

How I Learned to Stop Worrying and Love Statistics

Catherine A. Sugar, Ph.D.


UCLA Departments of Biostatistics, Statistics & Psychiatry
Semel Institute Statistics Core
csugar@ucla.edu

MTPCCR/Excito Webinar
October 12th, 2017



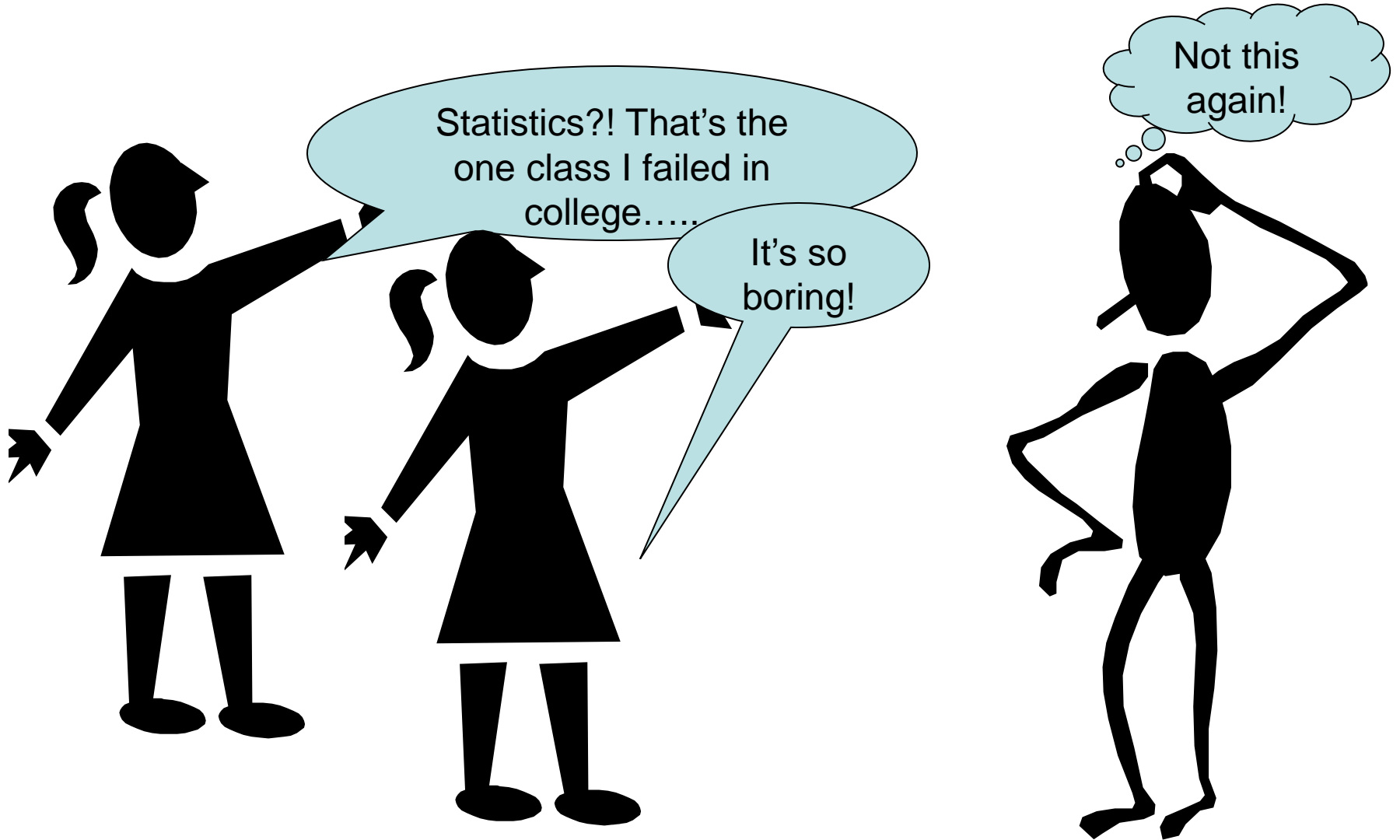
You Are Not Alone!

- Math anxiety in general and “statisticitis” in particular are very common.
- Every statistics professor I know has a set of jokes and stories aimed at this topic—in fact it was scary as I prepared this talk how many I could think of....

A newborn baby is lying in a hospital bed, covered with a white blanket. The baby's right hand is raised to their forehead. A light blue speech bubble with a black outline is positioned above the baby's head, containing the text "Please Mum no more statistics!".

Please Mum no
more statistics!

Statistics and the Cocktail Party



Where Does Stat Fear Originate?

- Some common (problematical!) memes:
 - The quantitative sciences are “hard” and you have to be “smart” or you will be “bad” at them.
 - Statistics is very abstract
 - Statistics is a very technical subject
- Notation-the jargon of math and statistics
- Uncertainty-people don't like it in general and that goes double in the quant world
- The counterfactual logical framework
- Teaching the calculations, not the concepts

The Reality

- As for any other field there is a range of abilities but most people can be perfectly good practitioners.
- You do not need to know advanced mathematics to be a successful applied user of statistics. All the notation can be translated into English.
- The uncertainty is where all the cool stuff lives.
- There is a systematic set of principles and techniques that you can learn which will give you confidence.
- In particular.....

Trust Yourself!

- Statistics is something you do every day without realizing it.
- Your scientific intuition will stand you in good stead.
- In fact, statistics is often just a way to formally quantify what you as an expert already know from looking at your data.

Statistics Is Not.....

