Chapter 12: Considerations in Designing Experiments

12.1 (See Example 12.1) Let $n_1 = \left(\frac{\sigma_1}{\sigma_1+\sigma_2}\right)n = \left(\frac{3}{3+5}\right)90$ = 33.75 or 34 and n_2 = 90 – 34 = 56.

12.2 (See Ex. 12.1). If n_1 = 34 and n_2 = 56, then

$$\sigma_{\bar{Y}_1-\bar{Y}_2} = \sqrt{\tfrac{9}{34}+\tfrac{25}{56}} = \sqrt{.7111}$$

In order to achieve this same bound with equal sample sizes, we must have

$$\sqrt{\tfrac{9}{n}+\tfrac{25}{n}} = \sqrt{.7111}$$

The solution is n = 47.8 or 48. Thus, it is necessary to have $n_1 = n_2 = 48$ so that the same amount of information is implied.

12.3 The length of a 95% CI is twice the margin of error:

$$2(1.96)\sqrt{\tfrac{9}{n_1}+\tfrac{25}{n_2}},$$

and this is required to be equal to two. In Ex. 12.1, we found $n_1 = (3/8)n$ and $n_1 = (5/8)n$, so substituting these values into the above and equating it to two, the solution is found to be n = 245.9. Thus, n_1 = 93 and n_2 = 154.

12.4 (Similar to Ex. 12.3) Here, the equation to solve is

$$2(1.96)\sqrt{\tfrac{9}{n_1}+\tfrac{25}{n_1}} = 2.$$

The solution is n_1 = 130.6 or 131, and the total sample size required is 131 + 131 = 262.

12.5 Refer to Section 12.2. The variance of the slope estimate is minimized (maximum information) when S_{xx} is as large as possible. This occurs when the data are as far away from $\bar{x}$ as possible. So, with n = 6, three rats should receive x = 2 units and three rats should receive x = 5 units.

12.6 When σ is known, a 95% CI for β is given by

$$\hat{\beta}_1 \pm z_{\alpha/2}\frac{\sigma}{\sqrt{S_{xx}}}.$$

Under the two methods, we calculate that S_{xx} = 13.5 for Method 1 and S_{xx} = 6.3 for Method 2. Thus, Method 2 will produce the longer interval. By computing the ratio of the margins of error for the two methods (Method 2 to Method 1), we obtain $\sqrt{\frac{13.5}{6.3}}$ = 1.464; thus Method 2 produces an interval that is 1.464 times as large as Method 1.

Under Method 2, suppose we take n measurements at each of the six dose levels. It is not difficult to show that now $S_{xx} = 6.3n$. So, in order for the intervals to be equivalent, we must have that $6.3n = 13.5$, and so n = 2.14. So, roughly twice as many observations are required.

12.7 Although it was assumed that the response variable Y is truly linear over the range of x, the experimenter has no way to verify this using Method 2. By assigning a few points at x = 3.5, the experimenter could check for curvature in the response function.

12.8 Checking for true linearity and constant error variance cannot be performed if the data points are spread out as far as possible.

12.9 **a.** Each half of the iron ore sample should be reasonably similar, and assuming the two methods are similar, the data pairs should be positively correlated.

b. Either analysis compares means. However, the paired analysis requires fewer ore samples and reduces the sample–to–sample variability.

12.10 The sample statistics are: $\bar{d} = -.0217$, $s_D^2 = .0008967$.

a. To test H_0: $\mu_D = 0$ vs. H_a: $\mu_D \neq 0$, the test statistic is $|t| = \frac{|-.0217|}{\sqrt{.0008967/6}} = 1.773$ with 5 degrees of freedom. Since $t_{.025} = 2.571$, H_0 is not rejected.

b. From Table 5, $.10 < p$–value $< .20$.

c. The 95% CI is $-.0217 \pm 2.571\sqrt{\frac{.0008967}{6}} = -.0217 \pm .0314$.

