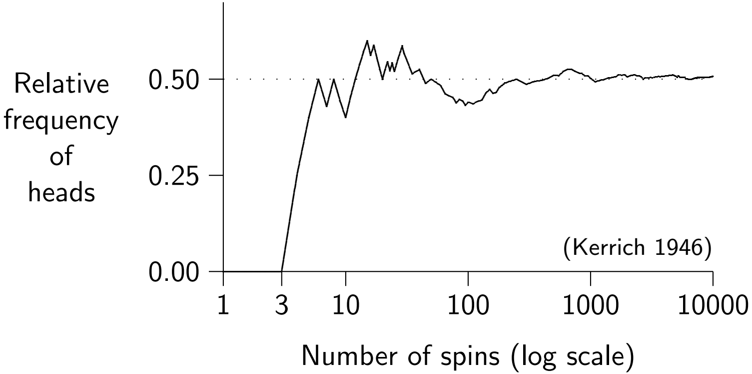
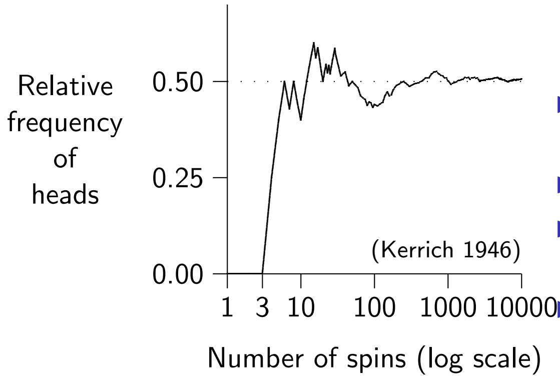
Probability

Alan R. Rogers

August 7, 2019


Relative frequency of heads
0.50
0.25
0.00
(Kerrich 1946)
1
3
10
100
1000
10000
Number of spins (log scale)

Probability and relative frequency in repeated trials



- rel. freq. of heads gradually approaches limiting value.
- Limiting value is the *probability* of heads
- Need not equal 1/2.
- We estimate probabilities from relative frequencies.
- We never know them exactly.

Kerrich's "urn" experiment

$$\left(\underline{\circ \quad \circ \quad \bullet \quad \bullet}\right)$$

- Urn contains 4 balls: 2 black and 2 white
- Mix them up.
- Draw one at random
- Draw a second *without* replacing first.
- Repeat 5000 times.

Results from Kerrich's urn experiment

First ball	Second ball: Black	Second ball: White	sum
Black	756	1689	2445
White	1688	867	2555
sum	2444	2556	5000

- If 1st ball is B, 2nd is likely to be W
- And vice versa

Model of Kerrich's urn experiment

Assumption: we are equally likely to draw any ball in urn.

1st Ball

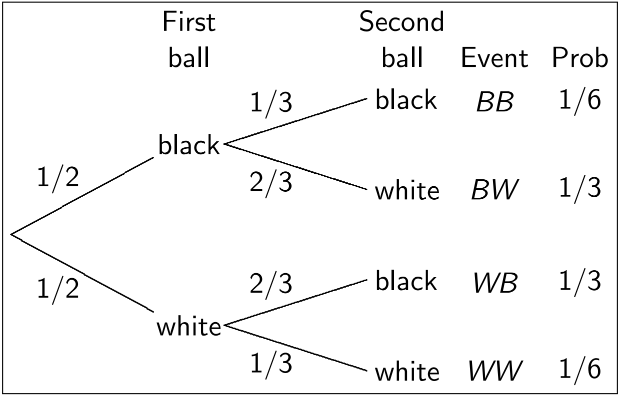
$$(\circ\ \circ\ \bullet\ \bullet)$$

We are equally likely to draw black or white

2nd Ball

First ball	Remaining balls	Prob. of black
$\bullet$	$(\circ\ \circ\ \bullet)$	$1/3$
$\circ$	$(\circ\ \bullet\ \bullet)$	$2/3$

2nd ball usually black if 1st was white, and vice versa.



