Exam in MSA251/TMS032 Experimental design and sampling.
March 17, 2021
Examiner: Marina Axelson-Fisk, tel. 031-772 4996.
Allowed aids: all aids except consulting another person

The maximum score is 30 points.
Grading limits for GU: 12p (G), 22p (VG), and for Chalmers: 12p (3), 18p (4), 24p (5).
Submit complete and well-motivated solutions for each problem.

1. Briefly explain the following concepts:

   a) Mixed effects model (2p)
   b) Split-plot design (2p)
   c) Stratified versus cluster sampling (2p)

2. In a 2^{5-2} fractional factorial design the influence of four different catalysts A, B, C, D and E on temperature was investigated. The following results were obtained

A	B	C	D	E	Temp increase
$-$	$-$	$-$	$-$	$+$	12
$+$	$-$	$-$	$+$	$+$	16
$-$	$+$	$-$	$+$	$+$	11
$+$	$+$	$-$	$-$	$+$	18
$-$	$-$	$+$	$+$	$+$	10
$+$	$-$	$+$	$-$	$+$	20
$-$	$+$	$+$	$-$	$+$	15
$+$	$+$	$+$	$+$	$+$	11

   a) What are the defining relations for this experiment? (1p)
   b) Give the confounding patterns of the main effects. (2p)
   c) What is the resolution of the design? (2p)
   d) Estimate the main effect of catalyst A and interaction AB. (2p)

3. In an experiment the amount of nickel and manganese in an alloy were changed in order to analyze the breaking strength of a component. A 3 × 2 design was created with two runs in each factor combination:

Ni (%)	Mn (%)	Strength (ft-lb)	
0	1	28	30
2	1	41	43
4	1	55	55
0	2	48	52
2	2	35	37
4	2	39	41

   The alloy expert of the company believes in a regression model with a mean, the two main effects, and the interaction.

   a) Formulate the model and state the assumptions needed. (2p)
   b) Fill in the missing values in the ANOVA table below: (2p)

Source	SS	df	MSE	F
Regression	21944	?	?	
Residual	?	?	?	
Total	22068	?		

c) Is the result significant? (1p)

d) Give an estimate of the error variance. (2p)

4. Suppose we draw 2 sampling units from a population having 3 units, y_1, y_2, y_3. Assume that the following estimator of the sample mean is used, depending upon which sampling method was used:

$$\tilde{y} = \begin{cases} \dfrac{y_1}{2} + \dfrac{y_2}{2} & \text{if sample is } (y_1, y_2) \\ \dfrac{y_1}{2} + \dfrac{2y_3}{3} & \text{if sample is } (y_1, y_3) \\ \dfrac{y_2}{2} + \dfrac{y_3}{3} & \text{if sample is } (y_2, y_3) \end{cases}$$

a) Verify whether $\tilde{y}$ is an unbiased estimator or not. (2p)

b) Compute the variance of $\tilde{y}$. (3p)

5. A sample to estimate the number of orchards of apple in a district in northern India is conducted. Four strata, A, B, C and D, of towns are formed, and the following data were collected

Stratum	Stratum size	Sample size	sample total	sample variance	S_h^2
A	275	15	49	9.64	8
B	146	10	152	88.84	2
C	93	12	169	205.91	15
D	62	11	362	192.69	20

a) Compute the allocations using equal, proportional, and optimal allocation. (2p)

b) Compute the variances of the estimated means. Which estimator is best? (3p)

Good luck!

Solutions:

1. Brief concept explanations:

   a) **Mixed effects model:** contains both fixed effect factors and random factors. In a fixed effect factor, the levels are fixed while in random factors the levels are random variables.

   b) **Split-plot design:** a generalization of factorial designs used when an experiment cannot be completely randomized, and the factors are randomized at two levels. The first factor is randomized across the whole plot (whole-plot treatment). The whole plot is then split into subplots, and the second factor is randomized over each subplot (subplot treatment). Since the randomization has two levels, the two treatments have separate experimental errors.

   c) **Stratified versus cluster sampling:** Stratified sampling is used when a heterogeneous population can be divided into homogeneous subgroups. Cluster sampling is used when the population can be divided into clusters, where the members within a cluster are heterogeneous, but the clusters are "homogeneous" in the sense that they contain the same type of heterogeneous group of members.

