Statistical Methods in Water Resources with R

MiRoVaGo

2020-01-17

Contents

Preface

This book is a re-coding of Statistical Methods in Water Resources by Helsel & Hirsch. As the use of R in the hydrology community steadily grows, translating Helsel & Hirsch material into R was an opportunity we could not miss. The content herein is a direct adaptation of the source material into R and its tools.

Chapter 1

Summarizing data

When determining how to appropriately analyze any collection of data, the first consideration must be the characteristics of the data themselves. Little is gained by employing analysis procedures which assume that the data possess characteristics which in fact they do not. The result of such false assumptions may be that the interpretations provided by the analysis are incorrect, or unnecessarily inconclusive. Therefore we begin this book with a discussion of the common characteristics of water resources data. These characteristics will determine the selection of appropriate data analysis procedures.

One of the most frequent tasks when analyzing data is to describe and summarize those data in forms which convey their important characteristics. "What is the sulfate concentration one might expect in rainfall at this location"? "How variable is hydraulic conductivity"? "What is the 100 year flood" (the 99th percentile of annual flood maxima)? Estimation of these and similar summary statistics are basic to understanding data. Characteristics often described include: a measure of the center of the data, a measure of spread or variability, a measure of the symmetry of the data distribution, and perhaps estimates of extremes such as some large or small percentile. This chapter discusses methods for summarizing or describing data.

This first chapter also quickly demonstrates one of the major themes of the book – the use of robust and resistant techniques. The reasons why one might prefer to use a resistant measure, such as the median, over a more classical measure such as the mean, are explained.

The data about which a statement or summary is to be made are called the **population**, or sometimes the **target population**. These might be concentrations in all waters of an aquifer or stream reach, or all streamflows over some time at a particular site. Rarely are all such data available to the scientist. It may be physically impossible to collect all data of interest (all the water in a stream over the study period), or it may just be financially impossible to collect them.

Instead, a subset of the data called the **sample** is selected and measured in such a way that conclusions about the sample may be extended to the entire population. Statistics computed from the sample are only inferences or estimates about characteristics of the population, such as location, spread, and skewness. Measures of location are usually the sample mean and sample median. Measures of spread include the sample standard deviation and sample interquartile range. Use of the term "sample" before each statistic explicitly demonstrates that these only estimate the population value, the population mean or median, etc. As sample estimates are far more common than measures based on the entire population, the term "mean" should be interpreted as the "sample mean", and similarly for other statistics used in this book. When population values are discussed they will be explicitly stated as such.

1.1 Characteristics of Water Resources Data

Data analyzed by the water resources scientist often have the following characteristics:

1. A lower bound of zero. No negative values are possible.
2. Presence of 'outliers', observations considerably higher or lower than most of the data, which infrequently but regularly occur. outliers on the high side are more common in water resources.
3. Positive skewness, due to items 1 and 2. An example of a skewed distribution, the lognormal distribution, is presented in figure **??**. Values of an observation on the horizontal axis are plotted against the frequency with which that value occurs. These density functions are like histograms of large data sets whose bars become infinitely narrow. Skewness can be expected when outlying values occur in only one direction.
4. Non-normal distribution of data, due to items 1 - 3 above. Figure **??** shows an important symmetric distribution, the normal. While many statistical tests assume data follow a normal distribution as in figure **??**, water resources data often look more like figure **??**. In addition, symmetry does not guarantee normality. Symmetric data with more observations at both extremes (heavy tails) than occurs for a normal distribution are also non-normal.
5. Data reported only as below or above some threshold (censored data). Examples include concentrations below one or more detection limits, annual flood stages known only to be lower than a level which would have caused a public record of the flood, and hydraulic heads known only to be above the land surface (artesian wells on old maps).
6. Seasonal patterns. Values tend to be higher or lower in certain seasons of the year.
7. Autocorrelation. Consecutive observations tend to be strongly correlated with each other. For the most common kind of autocorrelation in water

resources (positive autocorrelation), high values tend to follow high values and low values tend to follow low values.

8. Dependence on other uncontrolled variables. Values strongly covary with water discharge, hydraulic conductivity, sediment grain size, or some other variable.

Methods for analysis of water resources data, whether the simple summarization methods such as those in this chapter, or the more complex procedures of later chapters, should recognize these common characteristics.

1.2 Measures of Location

The mean and median are the two most commonly-used measures of location, though they are not the only measures available. What are the properties of these two measures, and when should one be employed over the other?

1.2.1 Classical Measure – the Mean

The mean ($\overline{\mathbf{X}}$) is computed as the sum of all data values $\mathbf{X_i}$, divided by the sample size n:

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \tag{1.1}$$

```
mean_x <- mean(x)
```

For data which are in one of k groups, equation (**??**) can be rewritten to show that the overall mean depends on the mean for each group, weighted by the number of observations n_i in each group:

$$\overline{X} = \sum_{i=1}^{n} \overline{X}_i \frac{n_i}{n} \tag{1.2}$$

where $\overline{\mathbf{X}}_\mathbf{i}$ is the mean for group i. The influence of any one observation X_j on the mean can be seen by placing all but that one observation in one "group", or

$$\begin{aligned} \overline{X} &= \overline{X}_{(j)} \frac{n-1}{n} + X_j \bullet \frac{1}{n} \\ &= \overline{X}_{(j)} + \left(X_j - \overline{X}_{(j)}\right) \bullet \frac{1}{n} \end{aligned} \tag{1.3}$$

where $\overline{\mathbf{X}}_{(\mathbf{j})}$ is the mean of all observations excluding X_j. Each observation's influence on the overall mean $\overline{\mathbf{X}}$ is $\left(\mathbf{X_j} - \overline{\mathbf{X}}_{(\mathbf{j})}\right)$, the distance between the observation and the mean excluding that observation. Thus all observations do not have the same influence on the mean. An 'outlier' observation, either high

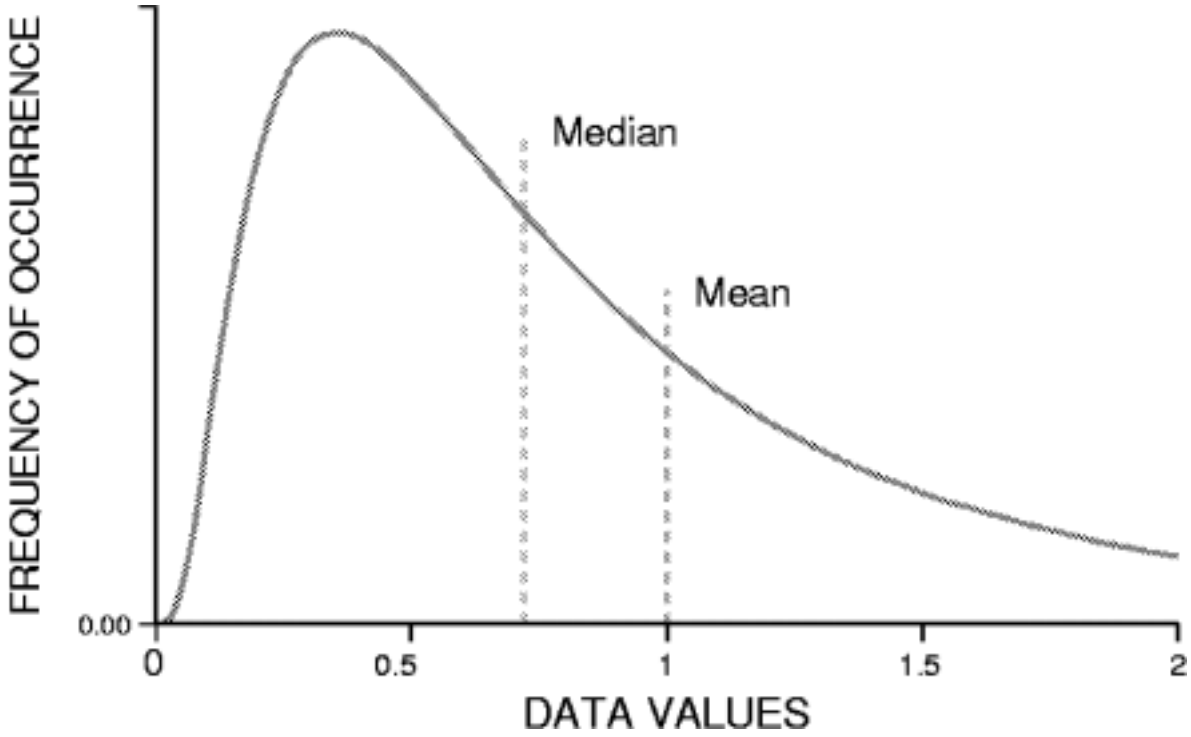


Figure 1.1: Density Function for a Lognormal Distribution

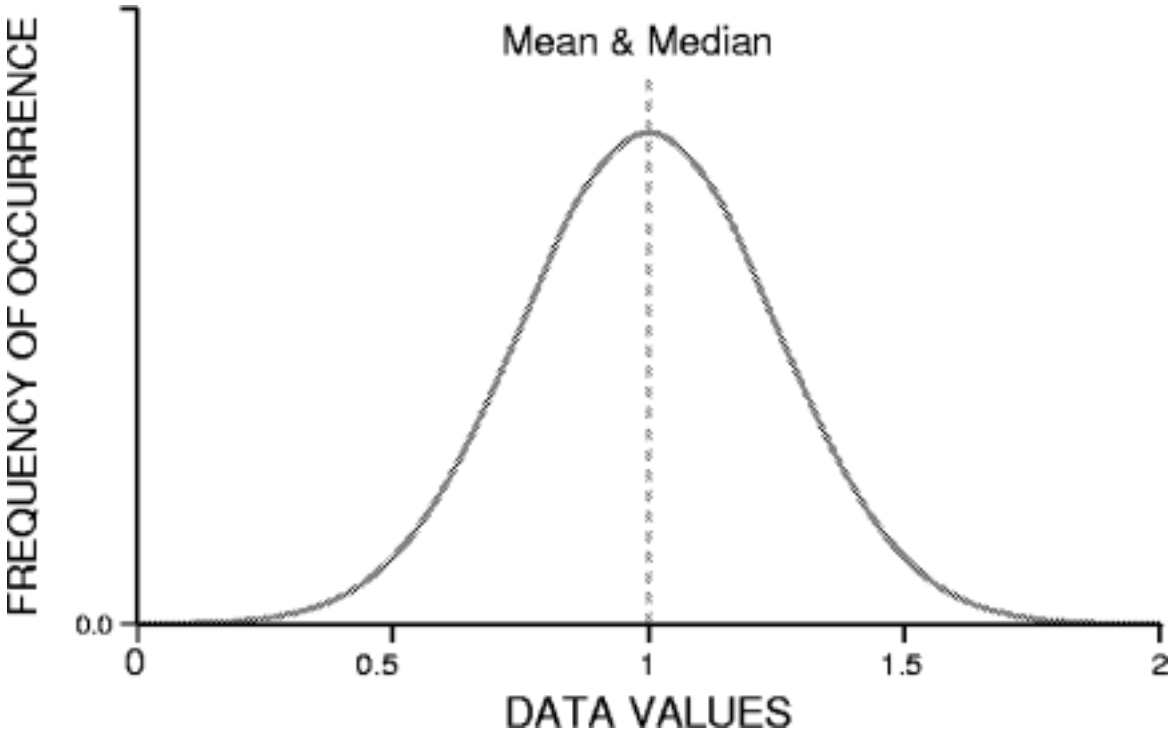


Figure 1.2: Density Function for a Normal Distribution

or low, has a much greater influence on the overall mean $\overline{\mathbf{X}}$ than does a more 'typical' observation, one closer to its $\overline{\mathbf{X}}_{(\mathbf{j})}$.

Another way of illustrating this influence is to realize that the mean is the balance point of the data, when each point is stacked on a number line (figure **??**). Data points further from the center exert a stronger downward force than those closer to the center. If one point near the center were removed, the balance point would only need a small adjustment to keep the data set in balance. But if one outlying value were removed, the balance point would shift dramatically (figure **??**). This sensitivity to the magnitudes of a small number of points in the data set defines why the mean is not a "resistant" measure of location. It is not resistant to changes in the presence of, or to changes in the magnitudes of, a few outlying observations.

When this strong influence of a few observations is desirable, the mean is an appropriate measure of center. This usually occurs when computing units of mass, such as the average concentration of sediment from several samples in a cross-section. Suppose that sediment concentrations closer to the river banks were much higher than those in the center. Waters represented by a bottle of high concentration would exert more influence (due to greater mass of sediment per volume) on the final concentration than waters of low or average concentration. This is entirely appropriate, as the same would occur if the stream itself were somehow mechanically mixed throughout its cross section.

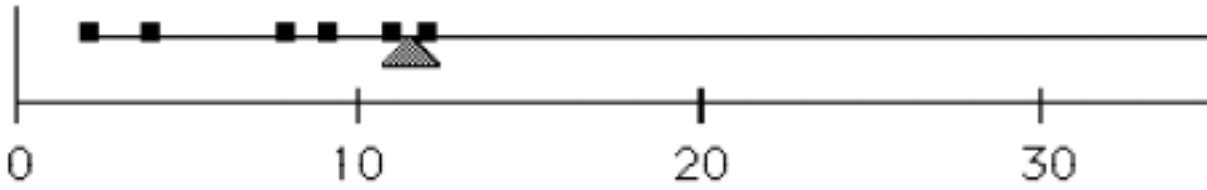


Figure 1.3: The mean (triangle) as balance point of a data set

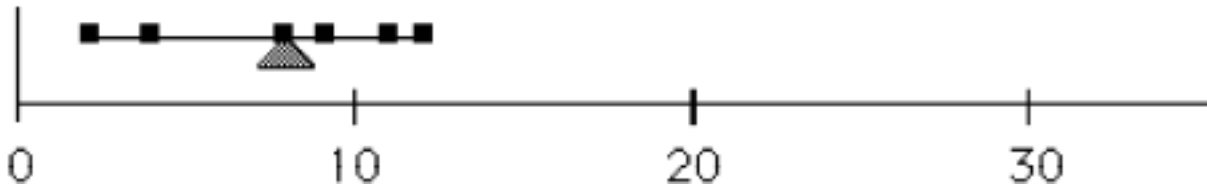


Figure 1.4: Shift of the mean downward after removal of outlier

1.2.2 Resistant Measure – the Median

The median, or 50th percentile $P_{0.50}$, is the central value of the distribution when the data are ranked in order of magnitude. For an odd number of observations,

the median is the data point which has an equal number of observations both above and below it. For an even number of observations, it is the average of the two central observations. To compute the median, first rank the observations from smallest to largest, so that x_1 is the smallest observation, up to x_n, the largest observation. Then

$$\begin{aligned} median\,(P_{0.50}) &= X_{(n+1)/2} && \text{when n is odd, and} \\ median\,(P_{0.50}) &= \frac{1}{2}\left(X_{(n/2)} + X_{(n/2)+1}\right) && \text{when n is even} \end{aligned} \tag{1.4}$$

```
median_x <- median(x)
```

The median is only minimally affected by the magnitude of a single observation, being determined solely by the relative order of observations. This resistance to the effect of a change in value or presence of outlying observations is often a desirable property. To demonstrate the resistance of the median, suppose the last value of the following data set (a) of 7 observations were multiplied by 10 to obtain data set (b):

```
a <- c(2,4,8,9,11,11,12)
mean(a)
```

```
## [1] 8.142857
```

```
median(a)
```

```
## [1] 9
```

```
b <- c(2,4,8,9,11,11,120)
mean(b)
```

```
## [1] 23.57143
```

```
median(b)
```

```
## [1] 9
```

The mean increases from 8.1 to 23.6. The median, the $\frac{(7+1)}{2}$th or 4th lowest data point, is unaffected by the change.

When a summary value is desired that is not strongly influenced by a few extreme observations, the median is preferable to the mean. One such example is the chemical concentration one might expect to find over many streams in a given region. Using the median, one stream with unusually high concentration has no greater effect on the estimate than one with low concentration. The mean concentration may be pulled towards the outlier, and be higher than concentrations found in most of the streams. Not so for the median.

1.2.3 Other Measures of Location

Three other measures of location are less frequently used: the mode, the geometric mean, and the trimmed mean. The mode is the most frequently observed value. It is the value having the highest bar in a histogram. It is far more applicable for grouped data, data which are recorded only as falling into a finite number of categories, than for continuous data. It is very easy to obtain, but a poor measure of location for continuous data, as its value often depends on the arbitrary grouping of those data.

