SOC 3811:
BASIC SOCIAL STATISTICS

John Robert ("Rob") Warren

Tuesdays 6:20-8:50pm

Scientific Research

Reality → Data → Knowledge

2954462187388221348
8711102682140050851
9291758426493659357
1797402968538728539
4401671369193970834
0165785879715135279
6365488407871392124
2920353540363160907
9546090581563252173
8998220343490929571
8159244365888845385
0229635205640107762
4385245611536981630
9196626803867353805
4956868092019177602
1078342587201850530
2250228297490991053
5575276026995705920
9224813120745604486
7096379166974664461
5350989184936109354

Health Research

Reality → Data → Knowledge

2954462187388221348
8711102682140050851
9291758426493659357
1797402968538728539
4401671369193970834
0165785879715135279
6365488407871392124
2920353540363160907
9546090581563252173
8998220343490929571
8159244365888845385
0229635205640107762
4385245611536981630
9196626803867353805
4956868092019177602
1078342587201850530
2250228297490991053
5575276026995705920
9224813120745604486
7096379166974664461
5350989184936109354

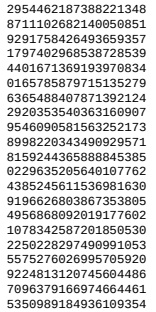
Health Research

Reality → Data → Knowledge

2954462187388221348
8711102682140050851
9291758426493659357
1797402968538728539
4401671369193970834
0165785879715135279
6365488407871392124
2920353540363160907
9546090581563252173
8998220343490929571
8159244365888845385
0229635205640107762
4385245611536981630
9196626803867353805
4956868092019177602
1078342587201850530
2250228297490991053
5575276026995705920
9224813120745604486
7096379166974664461
5350989184936109354

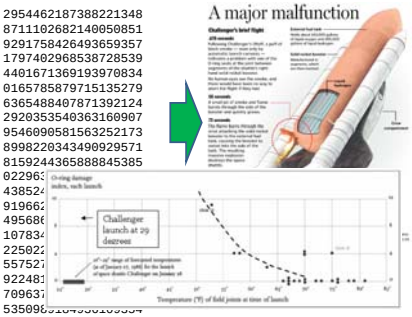
Consumer Research

Reality | Data | Knowledge



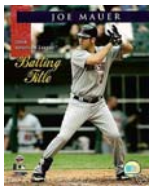
Aerospace Research

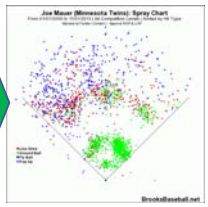
Reality | Data | Knowledge



Sports

Reality | Data | Knowledge



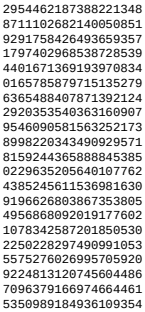


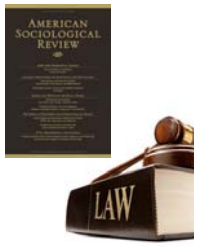
Business

Reality | Data | Knowledge

Social Sciences

Reality → Data → Knowledge





Reality → Data

Reality → Data

Unit of Observation
The type of object that we observe
(e.g., people, countries, classrooms)

Population
All of the individuals about whom we wish to draw conclusions

Sample
A selected subset of individuals from the population

Sampling
The process we use to select a sample from the population

Variable
Any attribute that can differ (or vary) from individual to individual

Worksheet

We want to study how eligible voters in our congressional district will vote in the upcoming election. There are 125,000 eligible voters. We randomly select 1,000 of them and mail them a survey that asks them who they will vote for in the presidential election; 161 people return those surveys.

1. What is the unit of observation?
2. What is the population?
3. What is the sample?
4. What is the sampling procedure?
5. What key variable(s) did we measure?

Worksheet

We want to understand how the recent wildfires in Colorado have affected moose living in Rocky Mountain National Park. To find out, we draw blood from the first 10 moose we find inside the park, and we measure the blood's cortisol level as a way to assess how stressed the moose are.

1. What is the unit of observation?
2. What is the population?
3. What is the sample?
4. What is the sampling procedure?
5. What key variable(s) did we measure?

