Lecture 8: Measure

Last time

We began using probability models as an “origin story” for how our data were generated -- We started very simply, assuming our data were n independent, identically distributed (IID) observations $X_1, \ldots, X_n$ from some probability function $f(x|\theta^*)$

This “true” function was assumed to be a member of a larger parametric family $f(x|\theta)$ indexed by a “parameter” θ -- We looked to our data $X_1, \ldots, X_n$ to form an “estimate” $\widehat{\theta}$ of θ^*

We saw two paradigms in the last lecture, Maximum Likelihood and the Method of Moments...

Parametric families

We saw the normal (location-scale family, parameters μ and σ)

$$\mathcal{F} = \left\{ \mathsf{f}(\mathsf{x}|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathsf{e}^{(\mathsf{x}-\mu)^2/2\sigma^2} : \mu \in \mathsf{R}, \sigma > 0 \right\}$$

and the binomial (parameter p)

$$\mathcal{F} = \left\{ \mathsf{f}(\mathsf{k}|\mathsf{p}) = \binom{\mathsf{n}}{\mathsf{k}} \mathsf{p}^{\mathsf{k}} (1-\mathsf{p})^{\mathsf{n}-\mathsf{k}} : \mathsf{p} \in [0,1] \right\}$$

Maximum likelihood

As its name suggests, we find the value of θ that maximizes the likelihood associated with our data $X_1, \ldots, X_n$

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

or, equivalently, the **the associated log-likelihood function**

$$l(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(x|\theta)$$

The score function and Fisher information

Given a likelihood function $\mathcal{L}(\theta)$ and a log-likelihood $\mathsf{l}(\theta) = \log \mathcal{L}(\theta)$, the score function is just the derivative

$$\mathsf{S}(\theta) = \mathsf{l}'(\theta)$$

Using this definition, the MLE is just the solution to $\mathsf{S}(\theta) = 0$

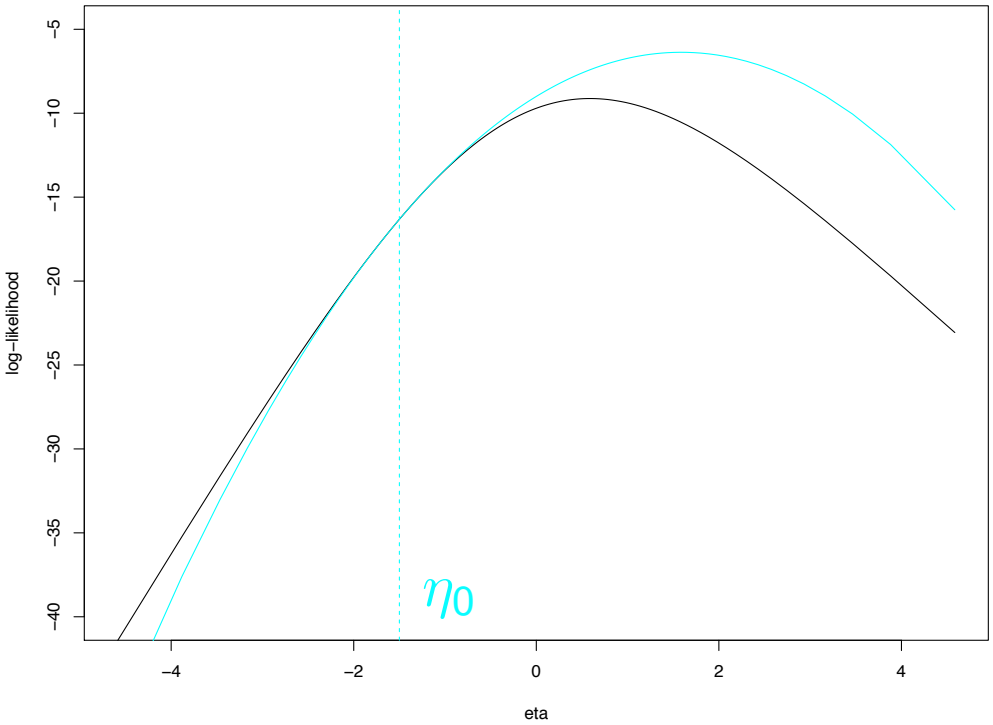
The (observed) Fisher information is defined to be the second derivative of the log-likelihood $\mathrm{I}(\theta) = -\mathsf{l}''(\theta)$

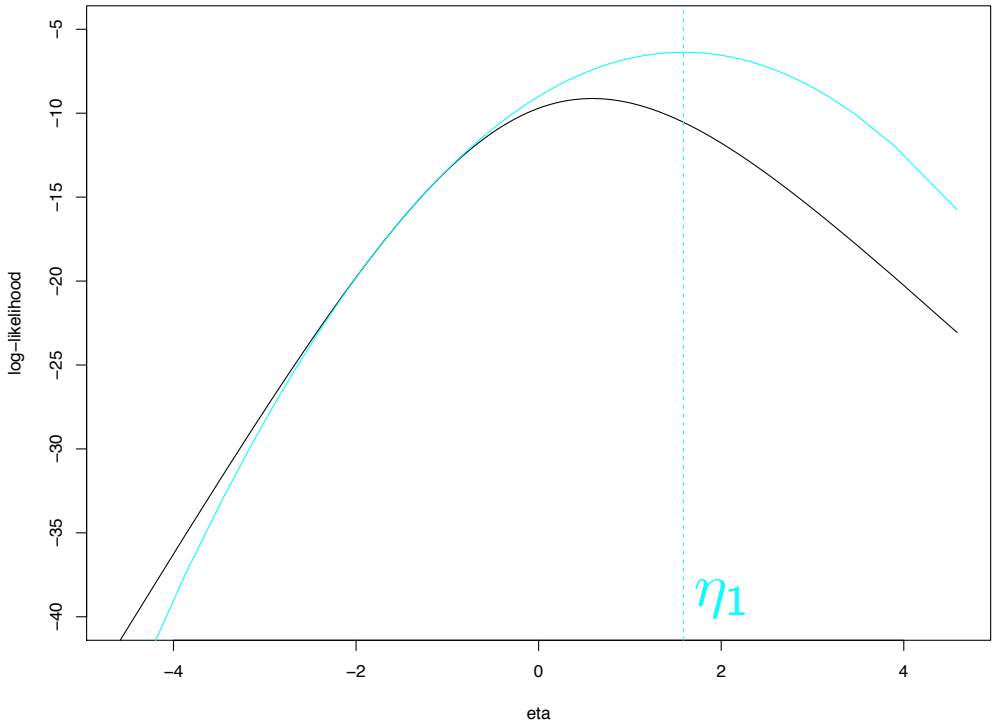
Numerical optimization

When we cannot find a closed-form expression for the MLE, we can appeal to simple Newton-Raphson iterations -- Last time, we derived the simple update rule

$$\theta_1 = \theta_0 + \frac{\mathrm{S}(\theta)}{\mathrm{I}(\theta)}$$

Here we start with an initial guess, form a quadratic approximation to the log-likelihood, find its maximum and update our guess


log−likelihood
−5
−10
−15
−20
−25
−30
−35
−40
−4
−2
0
2
4
eta
η_0


log−likelihood
−5
−10
−15
−20
−25
−30
−35
−40
−4
−2
0
2
4
eta
η_1

Exponential families

For a single parameter, members of this family take the form

$$f(x|\theta) = h(x)\, e^{\eta(\theta)T(x)-B(\theta)}$$

A number of the probability distributions in your textbook can be rewritten in this way -- Here are a few more

$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$	Normal
$\frac{e^{-\lambda}\lambda^x}{x!}$	Poisson
$\binom{m}{x} p^x (1-p)^{m-x}$	Binomial (with known m, the number of trials)
$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$	Gamma
$\binom{x+r-1}{x} (1-p)^r p^x$	Negative binomial (where r, the number of failures until a success, is known)
$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	Beta

Sufficiency

If our model comes from an exponential family, its log-likelihood can be written as follows (where C doesn't involve θ)

$$l(\theta) = \eta(\theta)T - nB(\theta) + C$$

where we let $T = \sum_{i=1}^{n} T(X_i)$

Given an exponential family, we call T the **natural sufficient statistic** for θ -- From the point of view of model fitting, we see that **maximizing the log-likelihood will involve the data only through T**

Method of moments

Suppose we have a two-parameter family, where $\theta = (\theta_1, \theta_2)$, and assume we can write these parameters in terms of the first two moments

$$\theta_1 = \mathrm{g}_1(\mu_1, \mu_2) \quad \text{and} \quad \theta_2 = \mathrm{g}_2(\mu_1, \mu_2)$$

where we define our theoretical moments $\mu_k = \mathrm{E}\, X^k$

Method of moments

Then, given data $X_1, \ldots, X_n$ we form the "empirical" the moments

$$\widehat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

and derive our estimates

$$\widehat{\theta}_1 = g_1(\widehat{\mu}_1, \widehat{\mu}_2) \quad \text{and} \quad \widehat{\theta}_2 = g_2(\widehat{\mu}_1, \widehat{\mu}_2)$$

Method of moments

We applied this technique for the normal family and found that our Method of Moments estimates matched our MLEs

We then considered the class of so-called normal mixtures, a flexible model that lets us describe distributions like the one on the right

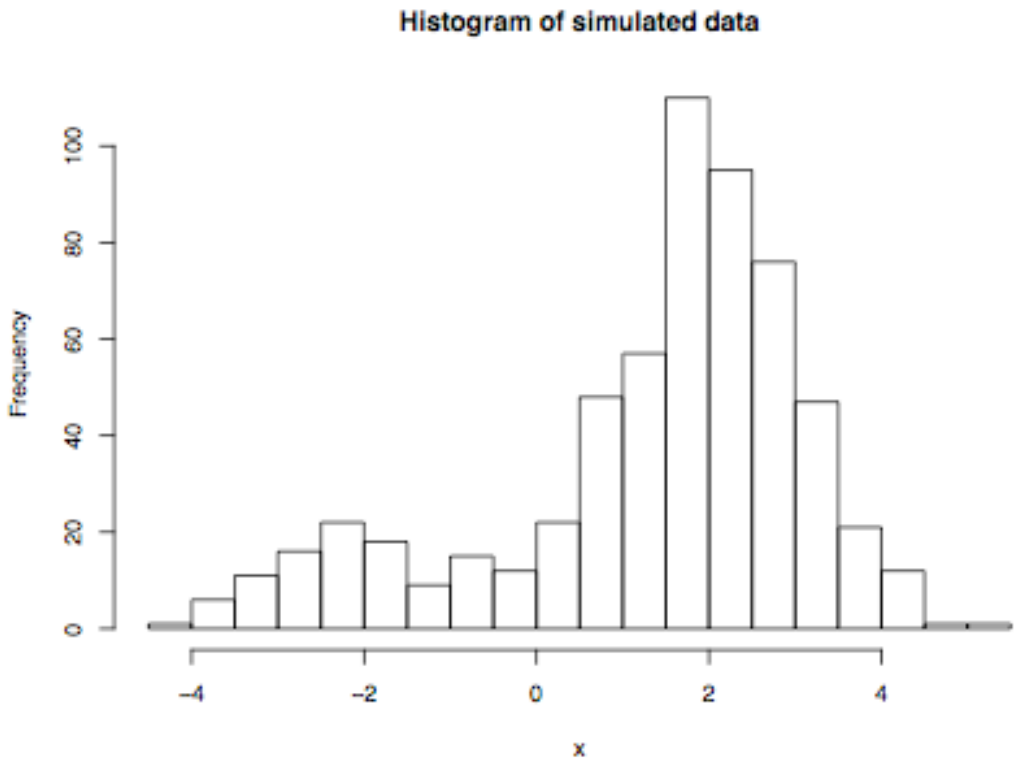
To generate these data we first flip a coin (pr α) -- Heads and we sample from

$$f_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{(x-\mu_1)^2/2\sigma_1^2}$$

tails and we sample from

$$f_2(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{(x-\mu_2)^2/2\sigma_2^2}$$

This simple scheme can generate data like that at the right -- Given $X_1, \ldots, X_n$, we want to estimate $\mu_1, \sigma_1, \mu_2, \sigma_2, \alpha$



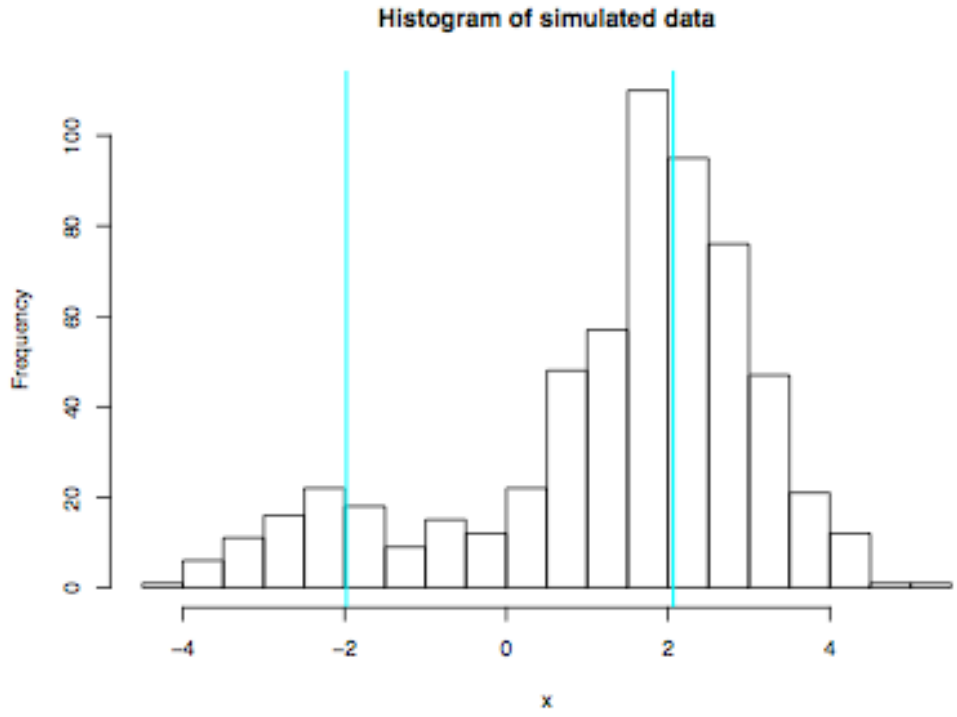
Method of moments

Under the assumption that the two components have the same standard deviation, we have four parameters to estimate

We then came up with four equations that related the first four moments to the four unknown parameters -- For example, the mean of our normal mixture is

$$\alpha\mu_1 + (1-\alpha)\mu_2$$

Four equations, four unknowns and a little tedious algebra and we have the solution at the right




```
# compute moments
xb = mean(x)
v2 = mean((x-xb)^2)
v3 = mean((x-xb)^3)
v4 = mean((x-xb)^4)
k4 = v4-3*v2^2

# form estimates
roots = polyroot(c(v3^2,k4,0,2))
v = Re(roots[2])  # here, 2nd root is negative
r = -v3/v
sig2 = v+v2
m1 = 0.5*(r-sqrt(r^2-4*v))
m2 = 0.5*(r+sqrt(r^2-4*v))
mu1 = m1+xb
mu2 = m2+xb
alpha = m2/(m2-m1)

# look at fit (!)
print(c(mu1,mu2,sig2,alpha))
hist(x)
abline(v=c(mu1,mu2),col=5,lwd=2)
```

Today

We are going to begin with a survey of a few views of probability -- We are just about to launch into a couple weeks of frequentist statistics and it's good for you to see some logic behind the style of reasoning

We'll then close with a discussion of the properties of estimators -- We'll run right up to the idea of confidence intervals, something we will pick up next time

By the way, we are now well into Chapter 7 of your book and will be making a lot of progress through it for the remainder of the quarter

Calculation versus interpretation

In what follows, we are going to try to keep distinct the mathematical rules for calculating with probabilities from their interpretation -- Probability is **a branch of mathematics with it's own rules** and most definitions of probability adhere to these rules

Quite aside from computation, however, we have the application of probability, **the interpretation of probability in our daily lives** -- What does it mean to say that event will occur with probability 1/3?

