Changing Times Call for Changing Stats

Danny Kaplan

USCOTS "Ignite" Session May 15, 2013

A Physicist Lapsed into Statistics

Rutherford (1871-1937)

DTK

- I studied physics, philosophy, and political science in college.
- I never took a statistics course.

Lord Rutherford

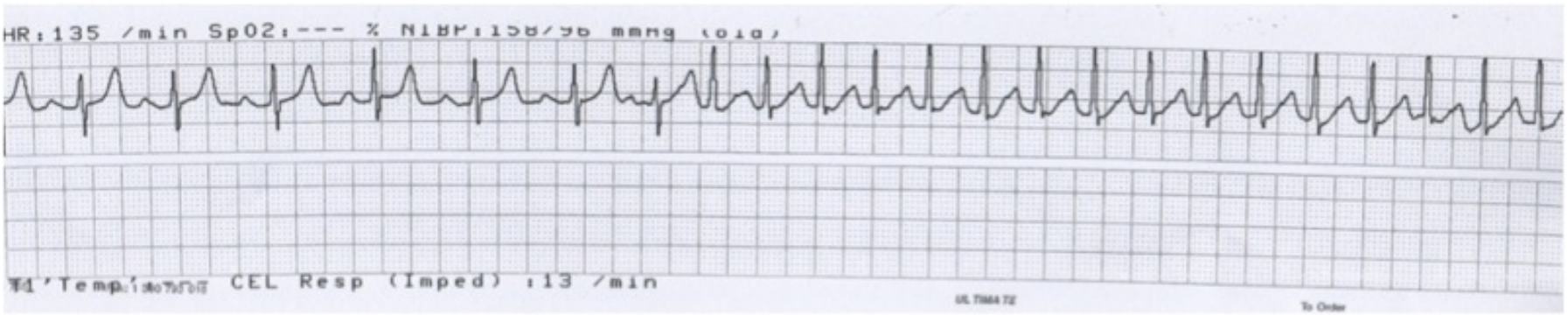
"If your experiment needs statistics, you ought to have done a better experiment."

Biomedical Engineering and Physiology

My Ph.D. work

Getting physiological information from signals

- ECG, EEG, Respiration and heart rate variability



- Signal processing
- Classification and Bayes' Rule, not hypothesis testing
- Sensitivity and specificity, not Type I and Type II

But You Need a CI to Publish

Sources for scientists:

- **Numerical Recipes**
- Brad Efron's **SIAM**

SIAM REVIEW
Vol. 21, No. 4, October 1979

0036-1445/79/2104-0002$01.00/0

COMPUTERS AND THE THEORY OF STATISTICS: THINKING THE UNTHINKABLE*

BRADLEY EFRON†

Abstract. This is a survey article concerning recent advances in certain areas of statistical theory, written for a mathematical audience with no background in statistics. The topics are chosen to illustrate a special point: how the advent of the high-speed computer has affected the development of statistical theory. The topics discussed include nonparametric methods, the jackknife, the bootstrap, cross-validation, error-rate estimation in discriminant analysis, robust estimation, the influence function, censored data, the EM algorithm, and Cox's likelihood function. The exposition is mainly by example, with only a little offered in the way of theoretical development.

1. Introduction. The editors have been kind enough to invite a survey article concerning what's new in the theory of statistics. Any answer to this question must be either incomplete or bewildering to the reader. Here I have tried to be incomplete, selecting my topics to illustrate a special point: how the advent of the high-speed computer has affected the theoretical structure of statistics.

"Yes" is the Correct Answer

Job Interview Question:

"Can you teach intro stats?"

Thank goodness I wasn't asked, "What's a t-test?"

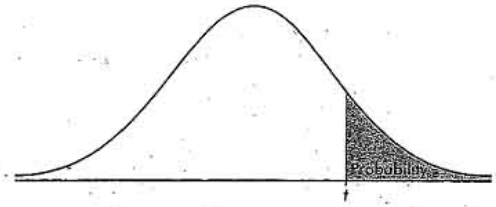


TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Resampling Seemed the Natural Way to Teach and Learn

I worked to make resampling easier and more accessible:

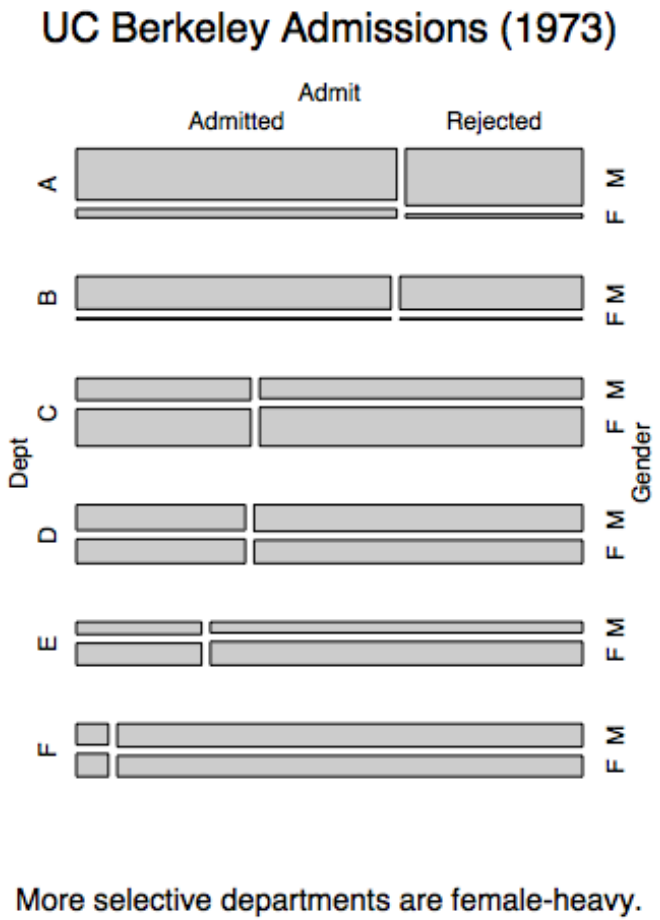
```
s1 = do(100)*median( Price, data=resample(houses) )
sd(s1)

## [1] 2495

s2 = do(100)*diff( median(Price~Fireplace,
                          data=resample(houses)))
sd(s2)

## [1] 4108
```

Simpson's Paradox and Philosophy



More selective departments are female-heavy.

I thought students should know something about inductive reasoning and its traps.

Simpson's Paradox seemed a good angle to give students an appropriate level of skepticism about induction.

The Wrong Lesson

What I was saying ...

Data don't speak, they inform our judgment.

Interpret data in the context of a whole system

What they were hearing ...

The data will say anything you want, depending on how you cut it.

Avoiding the Abstinence-Based Curriculum

Causation is often the issue. But ...

- Confounding is common
- Adjustment provides insight if not proof
- It's very common in the literature

I don't want students to be powerless about covariation and causation.

But how to do this?

"Never, ever, think outside the box."

Never, ever, think outside the box.

My epiphany

The links between modeling and geometry of subspaces.