Reality ➔ Data

Measurement

The process of ascertaining the value of a variable for an individual

Validity

The degree to which a measure of a concept is accurate

Reliability

The degree to which a measure of a concept is consistent

Reality ➔ Data

How do we turn a measure into data?

Concept: Wealth

Measure: Ask people for dollar value of all savings and investments

Possible Values: $0, $1, $2, $3, … to ∞

Data: 0 = $0
1 = $1
2 = $2
etc

← Values are in a logical order and the assigned values have a real numeric meaning

Reality ➔ Data

How do we turn a measure into data?

Concept: Healthiness

Measure: Ask people whether they are in excellent, good, fair, or poor health

Possible Values: excellent, good, fair, poor

Data: 1 = Excellent
2 = Good
3 = Fair
4 = Poor

← Values are in a logical order but the assigned values have no numeric meaning

Reality ➔ Data

How do we turn a measure into data?

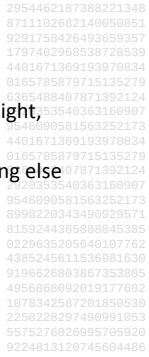
Concept: Sexual orientation

Measure: Ask people whether they are gay, straight, bisexual, or something else

Possible Values: gay, straight, bisexual, something else

Data: 1 = Gay
2 = Straight
3 = Bisexual
4 = Something else

Values are not in a logical order and the assigned values have no numeric meaning



Reality ➔ Data

Continuous Variables

Variables that categorize observations in terms of their numeric value of some attributes

In theory, continuous variables can take on all possible numerical values in a given interval

(e.g., IQ, wealth in dollars, height in inches)

Reality ➔ Data

Discrete Variables

Variables that categorize observations in terms of qualitative, non-numeric attributes

Nominal Variables

Discrete variables in which the categories cannot be put in order

(e.g., favorite color, state of birth)

Ordinal Variables

Discrete variables in which the categories can be put in order

(e.g., highest degree, level of agreement with a statement)

Reality ➔ Data

Variable Has An *Infinite* Number of Possible Values

➔It's a Continuous Variable

Variable Has a *Finite* Number of Possible Values…

➔ It's a Discrete Variable

…and the categories *can* be put in order.

➔ The Discrete Variable is Ordinal

…and the categories *cannot* be put in order.

➔ The Discrete Variable is Nominal

Worksheet

Are the following variables continuous or discrete? If discrete, are they nominal or ordinal?

1. Height
2. Hair color
3. Number of own children
4. Social class
5. Day of the week
6. Ever been arrested

Reality ➔ Data

Data file

The set of numeric values for each variable and for each observation

For nominal and ordinal variables, categories must be assigned (ultimately arbitrary) numeric values

Documentation (or **Codebook**)

A description of how units were sampled from the population, how variables are defined and coded, how measurements were made for each variable, how the data file is organized, etc.

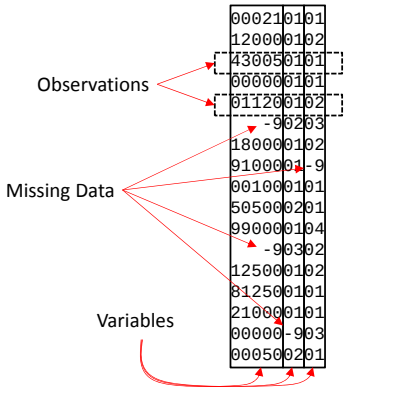
Missing data

A situation in which a numeric value is not available (for whatever reason) for a particular variable for a particular case

Reality ➔ Data

```
000210101
120000102
430050101
000000101
011200102
   -90203
180000102
910000102
001000101
505000201
990000104
   -90302
125000102
812500101
210000101
00000-903
000500201
```

Reality ➔ Data



Reality ➔ Data

Some Questions for
the Documentation:

1. What do rows and columns represent?
2. How were individuals selected?
   (Sampling procedure; Population covered)
3. How were variables measured?
   (Continuous? Categorical? Ordinal?
   What do values mean? Name of variables?)
4. Missing data?
   (How generated? What do values mean?)
5. How many observations should there be?
6. Measures valid? Reliable?
7. More…

```
000210101
120000102
430050101
000000101
011200102
   -90203
180000102
910000102
001000101
505000201
990000104
   -90302
125000102
812500101
210000101
00000-903
000500201
```

Reality ➔ Data

Statistical Software

Software that allows the researcher to read and manipulate a data file and to extract information from the data

Syntax

Computer programming code that instructs statistical software as to how to manipulate a data file and what statistical information to extract from the data

Output

Information produced as the result of applying syntax to a data file

Data ➔ Knowledge

How do we use data to…

…describe a sample?
…make inferences about populations?
…test existing theory or evaluate existing policy?
…inform the development of new theory or policy?

