Unit 2: Producing Data: Samples and Experiments

Chapter 5: Producing Data

5.1A: Designing Samples

Survey: - an exercise to ask a group of people about their responses to issues / products / preferences.

Experiment: - an exercise where treatments of various kinds are imposed on individuals to elicit a response.

Population: - the entire group of people who will be affected by the result of the experiment or information can be applied from a survey.

Census: - asking the survey questions to the **ENTIRE** population.

Sampling: - asking the survey questions to a **Sample (PART** of the population).

Sample Size: - the size of the sample compared to the rest of the population.

Confounding: - a lurking variable that has an effect on the response variable. The result at which renders an unclear relationship between the original exploratory variable and the measured response variable.

Statistical Inference: - a conclusion derived from an experiment or a survey that are both valid and reliable (or at least with fairly good statistical guarantee).

Example 1: Identify the population and explain whether the event (survey) is conducted from a sample or as a census.

a. Filling out the Student Information Demographic Forms.

Population: All students in a school

Because all students have to complete the form, it is a **census**.

b. Determining whether the American federal government should increase military spending.

Population: All American citizens who can legally vote.

Because the size of the population involved, a **sample** should be used.

c. Electing / Recalling the governor of the state of California.

Population: All California citizens who can legally vote.

Because it is an election, it should be a **census**.

d. Deciding whether Star Trek is better than Star Wars.

Population: All sci-fi fans who watches both Star Trek and Star Wars.

Because there are so many Star Wars and Star Trek fans, it can be done as a **sample**.

Sample Designs: - methods employed to choose the sample from a population.

1. **Non-Probability Sampling**: - sampling where a NON-RANDOM selection of the population is used.
   - it is important to identify these surveys are **bias**.

Bias: - when the survey will result in invalid (data that does not represent the population) or unreliable (same survey could not be repeated with the same result)

a. **Voluntary Response Sampling**: - the sample is chosen such that the favourable response will intentionally be resulted.
   - this includes giving away prizes for doing the survey, and samples where there is an over-representation of strong opinions (telephone call-ins, web surveys … etc).

b. **Convenience Sampling**: - the sample is taken from a place just because it is the surveyor closest position.

2. **Probability Sampling**: - sampling where a random selection of the population is used.
   - all individuals in the population have an equal chance to be selected for the sample.

a. **Simple Random Sampling (SRS)**: - all individuals in the population have an equal chance to be selected for the sample.

Table of Random Digits: - a table consisting of long strings of random digits from 0 to 9.
- there is an equal probability of any of the 0 to 9 digits to appear within each string, and all random digits do not relate to one another (completely independent).

Example 2: Using the row 128 of Table B: Random Digits at the back of the textbook, randomly select sample size of 5 students from a class of 26 people named A to Z.

1. Assign a two-digit number to each student.

01 A	04 D	07 G	10 J	13 M	16 P	19 S	22 V	25 Y
02 B	05 E	08 H	11 K	14 N	17 Q	20 T	23 W	26 Z
03 C	06 F	09 I	12 L	15 O	18 R	21 U	24 X	

2. Choose row 128 from the random digit table.

15689 14227 06565 14374 13352 49367 81982 87209

3. Group the strings of the random digits into groups of two digits. Select the first 5 groups that are within 01 to 26.

15 68 91 42 27 **06** 56 51 43 74 **13** 35 **24** 93 67 81 98 28 72 **09**

4. Identify the corresponding students in the sample

15-O 06-F 13-M 24-X 09-I

5.1A Assignment
pg. 248–249 #5.1 and 5.3
pg. 252–253 #5.5 and 5.7

5.1B: Other Sampling Designs

Other Probability Sampling

b. **Systematic Sampling**: - the sample is taken from every n^{th} element (units) of the population.

c. **Stratified Random Sampling**: - the population is divided into appropriate groups (strata), and then a random set of sample is taken from each group (stratum).

d. **Multistage (Clustered) Sampling**: - the population is divided into groups based on the nature of the survey questions and only a specific group (cluster) relevant to the question is sampled randomly.
- assumes each group has similar characteristic as the entire population.

Example 3: A survey needs to be designed to determine if all Santa Clara County Public Schools' Students should wear uniforms. Identify the population and classified the sampling techniques from the list below.

Population: All Santa Clara County Public Schools' Students, Parents, and Teachers.

