Part 3. Additional Concepts: Building on the Basics

I. Terms and Concepts

A. Focus

This section builds on the statistical terms and concepts presented in Part 2, Section III, providing some additional concepts and adding to some of the basic concepts presented earlier. Sometimes, the same examples given in Part 2 are further developed by adding details or alternative approaches. None of these terms or concepts require any specific educational background; they are simply part of the language and practice of statistics.

The information in this section will be of particular interest to planning group members who fit either or both of the following categories:

- Those who enjoy working with statistics, and would like to know more about terms and concepts which are particularly relevant to HIV Prevention Community Planning; and
- Those who expect to play a particularly active role in the review of statistical information brought before the planning group or one of its committees, and would like some additional tools to use in this process.

The section provides information on concepts related to:

- Sampling Issues; and
- Data Collection and Analysis, including:
  - Questionnaire Construction and Coding of Data;
  - Percentages;
  - Decimal Points and Rounding of Data;
  - Averages; and
  - Probability and Significance.

B. Sampling and Sample Sizes

Part 2 presented some basic sampling information. Some additional information can be helpful in reviewing statistics needed for needs assessment and priority setting — especially the concepts of stratified random sampling and oversampling (See Figure 1).

Using random sampling can allow you to assume that the sample is representative of the total population — but does not automatically enable you to separately analyze and compare findings for different segments of the population, such as dropouts versus in-school youth, males versus females, or persons from various racial and ethnic groups. To do this, you will generally need to do stratified random sampling to separately sample major population segments, and to oversample groups that are a small part of the total population. Unless you oversample, you will be able to make inferences for the total population, but not for the various subpopulations of interest to you — because the number of persons sampled from some subpopulations may be too small. Often, having separate information about a particular population segment will be very important to your planning process, for reasons such as the following:
That population segment is overrepresented among current AIDS cases;

That population segment represents a significant and perhaps growing proportion of the planning area population;

Current prevention activities do not seem to be reaching that population adequately; and/or

You don’t know enough about the population to know what kinds of prevention activities are most likely to be effective.

A stratified random sample which deliberately oversamples such a population segment will help address those needs.

Remember that the sampling procedure is only as good as the universe or total population from which the sample is drawn. This is important in designing your own studies and in using the results of existing research. For example, if your planning group wanted to know about the sexual attitudes and behaviors of young adults, the most obvious way to obtain such information would be to do a survey in places where large numbers of young adults can be found — such as high schools and colleges. However, if the sampling universe were high school seniors and college freshmen, it would not be representative of the entire young adult population, because it would exclude high school dropouts and other young adults who are in the workforce rather than in school. Because minorities and low income people are less likely to complete high school and go to college, this would probably mean that minorities and low-income people would be underrepresented. The sampling universe would be more complete if samples were drawn

Figure 1
Sampling Terms and Concepts
You already know that a sample is a group selected from a total population or universe with the expectation that studying the group will provide important information about the total population — which means that you would like to be able to generalize from the sample to the total population. to assume that the conclusions you reach about the sample are also true of the total population from which the sample was drawn. Therefore, you will probably want to draw a random sample from the population.

If you want to be able to analyze data not just for the entire population but also separately for subpopulations — such as specific racial/ethnic groups or men versus women — then you will probably want a stratified random sample. This means you will divide the total population into groups (strata) and then draw random samples from each of the groups. Often, you will need to oversample some groups (strata), in order to have a large enough sample for analysis and generalization.

