

Welcome to 6010!

Sumanta Basu

About the class

Staff

Fall 2019

Instructor: Sumanta Basu

TA: Shagun Gupta, Adam He, David Kent, Daniel Kipnis

Author of online supplements: Kara Karpman

Resources

Blackboard: <https://blackboard.cornell.edu>

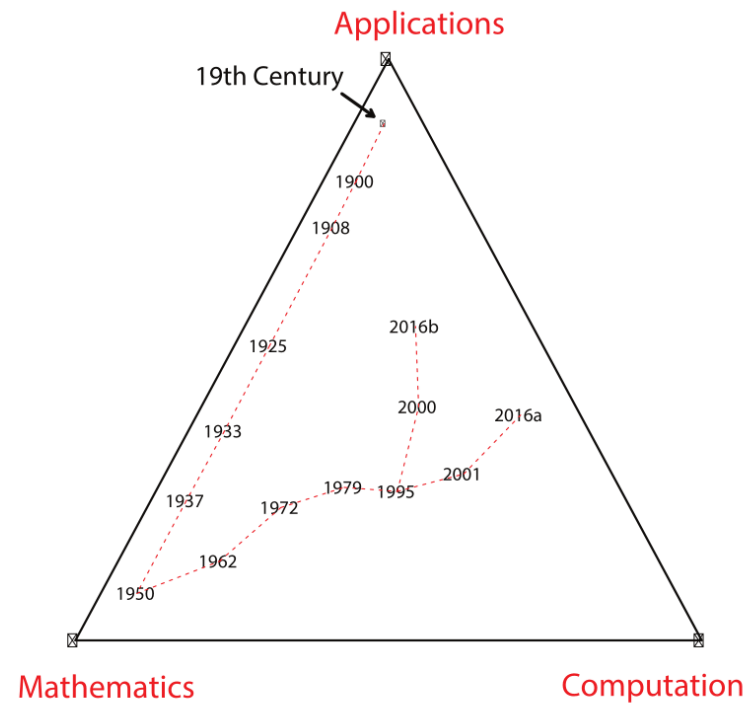
Piazza: piazza.com/cornell/fall2019/btry6010

Instructor OH: 1192 Comstock Hall, Tuesdays 5-6 pm

TA OH: Mon, Thu, Fri (5-6 pm), Wed 5:30-6:30 pm

Syllabus, textbook (youtube video, exercise, R examples),
supplements

Three Main Components



Development of the statistics discipline since the end of the nineteenth century, as discussed in the text.

Lectures: Applications ↔ Mathematics, Labs: Applications ↔ Computation

Homeworks

- Due Saturdays at 11:59 pm (counts only 10% of grade, but useful for exam prep)
- HW1 posted on blackboard, due September 7, 11:59 pm
- Getting software set up
- Three things to submit: pdf, R Markdown, screenshot image

Late Submission Policy (for all assignments):

- No late submission accepted, two HWs with lowest scores dropped in the end
- Make sure R Markdown and pdf are submitted to the right link. You have a 30 minute grace period after deadline to fix this if needed

Tips for succeeding in 6010

- Attend lecture and lab; read ahead of attending if possible
- Take advantage of office hours
- Make use of the textbook: reading, youtube videos, exercises
- Practice R: use resources on syllabus and blackboard
- Post doubts on piazza, work with a study partner

Why this class is special

- a step removed from subject matter
- focus on *the process* of research
- learn the thinking and the tools
- a great opportunity to meet future scientists in other fields!

Why learn statistics?


An (unfairly) narrow view


"If your experiment needs statistics, you ought to have done a better experiment." -Lord Ernest Rutherford

or

"I just need to know how to get a p-value < 0.05 so the referees will leave me alone." -Anonymous 6010 student

This was actually published

 **NCBI** [Resources](#) ☒ [How To](#) ☒


US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Search

Abstract

Send to:

[BMC Syst Biol.](#) 2011;5 Suppl 3:S4. doi: 10.1186/1752-0509-5-S3-S4. Epub 2011 Dec 23.

An integrative analysis of DNA methylation and RNA-Seq data for human heart, kidney and liver.

[Xie L](#)¹, [Weichel B](#), [Ohm JE](#), [Zhang K](#).

