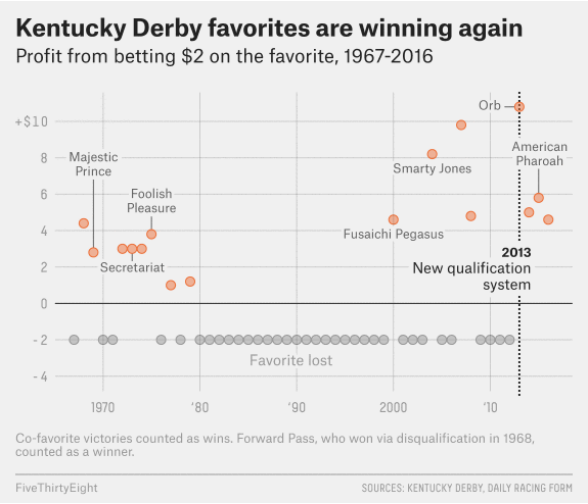
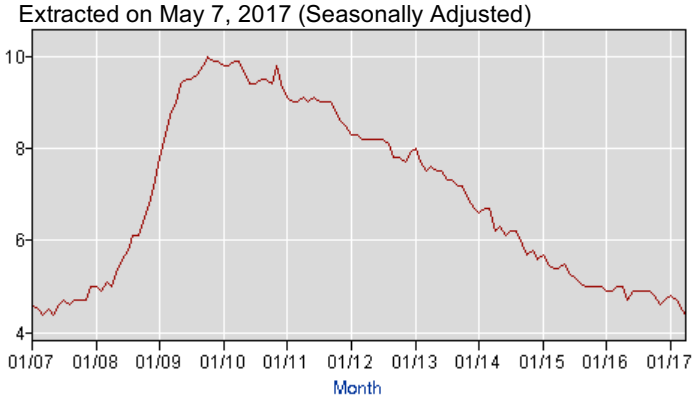
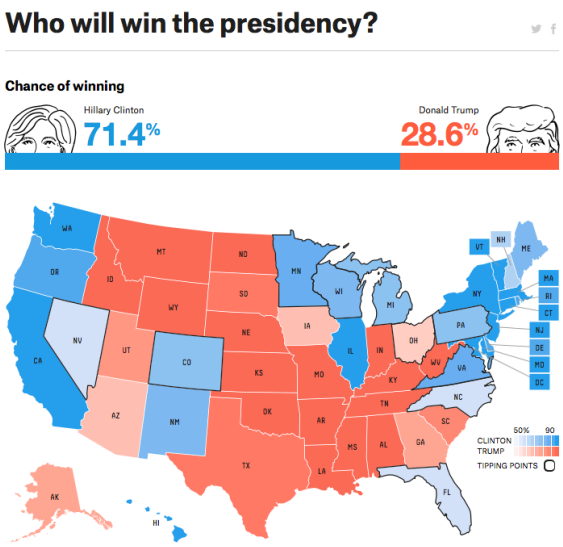
AP Statistics Summer Institute 2017

E-mail: dave_ferris@nobl.k12.in.us
Web site: noblestatman.com

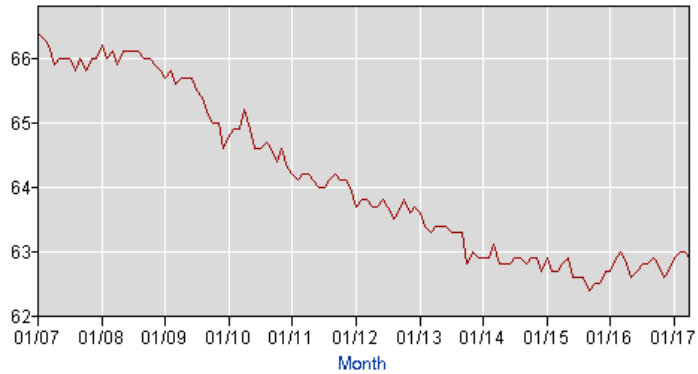


Current Population Survey: Unemployment Rate
Extracted on May 7, 2017 (Seasonally Adjusted)

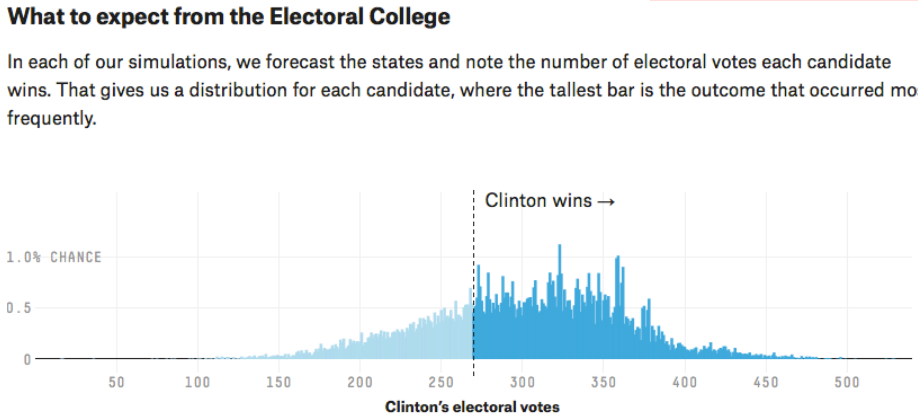


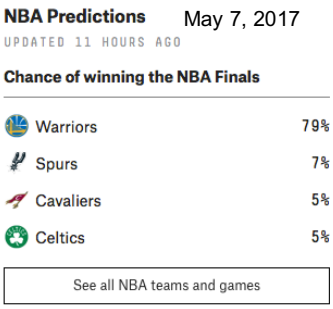


Current Population Survey: Labor Force Participation Rate
Extracted on May 7, 2017 (Seasonally Adjusted)



Questions about these graphs:
Does the graph prove that Derby favorites are winning?
What does the 71.4% mean?
Why is Participation Rate important with CPS numbers?
Were the NBA champs correctly predicted in May?





