Statistical Thinking at the Introductory Level

Deborah Nolan
University of California, Berkeley

Statistics and Mathematics

Adapted from Cobb and Moore, Mathematics, Statistics, and Teaching

Statistics is a

- Mathematical science
- Data science
- Computational science

- Statistics is not a subfield of mathematics
- Statistics makes essential use of mathematics

Aphorisms:

George Box:

- All models are wrong, but some are useful

George Cobb:

- In mathematics, context obscures structure. In data analysis, context provides meaning

Aphorisms:

David Moore:

- Mathematical theorems are true: statistical methods are sometimes effective when used with skill

Variability

- Need of statistics arises from the omnipresence of variability
- Repeated measurements on the same individual vary.
- Some times we want to find unusual individuals
- Other times we focus on the variation of measurements.
- Other times we want to detect systematic effects against the background noise of individual variation.

The role of context in statistics

Context

- Statistics requires a different kind of thinking
- Data are not just numbers
- Data are numbers with a context

Mathematics and Context

- Context is used for motivation
- Context is a source of problems
- Ultimately:
  - Context is the irrelevant detail that we ignore to reveal the hidden pure structure.
  - Context obscures structure.
  - Focus is on abstract patterns

Data Analysis and Context

- Focus is also on structure and patterns
- Ultimately,
  - The patterns have meaning,
  - The structure has value,
  - If the patterns make sense in the complementary threads of the story line
- Context provides meaning

Implications for teaching

Implications for teaching

- Need more than mathematical theory
- Need to understand the non-mathematical theory of statistics

AND

- Need real illustrations
- Need to use illustrations to develop critical judgment

Essential pieces of statistical analysis

- Design for data production
- Exploration for patterns and structure
- Formulation of models
- Application of methods
- Summarize results
- Interpretation of results

ASA/MAA Recommendations

- Almost any statistics course can be improved on by
  - More emphasis on data analysis
  - More emphasis on concepts
  - Fewer recipes
  - Less theory

ASA/MAA Recommendations

- Main focus of an introductory course should be on statistical thinking.
- Statistical Thinking includes:
  - The need for data
  - The importance of data production,
  - The omnipresence of variability,
  - The quantification and explanation of variability.

Statistics taught as magic

- Student is the sorcerer's apprentice
- Incantation has automatic effectiveness, e.g. renders a study publishable
- Apprentice is not meant to understand how the incantation works
- Follow the recipe exactly, better yet – use software

Counter Statistics as magic

- Retreating to mathematics does not solve the problem

Counter Statistics as magic

- Present an intellectual framework that
  - Makes sense of the collection of tools that statisticians use
  - Encourages flexible application of tools to solve problems.
  - Reasons from uncertain empirical data

Topics Today and Tomorrow

Workshop topics

1) Examples of real illustrations to develop a student's critical judgment
2) Computing technology has completely changed the practice of statistics and is a necessary tool
3) Box Model – A teaching strategy for learning how randomness in data production leads to inference

Workshop topics

4) Engaging in graphics early can establish good habits, prepare for design and inference, provide experience with data distributions, introduce concepts
5) Physical examples can give a pictoral grasp of a statistical concept and be effective in conveying ideas

Statistics Courses at Berkeley

Introductory Courses

- Quantitative Reasoning requirement for Arts and Humanities majors
- Statistics for Business majors
- Statistics for students with calculus background

Introductory Statistics Course

- 120 students
- First and second year students
- Undeclared majors, interested in economics and biological and physical science
- Calculus prerequisite for the course
- 3 hours of "lecture" a week
- 2 hours of lab – 25 students to the lab

- Three versions of the introductory course
- All focus on *statistical thinking*
- Philosophy of *Statistics,* Freedman, Pisani, and Purves
- This philosophy appears throughout the curriculum –
  - in the theoretical courses for the statistics major *and*
  - in the PhD level courses

Three stories:
Real illustrations to develop a student's critical judgment

Randomized Controlled Experiments

The HIP Trial

adapted from Statistical Models: Theory and Practice, Freedman

Background

- Breast cancer common malignancy among women
- If Detected early, then chance of successful treatment better
- Mammography – screening by X-ray

Does mammography speed up detection enough to matter?

