QUALIFYING EXAM SOLUTIONS
Statistical Methods
Fall 2012

1. The strength of concrete used in commercial construction tends to vary from one batch to another. The dataset collected the quality of concrete according to four different methods and four different types of concrete making procedures, where each combination had four replications. Therefore, there were totally 64 observations in the dataset. The **R** output of is given below.

```
> anova(g)
Analysis of Variance Table
Response: quality
                           Df Sum Sq Mean Sq  F value    Pr(>F)
factor(method)              3  388.4   129.5  45.1235 5.203e-14 ***
factor(type)                3 3633.5  1211.2 422.1067 < 2.2e-16 ***
factor(method):factor(type) 9   42.0     4.7   1.6262    0.1346
Residuals                  48  137.7     2.9
> summary(g1)
Call:
lm(formula = yy ~ factor(method) + factor(type))
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     29.9069     0.5873  50.926  < 2e-16 ***
factor(method)2  6.8706     0.6278  10.944 1.24e-15 ***
factor(method)3  3.2981     0.6278   5.253 2.31e-06 ***
factor(method)4  4.3288     0.6278   6.895 4.76e-09 ***
factor(type)2    8.6450     0.6278  13.770  < 2e-16 ***
factor(type)3   20.4613     0.6278  32.592  < 2e-16 ***
factor(type)4   14.3338     0.6278  22.832  < 2e-16 ***
```

   (a) Construct an ANOVA table for the model without interaction effect, which does not need to include the p-value of the effects.
   Solution:

Source	Df	SS	MS	F-value
method	3	388.4	129.5	41.063
type	3	3633.5	1211.2	384.124
Residual	57	179.7	3.2	
Total	63	4201.6		

   (b) Propose the main effect factor effect model (where the constraints are zero-sum). Based on the output of the baseline model, compute the estimates of parameters of the factor effect model.

Solution: The following factor effect model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

for $i, j, k = 1, 2, 3, 4$, where $\epsilon_{ijk} \sim^{iid} N(0, \sigma^2)$ and $\sum_{i=1}^{4} \alpha_i = \sum_{j=1}^{4} \beta_j = 0$. The estimates of parameters can be computed by $\hat{\sigma}^2 = MSE = 3.2$,

$$\hat{\mu} = 29.9069 - \frac{6.8706 + 3.2981 + 4.3288}{4} - \frac{8.6450 + 20.4613 + 14.3338}{4} = 15.4225,$$

$$\begin{aligned}
\hat{\alpha}_1 =& -\frac{6.8706 + 3.2981 + 4.3288}{4} = -3.6244 \\
\hat{\alpha}_2 =& 6.8706 - \frac{6.8706 + 3.2981 + 4.3288}{4} = 3.2462 \\
\hat{\alpha}_3 =& 3.2981 - \frac{6.8706 + 3.2981 + 4.3288}{4} = -0.3263 \\
\hat{\alpha}_4 =& 4.3288 - \frac{6.8706 + 3.2981 + 4.3288}{4} = 0.7044,
\end{aligned}$$

and

$$\begin{aligned}
\hat{\beta}_1 =& -\frac{8.6450 + 20.4613 + 14.3338}{4} = -10.8600 \\
\hat{\beta}_2 =& 8.6450 - \frac{8.6450 + 20.4613 + 14.3338}{4} = 2.2150 \\
\hat{\beta}_3 =& 20.4613 - \frac{8.6450 + 20.4613 + 14.3338}{4} = 9.6013 \\
\hat{\beta}_4 =& 14.3338 - \frac{8.6450 + 20.4613 + 14.3338}{4} = 3.4738.
\end{aligned}$$

(c) Compute the 95% Bonferroni joint confidence intervals for pair differences between the four levels of **method** in the main effect model, where $t_{0.0042,57} = 2.7305$.

