Mathematics Teachers' Circle, 2008
Good Statistics

Davar Khoshnevisan
University of Utah

March 10, 2008

1 A few general rules of thumb

- Avoid the use of canned procedures unless you understand well:
  - When they apply;
  - how they apply [a simple, though possibly, technical matter];
  - whether, or how, they apply to the problem at hand.
    [This is of fundamental importance.]
- Good statistical analysis begins by a careful examination of the problem:
  - What are the issues we are trying to understand?
    [Know the problem.]
  - What is the data, especially its quality? How was the data gathered? What does it mean? [Avoid bad sampling methods]
- Before making fancy computations:
  - Plot your data in various meaningful ways.
  - Understand your data [units, date structure, ...].
  - Identify procedures that might apply.
    [Pay some attention to whether they do.]
  - Compute, and make inferences if possible.
  - If not possible, then consult [other] experts.
- Serious problems deserve serious attention.

2 Case Study 1: Landon versus FDR

Today
Election Forecast

AMERICA SPEAKS
THE NATIONAL WEEKLY POLL of PUBLIC OPINION

Next Sunday
The Election in Review

Institute Forecasts the Re-election of Franklin D. Roosevelt, Gives Him 54% of Popular Vote, Minimum of 315 Electors

Major Party Percent Is 55.7; New York in F.D.R. 'Sure' Column

Election Forecast

1—The American Institute of Public Opinion predicts the re-election of Franklin D. Roosevelt and John N. Garner.

* * *

2—The Institute's latest presidential poll indicates that Roosevelt will receive approximately 56% of the major party vote (minor parties eliminated), to 44% for Alfred M. Landon and Frank Knox. In 1932 the President received 59.1% of the major party vote.

* * *

3—With minor parties included, President Roosevelt's percentage of the total popular vote will be approximately 54%, to 46% for Landon.

* * *

4—The President will receive a minimum of 315 electoral votes. The number necessary to win is 266. Should last-minute shifts in the group of states where the race is nip-and-tuck give this entire group to Roosevelt, he would receive more electoral votes than in 1932, when he polled 472.

* * *

5—William Lemke, candidate of the Union Party, will poll fewer than 1,500,000 popular votes, and carry no state.

* * *

6—Norman Thomas, Socialist candidate, will poll about half as many votes as in 1932, when he received 884,000.

Election Will Test Clashing Poll Methods

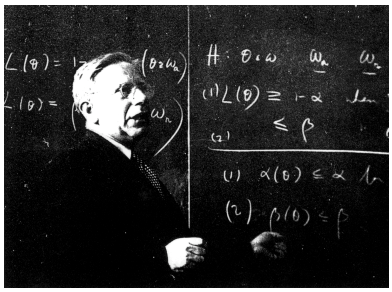
- President Franklin Delano Roosevelt (FDR) running for re-election for the United States' Presidency (1936)
- Main Republican candidate = Governor (Alf) Alfred Landon (R. KS)
- Recovery from The great Depression (9 million unemployed; real income had dropped by $\frac{1}{3}$ from 1929–1933, and was only just turning around)
- **Landon's Platform:** Economy in Government ("the spenders must go")
- **FDR's platform:** Deficit financing ("... balance the budget of the American people before ... the national government")
- Both candidates concerns with domestic affairs (Nazis arming Germany; Civil war in Spain reaching its climax)
- Most people thought FDR would win
- *Literary Digest* predicted 43% for FDR; **fact:** he received 62% of votes
- *Lit. Dig.* had correctly predicted the elections five times prior to this event!
- Largest gross prediction error in presidential races *ever*!
- *Lit. Dig.* went bankrupt soon after

A bit about the Literary Digest methodology:

- Mailed questionnaire to 10 million people (2.4 million responded)
- Addresses were chosen from:
  - Telephone books had phones)
  - club membership listings, ...
- George Gallup took a *random* sample of 50,000; predicted 54% for FDR
- Nowadays, random samples of 5,000 beat even those methods

Question: Why did Gallup's sample of size 50,000 beat that of Literary Digest (size 2.4 million!)?

3 Case Study 2: Wald's quick response



In the early-to-mid period of The Second World War, Abraham Wald (US Center for Naval Analyses and Columbia University) was asked to make a suggestion to the United States' Navy. The issue is simple to describe: At the time, the Navy enjoyed a certain surplus of armor plating for fighter planes. It wanted to know which parts of the aircrafts should received further armor plating for better protection. Wald made one of the following remarks. Can you identify his answer? Why?

