THE EFFICIENCY OF $\bar{X}$ WHEN LOG X IS HOMOSCEDASTIC: A M.S. THESIS RESEARCH PROBLEM

BU-433-M by November, 1972

D. S. Robson

Abstract

The transformation $y = \log x$ is widely applied in circumstances where the coefficient of variation $C_x = s/\bar{x}$ tends to remain constant for samples from different populations. When x rather than log x represents the scale of economic value, as when x = bushels per acre, an estimate of the mean μ_x is normally the most cogent summary statistic and is usually calculated as the arithmetic mean $\bar{x}$ of the untransformed measurements. We propose here to investigate the efficiency of both this robust estimator of μ_x and the conventional interval estimator $\bar{x} \pm ts_{\bar{x}}$ when the pdf of x belongs to either a gamma or a log-normal family satisfying σ_x/μ_x = constant.

Paper No. BU-433-M in the Biometrics Unit Mimeograph Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14850.

THE EFFICIENCY OF $\bar{X}$ WHEN LOG X IS HOMOSCEDASTIC: A M.S. THESIS RESEARCH PROBLEM

BU-433-M by November, 1972

D. S. Robson

In many fields of quantitative investigation the coefficient of variation $C_x = s/\bar{x}$ is informally employed as an indicator of quality control. Agronomic experiments with corn may reveal, for example, that the coefficient of variation in yield is typically, say, 15 percent ($C_x = .15$), and a corn experiment producing a C_x-value substantially different from .15 is then considered suspect.

Constancy of this ratio of experimental standard deviation to experimental mean implies that $\sigma_x^2 = C_x^2\mu_x^2$, and since

$$\sigma^2_{\log x} \doteq \frac{\sigma_x^2}{\mu_x^2} = C_x^2$$

then homoscedasticity (homogeneity of variance) would be approximately achieved by the transformation $y = \log x$. Though widely employed in data analysis the log transformation is sometimes avoided on the grounds that "bushels of corn and not log-bushels of corn represent the economic scale". Accepting this fact that the experimenter wishes to estimate the mean value of X, or $E(e^Y)$ rather than either $E(Y)$ or $e^{E(Y)}$, we here consider the question of efficiency of the simple, robust estimators $\bar{X}$ and $\bar{X} \pm 1.96s_{\bar{x}}$ compared to some parametric alternatives.

Two particular parametric alternatives which approximate a wide range of real situations are the gamma and log-normal models. If the pdf of X is a gamma density with scale parameter θ,

$$f_X(x) = \frac{x^{\alpha-1}e^{-x/\theta}}{\theta^{\alpha}\Gamma(\alpha)}$$

then the average value of X is given by $\mu_x = \alpha\theta$, and in the scalar family determined by a fixed but unknown α the variance σ_x^2 is proportional to μ_x^2, $\sigma_x^2 = \mu_x^2/\alpha$, and the transformation Y = log X would thus transform this scalar family into an approximately homoscedastic family. In this case the pair $(\bar{X},\bar{Y})$ is a minimal sufficient statistic, and $\bar{X}$ is an efficient estimator of μ_x. The asymptotically valid interval estimator $\bar{X} \pm 1.96s_{\bar{x}}$ has questionable finite sample properties, however, which could be revealed by comparison with the power function of an exactly valid interval estimator of μ_x. Questions to be answered are: how rapidly does $P(\bar{X} - 1.96s_{\bar{x}} < \mu_x < \bar{X} + 1.96s_{\bar{x}})$ approach the nominal and limiting value of .95 over the (α,θ) parameter space, and how does the power of this interval estimator compare to the power which could be achieved with the a priori knowledge that the pdf of X is a gamma density function?

If Y = log X is normally distributed then $\mu_x = e^{\mu_y + \frac{1}{2}\sigma_y^2}$ and

$$\sigma_x^2 = \mu_x^2(e^{\sigma_y^2} - 1).$$

Thus, if the family of normal distributions of Y is homoscedastic (σ_y^2 = constant) then the coefficient of variation C_x is constant, $C_x^2 = e^{\sigma_y^2} - 1$. In this case $\bar{X}$ is not an efficient estimator of μ_x since $E(\bar{X}|\bar{Y},s_Y^2) \neq \bar{X}$, so the use of this robustly unbiased estimator must entail some sacrifice in efficiency when compared to parametric alternatives. The above Blackwell estimator, $E(\bar{X}|\bar{Y},s_Y^2)$, if efficient is still impractical, and other parametric alternatives $\hat{\mu}(\bar{Y},s_Y^2)$ of simpler structure should also be examined for comparison with $\bar{X}$. Again, the rate of approach to asymptotic validity of the interval estimator $\bar{X} \pm 1.96s_{\bar{x}}$ should be examined and power comparisons made with exactly or more nearly valid estimators based on the log-normal parametric model.

In both of these parametric families the presence of a nuisance parameter complicates the estimation problems and ensures that numerical analytic techniques must play a prominent role in the investigation.