- Recording stuff for the football team
- Reading lists of numbers
- Entering data in a computer
- Performing endless calculations
- Figuring out ways to lie with numbers
- Abstract mathematics [I know—I did that!]
- Something to be afraid of.

Statistics Is.....

- Saving the space shuttle
- Understanding elections
- Identifying genes and other factors that cause diseases
- Figuring out whether a treatment works
- Running the census
- Trying to predict the stock market
- Analyzing college admissions data
- Setting insurance premiums
- Being an expert witness in criminal cases
- Deciding if a text was written by Shakespeare
- A good way to win the Nobel Prize in Economics
- And a million other interesting and useful things involving extracting information from data

Practical Steps: Part I

- Reinforce your mathematical background.
- Understand the basic conceptual framework underlying statistics.
- Learn the vocabulary-statistics is a language. There are basic descriptive templates you can use to interpret results.
- Be willing to play around with your data and become proficient with statistical graphics. If you have the right picture it will almost always tell you the right model (and the answer!)

Practical Steps: Part II

- Create flow charts to help you understand what techniques to apply in what situations and know what the underlying assumptions are.
- Let your intuition and scientific knowledge help you! Anchor your learning in the context of a particular problem or field you understand well. Know what are or are not reasonable answers for a particular problem.
- Find a good statistician with whom to collaborate.

What Math Do You Need?

- First, statistics is not just math! Many people are bad at math but good at statistics (and vice versa.)
- You can use most statistical tools with just a good command of high-school algebra and pre-calculus.
- A course or some reading on the basics of formal logic also doesn't hurt.
- Understanding functions (e.g. polynomials, logs, exponentials) and models (e.g. the equation that describes a straight line) are the keys.
- For a deeper methodological understanding calculus and linear algebra/matrices are useful.

The Conceptual Framework: Probability vs Statistics

- **Probability:** You know the underlying mechanism of a system and you reason from it about what will happen next. (Go from a population or model to a sample.)
- **Statistics:** You have some observed data and you try to infer from it how the underlying mechanism of your system works. (Go from a sample to the population or model.)

Major Uses of Statistics

- Fundamentally, statistics is about explaining variability/understanding uncertainty.
- Major sub-goals include
 - **Estimating** key population characteristics
 - **Testing** theories about the population
 - Trying to find **patterns** in data
 - Trying to understand the **nature of relationships** among variables (form, causality(?!))
 - Making **predictions**

Important Terms and Concepts

- **Statistical Inference:** The process of using data to draw conclusions about the population or system of interest. Key components are estimation, confidence intervals, hypothesis tests, and p-values
- **Statistical Models:** Equations (based on data) that describe the relationships among multiple variables while accounting for uncertainty. The simplest example is linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

A model is a simplified version of reality. There is a tradeoff between model complexity and accuracy

What Technique Should I Use?

It depends on

- The types of variables involved in the study
- The study design
- The assumptions you are willing to make (or can justify making)
- And most of all....what question you want to answer!

Some Fun Examples....

- The best way to learn statistics is to (surprise!) practice it. The more examples you do the more natural it will seem. You don't have to do all the calculations—the key is the interpretations.
- See the statistics in the world around you. The following are some examples that illustrate a variety of statistical concepts in an (I hope!) accessible way.....

Lady Justice Is...A Statistician?

- Our entire legal system is an analogy for hypothesis testing.
- Innocence and guilt are the null and alternative hypotheses.
- The evidence presented at trial is the data.
- The significance level, α , is the standard of reasonable doubt.
- You convict if the evidence puts guilt beyond a reasonable doubt (small p-value.)



Why NASA Needs Statisticians

The Mars Climate Orbiter:

- Wrong units = \$1 billion lost

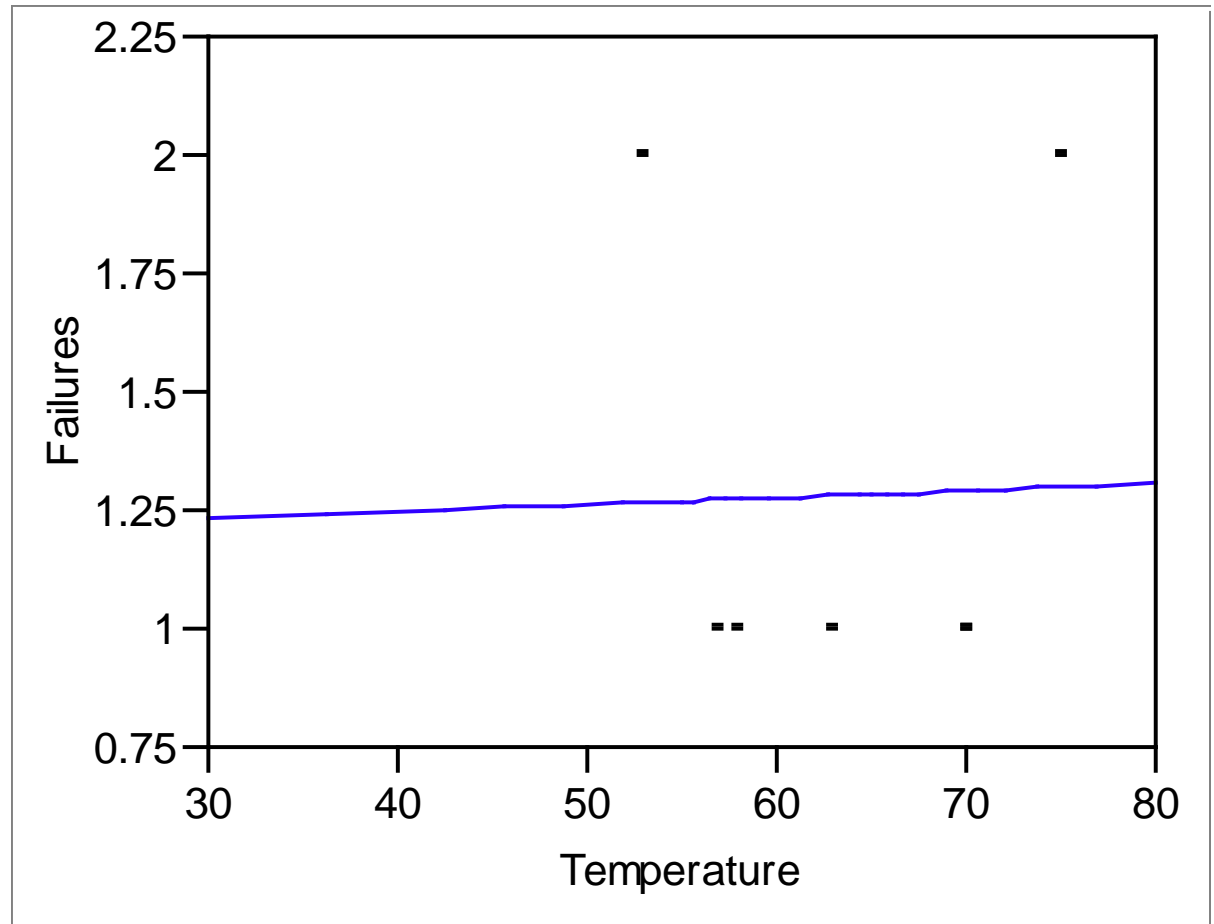
The Challenger Disaster:

