2. Exploratory Univariate Data Analysis

Environmental fields are awash with data. The first priority of any analysis is to simply spend time looking at the data in a variety of ways. Then, the next step is to reduce the dimensionality of the data sample by summarizing the data. Different types of data lend themselves to different approaches, so use examples from other studies or papers to gain ideas as to what might work for a particular data set. We'll begin by examining data samples (level of the Great Salt Lake and temperature and precipitation summarized for the state of Utah as a whole) using univariate techniques (i.e., the analysis of one variable is assumed to be independent of any other variables).

a. Examining time series of data

Geophysical data are usually collected sequentially, often at regularly spaced intervals. However, the data collection interval may change during the period of record, which introduces issues as far as how to summarize the data. Let's begin with the record of the level of the Great Salt Lake as a function of year from 1895 to 2010. Data are available from http://ut.water.usgs.gov/greatsaltlake/elevations/. Be sure to poke around that site to understand how and what the observation are. We'll also use estimates of annual precipitation and temperature for the state of Utah for the same period. These are available from http://www.wrcc.dri.edu/cgi-bin/divplot1_form.pl?4203. The files are available from the class web page (http://chpc utah.edu/~u0035056/5040/data/gsl_yr.txt and ut_ppt_yr.txt and ut_temp_yr.txt in the same directory).

First, look at the raw data files. The middle column of the lake level file is the number of observations. During the early years, the number of observations is only a couple per year. Recently, daily values of lake level are available. Hence, there may be some uncertainty about the lake level in the early part of the record relative to the later part simply on the basis of the methods used to record the observations (however, we'll see that the large serial dependence of lake level, i.e., that the lake level varies slowly, mitigates this problem to a large extent). The annual precipitation (in cm) and temperature (in C) are derived from Cooperative Observer reports, summarized into climate divisions, and then aggregated into the statewide average. There's a large number of steps and assumptions behind those calculations, so don't simply assume that these annual estimates are necessarily representative of the state as a whole.

In matlab, read the data by using the following:

- **[year,number,level] = textread('gsl_yr.txt','%f %f %f');**
- **[year,temp] = textread('ut_temp_yr,txt','%f %f');**
- **[year,ppt] = textread('ut_ppt_yr.txt','%f %f');**

Look at the data in each of the column vectors. Then, plot the time series using bar plots: **bar(year,level)**, **bar(year,temp)** , and **bar(year,ppt).** You'll need to play around a bit to reproduce the layouts of the following figures, (The graphs were cleaned up interactively to add labels, grids, change the axes, etc. You are expected to figure out how to do such things.)

As shown in Fig. 2.1, the lowest (highest) lake levels were observed during the 60's (80's). The trend during the past decade has been for lake level to be dropping. The serial dependence of the lake level data is evident, i.e., the value of lake level in one year is usually similar to that in adjacent years. A simple way to estimate the number of independent values in a sample is to draw subjectively line segments that reproduce the primary features of a time series as shown by the heavy line in Fig. 2.1. Then, the degrees of freedom is the number of line segments or ~13 in this case. Hence, even though there are 116 years in the sample, only about one in ten of those values are independent of the others.

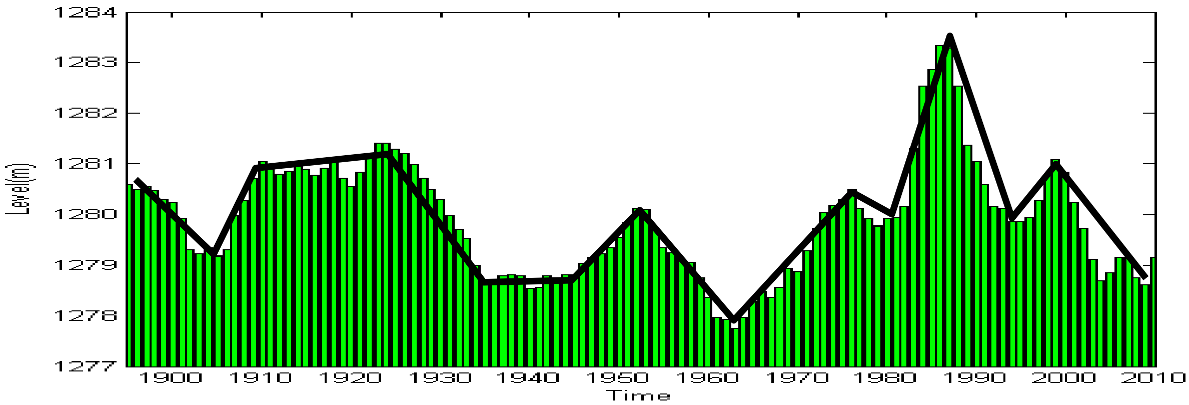


Figure 2.1. Level (m) of the Great Salt Lake. Bottom panel. Line segments added subjectively to estimate the number of independent values in the sample.

Now, let's examine the year-to-year changes in air temperature in the state of Utah. The temperature in Utah during the late 90's through the mid-part of the last decade have been warmer than that during any other decade during the past 100 years. Temperature during 1934 appears pretty unusual. The serial dependence from year to year is clearly less for temperature than for lake level, i.e., we have many more independent values in our sample of air temperature than we have for lake level, such that I subjectively estimated about 50 line segments would be required for the 116 year sample.