Health Insurance Plan in New York

- HIP – group medical practice with 700,000 members in 1960s
- Subjects: 62,000 women
  - Aged 40-64
  - Members of HIP
- Split at random into two groups

Treatments

- "Treatment": invitation to 4 rounds of annual screening
  - Clinical exam
  - Mammography
- Control: received usual health care

Results

	Group Size	Breast Cancer Number	Rate
Treatment	31,000	39	**1.3**
Control	31,000	63	**2.0**

Are these the right numbers to compare?
Not all women in the treatment group accepted treatment

Results

	Group Size	Breast Cancer Number	Rate
Treatment			
Screened	20,200	23	1.1
Refused	10,800	16	1.5
Total	31,000	39	1.3
Control	31,000	63	2.0

A different comparison: those who accepted screening to those who refused.

Results

	Group Size	Breast Cancer Number	Rate
Treatment			
Screened	20,200	23	1.1
Refused	10,800	16	1.5
Total	31,000	39	1.3
Control	31,000	63	2.0

Another comparison: those who accepted screening to those in control group.

Which comparison makes sense?

Considerations

- Investigators chose (at random) those to receive treatment
- Subjects themselves decided whether or not to accept treatment
- Comparing those who accept to those who refuse is an Observational Comparison

Differences between groups

- Richer and better-educated subjects more likely to accept invitation than those who were poorer and less well-educated
- Richer women are less vulnerable to most diseases, but breast cancer hits rich harder
- Social status is a confounding factor: a factor associated with the outcome and with the decision to accept screening

Which comparison?

- The comparison of those who accept treatment to those who refuse is biased against screening
- The comparison of those who accept treatment to those in the control group is also problematic because the control group includes women who would have refused screening

Results

	Group Size	Breast Cancer Number	Rate	All Other Number	Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

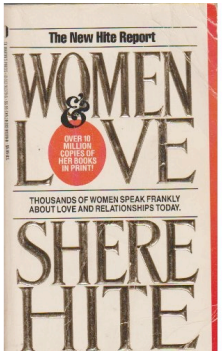
- Experimental comparison: between whole treatment group and whole control group
- Intention-to-treat analysis
- Effect of the invitation:
  - 63 – 39 = 24 lives saved
  - In relative terms 39/63 = 62%

Survey Sample

The Hite Report,

adapted from Sampling: Design and Analysis, by Lohr

Published in 1987



Survey of 4,500 women

Hite's sample & US population

Income

	Hite's Sample	U. S. population
under $2,000	19.0 %	18.3%
$2,000-$4,000	12.0 %	13.2%
$4,000-$6,000	12.5 %	12.2 %
$6,000-$8,000	10.0 %	9.7%
$8,000-$10,000	7.0 %	7.4%
$10,000-$12,500	8.0 %	8.8%
$12,500-$15,000	5.0 %	6.2%
$15,000-$20,000	10.0 %	9.8%
$20,000-$25,000	8.0 %	6.4%
$25,000 and over	8.5 %	8.2 %

Hite's sample and US population

Type of area

	Hite's Sample	U. S. population
Large city/urban	60 %	62 %
Rural	27%	26%
Small town	13 %	12 %

Race

	Hite's Sample	U. S. population
White	82.5%	83.0%
Black	13.0%	12.0%
Hispanic	1.8 %	1.5%
Asian	1.8%	2.0%

According to the Women Surveyed:

- 84% of women are “not satisfied emotionally with their relationships”.
- 70% of all women “married five or more years are having sex outside of their marriages”.
- 98% want to make basic changes in their relationships.
- 84% of women report forms of condescension from the men in their love relationships.

This survey appeared to be ground-breaking?
Or, did something go wrong?

Other Studies and Experts:

- Harris poll (1987) 89% say their relationship with their partner is satisfying.
- "Any question you asked that got 98% is either a wrong question or wrongly phrased" says Tom Smith of the National Opinion Research Center.
- Several other polls found 25-30% who are married have had or are having an extramarital affair.

Questionnaire

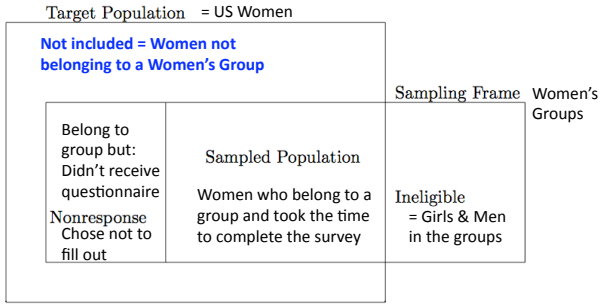
- Vague wording of questions, e.g. “***in love”***
- Leading questions:

Does your husband/lover see you as an equal? Or are there times when he seems to treat you as an inferior? Leave you out of the decisions? Act superior?