1. The side of the plane: It is the easiest to hit.
2. Look at the planes that returned safely. Armor-plate the parts that were hit on the returned places.
3. Look at the planes that returned safely. Armor-plate the parts that were not hit on the returned places.
4. Why, the most sensitive part, of course. That would be the engine now, wouldn't it?

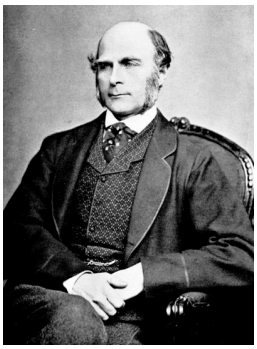
- Wald's answer helped saved thousands of lives.
- Wald himself died in an aircraft crash over India in 1950.

4 Case Study 3: Those hidden variables

Useful buzz words:
Correlation does not imply causation

4.1 A little history

Correlation was discovered by the founding father of modern genetics, Francis Galton, in 1888. Galton first dubbed it "co-relation." Nowadays, the proper term is "correlation coefficient," but is loosely referred to as "correlation," as well. That correlation is now a part of the English language is testimony to the depth of Galton's original ideas.



Correlation, together with its role in data analysis, were addressed subsequently and far greater depth by Galton and his mathematician colleague Karl Pearson around 1890.[1]

4.2 Formulas that changed the world

Typically we have paired data in a sample of size n:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

For example, suppose we are interested in the effect of dosage for a certain drug. Then, x_j could be the dosage administered to the jth subject; and y_j could be that subject's heart rate, measured after the drug was administered.

The Galton-Pearson formula for correlation coefficient [between the x's and the corresponding y's] is

$$\text{Correlation} = \frac{1}{n-1} \sum_{j=1}^{n} \left(\frac{x_j - \bar{x}}{\text{sd}(x)} \right) \left(\frac{y_j - \bar{y}}{\text{sd}(y)} \right),$$

where

$$\bar{x} := \frac{x_1 + \cdots + x_n}{n} \quad \text{and} \quad \bar{y} := \frac{y_1 + \cdots + y_n}{n}$$

are the respective averages of the x's and the y's, and

$$\text{sd}(x) := \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2} \quad \text{and} \quad \text{sd}(y) := \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2}$$

[1] Incidentally, Galton was a cousin of another great scientist of that time, Charles Darwin. Good genes and all that! I am aware of at least one descendent of that family today; he is a prominent mathematician in Canada.

are the corresponding standard deviations.

Warmup problem 1: Verify that correlation is a dimension-free number. For example, if the x's are measured in grams and the y's in beats-per-second, then the correlation between the x's and the y's does not depend on these units.

Warmup problem 2: Verify that the correlation between the x's and the y's is the same as the correlation between the y's and the x's.

Recall Cauchy's inequality:[2] For all real numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$$\left(\frac{a_1 b_1 + \cdots + a_n b_n}{n}\right)^2 \leq \left(\frac{a_1^2 + \cdots + a_n^2}{n}\right)\left(\frac{b_1^2 + \cdots + b_n^2}{n}\right).$$

You can find this in many basic mathematics textbooks, but this turns out to be a fundamental fact in mathematics, and we should not casually gloss over it.

We can understand Cauchy's inequality by considering it in the case that $n = 2$. In that case, we can easily evaluate

$$\left(\frac{a_1 b_1 + a_2 b_2}{2}\right)^2 = \frac{a_1^2 b_1^2 + a_2^2 b_2^2 + 2a_1 b_1 a_2 b_2}{4},$$

and

$$\left(\frac{a_1^2 + a_2^2}{2}\right)\left(\frac{b_1^2 + b_2^2}{2}\right) = \frac{a_1^2 b_1^2 + a_2^2 b_2^2 + a_1^2 b_2^2 + a_2^2 b_1^2}{4}.$$

Therefore, we compare the preceding 2 displays to find that Cauchy's inequality is equivalent to the following:

$$a_1^2 b_2^2 + a_2^2 b_1^2 \geq 2a_1 b_1 a_2 b_2.$$

You should check that this is another way to write the following inequality, which is valid tautologically:

$$(a_1 b_2 - a_2 b_1)^2 \geq 0.$$

It can be shown that the general case follows from the case $n = 2$ and mathematical induction. [Can you see how to do this?]