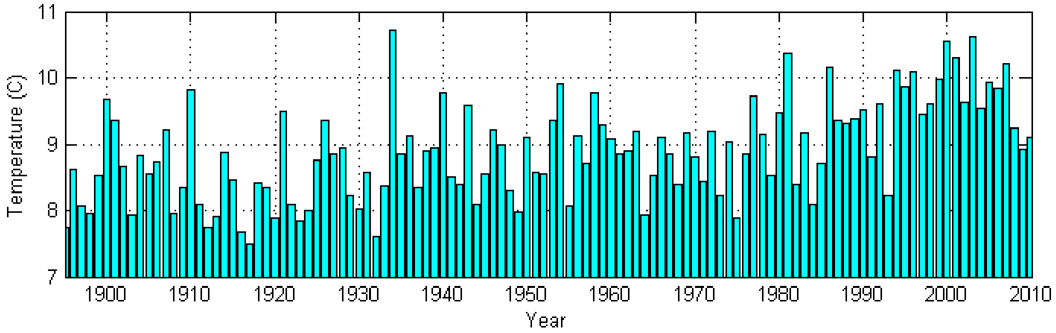


Figure 2.2. Annually averaged temperature (C) for the state of Utah

Finally, let's look at the annual total precipitation for the state of Utah. The wettest year for the state as a whole took place during 1941 and one of the driest was 1977. There are clearly some strings of years with greater than usual precipitation as well as drought episodes. The string of wet years in the early 1980's corresponds to the increase in lake level of the Great Salt Lake, for example. I'd guestimate about 40 line segments would be required to reproduce the major features of the time series, which would suggest that roughly every third year is independent of the others.

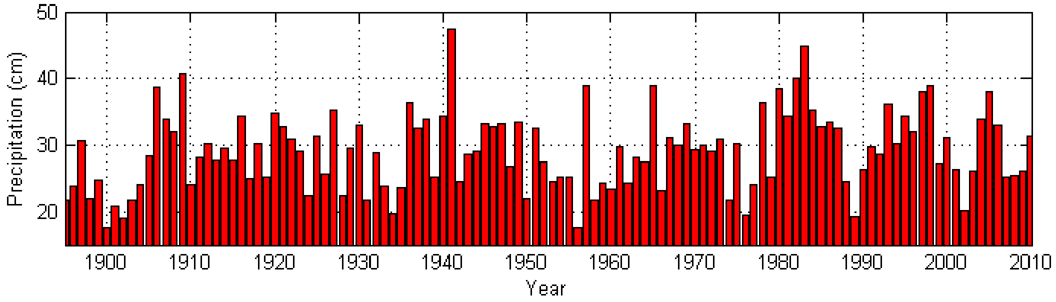


Figure 2.3. Annual precipitation (cm) for the state of Utah.

b. Data Distributions and Histograms

You should copy the matlab code **chapter_2_1.m** from the class web page into a directory of yours and examine it closely. Although that code is set up for all three variables (lake level and state temperature and precipitation), the notes in this section will describe the commands used for lake level only.

An obvious first step to examining a data set is to order them from smallest to largest:

- **levsort = sort(lev,'ascend');**

Take a moment and look at the resulting ordered data. The first (last) element is the lowest (highest) value and equal to 1277.8 m (1283.3 m) for lake level. Of course, computing the maxima, minima, and range (difference between the highest and lowest value) can be done:

- ***range_lev = range(level)*, or**
- **range_lev = max(level) – min(level)**

Hence, the lake has fluctuated in this sample over a range of 5.6 m. It is important to recognize that for a time series of environmental observations, the serial dependence of the data is lost when we sort by value. So, there is a tendency with sorted data to overemphasize the total number of values in a sample (116 years in our case) rather than how many independent values there may be (order 13).

Histograms are a convenient way to summarize the sorted data by aggregating them into bins ordered from smallest to largest. The most basic rules of thumb are simply to choose bin widths for histograms that give a relatively smooth appearing histogram or subdivide the range into convenient subintervals for labelling. The figure below on the left was generated by using:

- **x = 1277.5:1:1283.5;**
- **hist(lev,x);**

So, the lake level has been most commonly around 1280 m (left figure), but if we divide it up into .5 m intervals we see that values from 1278.5 to 1281 m are equally likely to occur. There are clearly some outliers, with a few years with levels greater than 1282.5 m and no years in the sample with values between1281.5 and 1282.5 m.

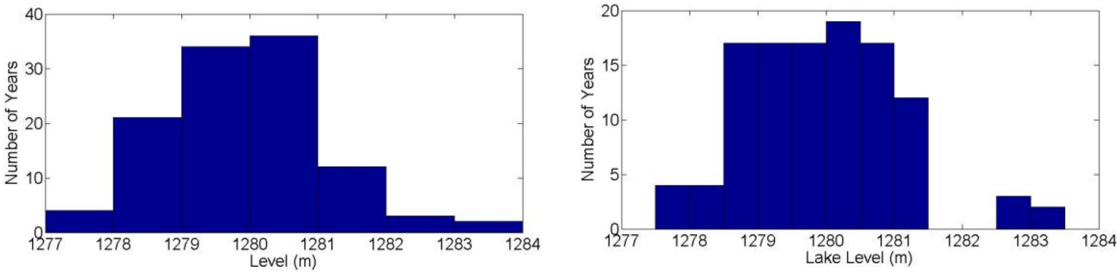


Figure 2.4. Histogram of lake level using 1m (left) and .5 m (right) intervals.

The matlab statistical toolbox has a distribution fitting tool that is quite useful. Type at the matlab command line ***dfittool*** and follow the directions. Pull down the data option and import the level data. You need to be able to reproduce the following figure.

