# MODULE 2

# FOUNDATIONAL DEFINITIONS

**Objectives:**

1. Describe what an individual is.
2. Describe what a population and a sample are and how they differ.
3. Describe what a parameter and a statistic are and how they differ.
4. Describe how a population, parameter, sample, and statistic are related.
5. Identify the individual, variable(s), population, parameter(s), sample, and statistic(s) (IVPPSS) in a given situation.
6. Identify variable types in context.

## Contents

STATISTICAL INFERENCE IS THE PROCESS of forming conclusions about the unknown parameters of a population by computing statistics from the individuals in a sample. As you can imagine from this definition, it is important that you understand the difference between a population and a sample and a parameter and a statistic before you can understand and appreciate the process of making statistical inferences. Before identifying these items, the individual and variable(s) of interest must also be identified. Understanding and identifying these six items is the focus of this module. Formal methods of inference are discussed beginning with Module 11.

Δ **Inference**: The process of forming conclusions about the unknown parameters of a population by computing statistics from the individuals in a sample.

The following hypothetical example is used throughout this module. Assume that interest is in determining the mean (or average) length of the 1015 fish in Square Lake (Figure 2.1). In "real life" you would not know how many fish are in this lake. However, for the purpose of illustrating important concepts in this module, it is assumed that all information for all 1015 fish in this lake is known.

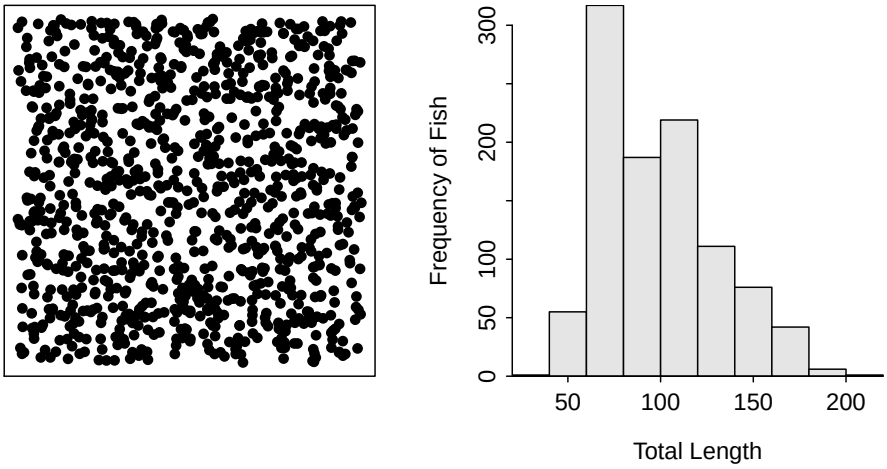


Figure 2.1. Schematic representation of individual fish (i.e., dots; **Left**) and histogram (**Right**) of the total length of the 1015 fish in Square Lake.

## 2.1 Definitions

The **individual** in a statistical analysis is one of the "items" to be examined by the researcher. Sometimes the individual is a person, but it may be an animal, a piece of wood, a site or location, a particular time, or an event. It is extremely important that you don't always visualize a person when you use the word individual in a statistical context. Synonyms for individual are unit, experimental unit (usually used in experiments), sampling unit (usually used in observational studies), case, and subject (usually used in studies involving humans). The individual of interest in the Square Lake example is an individual fish, because the researcher will collect a set of fish and examine each fish individually.

Δ **Individual**: One of the items examined by the researcher.

◇ **An individual is not necessarily a person.**

The **variable** is the characteristic of interest recorded about each individual. The variable of interest in the Square Lake example is the length of each fish. Note that in most "real life" studies the researcher will record more than one variable. For this example, the researcher may also record the fish's weight, sex, and age. Studies with one variable are called univariate studies, studies with two variables are bivariate studies, and studies with more than two variables are called multivariate studies.

**Δ Variable**: The characteristic of interest recorded about each individual.

A **population** is ALL individuals of interest. In the Square Lake example, the population is all 1015 fish in the lake. The population should be defined as thoroughly as possible including qualifiers as necessary. This example is simple because Square Lake is so well defined; however, as you will see in the review exercises, the population is often only well-defined by your choice of descriptors.

