3. 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

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. 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. 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 using ***dfittool***
- use the options in ***dfittool*** to see if there is 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.

There are many examples in the sample code **chapter_3.m**. Traditionally, applications of parametric distributions required lookup tables; statistics books are full of such (e.g. Appendix B of Wilks). However, matlab tools are available that eliminate the need for lookup tables.

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. 5.1 represents $\int_{.5}^{1} f(x)dx$. The cumulative distribution function (CDF) is the total probability below a threshold, hence, the total area to the left of a particular value: $F(X) = \Pr\{x \le X\} = \int_{-\infty}^{X} f(x)dx$. For example, for the CDF in Fig. 3.2, the cumulative probability of negative values is 50%. Also, it is useful to define X(F) as the value of the random 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.

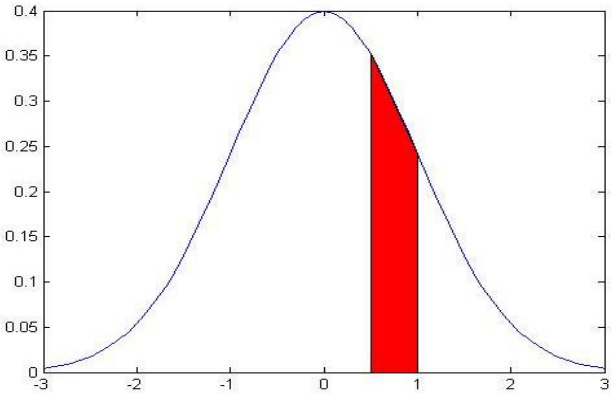

Fig. 3.1. Probability density function.

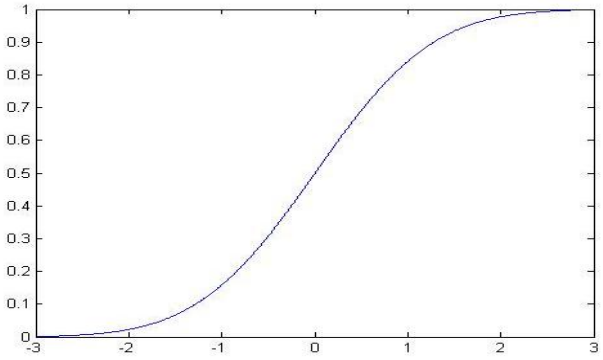

Fig. 3.2. Cumulative density function

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 it's 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 level annual record. Using ***dfittool***, the histogram is plotted in Fig 3.3 and a Gaussian (normal) distribution is fit 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 Fig. 3.4 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.

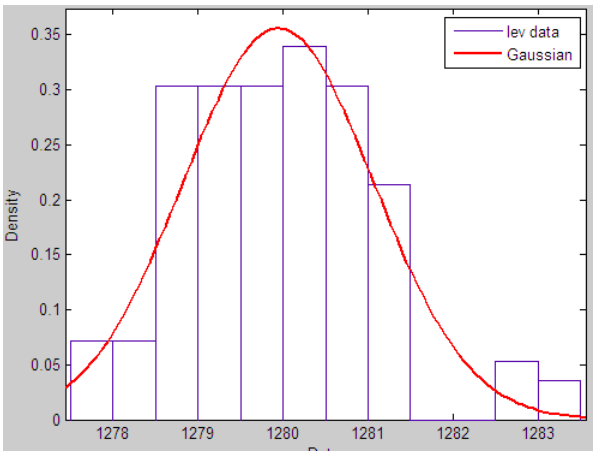

Figure 3.3. Gaussian fit to the annual level of the Great Salt Lake.