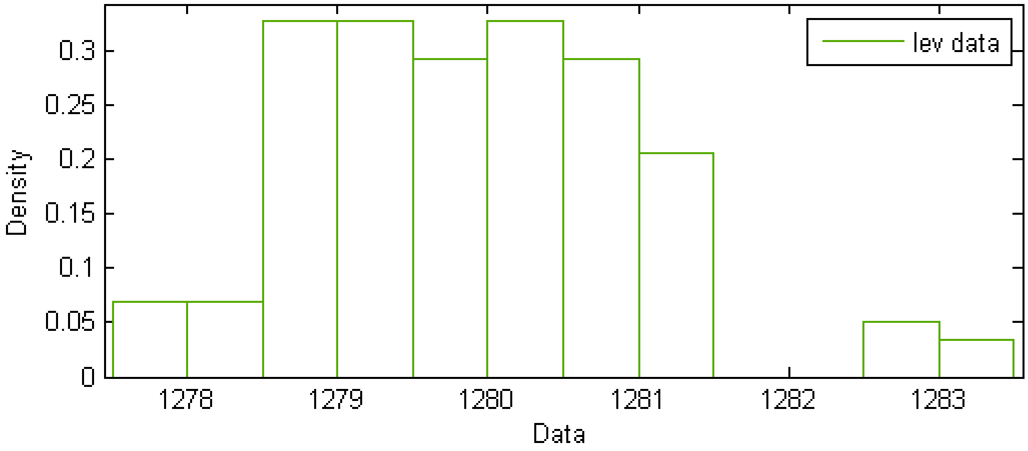


Figure 2.5. Histogram of lake level using the matlab dfittool utility.

The default settings led to binning the lake level into .5m bins, which is the same as the right panel in Fig. 2.4. The default vertical axis 'Density' can be a bit confusing: it is the fraction of the total number of values in a particular bin (e.g., 19/116 between 1280 and 1280.5 m) divided by the bin width (.5). So, the area of each box is the percentage contribution to the total sample (19/116 between 1280 and 1280.5 m). You can confirm that the sum of the area of all boxes adds up to 1. The dfittool has a bunch of other features, some of which we will use later on.

If the percentage of the total contributed by each bin in the histogram is added from smallest to greatest, then the cumulative frequency distribution is created. Toggle the display type in the ***dfittool*** to Cumulative Probability (CDF) or use ***cdfplot(lev)*** at the matlab prompt.

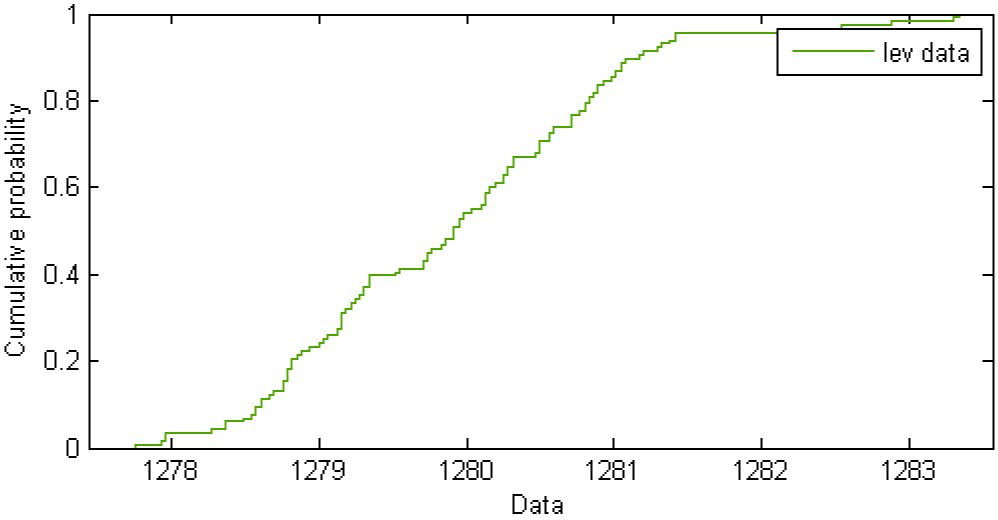


Figure 2.6. Cumulative frequency distribution of lake level using the matlab dfittool utility.

So, 50% of the time the lake level has been above 1280 m while 25% of the time the lake level has been below about 1279 m. If the cumulative probability was computed from a coarse histogram (such as if it was done from the values in Figs. 2.4 or 2.5), then there would be distinct transitions in the cumulative probability from one discrete value of GSL level to another. However, a probability is computed in Fig. 2.6 for each data value, which is why there is the fine scale chatter in the distribution.

c. Central value, spread, and symmetry

The characteristics of a sample of data are often summarized in terms of:

- central value (central tendency or typical value)
- spread (variation or dispersion about the central or typical value)
- symmetry (degree to which the values tend to be larger or smaller than the central value)

These quantities are often referred to as the first, second, and third moments of the data. There are also higher moments: the fourth moment is kurtosis that evaluates the degree to which a sample has a peak in its distribution or whether it is relatively flat. Whatever approaches we use to summarize the data, we want measures that are:

- *robust*- which means not overly sensitive to the characteristics of the entire sample of data values. In other words, we want the measure to perform reasonably well no matter how the data values are distributed.

- *resistant*- not unduly influenced by outliers in the sample. For example, the range is not resistant to outliers.