The mathematical framework lets us solve textbook problems (rolling dice or pulling cards from a well-shuffled deck), but the interpretation, what we mean by the term probability can be a different animal entirely

If statistics uses the language of probability to describe events in our lives, then **our interpretation of probability can influence how we reason about the world**

The emergence of probability

Ian Hacking, a historian and philosopher, believes that probability was "born" in 1660 -- That's precisely the time of **John Graunt and his work on the London bills**, work that will feed into the emergence of probability

Hacking notes that since that time, probability has had two faces: In one it is an **explanation for "stable" frequencies seen in the world**, while in another it is **a relation between a hypothesis and the evidence for it**

He notes that prior to 1660, a statement was said **to be probable if it could be attested to by an authority** and a kind of expert testimony was involved -- Over time that changed and slowly **Nature became a kind of expert**

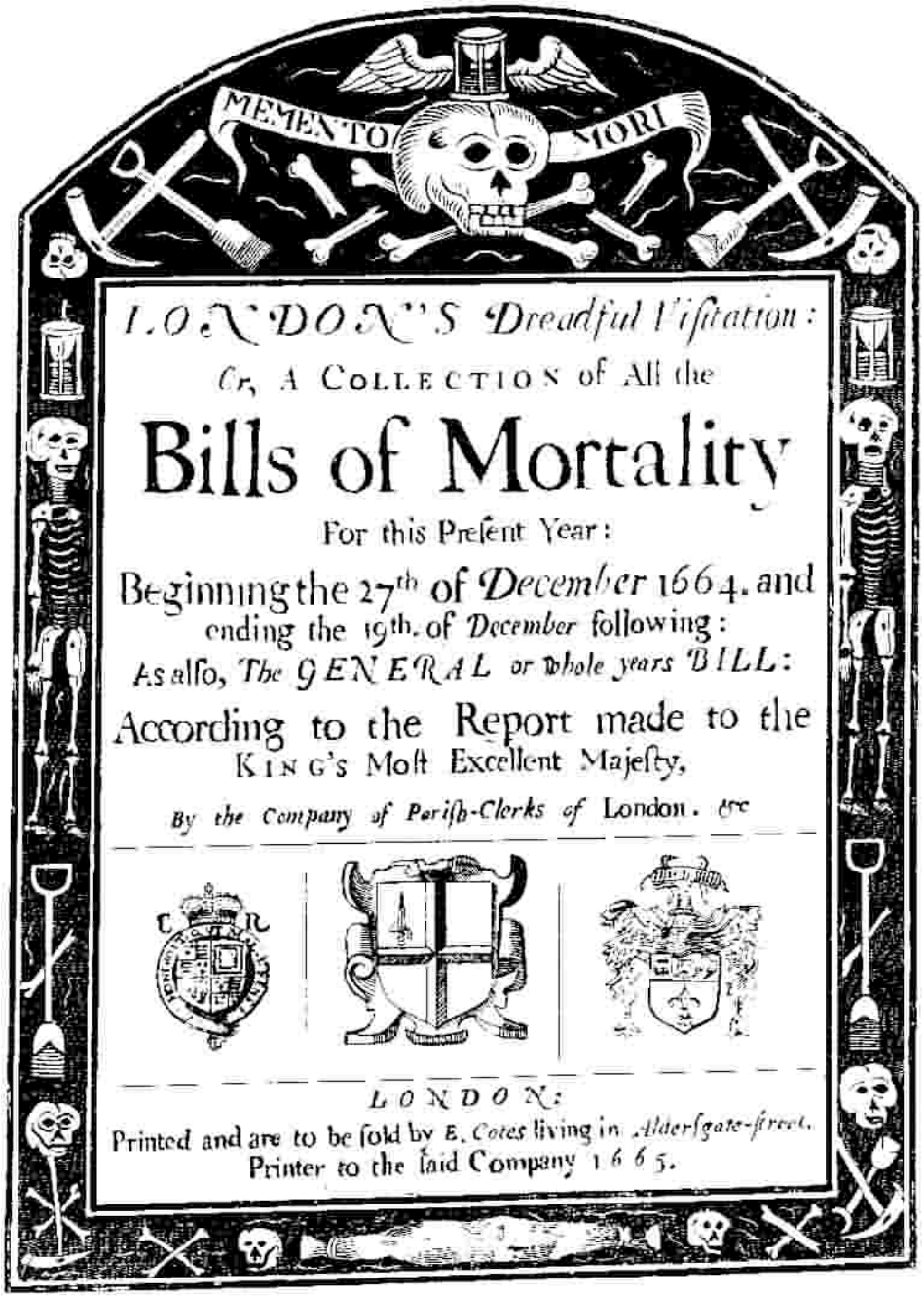
He writes *"It is here that we find* **the old notion of probability as testimony conjoined with that of frequency***. It is here that stable and law-like regularities become both observable and worthy of observation.* ***They are part of the technique of reading the true world****."*

The emergence of probability

He continues: "*A proposition was now probable, as we should say, if there was evidence for it, but in those days it was probable because it was testified to by the best authority.* ***Thus: to call something probable was still to invite the recitation of authority. But: since the authority was founded on natural signs, it was usually of a sort that was only "often to be trusted"****. Probability was communicated by what we should now call law-like regularities and frequencies.* ***Thus the connection of probability, namely testimony, with stable law-like frequencies is a result of the way in which the new concept of internal evidence came into being****."*

And so probability emerges with two faces -- **One related to hypothesis and evidence, and another related to stable frequencies seen in data about the world** (think about Graunt's birth and death records)

We'll see that these two views of probability, these two interpretations of the word, still exist today in modern statistical practice (Hacking says that **we've been swinging on a pendulum for centuries**) -- But first, let's go to the birth of probability, circa 1660..



LONDON's Dreadful Visitation:

Or, A COLLECTION of All the

Bills of Mortality

For this Present Year:

Beginning the 27th of *December* 1664. and ending the 19th. of *December* following:

As also, *The* GENERAL *or Whole years* BILL:

According to the Report made to the KING's Most Excellent Majesty,

By the Company of Parish-Clerks of London. &c

LONDON:

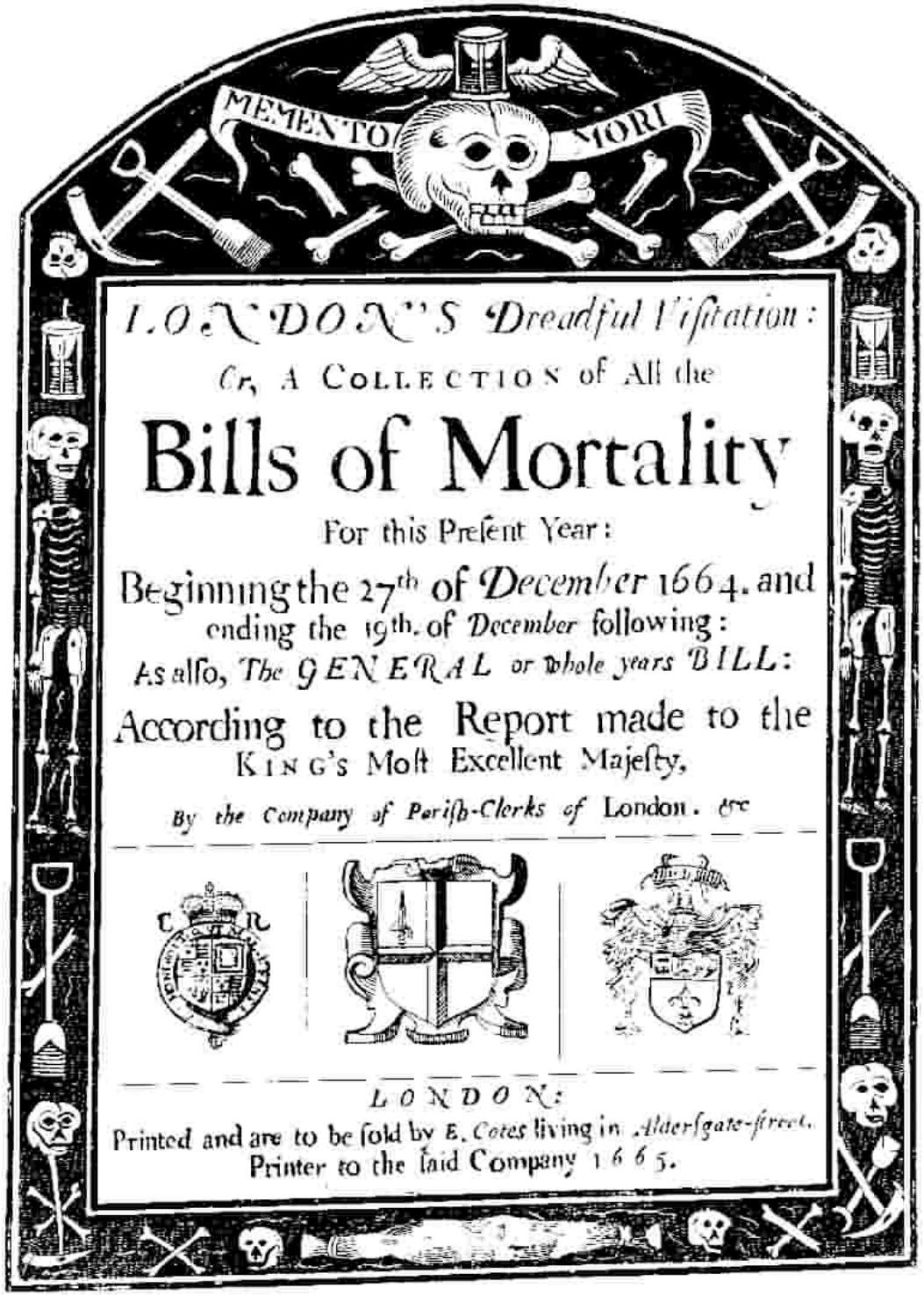
Printed and are to be sold by E. *Cotes* living in *Aldersgate-street*. Printer to the said Company 1665.

Bills of Mortality

In an effort to monitor the incidence of the plague, an injunction issued in 1538 on behalf of Henry VIII required the **registration of all burials and christenings in every English Parish**

The weekly Bills of Mortality were compiled from these registers, and were initially circulated only to government officials

The Bills were made available to the public in 1594, but were discontinued the next year when the plague abated; publication of the Bills resumed in 1603 when the plague broke out again



Taken from The World of Mathematics By James Roy Newman

Bills of mortality

In 1662, John Graunt, a successful London shopkeeper (who also had a taste for scholarship), published *Observations on the Bills of Mortality*, in which he reported systematic observations about the population in and around London

His observations were, in effect, generalizations about the patterns and regularities in births, deaths and migration in England

The Diseases, and Casualties this year being 1632.

Abortive, and Stilborn ..	445	Grief	11
Affrighted	1	Jaundies	43
Aged	628	Jawfaln	8
Ague	43	Impostume	74
Apoplex, and Meagrom	17	Kil'd by several accidents..	46
Bit with a mad dog.....	1	King's Evil.....	38
Bleeding	3	Lethargie	2
Bloody flux, scowring, and flux	348	Livergrown	87
Brused, Issues, sores, and ulcers, ...	28	Lunatique	5
Burnt, and Scalded.....	5	Made away themselves.....	15
Burst, and Rupture.....	9	Measles	80
Cancer, and Wolf.....	10	Murthered	7
Canker	1	Over-laid, and starved at nurse	7
Childbed	171	Palsie	25
Chrisomes, and Infants.....	2268	Piles.....	1
Cold, and Cough.....	55	Plague.....	8
Colick, Stone, and Strangury	56	Planet	13
Consumption	1797	Pleurisie, and Spleen.....	36
Convulsion	241	Purples, and spotted Feaver	38
Cut of the Stone.....	5	Quinsie	7
Dead in the street, and starved	6	Rising of the Lights.....	98
Dropsie, and Swelling.....	267	Sciatica	1
Drowned	34	Scurvey, and Itch.....	9
Executed, and prest to death	18	Suddenly	62
Falling Sickness.....	7	Surfet	86
Fever	1108	Swine Pox	6
Fistula	13	Teeth	470
Flocks, and small Pox.....	531	Thrush, and Sore mouth...	40
French Pox.....	12	Tympany	13
Gangrene	5	Tissick	34
Gout	4	Vomiting	1
		Worms	27

Christened { Males....4994, Females..4590, In all....9584 } Buried { Males....4932, Females..4603, In all....9535 } Whereof, of the Plague.8

Increased in the Burials in the 122 Parishes, and at the Pest-house this year..... 993

Decreased of the Plague in the 122 Parishes, and at the Pest-house this year..... 266

The Diseases, and Casualties this year being 1632.

Abortive, and Stilborn ..	445
Affrighted	1
Aged	628
Ague	43
Apoplex, and Meagrom	17
Bit with a mad dog.....	1
Bleeding	3
Bloody flux, scowring, and flux	348
Brused, Issues, sores, and ulcers, ...	28
Burnt, and Scalded.....	5
Burst, and Rupture.....	9
Cancer, and Wolf.....	10
Canker	1
Childbed	171
Chrisomes, and Infants.....	2268
Cold, and Cough.....	55
Colick, Stone, and Strangury	56
Consumption	1797
Convulsion	241
Cut of the Stone.....	5
Dead in the street, and starved	6
Grief	11
Jaundies	43
Jawfaln	8
Impostume	74
Kil'd by several accidents..	46
King's Evil.....	38
Lethargie	2
Livergrown	87
Lunatique	5
Made away themselves.....	15
Measles	80
Murthered	7
Over-laid, and starved at nurse	7
Palsie	25
Piles.....	1
Plague.....	8
Planet	13
Pleurisie, and Spleen.....	36
Purples, and spotted Feaver	38
Quinsie	7
Rising of the Lights.....	98
Sciatica	1
Scurvey, and Itch.....	9

Bills of Mortality

His innovation was to apply the scientific method to the study of populations; in Graunt's time science was largely limited to observations and descriptions of "naturally" occurring events

It is an early example of what has been termed **"political arithmetic,"** a practice that hoped to ground "official policy... in an understanding of the land and its inhabitants"

"Implicit in the use by political arithmeticians of social numbers was the belief that the wealth and strength of the state depended strongly on the number and character of its subjects"

THE CONCLUSION

It may be now asked, to what purpose tends all this laborious buzzling, and groping? To know,

1. The number of the People?
2. How many *Males*, and *Females*?
3. How many Married, and single?
4. How many *Teeming* Women?
5. How many of every *Septenary*, or *Decad* of years in *age*?
6. How many *Fighting* Men?
7. How much *London* is, and by what steps it hath increased?
8. In what time the housing is replenished after a *Plague*?
9. What proportion die of each general and perticular *Casualties*?
10. What years are Fruitfull, and Mortal, and in what Spaces, and Intervals, they follow each other?
11. In what proportion Men neglect the Orders of the *Church*, and *Sects* have increased?
12. The disproportion of Parishes?
13. Why the Burials in *London* exceed the Christnings, when the contrary is visible in the Country?

