1.2: Types of Biological Variables

Learning Objectives

- To identify the types of variables in an experiment in order to choose the correct method of analysis.

Introduction

One of the first steps in deciding which statistical test to use is determining what kinds of variables you have. When you know what the relevant variables are, what kind of variables they are, and what your null and alternative hypotheses are, it's usually pretty easy to figure out which test you should use. I classify variables into three types: measurement variables, nominal variables, and ranked variables. You'll see other names for these variable types and other ways of classifying variables in other statistics references, so try not to get confused.

You'll analyze similar experiments, with similar null and alternative hypotheses, completely differently depending on which of these three variable types are involved. For example, let's say you've measured *variable X* in a sample of 56 male and 67 female isopods (*Armadillidium vulgare*, commonly known as pillbugs or roly-polies), and your null hypothesis is "Male and female *A. vulgare* have the same values of variable X."

Fig. 1.2.1 Isopod crustacean (pillbug or roly-poly), Armadillidium vulgare

If variable X is width of the head in millimeters, it's a measurement variable, and you'd compare head width in males and females with a two-sample *t*–test or a one-way analysis of variance (anova). If variable X is a genotype (such as AA, Aa, or aa), it's a nominal variable, and you'd compare the genotype frequencies in males and females with a Fisher's exact test. If you shake the isopods until they roll up into little balls, then record which is the first isopod to unroll, the second to unroll, etc., it's a ranked variable and you'd compare unrolling time in males and females with a Kruskal–Wallis test.

Types of Variables

There are three main types of variables:

- Measurement variables, which are expressed as numbers (such as $3.7mm$)
- Nominal variables, which are expressed as names (such as "female")
- Ranked variables, which are expressed as positions (such as "third")

Measurement variables

Measurement variables are, as the name implies, things you can measure. An individual observation of a measurement variable is always a number. Examples include length, weight, pH, and bone density. Other names for them include "numeric" or "quantitative" variables.

Some authors divide measurement variables into two types. One type is continuous variables, such as length of an isopod's antenna, which in theory have an infinite number of possible values. The other is discrete (or meristic) variables, which only have whole number values; these are things you count, such as the number of spines on an isopod's antenna. The mathematical theories underlying statistical tests involving measurement variables assume that the variables are continuous. Luckily, these statistical tests work well on discrete measurement variables, so you usually don't need to worry about the difference between continuous and discrete measurement variables. The only exception would be if you have a very small number of possible values of a discrete variable, in which case you might want to treat it as a nominal variable instead.

When you have a measurement variable with a small number of values, it may not be clear whether it should be considered a measurement or a nominal variable. For example, let's say your isopods have 20 to 55 spines on their left antenna, and you want to

know whether the average number of spines on the left antenna is different between males and females. You should consider spine number to be a measurement variable and analyze the data using a two-sample *t*–test or a one-way anova. If there are only two different spine numbers—some isopods have 32 spines, and some have 33—you should treat spine number as a nominal variable, with the values "32" and "33" and compare the proportions of isopods with 32 or 33 spines in males and females using a Fisher's exact test of independence (or chi-square or *G*–test of independence, if your sample size is really big). The same is true for laboratory experiments; if you give your isopods food with 15 different mannose concentrations and then measure their growth rate, mannose concentration would be a measurement variable; if you give some isopods food with $5mM$ mannose, and the rest of the isopods get $25mM$ mannose, then mannose concentration would be a nominal variable.

But what if you design an experiment with three concentrations of mannose, or five, or seven? There is no rigid rule, and how you treat the variable will depend in part on your null and alternative hypotheses. If your alternative hypothesis is "different values of mannose have different rates of isopod growth," you could treat mannose concentration as a nominal variable. Even if there's some weird pattern of high growth on zero mannose, low growth on small amounts, high growth on intermediate amounts, and low growth on high amounts of mannose, a one-way anova could give a significant result. If your alternative hypothesis is "isopods grow faster with more mannose," it would be better to treat mannose concentration as a measurement variable, so you can do a regression.

The following rule of thumb can be used:

- a measurement variable with only two values should be treated as a nominal variable
- a measurement variable with six or more values should be treated as a measurement variable
- a measurement variable with three, four or five values does not exist

Of course, in the real world there are experiments with three, four or five values of a measurement variable. Simulation studies show that analyzing such *dependent* variables with the methods used for measurement variables works well (Fagerland et al. 2011). I am not aware of any research on the effect of treating *independent* variables with small numbers of values as measurement or nominal. Your decision about how to treat your variable will depend in part on your biological question. You may be able to avoid the ambiguity when you design the experiment—if you want to know whether a dependent variable is related to an independent variable that could be measurement, it's a good idea to have at least six values of the independent variable.

