Lecture 8: Measure for measure

Last time

We finished our discussion of the A/B tests performed by the New York Times -- Along the way we looked at how transformations can help us “see” data better and discussed different kinds of statistics we might use to test hypotheses with skewed data

We then talked about a randomization technique to assess the effect of preferential treatment due to ballot placement in the context of the 2003 California recall election

Today

We are going to talk about probability -- Every statistics class has to come to grips with that at some point, and today's the day

We will examine the basic axioms of probability, but they will come out by examining different interpretations of probability -- We will see that probability can be thought of in different ways, each with a different approach to inference

OK, buckle up!

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..


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

In lab, we used a simple "model" for how to randomly divide patients into two groups; suppose we write the names of all 107 patients on slips of paper, **toss them into a bag, and pull out 55 names**, these 55 being our treatment group

Reasoning classically, **any group of 55 names is equally likely**, and so our first problem is to figure out how many different groups of size 55 we can form from the list of 107 patients

The answer involves the so-called binomial coefficient, or, for a more romantic flair, Pascal's triangle; to motivate our solution, let's take a slightly simpler problem, where we have only **five patients and we want a treatment group of size three and a control group of size two**

So, how many ways can we form a treatment group of size three from a list of five patients?

Classical probability: Hill's tables

Suppose our patients are named Pascal, Fermat, Laplace, Bernoulli and Huygens (hey, it's classical probability, right?)

It's easier to **first think about forming sequences of names**: We have 5 choices for the first name, 4 for the second and 3 for the third; this gives us $5 \times 4 \times 3 = 60$ different sequences of names

While it's easy to count the number of sequences, it's not quite what we're after; below we list all 60 sequences, what do you notice?

```
BFH BFL BFP BHF BHL BHP BLF BLH BLP BPF BPH BPL
FBH FBL FBP FHB FHL FHP FLB FLH FLP FPB FPH FPL
HBF HBL HBP HFB HFL HFP HLB HLF HLP HPB HPF HPL
LBF LBH LBP LFB LFH LFP LHB LHF LHP LPB LPF LPH
PBF PBH PBL PFB PFH PFL PHB PHF PHL PLB PLF PLH
```

Classical probability: Hill's tables

What we really want are **sets of names, not sequences**, as th
names doesn't matter; repeating our reasoning from the previou
each set, say a treatment group with P, F and B, we have $3 \times$
different possible sequences

Therefore, the number of different groups (or sets) of size 3 tha
from our 5 patients is $(5 \times 4 \times 3)/(3 \times 2 \times 1) = 60/6 = 10$

Using the notation that $n! = n \times (n-1) \times, \cdots \times 2 \times 1$ (whe
we can rewrite this as $5!/3!\,2!$

Classical probability: Hill's tables

What we really want are **sets of names, not sequences**, as the order of the names doesn't matter when we enroll people into treatment or control -- Repeating our reasoning from the previous page, for each set, say a treatment group with P, F and B, we have 3 x 2 x 1 = 6 different sequences from the single set

Therefore, the number of different groups (or sets) of size 3 that we can form from our 5 patients is (5x4x3)/(3x2x1) = 60/6 = 10

Using the notation that n! = n x (n-1) x ... x 2 x 1 (where we define 0! = 1), we can rewrite 5 x 4 x 3 = (5 x 4 x 3 x 2 x 1)/(2 x 1) = 5!/2! so that the number of sets of size 3 we can form from 5 patients is 5!/(3! 2!)

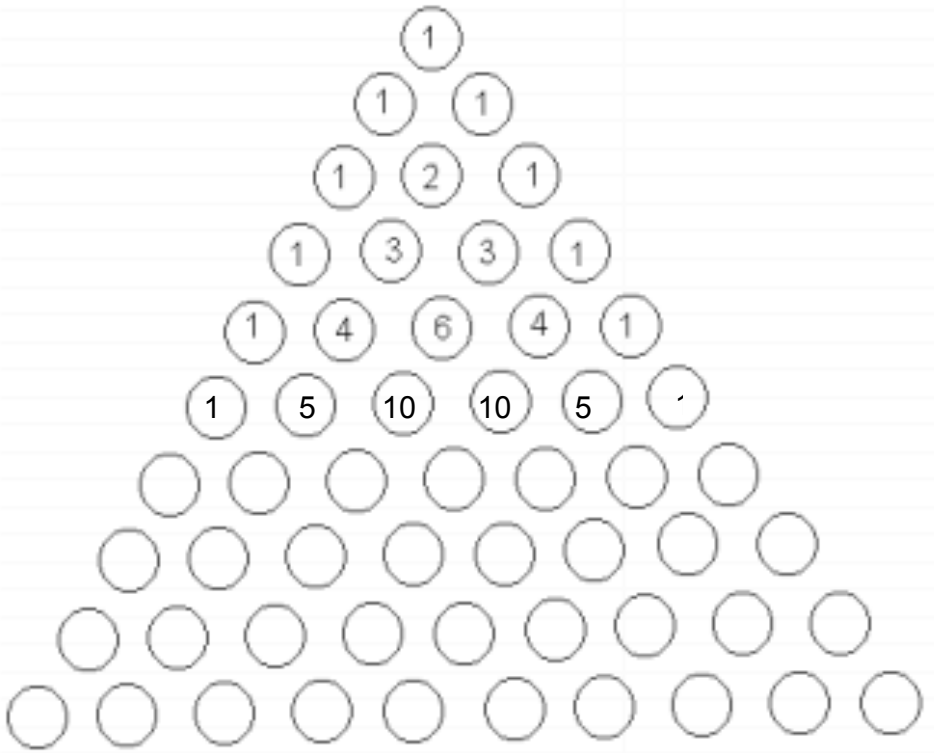
Classical probability: Hill's tables

In general, the number of ways we can select k items from a group of size n is given by the so-called **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

The stacked-notation here is read "choose" as in "n choose k"; the answer can also be read from Pascal's famous triangle

Recall, that the entry in each circle is obtained by adding the values if it's two parents, the two circles in the row above it, just to the right and left



Classical probability: Hill's tables

With this piece of technology in hand, we can go back to our counting; there are a total of

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

different ways that Hill's patients can be divided into treatment and control

But this grouping is not the end of the story; **remember that our interest is in the resulting tables**: Of those 107 patients, 18 died; for each of the divisions into treatment and control, how many of those 18 end up receiving Streptomycin and how many get just bed rest?

