Mid-Term for ISYE 6413 at GT

Was this exam helpful? YES NO +0 Helpful

Favorite Report

No description

Related Documents

Study Stream

P W.

Create a comment

NO PC

ISyE6413

First Midterm Examination February 19th, 2008

(Total : 50 points)

Name :

Problem	1	2	3	4	Total
Max Points	16	12	9	13	50
Your score					

Problem 1 (16 pts)

The Kenton Company wished to test 4 different package designs for a new breakfast cereal. Twenty stores, with approximately equal sales volumes, were selected as the experimental units. Each store was randomly assigned one of the package designs, with each package design assigned to 5 stores. A fire occurred in one store during the study period, so this store had to be dropped from the study. Sales, in number of cases, were observed for the study peroid, and the results are recorded in Table 1. The characteristics of 4 package designs are shown in Table 2. The data is analyzed using one-way layout with fixed effects.

Table 1: Number of Cases Sold by Stores for Each of Four Package Designs

	Package Design			
	1	2	3	4
	11	12	23	27
	17	10	20	33
	16	15	18	22
	14	19	17	26
	15	11		28
Mean	$\bar{y}_1 = 14.6$	$\bar{y}_2 = 13.4$	$\bar{y}_3 = 19.5$	$\bar{y}_4 = 27.2$
Standard Deviation	2.3	3.6	2.6	4.0

Table 2: Characteristics of Four Package Designs

Package Design	Characteristics
1	3 color, with cartoons
2	3 color, without cartoons
3	5 color, with cartoons
4	5 color, without cartoons

(a) (7 pts) Fill up the blanks in the ANOVA table. (Hint: the treatment sum of squares can be determined from the sample means and the mean-squared error can be calculated by pooling the sample variances.)

Table 3: One-way ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Between designs	3	588.2	196.1	18.59
Error	15	157.2	10.5	
Total	18	746.4		

(b) (3 pts) Conduct an F test to test at 0.05 level the null hypothesis that the four package designs have the same sales. Compute the p value of the observed F statistic.

$F_{3,15,0.05} = 3.29$, $F_{3,15,0.01} = 5.42$, $F = 18.59 > F_{3,15,0.05}$, reject the null hypothesis, and $p-value = F_{3,15}(x \geq 18.59) < 0.01$.

(c) (3 pts) Use the Tukey method to perform multiple comparisons of the four package designs at the 0.05 level.

$$\sqrt{\hat{\sigma}^2(\frac{1}{5}+\frac{1}{5}} = 2.05, \quad \sqrt{\hat{\sigma}^2(\frac{1}{5}+\frac{1}{4})} = 2.17, \quad \frac{1}{\sqrt{2}}q_{4,15,0.05} = \frac{1}{\sqrt{2}}4.08 = 2.88$$

$$\begin{aligned} t_{1,2} &= \frac{\bar{y}_{1.}-\bar{y}_{2.}}{2.05} = 0.59, & t_{1,3} &= \frac{\bar{y}_{1.}-\bar{y}_{3.}}{2.17} - 2.26 \\ t_{1,4} &= \frac{\bar{y}_{1.}-\bar{y}_{4.}}{2.05} = -6.15, & t_{2,3} &= \frac{\bar{y}_{2.}-\bar{y}_{3.}}{2.17} = -2.81 \\ t_{2,4} &= \frac{\bar{y}_{2.}-\bar{y}_{4.}}{2.05} = -6.73, & t_{3,4} &= \frac{\bar{y}_{3.}-\bar{y}_{4.}}{2.17} = -3.55. \end{aligned}$$

$|t_{1,4}|$, $|t_{2,4}|$, and $|t_{3,4}|$ are larger than $\frac{1}{\sqrt{2}}q_{4,15,0.05}$.

(d) (3 pts) To compare the mean sales for design 1 with average sales of all four designs, the contrast $L = \bar{y}_1 - \frac{\bar{y}_1+\bar{y}_2+\bar{y}_3+\bar{y}_4}{4}$ is computed. Obtain the p value of the computed contrast and test its significance at the 0.05 level. Comment on the difference between the mean sales for design 1 with average sales of all four designs. (Hint: linear interpolate between values from the provided table.)

$$\hat{L} = \frac{3}{4}\bar{y}_1 - \frac{1}{4}\bar{y}_2 - \frac{1}{4}\bar{y}_3 - \frac{1}{4}\bar{y}_4 = -4.075$$

$$\hat{T} = \frac{\hat{L}}{\hat{\sigma}\sqrt{(\frac{3}{4})^2\frac{1}{5} + (\frac{1}{4})^2\frac{1}{5} + (\frac{1}{4})^2\frac{1}{4} + (\frac{1}{4})^2\frac{1}{5}}} = \frac{-4.075}{1.27} = -3.21.$$

$$t_{15,0.0025} = 3.286, t_{15,0.005} = 2.947,$$

$$\frac{1}{2}(\text{p value}) = 0.005 + \frac{3.21 - 2.947}{3.286 - 2.947}(0.0025 - 0.005) = 0.0031.$$

$$\text{p value} = 0.006$$

Problem 2 (12 pts)

(a) (3 pts) Explain why the parameter η in the one-way random-effects model: $y_{ij} = \eta + \tau_i + \varepsilon_{ij}$, is called the population mean. Here ε_{ij}'s are independent error terms with $N(0, \sigma^2)$, τ_i are independent $N(0, \sigma_\tau^2)$, and τ_i and ε_{ij} are independent.
$E(y_{ij}) = \eta$.
In one-way random-effects model, τ_i are random effects, and $y_{i,j}$ are samples from the whole population with different random effects τ_i; in fixed-effects model, τ_i are fixed so that $y_{i,j}$, as observations with different fixed effects, cannot be samples of the whole population.

