Advanced Quantitative Methods for Environmental Scientists

Dale H Easley
http://daleeasley.com

April 2020

Note: This book is best read in landscape mode.

Table of Contents

Prologue: Why Statistics? 1

I Gathering your data 4

1 Sampling methods 5

2 What does your data represent? 9

II Getting to know your data 13

3 Your eyes see patterns, so plot your data 14

4 Bins, frequencies, and histograms 19

III Comparing two data sets of the same variable 22

5 Statistical tests 23

6 F-tests 28

7 T-tests 30

IV Analyzing two variables 33

8 Correlation 34

9 Regression 38

10 χ^2-test of goodness-of-fit 39

V Multiple variables 42

11 Spatial autocorrelation 43

12 Correlation matrices 48

VI Some further reading 49

List of Figures

1 1981 map of hurricane risk. . . . 2
2 Gridded, random, and stratified-random sampling locations. . . . 7
3 Waterbuck in Kenya. . . . 15
4 Histograms of two different sets of data with the same mean and standard deviation. . . . 16
5 Example histograms of grades in a class of 40 students. . . . 21
6 Ink blot number 4. . . . 23
7 The possible outcomes of rolling two dice. . . . 25
8 Incorrect linear-regression fit. . . . 27
9 Two population distributions comparing means and variance. . . . 29
10 The r^2 value for the fit of a line to data. . . . 36
11 Positive and negative r value for the fit of a line to data. . . . 37
12 Coastal wetlands in the tidal zone covered at least part of the time with water. . . . 40
13 Contours of data with (A) much persistence in space (high spatial autocorrelation) and (B) little persistence in space (low spatial autocorrelation). . . . 46
14 Example sample variogram showing the main components of nuggest-effect, range, and sill. The x-axis is lag, and the y-axis is quared differences. . . . 47

Statistics is a lively and fascinating subject,
but studying it is too often excruciatingly dull.

Wallis and Roberts, 1956, *Statistics, a New Approach*

I'm writing this during the height of the COVID-19 pandemic, not sure yet what comes next. But I'm certain that understanding statistics is more valuable than ever—what's the infection rate? the death rate? the recovery rate? Take a look at the article [here.] We need good data and good analysis to make decisions about reopening the society. If we open too soon, additional people will die who might have survived. If we delay unnecessarily, people may also suffer additionally.

We haven't been very good so far in pulling together. A lack of numerical literacy drives some of those conflicts. People ignore data, even deny reality, and make bad decisions because of it. They blame each other for things that should have been foreseen and prevented. They want certainty when all we can generate are probabilities. And they turn to superstition for explanations.

15 years ago, people I knew and loved suffered because of our failure to take seriously historical data indicating a significant risk of a major hurricane hitting New Orleans. A map prepared in 1981 by the Council on Environmental Quality, below, showed a 4% risk of a great hurricane in any given year—a 1 in 25 chance. (For an updated map see [here.]) 24 years later, Hurricane Katrina hit New Orleans. We predicted it was coming, but we still weren't prepared. And peopled died.

The map wasn't our only warning. At the University of New Orleans where I worked, I attended a presentation of a computer model that showed the city filling up like a bowl as its levees were over-topped. The speaker, from Louisiana State University, was just one of many who wanted New Orleans to be better prepared. But it seems easier to generate political will for cleanup than it is for preparation.

One of my former students, Paris Ervin, returned to New Orleans only to discover that his mother was still in her home, months after the hurricane, dead under a refrigerator, drowned. No matter what rebuilding we do, Paris' life is forever scarred. Paris' story is part of *When the Levees Broke* by Spike Lee.

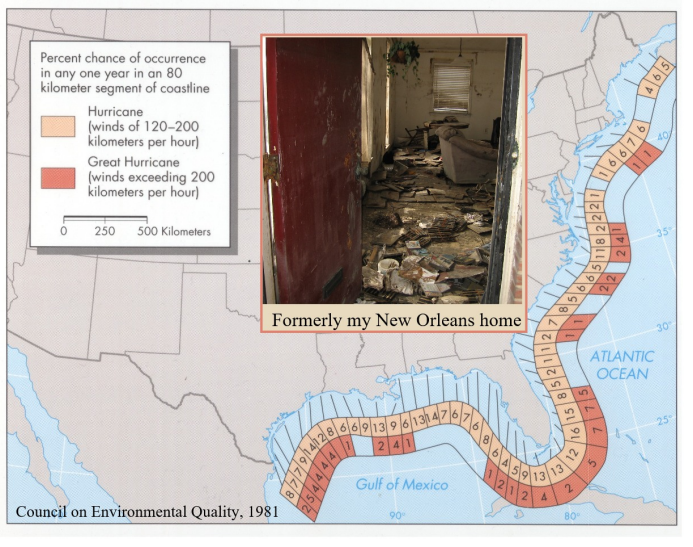

Figure 1: 1981 map of hurricane risk. Inset is the view through the door of my former home after 6.5 feet of water stood in it. The water line is above the window.

Statistics gives us tools for evaluating risks, not just what we imagine, fear, or wish for. The human tendency is to see what we want or fear. And though, no doubt, statistics can be used incorrectly, like any tool, it is a powerful means of separating out patterns in our observations from random noise. By doing our best to predict the future, including evaluating the quality of our predictions, we can improve decisions, even alter the future to one better than might have been.

To-do:

1. Think about your own life. When was a time where you ignored data or the opinion of experts, resulting in a less-than-optimal outcome? When was a time when experts turned out to be mistaken?

2. Given the COVID-19 pandemic occurring as I write this, what is an example where better use of scientific data might have saved lives?

3. I'm assuming you've had an introductory statistics course previously.

    - Terms like *mean* and *variance* should already be familiar, but it may have been a while since you used them. Go refresh your memory.
    - A number may represent a variety of types of measurements—nominal (labels, like brown/blue/red), ordinal (rank, like the Mohr's hardness scale), interval (having both rank and consistent difference between values, like the Celsius scale), and ratio (like degrees Kelvin, which has an absolute zero point.)

Part I
Gathering your data

1 Sampling methods

Before you can begin analyzing data, you have to gather it. Sometimes you are working with data from government sources or on-line databases. But no matter the source, you need an understanding of *how* your data was originally gathered so that you have a better understanding of its strengths and weaknesses, particularly its *bias*.