Strategies	Sampling Techniques
The surveyor sends out a questionnaire with all the Santa Clara County School Boards' Quarterly Newsletters, and the results are tabulated from those completed questionnaires that have been returned.	**Volunteer Response Sampling**
20 Santa Clara County Public Elementary, Middle and High Schools are picked randomly from the list of all Santa Clara County Public Schools. The surveyor then randomly selects 10% of the students, parents and teachers from those schools to ask the question to.	**Clustered**
The surveyor went to the neighbourhood Santa Clara County Public School closest to her house and selected every 10^{th} person who walks in or out of the school on any school day.	**Convenient**
The list of all Santa Clara county public schools' students, parents, and school teachers and their phone numbers are complied separately; the surveyor then asks every 10^{th} person on each list.	**Stratified**
A list of all Santa Clara county public schools' students, parents, and school teachers and their phone numbers are complied; the surveyor will ask every 10^{th} person on the list by phone.	**Systematic**
A group of Santa Clara county public schools' parents, students and teachers are randomly chosen to represent 10% of the population.	**Simple Random**

Example 3: The high school student council has decided to put in a CD jukebox in the cafeteria. A decision needs to be made about the amount and the type of music to be stock into the jukebox. The council decided to conduct a survey by having students select their three most popular type of music from a list below. Identify the population and suggest a strategy for each probability sampling technique.

Cafeteria CD Jukebox Survey:

Please pick up to 3 of your favourite type of music below: **Current Grade Level:** _____

_____ Pop (Top 40)	_____ Dance	_____ Hard Rock	_____ Heavy Metal	_____ Rap
_____ Hip Hop	_____ R & B	_____ House	_____ Trance	_____ Country
_____ Independent	_____ World	_____ Sound Tracks	_____ Classical	_____ 70's & 80's

Others (please specify): _____

Population: All students who use the cafeteria frequently.

Sampling Technique	Strategy
Simple Random	100 randomly selected students in the cafeteria are taken during the lunch hour and before school starts to complete the survey.
Systematic	Ask every 5th student that walks into the cafeteria during the lunch hour and before school start to complete the survey. (Unit of Arrangement - every 5th student.)
Stratified	Find out from the office the percentage of Grade 9, 10, 11 and 12 in the school. Students are randomly chosen in the cafeteria before school starts and during the lunch hour to complete the survey. Only 10 % of each group will be used for the final tabulation. (Stratum: Grade 9, 10, 11 and 12.)
Clustered	10 classes are selected randomly from the school during the 3rd period. All students who frequent the cafeteria before school start and during the lunch hour are to complete the survey.

Different Types of Bias

a. **Selection Bias**: - when the population is not represented properly in the sample.

Undercoverage: - when some groups in the population are left out of the samples.
(Example: homeless people, prisoners, and people who have no phones ≈ 7% to 8% of Americans)

b. **Response Bias**: - when the phrasing of the question will lead to a one-sided response **(wording effect)**.
- when respondents lie about their answers willingly in illegal or unpopular behaviours.
- when respondents purposely give socially acceptable responses due to the race or sex of interviewers.
- when respondents provide inaccurate responses due to faulty memories.

c. **Non-Response Bias**: - when a large part of the sample did not respond to a survey or refuse to co-operate.
(Example: In general, 30% of telephone survey sample do not wish to answer any questions.)

***Notes*: - Results of surveys can only be valid when the actual question posed, sampling design, amount of non-response, possible undercoverage, and date of the survey is published and analyzed to be adequate and unbiased.**

- The law of probability states that the results of experimental probabilities will be very similar to the theoretical probability given a large enough trials. In surveys employing probability sampling, it means that large sample size surveys will be more accurate and reliable (less margin of error) than surveys performed with smaller samples.

Example 4: Determine the type of bias and the inference that it will likely occur with the statement or statement below. Suggest a correction for each case.

a. 9 out of 10 dentists agree that Toothpaste A is way better than the Toothpaste B. Which toothpaste do you prefer?

(Response Bias with Wording Effect) The question leads people to believe that most dental experts prefer Toothpaste A. Therefore, the likely inference is that more people will answer Toothpaste A rather than Toothpaste B.

Correction: Which toothpaste do you prefer, Toothpaste A or Toothpaste B?

b. 100 people were randomly selected at the hockey arena after a hockey game to determine the percentage of Americans that watches hockey on a regular basis.

(Selection Bias) Most people coming out of a hockey game tend to like hockey. Therefore, the likely inference is that a large population will respond that they watch hockey on a regular basis.

Correction: Move the survey location to a place where the population is better represented (local shopping malls or systematic telephone survey).

c. To survey satisfaction levels of customers, a restaurant owner leave comment cards on the tables of her establishment for customers to voice their opinions.

