Data 102

Moritz Hardt
UC Berkeley, Spring 2020

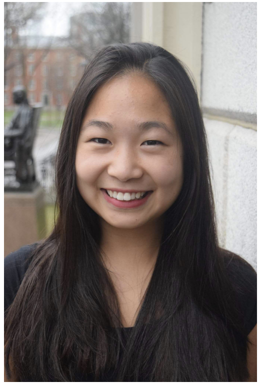
DS 102 team this semester

Prof. Jacob Steinhardt

Akosua Busia

Ashley Chien

Wenshuo Guo



Serena Wang

Clara Wong-Fannjiang

Building on tons of work by Michael Jordan, Fernando Perez and the whole Fall 2019 team.

Announcements

All class discussion on Piazza. Please be respectful and reasonable.

TAs do not answer questions by email. Available via Piazza/labs/OH

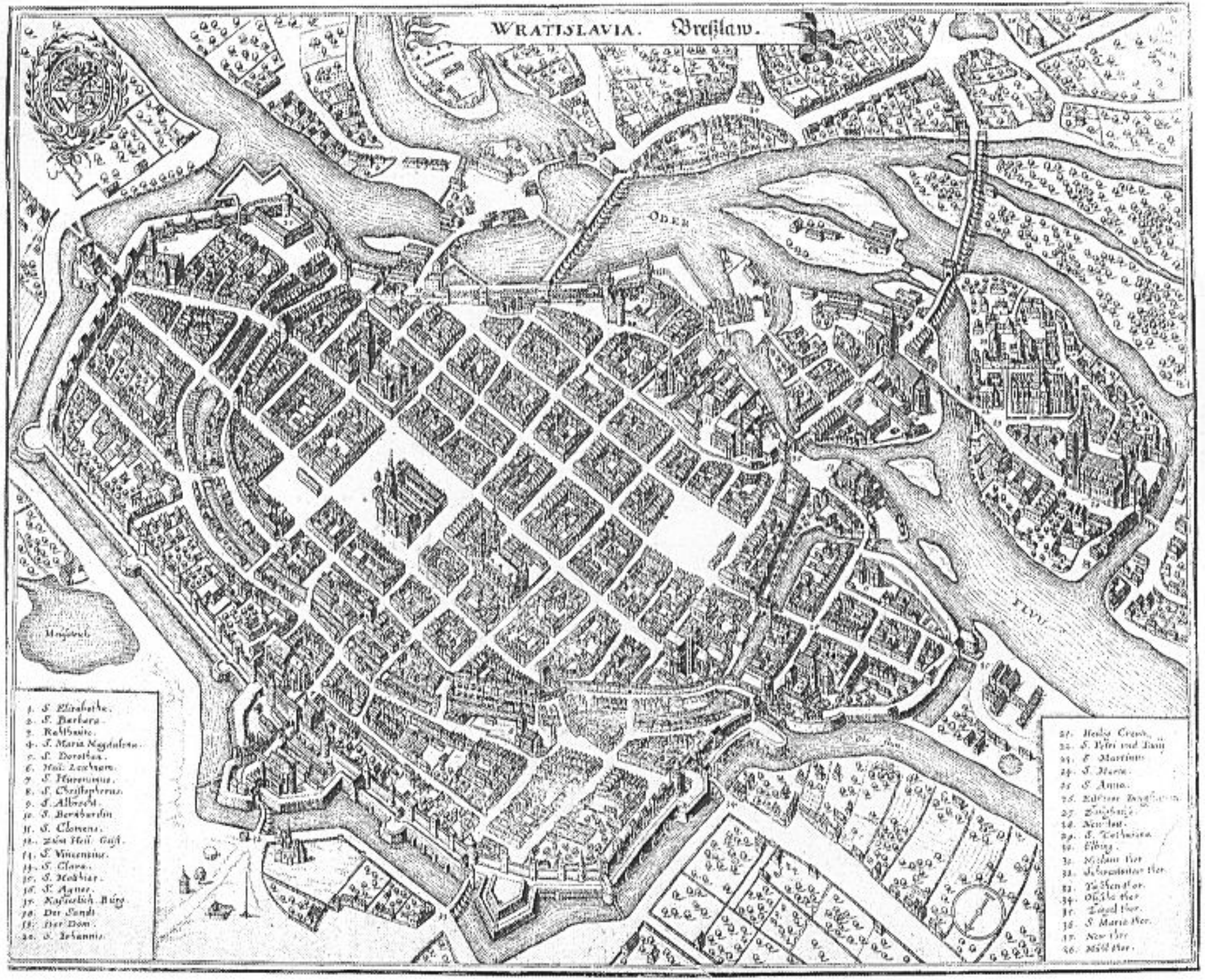
Enrollment cap of 160 is firm. Instructors cannot change anything about that.