Bias is a problem built into data before you ever start to analyze it. For example, let's suppose a scale isn't zeroed properly. All of your resulting weight measurements—the data you want to analyze—are off. But you'll never know that simply by analyzing the numbers.

Or suppose that your helper is taking collecting soil but doesn't like to walk far from his car. All the resulting soil measurements represent areas near a road. In Iowa, it isn't hard to imagine that they would have higher than average salt concentrations. But you'll never know that simply by analyzing the numbers.

Scientists have developed lots of methods to try to prevent bias in their data, including these:

Standardized techniques: By standardizing how material (evidence) is collected, how the sample is stored prior to processing, how the chemistry is analyzed, and how the results are reported, scientists attempt to reduce errors and bias.

Randomization: By making observations at random locations or times, the scientists' prejudices are minimized. In medical studies, it has been clear for years that the doctor's belief about the value of a medicine alters the effectiveness of a drug. (The gold standard is a double-blind test, where the patient doesn't know if she is receiving the drug or a placebo nor does the doctor know. Think *chloroquine for COVID-19.*)

Stratification: A geologic stratum is a distinct layer of sedimentary rock. In statistics, strata are subpopulations for which there are good reasons for expecting results may differ and that we should therefore sample differently. Consider, for example, gender and breast or prostate cancer.

It may be easier to visualize these approaches by applying them to sampling at locations in a field. Let's suppose we wanted to collect soil at 25 locations for our site. A standardized approach could include laying out a nice, even grid pattern, 5 by 5, at which to gather our soil, like the leftmost portion of figure (gridded) below.

Or we might choose purely random locations, as in the central portion of the figure. Finally, if we have reasons for separating the site, such as human alterations, we might place our sample locations like in the rightmost portion of the figure.

Regardless of how we geographically place our samples, we need to follow standard protocols (sealed containers, stored in coolers on ice, chain-of-custody documentation, EPA-certified lab analysis) so that our work will stand up in court of law. Yep, that's a good way to think of it—it's realistic that as an environmental scientist you may have to testify about your work. If lawyers can show that you broke with standard procedures at any point, your data may prove worthless.

A point about language

Statisticians and scientists often use the word *sample* differently. To a statistician, a sample is a collection of observations. To a scientist, to sample is to take a part of something bigger, perhaps a sample of water from

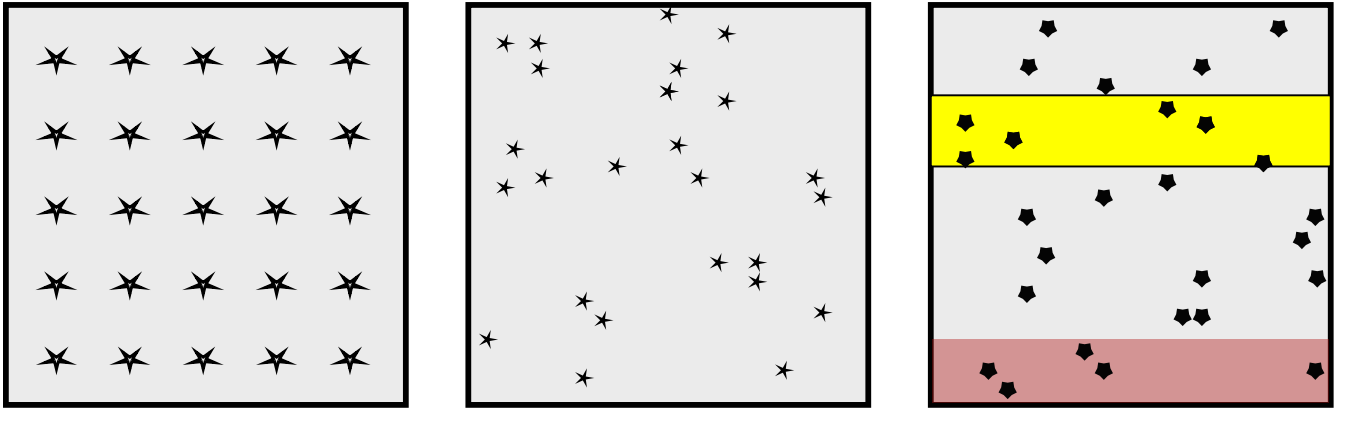

Figure 2: Gridded, random, and stratified-random sampling locations.

the Mississippi River. In both cases, *sample* refers to part of a bigger whole, but if you combine the two ideas carelessly, you come out speaking of a sample containing data from samples, a bit sloppy. One way some scientists and statisticians get around this problem is to speak of the sample set—the set of observations that make up the sample. But it can still be confusing, as English often is.

To-do:

1. Follow these instructions to add Excel's data-analysis toolpak.
2. The data-analysis tookpak includes an option for generating a variety of type of random numbers. If those numbers have no preferred clumping, they are *uniform.* Excel's toolpak allows us to create them for any interval and for multiple variables (columns), so an XY coordinate would be two variables (columns) of as many points as we specify. You'll need this for the problem below.
3. Try the first of the fire-ant assignments on the web site

2 What does your data represent?

In environmental science we rarely have all the data possible or even all the data we'd like to have. We have a sample of the data that hopefully represents well the entire population of data we could select from. Confusing? Let's try an example:

Suppose we wanted to know the current number of deer in Iowa with chronic wasting disease (CWD). For background on the problem, read [here] and [here]. It's impossible and a waste of money to test every deer in the state for CWD. What we'll have to do is figure out a way to test a few deer so that those deer represent as best as possible all deer.

One of the key things to consider is why the deer we test may not represent the entire population, such as

- if we test only dead deer, maybe deer with CWD were more likely to die, including being more likely to be shot by hunters because they are sickly, or
- CWD alters their behavior in ways that makes them more likely to be tested, or
- deer with CWD might wonder off by themselves to die without ever being tested, or ....

You get the idea, I hope. If we mess up the sampling, everything that follows may be wrong without us even recognizing it.

If we don't make mistakes or overlook flaws in our plan, we will end up with a *representative sample*. In that case, several things result:

- The mean (average) of our sample, $\overline{x}$, is the best guess for the mean of the population, μ.

