PSU Problem Solving Process

1. Position
2. Sense
3. Uncover
4. Solve
5. Build
6. Achieve

Data

- How and where do you get it?
  - Sources and methods
- What do you do with it once you get it?
  - Analysis and manipulation
- How do you use it once you’ve analyzed it?
  - Presenting data so others can understand what you know so well!

BA 301 Winter 2011 Data Analysis Lecture
Research & Data Gathering

- Primary Research – gathering new data to answer a specific question
  - The BA 301 demographic survey
- Secondary Research – collecting data that already exists from a variety of sources
  - An existing survey on airline travel from Mintel
- Pros & cons?

Primary Research

- Observe
  - Experiments
  - Focus groups
- Ask questions
  - Questionnaires
    - Online, by mail, in person, on the phone
    - Focus Groups
    - Consumer Panels
    - Online Networks

Secondary Research

- Internal sources
  - Existing documents
    - Policies, quality reports, emails and memos
    - Production records, HR records, sales records, etc.
  - Corporate databases
- External sources
  - Commercial surveys
  - Online and internet-based

Research Strategy

- What is a research strategy?
  - A systematic plan for tracking down sources for your research topic.
- There are many options:
  - Databases with access to articles in magazines, newspapers and journals (specialized and general).
  - Web search engines (Google) to locate relevant web sites, online articles and government publications.
  - Books, using the library’s online catalog.
Pyramid of Sources

General Encyclopedias
General Interest Magazines & Newspapers
Specialized Magazines
Trade Books
Gov’t Documents
Scholarly Books
Academic Journals

Source: Ballenger, 74

Research Strategy

- But, honestly, where’s the first place you go?

- What percentage of the information on the “Web” is visible to search engines like Google?
  - Some industry people say as little as 1%

- Where is the rest of it?
  - The DEEP web – tens of thousands of terabytes (estimate)

The DEEP (Invisible) Web

- What is the DEEP web?
  - Databases that are hidden – firewalls, password protected areas, technical restrictions, etc.
  - What can you find there?
    - Government databases, medical and academic journals, personal records, professional publications, proprietary research and “peer-reviewed sources”.
  - Can you get to some of this great data?
    - Yes – through the PSU library web site.

What is “Peer Review”?

- Peer reviewing or “refereeing” is the process of submitting an article, a piece of scholarly work, research or ideas to the scrutiny of others who are recognized experts on the subject.

- Publications that have not undergone peer review may be regarded with suspicion by scholars and professionals.
The Visible Web

- Search Engines
- Directories
- Archives
- Government
- News sites
- Online Reference Sources

Search Engines

- These take your search terms and seek matches among millions of web pages. Some are better than others, but none search the entire web.
  - Google, Yahoo, Bing.com, Ask.com, Wolfram Alpha, Altavista, HotBot, AlltheWeb
  - Metasearch
    - Clusty, Dogpile, Mamma, Search.com, Vivisimo, SurfWa.
- Check <www.searchenginewatch.com> for reviews of search engines.

Directories

- Unlike bot-based search engines, directories have a human touch – information specialists arrange sites by topic.
  - Infomine <http://infomine.ucr.edu>
  - Librarian’s Index <http://www.lii.org>
  - Scout Report Archives <http://scout.wisc.edu/archives>
  - WWW Virtual Library <http://www.vlib.org>
  - Yahoo <http://www.yahoo.com>
  - Even Google <http://directory.google.com/>

Archives

- Archives contain the texts of poems, books, speeches, political cartoons and historically significant documents.
  - American Memory <http://memory.loc.gov>
  - Scholar’s Lab <http://etext.lib.virginia.edu>
  - Internet History Sourcebooks <http://www.fordham.edu/halsall>
Government & News Sites

- For current topics, government and news sites can provide useful information.
  - Census Bureau <http://www.census.gov>
  - Fedstats <http://www.fedstats.gov>
  - CNN <http://www.cnn.com>
  - BBC <http://www.bbc.co.uk>
  - Fox News <http://www.foxnews.com>

Online Reference Sources

- Almanacs, directories and encyclopedias:
  - The Old Farmers Almanac <http://www.almanac.com/>
  - Baseball Almanac http://www.baseball-almanac.com/

Refining A Keyword Search

- Use quotation marks around words that are part of a phrase: “Exxon profit levels”.
- Use AND (sometimes a + sign) to connect words that must appear in a document: McDonalds AND nutrition.
- Use NOT in front of words that must not appear in a document: PSU NOT Penn State.
- Use OR if only one of the words must appear in a document: DeBeers OR "conflict diamond”.
- Use an asterisk as a substitute for letters that might vary: "marine biologist” (to find marine biologist or marine biology).
- Use parentheses to group a search expression and combine it with another: (hamburgers OR cheeseburger OR fat*) AND McDonalds.