Don't email instructors about class absence. Attendance is strongly encouraged but not mandatory.

Email Laura Imai (lauraimai@berkeley.edu) for enrollment related questions.

Hardt OH: Wed 4p-5p in 525 Soda Hall

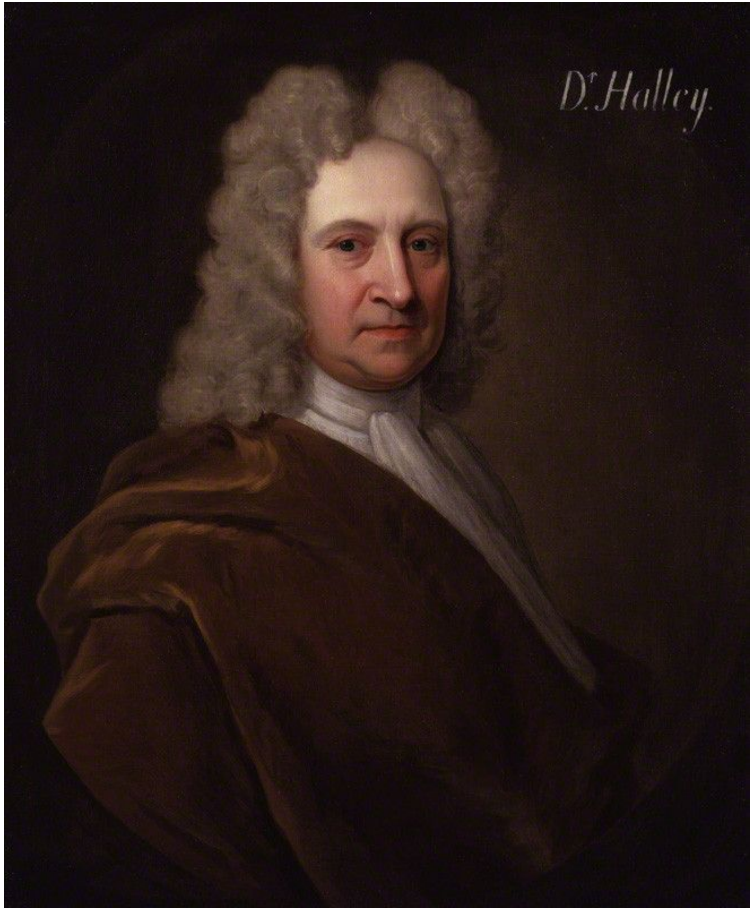
What’s data science?



Breslau
(now Wrocław)
ca 1660

Age. Curt.	Perſons.	Age. Curt.	Perſons	Age. Curt.	Perſons	Age. Curt.	Perſons	Age. Curt.	Perſons	Age. Curt.	Petſons
1	1000	8	680	15	628	22	586	29	539	36	481
2	855	9	670	16	622	23	579	30	531	37	472
3	798	10	661	17	616	24	573	31	523	38	463
4	760	11	653	18	610	25	567	32	515	39	454
5	732	12	646	19	604	26	560	33	507	40	445
6	710	13	640	20	598	27	553	34	499	41	436
7	692	14	634	21	592	28	546	35	490	42	427
Age. Curt	Perſons.	Age. Curt.	Perſons	Age. Curt.	Perſons	Age Curt	Perſons	Age. Curt.	Perſons	Age. Curt.	Perſons
43	417	50	346	57	272	64	202	71	131	78	58
44	407	51	335	58	262	65	192	72	120	79	49
45	397	52	324	59	252	66	182	73	109	80	41
46	387	53	313	60	242	67	172	74	98	81	34
47	377	54	302	61	232	68	162	75	88	82	28
48	367	55	292	62	222	69	152	76	78	83	23
49	357	56	282	63	212	70	142	77	68	84	20

Age.	Perſons.
7	5547
14	4584
21	4270
28	3964
35	3604
42	3178
49	2709
56	2194
63	1694
70	1204
77	692
84	253
100	107
	34000
	Sum Total.

