University of Gothenburg
Chalmers university of Technology

Mathematical sciences

MSA251 and TMS032 Experimental design and sampling

Examination
Date: August 21 2019
Time: 4 hours

Examiner: Anders Muszta.

On call: Felix Held, telephone 031 7725325.

Aids:
- Calculator;
- Course literature in bound format or printed pages from digital format;
- No handwritten notes are allowed in the course literature.
- Highlighting and unmarked post-it notes are allowed in the course literature.
- Course literature are the two books *Sampling Techniques* by William G. Cochran and *Design and Analysis of Experiments* by Douglas C. Montgomery.

Grading:
- The exam comprises five problem-sets that are awarded at most 32 points.
- Solutions should be legible and easy to follow.
- Solutions that are illegible run the risk of not being assessed.
- Notation that is introduced in the solution process must be accompanied with a definition.

TMS032: Grade 3 requires 40 percent of maximum possible points on the exam and grade 4 requires 60 percent and grade 5 requires 80 percent.

MSA251: Grade G requires 40 percent of maximum possible points on the exam and grade VG requires 65 percent.

Problem 1 (5 points)

a) Give a positive and a negative aspect with employing fractional factorial design. (2p)

b) Give a positive and negative aspect with employing factorial design. (2p)

c) Why should one examine statistical properties of residuals when conducting a regression analysis? (1p)

Problem 1

a)

Positive: Able to use limited resources to estimate interaction effects between factors.

Negative: Wasting resources when estimating irrelevant high- order interaction effects.

b)

Positive: Does not waste resources in estimating high-order interaction effects to the same extent as factorial design.

Negative: Introduces confounding between effects (aliasing).

c) To investigate whether data satisfies the assumptions required for ANOVA to be applicable:

1. Independent residuals
2. Constant variance
3. Normal distributed residuals.

Problem 2 (9 points)

For each of the following contexts suggest a possible sampling design as well as an appropriate statistical method of analysis and motivate your choices.
Each sub-problem can be awarded 3 points.

a) In a forest you investigate the effects of tree species and ground preparation on tree biomass. Parts of the forest grow around a lake.

b) Is there a connection between height and diameter of spruce trees in a forest? The forest grows on fairly even terrain with little topographical variation.

c) You are interested in studying all forests in Sweden and want to sample the forests appropriately due to restricted resources. You are particularly interested in whether the connection between height and diameter of trees in Swedish forests depend on tree species?

Problem 2

a) Two categorical explanatory variables (tree species and ground preparation) are thought to affect one continuous response variable (tree biomass). This suggest using a two-way ANOVA. Stratified sampling.

b) One continuous explanatory variable (tree diameter) is thought to affect one continuous response variable (tree height). This suggest using a simple linear regression analysis. Simple random sampling.

c) One continuous explanatory variable (tree diameter) and one categorical explanatory variable (tree species) are thought to affect one continuous response variable (tree height). This suggests using a multiple linear regression, with particular emphasis on the interaction effect of diameter and species on height. Two-stage cluster sampling.

Problem 3 (6 points)

Let $\{X_k\}_{k=1}^{n}$ denote independent observations of a random variable that is assumed to be uniformly distributed on some interval $[\theta, 1]$. The observations are to be used to estimate parameter θ via estimator $\hat{\theta}_n$ defined as

$$\hat{\theta}_n = \min_{1 \leq k \leq n} X_k \ .$$

a) Show that the probability density function of estimator $\hat{\theta}_n$ is

$$\frac{n}{(1-\theta)^n}(1-x)^{n-1} \ , \quad \theta \leq x \leq 1$$

by first determining the cumulative distribution function of estimator $\hat{\theta}_n$. (2p)

Problem 3 (continued)

b) Show that the bias of $\hat{\theta}_n$ is $\frac{1-\theta}{n+1}$. (1p)

c) Use $\hat{\theta}_n$ to create an unbiased estimator of θ. (1p)

d) Determine the mean square error of $\hat{\theta}_n$. (1p)

e) Show that estimator $\hat{\theta}_n$ is weakly consistent. (1p)

