3a Probability

a. Definitions

We are inundated with probabilities in environmental fields as well as society. The chance of rain is 50%- what does that mean? The chance of lung cancer in bald males who smoke is XX%. Probabilities should be defined carefully. We begin with some definitions.

- Event- set or class or group of possible uncertain outcomes. Rain/no rain. Temperature greater than 50^{o}F, etc.
- Elementary event- cannot be decomposed into other events
- Compound event- decomposable into 2 or more elementary events or other compound events
- Null event- that which cannot occur

Example: roll 6 sided die. (1) elementary event- 1 spot comes up; (2) compound event- odd number of spots comes up (1, 3, or 5); (3) null event- getting a 7 on a 6 sided die.

Will precipitation occur tomorrow? That is an elementary event if the only other choice is no precipitation. However, a compound event would be: will precipitation greater than 0.1 inch occur (it could rain more or could rain less or not at all) or will it snow or rain or both?

- S- Sample or event space. Set of all possible elementary events or the largest possible compound event
- Mutually exclusive- two events that cannot occur at the same time
- Mutually exclusive and collectively exhaustive events (MECE)- no more than 1 event can occur and at least one event will occur

b. Venn diagrams

Venn diagrams are a convenient way to display the sample space and make sense of the event outcomes that are possible. The NCDC storm event climatology (http://www.ncdc.noaa.gov/stormevents/) is a rich resource for examining weather events. I went through the reports for Salt Lake County from the NCDC Storm Event climatology and counted up the number of cases reported of winter and summer (convective) storms and those storms with property damage greater than $5000 during the thirteen year period 1993-2005. Now, I made some assumptions along the way as far as how to count events- some winter storms events may have been multiple day events, for example, and I counted lightning as being associated with convective storms. I ignored some iffy cases where it could have been a convective winter storm. Property damage has occurred from "other" storms and obviously the results might have been different if I used another $ damage threshold. In any event, there were a total of 142 winter storms and 83 summer storms as defined by me. 79 winter storms had damage in excess of $5000 while 25 summer storms had damage of similar amount. Given the nearly 5000 days during the 13 year record, these major weather events as defined by NCDC are not very common in Salt Lake County. The Venn diagram helps to highlight that winter storms are associated with

property damage more frequently than summer storms in Salt Lake County and, as defined here, winter and summer storms are obviously mutually exclusive.

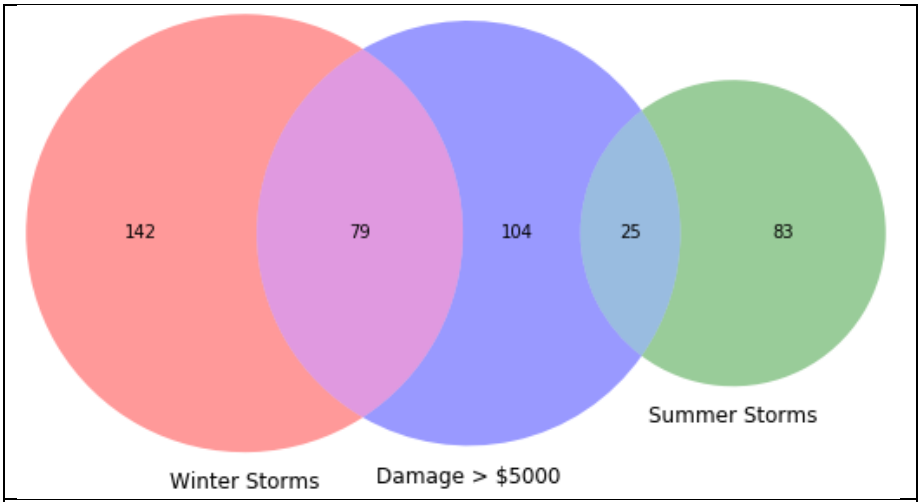


Figure 3.1. Venn diagram of major storms in Salt Lake County during 1993-2005.

Venn diagrams are useful for categorizing events that fall into clear categories and they don't need to be done in terms of circles. Consider Fig 3.2 that shows the possible MECE for a seasonal forecast of above/below normal temperature and precipitation for a specific location. All four possibilities are shown and the probability of each event will depend on the situation and location.

Seasonal Forecast Events

Temperature below Precipitation below	Temperature above Precipitation below
Temperature below Precipitation above	Temperature above Precipitation above

Figure 3.2. MECE possibilities for seasonal forecasts of temperature and precpitation anomalies for a specific location.

c. Probability Concepts

The following are pretty obvious, but when you get mathematicians involved, they have to have "axioms", lemmas, etc.

- probability of any event is nonnegative. In English: an event has to happen or else it is not an event
- probability of the compound event S is 1 or 100%. The probability that all events will happen is 1.
- probability that one or the other of two mutually exclusive events is the sum of their individual probabilities.

Definitions:

- Let E- event
- Pr{E}- probability of Event E; $0 \le \Pr\{E\} \le 1$
- Pr{E}=0 event does not occur
- Pr{E}=1 absolutely sure that event will occur

There are two approaches to probabilities: the frequency view and the Bayesian view. Which approach is used depends on the type of problem being investigated.

Frequency view- probability of an event is its relative frequency after many, many trials

- a- number of occurrences of E
- n- number of opportunities for E to take place
- a/n – relative frequency of event E occurring
- $Pr\{E\} \rightarrow a/n$ as $n \rightarrow \infty$
- Or a = outcomes = n Pr{E}

Examples: role a die. We expect the 6 spot to come up 1/6 times or 1 time every 6 opportunities. If we role the die 100 times, we expect the 6 to come up 16-17 times. What are the odds of drawing an ace? 4/52. So once every 13 times we expect to draw an ace. However, what we expect and what actually happens are clearly different things, that's where chance/randomness comes into play.

Bayesian view- probability represents the degree of belief or quantifiable judgement of a particular individual about an outcome of an uncertain event

- this approach recognizes that some events occur so rarely that there is no long-term probability estimate that are relevant
- Bookies make odds all the time based on their evaluation of the odds of winning for a particular team- it is not based on a large sample
- Two individuals can have different probabilities for same outcome

More concepts

- If event $\{E_2\}$ occurs whenever $\{E_1\}$ occurs, then $\{E_1\}$ is a subset of $\{E_2\}$
- Example: $\{E_1\}$- temperature below freezing; $\{E_2\}$- temperature below 50F, then $Pr\{E_1\} \leq Pr\{E_2\}$
- The complement of {E} is that event $\{E\}^c$ that does not occur
- $Pr\{E\}^c = 1 - Pr\{E\}$

What is the probability that $\{E_1\}$ and $\{E_2\}$ occur, that is, the intersection between the two events?

- $Pr\{E_1 \cap E_2\}$ = joint probability that $\{E_1\}$ and $\{E_2\}$ will occur *(3.c.1)*
- $Pr\{E_1 \cap E_2\} = 0$ if $\{E_1\}$and $\{E_2\}$are mutually exclusive
    - Example: if $\{E_1\}$is the occurrence of temperature below freezing and $\{E_2\}$ is the occurrence of temperature above 50^oF, then their joint probability is 0.

Let's return to the Venn diagram of the weather events in Salt Lake County. In the way that I went through the sample, the winter storms and convective storms are mutually exclusive, so there is no overlap between those two events. Assuming, that winter storms occur only during the winter half of the year and that convective storms occur only in the summer half (not great assumptions!), then the number of opportunities is order 180 days x 13 years= 2340 opportunities. Also, remember that some of the winter storms could be multiple day events, so I have some uncertainty and error in my results.

- $\{E_1\}$- occurrence of winter storms = 142
- $Pr\{E_1\}$= 142/2340 = .061

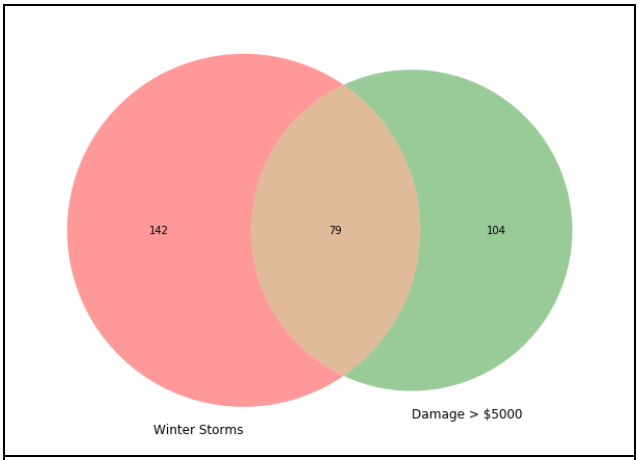


Figure 3.3. The overlap ($Pr\{E_1 \cap E_2\}$) has to be subtracted from the sum of the probabilities of each event occurring in order to determine if either occur ($Pr\{E_1 \cup E_2\}$).

- $Pr\{E_1\}^c = .939$
- $\{E_2\}$- occurrence of summer convective storms = 83
- $Pr\{E_2\} = 83/2340 = .035$
- $Pr\{E_2\}^c = .965$
- $\{E_3\}$- occurrence of property damage = 113
- $Pr\{E_3\} = 113/2340 = .048$
- $Pr\{E_3\}^c = .952$
- $Pr\{E_1 \cap E_2\} = 0$
- $Pr\{E_1 \cap E_3\} = 79/2340 = .034$
- $Pr\{E_2 \cap E_3\} = 25/2340 = .011$

What is the probability that $\{E_1\}$ OR $\{E_2\}$ will occur? That is, one the other, or both will occur. This is referred to as the "union" of the two events.

