

# **GAISE Workshop**

**Session 5**

**Nov. 9, 2005**

**1:00 – 3:00 pm**

**Brian Smith and Bob delMas**

**Using Real Data**

## **Sources of real data**

From the GAISE report:

“Using real data sets of interest to students is also a good way to engage them in thinking about the data and relevant statistical concepts.”

Some guidelines from the GAISE College Report:

- Make sure questions used with data sets are of interest to students – if no one cares about the questions, it’s not a good data set for the introductory class. (Example: physical measurements on species no one has heard of.) Note: Few data sets interest all students, so one should use data from a variety of contexts.
- Use class-generated data to formulate statistical questions and plan uses for the data before developing the questionnaire and collecting the data. (Example: ask questions likely to produce different shaped histograms, use interesting categorical variables to investigate relationships). It is important that data gathered from students in class not contain information that could be embarrassing to students and that students’ privacy is maintained.
- Get students to practice entering raw data using a small data set or a subset of data, rather than spending time entering a large data set. Make larger data sets available electronically.
- Use subsets of variables in different parts of the course, but integrate the same data sets throughout. (Example: do side-by-side boxplots to compare two groups, then later do two-sample  $t$ -tests on the same data. Use histograms to investigate shape, then later to verify conditions for hypothesis tests.)

The Appendix to the GAISE report includes examples of good ways (and examples of not so good ways) to use data in homework, projects, tests, etc.

## GAISE College Report, Appendix, Page 26

### D. Examples of naked, realistic and real

#### (1) Naked data (not recommended)

Find the least squares line for the data below. Use it to predict Y when  $X=5$ .

X	1	2	3	4	6	8
Y	3	4	6	7	14	20

*Critique: Made-up data with no context (not recommended). The problem is purely computational with no possibility of meaningful interpretation.*

#### (2) Realistic data (better, but still not the best)

The data below show the number of customers in each of six tables at a restaurant and the size of the tip left at each table at the end of the meal. Use the data to find a least squares line for predicting the size of the tip from the number of diners at the table. Use your result to predict the size of the tip at a table that has five diners.

Diners	1	2	3	4	6	8
Tip	\$3	\$4	\$6	\$7	\$14	\$20

*Critique: A context has been added which makes the exercise more appealing and shows students a practical use of statistics.*

#### (3) Real data (recommended)

The data below show the quiz scores (out of 20) and the grades on the midterm exam (out of 100) for a sample of eight students who took this course last semester. Use these data to find a least squares line for predicting the midterm score from the quiz score.

Assuming that the quiz and midterm are of equal difficulty this semester and the same linear relationship applies this year, what is the predicted grade on the midterm for a student who got a score of 17 on the quiz?

Quiz	20	15	13	18	18	20	14	16
Midterm	92	72	72	95	88	98	65	77

*Critique: Data are from a real situation that should be of interest to students taking the course.*

## Types of Real Data:

- Archival data
- Classroom-generated data
- Simulated data.

Occasionally a hypothetical data set, such as the famous Anscombe data, may be used to illustrate a specific concept. In the case of the Anscombe data we see how four data sets can have the same regression equation and the same correlation coefficient, but noticeably different scatterplots. Such artificial data sets should be used sparingly and only for the intended purpose of illustrating a particular concept.

Raw data sets can be found in a wide range of different sources:

- Textbooks
  - Many newer texts have good data sets e.g.
    - *The Basic Practice of Statistics* by David S. Moore
    - *Interactive Statistics* by Aliaga and Gunderson
    - *Practical Statistics by Example* by Sincich, Levine and Stephan
- Software packages
  - many software packages come with CDs packed with real data
    - Minitab
    - Fathom
    - Data desk
    - SPSS
    - JMP, etc.
- From a researcher in your institution (Use first hand research data or seek references to articles in scientific journals)
  - Biologists
  - Psychologists (delete identifying information)
  - Chemists, etc.
- From an administrative office in your institution
  - Student records, Financial aid, etc
  - Obtain IRB approval and/or have identifying information deleted
- Government Agencies
  - U.S. Census Bureau
  - U.S. Department of Labor Statistics
  - U.S. Department of Transportation Statistics
  - Statistics Canada
  - FedStats
  - Economic Statistics Briefing Room

- Polling Agencies
  - Gallup Polls
  - Roper Center for Public Opinion Research
  - Harris Interactive
  - Ipsos (Canada)
  - Pollara (Canada)
  - Compas (Canada)
  
- Find data in newspaper and magazine articles.
  - <http://news.yahoo.com/>
  - <http://www.nytimes.com/pages/health/index.html>
  - Conscript students in this task. Ask each student in the class to find one interesting source of data. Select a dataset that has broad based appeal and analyze it – or have students analyze it!
  - Use the result as a basis for class discussion. What sort of questions arise from the data set itself – have students think of new research questions!
    - What is the population of interest?
    - What kind of study was completed?
    - What is a major finding of the study?
    - Does the article claim or imply a casual connection? Is it warranted?

