ANSWERS TO COMMONLY ASKED QUESTIONS ABOUT R-EMAP SAMPLING DESIGNS AND DATA ANALYSES

Prepared for

Victor Serveiss
U.S. Environmental Protection Agency
Research Triangle Park, NC

Prepared by

Jon H. Volstad
Steve Weisberg

Versar, Inc.
Columbia, MD 21045

ar.d

Douglas Heimbuch
Harold Wilson
John Seibel

Coastal Environmental Services, Inc.
Linthicum, MD

March 1995

ANSWERS TO COMMONLY ASKED QUESTIONS ABOUT R-EMAP SAMPLING DESIGNS AND DATA ANALYSES

Prepared for

Victor Serveiss
U.S. Environmental Protection Agency
Research Triangle Park, NC

Prepared by

Jon H. Volstad
Steve Weisberg

Versar, Inc.
Columbia, MD 21045

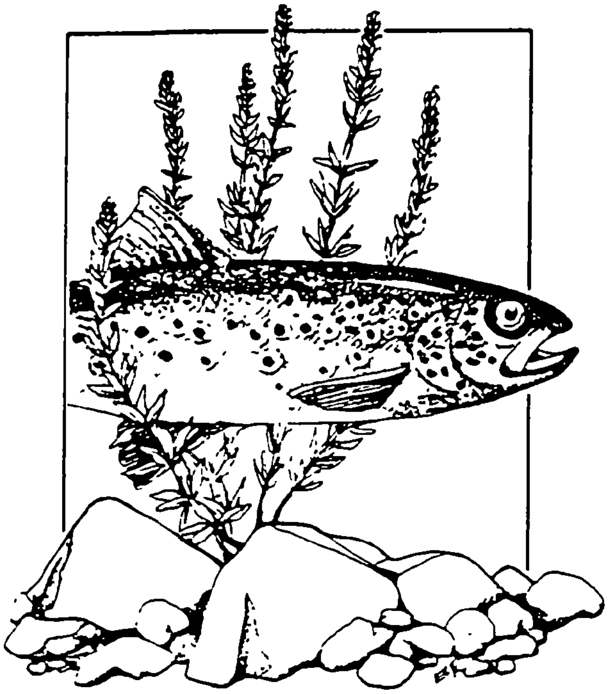
ar.d

Douglas Heimbuch
Harold Wilson
John Seibel

Coastal Environmental Services, Inc.
Linthicum, MD

March 1995

ANSWERS TO COMMONLY ASKED QUESTIONS ABOUT R-EMAP SAMPLING DESIGNS AND DATA ANALYSES



INTRODUCTION

The Environmental Monitoring and Assessment Program (EMAP) is an innovative, long-term research, and monitoring program designed to measure the current and changing conditions of the nation's ecological resources. EMAP achieves this goal by using statistical survey methods that allow scientists to assess the condition of large areas based on data collected from a representative sample of locations. Statistical survey methods are very efficient because they require sampling relatively few locations to make valid scientific statements about the condition of large areas (e.g., all wadable streams within an EPA Region).

Regional-EMAP (R-EMAP) is a partnership between EMAP, EPA Regional offices, states, and other federal agencies to adapt EMAP's broad-scale approach to produce ecological assessments at regional, state, and local levels. R-EMAP is based on the same statistical survey techniques used in EMAP, which have proven successful in many disciplines of science. Applying these techniques effectively requires recognizing several key principles of survey sampling and using specialized, although not difficult, data analysis methods.

This document provides a nontechnical overview of the survey sampling and data analysis concepts underlying R-EMAP projects. It is intended for regional resource managers who have had little statistical training, but who feel they would benefit from a better understanding of the statistical and scientific underpinnings of R-EMAP. Familiarity with these concepts is helpful for understanding the kinds of information R-EMAP can provide and appreciating the strengths of R-EMAP. Several additional documents are being prepared for scientists with some statistical training who may become involved in analyzing R-EMAP data.

This document is organized in two sections. The first section explains the general principles of survey sampling and its application to determining ecological condition. Terms such as target population, sampling frame, and random sampling are defined. The second section addresses questions frequently asked about the R-EMAP sampling design and data analysis methods. Throughout the document, the concepts of survey design are illustrated first with examples from everyday life, and then with examples from a typical R-EMAP study. The R-EMAP examples involve a stream study; however, the concepts are equally applicable to assessing the condition of other resources such as lakes, estuaries, wetlands, or forests.

