Apply Statistical Analysis

Michael Dang

University of Missouri - Kansas City

April 7, 2025

Assignment 3

Solutions are to be due on April 7th.
The source code can be found at: here

1. Solution.
   (a) The response variable = yields of the wheat measure in grams.
   (b) The response variable is a quantitative (numerical) variable and it is continuous.
   (c)
      - Because each plot was assigned one of 4 fertilizer treatments by experimenters.
      - The goal is to see what type of fertilizer change in wheat yield.
   (d)
      - Treatments = fertilizer treatments ($t = 4$)
      - Observational units = individual wheat plants ($OU = 40 \quad plants$)
      - Experimental units = the plots of land ($EU = 20 \quad plots$)
   (e) Assume the EUs are homogenous, (i.e. the plots of land are under similar weather conditions) □

2. Solutions
   (a)
      - $n = 4$ (number of observations per treatment or replications).
      - $t = 3$ (number of treatment)
      - $N = n \cdot t = 12$ (total number of observations)

(b)

- $$
\begin{aligned}
y_{i=1,\cdot} = \sum_{j=1}^{n_{i=1}=4} &= y_{11} + y_{12} + y_{13} + y_{14} \\
&= 1.83 + 2.01 + 1.94 + 1.79 \\
&= 7.569
\end{aligned}
$$

- $$
\begin{aligned}
\overline{y}_{i=1,\cdot} &= \frac{y_{i=1,\cdot}}{n_{i=1}} \\
&= \frac{7.569}{4} \\
&= 1.892
\end{aligned}
$$

- $$
\begin{aligned}
y_{i=2,\cdot} = \sum_{j=1}^{n_{i=2}=4} &= y_{21} + y_{22} + y_{23} + y_{24} \\
&= 1.74 + 1.68 + 1.85 + 1.72 \\
&= 6.989
\end{aligned}
$$

- $$
\begin{aligned}
\overline{y}_{i=2,\cdot} &= \frac{y_{i=2,\cdot}}{n_{i=2}} \\
&= \frac{6.989}{4} \\
&= 1.747
\end{aligned}
$$

- $$
\begin{aligned}
y_{i=3,\cdot} = \sum_{j=1}^{n_{i=3}=4} &= y_{31} + y_{32} + y_{33} + y_{34} \\
&= 1.53 + 1.60 + 1.56 + 1.62 \\
&= 6.31
\end{aligned}
$$

- $$
\begin{aligned}
\overline{y}_{i=3,\cdot} &= \frac{y_{i=3,\cdot}}{n_{i=3}} \\
&= \frac{6.31}{4} \\
&= 1.577
\end{aligned}
$$

(c)

- $$\begin{aligned} y_{\cdot,\cdot} &= \sum_{j=1}^{3} y_{i,\cdot} \\ &= y_{1,\cdot} + y_{2,\cdot} + y_{3,\cdot} \\ &= 7.569 + 6.989 + 6.31 \\ &= 20.868 \end{aligned}$$

- $$\begin{aligned} \bar{y}_{\cdot,\cdot} &= \frac{y_{\cdot,\cdot}}{N} \\ &= \frac{20.868}{12} \\ &= 1.738 \end{aligned}$$

(d)

- $$\begin{aligned} s_1^2 &= \frac{\sum_{j=1}^{n_1}(y_{1,j} - y_{1,\cdot})^2}{n_1 - 1} \\ &= \frac{(1.83 - 1.892)^2 + (2.01 - 1.892)^2 + (1.94 - 1.892)^2 + (1.79 - 1.892)^2}{4 - 1} \\ &= \frac{0.0304}{3} \\ &= 0.0101 \end{aligned}$$

- $$\begin{aligned} s_2^2 &= \frac{\sum_{j=1}^{n_2}(y_{2,j} - y_{2,\cdot})^2}{n_2 - 1} \\ &= \frac{(1.74 - 1.747)^2 + (1.68 - 1.747)^2 + (1.85 - 1.747)^2 + (1.72 - 1.747)^2}{4 - 1} \\ &= \frac{0.0158}{3} \\ &= 0.0052 \end{aligned}$$

