Session 8

SAMPLING THEORY

STATISTICS

STATISTICS ANALYTIC Sampling Theory

A **probability sampling** method is any method of sampling that utilizes some form of *random selection*. In order to have a random selection method, you must set up some process or procedure that assures that the different units in your population have equal probabilities of being chosen. Humans have long practiced various forms of random selection, such as picking a name out of a hat, or choosing the short straw. These days, we tend to use computers as the mechanism for generating random numbers as the basis for random selection.

An Introduction to Sampling Theory

The applet that comes with this WWW page is an interactive demonstration that will show the basics of sampling theory. Please read ahead to understand more about what this program does. For more information on the use of this applet see the bottom of this page.

A Quick Primer on Sampling Theory

The signals we use in the real world, such as our voices, are called "analog" signals. To process these signals in computers, we need to convert the signals to "digital" form. While an analog signal is continuous in both time and amplitude, a digital signal is discrete in both time and amplitude. To convert a signal from continuous time to discrete time, a process called sampling is used. The value of the signal is measured at certain intervals in time. Each measurement is referred to as a sample. (The analog signal is also quantized in amplitude, but that process is ignored in this demonstration. See the Analog to Digital Conversion page for more on that.)

When the continuous analog signal is sampled at a frequency F, the resulting discrete signal has more frequency components than did the analog signal. To be precise, the frequency components of the analog signal are repeated at the sample rate. That is, in the discrete frequency response they are seen at their original position, and are also seen centered around +/- F, and around +/- 2F, etc.

How many samples are necessary to ensure we are preserving the information contained in the signal? If the signal contains high frequency components, we will need to sample at a higher rate to avoid losing information that is in the signal. In general, to preserve the full information in the signal, it is necessary to sample at twice the maximum frequency of the signal. This is known as the Nyquist rate. The Sampling Theorem states that a signal can be exactly reproduced if it is sampled at a frequency F, where F is greater than twice the maximum frequency in the signal.

What happens if we sample the signal at a frequency that is lower that the Nyquist rate? When the signal is converted back into a continuous time signal, it will exhibit a phenomenon called ***aliasing***. Aliasing is the presence of unwanted components in the reconstructed signal. These components were not present when the original signal was sampled. In addition, some of the frequencies in the original signal may be lost in the reconstructed signal. Aliasing occurs because signal frequencies can overlap if the sampling frequency is too low. Frequencies "fold" around half the sampling frequency - which is why this frequency is often referred to as the folding frequency.

Sometimes the highest frequency components of a signal are simply noise, or do not contain useful information. To prevent aliasing of these frequencies, we can filter out these components before sampling the signal. Because we are filtering out high frequency components and letting lower frequency components through, this is known as low-pass filtering.

Demonstration of Sampling

The original signal in the applet below is composed of three sinusoid functions, each with a different frequency and amplitude. The example here has the frequencies 28 Hz, 84 Hz, and 140 Hz. Use the filtering control to filter out the higher frequency components. This filter is an ideal low-pass filter, meaning that it exactly preserves any frequencies below the cutoff frequency and completely attenuates any frequencies above the cutoff frequency.

Notice that if you leave all the components in the original signal and select a low sampling frequency, aliasing will occur. This aliasing will result in the reconstructed

signal not matching the original signal. However, you can try to limit the amount of aliasing by filtering out the higher frequencies in the signal. Also important to note is that once you are sampling at a rate above the Nyquist rate, further increases in the sampling frequency do not improve the quality of the reconstructed signal. This is true because of the ideal low-pass filter. In real-world applications, sampling at higher frequencies results in better reconstructed signals. However, higher sampling frequencies require faster converters and more storage. Therefore, engineers must weigh the advantages and disadvantages in each application, and be aware of the tradeoffs involved.

The importance of frequency domain plots in signal analysis cannot be understated. The three plots on the right side of the demonstration are all Fourier transform plots. It is easy to see the effects of changing the sampling frequency by looking at these transform plots. As the sampling frequency decreases, the signal separation also decreases. When the sampling frequency drops below the Nyquist rate, the frequencies will crossover and cause aliasing.

Experiment with the following applet in order to understand the effects of sampling and filtering.

Hypothesis testing

The basic idea of statistics is simple: you want to extrapolate from the data you have collected to make general conclusions. **Population** can be e.g. all the voters and **sample** the voters you polled. Population is characterized by parameters and sample is characterized by statistics. For each parameter we can find appropriate statistics. This is called **estimation**. Parameters are always fixed, statistics vary from sample to sample.

Statistical hypothesis is a statement about population. In case of parametric tests it is a statement about population parameter. The only way to decide whether this statement is 100% truth or false is to research whole population. Such a research is ineffective and sometimes impossible to perform. This is the reason why we research only the sample instead of the population. Process of the verification of the hypothesis based on samples is called hypothesis testing. The objective of testing is to decide whether observed difference in sample is only due to chance or statistically significant.

Steps in Hypothesis testing:

1) **Defining a null hypothesis**
The null hypothesis is usually an hypothesis of "no difference"

2) **Defining alternative hypothesis**
Alternative hypothesis is usually hypothesis of significant (not due to chance) difference

3) **Choosing alpha (significance level)**
Conventionally the 5% (less than 1 in 20 chance of being wrong) level has been used.

4) **Do the appropriate statistical test to compute the P value.**
A P value is the largest value of alpha that would result in the rejection of the null hypothesis for a particular set of data.

5) **Decision**
Compare calculated P-value with prechosen alpha.
If P value is less than the chosen significance level then you reject the null hypothesis i.e. accept that your sample gives reasonable evidence to support the alternative hypothesis.
If the P value is greater than the threshold, state that you "do not reject the null hypothesis" and that the difference is "not statistically significant". You cannot conclude that the null hypothesis is true. All you can do is conclude that you don't have sufficient evidence to reject the null hypothesis.

Possible outcomes in hypothesis testing:

	Decision	
Truth	H_0 not rejected	H_0 rejected
H_0 is true	Correct decision ($p = 1-\alpha$)	Type I error ($p = \alpha$)
H_0 is false	Type II error ($p = \beta$)	Correct decision ($p = 1-\beta$)

H_0: Null hypothesis
p: Probability
α: Significance level
1-*α*: Confidence level
1-β: Power

Inferring parameters for models of biological processes are a current challenge in systems biology, as is the related problem of comparing competing models that explain the data. In this work we apply Skilling's nested sampling to address both of these problems. Nested sampling is a Bayesian method for exploring parameter space that transforms a multi-dimensional integral to a 1D integration over likelihood space. This approach focusses on the computation of the marginal likelihood or evidence. The ratio of evidences of different models leads to the Bayes factor, which can be used for model comparison. We demonstrate how nested sampling can be used to reverse-engineer a system's behaviour whilst accounting for the uncertainty in the results. The effect of missing initial conditions of the variables as well as unknown parameters is investigated. We show how the evidence and the model ranking can change as a function of the available data. Furthermore, the addition of data from extra variables of the system can deliver more information for model comparison than increasing the data from one variable, thus providing a basis for experimental design

Some Definitions

Before I can explain the various probability methods we have to define some basic terms. These are:

- **N** = the number of cases in the sampling frame
- **n** = the number of cases in the sample
- ${}_{N}\mathbf{C}_{\mathbf{n}}$ = the number of combinations (subsets) of n from N
- **f** = n/N = the sampling fraction

That's it. With those terms defined we can begin to define the different probability sampling methods.

