

## How to find and use real-world data

Mark Harbison  
CMC3 Business Liaison  
7359 Rush River dr.  
Sacramento, CA 95831  
[harbism@scc.losrios.edu](mailto:harbism@scc.losrios.edu)

1. Use a Statistics Professor's webpage (**.edu**) for raw dataSets.
  - a. <http://lib.stat.cmu.edu/> = StatLib @ Carnegie Mellon U. (click on "Get Data").
    - i. the Data And Story Library (DASL = 114 datasets). pg. 2
    - ii. the Data Expo Archive (9 datasets). pg. 3
    - iii. the DataSet Archive (104 datasets). pg. 3-4
    - iv. the Journal of the Am. Statistical Assoc. Data Archive (17 JASA dataSets).
  - b. [www.ruf.rice.edu/~lane/case\\_studies/](http://www.ruf.rice.edu/~lane/case_studies/) = Rice U. Virtual Lab in Statistics.
    - i. highly-recommended Java Applets such as a CLT demo.
    - ii. Seven case studies include raw data.
  - c. <http://www-stat.stanford.edu/ElemStatLearn/> = Hastie, Tibshirani & Friedman textbook
2. Use a Statistics Organization's webpage (**.org**) for raw dataSets.
  - a. [www.amstat.org/publications/jse/jse\\_data\\_archive.html](http://www.amstat.org/publications/jse/jse_data_archive.html) pg. 5  
= the Journal of Statistics Education from the American Statistical Association.
  - b. [www.causeweb.org](http://www.causeweb.org) seems brand new.
3. Use a government webpage (**.gov**) for summarized data.  
(e.g. [www.census.gov](http://www.census.gov) , [www.ssa.gov](http://www.ssa.gov) , [www.cdc.gov](http://www.cdc.gov) , ... )
  - a. [www.fedStats.gov](http://www.fedStats.gov) has many statistics links. pg. 6
  - b. [www.firstGov.gov](http://www.firstGov.gov) has at least 474 agencies listed. pg. 7
4. Use a commercial webpage (**.com**) for summarized data.
  - a. [www.statistics.com](http://www.statistics.com) has 2556 links to other sites. pg. 8
  - b. [www.gallup.com](http://www.gallup.com) has up-to-date reports on interesting topics. pg. 9
5. **Interview** students to gather your own data. pg. 10
6. Use [www.Google.com](http://www.Google.com) to **search** for something new & specific. pg. 11-13
7. Use **printed items** if the computer is down.
  - a. books (e.g. almanacs)
  - b. journals
  - c. magazines
  - d. newspapers
8. If all else fails, then use somebody else's data (**Gene**)... pg. 14-17

Also, please send us suggestions of your favorite data sources. Thank you.

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

1st. chose a category (out of 24 categories) and then choose a DataSet (out of 114 DataSets)

**Categories DataSets**

<a href="#">Economics</a>	Billionaires 1992; Cities; Educational Spending; European Jobs; Companies; Emeralds; Oil Production; Faculty Salaries; Unemployment; Labor Force; CEO Salaries; Enrollment Forecast; Country Inflation; State Labor Law; OECD Econ. Dev.; Quarterback+Team Salaries; State Public Expenditures; Teacher Salary by State; TV Ads; Wages+Hours; Waste Run Up.		
<a href="#">Consumer</a>	Alcohol+Tobacco; Fish Prices; Shoppers; Ice Cream; Ag Econ; Predicting Appliance Sales; Beef Council; Predicting Retail Sales; Albuquerque Home Prices; Nambeware Manuf.; Magazine Ad Readability; Montana Outlook Poll; Refusals in Mortgage Lending; Quarterly Appliance Sales; Food Taste Test.		
<a href="#">Education</a>	Colleges; Education by Age; Nurses; Graduation; DRP Scores; Reading Scores; Scents; Instructor Behavior; Popular Kids (see example below).		
<a href="#">Health</a>	Age+Height; Brain size; Breast Cancer; Calcium; Cereals; Hearing; Fiber; Balance; Smoking+Cancer; Stepping; MA Lunatics; Nursing Homes.		
<a href="#">Government</a>	Draft Lottery; Votes; Parking Meter Theft; Highway Deaths; NYC Crime; Time Series.		
<a href="#">Automotive</a>	Cars; Auto Pollution Filter Noise; Batmobile; Passenger Car Mileage; U.S. Vehicle Weights '75-'90.		
<a href="#">Biology</a>	Eggs; Fisher's Iris; Flea Beetles; Medflies; Mercury in Bass; Termites; Cuckoo Eggs; Student t Distr.		
<a href="#">Sports</a>	Helium football; World Series; Home Runs; Men's Track; Olympics; Crews; Hitters '20-'50; Pitchers.		
<a href="#">Science</a>	Birthrates; Chromatography; Speed of Light; Rain Forest; US Crime.		
<a href="#">Miscellaneous</a>	Militiamen Chests; Distribution Patterns; Transformations.		
<a href="#">Psychology</a>	Friday the 13 <sup>th</sup> ; Fusion Times; Singers; Crawling.	<a href="#">Datasets</a>	Difference Tests.
<a href="#">Environment</a>	SMSA; Clouds; Refinery; US Temperatures.	<a href="#">Food</a>	Cheese; Hot Dogs.
<a href="#">Medical</a>	Heart Valves; Cancer Survival.	<a href="#">Energy</a>	Nuclear Plants.
<a href="#">Archeology</a>	Egyptian Skulls, Pottery.	<a href="#">Nature</a>	Wild Horses.
<a href="#">Europe</a>	Protein Consumption.	<a href="#">Nutrition</a>	Calories.
<a href="#">Physics</a>	Cavendish; Michelson.	<a href="#">Astronomy</a>	Hubble.
<a href="#">Geography</a>	Acorns; New Jersey.	<a href="#">Engineering</a>	Crash.

**DASL Example**

**Datafile Name:** Popular Kids

**Datafile Subjects:** [Psychology](#), [Social science](#)

**Story Names:** [Students' Goals](#) , [What Makes Kids Popular](#)

**Authorization:** Contact authors

**Reference:** Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children," *Research Quarterly for Exercise and Sport*, 63, 418-424

**Description:** Subjects were students in grades 4-6 from three school districts in Ingham and Clinton Counties, Michigan. Chase and Dummer stratified their sample, selecting students from urban, suburban, and rural school districts with approximately 1/3 of their sample coming from each district.