For example: Suppose you want to know the level of awareness and knowledge about HIV/AIDS among school children grades three through six in your city or county. Your total population will be all children enrolled in grades three, four, five, and six, in public and private schools. You will want to draw a random sample of these children. Suppose the total elementary school population is 65% White non-Hispanic, 15% Hispanic, 12% African American, 6% American Indian/Alaskan Native, and 2% Asian/Pacific Islander. Because you want to be able to determine differences by sex and race/ethnicity, you will stratify your sample by race/ethnicity and also by sex, being sure to get approximately equal numbers of boys and girls. Instead of drawing a sample that “fits” the percentage breakdown of the school (for example, 65% White non-Hispanic, 15% Hispanic, etc.), you will probably need to oversample the racial/ethnic minority groups in order to have enough children from each group in order to separately tabulate and analyze information about them, generalize to the total population of children of that race/ethnicity and grade, and be able to compare findings for children across racial/ethnic groups. Then you can use a weighting procedure on the results to reflect the percentage of the total school population that comes from each racial/ethnic group, to generalize the results to the whole school population.
from the enrollments of community colleges and participants of employment and human service programs not requiring high school graduation — and from groups of employed young adults. Specific efforts to include out-of-school youth and employed young adults would be very important. If your planning group is relying on existing research, be sure to seek out studies which deliberately included and oversampled population segments important in your planning.

C. Data Collection and Analysis

1. Questionnaire Construction and Coding of Data

   This guide does not attempt to address in detail the issues related to constructing survey instruments and collecting data. However, it is important to be aware of the distinction between closed-ended and open-ended questions, and their implications for coding and tabulation of data (See Figure 2).

   Research involving large samples usually relies heavily on fixed-choice, closed-ended questions, which are easy to pre-code and tabulate on a computer. If a question has a set of pre-determined responses, then the person responding can simply check the appropriate box — or the interviewer can check it — and the questionnaire can be conveniently computer-tabulated. However, a respondent who doesn’t like any of the fixed choices may not respond to the question, or may be forced to choose a response which does not accurately reflect the desired response; for this reason, such questions are sometimes called forced-choice. Sometimes a question will allow the respondent to select “other” as a response — and the tabulation may simply include the number or percent of respondents choosing this answer.

   One alternative which allows more freedom of response is the use of some open-ended questions. Such questions often have no stated response options, so the individual responding writes down, or tells the interviewer, a personal answer to the question. This enables each person responding to respond precisely as desired. These responses can be analyzed individually and used to provide a narrative description. However, particularly if the sample is large, they are often coded so they can be computer analyzed. Typically, the people responsible for the research will read through the responses to a particular question and develop a set of codes which they believe cover the range of responses received. Then analysts will read each response and code it according to the categories provided. It is very important that such coding be done carefully, so that it accurately reports what the person responding wanted to communicate, and so that it is sensitive to small but important response differences. It is also important that the coding process recognize and accurately reflect cultural variations in responses.

---

**Figure 2**

**Use of Coding**

Coding refers to the process of “translating” data from one format to another. Information collected through surveys or other research is often coded so that it can be entered into a computer to be tabulated and analyzed.

Answers to survey questions are often closed-ended and pre-coded — for example, each pre-determined possible response may have a number.

☐ Yes  ☐ No

Sometimes questions are open-ended, so that the person responds with a narrative rather than checking off one of a set of pre-determined responses. This may be done because you don’t want to limit responses by giving a fixed set of alternative answers, or because you really don’t know what kinds of answers to expect. In such situations, the coding may be done after the data are collected.
In contributing to successful questionnaire design and data collection, planning group members might want to help ensure that:

- Interview guides and questionnaires are constructed by individuals knowledgeable about issues of reliability and validity, so that the questions consistently measure what they were intended to measure (See Figure 3);

- Individuals from various population segments of importance to the study are involved in preparation of the instruments, to assure that questions are relevant and appropriate for each of these subpopulations;

- Instruments are pilot tested on a small group of individuals similar to those who will be in the sample, including people from the various subpopulations, to be sure that they are generating consistency and valid data from all these subpopulations;

- Individuals involved in interviews or other data collection activities are carefully selected, trained, and monitored, to assure appropriate diversity and skills;

- Appropriate arrangements are made to obtain information from individuals who may be limited-English-proficient; and

- Coding of information from open-ended questions is done by individuals sensitive to cultural and other subgroup differences.

Research based on these considerations is likely to generate useful information — which then will need to be appropriately analyzed and presented.

2. Use of Percents

Presenting data in the form of percentages is a basic component of data analysis, and is very useful because percents are easier to compare than numbers. However, they should be used with care and sometimes caution. It is particularly important to be aware of the size and source of the numbers behind the percents, and to be sure that comparisons using percents are done correctly.