12.11 Recall that $Var(\bar{D}) = \frac{1}{n}\left(\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2\right)$ given in this section.

a. This occurs when $\rho > 0$.

b. This occurs when $\rho = 0$.

c. This occurs when $\rho < 0$.

d. If the samples are negatively correlated, a matched–pairs experiment should not be performed. Otherwise, if it is possible, the matched–pairs experiment will have an associated variance that is equal or less than the variance associated with the independent samples experiment.

12.12 **a.** There are $2n - 2$ degrees of freedom for error.

b. There are $n - 1$ degrees of freedom for error.

c.

n	Independent samples	Matched–pairs
5	d.f. = 8, $t_{.025} = 2.306$	d.f. = 4, $t_{.025} = 2.776$
10	d.f. = 18, $t_{.025} = 2.101$	d.f. = 9, $t_{.025} = 2.262$
30	d.f. = 58, $t_{.025} = 1.96$	d.f. = 29, $t_{.025} = 2.045$

d. Since more observations are required for the independent samples design, this increases the degrees of freedom for error and thus shrinks the critical values used in confidence intervals and hypothesis tests.

12.13 A matched–pairs experiment is preferred since there could exist sample–to–sample variability when using independent samples (one person could be more prone to plaque buildup than another).

12.14 The sample statistics are: $\bar{d} = -.333$, $s_D^2 = 5.466$. To test H_0: $\mu_D = 0$ vs. H_a: $\mu_D < 0$, the test statistic is $t = \frac{-.333}{\sqrt{5.466/6}} = -.35$ with 5 degrees of freedom. From Table 5, p–value $> .1$ so H_0 is not rejected.

12.15 **a.** The sample statistics are: $\bar{d}$ = 1.5, s_D^2 = 2.571. To test H_0: $\mu_D = 0$ vs. H_a: $\mu_D \neq 0$, the test statistic is $|t| = \frac{|1.5|}{\sqrt{2.571/8}}$ = 2.65 with 7 degrees of freedom. Since $t_{.025}$ = 2.365, H_0 is rejected.

b. Notice that each technician's score is similar under both design A and B, but the technician's scores are not similar in general (some are high and some are low). Thus, pairing is important to screen out the variability among technicians.

c. We assumed that the population of differences follows a normal distribution, and that the sample used in the analysis was randomly selected.

12.16 The sample statistics are: $\bar{d}$ = –3.88, s_D^2 = 8.427.

a. To test H_0: $\mu_D = 0$ vs. H_a: $\mu_D < 0$, the test statistic is $t = \frac{-3.88}{\sqrt{8.427/15}}$ = –5.176 with 14 degrees of freedom. From Table 5, it is seen that p–value < .005, so H_0 is rejected when $\alpha = .01$.

b. A 95% CI is $-3.88 \pm 2.145\sqrt{8.427/15}$ = –3.88 ± 1.608.

c. Using the Initial Reading data, $\bar{y}$ = 36.926 and s^2 = 40.889. A 95% CI for the mean muck depth is $36.926 \pm 2.145\sqrt{40.889/15}$ = 36.926 ± 3.541.

d. Using the Later Reading data, $\bar{y}$ = 33.046 and s^2 = 35.517. A 95% CI for the mean much depth is $33.046 \pm 2.145\sqrt{35.517/15}$ = 33.046 ± 3.301.

e. For parts a and b, we assumed that the population of differences follows a normal distribution, and that the sample used in the analysis was randomly selected. For parts c and d, we assumed that the individual samples were randomly selected from two normal populations.

