Chapter 2

Using Surveys and Experimental Studies to Gather Data

2.1

- **a.** The explanatory variable is level of alcohol drinking. One possible confounding variable is smoking. Perhaps those who drink more often also tend to smoke more, which would impact incidence of lung cancer. To eliminate the effect of smoking, we could block the experiment into groups (e.g., nonsmokers, light smokers, heavy smokers).
- **b.** The explanatory variable is obesity. Two confounding variables are hypertension and diabetes. Both hypertension and diabetes contribute to coronary problems. To eliminate the effect of these two confounding variables, we could block the experiment into four groups (e.g., hypertension and diabetes, hypertension but no diabetes, diabetes but no hypertension, neither hypertension nor diabetes).

2.2

- **a.** The explanatory variable is the new blood clot medication. The confounding variable is the year in which patients were admitted to the hospital. Because those admitted to the hospital the previous year were not given the new blood clot medication, we cannot be sure that the medication is working or if something else is going on. We can eliminate the effects of this confounding by randomly assigning stroke patients to the new blood clot medication or a placebo.
- **b.** The explanatory variable is the software program. The confounding variable is whether students choose to stay after school for an hour to use the software on the school's computers. Those students who choose to stay after school to use the software on the school's computers may differ in some way from those students who do not choose to do so, and that difference may relate to their mathematical abilities. To eliminate the effect of the confounding variable, we could randomly assign some students to use the software on the school's computers during class time and the rest to stay in class and learn in a more traditional way.

2.3

Possible confounding factors include student-teacher ratios, expenditures per pupil, previous mathematics preparation, and access to technology in the inner city schools. Adding advanced mathematics courses to inner city schools will not solve the discrepancy between minority students and white students, since there are other factors at work.

2.4

There may be a difference in student-teacher ratios, expenditures per pupil, and previous preparation between the schools that have a foreign language requirement and schools that do not have a foreign language requirement.

2.5

The relative merits of the different types of sampling units depends on the availability of a sampling frame for individuals, the desired precision of the estimates from the sample to the population, and the budgetary and time constraints of the project.

2.6

She could conduct a stratified random sample in which the plants serve as the stratum. A simple random sample could then be selected within each plant. This would provide information concerning the differences between the plants along with the individual opinions of the employees.

2.7

The list of registered voters could be used as the sampling frame for selecting the persons to be included in the sample.

2.8

- **a.** No. The survey in which the interviewer showed the peanut butter should be more accurate because it does not rely on the respondent's memory of which brand was purchased.
- **b.** Both surveys may have survey nonresponse bias because an entire segment of the population (those not at home) cannot be contacted. Also, both surveys may have interviewer bias resulting from the way the question was posed (e.g., tone of voice). In the first survey, results may be biased by the respondent's ability to recall correctly which brand was purchased. The second survey may be biased by the respondent's unwillingness to show the interviewer the peanut butter jar (too intrusive), or by the respondent not recognizing that the peanut butter that was purchased was *low fat*.

2.9

- **a.** Alumni (men only?) who graduated from Yale in 1924.
- **b.** No. Alumni whose addresses were on file 25 years later would not necessarily be representative of their class.
- **c.** Alumni who <u>responded</u> to the mail survey would not necessarily be representative of those who were <u>sent</u> the questionnaires. Income figures may not be reported accurately (intentionally), or may be rounded off to the nearest $5,000, say, in a self-administered questionnaire.
- **d.** Rounding income responses would make the figure $25,111 unlikely. The fact that higher income respondents would be more likely to respond (bragging), and the fact that incomes are likely to be exaggerated, would tend to make the estimate too high.

2.10

- **a.** Simple random sampling.
- **b.** Stratified sampling.
- **c.** Cluster sampling.

2.11

a. Simple random sampling.
b. Stratified sampling.
c. Cluster sampling.

2.13

a. Stratified sampling. Stratify by job category and then take a random sample within each job category. Different job categories will use software applications differently, so this sampling strategy will allow us to investigate that.
b. Systematic random sampling. Sample every tenth patient (starting from a randomly selected patient from the first ten patients). Provided that there is no relationship between the type of patient and the order that the patients come into the emergency room, this will give us a representative sample.