(Non-Response Bias) The survey method will likely generate small number of response. And those who response most likely has negative comments about the service.

Correction: Have every 5th customers complete a small survey, as they were about to pay for their bill.

5.1B Assignment
pg. 256 #5.9; pg. 259 #5.11; pg. 261–262 #5.15 and 5.17

5.2A: Designing Experiment

Observational Study: - an exercise where the measurement of variables are from observation of individuals with no attempt to influence any responses.
- all survey samples are observational study.

Experiment: - an exercise where treatments of various kinds are imposed on individuals to elicit a response.
- again, all lurking variables have to be identified and controlled for the results of the experiments to be valid.

Terminology of an Experiment

1. **Experimental Units**: - the individuals in an experiment.
   - **Subjects**: - when the individuals are human beings.
2. **Treatment**: - the specific experimental condition being applied to the units.
3. **Factors**: - the explanatory variables in the experiment.
   - there can be more than one factor in an experiment. In such case, a two-way table is used to organize the results.
4. **Levels**: - specific values of each factor (measurement intervals).

Example 1: An experiment was performed on 80 college students on amount of sleeps and breakfast on subsequent memory recall ability during the morning. The experimenter decided on 2, 4, 6, and 8 hours of sleep with 200, 400, 600, 800 calories of breakfast. Each subject was than given a reading and asked to recall the facts in a short test.

a. What were the experimental units?
b. Identify the factor(s). Explain all levels involved if any.
c. What was the response variable?
d. Provide a possible observation table.

a. Experimental Units: 80 College Students (Human Subjects)
b. Factors:
1. Hours of Sleep (Levels: 2, 4, 6 and 8 Hours).
2. Calories of Breakfast (Levels: 200, 400, 600, 800 calories).
c. Response Variable: Amount of Recall.

	Breakfast Calories			
Hours of Sleep	**200 Calories**	**400 Calories**	**600 Calories**	**800 Calories**
2 Hours				
4 Hours				
6 Hours				
8 Hours				

Comparative Experiments: - experiments that involved at least two groups of experimental units. One group was given the treatment where the other was given no treatment (**Control Group**).
- was used mostly to eliminate any confounding situations with unknown lurking variables.

Placebo Effect: - when both groups of a comparative experiment (Treated and Control groups) yield similar results. This is an indication that the treatment does NOT cause the actual response observed.

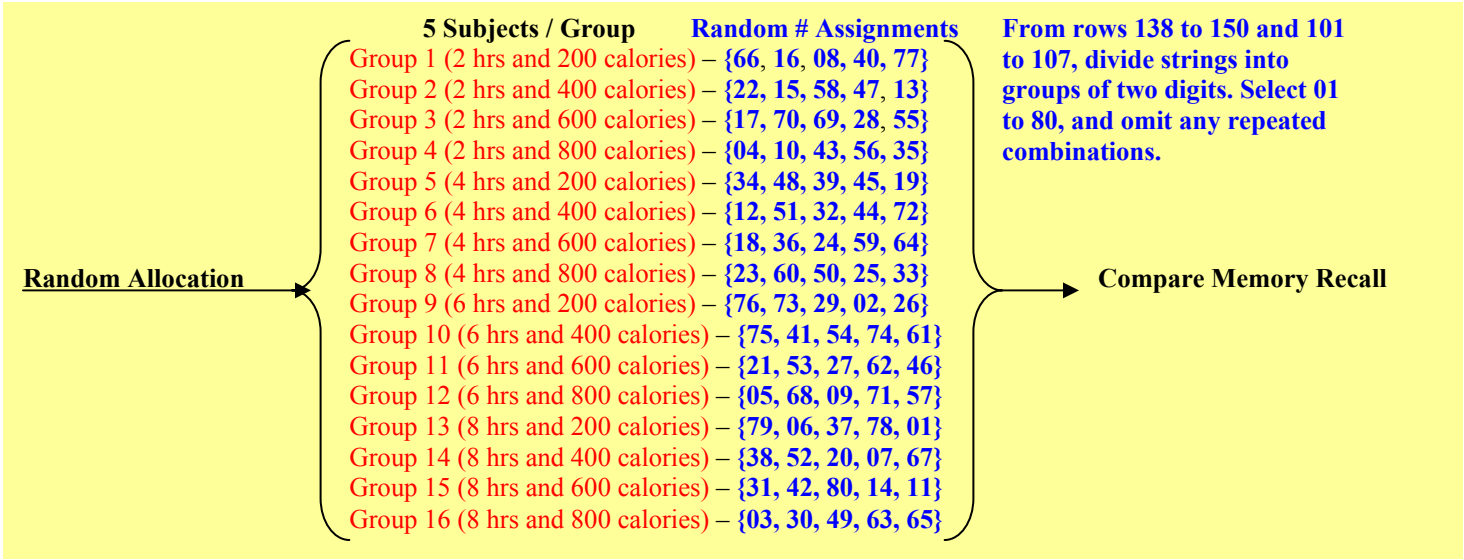
Example: An experimental drug was given to a group of patients with chronic migraine. A placebo (sugar pills) is given to another group of patients also suffering from chronic migraine. After two months, both groups indicate a general improvement of their conditions. Some in the control group even report that the "drug" had eliminated their migraine all together. This is a clear indication that the drug is not the cause of diminishing the pain associated with chronic migraine, let along eliminating the condition all together.