2. A 2^{5-2} design:

   a) The generators of the design are: $D = ABC$ and $E = ABCD$, so the defining relations become

$$I = ABCD = ABCDE = E$$

   b) Confounding patterns of main effects

$$\begin{aligned} A &= BCD = BCDE = AE \\ B &= ACD = ACDE = BE \\ C &= ABD = ABDE = CE \\ D &= ABC = ABCE = DE \\ E &= ABCD = ABCDE \end{aligned}$$

   c) Since main effects are confounded with second-order interactions, the resolution is III (3).

   d) Effect estimates of A and AB

$$A = \frac{1}{4}(-12 + 16 - 11 + 18 - 10 + 20 - 15 + 11) = \frac{17}{4} = 4.25$$

$$AB = \frac{1}{4}(12 - 16 - 11 + 18 + 10 - 20 - 15 + 11) = -\frac{11}{4} = -2.75$$

3. Breaking strength of nickel and manganese alloys:

   a) Regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \epsilon_i$$

   The random errors ϵ_i are assumed to be independent with distribution $N(0, \sigma^2)$.

   b) ANOVA

Source	SS	df	MSE	F
Regression	21944	**3**	**7314.67**	**471.91**
Residual	**124**	**8**	**15.5**	
Total	22068	**11**		

c) We are testing the hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_{12} = 0$$

The F-statistic is distributed as $F_{3,8}$ and for $\alpha = 0.05$ we get $F_{3,8} = 8.85$, and we reject H_0 since $F > F_{3,8}$. Our result is **very** significant.

d) The variance σ^2 can be estimated by MS_E. I.e. $\hat{\sigma}^2 = 15.5$.

4. Estimator $\tilde{y}$

a) is unbiased if $E[\tilde{y}] = \bar{y}$.

$$E[\tilde{y}] = \frac{1}{3}\left(\frac{y_1}{2} + \frac{y_2}{2}\right) + \frac{1}{3}\left(\frac{y_1}{2} + \frac{2y_3}{3}\right) + \frac{1}{3}\left(\frac{y_2}{2} + \frac{y_3}{3}\right) =$$
$$= \frac{1}{3}\left(\frac{2y_1}{2} + \frac{2y_2}{2} + \frac{3y_3}{3}\right) = \frac{1}{3}(y_1 + y_2 + y_3) = \bar{y}$$

so $\tilde{y}$ is an unbiased estimator of the mean.

b) Variance:

$$\mathrm{Var}(\tilde{y}) = E[(\tilde{y} - E[\tilde{y}])^2] = E[\tilde{y}^2] - (E[\tilde{y}])^2$$

So we need to compute $E[\tilde{y}^2]$.

$$E[\tilde{y}^2] = \frac{1}{3}\left(\left(\frac{y_1}{2} + \frac{y_2}{2}\right)^2 + \left(\frac{y_1}{2} + \frac{2y_3}{3}\right)^2 + \left(\frac{y_2}{2} + \frac{y_3}{3}\right)^2\right) =$$

…

$$\mathrm{Var}(\tilde{y}) = \frac{1}{18}\left(y_1^2 - y_1y_2 + y_2^2 - 2y_2y_3 + \frac{4}{3}y_3^2\right)$$

5. We add the assumption that we'll use the same sample size $n = 48$.

a) Equal allocation: $n/L = 48/4 = 12$

Proportional allocation:

$$n_h = N_h \frac{n}{N}$$

Optimal allocation:

$$n_h = n \frac{N_h S_h}{\sum N_h S_h}$$

	Equal	**Proportional**	**Optimal**
n_A	12	23	23
n_B	12	12	6
n_C	12	8	11
n_D	12	5	8

b) The variance in stratified sampling is given by

$$\mathrm{Var}(\bar{y}_{\mathrm{st}}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h)$$

Equal allocation

Stratum	N_h	n_h	f_h	W_h	S_h^2
A	275	12	0.044	0.477	8
B	146	12	0.082	0.253	2
C	93	12	0.129	0.161	15
D	62	12	0.194	0.107	20

With $n_h = n/L$ the variance formula reduces to

$$\mathrm{Var}(\bar{y}_{\mathrm{st}}) = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 = \cdots = 0.199$$

Proportional allocation

Stratum	N_h	n_h	f_h	W_h	S_h^2
A	275	23	0.084	0.477	8
B	146	12	0.082	0.253	2
C	93	8	0.086	0.161	15
D	62	5	0.081	0.107	20

With $n_h = N_h \frac{n}{N}$ the variance formula reduces to

$$\mathrm{Var}(\bar{y}_{\mathrm{st}}) = \frac{1 - f}{n}\sum_{h=1}^{L} W_h S_h^2 = \cdots = 0.170$$

Optimal allocation

Stratum	N_h	n_h	f_h	W_h	S_h^2
A	275	23	0.084	0.477	8
B	146	6	0.041	0.253	2
C	93	11	0.118	0.161	15
D	62	8	0.129	0.107	20

$$\mathrm{Var}(\bar{y}_{\mathrm{st}}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h) = \cdots = 0.150$$

Optimal allocation gets the lowest variance.