Beware of statistics that include percents but don’t tell you the size of the total population. Sometimes the number of total subjects is too small to make the statistics meaningful — the sample is not large enough to be representative of the entire population, so you cannot trust the findings.

It is important to ask about the numbers from which these percents were generated: beware of very small samples. What is the size of your study population? What is the size of your sample? How many cases does this represent? One helpful hint is to use what has been called “the rule of two.” A reporter who often reviews research studies suggests how to use that “rule”:
...If someone makes some numerical claim, I look at the numbers, then see how much I might change the finding by adding or subtracting two from any of the figures. For example, someone says there are five cases of cancer in a community. Would it seem meaningful if there were three?

Or if there were eight cases this year but four the year before — a 100 percent increase — I ask myself, “If I add two cases to last year’s total and subtract two from this year’s, is there a chance things haven’t changed, except by chance?” This approach will never supplant refined analysis. But by playing around with the numbers this way — I sometimes try three instead of two — a reporter can often spot a potential problem or error.*

Figure 4 shows how data based on percents can be misleading when based on very small numbers.

**Comparing percents is different from comparing numbers, because percents have already been changed from individual cases to proportions.** For example, suppose you want to understand the extent to

---

**Figure 4**

**Percents Based on Small Samples**

Suppose you received the following information:

*Of the individuals found to be HIV+ during the past three months at an STD clinic, the overwhelming majority were African American, although African Americans represented a small proportion of the total population tested.*

<table>
<thead>
<tr>
<th>Population Group</th>
<th>Percent of Total Persons Tested</th>
<th>Percent of HIV+ Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>White Non-Hispanic</td>
<td>70</td>
<td>33</td>
</tr>
<tr>
<td>African American</td>
<td>15</td>
<td>67</td>
</tr>
<tr>
<td>Hispanic</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**This could be a very important finding.** These statistics could summarize the situation if 2,000 people were tested (1,400 White non-Hispanics, 300 African Americans, and 300 Hispanics), and 100 were found to be HIV+, 67 of them African American.

**It could also be a finding from numbers too small to be meaningful statistically.** Suppose the actual number of people tested was 100 (70 White non-Hispanics, 15 African Americans, and 15 Hispanics), 3 were found to be HIV+, and 2 were African American. The number of HIV+ people — 3 — is too small to use in drawing any conclusions about the actual rates of HIV infection among these three groups in the whole population.

---

which people entering substance abuse treatment in your area are HIV+. Suppose all the programs tested new clients at entry, and 102 out of 803 new clients tested positive last year. Those numbers become much easier to understand as a percent — 13% — or proportion — about one in eight. But it is important to remember that 13% represents not 13 but 102 people. Figure 5 shows one of the most important considerations in working with percents: remembering the difference between percent and percentage point when making comparisons.

Percents are sometimes presented or explained incorrectly, resulting in an interpretation which is not accurate. For example, Figure 6 shows what happens if you confuse percents with percentage points. Once numbers have been converted into percents, the comparisons must be calculated differently.

---

**Figure 5**

**Percent versus Percentage Points**

Percents are often used to compare groups. For example, in 1990, the U.S. population was 75% White non-Hispanic, 12% African American, 9% Hispanic, 3% Asian/Pacific Islander, and 1% American Indian/Alaskan Native (Note that the percents should total 100% since the entire U.S. population = 100%).

If you compare the size of two groups based on percents, you can calculate the relative size of each group. If you compare groups using percentage points, you don’t have to do any calculations, but you also can’t explain the relative size of the two groups. To compare the size of the African American and Hispanic populations according to the 1990 Census, you can correctly say any of the following:

♦ Using percentage points:

  - The African American population was three percentage points larger than the Hispanic population in 1990, according to Census figures.

To calculate the percentage point difference, simply subtract one percent from the other — 12 minus 9 = 3. Be sure to state the difference not as a percent difference but as a percentage point difference. We cannot say that 12% minus 9% = 3%.

♦ Calculating the relative size of the two groups with percents:

  - The African American population was 33% larger than the Hispanic population in 1990, according to Census figures.

