Theory of Survey Sampling for STAT112

Charles Fleming

April 16, 2018

1 Design of Surveys

Some statisticians make a distinction between a survey and an experiment by pointing out that experimental variables are controllable, whereas, in a survey, they cannot be controlled. In an experiment, for example, of heating a cup of water in a microwave oven, the time of duration of heating is a controllable variable, and therefore, the resulting temperature of the water is controllable. On the other hand, in a survey, there are some people who may be unemployed, there may be some who are retired, and there may be others who have been involuntarily unemployed. These conditions cannot be controlled by the experimenter nor can the experimenter control the effect that employment has on a person's self-esteem. There are other statisticians who argue that a survey is just a special case of an experiment and we will use that understanding here.

Although the population which a researcher has defined contains all the information that he is seeking, getting that information constitutes the goal of performing an experiment. It is infeasible in practical matters to examine every element of the population even if the list of elements of the population is perfect. The determination of the sampling size given the constraint of limited financial resources in order to achieve a prescribed precision in the estimate is a difficult mathematical calculation. Before the sample is drawn, the design of the survey has already taken into account the specific mathematical formulas which will be appropriate in calculating the estimates.

A simple experiment will illustrate some of these concepts. The job is assigned to count the number of blades of grass which are alive in a suburban one acre lot. If the turf is well groomed and has uniform thickness, then a clever statistician could count the number of blades of grass in a 4" by 4" square and by means of proportions find the number of blades of grass in a square foot and then continue to find the number of blades in the lot. So, let $\mathcal{B}$ be the number of blades of grass in a 4" by 4" square. There are nine 4" by 4" squares in a square foot. There are 43560 square feet in an acre; therefore the number of blades of grass in a one acre lot is: $\mathcal{B}$(9)(43560)=392040$\mathcal{B}$.

Usually, however, the turf is blemished. There is probably a house, driveway, sidewalks on the lot. The turf which grows under trees will be thin. The natural approach would be to partition the lot into areas of similar composition and qualities. The driveway, sidewalks, and the house where no grass grows could be one partition. The area under the trees where the density of the turf is light might constitute another. The area in full sunlight could still be another area. By knowing the area of each partition of the lot and the number of blades of grass found in a 4” by 4” square, the grand total can be estimated by adding the estimated totals from each partition.

The lot is analogous to a population. But the population is not uniformly covered with the information which is being sought. Some elements might represent the poor, some the rich, others might represent the healthy or the infirm, and so on. The process of partitioning the population into areas of uniform composition, in order to facilitate the process of producing precise estimates is called stratification. When the population is stratified, it is stratified into strata. Each stratum is defined in such a way that its elements are uniformly similar. The numerous the strata, does not mean better precision. The balance between the right number of strata and obtaining the highest precision is a difficult problem which is usually solved by intuition and the method of trial and error.

A significant problem which faces experimenters of the social sciences is the necessity of using an artificial scale of measurements. Unlike the physical sciences like physics in which observations span a continuum of values, the social sciences almost always obtain data according to some artificial scales. They are given the name *nominal, ordinal, interval, ratio*. A *nominal* measurement is associated with a categorical attribute of a subject. For example, 1 for male, and 2 for female. Or 1 for employed; 2 for retired, 3 for involuntarily unemployed, and 4 for other.

An *ordinal* measurement applies to some kind of ranking. For example, 1 for strongly disagree, 2 for disagree, 3 for no opinion, 4 for agree, 5 for strongly agree. An important limitation of the ordinal measurement is that they show a relative ranking rather than one based on an absolute standard. Does a value of 5 mean that it is five times better and a 1? The scales could have been reversed: 5 for strongly disagree, 4 for disagree, 3 for no opinion, 2 for agree, and 1 for strongly agree. Sometimes a scale might include answers with no useful value like 6 for no response and 7 for not applicable. Obviously, 6 does not mean that a no response is twice as good as a no opinion.

An *interval* measurement is the kind which scientists and engineers use all the time when they measure such things as voltage and time. *Ratio* measurements are interval measurements which have been translated to the origin. Specifically, if $y = a + bx$ and $z = \alpha + \beta x$ then $y - a$ and $z - \alpha$ are ratio measurements because $\frac{y-a}{z-\alpha} = \frac{a}{\alpha}$ = a constant, i.e. the x disappears. The absolute temperature scale which is used in thermodynamics is a classic example of a ratio measurement. The Fahrenheit and centigrade (Celsius) temperature scales are interval measurements but Fahrenheit-456^oF and Celsius-273^oC are ratio measurements because they have been translated to absolute zero.

2 Content Analysis

Creating a scale for answers to questions is intuitive. Answers to questions for which an essay or "free-response" is expected require the identification of key ideas which appear in the answer. The answers for the question, *HOW DO YOU DEFINE HAPPINESS?*, could be as different with one another as there are people who supplied answers. In the collection of answers, the same themes or concepts might frequently occur so that they may be consolidated into a few general categories like good health, optimism, and good financial security. In analyzing the contents of answers coming from free response questions, common themes are identified. This process of identifying common themes in essays, magazine articles, speeches, and answers is called content analysis. Each category might be given a code which would facilitate the entry of the data into a computer for subsequent analysis. The insights which can be gained from an analysis of the data depends on the precision of the questions, the clear inter-relationships of the questions between themselves, the form of the scales, and the quality of the content analysis. Designing a survey is a science unto itself. Even such seemingly insignificant considerations like the font, color of paper, voice and appearance of the interviewer affect the quality of the responses. Molecules and electricity do not care what the scientist looks like but people who are being interviewed do care about such things and quite often fashion their answers according to the impressions which they make of the interview.

3 Assumptions

1. Independent observations

2. Unbiased data

3. Strata are disjoint

4. Identically distributed

5. Theory makes sense

6. Sample is representative of the population

4 Considerations when Planning a Survey

1. State the objective.

2. Precisely define the population in terms of space and time.

3. Construct the list also known as the sampling frame.

4. Develop the plan of sampling. Calculate optimum sampling size and optimum allocation across strata.
5. Choose the method of measurement.
6. Compose the survey instrument for taking the measurements.
7. Use of Focus groups.
8. Hire and train personnel.
9. Pre-test the survey.
10. Organize the enumerators, regional offices, and headquarters staff,
11. Organize the sets of data with such precautions as security and damage to the files.
12. Conduct the analysis of the data.
13. Method of publishing the results.

5 Sources of Error

1. Drawing a sample.
2. Lying by the respondent
3. Omission of data
4. Non-response
5. Data entry
6. Bad computer programming
7. Misunderstanding
8. Personal appearance, etiquette, and behavior
9. Inaccessible
10. Partial response
11. Sensitive information.

6 Estimation

Based on intuition, we reason that in the process of estimating a characteristic of a population, the information which is obtained from a survey can be generalized to the whole population provided that the sample is representative of the population.

In the process of estimating the number of blades of grass which could be seen on a one acre suburban lot, the number of blades of grass found in a $4" \times 4"$ square was expanded to the number of blades of grass in the one acre lot by knowing that there are $9(43560)$ $4" \times 4"$ squares in an acre. We reasoned that the number of blades of grass in that acre would be $9(43560)B$ where B is the counted number of blades of grass found in the $4" \times 4"$ square.

Another way to view the problem is to frame it in terms of probability. The probability of selecting any one $4" \times 4"$ square at random is $\frac{1}{9(43560)}$. We will denote the probability of the selection of a unit, π_i. Because all $4" \times 4"$ squares, in our example, have the same characteristics and that they are selected at random, then we may assert that $\pi_i = \pi_j$ for any i and j unit in the lot. That is, the probability of selecting a $4" \times 4"$ unit is uniformly distributed such that, if there are N units in the population, $\pi_i = \frac{1}{N} \; \forall i$. With respect to our example, the number of blades of grass in the one acre lot will be $\widehat{\tau} = 9(43560)B_i = \frac{B_i}{\frac{1}{9(4360)}} = \frac{B_i}{\pi_i}$ where B_i is the number of blades of grass in the i^{th} $4" \times 4"$ square. This example leads to the following theorem.

