Meteorology 5040
Environmental Statistics

1 Introduction

Statistics encompasses a broad range of subjects. The goal of this course is to review and apply methods to examine data. You may have some distrust regarding the applications of statistics to everyday life. Nearly every day some statistical study is reported in the media that is construed to prove the value of one substance or approach vs. another. A common public perception of statistics can be summarized as: statistics is useful if it confirms your biases and not useful if it requires you to question your beliefs. For example, consider the often misattributed quote: there are three types of lies- lies, damn lies, and statistics. But, let's be real- you rely on statistics all the time to *describe* events or *compare* outcomes (ERA, GPA), *infer* what's going to happen in the future based on past or current samples (polling for elections), *assessing risk* (is smoking bad for you?), and, what scientific research is all about, *identify* relationships within a large volume of data.

One of the goals of this course will be to emphasize that evaluating data requires you to check results and conclusions carefully. Could there be a problem with the data? What assumptions were made along the way to distill the data that might bias the results? Was the analysis approach really appropriate? What alternative explanations could there be? How statistically, as well as practically, significant are the results?

The bottom line is KISS- Keep It Simple Stupid. Most complicated statistical approaches are just that- complicated. My general rule is that if you have a really useful result, you should be able to see the glimmer of it in the raw data. The trick is to find the gold nugget in the gravel. Subjective evaluation of data is critical- if you have to manipulate and massage the data and then use some complicated analysis technique to distill the results, how useful will your results be? It may be useful if you are attempting to test and verify a well-defined hypothesis, but it may not have a direct practical application.

Abuses of statistics result from the common situation that if the only tool in your tool belt is a hammer, everything starts to look like a nail. Some statistical measures work better than others, depending on the goal of the research and the type of data. A goal of this course is to add a few tools to your tool belt- it will be up to you to figure out when is the right time to use them.

As an introductory example of environmental data, consider a drought index for Utah shown in Fig. 1.1 that is derived from a carbon isotope analysis of tree ring data near Alton Utah (near Bryce Canyon). Droughts lasting a decade or more are estimated by two or more consecutive negative index values and have been estimated to occur most recently during the 1960's, 1930's, and 1900's as well as many other periods during earlier centuries. There are many questions we need to ask before assuming this index is a useful estimate of recent climate behavior for our state. How was it constructed in terms of the physiologic response of trees to climate? How sensitive is the methodology to local weather conditions and is the index representative of climate conditions regionally? What corroboration is available during the instrumental record that can be used to assess its value? What other factors do you think should be considered?

a. Effective Research

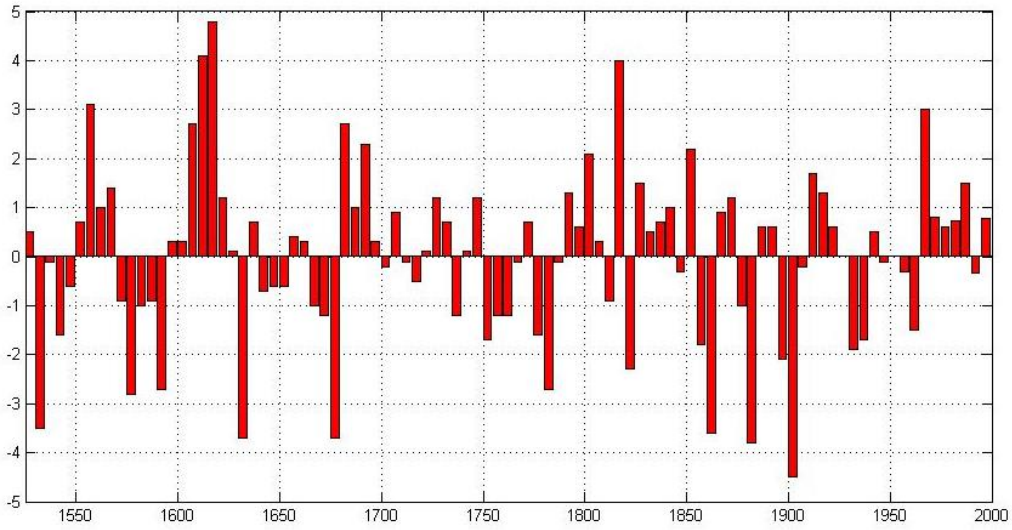

Figure 1.1. Utah drought index derived from an isotopic analysis of tree ring data near Alton Utah (see http://lwf.ncdc.noaa.gov/paleo/treering/isotope/iso-drought.html).