Case Space versus **Variable Space**

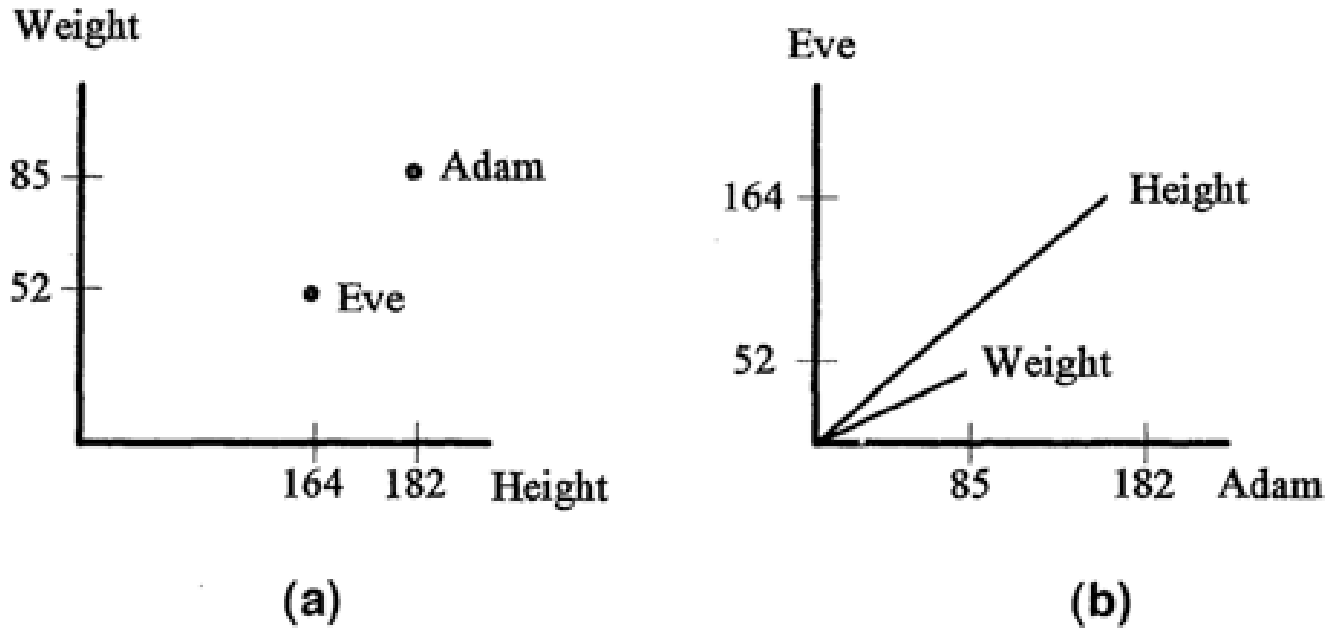
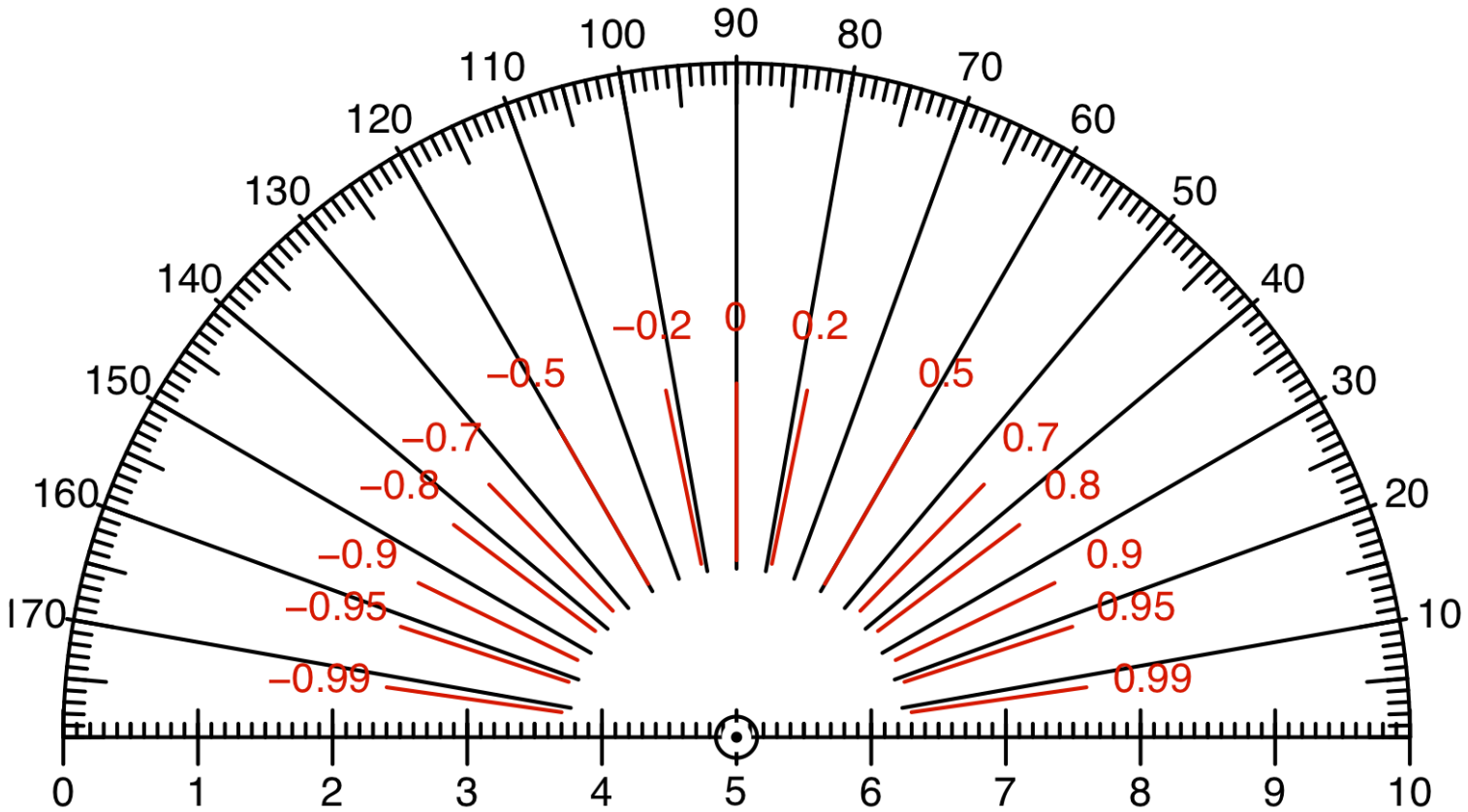


Figure 1. Two Geometrical Presentations of the Data in Table 1: (a) Variable-Axes and (b) Observation-Axes.

Correlation is an Angle

Statistics formulas (e.g. correlation) are based on linear algebra but fail to present the operations at a high level.



When it's hard, We're doing the wrong thing

The purpose of expertise should be to find ways to make it obvious.

A 7-year old eyeballs the p-value

A difficult theoretical question ...

Where does the t-distribution come from?

The back of the book!

Things Have Changed

In the 16 years since my job interview...

- Every student has access to a computer, in class and out.
- Software is free.
- Data is everywhere.
- People want to use data for decision-making, not just for publishing research.