Classical probability: Hill's tables

We'll answer this problem in general, but why don't we start by asking how many groups of 55 duplicate Hill's observed table; that is, how many of the zillions of groups have just four of the doomed patients?

To answer that we can again start counting; first off, there are $\binom{18}{4}$ ways to select 4 of the 18 doomed patients and $\binom{89}{51}$ ways to select the remaining 55-4 = 51 patients from the 89 that survived; each of these two parts can be mixed and matched to create a total of

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

different groups that exactly produce Hill's tables

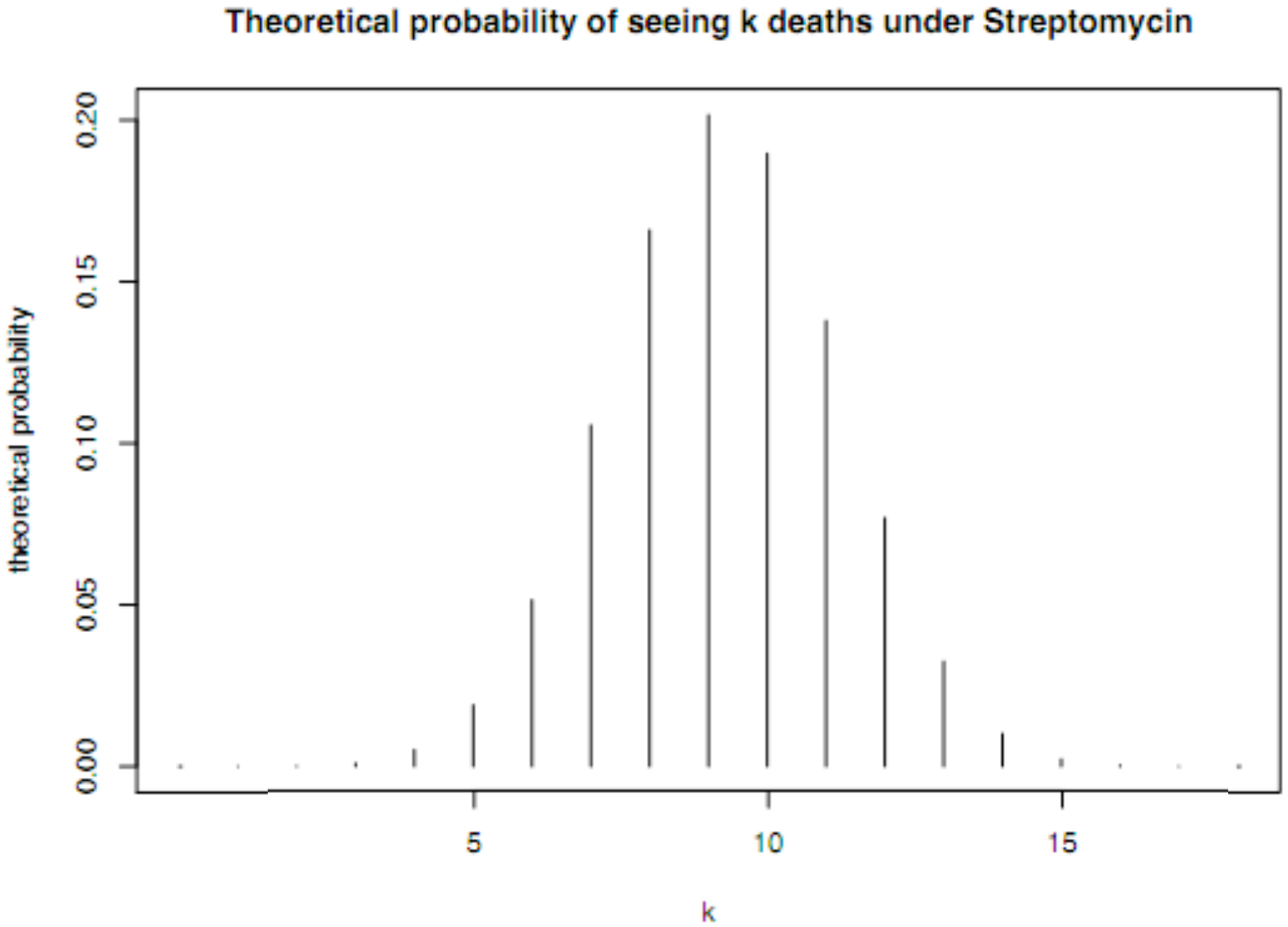
Classical probability: Hill's tables

Therefore, following the classical approach, we take **the number of outcomes we're interested in** (assignments of 55 out of 107 patients to the Streptomycin group that include 4 deaths) **and divide by the total number of outcomes** (the total number of ways we can form a treatment group of size 55 from 107 patients)

Numerically, this works out to $\frac{\binom{18}{4}\binom{89}{51}}{\binom{107}{55}} \approx 0.0052$

We can repeat this for any number, not just four, and the formula for the general probability of seeing k deaths in the Streptomycin group is

$$\frac{\binom{18}{k}\binom{89}{55-k}}{\binom{107}{55}}$$


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

Classical probability: Hill's tables

OK so that seemed like a lot of work, but here's the take away

1. **The style of reasoning:** This classical approach to probability hinges on starting with outcomes that are "equipossible"

2. **The relationship with combinatorics:** All of our calculations involved asking how many arrangements were possible

3. **The link with simulation:** In our previous lectures, we depended on the computer to perform a series of randomizations and we worked with the proportion of simulated tables having the characteristics we wanted

Classical probability

Your book (the chapter Probability) covers these models in some detail and it's worth going over that material -- In particular, read the derivation of the Binomial distribution

We will cover the Binomial in a lab at some point as well...