Solution: From the output, we have $s(\hat{\alpha}_i - \hat{\alpha}_{i'}) = 0.6278$ for all $i \neq i'$. Then, we have the formula

$$\hat{\alpha}_i - \hat{\alpha}_{i'} \pm t_{0.025/6,57} s(\hat{\alpha}_i - \hat{\alpha}_{i'}) = \hat{\alpha}_i - \hat{\alpha}_{i'} \pm 2.7305 \times 0.6278.$$

Then, the 95% Bonferroni joint confidence intervals for $\alpha_1 - \alpha_2$ is $[-8.5848, -5.1564]$, for $\alpha_1 - \alpha_3$ is $[-5.0123, -1.5839]$, for $\alpha_1 - \alpha_4$ is $[-6.0430, -2.6146]$, for $\beta_2 - \beta_3$ is $[-13.3256, -9.8261]$, for $\beta_2 - \beta_4$ is $[-7.1981, -3.6986]$, and for $\beta_3 - \beta_4$ is $[4.6182, 8.1177]$.

(d) Compute the 95% Tukey joint confidence intervals for the pair difference between the four levels of **type** in the main effect model, where $q_{0.05,4,57} = 3.40$.

Solution: We use the formula

$$\hat{\alpha}_i - \hat{\alpha}_{i'} \pm \frac{1}{\sqrt{2}} \times 3.40 \times 0.6278.$$

Then, the 95% Tukey joint confidence intervals for $\beta_1 - \beta_2$ is $[-10.1543, -6.6548]$, for $\beta_1 - \beta_3$ is $[-21.9706, -18.4711]$, for $\beta_1 - \beta_4$ is $[-15.8431, -12.3436]$, for $\beta_2 - \beta_3$ is $[1.8583, 5.2867]$, for $\beta_2 - \beta_4$ is $[0.8276, 4.2560]$, and for $\beta_3 - \beta_4$ is $[-2.7449, 0.6835]$.

(e) Propose a factor effect model with only the **type** main effect and compute the estimate of parameters in the model.

The model is

$$y_{ijk} = \mu + \beta_j + \epsilon_{ijk}$$

where $\epsilon \sim^{iid} N(0, \sigma^2)$ and $\sum_{j=1}^{4} \beta_j = 0$. The estimate of parameter is $\hat{\sigma}^2 = (388.4 + 179.7)/60 = 9.4683$ and $\mu = 15.4225$, $\hat{\beta}_1 = -10.8600$, $\hat{\beta}_2 = 2.2150$, $\hat{\beta}_2 = 9.6013$, and $\hat{\beta}_4 = 3.4738$.

2. The strength of fibers are major determinants of their quality. This was the focus of a study reported in a research article, which reported the data on fiber strength according to the fiber density and fiber types. The R output of a baseline model is given below, where **type2** and **type3** represent the second and third types, respectively.

```
Call:
lm(formula = strength ~ density * factor(type), data = fiber)
Coefficients:
              Estimate  Std.Error t value Pr(>|t|)
(Intercept)    2.8536     0.8414   3.392 0.000956 ***
density        2.9934     0.6150   4.868 3.66e-06 ***
type2          1.2357     1.2023   1.028 0.306226
type3          1.1055     1.2200   0.906 0.366756
density:type2  2.0998     0.9200   2.282 0.024316 *
density:type3  3.8074     0.9996   3.809 0.000227 ***
Residual standard error: 1.957 on 114 degrees of freedom
> round(summary(g)$cov.unscaled*summary(g)$sigma^2,4)
             (Intercept) density  type2   type3  density:type2 density:type3
(Intercept)   0.7079     -0.4811 -0.7079 -0.7079     0.4811       0.4811
density      -0.4811      0.3782  0.4811  0.4811    -0.3782      -0.3782
type2        -0.7079      0.4811  1.4456  0.7079    -1.0293      -0.4811
type3        -0.7079      0.4811  0.7079  1.4883    -0.4811      -1.1332
density:type2 0.4811     -0.3782 -1.0293 -0.4811     0.8463       0.3782
density:type3 0.4811     -0.3782 -0.4811 -1.1332     0.3782       0.9992
```

(a) Write down three regression lines according to the three types of asparagus.

Solution: The regression line for type 1 is

$$\hat{y} = 2.8536 + 2.9934 \times density.$$

The regression line for type 2 is

$$\begin{aligned}\hat{y} =& (2.8536 + 1.2357) + (2.9934 + 2.0998) \times density \\ =& 4.0893 + 5.0932 \times density.\end{aligned}$$

The regression line for type 3 is

$$\begin{aligned}\hat{y} =&(2.8536 + 1.1055) + (2.9934 + 3.8074) \times density\\ =&3.9591 + 6.8008 \times density.\end{aligned}$$

(b) Compute the standard errors of the estimates of parameters in the regression lines for the three types, respectively.