- The variance of our sample, s^2, is the best guess for the variance of the population, σ^2.
- We can determine using the Central Limit Theorem how well we know the population mean by looking at the sample variance and the number of observations making up the sample,

$$SE = \frac{s}{\sqrt{n}},$$

where SE is the standard error of the mean and n is the number of observations in our sample. And because the Central Limit Theorem says that the sample mean, $\overline{X}$, is normally distributed even if the observations aren't, then there's a 68% chance the population mean is within one SE of the sample mean and 95% probability that the population mean is with two standard deviations of the sample mean.

Taking a bigger sample won't change μ and shouldn't change s, but it does change n. Thus, it also changes how well we estimate our population mean. Let's look at a simple example:

Problem: Suppose we are trying to estimate the average height of adult giraffes in Kenya, and we observe $n = 100$ giraffes. If mean height is $\overline{x} = 18'$ and the standard deviation of height is $s = 2'$, how well do we know the population mean?

Solution: We calculate our standard error of the mean, SE, as

$$SE = \frac{2}{\sqrt{100}} = \frac{2}{10} = 0.2$$

.

Therefore, we can be 68% sure that μ is 18 ± 0.2, i.e., between 17.8 and 19.2 feet. We can also be 95% sure μ is between 18 ± 0.4, i.e., between 17.6' and 18.4'.

The Central Limit Theorem also gives us a tool to figure out how many observations we need in our sample in order to know the mean as closely as we want.

Problem: Suppose we wanted to know the population mean, μ, within a tenth of a foot either way with 95% certainty. If from the study above, we estimated $s = 2.0'$, then what would the number of observations, n, need to be?

Solution: If we want 95% certainty, we need to look within two SE of the mean. That means

$$2 \times SE = 2 \times \frac{2.0}{\sqrt{n}} = 0.1,$$

or

$$n = \left(\frac{4.0}{0.1}\right)^2,$$

so

$$n = 1,600$$

In other words, to make our uncertainty a fourth as small, we'll need 16 times as many observations.

Further thoughts: We started this section discussing the question of what our sample represented. Let's come back to that and think about the realities:

- Is there more than one type of giraffe? If so, did we distinguish between them when making our observations?
- We wanted an estimate for *adult* giraffes. How can we tell when they become adults? (I can't even be sure about humans.)

- Do they differ by gender? Do we care?
- Do they differ by ecosystem? If so, was it easier to sample out in the savanna than in the forests, and did that bias our sample?
- How do you measure the height of a giraffe anyway? I doubt we'll climb a ladder with a measuring tape. And were all the measurements done the same way? To the top of the skull? to the maximum top of the ears? to the top of those little bumpy things (ossicones)?
- And how do you make sure a giraffe stands up straight?

Hopefully, you can see that even what starts off as a pretty straightforward project can become rapidly complicated. As I mentioned before, that's one reason we develop standard procedures—to try to make sure that our data represent what we claim they represent.

To-do:

1. Use *Descriptive Statistics* in Excel's Data-analysis toolpak to determine mean, standard deviation, and standard error of the mean (SE) for the SpASizes and SpBSizes in the SpeciesPreTest.xls dataset on the class website.
2. Look up Excel's built-in functions for statistical formulas, such as *=AVERAGE(A1:A25)*.

Part II
Getting to know your data

3 Your eyes see patterns, so plot your data

Think back to what you've learned abot the scientific method:

- make observations (gather data),
- look for patterns,
- hypothesize (guess) why the patterns occur, and
- test predictions based on the implications of the hypothesis.
- If the predictions are strong, the hypothesis might become a theory. If not, refine the hypothesis.

We talked previously a bit about gathering data. Now we need to start looking for patterns. Fortunately, evolution has provided a great tool—your eye! It's quite valuable to tell the difference between something you might eat (Figure 3) and something that might eat you.

For example, two sets of data with the same mean and standard deviation are shown in histograms in Figure 4. Obviously, the two sets are quite different, but simply looking at descriptive statistics can be misleading. Use your eyes!

In order to taken advantage of your ability to recognize patterns, you need to become both skillful and knowledgeable about plotting data. Excel gives you quite a few options. See, for example, [here] and [here.]

Figure 3: Waterbuck in Kenya.

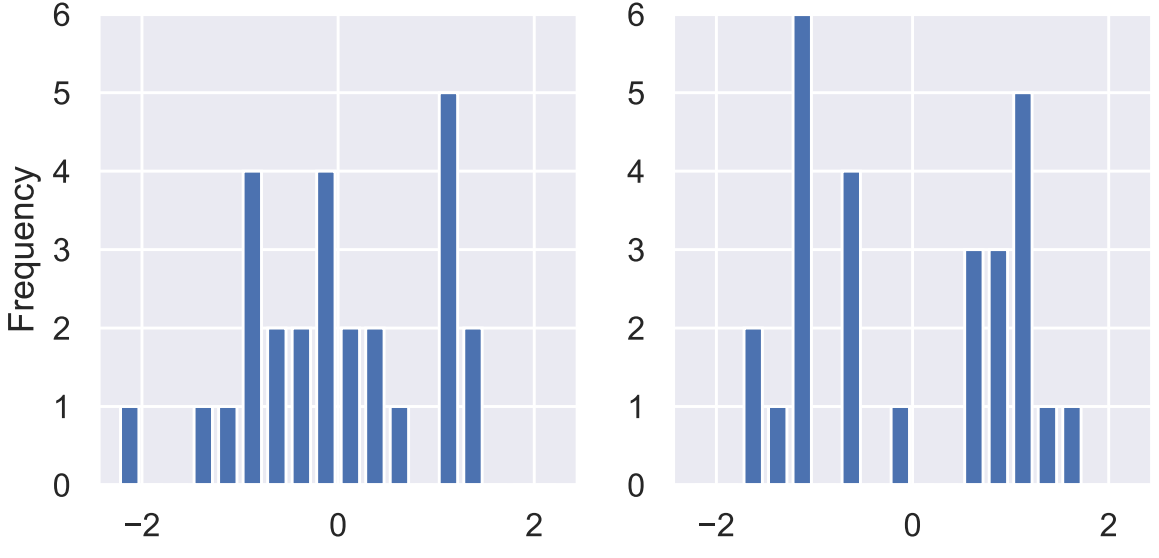

Figure 4: Histograms of two different sets of data with the same mean and standard deviation.