To this I might answer in general by saying, that those, who cannot apprehend the reason of these Enquiries, are unfit to trouble themselves to ask them.

I might answer by asking; Why so many have spent their times, and estates about the Art of making Gold? which, if it were much known, would onely exalt Silver into the place, which Gold now possesseth; and if it were known but to some one Person, the same single *Adeptus* could not, nay, durst not enjoy it, but must be either a Prisoner to some Prince, and Slave to some Voluptuary, or else skulk obscurely up and down for his privacie, and concealment.

I might Answer; That there is much pleasure in deducing so many abstruse, and unexpected inferences out of these poor despised Bills of *Mortality*; and in building upon that ground, which hath lain waste these eighty years. And there is pleasure in doing something new, though never so little, without pestering the World with voluminous Transcriptions.

But, I Answer more seriously; by complaining, That whereas the Art of Governing, and the true *Politiques*, is how to preserve the Subject in *Peace*, and *Plenty*, that men study onely that part of it, which teacheth

Bills of Mortality

At the right we exhibit another short excerpt from the beginning of *Observations*; notice the detail in describing how the data were collected

Using these data, Graunt also constructed the first known "life table," a numerical device summarizing mortality in terms of the number, percent and probability of living or dying throughout a lifetime

He then proposed that each country should develop similar tables for comparison and to construct a general law of mortality "a seminal moment in the origins of epidemiology"

The first of the continued weekly *Bills* of *Mortality* extant at the Parish-Clerks *Hall*, begins the 29. of *December*, 1603, being the first year of *James* his Reign; since when, a weekly Accompt hath been kept there of *Burials* and *Christnings*. It is true, There were *Bills* before, *viz.* for the years 1592, -93, -94. but so interrupted since, that I could not depend upon the sufficiencie of them, rather relying upon those Accompts which have been kept since, in order, as to all the uses I shall make of them.

I believe, that the rise of keeping these Accompts, was taken from the *Plague*: for the said *Bills* (for ought appears) first began in the said year 1592. being a time of great *Mortality*; And after some disuse, were resumed again in the year 1603, after the great *Plague* then happening likewise.

These *Bills* were Printed and published, not onely every week on *Thursdays*, but also a general Accompt of the whole Year was given in, upon the *Thursday* before *Christmas Day*: which said general Accompts have been presented in the several manners following, *viz.* from the Year 1603, to the Year 1624, *inclusive* . . .

We have hitherto described the several steps, whereby the *Bills* of *Mortality* are come up to their present state; we come next to shew how they are made, and composed, which is in this manner, *viz.* When any one dies, then, either by tolling, or ringing of a Bell, or by bespeaking of a Grave of the *Sexton*, the same is known to the *Searchers*, corresponding with the said *Sexton*.

The *Searchers* hereupon (who are antient Matrons, sworn to their Office) repair to the place, where the dead Corps lies, and by view of the same, and by other enquiries, they examine by what *Disease*, or *Casualty* the Corps died. Hereupon they make their Report to the *Parish-Clerk*, and he, every *Tuesday* night, carries in an Accompt of all the *Burials*, and *Christnings*, hapning that Week, to the *Clerk* of the *Hall*. On *Wednesday* the general Accompt is made up, and Printed, and on *Thursdays* published, and dispersed to the several Families, who will pay four shillings *per Annum* for them. . . .

Taken from The World of Mathematics By James Roy Newman

THE TABLE OF CASUALTIES.

The Years of our Lord	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1629	1630	1631	1632	1633	1634	1635	1636	1629 1630 1631 1632	1633 1634 1635 1636	1647 1648 1649 1650	1651 1652 1653 1654	1655 1656 1657 1658	1659	In 20 Years.
Abortive, and Bilborn	335	329	327	351	389	381	384	433	483	419	463	467	421	544	499	439	410	445	500	475	307	513	1793	2005	1342	1587	1831	1247	8559
Aged	915	815	889	696	780	834	864	974	743	891	869	1176	909	1095	579	711	661	671	704	613	794	714	2475	2814	3336	3452	3680	2377	15717
Ague, and Fever	1260	884	751	970	1038	1212	1282	1171	689	875	999	1800	2303	2148	956	1091	1115	1108	953	1279	1622	2360	4418	6215	5865	4905	4363	4010	21704
Apoplex, and Sidderly	68	74	64	74	106	111	118	86	92	102	113	138	91	67	22	36		17	24	35	26		75	85	280	422	445	177	1306
Bleach			1	3	7	2				1															4	9	1	1	15
Bluffed	4	1			6	6			4		5	5	3	8	13	8	10	13	6	4		4	54	14	5	12	14	16	99
Bleeding	3	2	5	1	3	4	3	2	7	3	5	4	7	2	5	2	5	4	4	3			26	7	11	12	19	17	65
Bloudy Flux, Scouring, and Flux	155	176	802	289	833	761	200	386	168	368	362	233	346	251	449	438	351	348	278	511	346	330	1587	1666	1412	2138	1161	1597	7818
Burnt, and Scalded	3	6	10	5	11	8	5	7	10	5	7	4	6	6	3	10	7	5	1	3	12	5	25	19	24	31	26	19	125
Calenture	1			1		2	1	1			3										1	3		4	2	4	3		13
Cancer, Gangrene, and Fiftula	26	29	31	19	31	33	36	17	73	31	24	35	63	52	20	14	23	28	27	30	24	30	85	112	105	157	150	114	609
Wolf				8																				8					8
Canker, Sore-mouth, and Thrush	66	28	54	42	68	51	53	72	44	81	19	27	71	68	6	4	4	1			5	74	35	79	190	244	161	133	689
Childbed	161	206	214	117	206	243	158	192	177	201	236	225	226	194	150	257	112	171	132	143	163	330	590	668	498	769	839	490	3364
Chrifomes, and Infants	1369	1254	1065	990	1237	1280	1050	1343	1089	1393	1161	1344	858	1113	2596	2378	2035	2268	2130	2315	2113	1895	9277	8451	4678	4900	4788	4519	32106
Colick, and Wind	103	71	85	82	76	102	80	101	85	120	113	179	116	167	48	57					37	50	205	87	341	359	497	247	1389
Cold, and Cough							41	36	21	58	30	31	33	24	10	58	51	55	45	54	50	57	174	207	00	77	140	41	598
Confumption, and Cough	2423	2200	2388	1988	2350	2410	2286	2868	2606	3184	2757	3610	2982	3414	1827	1910	1713	1797	1754	1955	2080	2477	5257	8266	8999	9916	10257	7197	44487
Convulion	684	491	530	495	569	653	606	828	702	1027	807	841	742	1031	52	87	18	141	221	386	418	709	498	1714	1198	2656	3577	1324	9073
Cramp			1														1	0	0	0	0	0	01	00	01	0	0	1	2
Cut of the Stone		2	1	3		1	1	2	4	1	1	5	46	48				5	1	5	1	2	5	10	6	4	13	47	38
Dropfy, and Tympany	185	434	421	508	444	556	617	704	660	706	631	931	646	872	235	251	279	280	266	250	329	389	048	1714	1518	2322	2982	1301	9621
Drowned	47	40	30	27	49	50	53	30	43	45	63	60	57	48	43	35	29	34	37	32	32	45	139	147	144	182	215	130	827
Exceflive drinking			2																									1	2
Executed	8	17	29	43	24	12	19	21	19	22	20	18	7	18	19	13	12	18	13	13	13	13	62	51	97	76	79	55	384
Fainted in a Bath					1																					1			1
Falling-Sickneſs	1	1	1	1		1	4	1	4	1	1		4	5	1	10	7	7	1	5	6	8	27	11	10	8	8	9	74
Flox, and Small pox	139	400	1190	184	525	1279	139	812	1294	823	815	409	1523	334	72	40	58	531	72	1354	293	127	701	1846	1913	2755	3381	1785	10576
Found dead in the Streets	6	6	9	8	7	9	14	4	3	4	9	11	2	6	18	33	26	6	13	8	24	24	83	69	29	14	27	19	143
French-Pox	28	29	15	18	21	20	20	20	29	23	25	53	51	31	17	22	22	22	7	17	12	22	53	48	80	82	130	83	392
Frighted	4	4	1			2			1	1				9	1			1				1	3	2	3	9	3	2	21
Gout	9	3	11	9	7	7	5	6	8	7	8	13	14	2	2	5	3	4	4	5	7	8	14	24	35	25	36	28	134
Grief	13	13	16	7	17	14	13	17	10	13	10	12	13	4	18	20	22	13	14	17	5	10	71	56	48	59	45	47	279
Hanged, and made away themſelves	11	10	13	14	9	14	15	9	14	16	24	18	11	36	8	8	6	15		3	8	7	37	18	48	47	72	32	222
Head-Ach		1	11	1		2	6	6	3	3	4	5	33	26							4	2	0	6	14	14	17	46	051
Jaundice	57	35	39	49	41	43	57	71	61	41	46	77	102	76	47	39	35	43	35	45	54	63	184	197	180	212	225	188	998
Jaw-falls	1	1			3		2	2			3	1			10	16	13	8	10	10	4	11	47	35	02	3	6	10	95
Impoſthume	75	61	65	59	80	105	79	90	92	122	80	134	105	96	58	76	71	74	50	62	73	130	282	315	260	354	428	228	1639
Itch		1																	10				00	10	02				11
Killed by ſeveral Accidents	27	57	39	94	47	45	57	58	52	43	52	47	55	47	54	55	47	46	49	41	51	60	202	201	217	207	194	148	1011
King's Evil	27	26	22	19	22	20	26	16	27	24	25	28	28	34	16	25	18	18	35	20	26	69	97	150	94	94	102	68	537
Lethargy	3	4	2	4	4	4	3	10	9	4	6	2	6	4	1		2	3	3		2	2	5	7	13	21	21	9	67
Leproſy			1									1		2	2						1	1	2	2	2	1		3	06
Liver-grown, Spleen, and Rickets	53	46	36	39	63	72	67	63	52	30	38	32	8	15	94	112	99	87	82	77	98	99	391	336	213	269	191	158	1621
Lunatique	12	18	6	11	7	11	9	12	6	7	13	5	14	14	6	11	6	5	4	2	2	5	28	13	47	39	31	26	158
Meagrom	12	13		5	8	6	6	14	3	6	7	6	5	4			24						22	24	22	30	36	22	05	132
Meaſles	5	92	3	33	33	62	8	52	11	153	13	80	6	74	42	2	3	80	22	33	27	12	127	83	133	153	259	52	737
Mother	2					1	1	2	2	3		3	1	8	1							3	01	3	2	4	8	02	18
Murdered	3	2	7	5	4	3	3	3	9	6	5	7	70	20			3	7		6	5	8	10	19	17	13	27	77	86
Overlayd, and ſtarved at Nurſe	25	32	36	28	28	29	30	36	38	53	44	50	46	43	4	10	13	7	8	14	10	14	34	46	111	133	225	86	529
Palſy	27	21	19	20	23	20	29	18	22	23	20	21	17	21	17	23	17	25	14	21	25	17	82	77	87	90	87	53	413
Plague	1597	611	67	15	23	16	6	16	9	6	4	14	36	14		1317	274	8		1		10400	1599	10401	4250	62	33	305	16184
Plague in the Guts				1		110	32		17	315	446		255	402									00	00	01	161	344	153	991
Pleurify	30	26	13	20	21	19	17	23	10	9	17	16	12	10	26	24	16	36	21		45	24	111	92	89	72	52	51	415
Poyſoned		3		7															2			2	00	4	10	00	00	00	14
Purples, and ſpotted Fever	145	47	43	65	54	60	75	89	56	52	56	126	368	146	32	58	58	38	24	125	245	397	156	791	300	278	290	243	1845
Quinſy, and Sore-throat	14	11	12	17	24	20	18	9	15	13	7	10	21	14	01	8	6	7	24	04	5	22	22	55	54	71	45	34	247
Rickets	150	224	216	190	260	329	219	372	347	458	317	476	441	521						14	49	50	00	113	780	1190	1598	857	3681
Mother, riſing of the Lights	150	92	115	120	134	138	135	178	166	212	203	228	210	249	44	72	99	98	60	84	72	104	309	220	777	585	809	369	2700
Rupture	16	7	7	6	7	16	7	15	11	20	19	18	11	28	2	6	4	9	4	3	10	13	21	30	36	45	68	3	203
Scal'd head	2				1				3																2	1	2		05
Scurvy	32	20	21	23	29	43	41	44	103	71	82	82	95	12	3	7	9		9		00	25	33	34	94	132	300	115	593
Smothered, and ſtifled			1													24							24		2			2	26
Sores, Ulcers, broken and bruiſed	15	17	17	16	26	32	25	32	23	34	40	47	61	48	23		20	48	19	19	22	29	91	89	65	113	144	141	504
Shot (Limbs													7	20														07	27
Spleen	12	17					13	11		6	2	5	7	7											29	26	13	07	68
Shingles													1						1					1				1	2
Starved		4	8	7	1	2	1	1	3	1	3	6	7	14									14		19	5	13	39	51
Stitch				1																					1				1
Stone, and Strangury	43	42	29	28	30	41	44	38	49	37	71	69	22	30				58	56	58	49	33	95	114	185	144	173	147	51	863
Sciatica														2					1	1		1	6	1	4					13
Stopping of the Stomach	29	29	30	33	53	87	66	107	94	145	129	177	186	214									6		6	321	295	247	216	869
Surfet	217	237	136	223	104	177	175	211	118	161	153	158	202	182	55	153	149	96	104	114	134	171	455	311	613	675	644	401	3094
Swine-Pox	4	4	1				1	6	1	1	1	1			5	1	4	6	1		10		23	11	11	3	3	10	57
Teeth, and Worms	767	597	340	598	709	905	691	1131	803	1198	878	1036	839	1008	440	500	335	420	432	454	539	1207	1751	2632	2502	3436	3305	1819	14036
Tiſſick	61	47													8	11	14	14	13	15	17		68	61	109			8	241
Thruſh											57	66			15	23	17	40	18	31	34		95	93			123	13	211
Vomiting	1	6	3	7	4	6	1	14	7	27	16	19	8	10	1	4	1	1	2	3	6	3	7	16	17	27	69	12	136
Worms	147	107	105	63	85	86	51								19	31	28	17	19	28	17		105	74	424	229		124	830
Wen	1		1		2	2			1		1	2	1	1			1		4				1	4	3	4	4	2	15
Sodainly															63	59	37	62	58	61	78	34	221	233				63	434

Bills of Mortality

Some view this work by Graunt as **the start of the science of demography;** others claim that he was the first epidemiologist; and still others see Observations as the initiation of statistics in that he attempted to interpret mass biological phenomena and social behavior from numerical data

While all of these might be true, political arithmetic, as a movement, was supplanted by statistics in France and Great Britain around the beginning of the nineteenth century; the term “statistics” being adopted from the German, where it was used to describe **“a science concerned with states”**


Blaise Pascal

Classical probability

It is often said that the era of mathematical probability (or, rather, the view of probability as a branch of mathematics) **started with an exchange of letters** between Blaise Pascal (1623-1662) and Pierre de Fermat (1601-1657) that took place in 1654

Their correspondence began with a question posed by Chevalier de Méré, “a gambler and a philosopher” known as the **“The problem of Points,”** one of a large class of so-called “division problems”

Their calculations applied **combinatorics** to questions about repeated gaming, and provided a framework for the so-called **classical approach to interpreting probability**

Classical probability

Here is de Méré's question:

> *Suppose two people, A and B, agree to play a series of fair games (think of tossing a coin) until one person has won a fixed number of games (say 6). They each have wagered the same amount of money, the intention being that the winner will be awarded the entire pot. But, suppose, for whatever reason, the series is prematurely terminated at which point A needs **a** more games to win, and B needs **b**. How should the stakes be divided?*

Seeing this problem for the first time, it's difficult approach, and to be fair, questions like it had been discussed for over a 100 years without a mathematical solution

If a=b, then it seems clear that the players should just divide the pot; but what if a=2 and b=3?

http://www.york.ac.uk/depts/maths/histstat/pascal.pdf

Classical probability

The solution hit upon by Pascal and Fermat is less about the history of the game as it was played before being interrupted, but instead considered **all the possible ways the game might have continued** if it had not been interrupted

Pascal and Fermat reasoned that the game would be over in a+b-1 further plays (possibly sooner); and that there were a total of 2^{a+b-1} possible outcomes (why?)