The geometric mean (GM) is often reported for positively skewed data sets. It is the mean of the logarithms, transformed back to their original units.

$$GM = exp\left(\overline{Y}\right), \quad \text{where } Y_i = \ln\left(X_i\right) \tag{1.5}$$

```
geometric_mean_x <- geoMean(x)
```

(in this book the natural, base e logarithm will be abbreviated **ln**, and its inverse e^x abbreviated **exp(x)**). For positively skewed data the geometric mean is usually quite close to the median. In fact, when the logarithms of the data are symmetric, the geometric mean is an unbiased estimate of the median. This is because the median and mean logarithms are equal, as in figure **??**. When transformed back to original units, the geometric mean continues to be an estimate for the median, but is not an estimate for the mean (figure **??**).

Compromises between the median and mean are available by trimming off several of the lowest and highest observations, and calculating the mean of what is left. Such estimates of location are not influenced by the most extreme (and perhaps anomalous) ends of the sample, as is the mean. Yet they allow the magnitudes of most of the values to affect the estimate, unlike the median. These estimators are called "trimmed means", and any desirable percentage of the data may be trimmed away. The most common trimming is to remove 25 percent of the data on each end – the resulting mean of the central 50 percent of data is commonly called the "trimmed mean", but is more precisely the 25 percent trimmed mean. A "0% trimmed mean" is the sample mean itself, while trimming all but 1 or 2 central values produces the median. Percentages of trimming should be explicitly stated when used. The trimmed mean is a resistant estimator of location, as it is not strongly influenced by outliers, and works well for a wide variety of distributional shapes (normal, lognormal, etc.). It may be considered a weighted mean, where data beyond the cutoff 'window' are given a weight of 0, and those within the window a weight of 1.0 (see figure **??**).

1.3 Measures of Spread

It is just as important to know how variable the data are as it is to know their general center or location. Variability is quantified by measures of spread.

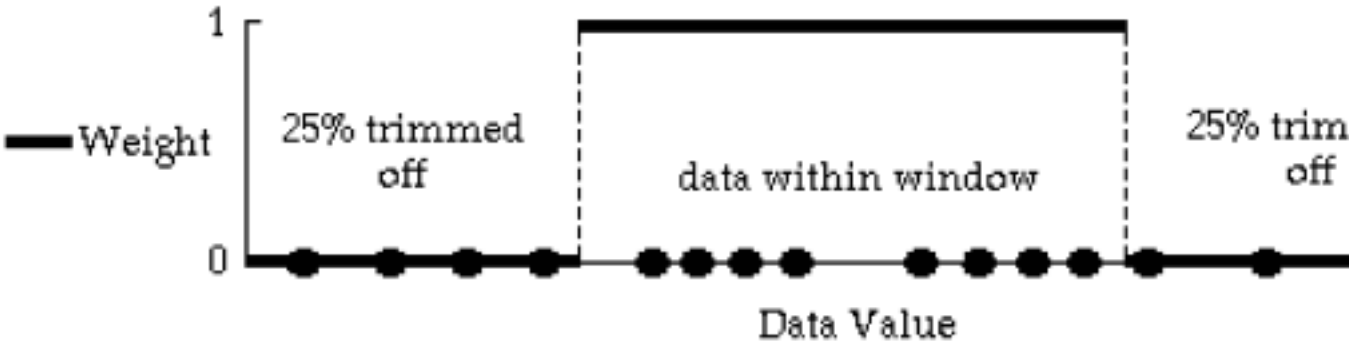


Figure 1.5: Window diagram for the trimmed mean

1.3.1 Classical Measures

The sample variance, and its square root the sample standard deviation, are the classical measures of spread. Like the mean, they are strongly influenced by outlying values.

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{(n-1)} \quad \text{sample variance} \tag{1.6}$$

```
variance_x <- var(x)
```

$$s = \sqrt{s^2} \quad \text{sample standard deviation} \tag{1.7}$$

```
standard_deviation_x <- sd(x)
```

They are computed using the squares of deviations of data from the mean, so that outliers influence their magnitudes even more so than for the mean. When outliers are present these measures are unstable and inflated. They may give the impression of much greater spread than is indicated by the majority of the data set.

1.3.2 Resistant Measures

The interquartile range (IQR) is the most commonly-used resistant measure of spread. It measures the range of the central 50 percent of the data, and is not influenced at all by the 25 percent on either end. It is therefore the width of the non-zero weight window for the trimmed mean of figure **??**.

The IQR is defined as the 75th percentile minus the 25th percentile. The 75th, 50th (median) and 25th percentiles split the data into four equal-sized quarters.

The 75th percentile ($P_{.75}$), also called the upper quartile, is a value which exceeds no more than 75 percent of the data and is exceeded by no more than 25 percent of the data. The 25th percentile ($P_{.25}$) or lower quartile is a value which exceeds no more than 25 percent of the data and is exceeded by no more than 75 percent. Consider a data set ordered from smallest to largest: X_i, i = 1,...n. Percentiles (P_j) are computed using equation (**??**)

$$P_j = X_{(n+1)\bullet j}$$

where n is the sample size of X_i, and

j is the fraction of data less than or equal to the percentile value (for the 25th, 50th and 75th percentiles, j = .25, .50, and .75).

(1.8)

```
interquartile_range_x <- IQR(x, type = 2)
# There are 9 different types you can specify. Type 2 seems to be the method used here
```

Non-integer values of (n+1) • j imply linear interpolation between adjacent values of X. For the example 1 data set given earlier, n = 7, and therefore the 25th percentile is $X_{(7+1)\bullet .25}$ or $X_2 = 4$, the second lowest observation. The 75th percentile is X_6, the 6th lowest observation, or 11. The IQR is therefore 11 − 4 = 7.

```
a
```

```
## [1]  2  4  8  9 11 11 12
```

```
IQR(a, type = 2)
```

```
## [1] 7
```

One resistant estimator of spread other than the IQR is the Median Absolute Deviation, or MAD. The MAD is computed by first listing the absolute value of all differences |d| between each observation and the median. The median of these absolute values is then the MAD.

$$MAD(X_i) = median|d_i|, \quad \text{where } d_i = X_i - median(X_i) \tag{1.9}$$

```
median_absolute_deviation_x <- mad(x, constant = 1)
```

Comparison of each estimate of spread for the Example 1 data set is as follows. When the last value is changed from 12 to 120, the standard deviation increases from 3.8 to 42.7. The IQR and the MAD remain exactly the same.

```
c(IQR(a, type = 2), var(a), mad(a, constant = 1))
```

```
## [1]  7.00000 14.47619  2.00000
```

```
c(IQR(b, type = 2), var(b), mad(b, constant = 1))
```

```
## [1]    7.000 1819.619    2.000
```

1.4 Measures of Skewness

Hydrologic data are typically skewed, meaning that data sets are not symmetric around the mean or median, with extreme values extending out longer in one direction. The density function for a lognormal distribution shown previously as figure **??** illustrates this skewness. When extreme values extend the right tail of the distribution, as they do with figure **??**, the data are said to be skewed to the right, or positively skewed. Left skewness, when the tail extends to the left, is called negative skew.

When data are skewed the mean is not expected to equal the median, but is pulled toward the tail of the distribution. Thus for positive skewness the mean exceeds more than 50 percent of the data, as in figure **??**. The standard deviation is also inflated by data in the tail. Therefore, tables of summary statistics which include only the mean and standard deviation or variance are of questionable value for water resources data, as those data often have positive skewness. The mean and standard deviation reported may not describe the majority of the data very well. Both will be inflated by outlying observations. Summary tables which include the median and other percentiles have far greater applicability to skewed data. Skewed data also call into question the applicability of hypothesis tests which are based on assumptions that the data have a normal distribution. These tests, called parametric tests, may be of questionable value when applied to water resources data, as the data are often neither normal nor even symmetric. Later chapters will discuss this in much detail, and suggest several solutions.

1.4.1 Classical Measure of Skewness

The coefficient of skewness (g) is the skewness measure used most often. It is the adjusted third moment divided by the cube of the standard deviation:

$$g = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(x_i - \overline{X})^3}{s^3} \tag{1.10}$$

```
library(e1071)
skewness_x <- skewness(x, type = 2)
# There are 3 different types you can specify. Type 2 seems to be the method used herein.
```

A right-skewed distribution has positive g; a left-skewed distribution has negative g. Again, the influence of a few outliers is important – an otherwise symmetric distribution having one outlier will produce a large (and possibly misleading) measure of skewness. For the example 1 data, the g skewness coefficient increases from −0.8 to 2.6 when the last data point is changed from 12 to 120.

```
e1071::skewness(a, type = 2)
```

```
## [1] -0.8398416
```

```
e1071::skewness(b, type = 2)
```

```
## [1] 2.610094
```

1.4.2 Resistant Measure of Skewness

A more resistant measure of skewness is the quartile skew coefficient qs (**?**):

$$qs = \frac{(P_{.75} - P_{.50}) - (P_{.50} - P_{.25})}{P_{.75} - P_{.25}} \tag{1.11}$$

```
quartiles_x <- quantile(x, prob = c(0.25, 0.50, 0.75), type = 2)
# There are 9 different types you can specify. Type 2 seems to be the method used here
quartile_skew_x <- ((quartiles_x[3] - quartiles_x[2]) - (quartiles_x[2] - quartiles_x[1
```

the difference in distances of the upper and lower quartiles from the median, divided by the IQR. A right-skewed distribution again has positive qs; a left-skewed distribution has negative qs. Similar to the trimmed mean and IQR, qs uses the central 50 percent of the data. For the example 1 data, qs = (11−9) − (9−4) / (11−4) = −0.43 both before and after alteration of the last data point. Note that this resistance may be a liability if sensitivity to a few observations is important.

```
quartiles_a <- quantile(a, prob = c(0.25, 0.50, 0.75), type = 2, names = FALSE)
quartile_skew_a <- ((quartiles_a[3] - quartiles_a[2]) - (quartiles_a[2] - quartiles_a[1
quartile_skew_a
```

```
## [1] -0.4285714
```

```
quartiles_b <- quantile(b, prob = c(0.25, 0.50, 0.75), type = 2, names = FALSE)
quartile_skew_b <- ((quartiles_b[3] - quartiles_b[2]) - (quartiles_b[2] - quartiles_b[1
quartile_skew_b
```

```
## [1] -0.4285714
```

1.5 Other Resistant Measures

Other percentiles may be used to produce a series of resistant measures of location, spread and skewness. For example, the 10 percent trimmed mean can be coupled with the range between the 10th and 90th percentiles as a measure of

spread, and a corresponding measure of skewness:

$$qs_{.10} = \frac{(P_{.90} - P_{.50}) - (P_{.50} - P_{.10})}{P_{.90} - P_{.10}} \tag{1.12}$$

```
quartiles_x <- quantile(x, prob = c(0.10, 0.50, 0.90), type = 2)
# There are 9 different types you can specify. Type 2 seems to be the method used herein.
quartile_skew_10_x <- ((quartiles_x[3] - quartiles_x[2]) - (quartiles_x[2] - quartiles_x[1])) / (
```

to produce a consistent series of resistant statistics. Geologists have used the 16th and 84th percentiles for many years to compute a similar series of robust measures of the distributions of sediment particles (**?**). However, measures based on quartiles have become generally standard, and other measures should be clearly defined prior to their use. The median, IQR, and quartile skew can be easily summarized graphically using a boxplot (see Chapter **??**) and are familiar to most data analysts.

1.6 Outliers

Outliers, observations whose values are quite different than others in the data set, often cause concern or alarm. They should not. They are often dealt with by throwing them away prior to describing data, or prior to some of the hypothesis test procedures of later chapters. Again, they should not. Outliers may be the most important points in the data set, and should be investigated further.

It is said that data on the Antarctic ozone "hole", an area of unusually low ozone concentrations, had been collected for approximately 10 years prior to its actual discovery. However, the automatic data checking routines during data processing included instructions on deleting "outliers". The definition of outliers was based on ozone concentrations found at mid-latitudes. Thus all of this unusual data was never seen or studied for some time. If outliers are deleted, the risk is taken of seeing only what is expected to be seen.

Outliers can have one of three causes:

1. a measurement or recording error.
2. an observation from a population not similar to that of most of the data, such as a flood caused by a dam break rather than by precipitation.
3. a rare event from a single population that is quite skewed.

The graphical methods of the Chapter **??** are very helpful in identifying outliers. Whenever outliers occur, first verify that no copying, decimal point, or other obvious error has been made. If not, it may not be possible to determine if the point is a valid one. The effort put into verification, such as re-running the sample in the laboratory, will depend on the benefit gained versus the cost of verification. Past events may not be able to be duplicated. If no error can

be detected and corrected, **outliers should not be discarded based solely on the fact that they appear unusual**. Outliers are often discarded in order to make the data nicely fit a preconceived theoretical distribution such as the normal. There is no reason to suppose that they should! The entire data set may arise from a skewed distribution, and taking logarithms or some other transformation may produce quite symmetrical data. Even if no transformation achieves symmetry, outliers need not be discarded. Rather than eliminating actual (and possibly very important) data in order to use analysis procedures requiring symmetry or normality, procedures which are resistant to outliers should instead be employed. If computing a mean appears of little value because of an outlier, the median has been shown to be a more appropriate measure of location for skewed data. If performing a t-test (described later) appears invalidated because of the non-normality of the data set, use a rank-sum test instead.

In short, let the data guide which analysis procedures are employed, rather than altering the data in order to use some procedure having requirements too restrictive for the situation at hand.

1.7 Transformations

Transformations are used for three purposes:

1. to make data more symmetric,
2. to make data more linear, and
3. to make data more constant in variance.

Some water resources scientists fear that by transforming data, results are derived which fit preconceived ideas. Therefore, transformations are methods to 'see what you want to see' about the data. But in reality, serious problems can occur when procedures assuming symmetry, linearity, or homoscedasticity (constant variance) are used on data which do not possess these required characteristics. Transformations can produce these characteristics, and thus the use of transformed variables meets an objective. Employment of a transformation is not merely an arbitrary choice.

One unit of measurement is no more valid a priori than any other. For example, the negative logarithm of hydrogen ion concentration, pH, is as valid a measurement system as hydrogen ion concentration itself. Transformations like the square root of depth to water at a well, or cube root of precipitation volume, should bear no more stigma than does pH. These measurement scales may be more appropriate for data analysis than are the original units. **?** has written an excellent article on hidden transformations, consistently taken for granted, which are in common use by everyone. Octaves in music are a logarithmic transform of frequency. Each time a piano is played a logarithmic transform is

employed! Similarly, the Richter scale for earthquakes, miles per gallon for gasoline consumption, f-stops for camera exposures, etc. all employ transformations. In the science of data analysis, the decision of which measurement scale to use should be determined by the data, not by preconceived criteria. The objectives for use of transformations are those of symmetry, linearity and homoscedasticity. In addition, the use of many resistant techniques such as percentiles and nonparametric test procedures (to be discussed later) are invariant to measurement scale. The results of a rank-sum test, the nonparametric equivalent of a t-test, will be exactly the same whether the original units or logarithms of those units are employed.

1.7.1 The Ladder of Powers

In order to make an asymmetric distribution become more symmetric, the data can be transformed or re-expressed into new units. These new units alter the distances between observations on a line plot. The effect is to either expand or contract the distances to extreme observations on one side of the median, making it look more like the other side. The most commonly-used transformation in water resources is the logarithm. Logs of water discharge, hydraulic conductivity, or concentration are often taken before statistical analyses are performed.

Transformations usually involve power functions of the form $y = x$, where x is the untransformed data, y the transformed data, and the power exponent. In figure **??** the values of are listed in the "ladder of powers" (**?**), a useful structure for determining a proper value of .

As can be seen from the ladder of powers, any transformations with less than 1 may be used to make right-skewed data more symmetric. Constructing a boxplot or Q-Q plot (see Chapter **??**) of the transformed data will indicate whether the transformation was appropriate. Should a logarithmic transformation overcompensate for right skewness and produce a slightly leftskewed distribution, a 'milder' transformation with closer to 1, such as a square-root or cuberoot transformation, should be employed instead. Transformations with > 1 will aid in making left-skewed data more symmetric. \begin{figure}

Use	θ	Transformation	Name
		•	
		•	
for (−)		•	
skewness	3	x^3	cube
	2	x^2	square
	1	x	original units
	1/2	$\sqrt{x}$	square root
	1/3	$\sqrt[3]{x}$	cube root
	0	log(x)	logarithm
for (+) skewness			
	−1/2	$-1/\sqrt{x}$	reciprocal root
	−1	−1/x	reciprocal
	−2	$-1/x^2$	
		•	
		•	
		•	

}

\caption{“LADDER OF POWERS” (modified from **?**)} \end{figure}
However, the tendency to search for the ‘best’ transformation should be avoided. For example, when dealing with several similar data sets, it is probably better to find one transformation which works reasonably well for all, rather than using slightly different ones for each. It must be remembered that each data set is a sample from a larger population, and another sample from the same population will likely indicate a slightly different ‘best’ transformation. Determination of ‘best’ in great precision is an approach that is rarely worth the effort.

