Statistics 5100 Exam #1 March 3rd, 2025 Instructor: Dr. Genschel

Good Luck! Name Sabrina Morse - Student A

Partner : Sam Olson

Please answer each question in the space provided, and **show all your work**. Credit cannot be given if work is not shown. Ask for extra paper if you need it. Good luck.

1. If I asked your best friend about one of your character strengths, what would they say?

---

2. **Technical and Conceptual.** A study was conducted to see if there are differences in the quality of steel produced by $m = 3$ machines. It is also felt that there may be differences in the feedstock obtained from $l = 3$ different suppliers. Nine samples of feedstock were selected from each supplier, and $n = 3$ samples were randomly assigned to each machine.

   In class we talked extensively about two different types of models to analyze data from this type of experiment. Most recently you encountered both models on Homework 5, Problem 1.

   (a) Name each statistical model. The order in which you name the models that does not matter!

   Model 1: Cell Means Model Model 2: Additive Model

   (b) For each, write out the statistical model to describe the quality y_{ijk} of the k^{th} piece of steel produced by the i^{th} machine using feedstock from supplier j.

   - Do **not** write the model in matrix or vector form!
   - Be sure to properly define all parameters, use clear and complete notation and state the **necessary assumptions** such that the model is also a **Gauss-Markov Model** with **Normal Errors**.
   - Your answer should look like one of the slides in the notes when we introduce a new statistical model.

   Model 1: $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$

   μ = combined mean from both factors

   α_i = machine main effect

   β_j = supplier main effect

   γ_{ij} = interaction effect between machine & supplier

   ϵ_{ijk} = random error

   Assumptions:

   - $\epsilon_{ijk} \sim N(0, \sigma^2)$
   - independence of random errors

Name _____

Model 2: $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

μ = overall mean

α_i = machine main effect

β_j = supplier main effect

ϵ_{ijk} = random error

Assumptions:

- $\epsilon_{ijk} \sim N(0, \sigma^2)$
- independence of random error
- no interaction between treatments

(c) **In the context of the data**, what is the main difference between both types of statistical models in terms of model complexity?

Cell means is more complex than additive model, since it includes a term for potential interaction between machines and suppliers

Name _____

3. Suppose that $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2, \boldsymbol{I})$, and let $q_i = \boldsymbol{y}^\top \boldsymbol{A}_i \boldsymbol{y}, i = 1, 2$ where

$$\boldsymbol{A}_1 = \frac{1}{3}\mathbf{1}\mathbf{1}^\top \quad \text{and} \quad \boldsymbol{A}_2 = \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Note that $\mathbf{1}$ is a vector of three ones, i.e., $\mathbf{1} = (1 \;\; 1 \;\; 1)^\top$.

(a) Determine the distribution of each quadratic form and explain how you know that this indeed is the distribution of $q_i, i = 1, 2$.

to use quadratic forms, A must be idempotent…

A_1: $A_1^2 = (\frac{1}{3}\mathbf{1}\mathbf{1}^\top)(\frac{1}{3}\mathbf{1}\mathbf{1}^\top) = \frac{1}{3}\mathbf{1}(1^\top\mathbf{1})\mathbf{1}^\top = \frac{1}{3}\mathbf{1}\mathbf{1}^\top = A_1$ ✓

A_2: $A_2^2 = \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = A_2$ ✓

$\frac{y^\top A z y}{\sigma^2} = \frac{y^\top A \sigma^2 I y}{\sigma^2} = y^\top A y \sim \sigma^2 \chi^2_m\left(\frac{\mu^\top A \mu}{2\sigma^2}\right)$ — will use this for form of quadratic eq.

quadratic form eq: $y^\top A y \sim \chi^2_m\left(\frac{\mu^\top A \mu}{2}\right)$ (m = rank of A)

$A_1 = \frac{1}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \Rightarrow$ rank $= 1$

$A_2 = \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 1/2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \Rightarrow$ rank $= 1$

$\therefore \; y^\top A_i y \sim \chi^2_1\left(\frac{\mu^\top A \mu}{2}\right)$ → now find ncp for each A_i

$\mu^\top = [\mu_1 \;\; \mu_2 \;\; \mu_3]$

A_1: $A_1\mu = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1+\mu_2+\mu_3 \\ \mu_1+\mu_2+\mu_3 \\ \mu_1+\mu_2+\mu_3 \end{bmatrix}$

