

# Comprehensive Exam Reviewers

Joseph S. Tabadaro, Jr.

10/19/2018

## Theory

### Sampling

1. Let  $Z_i = 1$  if unit  $i$  is in the sample and 0 otherwise, with  $P(Z_i = 1) = \pi_i$ . Let  $R_i = 1$  if unit  $i$  responds to the survey and 0 otherwise, with  $P(R_i = 1) = \phi_i$  and  $\Phi = \sum_{i=1}^N \phi_i/N$ . Assume  $R_i$  is independent of  $Z_i$  for each  $i = 1, 2, \dots, N$ . Let  $\bar{y}_R$ . Estimate the population mean  $\bar{Y}$  using only the respondents. Divide the sample into  $C$  weighting classes and define  $X_{ci} = 1$  if unit  $i$  is in class  $C$  and 0 otherwise. Let

$$\bar{\phi}_c = \frac{\sum_i \phi_i X_{ci}}{\sum_i X_{ci}}, \quad \hat{\phi}_c = \frac{\sum_{i=1}^N Z_i R_i X_{ci} w_i}{\sum_{i=1}^N Z_i X_{ci} w_i}, \quad \hat{y}_{wc} = \sum_{i=1}^n Z_i R_i w_i y_i \sum_{c=1}^C \frac{X_{ci}}{\hat{\phi}_c}. \quad (1)$$

Show that if the weighting classes are sufficiently large,

$$\text{Bias}(\hat{y}_{wc}) = \sum_{c=1}^C \sum_{i=1}^N \frac{X_{ci} \phi_i (y_i - \bar{Y}_c)}{\bar{\phi}_c} \quad (2)$$

Thus, the weighting class adjustments for non-response produce an approximately unbiased estimator if (a)  $\phi_i = \bar{\phi}$  for all units in class  $C$  (b)  $y_i = \bar{Y}_c$  for all units  $C$  or (c) within each class  $C$ , the propensity to respond is uncorrelated with  $y_i$ .

2. Show that for epsem, the Horvitz-Thompson formulas are equivalent to the SRS formulas. Let  $Z_i = 1$  if PSU  $i$  is in the sample and 0 otherwise.

$$\hat{y}_{SRS} = N\bar{y}, \quad \text{Var}(\hat{y}) = N^2(1-f) \frac{S^2}{n} \quad (3)$$

$$\hat{y}_{HT} = \sum_{i \in S} \frac{\hat{y}_i}{\pi_i} = \sum_{i=1}^A Z_i \frac{\hat{y}_i}{\pi_i} \quad (4)$$

$$\text{Var}(\hat{y}_{HT}) = \sum_{i=1}^A A\pi_i(1-\pi_i) \frac{y_i^2}{\pi_i^2} + \sum_{i=1}^A \sum_{k=1, i \neq k}^A (\pi_{ik} - \pi_i \pi_k) \frac{y_i}{\pi_i} \frac{y_k}{\pi_k} + \sum_{i=1}^A \frac{\text{Var}(\hat{y}_i)}{\pi_i} \quad (5)$$

3. Kuk proposed the following randomized response methods: Ask the respondent to answer one of two independent binary variables  $x_1$  and  $x_2$  with  $P(x_1 = 1) = \theta_1$  and  $P(x_2 = 1) = \theta_2$ . The probabilities are known. The respondent could tell you the value of  $x_1$  if she is in the sensitive class and  $x_2$  if she is not in the sensitive class. Suppose the true proportion of the persons in the sensitive class is  $p_S$ .

- (a) What is the probability that the respondent reports 1?
- (b) Using your answer to (a) give an estimator  $\hat{p}_S$  of  $p_S$ . What conditions must  $\theta_1$  and  $\theta_2$  satisfy?
- (c) What is  $\text{Var}(\hat{p}_S)$  if an SRS is taken?