**Some articles Brian has used recently for classroom discussion:**

*Tests of causal linkages between cannabis use and psychotic symptoms.*

David M. Fergusson, L. John Horwood & Elizabeth M. Ridder. *Addiction*, 100, 354-366

*Tattoos and Body Piercings as Indicators of Adolescent Risk-Taking Behaviors.*

Sean T. Carroll, Robert H. Riffenburgh, Timothy A. Roberts, and Elizabeth B. Myhre. *PEDIATRICS* Vol. 109 No. 6 June 2002, pp. 1021-1027

*A Simple Dataset for Demonstrating Common Distributions.* Peter K. Dunn

University of Southern Queensland. *Journal of Statistics Education* v.7, n.3 (1999)

*Mean, Median, and Skew: Correcting a Textbook Rule.* Paul T. von Hippel

The Ohio State University *Journal of Statistics Education* Volume 13, Number 2 (2005)

- Surveys or activities in class
  - Have students complete a first day survey
  - Have students collect data on campus, from their place of employment, from local libraries, municipal offices, hospitals, etc

First Day Survey from Bob delMas' Introductory Statistics Class

GC 1454 Statistics Class Survey

1. Who is your instructor?	<input type="radio"/> Suzanne Loch	<input type="radio"/> Bob delMas								
2. Which section are you in?	<input type="radio"/> 001	<input type="radio"/> 002	<input type="radio"/> 003	<input type="radio"/> 004	<input type="radio"/> 005					
3. What is your gender?	<input type="radio"/> MALE	<input type="radio"/> FEMALE								
4. What is your age in years? <input type="text"/> years of age										
5. How many siblings do you have including half- and step- ? <input type="text"/> siblings										
6. How many pairs of footwear do you personally own (this includes shoes, boots, sandals, clogs, etc.)? <input type="text"/> pairs of footwear										
7. How many body piercings do you have on your body? <input type="text"/> piercings										
8. Rate your intelligence level on a scale of 1 to 10 (where 1 = Brain Dead and 10 = Rocket Scientist)										
	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	<input type="radio"/> 6	<input type="radio"/> 7	<input type="radio"/> 8	<input type="radio"/> 9	<input type="radio"/> 10
9. Do you live?	<input type="radio"/> OFF CAMPUS	<input type="radio"/> ON CAMPUS								
10. What is your cumulative GPA? <input type="text"/>										
11. How many miles do you travel from your current home to campus each day, to the nearest mile? <input type="text"/> miles										
12. How many minutes does it take you to travel to school each day, on the average? <input type="text"/> minutes										
13. What type of transportation do you use most often to get to school?										
	<input type="radio"/> WALK	<input type="radio"/> CAR	<input type="radio"/> BUS	<input type="radio"/> BIKE	<input type="radio"/> OTHER					
14. How many minutes do you exercise each week, on the average? <input type="text"/> minutes										
15. How many credits are you taking this semester? <input type="text"/> credits										
16. How many hours per week do you study, on the average? <input type="text"/> hours										

17. What is your shoe size, to the nearest half?

<input type="radio"/> 4	<input type="radio"/> 4.5	<input type="radio"/> 5	<input type="radio"/> 5.5	<input type="radio"/> 6	<input type="radio"/> 6.5
<input type="radio"/> 7	<input type="radio"/> 7.5	<input type="radio"/> 8	<input type="radio"/> 8.5	<input type="radio"/> 9	<input type="radio"/> 9.5
<input type="radio"/> 10	<input type="radio"/> 10.5	<input type="radio"/> 11	<input type="radio"/> 11.5	<input type="radio"/> 12	<input type="radio"/> 12.5
<input type="radio"/> 13	<input type="radio"/> 13.5	<input type="radio"/> 14	<input type="radio"/> 14.5	<input type="radio"/> 15	<input type="radio"/> 15.5
<input type="radio"/> 16	<input type="radio"/> 16.5	<input type="radio"/> 17	<input type="radio"/> 17.5	<input type="radio"/> 18	<input type="radio"/> 18.5

18. What is your astrological sign?

<input type="radio"/> AQUARIUS	<input type="radio"/> LIBRA	<input type="radio"/> GEMINI	<input type="radio"/> TAURUS
<input type="radio"/> VIRGO	<input type="radio"/> CAPRICORN	<input type="radio"/> CANCER	<input type="radio"/> SCORPIO
<input type="radio"/> PISCES	<input type="radio"/> SAGITTARIUS	<input type="radio"/> ARIES	<input type="radio"/> LEO