- $Pr\{E_1 \cup E_2\} = Pr\{E_1\} + Pr\{E_2\} - Pr\{E_1 \cap E_2\}$ *(3.c.2)*
- As can be seen visually from the Venn diagram to the right, the joint probability is counted twice when the individual probabilities are summed, so it is subtracted once.

It is worthwhile to see how that is true algebraically as well. Add up each probability separately:

$Pr\{E_1 \cup E_2\} = Pr\{E_1\} - Pr\{E_1 \cap E_2\} + Pr\{E_1 \cap E_2\} + Pr\{E_2\} - Pr\{E_1 \cap E_2\}$ or
$Pr\{E_1 \cup E_2\} = Pr\{E_1\} + Pr\{E_2\} - Pr\{E_1 \cap E_2\}$

- If $\{E_1\}$ and $\{E_2\}$ are mutually exclusive, then $Pr\{E_1 \cup E_2\} = Pr\{E_1\} + Pr\{E_2\}$

A couple more identities that you should be able to visualize from a Venn diagram:

- $Pr\{(E_1 \cup E_2)^c\} = Pr\{E_1^c \cap E_2^c\}$
  - that is the area outside of both circles
- $Pr\{(E_1 \cap E_2)^c\} = Pr\{E_1^c \cup E_2^c\}$
  - this one is a little harder to visualize; it is everything outside of the intersection of the two events

Returning to the Salt Lake County storm data:

- $Pr\{E_1 \cup E_2\}$ = Pr{winter storms} + Pr{convective storms} - Pr{winter storms ∩ convective storms} = .096 since the two events are mutually exclusive the last term is 0
- $Pr\{E_1 \cup E_3\} = .061 + .048 - .034 = .082$ = Pr{winter storms or property damage or both}
- $Pr\{E_2 \cup E_3\} = .035 + .048 - .011 = .072$ = Pr{summer storms or property damage or both}

d. Conditional Probability

The storm data example for Salt Lake County indicates that some winter storms lead to expensive damage. So, given that a winter storm has occurred, what is the probability that damage has occurred? We are now limiting our sample to a smaller number of events, only the 142 winter storms. So, the probability is now the 79 damaging winter storms divided by the 142 total winter storms or 56%.

- Conditional probability: probability that $\{E_2\}$ will occur given that $\{E_1\}$ has occurred
- $Pr\{E_2 \mid E_1\} = Pr\{E_1 \cap E_2\} / Pr\{E_1\}$ *(3d.1)*

$\{E_1\}$ is called the conditioning event; if it doesn't happen, then we know nothing about the probability that $\{E_2\}$ will happen

Alternatively, we can write:

- $Pr\{E_1 \cap E_2\} = Pr\{E_2 \mid E_1\} \times Pr\{E_1\} = Pr\{E_1 \mid E_2\} \times Pr\{E_2\}$ *(3d.2)*

Whether we condition from the first or second event to determine the intersection of the two events is our choice and simply depends on the available data.

If two events are completely independent, such that the occurrence of nonoccurrence of one event does not affect the probability of the other, then

- $Pr\{E_2 \mid E_1\} = Pr\{E_2\}$ and $Pr\{E_1 \mid E_2\} = Pr\{E_1\}$

Then,

- $Pr\{E_1 \cap E_2\} = Pr\{E_1\} \times Pr\{E_2\}$ for independent events

If we have a fair coin, then the Pr{head} = .5. The second coin toss does not depend on the first, so Pr{10 heads in a row} = $.5^{10}$

Now for some simple examples using a standard deck of cards. The odds of drawing a specific card are: Pr{ace} = 4/52; Pr{10-K} = 16/52; Pr{2-9}= 32/52. Now deal two cards facedown. You haven't looked at the first card, so you don't know if you have already drawn a specific card. The odds for the second card are the same as if it was the first card. Then, the various probabilities are:

	First Card			
Second card	2-9	10-K	Ace	totals
2-9	37.9%	18.9%	4.7	61.5%
10-K	18.9%	9.5%	2.4%	30.8%
Ace	4.7	2.4%	.6%	7.7%
totals	61.5%	30.8%	7.7%	100%

What is the probability that you will get twenty-one? The probability for each card are independent events, so Pr{ ace ∩ 10-K} = 4/52 × 16/52 = 2.4% and Pr{10-K ∩ Ace} = 2.4%. So, there is a 4.8% chance in getting a twenty-one. Every 100 hands, you should get a twenty-one about 5 times (4.8%) times and about 38 hands with 2 cards being 2-9, etc. What is

probability of getting twenty-one from 2 successive deals if you reshuffle? Pr{2 twenty ones} = $.048^2 = 0.23\%$.

e. Persistence

Persistence is the existence of statistical dependence over time (or space), i.e., that once a phenomenon begins it does not necessarily end before the next observation time or, that the observations at one location are related to the observations at a nearby one. Observations from environmental fields should not be considered to be independent of one another unless care is taken to choose a sample taking into account spatial and temporal dependence.

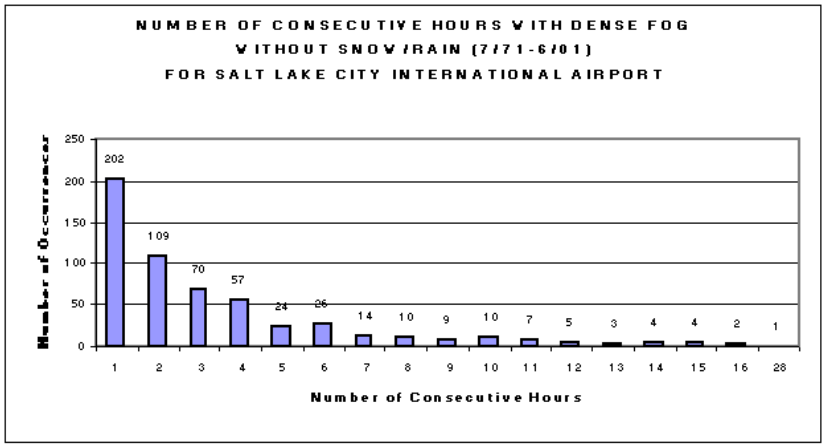


Figure 3.4. Fog climatology at the Salt Lake airport.

Consider the fog climatology shown in Fig. 3.4 that was created for the 2002 Olympics by Jonathan Slemmer. We wanted to provide the Olympic forecasters with some information on the likelihood of persistent heavy fog at the airport. If it happened during the Olympics, then there would have been a bunch of negative consequences with flight delays, etc. (fortunately, it didn't happen during that period). Dense fog doesn't happen very often. The sample here is large: 31 years x 365 days= 11315 days. Dense fog happened over a 2 hour period (Jonathan labels this 1 h of consecutive fog) on only 202 days. So Pr{ 1 hour of consecutive fog} = Pr{1}= 202/11315 = 1.8% of days. If we considered the number of hours in a day as well, then the probability that it would happen in a specific hour would be correspondingly less. Pr{2} = 109/11315 = .96% of days, etc.

The probability that 3 hours in a row of fog (2 consecutive hours) is going to happen is pretty unlikely = .96% for any given day. However, if 1 hour of consecutive fog has already happened, then the odds of it continuing are obviously much higher.

- Pr{2|1} = 109/202 = 53.9%, Pr{3|1} = 70/202 = 35%, Pr{10|5} = 41.6%, etc.

Persistence is a good statistical forecasting baseline. When my family asks me what the weather is going to be like tomorrow, without any other information available, I'm going to say whatever it is today. And, just to let you know, my family doesn't think I'm a very good forecaster. A forecaster adds value when the conditions are present that lead to change and those conditions are correctly recognized as such. While it may be useful to tell an airport operator that the current dense fog has a high likelihood of continuing, more value will be added if the forecaster has information available from which to diagnose when the fog is going to break up.

Later, we will examine ways to estimate the probability of rare events. For example, it is not particularly useful to develop probabilities on the occurrence of dense fog for over 20 consecutive hours based on the 1 event that has happened in our sample. Similarly, we can't wait around for a 100 years to estimate the occurrence of a once in a hundred year flood.

f. Forecast Verification

We'll touch on verification of forecasts from several angles. I'll introduce the difference here between a "measures oriented verification approach" (typically using well-established and often over-used performance metrics such as hit rate or threat score) vs. a "distributions-oriented approach" (where the empirical joint distributions of forecasts and verifying observations are generated).

Let's start with the simple approach that something happens or it doesn't and we forecast it to happen or not. We then count the number of cases for the four possibilities:

		Observed	Observed	**Forecast marginal totals**
		Yes	No	
Forecast	Yes	a	b	**a+b**
Forecast	No	c	d	**c+d**
	Observed marginal totals	**a+c**	**b+d**	**n=a+b+c+d sample size**

First, the marginal totals are important- they are how often something is observed or forecast (or not observed/not forecast). How often do we get the "right" forecasts and how often do we forecast them and they don't happen?