Effective research requires preparation. Environmental problems are complicated. It is a bit unrealistic to expect that a phenomenon that evolves in time and space with complex interactions between variables will have a single causal factor that is completely apparent through a quick analysis of data. There are several basic steps to effective research:

- **distill** a general interest in a subject into **a specific question/hypothesis** that can be evaluated. While we are all interested in global warming, it is a bit unrealistic to expect to examine all of the GCM data or long records of weather data in one study. Sometimes it is necessary to limit a study to what can be determined given the existing model or data assets. However, as your study progresses, you may find the need to search for other data that can help address the question of interest.

- **organize the data**. This can be the most tedious step, but the most critical. What data do you need? How should gaps of missing data be treated? What are the possible limitations of the data? Have you looked at the raw data enough to recognize what might be physically implausible values? Does the data make sense physically? Do you need to remove large signals that are not of interest to your question, i.e., seasonal or diurnal cycles, or model biases? You can avoid a lot of missteps by carefully preprocessing the data.

- **find relationship(s) among the data**. Your choices of analysis approaches are critical. Given your hypothesis and your data, what is the best way to proceed? It isn't going to be easy- as you proceed, you may end up having to reevaluate the goal of your research, find other data assets, or try other ways of analyzing the data.

- **examine the significance of your results**. There is statistical significance, which you need to address, and then there is practical significance. Statistical significance is important- all too often people over fit data and then their results will be useless when applied to an independent data set. However, I want you to recognize practical significance. Do your results make sense physically? Have you considered the amount of

spatial and temporal dependence in the underlying data set? That is, the variations in environmental observations at one location may not be that different from those nearby. And, observations at one time may not be that different from those before or after the time of interest. Does the relationship hold up in an independent data set? While your results may pass a statistical significant test, are they of any use?

- **review thoroughly what you have done and document your analysis and results.** Did you cut any corners during the data preparation or analysis? The scientific ethos requires that based on the information you provide, someone else can take the same data and derive the same results. It is now often required in order to publish a scientific study that the data used in that study must be made available to others to evaluate your results.

- **submit your results and study for independent evaluation.** Peer review is critical for providing an independent opinion of the merits of the analysis that you have completed. Accepting criticism of one's work can be difficult. Be wary of any study that has not been subjected to peer review. If a project does not require peer review, then ask colleagues for an honest appraisal of the study's results.

Be aware, however, that refereed statistical studies may have serious flaws. Recognize that there is a *publication bias* to report positive results, no matter how inconsequential, as opposed to negative results. The influence of El Nino on the snowpack in the Wasatch is weak at best, and there have been many papers attempting to explain those weak relationships. However, it would be virtually impossible to have a journal publish a paper with the primary conclusion being there is no relationship between El Nino and Wasatch snowpack.

b. Uncertainty

Uncertainty is at the core of this course. Uncertainty arises because:

1. we can never measure the environment with complete accuracy and precision,
2. the environment is a chaotic system, which is a maddening combination of randomness and order arising from the characteristics of a complex nonlinear system,
3. our understanding of the environmental system is imperfect, so physical (and certainly statistical) models do not capture the complete behavior of the system.

Always assume that ANY observations are uncertain. Was a human observer involved? Did the same observer take the observations each time? What automated equipment was used? What metadata (data describing the data) are available to explain how the observations were taken?

I'm a pessimist by nature. If someone asks about some unusual observation that they saw in MesoWest, the first thing I do is begin thinking about why that observation is likely messed up. Rarely do we know what the "true" state of the environment should be. An instrument can be calibrated to a standard in the laboratory, for example, distilled water in an ice bath should have a temperature of 0^{o}C (but what about impurities, dirt in the container, etc.?). However, I'm also more of an optimist than many- we can make sense of data even when imperfections in the data exist. If a wind sensor in a slot canyon only blows from the west or east, you may still be able to

determine when the front passes that station in most cases, even with the biases inherent in the data.