Simple Random Sampling

The simplest form of random sampling is called **simple random sampling**. Pretty tricky, huh? Here's the quick description of simple random sampling:

- **Objective**: To select *n* units out of *N* such that each ${}_{N}C_{n}$ has an equal chance of being selected.
- **Procedure**: Use a table of random numbers, a computer random number generator, or a mechanical device to select the sample.

A somewhat stilted, if accurate, definition. Let's see if we can make it a little more real. How do we select a simple random sample? Let's assume that we are doing some research with a small service agency that wishes to assess client's views of quality of service over the past year. First, we have to get the sampling frame organized. To accomplish this, we'll go through agency records to identify every client over the past 12 months. If we're lucky, the agency has good accurate computerized records and can quickly produce such a list. Then, we have to actually draw the sample. Decide on the number of clients you would like to have in the final sample. For the sake of the example, let's say you want to select 100 clients to survey and that there were 1000 clients over the past 12 months. Then, the sampling fraction is f = n/N = 100/1000 = .10 or 10%. Now, to actually draw the sample, you have several options. You could print off the list of 1000 clients, tear then into separate strips, put the strips in a hat, mix them up real good, close your eyes and pull out the first 100. But this mechanical procedure would be tedious and the quality of the sample would depend on how thoroughly you mixed them up and how randomly you reached in. Perhaps a better procedure would be to use the kind of ball machine that is popular with many of the state lotteries. You would need three sets of balls numbered 0 to 9, one set for each of the digits from 000 to 999 (if we select 000 we'll call that 1000). Number the list

of names from 1 to 1000 and then use the ball machine to select the three digits that selects each person. The obvious disadvantage here is that you need to get the ball machines. (Where do they make those things, anyway? Is there a ball machine industry?).

Neither of these mechanical procedures is very feasible and, with the development of inexpensive computers there is a much easier way. Here's a simple procedure that's especially useful if you have the names of the clients already on the computer. Many computer programs can generate a series of random numbers. Let's assume you can copy and paste the list of client names into a column in an EXCEL spreadsheet. Then, in the column right next to it paste the function =RAND() which is EXCEL's way of putting a random number between 0 and 1 in the cells. Then, sort both columns -- the list of names and the random number -- by the random numbers. This rearranges the list in random order from the lowest to the highest random number. Then, all you have to do is take the first hundred names in this sorted list. pretty simple. You could probably accomplish the whole thing in under a minute.

Simple random sampling is simple to accomplish and is easy to explain to others. Because simple random sampling is a fair way to select a sample, it is reasonable to generalize the results from the sample back to the population. Simple random sampling is not the most statistically efficient method of sampling and you may, just because of the luck of the draw, not get good representation of subgroups in a population. To deal with these issues, we have to turn to other sampling methods.

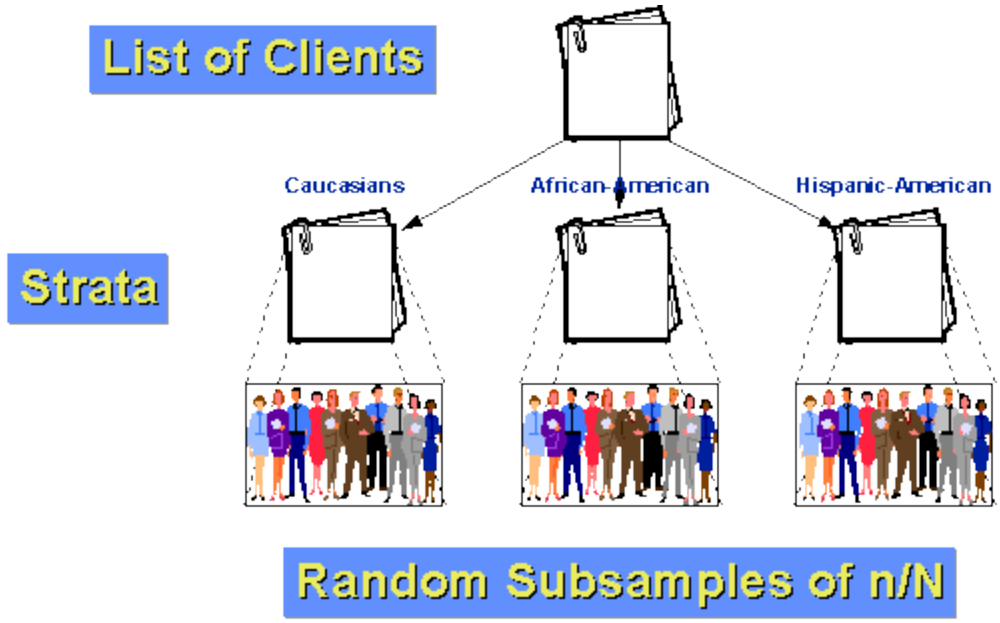
Stratified Random Sampling

Stratified Random Sampling, also sometimes called *proportional* or *quota* random sampling, involves dividing your population into homogeneous subgroups and then taking a simple random sample in each subgroup. In more formal terms:

Objective: Divide the population into non-overlapping groups (i.e., *strata*) N_1, N_2, N_3, ... N_i, such that $N_1 + N_2 + N_3 + ... + N_i = N$. Then do a simple random sample of f = n/N in each strata.

There are several major reasons why you might prefer stratified sampling over simple random sampling. First, it assures that you will be able to represent not only the overall population, but also key subgroups of the population, especially small minority groups. If you want to be able to talk about subgroups, this may be the only way to effectively assure you'll be able to. If the subgroup is extremely small, you can use different sampling fractions (f) within the different strata to randomly over-sample the small group (although you'll then have to weight the within-group estimates using the sampling fraction whenever you want overall population estimates). When we use the same sampling fraction within strata we are conducting *proportionate* stratified random sampling. When we use different sampling fractions in the strata, we call this *disproportionate* stratified random sampling. Second, stratified random sampling will generally have more statistical precision than simple random sampling. This will only be true if the strata or groups are homogeneous. If they are, we expect that the variability within-groups is lower than the variability for the population as a whole. Stratified sampling capitalizes on that fact.