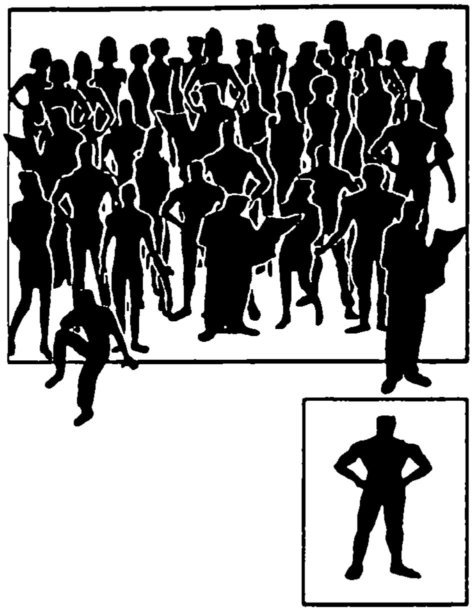
PRINCIPLES OF SURVEY DESIGN

There are two generally accepted data collection schemes for studying the characteristics of a population. The first is a census, which entails examining every unit in the population of interest. For most ecological studies, however, a census is impractical. For example, measuring fish assemblages everywhere to assess conditions within a watershed that has 1000 kilometers of stream would be prohibitively expensive.

A more practical approach for studying an extensive resource, such as a watershed, is to examine parts of it through probability (or random) sampling. Studies based on statistical samples rather than complete coverage (or enumeration) are referred to as sample surveys. Sample surveys are highly cost-effective, and the principles underlying such surveys are well developed and documented. The principles of survey design provide the basis for (a) selecting a subset of sampling units from which to collect data, and (b) choosing methods for analyzing the data.

One example of a sample survey is an opinion poll to estimate the percentage of eligible voters who plan to vote Democratic in a presidential election. Such opinion polls are based on interviews with only a small fraction of all eligible voters. Nevertheless, by using statistically sound survey methods, highly accurate estimates can be obtained by interviewing a representative sample of only around 1200 voters. If 700 of the polled voters plan to vote Democratic, then the fraction 700/1200, or 58 percent, is a reliable estimate of the percent of all voters who plan to vote Democratic.

A target population of enrolled students at a university. Sampling unit = individual student.

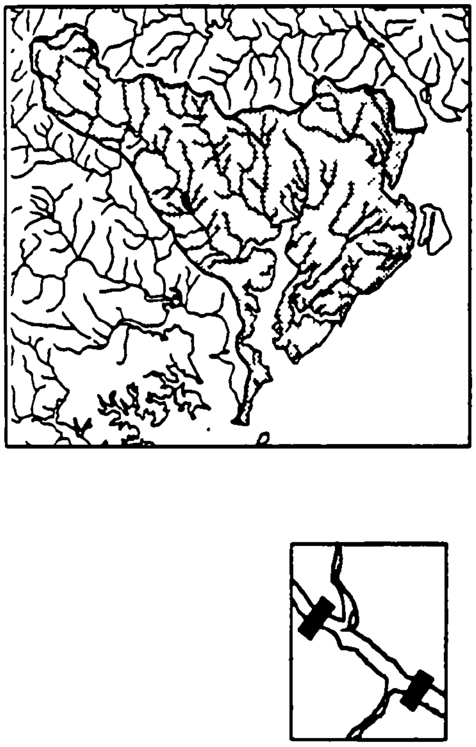


The approach used in conducting a R-EMAP stream survey is basically the same as in an opinion poll. Instead of collecting the opinions of a sample of people, a R-EMAP project might collect data about fish assemblages from a representative sample of point locations along the stream length of a watershed to determine the percent of kilometers of streams in which ecological conditions are degraded. If data are collected from plots of, say, 40 times the stream width in length at each of 40 randomly selected sites, and 16 of the 40 sites exhibit degraded conditions, then the estimated proportion of degraded stream kilometers in the watershed would be 40% (i.e., 16/40).

