QUALIFYING EXAM SOLUTIONS
Statistical Methods
Fall, 2009

1. (a) In this problem, we have $\bar{x} = 98.7143$ and $\bar{y} = 73.8429$. If we want the linear passes $(x^*, y^*) = (98.6, 73.8)$. For Model (1), we have

$$S_{xy} = \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) = 157.59$$

and

$$S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2 = 80.86.$$

Thus,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 1.9490.$$

and

$$\hat{\beta}_0 = \bar{y} - 1.9490\bar{x} = -118.55.$$

For Model (2), we have

$$S_{xy} = \sum_{i=1}^{n} (X_i - 98.6)(Y_i - 783.8) = 157.66$$

and

$$S_{xx} = \sum_{i=1}^{n} (X_i - 98.6)^2 = 81.04.$$

Then,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 1.9455.$$

(b) Compare the estimates in Model (1) and (2). Because $\hat{\beta}_1 \approx \hat{\beta}$, we believe the two models almost the same. Thus it is sufficient to consider Model (2) only.

(c) We have

$$\hat{V}(\hat{\beta}) = \frac{n\sigma^2}{(\sum_{i=1}^{n} X_i)^2} = 0.03499.$$

Thus, the t-value of $\hat{\beta}$ is

$$T = \frac{1.455}{\sqrt{0.03499}} = 7.778$$

which is larger than $t_{0.005,13} = 3.01$. Thus, we conclude $\beta \neq 0$.

(d) The fitted values are $\hat{y}_1 = 83.9166$, $\hat{y}_2 = 76.7182$, $\hat{y}_3 = 71.4654$, and $\hat{y}_4 = 70.4927$. The $SSLF$ is

$$\begin{aligned} SSLF =& 2(84.2 - 83.9166)^2 + 3(75.9 - 76.7182)^2 \\ &+ 4(71.3 - 71.4654)^2 + 5(70.5 - 70.4927)^2 \\ =& 2.2768. \end{aligned}$$

Then, the χ^2 statistic is

$$X^2 = \frac{SSLF}{0.08^2} = 355.75$$

which is too large (comparing to $\chi_{0.05,3} = 7.81$. Thus, we conclude there is a significant lack of fit in Model (2).

(e) We cannot use the previous $SSLF$ to reject both parts (c) and (d), because the lack of fit comparing Model (2) with the ANOVA model. It does not include the case $H_0 : \beta = 0$.

(f) We think the response at 100.1 are high leverage points because its difference to the expected value is much larger than the rest differences.

2. (a) The ANOVA model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, i = 1, \cdots, 10, j = 1, \cdots, 5, k = 1, 2, 3,$$

with $\sum_{i=1}^{10} \alpha_i = \sum_{j=1}^{5} \beta_j = \sum_{i=1}^{10}(\alpha\beta)_{ij} = \sum_{j=1}^{5}(\alpha\beta)_{ij} = 0$, where i is the index of the varieties and j is the index of locations, and $\epsilon_i \sim^{iid} N(0, \sigma^2)$. The degree of freedom for varieties is 9, for locations is 4, and for their interaction is 36. The ANOVA table is

Source	DF
Variety	9
Location	4
Variety:Location	36
Error	100
Total	149

(b) The assumption is the error term is iid $N(0, \sigma^2)$. We can use a residual plot to diagnose the assumption.

(c) Because the standard error is almost proportion to the average yield, we need to consider a squared root transformation.

(d) We can use the weighted least square method.

(e) The method of transformation does not need any degree of freedom for the estimate of weight. However, it may not easy to interpret.

(f) The Bonferroni method derived the $1-\alpha$-level joint confidence interval for the difference by using

$$\bar{y}_i - \bar{y}_{i'} \pm s(\bar{y}_i - \bar{y}_{i'}) \times t_{\alpha/90,100}.$$

The Tukey's method derives the $1 - \alpha$-level joint confidence interval by using

$$\bar{y}_i - \bar{y}_{i'} \pm s(\bar{y}_i - \bar{y}_{i'}) \times \frac{q_{1-\alpha/2,10,100}}{\sqrt{2}},$$

where $q_{\alpha,10,100}$ is the upper α quantile of the standard range distribution. Because the number of levels is large, we recommend to use the Tukey's method.

3. (a) Let n be the sample size and assume $n/2$ assigned standard and $n/2$ assigned new. Let $\bar{Y}_1$ and $\bar{Y}_2$ be their averages, respectively. Then $V(\bar{Y}_1 - \bar{Y}_2) = 4\sigma^2/n = 24/n$. We consider the one-sided test for

$$H_0 : \mu_1 \geq \mu_2 \leftrightarrow H_1 : \mu_1 < \mu_2.$$

Then, we reject H_0 if

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{24/n}} \leq -1.645 \Rightarrow \bar{Y}_1 - \bar{Y}_2 \leq -8.0588/\sqrt{n}.$$

If $\mu_1 - \mu_2 = -3$, then

$$\bar{Y}_1 - \bar{Y}_2 \sim N(3, \sqrt{\frac{24}{n}}).$$

Thus,

$$P(\bar{Y}_1 - \bar{Y}_2 \leq -\frac{8.0588}{\sqrt{n}}) \geq 0.9 \Rightarrow \Phi(\frac{-8.0588/\sqrt{n} - 3}{\sqrt{24/n}}) \geq 0.9$$

which implies

$$\frac{-8.0588/\sqrt{n} - 3}{\sqrt{24/n}} \geq 1.28 \Rightarrow \frac{14.3295}{\sqrt{n}} \leq 3 \Rightarrow n \geq 22.8.$$

Thus, we choose $n = 24$ since n should be even.