These questions motivate Sociology 3811 (and 5811 and 8811 and parallel courses elsewhere)

SYLLABUS

Worksheet

What questions do you have for me about this course, about its content, or about how it will be run?

BREAK

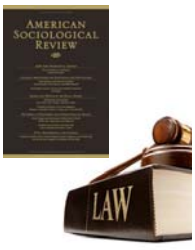
Data ➔ Knowledge

Reality

Data

```
29544621873882213 48
87111026821400508 51
92917584264936593 57
17974029685387285 39
44016713691939708 34
01657858797151352 79
63654884078713921 24
29203535403631609 07
95460905815632521 73
89982203434909295 71
81592443658888453 85
02296352056401077 62
43852456115369816 30
91966268038673538 05
49568680920191776 02
10783425872018505 30
22502282974909910 53
55752760269957059 20
92248131207456044 86
70963791669746644 61
53509891849361093 54
```

Knowledge



Distributions

Nominal and ordinal variables

"Distributions" describe how many cases are in each category of the variable

Continuous variables

"Distributions" describe how many cases have specific numeric values for that variable

Distributions

Distribution of the variable "health" (based on responses to survey question about overall health):

Value	Times Observed
Excellent	20
Good	10
Fair	4
Poor	1
Total	*35*

Distributions

Frequency distribution

A table that lists how frequently each value of a variable occurs in the data

Percentage distribution

A table that lists the percentage of times each value of a variable occurs in the data

Distributions

Distribution of the variable "health" (based on responses to survey question about overall health):

Value	Frequency	Percentage
Excellent	20	57%
Good	10	29%
Fair	4	11%
Poor	1	3%
Total	*35*	*100%*

Distributions

Question:

What if a continuous variable has too many individual values to make a frequency distribution table practical (or intelligible)?

Grouped distribution

A grouped distribution is a table that reports how frequently *ranges* of values occur

Distributions

Distribution of the variable "birth weight in grams:"

Birth Weight	Frequency	Percentage
499 or less	6,874	0.2%
500 - 999	24,527	0.6%
1000 - 1499	32,821	0.8%
1500 - 1999	68,940	1.6%
2000 - 2499	221,171	5.1%
2500 - 2999	797,339	18.5%
3000 - 3499	1,685,935	39.1%
3500 - 3999	1,143,273	26.5%
4000 - 4499	286,041	6.6%
4500 - 4999	40,188	0.9%
5000 – or more	4,534	0.1%

Distributions

Cumulative frequency distribution

A table that lists the number of cases in the data that fall at or below each value of a variable

Cumulative percentage distribution

A table that lists the percentage of cases in the data that fall at or below each value of a variable

Note: Cumulative distributions don't make sense for nominal variables (because the ordering of the values is arbitrary)

Distributions

Birth Weight	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
499 or less	6,874	0.2%	6,874	0.2%
500 - 999	24,527	0.6%	31,401	0.7%
1000 - 1499	32,821	0.8%	64,222	1.5%
1500 - 1999	68,940	1.6%	133,162	3.1%
2000 - 2499	221,171	5.1%	354,333	8.2%
2500 - 2999	797,339	18.5%	1,151,672	26.7%
3000 - 3499	1,685,935	39.1%	2,837,607	65.8%
3500 - 3999	1,143,273	26.5%	3,980,880	92.3%
4000 - 4499	286,041	6.6%	4,266,921	99.0%
4500 - 4999	40,188	0.9%	4,307,109	99.9%
5000 – or more	4,534	0.1%	4,311,643	100.0%

Table 4. Frequency Distribution of *TS* Articles by *TS* Contributors, 1990-1999 (N=442)

Number of Publications	Frequency	Percent
1	380	86.0
2	47	10.6
3	8	1.8
4	3	0.7
5	4	0.9

Note: The count includes authored and co-authored articles.