Warmup problem 3: Use the Cauchy inequality to verify that the correlation coefficient is a dimension-free number between -1 and 1.

[2] Frequently, credit is given to various combinations of Bienaymè, Bunyakovsky, Cauchy, and Schwarz [sometimes even Schwartz, who had nothing to do with this family of inequalities!]. The present form of the inequality is the first of its kind, and was found by Augustine–Louis Cauchy.

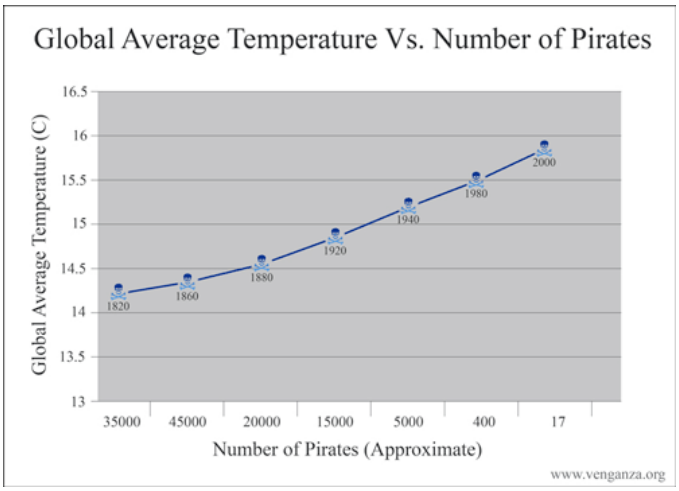
4.3 Introducing the Buick Estate Wagon

In a certain analysis to predict fuel consumption (in Gallons/100 miles) from the weight- and drive-ratio of cars, the *Buick Estate Wagon* shows up as an outlier. Remarkably enough, even though the *Buick Estate Wagon* has high fuel consumption, when compared to other vehicles, standard statistical methods predict its fuel consumption to be a lot more than what is observed in actual data. "Why does the Buick Estate Wagon consume less fuel than we expect?"

The answer is not available in the original data, but was discovered through additional research. These data were collected by Consumer's Union on a test track, rather than using the EPA test values for fuel efficiency, following the manufacturer's recommendations for each car's maintenance. Additional research revealed that, starting with this model year, Buick recommended a higher tire inflation pressure for the Buick Estate Wagon. The recommended inflation pressure level was higher than the level for other cars in the survey. Harder tires present less rolling resistance and improve gas mileage; therefore, the Buick Estate Wagon outperformed our expectations based on our regression model, which did not account for tire inflation pressure. In our model Tire Pressure is a lurking variable, variable that seems to help in predicting gas mileage but is not included in the model.

4.4 An amusing question

Do pirates cause global warming?



4.5 A more serious question

Is there sex bias in graduate admissions at UC Berkeley? Here is some real data from a few years ago (Z. Zhu, UNC):

- 8,442 men applied; about 44% admitted

- 4,321 women applied; about 35% admitted.

Question: Was there [9%] sex bias in graduate-school admissions?

The following breakdown by majors [referred to as majors "A" through "F"] might help you sort this out.

	Men		Women	
Major	#applicants	%admitted	#applicants	%admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

5 Further Reading: Start with Huff

The best place to start [along these lines] is the classic, "How to Lie with Statistics," by Darryl Huff (1954; Huff was then the editor for *Better Homes & Gardens!*). It is chock full of ideas that are original and fresh to this day. I looked up the most recent printing of this book on *Amazon.com*; it is currently being sold for a little under 10 dollars.

Caveat: The text was written in a different world; its style is old-fashioned, and the language out of date. If you find yourself getting offended by the anecdotes, then you are missing the very valid scientific ideas in this book. And that would be a pity.

Michael Steele (U Pennsylvania) has written a nice account of the history, as well as cultural-intellectual impact, of this wonderful little book. Steele's remarks appeared in the journal *Statistical Science* (2005). You can find a facsimile of Steele's article at his website under the following url:
`www-stat.wharton.upenn.edu/~steele/Publications/PDF/TN148.pdf`