Tree diagram for urn model

Kerrich's urn experiment: model versus data

Event	Theoretical probability	Observed relative frequency
BB	0.167	0.151
BW	0.333	0.338
WB	0.333	0.338
WW	0.167	0.173

Theory and observation are not identical, but they are close.

Why do we multiply along branches?

Conditional probability

- What is the conditional probability that the 2nd ball is white given that the first was black?
- $2/3$.
- Called a *conditional probability* and written

$$\Pr[\text{2nd ball white} | \text{1st one black}].$$

- "$|$" is pronounced "given."

Conditional relative frequencies

First ball	Second ball Black	White	sum
Black	**756**	$\boxed{\mathbf{1689}}$	**2445**
White	1688	867	2555
sum	2444	2556	5000

- On trials where the 1st ball was black, how often was the 2nd white?
- A fraction 1689/2445 of the time, or ≈ 0.69.

This is a conditional relative frequency. If the number of trials is large, this approximates a conditional probability.

The results of 20,000 throws with two dice (Wolf 1850, cited in Bulmer 1967)

	White							
Black	1	2	3	4	5	6	$\sum$	f
1	547	587	500	462	621	690	3407	.170
2	609	655	497	535	651	**684**	**3631**	.182
3	514	540	468	438	587	629	3176	.159
4	462	507	414	413	509	611	2916	.146
5	551	562	499	506	658	672	3448	.172
6	563	598	519	487	609	646	3422	.171
$\sum$:	3246	3449	2897	2841	3635	3932	20000	1.000
f:	.162	.172	.145	.142	.182	.197	1.000	

- What is the conditional frequency of $W6$ given $B2$?
- $684/3631 \approx 0.188$

Product rule for relative frequencies

How often did Kerrich get $B1$ and $W2$?

First ball	Second ball: Black	Second ball: White	sum
Black	756	$\boxed{\mathbf{1689}}$	2445
White	1688	867	2555
sum	2444	2556	**5000**

A fraction 1689/5000 of the time.

$$\frac{1689}{5000} = \frac{1689}{2445} \times \frac{2445}{5000}$$

$$\overbrace{\frac{1689}{5000}}^{f(B1\ \&\ W2)} = \overbrace{\frac{1689}{2545}}^{f(W2|B1)} \times \overbrace{\frac{2445}{5000}}^{f(B1)}$$

As N increases, relative frequencies (f) become probabilities.

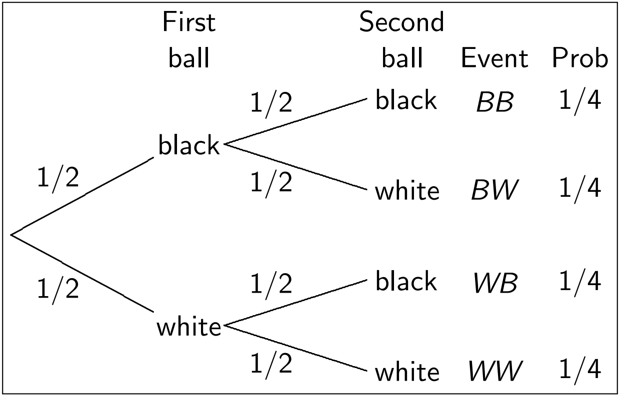
Product rule

The probability of A and B is

$$\Pr[A \,\&\, B] = \Pr[B|A]\,\Pr[A]$$

This is why we multiply along the branches of a tree diagram.

Statistical independence: sampling w/ replacement



$$\Pr[W_2|B_1] = \Pr[W_2|W_1] = \Pr[W_2] = 1/2$$

Sampling with replacement: model versus data

Event	Theoretical probability	Observed relative frequency
BB	0.25	0.254
BW	0.25	0.255
WB	0.25	0.252
WW	0.25	0.239

Data from computer simulation

Theory and observation are not identical, but they are very close.