Theorem 1 *Let $\widehat{\tau}$ be the estimate of the population total of some quantity, then $\widehat{\tau} = \frac{x_i}{\pi_i}$ where x_i is a measurement of a quantity in the i^{th} unit and π_i is the probability of selecting unit i.*

We can generalize Theorem 1 to estimate the population total from a sample of more than one element. Denote a sample of size n by $\mathcal{S}_n$ with elements denoted by l_i. That is, $\mathcal{S}_n = \{l_1,\ l_2,\ l_3, \ldots, l_n\}$.

Theorem 2

$$\widehat{\tau} = \sum_{i=1}^{n} \frac{x_i}{P(l_i \in \mathcal{S}_n)} \tag{1}$$

For example, suppose the sample has three elements like three members of Congress, and let $x_1,\ x_2,\ x_3$ be their reported taxable incomes. We want to estimate the total taxable income of the members of the Congress. The size of the population is 535. Let π_1 be the probability of selecting the first member of Congress in our sample of three. Let π_2 and π_3 be the probabilities of selecting the remaining two members of Congress in our sample of three.

Let us focus on Congressman, l_1. We want to find the probability that he will be selected for our sample, $\mathcal{S}$, of size three. That implies that he will have to be selected on the first draw or the second draw or on the third draw. If each element of $\mathcal{S}$ is selected without replacement with equal initial probability of selection, then the probability of selecting l_1 on the first draw will be $\frac{1}{N} = \frac{1}{535}$. Later in Section 7, we will see that the provision of selecting an element without replacement with equal initial probability of selection produces a simple result. Regardless of the stage of sampling, the probability of selection remains $\frac{1}{N}$. Consequently, the probability of selecting l_1 on the second draw will be $\frac{1}{N}$, and the probability of selecting l_1 on the third draw will, also, be $\frac{1}{N}$. The probability that Congressman, l_1, will be selected in our sample of size three will be: $P(l_1 \in \mathcal{S}_3) = \frac{1}{N} + \frac{1}{N} + \frac{1}{N} = \frac{3}{N} = \pi_1$. Likewise, $P(l_2 \in \mathcal{S}_3) = \frac{3}{N} = \pi_2$, and $P(l_3 \in \mathcal{S}_3) = \frac{3}{N} = \pi_3$. According to Theorem 2,

$$\widehat{\tau} = \frac{x_1}{\pi_1} + \frac{x_2}{\pi_2} + \frac{x_3}{\pi_3} = \frac{x_1}{\frac{3}{N}} + \frac{x_2}{\frac{3}{N}} + \frac{x_3}{\frac{3}{N}} = \frac{x_1 + x_2 + x_3}{\frac{3}{N}} = N\overline{x}$$

which agrees with our intuition. We are led to the next theorem.

Theorem 3 *If members of a sample of size n is drawn without replacement with equal initial probabilities selection from a list of size N, then the estimate of the population total is:*

$$\widehat{\tau} = N\overline{x}$$

Notice that the condition of uniform probability of selection stated in Theorem 3 produces a very simple formula for estimating the population total otherwise if the probabilities of selection are not uniformly distributed, then we must resort to Theorem 2 which usually produces extremely complex formulas. For example, suppose that there are only four members of Congress, that is: N=4 instead of N=535, and suppose the initial probabities are not equal but are the same ones shown in Table 2 found on page 15. The probability for Congressman, l_1, to be selected in the sample of size three, now, becomes:

$$\begin{aligned} P(l_1 \in S_3) &= P(X_1 = 1 \text{ } or \text{ } X_1 = 2 \text{ } or \text{ } X_1 = 3) \\ &= \frac{1}{3} + \frac{44}{135} + \frac{4606}{19305} \\ &= \frac{17333}{19305} \end{aligned}$$

Likewise, $P(l_2 \in \mathcal{S}_3) = \frac{38999}{42075}$ and $P(l_3 \in \mathcal{S}_3) = \frac{2593}{3672}$. According to Theorem 2, $\widehat{\tau} = \frac{x_1}{\frac{17333}{19305}} + \frac{x_2}{\frac{38999}{42075}} + \frac{x_3}{\frac{2593}{3672}}$. One can imagine, the extremely complicated formulas for drawing three members of Congress without replacement with unequal probabilities from a population of 535.

To avoid complicated formulas is the reason why simple random sampling is so popular even though it might not be the most efficient method of sampling.

6.1 Effect of Finite Population

By independence we mean that $P(A|B) = P(A)$ for events A and B. By identically distributed, the random variables have the same probability distribution which implies that they have the same means and variances, i.e. (μ, σ^2). Both properties are essential for deriving the properties of $\bar{x}$. Theorem 4 gives the expected value and variance of $\bar{x}$.

We note that by definition, $var(X) = E[(X - E[X])^2]$, and based on it $var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$ where the co-variance $cov(X, Y) = E[(X - E[X])(Y - E[Y])]$. When events A and B are independent, then $P(A \cap B) = P(A)P(B)$ so that $E[AB] = E[A]E[B]$ and $cov(X, Y) = E[(X - E[X])]E[(Y - E[Y])]$. But $E[(X - E[X])] = E[X] - E[X] = 0$; therefore, when X and Y are independent, $cov(X, Y) = 0$ which produces the simple result, $var(X + Y) = var(X) + var(Y)$. This is the main benefit of independent events and the highly desired property of statistical data like survey data. By virtue of i.i.d., then, we are able to prove the very simple Theorems 4 and 5.

Theorem 4 *If $X_1,\ X_2,\ \ldots,\ X_n$ are i.i.d. each with mean μ and variance σ^2, and $\bar{x} = \frac{X_1 + \cdots + X_n}{n}$, then*

$$E[\bar{x}] = \mu \textit{ and } var(\bar{x}) = \frac{\sigma^2}{n}$$

Theorem 5 *If $X_1,\ X_2,\ \ldots,\ X_n$ are i.i.d. each with mean μ and variance σ^2, and $s^2 = \frac{\sum_{i \in \mathcal{S}} (x_i - \bar{x})^2}{n-1}$, then*

$$E[s^2] = \sigma^2 \textit{ and } var(s^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) \textit{ where } \mu_4 = E[(X - \mu)^4]$$

From Theorem 4, we learned that if $X_1,\ X_2,\ \ldots,\ X_n$ are i.i.d. each with mean μ and variance σ^2, and $\bar{x} = \frac{X_1 + \cdots + X_n}{n}$, then

$$E[\bar{x}] = \mu \text{ and } var(\bar{x}) = \frac{\sigma^2}{n}$$

Theorem 4 applies to a population of infinite size. In surveys, however, the populations are finite, though they might be very large. A finite population imposes a constraint on the calculation of the variance in that if we were to conduct a census of a population we will have had examined every element in the population, consequently our estimates of the population mean and of the population variance will be known with certainty. That is, $var(\bar{x}) = var(\mu) = 0$ and not $var(\bar{x}) = \frac{\sigma^2}{n}$ according to Theorem 4.

This constraint which a finite population imposes on the variance is taken into account in Theorem 6.

Theorem 6 *If from a finite population of size N, $X_1,\ X_2,\ \ldots,\ X_n$ are i.i.d. each with mean μ and variance σ^2, and $\bar{x} = \frac{X_1 + \cdots + X_n}{n}$, then*

$$E[\bar{x}] = \mu \textit{ and } var(\bar{x}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

Accordingly, the variance of the estimated population total is expressed in Corollary 1

Corollary 1 *If from a finite population of size N, $X_1,\ X_2,\ \ldots,\ X_n$ are i.i.d. each with mean μ and variance σ^2, and $\bar{x} = \frac{X_1+\cdots+X_n}{n}$, then*

$$E[\widehat{\tau}] = \tau = N\bar{x} \textit{ and } var(\widehat{\tau}) = N^2\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

In both Theorem 6 and its Corollary 1, we still do not know μ and σ^2.