We need to distinguish between the usually unknown (outside of the laboratory) "truth" and actual measurements.

- True value- value of a quantity sought through measurement, but unknown usually in the field

All measurements from instruments will provide an estimate of the true value, with varying degrees of success. A number of measures are available to gauge the uncertainty of observations.

- Accuracy- difference in response between a standard and instrument in varying environmental conditions
    - a measure of how close a measurement is to the "true" value
    - high accuracy can be expensive
- Limitation- capability of an instrument to give accurate readings within a specific range. At extreme ranges, readings may be less accurate and hence, limited
- Precision- how well repeated measurements of some quantity agree with each other.
    - a precise instrument can be inaccurate

Consider the following shots at a "bulls eye", where the goal is to hit the X, the true value. While it is certainly best to have observations that are both accurate and precise, it may be too costly to do so.

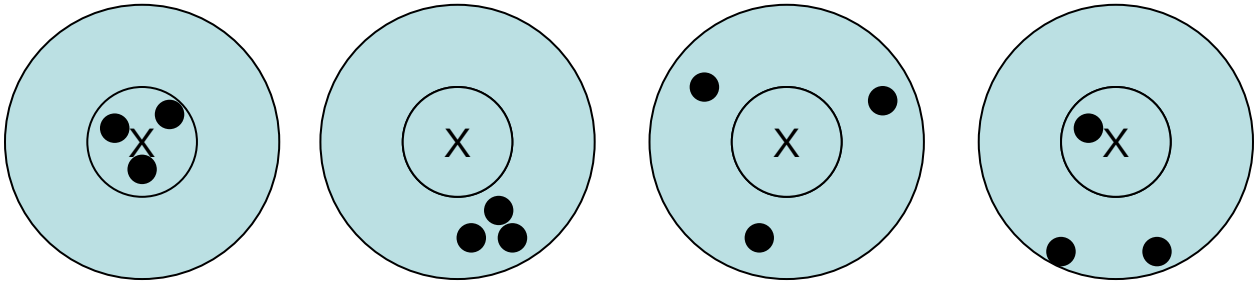

Figure 1.2. X is the "truth" while the filled circles reflect observations.

High accuracy	**Low accuracy**	**High accuracy**	**Low accuracy**
High precision	**High precision**	**Low precision**	**Low precision**
Small uncertainty	**Large uncertainty**	**Large uncertainty**	**Large uncertainty**

So, high precision cannot make up for inaccurate information ("*garbage in, garbage out*"). Many people end up confused when they hear the same wrong information from multiple sources (which often originate from a single highly unreliable source). *Judgment and integrity are critical* in all facets of life, particularly when evaluating the credibility of statistical results.

c. Systematic vs. Random Errors

The tendency in environmental fields is to overstate the amount of random error and underestimate systematic errors (or biases).

- Random- that which is not precisely predictable or determinable. Random errors arise physically due to sudden changes in the environment, due to turbulence or other processes. Random errors can also occur due to faulty equipment or observer carelessness.
- Systematic- errors arising from a consistent response of a measuring device to environmental conditions or faulty characteristics of instrumentation that occurs frequently. Systematic errors can vary as a function of weather regime: nonaspirated thermistors (thermometers over which there is no air blowing past mechanically) will tend to be affected by radiational cooling and heating when the winds are light, for example.

Look at the bulls eye plots above. High accuracy and low precision observations are an example of random errors while low accuracy and high precision observations are an example of systematic errors.

c. Population vs. Sample

In the same way that we rarely know what the "true" value of a variable should be, we never know the entire population of true values as the environmental conditions change in time or space. We hope that we choose a sample of observations for analysis such that each element in the population has an equal chance to be selected, that is our sample is *representative* of the larger (usually unknown population). For example, we don't know what the future values will be, so there are clear problems with any sample we choose. For example, the increasing trend of atmospheric carbon dioxide concentrations implies that a sample taken from the recent past can not reflect accurately the population of carbon dioxide concentrations that includes future values. When we choose a sample, we must try to avoid *selection bias* (cherry picking the environmental situations we study).