**Number of cases:** 478

**Variable Names:**

- |            |   |  |              |
|------------|---|--|--------------|
| 1. Gender: | Boy or girl   | 2. Grd.:   | 4, 5 or 6    |
| 3. Age:    | Age in years  | 4. Race:   | White, Other |
| 5. U/R:    | Urban or Rural or Suburban school district  |  |              |
| 6. School: | Brentwood Elem., Brentwood Middle, Ridge, Sand, Eureka, Brown, Main, Portage, Westdale Middle   |  |              |
| 7. Goals:  | Student's choice of personal goals: 1 = Make Good Grades, 2 = Be Popular, 3 = Be Good in Sports |  |              |
| 8. Grades: | Rank of "make good grades"  | (1=most important for popularity, 4=least important) |              |
| 9. Sports: | Rank of "being good at sports"  | (1=most important for popularity, 4=least important) |              |
| 10. Looks: | Rank of "being handsome or pretty"  | (1=most important for popularity, 4=least important) |              |
| 11. Money: | Rank of "having lots of money"  | (1=most important for popularity, 4=least important) |              |

**The Data:**

Gender	Grd.	Age	Race	U/R	School	Goals	Grades	Sports	Looks	Money
girl	4	9	White	Rural	Ridge	Grades	1	3	2	4
boy	4	9	Other	Rural	Elm	Popular	4	3	2	1
boy	6	12	Other	Suburban	Brentwood Middle	Grades	4	1	2	3
girl	4	9	Other	Urban	Main	Grades	2	4	1	3

girl 6 11 White Urban Westdale Middle Sports 3 2 1 4  
 .....

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

The data-expo archive currently contains:

- [1997](#) Hospital Benchmarking Data (Not available via e-mail. Use WWW or FTP instead).
- [1995](#) U.S. Colleges and Universities
- [1993](#) Oscillator time series & Breakfast Cereals
- [1991](#) Disease Data for Public Health Surveillance
- [1990](#) King Crab Data
- [1988](#) Baseball (**excerpt shown below**)
- [1986](#) Geometric Features of Pollen Grains
- [1983](#) Automobiles

Last modified: Fri Jul 13 16:50:58 EDT 2001 By Pantelis Vlachos

Data Expo Example (Baseball)

Hitter's name (n = 322)	Andy Allanson	Alan Ashby	Alvin Davis	Andres Galarraga
# of at-bats in 1986	293	315	479	496
# of hits in 1986	66	81	130	141
# of HRs in 1986	1	7	18	20
# of runs in 1986	30	24	66	65
# of RBIs in 1986	29	38	72	78
# of walks in 1986	14	39	76	37
# of years w/ MLB	1	14	3	11
career # of at-bats	293	3449	1624	5628
career # of hits	66	835	457	1575
career # of HRs	1	69	63	225
career # of runs	30	321	224	828
career # of RBIs	29	414	266	838
career # of walks	14	375	263	354
league at end of 1986	A	N	A	N
div. at end of 1986	E	W	W	E
team at end of 1986	Cle.	Hou.	Sea.	Mon.
position(s) during 1986	C	C	1B	RF
# of put outs in 1986	446	632	880	200
# of assists in 1986	33	43	82	11
# of errors in 1986	20	10	14	3
\$1000s salary on opening day of 1987	NA	475	480	500
league at start of 1987	A	N	A	N
team at the start of 1987	Cle.	Hou.	Sea.	Mon.

<http://lib.stat.cmu.edu/datasets> Top-16 Name List

Dataset name	# of downloads (as of Feb '05)
bodyfat % and circumference for men ( <b>see next page</b> )	1521
A. Agresti <b>textbook</b> data	1388
S. Weisberg <b>textbook</b> data	1372
arsenic levels in toenail measurements	1345
IQ, Brain size data for twins	1329
backache during pregnancy data	1255
clustering web pages, machine-learning	1129
automobile mpg, cylinders etc data	1095
Boston house-price data	1080
Foster, Stine and Waterman <b>textbook</b> data	1053
Sleep in Mammals: Eco-constitutional Correlates ( <b>see next page</b> )	1045

radiation measurements taken from a balloon 1018  
 J. Simonoff **textbook** data 894

**cmu.edu** DataSets Example (Sleep in Mammals)

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

Data from which conclusions were drawn in the article "Sleep in Mammals: Ecological and Constitutional Correlates" by Allison, T. and Cicchetti, D. (1976), *Science*, Nov.12, vol. 194, pp. 732-734.

p = predation index: (1 = least likely to be preyed upon) to (5 = most likely to be preyed upon)  
 s = sleep exposure index: (1 = least exposed = e.g. a well-protected den) to (5 = most exposed)  
 o = overall danger index: (1 = least danger from other animals) to (5 = most danger from other animals)

note: Missing values denoted by -999.0

**The Data (n = 62)**

Species	body(kg)	brain(kg)	sleep (hours/day)			max yrs	gest.days	p	s	o
			light	heavy	total					
Arctic squirrel	0.920	5.700	-999.0	-999.0	16.5	-999.0	25.0	5	2	3
Asian elephant	2547.000	4603.000	2.1	1.8	3.9	69.0	624.0	3	5	4
Baboon	10.550	179.500	9.1	.7	9.8	27.0	180.0	4	4	4
Cat	3.300	25.600	10.9	3.6	14.5	28.0	63.0	1	2	1
Cow	465.000	423.000	3.2	.7	3.9	30.0	281.0	5	5	5
Giraffe	529.000	680.000	-999.0	.3	-999.0	28.0	400.0	5	5	5
:										

**cmu.edu** DataSets Example (Bodyfat %)

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

The variables listed below, from left to right, are:

- Density determined from underwater weighing
- Percent body fat from Siri's (1956) equation
- Age (years)
- Weight (lbs)
- Height (inches)
- Neck circumference (cm)
- Chest circumference (cm)
- Abdomen 2 circumference (cm)
- Hip circumference (cm)
- Thigh circumference (cm)
- Knee circumference (cm)
- Ankle circumference (cm)
- Biceps (extended) circumference (cm)
- Forearm circumference (cm)
- Wrist circumference (cm)

The data were supplied by Dr. A. Garth Fisher who gave permission to freely distribute the data and use for non-commercial purposes.

Reference (more are listed on the webpage):

Siri, W.E. (1956), "Gross composition of the body", in *Advances in Biological and Medical Physics*, vol. IV, edited by J.H. Lawrence and C.A. Tobias, Academic Press, Inc., New York.

**The Data (n = 252)**

1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	17.1
1.0853	6.1	22	173.25	72.25	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2
1.0414	25.3	22	154.00	66.25	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6
1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2

.....

[www.amstat.org/publications/jse/jse\\_data\\_archive.html](http://www.amstat.org/publications/jse/jse_data_archive.html)**the American Statistical Association: JSE Data Archive (Journal of Stats Edu.)**

(34 from JSE articles) + (17 unpublished) = (51 total dataSets, 34 shown here)

(21 Census types) + (22 Sample types) + (8 other types)

where "Sample" includes: designed experiment, random sample, observational, measurement &amp; six-tuples.