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 $\mathrm{P}(\mathrm{A} \text{ or } \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$

These properties should be obvious from how we treated our "classical" examples

Some axioms

For example, if we were to think about tossing a single (six-sided) die, we could take a classical approach and assume that each side appears with probability 1/6 -- We could then talk about the probability of an even number of dots showing as being

$$P(\text{2 spots or 4 spots or 6 spots}) = P(\text{2 spots}) + P(\text{4 spots}) + P(\text{6 spots})$$

because the three events are disjoint -- Since each has probability 1/6, we would compute the chance of tossing an even number of spots as 3/6 = 1/2

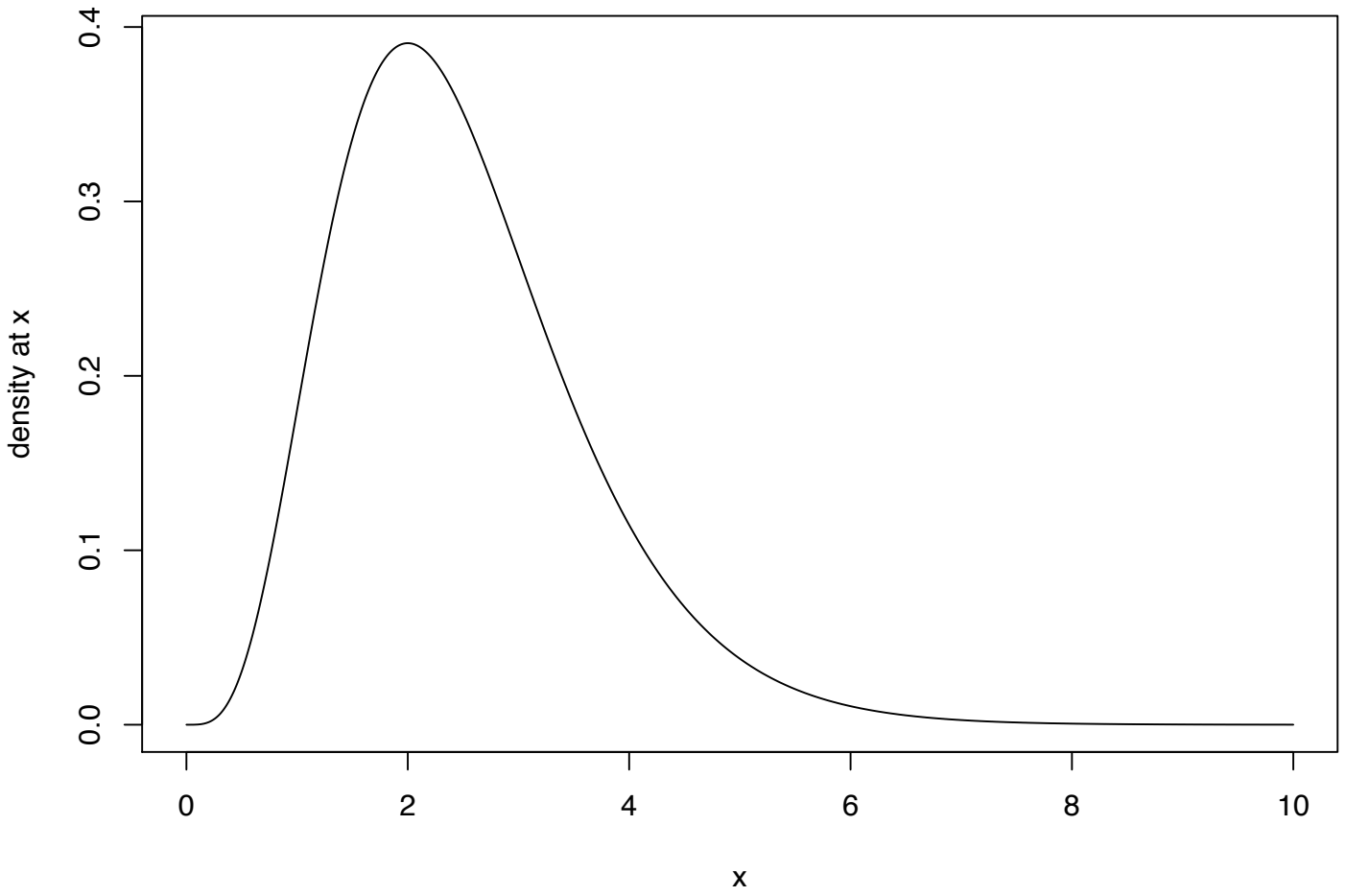
Again, the combinatoric approach lurking behind classical probability computations makes the rules on the previous page pretty clear