Exercises

1.1

Yields in wells penetrating rock units without fractures were measured by **?**, and are given below. Calculate the

a. mean
b. trimmed mean
c. geometric mean
d. median
e. compare these estimates of location. Why do they differ?

1.2

For the well yield data of exercise 1.1, calculate the

a. standard deviation
b. interquartile range
c. MAD
d. skew and quartile skew.

Discuss the differences between a through c.

1.3

Ammonia plus organic nitrogen (in mg/L) was measured in samples of precipitation by **?**. Some of their data are presented below. Compute summary statistics for these data. Which observation might be considered an outlier? How should this value affect the choice of summary statistics used

a. to compute the mass of nitrogen falling per square mile.
b. to compute a “typical” concentration and variability for these data?

Chapter 2

Graphical Data Analysis

Perhaps it seems odd that a chapter on graphics appears at the front of a text on statistical methods. We believe this is very appropriate, as graphs provide crucial information to the data analyst which is difficult to obtain in any other way. For example, figure **??** shows eight scatterplots, all of which have exactly the same correlation coefficient. Computing statistical measures without looking at a plot is an invitation to misunderstanding data, as figure **??** illustrates. Graphs provide visual summaries of data which more quickly and completely describe essential information than do tables of numbers.

Graphs are essential for two purposes:

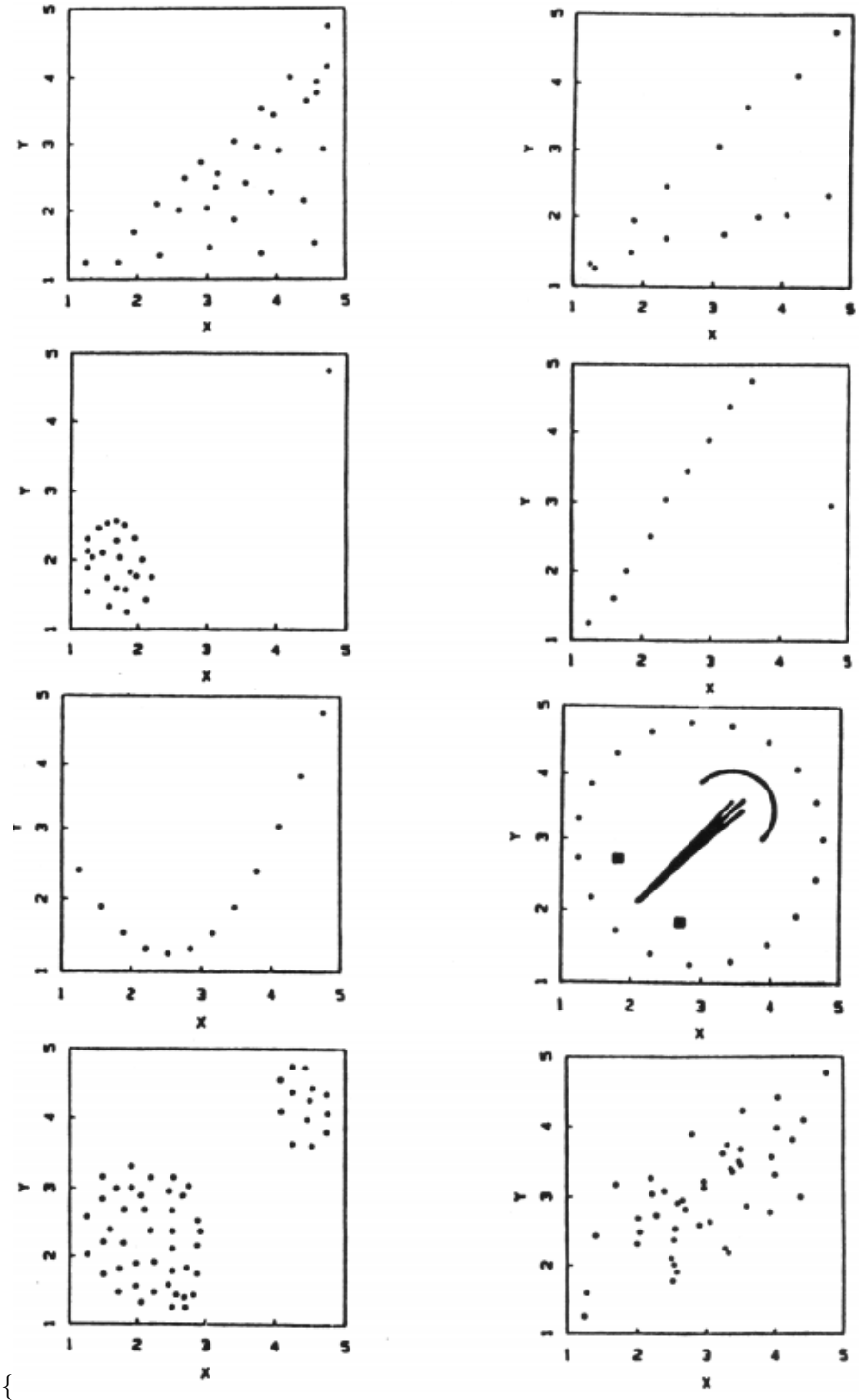
1. to provide insight for the analyst into the data under scrutiny, and
2. to illustrate important concepts when presenting the results to others.

The first of these tasks has been called exploratory data analysis (EDA), and is the subject of this chapter. EDA procedures often are (or should be) the 'first look' at data. Patterns and theories of how the system behaves are developed by observing the data through graphs. These are inductive procedures – the data are summarized rather than tested. Their results provide guidance for the selection of appropriate deductive hypothesis testing procedures.

Once an analysis is complete, the findings must be reported to others. Whether a written report or oral presentation, the analyst must convince the audience that the conclusions reached are supported by the data. No better way exists to do this than through graphics. Many of the same graphical methods which have concisely summarized the information for the analyst will also provide insight into the data for the reader or audience.

The chapter begins with a discussion of graphical methods for analysis of a single data set. Two methods are particularly useful: boxplots and probability plots. Their construction is presented in detail. Next, methods for comparison of two or more groups of data are discussed. Then bivariate plots

(scatterplots) are presented, with an especially useful enhancement called a smooth. The chapter ends with a discussion of plots appropriate for multivariate data. \begin{figure}


Y
X
1
2
3
4
5

{

}

\caption{Eight scatterplots all with correlation coefficient r = 0.70 (**?**)}
\end{figure} Throughout sections **??** and **??** two data sets will be used to compare and contrast the effectiveness of each graphical method. These are annual streamflow (in cubic feet per second, or cfs) for the Licking River at Catawba, Kentucky, from 1929 through 1983, and unit well yields (in gallons per minute per foot of water-bearing material) for valleys without fracturing in Virginia (**?**).

2.1 Graphical Analysis of Single Data Sets

2.1.1 Histograms

Histograms are familiar graphics, and their construction is detailed in numerous introductory texts on statistics. Bars are drawn whose height is the number n_i, or fraction n_i/n, of data falling into one of several categories or intervals (figures **??** & **??**). **?** suggest that, for a sample size of n, the number of intervals k should be the smallest integer such that 2^k n.

Histograms have one primary deficiency – their visual impression depends on the number of categories selected for the plot. For example, compare figure **??** with **??**. Both are histograms of the same data: annual streamflows for the Licking River. Comparisons of shape and similarity among these two figures and the many other possible histograms of the same data depend on the choice of bar widths and centers. False impressions that these are different distributions might be given by characteristics such as the gap around 6,250 cfs. It is seen in **??** but not in **??**.

Histograms are quite useful for depicting large differences in shape or symmetry, such as whether a data set appears symmetric or skewed. They cannot be used for more precise judgements such as depicting individual values. Thus from figure **??** the lowest flow is seen to be larger than 750 cfs, but might be as large as 2,250 cfs. More detail is given in **??**, but this lowest observed discharge is still only known to be somewhere between 500 to 1,000 cfs.

For data measured on a continuous scale (such as streamflow or concentration), histograms are not the best method for graphical analysis. The process of forcing continuous data into discrete categories may obscure important characteristics of the distribution. However, histograms are excellent when displaying data which have natural categories or groupings. Examples of such data would include the number of individual organisms found at a stream site grouped by species type, or the number of water-supply wells exceeding some critical yield grouped by geologic unit.

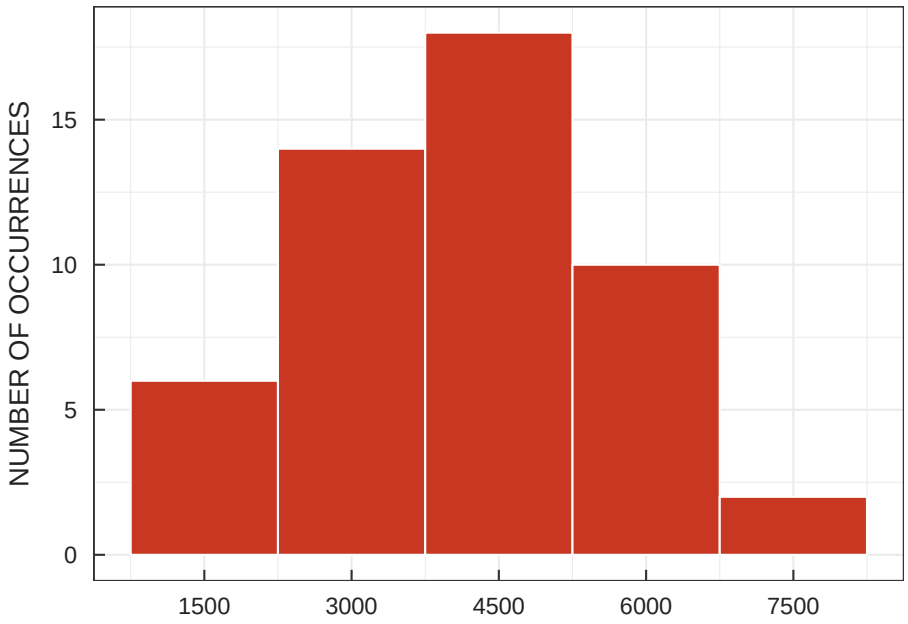


Figure 2.1: Histogram of annual streamflow for the Licking River

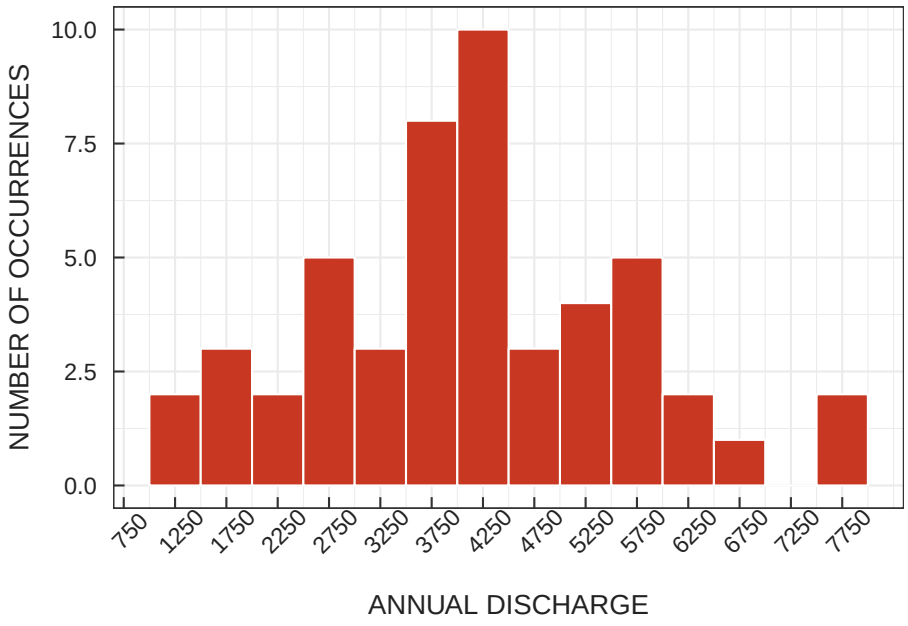


Figure 2.2: Second histogram of same data, but with different interval divisions

2.1.2 Stem and Leaf Diagrams

Figure 2.4 shows a stem and leaf (S-L) diagram for the Licking River streamflow data with the same divisions as in figure **??**. Stem and leaf diagrams are like histograms turned on their side, with data magnitudes to two significant digits presented rather than only bar heights. Individual values are easily found. The S-L profile is identical to the histogram and can similarly be used to judge shape and symmetry, but the numerical information adds greater detail. One S-L could function as both a table and a histogram for small data sets.

An S-L is constructed by dividing the range of the data into roughly 10 intervals, and placing the first digit corresponding to these intervals to the left of the vertical line. This is the 'stem', ranging from 1 to 8 (0 to 8000+ cfs) in figure 2.4*. Each observation is then represented by one digit to the right of the line (the 'leaves'), so that the number of leaves equals the number of observations falling into that interval. To provide more detail, figure 2.4* has two lines for each stem digit, split to allow 5 leaf digits per line (0-4 and 5-9). Here the first stem is the stem for leaves less than 5, and the second stem for leaves greater than or equal to 5. For example, in figure 2.4* three observations occur between 1500 and 2000 cfs, with values of 1800, 1900, and 1900 cfs.

The lowest flow is now seen to be between 1000 and 1500 cfs. The gap between 7,000 to 7,500 cfs is still evident, and now the numerical values of the three highest flows are presented. Comparisons between distributions still remain difficult using S-L plots, however, due to the required arbitrary choice of group boundaries.

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##   1 | 34
##   1 | 899
##   2 | 2
##   2 | 566789
##   3 | 013
##   3 | 5667789
##   4 | 000111222
##   4 | 5558
##   5 | 00034
##   5 | 66789
##   6 | 12
##   6 | 5
##   7 |
##   7 | 5
##   8 | 0
```

```
## [1] "Fig 2.4* Stem and Leaf Plot of Annual Streamflow"
```

2.1.3 Quantile Plots

Quantile plots visually portray the quantiles, or percentiles (which equal the quantiles times 100) of the distribution of sample data. Quantiles of importance such as the median are easily discerned (quantile, or cumulative frequency = 0.5). With experience, the spread and skewness of the data, as well as any bimodal character, can be examined. Quantile plots have three advantages:

1. Arbitrary categories are not required, as with histograms or S-L's.
2. All of the data are displayed, unlike a boxplot.
3. Every point has a distinct position, without overlap.

Figure **??** is a quantile plot of the streamflow data from figures **??** and **??**. Attributes of the data such as the gap between 6500 and 7500 cfs (indicated by the nearly horizontal line segment) are evident. The percent of data in the sample less than a given cfs value can be read from the graph with much greater accuracy than from a histogram.

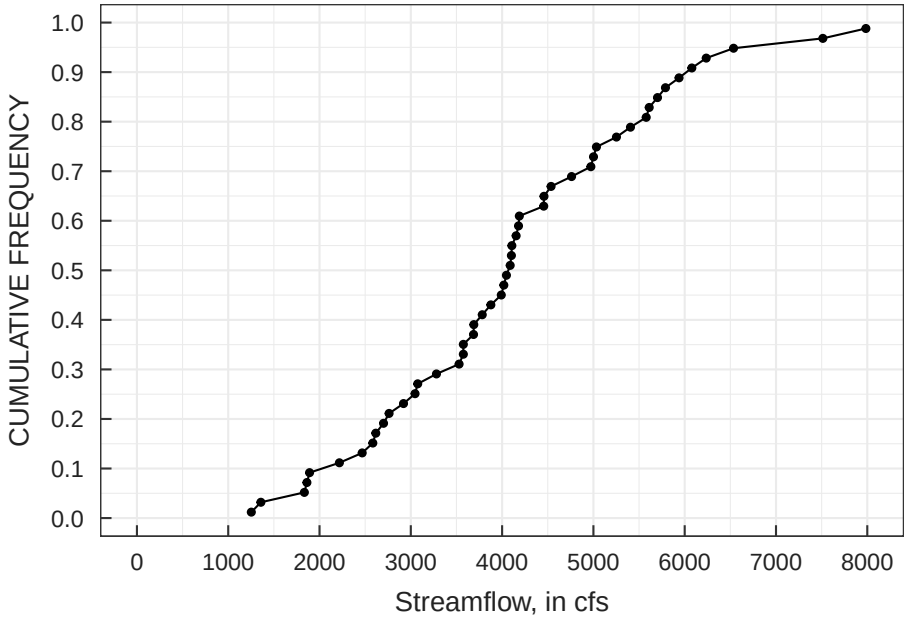


Figure 2.3: Quantile plot of the Licking River annual streamflow data

2.1.3.1 Construction of a quantile plot

To construct a quantile plot, the data are ranked from smallest to largest. The smallest data value is assigned a rank i = 1, while the largest receives a rank i = n, where n is the sample size of the data set. The data values themselves are plotted along one axis, usually the horizontal axis. On the other axis is the "plotting position", which is a function of the rank i and sample size n. As discussed in the next section, the Cunnane plotting position $p_i = (i-0.4)/(n+0.2)$ is used in this book. Below are listed the first and last 5 of the 50 data pairs used in construction of figure **??**. When tied data values are present, each is assigned a separate plotting position (the plotting positions are not averaged). In this way tied values are portrayed as a vertical "cliff" on the plot.