Also, because of the serial dependence of environmental data (observations collected consecutively will tend to be similar to one another depending on the time scale of the phenomenon being measured), it is often difficult to have each element of the sample be representative of the span of the observations possible in the population. Further, model errors are often such that samples from a model tend to be less variable than observed samples, so that a sample derived from model fields will not be representative of the observed population. Selecting the sample for analysis is a critical aspect of organizing the data and depends on the question to be addressed by the study. If we want to examine the frequency of occurrence of major snow storms at Salt Lake City, does it make sense to include data from the entire year or limit the analysis to data from the cool season only?

An obvious rule of thumb is that your sample should be large enough to capture the phenomenon of interest many times. "Degrees of freedom" refers to the number of independent elements in

the sample; the number of degrees of freedom is usually much smaller than the total number of members in the sample in environmental data sets. You may want to use only a fraction of the total data available to begin your analysis. Then, the remaining data will be an independent sample that you can use to evaluate and confirm your results. Alternatively, procedures can be used to rerun the analysis hundreds of times omitting randomly data each time in order to develop confidence in the results. All too often, people assume that their sample is drawn randomly from the population, when in reality, their sample grossly underestimates the variability inherent in the population.

To illustrate these points, consider Fig. 1.3. Assume that we are measuring some phenomenon for which the population consists of only 1 possible "true" value equal to -2. So the population mean μ (average of all of the true values) in this case is the same as each individual value and equal to -2. Then, we make a total of "n" repeated measurements, which is our sample. One of those measured values, x_i equals 2, so the measurement error for that specific case is 4. The bell curve shows the likelihood of any particular value in the sample. It is more likely in this made up example that we have a value near 0 and never measure values greater than the lower and upper limits. Then, the sample mean is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In our example, the sample mean is 0, so we have a systematic error or bias of 2. As the width of the bell curve narrows, the precision of the sample increases. As the mean of the sample shifts towards -2, then the accuracy of the sample increases.

It is also important to recognize when events may really be independent of one another. The probability of getting heads when you flip an unbiased coin is always 50%, no matter if you have had 10 heads in a row before.
Many times people confuse streaks, or "hot hands" as being real when they are not. Casinos stay in business because gamblers' perceptions differ from reality- a person's intuitive misunderstanding of causality (I'm wearing my lucky shoes) departs from well-founded odds of a likely (or unlikely) outcome. And, clusters of events do happen by chance- two people could be winning at the same blackjack table *by chance*, not because it is a lucky table. Pregnancy complications (an unlikely event for any one mother) might occur in one town at a higher rate than "normal" *by chance*, independent of local environmental conditions, simply because lots of pregnant women live in lots of towns.

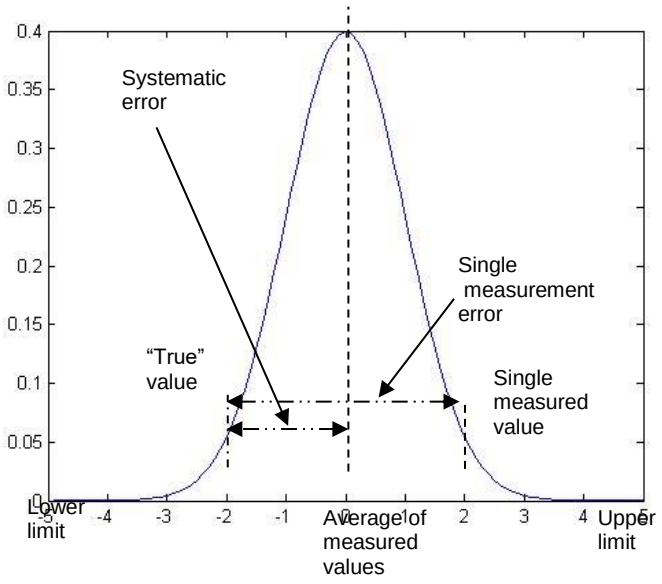

Figure 1.3. A sample of observations. The bell curve reflects the probability of a particular value in the sample.

d. Reducing Dimensionality

Statistical analysis of environmental data typically involves reducing the dimensionality of the data to a manageable size. Which variable(s) do we need to consider? Can we consider one variable (univariate analysis) or must we consider multiple variables (multivariate analysis). What time scales are we interested in? Hours, days, months, years? And, what region (local, regional, globally) or level in the vertical (surface, subsurface, upper air)? Are the data available on a spatial grid or at specific points?