STEPS FOR IMPLEMENTING A SAMPLE SURVEY

The *survey design* is a plan for selecting the sample appropriately so that it provides valid data for developing accurate estimates for the entire population or area of interest. Planning and executing a sample survey involves three primary steps: (1) creating a list of all units of the target population from which to select the sample, (2) selecting a random sample of units from this list, and (3) collecting data from the selected units. The same techniques used to select the sample of people to interview in an opinion poll are used to select the sample of sites from which to collect field data.

A target population of perennial, wadable streams in a watershed. Sampling unit = point location and associated plot.



Developing a Sampling Frame

Before the sample survey can be conducted, a clear, concise description of the *target population* is needed. In statistical terminology the target population (often shortened to "population") does not necessarily refer to a population of people. It could be a population of schools, area units of farm land, freshwater lakes, or length-segments of streams.

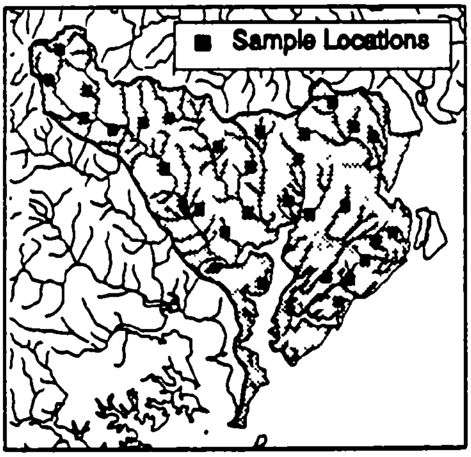
The list or map that identifies every unit within the population of interest is the *sampling frame*. Such a list is needed so that every individual member of the population can be identified unambiguously. The individual members of the target population whose characteristics are to be measured are the *sampling units*.

A random sample of students from the target population. The poll results in "yes" or "no" responses.

For example, if we were conducting a sample survey to estimate the percentage of students at a university who participate in intramural sports, the target population would consist of all the enrolled students. The individual students would be the sampling units, and the registrar's office could provide a list of students to serve as the sampling frame. We could draw a representative (random) sample of students from this list and interview them about their participation in sports. Their responses would be "yes" or "no." The percentage of interviewed students who participate in intramural sports would yield an estimate of the "true" percentage for all students.

For a stream survey, the target population might be all perennial, wadable streams in a watershed. The sampling unit is a point along the stream length, and an associated plot, e.g. 40 times the stream width in length. The response variable might be "degraded" or "non-degraded" based on measures of water quality. Conceptually, the collection of all possible point locations along these streams serve as a sampling frame, similar to the list of students in the previous example. The sampling frame for streams typically would be established by using U.S. Geological Survey stream reach files through a geographic information system (GIS).

A random sample of locations from the target population.



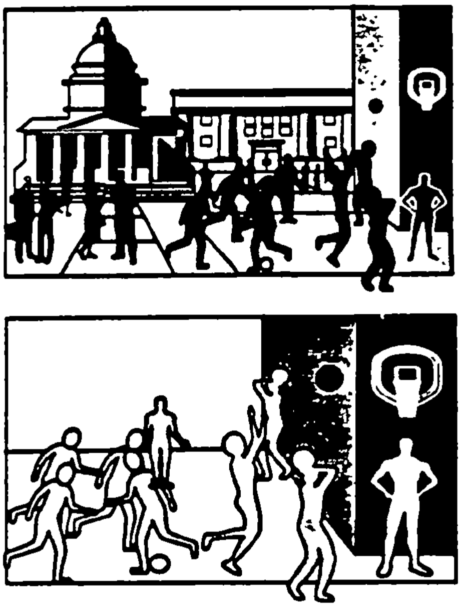
Selecting a Representative Sample

Survey sampling is intended to characterize the entire population of interest; therefore, all members of the target population must have a known chance of being included in the sample. Conducting an election poll by asking only your neighbors' opinions probably would not enable you to predict the outcome of a national election accurately.

Simple random selection ensures that the sample is representative because all members of the population have an equal chance of being selected. Random selection can be thought of as a kind of lottery drawing to determine which stream reaches, for example, are included in the sample. The selection is non-preferential towards any particular reach or group of reaches. One way to make a random selection would be to place uniquely numbered ping-pong balls (one for each sampling unit) into a drum, blindly mix the drum, and then blindly pick one ball corresponding to each stream reach (i.e., sampling unit) from which data are to be collected. In practice, computers are used to make the random selections. Either way, the result is a subset of sampling units randomly selected from the sampling frame.