19. What year are you in college?

<input type="radio"/> PSEO	<input type="radio"/> Freshman	<input type="radio"/> Sophomore	<input type="radio"/> Junior	<input type="radio"/> Senior
----------------------------	--------------------------------	---------------------------------	------------------------------	------------------------------

20. Where were you born?

If born in the USA, please select the **STATE** from the list below

<input type="radio"/> AL	<input type="radio"/> AK	<input type="radio"/> AR	<input type="radio"/> AZ	<input type="radio"/> CA	<input type="radio"/> CD	<input type="radio"/> CT	<input type="radio"/> DE	<input type="radio"/> FL	<input type="radio"/> GA
<input type="radio"/> HI	<input type="radio"/> IA	<input type="radio"/> ID	<input type="radio"/> IL	<input type="radio"/> IN	<input type="radio"/> KS	<input type="radio"/> KY	<input type="radio"/> LA	<input type="radio"/> MA	<input type="radio"/> MD
<input type="radio"/> ME	<input type="radio"/> MI	<input type="radio"/> MN	<input type="radio"/> MO	<input type="radio"/> MS	<input type="radio"/> MT	<input type="radio"/> NC	<input type="radio"/> ND	<input type="radio"/> NE	<input type="radio"/> NH
<input type="radio"/> NJ	<input type="radio"/> NM	<input type="radio"/> NV	<input type="radio"/> NY	<input type="radio"/> OH	<input type="radio"/> OK	<input type="radio"/> OR	<input type="radio"/> PA	<input type="radio"/> RI	<input type="radio"/> SC
<input type="radio"/> SD	<input type="radio"/> TN	<input type="radio"/> TX	<input type="radio"/> UT	<input type="radio"/> VA	<input type="radio"/> VT	<input type="radio"/> WA	<input type="radio"/> WI	<input type="radio"/> WV	<input type="radio"/> WY

If **NOT** born in the USA, please type the name of the **COUNTRY**.

I was born in

21. Your high school class rank is the percentage of students in your high school who had a GPA that is **LESS** than your high school GPA.

To the nearest 5%, what was your high school class rank?

<input type="radio"/> 5	<input type="radio"/> 10	<input type="radio"/> 15	<input type="radio"/> 20	<input type="radio"/> 25	<input type="radio"/> 30	<input type="radio"/> 35	<input type="radio"/> 40	<input type="radio"/> 45	<input type="radio"/> 50
<input type="radio"/> 55	<input type="radio"/> 60	<input type="radio"/> 65	<input type="radio"/> 70	<input type="radio"/> 75	<input type="radio"/> 80	<input type="radio"/> 85	<input type="radio"/> 90	<input type="radio"/> 95	<input type="radio"/> 99

22. Do you speak more than one language fluently?  YES  NO

## Examples of Activities that use the First Day Survey Data

---

### CATEGORICAL AND QUANTITATIVE MEASUREMENTS

Your group's task is to determine whether or not each question on the First Day Student Survey represents a CATEGORICAL or QUANTITATIVE measurement. You will have 10 to 15 minutes to work on this.

Here is some information to guide your decisions.

Each question on the survey represents one of two different types of measurement. The first type of measurement is **categorical**. An example would be to ask for a person's religious preference. This type of measurement is also referred to as **nominal** (i.e., you are naming something). With this type of question, a person's response falls into one of many categories (e.g., Catholic, Jewish, Lutheran, Baptist, etc.). You can count how many people fall into each category, but it doesn't make sense to calculate the average or perform other mathematical manipulations with the information. Note that this type of question might require something like "No Preference" as an option so that everyone can give an honest response.

The second type of measurement is **quantitative**. These measurements are referred to as quantitative in that numbers represent increasing amounts of a quantity. Quantitative measurements can be of two types.

The first type of quantitative measurement is called **ordinal**. These types of measurements can be placed in increasing or decreasing order, but you can't determine how much more or less one measurement is in comparison to another. A good example of an ordinal measurement is an athlete's place at the end of race: 1st, 2nd, 3rd, 4th, and so on. The measurements place the athletes in order, but the measurements do not let you determine how much faster the 1st place winner is in comparison to the athlete who finished in 2nd place.

The second type of quantitative measurement is called **continuous**: the measurements have no gaps, and you can determine the amount by which two measurements differ. For example, you can ask for a person's height or age. A person who is 30 years old is twice the age of someone who is 15. With this type of measurement it makes sense to ask about the average or mean. Other measurements of this type are money (e.g., checkbook balance), College GPA, and distance (e.g., number of miles a person lives from their hometown), just to name a few.

As a group, look at each question on the First Day Student Survey and decide whether it represents a **categorical** or **quantitative** measurement. Record your decisions for each question in the table provided below by circling either **Cat.** or **Quant.** for each question.

