QUALIFYING EXAMINATION — MATHEMATICAL STATISTICS

Part I, Morning — May 29, 1958

DO FIVE PROBLEMS

1. Show that the probability of the simultaneous realization of exactly m events among N events $E_1, E_2, \ldots, E_N$ is

$$S_m - \binom{m+1}{1} S_{m+1} + \binom{m+2}{2} S_{m+2} - \ldots + (-1)^{N-m} \binom{N}{m} S_N,$$

where

$$S_1 = \sum_i \Pr(E_i), \quad S_2 = \sum_i \sum_j \Pr(E_i E_j),$$

$$S_3 = \sum_i \sum_j \sum_k \Pr(E_i E_j E_k), \text{ etc.},$$

summation in the expressions for the S's being extended over $i, j, k, \ldots$ such that $1 \leq i < j < k < \ldots \leq N$; $\binom{N}{m}$ is the usual binomial coefficient $N!/\{m!(N-m)!\}$.

[Hint: Establish first the formula for $N = 3$, $m = 2$, with the aid of a diagram, to devise the method of proof for the general case.]

r objects are placed in n cells under the condition that all n^r arrangements are equally likely. By defining E_k $(k = 1, 2, \ldots, n)$ as the event 'cell number k empty', or otherwise, prove that the probability that exactly m cells are empty is

$$\binom{n}{m} \sum_{\nu=0}^{n-m} (-1)^{\nu} \binom{n-m}{\nu} \left(1 - \frac{m+\nu}{n}\right)^r .$$

2. State and prove the central limit theorem for the suitably standardized sum of a number of identically and independently distributed random variables. Under a certain growth hypothesis the final size of an organ may be regarded as the net result of a large number of infinitesimally small independent impulses, where the effect of each impulse (i.e. the increase in growth) is proportional to the size of the impulse as well as to the momentary size of the organ. Show that the final size has a logarithmic normal distribution.

3. Given that $\sum_i x_i^2$, where $x_1, x_2, \ldots, x_n$ are independent normal random variables with zero means and unit variances, and where $\sum_i$ denotes summation from $i = 1$ to n has a χ^2 distribution with n degrees of freedom, deduce the distributions of

(i) $\sum_i x_i^2 - (\sum_i a_i x_i)^2 / \sum_i a_i^2$, where the a_i are arbitrary constants.

(ii) $\sum_i w_i (z_i - \bar{z})^2$, where the z_i $(i = 1, \ldots, n)$ are independent normal random variables with common mean μ and variances σ^2/w_i, and where $\bar{z} = \sum_i w_i z_i / \sum_i w_i$.

Independent groups of observations x, y are such that those in the i^{th} group satisfy the equation $y = \alpha_i + \beta_i x + \epsilon$, the error ϵ being normal and independent with zero mean and variance σ^2. Let $S_i(xy)$, $S_i(x^2)$ denote the sum of products of x and y and the sum of squares of x respectively, for the i^{th} group, each variable being measured from its group mean; write $b_i = S_i(xy)/S_i(x^2)$. Then if the β_i are all equal to β, deduce the distribution of

$$\sum_i \left\{ S_i(x^2)(b_i - \bar{b})^2 \right\}, \text{ where } \bar{b} = \sum_i S_i(xy) / \sum_i S_i(x^2).$$

How would this be used to provide a test of the hypothesis of the equality of the β_i when σ^2 is known?

4. After the starting of a loom, the time to the first warp break has a distribution with frequency law

$$f(x)dx = e^{-x/\theta}\,dx/\theta \qquad (x > 0)\ ;$$

n looms are started simultaneously and the times to break recorded. Observations stop when the warp of r looms has broken, the recorded times being $x_1, x_2, \ldots, x_r$ $(x_1 \leq x_2 \leq \ldots \leq x_r)$. Show that the probability density function of $x_1, x_2, \ldots, x_r$ is

$$\frac{n!}{(n-r)!\,\theta^r} \exp\left\{-\frac{1}{\theta}[\sum_{i=1}^{r} x_i + (n-r)x_r]\right\}.$$

Hence find the maximum likelihood estimate $\hat{\theta}$ of θ, and show that it is an unbiased estimate with minimum variance θ^2/r. By obtaining the moment-generating function of $\hat{\theta}$, or otherwise, show that $\hat{\theta}$ is distributed as a multiple of χ^2 with $2r$ degrees of freedom.