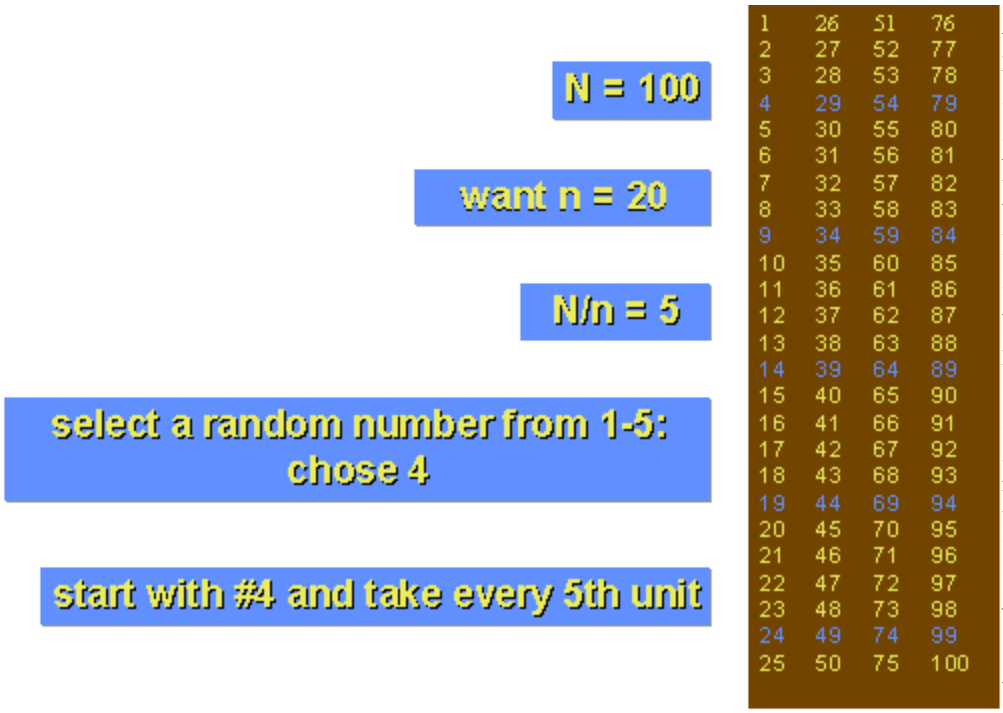
For example, let's say that the population of clients for our agency can be divided into three groups: Caucasian, African-American and Hispanic-American. Furthermore, let's assume that both the African-Americans and Hispanic-Americans are relatively small minorities of the clientele (10% and 5% respectively). If we just did a simple random sample of n=100 with a sampling fraction of 10%, we would expect by chance alone that we would only get 10 and 5 persons from each of our two smaller groups. And, by chance, we could get fewer than that! If we stratify, we can do better. First, let's determine how many people we want to have in each group. Let's say we still want to take a sample of 100 from the population of 1000 clients over the past year. But we think that in order to say anything about subgroups we will need at least 25 cases in each group. So, let's sample 50 Caucasians, 25 African-Americans, and 25 Hispanic-Americans. We know that 10% of the population, or 100 clients, are African-American. If we randomly sample 25 of these, we have a within-stratum sampling fraction of 25/100 = 25%. Similarly, we know that 5% or 50 clients are Hispanic-American. So our within-stratum sampling fraction will be 25/50 = 50%. Finally, by subtraction we know that there are 850 Caucasian clients. Our within-stratum sampling fraction for them is 50/850 = about 5.88%. Because the groups are more homogeneous within-group than across the population as a whole, we can expect greater statistical precision (less variance). And, because we stratified, we know we will have enough cases from each group to make meaningful subgroup inferences.



Systematic Random Sampling

Here are the steps you need to follow in order to achieve a **systematic random sample**:

- number the units in the population from 1 to N
- decide on the n (sample size) that you want or need
- k = N/n = the interval size
- randomly select an integer between 1 to k
- then take every kth unit



All of this will be much clearer with an example. Let's assume that we have a population that only has N=100 people in it and that you want to take a sample of n=20. To use systematic sampling, the population must be listed in a random order. The sampling fraction would be f = 20/100 = 20%. in this case, the interval size, k, is equal to N/n = 100/20 = 5. Now, select a random integer from 1 to 5. In our example, imagine that you chose 4. Now, to select the sample, start with the 4th unit in the list and take every k-th unit (every 5th, because k=5). You would be sampling units 4, 9, 14, 19, and so on to 100 and you would wind up with 20 units in your sample.

For this to work, it is essential that the units in the population are randomly ordered, at least with respect to the characteristics you are measuring. Why would you ever want to use systematic random sampling? For one thing, it is fairly easy to do. You only have to select a single random number to start things off. It may also be more precise than simple random sampling. Finally, in some situations there is simply no easier way to do random sampling. For instance, I once had to do a study that involved sampling from all the books in a library. Once selected, I would have to go to the shelf, locate the book, and record when it last circulated. I knew that I had a fairly good sampling frame in the form of the shelf list (which is a card catalog where the entries are arranged in the order they occur on the shelf). To do a simple random sample, I could have estimated the total number of books and generated random numbers to draw the sample; but how would I find book #74,329 easily if that is the number I selected? I couldn't very well count the cards until I came to 74,329! Stratifying wouldn't solve that problem either. For instance, I could have stratified by card catalog drawer and drawn a simple random sample within each drawer. But I'd still be stuck counting cards. Instead, I did a systematic random sample. I estimated the number of books in the entire collection. Let's imagine it was 100,000. I decided that I wanted to take a sample of 1000 for a sampling fraction of 1000/100,000 = 1%. To get the sampling interval k, I divided N/n = 100,000/1000 = 100. Then I selected a random integer between 1 and 100. Let's say I got 57. Next I did a little side study to determine how thick a thousand cards are in the card catalog (taking into account the varying ages of the cards). Let's say that on average I found that two cards that were separated by 100 cards were about .75 inches apart in the catalog drawer. That information gave me everything I needed to draw the sample. I counted to the 57th