(b) (3 pts) Argue, by using a concrete example, that the η in Part (a) is generally of interest (and that is why we gave a formula for its confidence interval) while the grand mean parameter η in the one-way fixed-effects model $y_{ij} = \eta + \tau_i + \varepsilon_{ij}$, is generally not of interest. Here y_{ij} is the jth observation with treatment i, $i = 1, \ldots, k; j = 1, \ldots, n_i$, $\tau_i =$ ith treatment effect, and $\varepsilon_{ij} =$ error, independent $N(0, \sigma^2)$.

Assume $y_{i,j}$ $i = 1, \ldots, 10$, $j = 1, \ldots, 20$, are the GPA's of 20 students randomly chosen from each of the 10 high schools. In the random-effects model, the 10 high schools randomly chosen from A city. From this one-way random-effects model we can obtain: (i) what is the average GPA of the high school students of city A; (ii) if there are significant differences on the education quality among the high schools of city A. In the one-way fixed-effects model, the 10 schools are fixed rather than randomly chosen from the all of the high schools, so $y_{i,j}$ are not the samples from the population of all the GPA's of high school students from city A. So we can not know the average GPA of the high school students of city A. But we can know if the 10 schools are significantly different from each other on their education quality and which one is better than the other.

(c) (2 pts) Now assume that in **Problem 1** the four package designs are just four sample designs from a large class of designs in which the experimenter is interested. If σ_τ^2 denotes the true design-to-design variance with respect to the sales (i.e., number of cases), write down, in terms of σ_τ^2, the null hypothesis that he would like to test.
H_0: $\sigma_\tau^2 = 0$

(d) (4 pts) Obtain an estimate of σ_τ^2 and a 95% confidence interval for the population mean for the sales (i.e., number of cases).
$n' = \frac{1}{4-1}\{\sum n_i - \frac{\sum n_i^2}{\sum n_i}\} = 4.7$. $\sigma_\tau^2 = \frac{MST-MSE}{n'} = 39.5$.
$\hat{\eta} = \bar{y}_{..} = 18.63$. $|T| = |\frac{\hat{\eta}}{\sqrt{196.1/(4.7*4)}}| = 6.45 > t_{3,0.025} = 3.182$.
The 95% confidence interval is $18.63 \pm 3.182\sqrt{196.1/(4.7*4)}$ = (8.35, 28.91).

Problem 3 (9 pts)

(a) (3 pts) The linear contrast (-1,0,1) and quadratic contrast (1,-2,1) can be used for any three-level quantitative factor and remain orthogonal. Do you think this statement is true? If not, give a 2-3 line explanation.
False, three levels need evenly spaced.

(b) (3 pts) For a regression problem, suppose that the "Scaled Lambda Plot" suggests that square transformation is needed (variance stabilizing transformation with $\lambda = 2$). Someone does the analysis with $z = f(y) = (y^{\lambda} - 1)/\lambda$ and another does with $z = f(y) = y^{\lambda}$. Which one is more appropriate? Explain.
Both are OK in the case of $\lambda = 2$. $f(y) = (y^{\lambda} - 1)/\lambda$ is a continuous family of transformation. $f(y) = y^{\lambda}$ is easy to compute.

(c) (3 pts) In a randomized block design with 4 blocks and 9 treatments, how many degrees of freedom are associated with the *residual* in the ANOVA table?
(b-1)(k-1) = (4-1)(9-1) = 24

Problem 4 (13 pts) Questions (a)-(d) involve the analysis of some data on cigarette consumption. The data were collected in 1970. The variable **perfem** is the percentage of females living in a state. Data are collected on each of the 50 states of the USA and Washington DC. The purpose of the analysis is to see how cigarette consumption is related to the other variables. A simple linear regression model was fitted trying to predict the cigarette consumption using the percentage of females and the following partial output was obtained:

```
Coefficients:
                 Value   Std. Error  t-Value  p value
(Intercept)   -93.4260   207.8126   -0.4496   0.6550
    perfem      4.2191     4.0777             0.3059

Residual standard error: 32.05 on 49 degrees of freedom
Multiple R-Squared: 0.02138
```

(a) (3 pts) Calculate the t-statistic for **perfem**. What (if any) is its unit? Comment on the effect of **perfem**.

$T = 4.2191/4.0777 = 1.03$. t-statistic has no unit. $p-value$ of t-statistic of **perfem** is larger, indicating the effect of **perfem** is not significant.

(b) (2 pts) The slope is not statistically significant but the cigarette consumption of males and females is known to differ. Offer the best explanation for this apparent contradiction.

perfem does not vary much from state to state - collinear with the intercept term.

(c) (3 pts) Suppose we replaced the percentage of females by the percentage of males in the above regression model. Give the regression coefficients for this new model. Would the standard error for the slope change? Would the standard error for the intercept change?

$y = -93.426 + 4.2191 \times (100 - x) = 328.484 - 4.2191x$. se of slope will not change, but se of the intercept will change.

(d) (3 pts) When we look at the ANOVA table for a regression model, it is possible that the p value reported by MINITAB is 0.000, which does not mean that it is actually zero, just extremely small. Would it be possible (for some other data) to have a p value of exactly zero in an F-statistic in a regression? Explain clearly.

For p-value to be zero, the F-statistic would need to be infinite. This can only happen if RSS of the model is zero - in other words, if the data fit the model perfectly.

(e) (2 pts) Let n = number of observations, and p = number of parameters in the regression model. Someone claims that setting $n = p$ will give a p value of exactly zero. Is the claim correct ?

setting n=p does not work since this results in a 0/0 in the denominator of the F-statistic undefined and not finite.