To-do:

1. Go [here] for weather data.
    - Search for *Daily Summaries, 1952-02-01 to present, Cities, Dubuque.*
    - Click on *View Full Details.*
    - Click on *Station List.*
    - Click on *Dubuque Lock and Dam 11.*
    - Click on *Add to Cart.*
    - Click on *View Cart.*
    - Choose *Custom GHCN-Daily CSV* as the *Output Format.*
    - Select the date range as 1942-01-01 to the present.
    - Click on *Continue.*
    - Select *Precipitation* and *Air Temperature.*
    - Click *Continue.*
    - Click *Submit Order.*
    - Note: The first time you try this, you will have to specify an email to which a link to the data is sent.
2. Once you obtain the data, open it in excel and start exploring.
    - A pivot chart and table is a good way to initially explore the data.

- Try plotting monthly averages.
- Try plotting XY (scatter) plots of time versus temperature.
- What else can you do?

4 Bins, frequencies, and histograms

In Figure 4, two histograms are used to represent sets of data. Histograms are graphical ways of representing data graphically where the heights of columns are proportional to the frequency of values within various ranges of value. That's too long of an explanation—let's use an example. To meet the core requirement for a lab science, a student can pass my Physical Geology course. To pass with a D requires at least a 60% grade. To make a C requires a 70%. Thus the *bin* for a D is between 60 and 70 (actually 69.5, as I round up.) The bin for a C is between 70 and 80. Get the idea? So if I looked at an entire class or 40 students, I might have 5 As, 8 Bs, 15 Cs, 10 Ds, and 2 Fs. I could plot these counts—the frequencies. Or I could plot the percentages—the relative frequencies, such as 8 out of 40 is 20%. The histograms would look like Figure 5.

Looking at histograms gives you considerably more information than just looking at the mean value. The teacher with his head in the oven and feet in the freezer on average feels fine. ☺ Similarly, in a class where half the students are making As and half Fs, the average is a C, but that's very different from the class discussed above. In terms of implications, a class with all As and Fs would tell me as a teacher that something needed to change: 1) Why were the two groups of students doing so radically different? 2) Was the issue with the class itself or with signing up for the class? 3) Was the grading meaningful? Regardless, such a distribution of grades is hard to justify for long. Looking at the histogram instead of just the mean gives me impetus to act.

To-do:

1. Use Excel's data analysis toolpak to plot a histogram of SpASizes in SpeciesPreTest.xls.

- First, leave the bin range blank. Excel will choose the ranges based on the number of points and the range of the data. Notice what it does.
- Enter your own set of bins by entering the upper boundary for each bin, starting at the lowest. For example, for grade bins, I might enter 59.5, 69.5, 79.5, and 89.5. Choose yours to fit the SpASizes set.

2. Determine how to convert your histogram from counts to percentages.

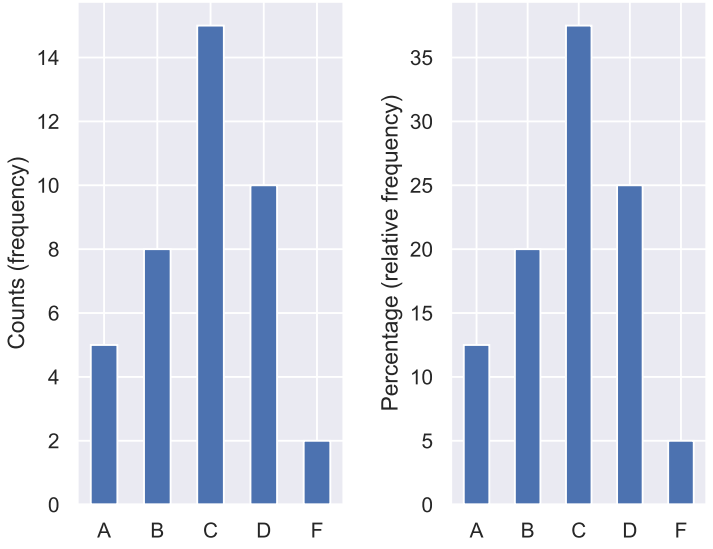

Figure 5: Example histograms of grades in a class of 40 students. Left: counts. Right: percentage

Part III

Comparing two data sets of the same variable

5 Statistical tests

All statistical tests come down at some point to this question: Is what I see likely to have occurred randomly, or is there a pattern present? If purely random, there is nothing significant about it. Take a look at 6. Does it look like anything to you? Those two eyes staring at me are creepy! What? You don't see them? You see a pelvis? Pervert!

Figure 6: Ink blot number 4 by Hermann Rorschach, Public Domain, from [here.]

Our *null hypothesis* is always that there is nothing significant there—no difference, no pattern, no impact of a different treatment. If I see that pelvis instead of a bunch of random ink on a sheet of paper, it most likely says more about my thinking than reality.

In order to say if something is improbable (unlikely to occur by chance alone), I need a decent sense of probability, which most of us lack. Clairvoyants and fortune-tellers use our ignorance to make a living, but we scientists shouldn't depend upon deception to pay our bills. Instead, we develop models of probability, such as the *normal distribution* or the *uniform distribution* to quantify the likelihood of events. We then use these models to determine probabilities. I'm sure by now that in some class you've calculated the probability of rolling a 7 with two dice. That calculation is based on a uniform distribution of the probability of each face of a die being equally likely to land facing up: a one is as likely as a two as a six, each coming up $\frac{1}{6}$th of the time. But when we roll two die, things get a bit more complicated, because a 2 is less likely than a 7 which is more likely than an 11 (Figure 7).

This works fine for the case where there are only a limited set of possible outcomes where we can list and calculate each. But in environmental science, it's far more common to have an uncountable set of possible outcomes, and in that case we have to depend on someone having done the calculus to determine the probabilities. Fortunately, nowadays we have plenty of resources, including computer programs, that will spit out the probabilities if we can only get things set up right. Excel reports these *p-values* directly in the output from the data-analysis toolpak, which we'll be using in the following sections.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Figure 7: The possible outcomes of rolling two dice. Notice that 12 and 2 only once in 36 outcomes, or $p = \frac{1}{36}$. But 7 occurs 6 in 36 outcomes, or $p = \frac{1}{6}$. Lucky 7, indeed.