Question	Type of Measure
1	Cat.    Quant.
2	Cat.    Quant.
3	Cat.    Quant.
4	Cat.    Quant.
5	Cat.    Quant.
6	Cat.    Quant.
7	Cat.    Quant.

Question	Type of Measure
8	Cat.    Quant.
9	Cat.    Quant.
10	Cat.    Quant.
11	Cat.    Quant.
12	Cat.    Quant.
13	Cat.    Quant.
14	Cat.    Quant.

Question	Type of Measure
15	Cat.    Quant.
16	Cat.    Quant.
17	Cat.    Quant.
18	Cat.    Quant.
19	Cat.    Quant.
20	Cat.    Quant.
21	Cat.    Quant.
22	Cat.    Quant.



## Creating Contingency Tables and Pie Charts with DataDesk

You can use DataDesk to look at the relationship between two categorical variables in a contingency table. You are going to use the data from the First Day Survey to look at two of the relationships from the previous activity.

The first relationship will look at the question “Is there a relationship between a person’s gender and how they typically get to school ?”

- To do this, open the **GC 1454 Survey.dsk** file in DataDesk.
- Once it is open, open up the **DATA** folder.
- Point to the **TRANSPORT** variable and click on it once. This will place a yellow **Y** on top of it to indicate that **TRANSPORT** has been selected.
- While holding down the **SHIFT** key on the keyboard, point to and click the **GENDER** variable. A large blue **X** should appear on top of the **GENDER** variable.
- Go to the **CALC** menu and select **Contingency Table** from the bottom of the list.

You should now see a table with M and F (Males and Females) as columns, and the different types of transportation (bike, car, walk, etc.) listed as rows along the left of the table. The contingency table presents counts, but you need to have percentages in order to compare the males and females. Consider the question: “**Is there a relationship between a person’s gender and how they typically get to school ?**” The question indicates that you want to compare males and females. Because **gender** is represented by the **columns** in the contingency table, you want DataDesk to compute **Column Percentages**.

To do this, click the **rectangle** in the upper left of the contingency table window. Select **Table Options...** from the drop down menu. In the Table Options window, check the box for **Percent of column total**. Then click the **OK** button.

Column percents are now printed below each count for each cell of the contingency table.

It may also help to have a graph in order to see if there are similarities and differences between the males and females. To create two pie charts of the **TRANSPORT** variable, one for males and the other for females, do the following:

- Make sure that **TRANSPORT** still has a yellow **Y** on top of it, and **GENDER** still has blue **X**. If not, click once on **TRANSPORT**, then hold down the **SHIFT** key and click once on **GENDER**.
- Go to the **MANIP** menu and select **Split into Variables by Group**.
- A new window titled **GENDER** will appear on the screen, with two icons, one labeled **M** and one labeled **F**. Hold down the **ALT** key on the keyboard and click once on the **F** icon, then once on the **M** icon so that both icons have a yellow **Y** on top.
- Go to the **PLOT** menu and select **Pie Charts**.


You should see two pie charts, one for females and one for males. If they are positioned on top of the contingency table, move them to the side.

Use the column percentages in the contingency table and the colored areas in the pie charts to write a comparison of females and males use of transportation to get to school AND determine if there appears to be a relationship between gender and type of transportation to school.

Once you have finished your discussion, use contingency tables and pie charts to answer the following question: **Do students who live on campus tend to use a different type of transportation compared to students who live off campus?**

## Box Plots Lab

Now that you have been introduced to the Data Desk program and to the class survey data, you will learn how to make and interpret boxplots using Data Desk.

**Note:** Any time you are finished looking at a table or graph, click the  box in the upper **LEFT** corner of the window to remove it from the screen.

### 1. Analysis of GPA

Start **ActivStats**, then click the icon for **Data Desk**. Once you are in Data Desk, go to the **FILE** menu and open the **GC 1454 Survey.dsk** data file in the **GC 1454 folder**.

- Click once on the variable **EXERCISE** so that it is marked with a yellow **Y**.
- Go to the **PLOT** menu and select **Boxplot Side by Side**.
- You can change the display to **Black on White** by selectin **Plot Options** from the **PLOT** menu.
- It will also be helpful to have summary statistics, so go to the **CALC** menu, select **Summaries** and then **Reports**.
- Click the **arrow** button in the upper left corner of the report and click **Select Summary Statistics**. Check the boxes for **Lower Percentile**, **Upper Percentile**, **Min**, and **Max**. Click the **OK** button.