4. Kalton and Anderson consider disproportionate stratified random sampling for estimating the mean of a characteristic  $y_i$  in a rare population. Let  $R_i = 1$  if person  $i$  is in the rare population and 0 otherwise. Stratum 1 contains  $N_1$  persons  $M_1$  of whom are in the rare population, stratum 2 contains  $N_2$  persons with  $M_2$  persons in the rare population. We wish to estimate the population mean  $\bar{y}_d = \sum_{i=1}^N R_i y_i / (M_1 + M_2)$  using stratified sampling. Sample  $n_1$  persons in stratum 1 and  $n_2$  persons in stratum 2.

- (a) Suppose  $A = M_1/(M_1 + M_2)$  is known. Let  $\bar{y}_d = A\bar{y}_1 + (1 - A)\bar{y}_2$  where  $\bar{y}_1$  and  $\bar{y}_2$  are the sample means of the rare population numbers in stratum 1 and 2, respectively. Show that if you ignore the finite population correction (fpc) and if the sample number of persons in the rare population in each stratum is sufficiently large, then

$$\text{Var}(\hat{y}_d) = \frac{A^2 S_1^2}{n_1 p_1} + \frac{(1 - A)^2 S_2^2}{n_2 p_2}, \quad (6)$$

where  $S_j^2$  is the variance of the  $y$  for the rare population members in stratum  $j$  and  $p_j = M_j/N_j$ ,  $j = 1, 2$ .

- (b) Suppose that  $S_1^2 = S_2^2$  and that the cost to sample each member is the same. Let  $f_2 = n_2/N_2$  be the sampling fraction in stratum 2, and write the sampling fraction in stratum 1 as  $f_1 = f_2$ . Show that the variance in (a) is minimized for a fixed sample size  $n$  when  $k = \sqrt{p_1/p_2}$ .

## Linear Model

1. Consider the design model  $\underline{y} = \underline{X}\underline{\beta} + \underline{\epsilon}$  given as

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix} \quad (7)$$

- (a) Is  $\underline{\beta}$  estimable? Is  $\underline{\mu}$  estimable? Show why or why not for each.  
(b) Give an example of  $\underline{\beta}$  estimable function and show that indeed it is estimable.
2. Discuss the significance of expected mean squares. Use the algebraic method to find  $E[MSTreatment]$  in model in # 1.
3. Show the effects of significant interaction effects in the test for the main effects for the two-factorial model.

## Statistical Inference

1. Consider the  $X$  be a single observation from the density  $f(x; \theta) = (1 + \theta)x^\theta I_{(0,1)}(x)$  where  $\theta > -1$ .
- (a) Find MP size  $\alpha$  test for  $H_0 : \theta = 0$  versus  $H_a : \theta = 1$ .  
(b) Give a UMP size test of  $H_0 : \theta \leq 0$  versus  $H_a : \theta > 0$ .
2. Consider a random variable  $X$  which takes on the values in the interval  $(0, 1)$ . Two possible probability density functions are being entertained for  $X$  namely:  $f_0(x) = 1$ , where  $0 < x < 1$  and  $f_1(x) = 2x$  where  $0 < x < 1$ . For testing  $H_0: X$  has pdf  $f_0$  versus  $H_a : X$  has pdf  $f_1$ .
- (a) Find the MP test of size  $\alpha$ .  
(b) Let  $\pi(\alpha) - \alpha$  be maximum, i.e., find the value of  $\alpha$  for which the power of MP level  $\alpha$  test exceeds  $\alpha$  by the greatest amount.
3. Let  $X$  be a random variable with pdf  $\text{Poisson}(\theta)$ , where  $\theta > 0$ , and let  $A = [0, \infty)$ . For any  $\theta > 0$  and  $a \in A$ , let the loss function be defined as  $L(\theta, a) = (\theta - a)^2/\theta$ . Give the prior distribution of  $\theta$  to be  $\text{Gamma}(\alpha, \beta)$ .
- (a) Find the posterior distribution of  $\theta$  given  $X = x$ .  
(b) Determine the Bayes estimator of  $\theta$  with respect to the given prior distribution.  
(c) Show that the UMVUE of  $\theta$  cannot be a Bayes estimator with respect to any prior distribution.