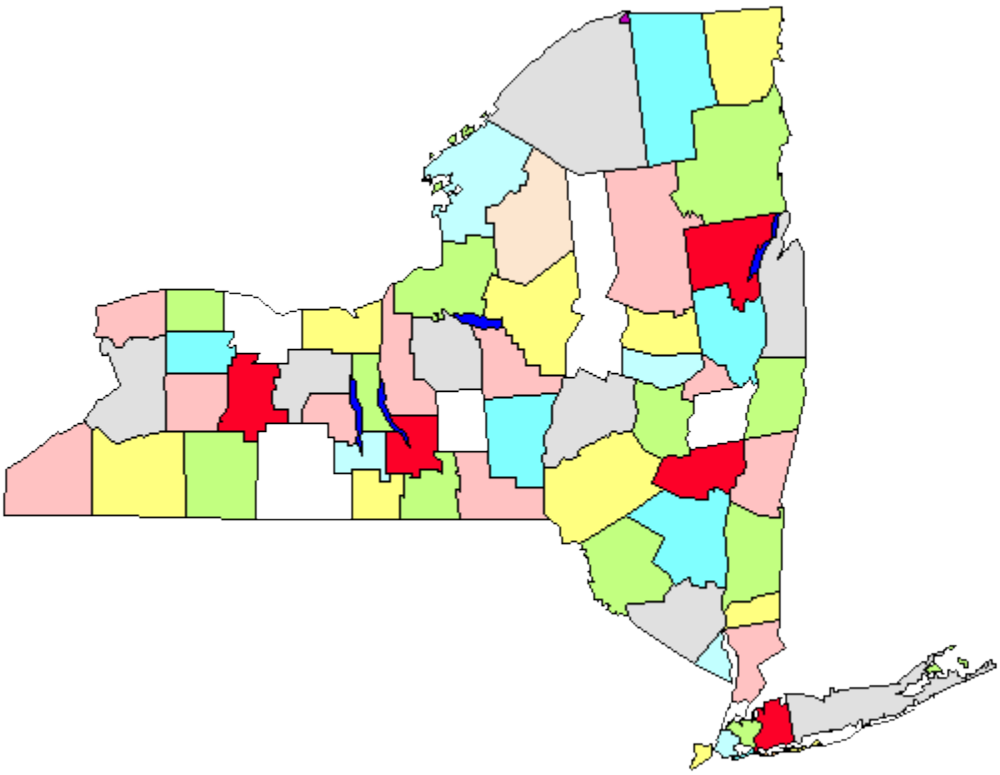
by hand and recorded the book information. Then, I took a compass. (Remember those from your high-school math class? They're the funny little metal instruments with a sharp pin on one end and a pencil on the other that you used to draw circles in geometry class.) Then I set the compass at .75", stuck the pin end in at the 57th card and pointed with the pencil end to the next card (approximately 100 books away). In this way, I approximated selecting the 157th, 257th, 357th, and so on. I was able to accomplish the entire selection procedure in very little time using this systematic random sampling approach. I'd probably still be there counting cards if I'd tried another random sampling method. (Okay, so I have no life. I got compensated nicely, I don't mind saying, for coming up with this scheme.)

Cluster (Area) Random Sampling

The problem with random sampling methods when we have to sample a population that's disbursed across a wide geographic region is that you will have to cover a lot of ground geographically in order to get to each of the units you sampled. Imagine taking a simple random sample of all the residents of New York State in order to conduct personal interviews. By the luck of the draw you will wind up with respondents who come from all over the state. Your interviewers are going to have a lot of traveling to do. It is for precisely this problem that **cluster or area random sampling** was invented.

In cluster sampling, we follow these steps:

- divide population into clusters (usually along geographic boundaries)
- randomly sample clusters
- measure all units within sampled clusters



For instance, in the figure we see a map of the counties in New York State. Let's say that we have to do a survey of town governments that will require us going to the towns personally. If we do a simple random sample state-wide we'll have to cover the entire state

geographically. Instead, we decide to do a cluster sampling of five counties (marked in red in the figure). Once these are selected, we go to *every* town government in the five areas. Clearly this strategy will help us to economize on our mileage. Cluster or area sampling, then, is useful in situations like this, and is done primarily for efficiency of administration. Note also, that we probably don't have to worry about using this approach if we are conducting a mail or telephone survey because it doesn't matter as much (or cost more or raise inefficiency) where we call or send letters to.

Multi-Stage Sampling

The four methods we've covered so far -- simple, stratified, systematic and cluster -- are the simplest random sampling strategies. In most real applied social research, we would use sampling methods that are considerably more complex than these simple variations. The most important principle here is that we can combine the simple methods described earlier in a variety of useful ways that help us address our sampling needs in the most efficient and effective manner possible. When we combine sampling methods, we call this **multi-stage sampling**.

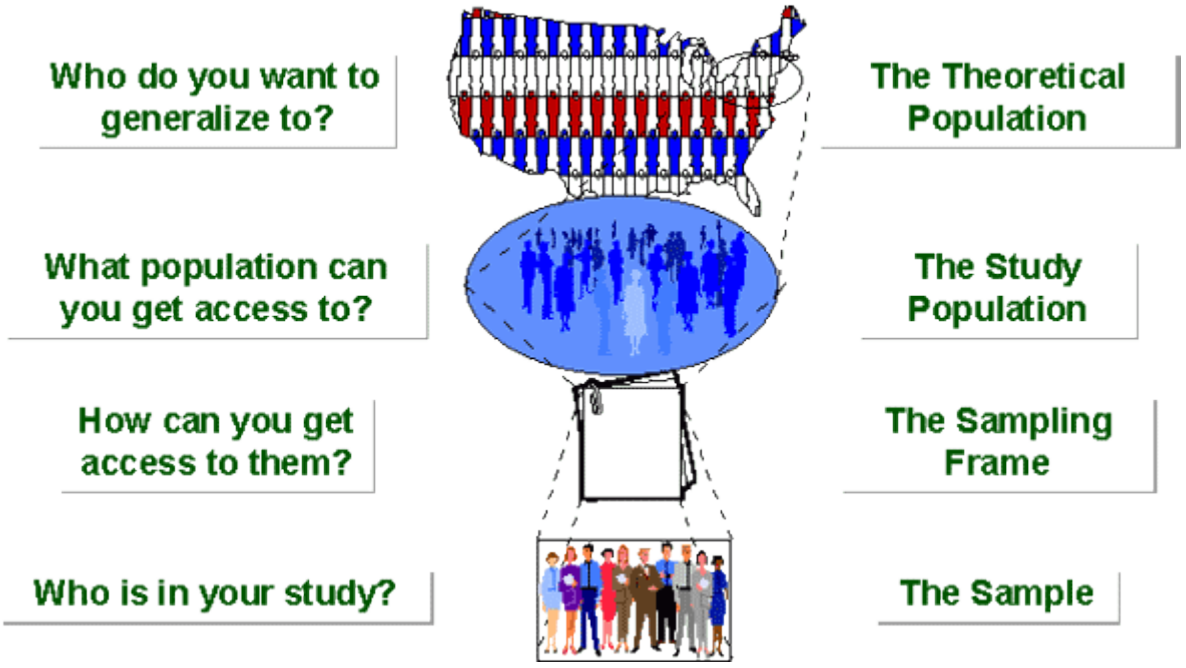
For example, consider the idea of sampling New York State residents for face-to-face interviews. Clearly we would want to do some type of cluster sampling as the first stage of the process. We might sample townships or census tracts throughout the state. But in cluster sampling we would then go on to measure everyone in the clusters we select. Even if we are sampling census tracts we may not be able to measure *everyone* who is in the census tract. So, we might set up a stratified sampling process within the clusters. In this case, we would have a two-stage sampling process with stratified samples within cluster samples. Or, consider the problem of sampling students in grade schools. We might begin with a national sample of school districts stratified by economics and educational level. Within selected districts, we might do a simple random sample of schools. Within schools, we might do a simple random sample of classes or grades. And, within classes, we might even do a simple random sample of students. In this case, we have three or four stages in the sampling process and we use both stratified and simple random sampling. By combining different sampling methods we are able to achieve a rich variety of probabilistic sampling methods that can be used in a wide range of social research contexts