FREQUENTLY ASKED QUESTIONS

Students polled at the entrance to the gymnasium are not representative of all students on the university campus.

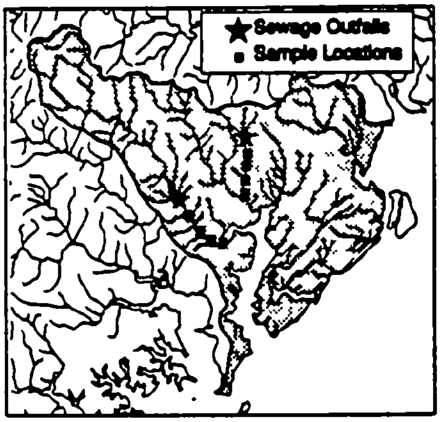


Upon thoughtful consideration of the sample survey approach, several questions may come to mind. This section answers several commonly asked questions. Some of them concern survey sampling, and some of them concern data analysis. These questions are addressed in fairly general terms. As noted in the introduction, additional technical detail will be available in a series of methods manuals.

Why is it so important to select sampling sites randomly?

The way we select the sample (i.e., choose the units from which to collect data) is crucial for obtaining accurate estimates of population parameters. We clearly would not get a good estimate of the percentage of all students at a university who participate in intramural sports if we polled students at the entrance to the gymnasium. This preferential sample would, most likely, include a much higher proportion of athletes than the general population of students.

A biased sample of locations from the target population of all streams in the shaded area.



Similarly in a stream study, preferential sampling occurs if the sample includes only sites downstream of sewage outfalls in a watershed where sewage outfalls affect only a small percentage of total stream length. This kind of sampling program may provide useful information about conditions downstream of sewage outfalls, but it will not produce estimates that accurately represent the condition of the whole watershed.

Preferential selection can be avoided by taking random samples. Simple random sampling ensures that no particular portion of the sampling frame (i.e., groups of students or kinds of river reaches) is favored. Within streams, the chance of selecting a sampling unit that has degraded ecological conditions would be proportional to the number of sampling units within the target population that have degraded conditions. For example, if 30% of the target population has degraded conditions, then on average 30% of the (randomly selected) units in the sample will exhibit degraded conditions. This property of random sampling allows estimates (based only on the sample) to be used to draw conclusions about the target population as a whole.

For 305b reports, I need to estimate the total number of stream miles in my EPA Region that are degraded. Can I do this from sample survey data?

The number of degraded stream miles can be calculated in two steps. First, the proportion of stream miles that are degraded is calculated as illustrated earlier. Then, that fraction is multiplied by the total number of stream miles in the population. The total number of stream miles is available from the sampling frame, which delineates all members of the target population.

Defining "degraded" is an important part of the calculation, regardless of whether it is for percent or absolute number of stream miles. "Degraded" can be defined if a *threshold value* or goal for each measurement variable can be established. Most of the variables measured in stream surveys, such as pH, have continuous ranges of response (e.g., between 1 and 14 for pH). Calculating the proportion of stream miles that are degraded requires converting this continuous data into binary, or yes/no (e.g., degraded or not degraded) form. The question of how many stream miles are degraded, therefore, must be rephrased to include a threshold value for the relevant measurement variable. For pH, the question might be rephrased as "What are the total number of stream miles in my Region with pH below 6.5?"

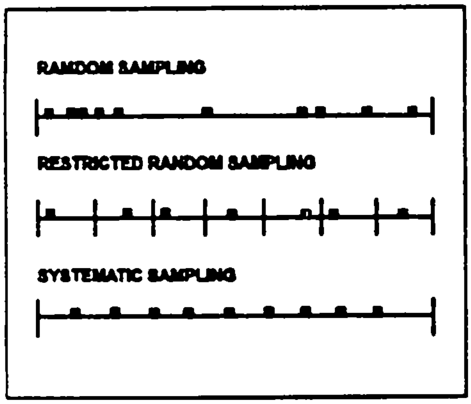
I am accustomed to seeing estimates of average condition instead of estimates of proportion. Can R-EMAP data be used to estimate average condition?