In environmental science, most of the time, we consider something *unlikely to have occurred by random chance* if its p-value is less than $\alpha = 0.05$. (We decide on α before we start and live with it. α is our *significance value.*) So remember these phrases:

- *Small p* means *big differences.*
- *Small p* means smaller than $p = 0.05$, most times.

However, you need remember that getting a small p-value doesn't say much of anything about whether you have picked the right probability model in the first place. Every model has assumptions built into it. You're responsible for knowing those assumptions. If you overlook one, you run the risk of making huge mistakes. For example, look at Figure 8. You can fit a line quite well, get a p-value that's tiny, and still have the wrong model.

In addition, it's important to remember that despite something being unlikely, it still occasionally happens. I was reminded of this fact while cleaning up from a weekend tubing trip on the Little Maquoketa River.

Several years ago, I bought several inner tubes to use on tubing trips with students. Having using some funds left by my mother, I named them the *Ina S. Easley Memorial Flotilla.* (The building where I work has a wing named after a rich donor's mother, so I also donated the *Ina S. Easley Memorial Toaster Oven.*) On last weekend's trip, my older daughter quickly put a hole in one of the tubes, from then on sharing a tube with her boyfriend. We got home late, and while I started the grill, my wife and a friend washed mud from a couple of kayaks and the six deflated tubes we'd used. Unfortunately, no one kept track of which tube had gone flat on its own.

This morning, I got out the inflater and started pumping up the inner tubes, looking for the one with a leak that needed patching. It ended up being the sixth tube I inflated. Sixth out of six. The odds of it being last were 1 in 6, or 0.167. Pretty small odds, yet it happened. Not a whole lot greater than the 0.05 we environmental scientists commonly use for decision-making.

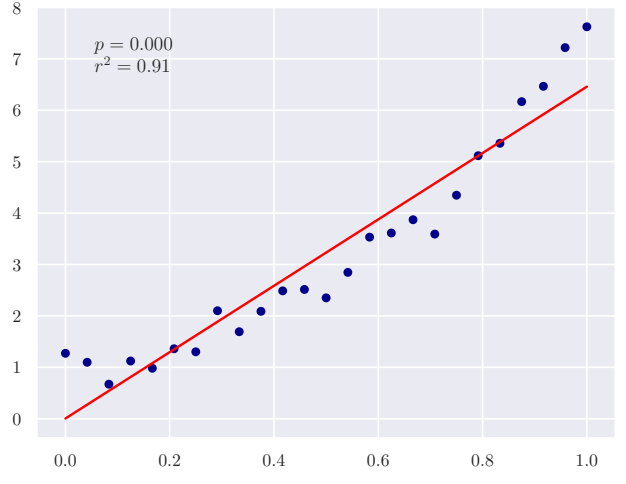

Figure 8: An extremely good fit of a line to data. However, a line is the wrong model. Notice the pattern of above-below-above the line.

6 F-tests

As we begin to compare means, gradients, and multiple populations, characterizing the variance will play a key role, and F-tests can be used to compare two *variances*. Why is that important? If you are trying to determine if two sets of data represent different populations, there are many ways of describing them. Though we often look at the mean (average) values, the variance is just as important (Figure 9).

For example, take a look at (C) in Figure 9. Even though the two populations have the same mean, they are clearly significantly different. Suppose the two graphs represent scores by two classes on a standardized test. The cheap and easy way to deal with the results is to just look at the means and say, "The two classes are doing as well as each other." But the tougher, more-honest approach would be to ask, "Why is there such wider spread in scores in one class than in the other?" We might look at differences in how the classes are taught, how students are chosen to be in the classes, or something happening in the classroom that is causing the spread. Regardless, we'll have to dig deeper if we want to understand what's going on.

In terms of next steps in this course, we will use t-tests for comparing means, and the choice of which t-test to use depends upon the variances of the two sample sets.

To-do:

1. Take a look [here] for using Excel to complete an F-test.

2. Be prepared to discuss at least three ways that variance in weather might be more important to an Iowa farmer than average weather.

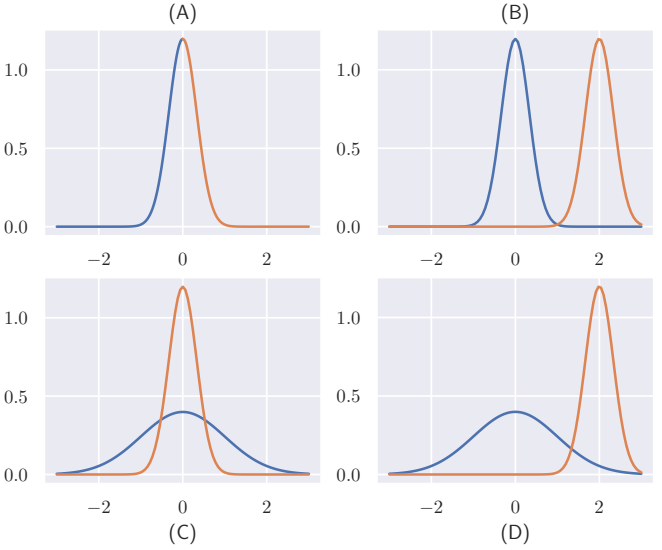

Figure 9: Two population distributions with (A) equal means and variances, (B) equal variances but different means, (C) equal means but different variances, and (D) different means and different variances.

7 T-tests

In your introductory statistics course, you probably spent a fair amount of time on z-tests. I'm going to ignore them—they assume you know the population variance, σ, which you almost never do in environmental-science applications. Instead, we will focus upon t-tests, similar in nearly every way to z-tests except that the variance is also estimated from your data as the sample variance, s^2.

Let's start off with a simple question: "Do black squirrels weigh the same as gray squirrels?" Simple, but not so easy to answer. Perhaps we'll put a trap next to our bird feeder that the squirrels love to steal from. Maybe we catch a few. How do we weigh them? String them up by the tails? Can you image the fight they'd put on? Maybe we put them to sleep first, perhaps by giving them a lecture on local geology. But maybe only the bully squirrels of the neighborhood venture toward our feeder. And they can the most seed and grow fat. Or only the young squirrels are foolish enough to venture into our trap. On our concerns go. But suppose somehow or other we gather weights for 10 gray squirrels and 6 black squirrels (Table 1). Now we are ready to see if the weights are different. Oh, darn—every single squirrel weighs something different, so "Yes, they do *not* weigh the same!"

