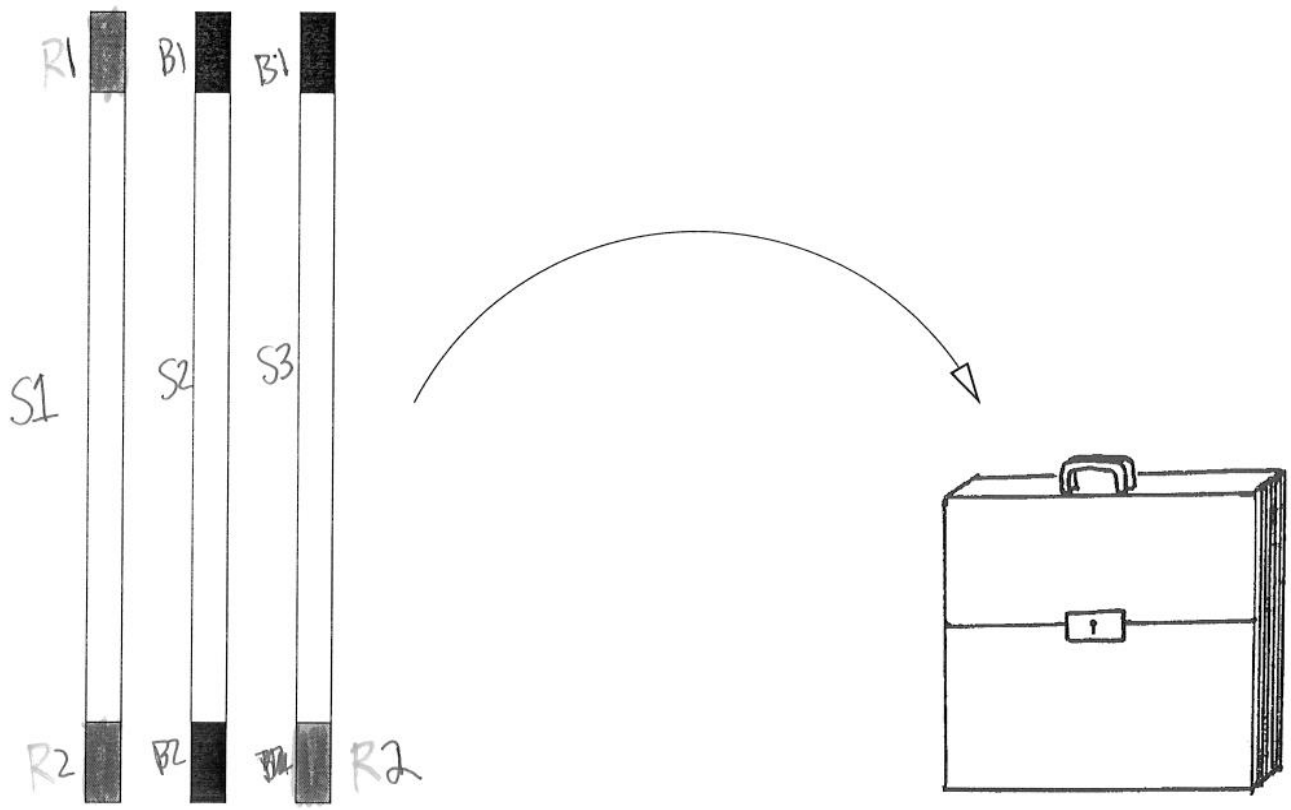
Take three sticks which have their ends colored and place them in a bag. The first stick has two red ends, the second has two black ends and the third stick has a red and a black end.



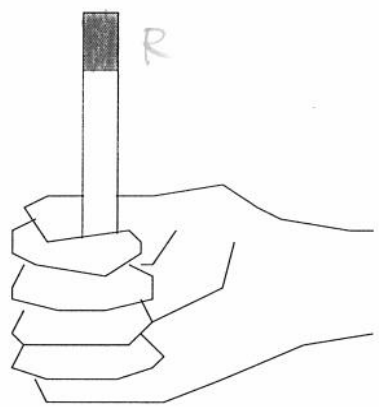
Take three sticks which have their ends colored and place them in a bag. The first stick has two red ends, the second has two black ends and the third stick has a red and a black end.