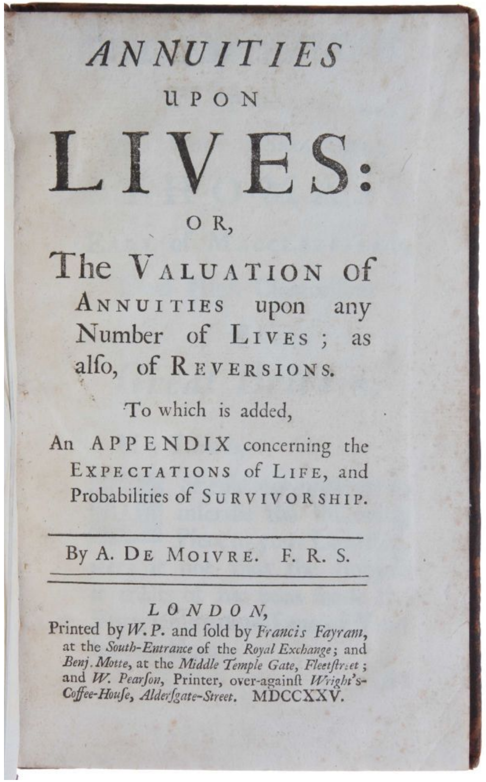


Edmond Halley 1656 - 1742

Halley's **life table** from 1693

based on data collected between 1687-1691

From data to decisions



ANNUITIES
UPON
LIVES:
OR,
The VALUATION of
ANNUITIES upon any
Number of LIVES; as
alſo, of REVERSIONS.
To which is added,
An APPENDIX concerning the
EXPECTATIONS of LIFE, and
Probabilities of SURVIVORSHIP.

By A. DE MOIVRE. F. R. S.

LONDON,
Printed by *W. P.* and ſold by *Francis Fayram*,
at the *South-Entrance* of the *Royal Exchange*; and
Benj. Motte, at the *Middle Temple Gate, Fleetſtreet*;
and *W. Pearſon*, Printer, over-againſt *Wright*'s-
Coffee-Houſe, Alderſgate-Street. MDCCXXV.

Halley's life table was then used to price **life annuities**

Price of annuity at age x is the *expected* sum of discounted fixed annual payment for the rest of person's life.

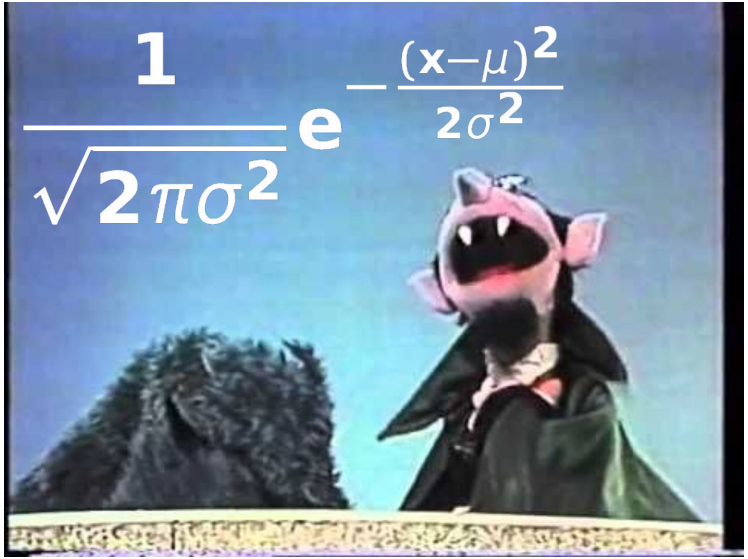
$$\text{Price at age } x = \sum_i \underbrace{p[\text{death at age } x+i]}_{\text{Halley's life expectancy model}}\ 0.95^i\ (\text{annual payout})$$