Author information

Abstract

BACKGROUND: Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation plays during normal cellular differentiation and tissue specification. We also need thorough, genome level studies to determine the meaning of methylation of individual CpG dinucleotides in terms of gene expression.

RESULTS: In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arrays and compared those to gene expression by RNA sequencing. We have identified unique signatures of global DNA methylation for human heart, kidney and liver, and showed that DNA methylation data can be used to correctly classify various tissues. It indicates that DNA methylation reflects tissue specificity and may play an important role in tissue differentiation. The integrative analysis of methylation and RNA-Seq data showed that gene methylation and its transcriptional levels were comprehensively correlated. The location of methylation markers in terms of distance to transcription start site and CpG island showed no effects on the regulation of gene expression by DNA methylation in normal tissues.

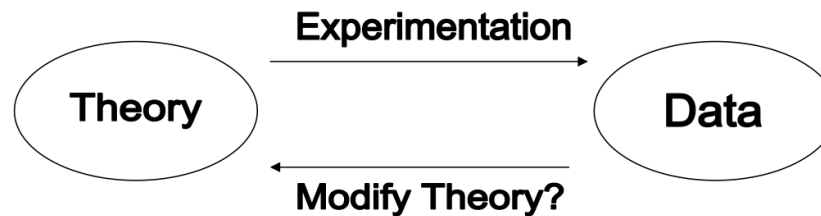
CONCLUSIONS: This study showed that an integrative analysis of methylation array and RNA-Seq data can be utilized to discover the global regulation of gene expression by DNA methylation and suggests that DNA methylation plays an important role in normal tissue differentiation via modulation of gene expression.

Why learn statistics?

“Statistics is the grammar of science.” -Karl Pearson

- statistics as **a way of thinking**
- how to make sound and convincing quantitative arguments
- develop **quantitative common sense**
- focus on **process**, procedure used (connects to reproducibility)
- embrace randomness

Statistics: Learning from Data



In general, statistics is concerned with

1. Systematic methods for data collection
2. Objective methods for data analysis and “inference”
3. Careful interpretation of results

Key steps in any statistical analysis

Sample Selection

Data Visualization

Data Summarization and exploratory analysis

Inference: from sample to population

Interpretation of results: association or causation?

Failure to address any of the above can lead to poor decision-making

Key questions to keep in mind

Collection of Data

- Will your study **generalize**: sample **represents** population?
- Correlation or causation: have you missed a **confounding factor**?
- Answer: Statistical frameworks for sampling, and design of experiments

Analysis of Data

- Are your results **due to chance alone**?
- If not, **how sure** are you?
- Answer: Statistical frameworks for measuring uncertainty