The boxplot provides a visual representation for median, lower and upper quartiles, and the low and high extremes of the distribution. You can use the information from the box plot and the summary statistics to determine the following values:

median \_\_\_\_\_

upper quartile \_\_\_\_\_ (same as the **lower ith %tile**)

lower quartile \_\_\_\_\_ (same as the **upper ith %tile**)

interquartile range \_\_\_\_\_

range \_\_\_\_\_ (value of the **Max** minus the **Min**)

About what percent of students exercise for 200 minutes or more each week? \_\_\_\_\_ %

### 2. GPA grouped by different characteristics.

You are going to create two boxplots, one for male students and another for female students.

- Check that the variable **GPA** is still marked with a yellow **Y**. If not, click on **GPA** once to mark it.
- Hold down the **SHIFT** key and click on the variable **GENDER**. This will mark it with an **X**. Data Desk uses an X to identify a **grouping variable** (a variable that can be used to split a dataset into groups).
- Go to the **PLOT** menu, but this time select **Boxplot y by x**. A window will open with two boxplots, one for females and one for males.
- You can change the display to **Black on White** from the **Plot Options** selection under the **PLOT** menu.

- Again it will be helpful to have summary statistics, but you want separate summaries for the females and males.
  - ◆ Go to the **CALC** menu, select **Summaries** and then **Reports by Groups**.
  - ◆ To get additional summary statistics, point to the arrow in the upper left of the window and select **Summary by Group Options**.
  - ◆ Select the **Lower** and **Upper percentiles**, as well as the **Min** and **Max**. Click the **OK** button.

You now have two boxplots that you can use to compare similarities and differences in the GPAs of males and females. Record the medians, quartiles, and interquartile ranges (IQR) for the two groups:

<b>GPA</b>	Median	Lower Quartile	Upper Quartile	Interquartile Range	Range
FEMALES	_____	_____	_____	_____	_____
MALES	_____	_____	_____	_____	_____

Describe the similarities and differences you see between the two boxplots.

3. Next, create side-by-side boxplots for the **Piercings** variable in order to compare the number of body piercings reported by members of each gender. Follow the same steps used above to create side-by-side boxplots for the GPA variable. The only thing you need to change is that **Piercings** is the **Y, Response** variable.

<b>PIERCINGS</b>	Median	Lower Quartile	Upper Quartile	Interquartile Range	Range
FEMALES	_____	_____	_____	_____	_____
MALES	_____	_____	_____	_____	_____

Do these two plots show any similarities or differences? Describe what you see in terms of the shapes (e.g., symmetric, skewed), the center and variability (i.e., the IQRs), and the amount of overlap in the two boxplots.

### Cleaning Real Data (Brian Smith)

A sample of data Brian collected from his Statistics Class: The table below shows 150 replies to the question:

**What is your height in centimeters (1 inch = 2.54 cm)?**

188	175	183	188	167.64
178cm	185	168	178	160
176	178	5 feet 6 inches	150	162
165	180.34	177.8	157	170
165	161	157.48	180	183
160.02	170	157.48	167	180
1meter63	183	182	168	154
170	1m89	356cm	162	169
147	162	155	165	142
163	1m77	175.26	156	183
185	184	180	160 cm	183cm
178	165	193	182	165
176	160	175	185	180
164cm	182.88	164	176	167,64
175	1,65 centimeters	170	157.48	158
180	174	166cm	177	181
169	158	188	170	178 cm
168	160	154	175 cm	172
163	162.56	162.56	160	185
171	163	165	170 cm	160
170	174	167	179	156
175	140	15	170	5'4
160.02	No idea	186 cm	169	185
193	181	180	170	180.34
1.57	175	152	169	170
172	180	193	190	163
178	173.5	160.02	183	178
168	183	175.26	178	5 feet and 17.78cm
13.97	160 centimeters	185	173	1.58
181	180 cm	179	175	175

356 cm = 140 ins  
= 11'8"

A very small person!!

Comma, not period

Mixed message - make up your mind!