To figure A's share, you should count how many of these outcomes see A winning and divide by the total -- This fraction is **the probability that A would have eventually won the game**

Classical probability

For example, suppose the game was meant to be played until either A or B had won 6 times; but suppose play is interrupted with A having won 4 games and B three (or a = 6 - 4 = 2 and b = 6 - 3 = 3)

Play could have continued for at most $2^{2+3-1} = 2^4 = 16$ more games, and all the possible outcomes are

AAAA AAAB AABA AABB ABAA ABAB ABBA ABBB
BAAA BAAB BABA BABB **BBAA** BBAB BBBA BBBB

Of these, 11 out of 16 favor A (bolded), so A should take 11/16 of the total and B should take 5/16

Classical probability

The answer was a significant conceptual advance; in addressing this problem, Pascal and Fermat provided **a recipe for calculating probabilities**, one that involves combinatorics (counting outcomes)

In their letters, they address other games of chance with a similar kind of approach, each time taking the **probability of an event as the number of possible outcomes that make up that event** (the number of outcomes that have B winning, for example) **divided by the total number of outcomes**

In forming his solution, Pascal makes use of his famous **triangle of binomial coefficients**, a fact we'll come back to shortly..

Classical probability

Classical probability (so-named because of it's "early and august pedigree") assigns probabilities equally to the possible outcomes that could occur in a given problem, so that **the classical probability of an event is simply the fraction of the total number of outcomes in which the event occurs**

To add yet another heavy-hitter to our list of distinguished mathematicians, Pierre-Simon Laplace (1749-1827) clearly describes the idea as follows (a statement which was later termed the **Principle of Indifference**)

> ***The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible****, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply* ***a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible.*** (1814, 1951 6-7)

This approach is well-adapted to games of chance (throwing dice, pulling cards from a well-shuffled deck) and is the basis for most of the probability problems in your textbook -- In these cases, **symmetry or the open physical process of drawing from a hat make the basic equally likely outcomes intuitively clear**

Classical probability: Hill's tables

In a previous lecture, we considered a simple process of re-randomization to study if the table of results Hill observed describing the effectiveness of Streptomycin could have occurred simply by chance -- Here are Hill's data again

Status	Treatment C	Treatment S	
Survived	38	51	89
Died	14	4	18
	52	55	107

Classical probability: Hill's tables

Under the null hypothesis that Streptomycin had no effect as a treatment for pulmonary tuberculosis, the 18 patients who died would have done so no matter how they were treated; this means **that the fact that we saw 14 deaths with bed rest and 4 with Streptomycin and bed rest was the result of Hill's randomization**

To test this idea, we had to get a sense of how often you would see the 14-4 split if we "re-randomized" Hill's trial; specifically, we took our 107 patients (89 who survived and 18 who died) and randomly assigned them to the control and Streptomycin groups

We re-randomized a number of times and looked at how Hill's result (4 deaths in the Streptomycin group) compared to what we saw through random assignment; if the two were very different, we had evidence that Streptomycin was helping to prevent death from pulmonary tuberculosis

At that point, **we relied on simulation** to generate different divisions of the patients into treatment and control; now, let's take a classical approach to working out the probabilities involved

Classical probability: Hill's tables

We recall that our simulations had us putting treatment and control labels into a "hat" and pulling them one at a time, assigning them to each patient -- In all we had 55 of the 107 labels indicating treatment meaning we have

$$\binom{107}{55} \approx 11,976,930,000,000,000,000,000,000,000,000$$

different ways to randomize our patients into treatment and control

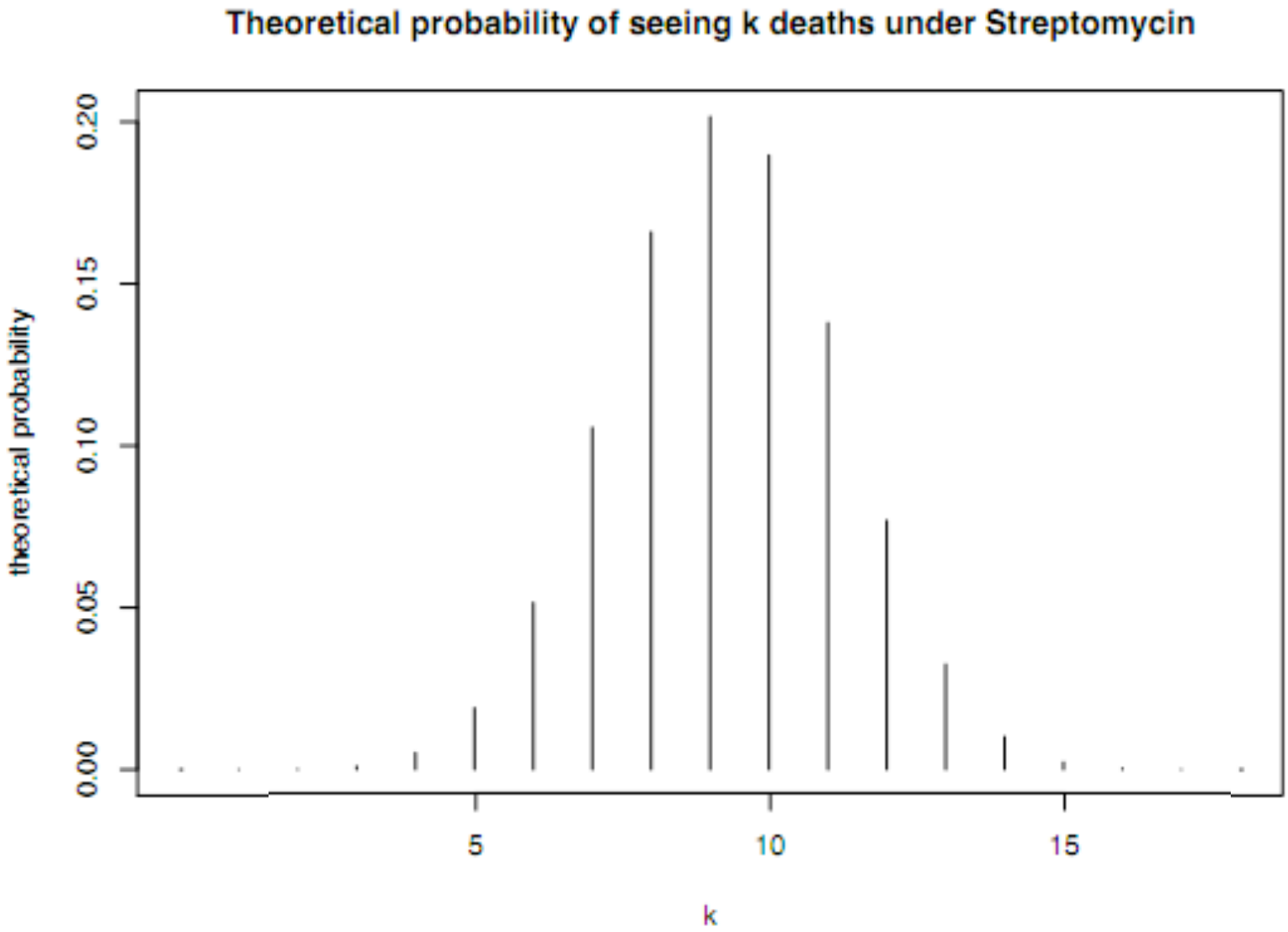
Classical probability: Hill's tables

Now, of these different configurations, how many saw 4 doomed patients randomized into the treatment group?

Recall, that there are $\binom{18}{4}$ ways to select 4 of the 18 patients, and $\binom{89}{51}$ ways to select the remaining 55-4 = 51 patients from the 89 that survived -- That means we have a total of

$$\binom{18}{4}\binom{89}{51}$$

re-randomizations that reproduce Hill's table exactly


Theoretical probability of seeing k deaths under Streptomycin
theoretical probability
0.00
0.05
0.10
0.15
0.20
5
10
15
k

Some axioms

With the classical view of probability, you can establish the basic mathematical framework or calculus for probability; let $\mathcal{X}$ be the set of **all possible outcomes** of some experiment or trial or situation we'd like to study, let A denote **an "event" or collection of outcomes** from $\mathcal{X}$; and finally let P(A) be the probability of A

1. The probability of A is a number between 0 and 1, $0 \leq \mathrm{P}(\mathrm{A}) \leq 1$
2. The probability that an outcome will occur is 1, $\mathrm{P}(\mathcal{X}) = 1$
3. If A and B have no outcomes in common (they're disjoint), then their probabilities add $\mathrm{P}(\mathrm{A} \text{ or } \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$

You'll remember these basic rules from your introductory probability classes -- Our goal here is not to rehash that material but to connect it to the development of probability as a mathematical exercise

Classical probability

While the classical model has all the ingredients for computing with probabilities, that is, it provides us with the basic calculus for probability; **as an interpretation of the concept of probability, the classical approach leaves a bit to be desired**

Ian Hacking puts it this way:

> *"The problems of real life are less tractable. We have stable mortality statistics, but who can ever tell the numbers of diseases? Who can enumerate the parts of the body that are attacked by disease? ... We have statistical regularities but no [fundamental set of outcomes]"*

The framework has limited scope -- It's not always appropriate to assign equal probabilities in more complex settings, and there are cases where the various outcomes are not obviously some known finite number

In addition, **many have criticized the underlying reasoning as circular** -- That is, saying events are "equipossible" already assumes equal probability

The first limit theorem

After the exchange between Pascal and Fermat was published, probability as a mathematical discipline really took off; starting in 1684, for example, Jakob Bernoulli (1654-1705) worked on a text entitled "*Ars Conjectandi*" (or, the *Art of Conjecture*)

JACOBI BERNOULLI,
Profess. Basil. & utriusque Societ. Reg. Scientiar.
Gall. & Prussiæ Sodal.
Mathematici Celeberrimi,

ARS CONJECTANDI,

OPUS POSTHUMUM.

Accedit

TRACTATUS
DE SERIEBUS INFINITIS,

Et Epistola Gallicè scripta
DE LUDO PILÆ
RETICULARIS.

BASILEÆ,
Impensis THURNISIORUM, Fratrum.
cIↃ IↃcc xiii.

For example, working with the axioms of probability, Bernoulli derived the first "limit theorem," a mathematical result often called **the law of averages or the (weak) law of large numbers**

The result is related to repeated trials; namely, if on each trial you have the same probability of success p, then the proportion of successes in n trials p_n "tends to" p

Keep in mind that this is a mathematical result, and idealization, a model; later when we talk about the binomial distribution we'll see that this is not so hard to prove

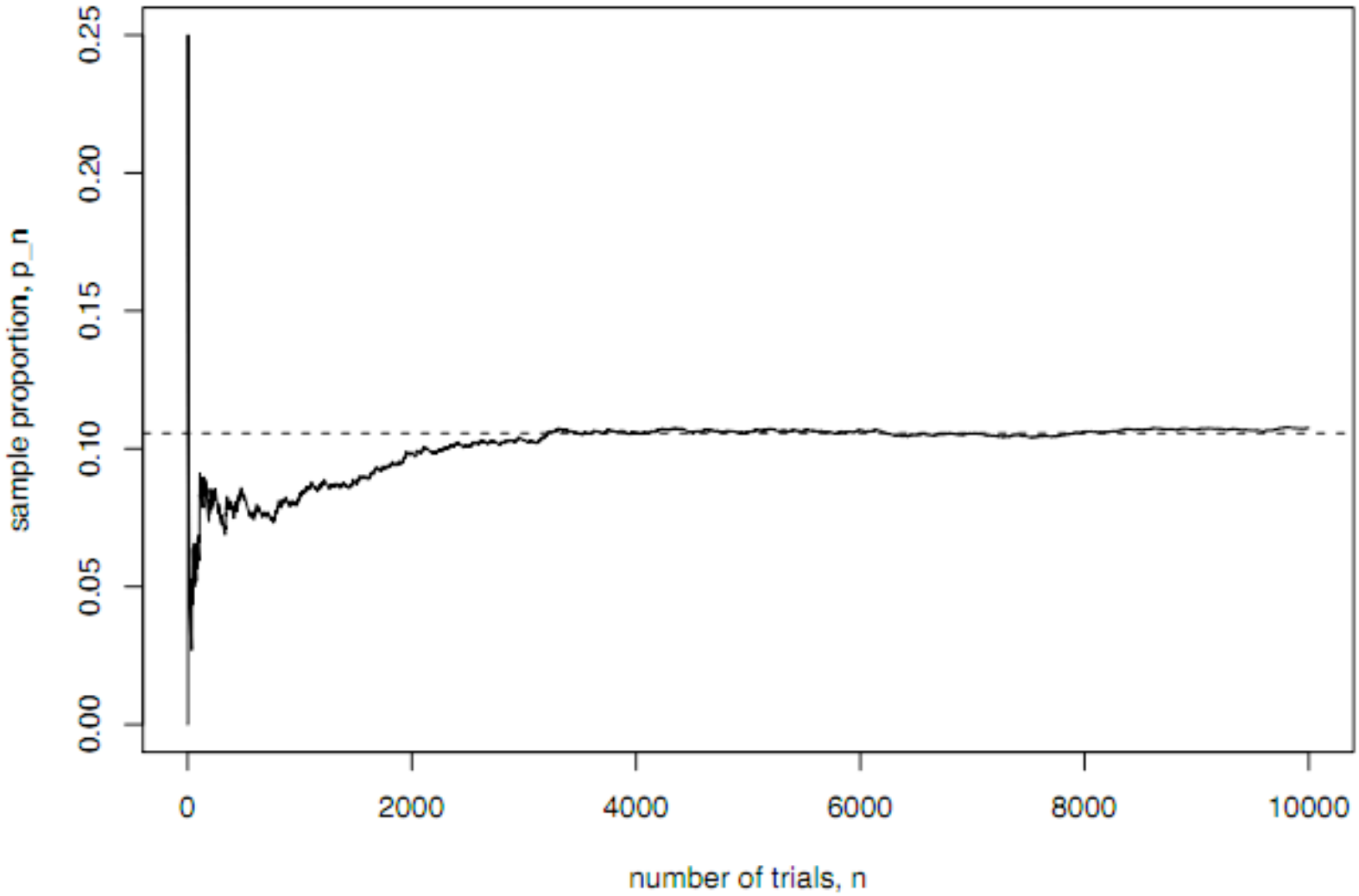
The law of large numbers

Think about the simple case of tossing a fair coin; if each flip is independent of the next (meaning that the result of one flip does not change the probabilities of seeing heads or tails on the next -- more on that shortly), then **the proportion of heads in n tosses should get close to 0.5 as n gets large**

