Lecture Handout 4-Hedrick ch2 (continued)

Announcements:

Try problem set 1 (answers on Compass)

Hedrick ch. 2: try problems 1-4, 6, 7, 11, 12, 15

Testing Hardy-Weinberg proportions

-Are observed numbers in the sample consistent with Hardy-Weinberg proportions in the population?

-Chi-square test is used (χ^2)

Chi-square distribution: a probability density function whose values range from zero to positive infinity.

- Very widely used, in tests of goodness of fit (does observed frequency differ from theoretical distribution) and in tests of independence (are paired observations on a contingency table independent of each other)

-Different distribution for differnt degrees of freedom
http://courses.ncssm.edu/math/Stat_Inst/PDFS/DFWalker.pdf

-Degrees of freedom to use: number of phenotypic classes (k), minus one, minus number of parameters estimated

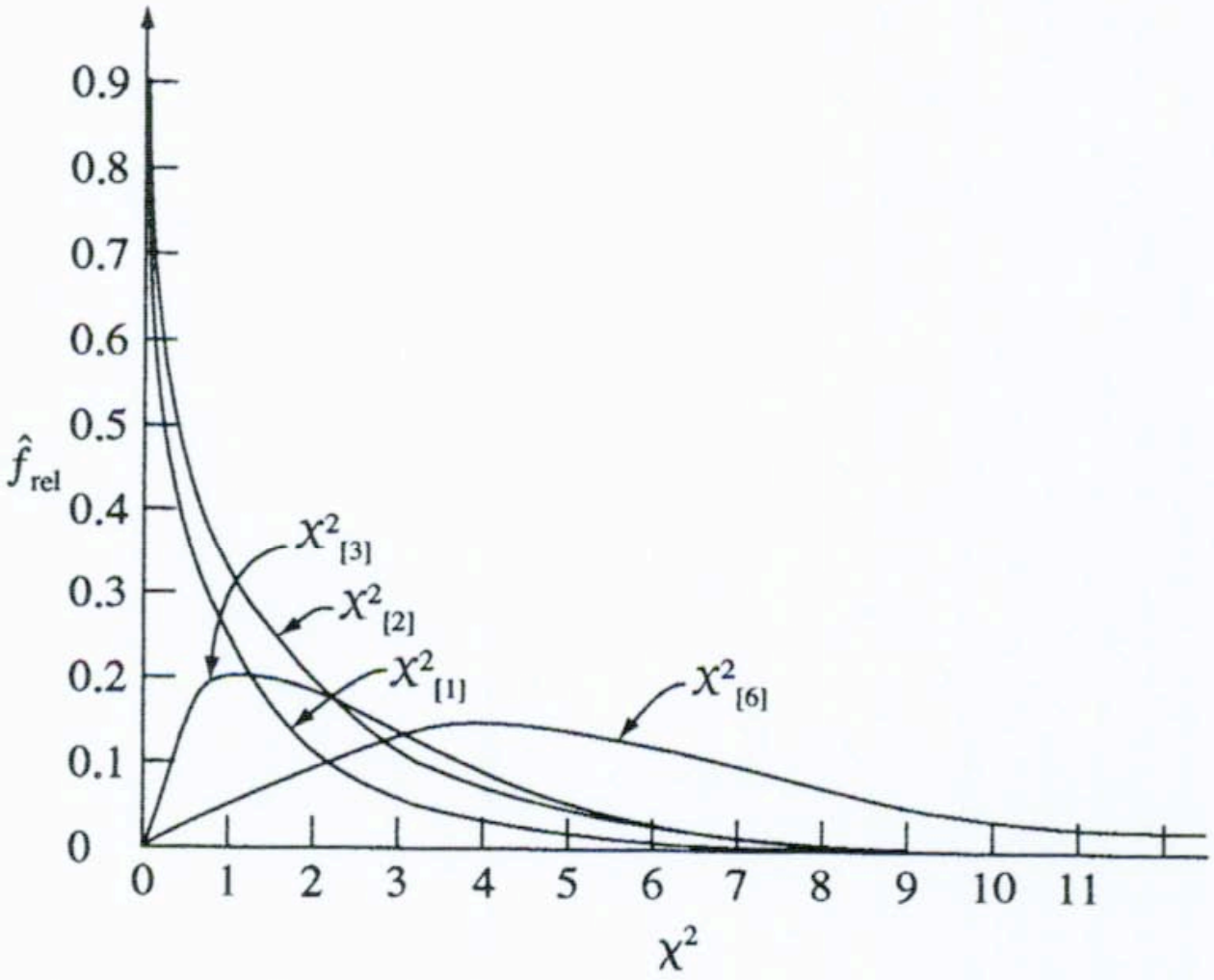
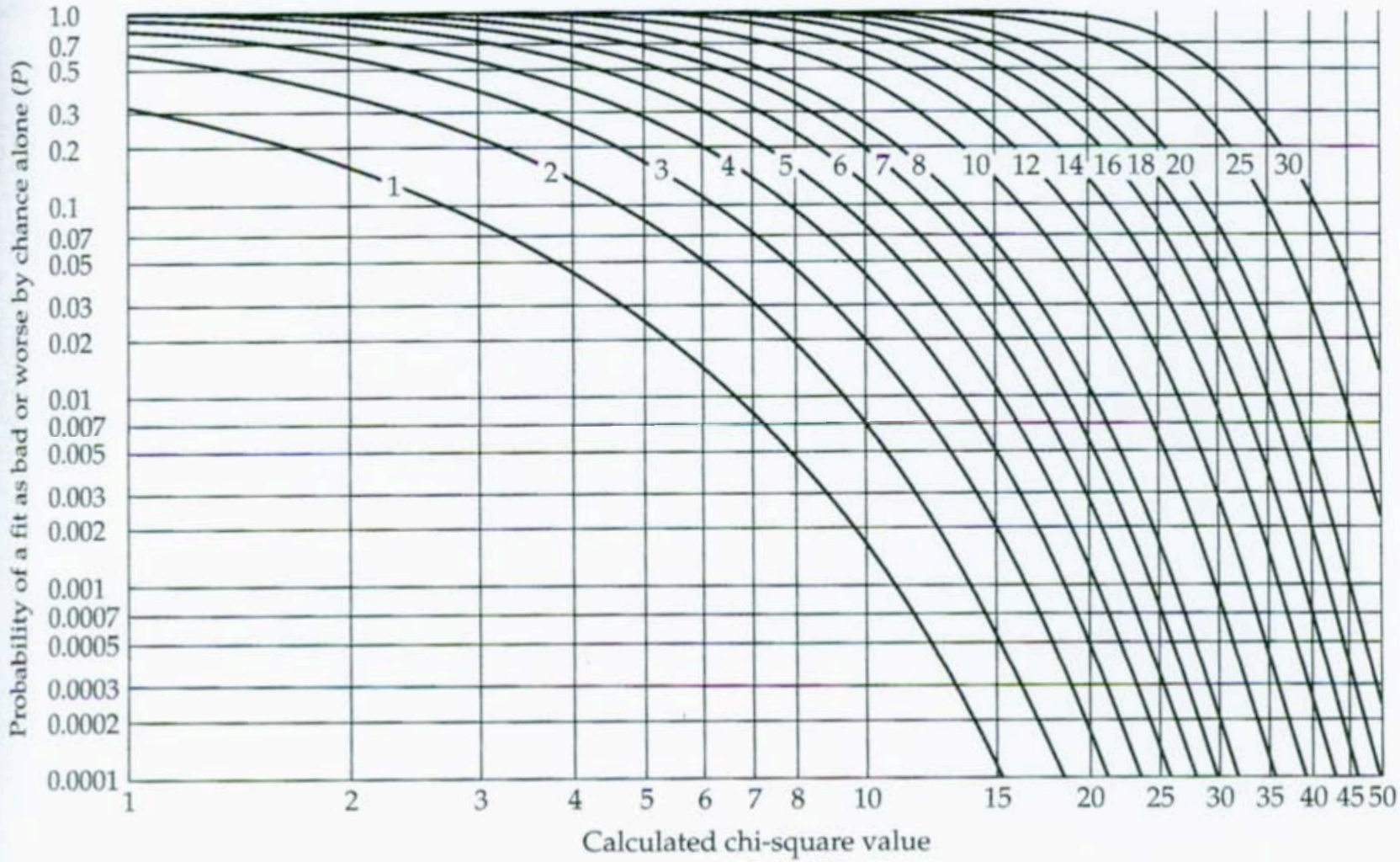