**Δ Population**: ALL individuals of interest.

A **parameter** is a summary computed from ALL individuals in a population. The term for the particular summary is usually preceded by the word "population." Parameters are ultimately what is of interest because interest is in all individuals in the population. However, in practice, parameters cannot be computed because the entire population cannot be "seen." In this hypothetical example, the parameters can be computed because all 1015 fish are accessible. In this example, the researchers were interested in the population mean length of all fish in Square Lake, which is 98.06 mm (Table 2.1).[1]

Table 2.1. Parameters for the total length of ALL 1015 fish in the Square Lake population.

| n | mean | sd | min | Q1 | median | Q3 | max |
|---|---|---|---|---|---|---|---|
| 1015 | 98.06 | 31.49 | 39 | 72 | 93 | 117 | 203 |

**Δ Parameter**: A summary of ALL individuals in a population.

**◇ Populations and parameters can generally not be "seen."**

The entire population cannot be "seen" in real life. Thus, a subset of the population is usually examined to learn something about the population. This subset is called a **sample**. The red dots in Figure 2.2 represent a random sample of n=50 fish from Square Lake (note that the sample size is usually denoted by n).

**Δ Sample**: A subset of the population examined by a researcher.

Summaries computed from individuals in a sample are called **statistics**. Specific names of statistics are preceded by "sample." The statistic of interest is always the same as the parameter of interest; i.e., the statistic describes the sample in the same way that the parameter describes the population. For example, if interest is in the population mean, then the sample mean would be computed.

Some statistics computed from the sample from Square Lake are shown in Table 2.2 and Figure 2.2. The sample mean of 100.04 mm is the best "guess" at the population mean. Not surprisingly from the discussion in Module 1, the sample mean does not perfectly equal the population mean.

[1] We will discuss how to compute and interpret each of these values in later modules.

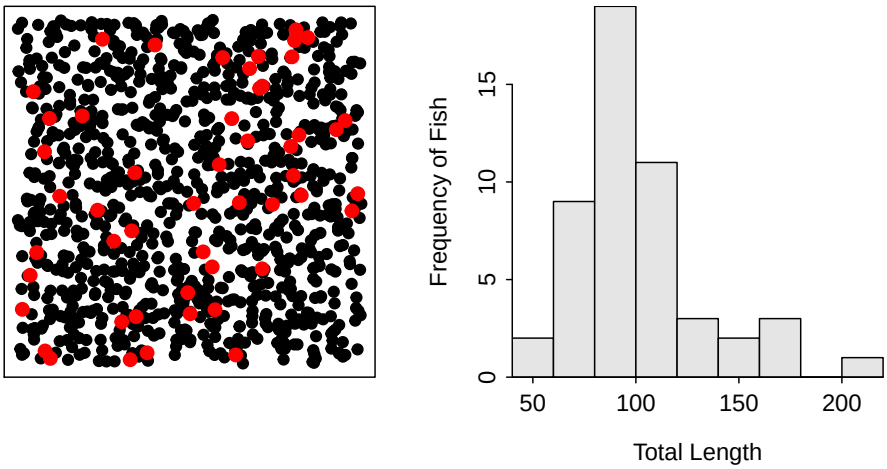


Figure 2.2. Schematic representation (**Left**) of a sample of 50 fish (i.e., red dots) from Square Lake and histogram (**Right**) of the total length of the 50 fish in this sample.

Table 2.2. Summary statistics for the total length of a sample of 50 fish from the Square Lake population.

| n | mean | sd | min | Q1 | median | Q3 | max |
|---|---|---|---|---|---|---|---|
| 50 | 100.04 | 31.94 | 49 | 81 | 91 | 118 | 203 |

**Δ Statistic**: A summary of all individuals in a sample.

## 2.2 Performing an IVPPSS

In each statistical analysis it is important that you determine the Individual, Variable, Population, Parameter, Sample, and Statistic (**IVPPSS**). First, determine what items you are actually going to look at; those are your individuals. Second, what are you going to record when you look at an individual; that is the variable. Third, the population is simply ALL of the individuals. Fourth, the parameter is a summary (e.g., mean or proportion) of the variable recorded from ALL of the individuals in the population.[2] Fifth, we usually cannot see all of the individuals in the population so only a few are examined; those few are the sample. Finally, the summary of the individuals in the sample is the statistic.

When performing an IVPPSS, keep in mind that parameters describe populations (note that they both start with "p") and statistics describe samples (note that they both start with "s"). This can also be looked at from another perspective. A sample is an estimate of the population and a statistic is an estimate of a parameter. Thus, the statistic has to be the same summary (mean or proportion) of the sample as the parameter is of the population.