```
##     idx       V1        p_i
##  1:   1 1253.083 0.01195219
##  2:   2 1357.667 0.03187251
##  3:   3 1833.891 0.05179283
##  4:   4 1861.817 0.07171315
##  5:   5 1891.042 0.09163347
##  6:  46 6078.083 0.90836653
##  7:  47 6235.858 0.92828685
##  8:  48 6535.075 0.94820717
##  9:  49 7513.200 0.96812749
## 10:  50 7984.617 0.98804781
```

Quantile plots are sample approximations of the cumulative distribution function (cdf) of a continuous random variable. The cdf for a normal distribution is shown in figure **??**. A second approximation is the sample (or empirical) cdf, which differs from quantile plots in its vertical scale. The vertical axis of a sample cdf is the probability i/n of being less than or equal to that observation. The largest observation has i/n = 1, and so has a zero probability of being exceeded. For samples (subsets) taken from a population, a nonzero probability of exceeding the largest value observed thus far should be recognized. This is done by using the plotting position, a value less than i/n, on the vertical axis of the quantile plot. As sample sizes increase, the quantile plot will more closely mimic the underlying population cdf.

2.1.3.1.1 Plotting positions

Variations of quantile plots are used frequently for three purposes:

1. to compare two or more data distributions (a Q-Q plot),
2. to compare data to a normal distribution (a probability plot), and
3. to calculate frequencies of exceedance (a flow-duration curve).

Unfortunately, different plotting positions have traditionally been used for

each of the above three purposes. It would be desirable instead to use one formula that is suitable for all three. Numerous plotting position formulas have been suggested, most having the general formula

$$p = (i - a)/(n + 1 - 2a)$$

where a varies from 0 to 0.5. Five of the most commonly-used formulas are:

	Reference	a	Formula
?		0	$i/(n+1)$
?		0.375	$(i-0.375)/(n+0.25)$
?		0.4	$(i-0.4)/(n+0.2)$
?		0.44	$(i-0.44)/(n+0.12)$
?		0.5	$(i-0.5)/n$

The Weibull formula has long been used by hydrologists in the United States for plotting flowduration and flood-frequency curves (**?**). It is used in Bulletin 17B, the standard reference for determining flood frequencies in the United States (**?**). The Blom formula is best for comparing data quantiles to those of a normal distribution in probability plots, though all of the above formulas except the Weibull are acceptable for that purpose (**?**). The Hazen formula is used by **?** for comparing two or more data sets using Q-Q plots.

Separate formulae could be used for the situations in which each is optimal. In this book we instead use one formula, the Cunnane formula given above, for all three purposes. We do this in an attempt to simplify. The Cunnane formula was chosen because

1. it is acceptable for normal probability plots, being very close to Blom.
2. it is used by Canadian and some European hydrologists for plotting flow-duration and flood-frequency curves. **?** presents the arguments for use of this formula over the Weibull when calculating exceedance probabilities.

For convenience when dealing with small sample sizes, table B1 of the Appendix presents Cunnane plotting positions for sample sizes n = 5 to 20.

2.1.4 Boxplots

A very useful and concise graphical display for summarizing the distribution of a data set is the boxplot (figure **??**). Boxplots provide visual summaries of

1. the center of the data (the median–the center line of the box)
2. the variation or spread (interquartile range–the box height)
3. the skewness (quartile skew–the relative size of box halves)
4. presence or absence of unusual values ("outside" and "far outside" values).

Boxplots are even more useful in comparing these attributes among several data sets.

Compare figures **??** and **??**, both of the Licking River data. Boxplots do not present all of the data, as do stem-and-leaf or quantile plots. Yet presenting all data may be more detail than is necessary, or even desirable. Boxplots do provide concise visual summaries of essential data characteristics. For example, the symmetry of the Licking River data is shown in figure **??** by the similar sizes of top and bottom box halves, and by the similar lengths of whiskers. In contrast, in figure **??** the taller top box halves and whiskers indicate a right-skewed distribution, the most commonly occurring shape for water resources data. Boxplots are often put side-by-side to visually compare and contrast groups of data.

Three commonly used versions of the boxplot are described as follows (figure **??**). Any of the three may appropriately be called a boxplot.

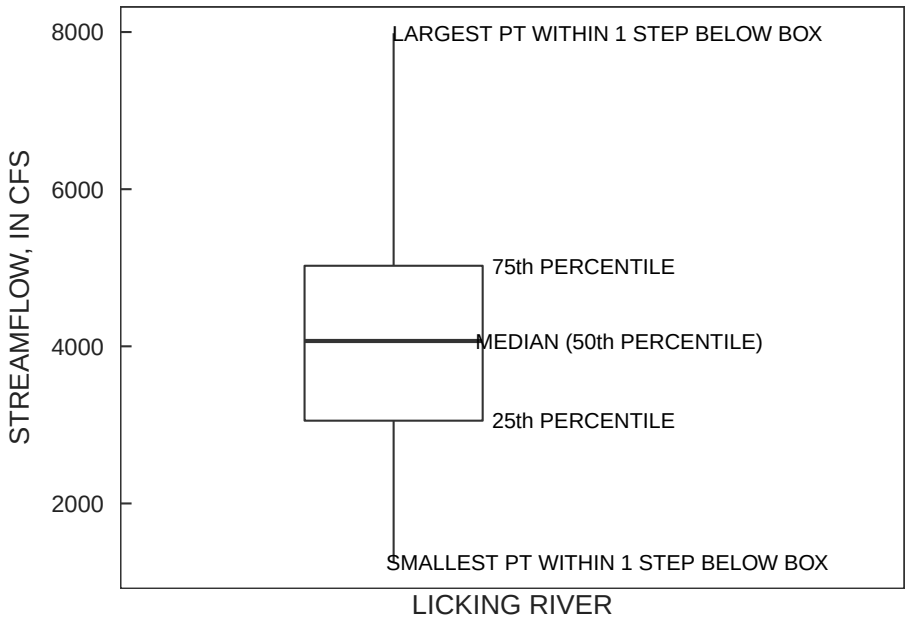


Figure 2.4: Boxplot of the Licking River annual streamflow data

2.1.4.1 Simple boxplot

The simple boxplot was originally called a "box-and-whisker" plot by **?**. It consists of a center line (the median) splitting a rectangle defined by the upper and lower hinges (very similar to quartiles – see appendix). Whiskers are lines

drawn from the ends of the box to the maximum and minimum of the data, as depicted in graph a of figure **??**.

2.1.4.2 Standard boxplot

Tukey's "schematic plot" has become the most commonly used version of a boxplot (middle graph in figure **??**), and will be the type of boxplot used throughout this book. With this standard boxplot, outlying values are distinguished from the rest of the plot. The box is as defined above. However, the whiskers are shortened to extend only to the last observation within one step beyond either end of the box ("adjacent values"). A step equals 1.5 times the height of the box (1.5 times the interquartile range). Observations between one and two steps from the box in either direction, if present, are plotted individually with an asterisk ("outside values"). Outside values occur fewer than once in 100 times for data from a normal distribution. Observations farther than two steps beyond the box, if present, are distinguished by plotting them with a small circle ("far-out values"). These occur fewer than once in 300,000 times for a normal distribution. The occurrence of outside or far-out values more frequently than expected gives a quick visual indication that data may not originate from a normal distribution.

2.1.4.3 Truncated boxplot

In a third version of the boxplot (left graph of figure **??**), the whiskers are drawn only to the 90th and 10th percentiles of the data set. The largest 10 percent and smallest 10 percent of the data are not shown. This version could easily be confused with the simple boxplot, as no data appear beyond the whiskers, and should be clearly defined as having eliminated the most extreme 20 percent of data. It should be used only when the extreme 20 percent of data are not of interest.

In a variation on the truncated boxplot, **?** plotted all observations beyond the 10th and 90th percentile-whiskers individually, calling this a "box graph". The weakness of this style of graph is that 10 percent of the data will always be plotted individually at each end, and so the plot is far less effective than a standard boxplot for defining and emphasizing unusual values.

Further detail on construction of boxplots may be found in the appendix, and in **?** and **?**.

2.1.5 Probability Plots

Probability plots are used to determine how well data fit a theoretical distribution, such as the normal, lognormal, or gamma distributions. This

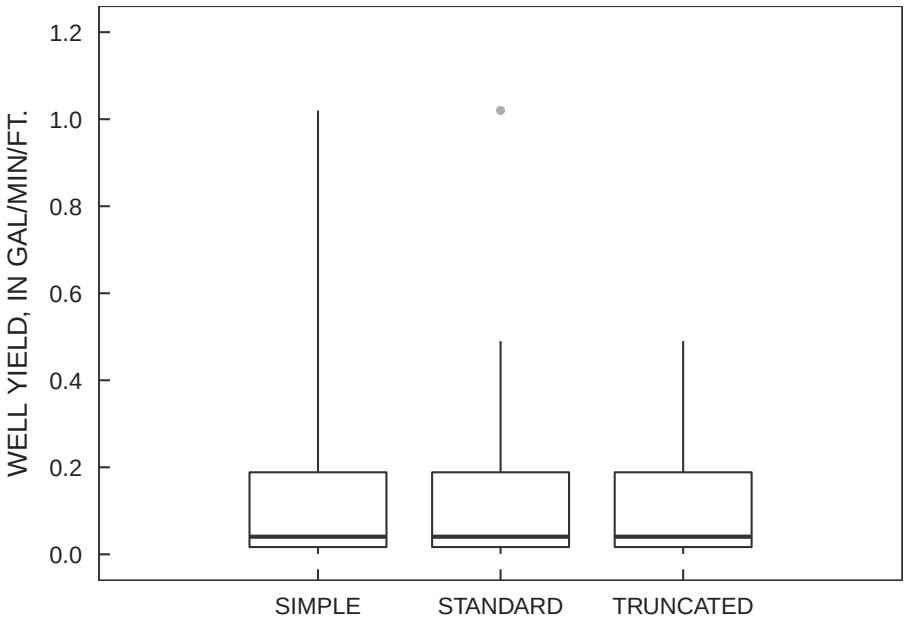


Figure 2.5: Three versions of the boxplot (unit well yield data).

could be attempted by visually comparing histograms of sample data to density curves of the theoretical distributions such as figures **??** and **??**. However, research into human perception has shown that departures from straight lines are discerned more easily than departures from curvilinear patterns. By expressing the theoretical distribution as a straight line, departures from the distribution are more easily perceived. This is what occurs with a probability plot.

To construct a probability plot, quantiles of sample data are plotted against quantiles of the standardized theoretical distribution. In figure **??**, quantiles from the quantile plot of the Licking River streamflow data (lower scale) are overlain with the S-shaped quantiles of the standard normal distribution (upper scale). For a given cumulative frequency (plotting position, p), quantiles from each curve are paired and plotted as one point on the probability plot, figure **??**. Note that quantiles of the data are simply the observation values themselves, the pth quantiles where $p = (i - 0.4)/(n + 0.2)$. Quantiles of the standard normal distribution are available in table form in most textbooks on statistics. Thus, for each observation, a pair of quantiles is plotted in figure **??** as one point. For example, the median ($p = 0.5$) equals 0 for the standard normal, and 4068 cfs for the Licking River data. The point (0,4068) is one point included in figure **??**. Data closely approximating the shape of the theoretical distribution, in this case a normal distribution, will plot near to a straight line.

To illustrate the construction of a probability plot in detail, data on unit well yields (y_i) from **?** will be plotted versus their normal quantiles (also called normal scores). The data are ranked from the smallest ($i = 1$) to largest (i=n), and their corresponding plotting positions $p_i = (i - 0.4)/(n + 0.2)$ calculated. Normal quantiles (Zp) for a given plotting position p_i may be obtained in one of three ways:

a. from a table of the standard normal distribution found in most statistics textbooks
b. from table B2 in the Appendix, which presents standard normal quantiles for the Cunnane plotting positions of table B1
c. from a computerized approximation to the inverse standard normal distribution available in many statistical packages, or as listed by **?**.

Entering the table with $p_i = .05$, for example, will provide a $Zp = -1.65$. Note that since the median of the standard normal distribution is 0, Zp will be symmetrical about the median, and only half of the Zp values must be looked up:

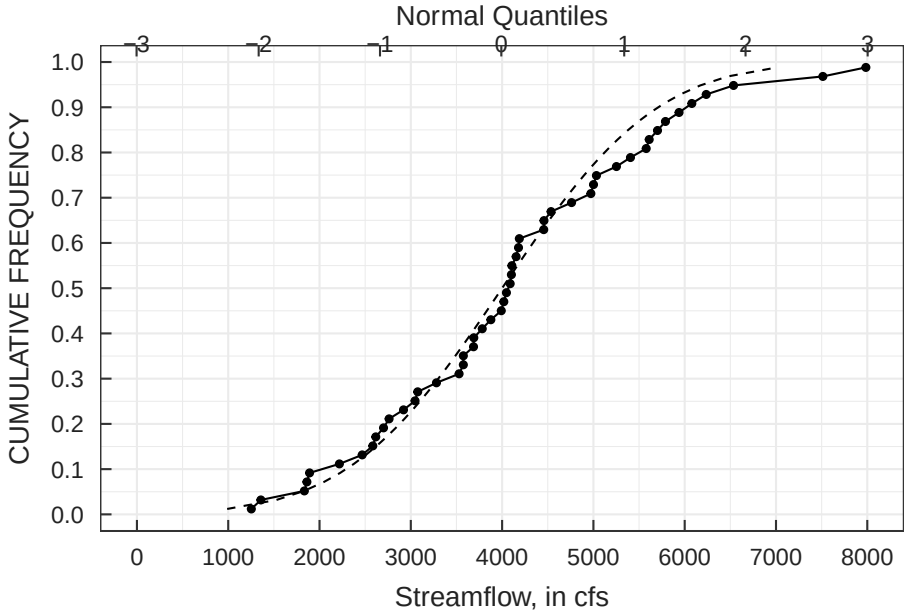


Figure 2.6: Overlay of Licking River and standard normal distribution quantile plots

```
##     unit_well_yield        p_i         zp
##  1:           0.001 0.04918033 -1.6528536
##  2:           0.003 0.13114754 -1.1209830
##  3:           0.007 0.21311475 -0.7956603
```

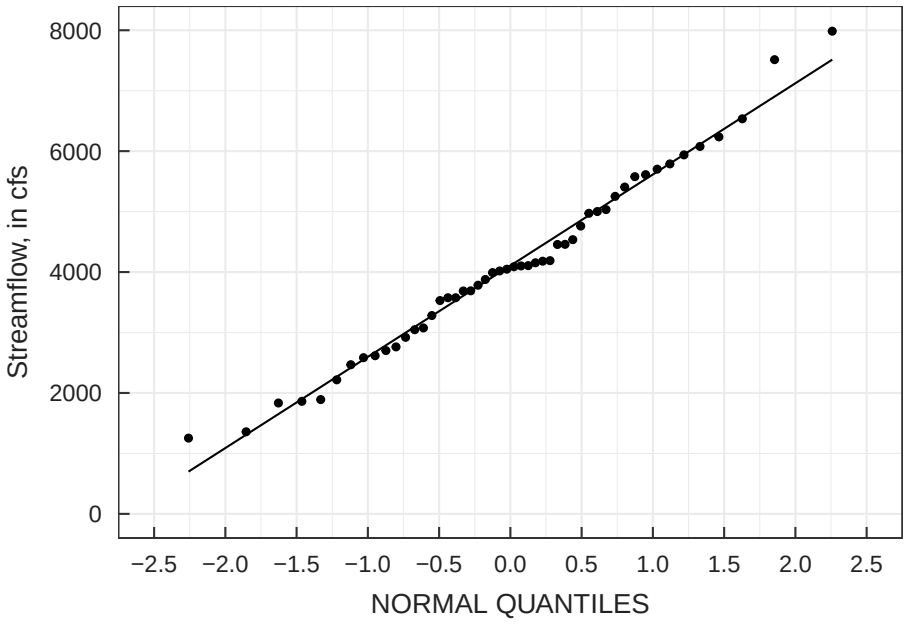


Figure 2.7: Probability plot of the Licking River data