NAME: Pricing the C's of Diamond Stones	NAME: 2004 New Car and Truck Data
TYPE: Observational Regression Analysis Data	TYPE: Sample
SIZE: 308 obs, 5 variables (JSE July 2001).	SIZE: 428 obs, 19 variables
NAME: 1993 New Car Data	NAME: 1997 University of Iowa Big Ten Basketball Data
TYPE: Sample	TYPE: Census
SIZE: 93 obs, 26 variables (JSE July 1993).	SIZE: 18 obs, 4 variables (JSE July 1998).
NAME: Time of Birth, Sex, and Birth Weight	NAME: The 1998 McGwire/Sosa Home Run Race
TYPE: Observational	TYPE: Census; Time series
SIZE: 44 obs, 4 variables (JSE Nov. 1999).	SIZE: 163 obs, 21 variables (JSE Nov. 1998).
NAME: Ball Bearing Reliability Data	NAME: Lotto 6/42 Selections
TYPE: Sample	TYPE: Six-tuples from {1, 2, 3, ..., 42}
SIZE: 210 records by 11 variables (JSE Nov. 2002).	SIZE: 762 obs, 7 variables (JSE Nov. 1999).
NAME: Are Baseball Salaries Based on Performance?	NAME: Career Stats for MLB Hall of Fame Eligibles
TYPE: Census	TYPE: Census
SIZE: 337 obs, 18 variables (JSE July 1998).	SIZE: 1419 obs, 27 variables (JSE July 2000).
NAME: BestBuy	NAME: MLB Attendance data 1969-2000
TYPE: Time series	TYPE: Census
SIZE: 48 obs with 3 variables, (JSE March 2003).	SIZE: 838 team/seasons, 10 variables (JSE July 2002).
NAME: a Longitudinal Study in South Africa	NAME: NFL Y2K PCA
TYPE: Census to show Simpson's Paradox	TYPE: Observational, Census
SIZE: 1590 obs, 3 variables (JSE Nov. 1999).	SIZE: 31 obs, 78 variables (JSE Nov. 2001).
NAME: Exploring Relationships in Body Dimensions	NAME: NFL Scores and Pointspreads
TYPE: Observational	TYPE: Census (All values from recent seasons)
SIZE: 507 Obs, 25 Variables (JSE July 2003).	SIZE: 235 or 251 obs, 8 variables (JSE Nov. 1997).
NAME: Cigarette data for an intro. to multiple regression	NAME: Normal Body Temp., Gender, and Heart Rate
TYPE: Sample	TYPE: Random sample
SIZE: 25 obs, 5 variables (JSE July 1994).	SIZE: 130 obs, 3 variables (JSE July 1996).
NAME: Diamond Ring Pricing Using Linear Regression	NAME: Drug Interaction
TYPE: Random Sample	TYPE: Designed experiment in humans
SIZE: 48 obs, 2 variables (JSE Nov. 1996).	SIZE: 44 lines, 8 variables (JSE March 2004).
NAME: The Draft Lotteries of 1970, 1971, and 1972	NAME: The Statistics of Poverty and Inequality
TYPE: Randomized	TYPE: Sample
SIZE: 1095 obs, 26 variables (JSE July 1997).	SIZE: 97 obs, 8 variables (JSE July 1995).
NAME: Electric Bill Data	NAME: Readability of Edu. Materials for Cancer Patients
TYPE: Sample	TYPE: Two independent samples
SIZE: 120 obs, 13 variables (JSE March 2003).	SIZE: 93 obs, 2 vars & freq. distributions (JSE July 1995).
NAME: An Experiential Approach to Integrating ANOVA	NAME: Equity in Public School Expenditures
TYPE: Designed experiment (JSE March 2002)	TYPE: Census
SIZE: four experiments (4 variables with {8, 16, 32, 64} obs).	SIZE: 50 obs, 8 variables (JSE July 1999).
NAME: Fitting % Body Fat to Simple Body Measurements	NAME: Televisions, Physicians, and Life Expectancy
TYPE: Sample	TYPE: Sample
SIZE: 252 obs, 19 variables (JSE March 1996).	SIZE: 40 obs, 6 variables (JSE Nov. 1994).
NAME: Data from the TV Game Show "Friend or Foe?"	NAME: Titanic Death Rates for an Unusual Episode
TYPE: Census	TYPE: Complete record for all of population at risk
SIZE: 454 obs, 14 variables (JSE Nov. 2004).	SIZE: 2201 obs, 4 variables (JSE Nov. 1995).
NAME: Sexual activity and lifespan of male fruitflies	NAME: The Tryptone Task
TYPE: Designed (almost factorial) experiment	TYPE: Designed experiment
SIZE: 125 obs, 5 variables (JSE July 1994).	SIZE: 30 obs, 9 variables (JSE July 2004).

NAME: What Does It Take to Heat a New Room?  
 TYPE: Time series  
 SIZE: 81 obs, 13 variables (JSE March 1998).

NAME: NFL Scores for 1998-2000  
 TYPE: Census (from '98-'00 regular seasons)  
 SIZE: 736 games, 7 variables

[www.fedstats.gov/toolkit.html](http://www.fedstats.gov/toolkit.html)

Many text and numeric databases are available from Federal agencies for policy analysis and general research. Web-based tools allow you to use your browser to view predefined reports and generate your own tables with data obtained through searches and queries of summary and microdata files.

The tools provided allow you to

- select information interactively (for example, making a selection from a scrollable list on a form),
- view the data in a variety of formats including HTML tables and graphics, and
- print or download results into spreadsheets.

Multi-agency = [Ferret](http://ferret.bls.census.gov/) (*Federal Electronic Research, Review, & Extraction Tool*)  
 (access to datasets from the Census Bureau, Bureau of Labor Statistics & Nat'l Center for Health Statistics)

*U.S. Census Bureau*

American FactFinder <http://factfinder.census.gov>  
 CenStats <http://censtats.census.gov>  
 TIGER Map Services <http://tiger.census.gov>

*Bureau of Economic Analysis*

Gross State Product (GSP) [www.bea.gov/bea/regional/gsp/](http://www.bea.gov/bea/regional/gsp/)  
 National Income and Product Accounts (NIPA) [www.bea.gov/bea/dn/nipaweb/index.asp](http://www.bea.gov/bea/dn/nipaweb/index.asp)  
 State Quarterly Personal Income [www.bea.gov/bea/regional/sqpi/](http://www.bea.gov/bea/regional/sqpi/)

*Bureau of Justice Statistics*

NACJD - Data Analysis System [www.icpsr.umich.edu/NACJD/](http://www.icpsr.umich.edu/NACJD/)  
 Federal Justice Statistics Resource Center <http://fjsrc.urban.org/>

*Bureau of Labor Statistics*

Selective Access [www.bls.gov/data/sa.htm](http://www.bls.gov/data/sa.htm)

*Economic Research Service, USDA*

Farm Business and Household Survey Data [www.ers.usda.gov/data/arms/](http://www.ers.usda.gov/data/arms/)

*Energy Information Administration*

Short-Term Energy Outlook Query System [http://tonto.eia.doe.gov/STEO\\_Query/app/](http://tonto.eia.doe.gov/STEO_Query/app/)

*Environmental Protection Agency (Data & Software)*

[www.epa.gov/epahome/Data.html](http://www.epa.gov/epahome/Data.html)

*National Agricultural Statistics Service*

[www.usda.gov/nass/](http://www.usda.gov/nass/)