Yes, estimates of average condition, such as the average pH in a watershed, provide valuable information and can be calculated with R-EMAP data as a simple mean. The principles of survey sampling, particularly the emphasis on selecting a representative sample, also apply to estimating a population mean. Just as an estimate of the percent of stream miles in a Region in which pH is below 6.5 is biased if data are collected only from sites downstream of sewage outfalls, so is the estimate of mean pH.

EMAP emphasizes estimating spatial extent (e.g., percent of river miles) because it has several advantages over estimating the mean. For instance, a Region with an average stream pH of 7 might be composed entirely of streams with a pH of 7; however, the same average would occur if half the streams have a pH of 6 and the other half a pH of 8. Estimating the spatial extent of the resource that fails to meet some standard (e.g., pH of at least 6.5) provides more information about the condition of the resource and is consistent with EPA initiatives to establish environmental goals and measure progress toward meeting them.

Many EMAP documents refer to hexagons in describing the sampling design. How are hexagons involved?

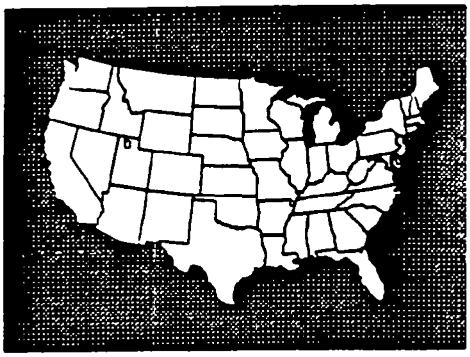
Distribution of sampling locations along a transect for different sampling schemes.

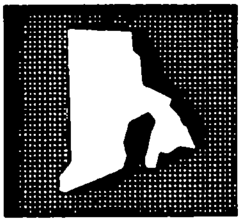


In geographic studies, such as a stream survey, it is often desirable to distribute samples throughout the study area. Often this is accomplished using a systematic design in which samples are placed at regular intervals. In EMAP, this is accomplished by a special kind of random sampling known as *restricted random sampling*. This type of random sampling has a systematic component. The systematic element causes the selected sampling units to be spread out geographically. The random element ensures that every sampling unit has an equal chance of being selected. The illustration at left compares the typical allocations of sampling units along a transect for random, restricted random, and systematic sampling designs.

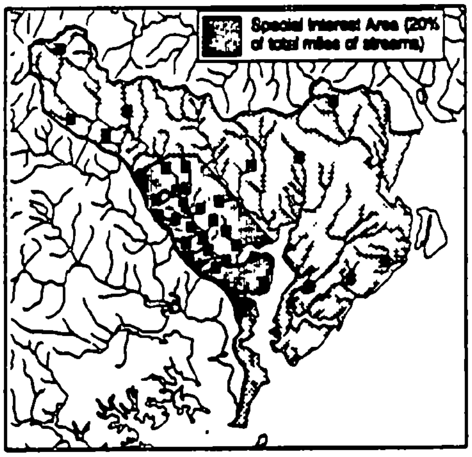
In EMAP, hexagons are used to add the systematic element to the design. The hexagonal grid is positioned randomly on the map of the target resource, and sampling units from within each grid cell are selected randomly. The grid ensures spatial separation of selected sampling units; randomization ensures that each sampling unit has an equal chance of being selected.

Target population: all eligible voters in all states. Area of special interest (stratum): voters in Rhode Island.





Target population: watershed with 1000 km of streams. Area of special interest (stratum): 200 km of streams.



EMAP documents suggest that the sampling design is "flexible to enhancement." What does this mean?

One goal of a sample survey may be to compare a subpopulation with the target population. For instance, an opinion poll might be used to determine if a higher percentage of the people living in Rhode Island are likely to vote Democratic than in the nation as a whole. Given its small size, Rhode Island probably would receive very little attention in a national poll if samples are allocated randomly. One way to achieve a sample of people in Rhode Island that is sufficient to make this comparison is to increase sampling effort for the nation as a whole until enough people from Rhode Island are included in the randomly selected national sample. This option is not very cost-effective because it requires considerable, unnecessary sampling effort in other areas to achieve a desired sample size in one small area.