Web Site List

RESOURCES, HANDOUTS, ACTIVITIES:

noblestatman.com ("**ehandouts**" and other resources)

AP Central (official **documents** and AP **Exam problems**)

apstatsmonkey.com (**clearinghouse** for many useful resources, including Best Practices Night at the AP Reading, FRAPPY's, etc.)

amstat.org (American Statistical Association)
(STEW **lesson plans** for activities and mini-projects; GAISE report: guidelines for statistical instruction)

TOOLS:

Against All Odds statistics **videos** (can stream for free--learner.org)

amstat.org (Census At School: **survey** and **student data**)

Rossman Chance **applets** (many good simulation applets)

onlinestatbook.com/stat_sim/sampling_dist/index.html (great sampling distribution demonstration **applet**)

getkahoot.com, quizlet.com, quizizz.com (engaging online review **games**)

StatCrunch (Stats **Software**: teacher account is free, students pay small fee)

StatKey (**simulation** website app)

tuvalabs.com (online **tool** for analyzing distributions and scatterplots)

gapminder.org (amazing online analysis **tool** of United Nations data)

Classifying Statistics Problems (ltcconline—**practice** at choosing the correct inference procedure)

ARTICLES and NEWS:

fivethirtyeight.com (great current **articles** with a statistical slant; engaging graphs)

www.causeweb.org/sbi (Simulation Based Inference discussions/**blog**. This is a "trending" topic among high school and college statistics teachers.)

thisisstatistics.org (engaging information on statistics as a **career**)

tylervigen.com/spurious-correlations (funny, non-causation relationships)

Perspectives on AP Statistics

By Chris Olsen
George Washington High School
Cedar Rapids, IA

Unlike the AP Statistics course description, which outlines the scope and nature of the course, this introduction focuses on the teaching and learning experience. The teachers who pioneered this program were deeply committed to it and excited about the benefits it could bring to their students. I believe their enthusiasm was rooted in (1) the discipline of statistics, (2) their experience with their students, and (3) the collective professional community they created.

Because the science, art, and practice of statistics differ significantly from other fields of mathematics, it is not surprising that this discipline is also taught differently. Among ecologists, there is a concept known as the "edge effect," the biologically active, interstitial region that forms a boundary --, for example, between a forest and a meadow. As the eminent statistician John Tukey noted, the field of statistics allows you to ***play in everyone else's backyard***. Statistics is positioned at the edge between the known and the unknown in all those backyards. Our classes are populated with students who possess a bewildering variety of interests, some of which are ***allegedly nonmathematical***. Statistics can encompass and expand those interests, and provide the methods and concepts for creatively extending knowledge in all of their backyards.

The AP Statistics classroom is nothing if not active. Students analyze data with calculators and computers, conduct classroom experiments, carry out individual and group projects, and perform simulations involving probabilistic concepts. AP Statistics students are active, engaged learners. ***Moreover, these students would not necessarily be enchanted by a traditional mathematics course***. The AP Statistics course not only accommodates students with a wide variety of interests, it also serves those with a relatively wide range of math proficiency. ***Discussion in an AP Statistics class is an activity for all students***. Group projects are less likely to be dominated by the most able student, and individuals can succeed by capitalizing on their individual interests. A more healthy learning and teaching environment is difficult to imagine.

It has long been a fact of life that AP Statistics teachers are ***lonely members*** of their math departments. The preservice preparation of most math teachers today does not include a statistics course, and high school statistics teachers have less opportunity to bounce ideas off their colleagues. (This phenomenon is also not unknown among statisticians teaching in some colleges.) Ironically, this isolation, together with the power of the Internet, has spawned what is possibly the most collegial resource available to high school teachers -- each other. From the very beginning of AP Statistics, an electronic discussion group (EDG) operating out of British Columbia attracted statistics teachers to ongoing discussions of content, philosophy, and pedagogy. This EDG, now operating under the aegis of the College Board®, created a synergistic network that has aided hundreds of high school teachers, as well as college and university professors who realize the importance of the precollege statistics curriculum. Sessions at the annual meetings of the National Council of Teachers of Mathematics, statistics institutes at the North Carolina School of Science and Mathematics, and a growing number of Web sites have been direct consequences of this long-range collegiality. Clearly, this is the way our profession ought to work -- and nobody has done it better than the AP Statistics teachers.

In the light of the first six years' experience, the AP Statistics phenomenon must be declared an incredible and enduring success. It is a success not merely by the numbers -- 173,944 students have taken the AP Statistics exam in the past six years[1] -- but because of the personal and professional experience of teachers like you, and the learning experience of students like yours.

[1] In 2016, approximately 208,000 students took the AP Statistics Exam (841 readers)

Why Take Statistics?

The Philosophical:
Statistics deals with how we can learn about the world from observations when those observations are fallible. It teaches us to work with randomness and make it a tool for discovering the unknown rather than something to fear. The reasoning of statistics is the foundation of scientific reasoning. When examined closely, it is amazing that we actually can draw conclusions from a random sample or fallible experiment. If nature were just a bit more intransigent, we'd be unable to get anywhere. There is something quite amazing and beautiful in reasoning and methods that tread so close to the edge of the impossible and yet are so fundamental to scientific progress.

The Political:
Your parents may have told you that their statistics course was the worst experience of their academic lives, but things are different now. The AP course is a modern course that focuses on data rather than on probability theory and combinatorics. We use calculators and computers and deal with real-world problems. We use discovery learning methods and multimedia materials. I guarantee that your experience will be nothing like what your parents suffered through.

The Practical:
Statistics is used by charlatans to fool the masses. Statistical literacy is good self-defense in a world of advertising and political claims that only seem to be based on data. This course will arm you with the necessary understanding so you can smell a statistically- disguised lie a mile away. And think of the money you'll save not playing the lottery.

The Crass:
1) Statistics is one of those courses that will get you a job and/or a promotion. When we survey the companies that hire our graduates (I teach at Cornell) they tell us that they'd like our students to know more statistics and more computing. When we survey our recent graduates (5-years out), they tell us that statistics was one of the most important courses to them in getting their current positions.

2) Statistics is required by virtually every social science major, engineering, pre-med, and many others. It is now required for most law degrees. AP Statistics credit is therefore a useful commodity -- it will save you time and money in college.

--Paul Velleman

AP Statistics Topic Outline

Following is an outline of the major topics covered by the AP Statistics Exam. The ordering here is intended to define the scope of the course but not necessarily the sequence. The percentages in parentheses for each content area indicate the coverage for that content area in the exam.

Full Document:
https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap-statistics-course-description.pdf

I. Exploring Data: Describing patterns and departures from patterns
(20%–30%)

Exploratory analysis of data makes use of graphical and numerical techniques to study patterns and departures from patterns. Emphasis should be placed on interpreting information from graphical and numerical displays and summaries.

A. Constructing and interpreting graphical displays of distributions of univariate data (dotplot, stemplot, histogram, cumulative frequency plot)
B. Summarizing distributions of univariate data
C. Comparing distributions of univariate data (dotplots, back-to-back stemplots, parallel boxplots)
D. Exploring bivariate data
E. Exploring categorical data

II. Sampling and Experimentation: Planning and conducting a study
(10%–15%)

Data must be collected according to a well-developed plan if valid information on a conjecture is to be obtained. This plan includes clarifying the question and deciding upon a method of data collection and analysis.

A. Overview of methods of data collection
B. Planning and conducting surveys
C. Planning and conducting experiments
D. Generalizability of results and types of conclusions that can be drawn from observational studies, experiments, and surveys

III. Anticipating Patterns: Exploring random phenomena using probability and simulation
(20%–30%)

Probability is the tool used for anticipating what the distribution of data should look like under a given model.