```
##  4:          0.020 0.29508197 -0.5385985
##  5:          0.030 0.37704918 -0.3132400
##  6:          0.040 0.45901639 -0.1029120
##  7:          0.041 0.54098361  0.1029120
##  8:          0.077 0.62295082  0.3132400
##  9:          0.100 0.70491803  0.5385985
## 10:          0.454 0.78688525  0.7956603
## 11:          0.490 0.86885246  1.1209830
## 12:          1.020 0.95081967  1.6528536
```

For comparison purposes, it is helpful to plot a reference straight line on the plot. The solid line on figure **??** is the normal distribution which has the same mean and standard deviation as do the sample data. This reference line is constructed by plotting $\overline{y}$ as the y intercept of the line ($Zp = 0$), so that the line is centered at the point $(0, \overline{y})$, the mean of both sets of quantiles. The standard deviation s is the slope of the line on a normal probability plot, as the quantiles of a standard normal distribution are in units of standard deviation. Thus the line connects the points $(0, \overline{y})$ and $(1, \overline{y} + s)$.

2.1.5.1 Probability paper

Specialized 'probability paper' is often used for probability plots. This paper simply retransforms the linear scale for quantiles of the standard distribution back into a nonlinear scale of plotting positions (figure **??**). There is no difference between the two versions except for the horizontal scale. With probability paper the horizontal axis can be directly interpreted as the percent probability of occurrence, the plotting position times 100. The linear quantile scale of figure **??** is sometimes included on probability paper as 'probits,' where a probit = normal quantile + 5.0. Probability paper is available for distributions other than the normal, but all are constructed the same way, using standardized quantiles of the theoretical distribution.

In figure **??** the lower horizontal scale results from sorting the data in increasing order, and assigning rank 1 to the smallest value. This is commonly done in water-quality and low-flow studies. Had the data been sorted in decreasing order, assigning rank 1 to the largest value as is done in flood-flow studies, the upper scale would result – the percent exceedance. Either horizontal scale may be obtained by subtracting the other from 100 percent.

2.1.5.2 Deviations from a linear pattern

If probability plots do not exhibit a linear pattern, their nonlinearity will indicate why the data do not fit the theoretical distribution. This is additional information that hypothesis tests for normality (described later) do not provide. Three typical conditions resulting in deviations from linearity are: asymmetry or skewness, outliers, and heavy tails of the distribution. These are discussed below.

Figure **??** is a probability plot of the base 10 logarithms of the Licking River data. The data are negatively (left) skewed. This is seen in figure **??** as a greater slope on the left-hand side of the plot, producing a slightly convex shape. Figure **??** shows a right-skewed distribution, the unit well yield data. The lower bound of zero, and the large slope on the right-hand side of the plot produces an overall concave shape. Thus probability plots can be used to indicate what type of transformation is needed to produce a more symmetric distribution. The degree of curvature gives some indication of the severity of skewness, and therefore the degree of transformation required.

Outliers appear on probability plots as departures from the pattern of the rest of the data. Figure **??** is a probability plot of the Licking River data, but the two largest observations have been altered (multiplied by 3). Compare figures **??** and **??**. Note that the majority of points in figure **??** still retain a linear pattern, with the two outliers offset from that pattern. Note that the straight line, a normal distribution with mean and standard deviation equal to those of

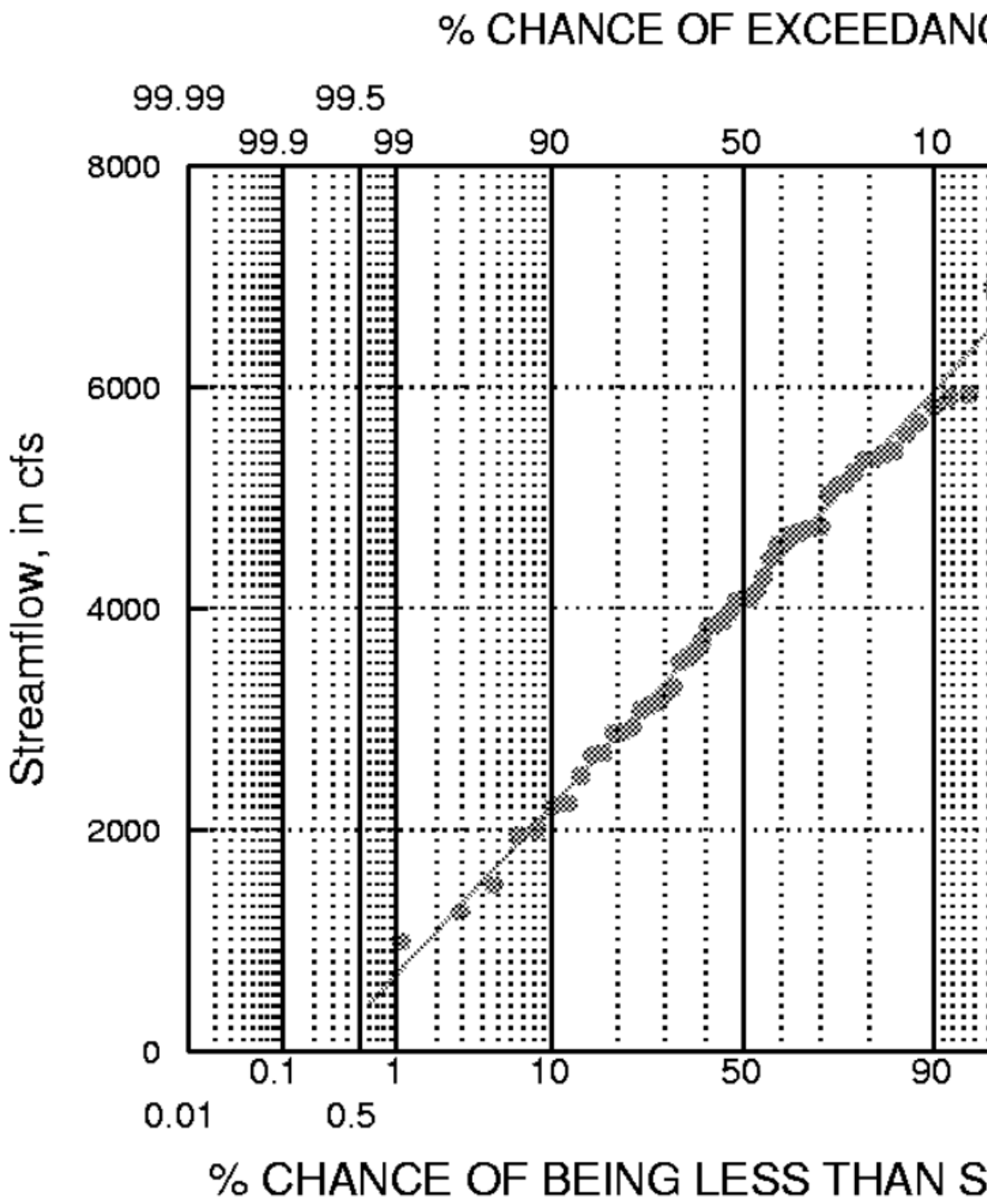


Figure 2.8: Probability plot of Licking River data on probability paper

the altered data, does not fit the data well. This is because the mean and standard deviation are inflated by the two outliers.

The third departure from linearity occurs when more data are present in both tails (areas furthest from the median) than would be expected for a normal distribution. Figure **??** is a probability plot of adjusted nitrate concentrations in precipitation from Wellston, Michigan (**?**). These data are actually residuals (departures) from a regression of log of nitrate concentration versus log of precipitation volume. A residual of 0 indicates that the concentration is exactly what would be expected for that volume, a positive residual more than what is expected, and negative less than expected. The data in figure **??** display a predominantly linear pattern, yet one not fit well by the theoretical normal shown as the solid line. Again this lack of fit indicates outliers are present. The outliers are data to the left which plot below the linear pattern, and those above the pattern to the right of the figure. Outliers occur on both ends in greater numbers than expected from a normal distribution. A boxplot for the data is shown in figure **??** for comparison. Note that both the box and whiskers are symmetric, and therefore no power transformation such as those in the "ladder of powers" would produce a more nearly normal distribution. Data may depart from a normal distribution not only in skewness, but by the number of extreme values. Excessive numbers of extreme values may cause significance levels of tests requiring the normality assumption to be in error. Therefore procedures which assume normality for their validity when applied to data of this type may produce quite inaccurate results.