Halley built the lookup table just by counting data

We now call these lookup tables *models*

and they’ve gotten bigger

Large tables with many columns require clever statistical interpolation and smoothing



333 years of consequential decisions from data

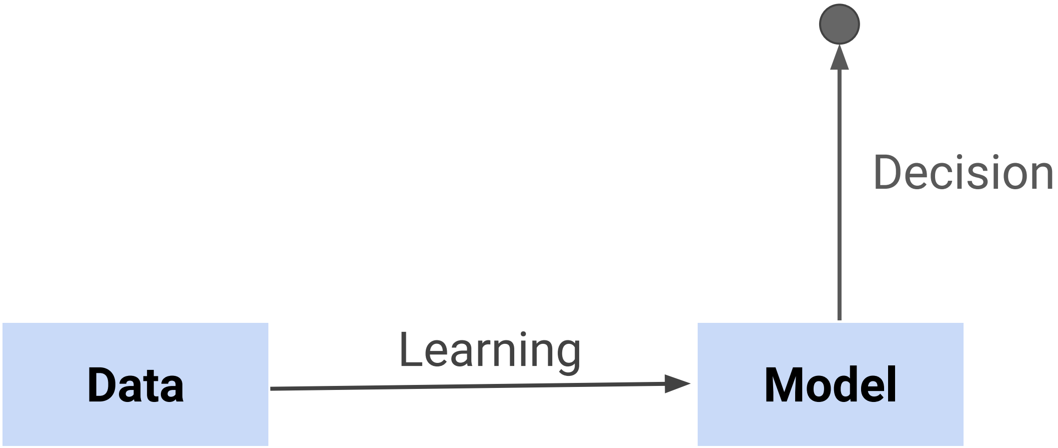
Halley built a **statistical model for decision making**

An approach used for centuries with varying degrees of rigor

20th century statistics formalized and vastly extended the approach

Current ML/AI wave pushes it into ever-increasing range of domains: *Health, finance, insurance, employment, education, criminal justice, policing*

The standard view of learning and decision making



First part of the class operates in this simple world view.

Context and consequences of decisions

“[T]echnologies are developed and used within a particular social, economic, and political context. They arise out of a social structure, they are grafted on to it, and they may reinforce it or destroy it, often in ways that are neither foreseen nor foreseeable.”

Ursula Franklin, 1989

“[C]ontext is not a passive medium but a dynamic counterpart. The responses of people, individually, and collectively, and the responses of nature are often underrated in the formulation of plans and predictions.”

Ursula Franklin, 1989

Early example of dynamics in decision making

In 1696, England's King William III seeks to tax wealth, but how to know one’s wealth?

Introduces tax based on **number of windows**

Idea spreads to France, Spain, Scotland

People adapt

One row of houses in Edinburgh featured no bedroom windows at all.

Tax revenues fell

Goodhart's law

"Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." - Charles Goodhart, 1975

Related:
Lucas critique 1976 in macroeconomics
Campbell's law 1979 in social sciences

Learning invites gaming

- Correlation is all you need for prediction
- Typically lots of features
- Features often easy to change
- Most learning problems aren't *causal* [Schölkopf et al. 2012]

Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment

Daniel Björkegren[1] and Darrell Grissen[2]

Features
Number of outgoing calls
Text response rate
Average airtime balance
Entropy of GPS coordinates

How can we identify *cause*?

How do we make decisions in changing environments?

What behavior do our decisions *incentivize*?

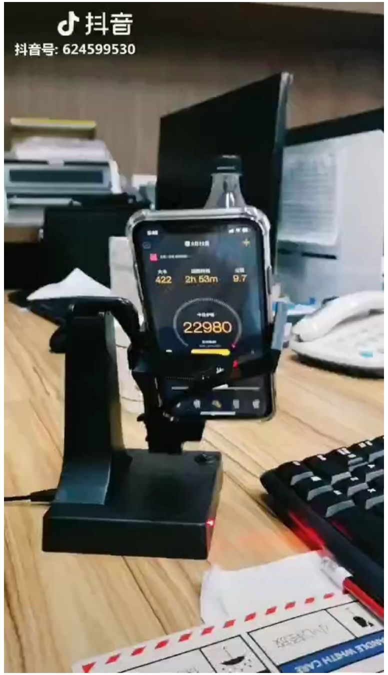
Get moving.