Unit of Observation: People who published an article in *Teaching Sociology* during the 1990s
Variable: Number of articles that they published in *Teaching Sociology* during the 1990s

Source: Marx, Jonathan Marx and Douglas Eckberg. 2005. "Teaching Scholarship during the 1990s: A Study of Authorship in *Teaching Sociology*." *Teaching Sociology* 33,:252-262

Table 4: Frequency Distribution of Variables

Variables	Frequency	Proportion (%)
Response Variable:		
Group		
Activists	49	4.12
Opportunists	56	4.71
Followers	47	3.96
Uncommitted	295	24.83
Non-participants	741	62.37
Predictor Variables:		
Familiarity of U.S.-China Relations		
Never Heard Of	32	2.66
Just Heard Of	237	19.67
Somewhat Familiar	473	39.25
Familiar	349	28.96
Very Familiar	114	9.46
Read *China Can Say No*		
Yes	463	38.33
No	745	61.67

Unit of Observation: Chinese college students
Variable: Various

Source: Zhiyuan Yu and Dingxin Zhao. 2006. "Differential Participation and the Nature of a Movement: A Study of the 1999 Anti-U.S. Beijing Student Demonstrations." *Social Forces* 84: 1755-1777

Table 1: Frequency Distribution of Petitions Filed under the Optional Protocol, 1976-1999

Petitions per Year	Frequency	Percent	Cumulative Percent
0	925	77.34	77.34
1	149	12.46	89.80
2	56	4.68	94.48
3	23	1.92	96.40
4	11	.92	97.32
5	11	.92	98.24
6	7	.59	98.83
7	2	.17	99.00
8	1	.08	99.08
9	1	.08	99.16
10	1	.08	99.25
11	1	.08	99.33
13	3	.25	99.58
14	1	.08	99.67
16	1	.08	99.75
19	1	.08	99.83
25	1	.08	99.92
30	1	.08	100.00

Note: "Frequency" gives the number of country-year observations.

Unit of Observation: Country-years (82 countries observed between 1976 and 1999; n=1,081)
Variable: Number of petitions to the U.N. Human Rights Committee

Source: Cole, Wade. 2006. "When All Else Fails: International Adjudication of Human Rights Abuse Claims, 1976-1999." *Social Forces* 84: 1909-1935

Worksheet

I surveyed 20 graduate students and asked them how many pets they own. Eight of them had no pets at all. Six had just one pet, five had two pets, and one had four

1. What is the unit of observation?
2. What is the variable?
3. Is the variable discrete or continuous? If discrete, is it ordinal or nominal?
4. Make a frequency distribution for this variable
5. Make a percentage distribution for this variable
6. Make a cumulative percentage distribution for this variable

Graphs

Visual images … graphs … often convey information better than tables of numbers

Which graph or visual image is "best" for depicting a distribution depends on the number of values of the variable, the type of variable it is, etc.

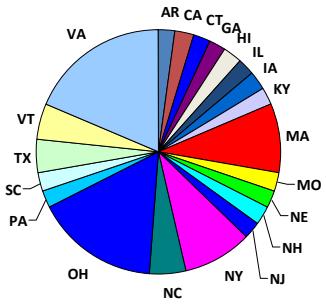
Graphs

State of birth of U.S. Presidents

State	Freq.	State	Freq.	State	Freq.
Arkansas	1	Kentucky	1	North Carolina	2
California	1	Massachusetts	4	Ohio	7
Connecticut	1	Missouri	1	Pennsylvania	1
Georgia	1	Nebraska	1	South Carolina	1
Hawaii	1	New Hampshire	1	Texas	2
Illinois	1	New Jersey	1	Vermont	2
Iowa	1	New York	4	Virginia	8