Another, preferable, alternative would be to divide the entire target population into two subpopulations, or *strata*. Voters in the United States could be stratified into (1) those living in Rhode Island, and (2) those living elsewhere. A simple random sample of desired size could then be selected from each of these groups. Statisticians refer to this as *stratified random sampling*. Stratified sampling designs can have any number of strata with a different level of sampling effort in each.

Stratified sampling could be used in a stream survey to enhance sampling effort in a watershed of special interest so that its condition could be compared with that of a larger area. In a study area with 1000 kilometers of streams, for example, an area of special interest may contain 200 kilometers of streams. If budget constraints limit the size of the total sample to 60 sampling units, 30 could be randomly selected from the special interest area, and 30 from the rest of the sampling frame. If simple random sampling is used, the area of special interest, which represents 20% of the area, will contain only about 12 of the 60 selected sampling units. A sample of 12 would be insufficient to estimate the condition of the special interest area reliably.

Doesn't enhancing the sampling intensity for an area of special interest bias the overall estimate?

No. Sampling units inside an area of special interest usually have a higher chance of being selected than sampling units outside the special interest area. Within each stratum, however, the chance of selecting any location is equal; therefore, a separate (unbiased) estimate can be computed for each stratum.

With stratified random sampling, estimates are generated first for individual strata, then the stratum-specific estimates are combined into an overall estimate for the whole target population. Stratum-specific estimates are combined by weighting each one by the fraction of all sampling units that are within the stratum. For the simple two-stratum example given above, the weights would be 200/1000 for stratum 1 and 800/1000 for stratum 2. So, if the stratum-specific estimates are 0.5 for stratum 1 and 0.25 for stratum 2, the overall estimate is 0.30 [(0.5 x 2/10) + (0.25 x 8/10)]. This approach ensures that the overall estimate is corrected for the intentional selection emphasis within a particular subpopulation.

EMAP's objectives state that estimates are made with known confidence. What is "known confidence"?

An estimate of a population parameter is of limited value without some indication of how confident one should be in it. Scientists typically describe the appropriate level of confidence in an estimate derived from a sample survey by defining *confidence limits* or *margins of error.* This description of statistical confidence is used frequently in reporting the results of opinion polls using statements such as "this poll has a margin of error of ± 4%". Provided random sampling is used, similar statements can be made about estimates from biological sample surveys.

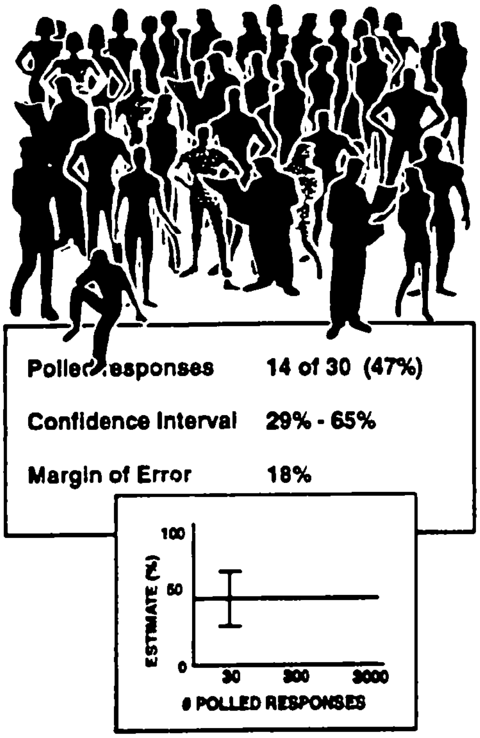
Sample surveys provide estimates that are used to make inferences about parameters for the population as a whole. Two types of estimates are commonly provided: the point estimate and the interval estimate. For example, the estimated proportion of voters that support a party is a point estimate. It is important to know how likely it is that such a point estimate deviates from the true population parameter by no more then a given amount. An interval estimate for a parameter is defined by upper and lower limits estimated from the sample values. A confidence interval is constructed so that the probability of the interval containing the parameter of interest can be specified. We do not know with certainty whether an individual interval, specified as a sample estimate plus minus a margin of error, includes the true population parameter. For repeated sampling, however, the estimated 95% confidence intervals would include the true parameter 95% of the times. The length of the confidence intervals is a measure of how precise the parameter is being estimated: a narrow interval signify high precision. The margin of error is often used for defining the upper and lower limits of the confidence interval, it is half the width of the confidence interval. Thus, if a poll estimates that 55% of the population will vote Democratic and the margin of error is ± 4%, then the estimated 95% confidence interval ranges from 51% to 59%.