But that's not really what we're getting at. We'd really like to know if there is a difference in the *average* of the two groups of squirrels. Choosing the right t-test to compare means requires us first to check the variances. For that, we use an F-test like in the last section. Give it a try with this squirrel data.

If you gave it a try, you got a p-value far larger than 0.05. Big p = small difference. Or, to put it in more statistical language, *we accept the null hypothesis of no significant difference in variances.*

We are now ready to choose a t-test. Because we ran the F-test with *no difference* results, we can run the t-test *with equal variances*. Give it a try. See any difference? any *significant* difference?

Gray	Black
463	517
480	494
499	450
425	550
470	488
485	546
460	
485	
547	
520	

Table 1: Weight in grams of Eastern gray squirrels and their black morph.

A special case: paired data

Sometimes our data come in pairs, often where we expect the two values to be the same or where we hope they are consistently different. For example, suppose we visit RV parks along rivers draining into the Mississippi River near Dubuque. We take water samples at each RV location, upstream and downstream, generating has a pair of data. Our null hypothesis is that there is no difference in water quality above and below the parks. Excel's t-test options include a *paired t-test.* We can use it without first running an F-test. However, as always, it's a good idea to graph our data, in this case a scatter plot of upstream versus downstream data. Perhaps a single location has a huge

effect. In addition, data from chemical analysis are coomonly recognized to have avariance that's proportional to the values—that is, if all data had a 10% uncertainty, the 10% of large concentrations would be large. Think about it. A common way of dealing with such data is to take the square root of the data to normalize it before running the t-test.

To-do:

1. If you didn't already, try the F- and t-tests with the squirrel data.
2. Do the t- and F-test practice [here.]

Part IV
Analyzing two variables

8 Correlation

As I write this, death statistics from the pandemic show that people with diabetes have a greater chance of dying if they become infected with Covid-19—diabetes and death by Covid-19 are *correlated.*[1] Likewise, the prevalence of diabetes decreases with wealth—they are *inversely correlated.*

A key phrase to remember is, "Correlation is not causation." Being poor doesn't cause diabetes. However, certain lifestyle behaviors, such as lack of exercise, can increase the chance of type-2 diabetes. And those behaviors may be linked to poverty, though not necessarily as a cause, perhaps even as a result.

Does this sound complicated? Well, it seems like everything related to humans is. But in environmental science, we are looking at human interactions with the environment, and we can't escape dealing with behavior. For example, hunting kills animals. But hunting also pays for protection of habitat, salaries for wildlife managers, and controlled management of populations. Kenya banned hunting in 1977, five years before I went there to teach math for two years. However, at multiple times since, poaching has killed far more animals than hunting once did. You might argue that poaching is wrong and hunting is, too, and two wrongs don't make a right. Whether I agreed with you or not is irrelevant—I am far more interested in effectiveness than ideological purity. And in terms of effectiveness, eliminating hunting may actually *decrease* animal populations and/or lead to greater suffering.

What we see in a discussion like that above is that the *application* of statistics is *not* value-free. I am a firm believer in the need for data-driven decisions and policies. We need to know what *is.* However, what is doesn't tell us what *ought to be,* independent of our values. And much of environmental science is driven by our values.

[1] Type 1 and Type 2 diabetes and COVID-19 related mortality in England: a whole population study. Emma Barron et al., accessed 24 May 2020.

r^2 values

When a model is fit to a set of data, a common way of characterizing the fit is with the r^2 value. Let's suppose we're fitting a line, a common model. If the line passes exactly through all the data, $r^2 = 1.0$. The line explains 100% of our data. If there is no benefit to fitting a line—if it explains nothing—then $r^2 = 0$. In between, r^2 represents the percentage of variation in our dataset explained by our model. Look at Figure 10.

The square root of r^2 is, of course, r, Usually referred to as the Person's-r, or, more simply, as the correlation coefficient. If r is positive, two variables are directly correlated. If negative, that areindrectly correlated. See Figure 11

To-do:

1. What is a *spurious correlation?* Take a look [here.]
2. What are five values that have driven major environmental policy and regulations. Be specific with your examples.
3. Try the exercisw [here.] Check the correlation between lead and zinc. What physical reason could cause them to be correlated? Does high lead content *cause* high zinc content? Explain.

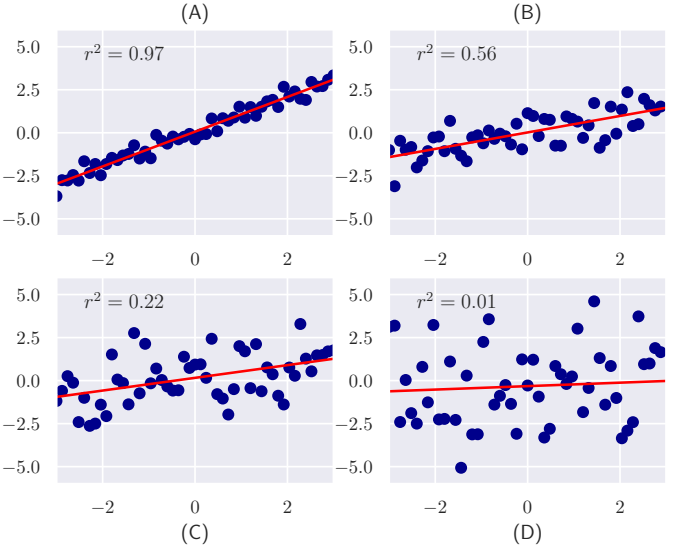

Figure 10: The r^2 value for the fit of a line to data, showing (A) perfect fit, (B) a very good fit, (C) a poor fit, and (D) a worthless fit.