FIGURE 7.13 *Frequency curves of χ^2-distributions for 1, 2, 3, and 6 degrees of freedom.*


Probability of a fit as bad or worse by chance alone (P)
1.0
0.7
0.5
0.3
0.2
0.1
0.07
0.05
0.03
0.02
0.01
0.007
0.005
0.003
0.002
0.001
0.0007
0.0005
0.0003
0.0002
0.0001
1
2
3
4
5
6
7
8
10
12
14
16
18
20
25
30
1
2
3
4
5
6
7
8
10
15
20
25
30
35
40
45
50
Calculated chi-square value

To determine if observations are consistent with Hardy-Weinberg predictions, calculate chi-square as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O - E)^2}{E}$$

Where O and E are the observed and expected numbers for each genotype class k.
IMPORTANT: use *actual numbers* not proportions

Use a chi-square distribution for the following degrees of freedom:

TABLE 2.13 Computation of the degrees of freedom for a number of genetic systems.

Systems	*Number of phenotypic classes*	*Number of parameters*	*Degrees of freedom*
Two alleles			
Codominance	3	1	1
Dominance	2	1	0
Three alleles			
Codominance	6	2	3
Dominant series	3	2	0
ABO, null allele	4	2*	1
***n* alleles**			
Codominance	$\frac{n(n+1)}{2}$	$n-1$	$\frac{n(n-1)}{2}$
Dominant series	n	$n-1$	0

*Because $p_1 = 1 - (p_2 + p_3)$, p_1 need not be estimated from the formula given in the text.

The value of chi-square is calculated for the data, and using the degrees of freedom, a chi-square table gives the probability that the observed numbers would deviate from expected numbers by more than chance. The calculated number must be less than this value to be consistent with Hardy-Weinberg:

Potentially useful chi square critical values.

Degrees of freedom	P value = .05
1	3.84
2	5.99
3	7.81
4	9.49

Some additional definitions and observations…

Population: “a group of interbreeding individuals that exist together in time and space”

Carrier: an unaffected individual heterozygote carrying a recessive allele for a disease gene or locus

allozymes: Variants of an enzyme that are coded for by different alleles at the same locus, but differ in amino acid sequence

-Usually detected through electrophoresis

-not the same as isozymes, 2 enzymes with same function

Overlapping generations: retain ca. 0.4 and 0.1 of an initial deviation from HW proportions after 1 and 2 generations, respectively.

Exact test: similar to a chi-square test, but avoids the approximations made by the latter. Used with small sample sizes.

Chi-square test requires sample size of ≥ 50, with each class size > 5; otherwise use the exact test

Permutation test, or "randomization test": non-parametric test of significance involving the shuffling of observed data to determine how unusual an observed outcome is. Typical steps:

1. Combine the observations from all the samples
2. Shuffle them and and redistribute them it in resamples of the same sizes as the original samples.
3. Record the statistic of interest.
4. Repeat many times
5. Determine how often the resampled statistic of interest is as extreme as the observed value of the same statistic.

The above procedure is a Monte Carlo permutation test, also called a sampled or approximate permutation test. A permutation test in which all possible shufflings are systematically used is an exhaustive permutation test.

STR: “short tandem repeat” or microsatellite

- polymorphic loci

- usually codominant

- consist of repeating units of 1-6 base pairs in length

-STRs with 10 or more copies are often used

-PCR using oligonucleotide primers

Null allele: present in organism, but cannot be detected; eg unamplified STR allele will not be detectable on a gel.