A. Probability
B. Combining independent random variables
C. The normal distribution
D. Sampling distributions

IV. Statistical Inference: Estimating population parameters and testing hypotheses
(30%–40%)

Statistical inference guides the selection of appropriate models.

A. Estimation (point estimators and confidence intervals)
B. Tests of significance

M&M's Activity:

My Guess:_____

Guesses:

Actual:

Describe the distribution of the guesses:

Compare and contrast the distributions of guesses vs. actual:

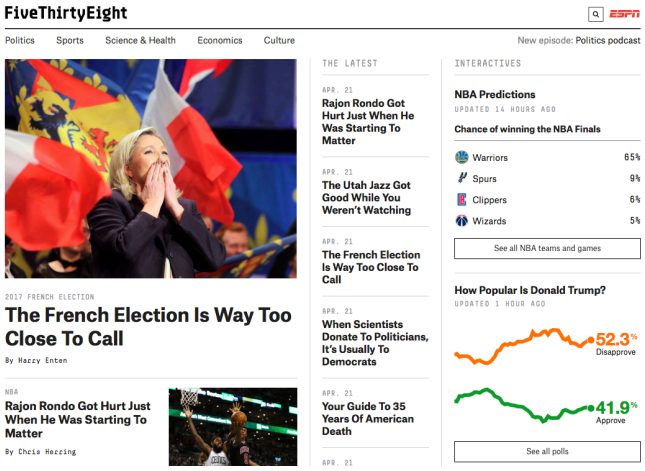
FIRST DAY OF SCHOOL IDEAS:

- Register and give CensusAtSchool survey: amstat.org/censusatschool

  (Your students' data can be downloaded as csv file for later use.)

- Fun Size M&M's: guess # in bag, # of colors, mean of distributions, etc.
- Show thisisstatistics.org
- Memory Game Activity (see noblestatman.com or TPS5e Ch. 4 resources)
- Show interesting/famous/current graphs, data stories, articles about stats and surveys

  FiveThirtyEight.com has great statistics stories and analysis:



- Look at some cool data sets: TuvaLabs.com, Census at School, gapminder.org, Bodyfat dataset, McDonald's menu, Arby's Menu, etc.

US Census at School Measurement Questions

Use Safari. Class ID: _____ Password: _____

The following questions require measurements. Please fill these out prior to taking the online survey.

4. How tall are you without your shoes on? Answer to the nearest centimeter.

_____

5. What is the length of your right foot (without your shoe on)? Answer to the nearest centimeter.

_____

6. What is your arm span? (Open arms wide and measure distance across your back from tip of right hand middle finger to tip of left hand middle finger.) Answer to the nearest centimeter.

_____

9. How long does it usually take you to travel to school? Answer to the nearest minute.

_____

14. What is the length of your left foot (without your shoe on)? Answer to the nearest centimeter.

_____

16. What is the length of your index finger (finger next to your thumb) on your left hand? Answer to the nearest centimeter.

_____

17. What is the length of your ring finger? (located between your middle finger and little finger) on your left hand? Answer to the nearest millimeter (there are 10 millimeters in one centimeter).

_____

26. How many hours of sleep do you usually get when you have school the next day?

_____

27. How many hours of sleep do you usually get when you don't have school the next day?

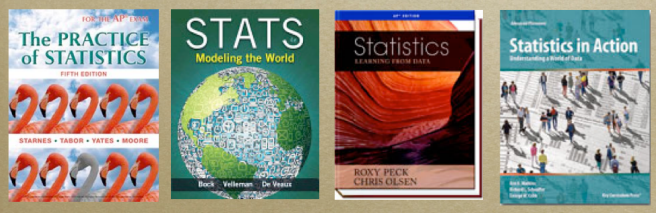
_____

Essential resources from apcentral.collegeboard.com:

- **Course Overview**
- **Full Course Description** (including AP StatsTopic Outline—later in this handout)
  - "Statistics is a discipline in which clear and complete communication is an essential skill."
  - "formulate cogent answers"
- **AP Statistics Teacher's Guide** (BIG pdf file)
- **AP Teacher Community** and e-mail discussion group (see green sidebar)
- **Audit information**
- **Special Focus**: Samping Dist'ns, Inference, Planning and Conducting a Study
  Excellent "extra" information about these three important topics

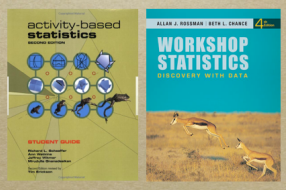
Textbooks:

- The Practice of Statistics 5e (Starnes, Yates, Moore)
- Stats: Modeling the World 4e (Bock, Velleman, Deveaux)
- Statistics: Learning From Data "1e" (Peck, Olsen)
- Statistics In Action 2e (Watkins, Schaeffer, Cobb)



Other Resources:

- www.apstatsmonkey.com
- Rossman/Chance applets
- TuvaLabs.com
- StatKey (online simulations and data analysis)
- AP Exams (FR) on AP Central (& scoring rubrics)
- Supplementals:
  - Activity-Based Statistics
  - Workshop Statistics
  - Exam Formulas (CB book)
- Fathom, StatCrunch, Google Forms



Survey Question Ideas

First Day of School Name:_____

Supply the data for the following categories/attributes/measurements.

1. First Name:
2. Last Name:
3. Gender:
4. Blood Type:
5. Age in Days:
6. Height in cm:
7. Belly button height above ground in cm.:
8. Foot length in cm.:
9. Wrist circumference in cm.:
10. Circumference of head in cm.:
11. Hand span in cm:
12. Length of leg (from hip bone to floor):
13. Shoe size:
14. Eye color:
15. Hair color:
16. Month of birth:
17. Day of birth (1–31):
18. Day of week on which you were born:
19. Number of sisters:
20. Number of brothers:
21. Are you left-handed or right-handed?
22. Number of texts sent in last 24 hours:
23. Choose and circle one of these numbers: 1 2 3 4 (may want to save for later)
24. Write down a number between 1 and 20: (may want to save for later)
25. Write down a color that begins with the letter "O:"
26. Write down the first word in the name of the school you attended in 5th grade:_____
27. Number of hours you have slept in the last 24 hours:
28. Amount of money in coins you have today:

The Kristen Gilbert Case

Kristen Gilbert worked as a nurse in the intensive care unit of the VA hospital in Northampton, Massachusetts in the 1990's. During her shifts, other nurses became suspicious that she was killing patients by injecting them with epinephrine, a heart stimulant. An analysis of 1641 eight-hour shifts was presented as evidence at her trial (Cobb and Gelbach, 2005[2]). Is there statistical evidence from this table that Gilbert murdered patients?