Sampling Terminology

As with anything else in life you have to learn the language of an area if you're going to ever hope to use it. Here, I want to introduce several different terms for the major groups that are involved in a sampling process and the role that each group plays in the logic of sampling.

The major question that motivates sampling in the first place is: "Who do you want to generalize to?" Or should it be: "To whom do you want to generalize?" In most social research we are interested in more than just the people who directly participate in our study. We would like to be able to talk in general terms and not be confined only to the people who are in our study. Now, there are times when we aren't very concerned about generalizing. Maybe we're just evaluating a program in a local agency and we don't care whether the program would work with other people in other places and at other times. In that case, sampling and generalizing might not be of

interest. In other cases, we would really like to be able to generalize almost universally. When psychologists do research, they are often interested in developing theories that would hold for all humans. But in most applied social research, we are interested in generalizing to specific groups. The group you wish to generalize to is often called the **population** in your study. This is the group you would like to sample from because this is the group you are interested in generalizing to. Let's imagine that you wish to generalize to urban homeless males between the ages of 30 and 50 in the United States. If that is the population of interest, you are likely to have a very hard time developing a reasonable sampling plan. You are probably not going to find an accurate listing of this population, and even if you did, you would almost certainly not be able to mount a national sample across hundreds of urban areas. So we probably should make a distinction between the population you would like to generalize to, and the population that will be accessible to you. We'll call the former the **theoretical population** and the latter the **accessible population**. In this example, the accessible population might be homeless males between the ages of 30 and 50 in six selected urban areas across the U.S.



Once you've identified the theoretical and accessible populations, you have to do one more thing before you can actually draw a sample -- you have to get a list of the members of the accessible population. (Or, you have to spell out in detail how you will contact them to assure representativeness). The listing of the accessible population from which you'll draw your sample is called the **sampling frame**. If you were doing a phone survey and selecting names from the telephone book, the book would be your sampling frame. That wouldn't be a great way to sample because significant subportions of the population either don't have a phone or have moved in or out of the area since the last book was printed. Notice that in this case, you might identify the area code and all three-digit prefixes within that area code and draw a sample simply by randomly dialing numbers (cleverly known as *random-digit-dialing*). In this case, the sampling frame is not a list *per se*, but is rather a procedure that you follow as the actual basis for sampling. Finally, you actually draw your sample (using one of the many sampling procedures).

The **sample** is the group of people who you select to be in your study. Notice that I didn't say that the sample was the group of people who are actually *in* your study. You may not be able to contact or recruit all of the people you actually sample, or some could drop out over the course of the study. The group that actually completes your study is a subsample of the sample -- it doesn't include nonrespondents or dropouts. The problem of nonresponse and its effects on a study will be addressed when discussing "mortality" threats to internal validity.

People often confuse what is meant by random selection with the idea of random assignment. You should make sure that you understand the distinction between random selection and random assignment.

At this point, you should appreciate that sampling is a difficult multi-step process and that there are lots of places you can go wrong. In fact, as we move from each step to the next in identifying a sample, there is the possibility of introducing systematic error or **bias**. For instance, even if you are able to identify perfectly the population of interest, you may not have access to all of them. And even if you do, you may not have a complete and accurate enumeration or sampling frame from which to select. And, even if you do, you may not draw the sample correctly or accurately. And, even if you do, they may not all come and they may not all stay. Depressed yet? This is a very difficult business indeed. At times like this I'm reminded of what Donald Campbell used to say (I'll paraphrase here): "Cousins to the amoeba, it's amazing that we know anything at all!"

The main idea of statistical inference is to take a random sample from a population and then to use the information from the sample to make inferences about particular population characteristics such as the mean (measure of central tendency), the standard deviation (measure of spread) or the proportion of units in the population that have a certain characteristic. Sampling saves money, time, and effort. Additionally, a sample can, in some cases, provide as much information as a corresponding study that would attempt to investigate an entire population-careful collection of data from a sample will often provide better information than a less careful study that tries to look at everything.

We must study the behavior of the mean of sample values from different specified populations. Because a sample examines only part of a population, the sample mean will not exactly equal the corresponding mean of the population. Thus, an important consideration for those planning and interpreting sampling results, is the degree to which sample estimates, such as the sample mean, will agree with the corresponding population characteristic.

In practice, only one sample is usually taken (in some cases such as "survey data analysis" a small "pilot sample" is used to test the data-gathering mechanisms and to get preliminary information for planning the main sampling scheme). However, for purposes of understanding the degree to which sample means will agree with the corresponding population mean, it is useful to consider what would happen if 10, or 50, or 100 separate sampling studies, of the same type, were conducted. How consistent would the results be across these different studies? If we could see that the results from each of the samples would be nearly the same (and nearly correct!), then we would have confidence in the single sample that will actually be used. On the other hand, seeing that answers from the repeated samples were too variable for the needed

accuracy would suggest that a different sampling plan (perhaps with a larger sample size) should be used.

A sampling distribution is used to describe the distribution of outcomes that one would observe from replication of a particular sampling plan.

Know that estimates computed from one sample will be different from estimates that would be computed from another sample.

Understand that estimates are expected to differ from the population characteristics (parameters) that we are trying to estimate, but that the properties of sampling distributions allow us to quantify, probabilistically, how they will differ.

Understand that different statistics have different sampling distributions with distribution shapes depending on (a) the specific statistic, (b) the sample size, and (c) the parent distribution.

Understand the relationship between sample size and the distribution of sample estimates.