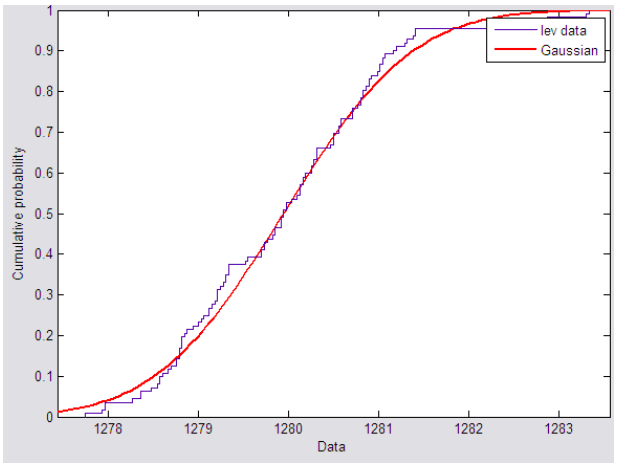

Figure 5.4. Cumulative probability distribution of Great Salt Lake level and Gaussian fit using the sample mean and variance.

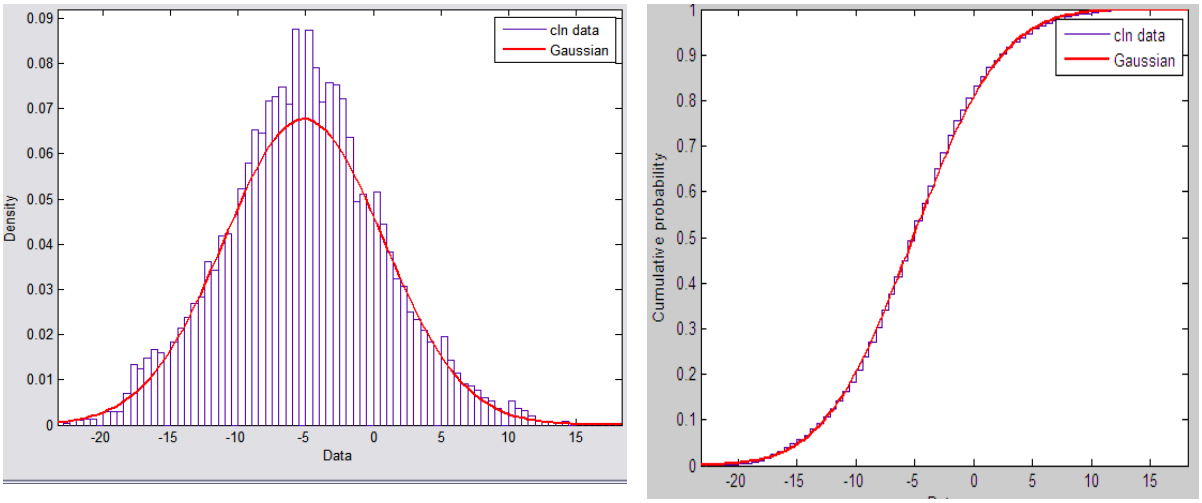

Figure 3.5. PDF and CDF of Alta-Collins hourly temperature data with Gaussian fit defined by the sample mean and variance

Let's now examine the hourly temperature values at Collins (CLN) near Alta during winter as shown in Fig. 3.5. 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.
Now, let's return to generic Gaussian distributions. Every variable can be transformed into standardized anomalies with mean 0 and variance 1. The matlab function **normspec** can be used to examine the total probability between specified limits, e.g., **p = normspec([-1,1],0,1).**

The leftmost panel of Fig. 3.6 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 (**p = normspec([-2,2],0,1)**) indicates that 95.5% of the total variance is within 2 standard deviations of the mean while the right figure (**p= normspec ([2,Inf],0,1)**) 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 **x = norminv([0.05 0.95],0,1)** returns ±1.65σ of the mean.