- Temperatures were below freezing and engineers were concerned the cold might crack the O-rings stabilizing the fuel tanks.
- A meeting was held the night before the launch to determine whether to go ahead
- The wrong analysis was performed and all the astronauts perished on live TV with millions of school-children watching.



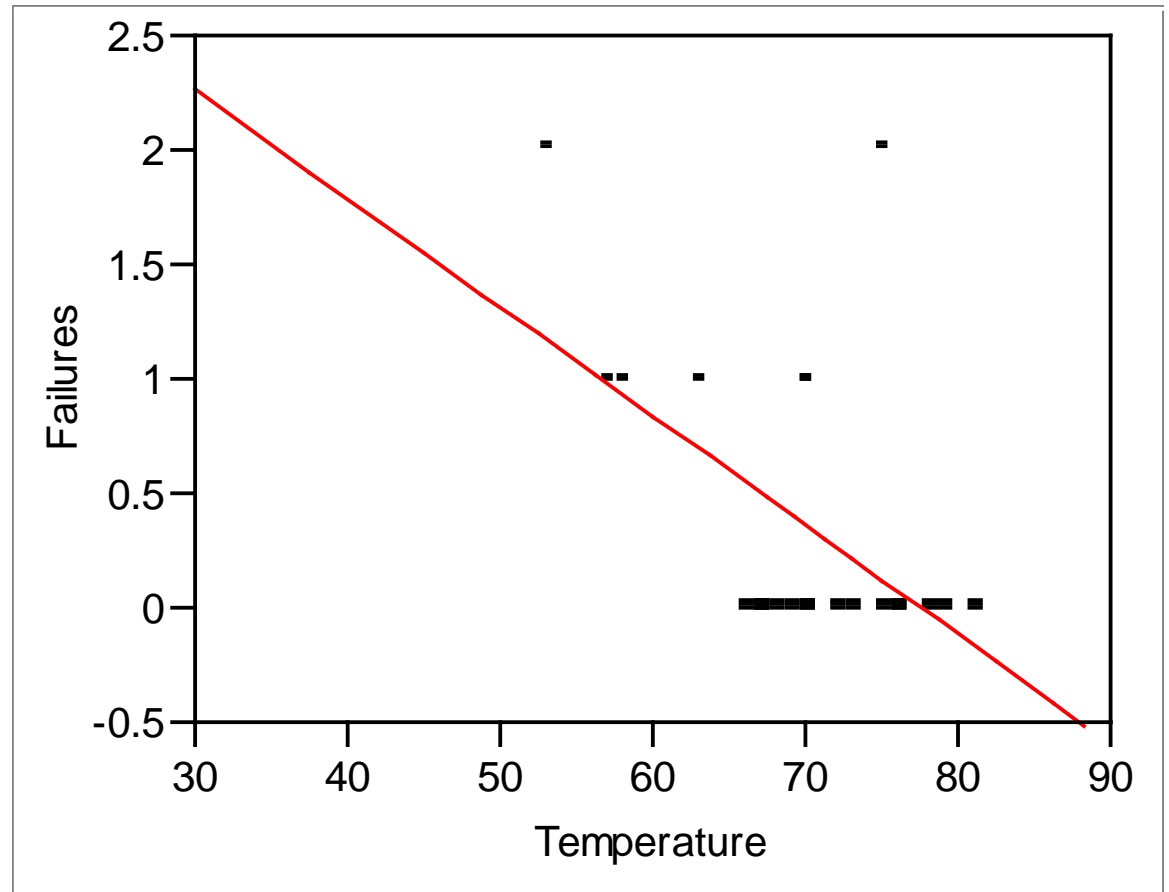
The Right Picture Is Worth....

The data used to decide whether to launch the Space Shuttle Challenger the night before it exploded.



.....7 Lives

- The data that should have been used.
- Conservative estimates put the probability of an explosion at 1/7.



What Model?