12.17 **a.** $E(Y_{ij}) = \mu_i + E(U_i) + E(\varepsilon_{ij}) = \mu_i$.

b. Each Y_{1j} involves the sum of a uniform and a normal random variable, and this convolution does not result in a normal random variable.

c. $\text{Cov}(Y_{1j}, Y_{2j}) = \text{Cov}(\mu_1 + U_j + \varepsilon_{1j}, \mu_2 + U_j + \varepsilon_{2j})$ = $\text{Cov}(\mu_1, \mu_2)$ + $\text{Cov}(U_j, U_j)$ + $\text{Cov}(\varepsilon_{1j}, \varepsilon_{2j})$ = 0 + $V(U_j)$ + 0 = 1/3.

d. Observe that $D_j = Y_{1j} - Y_{2j} = \mu_1 - \mu_2 + \varepsilon_{1j} - \varepsilon_{2j}$. Since the random errors are independent and follow a normal distribution, D_j is a normal random variable. Further, for $j \neq j'$, $\text{Cov}(D_j, D_{j'}) = 0$ since the two random variables are comprised of constants and independent normal variables. Thus, D_j and $D_{j'}$ are independent (recall that if two normal random variables are uncorrelated, they are also independent – but this is not true in general).

e. Provided that the distribution of U_j has a mean of zero and finite variance, the result will hold.

12.18 Use Table 12 and see Section 12.4 of the text.

12.19 Use Table 12 and see Section 12.4 of the text.

12.20 **a.** There are six treatments. One example would be the first catalyst and the first temperature setting.
b. After assigning the *n* experimental units to the treatments, the experimental units are numbered from 1 to n. Then, a random number table is used to select numbers until all experimental units have been selected.

12.21 Randomization avoids the possibility of bias introduced by a nonrandom selection of sample elements. Also, it provides a probabilistic basis for the selection of a sample.

12.22 Factors are independent experimental variables that the experimenter can control.

12.23 A treatment is a specific combination of factor levels used in an experiment.

12.24 Yes. Suppose that a plant biologist is comparing three soil types used for planting, where the response is the yield of a crop planted in the different soil types. Then, "soil type" is a factor variable. But, if the biologist is comparing the yields of different greenhouses, but each greenhouse used different soil types, then "soil type" is a nuisance variable.

12.25 <u>Increases accuracy of the experiment</u>: 1) selection of treatments, 2) choice of number of experimental units assigned to each treatment.
<u>Decreases the impact of extraneous sources of variability</u>: randomization; assigning treatments to experimental units.

12.26 There is a possibility of significant rat–to–rat variation. By applying all four dosages to tissue samples extracted from the same rat, the experimental error is reduced. This design is an example of a randomized block design.

12.27 In the Latin square design, each treatment appears in each row and each column exactly once. So, the design is:

B	*A*	*C*
C	*B*	*A*
A	*C*	*B*

12.28 A CI could be constructed for the specific population parameter, and the width of the CI gives the quantity of information.

12.29 A random sample of size *n* is a sample that was randomly selected from all possible (unique) samples of size *n* (constructed of observations from the population of interest) and each sample had an equal chance of being selected.

12.30 From Section 12.5, the choice of factor levels and the allocation of the experimental units to the treatments, as well as the total number of experimental units being used, affect the total quantity of information. Randomization and blocking can control these factors.

12.31 Given the model proposed in this exercise, we have the following:

a. $E(Y_{ij}) = \mu_i + E(P_i) + E(\varepsilon_{ij}) = \mu_i + 0 + 0 = \mu_i$.

b. Obviously, $E(\overline{Y}_i) = \mu_i$. Also, $V(\overline{Y}_i) = \frac{1}{n}V(Y_{ij}) = \frac{1}{n}\left[V(P_i) + V(\varepsilon_{ij})\right] = \frac{1}{n}\left[\sigma_P^2 + \sigma^2\right]$, since P_i and ε_{ij} are independent for all i, j.

c. From part b, $E(\overline{D}) = E(\overline{Y}_1) - E(\overline{Y}_2) = \mu_1 - \mu_2$. Now, to find $V(\overline{D})$, note that

$$\overline{D} = \tfrac{1}{n}\sum_{j=1}^{n} D_j = \mu_1 - \mu_2 + \tfrac{1}{n}\left[\sum_{j=1}^{n}\varepsilon_{1j} + \sum_{j=1}^{n}\varepsilon_{2j}\right].$$

Thus, since the ε_{ij} are independent, $V(\overline{D}) = \frac{1}{n^2}\left[\sum_{j=1}^{n}V(\varepsilon_{1j}) + \sum_{j=1}^{n}V(\varepsilon_{2j})\right] = 2\sigma^2/n$.