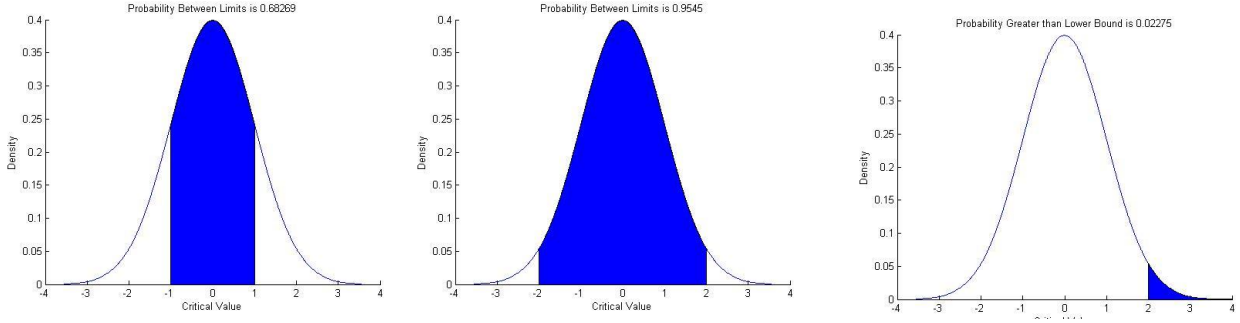

Figure 3.6. PDF's for the case when the sample mean is 0 and variance is 1.

c. Other parametric distributions

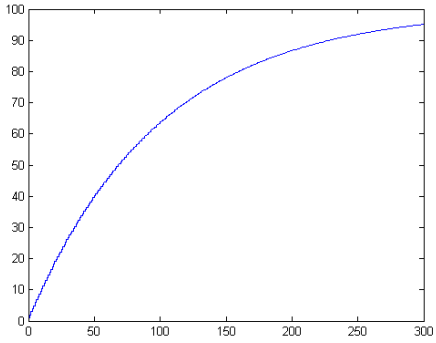

Figure 3.7 CDF as a function of years into the future for a 1 in a hundred year event assuming a geometric distribution.

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.

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). Using the following **x=0:300; y=geocdf(x,0.01); stairs(x,y);** yields Fig. 3.7, 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.

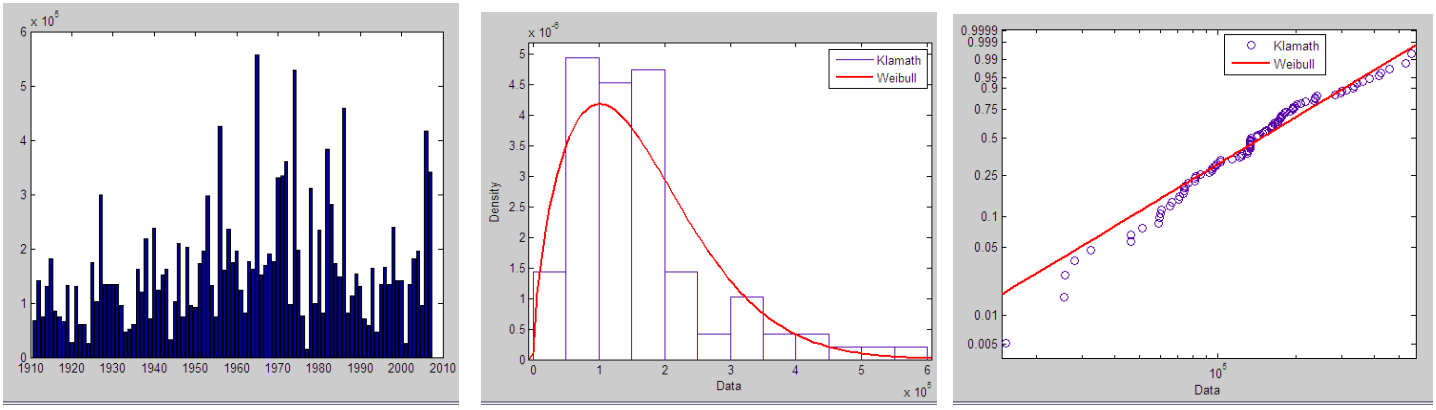

Figure 3.8. Peak streamflow (right panel) during the water year for the Klamath River, CA. The "100 year" flow was observed in December 1964 (1965 water year). PDF of peak streamflow and Weibull parametric fit to the data (center panel). Probability plot and Weibull parametric fit to the data (right panel).

As an example of evaluating the return period of extreme events, let's examine the peak streamflow record from the Klamath River for the 1910 to 2007 water years (Oct.-Sept.; hence December floods such as those in December 2005 are part of the 2006 water year) as shown in **chapter_3.m**. You should experiment with various fits- the Weibull fit to the data is shown here..