The IVPPSS process is illustrated for the following situation:

> *A University of New Hampshire graduate student (and Northland College alum) investigated habitat utilization by New England (Sylvilagus transitionalis) and Eastern (Sylvilagus floridanus) cottontail rabbits in eastern Maine in 2007. In a preliminary portion of his research he determined*

[2] Again, parameters generally cannot be computed because all of the individuals in the population can not be seen. Thus, the parameter is largely conceptual.

*the proportion of "rabbit patches" that were inhabited by New England cottontails. He examined 70 "patches" and found that 53 showed evidence of inhabitance by New England cottontails.*

- An individual is a rabbit patch in eastern Maine in 2007 (i.e., a rabbit patch is the "item" being sampled and examined).
- The variable is "evidence for New England cottontails or not (yes or no)" (i.e., the characteristic of each rabbit patch that was recorded).
- The population is ALL rabbit patches in eastern Maine in 2007.
- The parameter is the proportion of ALL rabbit patches in eastern Maine in 2007 that showed evidence for New England cottontails.[3]
- The sample is the 70 rabbit patches from eastern Maine in 2007 that were actually examined by the researcher.
- The statistic is the proportion of the 70 rabbit patches from eastern Maine in 2007 actually examined that showed evidence for New England cottontails. [In this case, the statistic would be 53/70 or 0.757.]

In some situations it may be easier to identifying the sample first. From this, and through the realization that a sample is always "of the individuals", it may be easier to identify the individual. This process is illustrated in the following example, with the items listed in the order identified rather than in the traditional IVPPSS order.

*The Duluth, MN Touristry Board is interested in the average number of raptors seen per year at Hawk Ridge.[4] To determine this value, they collected the total number of raptors seen in a sample of years from 1971-2003.*

- The sample is the 32 years between 1971-2003 at Hawk Ridge.
- An individual is a year (because a "sample of *years*" was taken) at Hawk Ridge.
- The variable recorded was the number of raptors seen in one year at Hawk Ridge.
- The population is ALL years at Hawk Ridge(this is a bit ambiguous but may be thought of as all years that Hawk Ridge has existed).
- The parameter is the average number of raptors seen per year in ALL years at Hawk Ridge.
- The statistic is the average number of raptors seen in the 1971-2003 sample of years at Hawk Ridge.

## Review Exercises

**2.1** My Dad owns 60 acres of timber (mostly Oak, Walnut, and Poplar) in Iowa. He wants to measure the mean diameter-at-breast-height (DBH) of the oak trees on his property. He measures the DBH of 75 randomly selected oak trees. Use this information to perform an IVPPSS. *Answer*

**2.2** I have a friend who wants to start a (fishing) bait store on the West end of Ashland. He wants to determine what proportion of Ashland residents who currently use the East end bait store would use a store in the West end if one existed. He sends out 5000 questionnaires and receives 2378 back from patrons of the East end store. Use this information to perform an IVPPSS. *Answer*

[3] Note that this population and parameter cannot actually be calculated but it is what the researcher wants to know.
[4] Information about Hawk Ridge is found here.

**2.3** I'm interested in developing a model to predict how many points an NBA starting basketball player scores. Therefore, I want to determine the relationship between points scored and height, speed (in the 40-yard dash), position, and minutes played. To identify this relationship I gather these data from 100 NBA starting basketball players. Use this information to perform an IVPPSS. *Answer*

**2.4** Pollsters wanted to determine the proportion of registered voters who approved of President Clinton's performance. They called 5000 randomly selected registered voters and ask 4123 of those (the rest weren't home, didn't answer, or hung up) "Do you approve of Pres. Clinton's performance?" Use this information to perform an IVPPSS. *Answer*

**2.5** You Might Be Interested To Know (YMBITK), the average level of mercury in newly-hatched goslings in the upper Midwest (MI, MN, ND, SD, WI). You obtained 20 goslings from resource agencies in each state. Use this information to perform an IVPPSS. *Answer*

**2.6** YMBITK, the proportion of NC students that think NC can become "the nation's leading environmental liberal arts college" in the next decade. You polled 124 students. Use this information to perform an IVPPSS. *Answer*