## Probability

1. Let  $\{X_n\}$  be iid with  $E(X_1) = 0$ . Let  $C_n$  be a bounded sequence of real number. Define the sequence of random variable  $\{Y_n\}$  as

$$Y_n = \begin{cases} X_n & \text{if } |X_n| \leq n \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

- if  $\frac{1}{n} (\sum C_i Y_i - E(\sum C_i Y_i)) \rightarrow 0$  as  $n \rightarrow \infty$ . Show that  $\frac{1}{n} \sum C_i Y_i$  converges almost surely to 0.
2. Let  $(X_n)$  be a sequence of independent random variables. Suppose  $X_n$  has uniform distribution on  $(-1/\sqrt{n}, 1/\sqrt{n})$ . Show that the Lindenberg's condition holds for the sequence and hence, that the CLT holds.
  3.  $\text{Var}(X_k) = k^{3/2}$ ,  $E(X_k) = 0$ . Determine whether or not the Weak Law of Large numbers (WLLN) holds for the sequence  $(X_k)$ .

## Applications

### Statistical Inference

1. Let  $X$  be a uniform random variable on the interval  $(0, \theta)$ . We observe one value of  $X$ , say  $x$  and want to test  $H_0 : \theta = 1$  versus  $H_a : \theta > 1$ . We decided to reject  $H_0$  if the sample exceeds 0.99. Compute the probability of type 1 error for this test. Draw the power function for this test. Label your plot clearly.
2. Using one (same) example only illustrate 4 different approaches in finding a sufficient statistic in estimating an unknown parameter  $\theta$  of  $f_X(\theta; x)$ .
3. Consider a random sample  $(x_1, x_2, \dots, x_n)$  from  $N(\mu, \sigma^2)$  where  $\sigma^2$  is known. Using two different methods show that the sample mean  $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$  is UMVUE for  $\mu$ . Describe two other methods of finding the UMVUE.

## Probability

1. Prove that  $\sum_{i=1}^n (Z_i - \bar{Z})^2 \sim \chi_{n-1}^2$ .
2. Find the number of tosses required so that the probability that the mean of tosses is between 0.4 and 0.6 is at least 0.9, assuming that the coin is fair. How do you make the probability as close to 0.9? Find the number of tosses required to make the probability as close to 0.9.
3. Let  $(X_1, X_2, \dots, X_n)$  be iid from  $f(x, \theta) = \theta^{-x} I_{(0, \infty)}(x)$  where  $\theta > 0$ . Compare asymptotic distribution of the sample mean  $(\bar{X})$  with the asymptotic distribution of the sample median  $X_{med}$ .
4. Let  $Y_1$  and  $Y_2$  be random samples of size 2 from uniform  $(0, 1)$ . Let  $y_1$  and  $y_2$  be the corresponding order statistics. For  $0 < y_2 < 1$ , find  $f_{y_1|y_2} = Y_{2(y_1|y_2)}$ , the conditional density of  $Y_1$  given  $Y_2 = y_2$ .

## Proposal for a research design

- a) If you are to conduct a survey for data collection, will you propose to use probability sampling?
- b) In choosing the appropriate respondents for the survey, identify the following:
  - Universal/population of the study
  - Sampling frame/sampling units and respondents
  - Variable to measure
- c) With your proposed design, describe how you will go about the analysis of the data by identifying estimation procedure to use and the test of hypothesis to conduct if there are any.
- d) Let us say we are to employ the principle of experimental design through pseudo-random experiment as part of the research methodology. Do you think it is appropriate in this study? Why and why not.
- e) If you are to apply the principle of experimental design, describe how you will achieve randomization, replication and local control in your research methodology.