May cause an apparent deficiency of heterozygotes

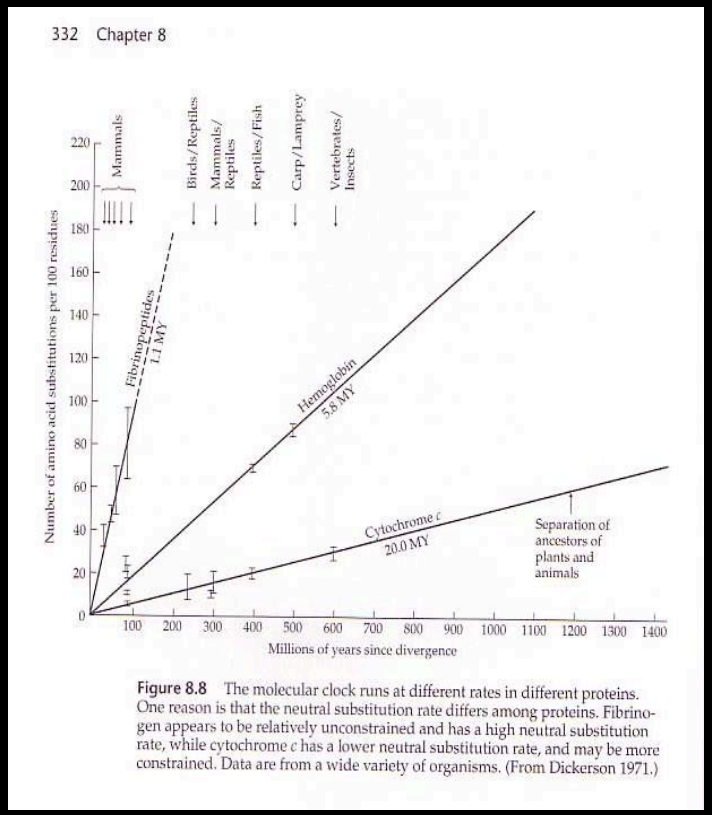


Figure 8.8 The molecular clock runs at different rates in different proteins. One reason is that the neutral substitution rate differs among proteins. Fibrinogen appears to be relatively unconstrained and has a high neutral substitution rate, while cytochrome *c* has a lower neutral substitution rate, and may be more constrained. Data are from a wide variety of organisms. (From Dickerson 1971.)