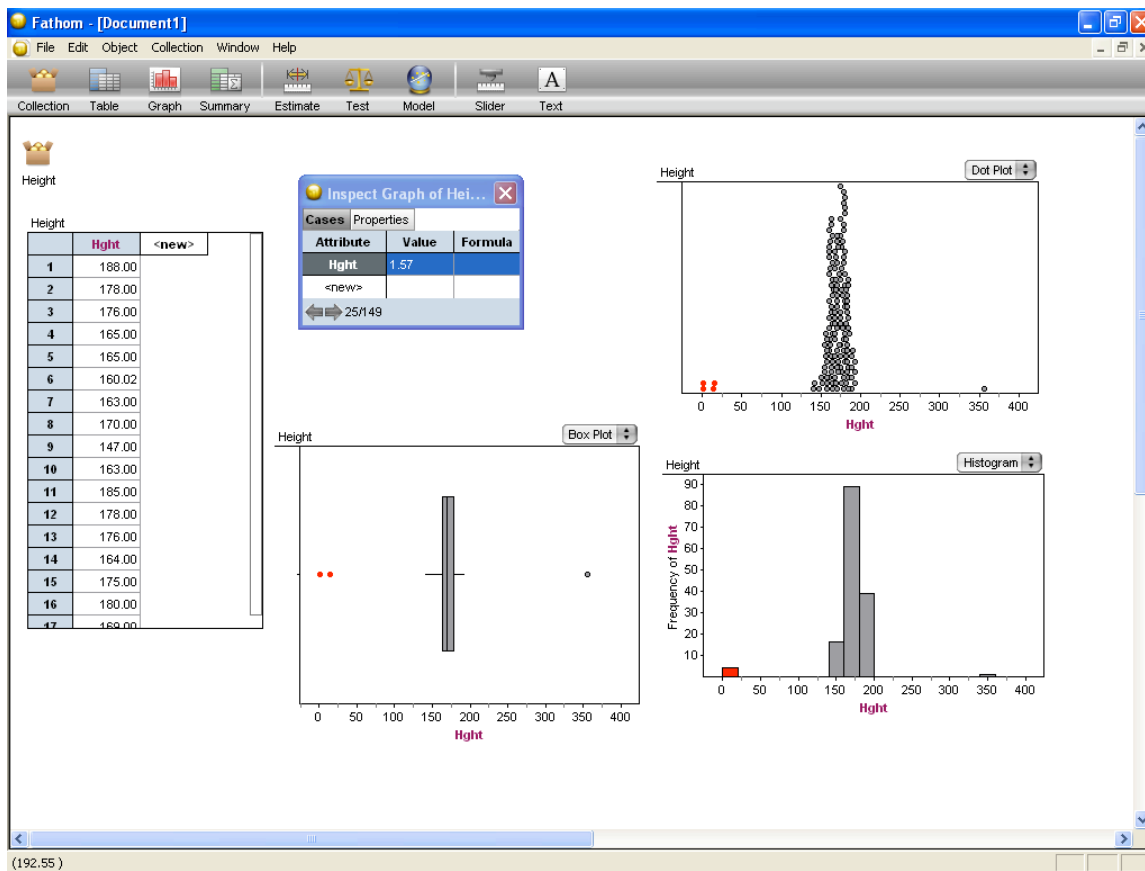
The raw data has 22 errors in 150 observations, or 14.7% error rate.

Note that there are two types of problems with the data. The first is the case of the overzealous students who insists on specifying the units, e.g. 164 cm. The problem with this is that "164 cm" is not a number recognized by Excel so that the entry has to be changed to "164" in order to be a valid Excel value.

The second problem is more serious and requires a judgment call by the analyst. For example, what does an entry of 356 cm mean? What does 13.97 mean? Is it 139.7, which would be possible (4'7"), or should it be treated as missing data?

This is a very valuable in-class exercise. Mostly we are given a data set and assume that all of the entries are correct and make sense. But data validation is necessary for every data set to ensure that statistical analyses are valid – otherwise we confront the GIGO problem – garbage in, garbage out.

The graph below shows three graphical representations in Fathom: a histogram, a dot plot and a box plot. In each case we see that there are outlier values, at both ends of the scale, that are clearly errors in data entry.



This is an important point: graphical displays will often prevent serious errors in statistical estimation by identifying invalid data entries.

### *Using Institutional Data: Constructing a Confidence Interval for a Proportion*

The General College traditionally has a large number of students of color enrolled. But, just how large is this group within the General College? A 95% confidence interval can be constructed to answer this question. Define a student of color as a person enrolled at the University of Minnesota who indicated that they are of a non-White ethnic background.

In fall 2003, there were 1912 students registered in the General College. At the top of the next page is a random sample of 100 students randomly selected from the 1912 General College students. Use the information above and the sample below to calculate a 95% confidence interval for the proportion of students of color among fall 2003 General College students.

(NOTE: XX indicates a student did not provide information on his or her ethnic background. Therefore, XX does **NOT** indicate a non-White ethnic background).

1. Calculate the sample proportion for the sample of 100 General College students.  
Sample Proportion =  $\hat{p}$  =
2. Does the sample meet the four conditions needed to assume that sample proportions will follow a normal distribution (for samples of size  $n = 100$  taken from the General College 2003 freshman population)? [see pages **433-434** for the four conditions]
  - a. Plausible Independence Condition?
  - b. Randomization Condition?
  - c. 10% Condition?
  - d. Success/Failure Condition?
3. If the sample of 100 General College students meets all four conditions, then you can calculate the 95% confidence interval. What is the **standard error** of the sampling distribution of  $\hat{p}$  for random samples of size 100 drawn from the population of fall 2003 GC freshmen? (see page **429** of the textbook)
4. What z-value would you use to construct a 95% confidence interval for this sample?  
 $z =$
6. Calculate the 95% confidence interval for the true value  $p$  for the population that the General College students came from. Use the example presented on pages **435-436**.
7. Using the 95% confidence interval you calculated, make a statement about the proportion of 2003 General College freshmen who were students of color (see pages **430-431** of the textbook).