PC = percent correct = $\frac{a+d}{n}$

FAR = false alarm ratio = $\frac{b}{a+b}$

We want the percent correct to be close to 1 and the false alarm ratio to be close to 0. But, what happens when we are verifying categorically forecasts of large precipitation amounts (say over an inch) at Salt Lake City? That doesn't happen very often, but the percent correct may be large because we may be very successful at forecasting when the precipitation is less than an inch (d). Then, to focus on the situations when it either was observed or forecast, the threat score (TS) or critical success index (CSI) is used:

TS = CSI = $\frac{a}{a+b+c}$

The threat score measures the number of correct "yes" forecasts relative to the total number of occasions on which the forecast was forecast or observed. Another metric commonly used is the probability of detection (POD) or hit rate (HR), which identifies how frequently an event is forecast relative to when it is observed:

POD = HR = $\frac{a}{a+c}$

Note when we are using one of the marginal totals in the denominator, we're computing a conditional probability. We can express the hit rate as: given that an event occurs, how often is it correctly forecast? The FAR is: given that an event is forecast, how often did it not happen?

Now let's look at an example using Matt Lammer's research on verifying forecasts made in support of prescribed burn and wildfire operations: http://meso1.chpc.utah.edu/jfsp/

On January 30, 2015, there were 77 forecasts issued in support of prescribed burns nationwide. How often did the forecasters anticipate high wind speeds (≥ 5m/s) later that afternoon to take place relative to what was observed that day?

		Observed	Observed	**Forecast Marginal totals**
		≥ 5m/s	<5 m/s	
Forecast	≥ 5m/s	11	6	**17**
Forecast	<5 m/s	16	44	**60**
	Observed Marginal totals	**27**	**50**	**77**

So, the PC= 71.4%; FAR= 35.3%; TS= 33.3%; and POD = 40.7% . How did they do? Clearly, the percent correct is high because they forecast correctly a lot of cases when the winds were light. The false alarms are not too bad, 6 of the 17 forecasts of high wind. But, the threat score is low because they missed a lot of cases when the winds were observed to be stronger than they were expecting.

Do these forecasts have skill? Statistical skill refers to the relative accuracy of a set of forecasts with respect to reference forecasts (random, persistent, or climatological, for example). The probability of a correct yes forecast by chance (meaning that the observations and forecasts are independent) is just the product of the marginal probabilities of the observations and forecasts:

Random correct yes forecast by chance = $\frac{(a+b)}{n}\frac{(a+c)}{n}$

Random correct no forecast by chance = $\frac{(b+d)}{n}\frac{(c+d)}{n}$

In our case, the odds of having a randomly correct yes forecast is low (7.7%) but the odds of having a randomly correct no forecast is pretty high (50.1%) since it is both observed and forecasted to not be windy frequently.

The most generic of skill scores is of the form: $SS = \frac{(correct\ forecasts - random\ correct\ forecasts)}{(total\ forecasts - random\ correct\ forecasts)}$

The Heidtke Skill Score is of this form and can be computed after some substitutions from the contingency table values as: $HSS = \frac{2(ad-bc)}{(a+c)(b+d)+(a+b)(b+d)}$

In our case, HSS= 31.4%, which is not particularly high and reflects that low wind speed forecasts don't require a lot of skill.

The measures-oriented metrics defined above are ok, but much information is lost by looking only at 2x2 contingency tables. Let's broaden the scope a bit and assume that an accurate forecast is one when the forecast wind speed is within 2 m/s of the observed forecast.
So, most frequently, the forecast errors are up to one m/s weaker than those observed. And, as we determined from the earlier metrics, there is a greater tendency to forecast the wind speeds to be lower than those observed on this particular day. However, we don't know from Fig. 3.4 whether the forecasters do a better job over some ranges of observed wind speeds than others. We can expand the contingency table concept to create a "distributions-oriented" approach to verification as shown in the following table. I've now arbitrarily decided an accurate forecast is within ±2 m/s of that observed and then keep track as well of the sign of the errors that exceed that limit. I've broken up the observed wind speeds into 3 categories as well. Now it is clearer that the forecasters tend to underforecast higher wind speeds and don't ever overforecast high wind speeds, only more moderate ones.

Table 3.1. Distribution-oriented verification of wind forecasts

	E_1	Observed	Observed	Observed	**Error Marginal totals**
E_2		≤3 m/s	3-6 m/s	≥6 m/s	
Error	≤ -2 m/s	0	10	11	**21**
Error	± 2 m/s	22	20	7	**49**
Error	> 2 m/s	0	7	0	**7**
	Observed Marginal totals	**22**	**37**	**18**	**77**

This is a MECE data set for this particular sample of forecasts issued on this single day. If we were to divide the counts in the interior bins by the sample size (77), then those interior bins would be joint probabilities, e.g., 26% of the forecasts were within 2 m/s when the wind speeds were between 3 and 6 m/s (20/77). A lot more information can be gleaned by considering the conditional probabilities as defined by 3c.1 and 3c.2. For example, given that the observed wind speed is greater than 6 m/s ($\Pr\{E_1\}$ = 18/77= 23.4%), the probability that the forecasters predict the winds to be too light $\Pr\{E_2 \mid E_1\}$ is:

$$\Pr\{E_2 \mid E_1\} = \Pr\{E_1 \cap E_2\} / \Pr\{E_1\} = ((11/77)/(18/77)) = 64.7\%$$

And, here's where the interpretation of conditional probabilities can get out of hand. On this particular day, given that the error is greater than +2 m/s, then the probability that the wind is in the 3-6 m/s category is 100%!!

$$\Pr\{E_2 \mid E_1\} = \Pr\{E_1 \cap E_2\} / \Pr\{E_1\} = ((7/77)/(7/77)) = 100\%$$

Hooray, we can say all forecasters tend to overforecast high winds when the winds are between 3-6 m/s (no- we can't).

Now, imagine only one in ten thousand people will get a particular disease- $Pr\{E_1\}$. But you hear on the news that 50% of the people that come down with the disease ate jello that day- $Pr\{E_2$- ate jello $| E_1\}$. Should you stop eating jello to avoid catching the disease? $Pr\{E_1 \cap E_2\} = Pr\{E_2 | E_1\} \times Pr\{E_1\} = .50 * .0001 = .005\%$ Don't focus on that eating jello seems to cause an alarming increase in risk; the more important issue is the low risk factor for this particular disease under any circumstance.

g. Summary

Probabilities are at the heart of modern weather forecasting as well as many other environmental applications. While many applications and users will continue to expect to hear on the radio what the temperature will be at 4 PM tomorrow, the underlying information from which a forecaster will base that specific number will likely be probabilistic information. For example, forecasters implicitly use conditional probabilities as part of the forecast preparation. Given the approaching front, and given that a specific model has a known bias in temperature in the prefrontal environment, they expect the temperature to be higher/lower than what would normally take place.

3b. Theoretical Distributions and Hypothesis Testing

a. Parametric and Empirical Probability Distributions

The empirical histograms and cumulative density distributions discussed in Chapter 2 have many applications but they are determined from a sample of the population. Parametric probability distributions are a theoretical construct using mathematical relationships to define populations with known properties. One or two parameters combined with the assumption that the population is composed of random events may be enough to define the occurrence of possible outcomes of an environmental phenomenon. By comparing parametric and empirical probability distributions, we can deduce additional information about the population from which a sample is taken. The advantages of applying parametric distributions include:

- compactness- we may be able to describe a critical aspect of a large data set in terms of a few parameters
- smoothing and interpolation- our data set may have gaps that can be filled using a theoretical distribution
- extrapolation- because environmental events of interest may occur rarely, our sample may not contain extreme events that could be estimated theoretically by extending what we know about less extreme events

But keep in mind that while parametric distributions have advantages, they also can instill a level of confidence about your understanding of a phenomenon out of proportion to what really can be known.

Roman letters (e.g., s- sample standard deviation) are used to define sample statistics while Greek letters (e.g., σ- population standard deviation) are used to define the population statistics.

Since parametric probability distributions are a theoretical construct that hopefully describes the population, the parameters used to define them are generally given by Greek letters.

Many environmental phenomena are discrete events: it either rains at a particular location or not; a tornado touches down or not; an earthquake happens in a location/time or it doesn't. There are a large number of parametric distributions (binomial, Poisson, etc.) appropriate for examining a data set of discrete events. Because of the limited time available in this course, we are not going to discuss discrete parametric distributions (see Wilks for further details). On the other hand, most environmental variables of interest can be defined as being continuous: whether it rains or not is part of a continuum of how much it rains; we can classify temperature above or below a threshold as a discrete event but temperature varies continuously over a wide range of values; earthquake intensity is defined continuously on the Richter scale. There are a suite of parametric distributions (Gaussian, lognormal, gamma, Weibull, etc.) that are relevant to continuous distributions.

It is important to recognize the steps involved in using parametric distributions:

- generate an empirical CDF
- determine a good match between the empirical CDF and a particular parametric distribution
- use the parameters from that parametric distribution to estimate the probabilities of values above or below a threshold, likelihood of extreme events, etc.

Traditionally, applications of parametric distributions required lookup tables; statistics books are full of such (e.g., Appendix B of Wilks).