To calculate the percent difference, decide which number is your base of comparison, subtract it from the other number, divide the resulting number by the base number, and multiply by 100.

In this case, Hispanics are the base population, so calculate 12 minus 9 = 3; 3 divided by 9 = .33; .33 x 100 = 33% larger.

  - The Hispanic population was 25% smaller than the African American population in 1990, according to Census figures.

In this case, African Americans are the base population. So subtract 9 minus 12 = -3; -3 divided by 12 = - .25; -.25 x 100 = 25% smaller.

It is important to use the correct “base” percent, since you get very different numbers with different “bases” — just remember that the “base” is the number which forms the basis for your comparison.
3. Decimal Points and Rounding of Data

When you use statistics, you can be very precise. You can present statistics not only using whole numbers but also showing parts of the whole. Or you can use rounding to present numbers in more convenient units. Figure 7 provides some information about decimal points and rounding.

![Figure 6](image)

**Figure 6**

**An Example: Confusing Percents with Percentage Points**

Suppose an analysis of unmet needs said the following:

*All the major racial/ethnic population groups in the Midurban metro area are being equitably served by current prevention programs. When the breakdown of prevention program clients is compared with the breakdown of the Midurban area population, the two are within 2% of each other:*

<table>
<thead>
<tr>
<th>Population Group</th>
<th>Percent in Planning Prevention Programs</th>
<th>Percent in Planning Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanics</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>African Americans</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>American Indians</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>White Non-Hispanics</td>
<td>64</td>
<td>66</td>
</tr>
</tbody>
</table>

In fact, both American Indians and Hispanics are underserved. The differences are not more than two percentage points, but are a great deal more than two percent. Use the percent of each group in the planning area population as your base number; subtract the percent in the prevention programs from the base number, and multiply by 100 to get a percent. If the percent is negative, the group is underrepresented in prevention programs; if the percent is positive, the group is overrepresented. American Indians are 50% underrepresented — they are 4% of the population (so the base number for comparison is 4) but only 2% of the prevention program clients (2 minus 4 = -2; -2 divided by 4 = -.5 x 100 = -50%). Hispanic Americans are 20% underrepresented; they are 10% of the population (so the base number here is 10), but only 8% of the clients (8 minus 10 = -2; -2 divided by 10 = -.2 x 100 = -20%). Similarly, African Americans are about 9% overrepresented (24 minus 22 = 2; 2 divided by 22 = 9%), and White non-Hispanics are about 3% overrepresented (66 minus 64 = 2; 2 divided by 64 = 3%). The difference here is important. If you use percentage points, it appears that all populations are being equitably served, if you use percentages (which is the appropriate statistic to use), then it appears that American Indians and Hispanics are underserved by current HIV prevention programs.

4. Averages

You already know that averages represent a very commonly used method of analyzing, understanding, and comparing statistics. Averages help in determining and comparing typical responses, scores, or values. Averages are particularly useful in understanding the relative demographics of various population groups; for example, various types of averages help in comparing racial and ethnic groups in terms of age, income, family size, unemployment, and other personal and family characteristics. There are several different kinds of averages, each of which is calculated differently, as shown in Figure 8. Be sure the average used is appropriate for presenting the kind of data involved.
Figure 7
Decimal Points and Rounding

In a statistic, the numbers to the left of the period or decimal point are whole numbers, with a value greater than 1. The numbers to the right of the decimal point have a value less than 1. When you compare statistics or present them in the form of percents, the data often do not come out in whole numbers.

For example, if you find that 10 of 80 clients in a particular substance abuse program are HIV+, you can say that this is "1 in 8," or convert it to a percent (1 divided by 8 x 100) and say 12.5%. Similarly, if 5 of the 80 report more than 10 years of injecting drug use, that would be 1 in 16 or 6.25% (1 divided by 16 x 100).

Usually, you will want to simplify comparisons by using the same number of “decimal places” (that is, numbers to the right of the decimal point) for all the statistics you are comparing. You do this by rounding — expressing numbers in shorter, more convenient units. You might want to round all the numbers to the nearest “whole” number — so there are no decimal places — or to the nearest tenth of a percent — so there is one decimal place.