To what extent can we "predict" the occurrence of the peak in December 1964? Obviously, during this nearly 100 year record, that event was the "one in a hundred year" event, since its value is at the 99th percentile in the right panel. More apparent in the right panel is that the 1977 drought was very anomalous with much lower peak flow than would be expected by the Weibull fit.

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 (use dfittool on the cln record) and create a quantile plot as shown in Fig. 3.9. Empirically, we could state that when the temperature is less than -15C, then that is "unusually" cold, since that is only observed to happen about 5% of the time. We can also use the information from the Gaussian fit, in this case the parameter estimates are μ = -5.1C and σ=5.9C. Then, using **normspec([μ-1.96σ, μ+1.96σ], μ, σ);** then the lower panel in Fig. 3.9 indicates that 95% of the time, temperatures randomly selected from a Gaussian distribution with that mean and standard deviation would fall between -16 and 6C. So, today's temperature at Collins is -20C. Your ski buddy says "Gee, it's really cold". Time for some hypothesis testing. Can you tell him it is unusually cold or not?

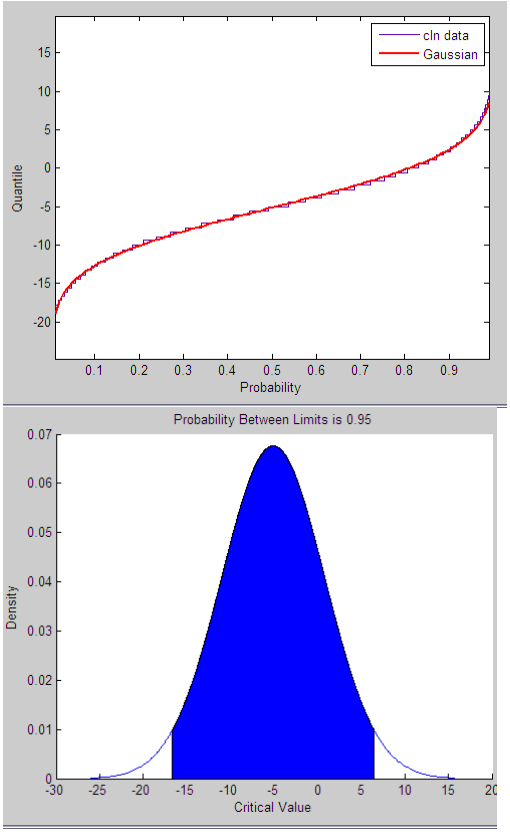

Figure 3.9. Top panel. Quantile plot of Alta-Collins temperature data. Bottom panel. Gaussian estimate using sample mean and standard deviation.

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 -5.1C. The temperatures associated with the shaded area (-16C to -6C) 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 -20C is outside of the shaded region.

As shown in Fig. 3.10, 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. 3.10. 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 -20C 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.

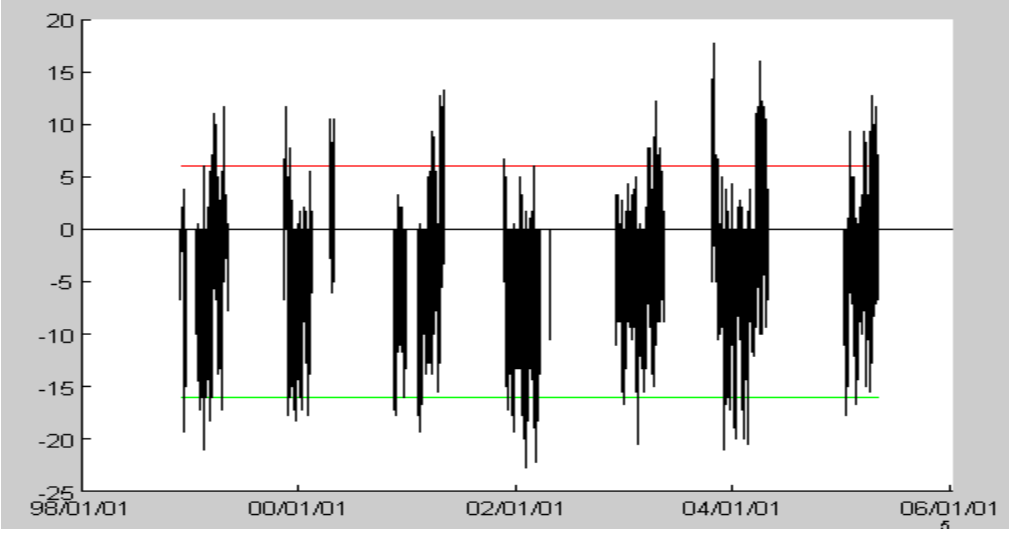

Figure 3.10 Time series of Alta Collins temperature (Nov-Apr) with confidence intervals