Histograms and cumulative frequency distributions help to define visually the central tendency, spread, and symmetry of the sample. Quantiles are defined as percentage thresholds of the data and are easily estimated from cumulative frequency distributions or can be computed as follows:

- q_{25} – lower quartile- 25% of the sample lies below that value and 75% lies above
- q_{50} – median- 50% of the sample lies below that value and 50% lies above
- **med_lev = median(lev);**
- q_{75} – upper quartile- 75% of the sample lies below that value and 25% lies above
- ***perc_lev = prctile(lev,[25,50,75]);***

The median is a very good measure of the central tendency of the data, i.e., the typical value. It tends to be robust and resistant. Terciles (thirds) and deciles (tenths) get used frequently as well. If the sample is small, then it is easy to go through an ordered list and count off where the percentage thresholds will lie. If the quantile falls between two values, then the average of the two adjacent values is used (i.e., if the sample contains 4 values, then the median is the average of the second and third value). Note that the ***prctile*** function does not require the data to be sorted first, i.e., you'll get the same percentiles if you determine them from the variables lev or perc_lev. Experiment by using terciles and deciles as well. In our case, the lake level has been below 1281 m roughly 90% of the time while it has been below 1279 m roughly 25% of the time.

Box and whiskers plots are a simple way to visualize the range, median, and quartiles of the data as well as outliers.

- ***boxplot(lev,'notch','on')***

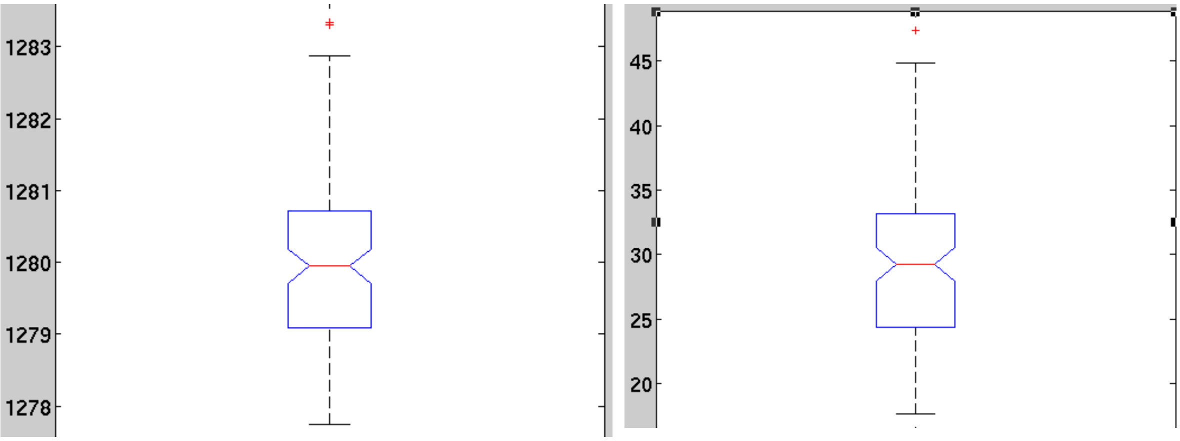


Figure 2.7. Boxplots of lake level (left) and Utah precipitation (right)

The center 'notched' line is the median. The top of the box is the upper quartile, the bottom of the box is the lower quartile. The 'whiskers' are defined in terms of the interquartile range (IQR):

- IQR = q_{75} - q_{25}
- ***iq = iqr(lev)***

The IQR tends to be a robust and resistant measure of spread or variation within the data set.

The top whisker is the median + 1.5 IQR and the bottom whisker is the median – 1.5 IQR. Finally, outliers above and below the whiskers are shown by plus symbols. In the case of the GSL level in the left panel, the high water years are viewed as outliers, which is obviously not the same as an erroneous value. It does point out that, relative to the values observed in other years, the high water years in the 80's were unusual. Similarly, the 1941 value is identified as an outlier for the annual precipitation sample in the right hand box and whisker plot. If you look at the Utah temperature box and whisker plot, there are no outliers.

A traditional measure of the central tendency or typical value of the sample is the mean, which is not a robust and reliant measure of the central value:

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *(2c.1)*
- **xbar = mean(lev)**

(For convenience in the sample code chapter_2_1.m, the variables lev, temp, and ppt are loaded into columns of variable 'array'. These notes will continue to refer to the lake level variable alone, which is the first column of array.) For the sample of lake levels, the mean and median are the same: 1279.9 m. But modify your data set by throwing in a bad value for the first element (e.g., 9999), which can commonly happen if the data file becomes corrupted for some reason. Obviously, the mean is now much higher, while the median remains unchanged.

The trimmed mean $\overline{X_\alpha}$ is a more resistant measure of central tendency, since a fraction of the high and low values are removed before the mean is calculated:

- $\overline{X_\alpha} = \frac{1}{n-2k}\sum_{i=k+n}^{n-k} x_i$ *(2c.2)*
- ***xbar_trim = trimmean(lev,10)*** where the 10 indicates that 10% of the top and bottom are removed before the mean is calculated

Now let's look at measures of spread. You should already be aware that the maxima, minima, and range are not robust and resistant measures of spread. The median absolute deviation (MAD) is a more robust and resistant measure of spread and uses all of the data rather than the central core of the data as with the IQR.

- MAD = median | x_i – $q_{.5}$ |

MAD is computed by taking the difference between each value and the sample median and then taking the median of that resulting sample.

- **ma = mad(lev)**

which is .89 m for the GSL level. The median absolute deviation tends to be a conservative measure of spread.

