Stat./For./Hort. 571
Midterm I, Fall 2005
Brief Solutions

1. This is a paired two-sample inference problem. Let Y_1 = aphid number at the bottom and Y_2 = aphid number at the top of a plant. Let $D = Y_2 - Y_1$ denote the difference of aphid numbers with $\mu_D = E(D)$. The observed differences are $10, 7, 7, -1, 7, 8, 5, 7, 7, 8$ with $n = 10, \bar{d} = 6.50, s_d^2 = 8.50$ ($s_d = 2.915$).

   (a) For testing $H_0 : \mu_D = 0$ versus $H_0 : \mu_D \neq 0$, use a paired two-sample t-test. The observed $t = \frac{\bar{d}-0}{\sqrt{s_d^2/n}} = \frac{6.50-0}{\sqrt{8.50/10}} = 7.05$. Compare with T distribution on df $= n - 1 = 9$, the p-value $2 \times P(T_9 \geq 7.05)$ is less than 2×0.001, which is 0.002. Reject H_0 at the 5% level and there is strong evidence against H_0.

   (b) The 90% confidence interval for μ_D is $\bar{d} \pm t_{\alpha/2,n-1}\sqrt{s_d^2/n}$ which is $6.50 \pm 1.833 \times \sqrt{8.50/10}$, or $6.50 \pm 1.69 = [4.81, 8.19]$.

2. (a) The assumptions made in the paired two-sample inference are: $D_1, D_2, \cdots, D_{10}$ are iid $N(\mu_D, \sigma^2)$. That is, the differences D_i's are independent of each other and they are from a normal distribution. The assumption of independence is potentially violated in the current experimental setting.

   (b) Let $\mu_D = \mu_2 - \mu_1$. Then the hypotheses of interest are $H_0 : \mu_D = 0$ versus $H_0 : \mu_D > 0$. Since $\alpha = 0.10$ and $\beta = 1 - 0.90 = 0.10$, we have $z_{0.10} = 1.282$. We also know that $\sigma^2 = 10$, $\mu_D^0 = 0$, and $\mu_D^A = 2$. Thus the desired sample size is

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_D^A - \mu_D^0)^2} = \frac{(1.282 + 1.282)^2 \times 10}{2^2} = 16.44 \uparrow 17.$$

3. This is a two-proportion inference problem.

   Here $y_A = 24, n_A = 50, y_B = 10, n_B = 60$, thus $\hat{p}_A = 24/50 = 0.48$ and $\hat{p}_B = 10/60 = 0.167$. An approximate 95% confidence interval for $p_A - p_B$ is

$$\hat{p}_A - \hat{p}_B \pm z_{0.025}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$
$$= (0.48 - 0.167) \pm 1.96 \times \sqrt{\frac{0.48 \times 0.52}{50} + \frac{0.167 \times 0.833}{60}}$$
$$= 0.313 \pm 0.168 = [0.145, 0.481]$$

   We may use the normal approximation, because $n_Ap_A = 24 > 5, n_Aq_A = 26 > 5, n_Bp_B = 10 > 5, n_Bq_B = 50 > 5$.

   (b) We assume that $Y_A \sim B(50, p_A), Y_B \sim B(60, p_B)$, Y_A and Y_B are independent, and for the normal approximation to be valid, we need $n_Ap_A, n_Aq_A, n_Bp_B, n_Bq_B$ to be at least 5. For Y_A to have a binomial distribution, we assume that the outcome of each lake is Bernoulli (contaminated or not), the contamination of lakes is independent, and the contamination rate is constant p_A for all 50 lakes (similarly Y_B).

4. Let Y denote the number of lakes that are contaminated among the $n = 40$ lakes and assume that $Y \sim B(40, p)$. The hypotheses of interest are $H_0 : p = 0.5$ versus $H_A : p \neq 0.5$. Let $\hat{p} = Y/40$ denote the proportion of contaminated lakes. Under H_0, $E(\hat{p}) = p = 0.5$ and $Var(\hat{p}) = pq/n = 0.5 \times 0.5/40$. We should reject H_0 if $\hat{p}$ is either much larger than 0.5 ($\hat{p} \geq c$) or much smaller than 0.5 ($\hat{p} \leq c^*$). Since $np = 20, nq = 20$ are both larger than 5, we may use normal approximation. By symmetry, we focus on c first:

$$P(\hat{p} \geq c) \approx P(Z \geq \frac{c - 0.5}{\sqrt{0.5 \times 0.5/40}}) = 0.05 = \alpha/2.$$

   Since $P(Z \geq 1.645) = 0.05$, we have

$$\frac{c - 0.5}{\sqrt{0.5 \times 0.5/40}} = 1.645$$

   and thus $c = 0.5 + 1.645 \times \sqrt{0.5 \times 0.5/40} = 0.5 + 0.13 = 0.63$. By symmetry, $c^* = 0.5 - 1.645 \times \sqrt{0.5 \times 0.5/40} = 0.5 - 0.13 = 0.37$. Thus the rejection region is $\hat{p} \leq 0.37$ or $\hat{p} \geq 0.63$.

5. (a) Use an independent two-sample t-test. The pooled variance is $s_p = \sqrt{(s_1^2 + s_2^2)/2} = 1.877$ and the observed $t = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{12.12-10.21}{1.877\sqrt{2/15}} = 2.78$. Compare with T distribution on df = 28, the p-value $2 \times P(T_{28} \geq 2.78)$ is between 2×0.001 and 2×0.005 (i.e. between 0.002 and 0.01). Reject H_0 at the 5% level and there is strong evidence against H_0.

   (b) We assume that Y_1's are iid $N(\mu_1, \sigma_1^2)$, Y_2's are iid $N(\mu_2, \sigma_2^2)$, Y_1 and Y_2 are independent, and the variances are equal $\sigma_1^2 = \sigma_2^2 = \sigma^2$. That is, we assume normal distribution, independence both within and across trt groups, and equal variance.

   (c) The ANOVA table looks like:

Source	df	SS	MS
Trt	1	27.20	27.20
Error	28	98.56	3.52
Total	29	125.76	–

   Note that MSError $= \sqrt{(s_1^2 + s_2^2)/2} = 3.52$ and thus SSError $= 3.52 \times 28 = 98.56$. You may compute SSTrt directly from $\bar{y}_1, \bar{y}_2$. Alternatively, from 5(a) $t = 2.78$, we have $f = t^2 = 2.78^2$ and MSTrt = MSError $\times f = 3.52 \times 2.78^2 = 27.20$. And SSTrt = MSTrt, because dfTrt = 1.

Grade Distribution

```
 100: 2
90-99:59
80-89:36
70-79:11          mean = 86.0, median = 90
60-69: 8
below: 3
```