We Should Change, Too

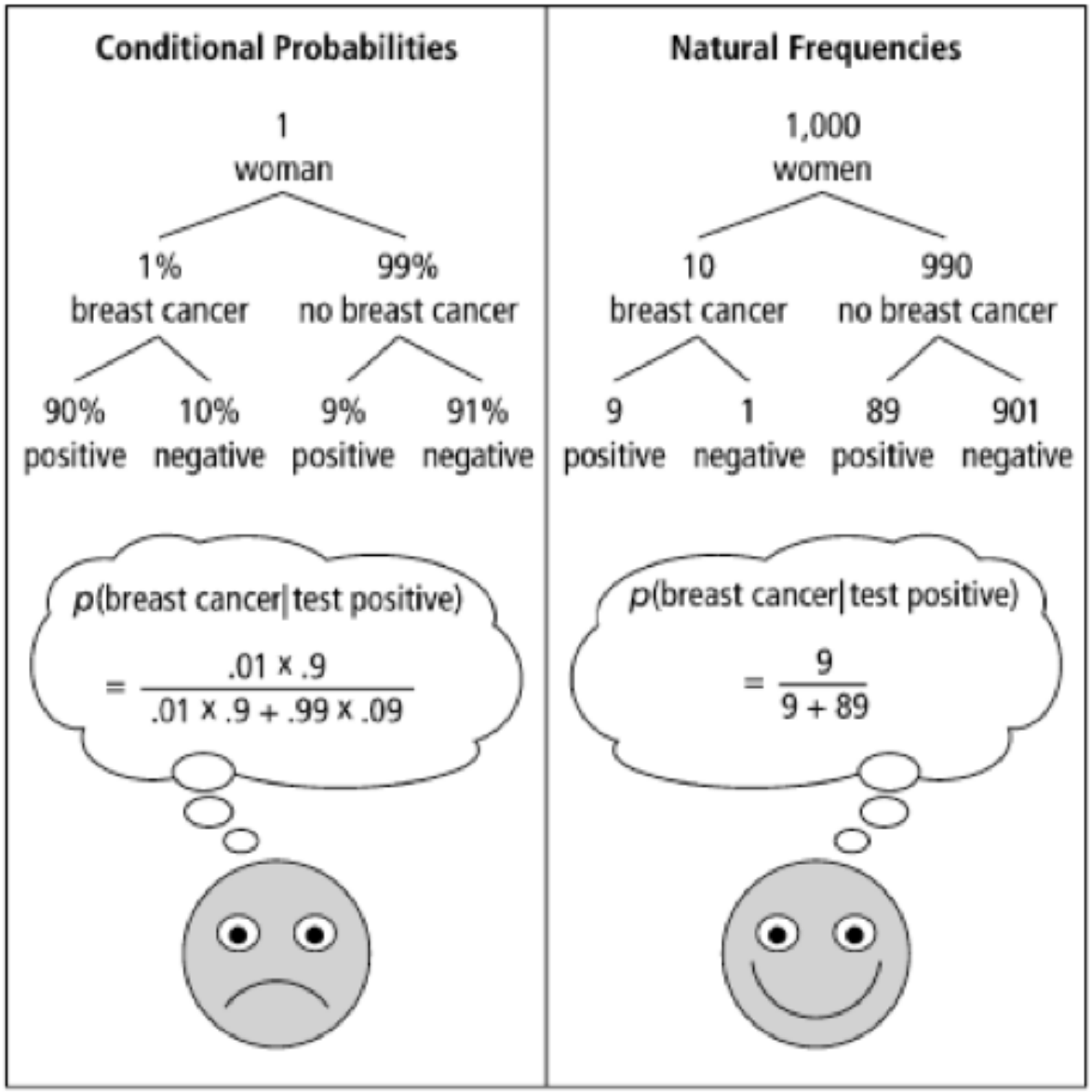
- Prepare students for technical computing.
- Mathematics should be about multiple variables.
- Make modeling central.

Statistics as the Science of Data

Insofar as we believe this, we should be helping students gain:

- The ability to access data.
- The ability to organize data.

Basic Concepts for Decision Making



Change at Macalester

- Redirecting Calculus to support statistics.
- Redirecting Statistics to support science.
- Redirecting Computation to support data.
- Teaching Quantitative Literacy in context: Epidemiology

Collaborative and Community Change

- Colleagues at Macalester
- The R Project
- Randy Pruim & Nick Horton and Project MOSAIC
- JJ Allaire and RStudio



The Golden Age of Statistics

Now is the golden age of statistics and scientific education.

- The demand is there.
- The technology is there.
- What's missing is the human capital.

We need a new organization of educational work

- More collaboration
- More ongoing training

Statistics with Confidence

Statistics is the quantitative/scientific/technical area that most strongly unites areas of research.

- This increases our responsibility to look after education generally.