Problem 3

a) The cumulative probability distribution function of $\hat{\theta}_n$ is determined by

$$\text{Prob}\left(\min_{1\leq k \leq n} X_k > x\right) = \prod_{k=1}^{n} \text{Prob}(X_k > x) = \left(\frac{1-x}{1-\theta}\right)^n$$

giving the cumulative distribution function

$$F(x) = \begin{cases} 1 & , \quad x > 1, \\ 1 - \left(\frac{1-x}{1-\theta}\right)^n & , \quad \theta \leq x \leq 1, \\ 0 & , \quad x < \theta \end{cases}$$

and its derivative is the desired probability density of $\hat{\theta}_n$.

$$f(x) = F'(x) = \frac{n}{(1-\theta)^n}(1-x)^{n-1} \quad 0 \leq x \leq \theta.$$

b) The expected value of $\hat{\theta}_n$ is the integral

$$\begin{aligned}\mathbb{E}(\hat{\theta}_n) &= \int x \cdot f(x)\,dx = \int_\theta^1 x \cdot \frac{n}{(1-\theta)^n}(1-x)^{n-1}\,dx \\ &= \{\text{Introduce } y = 1-x\} \\ &= \frac{n}{(1-\theta)^n}\int_0^{1-\theta} y^{n-1} - y^n\,dy = \frac{1+n\theta}{n+1}\end{aligned}$$

showing that $\hat{\theta}_n$ is a biased estimator of θ, and its bias is

$$\text{Bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta = \frac{1-\theta}{n+1}.$$

c) The expected value $\mathbb{E}((\hat{\theta}_n)^2)$ is the integral

$$\begin{aligned}\mathbb{E}(\hat{\theta}_n^2) &= \int x^2 \cdot f(x)\,dx = \int_\theta^1 x^2 \cdot \frac{n}{(1-\theta)^n}(1-x)^{n-1}\,dx \\ &= \{\text{Introduce } y = 1-x\} \\ &= \frac{n}{(1-\theta)^n}\int_0^{1-\theta} (1-2y+y^2)y^{n-1}\,dy \\ &= 1 - \frac{2n}{n+1}(1-\theta) + \frac{n}{n+2}(1-\theta)^2,\end{aligned}$$

and consequently the variance becomes

$$\begin{aligned}\text{Var}(\hat{\theta}_n) &= 1 - \frac{2n}{n+1}(1-\theta) + \frac{n}{n+2}(1-\theta)^2 - \left(1 - \frac{n(1-\theta)}{n+1}\right)^2 \\ &= \ldots = \frac{n(1-\theta)^2}{(n+1)^2(n+2)}\end{aligned}$$

resulting in the mean-square error $\sqrt{\text{Var}(\hat{\theta}_n)} = \frac{1-\theta}{n+1} \cdot \sqrt{1 - \frac{2}{n+2}}$.

d) Since the variance decreases to zero as sample size n increases, the estimator $\hat{\theta}_n$ is weakly consistent.

Problem 4 (7 points)

Consider constructing a 2^{7-3} fractional factorial design such that main effect A is not aliased with interaction effect AB.

a) Which are the generators of your design? (1p)

b) Which are the words in your defining relation? (1p)

c) What is the resolution of your design? (1p)

d) Which effects are aliased with the main effect B? (2p)
Is your conclusion consistent with the resolution of your design?

e) Which effects are aliased with the interaction effect AB? (2p)
Is your conclusion consistent with the resolution of your design?