It can be proven in a course on sampling that $E[s^2] = \left(\frac{N}{N-1}\right)\sigma^2$ for a finite population; therefore, $\widehat{\sigma^2} = \left(\frac{N-1}{N}\right)s^2$ is an unbiased estimator of σ^2, because we can now show, $E[\widehat{\sigma^2}] = \sigma^2$. By direct substitution into $\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$, we obtain an unbiased estimator of $var(\bar{x})$, namely

$$\widehat{var(\bar{x})}= \left(\frac{N-n}{N-1}\right)\frac{\widehat{\sigma^2}}{n} = \left(\frac{N-n}{N}\right)\frac{s^2}{n}$$

These results are summarized in Theorem 7.

Theorem 7 *If from a finite population of size N, $X_1,\ X_2,\ \ldots,\ X_n$ are i.i.d. each with mean μ and variance σ^2, and $\bar{x} = \frac{X_1+\cdots+X_n}{n}$, then*

1. $\widehat{E[\bar{x}]}= \bar{x}$ *and* $\widehat{var(\bar{x})}= \left(\frac{N-n}{N}\right)\frac{s^2}{n}$
2. $\widehat{E[\widehat{\tau}]}= N\bar{x}$ *and* $\widehat{var(\widehat{\tau})}= N^2\left(\frac{N-n}{N}\right)\frac{s^2}{n}$

Example 1 *Three 4" × 4" square areas were selected at random from a one acre suburban lot. The number of blades of grass were counted in each unit. The counts are: 90, 83, 96. N=9(43560)=392,040; $\bar{x} = 89.6667$; $s^2 = 42.330$; and s=6.5064.*

1. $\widehat{E[\bar{x}]}= 89.6667$
2. $\widehat{var(\bar{x})}= \left(\frac{392040-3}{392040}\right)\frac{42.330}{3} = 14.1111 \rightarrow \sqrt{\widehat{var(\bar{x})}}= 3.7564$
3. $\widehat{E[\widehat{\tau}]}= 392040(89.6667) = 35,152,920$
4. $\widehat{var(\widehat{\tau})}= 392040^2\left(\frac{392040-3}{392040}\right)\frac{42.330}{3} = 2.1687 \times 10^{12} \rightarrow \sqrt{\widehat{var(\widehat{\tau})}}= 1,472,683$

Assume that the Central Limit Theorem applies to this problem; therefore, the lower limit of a confidence interval will be for the sample mean $a = \bar{x} - t_{n-1,\frac{\alpha}{2}}\sqrt{\widehat{var(\bar{x})}}= 89.6667 - 1.96(3.756) = 82.30$ and $89.6667 + 1.96(3.756) = 97.02$, so that (82.30,97.02) is the 95% confidence interval for the population mean, μ.

Similarly for the population total, $a = \widehat{\tau} - t_{n-1,\frac{\alpha}{2}}\sqrt{\widehat{var(\tau)}}$= $32,266,461$ and the upper limit, b=38,039,379, so that the 95% confidence interval for the total number of blades of grass in the one acre lot is (32266461,38039379).

Definition 1 *The coefficient of variation is*

$$CV = \frac{\sqrt{var(estimate)}}{estimate}$$

Accordingly, for the preceding Example 1,

1. CV of $\bar{x} = \frac{\sqrt{\widehat{var(\bar{x})}}}{\widehat{x}} = \frac{3.7564}{89.6667} = .0418 = 4.18\%$

2. CV of $\widehat{\tau} = \frac{\sqrt{\widehat{var(\widehat{\tau})}}}{\widehat{\tau}} = \frac{1472683}{35152920} = .0418 = 4.18\%$

We see that the CV of $\bar{x}$ and the CV of $\widehat{\tau}$ are the same. This is because
CV of $\widehat{\tau} = \frac{\sqrt{\widehat{var(\widehat{\tau})}}}{\widehat{\tau}} = \frac{\sqrt{N^2\left(\frac{N-n}{N}\right)\frac{s^2}{n}}}{N\bar{x}} = \frac{\sqrt{\left(\frac{N-n}{N}\right)\frac{s^2}{n}}}{\bar{x}} = \frac{\sqrt{\widehat{var(\widehat{\mu})}}}{\bar{x}}$=CV of $\bar{x}$

6.2 Proportions

The example of estimating a population proportion as illustrated in Example 2 uses the property of a Bernoulli random variable which is characterized by being a mapping of an outcome taken from a sample space, Ω, of only two outcomes to the number line. Let $\Omega = \{\omega_1,\ \omega_2\}$. The two outcomes might be pass-fail, on-off, up-down, 0-1, success-failure. Define $X = 1$ to signify success and $X = 0$ to signify failure. Suppose $X_i \sim b(1,p)$, then $P(\{\omega \in \Omega | X(\omega) = 1\}) = p$ and $P(\{\omega \in \Omega | X(\omega) = 0\}) = 1 - p = q$. We know that $E[X] = p$ and $var(X) = pq$.

Theorem 8 *If from a finite population of size N, $X_1,\ X_2,\ \ldots,\ X_n$ are i.i.d. Bernoulli random variables with probability of success, p. and $\bar{x} = \frac{X_1+\cdots+X_n}{n}$, then*

1. $E[\bar{x}] = p$

2. $var(\bar{x}) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} = \left(\frac{N-n}{N}\right)\frac{pq}{n}$

3. $\widehat{E[\bar{x}]}$= $\widehat{p}$

4. $\widehat{var(\bar{x})}$= $\left(\frac{N-n}{N}\right)\frac{\widehat{p}\widehat{q}}{n-1}$

Example 2 *A salesman of a golfing supply house is curious to learn how many students at the local university use a one iron in one semester. He selected at random 100 students from the 13,000 student body and asked them whether or not the student had used a one iron at least once* *during the previous semester. Only one student reported that he had used a one iron the previous semester.*

Let

$$x_i = \begin{cases} 1 & \textit{if the student had used a one iron} \\ 0 & \textit{if a student did not use a one iron} \end{cases}$$

$\sum_{i=1}^{100} x_i$ *= total number of one irons. In this example,* $\sum_{i=1}^{100} x_i = 1$*; therefore,* $\widehat{p} = .01$ *and* $\widehat{q} = .99$*.*

We are informed that N=13,000. Consequently, $\widehat{\tau} = N\widehat{p} = 13000(.01) = 130$ *and*

$$\begin{aligned} \widehat{var(\widehat{p})} &= \left(\frac{12900}{13000}\right)\frac{(.01)(.99)}{99} = .00009925 \\ \sqrt{\widehat{var(\widehat{p})}} &= .0096 \end{aligned}$$

The 95% confidence interval of $\widehat{p}$ *is*

$$\begin{aligned} a &= \widehat{p} - 2\sqrt{\widehat{var(\widehat{p})}} = 0 \\ b &= \widehat{p} + 2\sqrt{\widehat{var(\widehat{p})}} = .0292 \end{aligned}$$

(0,.0292).

The confidence interval for the population total is N times that which is (0,249). The salesman is 95% confident that the number of students at the university who use a one iron is at most 249.

7 Sampling

We want i.i.d. to make the mathematics simple. Even with i.i.d., the formulas can be complicated. Nonetheless, to preserve the property of i.i.d. while drawing the sample will help make it representative of the population. The process of sampling is a key step in producing a successful survey. There are different methods of sampling to achieve that end most economically.

1. Suppose the names on the class roster of students are ordered according to the sex of the student and then within each group ordered by English grades. One scheme of selecting a sample of ten students to interview about some political opinion would be to draw from the roster the first ten men. Would this sample be representative of the population of GW students?

2. Suppose ten students are chosen by asking the first student to recommend a classmate who has a similar political opinion and then ask the second student to recommend another classmate and so on until ten students are selected. Would the responses of the survey be independent of each other?