5. In an investigation of the toxicity of a certain drug, a dose x_i was given to each of n_i animals in the i^{th} group $(i = 1, 2, \ldots, k)$, as a result of which r_i animals died in this group. If the probability of death for each animal and for a given dose x_i has the form $P_i = F(\alpha + \beta x_i)$, where F is a known cumulative distribution function and α and β unknown parameters, deduce the maximum likelihood equations for α and β:

$$\sum_i w_i(U_i - Y_i) = 0\ , \quad \sum_i w_i x_i(U_i - Y_i) = 0\ ,$$

where $Y_i = \alpha + \beta x_i$, $U_i = Y_i + (\frac{r_i}{n_i} - P_i)\frac{1}{F'(Y_i)}$, $w_i = \frac{n_i[F'(Y_i)]^2}{P_i(1-P_i)}$,

$$(F'(x) = \frac{dF(x)}{dx})\ ,$$

and where $\sum_i$ denotes summation from $i = 1$ to k.

Problem 5 continued

From the associated information matrix, show that

$$\operatorname{var} \hat{\alpha} \sim \frac{1}{\sum_i w_i} + \frac{\bar{x}^2}{S}, \quad \operatorname{cov}(\hat{\alpha}, \hat{\beta}) \sim -\frac{\bar{x}}{S}, \quad \operatorname{var} \hat{\beta} \sim \frac{1}{S},$$

where $\bar{x} = \sum_i w_i x_i / \sum_i w_i$, $S = \sum_i w_i (x_i - \bar{x})^2$.

Explain also how the χ^2 Distribution may be used to test the goodness of fit of the dosage-mortality relation $P = F(\alpha + \beta x)$ to the observed proportions of deaths at the various doses.

6. If $\bar{x}$ is the arithmetic mean of n independent observations from a population with mean μ, discuss the validity and meaning of the assertion: $\bar{x} \to \mu$ in probability as $n \to \infty$.

If the observations are all mutually intercorrelated, the correlation between every pair being ρ_n, can the assertion still be true when (i) $\rho_n = 1/n$, (ii) $\rho_n = 1/2$, (iii) $\rho_n < 0$?

7. A factory produces a standard article in batches of N. From each batch x articles are selected at random and tested to destruction simultaneously, and the whole batch is rejected if any defectives are found in the sample. The cost of testing is a per article tested, the cost of each good article rejected is b, and the loss entailed through each defective article in a batch that has passed the test is c. Write down the average cost of the inspection scheme for batches made up when the proportion of good articles being produced is g.

If g can be assumed to be distributed from batch to batch as

$$(m + 1)g^m dg \qquad (0 < g < 1)$$

where m is a known parameter, find the mean cost of sampling per batch and discuss the choice of an optimal value for x.

8. State, <u>without attempting to supply a proof</u>, Wald's fundamental identity. By means of it, or otherwise, obtain approximate expressions for the operating characteristic and average sample number functions for the probability ratio sequential test relevant to the discrimination between two hypotheses, corresponding to two probability density functions (or two probability functions) of known functional form with prescribed risks of errors of the two kinds. [Any additional assumptions should be clearly stated.]

The deviations along any two perpendicular directions of the points of impact of rounds fired from a certain gun, measured from the center of the target, are distributed normally and independently with zero means and equal but unknown standard deviations σ. The specifications laid down are as follows: the probability of classifying the gun as unsatisfactory should not exceed a small preassigned value α whenever $\sigma \leq \sigma_0$, and the probability of classifying the gun as satisfactory should not exceed a preassigned value β whenever $\sigma \geq \sigma_1$ $(\sigma_1 > \sigma_0)$. Give a sequential procedure for deciding whether the gun is satisfactory or not, and obtain the approximate average number of rounds necessary to arrive at a decision when $\sigma = \sigma_0$.

9. State the assumptions on which the standard analysis of a randomized block experiment depend. Explain why a transformation of the variable may be useful when these assumptions do not hold, indicating briefly the considerations involved in the choice of a suitable transformation. What transformation would you recommend when the variable is

(a) an estimate of variance from a normal sample with a fixed number of degrees of freedom,

(b) the proportion of occurrences of an event in a fixed number of independent trials?

In an experiment to compare two paints A and B, as possible rust-preventives for steel plates, a number of batches of twenty plates were used, a random ten from each batch receiving paint A, the rest B. After exposure to the air for a given time, the proportion in each group showing satisfactory rust-prevention was recorded. Explain how you would analyse the results.