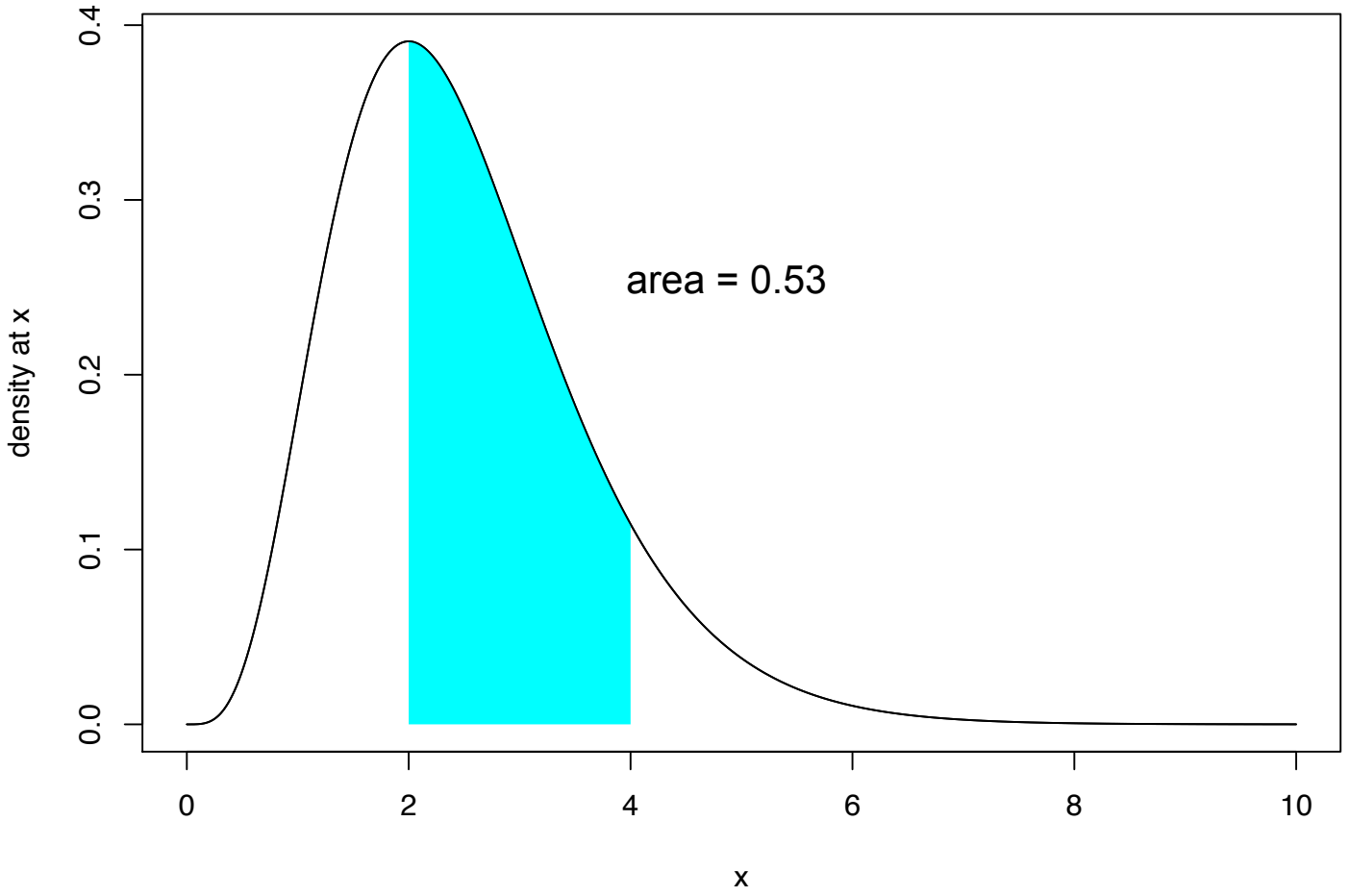
Continuous probability functions

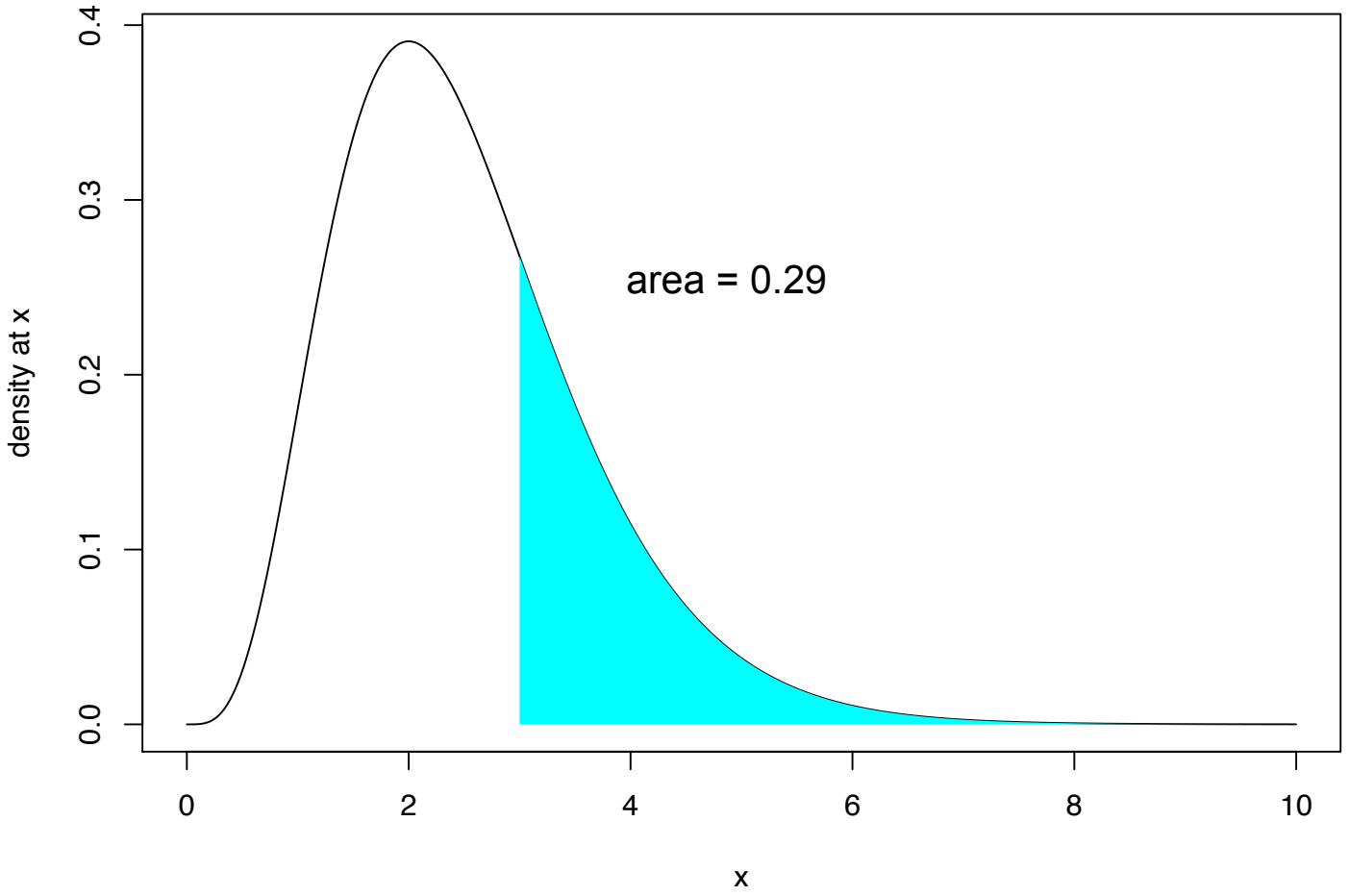
So far, we have worked with discrete events -- By that I mean we have 52 cards or 6 sides of a die or two heads of a coin and we are counting