When you “round” or “round off” numbers, you usually round up any number above 5 (because it is more than half) and round down any number below 5 (because it is less than half). Usually, 5 is also rounded up.

To round 12.5% to the nearest whole number, you would round up to 13%.
To round 6.25% to the nearest whole number, you would round down, to 6%.
To round to the nearest tenth of a percent (one decimal place), you would use 12.5% as is, and round up 6.25% to 6.3%.

Figure 8
Understanding “Averages”

Three different kinds of averages are used in statistics, and each one is calculated differently and has a different meaning. Suppose the planning group is looking at a report on a survey of young adults aged 18-30 in your community who participated in an HIV prevention program, to find out the “average” age at which these young people became sexually active.

Suppose there were 15 young people in the program, and they reported the following ages at the time of first sexual encounter:
18, 28, 12, 11, 30, 15, 14, 20, 16, 13, 19, 13, 14, 17, 14

The average could be 17, 15, or 14, depending on what kind of average was calculated. It could be:

◆ 17 (rounded up from 16.9), which is the arithmetic average or mean age. The mean is calculated by adding up all the individual “values” — the ages at first sexual encounter of all the young adults in the sample — and dividing by the number of values — the number of young adults surveyed.

◆ 15, which is the median. The median is the central value, the one that falls in the middle of all the values when they are placed in order from highest to lowest. To calculate the median, you would reorder the “values” as follows: 30, 28, 20, 19, 18, 17, 16, 15, 14, 14, 14, 13, 13, 12, 11, and then count down to the middle, in this case the eighth, “value” — 15.

◆ 14, which is the mode. The mode is the most frequently occurring “value”; four young adults were 14.

In this situation, the mode is not a good measure to use, because the number of cases is small, and usually, it is best to use the median rather than the mean when a few values are very different from the others. In this case, two people were much older than the others at the time of first sexual encounter. This makes the mean quite high at 17; nine of the 15 young adults were younger than age 17 when they became sexually active.

When you review statistics, ask what kind of average was used, and why. Be sure it gives an accurate picture of what is “typical” for that group.
Averages can be very helpful in understanding risk factors for HIV. You may want to know the median age at which young people become sexually active, and to be able to compare this across populations or areas. You may want to compare the mean scores of groups of prevention program clients on pre- and post-tests of their knowledge of HIV transmission. In comparing male and female answers to questions about sexual behaviors, such as why young adults report engaging in unprotected sex, the mode can tell you the most frequent response.

5. Probability and Significance

One of the more important — and technically complex — aspects of analysis is the determination of whether research findings are statistically significant. Studies which are not based on formal sampling and analysis procedures can provide extremely valuable information about those responding or participating in the study, but the findings cannot be generalized to a broader population. Similarly, if a study does not use some kind of mathematical test of significance, you will not know how confident you can be that reported results reflect real characteristics or differences, rather than those which occurred by chance. Figure 9 provides a brief summary of the concepts of significance and probability as they relate to research findings.

Because research findings will provide an important information base for the planning group’s work, it is important that you know the extent to which you can “trust” the research available to you. If the findings are not statistically significant, they may still provide useful insights — but you will use them with some caution. If they are statistically significant, they may serve as the foundation for priority setting with regard to target groups and types of prevention activities. For example, if a survey found differences in sexual practices between different groups of gay and bisexual men, you would want to know whether the differences were statistically significant before being confident that you should use these differences in recommending types of prevention activities for each of these groups.

You will rely on statisticians to carry out tests of significance, as appropriate, in any needs assessment or other research conducted for the planning group. However, you will want to be aware of the need for these tests and ask whether they have been conducted, just as you will want to ask questions about the sampling procedures used.

When you look at existing research findings, particularly studies using large samples and reporting quantitative data, ask whether the findings have been subjected to appropriate tests of significance. If the sampling approach was appropriate and the analysis was properly done, you can have confidence in the results. Similarly, if your planning group conducts or supports research, help ensure that the sampling and analysis techniques are sufficient to give you confidence in the findings.

