Test 3 Formula Sheet

STAT 021

Swarthmore College

Summary statistics

- Sample mean: $\frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample variance: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Population mean: For a random variable X, the expectation of X is

$$E(X) = \sum(\text{possibilities} \times \text{probabilities})$$

- Population variance: For a random variable X, the variance of X is

$$Var(X) = \sum((\text{possibilities} - E(X))^2 \times \text{probabilities})$$

Population proportions

- To find a $(1-\alpha)100\%$ CI for p: $\hat{p} \pm z^*_{\alpha/2} \times SE(\hat{p})$
- If p_0 is the true value of p, then $\frac{\hat{p}-p_0}{SE(\hat{p})} \sim N(0,1)$ for large enough n.
- To find a $(1-\alpha)100\%$ CI for $p_1 - p_2$: $\hat{p}_1 - \hat{p}_2 \pm z^*_{\alpha/2} \times SE(\hat{p}_1 - \hat{p}_2)$
- If $p_1 - p_2$ is the true difference in two independent population proportions, then $\frac{\hat{p}_1-\hat{p}_2}{SE(\hat{p}_1-\hat{p}_2)} \sim N(0,1)$ for large enough n_1 and n_2.

Population means

- To find a $(1-\alpha)100\%$ CI for μ: $\hat{p} \pm t^*_{(n-1),\alpha/2} \times SE(\bar{x})$
- If μ_0 is the true value of μ, then $\frac{\bar{x}-\mu_0}{SE(\bar{x})} \sim t_{(n-1)}$ for any $n > 2$.
- To find a $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ for two independent populations: $\bar{x}_1 - \bar{x}_2 \pm t^*_{\nu,\alpha/2} \times SE(\bar{x}_1 - \bar{x}_2)$, but the formula for ν is complicated and you don't need to know it.
- If $\mu_1 - \mu_2 = 0$ is the true difference in two independent population means, then $\frac{\bar{x}-\mu_0}{SE(\bar{x})} \sim t_\nu$ for any $n > 2$.

Linear Regression Formulas and Definitions

- MLR main effects model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad \text{and } \epsilon \sim N(0, \sigma^2)$$

- Fitted/estimated model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \hat{\beta}_k x_{k,i}$
- Residuals: $e_i = \hat{y}_i - y_i$
- For any predictor x_j, the variance inflation factor is: $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of multiple determination for a model to predict x_j using the other predictors in the model. (A VIF value larger than 5 corresponds to a value of over 0.8 for R_i^2.)
- To find a $(1-\alpha)100\%$ CI for β_j: $\hat{\beta}_j \pm t^*_{(n-k-1),\alpha/2} \times SE(\hat{\beta}_j)$
- If $\beta_j = 0$ is the true regression slope for the j^{th} predictor term, then $\frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim t_{(n-k-1)}$ for any $n > 2$.

ANOVA table for MLR models

For a MLR model with k *predictor terms*, the analysis of variance table for the model is:

Source	Deg of freedom	Sum of Squares	Mean Squares	F-statistic
Model	k	$SS_{Mod} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$MS_{Mod} = \frac{SS_{Mod}}{df_{Mod}}$	$\frac{MS_{Mod}}{MSE}$
Error	$n-k-1$	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{df_E}$	
Total	$n-1$	$SS_{Tot} = \sum_{i=1}^n (y_i - \bar{y})^2$		

Useful statistics

Statistics based on the sum of squares

- $R^2 = 1 - \frac{SSE}{SS_{Tot}} = \frac{SS_{Mod}}{SS_{Tot}}$ and $R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SS_{Tot}/(n-1)} = 1 - \frac{\hat{\sigma}_2}{s_Y^2}$
- Mallows's Cp: $C_p = \frac{SSE_m}{MSE_k} + 2(m+1) - n$, where $m < k$ and k is the number of predictor terms in the full model

Statistics that help identify influential data points

- Leverage: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$, for MLR models, data with moderate leverage have values $> 2k/n$ and those with extreme leverage have values $> 3k/n$, where k is the number of predictor terms.
- Standardized residuals: $stdres_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_i}}$, moderate values are > 2 and extreme values are > 3.
- Studentized residuals: $studres_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$, where $\hat{\sigma}_{(i)}$ is the estimated standard deviation of the error when the i^{th} data point is deleted; moderate values are > 2 and extreme values are > 3.
- Cook's distance: $D_i = \frac{(stdres_i)^2}{k+1}\left(\frac{h_i}{1-h_i}\right)$, moderate values are > 0.5 and extreme values are > 1.

ANOVA Model Forms

Group effects

For one categorical predictor variable with m different levels, the one-way ANOVA model in group effects form is:

$$Y = \mu + \alpha_j + \epsilon, \quad \text{where } j = 1, \ldots, m.$$

Group means

For one categorical predictor variable with m different levels, the one-way ANOVA model in group means form is:

$$Y = \mu_j + \epsilon, \quad \text{where } \mu_j = \mu + \alpha_j \text{ and } j = 1, \ldots, m.$$

MLR

For one categorical predictor variable with m different levels, the one-way ANOVA model in MLR form is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{m-1} x_{m-1} + \epsilon, \text{ where } x_j = \begin{cases} 1, & \text{if observation is in group } j \\ 0, & \text{otherwise} \end{cases} \quad \text{for } j = 1, \ldots, (m-1).$$

Fisher's LSD

If the overall F-test for model significance has a small enough p-value, we can use the following formula for a CI of the difference in means to determine which pairs of groups are significantly different from one another:

$$(\bar{y}_{Group1} - \bar{y}_{Group2}) \pm (t^* \cdot SE(\bar{y}_{Group1} - \bar{y}_{Group2})), \quad \text{where} \quad SE(\bar{y}_{Group1} - \bar{y}_{Group2}) = \sqrt{\hat{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

where t^* is the $\alpha/2^{th}$ quantile from a t-distribution with the model "error" degrees of freedom.