Randomization: - a process where the experimental units selected from the sample to form any groups is completely randomized by employing the use of impersonal chance.

Matching: - a method of matching experimental units with similar characteristics for both the treated and control groups.
- it is more often an inadequate attempt rather than helpful because many times there are too many lurking variables to isolate. Thereby, it is very difficult to ensure all characteristics are matched between both groups.

Completely Randomized Design: - an experimental design where all experimental units are randomly distributed for all the treatments.

Example 2: Using a diagram, describe a completely randomized design for the experiment in example 1 of this section. Indicate the size of the each treatment group. Starting with row 138 of the Table B: Random Digits at the back of the textbook, assign the subjects to each group of this experiment.



5.2A Assignment
pg. 268 #5.27 and 5.29; pg. 274 #5.31 and 5.33

5.2B: The Logic of Experimental Design

The Logic behind a Completely Randomized Comparative Design

1. Randomization of groups allows the elimination of the any personal chance in the selection process.
2. Comparative Design minimizes the possibility of bias from unknown lurking variables, which will yield a confounding situation.
3. Hence, the differences in the results must be due to the application of various treatments. The experiment can be called valid. The reliability of this experiment can be tested with other similar experiments. Possible cause-and-effect can now be hypothesized and be further studied.

Statistical Significance: - the result of a properly conducted completely randomized comparative experiment where the changes in the response variable is large enough, such that it is not caused by some chance occurrence, but rather from the application of various treatments.

Three Principles of Experimental Design:

1. **Control**: - by comparing various treatments, we can eliminate various lurking variables.
2. **Randomization**: - using impersonal chance (Random Digits Table and Random Digits Generator Computer Software) to assign experimental units into groups.
3. **Replications**: - the larger the size of experimental units in each group, the smaller the variation will result due to chance.

Other Bias and Weakness in an Experiment:

1. **Hidden Bias**: - when an experimenter subconsciously forms a bias towards a particular treatment in an experiment.
   - this may result in the attitude change towards the experimental units such that the selection of group is not completely randomized or response bias will result from wording effect.

Example: When the experimenter in a study is also the health care-giver of the patients taking part in the experiment, a possible factor of empathy might create a hidden bias to the type of treatment this patient will get.

Double-Blind Experiment: - when neither the experimenter and nor the subjects are aware what treatments are given or received.

2. **Lack of Realism**: - when the treatments do not replicate a realistic condition for the subjects.

Example: Measuring the adrenaline level of subjects in an emergency drill. (The subjects know that it is a drill and not a real emergency. Therefore, they would definitely be calmer in the drill than in the real thing).

Other Experimental Designs

1. **Block Design**: - when two distinct groups are suspected to yield very different outcomes. The experimental units are blocked and similar treatment conditions are applied to each block.

Example: Men and Women form different groups to study the response to a new romantic drama motion picture. It is a stereotype that most men like action movies and women like romantic drama shows. However, a block design can definitely show whether such stereotypes are in fact true without being labelled as a bias experimental design.

2. **Matched Pairs Design**: - when a block design just compared the results from two treatments. The two blocks however are matched as closely as possible.
   - this can include one experimental units is offered two treatments at separate time. The danger is that the order of the treatments given may influence in the response.
   - as mentioned before, by trying to matched all experimental units so their characteristics are as close as possible violates the rule of randomization. As such, matched pair designs are not usually statistically valid.

Example: Comparing two dishes in a cooking contest can depend on which dish the judge taste first. This is a matched pairs design because one experimental unit (the judge) is subjected to two treatments (two dishes). Besides, the judge knows who is the cook for each dish.