Understand that the variability in a sampling distribution can be reduced by increasing the sample size.

See that in large samples, many sampling distributions can be approximated with a normal distribution.

Variance and Standard Deviation

Deviations about the mean of a population is the basis for most of the statistical tests we will learn. Since we are measuring how widely a set of scores is dispersed about the mean we are measuring variability. We can calculate the deviations about the mean, and express it as variance or standard deviation. It is very important to have a firm grasp of this concept because it will be a central concept throughout the course.

Both variance and standard deviation measures variability within a distribution. Standard deviation is a number that indicates how much, on average, each of the values in the distribution deviates from the mean (or center) of the distribution. Keep in mind that variance measures the same thing as standard deviation (dispersion of scores in a distribution). Variance, however, is the average squared deviations about the mean. Thus, variance is the square of the standard deviation.

In terms of quality of goods/services, It is important to know that higher variation means lower quality. Measuring the size of variation and its source is the statistician's job, while fixing it is the job of the engineer or the manager. Quality products and services have low variation.

General Sampling Techniques

From the food you eat to the TV you watch, from political elections to school board actions, much of your life is regulated by the results of sample surveys. In the information age of today

and tomorrow, it is increasingly important that sample survey design and analysis be understood by many so as to produce good data for decision making and to recognize questionable data when it arises. Relevant topics are: Simple Random Sampling, Stratified Random Sampling, Cluster Sampling, Systematic Sampling, Ratio and Regression Estimation, Estimating a Population Size, Sampling a Continuum of Time, Area or Volume, Questionnaire Design, Errors in Surveys.

A sample is a group of units selected from a larger group (the population). By studying the sample it is hoped to draw valid conclusions about the larger group.

A sample is generally selected for study because the population is too large to study in its entirety. The sample should be representative of the general population. This is often best achieved by random sampling. Also, before collecting the sample, it is important that the researcher carefully and completely defines the population, including a description of the members to be included.

Random Sampling: Random sampling of size n from a population size N. Unbiased estimate for variance of $\bar{x}$is $Var(\bar{x}) = S^2(1-n/N)/n$, where n/N is the sampling fraction. For sampling fraction less than 10% the finite population correction factor (N-n)/(N-1) is almost 1.

The total T is estimated by N. $\bar{x}$, its variance is $N^2Var(\bar{x})$.

For 0, 1, (binary) type variables, variation in estimated proportion p is:

$$S^2 = p.(1-p).(1-n/N)/(n-1).$$

For ratio $r = \Sigma x_i/\Sigma y_i = \bar{x} / \bar{y}$, the variation for r is

$$[(N-n)(r^2S^2_x + S^2_y - 2\, r\, Cov(x, y)]/[n(N-1).\bar{x}^2].$$

Stratified Sampling: Stratified sampling can be used whenever the population can be partitioned into smaller sub-populations, each of, which is homogeneous according to the particular characteristic of interest.

$\bar{x}_s = \Sigma\ W_t.\ Bxar_t$, over t=1, 2, ..L (strata), and $\bar{x}_t$ is $\Sigma X_{it}/n_t$.

Its variance is:

$$\Sigma W^2_t\ /(N_t-n_t)S^2_t/[n_t(N_t-1)]$$

Population total T is estimated by N. $\bar{x}_s$, its variance is

$$\Sigma N^2_t(N_t-n_t)S^2_t/[n_t(N_t-1)].$$

Since the survey usually measures several attributes for each population member, it is impossible to find an allocation that is simultaneously optimal for each of those variables. Therefore, in such a case we use the popular method of allocation which use the same sampling fraction in each stratum. This yield optimal allocation given the variation of the strata are all the same.

Determination of sample sizes (n) with regard to binary data: Smallest integer greater than or equal to:

$$[t^2 N p(1-p)] / [t^2 p(1-p) + \alpha^2 (N-1)]$$

with N being the size of the total number of cases, n being the sample size, α the expected error, t being the value taken from the t distribution corresponding to a certain confidence interval, and p being the probability of an event.

Cross-Sectional Sampling: Cross-Sectional Study the observation of a defined population at a single point in time or time interval. Exposure and outcome are determined simultaneously.

Quota Sampling: Quota sampling is availability sampling, but with the constraint that proportionality by strata be preserved. Thus the interviewer will be told to interview so many white male smokers, so many black female nonsmokers, and so on, to improve the representatives of the sample. Maximum variation sampling is a variant of quota sampling, in which the researcher purposively and non-randomly tries to select a set of cases, which exhibit maximal differences on variables of interest. Further variations include extreme or deviant case sampling or typical case sampling.

What is a statistical instrument? A statistical instrument is any process that aim at describing a phenomena by using any instrument or device, however the results may be used as a control tool. Examples of statistical instruments are questionnaire and surveys sampling.

What is grab sampling technique? The grab sampling technique is to take a relatively small sample over a very short period of time, the result obtained are usually instantaneous. However, the **Passive Sampling** is a technique where a sampling device is used for an extended time under similar conditions. Depending on the desirable statistical investigation, the Passive Sampling may be a useful alternative or even more appropriate than grab sampling. However, a passive sampling technique needs to be developed and tested in the field.

In probability theory and statistics, a **discrete probability distribution** is a probability distribution characterized by a probability mass function. Thus, the distribution of a random variable *X* is discrete, and *X* is then called a **discrete random variable**, if

as *u* runs through the set of all possible values of *X*. It follows that such a random variable can assume only a finite or countably infinite number of values. That is, the possible values might

be listed, although the list might be infinite. For example, count observations such as the numbers of birds in flocks comprise only natural number values {0, 1, 2, ...}. By contrast, continuous observations such as the weights of birds comprise real number values and would typically be modeled by a continuous probability distribution such as the normal.

In cases more frequently considered, this set of possible values is a topologically discrete set in the sense that all its points are isolated points. But there are discrete random variables for which this countable set is dense on the real line (for example, a distribution over rational numbers).

Among the most well-known discrete probability distributions that are used for statistical modeling are the Poisson distribution, the Bernoulli distribution, the binomial distribution, the geometric distribution, and the negative binomial distribution. In addition, the discrete uniform distribution is commonly used in computer programs that make equal-probability random selections between a number of choices.

Alternative description

Equivalently to the above, a discrete random variable can be defined as a random variable whose cumulative distribution function (cdf) increases only by jump discontinuities—that is, its cdf increases only where it "jumps" to a higher value, and is constant between those jumps. The points where jumps occur are precisely the values which the random variable may take. The number of such jumps may be finite or countably infinite. The set of locations of such jumps need not be topologically discrete; for example, the cdf might jump at each rational number.

