#3D), and 1 map (#4C).
5. Remember to keep your writing clear, tight, and to the point. Think before you write. You do not need more than a few well-written sentences to answer any of the questions on the exam. (This will save you time!)
6. Point values for each question are included in [brackets] next to each question on the exam: there are 100 total possible points.
7. During the exam, please let us know if you find unclear printing, typing, or errors on the exam. We will not give any hints about how to answer a question. If you think a question is unclear, clearly state your assumptions on the exam and complete the question to the best of your ability.
8. The exams will be scored and returned to your mailboxes in the student common room. You will be notified *via* email whether you have placed out of 11.220.

Please read this statement and sign your name below.

I certify that I have neither given to nor received assistance from another person on this exam, and that I have used only simple arithmetic functions on my calculator.

Signature: _____

11.220: Quantitative Reasoning and Statistical Methods for Planning Testout – January 28, 2005 – Computing Portion

This part of the 11.220 testout uses data about the 2004 US Presidential Election.¹ The data include vote counts for Bush, Kerry, and Nader broken down by County and stored as attributes of a shapefile that can be viewed using ArcMap. The data also include a spreadsheet with county-level demographic, socio-economic, and environmental counts and factors. You will be asked to examine, create, and interpret descriptive statistics and thematic maps of the election results. You will also compute and interpret several measures of association between the election results and other county characteristics. Finally, you will use another dataset of major US cities to identify which counties contain major cities and compare their election results with those of the other counties.

The GIS files and datasets needed for this testout exam are available in the 11.220 class locker. They include:

- election04_county.shp** - a 'shapefile' of US Counties and Election 2004 vote counts
- election04_data_220s05.xls** - a spreadsheet with additional county-level data
- election04_data_220s05.mdb** - an MS-access database with the same data as the spreadsheet
- UScities.shp** - the ESRI sample 'shapefile' of major US cities
- 11.220_testout04_start_Z.mxd** - an ArcMap document that maps some of the election04 data

These data are accessible from any WinAthena PC in the class AFS locker. The shapefiles, spreadsheet, and access database are in Z:\athena.mit.edu\course\11\11.220\data\testout05\testout05_data and the ArcMap document is in Z:\athena.mit.edu\course\11\11.220\data\testout05. (We put the '_Z' at the end of the name of this ArcMap document to remind you that it accesses the two shapefiles from drive Z. Since the basemaps are read-only, you can leave it that way.)

*First, copy the whole 'Z:\athena.mit.edu\course\11\11.220\data\testout05' directory into C:\temp or some other scratch space on your local PC. Do your analyses using these locally stored datasets. By using the local copies you will speed up processing and avoid file sharing conflicts with your colleagues. In order to answer the following questions, the WinAthena machines in the computing lab have available for your use the following software packages: Excel, ArcMap, MS-Access, and SPSS. When you open the ArcMap document, you will see, in the Data Frame, three thematic maps using **UScities.shp** and the **election04_county.shp** shapefile. The top layer identifies major US cities (in the 48 contiguous states). The second layer is a thematic map shading the number of votes Bush received in each county (in the 48 contiguous states) using a red-to-blue color scale with 5 categories and 'natural break' classification. The bottom thematic map shades the same attribute field (Bush votes) using quantile classification. All the map layers are projected using a North_America_Albers_Equal_Area_Conic projection.*

[10 points] Question 1: (2 points each) – Some Election Results

- What is the maximum number of votes received by Ralph Nader in any one **county**:

- What is the state, county name, and FIPS code for that **county**? _____

- How many counties are there in Florida? _____

¹ The Election 2004 data were assembled and made available online by Anthony Robinson, an RA at the GeoVISTA Center at Penn State. See: <http://www.personal.psu.edu/users/a/c/acr181/election.html>

- What are the mean and standard deviation of the percentage of votes Bush received among Florida counties? Mean _____ standard deviation _____
- What is the largest number of votes received by George Bush in any one of the **48 contiguous US states**? _____ ? Which State: _____

[15 points] Question 2: Interpreting the Red and Blue Maps

Part 2A (4 points): In the ArcMap document, the 'Bad Election Map A' map looks strange. It shades votes for Bush using a natural-break classification. We all know that news reports tend to use red colors where Republicans won and blue colors where Democrats won. This map does use red for the counties where Bush wins the most votes and blue where he receives the least. Yet the map is mostly blue even though Bush won the election. Explain briefly why this is the case.