Some examples

Example: Sample Selection Bias

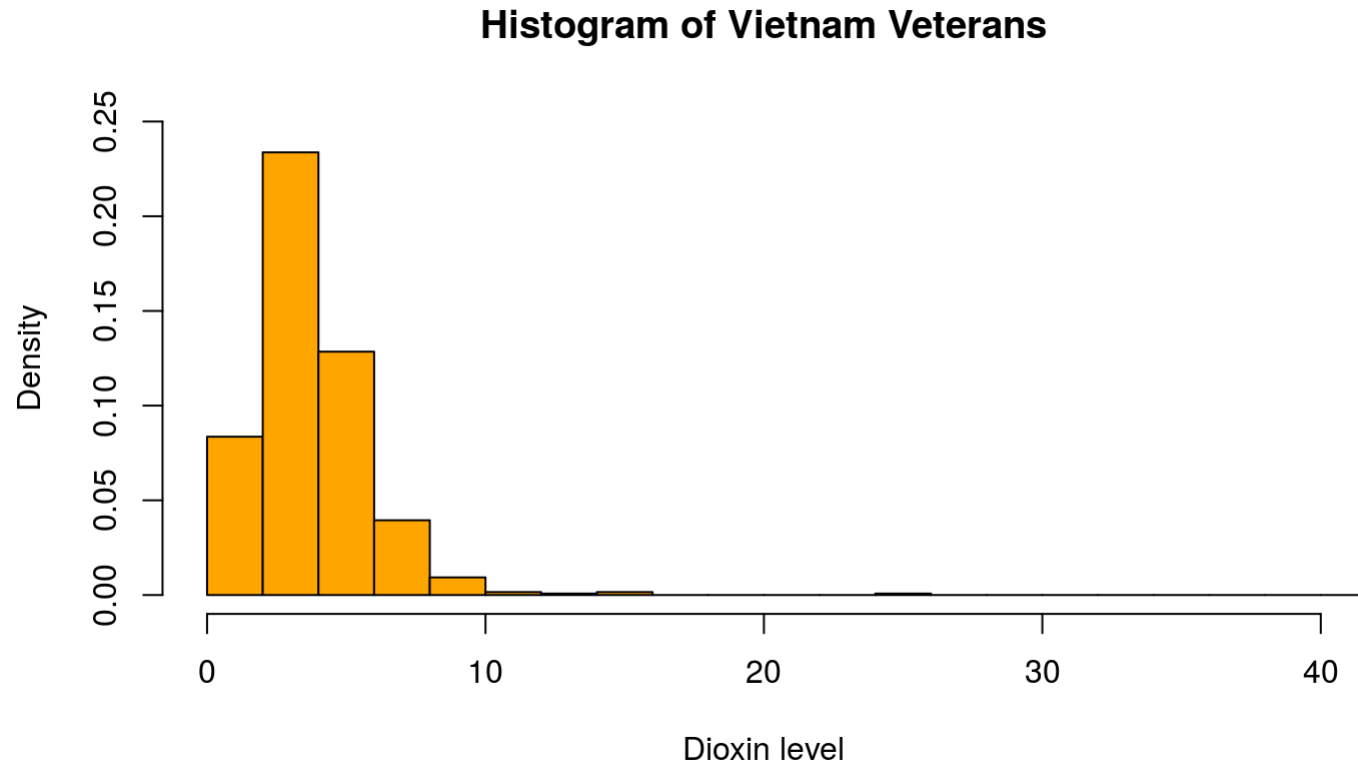


Exposure to Agent Orange

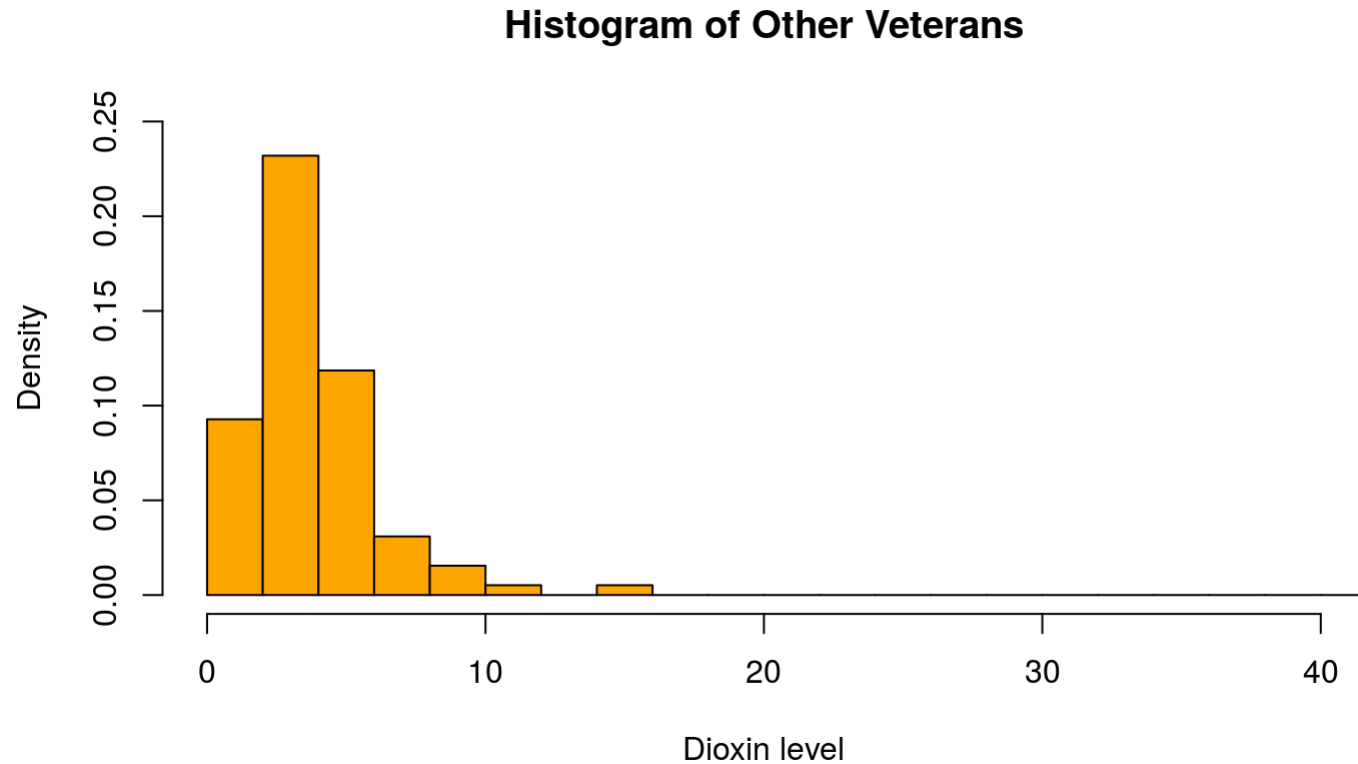
- During Vietnam War, US military used Agent Orange herbicide, 1962–1970
- [1987 study](#): Do Vietnam veterans have higher levels of dioxin than veterans who served elsewhere?