We begin by defining the probability density function (PDF) for a random continuous variable x as f(x), which is the theoretical analog of the histograms in Chapter 2. The sum of f(x) over all possible values of x is $\int_{-\infty}^{\infty} f(x)dx = 1$. As with the interpretation of integrals in general, think of the product $f(x)dx$ as the incremental contribution to the total probability. The shaded area shown in Fig. 3.5 represents $\int_{.5}^{1} f(x)dx$ and represents 15% of all the possible values. The cumulative distribution function (CDF) is the total probability below a threshold, hence, the total area to the left of a particular value:

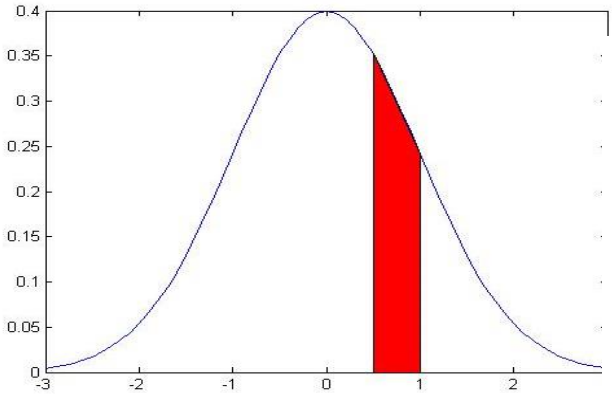


Fig. 3.5. Probability density function.

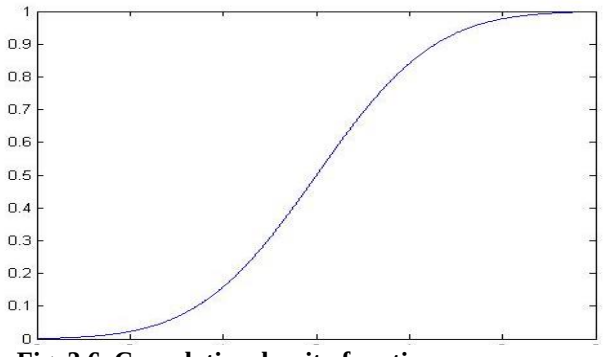


Fig. 3.6. Cumulative density function

$F(X) = \Pr\{x \le X\} = \int_{-\infty}^{X} f(x)dx$. For example, for the CDF in Fig. 3.6, the cumulative probability of negative values is 50%. Also, it is useful to define X(F) as the value of the variable corresponding to a particular cumulative probability, e.g., from the figure X(75%)=.66.

The function that defines all possible values of X(F) is referred to as the quantile function. The expected value, E, of a random variable or function *of a random variable* is the probability-weighted average of that variable or function.

- $E[g(x)] = \int_{-\infty}^{\infty} g(x) f(x)dx$

Consider this intuitively as weighting the values of g(x) by the probability of each value of x. A reminder of a few integral properties:

- for a constant c, $E[c] = c$ since the sum of f(x) over all values of x is simply 1
- for g(x)=x, $E[x] = \int_{-\infty}^{\infty} xf(x)dx = \mu$: μ is the mean of the distribution whose PDF is f(x)
- $E[cg(x)] = c\int_{-\infty}^{\infty} g(x) f(x)dx$
- The contribution to the total variance from a particular value of x is $g(x) = (x - E(x))^2$. So, the total variance is

  $Var[x] = E[g(x)] = \int_{-\infty}^{\infty} (x - E(x))^2 f(x)dx = \int_{-\infty}^{\infty} (x^2 f(x)dx - 2xE(x) f(x)dx + E(x)^2 f(x))dx$

  $= E(x^2) - (E(x))^2 = E(x^2) - \mu^2$

We'll use the above relationships for several different continuous parametric distributions.

b. Gaussian parametric distribution

Each parametric distribution that you are likely to use has a rich tradition in statistics, none more so than the Gaussian distribution. The PDF in the previous subsection is that of the Gaussian distribution defined by:

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ for $-\infty \le x \le \infty$

and its CDF is

- $F(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{X} \exp(-\frac{(x-\mu)^2}{2\sigma^2})dx$

The two parameters that define the Gaussian distribution are μ and σ. Confusion often crops up as a result of outdated statistical terminology; the Gaussian distribution is often referred to as the normal distribution. However, that does not mean that the Gaussian distribution is what everything should follow- it is just one possibility of many.

Let's return to the GSL monthy lake level record. The values for μ and σ are estimated from the histogram plotted in Fig 3.7 and a Gaussian (normal) distribution is then calculated using the

sample mean and variance. Visually, you should be able to tell that the Gaussian fit in this instance is not particularly good, since the lake level is skewed (i.e., there are a few events of high water levels that would not be expected given the typical values of lake level and its spread about the sample mean). Also, there are fewer low water years than expected from the Gaussian distribution. A plot of the quantile function for lake level shown in the lower panel of Fig. 3.7 affirms that empirically we observe more high water years and fewer low water years than would be expected according to a Gaussian distribution with the sample mean and variance.

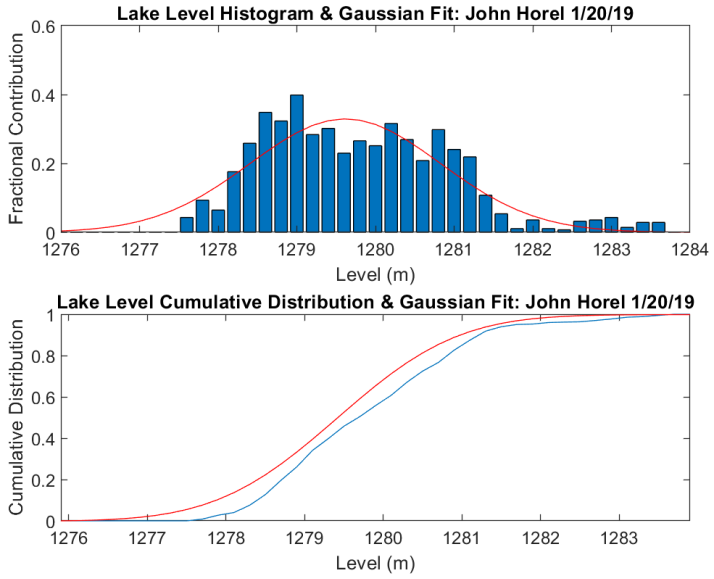


Figure 3.7. Gaussian fit to the monthly level of the Great Salt Lake in terms of its histogram (top) and cumulative distribution (bottom).

Let's examine the hourly temperature values at Collins (CLN) near Alta during winter from 1998-2005 as shown in Fig. 3.8. Although the Gaussian distribution underestimates the occurrence of temperature near the mean value, it appears that Collins winter temperature can be approximated by a Gaussian parametric distribution defined by the sample mean and variance. Note the occasional gaps in the histogram- the original data is in 1^{o}F intervals, so there are some 0.5^{o}C bins with no values.

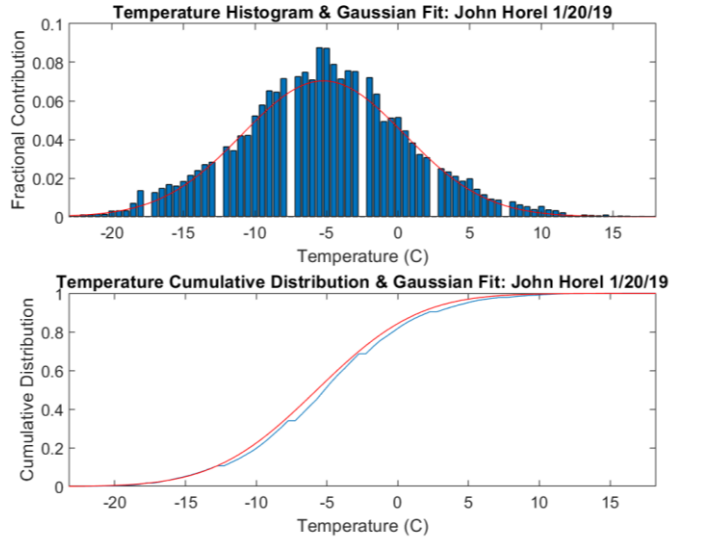


Figure 3.8. Gaussian fit to temperature at Alta in terms of its histogram (top) and cumulative distribution (bottom).

Now, let's return to generic Gaussian distributions. Every variable can be transformed into standardized anomalies with mean 0 and variance 1. The leftmost panel of Fig. 3.9 indicates that for an environmental variable for which the Gaussian is a good fit to its empirical PDF, then 68.3% of the total variance is within 1 standard deviation of the mean. The middle figure indicates that 95.5% of the total variance is within 2 standard deviations of the mean while the right figure defines that 2.3% of the time we would expect that a variable explained by a Gaussian distribution would be larger than 2 standard deviations of the mean. Alternatively, we can use the quantile function to determine the x values that correspond to a particular probability.

For example, if we are interested in the limits corresponding to 90% of the total variance, then that is equivalent to ±1.65σ of the mean.