The exact same result could be applied to our randomizations for Hill; suppose now our trial is a random assignment (a draw of 55 names from a bag) and "success" means we get a table with 7 deaths in the Streptomycin group

Then, if we repeat the "trial" many times and look at the proportion of assignments that result in 7 doomed patients going to the Streptomycin group, that number should be close to the actual probability (which we computed to be 0.11)

Here's one simulation...


sample proportion, p_n
0.00
0.05
0.10
0.15
0.20
0.25
0
2000
4000
6000
8000
10000
number of trials, n

Simulation

Bernoulli's theorem is a piece of mathematics -- It says that "in the limit," the classical value of a probability should emerge by repeated trials, and this is **entirely within the world of mathematics**

What we've done on the previous slide is to **replace a mathematical model for probability with one that lives in the computer** -- That is, we are replacing Bernoulli's mathematical idealization for a computerized one

In a previous lecture, we discussed how the computer can be used to **emulate random sequences that have properties predicted by the mathematics** -- The image on the previous page is another instantiation of that result

Simulation

Importantly, **this limit law suggests the rationale for our simulations** -- If we repeat trials often enough, then **the proportions we are computing from our simulations should approach the probabilities we can calculate mathematically**

In short, our simulation, if performed often enough, should provide us with answers (P-values, say) that are close to what we could work out mathematically -- The plot two slides back illustrates this for Hill's tables

The first limit law

The previous slides dealt mainly with mathematical abstraction, the calculus of probability, and its counterpart in simulation software; we now turn to **what all this means in terms of the interpretation of probability**

While much can be said about what precisely Bernoulli proved and what it meant for statistical inference in general, from the point of view of this lecture, his law helped people at the time make sense of **all the stable statistical frequencies that people were observing at the end of the 17th and the beginning of the 18th centuries**

Remember, this is when John Graunt was looking at the **Bills of Mortality,** observing a number of regularities in birth and death and marriage statistics, and even computing life tables proportions of people who died with different ailments -- This is also about the time that Arbuthnot made his own interesting comments on the Bills and the sex ratio

Bernoulli's theorem, a purely mathematical result, **seemed to tie the classical view of probability with the stable frequencies** observed by Graunt and Arbuthnot and others -- **If a random mechanism like coin flipping was at work in the world, then frequencies will be stable**

Aside: Arbuthnot

Divine Providence

John Arbuthnot, a physician to Queen Anne, used the christening records in the London Bills to support an argument for the existence of “Divine Providence”

While Arbuthnot’s larger point is certainly beyond the scope of this course, the article is interesting for us because it is widely regarded as **the first published statistical test of significance**

II. *An Argument for Divine Providence, taken from the conſtant Regularity obſerv'd in the Births of both Sexes. By Dr.* John Arbuthnott, *Phyſitian in Ordinary to Her Majeſty, and Fellow of the College of Phyſitians and the Royal Society.*

AMong innumerable Footſteps of Divine Providence to be found in the Works of Nature, there is a very remarkable one to be obſerved in the exact Ballance that is maintained, between the Numbers of Men and Women; for by this means it is provided, that the Species may never fail, nor periſh, ſince every Male may have its Female, and of a proportionable Age. This Equality of Males and Females is not the Effect of Chance but Divine Providence, working for a good End, which I thus demonſtrate:

Let there be a Die of Two ſides, M and F, (which denote Croſs and Pile), now to find all the Chances of any determinate Number of ſuch Dice, let the Binome M+F be raiſed to the Power, whoſe Exponent is the Number of Dice given; the Coefficients of the Terms will ſhew all the Chances ſought. For Example, in Two Dice of Two ſides M+F the Chances are $M^2+2\,MF+F^2$, that is, One Chance for M double, One for F double, and Two for M ſingle and F ſingle; in Four ſuch Dice there are Chances $M^4+4\,M^3F+6\,M^2F^2+4\,MF^3+F^4$, that is, One Chance for M quadruple, One for F quadruple, Four for triple M and ſingle F, Four for ſingle M and triple F, and Six for M double and F double; and univerſally, if the Number of Dice be n, all their Chances will be expreſſed in this Series

M^n+

Divine Providence

In his argument for “Divine Providence,” Arbuthnot considers the gender of babies born in London

While reflecting on the lives of men and women in 1710 England, he notes that men are subject to various “external Accidents” as they “must seek their Food with danger”

For Arbuthnot, these external accidents meant that to maintain a balance between men and women, Divine Providence would arrange for the birth of a larger proportion of boys than girls

Therefore, for Arbuthnot, to demonstrate that boys and girls were not born in equal proportion was to argue in favor of the existence of Divine Providence

the middle Term will not exactly give A's Chances, but his Chances will take in ſome of the Terms next the middle one, and will lean to one ſide or the other. But it is very improbable (if mere Chance govern'd) that they would never reach as far as the Extremities: But this Event is wiſely prevented by the wiſe Oeconomy of Nature; and to judge of the wiſdom of the Contrivance, we muſt obſerve that the external Accidents to which are Males ſubject (who muſt ſeek their Food with danger) do make a great havock of them, and that this loſs exceeds far that of the other Sex, occaſioned by Diſeaſes incident to it, as Experience convinces us. To repair that Loſs, provident Nature, by the Diſpoſal of its wiſe Creator, brings forth more Males than Females; and that in almoſt a conſtant proportion. This appears from the annexed Tables, which contain Obſervations for 82 Years of the Births in *London*. Now, to reduce the Whole to a Calculation, I propoſe this.

Problem. A lays againſt B, that every Year there ſhall be born more Males than Females: To find A's Lot, or the Value of his Expectation.

It is evident from what has been ſaid, that A's Lot for each Year is leſs than $\frac{1}{2}$; (but that the Argument may be ſtronger) let his Lot be equal to $\frac{1}{2}$ for one Year. If he undertakes to do the ſame thing 82 times running, his Lot will be $\frac{1}{2}|^{82}$, which will be found eaſily by the Table of Logarithms to be $\frac{1}{4836\ \text{oooo}\ \text{oooo}\ \text{oooo}\ \text{oooo}\ \text{oooo}\ \text{oooo}}$. But if A wager with B, not only that the Number of Males ſhall exceed that of Females, every Year, but that this Exceſs ſhall happen in a conſtant Proportion, and the Difference lye within fix'd limits; and this not only for 82 Years, but for Ages of Ages, and not only at *London*, but all over the World; (which 'tis highly probable is Fact, and deſigned that every Male may have a Female of the ſame Country and ſuitable Age) then A's Chance will be near an infinitely ſmall Quantity, at leaſt

leſs

Statistics and Divine Providence

To make his case, Arbuthnot starts with a simple **probability model** in which the sex of a baby is determined by the toss of a fair coin; that is, we see an “M” with probability 0.5 and “F” with probability 0.5*

Because the underlying mechanism is assumed to be **stochastic****, you expect to see fluctuations from year to year in the proportion of boys to girls; some years you will see more boys, in others, more girls

But because the gender of each is determined by the toss of a fair coin, **Arbuthnot reasoned that for any given year, the probability that boys outnumbered girls was again 0.5**

Arbuthnot then uses the christening records to “test” the hypothesis that boys and girls are born in equal proportion; or, rather that boys outnumber girls in a given year based on the toss of a fair coin

So, what do the data say?

* Arbuthnot actually refers to “a Die of Two sides, M and F

** Stochastic, from the Greek "Στόχος" which means "aim, guess", means of, relating to, or characterized by conjecture and randomness”

Anno.	Christened. Males.	Females.
1629	5218	4683
30	4858	4457
31	4422	4102
32	4994	4590
33	5158	4839
34	5035	4820
35	5106	4928
36	4917	4605
37	4703	4457
38	5359	4952
39	5366	4784
40	5518	5332
41	5470	5200
42	5460	4910
43	4793	4617
44	4107	3997
45	4047	3919
46	3768	3395
47	3796	3536

Anno.	Christened. Males.	Females.
1648	3363	3181
49	3079	2746
50	2890	2722
51	3231	2840
52	3220	2908
53	3196	2959
54	3441	3179
55	3655	3349
56	3668	3382
57	3396	3289
58	3157	3013
59	3209	2781
60	3724	3247
61	4748	4107
62	5216	4803
63	5411	4881
64	6041	5681
65	5114	4858
66	4678	4319

Anno.	Christened. Males.	Females.
1667	5616	5322
68	6073	5560
69	6506	5829
70	6278	5719
71	6449	6061
72	6443	6120
73	6073	5822
74	6113	5738
75	6058	5717
76	6552	5847
77	6423	6203
78	6568	6033
79	6247	6041
80	6548	6299
81	6822	6533
82	6909	6744
83	7577	7158
84	7575	7127
85	7484	7246
86	7575	7119
87	7737	7214
88	7487	7101

Anno.	Christened. Males.	Females.
1689	7604	7167
90	7909	7302
91	7662	7392
92	7602	7316
93	7676	7483
94	6985	6647
95	7263	6713
96	7632	7229
97	8062	7767
98	8426	7626
99	7911	7452
1700	7578	7061
1701	8102	7514
1702	8031	7656
1703	7765	7683
1704	6113	5738
1705	8366	7779
1706	7952	7417
1707	8379	7687
1708	8239	7623
1709	7840	7380
1710	7640	7288

Statistics and Divine Providence

Arbuthnot noticed that **in every of the 82 years from 1629 to 1710, there were more boys christened than girls**; while this might seem like a compelling enough observation on its own, Arbuthnot takes it farther

His idea was to compare this observation to the probability model he hypothesized for the data; that is, if boys outnumber girls in a given year based on the toss of a fair coin, what is the chance that we see 82 heads in 82 tosses?

For that matter, what is the chance that we would see any large number, say 70 or 80 heads out of 82 tosses?

Statistics and Divine Providence

Arbuthnot showed that the chance of seeing boys outnumbering girls year after year is extremely unlikely, approx. **1 in 4,836,000,000,000,000,000,000,000**

The idea, then, is that if the hypothesis that boys and girls are born in equal proportion is correct, then the christening data is extraordinarily unlikely; to give you some perspective, **the odds of winning the California Super Lotto Jackpot is 1 in 18,000,000,000**

With this calculation in hand, we would feel comfortable **abandoning our hypothesis that boys and girls are born in equal proportion**

Back to lecture

The frequentist view of probability

Which leads us to another view of probability -- In the frequentist view, we would say, for example, that an event has probability 1/3 if the event occurs about 1/3 of the time **in a long sequence of repetitions done under more or less ideal circumstances**

Of course the relationship between relative frequencies (proportions over many trials) and probability will come as **no surprise to people who actually play games of chance** -- Certainly people have been assigning odds on games for a long time, long before Bernoulli, Pascal and Fermat

And frankly, if you ask most practicing statisticians what we mean by probability and they aren't ready for a long conversation, this is the kind of answer you'll get -- **Certainly, over the years there have been many attempts to "verify" this interpretation of probability**, to extract some "objective" sense of probability by repeating trials a large number of times...

Passing time

In his “A Treatise on Probability,” the British Economist John Maynard Keynes discusses several attempts to verify the conclusions of Bernoulli’s Theorem -- He writes “**I record them because they have a good deal of historical and psychological interest, and because they satisfy a certain idle curiosity from which few students of probability are altogether free.”**

The French naturalist Count Buffon (1707-1788), who “assisted by a child tossing a coin into the air” recorded 2048 heads in 4040 flips (for a relative frequency of 0.507)

A similar experiment was carried out by **a student of the British mathematician De Morgan** (1806-1871) “for his own satisfaction” involving 4092 tosses, 2048 of which were heads (relative frequency of 0.500

Passing time

The Belgian mathematician/astronomer/statistician/sociologist Adolphe Quetelet (1796-1874) drew 4096 balls from an urn, replacing them each time, and recorded the result at different stages; in all, he drew 2066 white balls and 2030 black balls (relative frequency of 0.504)

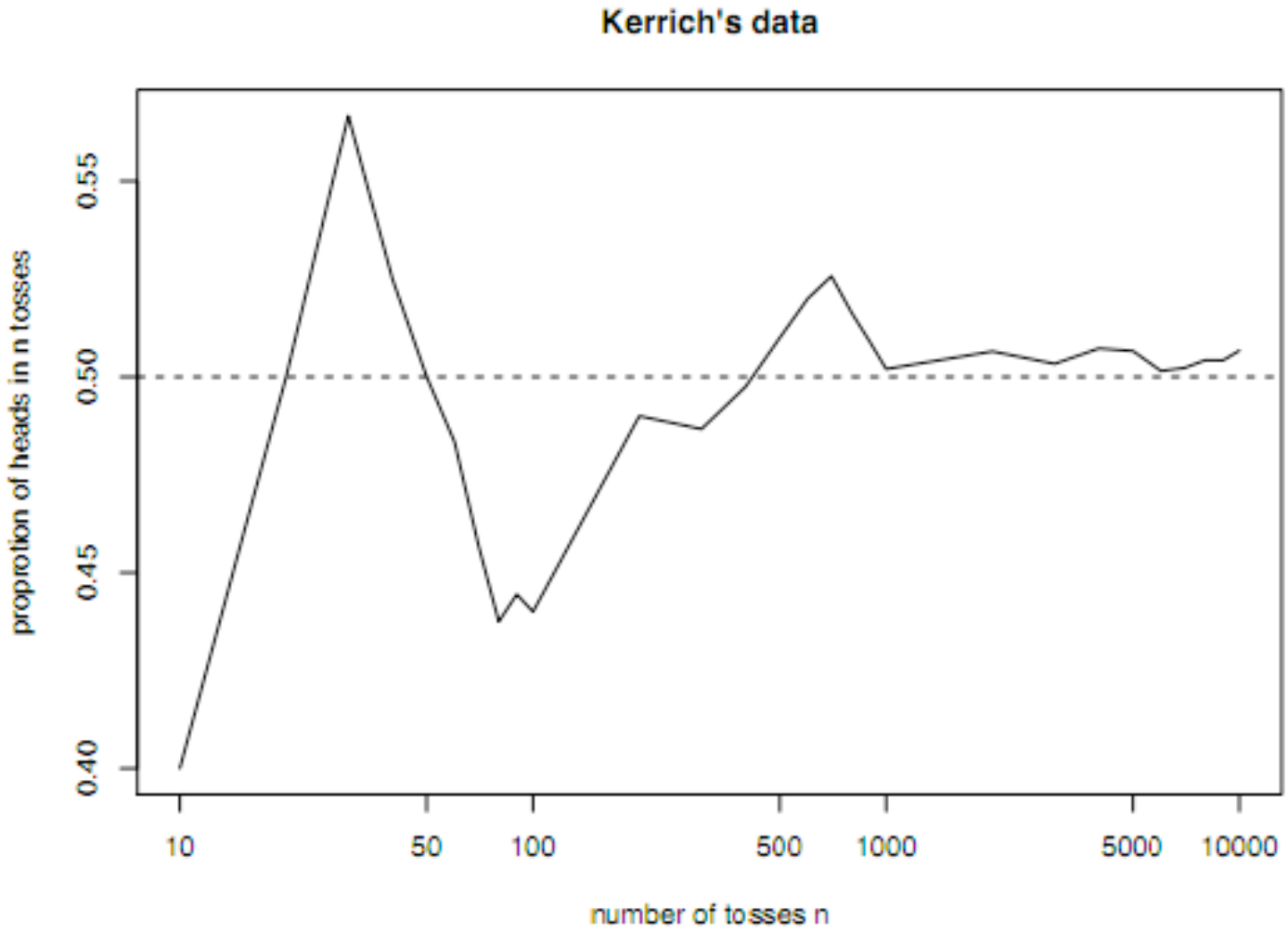
English economist W S Jevons (1835-1882) made 2048 throws of ten coins at a time; in all, he saw 20,480 tosses out of which 10,353 were heads (relative frequency of 0.506, although this is not quite the same kind of trial)