Part 2B (5 points): The 'Bad Election Map B' (using quantiles) isn't so blue, but it still looks strange. For example, many counties along the NorthEast and West coast are red even though Bush lost the states along these coasts. Explain briefly why there are lots of red counties in these areas even though Bush lost those states.

Part 2C (7 points): Explain briefly your choice of attribute field, symbology, and classification choice in order to display a map that presents a better indication of the geographic pattern of the voting results. (You do not need to turn in a map at this point - just explain what you would do and why.)

[30 points] Question 3: Associating Election Results and Poverty

Part 3A (2 points + 3 points): Prepare a boxplot for the county population density using the **pop00sqmil** variable. The boxplot looks strange. What is going on?

Part 3B (4 points): Prepare a histogram for the county population density using the **pop00sqmil** variable – after **excluding** those counties with **pop2000=0** and counties with densities *greater than or equal to* 750 people per square mile.

Part 3C (2+2+2 points): What are the mean _____ and standard deviation _____ of **pop00sqmil** among those counties with **pop2000 > 0** and with densities less than 750 persons per square mile? How many counties did you include in your computations? _____

Part 3D (4+4+3+4 points): Plot a scattergram showing the percentage of votes for Bush in each county in *California* vs. **pop00sqmil**. Be sure to **exclude** all those counties not in California or *with pop2000=0* or with population densities *greater than or equal to* 750 persons per square mile. What is the value of Pearson’s correlation between **bush_pct** and **pop00sqmil**? _____. Is the correlation between these two variables significant at the 0.05 level? _____. Briefly explain your reasons for saying ‘yes’ or ‘no’: _____

[20 points] Question 4: Big City Effects

Make visible the ‘USCities’ layer in the ArcMap document, ‘11.220_testout04_start_Z.mxd’ and restrict the cities so that only those larger cities in the continental US (i.e., the 48 contiguous states without Hawaii and Alaska) with 1990 population greater than 200,000 are shown.

Part 4A (3 points): How many cities in the continental US have 1990 population greater than 200,000? _____

Part 4B (5 points): Use ArcGIS to create a 50 kilometer buffer around these major cities. How many counties intersect these buffer zones surrounding the big cities? _____. Hint: Use ArcGIS’s ‘Select By Location...’ tool.)

Part 4C (12 points): Turn in a map that shows these buffer zones on top of a thematic map shading the counties in proportion to percentage of its population that is poor. (Use 5 categories and quantile classification.) Also, highlight those counties that overlap the big-city buffers. Be sure that the highlighted counties are visible on the map you turn in. To make the map more readable, zoom in to the Southeast portion of the United States – viz. show all of Florida and a few states north and west of Florida. (Note: In order to map **pctpoor**, you’ll need to join the county shapefile to the data table from the spreadsheet or MS-Access files.)

[25 points] Question 5: Testing for Significant Differences

Part 5A (8 points, 2 points each): Once again, select those counties that intersect the big-city buffers. What are the mean _____ and standard deviation _____ of the Bush percentages for these counties? What about for the counties outside the big-city buffers? Outside mean=_____ and outside standards deviation=_____ (In each case, be sure to exclude any counties that do not have any votes: that is, **Total=0**.)

Part 5B (5 points): Both means in Part 5A are **higher** than the overall percentage (about 51.5%) of all votes that were for Bush. How can this be the case? Explain briefly what is going on.

Part 5C (12 points): Use your answers to Part 5A to estimate the standard error for the difference in subgroup means of the Bush percentages for counties in and out of the buffer. Use the results to test the statistical significance of the observed difference in Bush percentages. Show your work. What can you conclude? That is, comment briefly on your interpretation of these results – in what way, if at all, does proximity to big cities seem to influence the Bush vote?

Note: When considering sampling variability for these data, focus on the simple case of differences among the county proportions for the share of votes going to Bush in each county (without trying to aggregate individual counts from the underlying vote tallies.)