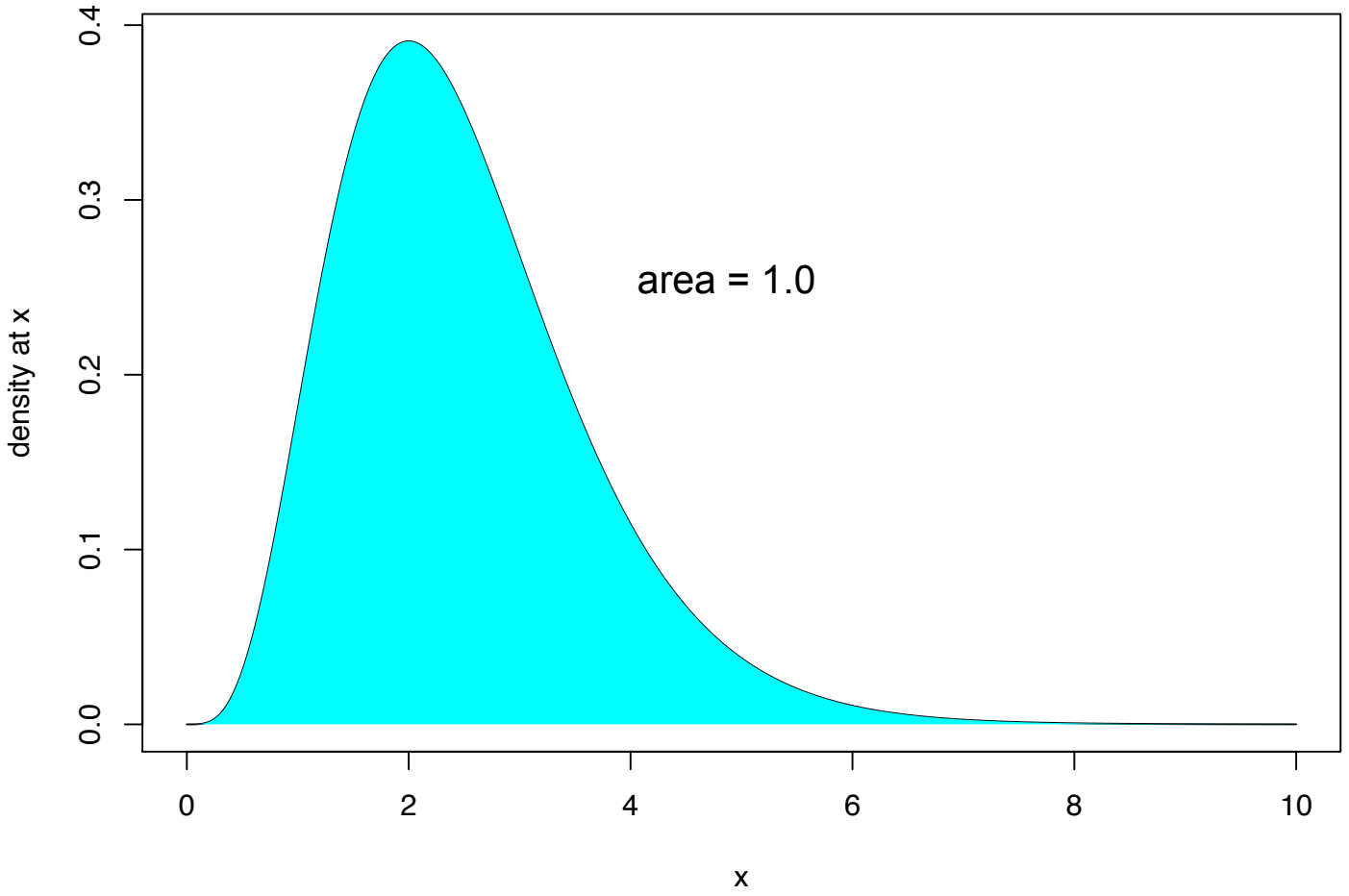
In many cases, we can observe, instead, **a continuous range of values** -- We have already seen this in our data analysis exercises if we talk about someone's height or weight (these can, in principle, be measured very accurately, producing data with many significant digits)

In these cases, our probability is specified not by **a function with discrete input**s (like P(H) = P(T) = 0.5) but instead **by a curve --** Probabilities of ranges are expressed as areas under the curve


density at x
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
x


density at x
0.0
0.1
0.2
0.3
0.4
area = 0.53
0
2
4
6
8
10
x


density at x
0.0
0.1
0.2
0.3
0.4
area = 0.29
0
2
4
6
8
10
x


density at x
0.0
0.1
0.2
0.3
0.4
area = 1.0
0
2
4
6
8
10
x

Continuous probability functions

We'll come back to this idea later in the quarter when we encounter the mother of all continuous distributions, the normal (or bell shaped) family -- For now, this concept is sufficient to motivate how to work with this object

Let's return to our overview of classical probability...