(b) Let (Y_{i1}, Y_{i2}) for $i = 1, 2, \cdots, n/2$ be measurement of the twin for standard and new methods. Let $d_i = Y_{i1} - Y_{i2}$. Then,

$$V(d_i) = V(Y_{i1} - Y_{i2}) = 6 + 6 - 2 \times 6 \times 0.8 = 2.4.$$

Then $V(\bar{d}) = 2.4/\sqrt{n/2} = 3.3941/\sqrt{n}$. Therefore, we reject $H_0 : \mu_1 \geq \mu_2$ and conclude $H_1 : \mu_1 < \mu_2$ if

$$\frac{\bar{d}}{3.3941/\sqrt{n}} < -1.645 \Rightarrow \bar{d} < -\frac{5.5833}{\sqrt{n}}.$$

Under $\mu_1 - \mu_2 = -3$, we have

$$\bar{d} \sim N(-3, \frac{3.3941}{\sqrt{n}}).$$

Then

$$P(\bar{d} \leq -\frac{5.5833}{\sqrt{n}}) \geq 0.9 \Rightarrow \Phi(\frac{-5.5833/\sqrt{n} + 3}{3.3941/\sqrt{n}}) \geq 0.9.$$

Thus, we have

$$\frac{-5.5833/\sqrt{n} + 3}{3.3941/\sqrt{n}} \geq 1.28 \Rightarrow \frac{8.9784}{\sqrt{n}} \leq 3 \Rightarrow n \geq 8.956.$$

Thus, we choose $n = 10$ because n must be even.

(c) Using twins can save the sample size and thus save the cost.

4. (a) We can used a partial regression method. For example, we fit (1) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ and derive the model residuals; (2) $X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ and derive the model residuals. We then fit the residuals of Model (1) with the residuals of Model (2). It can give us the estimate of β_3.

   (b) The ANCOVA model is preferred if there is another factor variable in the model, and it also considers the case when the slope of X is not one. The difference if the residual degrees of freedom because the ANOVA model forces the slope of X to be one.

   (c) The estimate of parameter will not change because x_0 is in the space spanned by $X(X'X)^{-1}X$. The standard errors of the estimates will change because the $(X'X)$ changes.

5. (a) We have

$$\hat{\lambda} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n t_i} = \frac{6}{207} = 0.02898.$$

   (b) The PDF if $f(t) = \lambda e^{-\lambda t}$

$$mrl(x) = \int_x^\infty (t - x)\lambda e^{-\lambda t} dt = \frac{1}{\lambda} e^{-\lambda x}.$$

   (c) Here we have

$$\hat{S}(30) = 0.4$$

   and based on the Greenwood's formula

$$\begin{aligned}\hat{V}(\hat{S}(30)) =& 0.4^2[\frac{1}{10 \times 9} + \frac{1}{9 \times 8} + \frac{1}{8 \times 7} + \frac{1}{7 \times 6} + \frac{1}{6 \times 5} + \frac{1}{5 \times 4}] \\ =& 0.024\end{aligned}$$

   Thus, the 95% confidence interval is

$$0.4 \pm 1.96 \times \sqrt{0.024} = [0.0964, 0.7036],$$

   which includes $e^{-0.02898 \times 30} = 0.4192$.

   (d) We may include a Weibull model with Remission Status as a parameter. The srvival function of the Weibull distribution is

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

   The model can be

$$-\log \gamma = \beta_0 + \beta_1 Trt + \beta_2 Rem$$

   which the scale parameter α determined by another procedure.

   (e) The model is

$$h(t) = h_0(t) e^{\beta_1 Trt + beta_2 Rem}.$$

   The model can estimate the effects of independent variables if the assumption holds, but it is less powerful than the model in (d) if the true model is Weibull.

6. (a) The model is

$$\log \lambda_{ij} = \mu + log(service) + \alpha_i + \beta_j + \gamma_k, i = 1, 2, 3, 4, 5, j = 1, 2, 3, 4, k = 1, 2,$$

where α_i is the index of type, β_j is the index of year, and γ_k is the index of period. We used the zero-sum constraint.

(b) The reason is because the occurrences of damage and services are proportion.

(c) Because $G^2 = 38.695 > \chi_{0.05,23} = 35.17$, the model does not fit the data well at 0.05 level.

(d) No, because G^2 is not too large comparing to its degree of freedom.

(e) In this case, the residual degrees of freedom is 26. Then

$$G_2^2 - G_1^2 = 42.329 - 38.695 = 3.634 < \chi^2_{0.05,3} = 7.81.$$

Therefore, we think this model is better than the previous model.