Solution: For the first type, the variance of the intercept is 0.7079, and the variance of the slope is 0.3782. For the second type, the variance of the intercept is $0.7079+1.4456-2(0.7079) = 0.7377$, and the variance of the slope is $0.3782+0.8463-2(0.3782) = 0.4681$. For the third type, the variance of the intercept is $0.7079 + 1.4883 - 2(0.7079) = 0.7804$, and the variance of the slope is $0.3782+0.9992-2(0.3782) = 0.6210$. Therefore, we have the standard errors of the intercepts are: 0.8414, 0.8589, and 0.8834, respectively. The standard errors of the intercepts are 0.6150, 0.6842, 0.7880 for the first, second, and the third types, respectively.

(c) Predict the value and standard error of the response when density equals 1.3 for the first type.

Solution: For the first type, the response is

$$\hat{y} = 2.8536 + 2.9934 \times 1.3 = 6.7450$$

and

$$s(\hat{y}) = \sqrt{0.7079 - 2 \times 1.3 \times 0.4811 + 1.3^2 \times 0.3782} = 0.0962.$$

(d) Predict the 95% confidence interval for the mean of the response in part (c), where $t_{0.975,114} = 1.98$.

Solution: The 95% confidence interval for the mean is

$$6.7450 \pm 1.98 \times 0.0962 = [6.5545, 6.9355].$$

(e) (2 points). Predict the 95% confidence interval for the observations of the response in part (c).

Solution: The 95% confidence interval for the observation is

$$6.7450 \pm 1.98 \times \sqrt{0.0962 + 1.95^2} = [2.8355, 10.6545].$$

3. The following table reports the data of the survival time in weeks of lips cancer patients.

Placebo	12 12 12+ 12+ 14 15 15+ 16 17 18 18+ 23+ 25+
Treatment	16 16+ 17 17+ 18 19 20 20+ 23 23+ 28 28+ 28+ 30 32+

(a) Compute the Kaplan-Meier estimate of the survival function for the treatment group.

Solution: The Kaplan-Meier estimate of the survival function for the treatment group is

$$\hat{S}(t) = \begin{cases} 1 & t < 16, \\ (1-15) = 0.933, & 16 \le t \le 17, \\ 0.933(1-1/13) = 0.862 & 17 \le t < 18, \\ 0.862(1-1/11) = 0.783 & 18 \le t \le 19, \\ 0.783(1-1/10) = 0.705 & 19 \le t < 20, \\ 0.705(1-1/9) = 0.627 & 20 \le t < 23, \\ 0.627(1-1/7) = 0.537 & 23 \le t < 28, \\ 0.537(1-1/5) = 0.430 & 28 \le t < 30, \\ 0.430(1-1/2) = 0.215 & t \ge 30, \end{cases}$$

(b) Assume the survival time follows an exponential distribution. Write down the model and compute the estimate of parameters for both treatment and placebo groups, respectively.

Solution: The survival function is

$$S(t) = e^{-\lambda t},$$

where λ is the reciprocal of the expected survival time. The MLE of λ is

$$\hat{\lambda} = \frac{\sum_{i=1}^{n_i} \delta_i}{\sum_{i=1}^{n_i} t_i}.$$

Put the data into the formula, we have $\hat{\lambda} = 0.03349$ for the placebo group and $\hat{\lambda} = 0.02388$ for the treatment group.

(c) Assume the survival time follows a Weibull distribution with PDF

$$f(t) = \alpha\lambda^{\alpha} t^{\alpha-1} e^{-\lambda^{\alpha} t^{\alpha}},$$

where the shape parameter is $\alpha = 0.5$. Derive the maximum likelihood estimate (MLE) of λ for both placebo and treatment groups, respectively.

