STATISTICS 244 **Problem Set 6** **9:30-10:50AM TTh**

Problem Set 6 due Thursday February 28

1. Suppose $X_1, X_2, X_3, ..., X_k$ are multinomial distributed, based upon n trials, with parameters $\theta_1, \theta_2, \theta_3, ..., \theta_k$, where $\Sigma\theta_i = 1$. (a) For an arbitrary i, j, i≠j, find $cov(X_i,X_j)$ and correlation ρ_{X_i,X_j}. (b) Suppose all $\theta_i = 1/k$. What is the covariance of X_i+X_j and X_i-X_j? [Hint for (a): What is the distribution of $Z = X_i + X_j$? What is $Var(X_i + X_j)$? Use your knowledge of $Var(X_i)$ and $Var(X_j)$ to find $cov(X_i,X_j)$ from this, and then find the correlation.]

2. Maximum Likelihood may not work well with many parameters and limited data. Consider this example. The data consist of n independent pairs (X_i,Y_i) where each pair X_i and Y_i are also independent with the same distribution $N(\theta_i, \sigma^2)$. The means are different from pair to pair but the variances are all the same; there are then n+1 parameters and 2n observations. It is not hard to show that the MLEs are $(X_i+Y_i)/2$ for the n means θ_i and the MLE for the variance σ^2 is $\Sigma(X_i - Y_i)^2/(4n)$ (this may look odd but it is simply the average of the n MLEs of the variances for the separate pairs, each of those based on only two observations). (You do not need to show the work for this problem, but I suggest you verify these are the MLEs.) (a) Show that the MLE of the variance is inconsistent, meaning that it is biased and as n increases the bias does not decrease even though the variance does decrease. You may use the facts that each difference $(X_i - Y_i)$ has a $N(0, 2\sigma^2)$ distribution and the variance of the MLE of the variance based upon one pair has itself variance = $\sigma^4/2$ (see notes for Lecture 11, slide 10 (Feb 19) with n=2). (b) Show that this problem is easily fixed by using a simple multiple of the MLE for σ^2. [This is usually called the Neyman-Scott example, but it was introduced earlier by Wald.]

3. Suppose we face a pattern recognition problem, where the data consist of a single set of pixels X (where there are 16 possible pixel patterns), and there are two possible patterns θ, "0" and "6". The model is that X has the probability function $p(x|\theta)$ depending on θ, given by the following table. Find the best test for "0" vs. "6" for which the chance of making the error of "6" when the pattern is "0" is no greater than 0.10. What is the power of this test?

pixel pattern # x:

$p(x\|\theta)$:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
θ: "0"	0	0	.02	.03	.02	.03	.02	.02	.08	.12	.02	.22	.02	.23	.02	.15
"6"	0	.03	.01	.13	.08	.12	0	0	.02	.20	.01	.17	.04	.11	0	.08

4. Suppose X has a $N(\mu, \sigma^2)$ distribution. (a) Find the Most Powerful test for testing at level $\alpha = 0.05$ the hypothesis H_0: $\mu = 6$ and $\sigma^2 = 4$ vs. H_1: $\mu = 9$ and $\sigma^2 = 4$. (b) Find the power of this test. (c) Suppose that instead of the above H_1, we have H_1: $\mu = \mu_1$ and $\sigma^2 = 4$, where $\mu_1 > 6$. Find and graph the power function.

5. <u>The Thick Coin</u>. Suppose a thick coin is tossed n = 100 time and the numbers of Heads (X), Tails (Y), and Edges (Z) that come up are counted, so that X + Y + Z = n = 100. Suppose that P(H) = P(T) = θ, and so P(E) = $1 - 2\theta$. (a) Find the form of the Most Powerful test of the hypothesis H_0: $\theta = 1/4$ vs H_1: $\theta = 1/3$, and express this test in terms of Z [e.g. $Z < C$ or $Z > C$]. (b) What, in terms of n and θ, is the approximate distribution of Z that you would find from the Central Limit Theorem? What for example is the approximate value of P($Z \leq 30$) when H_0 is true? (c) Find C when $\alpha = .05$. (d) Is this test Uniformly Most Powerful for H_0 vs H_1: $\theta > 1/4$?

6. The following problem arises in a study in ecology. A bird's nest is selected at random as a reference point (P) in a marsh, and the largest circle is determined that has the reference point P as the center and does not include (or just barely includes, if you count the boundary) the next nearest bird's nest (N). The area of that circle is used as a measure of the population of birds in the marsh; the larger the area is, the lower the population. If the nests are randomly scattered in the marsh, then according to one model the <u>radius</u> X of that largest circle will have a Gamma (α, β) distribution (see text Ch. 5, p. 5-20) with $\alpha = 3$ and $\beta = 1/\lambda$; that is, its probability density is given by:

$$f(x \mid \lambda) = \left(\frac{\lambda^3}{2}\right) x^2 e^{-\lambda x} \quad \text{for } x > 0$$

$$= 0 \quad \textit{otherwise}.$$

Consider the problem of testing H_0: $\lambda = 3$ vs H_1: $\lambda = 5$, based upon a sample $X_1, X_2, , X_n$. (a) Find the form of the Most Powerful test, and explain how you could the "Additional Fact" below to find an exact, explicit 0.05 level test for a given n. (b) Use the Central Limit Theorem to specify the test in (a) approximately. [You may use the fact that for a random variable X with a Gamma distribution with this parameterization, E(X) = α/λ and Var(X) = α/λ^2.]

<u>Additional Fact about the Gamma Distribution (see Chapter 5 of text):</u>
The Chi-squared distributions are a special case of the gamma distributions, with $\alpha = n/2$ and $\beta = 2$. If $X_1, X_2, X_3, ..., X_m$ are independent random variables, each with a gamma (α, β) distribution, then $Z = X_1 + X_2 + X_3 + ... + X_m$ can be shown to have a gamma ($m\alpha$, β) distribution, and the average $\overline{X} = Z/m$ has a gamma ($m\alpha$, β/m) distribution.