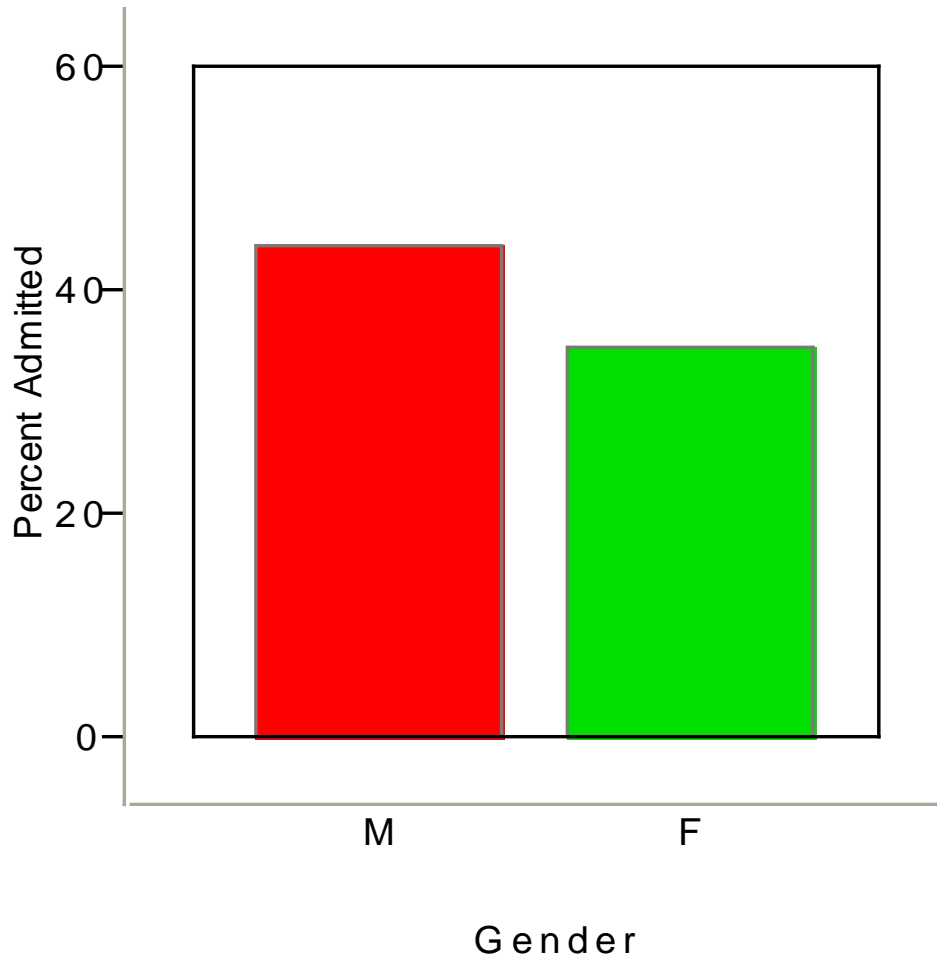
- The graph of the shuttle data clearly suggests the number or chance of O-ring failures goes up as temperature drops.
- How can we quantify this?
 - Probability of any failure (logistic model)
 - Treat the outcome as continuous (regression)
 - Number of failures (Poisson model or ordinal logistic model)

College Admissions



- Colleges and universities regularly come under fire for perceived unfairness in their admissions processes.
- There are many factors that affect who gets in and who doesn't. How can one understand whether there is a real problem?
- The following is an example of a complaint about gender discrimination in UC Berkeley graduate school admissions

Did Berkeley Favor Male Students?



44% of men who applied were admitted

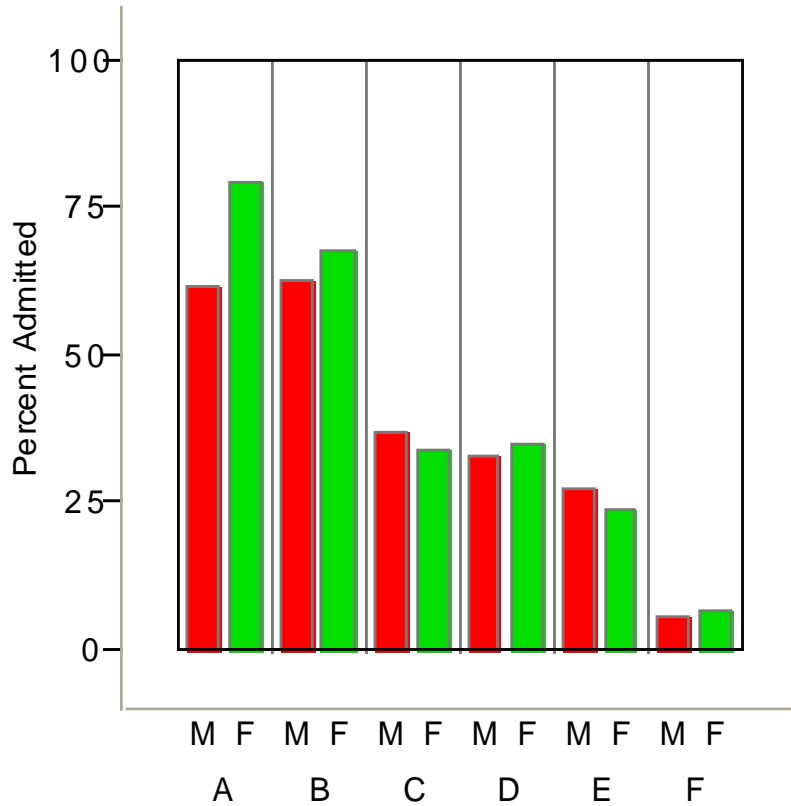
35% of women who applied were admitted

But.....