The standard deviation, s, is a common measure of spread that is not resistant to outliers or robust. The square of the standard deviation is the variance, s^2, and is called an unbiased estimate of the population variance (and s is referred to as an unbiased estimate of the population standard deviation):

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ *(3c.3)*
- **var0 = var(lev,0)**
- **std0 = std(lev,0)**

For the GSL level, the standard deviation is 1.12 m and the variance is 1.25 m^2. Hence, the standard deviation provides an indication that there is roughly 1 meter of variation or dispersion about the central value of 1280 m. It is very important to pay attention to the units: the standard deviation has the same units as the quantity itself, while the variance has those units squared.

Why is the variance s^2 calculated by dividing through by n-1 rather than n as we did when computing the mean? Consider a population with mean 0 and population variance σ^2. Then the variance of the population can be computed in the same manner as the population mean:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 = \overline{x^2} \quad (2c.4)$$

When we use a sample of n independent values (which may be a small sample), and if we compute:

$$s_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad (2c.5)$$

then $s_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{2}{n}\sum_{i=1}^{n}(x_i\overline{x}) + \frac{1}{n}\sum_{i=1}^{n}(\overline{x})^2$

Now we need some summation identities:

$$\sum_{i=1}^{n}a = na \text{ where a is a constant and } \sum_{i=1}^{n}ax_i = na\overline{x} \quad (2c.6)$$

Then $s_x{}^2 = \overline{x^2} - \frac{2}{n}\overline{x}\sum_{i=1}^{n}(x_i) + \frac{1}{n}\overline{x}^2 n = \overline{x^2} - 2\overline{x}\overline{x} + \overline{x}^2 = \overline{x^2} - \overline{x}^2$ *(2c.7)*

Let’s stop here a moment and recognize that the final relationship in 2c.7 is a convenient way to compute the variance that does not require removing the mean first and then summing. Instead, we can sum the squares of the values (first term) at the same time we are summing the values (second term) and later squaring the mean value.

Now, if we had a very large sample, the second term $\overline{x}^2$ should be zero, since the population mean is zero. But, for a small sample, it may not. Given that the sample is supposed to be comprised of independent values, then it can be shown that:

$\overline{x}^2 = \frac{1}{n}\overline{x^2}$ (you’ll see that this comes from the central value theorem later). As n gets big, then it will trend to zero. So, $s_x{}^2 = \overline{x^2} - \frac{1}{n}\overline{x^2} = \frac{n-1}{n}\overline{x^2} = \frac{n-1}{n}\sigma^2$

Comparing *(2c.3)* to the above, $s_x{}^2$ = (n-1)/n $s\ ^2$, which is why s^2 is an unbiased estimate of the population variance; $s_x{}^2$ is the *sample variance* and it is computed in matlab using the following:

- **var1 = var(lev,1)**
- **std1 = std(lev,1)**

The differences between the sample and population standard deviation are likely small if the sample size is large (more than 50 or so). However, we’ll see later that it can be important to differentiate between what we measure from a sample and what we estimate for the population.

Symmetry is a measure of the balance around the center value. Skewness (γ) is a nondimensional measure of asymmetry. If γ is negative, then data are spread more below the mean than above the mean. Skewness is neither robust nor reliant.

- $\gamma = \dfrac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$
- ***skew= skewness(lev)***

In the case of the GSL level, the skewness is .54, which indicates that there are larger departures above the mean, i.e., the large positive outliers “skew” the distribution of lake level. The annual precipitation is also positively skewed with a value of .33.

d. Transforming Data

The interpretation of a sample of data can be aided by transforming the data to a new variable. The simplest transformation is to remove the mean, in order to examine specifically the variability about a central value.

- x′= x - $\overline{x}$ = anomaly or departure from the mean
- **lev_a = lev – xbar;**

For example, the 116-year sample mean has been removed from the GSL levels in Fig. 2.8. Such a simple transformation can make a big difference as far as the interpretation of the data. In this instance, the anomalous period of the mid-1980’s stands out. The recent period appears comparable to that during much of the middle part of the last century.

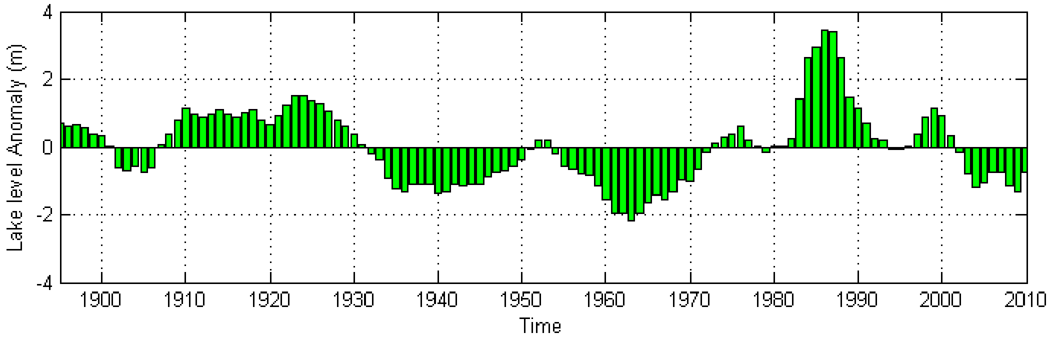


Figure 2.8. Departure (m) of lake level from 116 year sample mean.

Removing the sample mean from the Utah temperature record reveals in Figure 2.9 that the string of positive temperature anomalies centered around 2000 was unprecedented. In addition, the 1934 temperature anomaly was quite unusual for that period. Strings of years with above and below normal precipitation in Utah are also more clearly evident in Fig. 2.10 than when the raw time series is examined. Obviously, you can transform the data in a myriad of ways. What if we removed the 1971-2000 "climate normal" for temperature instead (Fig. 2.11)? Then, the period prior to 1980 appears to be nearly always below normal.