**2.7** YMBITK, the relationship between hours studied and GPA of students in the UW system (excluding UW-Madison). You interviewed 250 students from throughout the system. Use this information to perform an IVPPSS. *Answer*

**2.8** YMBITK, the average difference in salaries between the head coaches of men's and head coaches of women's basketball teams at Division I schools. You interviewed 73 head-coach pairs. Use this information to perform an IVPPSS. *Answer*

**2.9** YMBITK, the proportion of graduates from small private schools, who majored in Biology and who have been out of school for at least 5 years, that feel that statistics is an "important" course. You interviewed 1023 people. Use this information to perform an IVPPSS. *Answer*

**2.10** Scientist in Chivyrkui Bay on Lake Baikal (Owens and Pronin 2000) were interested, among other things, in determining the mean age of pike (*Esox lucius*) in the bay. They collected scales from 30 fish using gill nets and angling methods. Use this information to perform an IVPPSS. *Answer*

**2.11** The Eurasian ruffe is an exotic species of fish that is causing some alarm in fisheries biologists in the Great Lakes area (Maniak *et al.* 2000). A few of these biologists tested to see if a certain pheromone released by injured ruffe would repel other ruffe. If so, natural, or possibly synthetic, versions of this pheromone could be used to distract ruffe from areas in which they are causing damage. In their experiment, they observed ruffe held in aquaria divided into four sections. They recorded what proportion of 24 randomly-selected ruffe caught in the St. Louis River Harbor, and then held in the aquaria, left a section when the chemical was added to that section. Use this information to perform an IVPPSS. *Answer*

### 2.2.1 Sampling Variability (Revisited)

It is instructive to once again (see Module 1) consider how statistics differ among samples. Table 2.3 and Figure 2.3 show results from three more samples of n=50 fish from the Square Lake population. The means from all four samples (including the sample in Table 2.2 and Figure 2.2) were quite different from the known population mean of 98.06 mm. Similarly, all four histograms were similar in appearance but were slightly different in actual values. These results illustrate that a statistic (or sample) will only approximate the parameter (or population) and that statistics vary among samples. This **sampling variability** is one of the most important concepts in statistics and will be discussed in great detail beginning in Module 12.

Table 2.3. Summary statistics for the total length in three samples of 50 fish from the Square Lake population.

| n | mean | sd | min | Q1 | median | Q3 | max |
|---|---|---|---|---|---|---|---|
| 50 | 99.56 | 32.47 | 57 | 69 | 91 | 123 | 167 |
| 50 | 88.64 | 24.52 | 53 | 68 | 86 | 106 | 166 |
| 50 | 112.74 | 35.86 | 61 | 84 | 108 | 147 | 174 |

$\Delta$ **Sampling Variability**: The realization that no two samples are exactly alike. Thus, statistics computed from different samples will likely vary.

This example also illustrates that parameters are fixed values because populations don't change. If a population does change, then it is considered to be a different population. In the Square Lake example, if a fish is removed from the lake, then the lake would then be considered a different population of fish. Statistics, on the other hand, vary depending on the sample because each sample consists of different individuals that vary (i.e., sampling variability exists because natural variability exists).