- $$\begin{aligned} s_3^2 &= \frac{\sum_{j=1}^{n_3}(y_{3,j} - y_{3,\cdot})^2}{n_3 - 1} \\ &= \frac{(1.53 - 1.577)^2 + (1.6 - 1.577)^2 + (1.56 - 1.577)^2 + (1.62 - 1.577)^2}{4 - 1} \\ &= \frac{0.0048}{3} \\ &= 0.0016 \end{aligned}$$

- $$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{t=3}(n_i-1)s_i^2}{\sum_{i=1}^{t=3}(n_i-1)} \\ &= \frac{(4-1)\cdot(0.0101)+(4-1)\cdot(0.0052)+(4-1)\cdot(0.0016)}{12-3} \\ &= \frac{0.0506}{9} \\ &= 0.0056 \end{aligned}$$
- The estimate of $\sigma^2 = s^2 = 0.0056$
- $$y_{i,j} = \mu_i + \epsilon_{i,j}$$
  - where: $i = 1,2,3$ and $j = 1,2,3,4$
  - where:
    $y_{i,j}$ = the bulk density (g/cm^3) on the j^{th} tracts of land under the i^{th} treatment.
    μ_i = the mean bulk density (g/cm^3) under the i^{th} treatment.
    $\epsilon_{i,j}$ = random experimental error on the j^{th} tracts of land under the i^{th} treatment.

(e)

- $$\epsilon_{i,j} \overset{iid}{\sim} N(0,\sigma^2)$$
- The bulk densities under each treatment are from the Normal population.
- The population of bulk densities are independent.
- The variances of the bulk densities under each population are equal.

(f)

- $$\begin{aligned} SS_T = \sum_{i=1}^{3}\sum_{j=1}^{4}(y_{i,j}-\bar{y}_{\cdot,\cdot})^2 &= \sum_{i=1}^{3}[(y_{i,1}-\bar{y}_{\cdot,\cdot})^2+(y_{i,2}-\bar{y}_{\cdot,\cdot})^2+(y_{i,3}-\bar{y}_{\cdot,\cdot})^2+(y_{i,4}-\bar{y}_{\cdot,\cdot})^2] \\ &= (y_{1,1}-\bar{y}_{\cdot,\cdot})^2+(y_{1,2}-\bar{y}_{\cdot,\cdot})^2 \\ &+ (y_{1,3}-y_{\cdot,\cdot})^2+(y_{1,4}-\bar{y}_{\cdot,\cdot})^2+ \\ &... \\ &+ (y_{3,3}-\bar{y}_{\cdot,\cdot})^2+(y_{3,4}-\bar{y}_{\cdot,\cdot})^2 \\ &= (1.83-1.738)^2+(2.01-1.738)^2+ \\ &+ (1.94-1.738)^2+(1.79-1.738)^2+ \\ &... \\ &+ (1.56-1.738)^2+(1.62-1.738)^2 \\ &= 0.2501 \end{aligned}$$

- $SS_E = 0.0506$
-

$$\begin{aligned} MS_{trt} &= \frac{SS_{trt}}{t-1} \\ &= \frac{SS_T - SS_E}{t-1} \\ &= \frac{0.2501 - 0.0506}{3-1} \\ &= \frac{0.1994}{2} \\ &= 0.0997 \end{aligned}$$

-

$$\begin{aligned} MS_E &= \frac{SS_E}{N-t} \\ &= \frac{0.0506}{12-3} \\ &= 0.0056 \end{aligned}$$

- Hypothesis test:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu \quad vs \quad H_A : \text{at least one pair } \mu_i \neq \mu_j \quad for \quad i \neq j$$

- Assume $\alpha = 0.05$
- Observe test stat:

$$\begin{aligned} F_{teststat} &= \frac{MS_{trt}}{MS_E} \\ &= \frac{0.0997}{0.0056} \\ &= 17.805 \end{aligned}$$

- $F_{(t-1),(N-t)} = F_{2,9} = 4.256$
- Hence, the test stat falls into the rejection region.
- Therefore, reject H_0.
- Thus, with 95% confidence, we have enough evidence to support there is at least one pair difference in the mean of bulk density.