Start a healthy habit in the new year
get moving. Want to know the fastes
more. Studies have shown that low-i
increases fat burn. It also produces e
elevators that make you feel good! V

speed-walk with your friends or are training for the NYC Marathon, we recognize that a little healthy competition is good! Download the Oscar app to your phone and sync Apple Health or Google Fit to track your steps and earn $1 a day in Amazon® Gift Card rewards when you meet your step goals.

https://www.hioscar.com/faq/five-ways-to-have-your-healthiest-year-yet

But there are two ways of going about it



Are our decisions *fair*?

Machine Bias

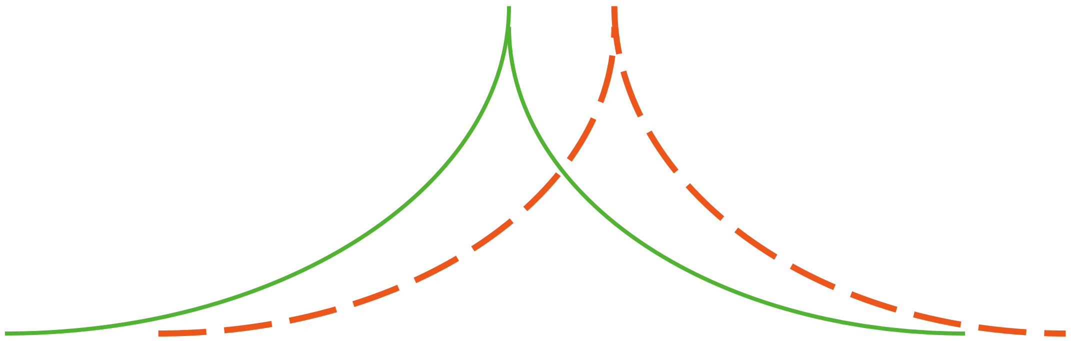
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

How do we respect individual privacy?

We’ll see a powerful tool called Differential Privacy



Decisions in the real world

Decisions feed into a social system of individuals, institutions, and markets

This changes how we ought to think about decision making in the first place

Decisions are consequential

Success of decision-making in the real world depends on context

Real-world decision-making is a *dynamic* problem

Looking ahead

A typical AI/machine learning class today will focus on *pattern recognition*

Data 102 focuses on decisions

Algorithmic decisions already are and will increasingly be deeply embedded in all kinds of sociotechnical systems.

You’ll learn some of the tools to maneuver this reality.