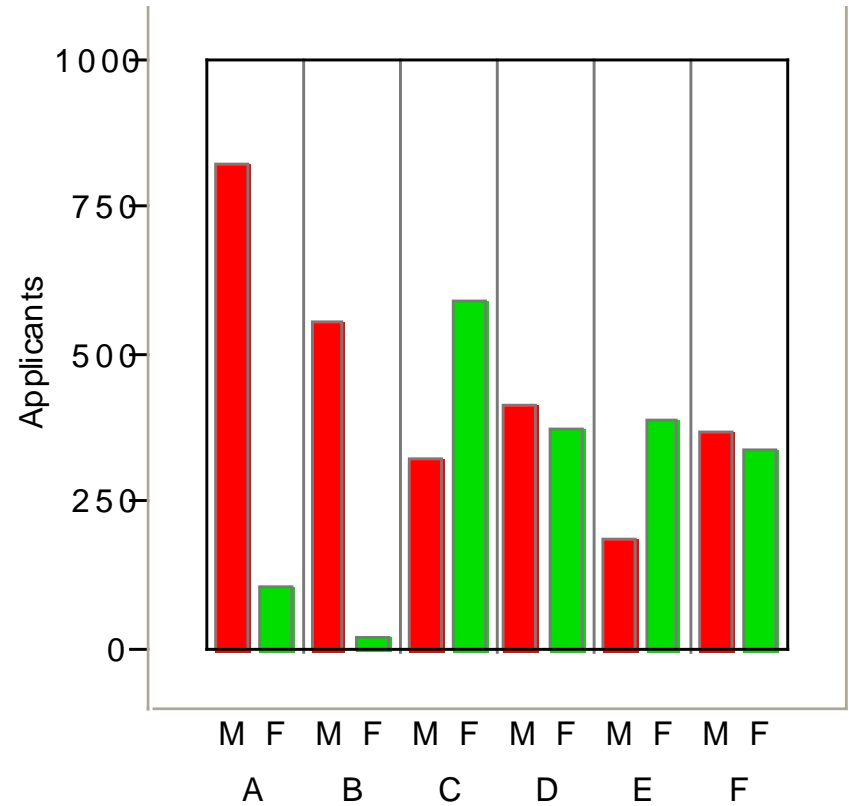
Gender  M  F

There Is A Confounding Factor:

Women Applied To More Competitive Departments!



Gender by Department



Gender by Department

Gender ■ M ■ F

Gender ■ M ■ F

What Model?

- Again, the correct set of graphics tells the story but how can we translate this into a statistical model?
- The outcome is admission (yes or no) which means a logistic model.
- We are interested in whether the probability of admission (p) is the same for men and women so gender is the main predictor.
- Department matters so we need to include it to and test whether there is an *incremental* effect of gender on top of department:

$$\text{logit}(p) = \beta_0 + \beta_1 \times \text{gender} + \beta_2 \times \text{DeptA} + \dots$$

- We could also look at interactions

Fun With Elections

- Elections are an ideal setting for illustrating many key statistical concepts in a way that fits with one's intuition.
- The following are a handful of recent (and not so recent) examples.

Dewey Defeats Truman

- Phone polls prior to the 1948 presidential election all showed the Republican candidate, Dewey, well ahead of Democrat Harry Truman



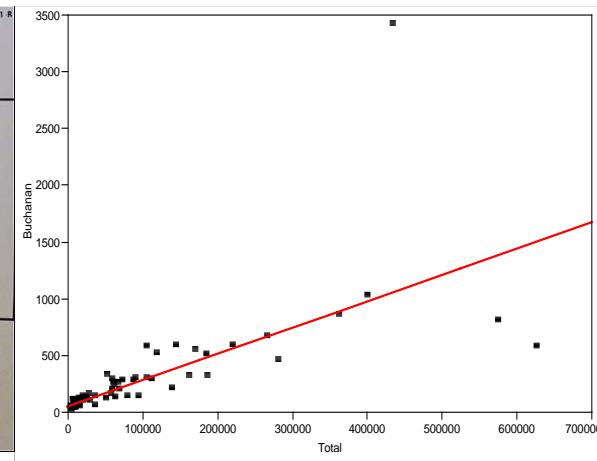
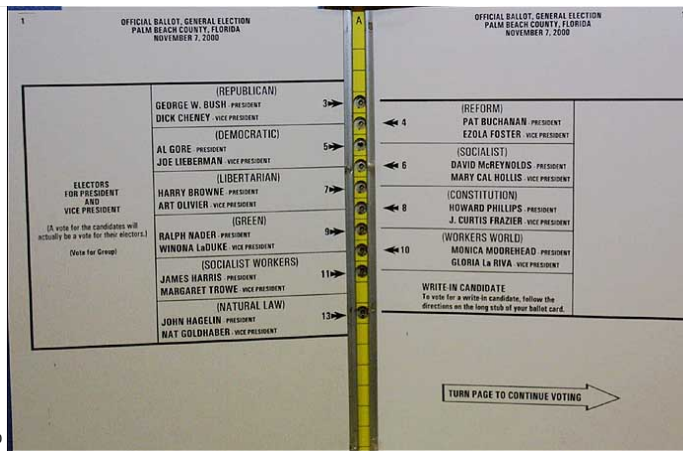
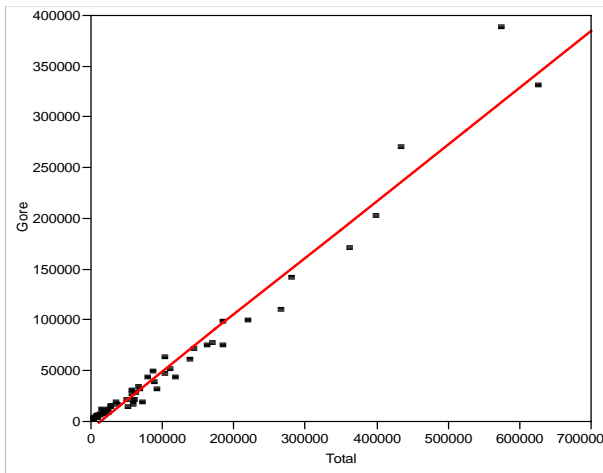
- On election night, exit polls showed Truman winning in a landslide. The networks didn't believe it and held up the results, but the exit polls were right. What went wrong with the earlier polls?
- It is critical to get a representative sample-or to adjust your results if you don't get one!
- Modern issues: Cell phones, increasing non-response, likely voter adjustments, etc.