3. Suppose American students responded to questions on politics based on American political tradition while foreign students responded based on their respective country's political traditions. Can one assume that the responses are identically distributed?

When designing a survey, these kinds of concerns must be addressed, in order to apply simple statistical formulas to the data otherwise a new statistical theory must be derived to accommodate the properties of the gathered survey data. In large government surveys like the United States Decennial Census, new statistical methods are derived as a result of new financial constraints or changing demographic characteristics of the country. The researchers, nonetheless, strive to design these surveys based on the concept of i.i.d. observations.

There are two fundamental methods of sampling:

1. Draw an element from the list for the sample and replace the drawn item to the list for possible selection again.

2. Draw an element from the list for the sample and do not replace the drawn item to the list so that it cannot be possibly selected again.

The first method is called sampling with replacement and the second method is call sampling without replacement. If the sampling is done at random, meaning that any sample drawn is equally likely to be selected, then the methods are called random sampling with replacement and random sampling without replacement. Bear in mind that not all survey samples are drawn at random. A popular method of drawing a sample is quota sampling performed in market research. In this method, an enumerator greets customers at a shopping mall entrance and asks them to participate in a survey. Of course, many customers will decline. The enumerator's job is done when he interviews a certain number of customers. How the customers are selected has some semblance of randomness, but quota samples are notorious for producing biased data. Yet, for the purposes of market research, they are deemed to be adequate and inexpensive.

Suppose a list has only four elements. Image that you are at a carnival arcade and this particular game makes you try to grab a prize using some mechanical apparatus. In the bin are four objects. Suppose that they are identical objects except that they are labeled, A, B, C, and D. Because they are physically identical, it would make sense that the probabilities of successfully grabbing the objects are the same. In other words, the process of selecting an object is based on equal initial probabilities of selection.

Suppose instead that the objects are different. Perhaps one is tiny and expensive and another is large and cheap. In this case, the probabilities of selection are not equal. We would say that the process of selecting an object is based on unequal initial probabilities of selection.

Survey samples are drawn without replacement so that we might randomly draw a sample with equal initial probabilities of selection without replacement or we might randomly draw a sample with unequal initial probabilities of selection without replacement.

Consider the first case where the sample is drawn from a list of four objects, A, B, C, and D, without replacement with equal initial probabilities of selection. Let A_1 be the event of drawing object A on the first draw, A_2 be the event of drawing object A on the second draw, and so on. By equal initial probabilities $P(A_1) = P(B_1) = P(C_1) = P(D_1) = \frac{1}{4}$. Suppose object B is selected in the first draw, that means that objects A, C, and D are left, because we do not replace the objects in this sampling scheme. On the second draw, $P(A_2) = \frac{1}{3}$ given that B was taken on the first draw or in other words $P(A_2|B_1) = \frac{1}{3}$. Suppose object C was selected on the second draw, that leaves objects A and D eligible to be selected on the third draw; therefore, $P(A_3|B_1, C_2) = \frac{1}{2}$ and finally if object D is selected on the third draw, there will only be object A left and the probability of selecting it on the fourth drawn will, obviously be 1 that is, $P(A_4|B_1, C_2, D_3) = 1$. We see that when selecting a sample without replacement, the probabilities change at each stage of the process. In the physical sciences where scientists study chemical reactions or metallurgical properties or measure the spectra of stars, they are dealing with essentially an infinite number of molecules or cosmic rays and sampling without replacement is an unimportant concern for them. In the social sciences, however, researchers deal with finite populations of people so that sampling without replacement poses a significant complication in doing a survey.

What is the probability that on the second draw, we draw object A? We want to find $P(A_2)$. To calculate the probability, we will use conditional probability.

$$\begin{aligned} P(A_2) &= P(A_2|A_1)P(A_1) + P(A_2|B_1)P(B_1) + P(A_2|C_1)P(C_1) + P(A_2|D_1)P(D_1) \\ &= 0 + \frac{P(A_1)}{1-P(B_1)}P(B_1) + \frac{P(A_1)}{1-P(C_1)}P(C_1) + \frac{P(A_1)}{1-P(D_1)}P(D_1) \end{aligned}$$

For example, suppose $P(A_1) = P(B_1) = P(C_1) = P(D_1) = \frac{1}{4}$, then

$$\begin{aligned} P(A_2) &= 0 + \frac{P(A_1)}{1-P(B_1)}P(B_1) + \frac{P(A_1)}{1-P(C_1)}P(C_1) + \frac{P(A_1)}{1-P(D_1)}P(D_1) \\ &= \left(\frac{\frac{1}{4}}{1-\frac{1}{4}}\right)\frac{1}{4} + \left(\frac{\frac{1}{4}}{1-\frac{1}{4}}\right)\frac{1}{4} + \left(\frac{\frac{1}{4}}{1-\frac{1}{4}}\right)\frac{1}{4} \\ &= \frac{1}{4} \end{aligned}$$

Suppose we draw a sample at random where the initial probabilities of selection are equal. If there are n objects, then $P(X_k = m) = \frac{1}{n}$ where $\{\omega \in \Omega | X(\omega)_k = m\}$ is the event that object k is selected on draw m. That is, $P(X_k = 1) = \frac{1}{n}$ implies $P(X_k = m) = \frac{1}{n}$ for all draws. In this nice case where the probabilities of selection regardless of the stage of drawing the object are the same $\frac{1}{n}$, we give the common name of simple random sampling. Simple random sampling is another name for sampling with equal initial probabilities without replacement.

8 More Details on Sampling

Table 1: Probability of Selecting an Element without Replacement with Equal Initial Probabilities

	Elements			
Draw	A	B	C	D
Initial	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
2	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
3	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
4	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Drawing a sample is the crucial operation of a survey. In that process, the size of the sample must first be determined according to the kind of sampling that will be performed, in order to achieve certain criteria. Generally, the size of a sample must be sufficiently large as to produce an estimate, the precision of which does not exceed a prescribed coefficient of variation (CV) while being as small as possible, in order to save money.

Simple random sampling is one of many ways of drawing a sample. Its popularity can be traced to that specific property whereby simple random sampling leads to the relatively simple derivation of a formula for an estimator. Typically, the estimator is for the population total, mean, variance or some ratio estimator. Simple random sampling is the common term that is used for the process of sampling from a finite population without replacement with *equal* initial probabilities. The outstanding feature of this way of sampling is that the probability of selecting an element is the same at each stage of the process. On the contrary, the probability of selecting an element does change with each draw when sampling is performed without replacement with *unequal* initial probabilities. The probabilities of selecting elements from a population of four objects with equal initial probabilities are shown in Table 1. Unlike Table 1 where all the entries are the same, the complicated fractions seen in Table 2 illustrate the consequences of sampling without replacement with unequal initial probabilities of selection.

To give an example which Table 2 might serve as an illustration, let the probability of selecting object C on the first draw be $\frac{1}{6}$ and let the probabilities of drawing the other objects on the first draw be as shown in Table 2. Suppose the identity of the first object that was drawn is not

Table 2: Probability of Selecting an Element without Replacement with Unequal Initial Probabilities

	Elements			
Draw	A	B	C	D
Initial	$\frac{1}{3}$	$\frac{2}{5}$	$\frac{1}{6}$	$\frac{1}{10}$
2	$\frac{44}{135}$	$\frac{73}{225}$	$\frac{23}{108}$	$\frac{41}{300}$
3	$\frac{4606}{19305}$	$\frac{8518}{42075}$	$\frac{1199}{3672}$	$\frac{1813}{7800}$
4	$\frac{1972}{19305}$	$\frac{3076}{42075}$	$\frac{1079}{3672}$	$\frac{4141}{7800}$

revealed and it is excluded from the sampling so that it cannot be selected again, then the probability of selecting object C on the second and only on the second draw is $\frac{23}{108}$. Suppose, once again, that the identities of the first two objects that were drawn are not revealed, and they are excluded from the sampling, then the probability of selecting object C on the third and only the third draw is $\frac{1199}{3672}$. With three out of four objects already drawn, one object remains. The probability that the remaining object is object C is $\frac{1079}{3672}$. Computing these numbers will be discussed later when there will be ample space to dwell on the abstruse nature of dealing with this kind of sampling. But, in the case where the initial probabilities of selection are equal, the probability of selecting any object at any stage of the process is $\frac{1}{4}$ as shown in Table 1. It is not difficult to reason that, in both cases, the sums of each row and each column equal 1.