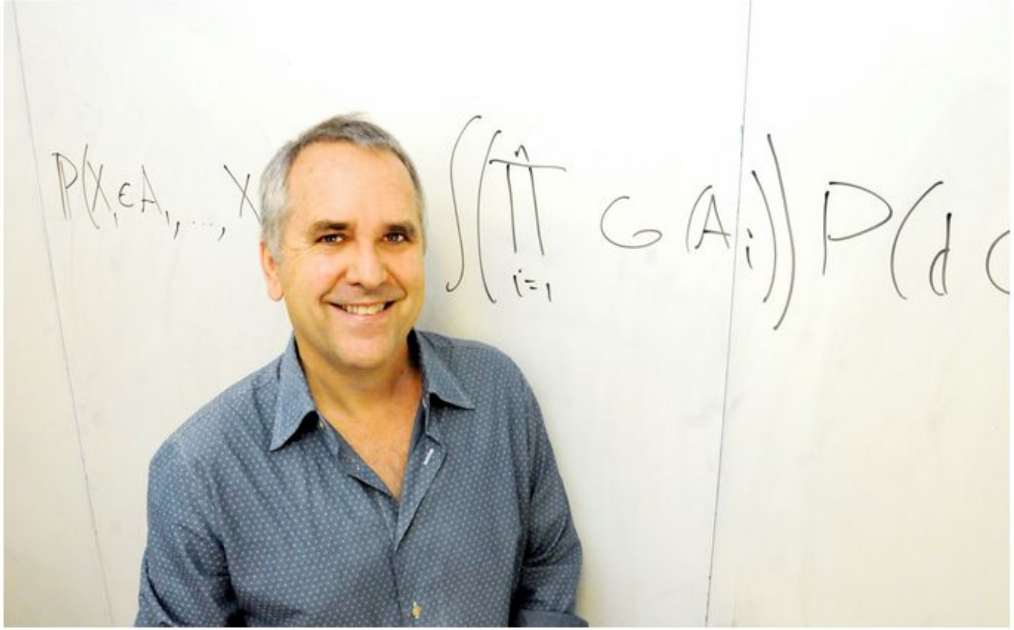


Photo credit: Peg Skorpinski

Artificial Intelligence — The Revolution Hasn't Happened Yet

Michael Jordan Follow

Apr 18, 2018 · 16 min read

Go ahead and read this article.

Back to the basics: Decision theory 101

The simplest setup

Reality is in one of two states 0, 1

Decision is also 0, 1

Decision x is the right one if reality is in state x

Classification: Cat vs Dog

Prediction: Rainfall vs sunshine

Hypothesis testing: Null vs Non-Null

Detection: Signal vs Noise

The basic two by two table

		Decision	
		0	1
Reality	0	TN	FP
	1	FN	TP

TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

Sometimes called “confusion matrix”, because it causes confusion

And then there's this…

False positive = Type 1 error

False negative = Type 2 error

Confusing them = Type 3 error

Being friends with people who use them = Type 4 error

I won't be using these names, since I already forgot which one is which

The basic two by two table

		Decision	
		0	1
Reality	0	TN	FP
	1	FN	TP

TN = True Negative

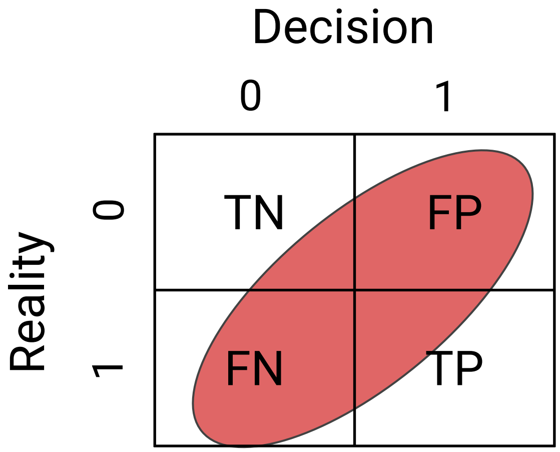
FP = False Positive

FN = False Negative

FP = True Positive

Think of these as good: Low cost or reward

The basic two by two table



TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

Think of these as bad: High cost or penalty

Examples

- Medical: 0 = no disease, 1 = disease
- Commerce: 0 = no fraud, 1 = fraud
- Physics: 0 = no Higgs boson, 1 = Higgs boson
- Social network: 0 = no link, 1 = link
- Self-driving car: 0 = no pedestrian, 1 = pedestrian
- Search: 0 = not relevant, 1 = relevant

Lots of complications arise in real settings

Towards a statistical framework

- Although the two-by-two table is useful conceptually, it’s not clear how to make use of it in a real problem, because we don’t know Reality
- We need to move towards a statistical framework, where we consider not just one decision, but a set of related decisions

Towards a statistical framework

- Imagine we not only make one decision, but we build a *decision-making algorithm*
- We want to evaluate the algorithm not just on one decision, but on a set of related decisions
- Concretely, we may have a collection of cases, where we repeatedly make a 0/1 decision
- Example: binary classification, hypothesis testing

Counting (reality, decision) pairs

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$N = n_{00} + n_{01} + n_{10} + n_{11}$