$\diamond$ **Parameters are fixed in value, while statistics vary in value.**

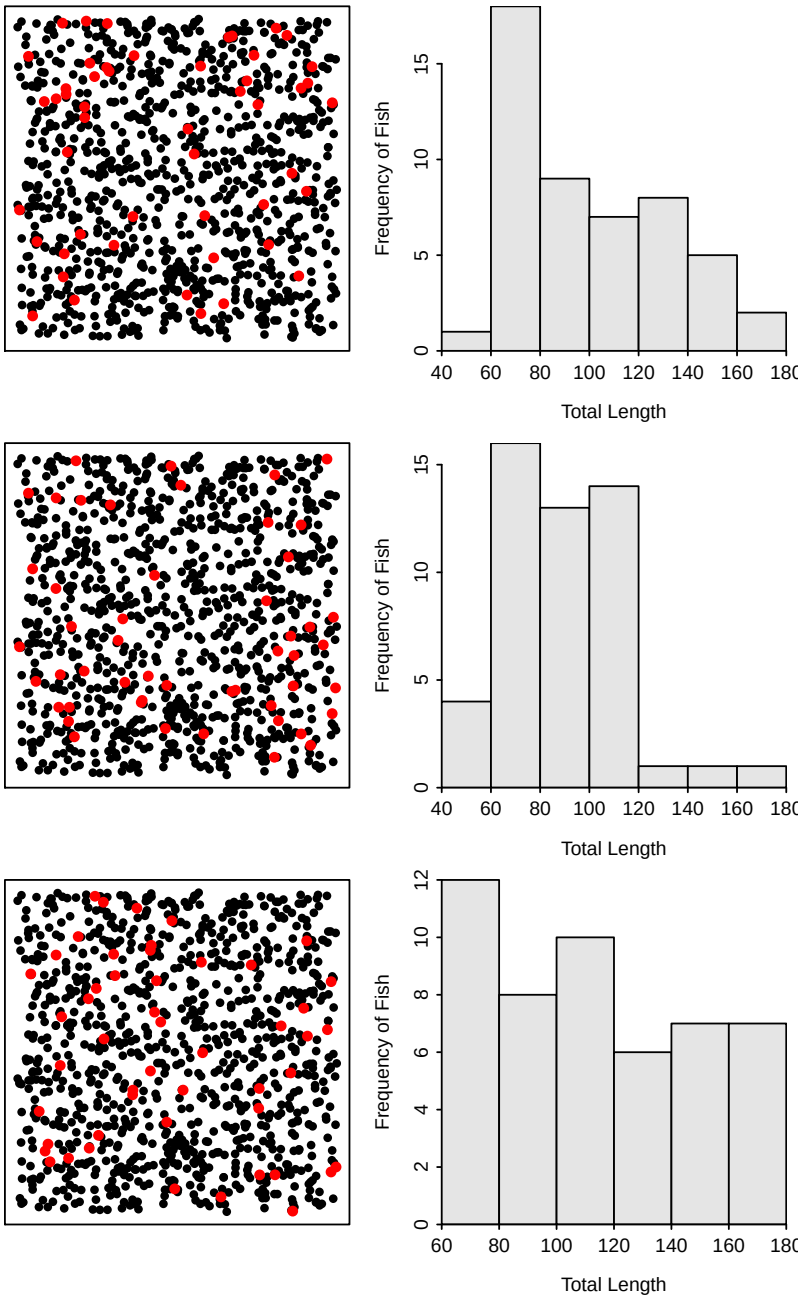


Figure 2.3. Schematic representation (**Left**) of three samples of 50 fish (i.e., red dots) from Square Lake and histograms (**Right**) of the total length of the 50 fish in each sample.

## 2.3 Variable Types

The type of statistic that can be calculated is dictated by the type of variable to be analyzed. For example, a sample mean (or average) can only be calculated for a quantitative variable (defined below). Thus, the type of that variable should be identifie immediately after performing an IVPPSS.

### 2.3.1 Variable Definitions

There are two main groups of variable types – quantitative and categorical (Figure 2.4). **Quantitative** variables are variables with numerical values for which it makes sense to do arithmetic operations (like adding or averaging). Synonyms for quantitative are measurement or numerical. **Categorical** variables are variables that record to which group or category an individual belongs. Synonyms for categorical are qualitative or attribute. Within each main type of variable are two subgroups (Figure 2.4).

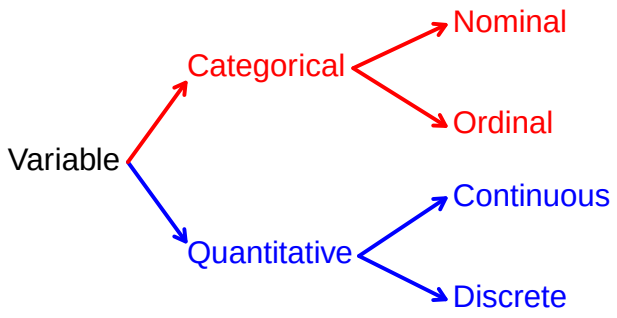


Figure 2.4. Schematic representation of the four types of variables.

The two types of quantitative variables are continuous and discrete variables. **Continuous** variables are quantitative variables that have an uncountable number of values. In other words, a potential value DOES exist between every pair of values of a continuous variable. **Discrete** variables are quantitative variables that have a countable number of values. Stated differently, a potential value DOES NOT exist between every pair of values of a discrete variable. Typically, but not always, discrete variables are counts of items.