One can imagine the insurmountable number of calculations that a sample of the size which is typically drawn for a survey would required, in order to find the probability of selecting a single element, if the initial probabilities were not equal.

9 Stratification

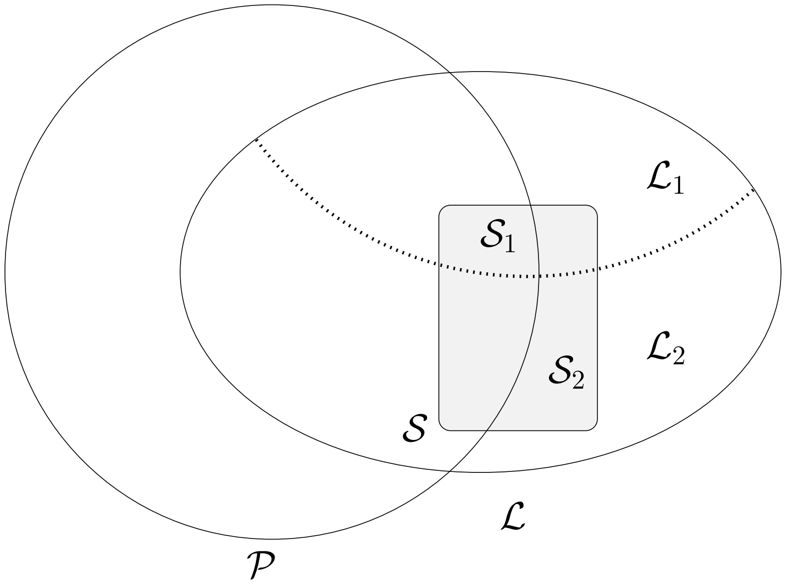

Figure 1

By referring to Figure 1, suppose a list, $\mathcal{L}$ is stratified into two strata, $\mathcal{L}_1$ and $\mathcal{L}_2$. Specifically, suppose $\mathcal{L}_1 = \{i \in \mathcal{L} | x_i \leqslant 30\}$ and $\mathcal{L}_2 = \{i \in \mathcal{L} | x_i > 30\}$ We see that the stratification depends on the number 30. If that number is changed, then the stratification will, also, change. If the boundary of $\mathcal{L}_1$ and $\mathcal{L}_2$ changes, then $\mathcal{S}_1$ and $\mathcal{S}_2$ will have to be different. Choosing the defining boundaries of strata is usually done by trial and error in order to produce the smallest variance for a given sampling size, for the main purpose of stratification is to reduce the variance of an estimate.

Typically, by means of simple random sampling, a sample, $\mathcal{S}_1$, is drawn from $\mathcal{L}_1$ and likewise a second sample, $\mathcal{S}_2$, is drawn from $\mathcal{L}_2$. We note that $\mathcal{L}_1$ and $\mathcal{L}_2$ are disjoint. The property of disjoint strata greatly simplifies the mathematics, because if A and B are disjoint events, then $P(A \cup B) = P(A) + P(B)$ which implies that $var(X + Y) = var(X) + var(Y)$.

Based on Theorem 3, $\widehat{\tau_i} = N_i \overline{y_i}$ is the population total of stratum i. The sum of them over all strata will produce an estimate of the population total, that is,

$$\widehat{\tau} = \sum_{i=1}^{h} N_i \overline{y_i}$$

The estimate of the population mean, μ, is

$$\widehat{\mu} = \frac{\widehat{\tau}}{N} = \frac{\sum_{i=1}^{h} N_i \overline{y_i}}{N}$$

which is the weighted average of the strata totals by strata sizes.

Theorem 9 on making estimates based on a stratified list immediately follows the preceding discussion.

Theorem 9 *Let y_{ij} be the measurement taken of element j in stratum i. Suppose there are h strata with corresponding size N_i and n_i for stratum i of the sample size and denote $\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$. If the strata are disjoint, the observations are i.i.d. and simple random sampling is used to draw the sample from each stratum, then*

$$\begin{aligned}
\widehat{\mu} = \bar{y} &= \frac{N_1\bar{y_1} + \ldots + N_h\bar{y_h}}{N} && (2)\\
\widehat{var(\bar{y})} &= \frac{1}{N^2}\sum_{i=1}^{h} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i} && (3)\\
\widehat{\tau} &= N_1\bar{y_1} + \ldots + N_r\bar{y_r} && (4)\\
\widehat{var(\widehat{\tau})} &= \sum_{i=1}^{h} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i} && (5)
\end{aligned}$$

9.1 Example Involving Stratified Sampling

$\mathcal{S}_1 = \{10,\ 15,\ 9,\ 13,\ 20,\ 16\}$
$\mathcal{S}_2 = \{105,\ 200,\ 150\}$
$\mathcal{S} = \{10,\ 15,\ 9,\ 13,\ 20,\ 16,\ 105,\ 200,\ 150\}$ where $\mathcal{S}$ would have been the sample if the list had not been stratified.

For this example, suppose the size of $\mathcal{L}_1$ is 970 and the size of $\mathcal{L}_2$ is 30, so that the size of the entire list,$\mathcal{L}$ is 1,000.

Table 3: Stratified Sample

	Sample of Stratum 1	Sample of Stratum 2	$\mathcal{S}$
List Size	$N_1 = 970$	$N_2 = 30$	$N = 1,000$
Sampling Size	$n_1 = 6$	$n_2 = 3$	$n = 9$
Sample Mean	$\bar{x}_1 = 13.833$	$\bar{x}_2 = 151.66$	$\bar{x} = 59.77$
Sample Variance	$s_1^2 = 4.07^2 = 16.5649$	$s_2^2 = 47.52^2 = 2258.15$	$s^2 = 72.9^2 = 5314.41$
CV			.4047

We see in Table 3 that the variance of the sample drawn from the unstratified list is larger than either sample variance gotten from the stratified samples, $\mathcal{S}_1$ and $\mathcal{S}_2$.

The estimate of the population mean is

$$\widehat{\mu} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N} \tag{6}$$

And because the strata are disjoint,

$$\begin{aligned}\widehat{var(\widehat{\mu})} &= \frac{1}{N^2}\left(N_1^2 var(\bar{x}_1) + N_2^2 var(\bar{x}_2)\right) \\ &= \frac{1}{N^2}\left(N_1^2\left(\frac{N_1-n_1}{N_1}\right)\frac{s_1^2}{n_1} + N_2^2\left(\frac{N_2-n_2}{N_2}\right)\frac{s_2^2}{n_2}\right)\end{aligned} \tag{7}$$

According to equation (6), $\widehat{\mu} = \frac{970(13.833)+30(151.66)}{1000} = 17.96$ and according to equation (7)

$$\begin{aligned}\widehat{var(\widehat{\mu})} &= \frac{1}{1000^2}\left(970^2\left(\frac{970-6}{970}\right)\frac{4.07^2}{6} + 30^2\left(\frac{30-3}{30}\right)\frac{47.52^2}{3}\right) \\ &= 3.19128\end{aligned}$$

So that $CV = \frac{\sqrt{3.19128}}{17.96} = .099$

Suppose the list had not been stratified, then $\bar{x} = 59.77$ and $s^2 = 5314.41$ and, by Theorem 7, $\widehat{\mu} = 59.77$ and $\widehat{\sigma^2} = \frac{N-n}{N}\frac{s^2}{n} = \frac{1000-9}{1000}\frac{5314.41}{9} = 585.1756$. The corresponding $CV = \frac{\sqrt{585.1756}}{59.77} = \frac{24.1904}{59.77} = .4047$ We see that that CV based on the stratified list, CV=.099, is four times smaller than the CV of $\widehat{\mu}$ based on the non-stratified list, CV=.4047.