Source: [Ramsey & Schafer \(2002\)](#)

Exposure to Agent Orange



Exposure to Agent Orange



Exposure to Agent Orange

1987 study's conclusion:

"This study is consistent with other studies and suggests that most US Army ground troops who served in Vietnam were not heavily exposed to TCDD, except perhaps men whose jobs involved handling herbicides."

Question: Do you agree with the conclusion?

A. Yes

B. No

What if having a high dioxin level means you die young?

->

Example: Correlation or
Causation?

Is Sitting a Killer?

[NY Times 2016:](#)

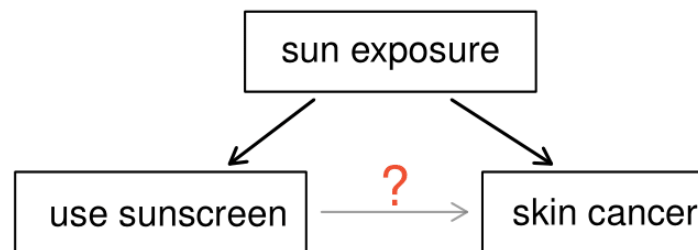
Should you be standing right now?



Sunscreen and Skin Cancer

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

- Source: Textbook Section 1.4.1



Why is Association \neq Causation?

- **Confounding factors**
- Systematic difference between treatment and control groups
- often exists in observational studies

Accounting for (more) confounding factors can lead to more accurate conclusion

Why not always do this?