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

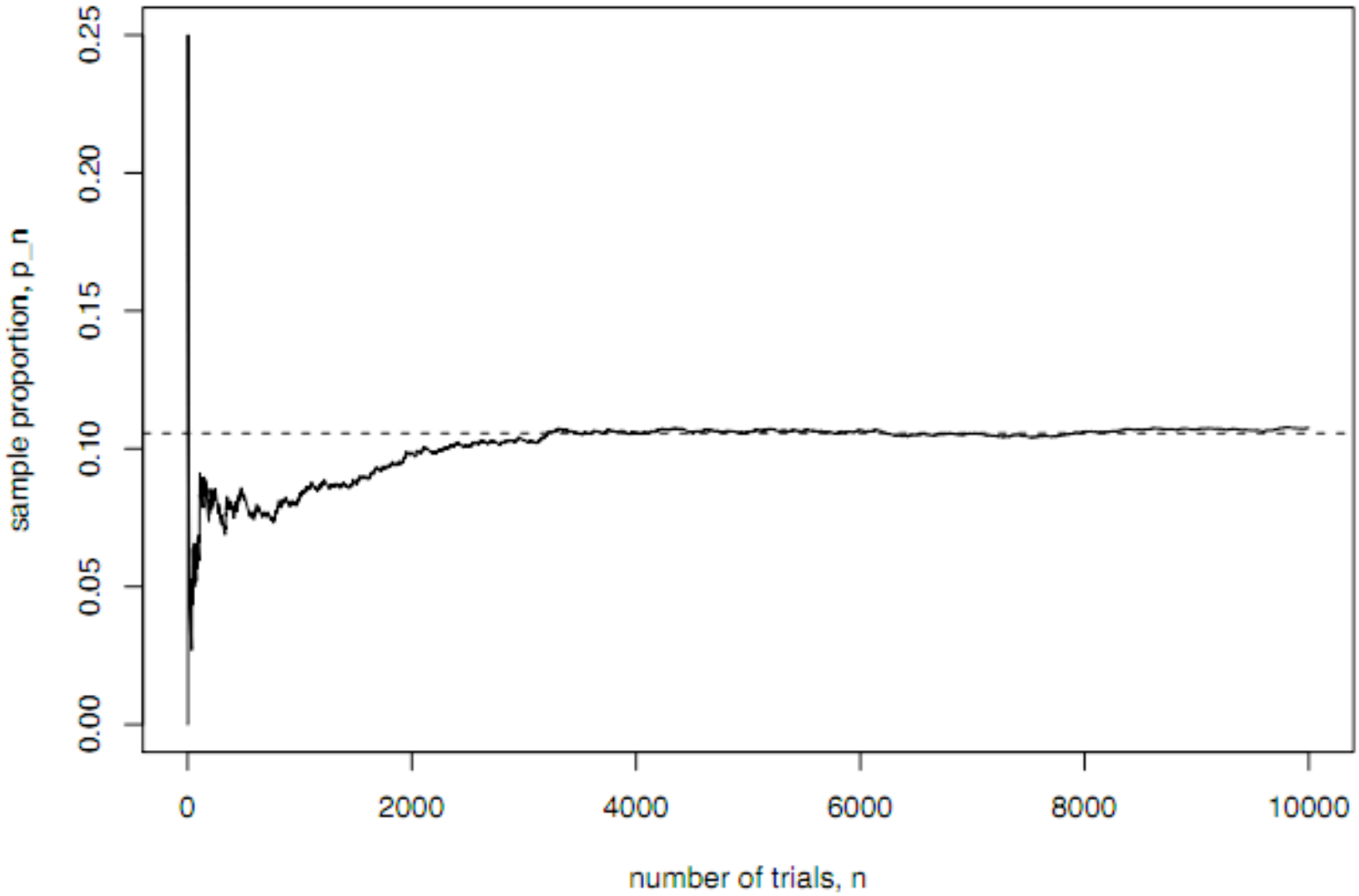
The law of large numbers

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**

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)

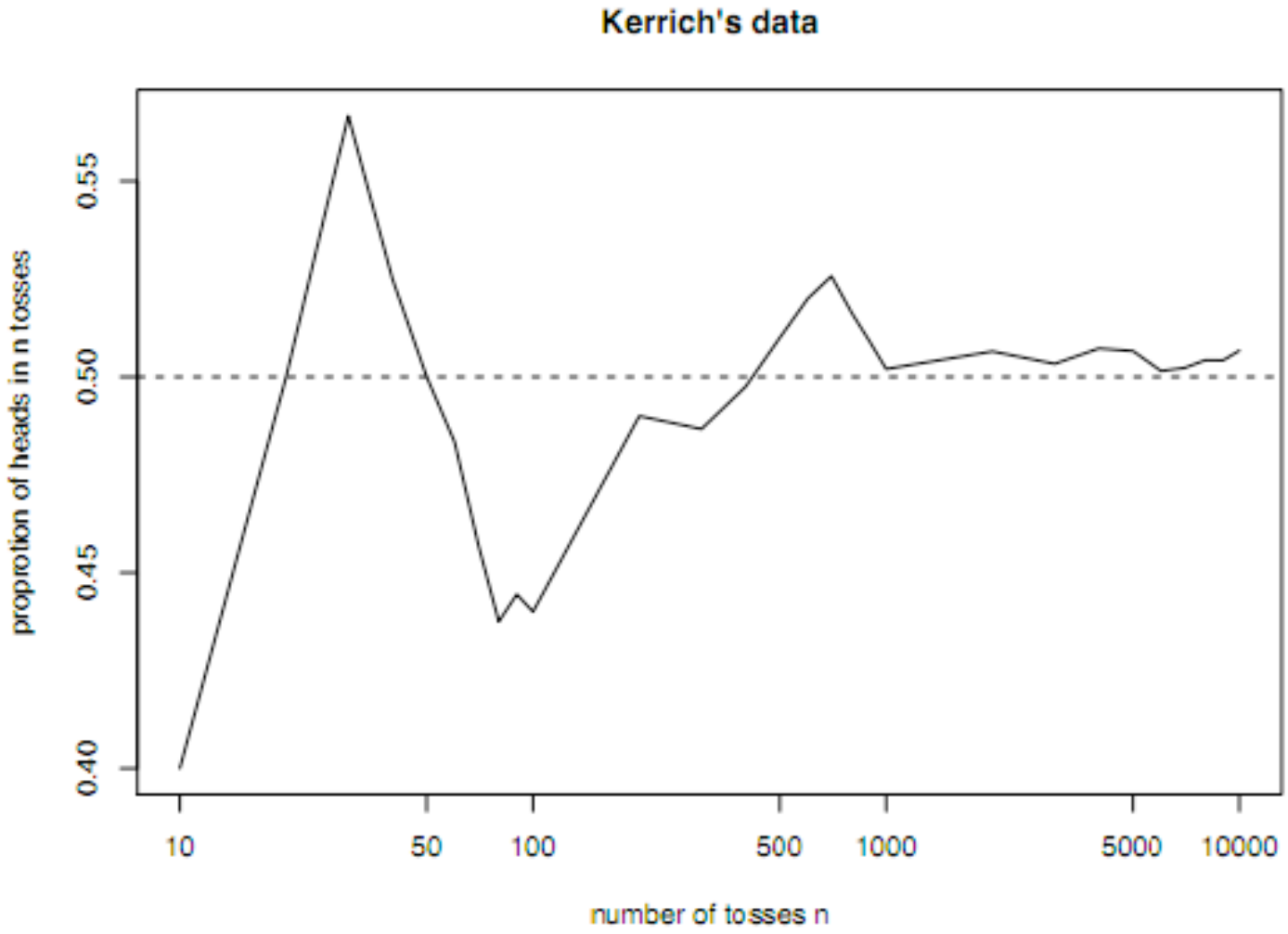
English economist W S Jevons (1835-1882) made 2048 throws of ten coins at a time; in all, he saw 20,480 tosses out of which 10,353 were heads (relative frequency of 0.506, although this is not quite the same kind of trial)

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**

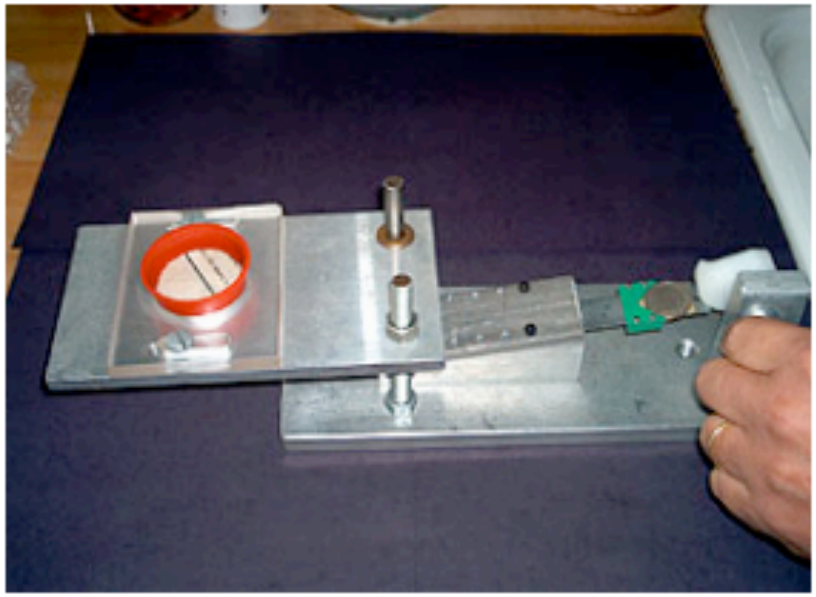
In 1850 **the Swiss astronomer Wolf** rolled one white and one red die 20,000 times- - Keynes writes “the relative frequency of the different combinations was very different from what theory would predict... an explanation is easily found... **the dice must have been very irregular**...” and he concludes “**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...

