Data you generate or collect for your project

P2

(revised from Jim Noble, In-Thinking Site)

In order to carry out your project you will need to have some information, measurements or data to help you do it. Once chosen, you then need to be very specific about what that information will be involved your situation. You cannot use data from another class or another project. Before you make decisions, you should consider the following:

- **At least a portion of your data should be quatitative -** categorical fields are important as well but not enough on their own. (Hair color and gender are both only categorical. Try to include at least include some other numerical data to complement it).

- **It is good, but not required, to use your information to generate new information** (for example, you could collect #of crimes in a state and the population of a state. Perhaps then you could calculate #of crimes per capita which could be one of the main variables of your project)

Primary Data (Data you personally collect)

Data from questionnaires, results of some experiments, measurements, or calculations, etc

The major advantage of collecting primary data for projects is that it gives you total control over what information you will have. Primary data is information that you personally collect (or measure, etc). You do not have to worry about whether you will find what you want. This allows you to formulate your ideas and then decide what information you need to collect in order to investigate your topic. Collecting your own data can be somewhat time consuming but be very rewarding and interesting.

A. Questionnaires

If you plan a survey, keep in mind:

- ✓ You can't go back and add/change questions once you have done it.

- ✓ Where possible, surveys should involve some kind of measurement rather than opinion. If not possible, so be it.

- ✓ Information collected experimentally or with surveys needs to be collected under defined and consistent conditions and you would be required to describe these conditions. On your project you will have to include details of how you attempted to collect from a random sample of respondents which will lend to an unbiased data set.

    - **Technology -** data collection tools such as 'Google forms' and 'Survey Monkey' take a huge amount of the legwork out of the collection and organization of responses. You still have to make the key strategic decisions relating to the questions and how they are answered. The process of making the questionnaires helps with these decisions too because you have to decide how the question will be answered in advance. You can complete a questionnaire and then man a station with a one or more computers during breaks and lunchtimes while students fill in and submit the answers. E-mail allows the questionnaire to reach

much further afield and helps question bigger samples. Once submitted the responses are automatically compiled in a spreadsheet.

- **Numerical responses -** As mentioned above, keep the balance of numerical and categorical data right and relevant to make sure they have enough numbers to perform analysis. You can do this by asking questions with numerical answers.

- **Test Your Questionnaire -** Spend time testing your questionnaire. It may take several drafts before you have the questionnaire you want. Ask others for feedback and suggestions. You might try a small sample of 10 students to get a preliminary idea about their idea. Then you can perhaps adjust the questions accordingly.

B. Experiments and measurements. *Sample ideas:*

- **human proportions**, when one measures different dimensions of the human body on a sample of students
- **human testing**, where different students are measured and timed doing different activities from running and jumping, to weight lifting or even the time it takes to complete different puzzles

There are many more experimental ideas that can lead to statistical or modeling projects. The key issues with this type of information collection are:

Secondary Data

Using secondary data can have the advantage both of saving huge amounts of time and allowing for considerably larger data sets from broader sources that can lead to more interesting investigations, conclusions and interpretations. The Internet is an incredible source of such information and potentially a major tool. Many projects would simply not be possible by using primary data only and so the use of secondary data does really open up possibilities that may help you to find an idea you are really interested in.

If you find data on the internet, hopefully it can be easily transferred into a spreadsheet for huge time saving and flexibility benefits. This may be possible by copying and pasting into a spreadsheet program like Excel or Sheets.

Sources

Sources have to, of course, be cited and referenced. Equally, you should spend some time questioning the reliability of the chosen source buy examining how and when the data was collected and what the associated limitations might be. This can make for interesting discussion in your project and should not necessarily stop you using the source. One of the biggest problems occurs when one uses multiple sources to build a single data set. The conditions under which the data was originally collected become increasingly relevant if the two data sets are to be considered together. Again, this should be looked at and written about, without stopping you using the multiple sources. The other associated snag with multiple sources is that you can only really use the field headings that the two sources have in common and so this should be taken into account when considering the different sources. Ultimately the choice of different sources used can affect how much time is actually saved by using secondary data!

A further warning on secondary data is that it is often difficult to find 'Raw Data' that is published. You are much more likely to find somebody's summary statistics that hint at the existence of the raw data you are looking for. Be aware of this distinction.

Combination of Primary and Secondary Data

Combining could lead to a project with possibilities for rich interpretations.