- Limited sample capacity

A Classic Example

John Snow and the Broad Street Pump

An earliest example of controlled experiment

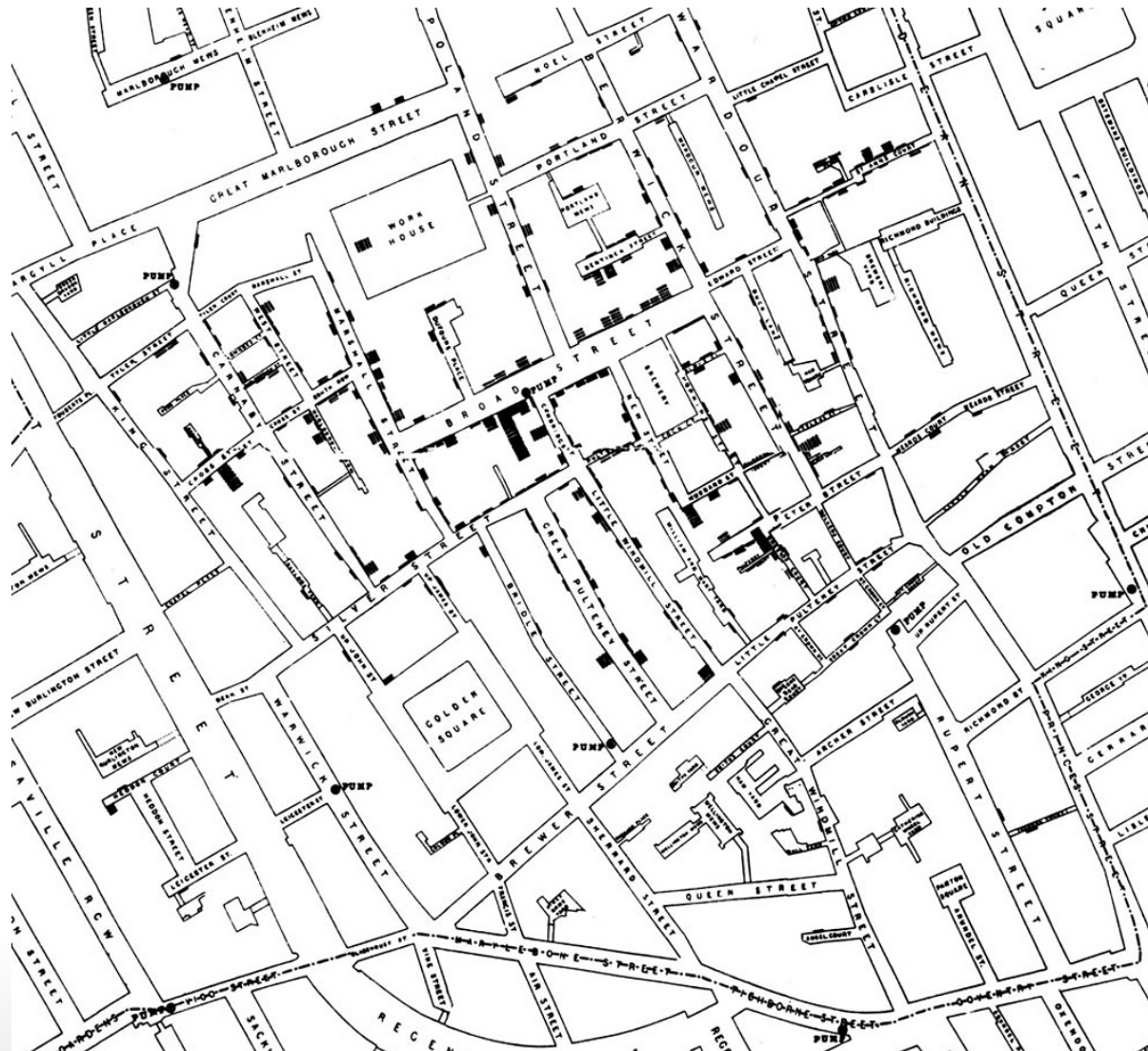
- Cholera outbreak in 18th century London
- Leading theory of cause: **miasma** (bad smell)
- Common remedy: hold sweet-smelling things to nose

John Snow, a doctor and scientist, suspected water contamination by sewage

Reference: <http://data8.org/>

John Snow (1813-1858)