Now, reach into this bag (no peeking) and grasp one of the sticks by an end so that the other end is showing and pull the stick out. Say that a red end is showing.

What color is most likely clasped in your fist?

Is the answer?

A) red
B) black
C) red/black are equally likely
D) don't know/care



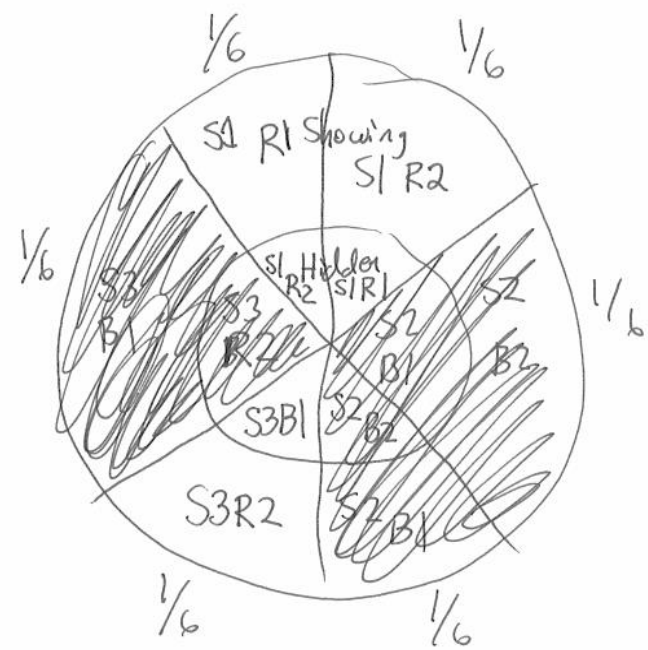


TABLE OF ENGLISH BILETTER CONDITIONAL PROBABILITIES

$$P(\text{next letter} = H \mid \text{prev letter} = C) = \frac{1277}{10{,}000}$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	11	193	388	469	20	100	233	20	480	20	103	1052	281	1878	8	222	0	1180	1001	1574	137	212	57	26	312	23
B	932	57	16	8	3220	0	0	0	605	57	0	1243	49	0	965	0	0	662	229	49	727	16	0	0	1165	0
C	1202	0	196	4	1707	0	0	1277	761	0	324	369	15	11	2283	0	4	426	87	893	347	0	0	0	94	0
D	1044	20	26	218	3778	7	132	7	1803	33	0	125	178	53	733	0	7	324	495	13	601	99	40	0	264	0
E	660	36	433	1195	438	142	125	21	158	5	36	456	340	1382	40	192	34	1927	1231	404	48	215	205	152	121	4
F	838	0	0	0	1283	924	0	0	1608	0	0	299	9	9	2788	0	0	1215	26	496	462	0	0	0	43	0
G	1078	0	0	18	2393	0	177	1281	839	0	0	203	27	451	1140	0	0	1325	256	247	512	0	0	0	53	0
H	1770	5	14	8	5624	0	0	5	1168	0	0	16	16	38	786	0	0	153	27	233	85	0	11	0	41	0
I	380	82	767	459	437	129	280	2	16	0	50	567	297	2497	893	100	8	342	1194	1135	11	250	0	23	2	79
J	1259	0	0	0	1818	0	0	0	350	0	0	0	0	0	3147	0	0	70	0	0	3356	0	0	0	0	0
K	395	28	0	28	5283	28	0	198	1582	0	113	198	28	565	198	0	0	85	1102	28	28	0	0	0	113	0
L	1342	19	22	736	1918	105	108	0	1521	0	79	1413	82	4	778	41	0	34	389	254	269	56	11	0	819	0
M	1823	337	26	0	2976	10	0	0	1345	0	0	10	654	42	1246	722	0	26	244	5	337	5	0	0	192	0
N	550	4	621	1681	1212	102	1391	13	665	9	66	73	104	194	528	4	7	11	751	1641	124	68	18	2	157	4
O	85	101	162	231	37	1299	82	25	92	14	78	416	706	2191	222	292	0	1531	357	396	947	334	345	12	41	4
P	1358	0	6	0	1747	0	0	237	423	0	0	812	73	6	1511	581	0	2305	180	287	457	0	0	0	17	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10000	0	0	0	0	0
R	1026	33	172	282	2795	31	175	17	1181	0	205	164	303	325	1114	55	0	212	655	596	192	142	17	2	306	0
S	604	12	284	27	1795	24	0	561	1177	0	91	145	112	21	706	386	9	27	836	2484	579	0	39	0	81	0
T	619	3	36	2	1417	7	2	3511	1406	0	0	101	44	15	1228	3	0	479	418	213	195	5	88	0	203	5
U	344	415	491	243	434	52	382	10	258	0	14	1097	329	1518	19	386	0	1460	1221	1255	29	14	0	10	14	5
V	749	0	0	23	6013	0	0	0	2568	0	0	0	12	0	530	0	0	0	23	0	12	12	0	0	58	0
W	2290	8	0	32	1942	0	0	1422	2104	0	0	41	0	357	1292	0	0	106	366	16	0	0	0	0	24	0
X	672	0	1119	0	1269	0	0	75	1119	0	0	0	75	0	75	3507	0	0	0	1716	0	0	0	373	0	0
Y	586	34	103	69	2898	0	0	0	691	0	34	172	379	172	2208	310	0	310	1518	172	138	0	103	0	69	34
Z	2278	0	0	0	4557	0	0	0	2152	0	0	127	0	0	506	0	0	0	0	0	127	0	0	0	0	253