Sum rule: Pr[black 4 or white 5 (or both)]

	White						
Black	1	2	3	4	5	6	$\sum$
1	547	587	500	462	**621**	690	3407
2	609	655	497	535	**651**	684	3631
3	514	540	468	438	**587**	629	3176
4	**462**	**507**	**414**	**413**	$\boxed{\mathbf{509}}$	**611**	2916
5	551	562	499	506	**658**	672	3448
6	563	598	519	487	**609**	646	3422
$\sum$:	3246	3449	2897	2841	3635	3932	*20000*

Relative frequency is the sum of the bold-face values divided by 20,000.

$$f[b4 \text{ or } w5] = \overbrace{\frac{2916}{20000}}^{f[b4]} + \overbrace{\frac{3635}{20000}}^{f[w5]} - \overbrace{\frac{509}{20000}}^{f[b4\ \&\ w5]}$$

Sum rule for probabilities

$$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A \;\&\; B]$$

Sum rule again: Pr[white 3 or white 5]

For mutually exclusive events, there is nothing to subtract.

	White						
Black	1	2	3	4	5	6	$\sum$
1	547	587	**500**	462	**621**	690	3407
2	609	655	**497**	535	**651**	684	3631
3	514	540	**468**	438	**587**	629	3176
4	462	507	**414**	413	**509**	611	2916
5	551	562	**499**	506	**658**	672	3448
6	563	598	**519**	487	**609**	646	3422
$\sum$:	3246	3449	2897	2841	3635	3932	*20000*

What is rel. frq. of white 3 or white 5?

$$f[w4 \text{ or } w5] = \overbrace{\frac{2897}{20000}}^{f[w4]} + \overbrace{\frac{3635}{20000}}^{f[w5]}$$

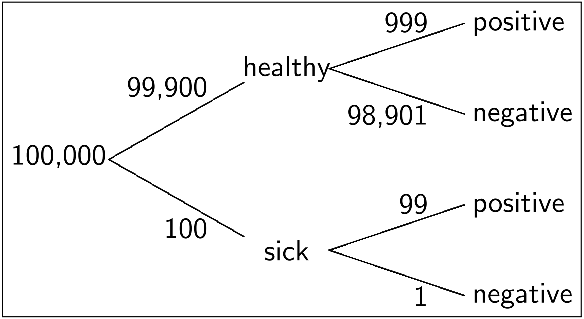
Bayes's rule

Problem: Our emphasis has been on the probability of an outcome given a hypothesis. But we often want to know the probability of the hypothesis, given the outcome.

Example: The probability the patient is sick given a positive result on some test.

Suppose that 0.1% of people have some disease. When tested for the disease 99% of sick people test positive, but so do 1% of well people. What fraction of those with positive results are really sick?

Bayes's rule in terms of counts



What fraction of those who test positive are really sick?

$$\frac{99}{99+999} \approx 0.09 \qquad \text{Fewer than 1 in 10!}$$

Bayes's rule in terms of probabilities

Recall the multiplication law:

$$\Pr[A \& B] = \Pr[B]\Pr[A|B] = \Pr[A]\Pr[B|A]$$

Divide through by $\Pr[B]$:

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]} \quad \text{(Bayes's rule)}$$

Allows us to calculate $\Pr[A|B]$ from $\Pr[B|A]$.

Back to example

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]} \quad \text{(Bayes's rule)}$$

A: patient is sick. $\Pr[A] = 1/1000$.

B: patient tested positive.
$\Pr[B] = (999 + 99)/100000 = 1098/100000$.

$\Pr[\text{testing positive if sick}]$ is $\Pr[B|A] = 99/100$.

Using Bayes's rule,

$$\Pr[A|B] = \frac{1/1000 \times 99/100}{1098/100000} = \frac{99}{1098} \approx 0.09$$

This is the same answer we got using counts.

Summary

Sum rule

$$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A \,\&\, B]$$

Product rule

$$\Pr[A \,\&\, B] = \Pr[A]\Pr[B|A]$$

Bayes's rule

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]}$$