E.g. n_{11} = number of true positives

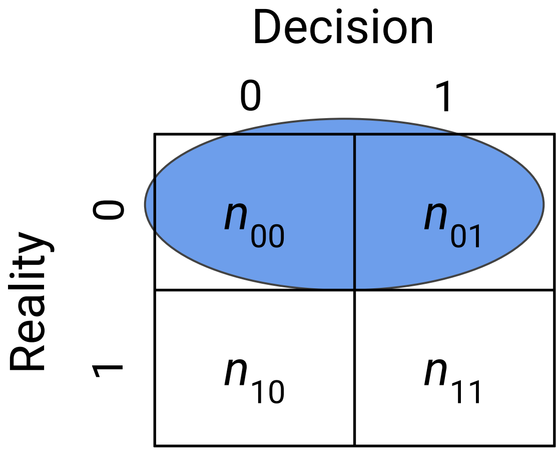
Counting (reality, decision) pairs *row-wise*

		Decision 0	Decision 1
Reality	0	n_{00}	n_{01}
Reality	1	n_{10}	n_{11}

true positives rate $\dfrac{n_{11}}{n_{10}+n_{11}}$

Sensitivity, power, recall

Counting (reality, decision) pairs *row-wise*



true negative rate $\dfrac{n_{00}}{n_{00} + n_{01}}$

specificity, selectivity

Probability view

Imagine you're cases are drawn from a distribution

True positive rate: Pr(Decision = 1 | Reality = 1)

True negative rate: Pr(Decision = 0 | Reality = 0)

The count table can be computed from a finite sample

How well we can estimate the distribution quantities from a finite sample depends on *prevalence* of positive and negative cases.

Probability view

Imagine you're cases are drawn from a distribution

True positive rate: Pr(Decision = 1 | Reality = 1)

True negative rate: Pr(Decision = 0 | Reality = 0)

False negative rate: Pr(Decision = 0 | Reality = 1) = 1 - Pr(Decision = 1 | Reality = 1)

False positive rate: Pr(Decision = 1 | Reality = 0) = 1 - Pr(Decision = 0 | Reality = 0)

What we want

Ideally, we want high true positive rate and high true negative rate.

But there’s a trade-off.

Example: Pearson's 1894 problem

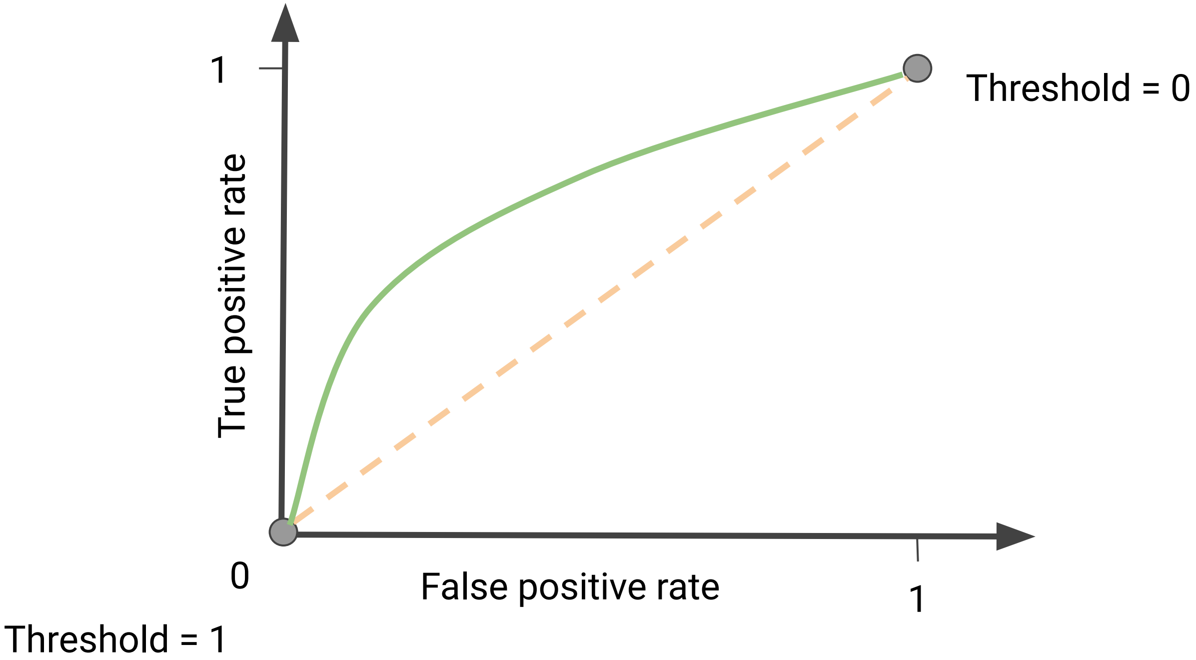
Decide if crab is male (0) or female (1)