$\mu^\top A_1 \mu = [\mu_1 \;\; \mu_2 \;\; \mu_3]\begin{bmatrix} \mu_1+\mu_2+\mu_3 \\ \mu_1+\mu_2+\mu_3 \\ \mu_1+\mu_2+\mu_3 \end{bmatrix} = (\mu_1+\mu_2+\mu_3)^2$

so $y^\top A_1 y \sim \sigma^2 \chi^2_1\left(\frac{(\mu_1+\mu_2+\mu_3)^2}{6\sigma^2}\right)$

A_2: $A_2\mu = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1-\mu_2 \\ \mu_2-\mu_1 \\ 0 \end{bmatrix}$

$\mu^\top A_2 \mu = [\mu_1 \;\; \mu_2 \;\; \mu_3]\begin{bmatrix} \mu_1-\mu_2 \\ \mu_2-\mu_1 \\ 0 \end{bmatrix} = \mu_1(\mu_1-\mu_2) + \mu_1(\mu_2-\mu_1) + \mu_2(\mu_1-\mu_2) + \mu_2(\mu_2-\mu_1)$

$= \mu_1^2 - \mu_1\mu_2 + \mu_1\mu_2 - \mu_1^2 + \mu_2\mu_1 - \mu_2^2 + \mu_2^2 - \mu_2\mu_1 = 0$

so $y^\top A_2 y \sim \sigma^2 \chi^2_1(0)$

we know these are distributions of q_i for $i = 1, 2$ since the normality of y ensures the quadratic form will follow χ^2 distribution of rank m & ncp of $\frac{\mu^\top A_i \mu}{2}$

(b) Show that q_1 and q_2 are independent.

show $A_1 \perp A_2$ = independence of q_1 & $q_2 \Rightarrow A_1 \cdot A_2 = 0$ means orthogonal

$$\begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \cdot \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = 1/3 - 1/3 + 0 - 1/3 + 1/3 + 0 + 0 + 0 + 0 = 0$$

since $A_1 \cdot A_2 = 0 \Rightarrow q_1$ & q_2 are independent ✓

Name _____

4. Consider a one-way ANOVA model with two levels and two observations at each level,

$$\mathrm{E}(y_{ij}) = \mu + \alpha_i, \quad i, j = 1, 2$$

(a) Is α_1 estimable? Show work to justify why it is or is not estimable.

(b) Provide a quantity that is estimable._____

The remaining questions refer to any general linear model as discussed in class. Thus, provide answers for a general $\boldsymbol{X}$ instead of referring to the particular $\boldsymbol{X}$ defined above.

(c) **True** or **False** Circle the appropriate choice. The expected value of any observation is only estimable when $\boldsymbol{X}$ has full column rank.

(d) The set of vectors $\boldsymbol{c}$ for which $\boldsymbol{c}^\top\boldsymbol{\beta}$ is estimable forms a vector space. Specify the vector space.

Answer: _____

(e) **Fill in the blank**.
The column rank of a model matrix $\boldsymbol{X}$ is always _____
the number of linearly independent vectors that span the vector space in part (d).

(f) What is the relationship/connection between the column rank of $\boldsymbol{X}$ and the estimability of $\boldsymbol{\beta}$? Answer using a short sentence.

Name _____

5. Consider the following linear model with $n = 5$ observations.

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$$

Note that the columns of $\boldsymbol{X}$ are perpendicular so that $\boldsymbol{X}^\top \boldsymbol{X}$ is **diagonal**.

(a) In a Gauss-Markov version of this model, which of the parameters, $\beta_1, \beta_2, \beta_3$ can be estimated with greatest precision? Explain carefully.

(b) Suppose $\boldsymbol{y}$ is such that SSE=3 and $\widehat{\boldsymbol{\beta}} = (5 \;\; 6 \;\; 2)^\top$. Consider an analysis under the Gauss Markov model with Normal errors and the following two null hypotheses:

$$\mathrm{H}_{0,1} : \mathrm{E}(y_1) = \mathrm{E}(y_2) \qquad \text{and} \qquad \mathrm{H}_{0,2} : \mathrm{E}(y_1) = \mathrm{E}(y_3)$$

i. Write $\mathrm{H}_{0,1}$ and $\mathrm{H}_{0,2}$ in testable form $\mathrm{H}_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$ by identifying an appropriate matrix $\boldsymbol{C}$. (**Hint:** Start out by expressing the each expected value as a function of $\boldsymbol{\beta}$ given $\boldsymbol{X}$ and $\boldsymbol{\beta}$ as defined above.)

ii. Based on $\boldsymbol{C}$ compute an F statistic for testing H_0 (you need not do the arithmetic, but plug correct numbers into a correct formula).

iii. Specify the reference distribution of F under the null hypothesis.