*National Center for Education Statistics*

Common Core of Data (CCD) 'Build a Table' Tool <http://nces.ed.gov/ccd/bat/>  
 The Data Analysis System (DAS) <http://nces.ed.gov/das/>  
 IPEDS College Opportunities On-Line – COOL <http://nces.ed.gov/ipeds/cool/>  
 National Assessment of Educational Progress [.../nationsreportcard/naepdata/](http://nces.ed.gov/ipeds/cool/.../nationsreportcard/naepdata/)

*National Center for Health Statistics (Center for Disease Control & Prevention)*

CDC Wonder <http://wonder.cdc.gov/>

*Nat'l Science Foundation (Division of Science Resources Statistics)*

SESTAT <http://srsstats.sbe.nsf.gov/>  
 WebCaspar <http://caspar.nsf.gov/>

... and from [www.fedstats.gov/policy/](http://www.fedstats.gov/policy/) FY2005.pdf

Table 3. Estimated Agency Purchases of Statistical Services, FY 2005 (In millions of dollars)

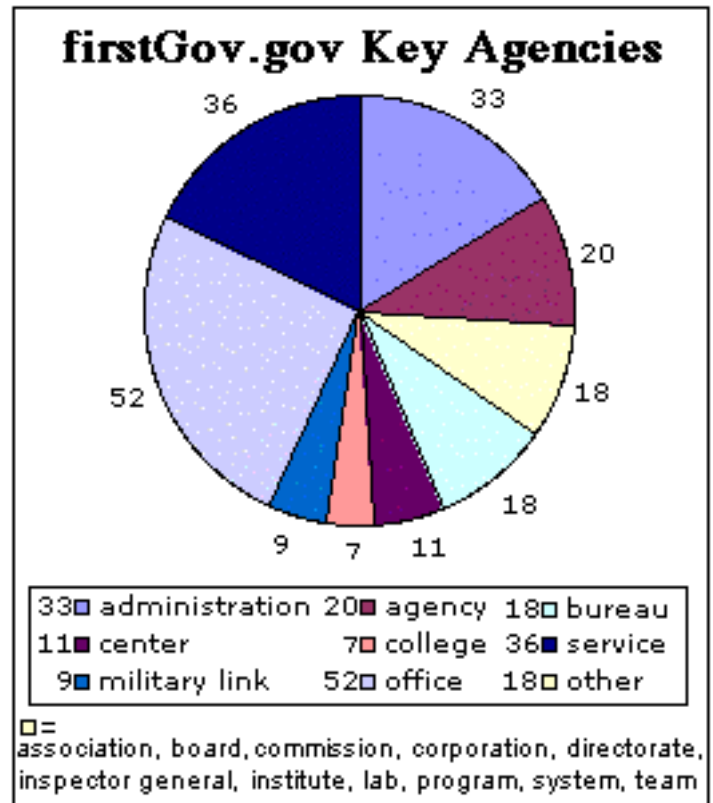
Agency Name	Direct Funding	Total Purchases	State/Local Gov't's	Private Sector	Other Federal Agencies
National Institutes of Health	802.7	409.3	0.0	387.8	21.5

Centers for Disease Control and Prevention (without NCHS)	375.0	249.6	120.3	111.4	18.0
Bureau of Labor Statistics	534.0	190.0	97.0	17.0	76.0
National Center for Education Statistics	186.5	184.9	2.0	167.6	15.3
National Center for Health Statistics	149.6	120.9	16.0	59.2	45.7
⋮	⋮	⋮	⋮	⋮	⋮
Agricultural Research Service	5.2	2.0	0.0	0.0	2.0
Total	4769.3	2095.8	424.0	1333.6	338.3

There are 474 links on the firstGov.gov “A-Z Agency” list (37 Administrations, 6 Advisories, 20 Authorities, ..., & 1 Trust).

Of the 15 Departments (Agriculture, Commerce, ..., Vet.Affairs), consider just the links to the State Department:

1. the Office of Arms Control & Int’l Security
  - a. the Bureau of Arms Control
  - b. the Bureau of Political-Military Affairs
  - c. the Bureau of Nonproliferation
  - d. the Bureau of Verification and Compliance
  - e. the Foster Fellows Visiting Scholars Program
2. the Office of Econ, Bus. and Ag. Affairs
  - a. the Bureau of Economic and Business Affairs
3. the Office of Management
  - a. the Bureau of Administration
  - b. the Bureau of Consular Affairs
  - c. the Bureau of Diplomatic Security
  - d. the Bureau of Human Resources
  - e. the Bureau of Info. Resource Management
  - f. the Bureau of Overseas Buildings Operations
  - g. the Director of Diplomatic Reception Rooms
  - h. the Foreign Service Institute
  - i. the Office of Management Policy
  - j. the Office of Medical Services
  - k. the White House Liaison
  - l. the Office of Rightsizing Overseas Presence
4. the Office of Political Affairs
  - a. the Bureau of African Affairs
  - b. the Bureau of E. Asian & Pacific Affairs
  - c. the Bureau of European & Eurasian Affairs
  - d. the Bureau of Near Eastern Affairs
  - e. the Bureau of S. Asian Affairs
  - f. the Bureau of W. Hemisphere Affairs
  - g. the Country Offices of the Department of State
  - h. the Bureau of International Organization Affairs (IO)
5. the Office of Global Affairs
  - a. the Bureau of Democracy, Human Rights, and Labor
  - b. the Bureau of International Narcotics and Law Enforcement Affairs
  - c. the Bureau of Population, Refugees, and Migration
  - d. the Office of International Women’s Issues
  - e. the Office of the Science & Technology Adviser
  - f. the Office to Monitor & Combat Trafficking in Persons
6. the Office of Public Diplomacy & Public Affairs
  - a. the Office of Policy, Planning & Resources for Public Diplomacy & Public Affairs



note: “Key Agencies” ≠ “A-Z Agency list”

- b. the Bureau of Public Affairs
- c. the Bureau of Educational & Cultural Affairs
- d. the Bureau of International Information Programs
- e. the Advisory Commission on Public Diplomacy
- 7. the Office of the U.S. Mission to the United Nations
  - a. the U.S. Ambassador to the U.N.
  - b. the U.S. Permanent Representative to the U.N.
  - c. the Ambassador for Management and Reform
  - d. the Ambassador for Special Political Affairs
  - e. the Ambassador to the Economic and Social Council

**Conclusion: there are hundreds of .gov bureaus (e.g. Census) which sometimes provide free data and sometimes charge for data.**

[www.statistics.com](http://www.statistics.com) “browse for data” section (2556 Links)

**(main topic Links) + (subtopic Links)**

**Agriculture (78+36=114)**

- Farming (24)
- Food Aid (6)
- Forestry (2)
- Historical (4)

**Business/Economics (140+382=522)**

- Commerce (20)
- Company Rankings (8)
- Currency (10)
- Emerging Markets (6)
- Employment (47)
- Entertainment (44)
- Fishing (3)
- Forestry (6)
- Historical (5)
- Housing (23)
- Income & Wealth (36)
- Industry (48)
- Insurance (3)
- Retail and Wholesale (38)
- Stock Exchange (34)
- Tourism (15)
- Trade (36)