Observe ratio R of forehead breadth to body length

Decision = 1 if $R >$ threshold and 0 if $R \leq$ threshold

Each setting of threshold gives us a different decision rule

The trade-off curve (also called ROC curve)



Neyman-Pearson formulation (1932)

Constrained optimization:

Maximize true positive rate

s.t. false positive rate ≤ some fixed number (e.g. 0.05)

Fruitful idea, sometimes the right thing to do, but not “written in stone”

Counting cases *column-wise*

Decision

Reality		0	1
	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

Pr(Reality = 1 | Decision = 0)

false omission rate $\dfrac{n_{10}}{n_{00} + n_{10}}$

Counting cases *column-wise*

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

Pr(Reality = 0 | Decision = 1)

false discovery rate $\dfrac{n_{01}}{n_{01} + n_{11}}$

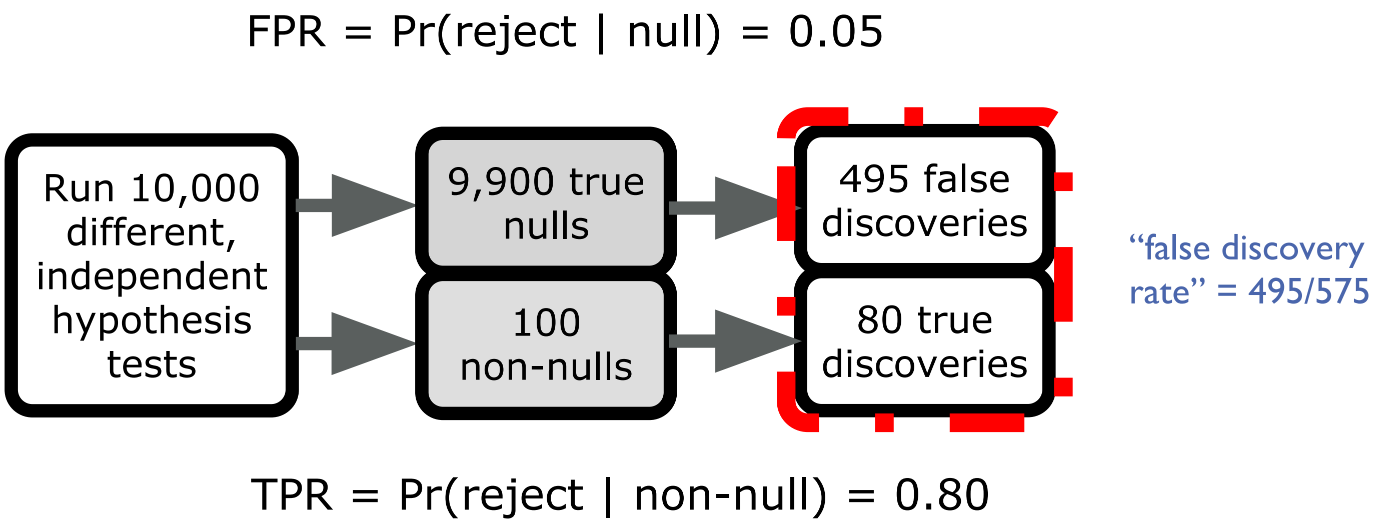
Hypothesis tests as decision making

Hypothesis H

Reality: Null hypothesis is true (0), null hypothesis is false (1)

Decision: Accept null hypothesis (0), Reject null hypothesis (1)

False discovery rate in hypothesis testing



Note: We're again not being rigorous at this point; FDR is actually an expectation of this proportion. We'll do it right later.