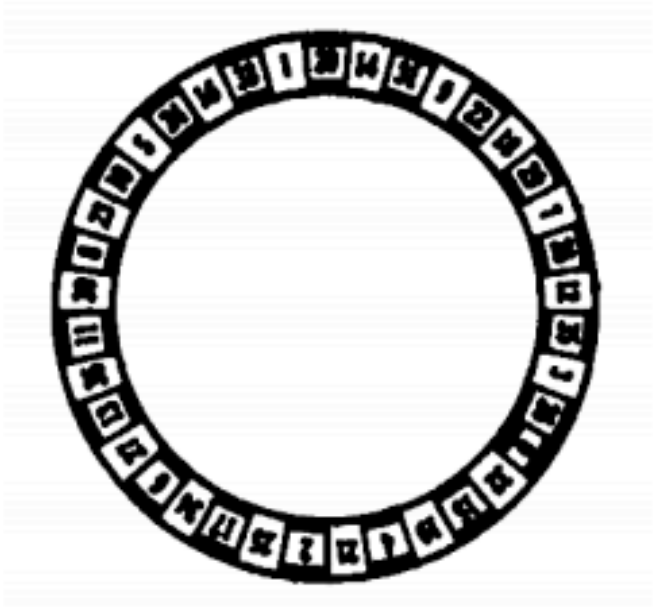
A device that will consistently flip a coin the same way; this machine was made for Persi Diaconis, a well-known Stanford statistician

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

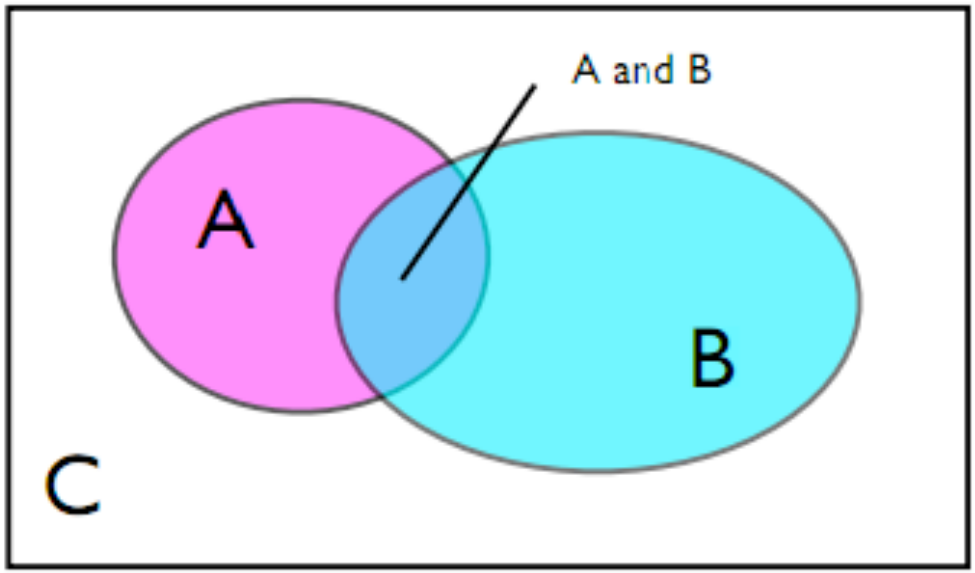
The third view of probability we will talk about is more in line with Hacking's use of the term probability as a "relation between a hypothesis and the evidence for it"

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.

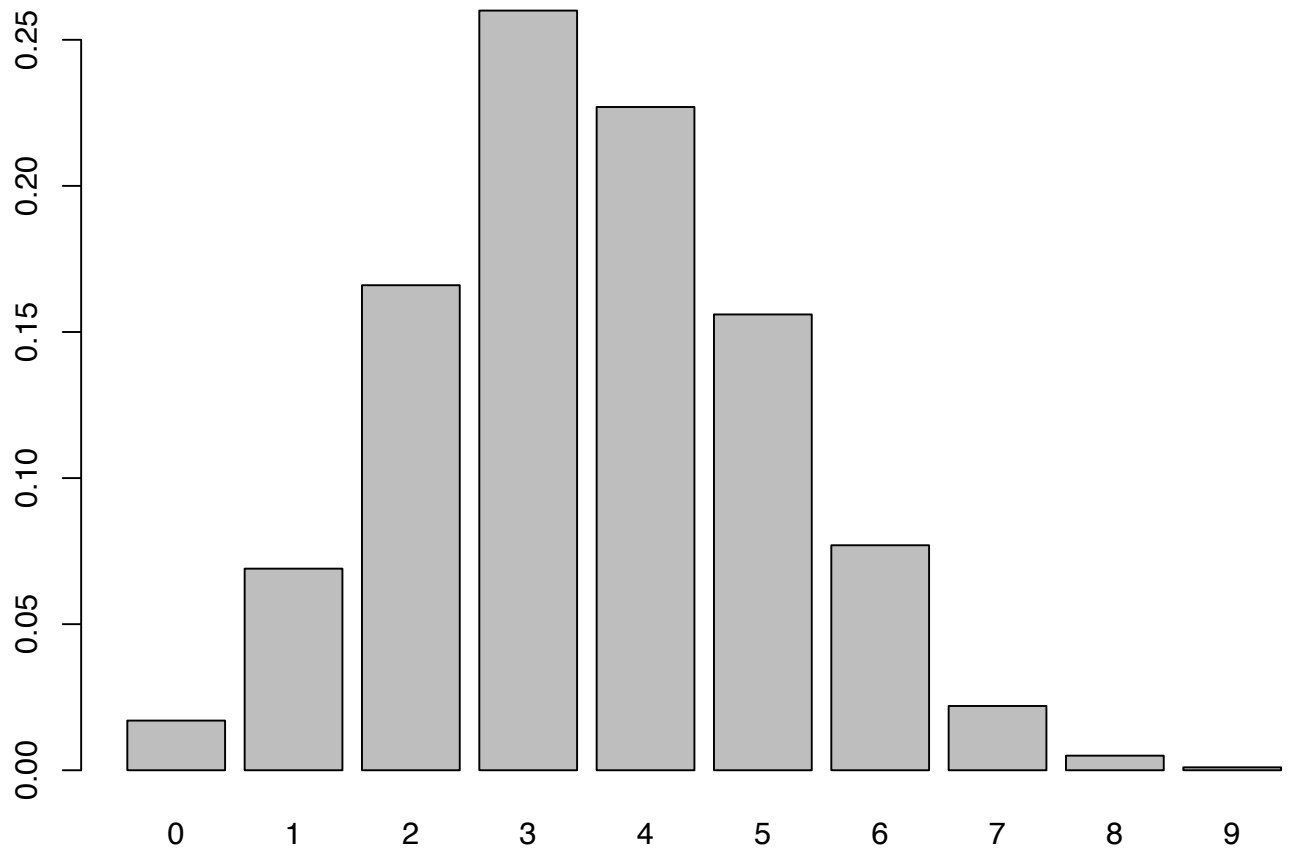
In general, the subjective view provides an elegant way to draw conclusions from data; it involves fundamentally different kinds of inferences than we will be making in this class... although maybe one day toward the end of the quarter we'll take a look at what this has to offer