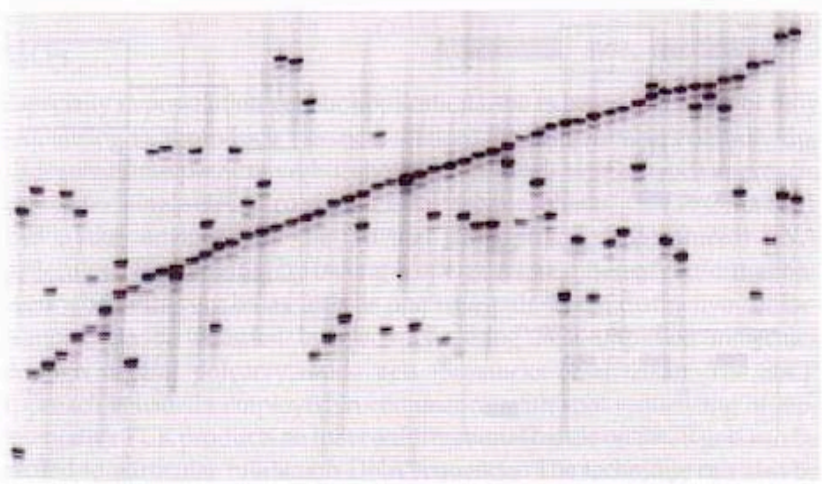
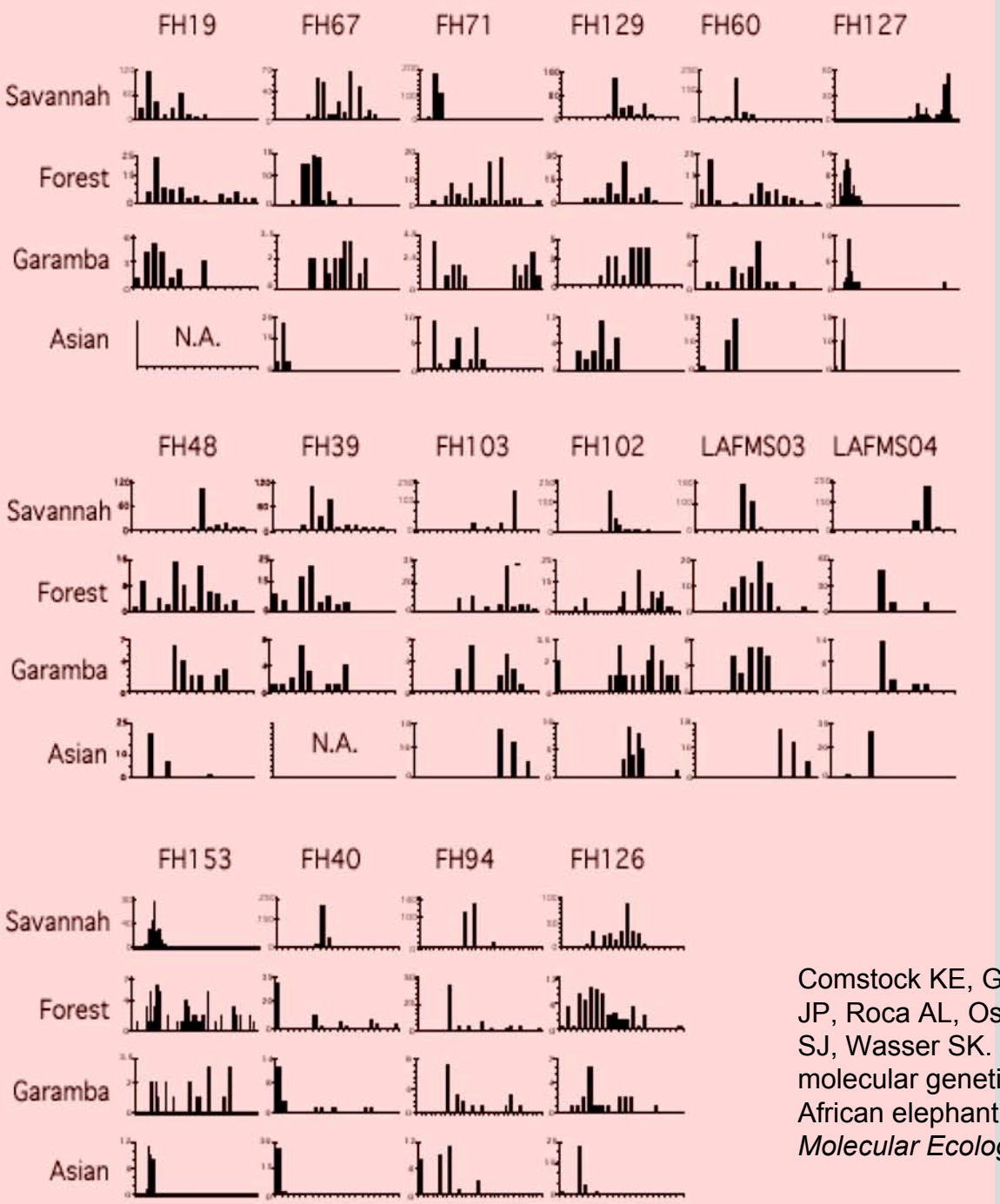


Figure 3.19 Extensive polymorphism at a microsatellite locus. Shown is an autoradiograph of a gel displaying more than 50 alleles in a local population of deer mice (unpublished data, Avise laboratory). All individuals shown are heterozygous (each displays two primary bands). Specimens were purposely arranged from left to right such that one of the two alleles in each contributes to a steplike series of consecutive-sized alleles. Actually, most microsatellite assays today involve the use of fluorescent dyes and computer gel scans to score alleles and genotypes. Several microsatellite loci can sometimes be "multiplexed" and run on the same gel, with their alleles distinguished by use of a different fluorescent dye for each.

ACGTCACACACACACACACACACACACACACACACACAGATT

ACGTCACACACACACACACACACACACACACACACACAGATT

short tandem repeats (STRs or microsatellites)



Comstock KE, Georgiadis N, Slattery JP, Roca AL, Ostrander EA, O'Brien SJ, Wasser SK. (2002) Patterns of molecular genetic variation among African elephant populations. *Molecular Ecology*.

Bonferroni correction: lowering of the p value at which a hypothesis is accepted, when one is testing multiple hypotheses.

Typically, $p \leq 0.05$ is the level of significance. This means that 1 out of every 20 hypothesis tests will appear to be significant at the $\alpha = 0.05$ level purely due to chance.

When testing n multiple independent hypotheses, typically the level of statistical significance is set to 0.05 X 1/n