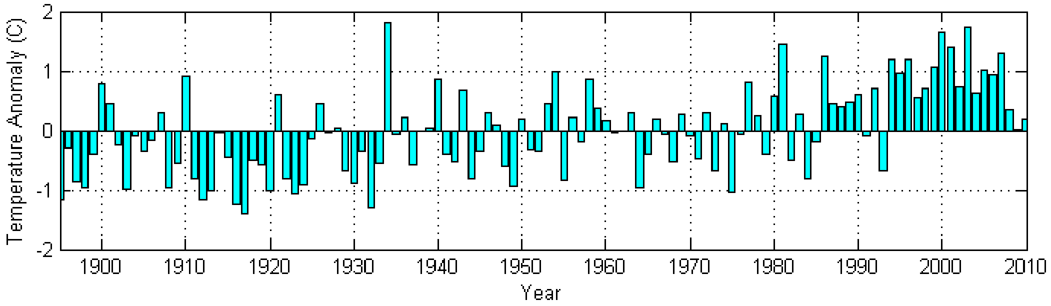


Figure 2.9. Anomalies (C) of Utah temperature relative to the 116-year sample mean.

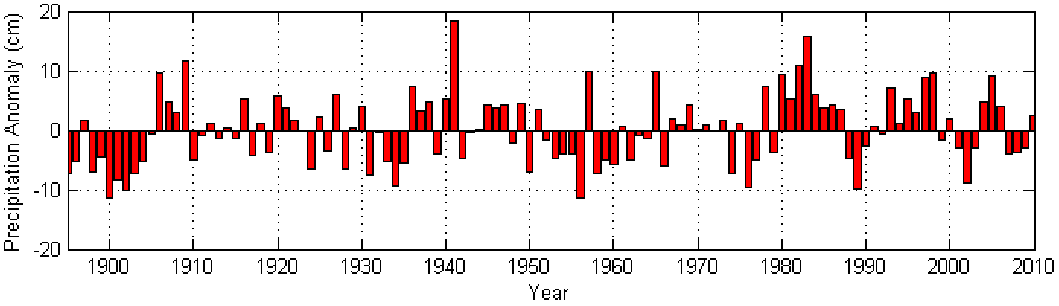


Figure 2.10 Anomalies (cm) of Utah precipitation relative to the 116-year sample mean.

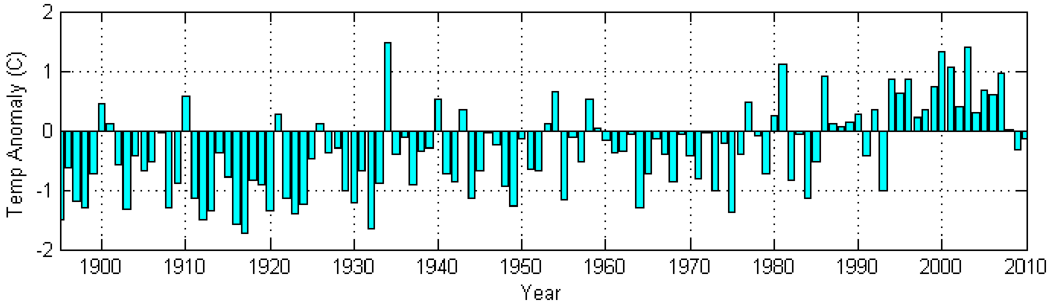


Figure 2.11. Temperature anomalies (C) relative to 1971-2000 climate normal.

In many environmental applications, quasi-periodic behavior due to solar forcing (diurnal or seasonal) can overwhelm the signal of interest. If you are interested in how much warmer each month of 2010 is compared to 2009, then the fact that January is always colder than July may not be relevant. Let's examine the monthly Great Salt Lake level data since 1904. You will need to download the data file (gsl_mon.txt) and the matlab program (chapter_2_2.m). Run the script and it will generate the following figures. You should examine carefully all of the matrices generated along the way.

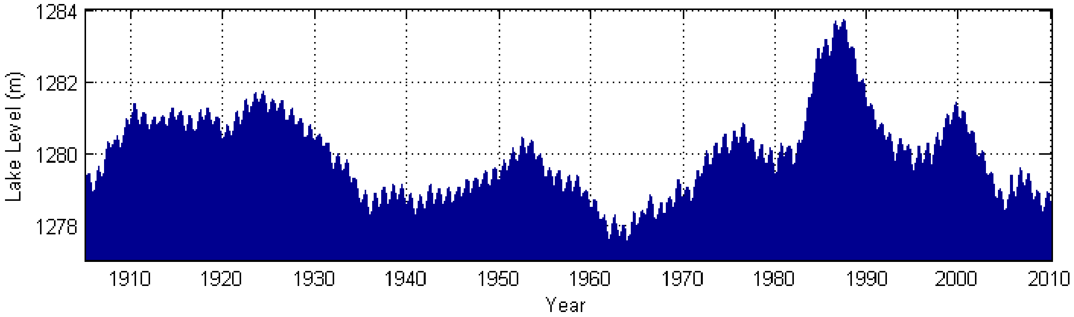


Figure 2.12. Monthly lake level (m).

Regular seasonal oscillations are clearly evident in Fig. 2.12, and, as shown in Fig. 2.13, the lake level peaks typically in June (after the spring runoff period) and is lowest in the fall. Using the approach that the number of independent samples can be defined by the number of line segments required, do we now have 208 degrees of freedom (one for each year's rise and fall)? NOOOO! We have 2 relevant time scales: the annual cycle as shown in Fig. 2.13 and the small number of multi-year fluctuations already described using the annually-averaged data. .