Around 1900, the English statistician Karl Pearson (1857-1936) made two heroic studies; the first involved 12,000 tosses (relative frequency of heads 0.52) and the second 24,000 times (12,012 of which landed heads for a relative frequency of 0.501)

While imprisoned by the Germans during World War II, **the South African mathematician John Kerrich** tossed a coin 10,000 times, 5067 of them heads (this gives a relative frequency of 0.5067 -- while interned, he also recorded a monograph **"An Experimental Introduction to the Theory of Probability"**

Passing time

Kerrich's data show the same pattern in relative frequencies that we observed from our computer simulation; as we repeat the trial over and over, the proportion of successes "settles down"



Passing time

Prisoners of war and 19th century intellectuals are not the only people to have tested the frequency notion of statistics; as it is the dominant interpretation of probability covered in introductory statistics textbooks, **students of statistics are routinely forced to participate**

At the right we have the results from several semesters of an introductory statistics class taught by Robin Lock at St. Lawrence University; he actually has his students record data on flips, spins and tips

"*I have my students do a lab on this each semester. They do 100 flips, around 70 spins and 50 tips each - so I've accumulated lots of data, but I'm never completely sure about the reliability of the data. They do the trials on their own outside of class so I can't monitor how carefully they follow the instructions*"

Flip (H)	Trials	prop	Semester
1079	2100	0.51	Fall 97A
1121	2260	0.50	Fall 97B
1071	2200	0.49	Spring 98
1093	2200	0.50	Fall 98A
1041	2000	0.52	Fall 98B
1232	2400	0.51	Fall 98C
802	1500	0.53	Spring 99
1002	2005	0.50	Fall 99A
1070	2200	0.49	Fall 99B
1000	2000	0.50	Spring 00
1021	2050	0.50	Fall 00A
984	1900	0.52	Fall 00B
1036	1900	0.55	Fall 01A
1157	2300	0.50	Fall 01B
14709	**29015**	**0.507**	**Combined**

www.dartmouth.edu/~chance/chance_news/for_chance_news/ChanceNews10.03/Robin.Lock.data.htm

Passing time

While many of these attempts seemed to behave according to frequentist expectations, some did not not; **when things went wrong, the chief culprit was the experimental setup**

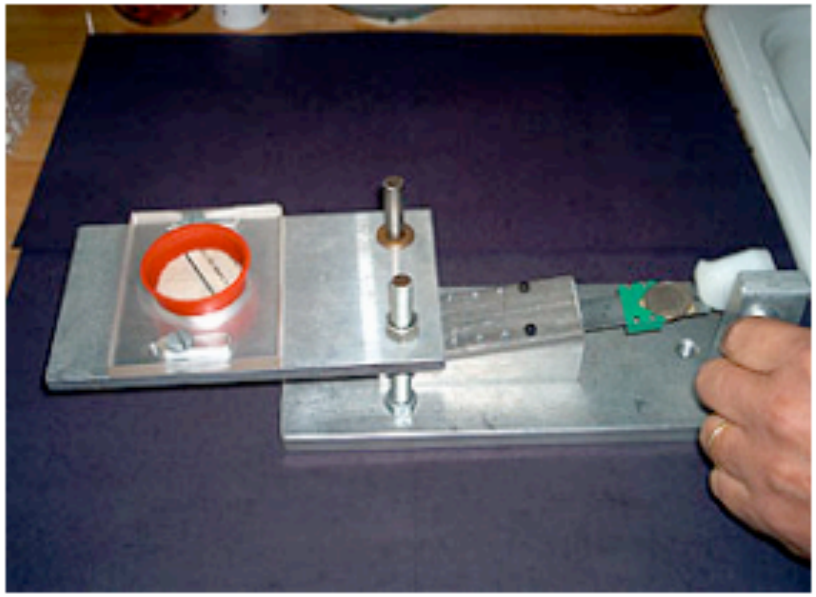
In 1850 **the Swiss astronomer Wolf** rolled one white and one red die 20,000 times- - Keynes writes “the relative frequency of the different combinations was very different from what theory would predict... an explanation is easily found... **the dice must have been very irregular**...” and he concludes “**This, then is the sole conclusion of these immensely laborious experiments -- that Wolf’s dice were very ill made**”

Ten years later, Wolf tried it again, this time with four die (white, yellow, red and blue), recording results from 280,000 tosses; “**It is not clear that Wolf had any well-defined object in view in making these records**... but they afford a wonderful example of **the pure love of experiment and observation**

And what are we to make of...

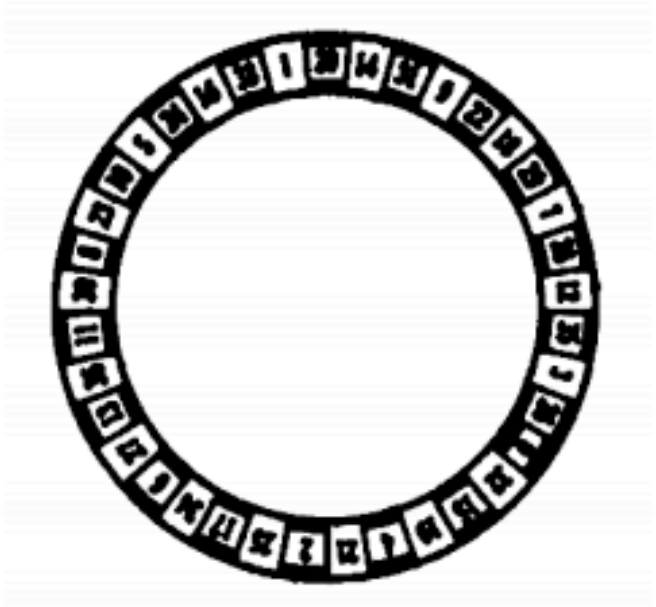
A device that will consistently flip a coin the same way; this machine was made for Persi Diaconis, a well-known Stanford statistician

Diaconis studied the dynamics of a coin toss, mapping out the right initial conditions so that the coin always lands heads up



Passing time

In the 1890s, Karl Pearson looked to a roulette table in Monte Carlo to amass experimental evidence for the "Laws of Chance" to provide material for a series of popular lectures he gave in 1893; he supervised 16,141 spins (a relative frequency of 0.5015 reds)



While the split of red/black "matched" what the theory anticipated, the **proportion of times each separate number appeared was off;** on observing the discrepancy, he commented

"I did not immediately assume that the laws of chance did not apply to Monte Carlo roulette, but I considered myself unfortunate to have hit upon a month of roulette which was so improbable in its characteristics that it would only occur, on the average, once in 167,000 years of continuous roulette playing."

Passing time

He then went on to test another feature of the data, **the relative frequency of the runs of a common colo**r -- Under the "null model" of red and black occurring independently from one play to the next, you can work out the theoretical probability of seeing runs of different lengths and here Pearson finds too few short runs and too many alternations (R,B,R,B...), deviations so severe that

> *"If Monte Carlo roulette had gone on since the beginning of geologic time on this earth, we should not have expected such an occurrence as this... to have occurred once... The man of science may proudly predict the results of tossing halfpence, but the Monte Carlo roulette confounds his theories and mocks at this laws!"*

While these results are interesting to recall, there is an idea here that we'll return to a couple times -- When thinking about whether or not data are consistent with a model, we might do well to consider various other test statistics

Back to the computer

I bring up the ideas from Pearson because to say that a series of numbers “is random” actually **implies a large number of patterns that we can anticipate mathematically**; just as Bernoulli described the behavior of relative frequencies, other characteristics like run lengths might also be used

These are the kinds of tests that people put computer-based random number generators through; **as more and more of statistical practice depends on simulation, the “quality” of our random numbers is something to consider!**

As I said before, however, R and its pseudo-random number generators is capable of supporting your simulations, your computational exploration of results from probability; that is, **when you can’t prove something, simulate!**

The frequentist view of probability

At a technical level, strict frequentists view probability as arising from a sequence of identical trials; there is a problem lurking, however, in determining how long we should go

Sure, probabilities may seem stable after 10,000 trials, but who is to say that they won't change farther in the series? In some sense, frequentists are led to imagining not just a long sequence of trials but an infinitely long one

The frequentist view also provides us with no real ability to reason about singular events like the election of Obama or whether or not a particular person will get cancer; there is no imaginary infinite sequence of trials here

The frequentist view of probability

John Venn (1834-1923, of Venn diagram fame) is often credited as the founder of the frequency view of probability; in his book called the *Logic of Chance,* he writes the **fundamental conception is that of a series** which 'combines the individual irregularity with aggregate regularity'

With his concept of a series, Venn imagines a population (say, the repeated throws of a die or patients with a particular disease); the probability of an event is the relative frequency in the series; probability has no meaning except in connection with such a series, and **any probability must refer to a series**

But how long should this series be? Consider the probability that a particular coin, when tossed, will land heads; how many times do we have to flip it before we know what it's probability is? Basing probability on frequencies involves infinite limiting procedures, experiments that we could not carry out, even in principle

The frequentist view of probability

While we might think of him as the founder of the frequentist view, it was one he struggled with; Venn, understood that **the idea of an infinite series was problematic**

To Venn, a series had **a kind of identity -- probability statements were assertions about classes of things**, but he understood that things in the world change; he wrote about species changing, weather patterns changing, the world evolving and so any ideas of an infinitely long sequence (in time or in numbers of items) were hard to justify as a basis for defining probability

To be able to reason mathematically about probabilities, Venn sidesteps his concerns about the appropriate "reference class" or population in which to compute a real probability, and invents **'substitute series' that 'must be regarded as indefinitely extensive in point of number and duration'**

Frequentist statistics

So far, the inferential procedures we have studied (re-randomization and P-values) are based on **the frequentist notion of probability** --They refer to an (**imaginary) set of possible alternative outcomes that could have happened had we repeated the experiment many times**

D.R. Cox puts it this way

> *In the first so-called frequentist approach, we ... use probability as representing a long-run frequency... [W]e measure uncertainty via procedures such as confidence limits and significance levels (P-values), whose behaviour ... is assessed by* ***considering hypothetically how they perform when used repeatedly under the same conditions.*** *The performance may be studied analytically or by computer simulation.*

> *In that, the procedure is calibrated by what happens when it is used, it is no different from other measuring devices.*

The subjective view

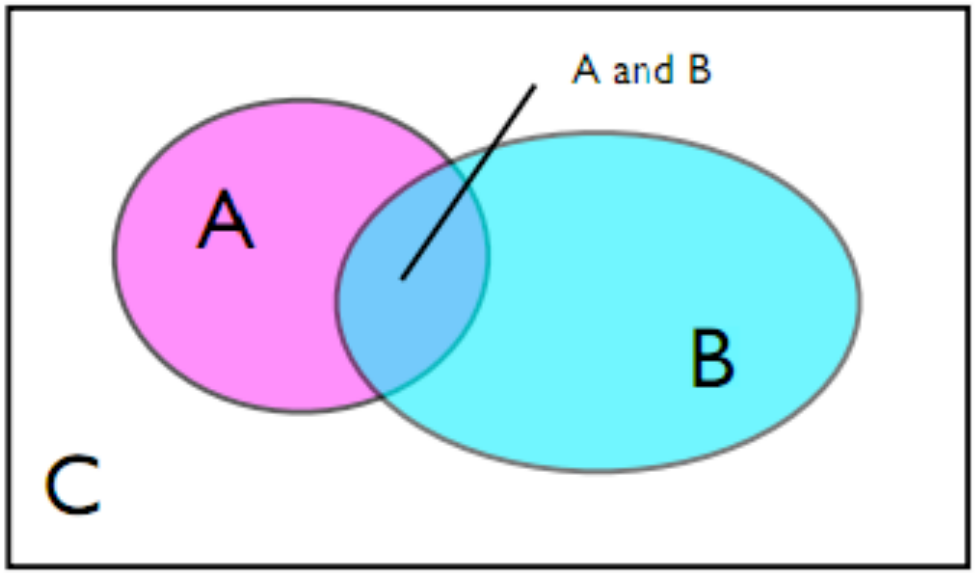
The third view of probability we will talk about is more in line with Hacking's use of the term probability as a "relation between a hypothesis and the evidence for it"

It has its roots in a result by the Rev. T. Bayes, published in 1763 after his death; Bayes Theorem is a simple fact about conditional probabilities, that we will define next

The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening.

The subjective view

Let's ground this a little, here's a Venn diagram (same Venn, different dogma); and here is a paraphrasing of a lecture by de Finetti in 1979 (more on him in a moment)



Let us let C represent everything we initially know to be true. Suppose that the diagram is to scale: in this way, taking the area of C as a unit, the area of the other events represents the respective probabilities. When a new piece of evidence B is acquired, B rules out all the parts that lie outside itself (that is, logically incompatible with B). Hence the part of A that is compatible with the new information is the "intersection" or "A and B". By normalizing, putting the area of B equal to 1, we obtain the new probabilities. Thus the probability of A given B will equal the area of "A and B" divided by the area of "B" If this area remained unchanged, then we would say that the events are independent.

