Types of data in biology

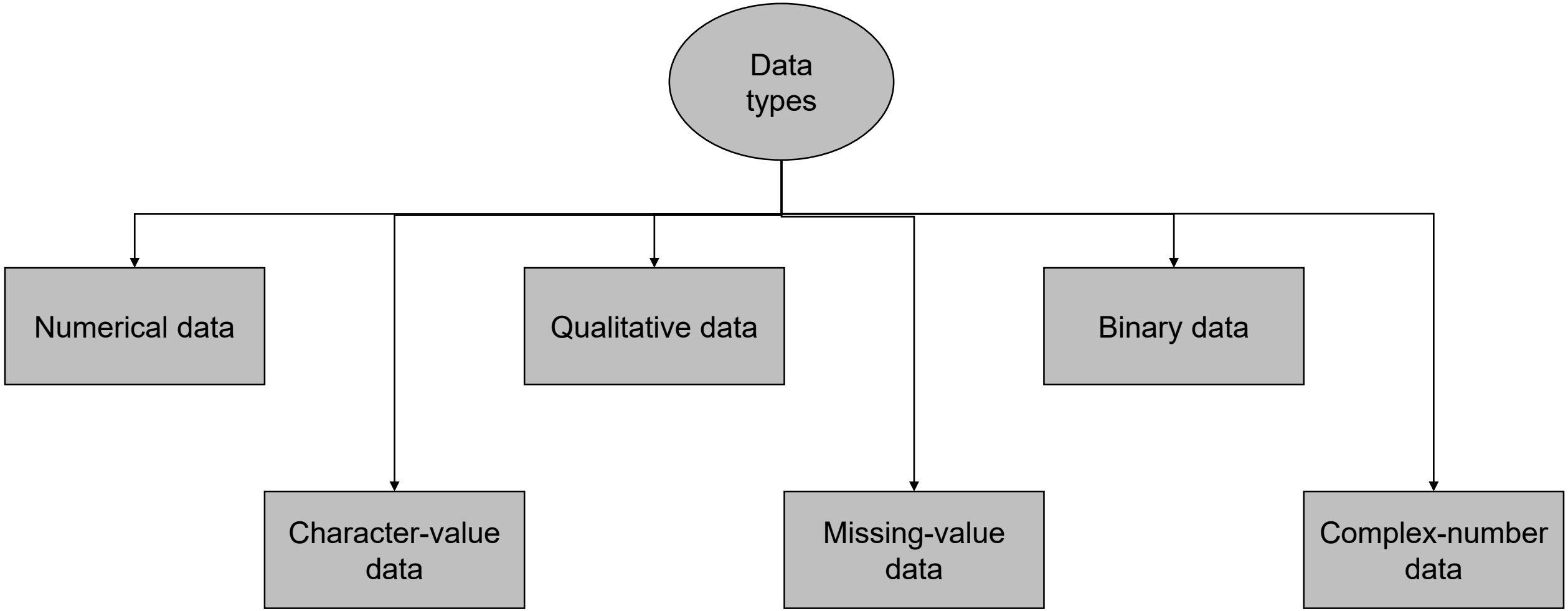
Types of Data

In this topic, we will discuss the different kinds of biological data that are typically collected by most biologists. Although there may be exhaustive list of various types of data, few related and/or frequently encountered data types will be discussed.

1. Data classes

In biology, one may encounter many kinds of data, and depending on which kind, the type of statistical analysis are used.

Types of Data



Types of Data

1.1. Numerical data

Numerical data are quantitative in nature. They represent things that can be objectively counted, measured or calculated. It represents values that can be measured and put into a logical order.

For example:

Height (in cm, women(men)): 50(50), 75(75), 85(85), 95(95), 100(102), 108(109), 115(115), 120(121), 128(128), 133(133), 138(138), 144(144), 150(150), 157(156), 159(164), 160(170), 162(173), 162(175), 163(176), 163(176), 163(177).

Weight (in kg, women(men) of 10 years and above): 35(35), 38(38), 40(41), 43(44), 45(47), 48(50), 50(53), 53(56), 55(59), 58(62), 60(65), 63(68), 65(71), 68(74), 70(77), 73(80), 75(83), 78(86), 80(89), 83(92), 85(95), 86(98), 90(101).