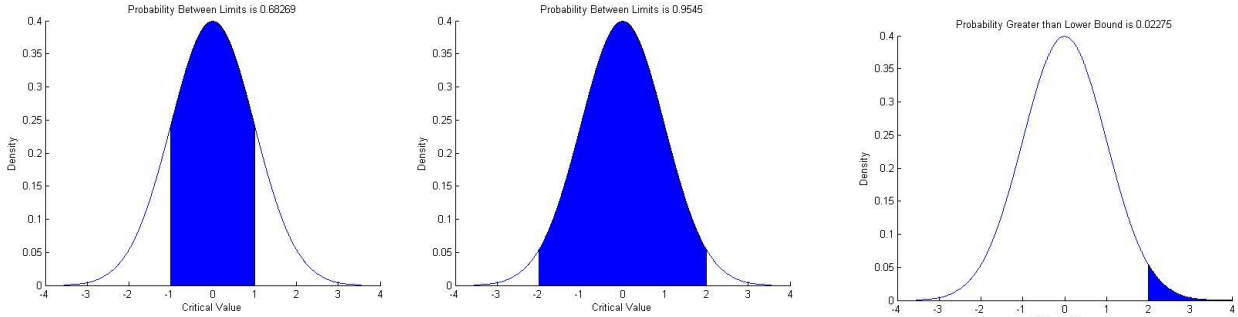


Figure 3.9. PDF's for the case when the sample mean is 0 and variance is 1.

c. Other parametric distributions

Many environmental variables (e.g., wind speed and rainfall) are decidedly skewed to the right in part because values are nonnegative. The gamma distribution with 3 parameters is quite versatile for such situations. Other variables (e.g., wind direction, relative humidity) are constrained at both ends for which the beta distribution with 2 parameters is an appropriate choice.

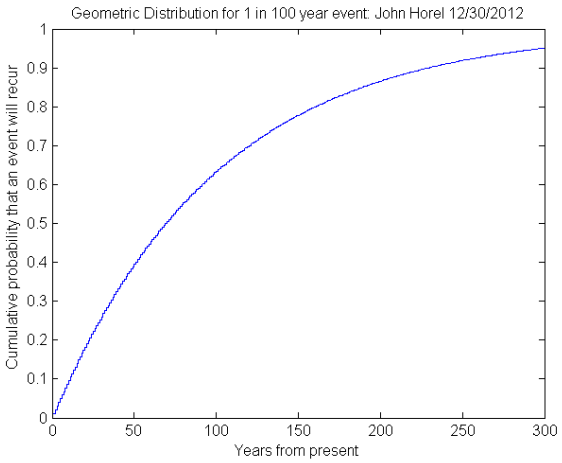


Figure 3.10. Cumulative distribution for the recurrence of a rare event- in this case a one in hundred year event assuming a geometric distribution.

Of interest for many applications, is to examine parametric distributions of extreme values, i.e., the rare events for continuous variables. There are a number of variants of theoretical distributions to describe extreme events: Gumbell, Fischer-Tippet, and Weibull, among others. However, these theoretical distributions assume random events that may not be appropriate for environmental events that often occur serially, e.g., an extreme heat wave typically will last several days in succession. If sufficient data are available, then the empirical PDF can be used to estimate the probability of rare events.

Extreme values are often defined to estimate the annual probabilities of damaging events such as heavy rains or high winds. The recurrence of extreme events is frequently defined in terms of the return period, i.e., 100 year floods, etc. However, there is no guarantee that a 100-year event will happen in the next 100 years. The probability of a 1 in a 100 year random event is Pr{0.01}. The geometric distribution specifies probabilities for the number of trials required until the next success (see Wilks). Fig. 3.10 shows the cumulative probability of the period until the next 100 year event. In other words, if the probability of a 100-year event is 0.01, then there is only a 63% chance that it will happen in the next 100 years after the last event and there is still a 12% chance that it will not happen in 200 years. If the probability of a rare event increases to 2%, then there is a 12% chance that it will not happen in 100 years.

As an example of evaluating the return period of extreme events, let's examine the peak streamflow record from the Klamath River for the 1911 to 2018 water years (Oct.-Sept.; hence December floods such as those in December 2005 are part of the 2006 water year) as shown in Fig 3.11. To what extent can we estimate the occurrence of extremely high peak flows on the Klamath River by a parametric fit to the data? We have an advantage here since we can estimate empirically what a one in a hundred year event is, since we have a record of at least a hundred years. That estimate is the blue line in Fig 3.11 determined as the 99th percentile from the empirical CDF that is shown as the blue curve in the right panel of Fig. 3.12. People often derive one in a hundred year events from records of 20 (or less) years based on parametric fits. We'll use this example to show how that can be done, but why this approach might not give a realistic answer.

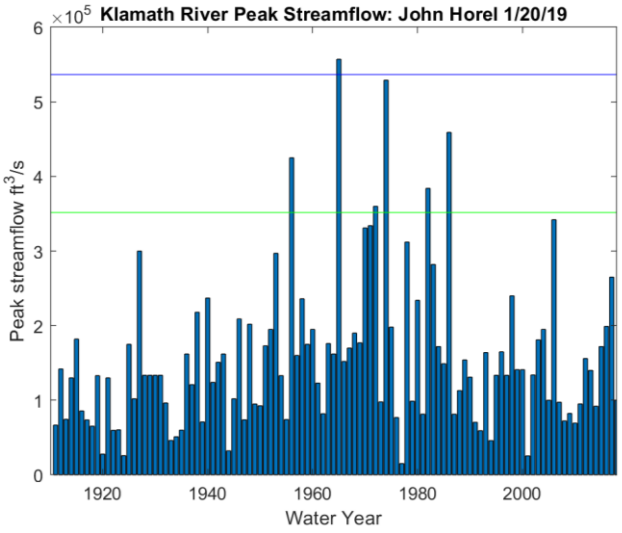


Figure 3.11. Peak streamflow during the water year for the Klamath River, CA. The "100 year" flow was observed in December 1964 (1965 water year). The green line is the Weibull estimate of the peak streamflow associated with one in hundred year event.

As shown in Fig. 3.12, a Gaussian parametric fit is a poor choice in this instance to describe the peak streamflow as it would estimate many more low peak flows than observed and fewer high peak flows. The Weibull fit does a better (but not perfect) job at capturing the skewed nature of the peak streamflow. The Weibull fit underestimates the occurrence of the really high peak streamflow events.

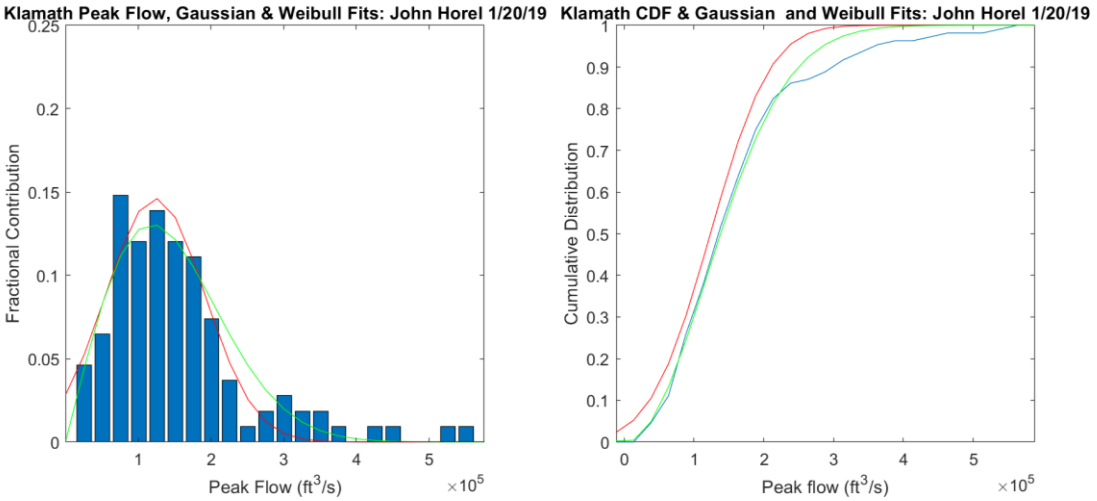
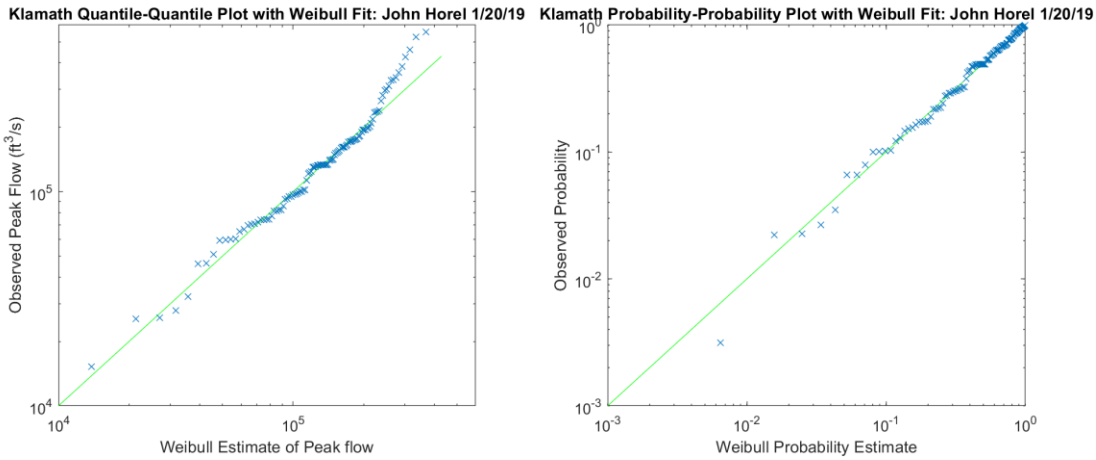


Figure 3.12. Empirical histogram (left) and CDF (right) of Klamath peak streamflow in blue. Red (green) curves denote Gaussian (Weibull) parametric fits to the data.