	Death on shift?		
Gilbert Present?	Yes	No	**Total**
Yes	40	217	**257**
No	34	1350	**1384**
Total:	**74**	**1567**	**1641**

[2] http://www.stat.ucla.edu/~nchristo/statistics100B/article.pdf

Using StatKey simulation

1. Open StatKey: http://lock5stat.com/statkey/
2. Choose Test for Difference in Proportions
3. Choose Edit Data, and fill in with 40, 257, 34, 1384. Click OK.
4. Generate 1 sample. What is the simulation doing?

   What do the numbers in the Randomization Sample table represent?

5. Generate 9 more samples. What do the 10 dots in the graph represent?

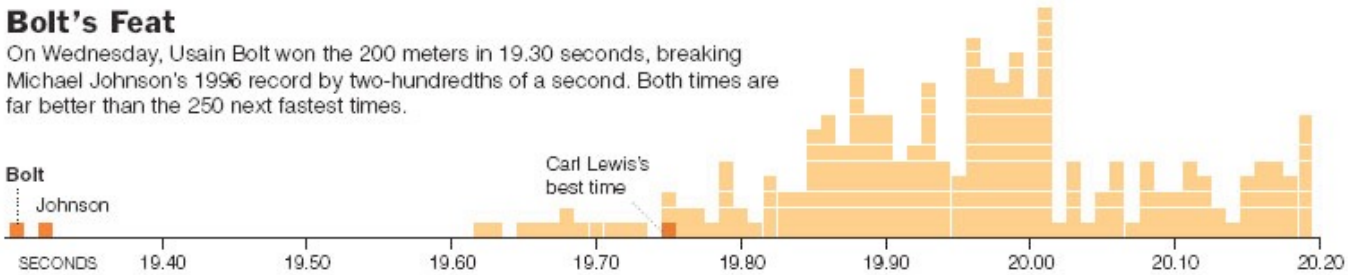
6. Generate 1000 Samples. Describe the distribution that is produced (shape, center, spread).

7. What does this distribution of 1010 dots tell us about the original Kristen Gilbert data?

8. Does this data suggest that Kristen murdered patients? Explain.

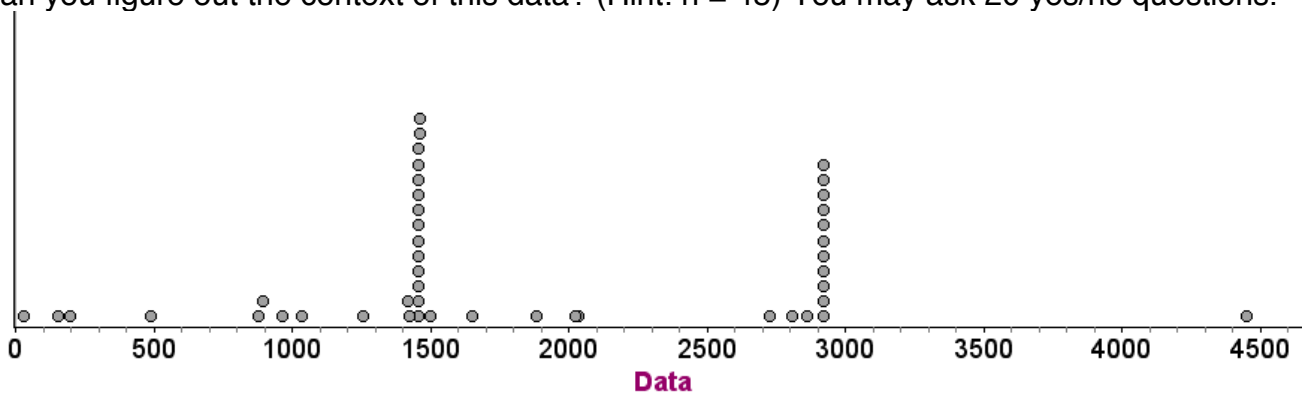
Things that come up make great class openers!

From the 2008 Summer Olympics:

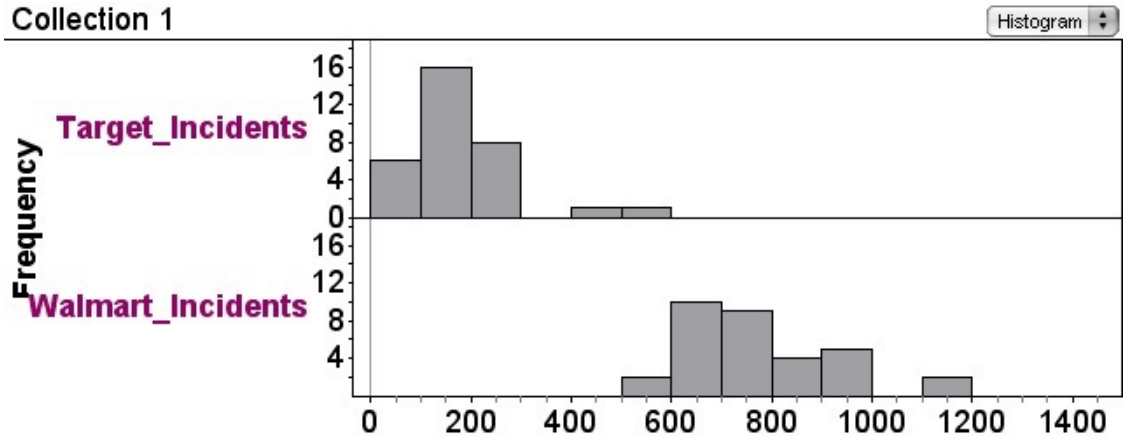


Outliers and Inliers

Can you figure out the context of this data? (Hint: n = 45) You may ask 20 yes/no questions.

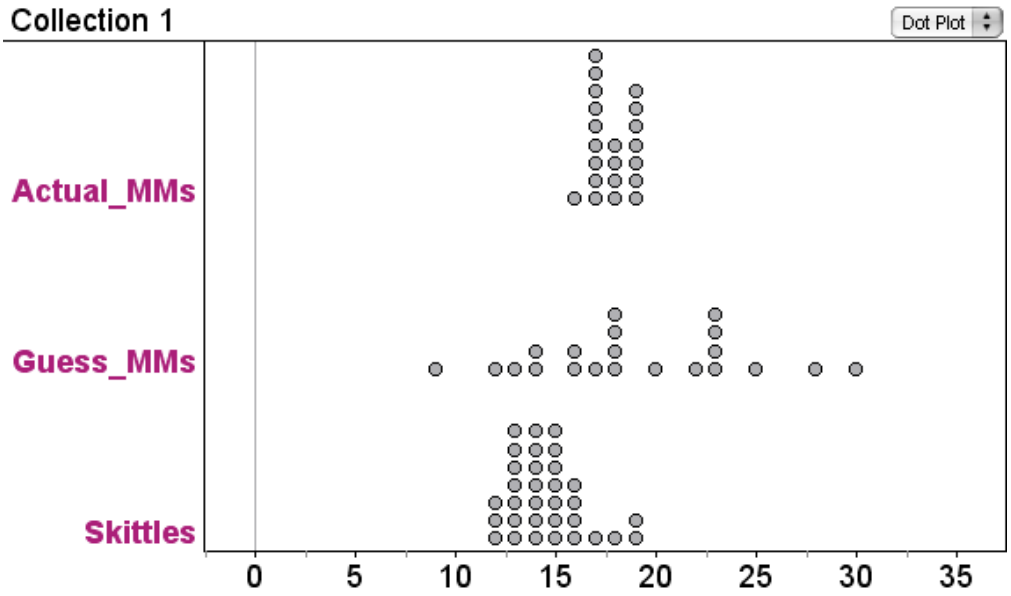


Is Wal-Mart Safe?[3]



[3] See notes at the end of this handout for details on how the data was collected.