**Crime (48+65=113)**

- Computer (5)
- Drugs (6)
- Justice (26)
- Juvenile (8)
- Law Enforcement (5)
- Regional Statistics (11)
- Terrorism (2)
- Victims (2)

**Education (39+43=82)**

- Adult (9)
- College/University (6)
- Libraries (7)
- Regional Statistics (21)

**Environment (27+92=119)**

- Air Pollution (11)
- Astronomy (1)
- Energy (25)

Hazardous Waste (9)

- Land Use (3)
- Natural Disasters (8)
- Oceanography (8)
- Water (3)
- Weather (13)
- Wildlife (11)

**Government (18+231=249)**

- Aid (10)
- Finance (70)
- Politics (40)
- Regional Statistics (19)
- Services (14)
- Social Statistics (78)

**Health (121+275=396)**

- Contraception & Sex Education (11)
- Diseases (62)
- Drinking Water (5)
- Drugs & Alcohol (31)
- Historical (1)
- Insurance (8)
- Life Expectancy (4)
- Nutrition (11)
- Pregnancy & Children (47)
- Regional Statistics (95)

**Market Research & Demographics (100+133=233)**

- Business (21)
- Geography (10)
- Homes (2)
- Language (3)
- Population (86)
- Religions (11)

**Military (5+54=59)**

- Air Forces (4)
- Arms Trade (3)
- Forces (14)
- Historical (5)
- Military Spending (10)
- Veterans (1)
- Warfare (17)

**Opinion Surveys (28+67=95)**

- Crime (4)
- Economics (4)

Health (9)  
 Politics (21)  
 Regional (6)  
 Science & Technology (4)  
 Social Issues (19)

**Science & Technology (9+31=40)**

Astronomy (4)  
 Careers (2)  
 Chemistry (3)  
 Physics (6)  
 Research & Development (16)

**Sports (26+267=293)**

Baseball (99)

Basketball (90)  
 Football (40)  
 Golf (10)  
 Hockey (7)  
 Olympic Games (5)  
 Soccer (16)  
 Women's Pro (0)

**Transportation & Communication (55+186=241)**

Airlines & Airports (10)  
 Automotive (55)  
 Internet (112)  
 Phone (3)  
 Railroads (6)

Sample [www.gallup.com](http://www.gallup.com) report.

February 16, 2004

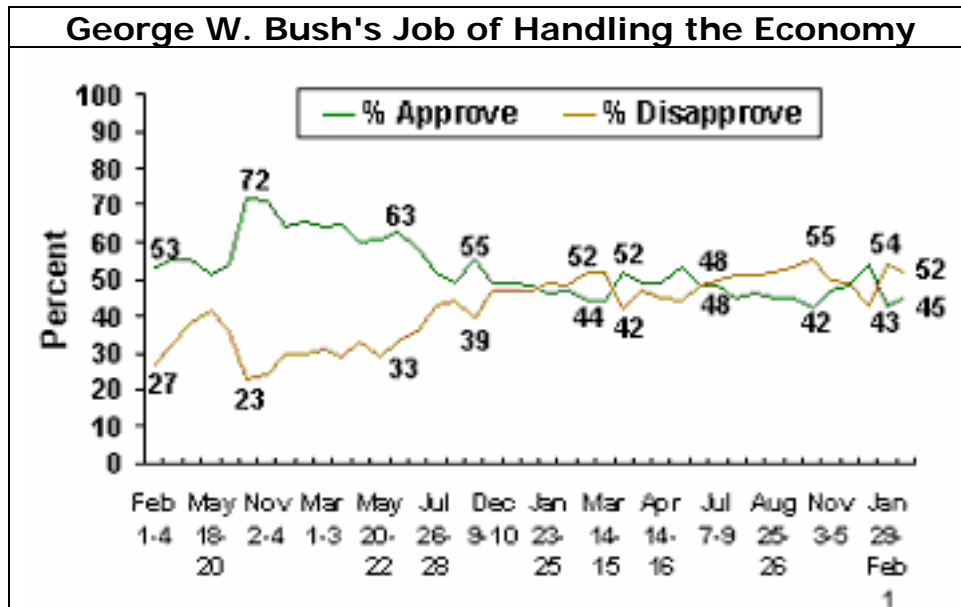
**Public Grows More Wary About Economy in February**  
 Percentage mentioning "jobs" as top problem at highest point in 10 years

by Dennis Jacobe

GALLUP NEWS SERVICE

**Approval of President's Handling of Economy Is Down**

Declining consumer confidence and increasing job fears have led to a reversal in the public's approval of the way Bush is handling the economy. In early January, 54% approved of Bush's performance on the economy and 43% disapproved. Today, only 45% approve, while 52% disapprove.



**Survey Methods**

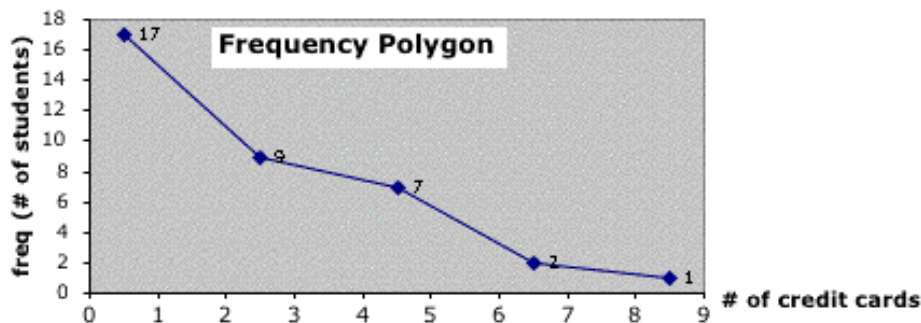
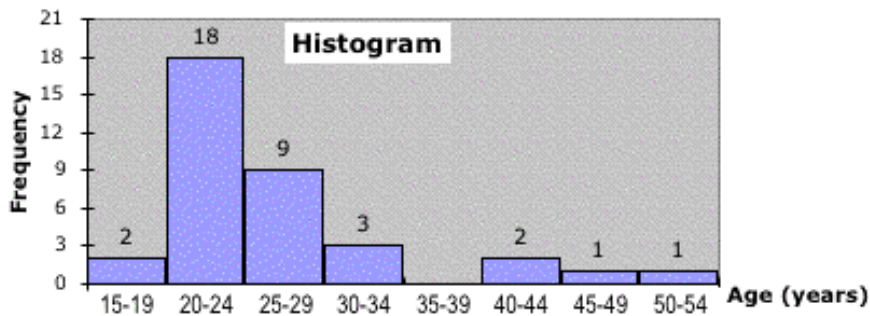
Results are based on telephone interviews with 1,029 national adults, aged 18 and older, conducted Jan. 2-5, 2004, and on telephone interviews with 1,002 national adults, aged 18 and older, conducted Feb. 9-12, 2004. For results based on these total samples, **one can say with 95% confidence** that the margin of sampling error is  $\pm 3$  percentage points. (Bold not in original)

Copyright © 2004 The Gallup Organization, Princeton, NJ. All rights reserved.