e. Hypothesis testing of means

Let's return to the annual precipitation in Utah. You'll examine as part of the homework some of the ways droughts are defined, but let's use here a completely arbitrary definition of a drought: that the average annual precipitation anomaly over a 5 year period differs substantively from zero. Consider the last 5 years with precipitation anomaly values of 3.9, -3.9, -3.8, -2.9, 2.4, which has a mean value of -.9 and a standard deviation among the 5 members of 3.9 cm. The standard deviation within the entire 6 year sample is 5.8 cm.

One expectation might be that the mean precipitation anomaly during the 5 years 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 5-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 five years within the past 116 years.
- Compare the observed test statistic (the composite mean value of -.9 cm 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.

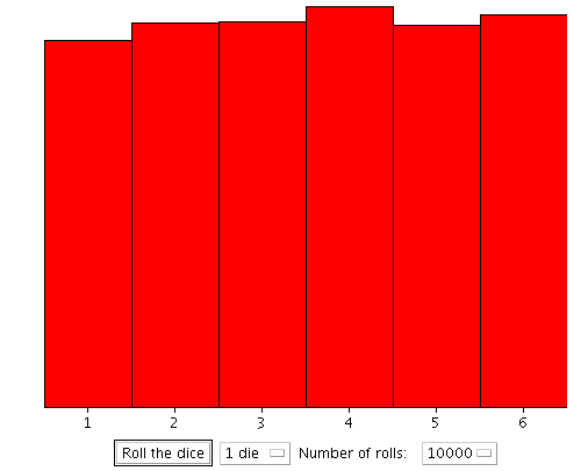

Figure 3.11. Top panel. Histogram based on rolling 1 die 10000 times. Bottom panel. Histogram based on rolling 5 dice 10000 times.

f. Central limit theorem and student-t test

Now we have to consider one of the reasons the Gaussian distribution is used so much. Run the java applet available from http://www.stat.sc.edu/~west/javahtml/CLT.html . First, roll 1 die 10,000 times. The red bars of roughly equal height show that the chance of getting any one number from 1-6 is roughly the same. Now roll 5 dice 10,000 times. In other words, we are taking the sum (or it could be the average) of 5 events. Note that we end up with a

Gaussian distribution. The odds of getting only a total count of 5 or 30 are small; most frequently we will get something around 17-18. *The central limit theorem states that the sum (or mean) of a sample (5 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.

Assume that the height anomalies have the distribution shown in the upper panel of Fig. 5.15, which corresponds to an assumed population standard deviation of 5.8 cm (which corresponds to the 116-year sample standard deviation). There is a 95% chance that the precipitation anomaly will lie between ±11.5 cm. We now take 5 values and average them. If we selected 5 years at random from the population many times, then according to the central limit theorem, we'd end up with the bottom panel. There is a 95% chance that the 5-year sample mean would lie between ±5.6 cm. In other words, it becomes less likely to have an extreme 5-year mean ("a drought" according to this lame definition) than just to have one extreme dry year.