5.2B Assignment
pg. 277 #5.35, 5.37; pg. 279 #5.39; pg. 282 #5.41

5.3: Simulating Experiments

Simulation: - an experiment carried out using the imitation of chance behaviours utilizing tools such as coin flips, dice rolls, and spinners … etc.

Probability Model: - a model that states an experiment carried out by simulation will yields result similar to ones carried out in reality given that it is a randomized experiment performed over many trials.

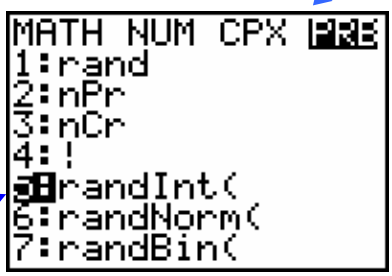
randInt (Random Integer): generates and displays a random integer from a specified range and number of trials.

randInt (Lower Integer Limit, Upper Integer Limit, Number of Trials)

If omitted, it is assumed one trial

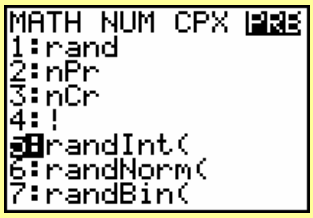
To access randInt:

1. Press **MATH**
2. Use ▶ to access PRB
3. Select Option 5

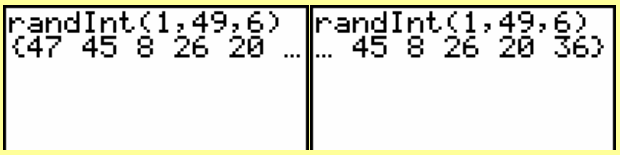


Example 1: Generate 6 integers out of 1 to 49 for the next Lotto draw.

1. Press **MATH**
2. Use ▶ to access PRB
3. Select Option 5



4. randInt(1, 49, 6) **ENTER**



The Random Integers are (47, 45, 8, 26, 20 and 36)

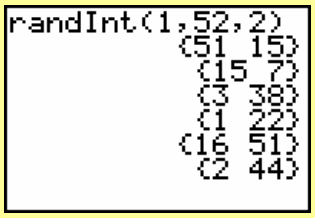
Note: **The random function of the calculator depends on its seed value. Unless the seed value is reset, even calculators with the same model may yield different results.**

Example 2: Simulate drawing 2 cards out of a deck of 52 cards that will be a black jack (10 or Picture with ace) 25 rounds with replacement.

Since there are 52 cards in a deck with 4 suits (diamonds, hearts, spades and clubs), we will let 1 to 52 represents the 52 cards. Using the following table to assign the cards,

Number	Card	Number	Card	Number	Card	Number	Card
1	♦ A	14	♥ A	27	♠ A	40	♣ A
2	♦ 2	15	♥ 2	28	♠ 2	41	♣ 2
3	♦ 3	16	♥ 3	29	♠ 3	42	♣ 3
4	♦ 4	17	♥ 4	30	♠ 4	43	♣ 4
5	♦ 5	18	♥ 5	31	♠ 5	44	♣ 5
6	♦ 6	19	♥ 6	32	♠ 6	45	♣ 6
7	♦ 7	20	♥ 7	33	♠ 7	46	♣ 7
8	♦ 8	21	♥ 8	34	♠ 8	47	♣ 8
9	♦ 9	22	♥ 9	35	♠ 9	48	♣ 9
10	♦ 10	23	♥ 10	36	♠ 10	49	♣ 10
11	♦ J	24	♥ J	37	♠ J	50	♣ J
12	♦ Q	25	♥ Q	38	♠ Q	51	♣ Q
13	♦ K	26	♥ K	39	♠ K	52	♣ K

We will use the command line randInt(1, 52, 2) and repeat 25 times. Writing down the results will be like simulating the drawing of 2 cards out of a deck 25 times.



For example, the first six draws as shown on the left represent {♣Q and ♥2}, {♥2 and ♦7}, {♦3 and ♠Q}, {♦A and ♥9}, {♥3 and ♣Q}, and {♦2 and ♣5}.

5.3 Assignment
pg. 293 #5.57;
pg. 296 #5.59;
pg. 297 #5.61

Chapter 5 Review
pg. 262–263 #5.18 to 5.21; pg. 264 #5.23
pg. 284–285 #5.45, 5.47, 5.49
pg. 297 #5.62, 5.63
pg. 300–304 #5.67, 5.73, 5.75, 5.79, 5.81

