Lecture 16 chi-square test (continued)

- Suppose 160 pairs of consecutive nucleotides are selected at random .
- Are data compatible with the independent occurrence assumption?

	A	T	G	C
A	15	10	13	7
T	10	13	7	10
G	10	10	10	10
C	5	12	10	8

Independence implies joint probability equals product of marginal probabilities

- Let P(first nucleotide = A)= P_{A1}
- P(first nucleotide = T)= P_{T1} and so on
- Let P (second nucleotide = A)= P_{A2}
- P(second nucleotide = T)= P_{T2} and so on
- P(AA)= P_{A1} P_{A2}
- P(AT) = P_{A1} P_{T2}
- We do not assume $P_{A1} = P_{A2}$ and so on

Expected value in ( ) ; df = (# of rows -1)(# of columns -1)

	A	T	G	C
A	15 (11.25)	10 (12.66)	13 (11.25)	7 (9.84)
T	10 (10)	13 (11.25)	7 (10)	10 (8.75)
G	10 (10)	10 (11.25)	10 (10)	10 (8.75)
C	5 (8.75)	12 (9.84)	10 (8.75)	8 (7.66)

Pearson's chi-square statistic= 166.8 > 27.88. P-value<.001

Simple or composite hypothesis

- Simple : parameters are completely specified
- Composite : parameters are not specified and have to be estimated from the data

- Loss of 1 degree of freedom per parameter estimated
- Number of parameters estimated = (# of rows -1)+
- (# of columns -1)
- So the df for chi-square test is #of cells -1 - (#of rows -1) - (# of columns -1) = (#of row -1)(#of col -1)

Test of independence in a contingency table

Are SARS death rates independent of countries ? Data from LA -times , as of Monday 5.pm. (Wednesday, from April 30, 2003)

	China	Hong Kong	Singapore	Canada	others
cases	3303	1557	199	344	243
death	148	138	23	21	11

Df = 1 times 4 = 4; but wait,

convert to death - alive table first

d.f. = 4

	China					total
death	148 (199.5)	138 (94)	23 (12)	21 (20.8)	11 (14.7)	341
alive	3155 (3103.5)	1419 (1463)	176 (187)	323 (323.2)	232 (228.3)	5305
total	3303	1557	199	344	243	5646

Pearson's Chi-square statistic =47.67 > 18.47; P-value<.001, reject null hypothesis, data incompatible with independence assumption