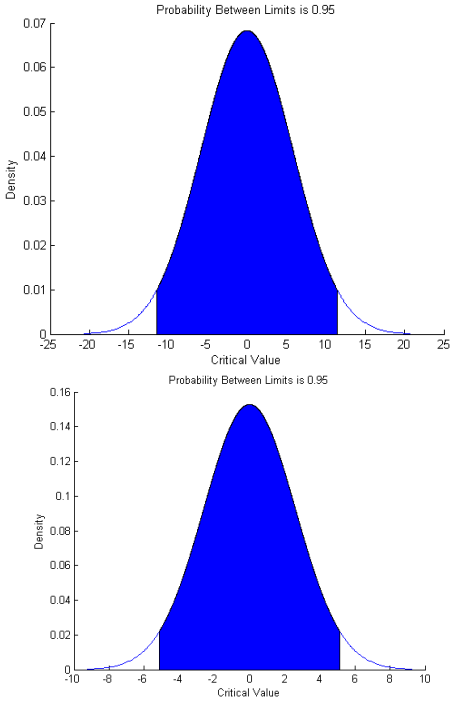

Figure 3.12. Gaussian distribution with standard deviation equal to 5.8 cm (top panel) and 2.6 cm (bottom panel).

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. If we truly knew that the population standard deviation was 5.8 cm as assumed in the top panel of Fig. 3.12, then we would determine that we could NOT reject the null hypothesis at the 5% level, since the sample mean during the last 5 years of -.9 cm lies within the shaded area in the lower panel.

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.

The chapter_3.m code loops over all 5-year samples in our record to see which periods might be considered droughts. Within the loop is the statement:

- **[h,p,ci,stat]= ttest(valy,0,.05,'left');**
- where on input valy is the vector of values in each 5-year sample
- 0 is the mean value for the null hypothesis
- .05 is the significance level chosen (5%)
- and 'left' indicates that we are assuming that we have ruled out that large positive anomalies are relevant (the other options are 'both' a two-tailed test and 'right' where we rule out large negative anomalies, i.e., look for wet periods)
- where on output h is a flag, 0 means the null hypothesis can not be rejected, 1 means it can be rejected
- p is the significance level corresponding to the t value, the smaller the number the better.
- ci is the confidence interval
- stat- is an array that returns the value of the t statistic, the number of degrees of freedom, and the estimated population standard deviation.

The top panel of Fig. 3.13 shows the center year of 5-year periods for which h = 1, i.e., the null

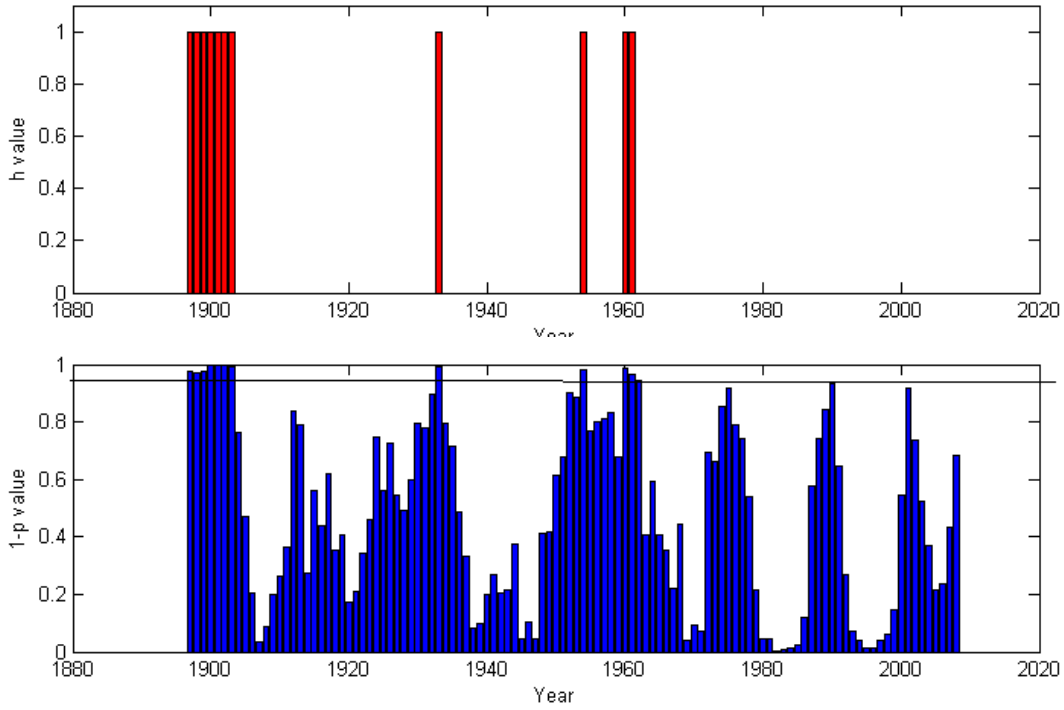

Figure 3.13. Top panel. 5-year periods which by the definition used here would be defined as droughts. Lower panel. Values of 1-p, which shows the 5-year periods when the null-hypothesis can be rejected at the 5% level.