### Sample Results from Students' Data

ID #	meas. ht.	gender	eyes	age	reported ht.	# of credit cards	exercise	# of units	# hr/wk work	smoke	color blind	handed	diff.	smoke
05270	61.5	f	brown	20	61	3	n	17	32	n	n	l	0.5	3 y
08182	61.5	f	brown	20	62	0	y	15	20	n	n	r	-1	33 n
62332	65	f	green	20	64	0	n	12	36	n	n	r	1	-handed
11155	63.5	f	dark brown	21	63	2	y	13	24	y	n	r	0.5	1 ambi.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	3 left
.	.	.	.	.	.	.	.	.	.	.	.	.	.	32 rt

Age (yrs)		# credit cards		# academic units		gender	diff.
Mean	26.1	Mean	2.4	Mean	9.35	21 f	0.7
Median	24	Median	2	Median	9	15 m	0.6
Mode	20	Mode	1	Mode	7	36 total	
Standard Deviation	8.26	Standard Deviation	2.21	Standard Deviation	4.29		
Sample Variance	68.3	Sample Variance	4.88	Sample Variance	18.4		
Range	38	Range	9	Range	14		
Minimum	16	Minimum	0	Minimum	3		
Maximum	54	Maximum	9	Maximum	17		
Sum	939	Sum	88	Sum	337		
Count	36	Count	36	Count	36		



Some of the best **Google.com** results that I've found:

**An Astrologer who knows Statistics** [www.astrodatabank.com/AS/JoseBecerra.htm](http://www.astrodatabank.com/AS/JoseBecerra.htm)

... There is an important difference between causation and association. I do not think that the field of astrology is in the position of proving causation. ...even the most conclusive observational evidence does not absolutely prove causation. Only an experimental design (double-blind randomized trial) can tackle the issue of causation.

...What are adequate measures of association? Before considering them, I would like to point out the importance of choosing the “right questions.” Few people would pretend that astrological influences explain 100% of the occurrence of an outcome. Therefore, the right question would not be if planet A conjunct planet B in sector X would explain outcome C, but, accepting a multifactorial causation, how much weight should be attributed to astrological factors alone in the etiology of outcome C. The fact that astrological factors may have low relevance in certain outcomes does not disprove astrology. Such evidence rather qualifies the conditions on which astrological influences operate.

The measure of association most widely used – the one almost exclusively used in most astrological research – is the Chi square test. It is usually reported as a “p value”: the probability that the association found may be due to chance. A low p value (inferior to 0.05) indicates that the association found has less than 2 in 20 chances [sic] of being spurious. ...

**U.S. Newborn Name Popularity** [www.ssa.gov/OACT/babynames](http://www.ssa.gov/OACT/babynames)

	Male	Number	Female	Number	
		Rank name		males name	females
1	Jacob	30122	Emily	24262	
2	Michael	28119	Madison	21546	
3	Joshua	25859	Hannah	18559	
4	Matthew	24831	Emma	16324	
5	Ethan	21949	Alexis	15411	
6	Joseph	21766	Ashley	15217	
7	Andrew	21696	Abigail	15155	
8	Christopher	21676	Sarah	14564	
9	Daniel	21186	Samantha	14540	
10	Nicholas	21148	Olivia	14481	
	top-10 boys	237 842	top-10 girls	170 059	
	4 019 280 U.S. births		in 2002 re: cdc webpage		
	-3 847 199 SSNs issued		in 2002 re: ssa letter to mh		
	172 081 babies w/o SSNs		in 2002 (4.28%)		
	28 035 babies born in 2000 died w/in 1 year		re: cdc webpage		

**Regional Mean & SD Temperatures** Search for files: [clim85\\_TEMP01.pdf](#)  
 and [clim85\\_prctp02.pdf](#)

CA region	Apr. Mean Temp (°F)	Apr. StDev Temp. (°F)	July Mean Temp (°F)	July StDev Temp. (°F)
NE Interior	42.0	3.5	64.2	2.3
Sacramento Valley	52.2	3.3	72.4	2.1
N. Coast	52.4	2.6	66.9	1.4
Central Coast	55.6	2.2	65.7	1.3
San Joaquin Valley	56.4	3.2	76.6	2.1
S. Coast	58.0	2.6	71.7	1.8
SE Deserts	61.7	3.9	84.6	1.9

More **Google.com** results that I've found:

[www.epa.gov/history/topics/fuel/01.htm](http://www.epa.gov/history/topics/fuel/01.htm)

Vehicle Class	Highest Fuel Economy (mpg)	Lowest Fuel Economy (mpg)
Two Seater	Honda Insight 61/70	Ferrari 550 Maranello 8/13
Subcompact Car	Volkswagon New Beetle (diesel) 42/49	Ferrari 456 MGT/MGTA 10/15
Compact Car	Volkswagon Jetta/Golf (diesel) 42/49	BMW 540i (automatic trans.) 15/21
Minicompact Car	Audi TT Coupe 22/31	Aston Martin DB-7 Vantage Volante 11/18
Midsized Car	Mazda 626 26/32	BMW 740i, 740i Sport 15/21
Large Car	Toyota Avalon 21/29	BMW 750i, 750i Protection 13/20
Small Station Wagons	Suzuki Esteem Wagon 30/36	BMW 540i Sport Wagon 15/20
Midsized Station Wagons	Ford Focus Station Wagon 26/33	Audi A6 Avant Quattro 17/24
Sport Utility Vehicle (2WD)	Chevrolet Tracker/Suzuki Vitara 25/28	Dodge Durango 2 WD 12/17
Sport Utility Vehicle (4WD)	Chevrolet Tracker/Suzuki Vitara 25/27	Land Rover Range Rover 12/15
Minivan	Dodge Caravan/Plymouth Voyager 20/26	Volkswagon Eurovan 15/20
Small Pickup Truck	Chevrolet S10 Pickup/GMC Sonoma 23/29	Chevrolet S10 Pickup 2WD versions 17/22
Standard Pickup Truck	Ford Ranger Pickup/Mazda 2500 22/27	Dodge Dakota Pickup 4WD 12/16
Van, Cargo Type	Chevrolet Astro/GMC Safari 16/22	Dodge B1500 Van 2WD 12/17
Van, Passenger Type	Chevrolet Astro/GMC Safari 16/20	Dodge B2500 Wagon 2WD 12/15