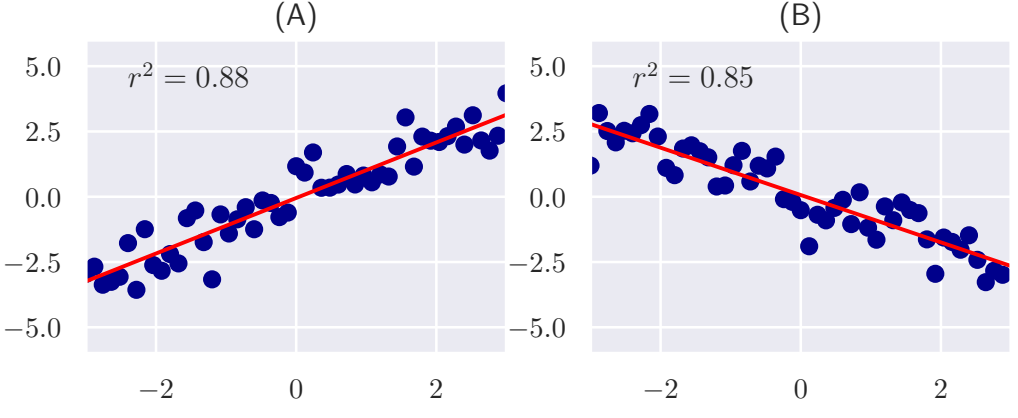

Figure 11: The r value for the fit of a line to data, showing (A) positive (direct) correlation and (B) negative (indirect) correlation.

9 Regression

The difference between correlation and regression is mainly in our thinking about our variables:

- In correlation studies, we are looking for patterns in how variables change together.
- In regression, we are looking at *cause and effect*. We examine the values of independ variables in terms of their abilty to predict the value of independent variables.

Both look for what patterns are present, but regression asserts that some things measured *cause* the other things measure to vary. It's up to the scientist to try to establish how the relationship works.

To-do:

1. Take a look [here.]
2. Watch the video [here.]
3. Do the assignment using the data set [here.]

10 χ^2-test of goodness-of-fit

We use a lot of models in science and statistics. We need a way to estimate how well our models fit our data. The χ^2-test is our tool. For example, we might be interested in whether Covid-19 kills members of all age groups equally (a uniform distribution). If true, expected number of deaths as a percentage of those infected should be roughly equal (assuming we can determine the number infected.) For each of those groups, we can calculate the number *expected* to die versus the number *observed* dying. This difference between expected and observed is the foundation of the χ^2-test.

Take a look at the presentation [here] and the Excel worksheet [here.] We are trying to determine if a species has a preferred habitat, something hunters have been doing since before recorded time. However, in this case, we're studying a type of sea grass. We'd like to know whether the tidal zones impact its abundance.

In Figure 12, the upper zone (A) between the average high tide and the annual high tide is covered with water *occassionally*. On a average day, it is not inundated. Zones B and C, on an average day are covered with water some of the time, though we would expect B to be covered more of the time.

Let's suppose that the area of beach for A, B, and C are all equal. (If they weren't we'd need to adjust for it—if B were twice as long, we'd expect twice as much sea grass.) We could go to the site, layout a transect perpendicular to the coast and a specified width, and then measure how much sea grass is with our area, perhaps by weight, number of clumps, area covered, or some other metric, depending on how it grows. Suppose in area A, we find 40 units, in area B, we find 75 units, and in area C, we find 125 Units. It looks like our sea grass prefers growing below mean tidal level, but we need to see if our measurements are statistically significant. For this, we use the χ_2-squared test.

Unfortunately, Excel's data analysis toolpak does not have a tool for performing χ^2 tests. However, it does have a function that we'll use after we do some calculations. We begin by setting up our data in like the Table2 below:

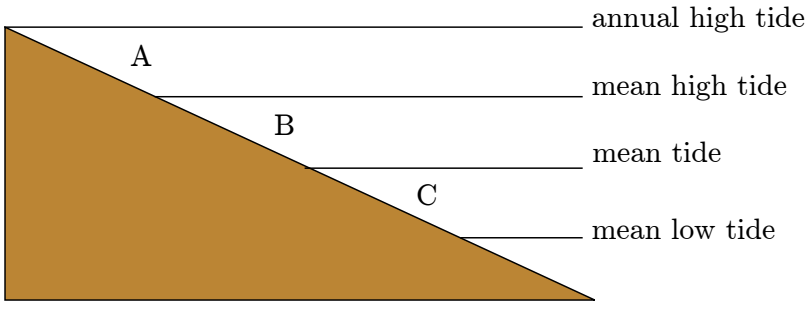

Figure 12: Coastal wetlands in the tidal zone covered at least part of the time with water.

Once we have completed these calculations, we use Excel's function, $CHISQ.DIST.RT(ChiSq, dof)$, where $ChiSq$ is the value in the bottom-right of our table, and dof is the degrees of freedom (the number of categories of observed values - 1). So for this problem, our formula evaluates $CHISQ.DIST.RT(25.3125, 2)$ and returns a p-value of 0.005. Therefore, we can reject our null hypothesis that, *The species of sea grass we observed is uniformly distributed across tidal zones.*

We can use the χ^2-test for checking how our data matches any distribution, including the normal distribution. For that test, we create bins, like for histograms, and compare the expected number to the observed number. The difference from what we did above is that our expectation depends on how close to the mean our bins are located.

Zone	**Observed**	**Expected**	$\frac{(O-E)^2}{E}$
(A) upper	40	80	20
(B) middle	75	80	0.3125
(C) lower	125	80	25.3125
Total	**240**	**240**	**45.625**

Table 2: Table for calculating whether our sea-grass measurements are uniformly distributed. (Data in black; calculated values in green.)

Take a look back at Figure 4. The data on the right is clearly not normally distributed, but it's not so clear with the data on the left.

To-do:

1. Perform a χ^2-test for the data in Figure 4. What are your p-values for each?
2. If you haven't already, complete the worksheet [here.]
3. Think of at least one real-world example where you might use the χ^2-test.

Part V
Multiple variables

11 Spatial autocorrelation

The concept of *spatial autocorrelation* sounds incomprehensible at first, but it is really very easily understood intuitively by environmental scientists. The idea is this: Often, going a short distance doesn't change things much. For example, if you're in Des Moines, traveling a few miles isn't likely to change your elevation much—elevation is strongly autocorrelated in Iowa. But travel to Colorado and those same few miles can mean thousands of feet of elevation change. This *persistence in space* is very valuable to environmental scientists, geographers, petroleum geologists, and mining engineers.