hypothesis could be rejected at the 5% confidence level. The lower panel shows 1-p, which simply confirms why the 5-year periods identified in the upper panel can be assumed to differ from the population mean of 0 cm- only those years have values of 1-p greater than 0.95. If we are willing to accept a higher risk of falsely rejecting the null hypothesis, then we could use an α value of say 0.10 and thereby identify more drought episodes.

In the simple example used here, the test of the sample mean is a one-sided 'left' test (we're only interested in droughts). A two-sided test would require the alternative hypothesis to be that the composite mean anomaly is simply nonzero (either positive or negative). This weaker alternative hypothesis implies that the sample value of -.9 cm must be even further from 0 in the null distribution (a smaller p value), since we can reject the null hypothesis only if the rejection level (α level) is smaller by a factor of two as a result of the area on both sides of the null distribution's mean.

k. 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 statitiscal analysis.

Check Your Understanding

1. Run the chapter_3.m code and carefully go through what each part is doing.
2. Assume that the situation is a bit unusual and that the probability of any temperature from 0 to 29C being observed is exactly the same (i.e., there are 30 possible outcomes and the likelihood that any particular one is observed is identical). Determine f(x) in this instance and plot its PDF. Determine F(X) in this instance and plot the CDF.
3. Assuming that a population can be described by a Gaussian distribution, determine the cumulative probability of the values greater than 1, 2, 3, 4, and 5 standard deviations from the mean (hint use normspec).
4. Use the matlab widget ***randtool*** to explore parametric fits. (a) use the Normal (Gaussian) parametric fit and increase the sample size by powers of 10 from 10 to 1000000. Resample several times at each setting to get a sense of how sensitive the distribution is to sample size. At what point do you consistently get a smooth distribution of values? What does that tell you about the likelihood that any particular sample of say a hundred values will conform exactly to a Gaussian distribution? (b) Using a sample size of 10000, cycle through six of the possible parametric fits (there are 22 of them) and describe in a couple sentences what the PDFs look like for the default settings AND what sort of

environmental parameter might be described by this parametric fit (say don't know if you can't think of one, some are easier than others, i.e., if the parametric fit looks positively skewed then what environmental parameters are frequently like that, etc.).

5. The chapter_3.m code provides access to observations of temperature at Alta from five different stations (only CLN was used in the notes and in the code). You now will examine all five temperature records. Begin by loading all five records into the dfittool. Note that the AGD record appears to be affected by unrealistically high temperature values. Create an exclusion rule to eliminate temperature >15C and use that exclusion rule in all subsequent fits. a) Create a normal distribution fit to each of the five PDFs. Label them fit AGD, fit ALT, etc. b) Create a table of the estimated population mean and standard deviation from the normal parametric fits. c) Determine the temperature values that bracket 95% of the Gaussian distribution for each of the five samples using the function ***norminv.*** Include these values in your table begun in part b. d) Using the estimated population means and standard deviations for ATB and AGD only, use normspec to plot the areas in the histograms for temperatures above freezing. What is the probability of the temperature being greater than 0C at ATB and AGD based on their parametric fit?
6. Read the article by Cerveny et al. (June 2007, BAMS, http://ams.allenpress.com/archive/1520-0477/88/6/pdf/i1520-0477-88-6-853.pdf) on extreme weather records. Describe briefly some of the issues associated with reporting extreme weather records and the recommendations the authors have for how to improve that reporting (other than creating a committee, ugh).
7. The definition of drought used in this Chapter is not a good one. Defining drought episodes has scientific, economic, and societal impacts. Using online resources (such as http://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/) discuss briefly at least one metric used to define droughts. Be sure to discuss the strengths and weaknesses of that metric discussed by the source.