Another way of examining the "goodness" of a parametric fit is to look at quantile-quantile or probability-probability plots (Fig. 3.13). If the Weibull parametric fit was perfect, then all the

blue crosses (observed values) would lie along the green lines. In this case, the Weibull parametric fit underestimates the extremely high peak flows (blue crosses above the green line in the left figure). In terms of probability of occurrence (right panel), the Weibull fit seems better but would overestimate the very rare occurrences of the lowest observed flow in 1977.

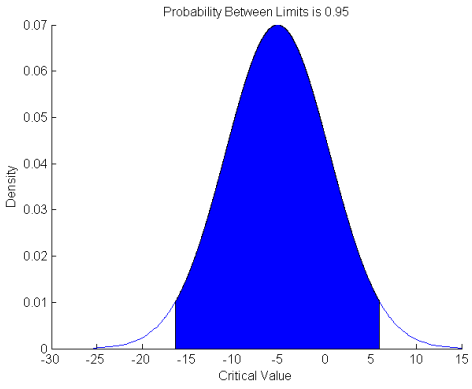
Figure 3.13. Quantile (left) and Probability (right) plots of observed Klamath peak streamflow in blue. Green straight lines show where the observed values should lie along if the Weibull parametric fit the data perfectly.



The green line in Fig. 3.11 is the Weibull parametric fit estimate of the peak streamflow associated with a one in a hundred year event (99th percentile). Because the Weibull approximation underestimates the very high peak streamflows, it is clear from Fig. 3.11 that 6 years would be classified as "one in a hundred year events", a clear overestimate relative to what has transpired. If we didn't have a hundred+ year record, we wouldn't necessarily know that.

d. Hypothesis testing and confidence intervals

People's perception of what is unusual often is heavily weighted by what has happened recently. "This storm was much stronger than anything before" or "I've never felt it be so cold". How can we provide information on whether something is truly extreme? Consider the Collins temperature record again. Since the Gaussian parametric fit captures the essence of the Collins data (see Fig. 3.8), we can use the information from the Gaussian fit, in this case the parameter estimates of $\mu = -5.2^oC$ and $\sigma=5.7^oC$. Then, Fig. 3.14



Shaded area denotes 95% of the values.

indicates that 95% of the time, temperatures randomly selected from a Gaussian distribution with that mean and standard deviation would fall between -16.4 and 6ºC. So, today's temperature at Collins is -20ºC. Your ski buddy says "Gee, it's really cold". Time for some hypothesis testing. Can you tell him it is unusually cold or not?

So, we define a null hypothesis that we hope to reject: today's temperature does not differ significantly from the mean temperature at Collins, namely -5.2ºC. The temperatures bounded by the red and green lines (-16.4ºC to -6ºC) contain the range of values for which we cannot reject the null hypothesis. Based on our sample of temperature at Collins and today's temperature, we can reject the null hypothesis accepting a 5% (1 in 20) risk that we are rejecting the null hypothesis incorrectly since -20ºC is outside of the bounded area.

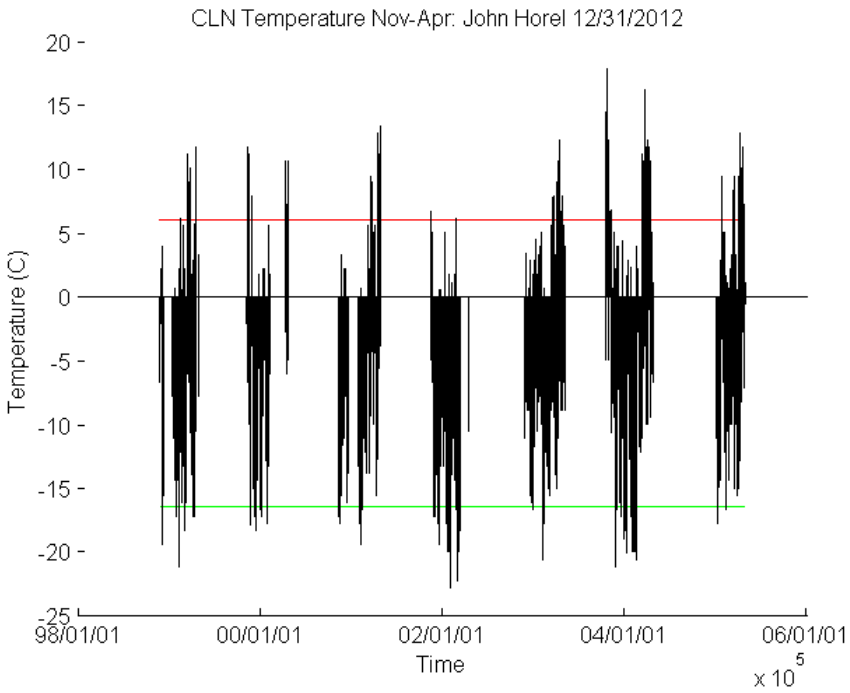


Figure 3.15. Time series of Alta Collins temperature (Nov-Apr) with 95% confidence intervals (red and green lines)

As shown in Fig. 3.15, time series of environmental data, such as Collins' temperature, are often depicted with upper and lower limits or "confidence intervals". These confidence intervals can be defined by the parameter estimates of Gaussian fits to the sample data, i.e., each specific value is shown relative to $\mu \pm 1.96\sigma$ in Fig. 3.15. We are assuming then that a random distribution with that mean and standard deviation would have 95% of the values between the red and green lines. In this instance, plotting all of the Collins data from November to April immediately tells us that the way we set up the hypothesis test is not very good. The high temperatures only occur at certain times of the year, when your buddy is less likely to be skiing. We should have limited our sample perhaps to only temperature during the core winter months- for that sample, the -20ºC might not be so unusual. Confidence intervals can be defined from other parametric fits as well, to express the degree to which specific data compares to the theoretical distributions.

e. Hypothesis testing of means

Let's return to the annual precipitation in Utah and use a completely arbitrary definition of a drought: that the average annual precipitation anomaly over a 3 year period differs substantively from zero. We will evaluate strings of the 3 year periods and try to objectively compare each three year period to the others.

One expectation might be that the mean precipitation anomaly during any of the 3 year periods is 0 - this would be the null hypothesis. The null hypothesis, H_0, defines a frame of reference against which to judge an alternative hypothesis, H_A, which in this instance could be "the mean precipitation anomaly during the past five years is not zero".

The steps required for a hypothesis test are:

- identify a test statistic that is appropriate to the data and question at hand. The test statistic is computed from the sample data values. In this example, the 3-year sample mean will be the test statistic, but we'll also need to use the sample variance as well.
- Define a null hypothesis that we hope to reject. In this case, the null hypothesis is that the sample mean is 0.
- Define an alternative hypothesis. In this case, the sample mean is negative.
- Estimate the null distribution, which is the sampling distribution of the test statistic if the null hypothesis is true. It is very important to recognize that we need to know the sampling properties of the test statistic. That is, the sample mean could be drawn from a Gaussian parametric distribution, another parametric distribution or even we could define the sampling distribution of the mean empirically by randomly sampling over and over taking three years within the past 124 years.
- Compare the observed test statistic (the composite mean value of each 3-year period to the null distribution. Either:
    - the null hypothesis is rejected as too unlikely to have been true if the test statistic falls in an improbable region of the null distribution, i.e., the probability that the test statistic has that particular value in the null distribution is small, or,
    - the null hypothesis is not rejected since the test statistic falls within the values that are relatively common to the null distribution.

Not rejecting H_0 does not mean that the null hypothesis is true; rather, there is insufficient evidence to reject H_0. The null hypothesis is rejected if the probability, p, of the observed test statistic in the null distribution is less than or equal to a specified significance (or rejection) level denoted as the α level. Usually, 1% or 5% significance levels are used, i.e., if the odds of the test statistic occurring in the null distribution are less than 1% or 5%, then we often reject the null hypothesis. Depending on how the alternative hypothesis is framed, rejecting the null hypothesis may be equivalent to accepting the alternative hypothesis; however, there may be many possible alternative hypotheses. The first step of any significance testing is to set an appropriate α level to reject the null hypothesis. In other words, you must first set a threshold, such as 1% that denotes a 1 in 100 chance that you are accepting the risk of rejecting the null hypothesis incorrectly. This 1% risk is a Type I category error of a false rejection of the null hypothesis.

f. Central limit theorem and student-t test

Now we consider one of the reasons the Gaussian distribution is used so much. First, roll 1 six-sided die 10,000 times. That's a population. The bars of roughly equal height in Figure 3.16a show that the chance of getting any one number from 1-6 is basically the same in that population (but as shown earlier they are not identical odds). Now roll 6 dice 10,000 times (Fig. 3.16b). In other words, we have a population of 10,000 samples of 6 events and can determine each

sample's sum, mean, or variance separately. Similarly, the mean of all those individual sample means is around 3.5. The odds of getting only a total count of 6 or 36 are small; most frequently we will get something around 21. Note that we end up with a Gaussian distribution. *The central limit theorem states that the sum (or mean) of a sample (6 dice) will have a Gaussian distribution even if the original distribution (one die) does not have a Gaussian distribution, especially as the sample size increases.* In other words, $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ where $\sigma_{\bar{x}}$ is the standard deviation of the sample means, σ is the standard deviation of the original population, and n is the sample size. For a large population, such as in this example, the standard deviation of the population is roughly 1.7 and that of the 6 member samples drawn from the population is roughly 0.7, which is what should be expected.