[www.the-numbers.com/charts/thisweek.html](http://www.the-numbers.com/charts/thisweek.html) (top 10 out of 81) movies

new rank	old rank	(Dec 12, 2003) movie name	Gross this week	change	# of theaters	\$ per theater	total gross	Days
1	(new)	Something's Gotta Give	\$16,064,723		2,677	\$6,001	\$ 16,064,723	3
2	-1	Last Samurai, The	\$14,087,074	-41.96%	2,908	\$4,844	\$ 46,874,330	10
3	(new)	Stuck On You	\$ 9,411,055		3,003	\$3,134	\$ 9,411,055	3
4	(new)	Love Don't Cost a Thing	\$ 6,315,311		1,844	\$3,425	\$ 6,315,311	3
5	-3	Haunted Mansion, The	\$ 6,139,023	-34.65%	3,001	\$2,046	\$ 53,749,464	19
6	-4	Elf	\$ 6,017,341	-25.03%	2,876	\$2,092	\$147,507,398	38
7	-6	Bad Santa	\$ 6,012,550	-14.28%	2,540	\$2,367	\$ 35,715,007	19
8	-2	Honey	\$ 4,860,975	-62.19%	1,972	\$2,465	\$ 19,776,370	10
9	-5	Cat in the Hat, The	\$ 4,166,590	-41.66%	2,955	\$1,410	\$ 90,728,185	24
10	-7	Gothika	\$ 2,725,221	-48.09%	1,806	\$1,509	\$ 53,933,915	24

[www.infoplease.com/ipa/A0884486.html](http://www.infoplease.com/ipa/A0884486.html) (top 10 only)

Rank	Place name	Population		Change (1990 to 2000)	
		April 1, 2000	April 1, 1990	Number	Percent
1.	Augusta-Richmond County, Ga.	199,775	44,639	155,136	347.5%
2.	Gilbert, Ariz.	109,697	29,188	80,509	275.8
3.	Vancouver, Wash.	143,560	46,380	97,180	209.5
4.	Henderson, Nev.	175,381	64,942	110,439	170.1
5.	North Las Vegas, Nev.	115,488	47,707	67,781	142.1
6.	Athens-Clark County, Ga.	101,489	45,734	55,755	121.9
7.	Peoria, Ariz.	108,364	50,618	57,746	114.1
8.	Pembroke Pines, Fla.	137,427	65,452	71,975	110.0
9.	Chandler, Ariz.	176,581	90,533	86,048	95.0
10.	Las Vegas, Nev.	478,434	258,295	220,139	85.2

The last page of **Google.com** results that I've found:

www.cdc.gov search for “anthropometric” data tables

Table 1. Weight in pounds for persons 3 months-19 years - of examined persons, mean, standard error of the mean, and selected percentiles, by sex and age: United States, 1988-1994

Sex and age	Number of examined persons	Mean	Standard error of the mean	Selected percentile									
				5th	10th	15th	25th	50th	75th	85th	90th	95th	
<b>Male</b>													
3-5 months .....	290	16.1	0.21	*	13.9	14.3	14.7	16.1	17.3	18.0	18.8	*	
6-8 months .....	320	19.3	0.24	*	16.3	16.8	17.5	19.1	20.8	22.0	22.8	*	
9-11 months .....	277	21.1	0.23	*	18.2	18.6	19.4	21.1	22.6	23.3	24.0	*	
1 year .....	663	25.4	0.23		20.5	21.2	21.9	23.2	25.3	27.2	28.3	29.7	31.2
2 years .....	645	29.9	0.24		24.9	25.9	26.6	27.5	29.7	31.8	33.1	34.7	35.9
3 years .....	516	34.7	0.45		28.2	29.2	30.1	31.7	34.0	37.3	38.5	39.5	41.1
4 years .....	549	38.8	0.41		31.6	33.1	33.9	35.2	38.2	41.5	43.3	45.2	48.0
⋮													
19 years ...													
<b>Female</b>													
3-5 months .....	308	14.8	0.20	*	12.4	12.9	13.3	14.6	16.2	16.9	17.4	*	
6-8 months .....	264	17.9	0.28	*	*	15.8	16.4	17.4	19.0	19.9	*	*	
9-11 months .....	315	19.8	0.21	*	17.1	17.5	18.0	19.6	21.2	22.2	22.9	*	
1 year .....	647	23.9	0.21		19.6	20.1	20.8	21.7	23.4	25.9	27.3	28.2	29.4
2 years .....	624	29.0	0.27		23.7	24.5	25.0	26.3	28.7	31.2	32.7	33.9	35.8
3 years .....	587	33.9	0.35		27.3	28.7	29.5	30.8	33.3	36.1	38.1	39.1	41.9
4 years .....	537	39.3	0.59	*	31.9	33.1	34.8	38.1	41.8	44.1	45.4	*	
⋮													
19 years ...													

Note: pregnant women are excluded.

the Insurance Institute for Highway Safety

[www.hwysafety.org/safety\\_facts/fatality\\_facts.motorcyl.htm](http://www.hwysafety.org/safety_facts/fatality_facts.motorcyl.htm)

Month	Percent of U.S. 2002 motorcycle deaths		Total U.S. deaths 2002
January	male	female	
January	3	1	4
February	4	5	9
March	145	17	162
April	421	32	453
May	371	22	393
June	352	29	381
July	327	54	381
August	338	44	382
September	331	39	370
October	271	33	304
November	159	19	178
December	60	6	66
<b>Total</b>	<b>2,860</b>	<b>302</b>	<b>3,162</b>

Time of Day	% of m-cycle deaths		Total
5:00 am – 9 am	31	7	38
9:00 am – 3 pm	25	23	48
3:00 pm – 9 pm	14	41	55
9:00 pm – 3 am	3	27	30

## Sellers/Harbison "A First Course in Statistics": Data Set on Cancer

"While the risks of lung cancer associated with the use of tobacco products is well established, less than 20% of heavy smokers will develop lung cancer in their lifetime. To examine the role of host factors in the pathogenesis of the disease, a study was conducted in southern Louisiana in the early 1980's to determine if relatives of lung cancer patients demonstrated higher rates of lung cancer than could be expected. The findings of this research, conducted by Dr. Wee Lock Ooi and Dr Henry Rothschild at Louisiana State University, were published in the *Journal of the National Cancer Institute* in 1986 (vol 76, pages 217-222). The data suggested that some families demonstrate a predisposition for lung cancer (a 2.4-fold greater risk after allowing for personal smoking habits). Additional analyses by Dr. Thomas A. Sellers demonstrated that this familiar risk extended to cancers at other sites (*American Journal of Epidemiology*, 1987; 126:237-246), and that the pattern of lung cancer in these families was consistent with Mendelian inheritance of a gene that produced an earlier age of onset lung cancer (*Journal of the National Cancer Institute*, 1990; 82:1272-1279). The data found in this textbook represent a randomly selected subset of the total study sample, and were kindly provided by Dr. Henry Rothschild."

<b>Parameters</b>	N = 2,538 data points			
<u>Age Data</u>	<u>Sex Data</u>		<u>Cancer Data</u>	
$\mu = 54.9$ years	1281 Males	$P_m = 50.5\%$	374 have cancer	$P_c = 14.7\%$
/				
Men				
\				
$\sigma = 18.2$ years	1257 Females	$P_f = 49.5\%$	2164 do not	$P_n = 85.3\%$
$\mu = 57.1$ years				
/				
Women	<u>Smoker Data</u>			
\				
$\sigma = 19.6$ years	1319 Smokers	$P_s = 52.0\%$		
	1219 Non-smokers	$P_n = 48.0\%$		

For men and women:

$\mu = 56.2$  years  
 $\sigma = 18.76$  years

To obtain a random sample, we suggest using random numbers. Some examples and exercises in Chapter One outline a procedure for using the table of random numbers in the appendix for obtaining the random sample. Unfortunately, the random numbers greater than 2600 cannot be paired with specific data points in the data set. As a consequence, many of the random numbers must be eliminated.