Research Hint

- Use the references and citations from one source to further your search.
Evaluating Online Sources

- Always keep your purpose in mind.
- Favor governmental and educational over commercial – how do you know?
- Favor authored over those without.
- Favor those available also in print.
- Favor those with recent updates. How recent?
- Favor those that document their claims.

Citing Your Sources

Why is this important?

- We must give credit to others for their ideas – avoid plagiarism.
  - Society requires a well-informed citizenry.
  - Society must maintain high standards in circulated and published materials.
  - These materials affect opinion and action.
  - Responsible writers take great care to specify when they refer to the work of others – readers can check the source.
  - Plagiarism breaks the “trust” and breeds skepticism.

When must you cite?

- Whenever you quote from an original source.
- Whenever you borrow ideas from an original source, even when you express them in your own words by paraphrasing or summarizing.
- Whenever you borrow factual information from a source that is not common knowledge.

Using The MLA Format

- MLA gives author’s full name on first mention, and omits the date.
- MLA allows for two ways to introduce cited material:
  - Brandon Conran argues that the story is written from “a bifocal point of view” (111).
  - The story is written from a bifocal point of view (Conran, 111).
Some Basic Rules

- The Bibliography starts on a new page.
- Double-space each line and between citations.
- Indent the second and subsequent lines of citations by five spaces (hanging indent).
- If citing an article you found on the web, but was originally in print form, provide enough info so that the reader can access it in either form.
- Arrange the list alphabetically by author.
- There are tons of sources for help:

Citing Periodical Print Pubs

- This list shows most of the possible components in an entry for an article from a print magazine:
  - Author’s name
  - Title of the article (in quotation marks)
  - Name of the periodical (italicized)
  - Series number or name (if relevant)
  - Volume number (for a scholarly journal)
  - Issue number (if available, for scholarly journal)
  - Date of publication (for a scholarly journal, the year; for other periodicals, the day, month, and year, as available)
  - Inclusive page numbers
  - Medium of publication (Print)
  - Supplementary information (see MLA Guide)

Print Pub Example

- An article from Business Week:

Online Only Periodical Example

- An online article from Newsweek:


Web Site Example

- General:
  Lastname, Firstname. “Article Title” Site Name. Organization Name, Article Date. Web. Date of Access.

- No Author/No Date:

When Using A Library Database

- General Rule:
  □ Must include both article/publication and library/database information.


Some Other Tools

- Outlining sites:
  □ http://www.loosestitch.com
  □ http://www.ioutliner.com

- Tracking research links:
  □ http://del.icio.us

- Citations:
  □ http://www.bibme.org
  □ http://www.ottobib.com
  □ http://www.citationmachine.net
Research Activity

- You are a management consultant hired to help Max with his video store business in Missoula.
  - Develop a research strategy in support of your consulting activities. What information would help with your efforts? How would you go about getting it?
  - Find a research report on Mintel that might be useful in your work for Max. Show me the title of the report along with one interesting piece of information that you think would be relevant for your work.
  - Find one online “refereed” source about the video rental business. Provide the Ulrich’s page proving the peer-reviewed status.
  - Blockbuster is a key competitor for Max. Find an online SWOT analysis for Blockbuster.
  - What are some key demographics for Missoula that might impact Max’s business?

Term Paper Hints

- Follow the format and watch your grammar, spelling and “neatness”
- Connect one section to another – support your conclusions and recommendations
- Have enough research sources and make sure they are “good”
- Use the tools reviewed in class – decision matrix, decision criteria, etc.
- Brainstorm and present more than one solution, and present a clear rationale for your choice
- Don’t forget to explicitly state the problem
- Lastly, make sure there is a specific, single recommendation, and that it is stated clearly!

Term Paper Grading

- Student Reviews/Critical Thinking (1)
- Research/Citations (4)
- Format/Grammar/Spelling (4)
- Problem Statement (4)
- Decision Making/Solutions (4)
- Data Analysis/Tools (4)
- Logic/Argument/Process (4)
Sample Problem Statements

- Traveller’s Insurance
  - Traveller’s is facing a steep decline in net income.