Standard Deviation:



Actual: mean= _____

SD = _____

Guess: mean= _____

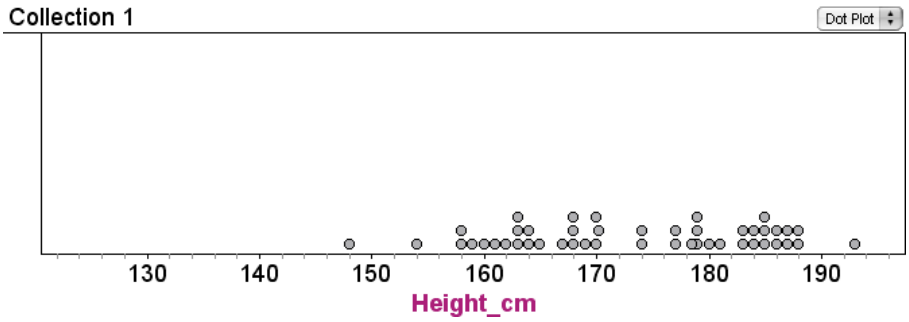
SD = _____

Skittles: mean= _____

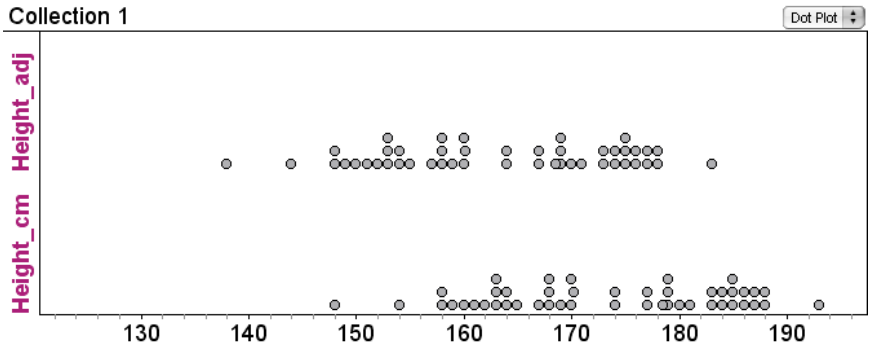
SD = _____

Transforming Data

Below is the distribution of student heights in cm.



1. What if we discovered that the tape measure used to measure height had 10cm cut off (i.e. it started at 10 instead of 0)? We need to subtract 10cm from each student's height. How would this transformation affect the mean and SD?



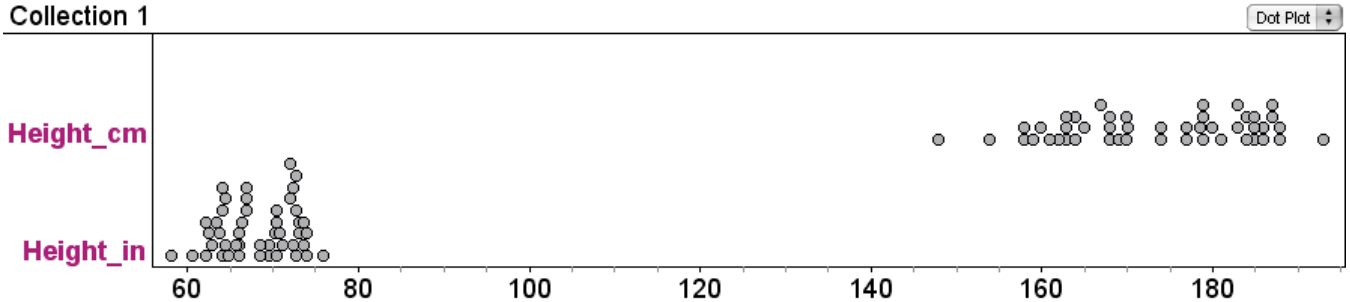
Mean = _____

SD = _____

2. What if we had to convert each height to inches and recalculate the mean and SD? What is the new mean and SD?

Mean = _____

SD = _____



Categorical Data

Below are some data and questions from the Titanic disaster.

http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic.html
http://www.encyclopedia-titanica.org/

1. Give the marginal distribution of class (in %'s).

Titanic Passengers

		Survived		Row Summary
		No	Yes	
Class	1st	129	193	322
	2nd	161	119	280
	3rd	574	137	711
Column Summary		864	449	1313

S1 = count ( )

2. Give the conditional distribution of survival by class (in percent).

3. Of the first class passengers, what percent survived? P(survived | 1st class) = _____

4. Of the survivors, what percent were 1st class? P(1st class | survived) = _____

5. What percent of passengers were 1st class? _____

6. What percent of passengers were either first class or survived? _____

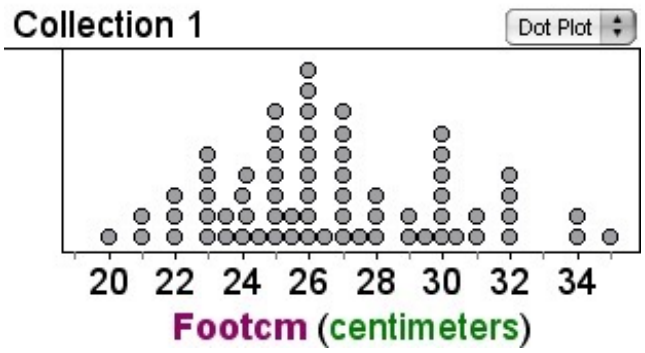
7. Do survivability and class appear to be independent? Explain.