10 Determining an Appropriate Sampling Size

Given a set of experimental data and the method from which the set of data was obtained, the sample mean is: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ and the sample variance is: $s^2 = \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n-1}$. They are common descriptive statistics of the data. If it is assumed that the measurements are distributed as a Normal distribution i.e., $x_i \sim N(\mu, \sigma^2)$, then the estimate of the population mean is: $\widehat{\mu} = \bar{x}$; the estimate of the population variance is: $\widehat{\sigma}^2 = s^2$ provided that the measurements are taken independently and they are not biased. Based on the data and the assumption that the measurements can be adequately described by a Normal distribution, then the $100(1-\alpha)\%$ confidence interval for μ is:

$$\left(\bar{x} - \frac{s}{\sqrt{n}}t_{n-1,\frac{\alpha}{2}}, \bar{x} + \frac{s}{\sqrt{n}}t_{n-1,\frac{\alpha}{2}}\right)$$

Sometimes, the distribution of x_i is unknown in which case the confidence interval is an approximate one as justified by the Central Limit Theorem. In order to avoid controversy over the ambiguity of the kind of distribution which is presumed to describe the measurements, another measure of precision is used and that is called the coefficient of variation (CV).

The length of a confidence interval is: $L = \frac{2s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}$. The relative length of the confidence interval to the estimated mean is:

$$\frac{L}{\bar{x}} = \frac{\frac{2s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}}{\bar{x}}$$

The essential part of the relative length is the quotient: $\frac{\frac{s}{\sqrt{n}}}{\bar{x}}$. It is called the *coefficient of variation* (CV) of $\bar{x}$. Equivalently, in terms of the length of a confidence interval,

$$CV = \frac{L}{2\bar{x} t_{n-1,\frac{\alpha}{2}}}$$

Its appeal stems from not having to specify a probability distribution which governs the measurements. The complexities, particularly found in surveys, in describing the nature of the measurements prohibit an easy derivation of a quantile like the t quantile which constitutes an essential component in constructing a confidence interval. In these situations, confidence intervals are not constructed rather the CV provides the measure of precision for an estimate. In general, $CV = \frac{\sqrt{var(est)}}{est}$ for some estimate, *est*.

10.1 Case I

Let $CV = \frac{s}{\sqrt{n}\bar{x}}$ and suppose $\bar{x}$ and s were found from a previous experiment, then when the desired CV is specified by the manager of the project, the size of the sample for the new experiment can be solved algebraically i.e.,

$$n = \left(\frac{s}{CV\bar{x}}\right)^2 = \left(\frac{2st_{n-1,\frac{\alpha}{2}}}{L}\right)^2 \tag{8}$$

As this discussion reveals, in order to find the right size of a sample, the functional relationship of the variance on n must be known. Therefore, when designing an experiment, its theory must have the formula for the estimated variance of the statistic which is being sought.

Example 3 *A previous study which examined the likelihood that a woman's enrollment in welfare was caused by abusive physical treatment from her common-law husband, a sociologist found that the proportion of such women who received a high school diploma is* $\bar{x} = .69$ *and s=.46. Find the size of a sample of these poor unmarried women who received only a high school diploma such that CV=2%.*

$$n = \left(\frac{.46}{(.02)(.69)}\right)^2 = 1111$$

According to the sociologist's previous study, the size of the sample which had been drawn for that preliminary study was only 242. We would advise the sociologist that, if he wants a more precise estimate, he must draw a much larger sample than 242 for his next survey. A calculation of the sampling size enables a researcher to reckon the cost of doing the larger subsequent study while at the same time being confident that the resulting estimate will achieve the prescribed precision which his research requires.

10.2 Case II

In the natural sciences, the size of the population is infinite, but in the social sciences and business, the size of the population is usually finite. In the case of drawing a sample from a population of finite size, the formula for the variance of $\bar{x}$ must take into account the size of the population, N, as can be seen in equation (9) for the estimated variance of $\bar{x}$.

$$\widehat{var(\bar{x})} = \frac{N-n}{N}\frac{s^2}{n} \tag{9}$$

We note in passing that as $N \to \infty$, then $\widehat{var(\bar{x})} = \frac{s^2}{n}$. The formula given by equation (8) applies to the case of an infinite population. For populations of finite size, the CV becomes $CV = \frac{\sqrt{(\frac{N-n}{N})\frac{s^2}{n}}}{\bar{x}}$. Solving for n produces the appropriate formula for finding the right sampling size for a finite population.

$$n = \frac{s^2}{CV^2\bar{x}^2 + \frac{s^2}{N}}$$

To check our answer with formula (8), we observe that when $N \to \infty$, $n = \left(\frac{s}{CV\bar{x}}\right)^2$ which was used earlier.

Example 4 *Assume that the size of the population of poor women suitable for this research study is 100,000, find the size of a sample which will produce a sample mean with a CV of 2%. From a previous research study, it was found that $\bar{x} = .69$ and that s=.46; therefore,*

$$n = \frac{.46^2}{.02^2(.69)^2 + \frac{.46^2}{100000}} = 1099$$

The sampling size is smaller in the case of a finite population than in the first example, because the factor, $\frac{N-n}{N}$, which appears in equation (9), is less than one. That factor drives the calculation of the sampling size to make an estimated variance of the sample mean conform to a finite population.

10.3 Case III

A behavioral experiment suffers from unco-operative subjects. To incorporate that aspect of a survey into the calculation of finding an appropriate sampling size of a sample, let ρ denote the rate of meeting co-operative subjects who will provide useful responses. Ideally, ρ should equal one. In other words, there should be 100% co-operation, but in practice ρ might equal 80 percent i.e., 20 percent of those who are approached for an interview refuse to co-operate or are unable to participate. ρ is called the response rate. Often, survey statisticians talk about the non-response rate which is equivalent to $1 - \rho$. The formula for the variance taking into account the response rate now becomes:

$$CV = \frac{\sqrt{\left(\frac{N-\rho n}{N}\right)\frac{s^2}{\rho n}}}{\bar{x}}$$

When that equation is solved for n:

$$n = \frac{\frac{s^2}{\rho}}{CV^2\bar{x}^2 + \frac{s^2}{N\rho}} \tag{10}$$

Example 5 *Use the same data as before in Example 4, but assume that the rate of response from the women is .60. Find the sampling size, n.*

$$n = \frac{.46^2/.60}{(.02)^2(.69)^2 + \frac{.46^2}{(100000)(.60)}} = 1818$$

This sampling size takes into account the information from a previous study, the size of the population, and the rate of useful responses. A comprehensive theory of determining a sampling size falls under the discipline of the design of experiments. Suppose it costs $10,000,000, for example, to make one observation as in measuring the accuracy of a launched submarine ballistic missile. Or consider that some surveys which are sponsored by the National Institute of Health cost $200,000,000 and last 30 years. Finding the right sampling size which will produce sufficient and informative data over the span of 30 years requires the opinion of many experts in addition to complicated formulas, if any can be derived.

Determining an appropriate sampling size is not easy. If knowledge is obtained from a previous experiment or extensive professional experience is available, then certain characteristics of the population may be estimated from which a sampling size can be computed. Lacking expert knowledge and information which might be available from previous studies, an experimenter will probably be inclined to do some preliminary experiments in order to gain some knowledge of the population after which a more general, complete, and expensive experiment will follow.