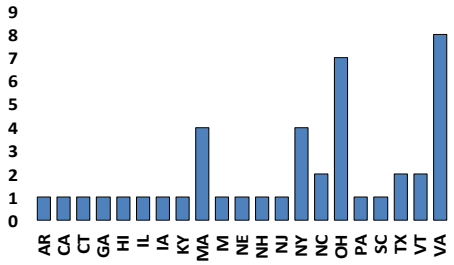
Graphs

State of birth of U.S. Presidents



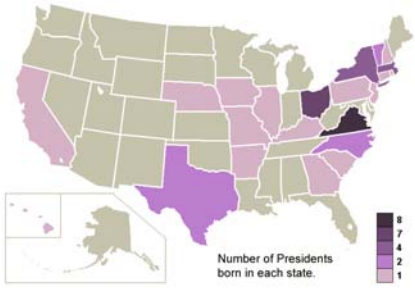
Graphs

State of birth of U.S. Presidents



Graphs

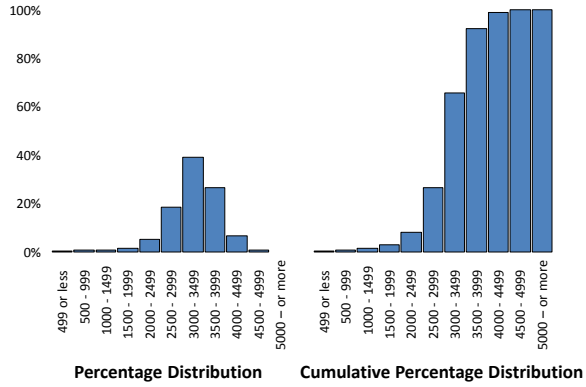
State of birth of U.S. Presidents



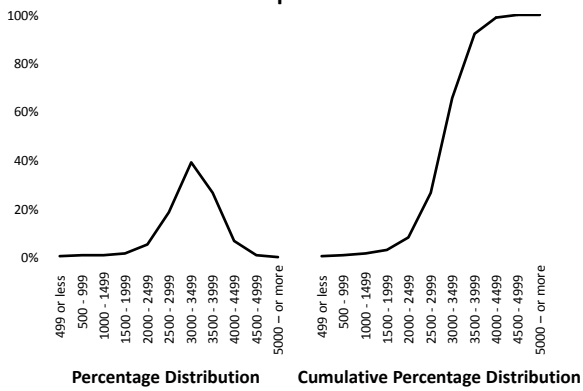
Graphs

Birth Weight	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
499 or less	6,874	0.2%	6,874	0.2%
500 - 999	24,527	0.6%	31,401	0.7%
1000 - 1499	32,821	0.8%	64,222	1.5%
1500 - 1999	68,940	1.6%	133,162	3.1%
2000 - 2499	221,171	5.1%	354,333	8.2%
2500 - 2999	797,339	18.5%	1,151,672	26.7%
3000 - 3499	1,685,935	39.1%	2,837,607	65.8%
3500 - 3999	1,143,273	26.5%	3,980,880	92.3%
4000 - 4499	286,041	6.6%	4,266,921	99.0%
4500 - 4999	40,188	0.9%	4,307,109	99.9%
5000 – or more	4,534	0.1%	4,311,643	100.0%

Graphs



Graphs



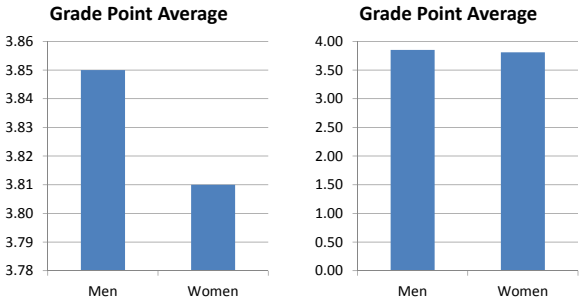
Graphs

Graphical representations of distributions can be misleading, either intentionally or unintentionally

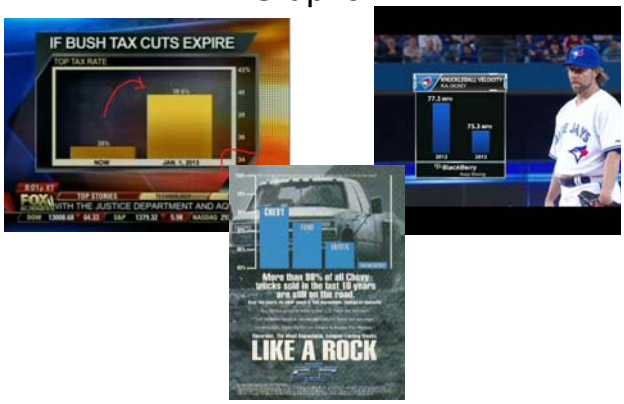
Most of the time, misleading graphs are misleading because of visual design elements