Consequently, a discrete probability distribution is often represented as a generalized probability density function involving Dirac delta functions, which substantially unifies the treatment of continuous and discrete distributions. This is especially useful when dealing with probability distributions involving both a continuous and a discrete part.

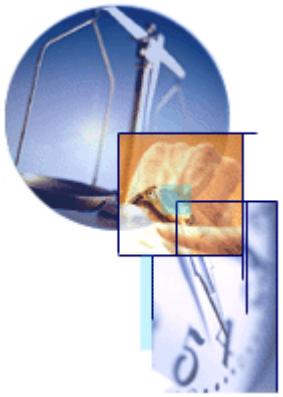
Representation in terms of indicator functions

For a discrete random variable X, let u_0, u_1, ... be the values it can take with non-zero probability. Denote

These are disjoint sets, and by formula (1)

It follows that the probability that X takes any value except for u_0, u_1, ... is zero, and thus one can write X as

except on a set of probability zero, where 1_A is the indicator function of A. This may serve as an alternative definition of discrete random variables.



A variable is continuous if the range of possible values for that variable falls along a continuum. You probably recall from the Discrete Probability Distributions section of this course that discrete random variables are measured in whole units, such as the number of people attending a ball game, the number of cookies in a package, or the number of cars assembled during one production shift. Continuous random variables are measured along a continuum, such as the loudness of cheering at a ball game, the weight of cookies in a package, or the time required to assemble a car.

A continuous probability distribution illustrates the complete range of values a continuous random variable can take on, as well as the probabilities associated with that range of values. A continuous probability distribution is important in predicting the likelihood of an event within a certain range of values.

As an example, consider temperature. Temperature is a continuous random variable because its possible values fall along a continuum. Automobile manufacturers want to be sure that the cold temperatures of northern climates do not cause fractures in vehicle parts. In this example, the temperature at which such fractures occur is a continuous random variable. For these manufacturers, determining the exact temperature at which vehicle parts develop fractures is impossible, as there are an infinite number of fractions of degree between, for example, –30 degrees and –35 degrees.

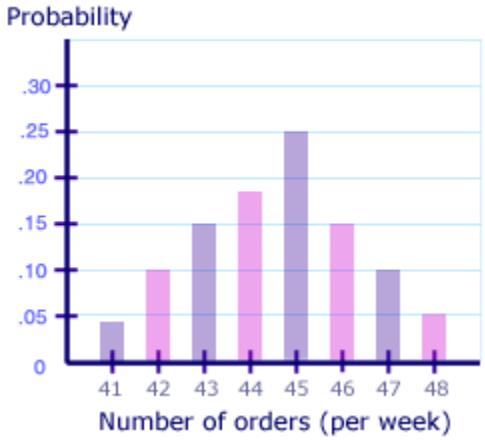
Although the specific temperature fracture point is impossible to discover, automobile manufacturers can determine the probability that fractures will occur within the range of temperatures experienced during a typical northern winter. The manufacturers would use a continuous probability distribution to determine the probability of fractures at or below a given temperature.

When you have completed the Continuous Probability Distributions section, you should be able to

- interpret a continuous probability distribution
- identify a normal distribution and explain the significance of a normal distribution's mean and standard deviation
- calculate a random variable's *Z* score and determine probabilities based on that *Z* score
- use a *Z* distribution table

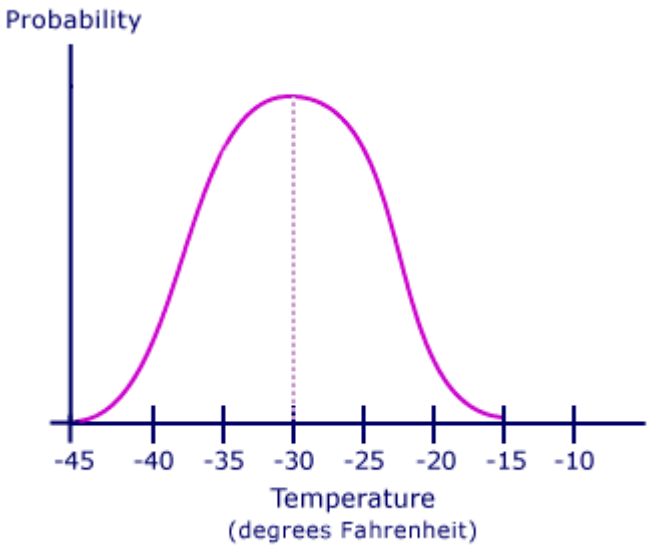
Continuous Probability Distributions: Continuous Random Variables

You probably recall from the Discrete Probability Distributions section of this course that a *discrete* random variable assumes values that are separate and distinct. The probability distribution for a discrete random variable, such as the number of orders taken in one week, may look like the following graph.

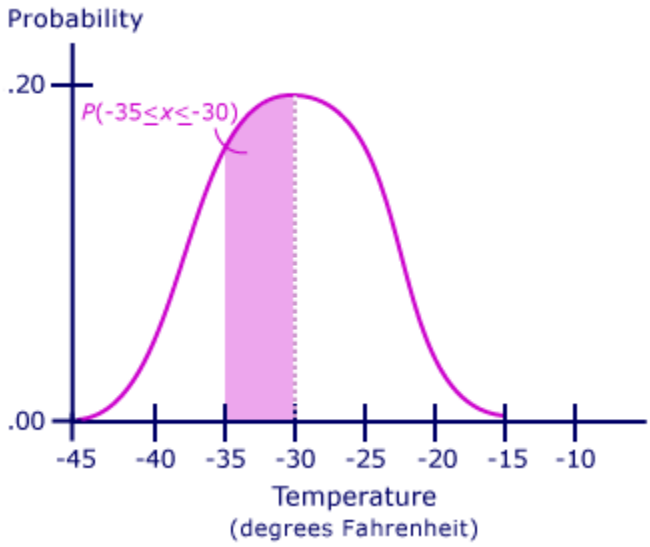


Alternatively, a *continuous* random variable can assume a range of values that falls along a continuum. The probability distribution for a continuous random variable can be represented by a curve that spans the range of values that the variable can assume. The curve is continuous and will not have distinct segments as in the discrete distribution above.

Consider the continuous probability distribution below, which illustrates the range of temperatures at which auto parts fractures occur. The distribution shows that fractures occur between approximately –45 and –15 degrees Fahrenheit.



A continuous probability distribution can be used to determine the probability of a variable falling between any two chosen values within the range of the distribution. The automobile manufacturer uses the graph below to analyze the probability of the temperature being between –35 and –30 degrees Fahrenheit when parts fractures occur, which is indicated by the shaded area under the curve.