What Is the Margin Of Error?

- Before every election we hear poll results like “Clinton leads Trump 53% to 47% with a margin of error of $\pm 4\%$.”
- We are interested in p , the proportion of people in the population who support Clinton
- The margin of error simple gives us a **confidence interval**—a range we are 95% sure includes the true value of p .
- National polls tend to be more accurate (bigger) than state/local polls and are done more often.

What Happened In Florida?

The 2000 presidential election is a source of endless statistical (and other) lessons! In Palm Beach (home of the butterfly ballot) the Buchanan vote was so significantly out of line with that in other counties that one can be statistically certain something unusual occurred. This point is called an “outlier.”

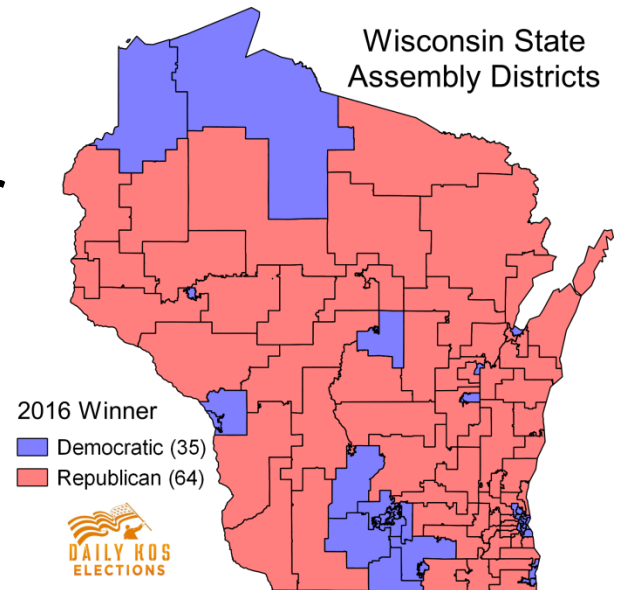


Recalling the Recall

- California law requires obtaining large numbers of voter signatures to put a measure on the ballot for an election.
- For the recall of Governor Gray Davis in 2003, LA county had to verify over 330,000 signatures, an impossible task given the short time period.
- The number of valid signatures was estimated by checking a random sample of 10,000 and extrapolating. 82% of the signatures in the sample were legitimate.

Gerrymandering

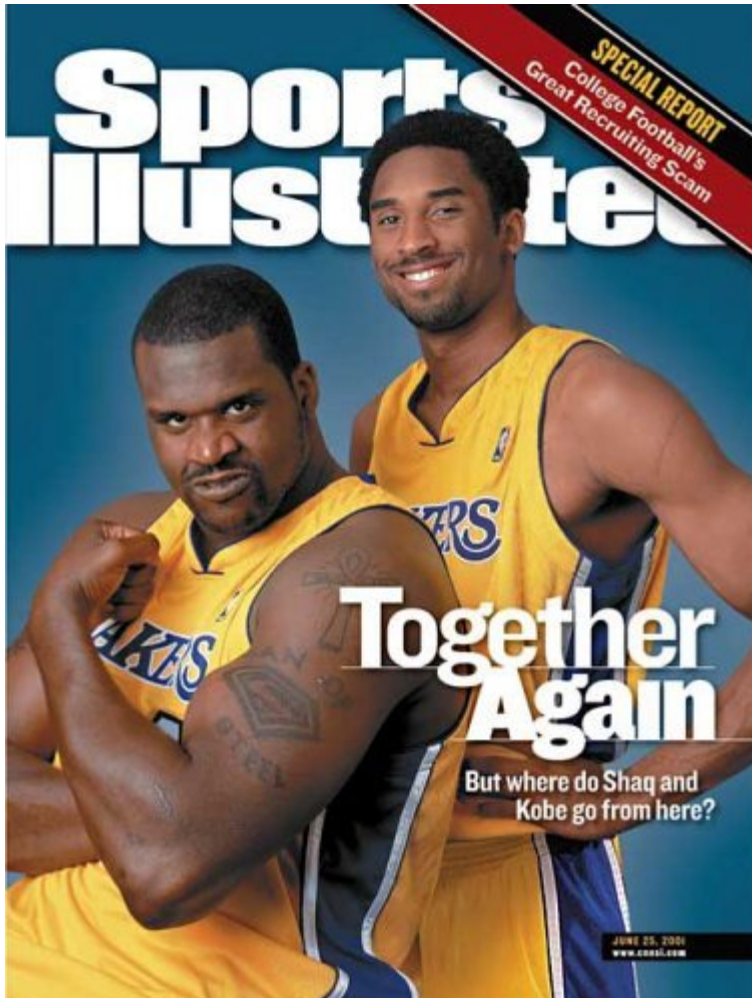
- Manipulating boundaries of electoral districts to favor or disfavor particular types of voter (→bias/uneven weighting).
- Now done via sophisticated computer modeling.
- Gill vs Whitford
 - Partisan gerrymandering case from Wisconsin, now at the Supreme Court
 - Republicans won 64 out of the 99 seats with only 52% of the statewide vote
 - A new statistic, called the `efficiency gap`, is being proposed to quantify gerrymandering