Always think critically about graphs

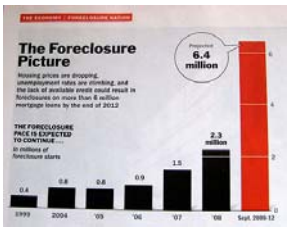
Graphs

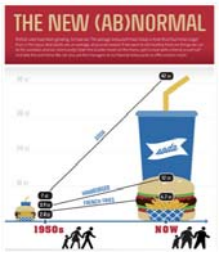


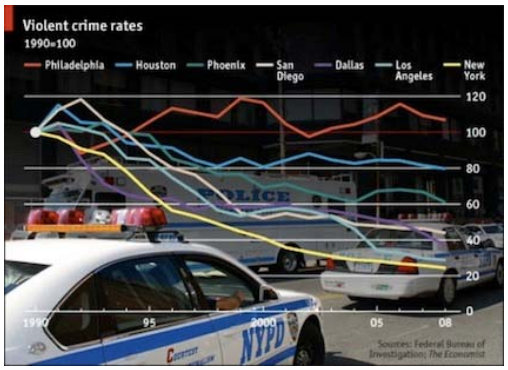
Graphs



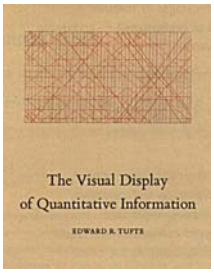
Graphs

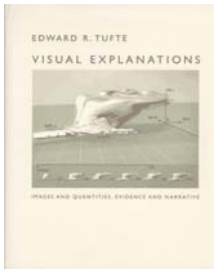






Graphs



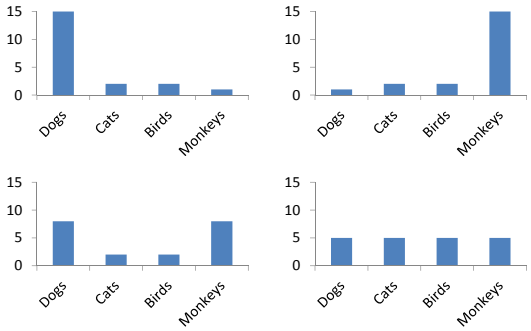


Depicting Distributions

What can we learn from graphical displays of the distribution of **nominal** variables?

1. Where are most of the cases located in the distribution?
2. How much variability is there?

Hypothetical distributions of the nominal variable "People's Favorite Pets"



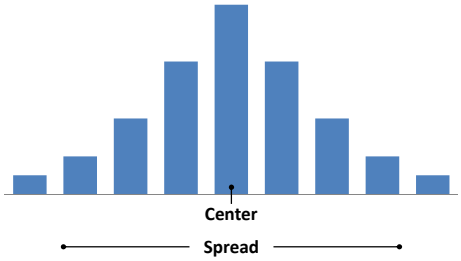
Depicting Distributions

What can we learn from graphical displays of the distribution of ordinal or continuous variables?

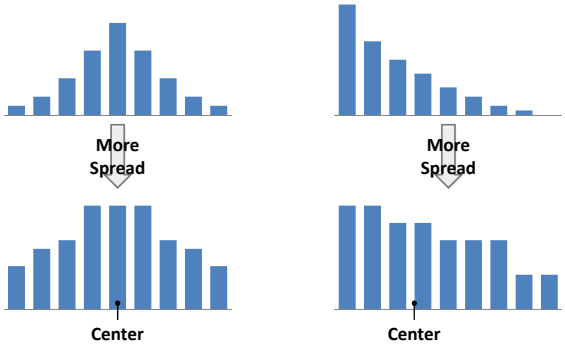
1. Where is the center of the distribution?
2. How much spread is there around that center?
3. Are there outliers?
4. Is the distribution symmetric or skewed?