Solution: The survival function is

$$S(t) = e^{-\lambda^{\alpha} t^{\alpha}}.$$

The loglikelihood function is

$$\begin{aligned} \ell(\lambda) &= \log[\prod_{i=1}^{n} f^{\delta_i}(t_i) S^{1-\delta_i}(t_i)] \\ &= \log \sum_{i=1}^{n} [(\alpha\lambda^{\alpha} t_i^{\alpha-1})^{\delta_i} e^{-\lambda^{\alpha} t_i^{\alpha}}] \\ &= \log \alpha \sum_{i=1}^{n} \delta_i + \alpha \log \lambda \sum_{i=1}^{n} \delta_i + (\alpha - 1) \sum_{i=1}^{n} \delta_i \log t_i - \lambda^{\alpha} \sum_{i=1}^{n} t_i^{\alpha}. \end{aligned}$$

Then,

$$\ell'(\lambda) = \frac{\alpha \sum_{i=1}^{n} \delta_i}{\lambda} - \alpha\lambda^{\alpha-1} \sum_{i=1}^{n} t_i^{\alpha} = 0 \Rightarrow \hat{\lambda} = (\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i^{\alpha}})^{1/\alpha}.$$

Then, we have $\hat{\lambda} = 0.01830$ for the placebo group and $\hat{\lambda} = 0.01292$ for the treatment group.

(d) Consider the cox proportional hazard model given by

$$h(t) = h_0(t)e^{\beta}$$

where $h_0(t)$ is the hazard function of the placebo group. Assume the estimate of β is given by $\hat{\beta} = -1.07$. Suppose the estimates of the survival function of the placebo group at $t = 12, 16, 20$ are given by $\hat{S}_0(12) = 0.8927$, $\hat{S}_0(16) = 0.6402$, and $\hat{S}_0(20) = 0.2678$. Compute the $\hat{S}(12)$, $\hat{S}(16)$ and $\hat{S}(20)$ for the treatment group.

Solution: The estimates of the survival function are given by $\hat{S}(12) = [S_0(12)]^{e^{-1.07}} = 0.9618$, $\hat{S}(16) = [S_0(16)]^{e^{-1.07}} = 0.8682$, and $\hat{S}(20) = [S_0(20)]^{e^{-1.07}} = 0.6364$.

4. Problem 4.

(a) The departments more women applied had low admission rates. This is called the Simptson Paradox.

(b) Let p_i be the rate for men and q_i be the rate for women. The null hypothesis is $H_0 : p_i = q_i$ versus $H_1 : p_i \neq q_i$ for $i = 1, 2, 3, 4, 5, 6$. We can propose a Pearson χ^2 test. The test statistic is

$$X^2 = \sum_{i=1}^{6}\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{(n_{ijk} - \hat{n}_{ijk})^2}{\hat{n}_{ijk}}.$$

The values of n_{ijk} and $\hat{n}_{ijk}$ can be computed. For example, when $i = 1$, we have $n_{111} = 825(0.62) = 511.5$, $n_{112} = 825(1 - 0.62) = 313.5$, $n_{121} = 108(0.82) = 88.56$, $n_{122} = 108(1-0.82) = 19.44$. Note that the rate in department A is (511.5+88.56)/(825+108) = 0.643. We have $\hat{n}_{111} = 825(0.643) = 530.5$, $\hat{n}_{112} = 825(1 - 0.643) = 294.5$, $\hat{n}_{121} = 108(0.643) = 69.44$, $\hat{n}_{122} = 108(1 - 0.643) = 38.56$. Finally, we have $X^2 = 25.87$. Under the null hypothesis, it follows χ^2_6.

5. Problem 5. The model is

$$y_{ijk} = \mu + trt_i + blk_j + \epsilon_{ijk},$$

where $i = 1, 2, 3, 4, 5$ are treatment effects and $j = 1, \cdots, 20$ are block (location) effects, $k = 1, 2$ are replications. We assume $blk_j \sim^{iid} N(0, \sigma_b^2)$ and $\epsilon_{ijk} \sim^{iid} N(0, \sigma^2)$.

i The DFs are 4 on numerator and $200 - 19 - 4 - 1 = 176$ on the denominator.

ii The percent is $1.0632/(1.0632 + 2.3297) = 31.33\%$ in one observation.

iii The standard is

$$V(\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}) = \frac{1}{10}(\hat{\sigma}_b^2 + \frac{\hat{\sigma}^2}{2}) = 0.2228.$$

iv It has a problem. The variance of the first ten was larger than the variance of the next ten. We need to involve a new variable for the variance. Note that the new variable is nested in the block effects. We may add **Group=** option in the **SAS** code.