Problem 4

a) Consider a full 2^4 factorial design with effects

$$\begin{array}{cccccc} A & B & C & D & & \\ AB & AC & AD & BC & BD & CD \\ ABC & ABD & ACD & BCD & ABC & ABCD \end{array}$$

and select 3 generators from the *same* high-order family to obtain a fractional design with maximum possible resolution.

b) Select e.g. $E = ABC$ and $F = ABD$ and $G = ACD$ giving the identities $1_1 = ABCE$ and $1_2 = ABDF$ and $1_3 = ACDG$ giving the additional identities $1_1 \cdot 1_2 = CDEF$ and $1_1 \cdot 1_3 = BDEG$ and $1_2 \cdot 1_3 = BCFG$ and $1_1 \cdot 1_2 \cdot 1_3 = AEFG$.

c) The defining relation of our design is

$$\begin{aligned} 1 &= ABCE = ABDF = ACDG = CDEF \\ &= BDEG = BCFG = AEFG \end{aligned}$$

with the seven indicated four-letter words. Our design has resolution IV since this is the length of the shortest word in the defining relation.

d) The defining relation gives the following aliases to main effect B.

$$\begin{aligned} B &= ACE = ABCDG = ADF = DEG \\ &= BCDEF = CFG = ABEFG. \end{aligned}$$

Main effect B is not confounded with any of the other main effects nor with any of the second-order interaction effects, as is to be expected from a design of resolution IV.

e) The defining relation gives the following aliases to second-order interaction effect AB.

$$\begin{aligned} AB &= CE = DF = BCDG = ABCDEF \\ &= ADEG = ACFG = BEFG \end{aligned}$$

Second-order interaction effect AB is not confounded with any of the main effects, but is confounded with some of the second-order interaction effects, as is to be expected from a design of resolution IV.

Problem 5 (5 points)

a) Show that when sampling from a finite population using any sampling design, the sum of all first-order inclusion probabilities of the population equals the size of the sample. (2p)

b) Show that Horvitz-Thompson estimators of the population mean are unbiased estimators. (2p)

c) Give an example of a Horvitz-Thompson estimator of the population total when picking a sample of size 3 from a population of size 6. (1p)

Problem 5

a) From a population of size N we collect n objects into a sample (a random set S) according to a sampling design whose first-order inclusion probabilities are $\mathrm{Prob}(S \ni k) = \pi_k$. Each population object (k) is either included ($S \ni k$) in or discarded ($S \not\ni k$) from the sample.

$$n = \sum_{k=1}^{N} 1_{\{S \ni k\}}.$$

Taking the expected value of this constant random variable using the specified sampling design results in the desired formula.

$$n = \mathbb{E}(n) = \sum_{k=1}^{N} \mathbb{E}(1_{\{S \ni k\}}) = \sum_{k=1}^{N} \pi_k.$$

b) Population mean is $\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} y_k$ and Horvitz-Thompson estimators of population mean are $\bar{y}_{HT} = \frac{1}{N}\sum_{k=1}^{N}(y_k/\pi_k)1_{\{S\ni k\}}$. The expected value of the estimators equal the population mean.

$$\begin{aligned}\mathbb{E}(\bar{y}_{HT}) &= \frac{1}{N}\sum_{k=1}^{N}(y_k/\pi_k)\mathbb{E}(1_{\{S\ni k\}}) = \frac{1}{N}\sum_{k=1}^{N}(y_k/\pi_k)\pi_k \\ &= \frac{1}{N}\sum_{k=1}^{N} y_k = \bar{Y}.\end{aligned}$$

c) For Horvitz-Thompson estimators to be defined the sample design must not assign zero first-order inclusion probability to any population object, i.e. $\pi_k > 0$ for all $k \in \{1, 2, \ldots, 6\}$. Summing all inclusion probabilities must equal the sample size, $3 = \sum_{k=1}^{6}\pi_k$.

SRS If a simple random sample is selected then all π_k are equal and the constraints force the conclusion $\pi_k = 3/6 = 1/2$ resulting in the Horvitz-Thompson estimator of population total

$$y_{HT} = \sum_{k=1}^{6} 2y_k 1_{\{S\ni k\}}.$$

Strat An example of a stratified sampling design is $\pi_1 = \pi_2 = 1$ and $\pi_3 = \pi_4 = \pi_5 = \pi_6 = 0.25$ resulting in the estimator

$$y_{HT} = y_1 + y_2 + \sum_{k=3}^{6} 4y_k 1_{\{S\ni k\}}.$$