Snow's Map of Cholera Outbreak

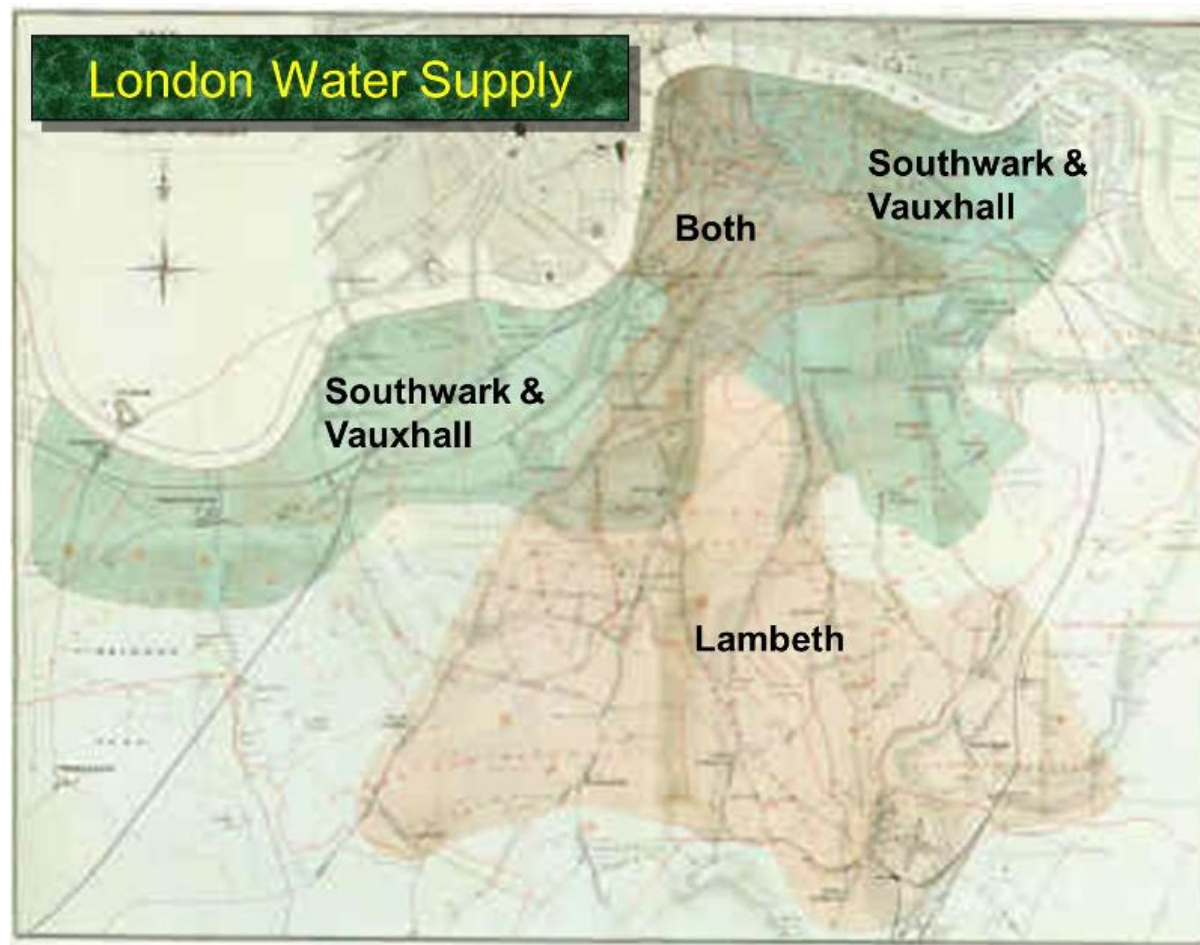


John Snow Pub (London, 2017)



Note the missing handle on the pump :)

Two Water Suppliers in London



Snow's Experiment

“...there is no difference whatever in the houses or the people receiving the supply of the two Water companies, or in any of the physical conditions with which they are surrounded ...”

- The groups were similar except for their water suppliers
- One of the earliest attempt to account for confounding factors

Snow's Experiment

Water Company	# Houses Served	# Cholera Deaths	Death Rate per 10,000 Houses
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
The Rest of London	256,423	1,422	59

About 10 times higher incidence of Cholera Death for Southwark & Vauxhall

Some Take-Aways from the examples

- Agent Orange: *How you sample* data is crucial.
- Sitting: Headlines are often *causal* regardless of what was established.
- John Snow: Accounting for confounding factors brings us *closer to* causality

Example: Are your results due to chance alone?

Angel of Death: United States v. Kristen Gilbert

- Kristin Gilbert was a respected nurse at a Massachusetts hospital (1990 - 1996)
- Some of her colleagues noticed **higher incidence** of cardiac events during her shifts
- A grand jury needed to assess, based on available evidence, whether there is *probable cause* that she committed a crime

The Data

Gilbert Present	Number of Shifts w/ Death	Number of Shifts w/ No Deaths
Yes	40	217
No	34	1350

Question: Would you think there is *probable cause*, i.e. these results are **NOT** due to chance alone?

A. Yes

B. No

Question: How sure are you, i.e. what are the chances that this is just a coincidence?

Reading: blackboard folder 'suppelement/Gilbert'