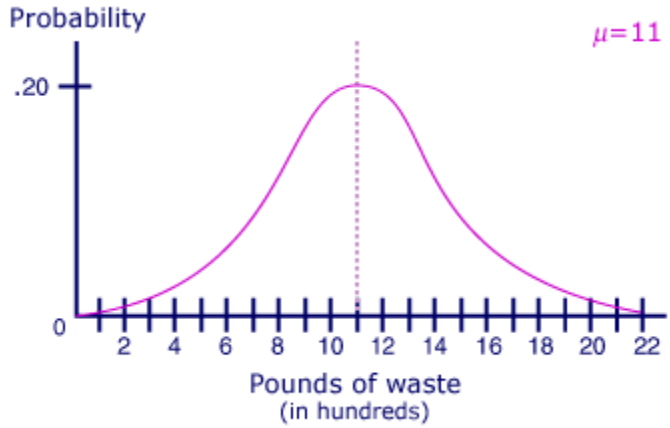
Integral calculus is used to find this area and calculate the probability, but that is beyond the scope of this course.

Consider a second business example: Each week, a manufacturer of plastic products generates a varying amount of waste during production. In order to

maintain production schedules and avoid problems of underutilized plant capacity, the manufacturer must have a sufficient amount of raw material at its facility to compensate for the waste generated during production.

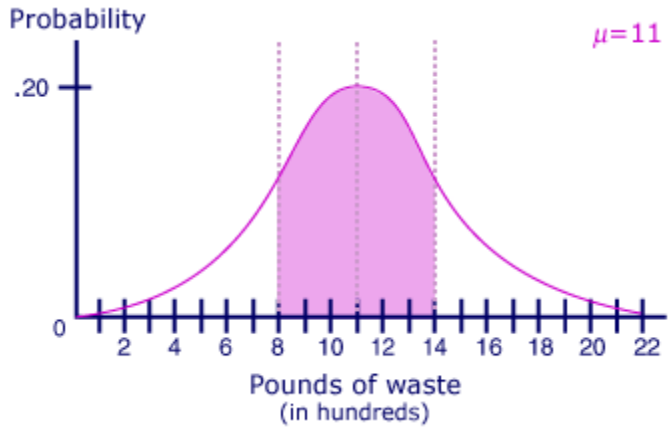
The company's operations executive needs to know the amount of waste generated each week. He will use this information to predict the probability associated with various levels of waste. This will allow him to determine the correct level of raw materials to have on hand for a production run.

The operations executive uses past data to construct a probability density function of the amount of waste generated each week. This distribution is illustrated below.



Using this probability distribution, the executive can see that the amount of waste ranges from 0 to 2,200 pounds, with the most probable amount of waste being approximately 1,100 pounds.

If the executive wanted to determine the probability that the level of waste will be between 800 and 1,400 pounds, he could calculate the area under the curve between 800 and 1,400. This area under the curve that he wishes to calculate is indicated by the shaded portion in the graph below.



Notice that the two continuous probability distributions you've seen here have similar shapes. Both are approximations of a useful distribution called the normal distribution, which will be examined in detail in the Normal Distribution portion of this course.

1. Classify each of the following random variables as discrete or continuous.

a. the number of people waiting in line at a checkout counter

b. the size of an intake valve on a 1965 Ford Mustang

c. the color of a macaroni-and-cheese dinner

d. the time it takes a customer to choose a product

Solution 1

2. How is it possible to measure the probability of temperatures, weights, times, or levels of satisfaction if there are an infinite number of values that the variable can take on?

Solution 2

3. Is it possible for a continuous random variable to have a binomial

distribution? Why or why not?

Inferential Statistics

With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

Here, I concentrate on inferential statistics that are useful in experimental and quasi-experimental research design or in program outcome evaluation. Perhaps one of the simplest inferential test is used when you want to compare the average performance of two groups on a single measure to see if there is a difference. You might want to know whether eighth-grade boys and girls differ in math test scores or whether a program group differs on the outcome measure from a control group. Whenever you wish to compare the average performance between two groups you should consider the t-test for differences between groups.

Most of the major inferential statistics come from a general family of statistical models known as the General Linear Model. This includes the t-test, Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA), regression analysis, and many of the multivariate methods like factor analysis, multidimensional scaling, cluster analysis, discriminant function analysis, and so on. Given the importance of the General Linear Model, it's a good idea for any serious social researcher to become familiar with its workings. The discussion of the General Linear Model here is very elementary and only considers the simplest straight-line model. However, it will get you familiar with the idea of the linear model and help prepare you for the more complex analyses described below.

One of the keys to understanding how groups are compared is embodied in the notion of the "dummy" variable. The name doesn't suggest that we are using variables that aren't very smart or, even worse, that the analyst who uses them is a "dummy"! Perhaps these variables would be better described as "proxy" variables. Essentially a dummy variable is one that uses discrete numbers, usually 0 and 1, to represent different groups in your study. Dummy variables are a simple idea that enable some pretty complicated things to happen. For instance, by including a simple dummy variable in an model, I can model two separate lines (one for each treatment group) with a single

equation. To see how this works, check out the discussion on dummy variables.

One of the most important analyses in program outcome evaluations involves comparing the program and non-program group on the outcome variable or variables. How we do this depends on the research design we use. research designs are divided into two major types of designs: experimental and quasi-experimental. Because the analyses differ for each, they are presented separately.

Experimental Analysis. The simple two-group posttest-only randomized experiment is usually analyzed with the simple t-test or one-way ANOVA. The factorial experimental designs are usually analyzed with the Analysis of Variance (ANOVA) Model. Randomized Block Designs use a special form of ANOVA blocking model that uses dummy-coded variables to represent the blocks. The Analysis of Covariance Experimental Design uses, not surprisingly, the Analysis of Covariance statistical model.

Quasi-Experimental Analysis. The quasi-experimental designs differ from the experimental ones in that they don't use random assignment to assign units (e.g., people) to program groups. The lack of random assignment in these designs tends to complicate their analysis considerably. For example, to analyze the Nonequivalent Groups Design (NEGD) we have to adjust the pretest scores for measurement error in what is often called a Reliability-Corrected Analysis of Covariance model. In the Regression-Discontinuity Design, we need to be especially concerned about curvilinearity and model misspecification. Consequently, we tend to use a conservative analysis approach that is based on polynomial regression that starts by overfitting the likely true function and then reducing the model based on the results. The Regression Point Displacement Design has only a single treated unit. Nevertheless, the analysis of the RPD design is based directly on the traditional ANCOVA model.

When you've investigated these various analytic models, you'll see that they all come from the same family -- the General Linear Model. An understanding of that model will go a long way to introducing you to the intricacies of data analysis in applied and social research contexts.