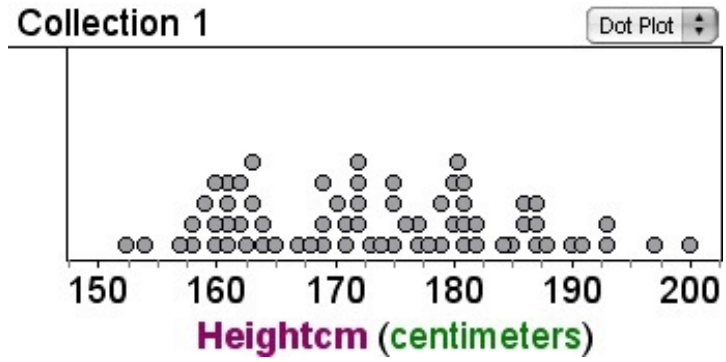
The Normal Model for Data:

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(Note difference between “math world” and “stats world.”)

Which is more extreme, a foot length of 32 cm or a height of 190 cm?

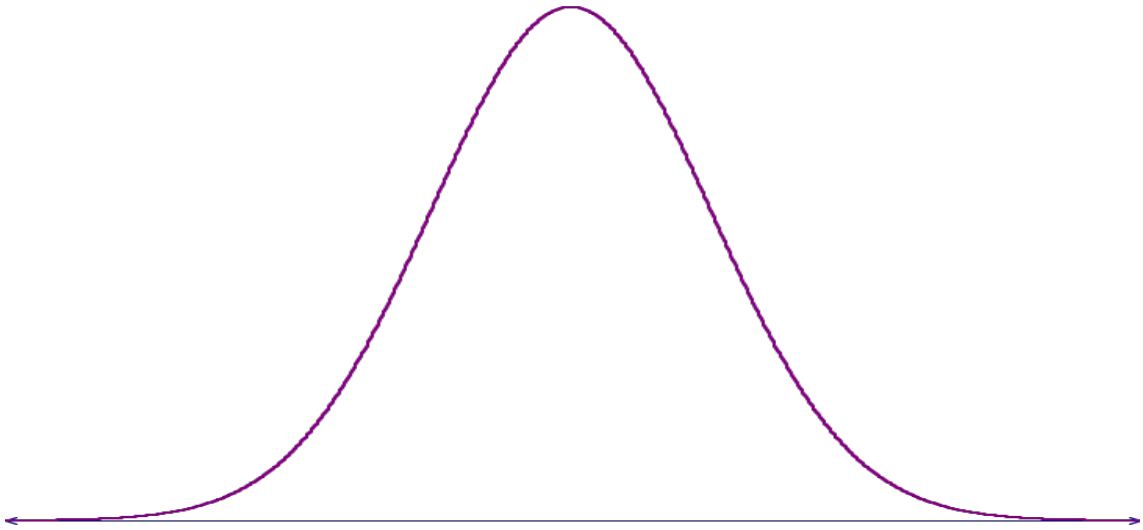




Heightcm	173.597 cm 11.36 cm
Footcm	26.6964 cm 3.37966 cm

S1 = mean ( )
S2 = stdDev ( )

Common exercises using the normal model:



1. Find the z-score for a student who scored 660 on the SAT Verbal, where the mean is 505 and the standard deviation is 110.

2. What percentile did this student score?

3. What SAT Verbal score represents the first (lower) quartile?

4. Approximately what percent of students scored between 450 and 560?

5. What score would a student have to earn to be in the top 5% of all SAT Verbal scorers?

6. Approximately what percent of students scored between 395 and 615?

7. Approximately what percent of student scored within two standard deviations of 505?

Age Guessing Activity

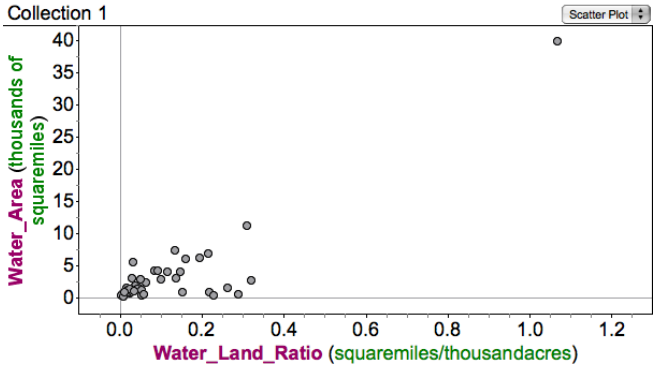
1. Guess the ages of the following people, and put your guess in this column: ↓

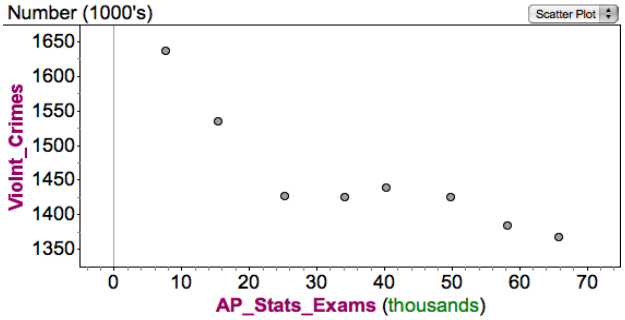
Name:	Actual Age:	Your Guess:
Donald Trump	_____	_____
Nate Silver	_____	_____
Bill Gates	_____	_____
Johnny Depp	_____	_____
Adele	_____	_____
Alex Trebek	_____	_____
Daisy Ridley	_____	_____
Miley Cyrus	_____	_____
Tom Brady	_____	_____
J. K. Rowling	_____	_____
Mick Jagger	_____	_____
Mark Zuckerberg	_____	_____

2. Put the actual ages of each person in the first column.
3. Type both lists into your calculator. (Actual ages in L1, your guesses in L2.)
4. Make a scatterplot for these two lists. (x-axis is L1, y-axis is L2)
5. Calculate r: _____
6. Describe below what you discovered about your age guesses:

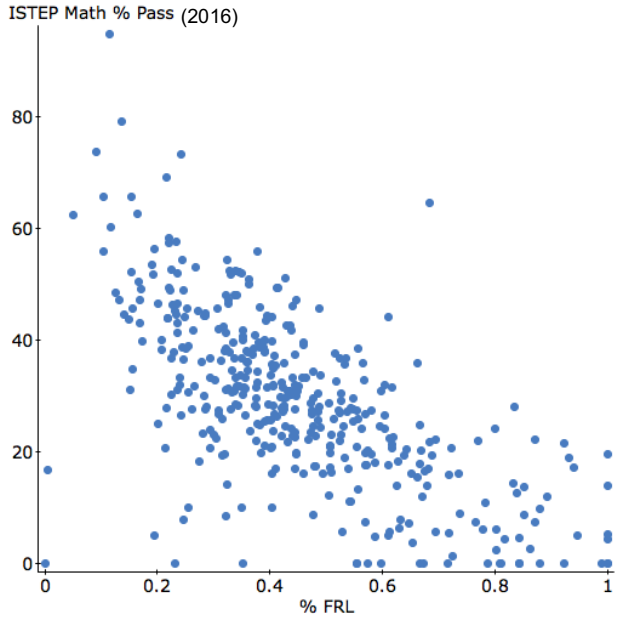
Some Scatterplot Examples:

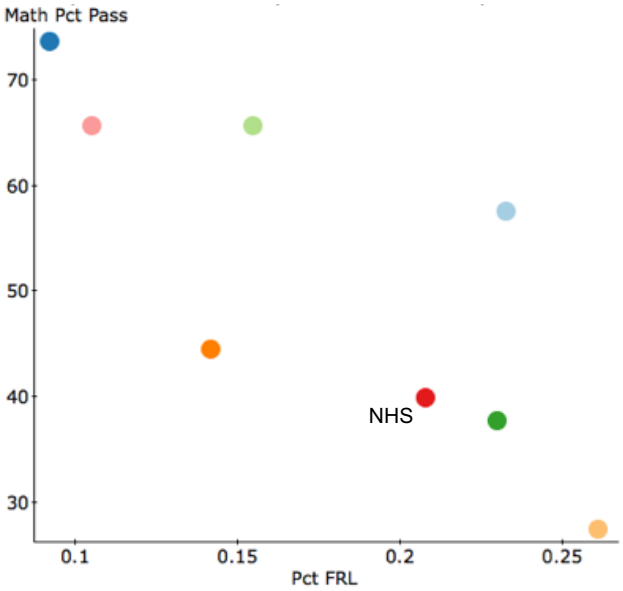
Great for discussion. The entire set of scatterplots is on noblestatman.com





Student performance data: Indiana vs. Hamilton Co. Schools:





Which school should we visit to inquire about their instructional interventions?

Commonly Asked Regression Questions

(as seen on AP Statistics exams…can be used with any bivariate data set)

1. Describe the association in context.

2. Is a linear model appropriate to describe this relationship? Explain.

3. Write the equation for the linear model on this data.

4. Explain the meaning of the slope in this linear model

5. Explain the meaning of the y-intercept in this linear model

6. Find the value and explain the meaning of the correlation coefficient.

7. Find the value of and interpret r-squared

8. a. Using the linear model, predict _____ when _____ is _____.

   b. Is the residual for this data point positive or negative? Is the model over- or underestimating? Explain.

9. Comment on any outliers present. Fully describe their effect on the analysis, if any.

10. Interpret regression and model information from a computer printout.

Starburst Grab Activity

NAME_____ Guess # Starbursts you can grab: _____

During Homecoming Week, your teacher is planning to give away a handful of Starburst candies to one lucky student in each class. To know how many candies to buy, it might be good to find out how many candies the typical student can grab with one hand. Naturally, we should collect some data! Each student will grab a handful of Starbursts and record the number grabbed.

1. Come to the container and grab one handful of Starbursts. You must shake your handful twice to allow any loose candies to fall out. After two shakes, this is your official sample. Count how many you grabbed and tally on the board and in a computer/calculator.

2. Based on the *class sample*, how many Starbursts should be bought per class so to have enough for the winner to grab? Explain your choice.

3. There is some variation in the number of Starbursts grabbed.

   a) Name some statistics that could be calculated to measure the variation in this data.

   b) Which one is the best for this data? Explain.

4. One measure of the total amount of variation is called Sum of Squares Total (SST). This is calculated by measuring the distances from each data point to the mean, squaring these distances, and adding them all up: $SST = \sum (x_i - \bar{x})^2$ Calculate the SST for this data.

SST = _____

5. The amount of total variation in this data makes it difficult to predict the number of Starbursts for a particular student. So now we will consider how improve our ability to predict. Perhaps the winning students should be picked first, *before* the teacher goes to Wal-Mart to buy Starbursts. What could be measured on each student to help accurately estimate the total amount of Starbursts needed? Pick three possible measurements, and explain each choice.

   #1:

   #2:

   #3:

6. Decide as a class on the one variable that we should use to help with our prediction. Then create an appropriate model that relates this new variable to the # of Starbursts grabbed.

   After the model is created, calculate the amount of variation that "still remains." In other words, the total amount of variation remaining between the actual # of Starbursts grabbed and the predicted # of Starbursts grabbed according to this new model. This statistic is called the Sum of the Squares of the Errors (Residuals), or SSE.

   SSE = _____

7. Go back and find the SST from #4. Notice that the SSE is smaller. In other words, by using a model on a second (explanatory) variable, we have been able to eliminate or account for some of the variation in the # of Starbursts grabbed. Calculate the percent of the SST that was "accounted for."

   The % of variation in the number of Starbursts that has been accounted for by the

   _____ model on _____ = _____%

   This number is called _____. (Yes, it really is _____ ^ ____.)

Starburst Grab Activity Teacher Notes:

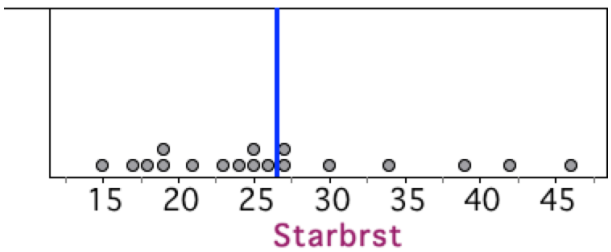
1. Have students guess how many Starbursts they can grab with one hand: _____

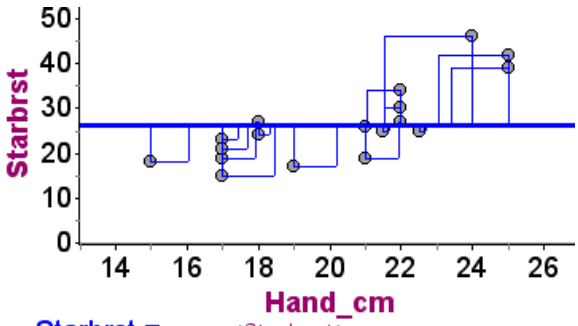
2. Then tell the story on the sheet @ Homecoming…

Sample Data:

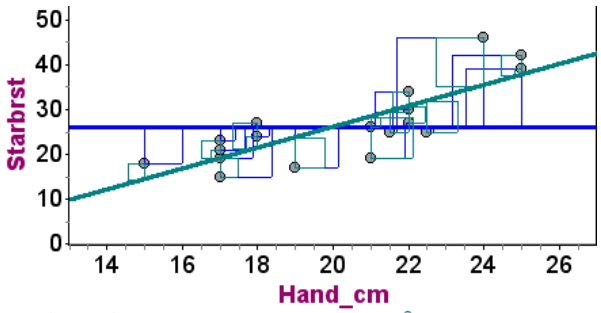
Mean = 26.5 candies, SST = 1306, SSE = 451.9, r-squared = 0.65

