2028: Basic Statistical Methods
Sample Final Review Problems: Solutions

1. **Effective marketing.**

   (a) We derive the 95% confidence interval for the difference in the proportion parameters $p_1 - p_2$ where p_1 is the proportion of internet users who completed a college and p_2 is the proportion of nonusers who completed a college.

Users	Nonusers
x =643	y = 349
n=1132	m = 852
$\widehat{p}_1 = 0.568$	$\widehat{p}_2 = 0.409$

   An approximate 95% confidence interval for $p_1 - p_2$ is

$$\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}} = (0.1145, 0.2022)$$

   Since zero is not in the interval and the interval contains only positive values, we conclude that the difference of the two proportion parameters is positive and therefore, users and nonusers differ significantly in the proportion of college graduates.

   (b) We define p_1 be the proportion of internet users with higher income that \$50,000 and p_2 be the proportion of nonusers with higher income that \$50,000. We test

$$H_0 : \ p_1 = p_2 \text{ vs } H_1 : \ p_1 \neq p_2$$

   based on the following data

Users	Nonusers
x =378	y =200
n=871	m = 677
$\widehat{p}_1 = 0.434$	$\widehat{p}_2 = 0.295$

   The test statistic is

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})(1/n + 1/m)}} = 5.59$$

   where $\widehat{p}$ is the pooled proportion parameter computed as follows

$$\widehat{p} = \frac{x+y}{n+m} = \frac{378+200}{871+677}$$

   The p-value is $2P(Z > 5.59) \approx 0$. Therefore, there is a significant difference in the proportion of users and nonusers with income of \$50,000 or more.

2. **No-credit Card Fee.** An important note here is that the bank is interested whether there is an increase in the mean of the amount charged. This suggests using one-sided procedures. Denote X the difference in amount charged for a customer of the bank. For the $n = 500$ sample of costumers, $\bar{x} = 265$ and $s = 167$.

   a-b Denote μ the mean increase. The hypothesis test is $H_0 : \mu = 0$ vs. $H_a : \mu > 0$. We will answer (a) and (b) together. Given that we are interested in testing a one-sided hypothesis test, we also estimate a one-sided confidence with lower bound (the bank would like to find whether the lower bound is positive.) The 95% confidence interval is then

$$(\bar{x} - z_{.05}\frac{s}{n}, \infty) = (252.6926, \infty).$$

   Based on this confidence interval, we conclude that there is a positive difference in the amount charged at a significance level $\alpha = 0.05$ since the confidence interval does not contain the value zero. In fact, we conclude that the mean increase is at least \$252.

   c The power is

$$P\left(\frac{\bar{X}}{s/\sqrt{n}} > z_{.01} | \mu = 100\right) = P\left(\frac{\bar{X} - 100}{s/\sqrt{n}} + \frac{100}{s/\sqrt{n}} > z_{.01}\right)$$
$$= P\left(Z > -\frac{100}{167/\sqrt{100}} + z_{.01}\right) = P(Z > -3.66) = .99$$

   which suggest the with high accuracy, the bank can rely on an increase of \$100 in the amount charged per costumer.

3. **Blood sample tests.**

   Because we test two types of endothelial monolayer on the same samples, this is a paired sampling.

   We test

$$H_0 : \ \mu_A = \mu_B \text{ vs. } H_A : \ \mu_A \neq \mu_B$$

   where μ_A is the mean number of adherent red blood cells under type A and μ_B is the mean number of adherent red blood cells under type B. Because this is a paring sample, to test this hypothesis problem, we will take the difference between observations and apply one sample hypothesis testing procedure on the differences.

   The differences are $z = (-2, 0, 0, -6, 1, 1, 4, 8, -6, 6, -2, -4, -2, -17)$ and the sample mean and the sample standard deviation of the differences are $\bar{z} = -1.357$ and $s_z = 6.314$, respectively. The test statistic is

$$t = \frac{-1.357 - 0}{6.314/\sqrt{14}} = -0.804.$$

   The p-value is $2 \times P(X > 0.804)$ where X has a t-distribution with $14 - 1 = 13$ degrees of freedom. Based on t-distribution table, the p-value is larger than .1 and therefore, we do not reject the null hypothesis. It is plausible that the stimulation conditions do not affect the adhesion of red blood cells.