Something that could be measured is a measurement variable, even when you set the values. For example, if you grow isopods with one batch of food containing $10mM$ mannose, another batch of food with $20mM$ mannose, another batch with $30mM$ mannose, etc. up to $100mM$ mannose, the different mannose concentrations are a measurement variable, even though you made the food and set the mannose concentration yourself.

Be careful when you count something, as it is sometimes a nominal variable and sometimes a measurement variable. For example, the number of bacteria colonies on a plate is a measurement variable; you count the number of colonies, and there are 87 colonies on one plate, 92 on another plate, etc. Each plate would have one data point, the number of colonies; that's a number, so it's a measurement variable. However, if the plate has red and white bacteria colonies and you count the number of each, it is a nominal variable. Now, each colony is a separate data point with one of two values of the variable, "red" or "white"; because that's a word, not a number, it's a nominal variable. In this case, you might summarize the nominal data with a number (the percentage of colonies that are red), but the underlying data are still nominal.

Ratios

Sometimes you can simplify your statistical analysis by taking the ratio of two measurement variables. For example, if you want to know whether male isopods have bigger heads, relative to body size, than female isopods, you could take the ratio of head width to body length for each isopod, and compare the mean ratios of males and females using a two-sample *t*–test. However, this assumes that the ratio is the same for different body sizes. We know that's not true for humans—the head size/body size ratio in babies is freakishly large, compared to adults—so you should look at the regression of head width on body length and make sure the regression line goes pretty close to the origin, as a straight regression line through the origin means the ratios stay the same for different values of the X variable. If the regression line doesn't go near the origin, it would be better to keep the two variables separate instead of calculating a ratio, and compare the regression line of head width on body length in males to that in females using an analysis of covariance.

Circular variables

One special kind of measurement variable is a circular variable. These have the property that the highest value and the lowest value are right next to each other; often, the zero point is completely arbitrary. The most common circular variables in biology are time of day, time of year, and compass direction. If you measure time of year in days, Day 1 could be January 1, or the spring equinox, or your birthday; whichever day you pick, Day 1 is adjacent to Day 2 on one side and Day 365 on the other.

If you are only considering part of the circle, a circular variable becomes a regular measurement variable. For example, if you're doing a polynomial regression of bear attacks vs. time of the year in Yellowstone National Park, you could treat "month" as a measurement variable, with March as 1 and November as 9; you wouldn't have to worry that February (month 12) is next to March, because bears are hibernating in December through February, and you would ignore those three months.

However, if your variable really is circular, there are special, very obscure statistical tests designed just for circular data; chapters 26 and 27 in Zar (1999) are a good place to start.

Nominal variables

Nominal variables classify observations into discrete categories. Examples of nominal variables include sex (the possible values are male or female), genotype (values are AA, Aa, or aa), or ankle condition (values are normal, sprained, torn ligament, or broken). A good rule of thumb is that an individual observation of a nominal variable can be expressed as a word, not a number. If you have just two values of what would normally be a measurement variable, it's nominal instead: think of it as "present" vs. "absent" or "low" vs. "high." Nominal variables are often used to divide individuals up into categories, so that other variables may be compared among the categories. In the comparison of head width in male vs. female isopods, the isopods are classified by sex, a nominal variable, and the measurement variable head width is compared between the sexes.

Nominal variables are also called categorical, discrete, qualitative, or attribute variables. "Categorical" is a more common name than "nominal," but some authors use "categorical" to include both what I'm calling "nominal" and what I'm calling "ranked," while other authors use "categorical" just for what I'm calling nominal variables. I'll stick with "nominal" to avoid this ambiguity.

Nominal variables are often summarized as proportions or percentages. For example, if you count the number of male and female *A. vulgare* in a sample from Newark and a sample from Baltimore, you might say that 52.3% of the isopods in Newark and 62.1% of the isopods in Baltimore are female. These percentages may look like a measurement variable, but they really represent a nominal variable, sex. You determined the value of the nominal variable (male or female) on 65 isopods from Newark, of which 34 were female and 31 were male. You might plot 52.3% on a graph as a simple way of summarizing the data, but you should use the 34 female and 31 male numbers in all statistical tests.