Let's look at a couple of images first. In Figure 13, notice how much smoother (A) is than (B).

Geostatisticians have developed a variety of means of characterizing spatial autocorrelation. One of the early and still widely used approaches is the *variogram*. The variogram (Figure 14) grew out of mining applications and their specific needs, but it has found uses in many other areas. Here are the basic ideas:

Management decisions: When running a mining operation, the scale at which samples are made and the scale at which processing decisions are made are often quite different. A rock sample to analyze might be the size of your fist, but the load to ship to the refinery or waste heap might be the size of a huge truck. Similarly, in the environmental field, we might take a few soil samples, each fitting into a ziplock bag, and need to decide whether a site is too contaminated to allow public access.

Nugget-effect: If we were mining for gold, a single pure nugget in our sample would give a very high (and unreasonable) estimate for the amount of gold present at a site. If we'd moved just a short distance away to where that nugget was not included in our sample, our estimate of the total gold on-site would be radically lower. This short-distance variation is called the *nugget-effect*.

Range: At some distance, the *range*, our ability to improve our estimate based on a measurement at our current location decreases to near zero. If you were in Colorado, moving a hundred miles could change your elevation radically. In Iowa, moving a hundred miles from Des Moines would lead to a few tens of feet of elevation change, in most cases. Thus, the range for elevation in Iowa is much longer than it is in Colorado.

Sill: As we said earlier, it is important to characterize the variance for a sample. However, it's common for geologists and environmental scientists to take more measurements near a place they find interesting—e.g., a high concentration of gold or a contaminant. But if high values are more likely to be found near other high values and vice-versa, then this sampling approach biases our estimates of both the mean and variance. To have meaningful estimates, we have to account for this spatial autocorrelation, and the sill is our best guess at the overall variance, unbiased by clumping of our observation points.

So, we can say a few things based upon the variogram we develop from our data:

1. Data whose variogram has a long range will vary more smoothly than those data with a variogram with a shorter range.

2. Data with a large sill and nugget-effect will have a larger range of values.

3. The larger the nugget effect is relative to the sill, the more uncertain estimates are for a short distance away.

Any time you've dealt with a contour map or heat map, realize that built into its construction are assumptions or models of how things vary in space. If observations have not been made closely spaced at least part of the time, the resulting smoothness may be misleading.

To-do:

1. Read the article [here.] Be prepared to discuss it. In particular,
    - How does the chemical analysis tie to grainsize?
    - How are hydrology and mineralogy distribution linked?
    - How does the opening of the spillway change the distribution of chemicals? Why?
2. Provide an example scenario where charactering spatial variation could save your client money on a project for which you are an environmental consultant.

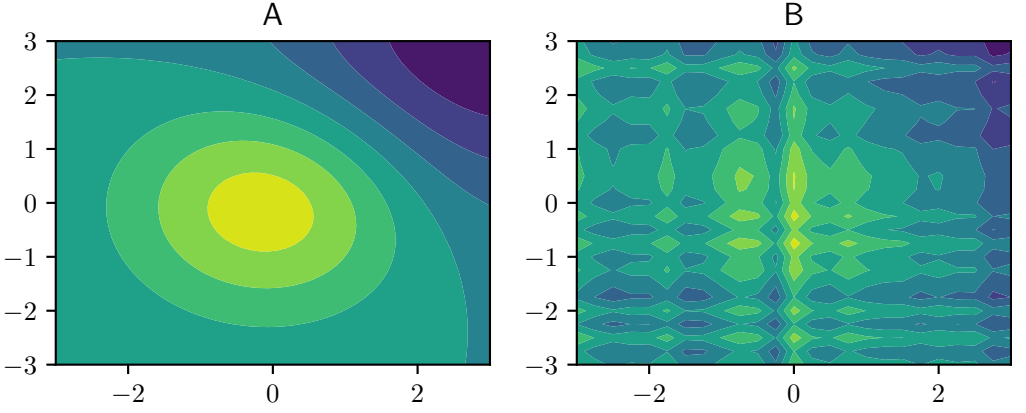

Figure 13: Contours of data with (A) much persistence in space (high spatial autocorrelation) and (B) little persistence in space (low spatial autocorrelation).

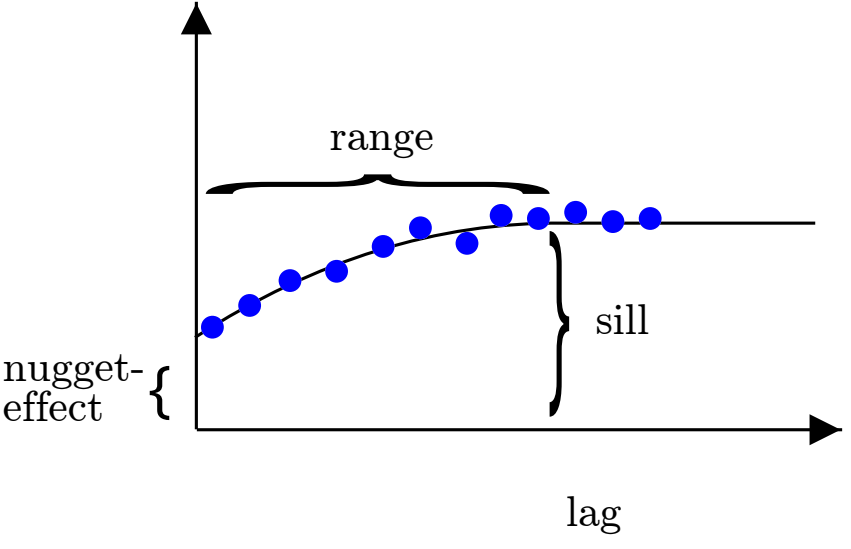

Figure 14: Example sample variogram showing the main components of nuggest-effect, range, and sill. The x-axis is lag, and the y-axis is quared differences.

12 Correlation matrices

As this pandemic progresses, we've seen that people with underlying health problems, such as diabetes, are more likely to die from Covid-19—diabetes and death from Covid-19 are correlated. Is diabetes the strongest predictor? Are there others? Looking at a variety of *risk factors* may help us determine who to keep an especially close eye on—who we may want to go the farthest out of our way to protect.

Part VI

Some further reading