Statistics often is misinterpreted as bookkeeping. What is the warmest temperature on record at Salt Lake City? What is the biggest snow storm at Alta? To even begin to answer those questions raises a number of issues. What period of record are we considering? What instrumentation? What about siting issues: where are the observations located (airport, downtown, Alta, etc.)?

We do need to distinguish between weather and climate:

- weather- state of the atmosphere and, more broadly, state of the environment
- climate- aggregate summary of the weather, or, more broadly, aggregate summary of the environment

How we go about aggregating the sequence of states of the environment is a critical issue. Much of the present climate framework in the U.S. is tied to somewhat outdated practices: climate normals are defined in terms of 30-year summaries that shift over time. "Normal" refers to the mean, the average of the values observed over time. It is designed to tell someone what the typical weather for a location might be. But, we'll see later that the mean is not always a good measure of the typical weather- it is sensitive to occasional outliers (either real ones or ones that arise from failing to identify erroneous observations). Focusing on the extremes (record highs and lows) can also be misleading. Is it critical if the temperature drops in a few minute period to some really low value, or is it more important if it stays at a slightly higher value for several hours? Many of our current practices of recording mean and extreme conditions is simply an attempt to reduce the dimensionality of the volume of data available. However, with the sorts of computing resources now available, we should not feel so constrained to follow such practices, and, at the least, be wary of how those practices may constrain our understanding of the way the environment works.

Statistical approaches are useful for weather and climate because:

- both are controlled by innumerable factors, which we hope to segregate into a few critical factors from the rest that, for the most part, simply contribute to background noise
- the characteristics of the system include linearly unstable processes that cause growth of small features into larger ones (a topic covered in other classes)
- the characteristics of the system (dynamics, thermodynamics) are nonlinear and include discrete step functions (i.e., rain/no rain) that can lead to the amplification of small errors into large ones
- the system is dissipative, which guarantees "stationarity", i.e., the climate system will remain stable and not run away from the current state. Global warming is not going to cause earth to turn into Venus, for example.

So, if we want to predict some future outcome, what variables must we consider? What locations? Over what time interval? Dealing with location (horizontal and vertical), time, and variable simultaneously is best left to numerical models of the entire climate system. The nonlinearities and instabilities make the environment unpredictable after characteristic times that differ among the various subcomponents (atmosphere, ocean, ice, etc.). The combination of noise and damping in the environmental system makes statistical predictive approaches credible. Deterministic approaches to forecasting future states of the environment (i.e., that the system is known and predictable at short lead times once the initial state is specified) may be less accurate than statistical approaches, which typically presume that there are a range of likely outcomes given the uncertainty inherent in the system.

d. Descriptive vs. Inferential Statistics

The distinction between descriptive and inferential statistics ties together many of the above points:

- Descriptive- organization and summarization of data, which may include using statistical models to interpret the volumes of data

- Inferential- figuring out why the environmental system behaves the way it does using data.

No matter what your career goals may be now, I expect that you will be faced many times with the task to extract information from environmental data. None of you are likely to develop a new coupled model of the earth-atmosphere system from scratch. All of you will likely be handed chunks of the vast environmental observational data that has been and will be collected during your career. An effective evaluation of data entails the use of statistics descriptively and inferentially. The process of data analysis requires brute force detective work: being organized and thorough as well as following through and experimenting with different approaches to examine the data. It also requires insight and intuition into the workings of the environment: you must understand the goals of your investigation and the flexibility to adapt as your understanding improves during preliminary analyses.

Statistical inference is the goal- we want to draw meaningful conclusions from the data. That means we need to have a plan and an idea about what to expect when we analyze the data. We need hypotheses. However, *statistical analyses can never prove anything- but, we can reject incorrect hypotheses*.