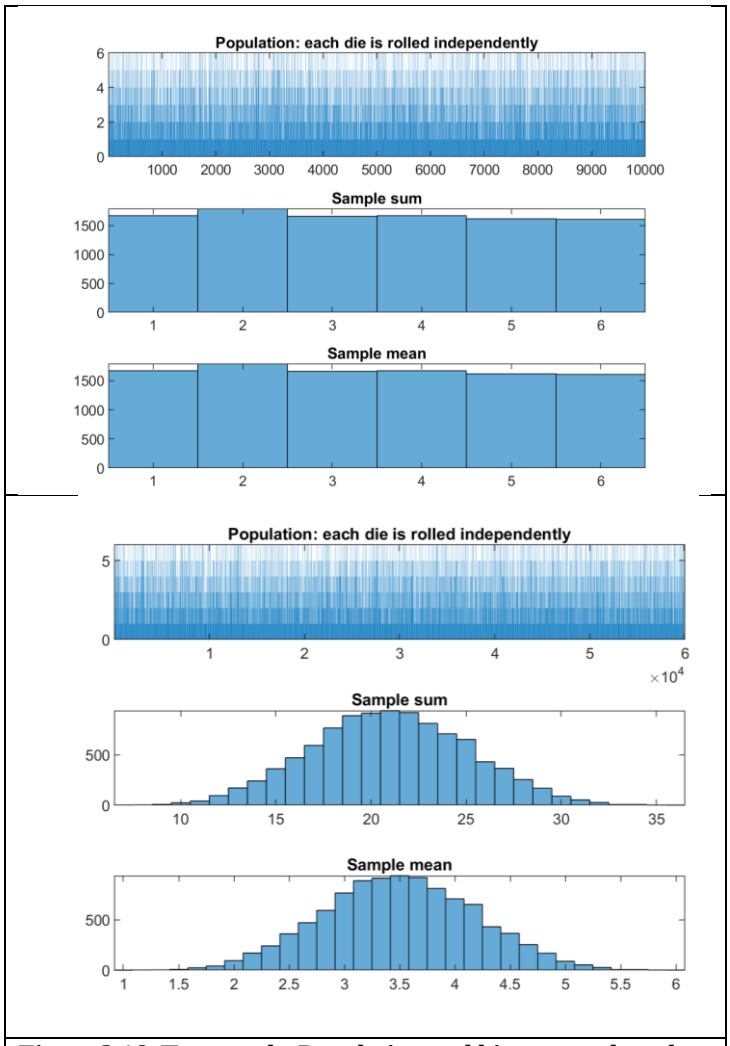
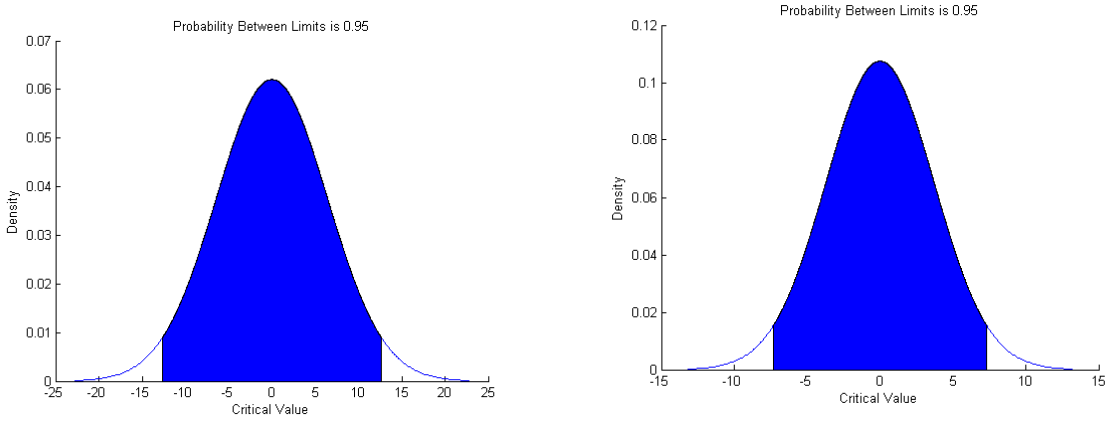


Figure 3.16. Top panels. Population and histograms based on rolling 1 die 10000 times. Bottom panels. Population and histograms based on rolling 6 dice 10000 times.

Let's return to our example attempting to determine 3-year drought periods. The sample standard deviation is 6.4 cm for the annual precipitation over the 124 years. We could randomly obtain the Gaussian distribution shown in the left panel of Fig. 3.17 with a standard deviation of 6.3 cm about the anomaly mean of 0. There is a 95% chance that the precipitation anomaly will lie between ±12.4 cm. We now randomly take 3 values and average them. If we selected 3 years at random from the population many times, then according to the central limit theorem, we'd end up with the right panel. There is a 95% chance that the 3-year sample mean would lie between ±7.2 cm. In other words, it becomes less likely to have an extreme 3-year mean ("a drought" according to this lame definition) than just to have one extreme dry year.

We use the central limit theorem as a way to determine whether a mean from a particular sample differs significantly from the mean we specify as being appropriate for the null hypothesis



assuming that we know something about the population variance. Assume for the moment that the population standard deviation was 6.3 cm as assumed in the left panel of Fig. 3.17. In the last 3 years, the annual precipitation anomalies are -.1, 1.5, .2 cm so the mean anomaly over the 3 years is .53 cm, obviously not a drought situation of late. Then we would determine that we could NOT reject the null hypothesis at the 5% level, since the sample mean during the last 3 years of .53 cm first is positive and also lies within the shaded area in the lower panel. If we go back to 1900-1902 when the precipitation

Figure 3.17. Gaussian distribution with standard deviation equal to 6 cm (left panel) and 2.7 cm (right panel).

departures from normal were -10,-5.7, and -8.3 cm, then the 3-year average is -8 cm, which is lower than the -7.3 cm limit associated with our 5% threshold to reject the null hypothesis. However, we just "cherry-picked" a case- did I have any reason ahead of time to look at 1900-1902? NO- I ran the analysis and then went searching one of the cases that meet my definition of "significance"- that is an aposteriori approach (after the fact).

Usually we only have an estimate of the population variance from our sample. Then, as already discussed in Chapter 2, the sample standard deviation $s_x = \sqrt{\frac{n-1}{n}}\sigma$ or $\sigma_{\bar{x}} = s_x / \sqrt{n-1}$. The degrees of freedom is n-1, which is a reminder that the sample can be described by the mean (1 value) plus n-1 others.

The Student's t test is a way to determine whether the null hypothesis can be rejected. The name "Student's t" comes from an employee of the Guiness brewery who had to submit his paper as "Student" anonymously to a journal. The t value is defined as: $t = (\bar{x} - \mu)\sqrt{n-1} / s_x$, which can be shown to be normally distributed for large numbers of degrees of freedom (n-1 greater than 30 or so). There are a variety of ways to grasp the meaning of the t statistic. Perhaps the simplest is to visualize the numerator as the 'signal', the difference between the sample and null hypothesis means times the number of members of the sample, and the denominator as the 'noise', the variability within the sample. As the value of t gets larger, our confidence in rejecting the null hypothesis that the mean of the sample is zero gets higher. The t value is large if: (1) the spread between the sample mean and the null value mean is large, (2) the number of members in the sample is large, (3) the variability in the sample is small.

So, let's loop over all 3-year samples in our record to see which periods might be considered droughts. The top panel of Fig. 3.18 shows the yearly precipitation anomalies. We want to know which 3-year periods can be classified as droughts and have some confidence that calling them a drought is not just due to chance. The 2nd panel defines the statistical significance, p, for rejecting the null hypothesis, which is there is no difference between the 3-year mean anomaly and zero. We want this value to be low, for example, p<5%. The center of the 3 years when the value lies above the 5% line (the scale has been reversed) reflect the cases where the null hypothesis can be rejected, i.e., accepting a 5% risk in those situations that classifying them as droughts could simply have happened by chance. The 3-year periods with very high values of p are ones where it is clearly not possible to reject the null hypothesis. The third panel flags those 3-year periods for which the null hypothesis could be rejected. If we are willing to accept a higher risk of falsely rejecting the null hypothesis, then we could use a higher threshold of say 0.10 and thereby identify more drought episodes. Note that our 1900-1902 period is identified as being one of the periods for which we could reject the null hypothesis.

Now for another caveat- we had 124 opportunities to reject the null hypothesis. So, 5% of 124 is 6, which means we might expect purely by chance that we would have 6 drought periods lasting 3 years. We found 10 (but there were really only 6 independent events), so we are teetering on the edge of not really finding much useful about drought periods in Utah.