```
## Warning: Removed 6 rows containing missing values (geom_path).
```

2.1.5.3 Probability plots for comparing among distributions

In addition to comparisons to a normal distribution, quantiles may be computed and probability plots constructed for any two-parameter distribution. The distribution which causes data to be most like a straight line on its probability plot is the one which most closely resembles the distributional shape of the data. Data may be compared to a two-parameter lognormal distribution by simply plotting the logarithms of the data as the data quantiles, as was done in figure **??**. **?** demonstrated the construction of probability plots for the Gumbel (extreme-value) distribution, which is sometimes employed for flood-flow studies. **?** cover the use of probability plots for the two-parameter Weibull distribution, used in fitting low-flow data. Again, the best fit is obtained with the distribution which most closely produces a linear plot. In both references, the use of a test of significance called the probability plot correlation coefficient augmented the visual determination of linearity on the plot. This test will be covered in detail in Chapter 4.

Use of three-parameter distributions can also be indicated by probability plots.

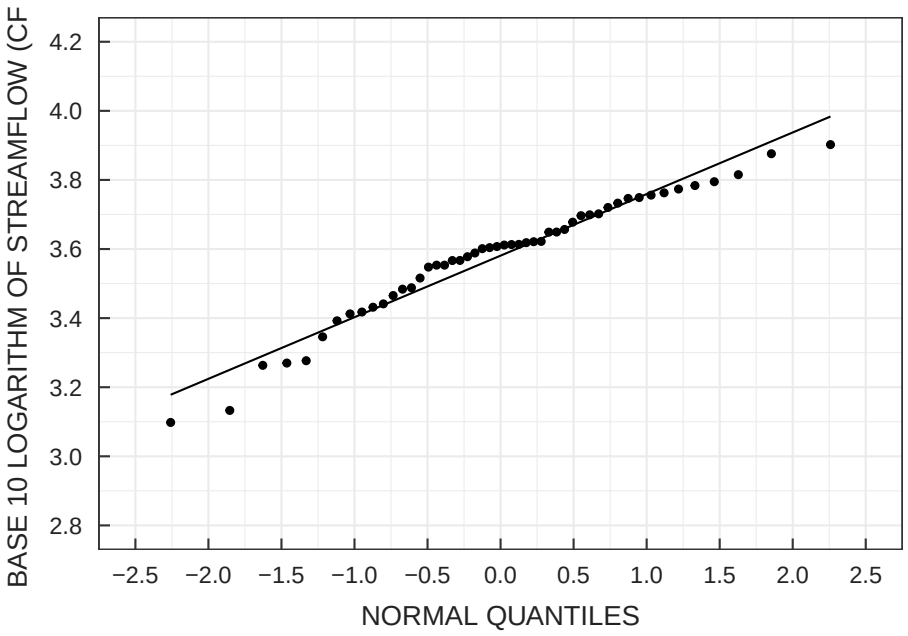


Figure 2.9: Probability plot of a left-skewed distribution (logs of Licking River data)

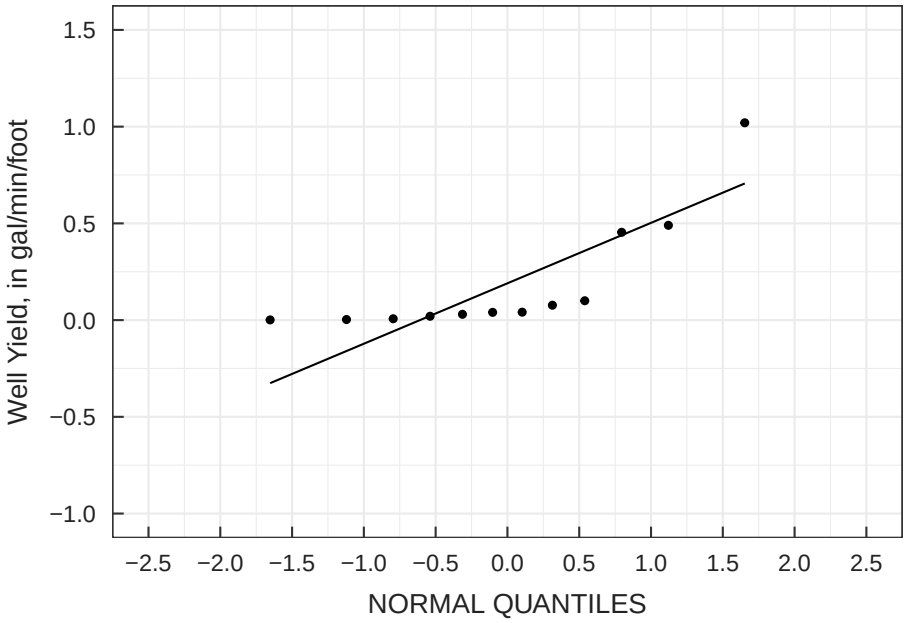


Figure 2.10: Probability plot of a right-skewed distribution (unit well yields)

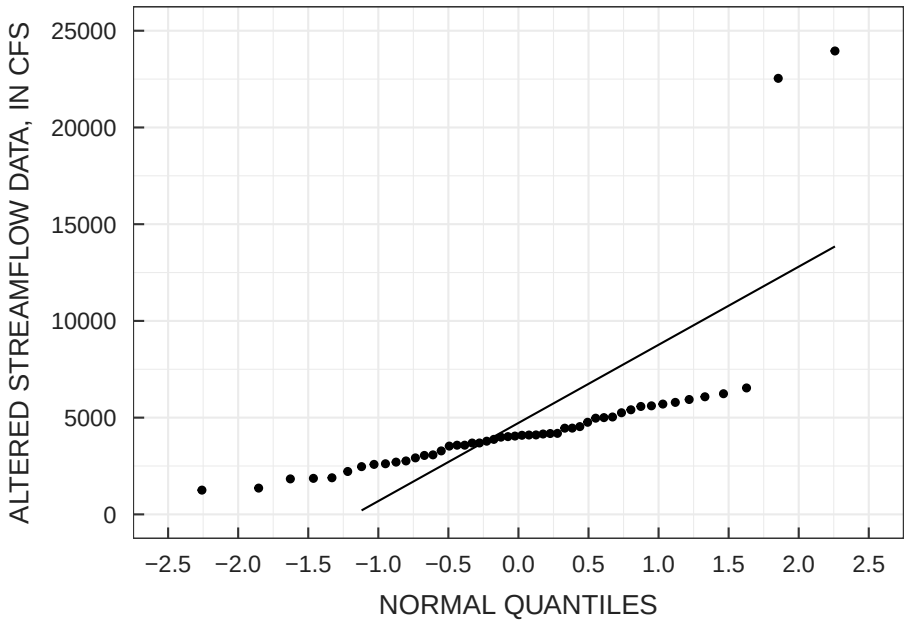


Figure 2.11: Probability plot of data with high outliers

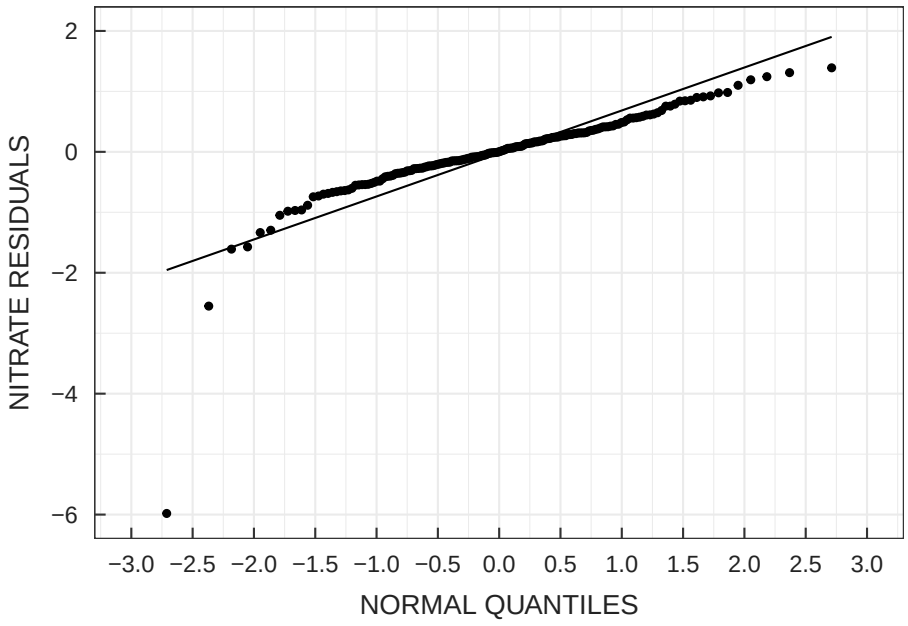


Figure 2.12: Probability plot of a heavy-tailed data set

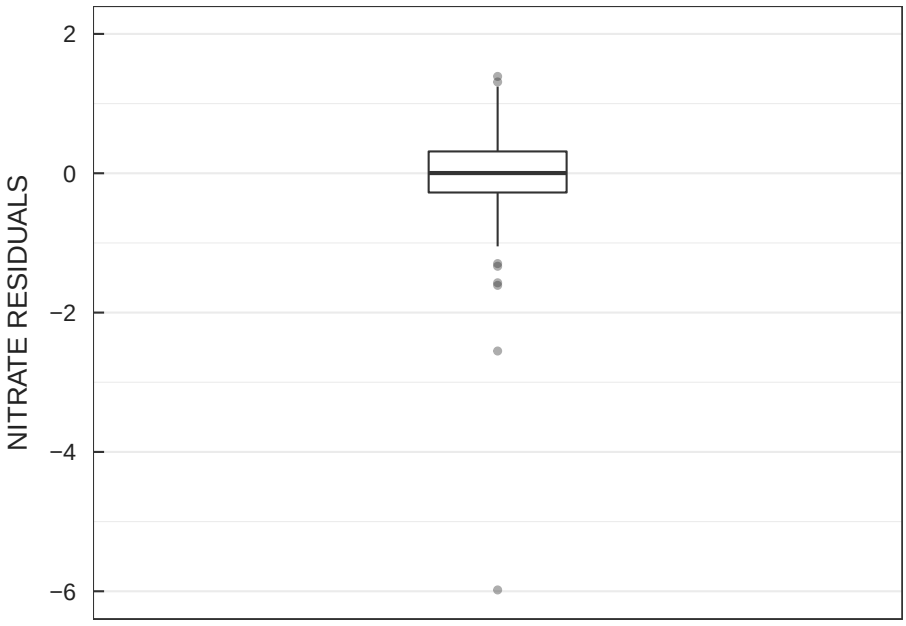


Figure 2.13: Boxplot of a heavy-tailed data set

For example, if significant right-skewness remains after logarithms are taken, the resulting concave shape on a lognormal probability plot indicates that a log-Pearson III distribution would better fit the data. **?** demonstrate the construction of a probability plot for the log-Pearson III distribution using a Wilson-Hilferty transformation.

2.2 Graphical Comparisons of Two or More Data Sets

Each of the graphical methods discussed thus far can be, and have been, used for comparing more than one group of data. However, each is not equally effective. As the following sections show, histograms are not capable of providing visual comparisons between data sets at the same level of detail as boxplots or probability plots. Boxplots excel in clarity and easy discrimination of important distributional characteristics, even for comparisons between many groups of data. A newer type of plot, the quantile-quantile (Q-Q) plot, provides additional information about the relationship between two data sets.

Each graphic will be developed for the same data set, a comparison of unit well yields in Virginia (**?**). These are small data sets: 13 wells are from valleys

underlain by fractured rocks, and 12 wells from valleys underlain by unfractured rocks.

2.2.1 Histograms

Figure **??** presents histograms for the two sets of well yield data. The right-skewness of each data set is easily seen, but it is difficult to discern whether any differences exist between them. Histograms do not provide a good visual picture of the centers of the distributions, and only a slightly better comparison of spreads. Positioning histograms side-by-side instead of one above the other provide even less ability to compare data, as the data axes would not be aligned. Unfortunately, this is commonly done. Also common are overlapping histograms, such as in figure **??**. Overlapping histograms provide poor visual discrimination between multiple data sets.

2.2.2 Dot and Line Plots of Means, Standard Deviations

Figure **??** is a "dot and line" plot often used to represent the mean and standard deviation (or standard error) of data sets. Each dot is the mean of the data set. The bars extend to plus and minus either one standard deviation (shown), or plus and minus one or more standard errors ($s.e. = s/\sqrt{n}$), beyond the mean. This plot displays differences in mean yields, but little else. No information on the symmetry of the data or presence of outliers is available. Because of this, there is not much information given on the spread of the data, as the standard deviation may describe the spread of most of the data, or may be strongly influenced by skewness and a few outliers.

To emphasize the deficiencies of dot and line plots such as these, figure **??** presents three data sets with very different characteristics. The first is a uniform distribution of values between 0 and 20. It is symmetric. The second is a right-skewed data set with outliers. The third is a bimodal distribution, also symmetric. All three have a mean of 10 and standard deviation of 6.63. Therefore each of the three would be represented by the same dot and line plot, shown at the right of the figure.

Dot and line plots are useful only when the data are actually symmetric. If skewness or outliers are present, as with data set 2, neither the plots (or a table of means and standard deviations) indicate their presence. Even for symmetric distributions, differences such as those between data sets 1 and 3 will not be evident. Far better graphical methods are available.

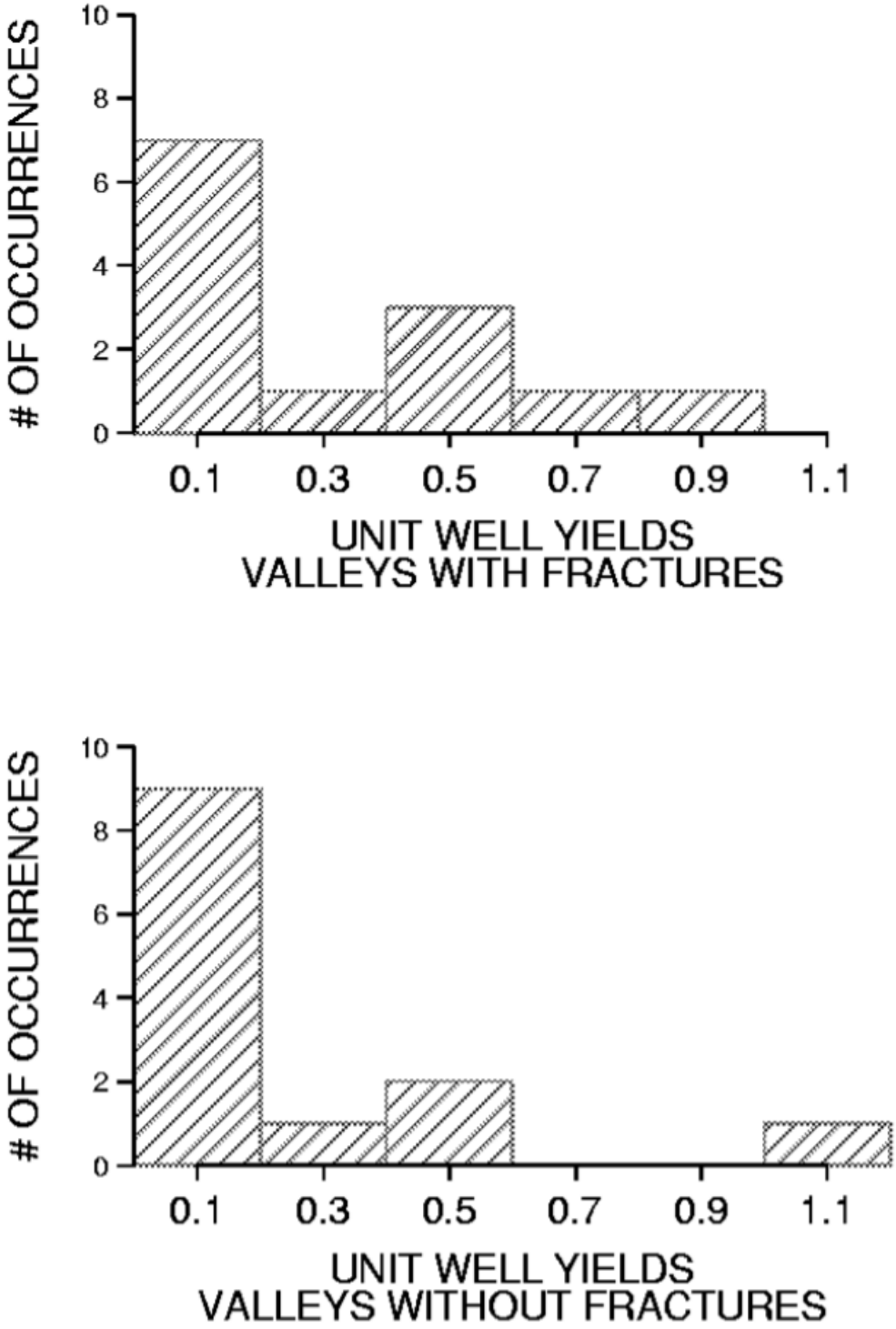


Figure 2.14: Histogram of the unit well yield data

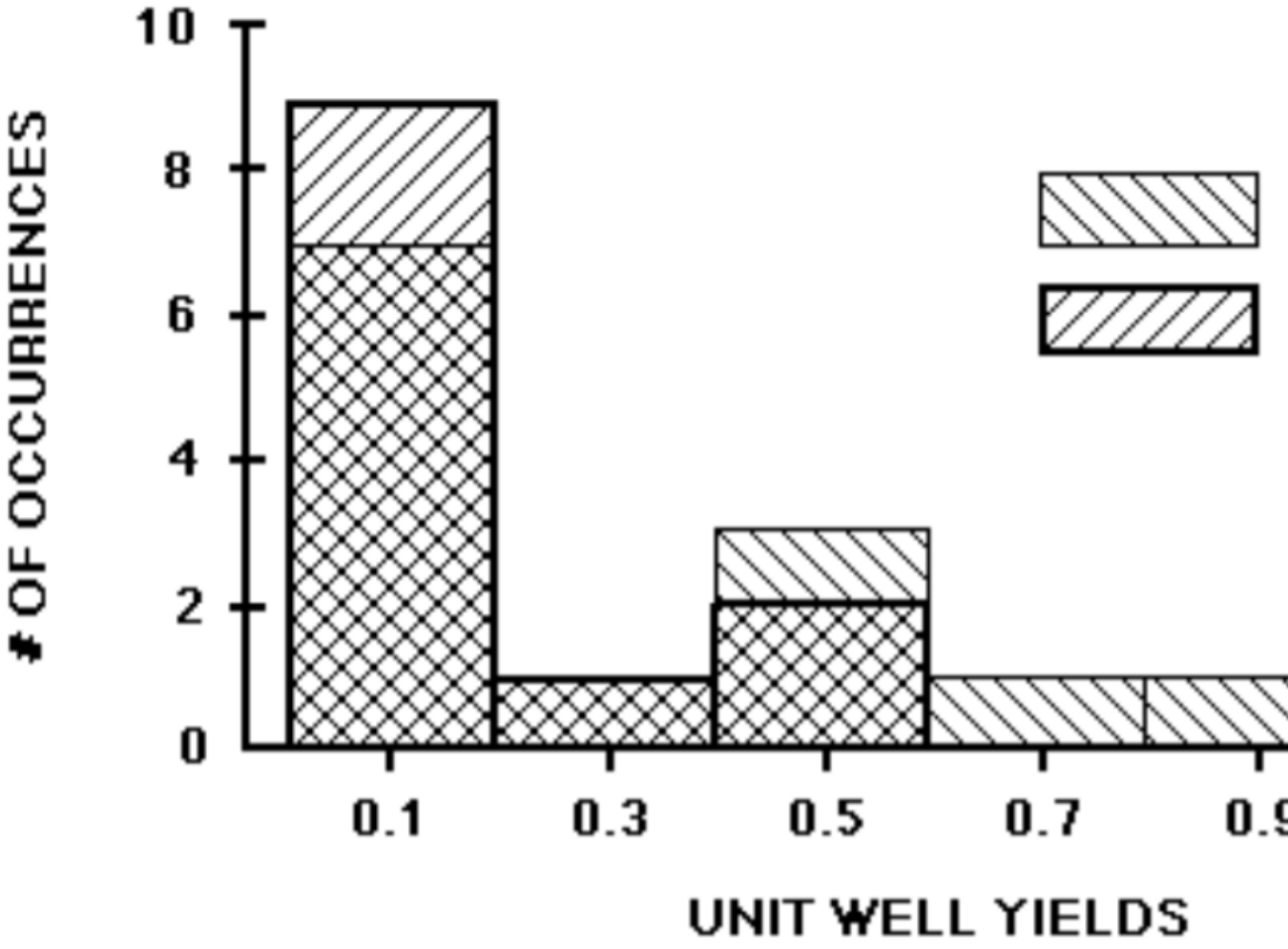


Figure 2.15: Overlapping histograms of the unit well yield data