- *Collect primary data and find a suitable secondary data source with which to compare your findings*
- *Use either the primary data or secondary data to make calculations that generate new field headings.*

Sufficient Quality & Quantity

Data collection is crucial to a good project. If the information is not sufficient in ***quality*** or ***quantity***, then the rest of the project will probably suffer accordingly. The definition of these two words will vary depending on the projects in question and so it is hard to give a hard and fast guideline on what is sufficient. What the two words also imply though is the important consideration of the nature of the information. It is very difficult (and not appropriate) to perform mathematical analysis without enough data and equally difficult to search for interesting conclusions without having multi-variable and rich data.

Quantity

Sample size ---How many measurements are needed? There is a nominal figure of 50 as a minimum for scatter plots (and correlation and LSRL) but that can be limiting, depending on the nature of the project. Running a single Chi-square test of independence would likely require a 100 pieces of data even for the smallest of contingency tables (2 by 2). You most likely want to target at least 100 to 200 pieces of data or even more in some cases. Moral of the story…always shoot for more data than you think you might need.

Quality

- **Avoid having "limiting" data –** If you are going to investigate winning percentage in the NBA, you would not want to only include data from a single year.
- **Must not be one dimensional** (ie temperature change over the last 50 years, lots of numbers but they are all temperatures).

Comparisons

More often than not statistics have greater meaning when compared to the same statistic for a different data set. For example, it is always interesting to compare the average cancer rates in a certain region to the world average or those results from previous years in the same region.

Perhaps you can look for ways to categorize your information so that such comparisons can be made and so this implies thought before collection! This also helps you make interesting interpretations.

Some past students have struck a balance between categorical and quantitative data. They made sure that there is a categorical component to their numerical data. This allows interesting ways so that the associated

numerical data can be compared accordingly. For example, someone might love baseball and is looking at factors that could possibly influence winning percent

Modeling

There are numerous possibilities for projects on modeling, where information collected is modeled by a function, examples include,

- progress of athletic records - exponential growth and decay
- paths of projectiles - quadratics
- temperature fluctuations - sine waves

These can make for some very interesting projects, but one should, again, focus on making sure they have a systematic approach to collecting the information and their one's choice of models. Below is an example of what is meant based on the ideas above.

- *When looking at athletic events, how do you choose the events to look at? Would the 100m, 200m, 400m, 800m, 1500m be more interesting to look at as a sequence than a random collection of events.*
- *Consider the quadratic path of a rugby ball being kicked between the posts from different parts of the pitch. How are the positions chosen? What are the common conditions? How can one introduce a systematic sequence in to the way they collect such information?*
- *When comparing the temperature fluctuations of different cities, how does one choose the cities? What is the logical progression from one comparison to the next?*

With such considerations, these projects can reach some lofty heights and it can be difficult, but it is important that one get as much potential from the information collection as possible even if that potential is not eventually realized.

A Mishmash of Project Ideas and data sources

Existing Data Sources and Project Ideas General Data Sites:

- http://www.cdc.gov/DataStatistics/ CDC
- http://unstats.un.org/unsd/databases.htm United Nations Statistics Division (UNSD)
- http://www.clinicalcorrelations.org/ Clinical Correlations
- http://faostat.fao.org/ Food and Agriculture Organization of the United Nations (FAOSTAT)
- http://www.who.int/research/en/ World Health Organization of the United Nations (disease and healthcare)
- http://www.fedstats.gov/ Federal Government Statistics
- http://www.eeps.com/zoo/index.html Data Zoo
- http://www.dartmouth.edu/~chance/teaching_aids/data.html
- DASL The Dataset and Story Library - a collection of datasets and related documentation (stories)
- http://www.keypress.com/x2814.xml - Fathom Data Sites
- http://www.keypress.com/x3894.xml - Fathom Data Sites
- http://www.census.gov/main/www/a2z/ U.S. Census Bureau A to Z:
- http://www.census.gov/compendia/statab/ us stat abstract:
- http://www.census.gov/population/international/ world population
- http://www.census.gov/population/www/popclockus.html U.S. Population:

Environment /Climate / Energy / Engineering Data Sources

a) Which are more accurate for a particular chosen city: the predicted high or predicted low temperatures?
b) the rainfall one year correlated positively, negatively, or not at all with the rainfall the previous year?
c) Is wind energy effective and practical?
d) Are electric vehicles helping to save the U.S. from importing oil?
e) Global warming – summer of 2011 is described as the second hottest summer ever on record.
f) Are antibiotics being administered to our food supply causing germ resistance to antibiotics in people?