SRS of 100 Fall 2003 General College Students

AI=Native American  
CH=Chicano/Hispanic

AS=Asian American  
WH=White

BL=African American  
XX=Missing

AI	AS	AS	AS	WH	WH	WH	WH
AI	AS	AS	AS	WH	WH	WH	WH
AS	AS	AS	AS	WH	WH	WH	WH
AS	AS	AS	AS	WH	WH	WH	WH
AS	AS	AS	BL	WH	WH	WH	WH
AS	AS	AS	BL	WH	WH	WH	WH
AS	AS	AS	BL	WH	WH	WH	XX
AS	AS	AS	BL	WH	WH	WH	XX
AS	AS	AS	CH	WH	WH	WH	XX
AS	AS	AS	CH	WH	WH	WH	
AS	AS	AS	WH	WH	WH	WH	
AS	AS	AS	WH	WH	WH	WH	
AS	AS	AS	WH	WH	WH	WH	

### *One-Sample Test for the Proportion: Students of Color in GC*

Previously, you conducted an activity to estimate the proportion of Students of Color who enrolled in the General College fall 2003. During that same semester, the proportion of students of color among all undergraduates NOT enrolled in the General College was .141 (or 14.1% students of color). Conduct a hypothesis test to determine if the proportion of Students of Color in the General College differed significantly from the proportion among University of Minnesota undergraduates in fall 2003.

1. Determine if you should use a one-sided or two-sided test (see page **455** of the textbook), then state the null ( $H_O$ ) and alternative ( $H_A$ ) hypotheses.
2. Are all conditions met for modeling the sampling distribution with a Normal model (see pages **433-434** and page **454** of the textbook)?
3. If the conditions for a normal model are met, carry out the mechanics for the test (calculate the standard error based on the **null hypothesis proportion**, calculate the z-value of the test statistic based on the **sample proportion**, and determine the p-value of the test statistic – see pages **456-458** for an example).
4. Draw an appropriate conclusion about the null hypothesis. Be sure to state your conclusion within the context of the problem.
5. Use the 95% confidence interval you calculated previously for the true proportion of Students of Color in the General College. Interpret the confidence interval within the context of the problem (see pages **459-462** of the textbook).



## **Online Resources for Finding Data**

### Data and Story Library (DASL)

<http://www.stat.cmu.edu/DASL/>

### StatLib (Carnegie Mellon University)

<http://lib.stat.cmu.edu/>

### Chance Database

<http://www.dartmouth.edu/~chance/>

### OzData: Australasian Data and Story Library

<http://www.maths.uq.oz.au/~gks/data/index.htm>

### Case Studies: UCLA

<http://www.stat.ucla.edu/cases/>

The Data Applet: (Graphical analysis of data sets – includes both the data sets and the software to analyze them)