4. **Dwarf Plants**

(a) We want to test:

$$H_0 : p = .75 \text{ vs. } H_1 : p \neq .75.$$

The observed test statistic value is:

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{131 - 200(.75)}{\sqrt{200(.75)(1-.75)}} = -3.103.$$

At $\alpha = .05$, the critical value is $z_{\alpha/2} = z_{0.025} = 1.960 < |z| = 3.103$. Therefore, the null hypothesis is not plausible and the observation strongly contradict the genetic model.

5. **Java against Perl.**

(a) Only β_0 is "statistical" zero because its p-value for testing $H_0 : \beta_0 = 0$ using t-test is .964. Even though the two coefficients are in the same range, they have different standard errors: $se(\beta_0) = 1.523 >> se(\beta_1) = 0.02506$.

(b) The observed t-test value for $H_0 : \beta_1 = 1.3$ becomes:

$$T_0 = \frac{\widehat{\beta}_1 - 1.3}{se(\widehat{\beta}_1)} = \frac{0.07936 - 1.3}{0.02506} = -48.7$$

and the critical t-point for t-distribution with $df = n - 2 = 4$ is $t_{.05,4} = 2.132$. Evidently, $T_0 << -t_{.05,4}$ so we do reject the null hypothesis.

(c) A $1 - \alpha$ *prediction* confidence interval for a future y^* at x^* is:

$$\widehat{\beta}_0 + \widehat{\beta}_1 x^* \pm t_{\alpha/2,n-2}\widehat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}\right)}.$$

In the formula above, we need to input the estimated values for the regression coefficients, the standard error standard error ($\sqrt{MSE} = 1.999$ from the R output), the sample mean over the predictor values $\bar{x} = 51.33$ and S_{XX}. We may compute S_{XX} from using its formula $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ or by using the R output as follows. We know that the standard error for $\widehat{\beta}_1$ is $\sqrt{MSE/S_{XX}}$ which is equal to $se(\widehat{\beta}_1) = 0.02506$, and therefore

$$S_{XX} = \frac{MSE}{se(\widehat{\beta}_1)^2} = \frac{1.999^2}{0.025^2} = 6363.022.$$

Therefore, the 90% prediction interval for the result of this test is computed as follows

$$-0.07386 + 0.07936 \times 50 \pm 1.999\sqrt{1 + 1/6 + (51.33 - 50)^2/6363.022} = (-0.7094211, 8.4977011)$$

Since 3.894 is in the confidence interval, it is plausible that the CP time to be equal to 3.894.

6. **Compression Strength.**

Solutions: Compression Strength.

(a) The complete ANOVA table is:

	Df	Sum Sq	Mean Sq
Treatments	3	127375	42458
Residuals	20	33839	1692

(b) We perform an F-test for testing the null hypothesis of equal means: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.
The F-statistic is:

$$F = \frac{MSSTr}{MSE}$$

and the F-value for these data is:

$$F - value = \frac{42458}{1692} = 25.09.$$

The F-statistic is F-distributed with $(3, 20)$ degrees of freedom for the observed data, so the p-value is:

$$p - value = P(F_{(3,20)} > F_0) \approx 0$$

and we reject H_0.

(c) There is a substantial amount of overlap among observations of the first three types of boxes, but comparison strengths for the fourth type appear considerably smaller than for the other three types. This supports our conclusion that H_0 is not true.

(d) The pairwise CI for the difference in means $\mu_i - \mu_j$ is:

$$\left(\bar{x}_i - \bar{x}_j \pm \frac{sq_{0.05,k,\nu}}{\sqrt{2}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \right)$$

We need the sample means: $\bar{x}_1 = 713$, $\bar{x}_2 = 756.93$, $\bar{x}_3 = 698.07$,$\bar{x}_4 = 562.02$, and MSE to estimate $s = \sqrt{MSE}$. But the ANOVA table gives the value of $MSE = 1692$, and thus $s = 41.13$. It follows that the pairwise CI are:

$$\begin{aligned}
\text{for } \mu_2 - \mu_1 : &\ (-22.53671, 110.403377) \\
\text{for } \mu_3 - \mu_1 : &\ (-81.40338, 51.536711) \\
\text{for } \mu_4 - \mu_1 : &\ (-217.45338, -84.513289) \\
\text{for } \mu_3 - \mu_2 : &\ (-125.33671, 7.603377) \\
\text{for } \mu_4 - \mu_2 : &\ (-261.38671, -128.446623) \\
\text{for } \mu_4 - \mu_3 : &\ (-202.52004, -69.579956)
\end{aligned}$$

Comparing the mean of the fourth group with the means of the other three groups, the corresponding pairwise CI's contain only negative values which is an indication of the fact that μ_4 is lower than any of the other three means.