A great advantage of using a random sampling design is that statisticians have developed procedures for calculating confidence intervals for the estimates. For most R-EMAP projects, in which the goal is to estimate the proportion of the resource that is degraded, a standard probability distribution known as the *binomial distribution* can be used to determine the upper and lower bounds of confidence intervals.

What are the most important factors affecting the size of the confidence interval?

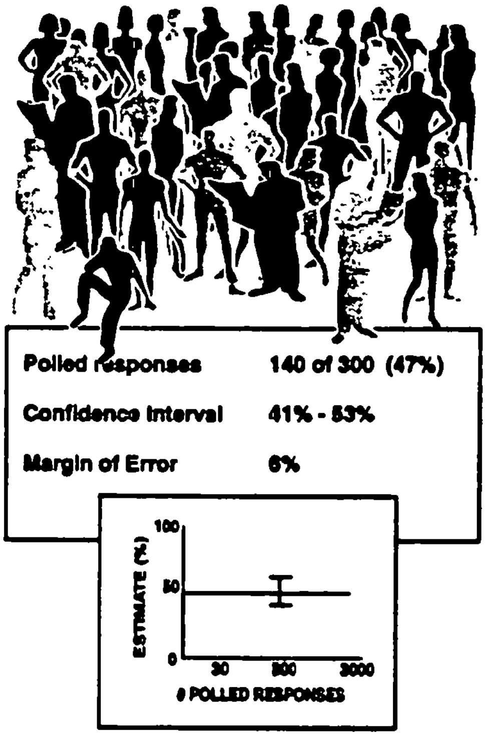
The sample size (# of sampling units collected) and the proportion of yes answers are the primary factors affecting the size of the confidence interval with binary (yes/no) data. The effect of sample size can be illustrated with a pre-election poll of voters. If only 30 people are sampled, and 14 indicate that they will vote Democratic, it would be unwise to predict the winner. With such a small sample size, the margin of error would be about ± 18% for a 95% confidence interval. The degree of confidence would be higher if 140 people out of a sample of 300 say they will vote Democratic (47% ± 6%), and higher still if 1400 people out of a sample of 3000 say they will vote Democratic (47% ± 2%). In this example, the estimated proportion of sampled voters

Percent of Democratic voters estimated from a sample of 30; note the wide confidence interval.



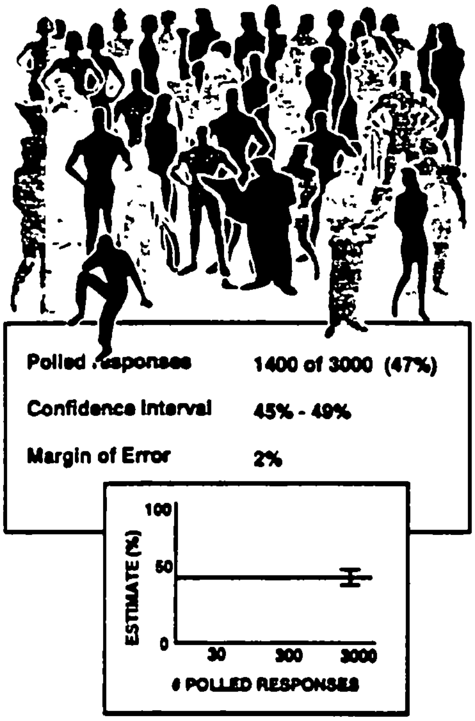
Polled responses	14 of 30 (47%)
Confidence Interval	29% - 65%
Margin of Error	18%

A sample of 300 people produces a better estimate; the confidence interval is narrower.