To compute this table a typical example of english text was chosen. Then the first row of the table was obtained by recording, for each of 10.000 occurrences of the letter A, the letter that immediately followed it. Thus the entry 469 in the column indexed by D means that in this sample of 10.000 occurrences of the letter A the letter D was observed to immediately follow A exatly 469 times. The same procedure was repeated for each of the letters of the alphabet. We see that in 10.000 occcurrences of Q, the letter U followed it all 10.000 times (not surprising!).
To get the conditional probability P[next lett=Y I prec lett = X] you simply look at the entry in row X and column Y and divide by 10.000.

Thus

P[next letter=E I preceeding letter = R] = 2795/10.000= .27

2. Suppose we have a computer program which generates words from the alphabet A,B,C,D according to the following procedure:

Pick the first letter according to the single frequency table given below then constructs each additional letter using the table of conditional biletter frequencies given below.

a) Calculate the probability that the program produces the word "DACB"

b) Determine the 2 letter word that has the highest probability.

Single letter table

A	10
B	9
C	12
D	9

Biletter table

	A	B	C	D
A	0	3	1	0
B	2	1	4	0
C	0	2	0	3
D	2	1	0	1

$$P(\text{word}=DACB) = P(\text{first}=D)\cdot P(\text{second}=A \mid \text{first}=D)\cdot P(\text{third}=C \mid \cancel{\text{first}=D},\ \text{second}=A)$$

$$\cdot P(\text{fourth}=B \mid \cancel{\text{first}=D},\ \cancel{\text{second}=A},\ \text{third}=C)$$

$$= \frac{9}{40}\cdot\frac{2}{4}\cdot\frac{1}{4}\cdot\frac{2}{5} = \frac{9}{800}$$

$$P\left(\begin{matrix}\text{2 letter}\\ \text{word}\end{matrix} = CD\right) = \frac{12}{40}\cdot\frac{3}{5} = \cancel{\frac{36}{200}} \qquad P(\text{word}=BC) = \frac{9}{40}\cdot\frac{\cancel{4}}{7} = \cancel{\frac{36}{280}}$$

$$P(\text{word}=AB) = \frac{10}{40}\cdot\frac{3}{4} = \frac{30}{160} \qquad P(\text{word}=DA) = \frac{9}{40}\cdot\frac{2}{4} = \cancel{\frac{9}{80}}$$

The index of coincidence is defined as

$$I_c = \frac{\text{number of pairs of equal letters in ciphertext}}{\text{the total number of pairs of letters}}$$

That is if we set

ABBAABBCAB $\frac{N!}{K!(N-K)!} = \binom{N}{K}$ = # of ways of choosing K things from a set of size N

- N_α = the number of occurrences of the letter α in the cyphertext

$\binom{N_\alpha}{2} = \frac{N_\alpha(N_\alpha - 1)}{2}$ = # of pairs of letters that $= \alpha$

- $$D_c = \sum_{\alpha=A}^{Z} \binom{N_\alpha}{2}$$

  D_c represents the number of pairs of equal letters in the cyphertext.