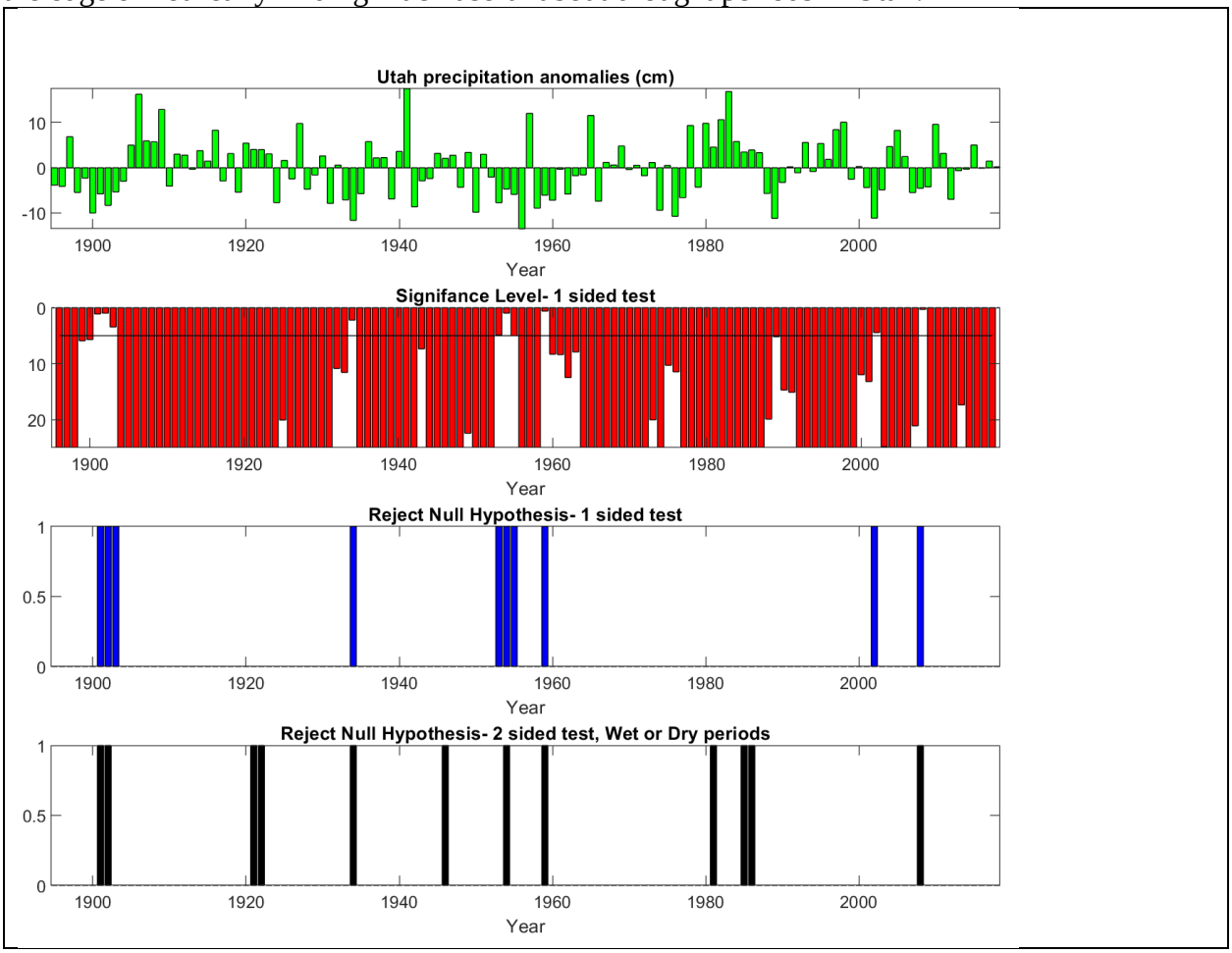


Figure 3.18. Top panel. Yearly precipitation anomalies. 2nd panel. Values of p, which shows the 3-year periods when the risk of incorrectly rejecting the null-hypothesis is less than 5% for a 1-sided left t-test. 3rd panel. Values of the flag h where h=1 indicates the null hypothesis can be rejected at the 5% confidence level for a 1-sided left t-test. Bottom panel. Values of the flag h where h=1 indicates the null hypothesis can be rejected at the 5% confidence level for a 2-sided t-test (both wet and dry events).

So far in this simple example, the test of the sample mean is a one-sided 'left' test (we're only interested in droughts). A two-sided test would require an alternative hypothesis that the 3-year mean anomaly is simply nonzero (either positive or negative). We're now interested in both "droughts" and "wet" periods- 3-year periods when the average is greater than zero. This weaker alternative hypothesis implies that any of the 3-year mean values must be even further from 0 (a smaller p value), i.e., a 2.5% chance for drought periods and a 2.5% chance for wet periods. It may seem somewhat paradoxical that assuming a weaker alternative hypothesis (both wet and dry) leads to a tougher obstacle to reject the null hypothesis. This becomes evident in the bottom panel of Fig 3.18 since the null hypothesis can be rejected for only a smaller number of really strong "drought" periods. But, now we can reject the null hypothesis for a number of 3-year wet periods, including the early 1980's. But, note that the "signal" (the mean of the 3 values) is evaluated relative to the "noise", the variance within the sample. So, it is more likely to be able to reject the null hypothesis when the variability within the sample is small. Note that the three-year period centered on the high precipitation year in 1984 is not a "wet" period because of the larger sample variability.

Figure 3.19 repeats the same analysis assessing wet and dry periods over 9 year samples. The main message is there have only been long dry periods state wide around 1900 and in the 1950's and wet periods in the 1910's and 1980's at least when evaluated using the student-t test.

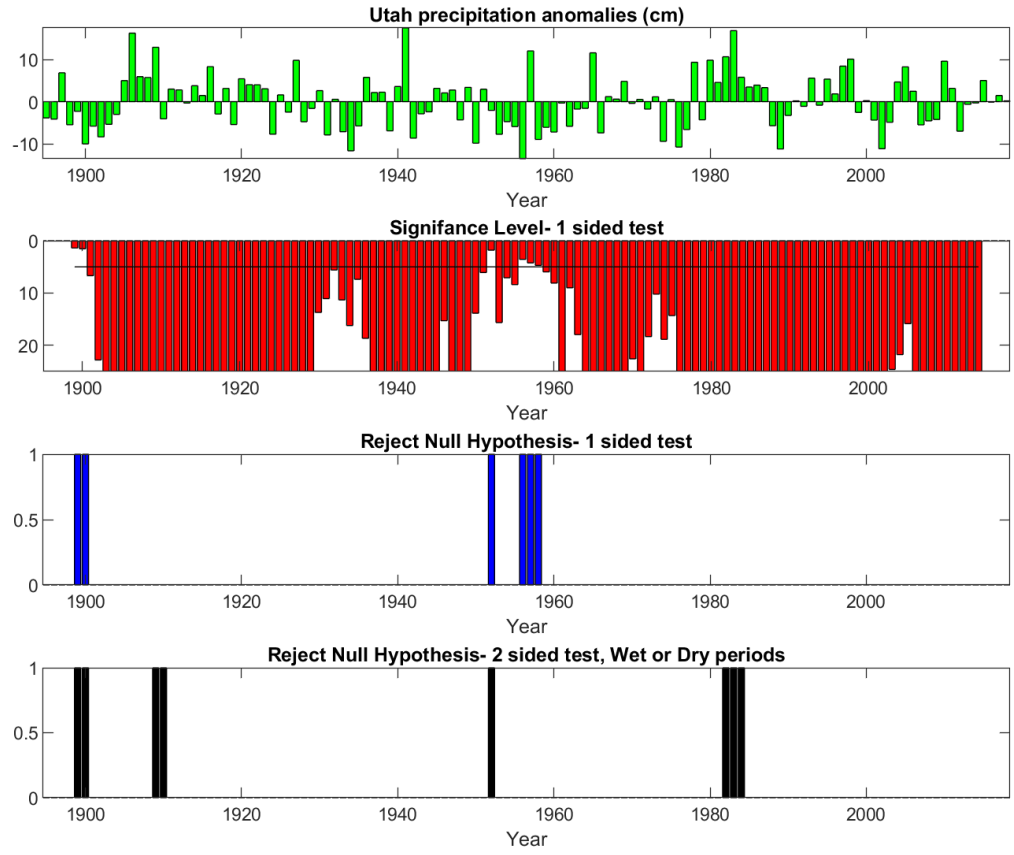


Figure 3.19. Top panel. Yearly precipitation anomalies. 2nd panel. Values of p, which shows the 9-year periods when the risk of incorrectly rejecting the null-hypothesis is less than 5% for a 1-sided left t-test. 3rd panel. Values of the flag h where h=1 indicates the null hypothesis can be rejected at the 5% confidence level for a 1-sided left t-test. Bottom panel. Values of the flag h where h=1 indicates the null hypothesis can be rejected at the 5% confidence level for a 2-sided t-test (both wet and dry events).

g. Summary

The exploratory data techniques developed in Chapter 2 are simply that: exploratory. Research involves defining a testable hypothesis and demonstrating that any statistical test of that hypothesis meets basic standards. Typical failings of many studies include: (1) ignoring serial correlation in environmental time series that reduces the estimates of the number of degrees of freedom and (2) ignoring spatial correlation in environmental fields that increases the number of trials that are being determined simultaneously. The latter inflates the opportunities for the null hypothesis to be rejected falsely. Use common sense. Be very conservative in estimating the degrees of freedom temporally and spatially. Avoid attributing confidence to a desired result when similar relationships are showing up far removed from your area of interest for no obvious reason. The best methods for testing a hypothesis rely heavily on independent evaluation using additional data not used in the original statistical analysis.