- Long survey with 127 essay questions, many with several parts

Selection Bias

- Questionnaires mailed to women's groups including professional women's organizations, counseling centers, church societies, … Non-response high: 100,000 questionnaires mailed, 4.5% returned
- Self-selected sample



Hypothesis Testing

United States v. Kristen Gilbert

(adapted from Cobb and Gehlbach, "Statistics in the Courtroom", in *Statistics: A Guide to the Unknown*)

Kristen Gilbert

- Born in 1967 in Fall River, MA
- Graduated high school at 16
- Graduated from Greenfield CC, and received certification as a registered nurse in 1988.
- In 1989, she joined the VA Medical Clinic in Northampton, MA.

VA Medical Clinic

Gilbert established a reputation of being particularly good in crisis

When a patient went into cardiac arrest, she would:

- Sound the code blue alarm
- Stay calm
- Administer a shot of epinephrine to restart the heart

Suspicions

- By the mid 1990's, other nurses had become suspicious of Gilbert.
- It seemed there were too many code blue calls, too many crises when Gilbert was on the ward.
- An initial VA report found that the number of deaths were consistent with patterns at other VA hospitals.
- The suspicions of the staff remained.

Suspicions

- The staff brought their concerns again to the administration of the VA Clinic.
- They hired a statistician as a consultant to look into the situation (Gehlbach)
- His findings agreed with the staff concerns.

Assistant U.S. Attorney Welch convened a grand jury in 1998 to hear the evidence against Gilbert.

Grand Jury

- A grand jury determines whether there is enough evidence for a trial.
- The grand jury examines evidence and issues an indictment, a formal accusation that a person has committed a crime.

Evidence Considered

- **Motivation** – Gilbert liked the thrill of a crisis, needed the recognition, and wanted to impress her boyfriend who also worked at the VA Clinic.
- **Testimony of co-workers** about access Gilbert had to epinephrine.
- **Testimony of a physician** about the symptoms of the men (healthy, middle-aged, not typical candidates for cardiac arrest).

Convincing?

- No one had seen Gilbert give fatal injections.
- A major part of the evidence was **statistical**.

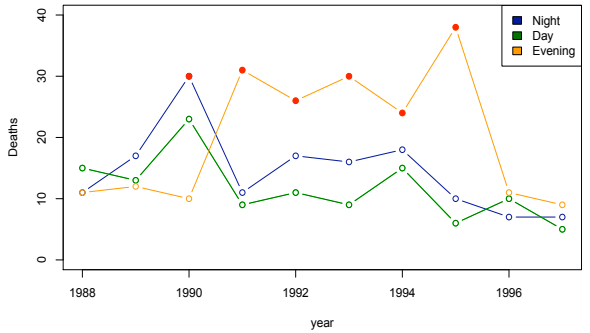
QUESTION:

Were there so many excess deaths when Gilbert was present as to be suspicious in the eyes of science?

Gelbach's Testimony

- Pattern of deaths, by shift and by year on the medical ward where Gilbert worked
- Variability in a chance process
- Statistical test for whether the pattern linking the excess deaths to Gilbert's presence on the ward was too extreme to be regarded as ordinary, expected variability

Pattern in Deaths



Pattern in Deaths

- There is a clear pattern associating Gilbert's presence with excess deaths
- However, the pattern *might* be nothing more than the result of ordinary, expectable **variation**.

Statistical Test

- Consider the 18 months leading up to Feb 1996, when Gilbert went on medical leave.
- There were 547 days in this 18 month period and 3 shifts a day, for a total of 1641 shifts
- For each shift, we have whether or not Gilbert worked the shift and whether or not there was a death on the shift.

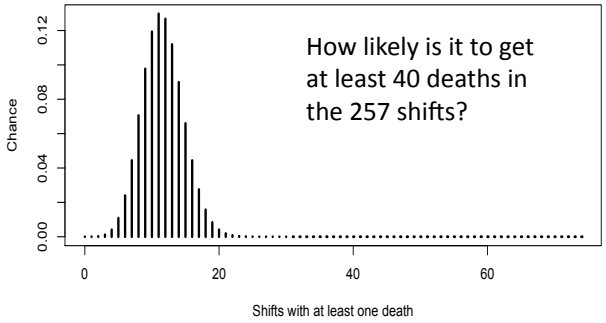
A Statistical Test

	DEATH ON SHIFT?		
GILBERT PRESENT?	Yes	No	Total
Yes	40	217	257
No	34	1350	1384
Total	74	1567	1641

- The table summarizes records.
- In the 1641 shifts there were 74 shifts for which there was at least one death.
- On 40 of the 74 shifts, Gilbert was working.