- then $I_c = \frac{D_c}{\binom{N}{2}}$
- where N = the number of letters in the cyphertext

The index of coincidence is invariant under monoalphabetic cyphers and we estimate under this condition that $N_\alpha = N * p_{\sigma(\alpha)}$ for some permutation of the alphabet σ and so

Where p_α = probability that α occurs in English.

$$
\begin{aligned}
I_c &= \frac{\sum_{\alpha=A}^{Z}(N_\alpha^2 - N_\alpha)}{N(N-1)} \\
&\approx \frac{N^2(\sum_{\alpha=A}^{Z} p_\alpha^2) - N}{N(N-1)} \\
&= \frac{N(.065) - 1}{N-1} \\
&\approx \underline{.065}
\end{aligned}
$$

Under monoalphabetic substitution.

$$\sum_{\alpha=A}^{Z} N_\alpha^2 \cancel{\approx} \sum_{\alpha=A}^{Z} N_{\sigma(\alpha)}^2$$

$N_A \approx N \cdot p_A \quad N_B \approx N \cdot p_B \quad N_C = N \cdot p_C$ - - - -

Monoalphabetic

$N_A \approx N \cdot p_{\sigma(A)} \quad N_B \approx N \cdot p_{\sigma(B)} \quad N_C = N \cdot p_{\sigma(C)}$

Now, we calculate that

I_c for i^{th} column $\sim$ same calculation for M letters in column

$$(.065) = \frac{\sum_{\alpha=A}^{Z} \binom{M_\alpha^{(i)}}{2}}{\binom{M}{2}}$$

$$(.065)\left(\frac{(M)(M-1)}{2}\right) = \sum_{\alpha=A}^{Z} \binom{M_\alpha^{(i)}}{2}$$

$$\begin{aligned}
2D_c &= \sum_{i=1}^{p} \sum_{\alpha=A}^{Z} M_\alpha^{(i)}(M_\alpha^{(i)} - 1) + \sum_{i=1}^{p} \sum_{j=i+1}^{p} \sum_{\alpha=A}^{Z} M_\alpha^{(i)} M_\alpha^{(j)} \\
&\approx M^2 p(.065) - pM + M^2(.038)p(p-1) \\
&= \frac{N^2}{p}(.027) - N + N^2(.038)
\end{aligned}$$

Lets see how accurate this is (it gives an approximation to the period, not the actual period) with text that contains about 21K letters. We use the same text and vigenere cipher with period 3 through 7.

`indcoin < plaintext`
Index of coincidence : 0.063616
Estimate of the period : 1.052158

- `indcoin < cyphertextvig3`
  Index of coincidence : 0.044720
  Estimate of the period : 3.990527
- `indcoin < cyphertextvig4`
  Index of coincidence : 0.042903
  Estimate of the period : 5.455495
- `indcoin < cyphertextvig5`
  Index of coincidence : 0.042236
  Estimate of the period : 6.304608
- `indcoin < cyphertextvig6`
  Index of coincidence : 0.041899
  Estimate of the period : 6.842702

Even with 21,000 letters the estimate of P is kind of off

$i=1$	$i=2$	$i=3$	$i=4$
D	A	R	T
N	O	W	I
S	T	H	E
T	I	M	E
F	O	R	A
L	L	G	O
O	D		

$P = 4$

$$\#\text{ of equal letters in column } i = \sum_{a=A}^{Z} \binom{M_a^{(i)}}{2}$$

$M_A^{(1)}$ = # of A's in the first group

$M_B^{(1)}$ = # of B's in first Group

⋮

each of groups has M letters

total # of letters $N = M \cdot P$