2.13

a. Stratified sampling. We should stratify by type of degree and then sample 5% of the alumni within each degree type. This method will allow us to examine the employment status for each degree type and compare among them.
b. Simple random sampling. Once we find 100 containers we will stop. Still it will be difficult to get a completely random sample. However, since we don't know the locations of the containers, it would be difficult to use either a stratified or cluster sample.

2.14

a. Water temperature and Type of hardener
b. Water temperature: 175 °F and 200 °F; Type of hardener: H_1, H_2, H_3
c. Manufacturing plants
d. Plastic pipe
e. Location on Plastic pipe
f. 2 pipes per treatment
g. 6 treatments:
(175 °F, H_1), (175 °F, H_2), (175 °F, H_3), (200 °F, H_1), (200 °F, H_2), (200 °F, H_3)

2.15

a.

- Factors: Location in orchard, Location on tree, Time of year
- Factor levels: Location in orchard – 8 sections; Location on tree – top, middle, bottom; Time of year – October, November, December, January, February, March, April, May
- Blocks: none
- Experimental units: Location on tree during one of the 8 months
- Measurement units: oranges
- Replications: For each section, time of year, and location on tree, there is one experimental unit, hence 1 replication.
- Treatments: 192 combinations of 8 sections, 8 months, and 3 locations on tree –
- (S_i, M_j, L_k), for $i = 1,...,8$; $j = 1,...,8$; $k = 1,2,3$

b.
- Factors: Type of treatment
- Factor levels: T_1, T_2
- Blocks: Hospitals
- Experimental units: Wards
- Measurement units: Patients
- Replications: 2 wards per treatment in each of the 8 hospitals
- Treatments: T_1, T_2

c.
- Factors: Type of treatment
- Factor levels: T_1, T_2
- Blocks: Hospitals, Wards
- Experimental units: Patients
- Measurement units: Patients
- Replications: 2 patients per treatment in each of the ward/hospital combinations
- Treatments: T_1, T_2

d.
- Factors: Type of school
- Factor levels: Public; Private – non-parochial; Parochial
- Blocks: Geographic region
- Experimental units: Classrooms
- Measurement units: Students in classrooms
- Replications: 2 classrooms per each type of school in each of the city/region combinations
- Treatments: Public; Private – non-parochial; Parochial

2.16

a. Factors: Temperature, Type of seafood
b. Factor levels: Temperature (0 °C, 5 °C, 10 °C); Type of seafood (oysters, mussels)
c. Blocks: None
d. Experimental units: Package of seafood
e. Measurement units: Sample from package
f. Replications: 3 packages per temperature
g. Treatments: (0 °C, oysters), (5 °C, oysters), (10 °C, oysters), (0 °C, mussels), (5 °C, mussels), (10 °C, mussels)

2.17

a. Randomized complete block design with blocking variable (5 farms) and 48 treatments in a 3 × 4 × 4 factorial structure.
b. Completely randomized design with 10 treatments (software packages) and 3 replications of each treatment.
c. Latin square design with blocking variables (position in kiln, day), each having 8 levels. The treatment structure is a 2 × 4 factorial structure (type of glaze, thickness).

2.18

a. Design B. The experimental units are not homogeneous since one group of consumers gives uniformly low scores and another group gives uniformly high scores, no matter what recipe is used. Using design A, it is possible to have a group of consumers that gives mostly low scores randomly assigned to a particular recipe. This would bias this particular recipe. Using design B, the experimental error would be reduced since each consumer would evaluate each recipe. That is, each consumer is a block and each of the treatments (recipes) is observed in each block. This results in having each recipe subjected to consumers who give low scores and to consumers who give high scores.

b. This would not be a problem for either design. In design A, each of the remaining 4 recipes would still be observed by 20 consumers. In design B, each consumer would still evaluate each of the 4 remaining recipes.

2.19

a. "Employee" should refer to anyone who is eligible for *sick days*.

b. Use payroll records. Stratify by employee categories (full-time, part-time, etc.), employment location (plant, city, etc.), or other relevant subgroup categories. Consider systematic selection within categories.

c. Sex (women more likely to be care givers), age (younger workers less likely to have elderly relatives), whether or not they care for elderly relatives now or anticipate doing so in the near future, how many hours of care they (would) provide (to define "substantial"), etc. The company might want to explore alternative work arrangements, such as flex-time, offering employees 4 ten-hour days, cutting back to 3/4-time to allow more time to care for relatives, etc., or other options that might be mutually beneficial and provide alternatives to taking sick days.