- http://www.ncdc.noaa.gov/oa/ncdc.html National Center for Climatic Data (NCDC)
- http://www.noaa.gov/index.html National Oceanic Atmospheric Administration (NOAA)
- http://www.epa.gov/enviro/html/ef_overview.html U.S. Environmental Protection Agency: (EPA)
- http://www.nrel.gov/rredc/ National Renewable Energy Laboratory: (NREL)
- http://www.afdc.energy.gov/afdc/ Alternative Fuels and Advanced Vehicles Data Center
- http://www.unep.org/ United Nations Environment Program (UNEP)
- http://www.ipcc.ch/ Intergovernmental Panel On Climate Change (IPCC)
- http://www.nhc.noaa.gov/ National Hurricane Center
- http://www.weather.gov/ National Weather Service
- http://swera.unep.net/index.php?id=7 Data for Solar and Wind Renewable Energy (SWERA)
- http://www.seattlecentral.edu/qelp/index.html Quantitative Environmental Learning Project (QELP)

Medicine & Public Health

Epidemiology is the study of health patterns. It is cornerstone method of public health research. PH research helps governments determine policy to help people become / stay healthier, and prevent disease. Method is identifying risk factors and trying to find associations between factors, or cause. Epidemiologists work on designing studies and collect data for statistical analysis. Outbreak investigation, disease surveillance, screening, bio monitoring, and conducting clinical trials are all public health functions.

1. Is there a correlation between what state a person lives in and their risk for getting cancer?
2. Is there a correlation between the number of TV sets per home and obesity rates?
3. What day of the week are there more deaths recorded?
4. Do states with lower speed limits have fewer fatalities?
5. Towns with High schools that have earlier starting times report that students have more accidents than students attending a school with a later starting time. Can you find data to verify this?
6. Do more car accidents happen on one particular day of the year such as the first day of Standard time after going off Daylight Savings time?
7. is there a correlation between malaria and sickle cell anemia
8. is there correlation / causation between root canal and cancer?
9. Do Vaccines cause Autism?
10. Is There a Correlation Between Autism Rates and vaccinations?
11. Relationship Between Maternal Age and Incidences of Down Syndrome
12. Caffiene and diabetes
13. is there a link between breast cancer and abortion
14. is there correlation between lymphoma and depression
15. is there correlation between gmos and cancer

16. is there a correlation between pesticides and cancer
17. is there correlation between marrow and spongy bone
18. is there correlation between stem cells and leukemia
19. is there correlation between stem cells lymphoma
20. is there a correlation between the complexity of an organism and the number of chromosomes
21. is there a correlation between spinal cord damage and paralysis?
22. is there a correlation between cord blood and stem cells
23. is there a correlation between pesticides and IQ?
24. is there a correlation between fast food and obesity
25. Is There a Correlation Between Autism Rates and Family Income Level?
26. Is is there a correlation between malaria and sickle cell anemia
27. there a correlation between daylight savings time (DST) and increased car accidents?
28. is there correlation / causation between caffeine and diabetes?
29. Is there a correlation / causation between taking multivitamins and benefits to long-term health?
30. Is there correlation / causation between dark chocolate and positive health outcomes?
31. Is there correlation between bad weather and arthritis?
32. Is there a correlation between getting tattoos and a skin or staph infection or other ill effect?
33. Do sales of large sugary drinks of 16 oz. Or more promote obesity?
34. Does nationwide morbidity increase in July when medical interns are placed in new positions at hospitals?
35. Is there a correlation between 3-D imaging (instead of digital imaging) and increased cancer detection rates and decreases false positives?

Medicine & Public Health-continued

36. Is there a correlation between inhaled insulin (instead of injectable) and improved diabetes symptoms?
37. Is there a correlation between excessive coffee/caffeine intake and irregular hearbeat? (atrial fibrillation)
38. Is there correlation between e-cigarette use and smoking cessation?
39. Is there a correlation between running and knee arthritis?
40. Is there a correlation between artificial sweeteners (instead of sugar) and positive health outcomes?
41. Is there correlation with eating read meat and negative health outcomes?