- Ford
  - Ford has lost market share in the US – from 17.4% to 14% - in 2009 and needs to reorganize many components to stay viable.

- Pfizer
  - Pfizer has experienced significant loss in regards to patent expiration.

- Microsoft
  - Microsoft’s smartphone OS design is uninspired.

PSU Problem Solving Process

1. Position
2. Sense
3. Uncover
4. Solve
5. Build
6. Achieve

It’s Okay To Be A Cynic

- There are three kinds of lies: lies, damned lies, and statistics. – Benjamin Disraeli
- Smoking is one of the leading causes of statistics. – Fletcher Knebel
- USA Today has come out with a new survey – apparently, three out of every four people make up 75% of the population. – David Letterman
Data And Information

What’s the difference?
- Raw numbers, facts and figures are data. Alone, a collection of data means nothing.
  - (145, 65), (215, 66), (250, 59), (244, 60)
- Information is data with a semantic association. Someone has taken the data and made it meaningful.
  - Weights and heights
  - What would you conclude about this group?

Qualitative Data
- Can’t necessarily be analyzed statistically
- Can be interpreted and understood through individual and group review and discussion
- Focus groups provide qualitative data – e.g., how people feel about a product

Quantitative Data
- Numerical data – can be counted and statistically analyzed
- Percentage of population 20 to 25 years old?

Data Types

Cross-Sectional Data
- Data collected at the same point in time.
- Generally info is collected on more than one variable (e.g., age, weight)

Time Series Data
- Data collected about one or more variables over multiple time periods
- For example - stock price over time.

Qualitative Data
- Data collected at the same point in time.
- Generally info is collected on more than one variable (e.g., age, weight)

Time Series Data
- Data collected about one or more variables over multiple time periods
- For example - stock price over time.

Working With Quantitative Data

Statistics
- A collection of tools and techniques for describing, organizing and interpreting data.
- They help you to extract useful information from data.

What are some of the basic techniques?
- Simple counting...
- Grouping and frequencies...
- You can calculate averages...

You can do other more complex statistical tests...
- Regression, t-test, chi-squared, etc.
Basic Calculations

- First – use simple stats to describe the data
  - Descriptive statistics
- What is the simplest calculation you can do to describe a group?
  - The one value that best describes a group of descriptors - scores or numbers
  - A measure of central tendency
- The average! (there are three)
  - Mean, Median and Mode

Central Tendency - Average

- Mean
  - The sum of all the values in the group divided by the total number of the values in the group
- Median
  - Different from the mean – the middle value in the group
- Mode
  - No formula – the value which occurs most frequently

Statistical Tools

- Basic math skills
- A calculator
- Specialized database analysis tools
  - dBase, Microsoft Access
  - SPSS and other statistical packages
Which One Do I Choose?

- It depends on the characteristics of the group
  - When there are a small number of extreme values in the group, the median is better than the mean.
- Ten people on a bus – mean income of $50,000.
  - Joe Blow gets off, Bill Gates gets on, what’s the mean now?
    - Around $50,000,000
    - The median is probably still around $50,000. But, what best describes the group?

More On Central Tendency

- The average alone doesn’t tell you enough
- You need to know more about how the values in the group vary from the mean
- Standard Deviation
  - The average distance from the mean
- Variance
  - The square of the deviation
- Why do we care?

Example

- 7, 6, 3, 3, 1
  - Mean – 4
  - Standard Deviation – 2.449
- 3, 4, 4, 5, 4
  - Mean – 4
  - Standard Deviation – 0.707
- 4, 4, 4, 4, 4
  - Mean – 4
  - Standard Deviation – 0.000

Plots Or Diagrams

- Why?
  - To understand the possible relationships between variables
- Plot or draw values on an X versus Y graph
  - e.g., plot age on the X-axis and cups of coffee per day on the Y-axis to see if there is a relationship
- How would you use this data?
- Tools like Excel make this very easy!
Data Relationships