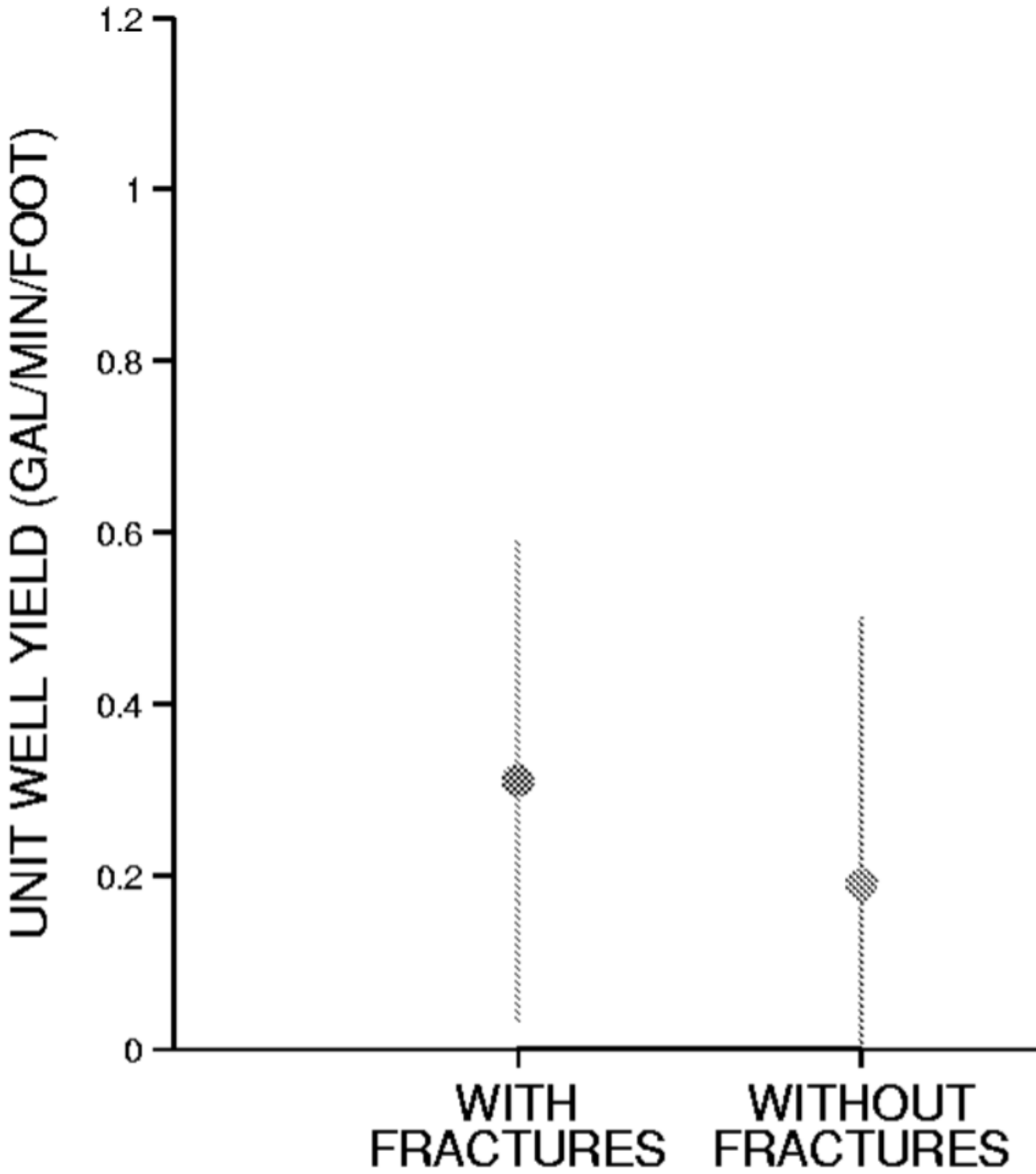


Figure 2.16: Dot and line plot for the unit well yield data

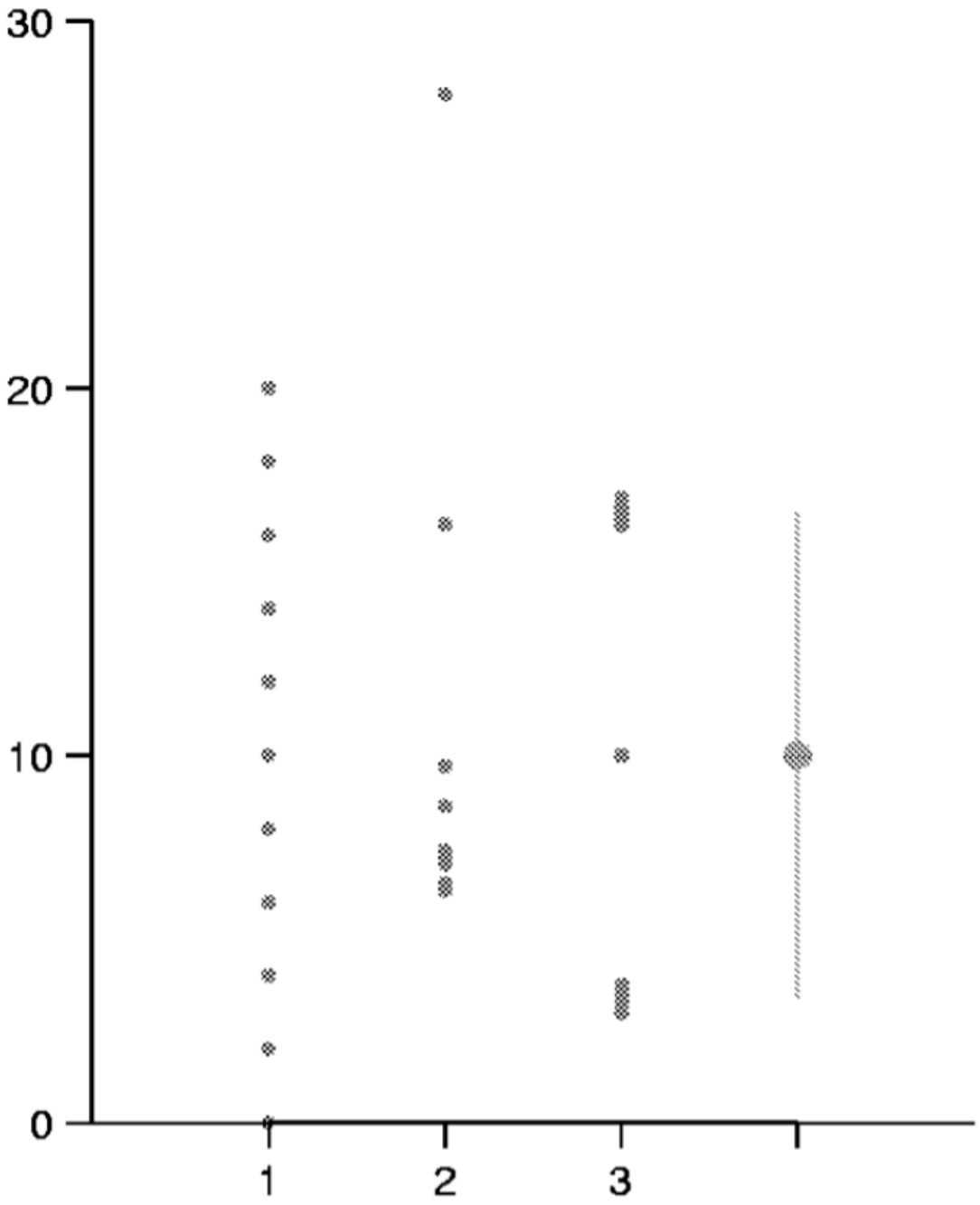


Figure 2.17: Number lines of 3 dissimilar groups of data, all having an identical dot and line plot (shown at right)

2.2.3 Boxplots

Figure **??** presents boxplots of the well yield data. The median well yield is seen to be higher for the areas with fractures. The IQR of wells with fractures is slightly larger than that for wells without, and the highest value for each group is similar. Both data sets are seen to be rightskewed. Thus a large amount of information is contained in this very concise illustration. The mean yield, particularly for wells without fractures, is undoubtedly inflated due to skewness, and differences between the two groups of data will in general be larger than indicated by the differences in their mean values.

In figure **??**, boxplots of the three data sets given in figure **??** are presented. The skewness of data set 2 is clear, as is the symmetry of 1 and 3. The difference in shape between 1 and 3 is evident. The minute whiskers of data set 3 illustrate that over 25 percent of the data are located essentially at the upper and lower quartiles – a bimodal distribution.

The characteristics which make boxplots useful for inspecting a single data set make them even more useful for comparing multiple data sets. They are valuable guides in determining whether central values, spread, and symmetry differ among groups of data. They will be used in later chapters to guide whether tests based on assumptions of normality may be employed. The essential characteristics of numerous groups of data may be displayed in a small space. For example, the 20 boxplots of figure **??** were used by **?** to illustrate the source of ammonia nitrogen on a section of the Detroit River. The Windmill Point Transect is upstream of the U.S. city of Detroit, while the Fermi Transect is below the city. Note the marked changes in concentration (the median lines of the boxplots) and variability (the widths of the boxes) on the Michigan side of the river downstream of Detroit. A lot of information on streamwater quality is succinctly summarized in this relatively small figure.
\begin{figure}

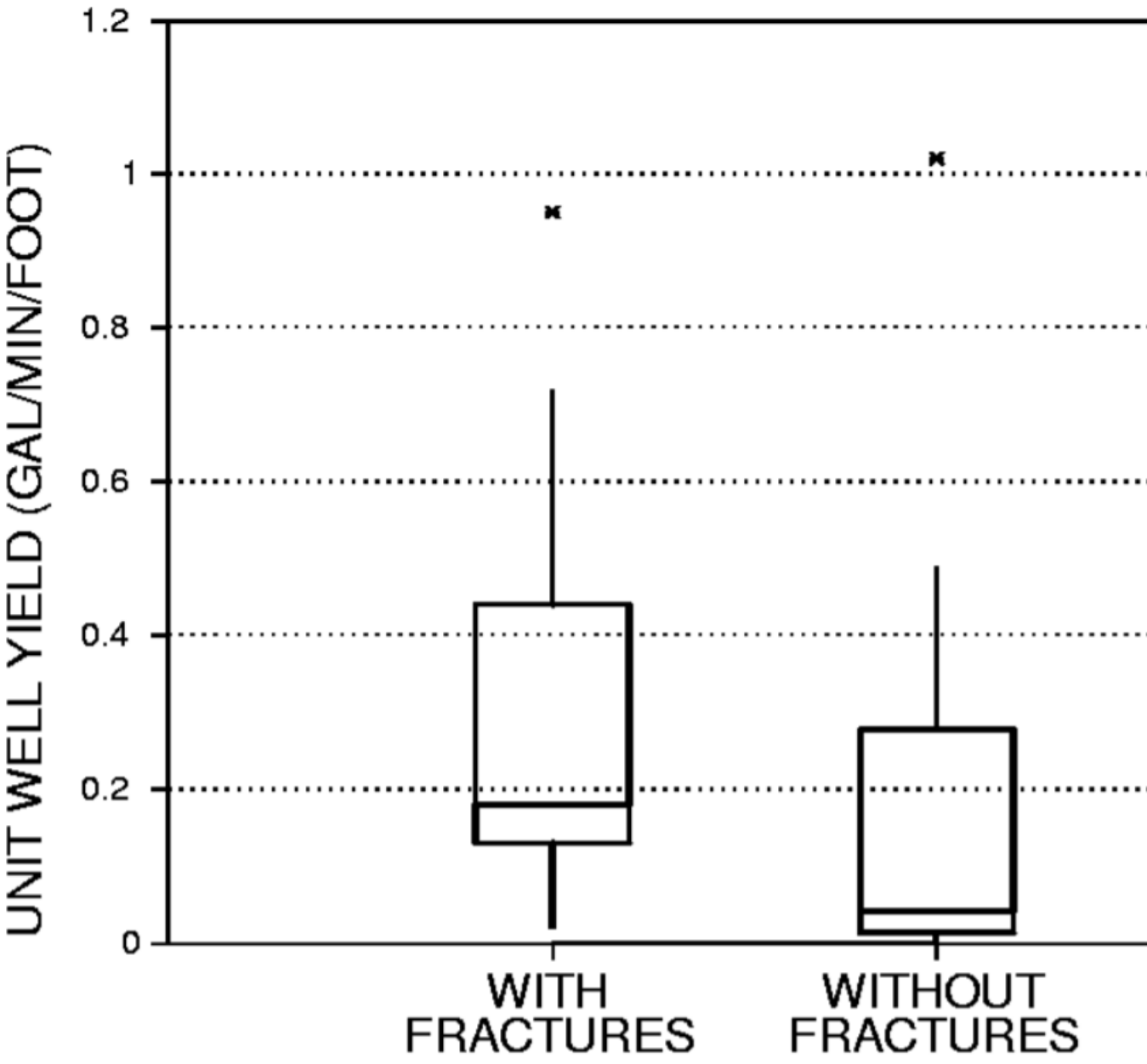


Figure 2.18: Boxplots of the unit well yield data

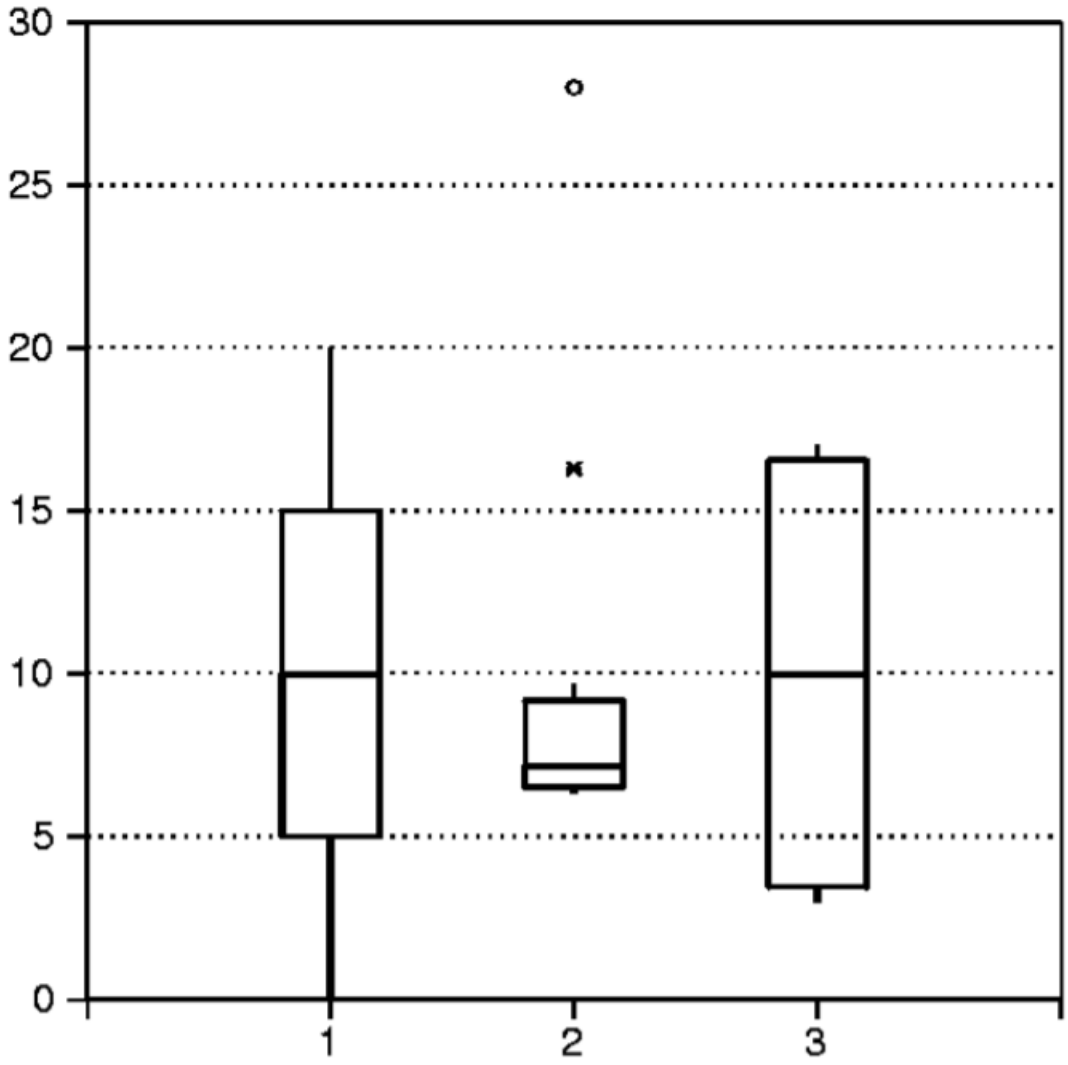
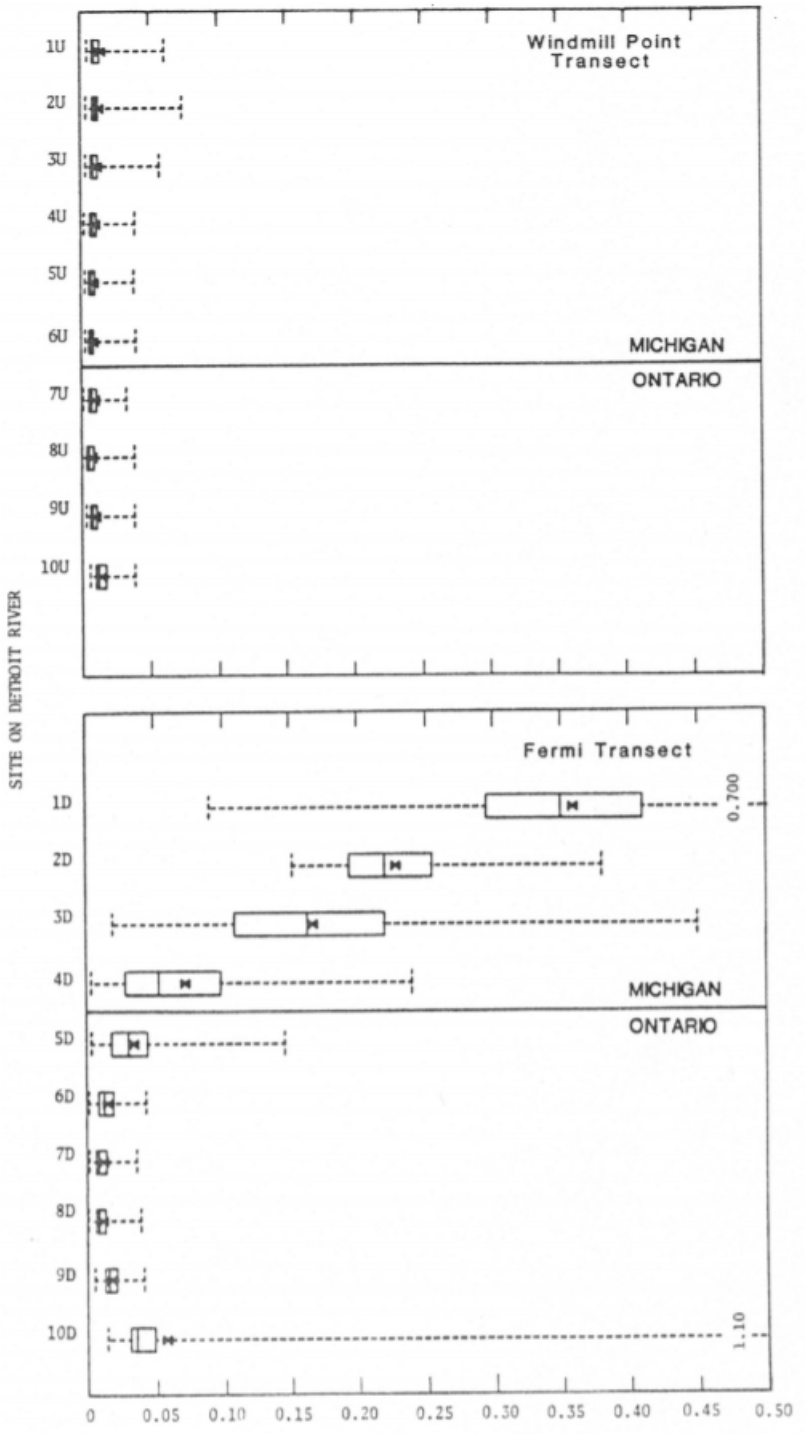


Figure 2.19: Boxplots of the 3 dissimilar groups of data shown in figure **??**


Windmill Point
Transect
1U
2U
3U
4U
5U
6U
MICHIGAN
ONTARIO
7U
8U
9U
10U
SITE ON DETROIT RIVER
Fermi Transect
1D
2D
3D
4D
MICHIGAN
ONTARIO
5D
6D
7D
8D
9D
10D
0.700
1.10
0
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50

{

}

\caption{Boxplots of total ammonia nitrogen concentrations (mg/L as N) at two transects on the Detroit River (from **?**)} \end{figure} ### Probability Plots

Probability plots are also useful graphics for comparing groups of data. Characteristics evident in boxplots are also seen using probability plots, though in a different format. Comparisons of each quantile, not just the boxplot quartiles, can be made. The straightness of each data set also allows quick comparisons to conformity with the theoretical distribution.

Figure **??** is a probability plot of the two well yield data sets. The right-skewness of each data set is shown by their concave shapes. Wells without fractures have greater skewness as shown by their greater concavity on the plot. Quantiles of the wells with fractures are higher than those without, indicating generally higher yields. Figure **??** shows that the lowest yields in each group are similar, as both data sets approach zero yield. Also seen are the similarity in the highest yield for each group, due to the outlier for the without fractures group. Comparisons between median values are simple to do – just travel up the normal quantile = 0 line. Comparisons of spreads are more difficult – the slopes of each data set display their spread.

In general, boxplots summarize the differences between data groups in a manner more quickly discerned by the viewer. When comparisons to a particular theoretical distribution such as the normal are important, or comparisons between quantiles other than the quartiles are necessary, probability plots are useful graphics. Either have many advantages over histograms or dot and line plots.

2.2.4 Q-Q Plots

Direct comparisons can be made between two data sets by graphing the quantiles (percentiles) of one versus the quantiles (percentiles) of the second. This is called a quantile-quantile or Q-Q plot (**?**). If the two data sets came from the same distribution, the quantile pairs would plot along a straight line with $Yp = Xp$, where p is the plotting position and Yp is the pth quantile of Y. In this case it would be said that the median, the quartiles, the 10th and 90th percentiles, etc., of the two data sets were equal. If one data set had the same shape as the second, differing only by an additive amount (each quantile was 5 units higher than for the other data set, for example), the quantile pairs would fall along a line parallel to but offset from the $Yp = Xp$ line, also with slope = 1. If the data sets differed by a multiplicative constant ($Yp = 5 \bullet Xp$, for example), the quantile pairs would lie along a straight line with slope equal to the multiplicative constant. More complex relationships will result in pairs of quantiles which do not lie along a straight line. The question of whether or