A calculator can also be used to obtain random numbers, but care must again be exercised to eliminate numbers that exceed the domain of the data set.

A "quick and dirty method" that is not as scientific, but actually as reliable as the random number method is the "flip and stab method." Namely:

- Step 1.** "Flip" to a page in the data set.
- Step 2.** With eyes shut, "stab" a data point with your index finger.
- Step 3.** Extract the needed information from the "randomly selected" data point.

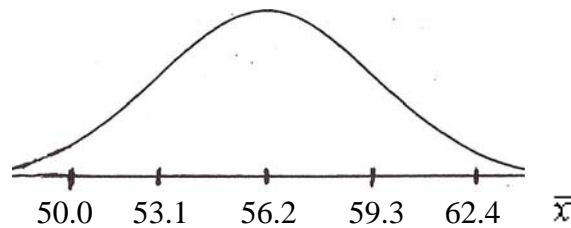
To get a reasonably small standard error, the sample size should be as large as possible. Most students will accept a sample size around 75 as not excessively large. Such a sample size will

yield a standard error in the 5% - 6% range.

### ACTIVITIES USING THE DATA SET

1. Have students sample  $n = 36$  data points. For the sampled data, compute  $\bar{x}$  and  $s$ .

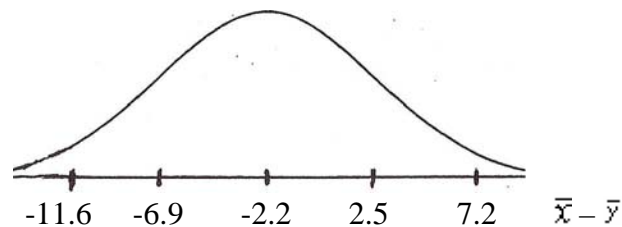
For the population of ages,  $\mu \approx 56.2$  and  $\sigma \approx 18.76$ . Thus,  $\mu_{\bar{x}} = 56.2$  and  $\sigma_{\bar{x}} = 3.12... \approx 3.1$ .



List the values of  $\bar{x}$  and  $s$  on the board, and circle the data points within one standard error (i.e. between 53.1 and 59.3). Based on CLT and standard normal distribution, about 68% of all  $\bar{x}$ 's should be on the interval. Furthermore, about 95% of all  $\bar{x}$ 's should be between 50.0 and 62.4. Only 5% of the  $\bar{x}$ 's should be "deviates," outside the two standard error interval.

2. As an added bonus, have each student compute an estimated standard error using their individual sample standard deviations. This activity shows that  $s$  can be used to approximate an unknown  $\sigma$  in applied problems.
3. Have students use their sample data ( $\bar{x}$  and  $s$ ) to construct confidence intervals for  $\mu = 56.2$  years. When the results are checked for the entire class, approximately  $\alpha\%$  of the intervals will not contain the (known) mean. This activity emphasizes the meaning of a confidence interval; namely, that a 90% confidence will correctly contain an unknown mean about 90% of the time, and will fail to contain the mean about 10% of the time. There is no probability associated with a confidence coefficient.
4. Have students use their sample data ( $\bar{x}$  and  $s$ ) to test  $H_0: \mu = 56.2$  years. Use any alternative hypothesis and  $\alpha$  that you want to. For the entire class, a certain percent will reject  $H_0$ , even though  $H_0$  is known to be true. Thus each erroneous rejection of  $H_0$  will result in a Type I error.
5. There are 1281 males and 1257 females in the Data Set. To investigate the  $\bar{X} - \bar{Y}$  random variable distribution, have students take two large samples of both genders. For example, let  $n_1 = 32$  males and  $n_2 = 34$  females, with  $\bar{x}$  = (mean sample age for males) and  $\bar{y}$  = (mean sample age for females). For the random variable  $\bar{X} - \bar{Y}$ :

$$\mu_{\bar{x}-\bar{y}} = 54.9 - 57.1 = -2.2 \text{ yrs} \quad \& \quad \sigma_{\bar{x}-\bar{y}} = \sqrt{18.2^2 \div 32 + 19.6^2 \div 34} = 4.65... \approx 4.7 \text{ years.}$$

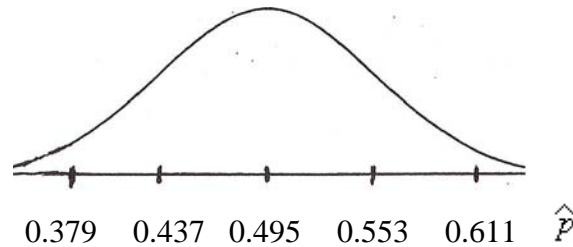


For the class, about 68% of  $(\bar{x} - \bar{y})$ 's will be on the interval  $(-6.9, 2.5)$ , and 95% will be on the interval  $(-11.6, 7.2)$ . Furthermore, each student can use their sample

data to construct confidence intervals for  $\mu_1 - \mu_2$  or carry out a hypothesis test for  $H_0: \mu_1 - \mu_2 = 0$ . Results typically follow the predicted results based on the confidence coefficient or level of significance.

6. Have students sample 75 data points and compute  $\hat{P}_f$ , the proportion of females in the sample. For the probability distribution of  $\hat{P}$ :

$$\mu_{\hat{P}} = 0.495 \quad \& \quad \sigma_{\hat{P}} = 0.0577\dots \approx 0.058$$



Compare the class results with the predicted results, namely:

- 68% of the  $\hat{P}_f$ 's within one standard error
- 95% of the  $\hat{P}_f$ 's within two standard errors.

In addition, have students compute the p-value for a test of  $H_0: P_f = 0.495$  versus your choice of a suitable alternative hypothesis.

7. As an alternative activity to #6, suppose that  $P_f$  (female) is unknown, then use the sample results to construct a confidence interval. For the class results, count the number of intervals that failed to contain the actual value of 0.495 for  $P_f$ .

Or test  $H_0: P_f = 0.495$  using the sample results. Choose any level of significance (your choice), then check the number of students who committed a type-I error with the predicted number based on the level of significance used on the test.

8. Repeat the activities of #6 & #7 using the “smoker” data, instead of “female” data.
9. Repeat the activities of #6 & #7 using the “cancer” data, instead of “female” data.
10. Activities suggested on #5 (age differences, by gender) for the random variable  $\bar{X} - \bar{Y}$  can also be carried out for  $\hat{P}_1 - \hat{P}_2$  for gender differences using any one of the characteristics examined by the data. Two sets of data would now be needed and suitable sample sizes to keep confidence intervals within reasonable lengths. Additionally, the sample data can be used to test  $H_0: P_1 - P_2 = 0$  versus any alternative hypothesis and any level of significance.