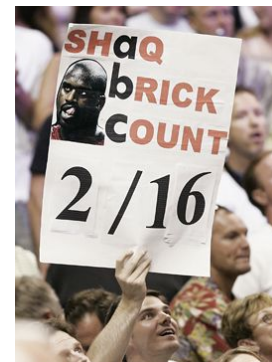
Statistics and the Census

- It is widely believed that the census under-counts certain groups of people such as the homeless
- Since much federal funding is allocated based on the census numbers, urban areas felt they were disadvantaged and sued the U.S. government
- It was proposed that statistical sampling techniques be used to “adjust” the census
- Statistical experts served as witnesses for both sides in the trials, debating the accuracy with which one could model the undercount

Shaq and Kobe



- Kobe Bryant was a 0.837 career free throw shooter.
- Shaq O'Neal was a 0.582 career free throw shooter

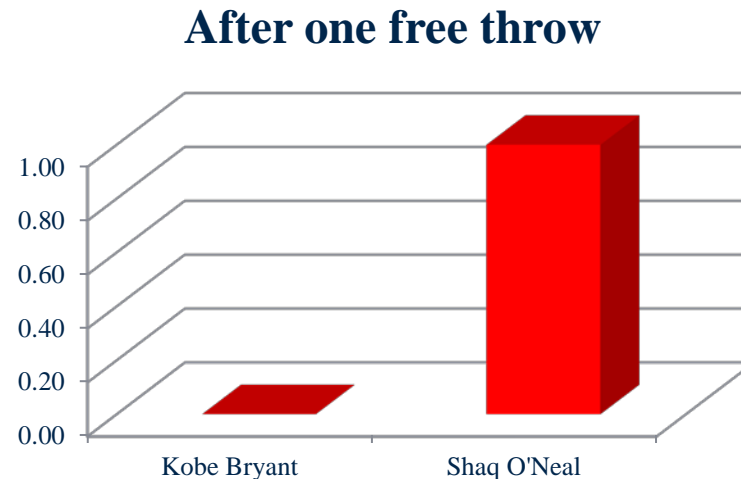


Lakers Play Cleveland



- On January 19, 2000 the Lakers played Cleveland on the way to the first of three championships.
- Shaq hits his first free throw but Kobe misses his first.
- Is this convincing evidence that Shaq is a better free throw shooter?

- Sample size is only 2.