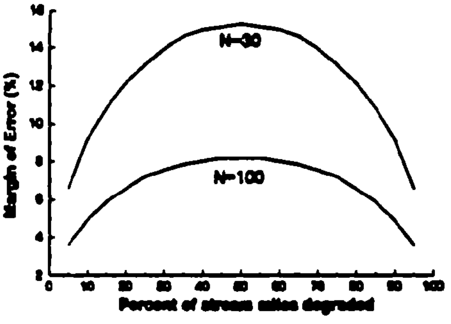


Polled responses	140 of 300 (47%)
Confidence Interval	41% - 53%
Margin of Error	6%

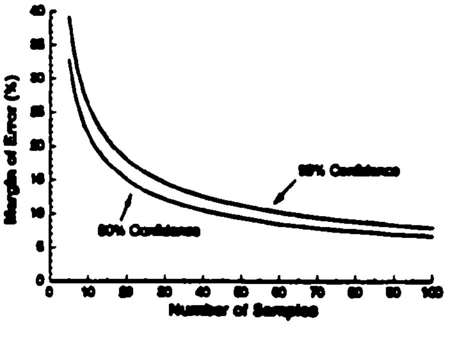
A sample of 3000 people produces a very accurate estimate, with a narrow confidence interval.



Margin of error as a function of the percent yes responses for fixed sample sizes of 30 and 100 (90% confidence interval).



Plot of margin of error versus sample size when 20% of the population is in the YES category (P = 0.2).



who will vote Democratic stays the same (p =47%), but the width of the confidence interval decreases with increasing sample size.

Confidence intervals for estimated percentages (p) are affected to a lesser degree by the proportion of yes answers (P) in the population. The widest confidence interval occurs for P equal to 50%. For values of P ranging from 20% to 80%, the margin of error will not vary much with P; it will be determined mainly by the sample size. The fact that there is a maximum margin of error for binomial estimates of proportions is very useful for planning a survey. If we plan for the worst case (i.e., when half of the population is in the yes category) we can select a sample size that ensures that the confidence interval for P will be smaller than a specified limit.

Doesn't the size of the target population affect confidence in the estimates?

The size of the target population theoretically affects the precision of the estimates. For most sample surveys, however, the effect is negligible because the sampled fraction of the target population is so small. When the sampled fraction is small, the size of the sample rather than the size of the target population determines the precision of the estimate. Polling 1000 people in the state of Rhode Island, for example, would yield as precise an estimate as polling 1000 people in the state of Texas. In both cases, a very small proportion of the total population is polled.

If the sample includes a large proportion of the population, in contrast, the accuracy of the estimate is improved. For instance, if a local town has a population of 1400 people, then a sample of 1200 people would produce a substantially more accurate estimate than a sample of 1200 people from a population of 100 million. As the size of the sample approaches the size of the population, statisticians adjust the confidence interval using the *finite population correction factor.* In practice, however, most sampling efforts don't sample a large enough fraction of the population for this correction factor to become important. That is why pollsters interview approximately the same number of people for a local election as for a presidential election.

For R-EMAP projects, the fraction of the population that is sampled is generally very small. Fish assemblages, for example, are generally sampled from 100-meter segments. If 50 such samples are collected from a Region with 1000 miles of streams, the sampled fraction is .0031.

CLOSING COMMENTS

The approaches and concepts described in this overview document are generally applicable to all R-EMAP projects. They are appropriate whether the purpose of sampling is to estimate the proportion of the number of resource units (e.g., numbers of lakes), the proportion of total length of a resource (e.g., miles of streams), the proportion of area of a resource (e.g., square miles of an estuary), or the proportion of volume of a resource (e.g., cubic meters of one of the Great Lakes). The approaches and concepts can be applied without modification to each of these situations.

This overview document purposefully was written non-technically; it does not contain enough detail to help someone analyze data. Three companion documents are being prepared to provide additional technical detail about recommended methods. These manuals describe data analysis methods (1) for assessing status (e.g., proportion of area with degraded conditions), (2) for assessing differences in proportions between two sub-populations of interest (e.g., deep versus shallow areas, two different states, two different stream orders), and (3) for assessing long-term trends. The methods manuals are intended for scientists with some statistical training. Technical documentation targeted for statisticians is available from the EMAP Statistics and Design Team in Corvallis, Oregon.

BIBLIOGRAPHY

Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. John Wiley and Sons. New York.

Gilbert, R. O. 1987. *Statistical Methods for Environmental Monitoring*. Van Nostrand Reinhold. New York.

Jessen, R. J. 1978. *Statistical Survey Techniques*. John Wiley and Sons. New York.

Stuart, A. 1984. *The Ideas of Sampling*. MacMillan Publishing Company. New York.