<table>
<thead>
<tr>
<th>House Size</th>
<th>House Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>235</td>
</tr>
<tr>
<td>18</td>
<td>229</td>
</tr>
<tr>
<td>26</td>
<td>355</td>
</tr>
<tr>
<td>20</td>
<td>261</td>
</tr>
<tr>
<td>22</td>
<td>234</td>
</tr>
<tr>
<td>14</td>
<td>216</td>
</tr>
<tr>
<td>33</td>
<td>308</td>
</tr>
<tr>
<td>28</td>
<td>306</td>
</tr>
<tr>
<td>23</td>
<td>289</td>
</tr>
<tr>
<td>20</td>
<td>204</td>
</tr>
<tr>
<td>27</td>
<td>265</td>
</tr>
<tr>
<td>18</td>
<td>195</td>
</tr>
</tbody>
</table>

Scatter Plot

Correlations

- Very simply – does the value of one variable (like GPA) change when the value of another variable changes (like age)?
- Is there a relationship?
- Correlation coefficients indicate the strength of that relationship
  - -1.0 to +1.0 (the absolute value is what matters)
  - e.g., -0.9 is better than +0.1
  - The correlation for house size/price is 0.76

Rules Of Thumb

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Level of Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8 to 1.0</td>
<td>Very Strong Relationship</td>
</tr>
<tr>
<td>0.6 to 0.8</td>
<td>Strong Relationship</td>
</tr>
<tr>
<td>0.4 to 0.6</td>
<td>Moderate Relationship</td>
</tr>
<tr>
<td>0.2 to 0.4</td>
<td>Weak Relationship</td>
</tr>
<tr>
<td>0.0 to 0.2</td>
<td>Weak or No Relationship</td>
</tr>
</tbody>
</table>

- Example:
  - Correlation between Level of Income and:
    - Level of Education, 0.574
    - Attitude Toward Voting, -0.08
    - Recently Voted, -0.291

What Is Regression Analysis?

- Remember linear functions in Algebra?
- Drawing a line on an X versus Y graph?
- Regression analysis tries to fit a line to a bunch of data
- Why do we care?
  - If you know what the line is (the linear equation) you can do some predicting
Excel Gives You a Bunch of Numbers?

- Two of them are important:
  - R-square
    - Is like the correlation coefficient – numbers close to an absolute value of 1.0 are better – that shows a better linear relationship
  - Significance of F
    - Numbers <0.05 show that there is small likelihood that the relationship between the two happened purely by chance
  - Use Excel to play around with these tools…

Other Statistics

- t-test
  - Determining the significance of differences between two independent groups
- ANOVA
  - Analysis of Variance – a whole bunch of different tools for analyzing the differences between means of different groups
- Chi-squared (Goodness of Fit)
  - A test for comparing what you observe against what you expect

Using Excel For Analysis

Conclusions

- Getting the data is generally not the problem – analyzing it and using it to make good decisions is the problem.
- Use the tools available to you, but don’t overanalyze.
- Think about the questions you want to answer, and the important stakeholders.
- Can you make the data say what you want it to say?
Critical Thinking

- 51% of women are now living without spouse.  
  – New York Times
- Conclusion – marriage is threatened in the U.S.
- The Times got to 51% by including 2.4 million American females over 15 (of 117 million) who are married but not living with their husbands. It also counts widows not living with their husbands (geez – they’re dead!)
- We spend $50 billion per year on weddings.

Sample Problem Statements

- Traveller’s Insurance
  - Traveller’s is facing a steep decline in net income.
  - Disruption in Traveller’s sales channel and a weak economy have caused a 9.5% decline in Net Income from Q1 2009 to Q1 2010.
- Ford
  - Ford has lost market share in the US – from 17.4% to 14% - in 2009 and needs to reorganize many components to stay viable.
  - Disappointing consumer response to new car introductions has resulted in a 3.4% loss in market share from 2008 to 2009 in the US.

Sample Problem Statements

- Amazon
  - A ten percent reduction in average brick and mortar bookstore pricing from 2009 to 2010 has contributed to an x% drop in Amazon book sales.
- Sprint/Nextel
  - Poor customer service, ongoing weakness in the economy and inferior handset selection have caused a x% decrease in customer retention rates.
- Google
  - Google has seen a y% increase in copyright infringement lawsuits from 2005 to 2010 due to recent court opinions and more clarity about intellectual property law in the digital age.

PSU Problem Solving Process

1. Position
2. Sense
3. Uncover
4. Solve
5. Build
6. Achieve
PSU Problem Solving Process

- How and where do you get it?
  - Sources and methods
- What do you do with it once you get it?
  - Analysis and manipulation
- How do you use it once you've analyzed it?
  - Presenting data so others can understand what you know so well!