More axioms

We don't want to give the impression that conditional probability is a construction unique to the subjective view; on the contrary, we can add it to our list of axioms we had some time back (next page)

The notion of conditional probability gives us a mathematical way to express independence of events; if the conditional P(A) = P(A|B) then A and B are said to be independent -- this also means that since P(A|B) = P(A and B)/P(B), for independent events P(A and B) = P(A)P(B)

A more complete set of axioms

Below we list a more complete set of axioms for the calculus of probability; let $\mathcal{X}$ be the set of all possible outcomes of some experiment or trial or situation we'd like to study, let A denote an "event" or collection of outcomes from $\mathcal{X}$; and finally let P(A) be the probability of A

1. The probability of A is a number between 0 and 1, $0 \leq P(A) \leq 1$
2. The probability that an outcome will occur is 1, $P(\mathcal{X}) = 1$
3. If A and B have no outcomes in common (they're disjoint), then their probabilities add $P(A \text{ or } B) = P(A) + P(B)$
4. Conditional probability:
$$P(A|B) = P(A \text{ and } B)/P(B) \quad \text{or} \quad P(A \text{ and } B) = P(A|B)P(B)$$
5. The law of total probability: $P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_J)P(B_J)$ where $B_1, B_2, \ldots, B_J$ are all disjoint and their union is $\mathcal{X}$

The subjective view

As we mentioned at the beginning of the lecture, probability moves between two poles, as a framework for reasoning and as a "stable law" for frequencies

The subjective view rejects the interpretation of probability as a physical feature of the world and interprets probability as a statement about an individual's state of knowledge; Persi Diaconis at Stanford says "Coins don't have probabilities, people have probabilities"

While we could provide a fairly long history of how the first idea developed, we'll focus instead on one of the main proponents, Bruno de Finetti (1906-1985); he began his *"Theory of Probability"* with the statement "Probability does not exist"

The subjective view

To de Finetti, the definition of probability and its evaluation are two different things; he takes issue with the frequentists and the classical probabilists who seem to conflate these two and in so doing embrace a "rigid" attitude toward probability

By contrast, subjectivism maintains **a distinction between definition and evaluation;** probability is defined as the degree of belief "as actually held by someone, on the ground of his whole knowledge, experience, information" regarding an event whose outcome is uncertain

De Finetti writes:

> *The subjective theory... does not content that the opinions about probability are uniquely determined and justifiable. Probability does not correspond to a self-proclaimed "rational" belief but to the effective personal belief of anyone...*

He contends that

> *"every probability evaluation essentially depends on two components: (1) the objective component, consisting of the evidence of known data and facts; and (2) the subjective component, consisting of the opinion concerning unknown facts based on known evidence*

The subjective view

It is in the way that the subjectivists incorporate personal beliefs in the evaluation of probability that makes the framework unique; De Finetti shows that as long as your beliefs are "coherent" in some sense, they can be expressed in terms of a mathematical quantity, a probability distribution (the exact notion of coherence has to do with betting and you making decisions based on your beliefs that are not sure to loose money)

> *"The conceptual basis for this numerical measure will be seen to derive from the formal rules governing quantitative, coherent preferences, irrespective of the nature of the uncertain events under consideration. This is in vivid contrast to what are sometimes called the classical and frequency approaches to defining numerical measures of uncertainty, where the existence of symmetries and the possibility of indefinite replication, respectively, play fundamental roles in defining the concepts for restricted classes of events"*

In the subjective view, as you collect data, you update your beliefs using the conditioning argument we read a few slides back; in fact, de Finetti (his so-called Representation Theorem) shows that this view allows one to build up an entire mathematical framework from which we can interpret why stable frequencies might occur, how to think about models... the works!

The subjective view

Here is the connection to Bayes; one version of Bayes theorem is a direct consequence of the definition of conditional probability

$$P(A)\,P(B|A) = P(A \text{ and } B) = P(B)\,P(A|B)$$

If we take A to be our data (relabel it D) and B to be some hypothesis about the world (call it H) then we can rewrite this as

$$P(H|D) = P(D|H)\,P(H) / P(D)$$

This describes how our initial beliefs about H, P(H), are transformed in the face of data to P(H|D) using, in part, what we expect to see from data if H were true.

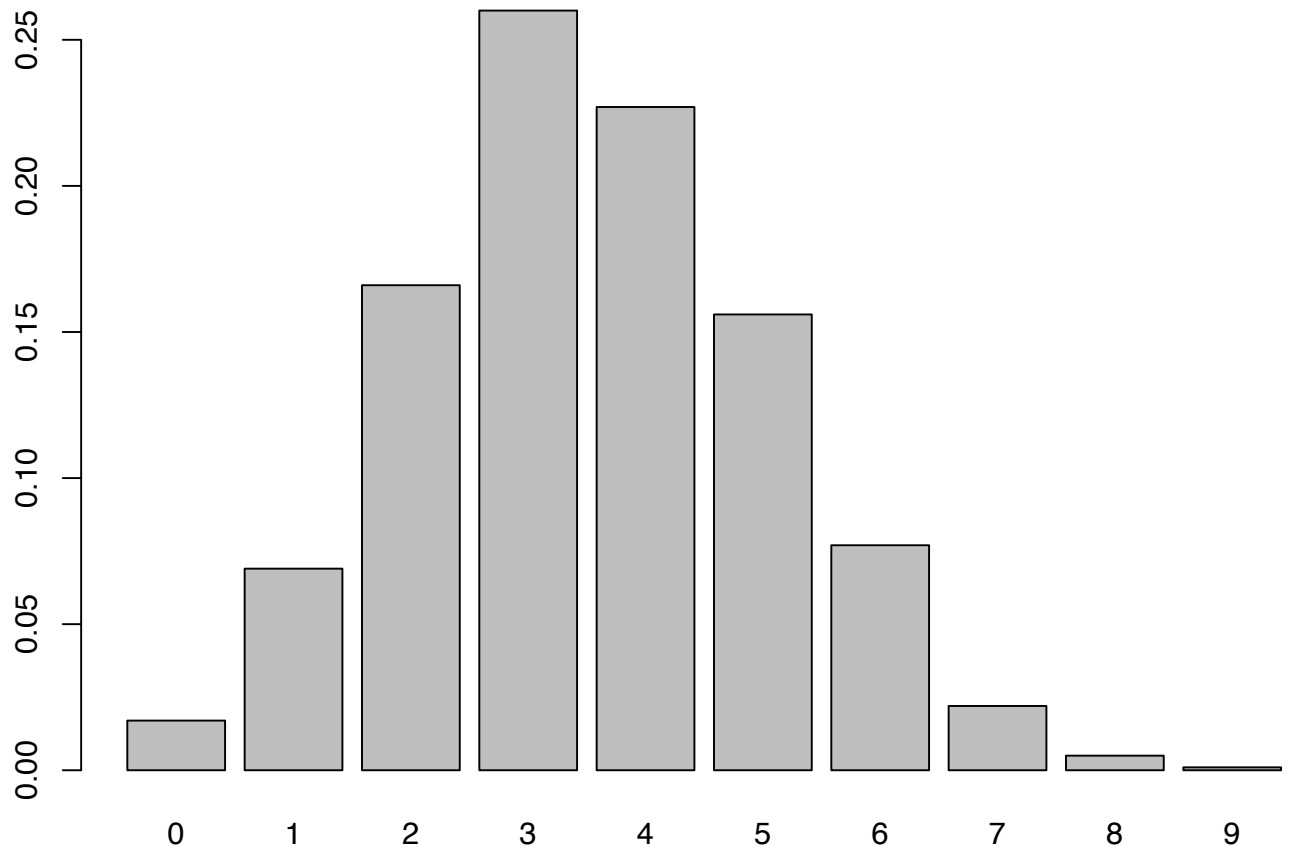
In general, the subjective view provides an elegant way to draw conclusions from data; it involves fundamentally different kinds of inferences than we will be making in this class... although maybe one day toward the end of the quarter we'll take a look at what this has to offer

A simple example

Suppose we want to test a new therapy for some disease, **the standard treatment for which has a historical success rate of 35%** -- To examine the effectiveness of the new therapy we prescribe it to 10 patients and see whether they improve or not

In a frequentist framework, we would consider the “null hypothesis” that the new therapy is the same (or at least, not better) as the traditional treatment -- **Under the null our experimental results should look like 10 coin tosses with success probability (a patient getting better) being 0.35**

We can simulate (a la Arbuthnot) or use a mathematical result about the Binomial distribution to compute the distribution for the number of patients seeing improvement under this model



mathematical table using pbinom in R (more later)

0	1	2	3	4	5	6	7	8	9	10
0.013	0.072	0.176	0.252	0.238	0.154	0.069	0.021	0.004	0.001	0.000

The frequentist approach

At the end of the experiment, suppose we see 7 patients improve -- We can use our simulations or the mathematical table to compute the probability of seeing 7 or more successes if the chance of a success is p=0.35

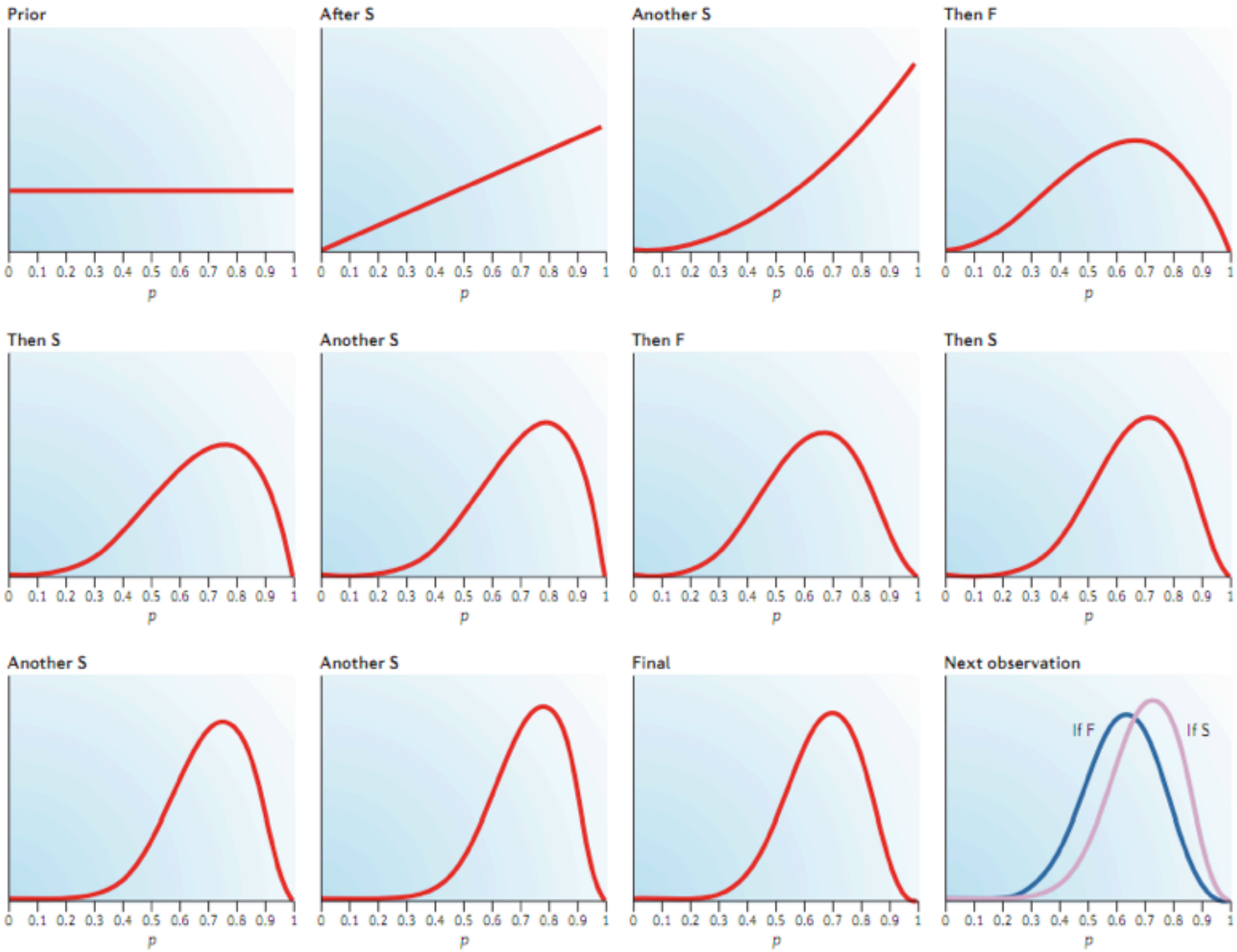
In this case, we sum up 0.021+0.004+0.001 = 0.026 to compute our P-value and we would reject the null hypothesis at the 0.05 level -- All pretty standard at this point

The Bayesian approach

In the Bayesian framework, **we use probability to express our uncertainty about unknown quantities, in this case the probability p** that someone improves on the new therapy

Our “prior” assessment might be one of complete ignorance -- We have no idea what value p might be except that it is in the interval [0,1], leading us to the uniform distribution

Starting from here, we can introduce data and update our beliefs about p -- Suppose that the experiment resulted in a sequence of successes and failures SSFSSFSSSF (again, 7 successes and 3 failures, but now listed in order of occurrence), then on the next page, **we update our beliefs sequentially**


Prior
After S
Another S
Then F
Then S
Another S
Then F
Then S
Another S
Another S
Final
Next observation
If F
If S
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
p

The Bayesian approach

While the mechanics of this are still opaque, the idea is clear -- **As we collect data, our notion of what values p can take become more and more sharp**, in this case coalescing on 0.7 (because we observed 7 out of 10 successes)

If we wanted to evaluate whether the new therapy was no better than the existing treatment, we would simply **compute the weight our current beliefs assign to the region to the left of 0.35** -- Literally we would find the area under the curve in the “Final” box to the left of 0.35... it is just 0.014

This number functions a bit like a P-value, but notice that the interpretation is very very different -- **The P-value references an imaginary set of experiments** while this approach **attempts to assess the evidence (data and prior beliefs) to support the idea** that p is less than or equal to 0.35

More refined notions

In the next few lectures, we'll see more refined frequentist constructions besides hypothesis testing -- Confidence intervals, for example, will express our evolving uncertainty as we collect more data

More on that in the next lecture...

A simple example

My point here is not to dwell on Bayesian statistics, but instead to demonstrate that **probability can be interpreted in different ways and that your view of the concept will change the way you reason with data**

For the most part, this class will remain with the frequentist tradition (as we have so much invested in randomization already!) but it's worth seeing other frameworks!

A computational view

There is one last view of probability I'd like to mention; at the left we have two pictures of Andrey Kolmogorov (1903-1987); he was a Russian who in the 1930s produced an axiomization of probability theory, a mathematical framework grounded in "measure theory"

Early in his career, he was a frequentist, believing that one should interpret probabilities as the result of long-run proportions of events in identical and independent trials

Toward the end of his life, he had a change of heart, and wanted to be able to speak about probability in finite terms; this led him to a somewhat remarkable line of reasoning

A computational view

Let's consider programs that "print" strings of numbers (let's say 0's and 1's for simplicity) -- Now, given a string, let's think about **the shortest program that would print the string**

If the string has a high degree of regularity

0101010101010101010101010101010101010101

then it can be printed with a short program (you just need to say **print "01" 20 times**) -- If the string has less structure,

10110110011111011100011101000100101000 11

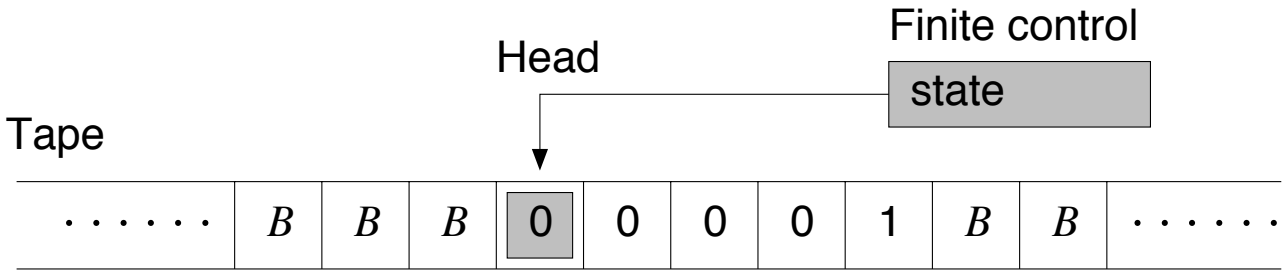
we might require a longer program

You can always have at least one program that does the job if you provide the actual string to the program and say

print "10110110011111011100011101000100101000 11"

A computational view

This feels very abstract and there are lots of questions to answer about the computer language you're using and what the commands might look like and so on -- Kolmogorov worked with **the mathematical abstraction of a computer known as a Turing machine**



This machine reads from a tape divided into cells and containing data (0 or 1 or a blank) and then takes action depending on its current state (taken from a set of states Q) -- The actions include moving the head, writing something to the tape or changing its state

The actions are specified in **a transition function** which you can think of as a program -- In 1936 A. M. Turing proposed the Turing machine as a model of **``any possible computation."**

A computational view

Kolmogorov linked this idea to a notion of what he called **universal probability** -- Simply, if we let L(s) be the length of the shortest program to print a string s (some pattern of 0's and 1's say), then $2^{-L(s)}$ can be thought of as its probability

Interestingly, this definition **embeds "Occam's razor",** the idea that the simplest explanation for an event is likely correct -- **Simple strings are much more probable than more complex strings and the most complex strings are considered "random"**

A computational view

The ideas behind this computational view are embedded in a number of statistical tools that try to choose between models for data -- They are a bit beyond the scope of this course, but the general framework should be intuitive

I bring this up now just to hint at the fact that probability is a fairly rich topic on its own and there are interesting approaches to the subject, each of which come with notions of inference and “randomness”

Back to estimation!