Further, since $\overline{D}$ is a linear combination of normal random variables, it is also normally distributed.

12.32 From Exercise 12.31, clearly $\dfrac{\overline{D} - (\mu_1 - \mu_2)}{\sqrt{2\sigma^2/n}}$ has a standard normal distribution. In addition, since $D_1, \ldots, D_n$ are independent normal random variables with mean $\mu_1 - \mu_2$ and variance $2\sigma^2$, the quantity

$$W = \frac{(n-1)S_D^2}{2\sigma^2} = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{2\sigma^2}$$

is chi–square with $\nu = n - 1$ degrees of freedom. Therefore, by Definition 7.2 and under H_0: $\mu_1 - \mu_2 = 0$,

$$\frac{Z}{\sqrt{W/\nu}} = \frac{\overline{D}}{S_D/\sqrt{n}}$$

has a t–distribution with $n - 1$ degrees of freedom.

12.33 Using similar methods as in Ex. 12.31, we find that for this model,

$$V(\overline{D}) = \tfrac{1}{n^2}\sum_{j=1}^{n}\left[V(P_{1j}) + V(P_{2j}) + V(\varepsilon_{1j}) + V(\varepsilon_{2j})\right] = \tfrac{1}{n}\left[2\sigma_P^2 + 2\sigma^2\right] > \tfrac{1}{n}2\sigma^2.$$

Thus, the variance is larger with the completely randomized design, since the unwanted variation due to pairing is not eliminated.

12.34 The sample statistics are: $\overline{d} = -.062727$, $s_D^2 = .012862$.

a. We expect the observations to be positively correlated since (assuming the people are honest) jobs that are estimated to take a long time actually take a long time when processed. Similar for jobs that are estimated to take a small amount of processor time.

b. To test H_0: $\mu_D = 0$ vs. H_a: $\mu_D < 0$, the test statistic is $t = \frac{-.0627272}{\sqrt{.012862/11}} = -1.834$ with 10 degrees of freedom. Since $-t_{.10} = -1.362$, H_0 is rejected: there is evidence that the customers tend to underestimate the processor time.

c. From Table 5, we have that .025 < p–value < .05.

d. A 90% CI for $\mu_D = \mu_1 - \mu_2$, is $-.062727 \pm 1.812\sqrt{.012862/11} = -.063 \pm .062$ or $(-.125, -.001)$.

12.35 The sample statistics are: $\bar{d} = -1.58$, $s_D^2 = .667$.

a. To test H_0: $\mu_D = 0$ vs. H_a: $\mu_D \neq 0$, the test statistic is $|t| = \frac{|-1.58|}{\sqrt{.667/5}} = 4.326$ with 4 degrees of freedom. From Table 5, we can see that $.01 < p$–value $< .025$, so H_0 would be rejected for any $\alpha \geq .025$.

b. A 95% CI is given by $-1.58 \pm 2.776\sqrt{.667/5} = -1.58 \pm 1.014$ or $(-2.594, -.566)$.

c. We will use the estimate of the variance of paired differences. Also, since the required sample will (probably) be large, we will use the critical value from the standard normal distribution. Our requirement is then:

$$.2 = z_{.025}\sqrt{\frac{\sigma_D^2}{n}} \approx 1.96\sqrt{\frac{.667}{n}}.$$

The solution is $n = 64.059$, or 65 observations (pairs) are necessary.