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

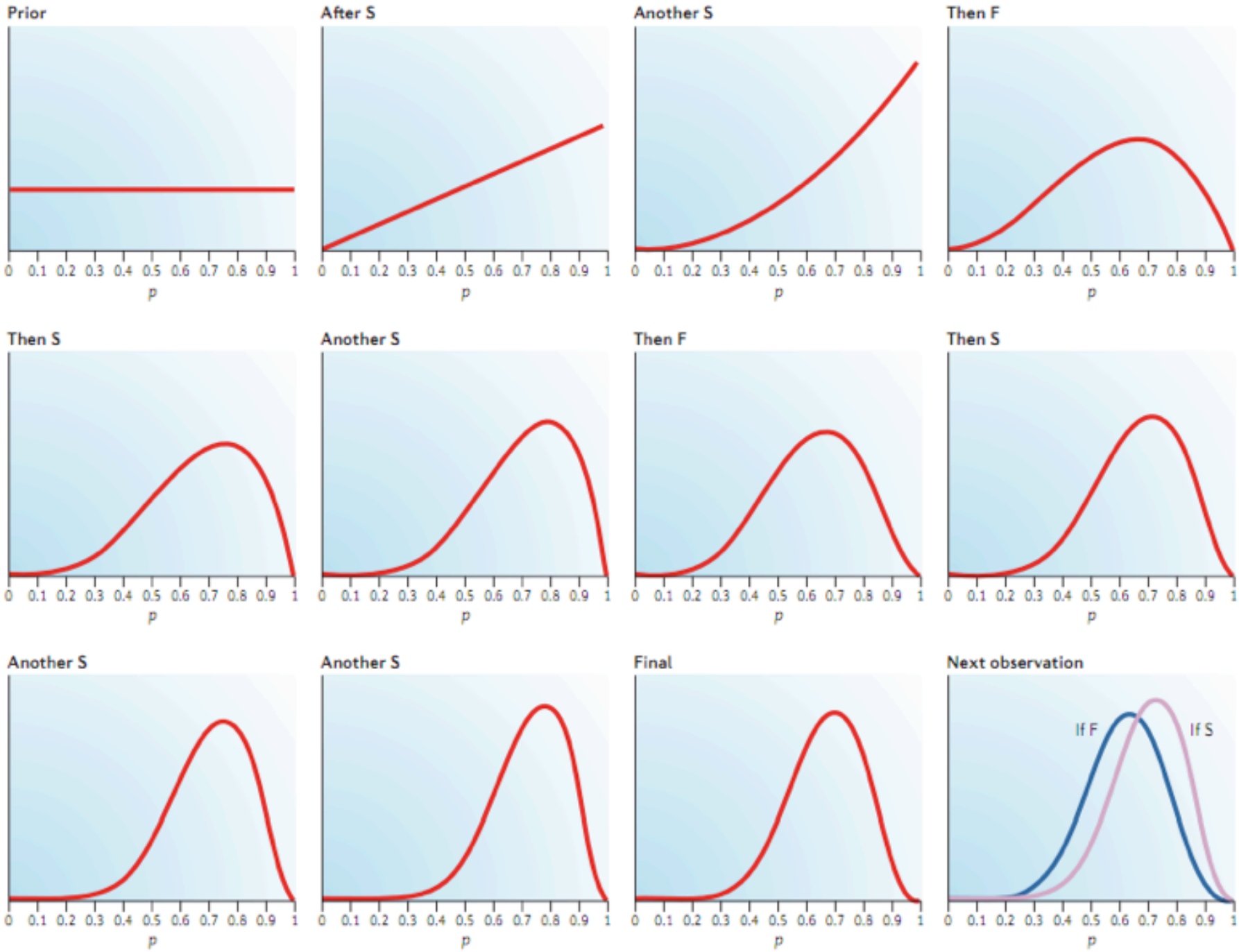
In this case, we sum up 0.021+0.004+0.001 = 0.026 to compute our P-value and we would reject the null hypothesis at the 0.05 level -- All pretty standard at this point

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

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

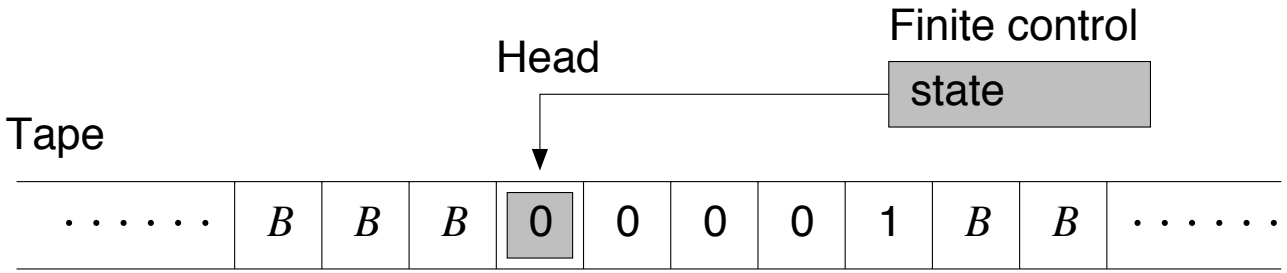
we might require a longer program

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”