Having gone through several views of probability and their impact for inference, we'll now return to our regularly scheduled frequentist program -- We will return to Bayesian analysis later in the term

Last time we examined two strategies for estimating in the context of a parametric family -- We'll now talk a bit about how one selects an estimator, providing some vocabulary around their properties

Properties of estimators

Suppose we are given a sample $X_1, \ldots, X_n$ of size n that are independent draws from some distribution $f(x|\theta^*)$ that's part of a parametric family $f(x|\theta)$

An estimate $\widehat{\theta}$ of θ^* is just some function of $X_1, \ldots, X_n$, or in symbols $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ -- We view $\widehat{\theta}$ as a random variable in the sense that each time we repeat our experiments, we would collect another sample of data, producing a different estimate

We refer to the distribution of $\widehat{\theta}$ over these repeated experiments as its **sampling distribution --** A frequentist tool for "calibrating" (recall the earlier quote from Cox) what happens when the estimate is used

Unbiasedness

First, we consider the “center” of the sampling distribution and define the bias of an estimate to be

$$\text{bias}(\widehat{\theta}) = \text{E}\widehat{\theta} - \theta^*$$

where the expectation is with respect to the sampling distribution of $\widehat{\theta}$ -- We say that an estimate is unbiased if $\text{bias}(\widehat{\theta}) = 0$ or $\text{E}\widehat{\theta} = \theta^*$

Variance

We can also consider the variance (or spread) of an estimate across repeated trials -- That is, how wide is the sampling distribution?

The standard deviation of $\widehat{\theta}$ is called its standard error and is defined as

$$\mathsf{se}(\widehat{\theta}) = \sqrt{\mathsf{var}(\widehat{\theta})}$$

Efficiency

If $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are two unbiased estimators of θ^* with variances $\mathsf{var}(\widehat{\theta}_1)$ and $\mathsf{var}(\widehat{\theta}_2)$, we say that $\widehat{\theta}_1$ is more efficient than $\widehat{\theta}_2$ if $\mathsf{var}(\widehat{\theta}_1) < \mathsf{var}(\widehat{\theta}_2)$

Mean squared error

When we judge the reasonableness of an estimator, we often combine both (squared) bias and variance into one measure, the mean squared error

$$\mathrm{MSE} = \mathrm{E}(\widehat{\theta} - \theta^*)^2$$

To see how the center and spread of the sampling distribution come into play, we can write down a little algebra

$$\begin{aligned}
\mathrm{MSE} &= \mathrm{E}(\widehat{\theta} - \theta^*)^2 \\
&= \mathrm{E}(\widehat{\theta} - \mathrm{E}\widehat{\theta} + \mathrm{E}\widehat{\theta} - \theta^*)^2 \\
&= \mathrm{E}(\widehat{\theta} - \mathrm{E}\widehat{\theta})^2 + (\mathrm{E}\widehat{\theta} - \theta^*)^2 + 2(\mathrm{E}\widehat{\theta} - \theta^*)\mathrm{E}(\widehat{\theta} - \mathrm{E}\widehat{\theta}) \\
&= \mathrm{var}(\widehat{\theta}) + \mathrm{bias}^2(\widehat{\theta})
\end{aligned}$$

Properties of estimators

We say that an estimator is consistent if, as n gets large, its sampling distribution concentrates around the true value θ^*

Recall the following from your probability course -- A sequence of random variables $Z_1, Z_2, Z_3, \ldots,$ is said to converge in probability to another random variable Z, written $Z_i \xrightarrow{P} Z$, if, for every $\epsilon > 0$

$$P(|Z_i - Z| > \epsilon) \to 0$$

as $i \to \infty$

Consistency

To apply this, let's write $\widehat{\theta}_n$ to make explicit the fact that our estimate depends on n data points -- We say that $\widehat{\theta}_n$ is consistent if it converges in probability to θ^*(or, rather, a random variable that takes on the value θ^* with probability 1)

It is possible to show that if both the bias and the standard error of an estimate tend to zero as we collect more data (or, rather, the MSE tends to zero as $n \to 0$), then the estimate is consistent

Estimating the point of symmetry

Let's assume that our data $X_1, \ldots, X_n$ were drawn from some univariate distribution f that is symmetric -- It's a pretty weak assumption compared to the parametric forms we've been working with

For a symmetric distribution, we know the theoretical mean and median of f are the same and are the point of symmetry of f -- When given data $X_1, \ldots, X_n$ however, the sample mean and median will not be the same, leading us to ask which is better?

Example: Means and the WLLN

We can establish consistency of the sample mean using the so-called weak law of large numbers: If $Z_1, \ldots, Z_i$ are independent draws from the same distribution having mean μ, the sample mean $\overline{Z} \xrightarrow{P} \mu$ as $i \to \infty$

Example: Medians

Note that the weak law of large numbers implies that the sample mean is a consistent estimate of the “population” mean -- We didn’t have to put a lot of modeling assumptions for this to happen

Now, another good estimate we’ve worked with when looking at the CDC data is the median -- Recall for the normal distribution the mean and median are the same

Let’s consider the consistency of the median -- Assume we have data $X_1, \ldots, X_n$ from some distribution f with median $\widetilde{\mu}$ and let $\widetilde{X}$ denote the sample median

To make things easy, let’s assume also that we have an odd number of points (n odd) so that the sample median is just the (n+1)/2 element in the list of sorted data

Example: Medians

To prove consistency, let's take $\epsilon > 0$ and consider

$$\begin{aligned} \mathrm{P}(\widetilde{\mathrm{X}} - \widetilde{\mu} > \epsilon) &= \mathrm{P}(\widetilde{\mathrm{X}} > \widetilde{\mu} + \epsilon) \\ &= \mathrm{P}(\text{at least } (\mathrm{n}+1)/2 \text{ of the } \mathrm{X}_\mathrm{i}'\mathrm{s} \text{ are bigger than } \widetilde{\mu} + \epsilon) \end{aligned}$$

Let S_n denote the number of sample points $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}$ that are larger than $\widetilde{\mu} + \epsilon$ -- That means S_n has a binomial distribution (n,p) where

$$\mathrm{p} = \mathrm{P}(\mathrm{X}_\mathrm{i} > \widetilde{\mu} + \epsilon) < 0.5$$

Example: Medians

Substituting this into our starting equation (and assuming we have an odd number of samples) we find that

$$\begin{aligned}
P(\widetilde{X} - \widetilde{\mu} > \epsilon) &= P(S_n > (n+1)/2) \\
&= P(S_n - np > (n+1)/2 - np) \\
&= P(S_n - np > n(1/2 - p) + 1/2) \\
&< P(S_n - np > n(1/2 - p)) \\
&< \frac{np(1-p)}{[n(1/2-p)]^2} \\
&\to 0 \quad \text{as } n \to \infty
\end{aligned}$$

where we have invoked Chebychev's inequality -- A similar argument can be formed for $P(\widetilde{X} - \widetilde{\mu} < -\epsilon)$, giving us consistency

Comparing consistent estimators

The sample median is a consistent estimate of the population median and the sample mean is a consistent estimate of the population mean -- When our data-generating distribution is symmetric, that means that both the sample mean and sample median are estimating the same quantity

So for a symmetric f, we have two consistent estimators -- How do we chose between them?

Example: Means and the CLT

Given a sample $X_1, \ldots, X_n$ of independent draws from a distribution with mean μ and standard deviation σ, we know that the sample mean $\overline{X}$ has mean μ and standard deviation $\sigma/\sqrt{n}$

The Central Limit Theorem states that

$$Z_n = \frac{\overline{X} - \mu}{\sqrt{\text{var}(\overline{X})}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \xrightarrow{D} Z$$

where Z has a standard normal (mean zero, standard deviation one) distribution

Example: Means and the CLT

To make this precise (as we had to do with convergence in probability) we say that a sequence of random variables $Z_1, Z_2, \ldots$ converges in distribution to Z if

$$\lim_{n \to \infty} F_n(x) = F(x)$$

where F_n is the CDF of Z_n and F is the CDF of Z, at all points where F_n is continuous

Example: Means and the CLT

The CLT implies that the $\sqrt{n}(\overline{X} - \mu)$ has a normal limiting distribution with mean zero and variance σ^2

What about the median?

Example: Medians

Now, given a sample $X_1, \ldots, X_n$ that come from a distribution f, can be shown that $\sqrt{n}(\widetilde{X} - \widetilde{\mu})$ also has a limiting normal distribution having zero mean but with variance $1/[2f(\widetilde{\mu})]^2$

Suppose our our data come from a normal distribution; that is, suppose f is a gaussian

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\widetilde{\mu})/2\sigma^2}$$

where we have inserted $\widetilde{\mu}$ since the mean and median are the same for this distribution

Therefore, $f(\widetilde{\mu}) = 1/\sqrt{2\pi\sigma^2}$ so $\sqrt{n}(\widetilde{X} - \widetilde{\mu})$ has a limiting normal distribution with mean zero and variance $\pi\sigma^2/2$

Example: Medians

So, if we use the mean to estimate the center of a distribution, we have an asymptotic variance of σ^2 ; if we use the median the asymptotic variance is $1/[2f(\widetilde{\mu})]^2$

In the normal case, the latter expression becomes $\pi\sigma^2/2$; we can then compute the so-called asymptotic relative efficiency between using the median and the mean for data that come from a normal family

$$\frac{\sigma^2}{1/[2f(\widetilde{\mu})]^2} = \frac{2}{\pi} = 0.637$$

This means that if our data really come from a normal distribution, we're better off using the sample mean instead of the sample median

Example: Medians

Now consider a contaminated normal family that's often used in so-called robustness studies; Tukey (1960) considered data generated by the normal mixture

$$f(x) = (1-\epsilon)N(x;0,1) + \epsilon N(x;0,\tau)$$

This family allows one to "contaminate" a standard normal distribution (first component) with some outliers (second component)

If we had observations solely from a normal distribution, then we know the sample mean (the MLE) is an efficient estimate; but if we start to introduce outliers, what happens?

Example: Medians

Given data from the contaminated distribution

$$f(x) = (1-\epsilon)N(x;0,1) + \epsilon N(x;0,\tau)$$

we know that the variance of this mixture is given by $\sigma^2 = (1-\epsilon) + \epsilon\tau^2$; also, the median of this family is 0 so that

$$f(0) = \frac{1}{\sqrt{2\pi}}\left(1 - \epsilon + \frac{\epsilon}{\tau}\right)$$

Therefore, the relative efficiency between the mean and the median is given by

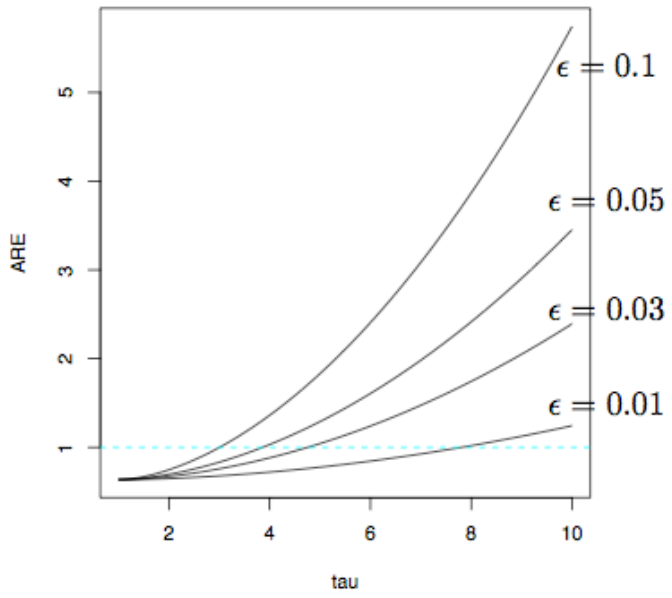
$$\frac{(1-\epsilon)+\epsilon\tau^2}{1/[2f(0)]^2} = \frac{2}{\pi}[(1-\epsilon)+\epsilon\tau^2]\left(1-\epsilon+\frac{\epsilon}{\tau}\right)^2$$

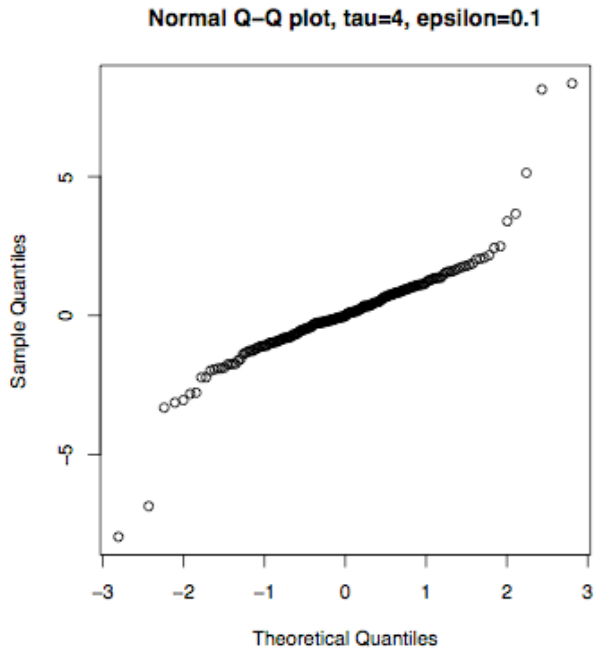
Example: Medians

At the left we have plots of the asymptotic relative efficiency for four values of ϵ and τ ranging from 2 to 10

We also have a Q-Q plot for one member of the family $\epsilon = 0.1, \tau = 4$ that has a relative efficiency of 1.36

In this case, the median outperforms the mean; notice the effect of the observations from normal with greater spread





Example: Medians

With this mixture device, we can see clearly the tradeoff between the mean and the median

Next time we will return to estimation in the context of parametric models and examine the performance of the MLE