It may help to understand the difference between measurement and nominal variables if you imagine recording each observation in a lab notebook. If you are measuring head widths of isopods, an individual observation might be "$3.41mm$." That is clearly a measurement variable. An individual observation of sex might be "female," which clearly is a nominal variable. Even if you don't record the sex of each isopod individually, but just counted the number of males and females and wrote those two numbers down, the underlying variable is a series of observations of "male" and "female."

Ranked variables

Ranked variables, also called ordinal variables, are those for which the individual observations can be put in order from smallest to largest, even though the exact values are unknown. If you shake a bunch of *A. vulgare* up, they roll into balls, then after a little while start to unroll and walk around. If you wanted to know whether males and females unrolled at the same time, but your stopwatch was broken, you could pick up the first isopod to unroll and put it in a vial marked "first," pick up the second to unroll and put it in a vial marked "second," and so on, then sex the isopods after they've all unrolled. You wouldn't have the exact time that each isopod stayed rolled up (that would be a measurement variable), but you would have the isopods in order from first to unroll to last to unroll, which is a ranked variable. While a nominal variable is recorded as a word (such as "male") and a measurement variable is recorded as a number (such as "4.53"), a ranked variable can be recorded as a rank (such as "seventh").

You could do a lifetime of biology and never use a true ranked variable. When I write an exam question involving ranked variables, it's usually some ridiculous scenario like "Imagine you're on a desert island with no ruler, and you want to do statistics on the size of coconuts. You line them up from smallest to largest...." For a homework assignment, I ask students to pick a paper from their favorite biological journal and identify all the variables, and anyone who finds a ranked variable gets a donut; I've had to buy four donuts in 13 years. The only common biological ranked variables I can think of are dominance hierarchies in behavioral biology

(see the dog example on the Kruskal-Wallis page) and developmental stages, such as the different instars that molting insects pass through.

The main reason that ranked variables are important is that the statistical tests designed for ranked variables (called "non-parametric tests") make fewer assumptions about the data than the statistical tests designed for measurement variables. Thus the most common use of ranked variables involves converting a measurement variable to ranks, then analyzing it using a non-parametric test. For example, let's say you recorded the time that each isopod stayed rolled up, and that most of them unrolled after one or two minutes. Two isopods, who happened to be male, stayed rolled up for 30 minutes. If you analyzed the data using a test designed for a measurement variable, those two sleepy isopods would cause the average time for males to be much greater than for females, and the difference might look statistically significant. When converted to ranks and analyzed using a non-parametric test, the last and next-to-last isopods would have much less influence on the overall result, and you would be less likely to get a misleadingly "significant" result if there really isn't a difference between males and females.

Some variables are impossible to measure objectively with instruments, so people are asked to give a subjective rating. For example, pain is often measured by asking a person to put a mark on a $10cm$ scale, where $0cm$ is "no pain" and $10cm$ is "worst possible pain." This is *not* a ranked variable; it is a measurement variable, even though the "measuring" is done by the person's brain. For the purpose of statistics, the important thing is that it is measured on an "interval scale"; ideally, the difference between pain rated 2 and 3 is the same as the difference between pain rated 7 and 8. Pain would be a ranked variable if the pains at different times were compared with each other; for example, if someone kept a pain diary and then at the end of the week said "Tuesday was the worst pain, Thursday was second worst, Wednesday was third, etc...." These rankings are not an interval scale; the difference between Tuesday and Thursday may be much bigger, or much smaller, than the difference between Thursday and Wednesday.

Just like with measurement variables, if there are a very small number of possible values for a ranked variable, it would be better to treat it as a nominal variable. For example, if you make a honeybee sting people on one arm and a yellowjacket sting people on the other arm, then ask them "Was the honeybee sting the most painful or the second most painful?", you are asking them for the rank of each sting. But you should treat the data as a nominal variable, one which has three values ("honeybee is worse" or "yellowjacket is worse" or "subject is so mad at your stupid, painful experiment that they refuse to answer").

Categorizing

It is possible to convert a measurement variable to a nominal variable, dividing individuals up into a two or more classes based on ranges of the variable. For example, if you are studying the relationship between levels of HDL (the "good cholesterol") and blood pressure, you could measure the HDL level, then divide people into two groups, "low HDL" (less than $40mg/dl$) and "normal HDL" (40 or more mg/dl) and compare the mean blood pressures of the two groups, using a nice simple two-sample *t*–test.