In court, you are expected to be assumed initially to be innocent (not guilty). That is the *null hypothesis*, you are not guilty. The prosecutor's job is to get the jury to reject that you are innocent and provide an alternative hypothesis that you are guilty beyond a reasonable doubt. The defending attorney's job is not necessarily to prove you are innocent, all that is necessary is to come up with any of a multitude of alternative hypotheses, such as someone else did it or raise the uncertainty about your level of guilt above reasonable doubt. We will see that "beyond a

reasonable doubt" for a statistical inference is related to how often we might expect something might happen by chance- once in 20 cases, once in one hundred cases, etc.

We have to be very careful about how we infer a meaningful conclusion. If we want to be really careful and only reject the null hypothesis at a high level of certainty (avoid false positives by saying it can only happen once every ten thousand cases), then there are going to be more times when we should have rejected the null hypothesis and didn't (false negatives). In other words, guilty people will go free more often to insure that no innocent person is sent to jail. Perhaps, a less threatening example is designing a spam filter. Your spam filter begins by assuming every email is not spam. You should be resigned to let through a few unwanted emails to avoid blocking an important email.

e. Putting Statistics to Work

This is not intended to be a programming class. We will use the computer lab to facilitate understanding how to apply statistics to environmental problems. However, an objective of this course is to help improve your programming skills to evaluate data. Employers expect people to be able to work independently and solve problems.

Nearly everyone should use Matlab. However, some of you may want to complete some of the assignments using Python. My recommendation is that nearly all undergraduates should use Matlab as it tends to be an easier computing environment. Graduate students in the first half can do some of the assignments in either Python or Matlab, but the second half of the course requires using Matlab. Bottom line- Matlab is recommended for nearly everyone this year.

The text, *Matlab Recipes for Earth Sciences* by Martin Trauth, is a great resource and the new 4th edition is quite good. Don't try to take this course without using it. It is available for free from the Mariott Library. Go through Chapters 1 and 2 immediately. It is important that you first understand what you need to do. Then, take advantage of Matlab to relieve the tedium of doing the calculations. That may often require you to do some of the calculations manually and then make sure and verify what you get from Matlab matches that.

If you decide to use Python for some of the assignments, consider getting the book, *Python Programming and Visualization for Scientists* by Alex DeCaria. It is just a general programming book. Graduate students more familiar with R can also use that language.

Share code features with others (and us!), if you do so. *However, using someone else's code is plagiarism. If you work in a group, make the final product yours, not someone else's code copied verbatim.*

As with all programming, there are good ways to approach an assignment, as well as very frustrating approaches. Here's a few reminders:

1. understand what you are expected to do. Ask questions and get clarification before trying to write code.

2. look at the example code, run it interactively, and understand what it does. You won't be expected to start from scratch. Again, ask questions if something is not clear and pay close attention to the details of the example code.
3. Read the notes and text and look online for techniques and tricks if you're stuck. There are lots of resources to solve problems.
4. Don't assume there is only one way to do something, but, recognize I am generally expecting a particular approach should be used. Let me know if you've found an alternative approach that works.

Programming is an iterative process. You do something, it doesn't work, you make a change, and it still doesn't work. Then what? First, when did it stop working? Did the example code do what it was supposed to do? Then, what did you change? Have you looked at the variables? Did you look at some intermediate values? Don't assume that just because you didn't get an error message that everything is coded correctly. Check your results. Do they make sense? Don't expect us to debug your code. Look very carefully at what you are doing and then ask others in the class if you are still struggling.

e. Navigating Online, Communicating and using your own laptop

The syllabus and assignments are online in Canvas: https://utah.instructure.com/courses/541131 Canvas is good to handle assignments and know what is due and when. Downloading code from Canvas can be awkward, so I will also be using a github repository: https://github.com/johnhorel/atmos_5040_2019 . The first thing you will do each class period will be to download the repository to your classroom Mac. Then you will have the example code and class notes at your fingertips. You will want to upload to Ubox or a USB stick the codes you develop in class.

I will be using Slack to provide quick updates to everyone or chat individually with you. The Slack workspace for this class is: https://atmos-5040.slack.com/ You will receive an invite early in the course about that.

The student Matlab license is a good deal ($30 annually) if you want to be able to complete assignments away from UU computer labs. You can also follow the instructions in the github files to configure Python on your computer. FYI- that is complicated.