Hand_cm	# Starbst
22.0	34
21.0	19
21.5	25
17.0	23
15.0	18
18.0	27
22.0	27
18.0	24
25.0	39
17.0	19
22.0	30
21.0	26
24.0	46
22.5	25
25.0	42
17.0	15
17.0	21
19.0	17

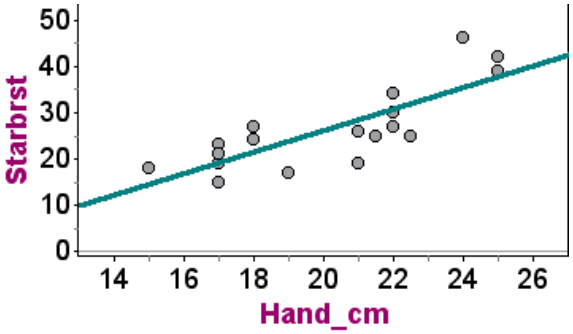




— Starbrst = mean (Starbrst)
Sum of squares = 1306



— Starbrst = -20.6 + 2.33Hand_cm; r^2 = 0.65
Sum of squares = 451.9
— Starbrst = mean (Starbrst)
Sum of squares = 1306



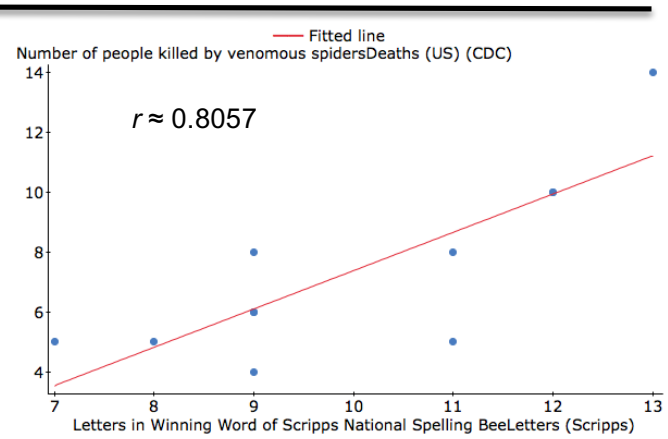
— Starbrst = -20.6 + 2.33Hand_cm; r^2 = 0.65

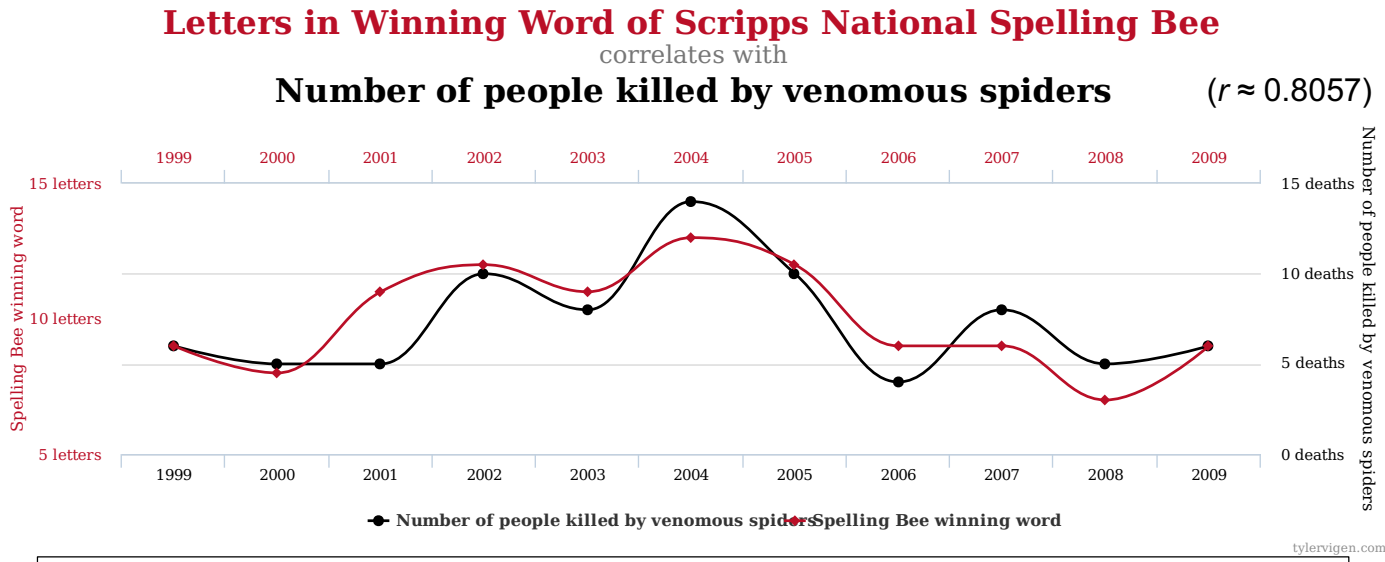
Age Guessing Key:

Name:	Actual BD:	AGE:
Donald Trump	6/14/46	_____
Nate Silver	1/13/78	_____
Bill Gates	10/28/55	_____
Johnny Depp	6/9/63	_____
Adele	5/5/88	_____
Alex Trebek	7/22/40	_____
Daisy Ridley	4/10/92	_____
Miley Cyrus	11/23/92	_____
Tom Brady	8/3/77	_____
J. K. Rowling	7/31/65	_____
Mick Jagger	7/26/43	_____
Mark Zuckerberg	5/14/84	_____

Wal-Mart vs. Target:

This phase of the study focused on the 50 Wal-Mart stores out of the 460 analyzed stores that experienced the "highest rate" of reported police incidents in 2004. Target stores chosen for the comparative analysis were within a 10-mile radius of the 50 "high incident" Wal-Mart stores. Of these 50 "high incident" Wal-Mart stores, three stores did not have a Target within 10 miles, leaving a sample of 47 Wal-Mart stores for further analysis. Because of further data restrictions, the sample for comparison was limited to 32 Wal-Mart stores and 30 nearby Target stores.





Graphs above show two ways of graphing the same data: one uses time and two different y-axes. The other shows only the two variables of interest. From tylervigen.com/spurious-correlations

Sample Data Sets for Regression (from real students):

Starburst Grab Activity:

Hand Width	Starbursts
21.5	30
21.0	28
19.5	22
19.5	32
20.0	37
19.0	20
22.0	26
19.0	19
21.0	22
23.0	51
23.0	43
18.0	25
20.1	27
21.0	36
19.5	27
19.5	32
21.5	36
18.0	32
22.5	25
18.5	25
20.5	31
19.5	35

Counting Steps:

LegLength	LongWalk
95.00	228
111.26	187
108.00	199
94.00	235
98.00	231
111.00	196
103.00	236
119.00	187
107.00	192
113.00	183
93.00	213
107.00	217
91.00	252
108.00	217
100.00	248
95.00	246
100.00	254
116.00	214
83.00	240
96.00	203
113.00	217