Age: 0, 10, 20, 30, 40, 50, 55, 60, 65, 67, 70, 73, 75, 78, 80, 82, 85, 88, 100.

Number of classes attended: 0, 3, 5, 7, 11, 14, 17, 20.

Protein length: 50, 60, 77, 150, 230, 310, 400, 510.

Alignment scores, query coverage, sequence identity, e-values, etc.

Types of Data

1.1.1. Discrete data

Integer data (discrete numbers or whole numbers), such as counts. Integer data usually answer the question, “how many?”

For example:
Number of children in a family: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.

Number of classes attended: 0, 3, 5, 7, 11, 14, 17, 20.

Number of positive amino acids in a protein: 0, 1, 2, 4, 5, 6, 10, 17, 23, 45.

Number of mutations in a gene over time: 0, 1, 2, 3, 4, 5, 6, 10.

Types of Data

1.1.2. Continuous data

These usually represent measured ‘things,’ such as something’s heat content (temperature, measured in degrees Celsius) or distance (measured in meters or similar), etc.

They can be rational numbers including integers and fractions, but typically they have an infinite number of ‘steps’ that depends on rounding (they can even be rounded to whole integers) or considerations such as measurement precision and accuracy.

Often, continuous data have upper and lower bounds that depend on the characteristics of the phenomenon being studied or the measurement being taken.

The kinds of summaries that lend themselves to continuous data are: Frequency distributions, Relative frequency distributions, Cumulative frequency distributions, Bar graphs, Box plots, Scatter plots.

For example:
Protein weight (in kDa): 5.0, 5.4, 6.5, 7.3, 8.0, 14.5, 20.3, etc.
Query coverage (%): 40.3, 60, 63.2, 71.4, 80, 83.5, 91.4, etc.
Sequence identity (%): 10, 14.5, 24.6, 30.3, 40.4, 50, 61.5, etc.
e-values: 0.00001, 0.0003, 0.0034, 0.0005, 0.007, etc.

Types of Data

1.1.2. Continuous data

Continuous data is further divided into two categories: Interval and Ratio.

Interval data – interval data type refers to data that can be measured only along a scale at equal distance from each other. It is always in the form of numerical values but cannot be multiplied or divided. They can be added or subtracted, but there is no true zero on an interval scale.

In research, interval data is essential because it can support most statistical tests.

For Example:
Body temperature can be measured in degree Celsius and degree Fahrenheit and neither of them can be 0.

Ratio data – unlike interval data, ratio data has zero point. Being similar to interval data, zero point is the only difference they have.

For Example:
In the body temperature, the zero point temperature can be measured in Kelvin.
This is used generally in Phylogenetic analysis.

Types of Data

Characteristics of Numerical Data

- Numerical data has two **categories**: discrete data and continuous data, where the latter is further classified into interval data and ratio data.
- Numerical data is **quantitative** in nature as it takes quantitative values for data.
- Numerical data allows us to perform **arithmetic operations** on them like add and subtract. It can also use any statistical analysis calculations.
- It can be **estimated and enumerated**. When the numerical data is precise, it is enumerated, or else it is estimated.
- The **interval difference** between each numerical data when put on a number scale, comes out to be equal. A clock, a thermometer are perfect examples for this.
- Numerical data can be **analyzed using two methods**: descriptive and inferential analysis.
- Numerical data makes it **easy to be visualized**. It uses data visualization techniques like scatter plot, dot plot, stem and leaf graphs, box plots, ogive graphs, histograms, etc.

Types of Data

1.1.3. Dates

Dates are a special class of continuous data, and there are many different representations of the date classes. This is a complex group of data, and we will not cover much of it in this course.

Types of Data

1.2. Qualitative data

Qualitative data may be well-defined categories or they may be subjective, and generally include descriptive words for classes (e.g. mineral, animal, plant) or rankings (e.g. good, better, best).

1.2.1. Categorical data

Categorical data are qualitative characteristics of individuals that do not have magnitude on a numerical scale. Because there are categories, the number of members belonging to each of the categories can be counted.

For example:

- There are three red flowers, 66 purple flowers and 13 yellow flowers.
- Survival (alive or dead).
- Sex chromosome genotype (e.g., XX, XY, XO, XXY, or XYY).
- Method of disease transmission (e.g., water, air, animal vector, or direct contact).
- Predominant language spoken (e.g., English, Mandarin, Spanish, Indonesian, etc.).
- Life stage (e.g., egg, larva, juvenile, sub adult, or adult).
- Snakebite severity score (e.g., minimal severity, moderate severity, or very severe).
- Class size (e.g., small, medium, or large).
- Homology vs similarity relationship (<, =, >)?