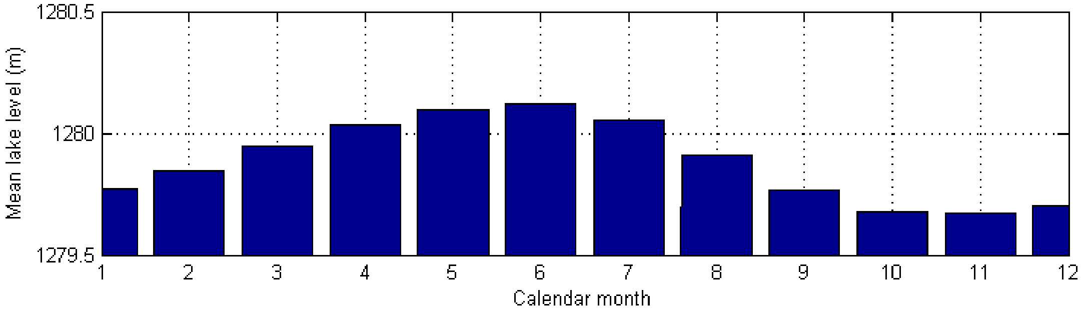


Figure 2.13. Lake level (m) averaged separately for each calendar month over the 108 years.

Since the variability in environmental data often differs during the year (e.g., weather systems moving across the midlatitudes are more frequent in winter than in summer), it is often appropriate when using data from all seasons to "normalize" the anomalies so that the variability in winter and summer receive similar weight. Hence, the value x_{ij} for year j and month i is subtracted from the mean for that month $\overline{x_i}$ and divided by the sample standard deviation s_{xi}, i.e.,

- $$z_{ij} = \frac{x_{ij} - \overline{x_i}}{s_{xi}} = \frac{x'_{ij}}{s_{xi}}$$

In other words, a nondimensional time series is generated by creating 'standardized' or 'normalized' anomalies'. As shown in Fig. 2.14, the lake level was three standard deviations above normal in 1986 and approached 2 standard deviations below normal in 1963. Even though we have a sample size of 108*12 values, it is pretty clear from Fig. 2.14 that there remain less than 20 independent samples in this time series, if we ignore some of the minor bumps and wiggles.

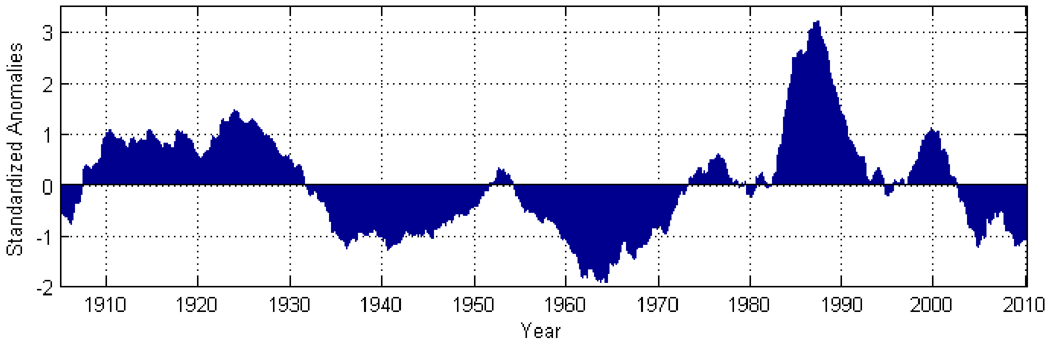
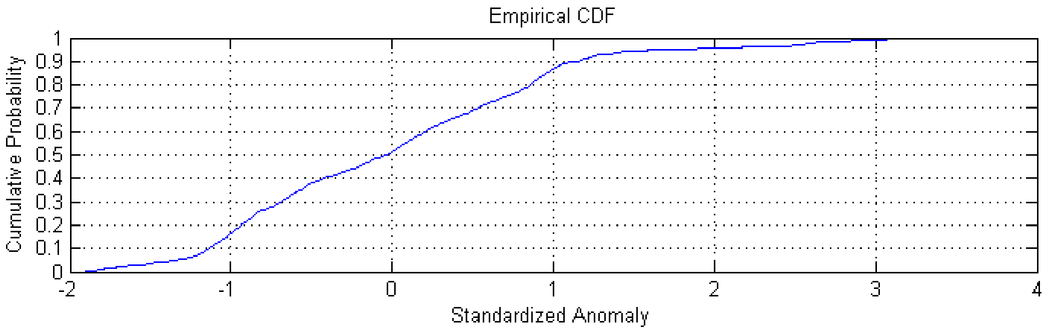


Figure
2.14. **Figure 2.15. Cumulative frequency distribution for monthly values of lake level.**
Standardized monthly anomalies of lake level



One of the advantages for appropriately transforming environmental data at higher sampling intervals (e.g., monthly vs. annually-averaged values) is that we can get an improved probability density distribution, as shown in Fig. 2.15. Compare this CDF to the one computed from the annual means. In terms of standardized anomalies, the median is 0, while the monthly lake level is greater than 1 standard deviation above normal during only 10% of the months yet the lake level is less than 1 standard deviation below normal during 20% of the months. However, keep in mind that even though the probability distribution is smoother because of the larger sample size, the number of independent values remains low.