Is that more than you would expect?

P(at least 40) < 1 in a trillion



Grand Jury

- The grand jury found the statistical evidence persuasive
- Gilbert was indicted
- The VA hospital is legal property of the federal government so it was a federal indictment.
- Trial would be in federal district court, and the death penalty would be a possible sentence, if found guilty.

The Trial

- The **petit jury** (or **trial jury**) hears the evidence in a trial as presented by both the plaintiff and the defendant.
- After hearing the evidence, the group retires for deliberation, to consider a verdict.
- The majority required for a guilty verdict was a simple majority (7 out of 12). A unanimous verdict for the sentence was needed for the death penalty .

Expert Witness

The court system allows expert testimony when the evidence involves specialized technical or scientific issues that go beyond what jurors would ordinarily be familiar with.

The experts help the jury understand the evidence better.

Dueling Expert Witnesses

- Supreme Court has set guidelines at making sure unscientific testimony is not admitted.
- If there is expert testimony on one side, attorneys for the other side sometimes hire another expert who will disagree and "cancel out" the other expert.

QUESTION:
Should the trial jury be allowed to hear the statistical evidence?

Report to the Judge

- Was the statistical analysis done correctly?
- What does the probability calculation NOT tell you?
  - Association vs Causation: Conclusions drawn from an Observational Study vs a Designed Experiment
  - Prosecutor's Fallacy: Probability computed under assumptions of innocence

Was the statistical analysis done correctly?

- The Defense Statistician (Cobb) agreed with the analysis performed by the Prosecution Statistician for the grand jury.
- The number of excess deaths was an extremely unlikely outcome due to chance variation.
- The pattern of deaths justified the indictment.

Association vs Causation

- In a well-designed experiment, e.g. the Salk Field trials, the only difference is in the treatment, all other possible causes of an effect have been eliminated.
- The tiny probability rules out chance variation, and the conclusion is that the difference is "real".
- In a designed experiment, we can conclude that the explanation for the observed difference is the treatment.

Association vs Causation

- This was NOT a randomized controlled experiment. (Gilbert's presence on the ward would have had to be assigned using a chance device).
- We can conclude that the difference is not due to chance variation, but the tiny probability does not provide an explanation for what happened.
- There could be other possible explanations.

Prosecutor's Fallacy

- The probability of 1 in a trillion was computed **assuming** that a result is due to chance variation.
- We compute the chance of getting a result as extreme as the one observed.
- If it is very rare (tiny probability), we conclude that it is not reasonable to think that random variation is the cause.
- This logic says **nothing** about other causes

Prosecutor's Fallacy – slippery logic

- Suppose Gilbert is innocent and the deaths behave in a chance-like way. The probability is less than 1 in a trillion that you would see so many excess deaths on Gilbert's shifts.
- If Gilbert is innocent, then it would be almost impossible to get so many excess deaths.
- With this many excess deaths, the chance is less than one in a trillion that Gilbert is innocent. **FALSE LOGIC**

Conclusion

- It is very easy to be tempted by the false logic (that the probability is the chance of innocence).
- Judge ruled that statistical evidence should not be allowed at trial.
- Jury found Gilbert guilty on 3 counts of first-degree murder and 2 counts of attempted murder.
- Jury voted 8-4 for a death penalty.
- Gilbert was given life in prison without possibility of parole.

Exercises to try

Considerations

- Provide an answer in "plain English" or "plain Chinese"
- Identify the core statistical thinking concept
- How is the math insufficient for answering the question?

Discussion

- What are the challenges to teaching this way?

- Hard to teach these concepts
- Hard to grade student work
- Context can depend on cultural background
- Often the examples are very reductionist
- This can lead to an over-critical approach

Discussion

- What are the benefits to teaching this way?

- See that there is more to statistics than manipulation of formula
- Gain practice in statistical thinking
- Seeing many examples will help when confront new problems

Statistics should be taught as statistics

Resources

- *Statistics*, Freedman, Pisani, Purves

- *Statistics a Guide to the Unknown*, Mosteller et al, edition
- *Statistics a Guide to the Unknown*, Peck et all, edition
- *Stat Labs: Theory though Applications*, Speed & Nolan