STAT 521: Midterm Exam 2 **Name:** Jae-Kwang Kim

Problem 1: (20 pts)

Assume a finite population of size $N = 10,000$. From the finite population, suppose that we select a simple random sample of size $n = 13$ with observation (x_i, y_i) to obtain

$$\hat{y} = 6.5 + 0.5x$$

as the regression fit from the sample. The mean squared error of the regression fit is 1.0. The sample correlation between x and y is 0.7. The sample mean of x is 4.0 and the population mean of x is 5.0.

(a) Compute the confidence interval of HT estimator of the mean of y.

$\hat{\bar{Y}}_{HT} = 6.5 + 0.5 \times 4 = 8.5$

$\hat{V}(\hat{\bar{Y}}_{HT}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S_y^2 = \frac{1}{n}\left(1 - \frac{n}{N}\right) \frac{MSE}{1 - r^2} \doteq \frac{1}{13} \times \frac{1}{1 - 0.7^2} = 0.151$

(b) Compute the confidence interval of the regression estimator of the mean of y.

$\hat{\bar{Y}}_{reg} = 6.5 + 0.5 \times 5 = 9$

$\hat{V}(\hat{\bar{Y}}_{reg}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S_e^2 = \frac{1}{n}\left(1 - \frac{n}{N}\right) MSE \doteq \frac{1}{13} \times 1 = 0.077$

Problem 2: (30 pt)

Consider a (sample random) sample of size n from a bivariate finite population of (x_i, y_i). Let $(\bar{X}, \bar{Y})$ be the population mean of (x_i, y_i) and (S_x^2, S_y^2) be the population variance of (x_i, y_i). Also, let ρ be the population correlation coefficient of (x_i, y_i). Assume that $\bar{X}, S_x^2, S_y^2$ and ρ are known and we are only interested in estimating $\theta = \bar{Y}$. Note that we can obtain $(\bar{x}_n, \bar{y}_n)$, the sample mean of (x_i, y_i), from the sample.

For estimating $\theta = \bar{Y}$, suppose that we are going to use the following estimator:

$$\hat{\theta}_p = \frac{\bar{x}_n}{\bar{X}} \bar{y}_n$$

(a) Express the bias of $\hat{\theta}_p$ as a function of population quantities up to order $1/n$.

$\hat{\theta}_p = \bar{Y} + R(\bar{x}_n - \bar{X}) + (\bar{y}_n - \bar{Y}) + \frac{1}{\bar{X}}(\bar{x}_n - \bar{X})(\bar{y}_n - \bar{Y}) - \frac{1}{\bar{X}} \cdot R \cdot (\bar{x} - \bar{X})^2 + o_p(\frac{1}{n})$

$\Rightarrow E[\hat{\theta}_p] = \bar{Y} + \frac{1}{\bar{X}} \cdot \frac{1}{n}[S_{xy} - R S_{xx}] + o(\frac{1}{n})$

$\therefore \text{Bias}(\hat{\theta}_p) = \frac{1}{n} \cdot \frac{1}{\bar{X}} \cdot (S_{xy} - R S_{xx})$, where $R = \bar{Y}/\bar{X}$.

(b) Discuss how to develop a bias-corrected estimator of $\hat{\theta}_p$. (That is, the bias of this estimator will be smaller order than the bias of $\hat{\theta}$ above.)

$\hat{\theta}_{BC} = \hat{\theta}_p - \widehat{\text{Bias}(\hat{\theta}_p)}$

$= \hat{\theta}_p - \frac{1}{n} \cdot \frac{1}{\bar{x}} [\hat{\rho} S_x S_y - \hat{R} S_x S_x^2]$, $\hat{R} = \frac{\bar{y}_n}{\bar{x}_n}$.

(c) Express the asymptotic variance of $\hat{\theta}_p$ and discuss when it is preferable to $\bar{y}_n$.

$\text{Var}(\hat{\theta}_p) \cong \text{Var}(\bar{y}_n + R \bar{x}_n)$

$= \frac{1}{n}(1-f)(S_y^2 + R^2 \rho S_x S_y + S_x^2)$

$\therefore \text{Var}(\hat{\theta}_p) \leq \text{Var}(\bar{y}_n)$

$\Leftrightarrow \rho \leq -\frac{1}{2} \frac{S_x/\bar{X}}{S_y/\bar{Y}}$ if $R > 0$

$\rho \geq -\frac{1}{2} \frac{S_x/\bar{X}}{S_y/\bar{Y}}$ if $R < 0$

Problem 3:(20 pts)

We are interested in estimating the proportion of married students at ISU with population size $N = 10,000$. Suppose that we have a simple random sample of size $n = 100$ and the result is as follows.

	Sample Size	Sample Proportion of Married Students
Male	60	0.3
Female	40	0.4

It is known that the population proportion of male students is 50 %.

(a) Compute the confidence interval for the proportion of married students at ISU using the post-stratified estimator with gender being the poststratum.

$\hat{P}_{post} = 0.5 \times 0.3 + 0.5 \times 0.4 = 0.35$

$\hat{V}(\hat{P}_{post}) = \frac{1}{n}(1-f) \cdot \frac{n}{n-1}(0.5 \times 0.3 \times 0.7 + 0.5 \times 0.4 \times 0.6) \doteq 0.00225$

(b) Compute the reduction of variance due to poststratum compared to the HT estimator in the estimation of the proportion of married students at ISU.

$\hat{V}(\hat{P}_{HT}) = \frac{1}{n}(1-f)\frac{n}{n-1} \times 0.34 \times (1-0.34) = 0.00224$

$\hat{P}_{HT} = \frac{1}{100} \times (60 \times 0.3 + 40 \times 0.4) = 0.34$

$\frac{\hat{V}(\hat{P}_{post})}{\hat{V}(\hat{P}_{HT})} \cong 1$ No reduction

Problem 4: (30 pts)

An experienced investigator makes an eye estimate of the number of dwellings, x_i, on each block of a city containing N blocks. Then a sample of n blocks is selected by SRS of blocks and the true numbers, y, are obtained. Consider the following estimator of the total $Y = \sum_{i=1}^{N} y_i$.

$$\hat{Y}_d = \frac{N}{n} \sum_{i \in A} \{y_i + (\bar{X} - x_i)\}$$

where $\bar{X} = N^{-1} \sum_{i=1}^{N} x_i$. Let $B = S_{xy}/S_x^2$ be the population slope for the regression of y on x.

(a) Find the mean and variance of $\hat{Y}_d$.

$E(\hat{Y}_d) = Y$

$V(\hat{Y}_d) = \frac{N^2}{n}(1-f)(S_y^2 + S_x^2 - 2S_{xy})$

(b) Under what conditions of B, it is better to use $\hat{Y}_d$ than the HT estimator of Y?

$B = \frac{S_{xy}}{S_x^2} \geq \frac{1}{2}$

(c) Instead of $\hat{Y}_d$, consider the following class of estimators.

$$\hat{Y}_c = \frac{N}{n} \sum_{i \in A} \{y_i + c\,(\bar{X} - x_i)\}$$

Find the value of c that minimizes the variance of $\hat{Y}_c$.

$V(\hat{Y}_c) = \frac{N^2}{n}(1-f)(S_y^2 + c^2 S_x^2 - 2c S_{xy})$

$\therefore$ minimized at $c = \frac{S_{xy}}{S_x^2}$.