2.20

a. Each state agency and some federal agencies have records of licensed physicians, professional corporations, facility licenses, etc. Professional organizations such as the American Medical Association, American Hospital Administrators Association, etc., may have such lists, but they may not be as complete as licensing records.

b. What nursing specialties are available at this time at the physician's offices or medical facilities? What medical specialties/facilities do they anticipate adding or expanding? What staffing requirements are unfilled at this time or may become available when expansion occurs? What is the growth/expansion time frame?

c. Licensing boards may have this information. Many professional organizations have special categories for members who are unemployed, retired, working in fields not directly related to nursing, students who are continuing their education, etc.

d. Population growth estimates may be available from the Census Bureau, university economic growth research, bank research studies (prevailing and anticipated load patterns), etc. Health risk factors and location information would be available from state health departments, the EPA, epidemiological studies, etc.

e. Licensing information should be stratified by facility type, size, physician's specialty, etc., prior to sampling.

2.21

If phosphorous first: [P,N]
[10,40], [10,50], [10,60], then [20,60], [30,60] or
[20,40], [20,50], [20,60], then [10,60], [30,60] or
[30,40], [30,50], [30,60], then [10,60], [10,60]

If nitrogen first: [N,P]
[40,10], [40,20], [40,30], then [50,30], [60,30] or
[50,10], [50,20], [50,30], then [40,30], [60,30] or
[60,10], [60,20], [60,30], then [40,30], [50,30]

2.22

	Factor 2		
Factor 1	I	II	III
A	25	45	65
B	10	30	50

2.23

a. Group dogs by sex and age:

Group	Dog
Young female	2, 7, 13, 14
Young male	3, 5, 6, 16
Old female	1, 9, 10, 11
Old male	4, 8, 12, 15

b. Generate a random permutation of the numbers 1 to 16:
15 7 4 11 3 13 8 1 12 16 2 5 6 10 9 14
Go through the list and the first two numbers that appear in each of the four groups receive treatment L_1 and the other two receive treatment L_2.

Group	Dog-Treatment
Young female	2-L_2, 7-L_1, 13, 14-L_2
Young male	3-L_1, 5-L_2, 6-L_1, 16-L_2
Old female	1-L_1, 9-L_2, 10-L_2, 11-L_1
Old male	4-L_1, 8-L_2, 12-L_2, 15-L_1

2.24

a. Bake one cake from each recipe in the oven at the same time. Repeat this procedure *r* times. The baking period is a block with the four treatments (recipes) appearing once in each block. The four recipes should be randomly assigned to the four positions, one cake per position. Repeat this procedure *r* times.

b. If position in the oven is important, then position in the oven is a second blocking factor along with the baking period. Thus, we have a Latin square design. To have $r = 4$, we would need to have each recipe appear in each position exactly once within each of four baking periods. For example:

Period 1		Period 2		Period 3		Period 4	
R_1	R_2	R_4	R_1	R_3	R_4	R_2	R_3
R_3	R_4	R_2	R_3	R_1	R_2	R_4	R_1

c. We now have an incompleteness in the blocking variable period since only four of the five recipes can be observed in each period. In order to achieve some level of balance in the design, we need to select enough periods in order that each recipe appears the same number of times in each period and the same total number of times in the complete experiment. For example, suppose we wanted to observe each recipe $r = 4$ times in the experiment. If would be necessary to have 5 periods in order to observe each recipe 4 times in each of the 4 positions with exactly 4 recipes observed in each of the 5 periods.

Period 1		Period 2		Period 3		Period 4		Period 5	
R_1	R_2	R_5	R_1	R_4	R_5	R_3	R_4	R_2	R_3
R_3	R_4	R_2	R_3	R_1	R_2	R_5	R_1	R_4	R_5

2.25
Discussion question; answers will vary.

2.26
Discussion question; answers will vary.

2.27
Discussion question; answers will vary.

2.28
Discussion question; answers will vary.

2.29
Discussion question; answers will vary.