Converting measurement variables to nominal variables ("dichotomizing" if you split into two groups, "categorizing" in general) is common in epidemiology, psychology, and some other fields. However, there are several problems with categorizing measurement variables (MacCallum et al. 2002). One problem is that you'd be discarding a lot of information; in our blood pressure example, you'd be lumping together everyone with HDL from 0 to $39mg/dl$ into one group. This reduces your statistical power, decreasing your chances of finding a relationship between the two variables if there really is one. Another problem is that it would be easy to consciously or subconsciously choose the dividing line ("cutpoint") between low and normal HDL that gave an "interesting" result. For example, if you did the experiment thinking that low HDL caused high blood pressure, and a couple of people with HDL between 40 and 45 happened to have high blood pressure, you might put the dividing line between low and normal at $45mg/dl$. This would be cheating, because it would increase the chance of getting a "significant" difference if there really isn't one.

To illustrate the problem with categorizing, let's say you wanted to know whether tall basketball players weigh more than short players. Here's data for the 2012-2013 men's basketball team at Morgan State University:

Height (inches)	Weight (pounds)
69	180
72	185
74	170
74	190

74	220
76	200
77	190
77	225
78	215
78	225
80	210
81	208
81	220
86	270

Table 1.2.1 2012-2013 men's basketball team at Morgan State University

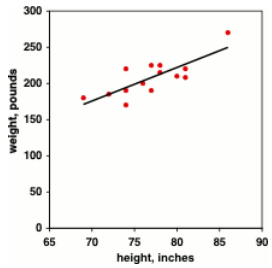

Fig. 1.2.2 Height and weight of the Morgan State University men's basketball players

If you keep both variables as measurement variables and analyze using linear regression, you get a P value of 0.0007; the relationship is highly significant. Tall basketball players really are heavier, as is obvious from the graph. However, if you divide the heights into two categories, "short" (77 inches or less) and "tall" (more than 77 inches) and compare the mean weights of the two groups using a two-sample *t*–test, the P value is 0.043, which is barely significant at the usual $P < 0.05$ level. And if you also divide the weights into two categories, "light" (210 pounds and less) and "heavy" (greater than 210 pounds), you get 6 who are short and light, 2 who are short and heavy, 2 who are tall and light, and 4 who are tall and heavy. The proportion of short people who are heavy is *not* significantly different from the proportion of tall people who are heavy, when analyzed using Fisher's exact test ($P = 0.28$). So by categorizing both measurement variables, you have made an obvious, highly significant relationship between height and weight become completely non-significant. This is not a good thing. I think it's better for most biological experiments if you don't categorize.

Likert items

Social scientists like to use Likert items. They'll present a statement like:

"It's important for all biologists to learn statistics"

and ask people to choose

- 1=Strongly Disagree
- 2=Disagree
- 3=Neither Agree nor Disagree
- 4=Agree,
- 5=Strongly Agree.

Sometimes they use seven values instead of five, by adding "Very Strongly Disagree" and "Very Strongly Agree"; and sometimes people are asked to rate their strength of agreement on a 9 or 11-point scale. Similar questions may have answers such as

- 1=Never
- 2=Rarely

- 3=Sometimes
- 4=Often
- 5=Always

Strictly speaking, a Likert scale is the result of adding together the scores on several Likert items. Often, however, a single Likert item is called a Likert scale.

There is a lot of controversy about how to analyze a Likert item. One option is to treat it as a nominal variable with five (or seven, or however many) items. The data would then be summarized by the proportion of people giving each answer, and analyzed using chi-square or *G*–tests. However, this ignores the fact that the values go in order from least agreement to most, which is pretty important information. The other options are to treat it as a ranked variable or a measurement variable.

Treating a Likert item as a measurement variable lets you summarize the data using a mean and standard deviation, and analyze the data using the familiar parametric tests such as anova and regression. One argument against treating a Likert item as a measurement variable is that the data have a small number of values that are unlikely to be normally distributed, but the statistical tests used on measurement variables are not very sensitive to deviations from normality, and simulations have shown that tests for measurement variables work well even with small numbers of values (Fagerland et al. 2011).