Variables such as wind speed and precipitation when examined from hour to hour or day to day tend to be strongly positively skewed. In other words, their distributions tend to be asymmetrical since no values are possible below 0, many values may be close to 0, and then occasional extreme values are possible. A simple transformation for wind speed or precipitation is to take the square root of the values, first, then remove the mean and normalize. For example, let

$y_{ij} = \sqrt{x_{ij}}$ and then $z_{ij} = \dfrac{y_{ij} - \overline{y_i}}{s_{yi}}$

The annual Utah precipitation time series exhibits considerable year-to-year variability compared to the GSL level time series. A common transformation is to smooth a time series by redefining each element of the time series in terms of an aggregate of nearby values within a specified “filter window”. Running means should be avoided because they tend to shift peaks and smooth well-defined jumps in the data. Other simple weighting schemes will be shown later that can be used to filter out or retain specific components of the underlying time scales in the data (i.e., high pass, low pass, and band pass filters).

One of the better smoother operators is a median smoother, where the median of values within a data window replace each element of the time series. The median smoother does a better job at maintaining sharp jumps in the data. Consider a portion of a time series on which a 3 point running mean and running median are applied as shown in Fig. 2.16. The original time series is the heavy line, i.e., the values are 0 from point 1-5 and then the data jumps to 2 at point 6. A three-point (or 5 point) median filter exactly matches the original data in this instance, which is sensible. However, a 3-point running mean (thin line) will smooth out the jump and lose the useful information that the sudden jump occurred at point 6.

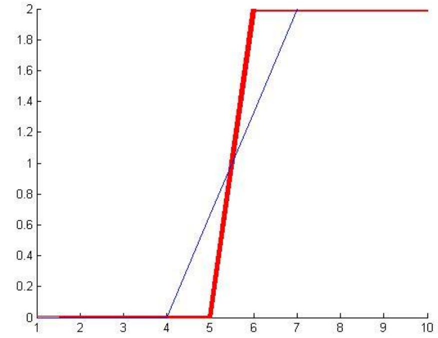


Figure 2.16. Example of original and median smoother (red line) and running mean smoother (blue line).

As a general rule, it is better to aggregate the data within a small window and apply the smoother several times rather than aggregating the data within a large window and only applying the smoother once. Figure 2.17 shows the annual Utah precipitation anomalies after three passes with a 3-year median smoother (heavy green line). A yearly spike such as in 1941 is lost as “noise” while the anomalous string of wet years during the early 1980’s begins to stand out. The median smoother is supplied as a matlab function that must be in your working directory (see **median_smooth.m**).

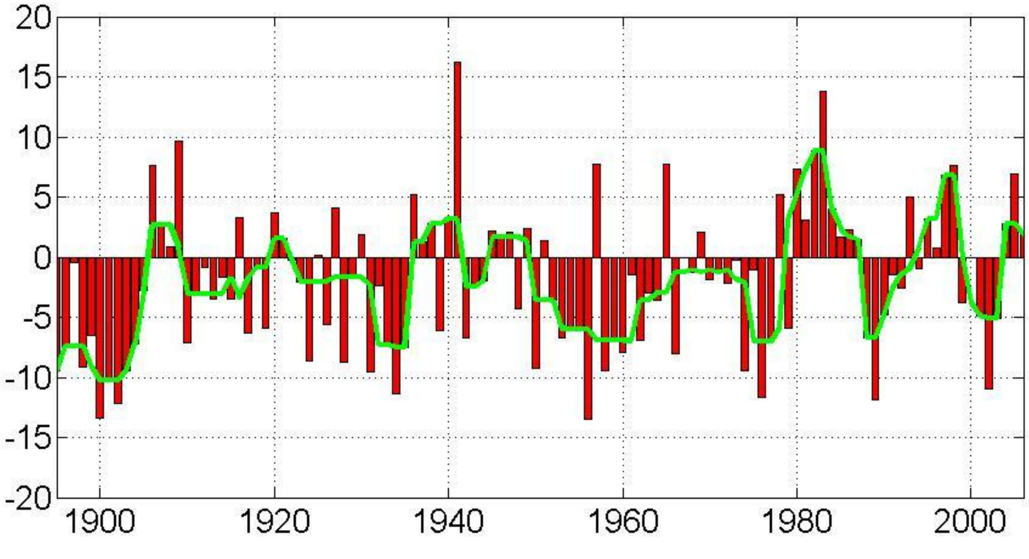


Figure 2.17. Utah annual precipitation anomalies (red bars) and the data smoothed by three passes of a 3-point median smoother (green line).

f. Check Your Understanding

1. A time series consists of the values ordered as follows: 0, 4, -3, -4, -1, 0, 1, -1, -3, -3. Compute the range, mean, quartiles, MAD, and IQR. Draw a box and whiskers plot for this time series.
2. Using the sample in #1, compute the sample standard deviation and variance. Then, determine the anomalies relative to the sample mean, and standardized anomalies.
3. Smooth the data in #1 with a single pass of a 3-point running smoother and a 3-point median smoother. The first and last values remain their original values. What does each smoothing technique do to the time series?
4. Review very carefully the two codes used in this chapter. Run them. Reproduce all figures in this chapter as they appear, not as the codes generate them. Turn those figures in.
5. Was it really so hot in Utah in 1934 and wet in 1941? Which parts were and which may have been less so? Verify from online resources that these outliers are real. Try to track down from historical accounts how people perceived those record events (i.e., did they recognize them at the time as being something unusual?).
6. Create a cumulative histogram of lake level using the bin intervals of Fig. 2.5. The y axis will vary between 0 and the total number of years (rather than a cumulative probability where the y axis varies between 0 and 1). Simply sum the count of values below each bin threshold.