□

3. Solutions

   (a) • The SAS output is in Figure 1:

The GLM Procedure

Dependent Variable: solid_bulk_density

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.19886667	0.09943333	17.47	0.0008
Error	9	0.05122500	0.00569167		
Corrected Total	11	0.25009167			

Figure 1: SAS Output, ANOVA table

(b)
- Checking the equal variance assumption:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2 \quad vs \quad H_A : \text{at least one pair } \sigma_i \neq \sigma_j \quad for \quad i \neq j$$

- Levene's test p-value: $0.1389 > 0.05 \quad (\alpha)$
- Hence, don't reject H_0
- Therefore, the equal variance assumption is satisfied.

The GLM Procedure

Levene's Test for Homogeneity of solid_bulk_density Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
treatment	2	0.00510	0.00255	2.48	0.1389
Error	9	0.00927	0.00103		

Figure 2: SAS Output for Levene's p-value

- Checking the Normality assumption:
- K-S test p-value: $0.15 > 0.05 \quad (\alpha)$
- Hence, the normality assumption is satisfied.

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.11866211	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.01998963	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.16011608	Pr > A-Sq	>0.250

Figure 3: SAS Output for K-S's p-value

(c) Since, $0.0008 < 0.05$ (p-value $< \alpha$ respectively, from Figure 1), hence with 95% confidence, we don't have enough evidence to support there at least one pair difference in mean of bulk density, i.e. $(\mu_1 = \mu_2 = \mu_3 = \mu)$

(d) Since we consider $\alpha = 0.05$, hence with 95% confidence, the only difference is in the mean of continuous grazing and the mean of two-week grazing with one-week rest, i.e. $(\mu_1 \neq \mu_3)$. If we consider $\alpha = 0.01$ then with 99% confidence, there are differences in the mean of continuous grazing and the mean of two-week grazing with one-week rest, i.e. $(\mu_1 \neq \mu_2)$; and the mean continuous grazing and the mean of two-week grazing with two-week rest, i.e. $(\mu_1 \neq \mu_2)$.

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

treatment	solid_bulk_density LSMEAN	LSMEAN Number
1_week_r	1.74750000	1
2_week_r	1.57750000	2
continou	1.89250000	3

Least Squares Means for effect treatment
Pr > \|t\| for H0: LSMean(i)=LSMean(j)

Dependent Variable: solid_bulk_density

i/j	1	2	3
1		0.0270	0.0561
2	0.0270		0.0006
3	0.0561	0.0006	

Figure 4: SAS Output for pairwise comparison

□

4. Solutions

(a)

Source	df	SS	MS	F
Treatments	3	126	42	$\frac{42}{16} = 2.625$
Error	20	320	16	
Total	23	446		

(b) The number of treatment is: $t - 1 = 3 \rightarrow t = 4$

(c)
- Total number of observations: $N - 1 = 23 \rightarrow N = 24$
- Number of replications per treatment: $n = \frac{N}{t} = \frac{24}{4} = 6$

(d)
- Assume: $\alpha = 0.05$
- Hypothesis test:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu \quad vs \quad H_A : \text{at least one pair } \mu_i \neq \mu_j \quad for \quad i \neq j$$

- Observe test stat: $F_{teststat} = 2.625$
- $F_{(t-1),(N-t)} = F_{3,20} = 3.098$
- Hence, the test stat doesn't fall into the rejection region.
- Therefore, fail to reject H_0.
- Thus, with 95% confidence, we have enough evidence to support the mean responses to each treatment are the same.

$\square$