- www.cdc.gov/nchs/ CDC - National Center for Health Statistics Homepage ---
- http://www.cdc.gov/az/ Centers for Disease Control topics A to Z (CDC)
- http://www.cdc.gov/DataStatistics/ CDC Data & Statistics
- http://www.americashealthrankings.org/ America's Health Rankings
- https://dawninfo.samhsa.gov/default.asp Drug Awareness Warning Network (DAWN):
- http://progressreport.cancer.gov/ Cancer Trends Progress Report:
- http://www.who.int/phe/health_topics/en World Health Organization (WHO) ---
- http://www.nih.gov/ National Institute of Health (NIH)
- http://www.cancer.gov/aboutnci/cis/page1 National Cancer Institute (NCI)
- http://www.ncbi.nlm.nih.gov/pubmed PubMed.Gov ---
- http://phpartners.org/health_stats.html Health Data Tools and Statistics:
- http://mchb.hrsa.gov/mchirc/chusa/ U.S. Child Health Statistics:

Political and Social Science / Economics / Psychology /Education

1. Is there correlation between Kindness and intelligence?
2. is there a correlation between crime and poverty?
3. Is there correlation / causation between intelligence and happiness
4. is there a correlation between intelligence and depression
5. is there correlation between marijuana and schizophrenia
6. Do Exit polls accurately predict winners of elections?
7. Has capital punishment reduced the amount of violent crime in Tx. or other state?
8. Do normally patient people become impatient behind the wheel?
9. Is there correlation / causation between gender and schizophrenia?
10. Is there a link between smoking and intelligence?
11. Is there correlation / causation between gender and elevated stress levels?
12. What are effects of video games on growth / development of individuals and society?
13. Is there correlation / causation between school uniforms and higher achievement levels?
14. is there correlation between marijuana and criminal activity
15. is there correlation between marijuana use and mental illness
16. is there correlation between legalized prostitution and decreased sexual violence?
17. is there correlation between social media and grades
18. Is there a correlation between marriage and longer life expectancy?
19. is there a correlation between mass shootings and population
20. is there a correlation between mass shootings and time of year
21. is there a correlation between mass shootings and age of shooter?

22. is there a correlation between mass shootings and decade?
23. Is there a correlation between IQ and political party
24. Is there a correlation between IQ and SAT scores?
25. is there a correlation between education and political party
26. is there a correlation between intelligence and political affiliation
27. is there a correlation between gun ownership and crime
28. is there a correlation between deficit and national debt
29. is there a correlation between federal budget deficit and trade deficit
30. Religion and IQ: is there a correlation?
31. Are suicide rates higher for certain Religions?

o http://pewresearch.org/
o http://www.gallup.com/poll/election.aspx
o http://fivethirtyeight.blogs.nytimes.com/
o http://www.hivaidssurveillancedb.org/HIVDB/ Aids Surveillance Data Base:
o http://www.apa.org/topics/obesity/index.aspx American Psychological Association
o http://www.esa.doc.gov/about-economic-indicators U.S.EconomicsStatistics Association
o http://www.bls.gov/data/ U.S. Department of Labor Bureau of Labor Statistics
o http://research.stlouisfed.org/fred2/ Federal Reserve Economic Data (FRED)

Political and Social Science / Economics / Psychology /Education….CONTINUED

- http://www.econdata.net/ EconData.net

o http://www.fedstats.gov/ Federal Statistics (FEDSTATA)
o http://www.brillig.com/debt_clock/ National Debt Clock and Information
o http://www.worldbank.org/ World Bank:
o http://www.imf.org/external/data.htm International Monetary Fund Data and Statistics
o https://www.cia.gov/library/publications/the-world-factbook/ CIA World Fact Book
o http://Fivethirtyeight.com http://fivethirtyeight.blogs.nytimes.com/ **FiveThirtyEight** is a polling aggregation
o http://dawninfo.samhsa.gov/ Drug Awareness Warning Network (DAWN)
o http://polisci.lsa.umich.edu/grad/comparative/data.htm
o http://www.icpsr.umich.edu/icpsrweb/ICPSR/ InterUniversity Consortium for Political and Social Research:
o http://politicaldata.com/Pages/Index.aspx Political Data Inc.
o http://einstein.library.emory.edu/international_socsci.html International Statistical and Electoral Resources
o