Later we will discuss numeric summaries of these aspects of distributions, but never underestimate how much can be learned by looking at pictures

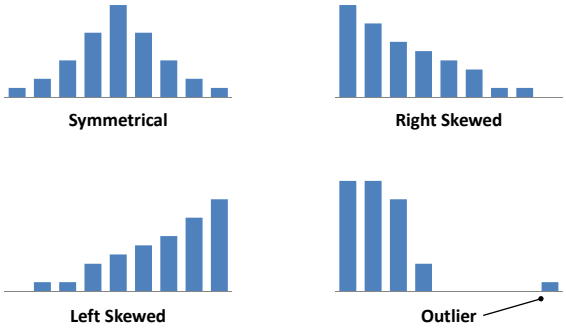
Depicting Distributions

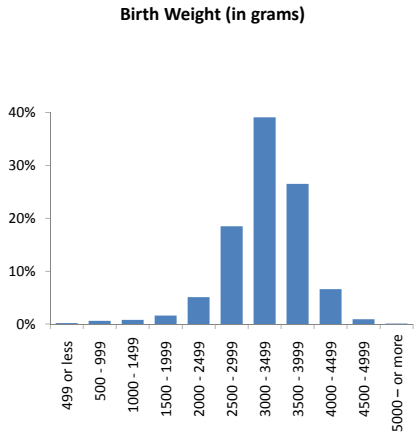


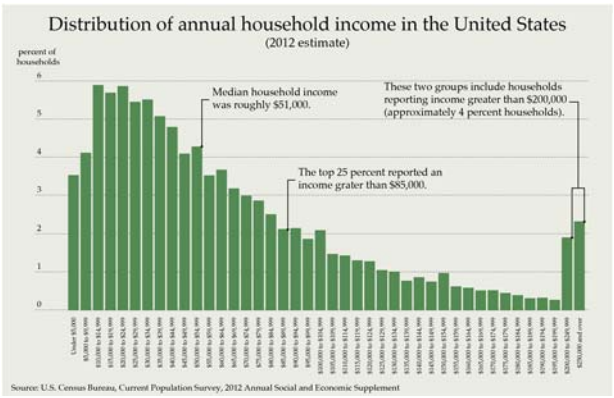
Depicting Distributions

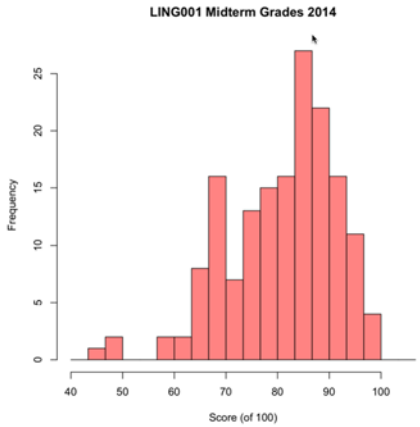


Depicting Distributions



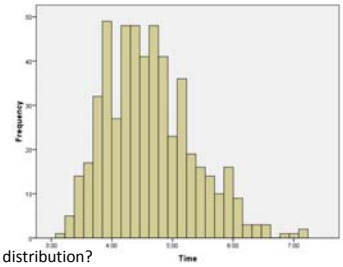






Worksheet

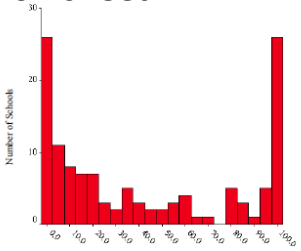
Finish Times for Women in the 2014 Minneapolis Marathon



1. Where is the <u>center</u> of this distribution?
2. What can you say about how much <u>spread</u> there is around that center?
3. Are there <u>outliers</u>?
4. Is the distribution skewed or symmetric?