12.36 The sample statistics are: $\bar{d} = 106.9$, $s_D^2 = 1364.989$.

a. Each subject is presented each sign in random order. If the subject's reaction time is (in general) high, both responses should be high. If the subject's reaction time is (in general) low. both responses should be low. Because of the subject–to–subject variability, the matched pairs design can eliminate this extraneous source of variation.

b. To test H_0: $\mu_D = 0$ vs. H_a: $\mu_D \neq 0$, the test statistic is $|t| = \frac{|106.0|}{\sqrt{1364.989/10}} = 9.15$ with 9 degrees of freedom. Since t.025 = 2.262, H_0 is rejected.

c. From Table 5, we see that p–value $< 2(.005) = .01$.

d. The 95% CI is given by $106.9 \pm 2.262\sqrt{1364.989/10} = 106.9 \pm 26.428$ or (80.472, 133.328).

12.37 There are nk_1 points at $x = -1$, nk_2 at $x = 0$, and nk_3 points at $x = 1$. The design matrix $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{X} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & -1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 1 \end{bmatrix}; \text{ thus } \boldsymbol{X'X} = \begin{bmatrix} n & n(k_3 - k_1) & n(k_1 + k_3) \\ n(k_3 - k_1) & n(k_1 + k_3) & n(k_3 - k_1) \\ n(k_1 + k_3) & n(k_3 - k_1) & n(k_1 + k_3) \end{bmatrix} = n\begin{bmatrix} 1 & b & a \\ b & a & b \\ a & b & a \end{bmatrix} = n\boldsymbol{A},$$

where $a = k_1 + k_3$ and $b = k_3 - k_1$.

Now, the goal is to minimize $V(\hat{\beta}_2) = \sigma^2 c_{22}$, where c_{22} is the (3, 3) element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. To calculate $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, note that it can be expressed as

$\boldsymbol{A}^{-1} = \frac{1}{n\det(\boldsymbol{A})}\begin{bmatrix} a^2 - b^2 & 0 & b^2 - a^2 \\ 0 & a - a^2 & ab - b \\ b^2 - a^2 & ab - b & a - b^2 \end{bmatrix}$, and (the student should verify) the determinant of $\boldsymbol{A}$ simplifies to $\det(\boldsymbol{A}) = 4k_1k_2k_3$. Hence,

$$V(\hat{\beta}_2) = \sigma^2 \frac{a - b^2}{4nk_1k_2k_3} = \frac{\sigma^2}{n}\left(\frac{k_1 + k_3 - (k_3 - k_1)^2}{4k_1k_2k_3}\right).$$

We must minimize

$$Q = \frac{k_1 + k_3 - (k_3 - k_1)^2}{4k_1k_2k_3} = \frac{k_1 + k_3 - \left[(k_3 + k_1)^2 - 4k_1k_3\right]}{4k_1k_2k_3} = \frac{(k_1 + k_3)\left[1 - k_1 - k_3\right]}{4k_1k_2k_3} - \frac{4k_1k_3}{4k_1k_2k_3}$$

$$= \frac{k_1 + k_3}{4k_1k_3} - \frac{1}{k_2} = \frac{k_1 + k_3}{4k_1k_3} - \frac{1}{1 - k_1 - k_3}.$$

So, with $Q = \frac{k_1 + k_3}{4k_1k_3} - \frac{1}{1 - k_1 - k_3}$, we can differentiate this with respect to k_1 and k_3 and set these equal to zero. The two equations are:

$$\begin{aligned} 4k_1^2 &= (1 - k_1 - k_3)^2 \qquad (*) \\ 4k_3^2 &= (1 - k_1 - k_3)^2 \end{aligned}.$$

Since k_1, k_2, and k_3 are all positive, $k_1 = k_3$ by symmetry of the above equations and therefore by (*), $4k_1^2 = (1 - 2k_1)^2$ so that $k_1 = k_3 = .25$. Thus, $k_2 = .50$.