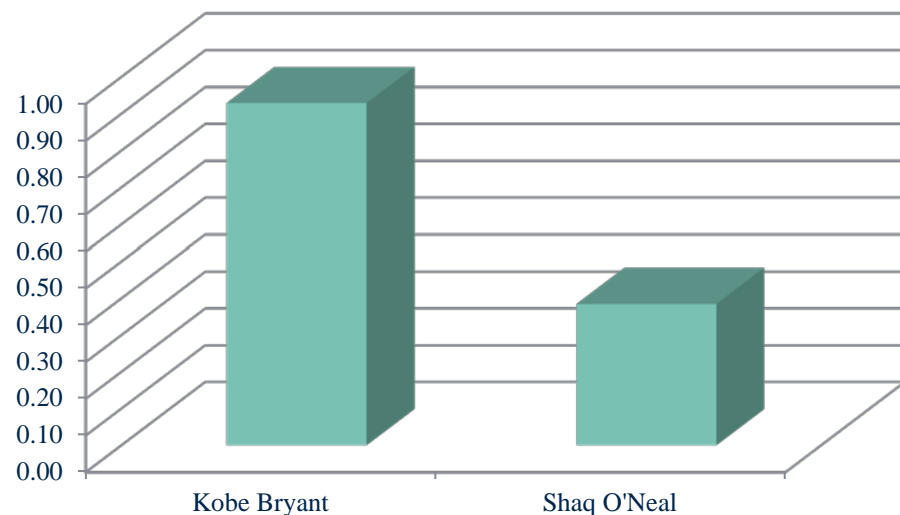
Lakers 95, Cleveland 86



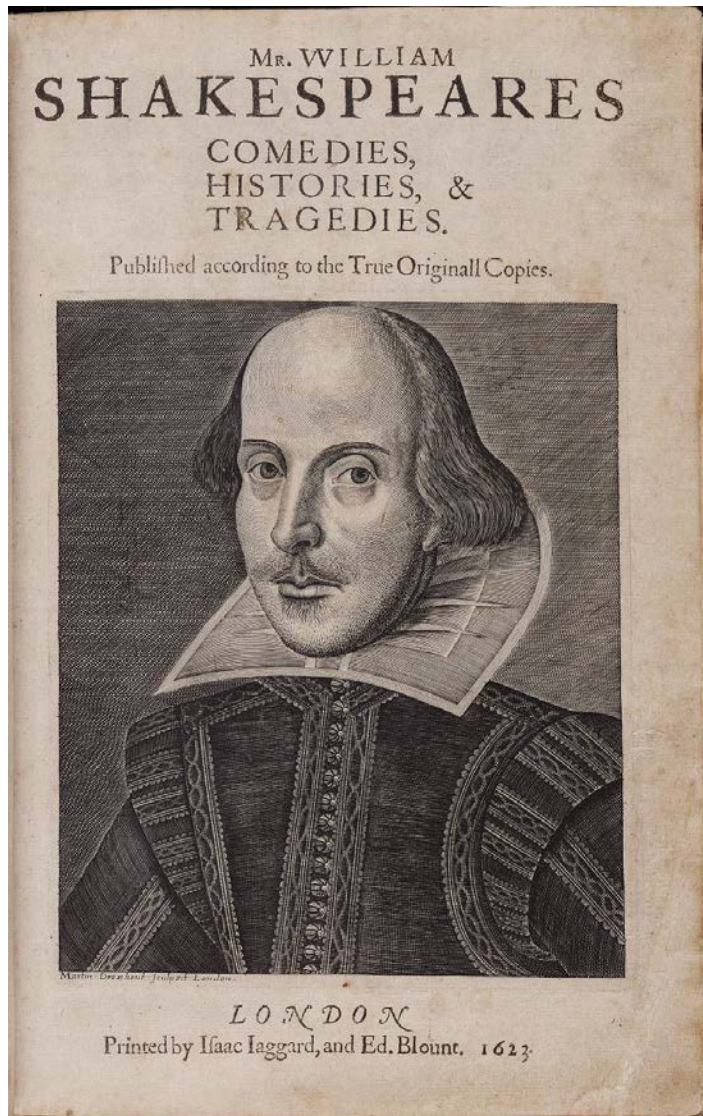
- By the end of the game Kobe has hit 13/14 (0.929) but Shaq has hit only 5/13 (0.385).
- Is this convincing evidence that Kobe is a better free throw shooter?

At end of game

- Now sample size is 27.
- P-value is 1 in 370!



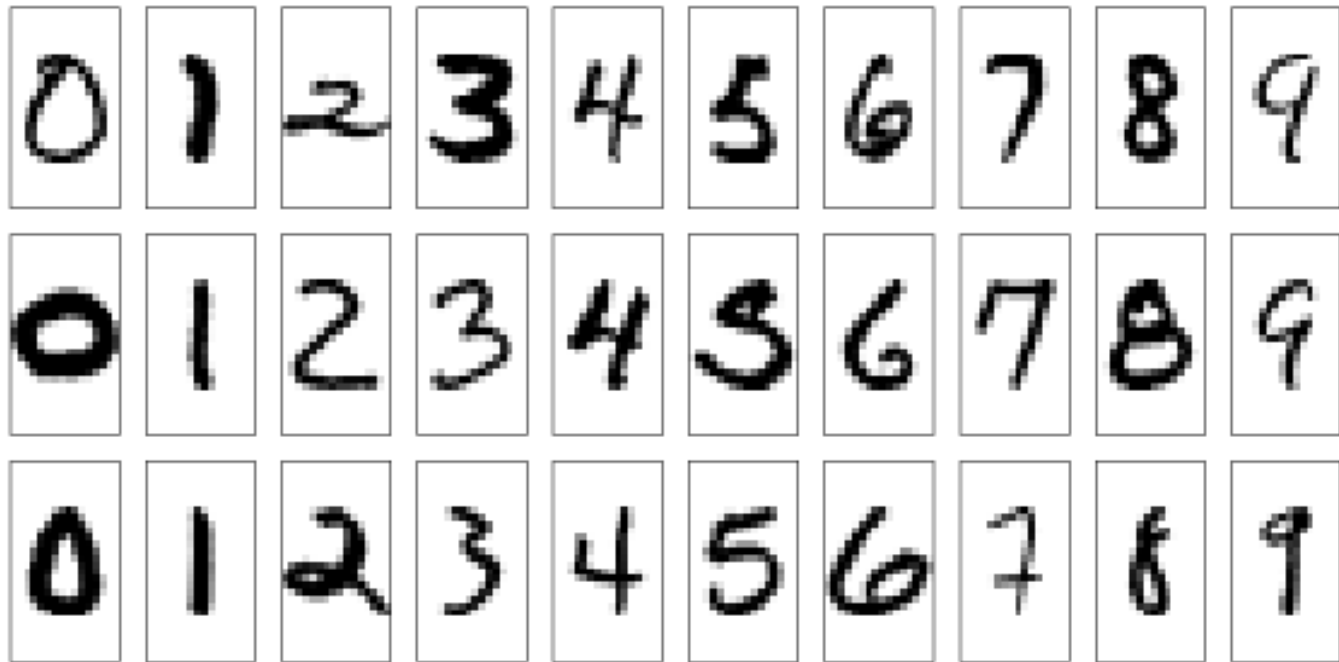
Statistics and Shakespeare:



- Authors use words and phrases in characteristic ways
- Statistical techniques can be used to analyze the frequency of word combinations in an author's known works to see if they match those in a newly discovered manuscript

Statistics and Pattern Recognition

- The US Postal Service uses “classification” techniques to automate the reading of zip-codes



- Similar techniques can be used to identify tumor types in medical scans

Conclusion

- Statistics is not as different as people think from other disciplines.
- Once you learn the language and the logic you will feel comfortable using it just like all the other implements in your scientific toolbox.
- But the best way to get over statistics anxiety is to discover that it is fun (really)! A wonderful source if you want more examples (with a very fitting title) is Hans Rosling's site

<http://www.gapminder.org/videos/the-joy-of-stats/>