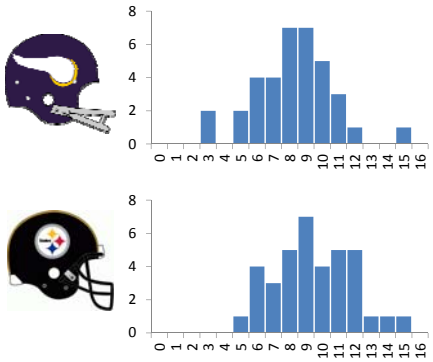
Worksheet

Percent of Students in Cook County High Schools Who are African American, 2002



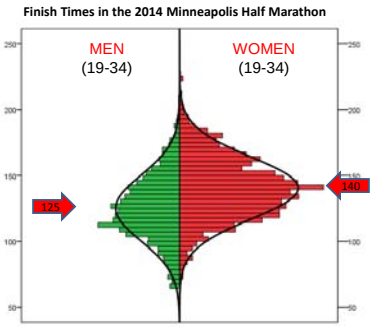
1. Where is the <u>center</u> of this distribution?
2. What can you say about how much <u>spread</u> there is around that center?
3. Are there <u>outliers</u>?
4. Is the distribution skewed or symmetric?

Learning from Distributions



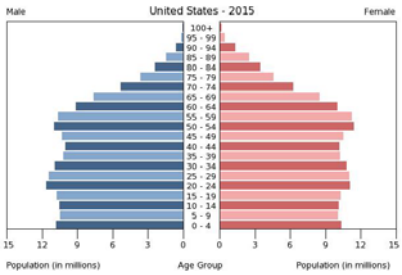
Wins by Season
1978-2014
(except 1982)

Learning from Distributions



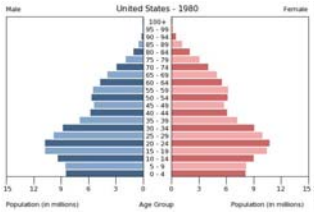
Learning from Distributions

2015 U.S. Population Pyramid
(http://www.census.gov/ipc/www/idbpyr.html)

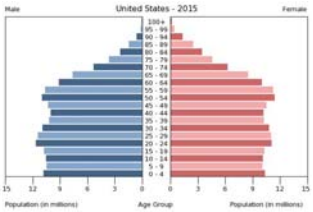


Learning from Distributions

United States **1980**

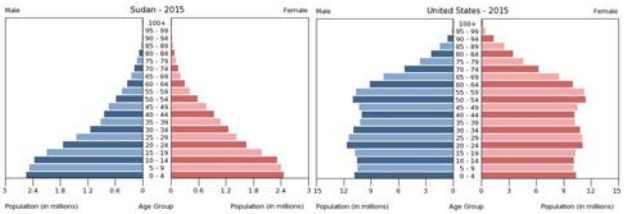


United States **2015**



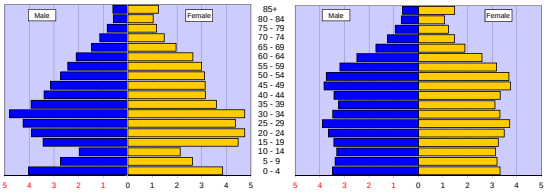
Learning from Distributions

Sudan **2015** United States **2015**

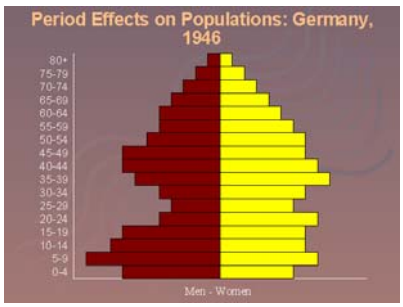


Learning from Distributions

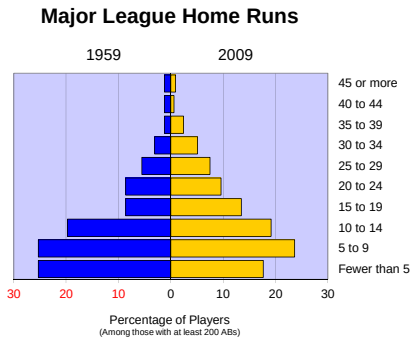
Washington, D.C., 2010 Minnesota, 2010



Learning from Distributions



Depicting Distributions



Want More?

David Lane's Text

http://onlinestatbook.com/2/graphing_distributions/graphing_distributions.html

Richard Lowry's Text

http://vassarstats.net/textbook/ch2pt1.html

Gerard Dallal's Text

http://www.jerrydallal.com/LHSP/plots.htm

Will Hopkins' Text

http://www.sportsci.org/resource/stats/summarize.html