Presenting Your Data

- Why use charts and graphs?
  - Visualization enhances comprehension, enhances analysis
  - Decisions made 25% more quickly when viewing data graphically
  - It's easier to see relationships

<table>
<thead>
<tr>
<th>Driver Error</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tailgating</td>
<td>13424</td>
</tr>
<tr>
<td>Speeding</td>
<td>7477</td>
</tr>
<tr>
<td>Failure To Yield</td>
<td>7436</td>
</tr>
<tr>
<td>Dangerous Left Turn</td>
<td>2463</td>
</tr>
<tr>
<td>Improper Lane Change</td>
<td>2059</td>
</tr>
<tr>
<td>Running Red Light</td>
<td>1882</td>
</tr>
<tr>
<td>Improper Backing</td>
<td>848</td>
</tr>
<tr>
<td>Running Stop Sign</td>
<td>807</td>
</tr>
<tr>
<td>Failure To Slow</td>
<td>750</td>
</tr>
<tr>
<td>Turn From Wrong Lane</td>
<td>509</td>
</tr>
</tbody>
</table>

Cross-sectional or time series?

2005 Accident Causes

- Tailgating
- Speeding
- Failure To Yield
- Dangerous Left Turn
- Improper Lane Change
- Running Red Light
- Improper Backing
- Running Stop Sign
- Failure To Slow
- Turn From Wrong Lane
Choose The Best!

Plotting And Graphing Basics

- What are the variables?
  - Characteristics of a sample or population (age, car brand, etc.)
- What is the data?
  - Values or counts of the variable from observations
- What type of data & what type of graph?
  - Trend graphs (generally quantitative)
  - Relative size graphs (categories counts/qualitative)
  - Composition graphs (counts/percentage breakdown)
- What do you want to communicate?
Principles of Good Design

- High data to chart ratio.
- The right graph for the right data.
  - Most Excel choices are poor choices.
    - What's a donut chart for, anyway?
- Make sure it's complete.
  - Labels.
  - Titles.
- Think about the overall presentation.

Trend Graphs

- Scatter plots and line graphs.
- Often used when you want to emphasize a trend over time.
- Put the information where the reader expects it to be!
  - Time goes on the X axis.
- Use columns when Y is categorical!

Relative Size Graphs

- Used to compare relative sizes of different variable categories.
  - Columns (vertical) are better than bars (horizontal).
  - Can be used with one or more variables (e.g., class, car brand) – try pivot charts rather than building a table.
  - Columns should be equal width.
  - Don’t overuse colors and take care with legends.

Composition Graphs

- This is where pie charts are often misused.
  - Readers often have trouble with angles…
  - Try a segmented column with important segments at the top or bottom.
  - Consider whether a relative size graph is better!
Pivot Charts For Analysis

Other Suggestions

- When to use a bar chart?
  - When you have “many” or “long” bars

Six Ways To A Great Chart

- A graph should communicate only one idea
- Minimize chart or graph junk
- Plan out your chart before you create the final copy
- Label everything so nothing is left to the misunderstanding of the audience
- Keep things balanced
- A chart alone should convey what you say

Common Errors

- The wrong type of chart
  - Line for time series, columns for categories
- Missing text
- Inconsistent scale
- Keep zero at the bottom
- 3D when it doesn’t add value
- Images at 2X have 4X the area
- $'s not adjusted for inflation
- More than two or three significant digits
Bad Charts

Sometimes A Table Is Best

Don't Use 3D Because You Can

When We Didn't Have Color…
Final Test – Choose a Chart

- The proportion of freshman, sophomores, juniors and seniors at PSU
  - Pie Chart
- Change in GPA over three quarters
  - Line Chart
- Number of applicants for four different jobs
  - Column

More on Critical Thinking

Job Trends

Critical Thinking Needed!
Activity C

- Look at the poorly formatted data in the Activity C handout. What does it represent?
  - Spending data for the 03/04 school year for photocopies.
- You are responsible for organizing and presenting this data for the University's budget director. What do you think that person cares about? What would they want to see?
- Develop a list of weaknesses in the way the data is presented – be thorough.
- Read question 3 – create a new table based on these principles.
- Create one PowerPoint slide with a graph that best presents the data (not a table). Email it to me by midnight, Sunday, July 25th.