Continuous and discrete variables are easily distinguished by determining if it is possible for a value to exist between every two values of the variable. For example, can there be between 2 and 3 ducks on a pond? No! Thus, the number of ducks is a discrete variable. Alternatively, can a duck weigh between 2 and 3 kg? Yes! Can it weigh between 2 and 2.1 kg? Yes! Can it weigh between 2 and 2.01 kg? Yes! You can see that this line of questions could continue forever; thus, duck weight is a continuous variable.

**Δ Discrete Variable**: A quantitative variable that can assume a countable number of values.

**Δ Continuous Variable**: A quantitative variable that can assume an uncountable number of values.

◇ **A quantitative variable is continuous if a possible value exists between every two values of the variable; otherwise, it is discrete.**

The two types of categorical variables are ordinal and nominal. **Ordinal** variables are categorical variables where a natural order or ranking exists among the categories. **Nominal** variables are categorical variables where no order or ranking exists among the categories.

Ordinal and nominal variables are easily distinguished by determining if the order of the categories matters. For example, suppose that a researcher recorded a subjective measure of condition (i.e., poor, average, excellent) and the species of each duck. Order matters with the condition variable – i.e., condition improves from the first (poor) to the last category (excellent) – and some reorderings of the categories would not make sense – i.e., average, poor, excellent does not make sense. Thus, condition is an ordinal variable. In contrast, species (e.g., mallard, redhead, canvasback, and wood duck) is a nominal variable because there is no inherent order among the categories (i.e., any reordering of the categories also "makes sense").

Δ **Ordinal Variable**: A categorical variable for which a natural order exists among the categories.

Δ **Nominal Variable**: A categorical variable for which a natural order DOES NOT exist among the categories.

◇ **Remember that ordinal means that an order among the categories exists (note "ord" in both ordinal and order).**

The following are some issues to consider when identifying the type of a variable:

1. The categories of a categorical variable are sometimes labeled with numbers. For example, 1="Poor", 3="Fair", and 5="Good". Don't let this fool you into calling the variable quantitative.
2. Rankings, ratings, and preferences are ordinal (categorical) variables.
3. Counts of numbers are discrete (quantitative) variables.
4. Measurements are typically continuous (quantitative) variables.
5. It does not matter how precisely quantiative variables are recorded when deciding if the variable is continuous or discrete. For example, the weight of the duck might have been recorded to the nearest kg. However, this was just a choice that was made, the actual values can be continuously finer than kg and, thus, weight is a continuous variable.
6. Categorical variables that consist of only two levels or categories will be labeled as a nominal variable (because any order of the groups makes sense). This type of variable is also often called "binomial.".
7. Do not confuse "what type of variable" (answer is one of "continuous", "discrete", "nominal", or "ordinal") with "what type of variability" (answer is "natural" or "sampling") questions.

◇ **"What type of variable is ...?" is a different question than "what type of variability is ...?" Be careful to note the word difference (i.e., "variable" versus "variability" when answering these questions.**

◇ **The precision to which a quantitative variable was recorded does not determine whether it is continuous or discrete. How precisely the variable COULD have been recorded is the important consideration.**

## Review Exercises

**2.12** What type of variable is the number of ducks found at the "Hot Pond" every morning? *Answer*

**2.13** What type of variable is the genotype (AA, Aa, aa) of a particular species of sunflower? *Answer*

**2.14** What type of variable is the length of petals on individual flowers? *Answer*

**2.15** What type of variable is the number of seeds produced by an individual sunflower? *Answer*

**2.16** What type of variable is the "quality" of the seeds produced by an individual plant ("quality" is recorded as 1=poor, 2=low, 3=good, and 4=excellent)? *Answer*

**2.17** What type of variable is student rankings ("Excellent", "Very Good", "Good", "Fair", "Poor") of a professor's abilities? *Answer*

**2.18** What type of variable is whether an account is valid or invalid? *Answer*

**2.19** What type of variable is the number of defects produced by a machine? *Answer*

**2.20** What type of variable is the ounces of cola in a sample of 100 bottles? *Answer*

**2.21** What type of variable is the sex of fish collected from a lake? *Answer*

**2.22** What type of variable is the number of legs on frogs collected in Bayfield County? *Answer*

**2.23** What type of variable is the frequency (mhz) of a bullfrog's "croak"? *Answer*

**2.24** What type of variable is the number of incorporated towns in a county? *Answer*

**2.25** What type of variable is the qualitative size of least weasels (small, medium, large)? *Answer*