<http://www.stat.uiuc.edu/courses/stat100/java/DataApplet.html>

### Finding data on the Internet (NilesOnLine)

<http://nilesonline.com/data/>

### Time Series Data Sets

<http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/>

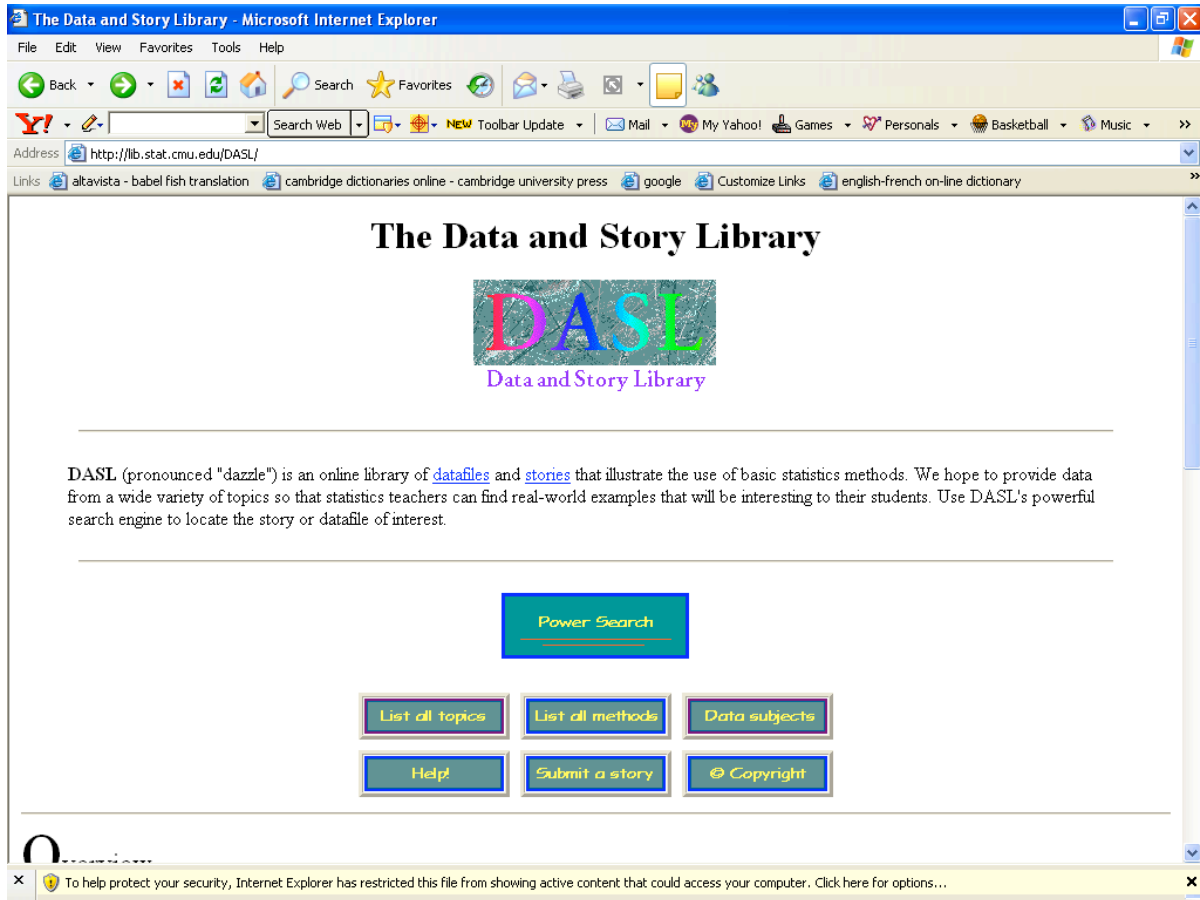
### Economic Time Series Page: Economagic

<http://www.economagic.com/>

### Quantitative Environment Learning Project (QELP)

<http://www.seattlecentral.org/qelp/Data.html>

Let's spend a few minutes investigating DASL – Data and Story Library  
<http://www.stat.cmu.edu/DASL/>



Click on **List all topics** button. From the list that appears, click on **Economics**, and then click on item number 28 [Billionaires 1992 Story](#), and you will see the information below:

**Datafile Name:** Billionaires 92

**Datafile Subjects:** [Economics](#)

**Story Names:** [Billionaires 1992](#)

**Reference:** *Fortune*, September 7, 1992. "The Billionaires." pp. 98-138.

**Authorization:** free use

**Description:** Fortune magazine publishes the list of billionaires annually. The 1992 list included 233 individuals or families. Their wealth, age and geographic location (Asia, Europe, Middle East, United States or Other) is reported.

**Number of cases:** 233

**Variable Names:**

1. wealth: Wealth of family or individual in billions of dollars
2. age: Age in years (for families it is the maximum age of family members)

The DASL site includes a discussion of the nature of the data set, as well as the types of graphs and statistical procedures that would be appropriate for analyzing the data.

**Methods:** [ANCOVA](#) , [Distribution](#) , [Interaction](#) , [Transformation](#)

The variable 'wealth' is right skewed, so the mean exceeds the median. There is a question of how we might look at the wealth distribution. We have here the upper tail of the distribution of the wealth of all humans. common assumption about the distribution of wealth is that its logarithm is roughly Normal. If so, then the logarithm of these wealth values will still be skewed. Experimentation shows that '1/wealth' is more nearly symmetric.

A Pie chart or bar chart can indicate how the billionares are distributed around the world.

A scatterplot of '1/wealth' vs 'age' shows very little trend. Fitting lines individually by part of the world, shows a negative trend in '1/wealth' vs. 'age' for the Mideast that stands out from the other parts of the world. This makes sense because a negative association between '1/wealth' and 'age' implies a positive relationship between 'wealth' and 'age'.

Fitting a model with the continuous explanatory variable 'age' and the discrete explanatory variable 'region' is an analysis of covariance (ANCOVA) with parallel regression lines for each region, each with a different intercept but the same slope of 'age'. Adding an 'age\*region' interaction allows the slope for each regression line to be different.

**Image:** Scatterplot of '1/wealth' vs. 'age' with regression lines colored by 'region'