II. Reviewing Statistics

Every planning group member should feel comfortable in questioning statistics — whether they come from existing research or from studies conducted for the planning group. In helping with this process, consider the following:

With regard to research conducted by or for the planning group:

- **Become involved at the design level.** Once the data have been collected, it is too late to make major changes in the research. Most studies include detailed plans including how the data will be collected and analyzed; review the plans to be sure that important issues are being addressed.
Figure 9
Probability and Significance

As described in Part 2, significance means that a research finding is meaningful or important. A research finding is considered to be statistically significant if there is only a small probability that the observed result could have occurred by chance alone, rather than measuring something that is real — such as a real difference between the behaviors of two populations, or a real effect of an HIV/AIDS prevention intervention.

Typically, statistical results are considered to be significant if there is less than a 5% chance — 5 times out of 100 — that the observed difference or relationship would be found by chance alone. In such situations, the probability value or \( p \) value is said to be less than \( .05 \) \((<.05)\). A more stringent \( p \) value is \( <.01 \), which means that there is less than a 1% chance that the observed difference or relationship occurred by chance alone.

The \( p \) value is calculated to determine whether results are statistically significant, and should be presented in statistical reports along with the findings, so that anyone using the data will know how much confidence to place in the reported relationships and analysis results.

- Ask the researchers to describe the planned sampling procedure in non-technical terms, so you can be sure it will generate data that meet the planning group’s needs; for example, find out the sampling universe, whether a stratified random sample is planned, the size of the sample, the way in which people in the sample will be contacted, etc.; this will let you identify whether the procedures were appropriate for fully including specific populations.

- Ask that minorities or other defined populations of particular interest be oversampled; this ensures large enough subsamples for meaningful analysis of each population group. Be sure that the researchers are using a sampling universe which can provide a representative sample — for example, if the population is young adults, not doing a survey only in schools but also locating dropouts.

- Be sure that all relevant population groups are included in research based on focus groups or other informal groups of respondents; be prepared to suggest ways to identify people from groups that might otherwise be excluded or underrepresented. You may know community-based organizations which have large numbers of program participants from these groups.

- Ask to review draft data collection forms, such as interview guides, questionnaires, or focus group scripts, for both content and relevance to specific populations.

- Ask about analysis plans. Be sure that separate analyses and comparisons are planned by subpopulation. Ask how qualitative information will be coded and analyzed. See whether tests of significance are planned.

- Find out how the data will be presented. Be sure all important information will be analyzed and presented, that qualitative information will be handled appropriately, and that findings will be reported in a user-friendly language and format.

With regard to existing studies and statistics — data not collected by or for the planning group — try to learn all you can about how the study was designed and carried out. Usually, this information is included in the study report. Sometimes, it may be possible to ask the researchers who did the study to make a
presentation at a planning group or committee meeting, or someone might contact them to obtain information not available in the study report or data tables. Among the questions you might want to ask are the following:

- Were the findings presented as representative of a larger population, or are they limited to the individuals actually included in the study? If the findings are generalized, do the methods used justify this?

- What was the sampling universe? If the findings are being generalized to a total population, what is the population and how was it defined and sampled?

- What sampling procedure was used? Was a stratified random sample drawn? Were some populations oversampled?

- Did the study succeed in getting data from the proposed sample? How were people in the sample contacted? How many people were surveyed, compared to the proposed sample size — were response rates lower for particular populations?

- Who collected the data and how? Were people from various population segments involved? If language-minority groups were sampled, was the data collection form translated into the appropriate languages, and were bilingual personnel used in data collection? What kind of quality control was done to assure that the data were collected properly?

- How were the data analyzed? Was there any coding of open-ended questions? Were people from appropriate cultural backgrounds involved?

- What tests of significance were conducted? What did these tests indicate?

- What major limitations are reported by the researchers?

As you review data tables and research reports within the HIV Prevention Community Planning process, you will become more and more familiar with the issues and questions of greatest importance to your work. Your basic purpose is clear: you want to understand what the research says, the extent to which the findings are truly representative of various population groups, and the confidence you can have that the reported findings genuinely reflect the various population segments which make up your community planning area.