6. Consider a completely randomized experiment in which a total of 10 freshly cut Gerber daisies were placed into 10 vases (one daisy per vase). The Gerber daisies were randomly assigned to five treatment groups with two Gerber daisies in each treatment group. The treatment corresponds to the amount of a chemical compound added to the water in each vase. Of interest is the longevity of the Gerber daisies measured in days.

Treatment	1	2	3	4	5
Amount of compound (g)	0	2	4	10	16

Suppose for $i = 1, ..., 5$ and $j = 1, 2$, y_{ij} denotes the longevity in days of the study of the j^{th} Gerber daisy from treatment group i. Furthermore, suppose

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

where the μ_i are unknown parameters and the ε_{ij} terms are $\mathcal{N}(0, \sigma^2)$ for some unknown $\sigma^2 > 0$.

Use the R code and partial output provided with this exam to answer the following questions.

Name _____

(a) For the first model fit in R, called `M1`, specify the model matrix $\boldsymbol{X}$ used by R.

(b) Consider the following information from the output associated with model `M1`:

```
F-statistic: 30.84 on 4 and 5 DF,  p-value: 0.001019
```

Specify the null and alternative hypothesis associated with this test:

(c) Provide the BLUE of μ_2: _____

(d) What is the standard error of the BLUE of μ_2? _____

(e) Provide the BLUE of $\mu_1 - \mu_2$: _____

(f) What is the standard error of the BLUE of $\mu_1 - \mu_2$? _____

(g) What is the value of $\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P_1})\boldsymbol{y}$, where $\boldsymbol{y}$ denotes the vector containing the values of `longevity`?

Answer: _____

(h) Provide the value of the F-statistic, numerator and denominator df, and the p-value associated with the following ANOVA table:

```
> anova(M1)
Analysis of Variance Table

Response: longevity
          Df    Sum Sq   Mean Sq   F value   Pr(>F)
amt       ___   94.398             ___       ___
Residuals ___    3.826
```

Name _____

(i) Look at the output associated with Model 2, `M2`.

i. Fill in the missing entries in the ANOVA table produced by the R command `anova(M2, M1)`.

```
> anova(M2, M1)
Analysis of Variance Table

Model 1: longevity ~ amount
Model 2: longevity ~ amt

  Res.Df     RSS      Df       Sum of Sq        F         Pr(>F)
1   ___      ___
2   ___      ___      ___      ___ (*)           ___         0.006412 **
```

ii. Provide an interpretation of Sum of Squares in part (i). This is the value denoted by `(*)`.

iii. Provide a conclusion in the context of the data about the null hypothesis that is tested in part (i).

R code and output for Question 6: Name _____

```
> amount<-c(0, 0, 2, 2, 4, 4, 10, 10,  16, 16)
> plot(amount, longevity, ylim = c(0,14))
> amt=as.factor(amount)
> M1<-lm(longevity~amt)
> summary(M1)

Call:
lm(formula = longevity ~ amt)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3200     0.6185   3.751 0.013281 *
amt2          2.4100     0.8747   2.755 0.040063 *
amt4          6.0850     0.8747   6.957 0.000943 ***
amt10         8.2400     0.8747   9.420 0.000227 ***
amt16         7.0300     0.8747   8.037 0.000482 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8747 on 5 degrees of freedom
Multiple R-squared:  0.9611,Adjusted R-squared:  0.9299
F-statistic: 30.84 on 4 and 5 DF,  p-value: 0.001019

> anova(M1)
Analysis of Variance Table

Response: longevity
          Df Sum Sq Mean Sq F value   Pr(>F)
amt          94.398
Residuals     3.826
---

> is.numeric(amount)
[1] TRUE

> M2<-lm(longevity~amount)

> anova(M2)
Analysis of Variance Table

Response: longevity
          Df Sum Sq Mean Sq F value   Pr(>F)
amount       60.467
Residuals    37.757
```