Sports / Entertainment Data Sources

a) http://www.stats.com/ Sports Data:
b) http://baseball1.com/ website features baseball statistics -beginning 1871. Includes player salary data
c) http://www.cbssports.com/mlb/stats CBS Sports Data: - fathom
d) http://sabr.org/ Saber metrics

e) NFL Quarterback Passer Rating Data http://www.dartmouth.edu/~chance/teaching_aids/data/NCAA.html
f) http://www.usatoday.com/life/television/nielsen.htm Neilson Television Ratings:

Websites with Information about Global Warming

http://www.ipcc.ch

Intergovernmental Panel on Climate Change: 'WG I Climate Change 2001 The Scientific Basis' - Summary for Policymakers and Technical Report
http://archive.greenpeace.org/~climate/science/reports/fossil.pdf

Greenpeace Report: 'Fossil Fuels and Climate Protection: Carbon Logic'
http://www.meto.govt.uk/research/hadleycentre/pubs/brochures/B2000/index.html

Hadley Centre: 'COP8 - Climate Change'
http://www.giss.nasa.gov/research/intro/hansen_04

NASA Goddard Institute for Space Studies: 'A Common Sense Climate Index: Is Climate Changing Noticeably?' (with links to other information)
http://www.epa.gov/globalwarming

US Environmental Protection Agency: short on-line summary with links to more detailed reports.
http://uspirg.org/reports/flirtingwithdisaster01.pdf

US Public Interest Research Groups report: 'Flirting with Disaster'
http://www.ucsusa.org/index.html

Union of Concerned Scientists – Global Warming with links to information and charts
http://users.erols.com/dhoyt1/index.html

'Greenhouse Warming: Fact, Hypothesis, or Myth? - A look at the scientific evidence.'
http://www.climateark.org/vital

Climate Change Overview - Climate graphics with interpretations.
A Resource for Free-standing Mathematics Qualifications Global Warming
Ó The Nuffield Foundation 2
http://www.whrc.org/globalwarming/warmingearth.htm

US Woods Hole Research Centre 'The Warming of the Earth - A beginner's guide to understanding the issue of global warming'.
http://www.pbs.org/wgbh/warming

Nova & Frontline: 'What's Up with the Weather?' – On-line information including 'graphs tell the story'.
http://www.abd.org.uk

'Climate Change Truths' from the Association of British Drivers
http://www.accesstoenergy.com/ate/9711/ate_1197.pdf

Article in Access to Energy Newsletter
http://www.greeningearthsociety.org/Articles/2000/surface1.htm

Report from the Greening Earth Society
http://www.cato.org/pubs/pas/pa329a.pdf

Policy Analysis by Patrick J Michaels of the Cato Institute
http://www.cato.org/pubs/regulation/regv23n3/michaels.pdf

More Project ideas

2. What is the relationship between a person's weight classification in Rowing and their times on an ergometer (rowing machine)?
3. Does Racial Bias exist in North Carolina's traffic enforcement?
4. Is there a correlation between Music and math?
5. How Does Age Play a Role in the Tony Awards Winner for Best Actress in a Musical?
6. Caffeine Consumption And Sports Achievement
7. Does playing a musical instrument raise SAT scores?
8. Energy Drinks and Your GPA, is there a cause and effect?
9. Population Growth
10. A Comparison of the Agricultural Economy of a Developed and a Developing Country
11. The Resilience of the Baseball and Its Effects on Home Run Production
12. Is There a Relationship Between the Length of a Turbine Blade and Voltage?
13. Is home court/ field a better predictor for high school, college, or pro sports?
14. Is home court/field a better predictor for Football, Basketball, Baseball, Hockey?
15. Is there a correlation between gender and suicide rates?
16. Is there an association between music genre and higher SAT scores?
17. Is there a correlation between strict gun laws and homicide or mass shooting rates?
18. Is there a correlation between brain size and intelligence?
19. Is there correlation / causation between gas prices and obesity?
20. Is there correlation / causation between happiness and income?
21. Is there correlation between Height and pulse rate?
22. Is there a rise in climate change and tick-borne diseases?
23. Is an oral vaccine effective for cholera outbreaks?
24. is there a correlation between vitamin d deficiency and obesity?
25. Is there correlation / causation between zinc supplements and curing common cold?
26. Is their correlation / causation between medications and curing obesity?
27. Is correlation between preadmission testing for MSRA and decreased MRSA

infections after hospital stay?
28.Is there correlation between antidepressant use and increased risk of suicide?
29.Is there a correlation between prayer and improved- positive health outcomes?
30.Is there a correlation between rising global temperature / climate change and Antarctic sea ice melt?