Types of Data

1.2.1. Categorical data

The categories cannot be ranked relative to each other. In the example above, no value judgement can be assigned to the different colors.

It is not better to be red than it is to be purple. There are just fewer red flowers than purple ones. Contrast to this is another kind of categorical data called ‘ordinal data’.

A categorical variable is *nominal* if the different categories have no inherent order. Nominal means “name.” Sex chromosome genotype, method of disease transmission, and predominant language spoken are nominal variables.

In contrast, the values of an *ordinal* categorical variable can be ordered. Unlike numerical data, the magnitude of the difference between consecutive values is not known. Ordinal means “having an order.” Life stage, snakebite severity score, and size class are ordinal categorical variables.

The kinds of summaries that lend themselves to categorical data are: Frequency distributions, Relative frequency distributions, Bar graphs, Pie graphs, Category statistics.

Types of Data

1.2.1.1 Nominal data

Nominal data is qualitative data used to name or label variables without providing numeric values. It is the most straightforward type of measurement scale. Nominal variables are labeled into categories that do not overlap. Unlike other data types, nominal data cannot be ordered or measured; it does not have equal spacing between values or a true zero value.

For example:

- For the nominal variable of preferred mode of transportation, you may have the categories of car, bus, train, tram or bicycle.
- Drug categories: antibiotics, antivirals, antifungals, etc.

Characteristics of Nominal Data

- Nominal data are categorical, the categories being mutually exclusive without any overlap.
- The categories of nominal data are purely descriptive, that is, they do not possess any quantitative or numeric value. Nominal data can never be quantified
- Nominal data cannot be put into any definite order or hierarchy. None of the categories can be greater than or worth more than one another.
- The mean of nominal data cannot be calculated even if the data is arranged in alphabetical order.
- The mode is the only measure of central tendency for nominal data.
- In most cases, nominal data is alphabetical.

Types of Data

1.2.1.2 Ordinal data

This is a type of categorical data where the classes are ordered (a synonym is “ranked”), typically from low to high (or *vice versa*), but where the magnitude between the ordered classes cannot be precisely measured or quantified.

In other words, the difference between them is somewhat subjective (i.e. it is qualitative rather than quantitative). These data are on an ordinal scale.

The data may be entered as descriptive character strings (i.e. as words), or they may have been translated to an ordered vector of integers; for example, “1” for terrible, “2” for so-so, “3” for average, “4” for good and “5” for brilliant.

Irrespective of how the data are present in the data frame, computationally (for some calculations) they are treated as an ordered sequence of integers, but they are simultaneously treated as categories (say, where the number of responses that report “so-so” can be counted).

Ordinal data usually answer questions such as, “how many categories can the phenomenon be divided into, and how does each category rank with respect to the others?”.

Types of Data

1.3. Binary data

Right or wrong? True or false? Accept or reject? Black or white? Positive or negative? Good or bad? You get the idea… In other words, these are observations or responses that can take only one of two mutually exclusive outcomes.

For example:
Whether two gene are homologous to each other?
Whether drug is effective or not?
Whether a system is present or not?

1.4. Character-value data

As the name implies, these are not numbers. Rather, they are human words that have found their way. In biology we most commonly encounter character values when we have a list of things, such as sites or species. These values will often be used as categorical or ordinal data.

Types of Data

1.5. Missing values

Unfortunately, one of the most reliable aspects of any biological dataset is that it will contain some missing data. But how can something contain missing data?

One could be forgiven for assuming that if the data are missing, then they obviously aren't contained in the dataset.

To better understand this concept we must think back to the principles of tidy data. Every observation must be in a row, and every column in that row must contain a value.

The combination of multiple observations then makes up our matrix of data. Because data are therefore presented in a two-dimensional format, any missing values from an observation will need to have an empty place-holder to ensure the consistency of the matrix.

These are what we are referring to when we speak of "missing values."

Types of Data

1.6. Complex numbers

Experiments in biology involving waves and motions generally use complex numbers to solve the equations such as Fourier Transforms, spectroscopy, etc.

Thank You