Although the use of the coefficient of variation avoids controversy which could arise when attempts are made to ascribe a probability distribution to an observation it might, nevertheless, be desirable to determine that size of a sample which will produce a confidence interval of a certain prescribed length. In that case, we will substitute for CV in equation (10) the formula $CV = \frac{L}{2\bar{x}t_{n-1,\frac{\alpha}{2}}}$. This formula was obtained by combining $CV = \frac{\frac{s}{\sqrt{n}}}{\bar{x}}$ and the length of the confidence interval, $L = \frac{2s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}$. After performing the substitution for CV, equation (10) can be written as shown in equation (11) as a function of the length of the confidence interval and the quantile for the presumed distribution which describes measurement, x_i.

$$n = \frac{\frac{s^2}{\rho}}{\frac{L^2}{4t^2_{n-1,\frac{\alpha}{2}}} + \frac{s^2}{N\rho}} \tag{11}$$

Suppose that the information which is given in Example 7 was the product of a cursory review of 100 grade point averages such that $\bar{x} = 3.5$ and $s = .5$.

Example 6 *From 1,000 patient records at a hospital, a sample of size 20 is drawn at random.*

Random meaning that any possible combination of 20 records, i.e. $\binom{1000}{20} = 3.394828 \times 10^{41}$ *is equally likely to be drawn.*

Use a table of random numbers or divide the set of 1,000 records into 50 groups pick a random starting point, #23 in the first group. Pick every 50^{th} *record starting from that one, so #23, #73, #123, ... This is an example of systematic sampling.*

Looking at some account billing, based on the sample, $\bar{x} = \$94.22$ *and* $s^2 = 445.21$. *Estimate* μ*, the population mean.*

$$\begin{aligned}
n &= 20 \\
\widehat{\mu} &= \$94.22 \\
\widehat{var(\widehat{\mu})} &= \left(\frac{N-n}{N}\right)\frac{s^2}{n} \\
&= \left(\frac{1000-20}{1000}\right)\frac{445.21}{20} = 21.815 \\
\sqrt{\widehat{var(\widehat{\mu})}} &= \sqrt{21.815} = 4.670 \\
CV &= \frac{4.670}{94.22} = 4.9\%
\end{aligned}$$

Question 1 What is the total hospital bill?

Answer 1

$$\widehat{\tau} = N\bar{x} = 1000(94.22) = \$94,220$$

$$\begin{aligned}\widehat{var(\widehat{\tau})} &= N^2\frac{N-n}{N}\frac{s^2}{n} \\ &= 1000^2\left(\frac{1000-20}{1000}\right)\frac{445.21}{20} = 21,815,290 \\ \sqrt{\widehat{var(\widehat{\tau})}} &= \sqrt{21,815,290} = 4670.68 \\ CV &= \frac{4670.68}{94220} = 4.9\%\end{aligned}$$

Question 2 Find the sampling size, n, to achieve the prescribed CV=2%. Assume $\rho = 1$.

Answer 2

$$n = \frac{\frac{s^2}{\rho}}{CV^2\bar{x}^2 + \frac{s^2}{N\rho}} = \frac{\frac{445.21}{1}}{.02^2(94.22^2) + \frac{445.21}{1000}} = 111.4$$

A sampling size of 112 should be large enough to obtain a CV of 2%.

Example 7 *The grade point averages (GPA) of a sample of 100 students were obtained. Denote the GPA of a student by X_i. From the data, it was found that $\bar{x} = 3.5$ and $s = .5$. Find the 90% confidence interval about the population mean.*

The problem requires the computation of the 90% confidence interval. Upon doing the calculation, the 90% confidence interval is given by (3.417, 3.583). The length of this confidence interval is: L=3.583-3.417=.166 and it corresponds to a CV of $CV = \frac{L}{2\bar{x}t_{n-1,\frac{\alpha}{2}}} = (3.583 - 3.417)/(2 * 3.5 * 1.66) = .142$. Because the researcher does not have access to the registrar's records, he plans instead to interview students, for the purpose of learning their grade point averages. The researcher knows that 13,000 students are enrolled in the university but from prior experience he knows that about 30% of the students whom he will interview will refuse to co-operate. He deemed the length of the confidence interval which was obtained from his preliminary study to be too long; consequently, to suit the requirements of his research, the researcher wants to interview enough students to produce a 90% confidence interval such that the length of the resulting confidence interval will be L=.1. The rate of refusal must be converted to the rate of response: $\rho = 1 - .3 = .7$. All the necessary components of equation (11) are now known except for $t_{n-1,\frac{\alpha}{2}}$. Since we do not know n, we cannot find a value for $t_{n-1,\frac{\alpha}{2}}$. To circumvent this difficulty, we instead assume the worst case by taking $n = \infty$, that is, we will use $t_{\infty,.05} = 1.644$. In the example under consideration,

$$n = \frac{\frac{.5^2}{.7}}{\frac{.1^2}{4(1.644)^2} + \frac{.5^2}{(13{,}000).7}} = 373.1983 \approx 374 \tag{12}$$

Upon reflecting on equation (11) and on the preceding calculation of n, it becomes obvious that n is actually a random variable, because $\bar{x}$, s^2, and ρ are all estimates based on previous studies. The implication is that n is not a deterministic quantity. It is an educated guess. For this reason, among others, the designing of an informative experiment at the least cost is a great challenge. To give an example of the computational challenge which statisticians face in designing a typical operational government stratified multivariate survey, the formula for finding an optimal sampling size is:

$$\begin{gathered} min \sum_{h \in strata} c_h n_h \\ \ni \\ \sum_{h \in strata} N_h(N_h - \widehat{\rho}_h n_h) \frac{s^2_{hk}}{\widehat{\rho}_h n_h} \leqslant (CV_k \, \widehat{\tau}_k)^2 \qquad \forall k \\ 0 \leqslant n_h \leqslant N_h \end{gathered}$$

It illustrates the complexity of just one dimension of designing a large experiment. In essence, we have come full circle in the sense that all of the assigned problems presented in this course have been predicated on having good data already given to us; however, sets of data can only come from well designed experiments which, in turn, depend on previously conducted studies, and so on. In practice, little inexpensive experiments lead to more extensive experiments which lead to pilot studies which lead ultimately to full scale production. To accomplish that progression of complexity most efficiently and accurately, the science of statistics provides the indispensable methods which we have only briefly described.

Appendix A: Estimated Variance of the Population Proportional

We formulate an estimator of the population total by defining the Bernoulli random variable, X_i, such that

$$X_i = \begin{cases} 1 & \text{if the subject is measured} \\ 0 & \text{otherwise} \end{cases}$$

Because X_i is either 0 or 1, X_i^2 is, also, 0 or 1; therefore, $\sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i$. This is a useful identity when manipulating the algebra of s^2.

By definition of s^2,

$$s^2 = \frac{\sum\limits_{i \in \mathcal{S}} (X_i - \bar{X})^2}{n-1}$$

By simple algebra, $s^2 = \frac{\sum\limits_{i \in \mathcal{S}} X_i^2 - n\bar{X}^2}{n-1}$. Because $X_i^2 = X_i$ as noted above, $s^2 = \frac{\sum\limits_{i \in \mathcal{S}} X_i - n\bar{X}^2}{n-1}$. Then s^2 in the case of Bernoulli random variables becomes $s^2 = \frac{n\bar{X} - n\bar{X}^2}{n-1}$. Since $\widehat{p} = \bar{X}$, then $s^2 = \frac{n\hat{p} - n\hat{p}^2}{n-1} = \frac{n\widehat{p}\widehat{q}}{n-1}$.

We take expectations of the last expression to get: $\frac{n}{n-1} E[\widehat{p}\widehat{q}] = E[s^2] = \left(\frac{N}{N-1}\right) \sigma^2$. We solve for σ^2 to get $\sigma^2 = \left(\frac{N-1}{N}\right) \left(\frac{n}{n-1}\right) E[\widehat{p}\widehat{q}]$. Or we may assert that $\widehat{\sigma^2} = \left(\frac{N-1}{N}\right) \left(\frac{n}{n-1}\right) \widehat{p}\widehat{q}$.