A bigger issue is that the answers on a Likert item are just crude subdivisions of some underlying measure of feeling, and the difference between "Strongly Disagree" and "Disagree" may not be the same size as the difference between "Disagree" and "Neither Agree nor Disagree"; in other words, the responses are not a true "interval" variable. As an analogy, imagine you asked a bunch of college students:

"How much TV they watch in a typical week"

and you give them the choices of

- 0=None
- 1=A Little
- 2=A Moderate Amount
- 3=A Lot
- 4=Too Much

If the people who said "A Little" watch one or two hours a week, the people who said "A Moderate Amount" watch three to nine hours a week, and the people who said "A Lot" watch 10 to 20 hours a week, then the difference between "None" and "A Little" is a lot smaller than the difference between "A Moderate Amount" and "A Lot." That would make your $0-4$ point scale not be an interval variable. If your data actually were in hours, then the difference between 0 hours and 1 hour is the same size as the difference between 19 hours and 20 hours; "hours" would be an interval variable.

Personally, I don't see how treating values of a Likert item as a measurement variable will cause any statistical problems. It is, in essence, a data transformation: applying a mathematical function to one variable to come up with a new variable. In chemistry, pH is the base-10 log of the reciprocal of the hydrogen activity, so the difference in hydrogen activity between a pH 5 and pH 6 solution is much bigger than the difference between pH 8 and pH 9. But I don't think anyone would object to treating pH as a measurement variable. Converting $25-44$ on some underlying "agreeicity index" to "2" and converting $45-54$ to "3" doesn't seem much different from converting hydrogen activity to pH, or micropascals of sound to decibels, or squaring a person's height to calculate body mass index.

The impression I get, from briefly glancing at the literature, is that many of the people who use Likert items in their research treat them as measurement variables, while most statisticians think this is outrageously incorrect. I think treating them as measurement variables has several advantages, but you should carefully consider the practice in your particular field; it's always better if you're speaking the same statistical language as your peers. Because there is disagreement, you should include the number of people giving each response in your publications; this will provide all the information that other researchers need to analyze your data using the technique they prefer.

All of the above applies to statistics done on a single Likert item. The usual practice is to add together a bunch of Likert items into a Likert scale; a political scientist might add the scores on Likert questions about abortion, gun control, taxes, the environment, etc. and come up with a 100-point liberal vs. conservative scale. Once a number of Likert items are added together to make a Likert scale, there seems to be less objection to treating the sum as a measurement variable; even some statisticians are okay with that.

Independent and dependent variables

Another way to classify variables is as independent or dependent variables. An independent variable (also known as a predictor, explanatory, or exposure variable) is a variable that you think may cause a change in a dependent variable (also known as an outcome or response variable). For example, if you grow isopods with 10 different mannose concentrations in their food and measure their growth rate, the mannose concentration is an independent variable and the growth rate is a dependent variable, because you think that different mannose concentrations may cause different growth rates. Any of the three variable types (measurement, nominal or ranked) can be either independent or dependent. For example, if you want to know whether sex affects body temperature in mice, sex would be an independent variable and temperature would be a dependent variable. If you wanted to know whether the incubation temperature of eggs affects sex in turtles, temperature would be the independent variable and sex would be the dependent variable.

As you'll see in the descriptions of particular statistical tests, sometimes it is important to decide which is the independent and which is the dependent variable; it will determine whether you should analyze your data with a two-sample *t*–test or simple logistic regression, for example. Other times you don't need to decide whether a variable is independent or dependent. For example, if you measure the nitrogen content of soil and the density of dandelion plants, you might think that nitrogen content is an independent variable and dandelion density is a dependent variable; you'd be thinking that nitrogen content might affect where dandelion plants live. But maybe dandelions use a lot of nitrogen from the soil, so it's dandelion density that should be the independent variable. Or maybe some third variable that you didn't measure, such as moisture content, affects both nitrogen content and dandelion density. For your initial experiment, which you would analyze using correlation, you wouldn't need to classify nitrogen content or dandelion density as independent or dependent. If you found an association between the two variables, you would probably want to follow up with experiments in which you manipulated nitrogen content (making it an independent variable) and observed dandelion density (making it a dependent variable), and other experiments in which you manipulated dandelion density (making it an independent variable) and observed the change in nitrogen content (making it the dependent variable).

References

1. Fagerland, M. W., L. Sandvik, and P. Mowinckel. 2011. Parametric methods outperformed non-parametric methods in comparisons of discrete numerical variables. BMC Medical Research Methodology 11: 44.
2. MacCallum, R. C., S. B. Zhang, K. J. Preacher, and D. D. Rucker. 2002. On the practice of dichotomization of quantitative variables. Psychological Methods 7: 19-40.
3. Zar, J.H. 1999. Biostatistical analysis. 4th edition. Prentice Hall, Upper Saddle River, NJ.
4. Picture of isopod from Australian Insect Common Names