Based on our work with $\bar{X}$, we know that $\widehat{var(\bar{x})} = \left(\frac{N-n}{N-1}\right) \frac{\widehat{\sigma^2}}{n}$. See page 8. But $\widehat{p} = \bar{X}$; therefore,

$$\widehat{var(\widehat{p})} = \left(\frac{N-n}{N-1}\right) \frac{\widehat{\sigma^2}}{n} \tag{13}$$

From above, we derived that $\widehat{\sigma^2} = \left(\frac{N-1}{N}\right) \left(\frac{n}{n-1}\right) \widehat{p}\widehat{q}$ which we substitute into equation (13) to get our final result.

$$\widehat{var(\widehat{p})} = \left(\frac{N-n}{N}\right) \frac{\widehat{p}\widehat{q}}{n-1} \tag{14}$$

Appendix B: Probability of Selecting an Element

The formula for computing the probability that object, k, taken from a population of size, N, will be selected on the m^{th} draw is quite complicated. The formulas for each successive stage of the sampling for a simple example will reveal the patterns in the formula in the general case. To that end, let p_{km} be the probability of drawing object k from a population of four objects on exactly the m^{th} draw with initial probabilities p_1, p_2, p_3, and p_4 like the ones illustrated in Table 2. The probability of drawing object C, for example, on the first draw is: $p_{31} = p_3 = \frac{1}{6}$; on the second draw,

$$p_{32} = \frac{p_3}{1-p_1}p_1 + \frac{p_3}{1-p_2}p_2 + \frac{p_3}{1-p_4}p_4 = \frac{23}{108}$$

on the third draw,

$$\begin{aligned} p_{33} &= \frac{p_3}{1-p_1-p_2}\left(p_1\frac{p_2}{1-p_1} + p_2\frac{p_1}{1-p_2}\right) + \frac{p_3}{1-p_1-p_4}\left(p_1\frac{p_4}{1-p_1} + p_4\frac{p_1}{1-p_4}\right) \\ &\quad + \frac{p_3}{1-p_2-p_4}\left(p_2\frac{p_4}{1-p_2} + p_4\frac{p_2}{1-p_4}\right) = \frac{1199}{3672} \end{aligned}$$

on the fourth draw,

$$\begin{aligned} p_{34} &= \frac{p_3}{1-p_1-p_2-p_4}\left(p_1\frac{p_2}{1-p_1}\frac{p_4}{1-p_1-p_2} + p_1\frac{p_4}{1-p_1}\frac{p_2}{1-p_1-p_4}\right. \\ &\quad + p_2\frac{p_1}{1-p_2}\frac{p_4}{1-p_1-p_2} + p_2\frac{p_4}{1-p_2}\frac{p_1}{1-p_2-p_4} + p_4\frac{p_1}{1-p_4}\frac{p_2}{1-p_1-p_4} \\ &\quad \left. + p_4\frac{p_2}{1-p_4}\frac{p_1}{1-p_1-p_4}\right) = \frac{1079}{3672} \end{aligned}$$

Suppose that if instead of four objects, there are N objects in the sampling frame, then the formulas become the following:

$$p_{k1} = \overbrace{\underbrace{p_k}_{\binom{N-1}{0}\ terms}}^{0!\ term}$$

$$p_{k2} = \underbrace{\frac{p_k}{1-p_1}\overbrace{p_1}^{1!\ term} + \cdots + \frac{p_k}{1-p_N}p_N}_{\binom{N-1}{1}\ terms}$$

$$p_{k3} = \underbrace{\frac{p_k}{1-p_1-p_2}\overbrace{\left(p_1\frac{p_2}{1-p_1} + p_2\frac{p_1}{1-p_2}\right)}^{2!\ terms} + \cdots + \frac{p_k}{1-p_i-p_j}\left(p_i\frac{p_j}{1-p_i} + p_j\frac{p_i}{1-p_j}\right) + \cdots}_{\binom{N-1}{2}\ terms}$$

$$p_{k4} = \underbrace{\frac{p_k}{1-p_1-p_2-p_3}\overbrace{\left(p_1\frac{p_2}{1-p_1}\frac{p_3}{1-p_1-p_2}+\cdots+p_3\frac{p_2}{1-p_3}\frac{p_1}{1-p_3-p_2}\right)}^{3!\ terms}+\cdots}_{\binom{N-1}{3}\ terms}$$

$$p_{km} = \underbrace{\frac{p_k}{1-p_1-p_2-\cdots-p_{m-1}}\overbrace{\left(p_1\frac{p_2}{1-p_1}\frac{p_3}{1-p_1-p_2}\cdots\frac{p_{m-1}}{1-p_1-p_2-\cdots-p_{m-2}}+\cdots\right.}^{(m-1)!\ terms}+\cdots}_{\binom{N-1}{m-1}\ terms}$$

To write the last equation more compactly, some special notation must be invented. Let S_n denote the symmetric group on n letters. Each element of S_n represents a unique permutation of n letters. There are $n!$ such permutations represented by S_n. For example, there are six permutations that comprise the group, $S_3 = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6\}$. They are defined by: $\sigma_1(123) = 123$; $\sigma_2(123) = 213$; $\sigma_3(123) = 231$; $\sigma_4(123) = 321$; $\sigma_5(123) = 132$; $\sigma_6(123) = 312$.

There are $\binom{n}{k}$ ways to choose k objects from a set of size n. Let $C^n_m(k)$ be the collection of the subsets of size m that can be drawn from $\{1, 2, \ldots, n\}$ minus element, k. Suppose, for example, that there are 5 objects, $\{1, 2, 3, 4, 5\}$, then $C^5_3(2) = \{134, 135, 145, 345\}$. The cardinality of $C^N_m(k)$ is $|C^N_m(k)| = \binom{N-1}{m}$. Let the $(N-1) \times 1$ vector, **i**, be an element of $C^N_m(k)$, then the general expression for p_{km} can be written as:

$$p_{km} = p_k \sum_{\mathbf{i}\in C^N_{m-1}(k)} \frac{1}{1-p_{i_1}-p_{i_2}-\cdots-p_{i_{m-1}}}\left(\sum_{\sigma\in S_{m-1}} p_{\sigma(i_1)}\frac{p_{\sigma(i_2)}}{1-p_{\sigma(i_1)}}\cdots\frac{p_{\sigma(i_{m-1})}}{1-p_{\sigma(i_1)}-\cdots-p_{\sigma(i_{m-1})}}\right) \quad (15)$$

By having p_{km} written in a more compact notation, it will be much easier to prove Theorem 10.

Theorem 10 *If the initial probabilities are equal when drawing a sample randomly without replacement from a finite population of size N, then the probability of drawing element, k, on the* m^{th} *draw is:* $p_{km} = \frac{1}{N} \quad \forall m \leqslant N$

Proof: By hypothesis, the initial probabilities are equal, that is: $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$. In other words, regardless of the permutation that would be done on the indices, $p_{\sigma(i)} = \frac{1}{N}$. Substituting these values into equation (15) gives:

$$\begin{aligned}
p_{km} &= \frac{1}{N} \sum_{\mathbf{i} \in C^N_{m-1}(k)} \left(\frac{1}{1 - \frac{m-1}{N}} \right) \left(\sum_{\sigma \in S_{m-1}} \frac{1}{N} \frac{\frac{1}{N}}{1 - \frac{1}{N}} \frac{\frac{1}{N}}{1 - \frac{1}{N} - \frac{1}{N}} \cdots \frac{\frac{1}{N}}{1 - \frac{m-2}{N}} \right) \\
&= |C^N_{m-1}(k)| \left(\frac{\frac{1}{N}}{1 - \frac{m-1}{N}} \right) |S_{m-1}| \frac{1}{N} \frac{1}{N-1} \frac{1}{N-2} \cdots \frac{1}{N-m+2} \\
&= \binom{N-1}{m-1} \frac{1}{N-m+1} (m-1)! \frac{(N-m+1)!}{N!} \\
&= \frac{(N-1)!}{(m-1)!(N-m)!} (m-1)! \frac{(N-m)!}{N!} \\
&= \frac{1}{N}
\end{aligned}$$

■