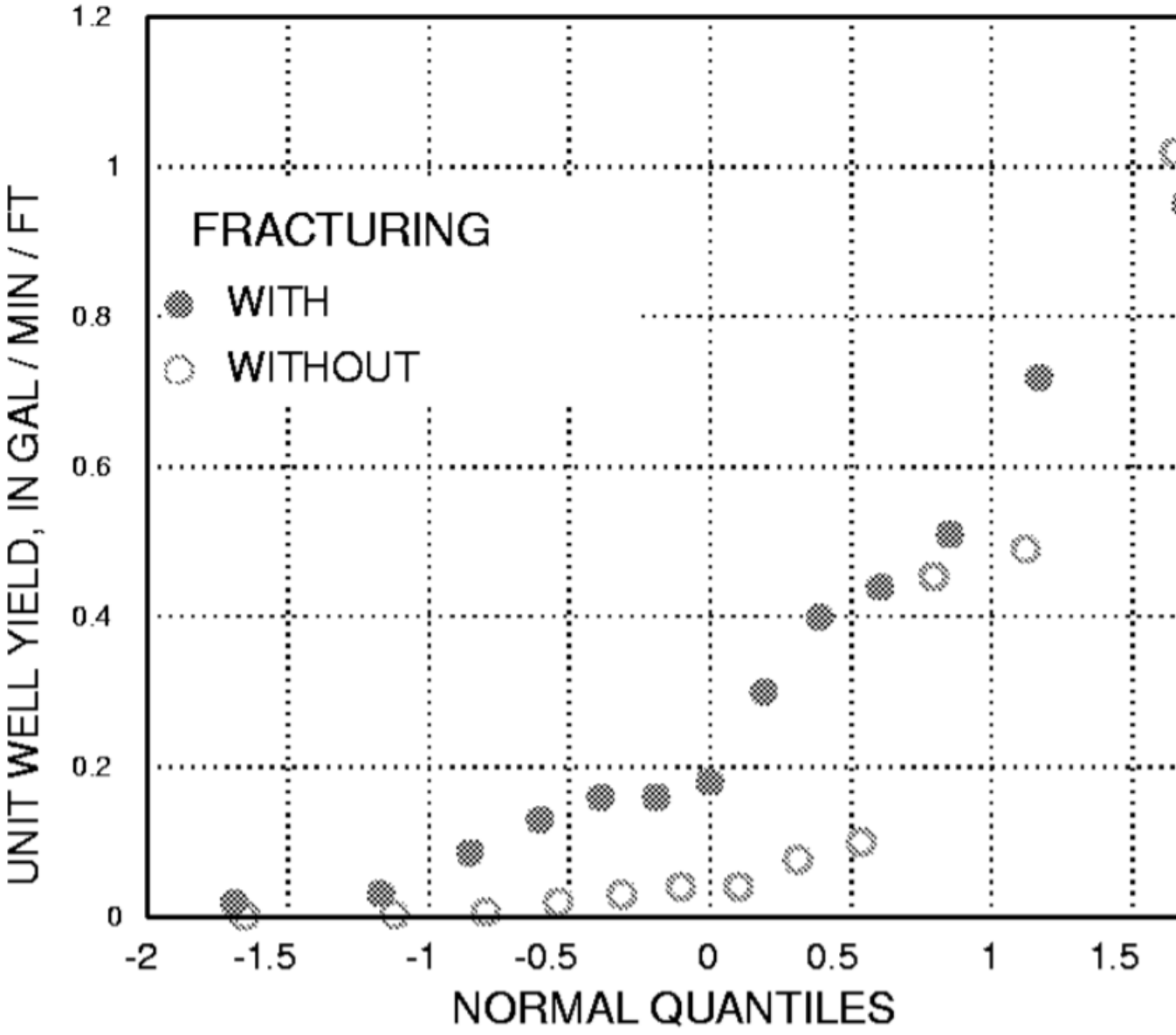


Figure 2.20: Probability plot of the unit well yield data

not data sets differ by additive or multiplicative relationships will become important when hypothesis testing is conducted.

Figure **??** is a Q-Q plot of the well yield data. Several aspects of the relationship between the two data sets are immediately seen. First, the lowest 9 quantile pairs appear to fall along a straight line with a slope greater than 1, not parallel to the $Yp = Xp$ line shown as a reference. This indicates a multiplicative relation between the data, with $Y \cong 4.4 \bullet X$, where 4.4 is the slope of those data on the plot. Therefore, the yields with fractures are generally 4.4 times those without fractures for the lowest 75 percent of the data. The 3 highest quantile pairs return near to the $Y = X$ line, indicating that the higher yields in the two data sets approach being equal. The hydrologist might be able to explain this phenomenon, such as higher yielding wells are deeper and less dependent on fracturing, or that some of the wells were misclassified, etc. Therefore the Q-Q plot becomes a valuable tool in understanding the relationships between data sets prior to performing any hypothesis tests.

2.2.4.1 Construction of Q-Q plots

Q-Q plots are similar to probability plots. Now instead of plotting data quantiles from one group against quantiles of a theoretical distribution such as the normal, they are plotted against quantiles of a second data group.

When sample sizes of the two groups are identical, the x's and y's can be ranked separately, and the Q-Q plot is simply a scatterplot of the ordered data pairs $(x_1, y_1) \cdots (x_n, y_n)$. When sample sizes are not equal, consider n to be the sample size of the smaller data set and m to be the sample size of the larger data set. The data values from the smaller data set are its pth quantiles, where $p = (i - 0.4)/(n + 0.2)$. The n corresponding quantiles for the larger data set are interpolated values which divide the larger data set into n equally-spaced parts. The following example illustrates the procedure.

For the well yield data, the 12 values without fractures designated $x_i, i = 1, \cdots n$ are themselves the sample quantiles for the smaller data set. Repeating the without fractures data given earlier in the chapter:

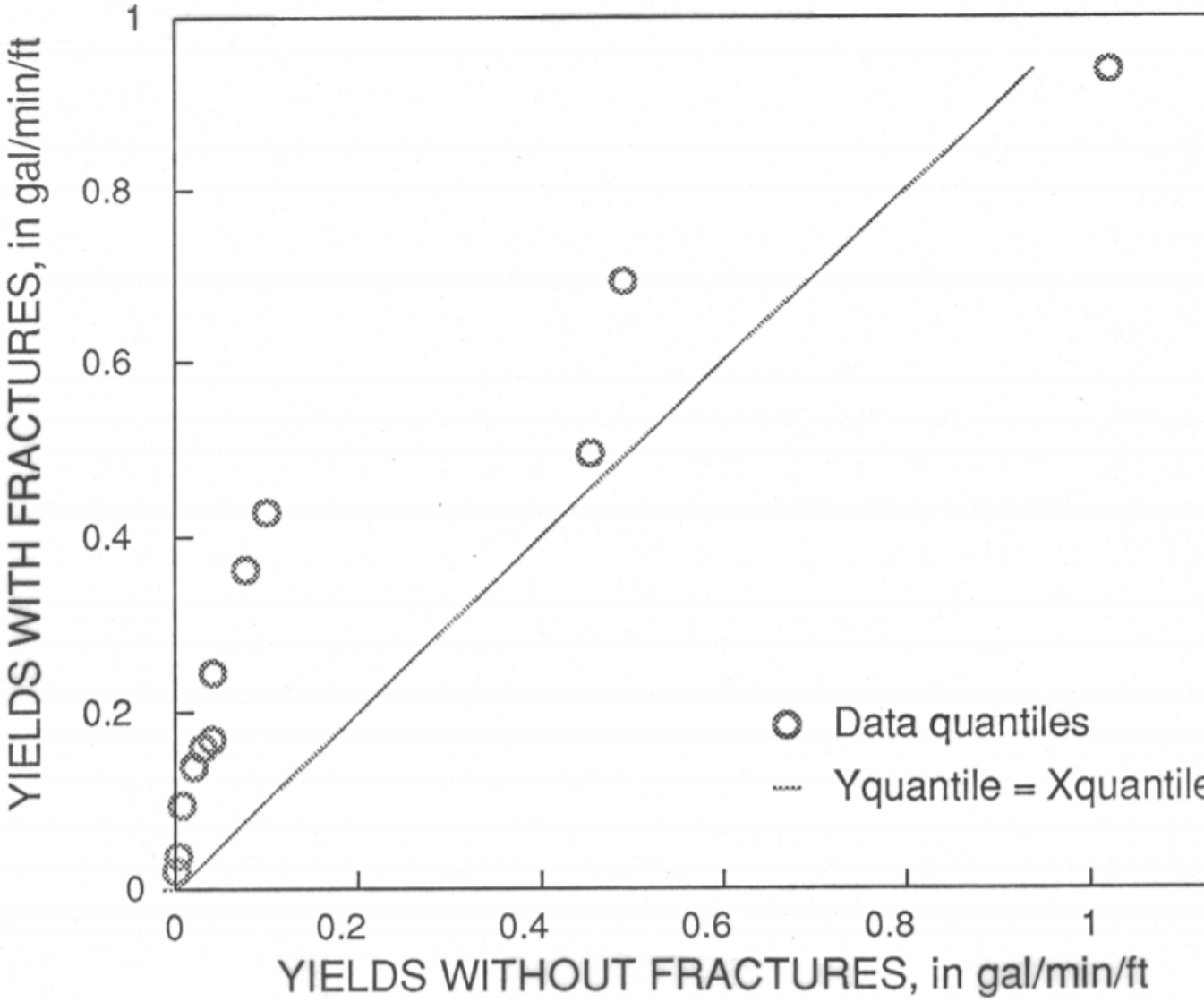


Figure 2.21: Q-Q plot of the unit well yield data

Unit well yields, in gal / min / ft (Wrig

x_i = yield without fractures p_i = plotting position

i	x_i	p_i	y_j	i	x_i	p_i	y_j
1	0.001	.05	-	5	0.030	.38	-
2	0.003	.13	-	6	0.040	.46	-
3	0.007	.21	-	7	0.041	.54	-
4	0.020	.30	-	8	0.077	.62	-

The .05 quantile (5th percentile) value of 0.001, for example, is to be paired on the Q-Q plot with the .05 quantile of the yields with fractures. To compute the corresponding y quantiles for the second data set, $p = (j - 0.4)/(m + 0.2)$, and therefore j must be:

$$\frac{(j-0.4)}{(m+0.2)} = \frac{(i-0.4)}{(n+0.2)} \quad \text{, or}$$
$$j = \frac{(m+0.2) \bullet (i-0.4)}{(n+0.2)} + 0.4 \tag{2.1}$$

If j is an integer, the data value y_j itself is plotted versus x_i. Usually, however, j will lie between two integers, and the y quantile must be linearly interpolated between the y data corresponding to the ranks on either side of j:

$$y_j = y_{j'} + (j - j') \bullet \left(y_{(j'+1)} - y_{j'}\right) \quad \text{where } j' = \text{integer(j)} \tag{2.2}$$

For example, the well yield data with fractures are the following:

0.020 0.031 0.086 0.130 0.160 0.160 0.180 0.300 0.400 0.440 0.510 0.720 0.950

Therefore $n = 12$ $m = 13$ and from eq. (**??**), $j = 1.08i - 0.03$.
The first of the 12 quantiles to be computed for the data with fractures is then:

$$i = 1 \quad j = 1.05 \quad j' = 1 \qquad y_j = y_1 + 0.05 \bullet (y_2 - y_1)$$
$$y_j = 0.020 + 0.05 \bullet (0.031 + 0.020) \tag{2.3}$$
$$y_j = 0.021$$

All 12 quantiles are similarly interpolated:

i	j	interpolated y_j	i	j
1	1.05	0.021	7	7.53
2	2.13	0.038	8	8.61
3	3.21	0.095	9	9.69
4	4.29	0.139	10	10.77
5	5.37	0.160	11	11.85
6	6.45	0.169	12	12.93

These interpolated values are added to the table of quantiles given previously:

x_i = yields without fractures		p_i = plotting position					y_j =
i	x_i	p_i	y_j	i	x_i	p_i	y_j	i
1	0.001	.05	0.021	5	.030	.38	0.160	9
2	0.003	.13	0.038	6	.040	.46	0.169	10
3	0.007	.21	0.095	7	.041	.54	0.245	11
4	0.020	.30	0.139	8	.077	.62	0.362	12

These (x_i, y_j) pairs are the circles which were plotted in figure **??**.

2.3 Scatterplots and Enhancements

The two-dimensional scatterplot is one of the most familiar graphical methods for data analysis. It illustrates the relationship between two variables. Of usual interest is whether that relationship appears to be linear or curved, whether different groups of data lie in separate regions of the scatterplot, and whether the variability or spread is constant over the range of data. In each case, an enhancement called a "smooth" enables the viewer to resolve these issues with greater clarity than would be possible using the scatterplot alone. The following sections discuss these three uses of the scatterplot, and the

enhancements available for each use.

2.3.1 Evaluating Linearity

Figure **??** is a scatterplot of the mass load of transported sand versus stream discharge for the Colorado River at Lees Ferry, Colorado, during 1949-1964. Are these data sufficiently linear to fit a linear regression to them, or should some other term or transformation be included in order to account for curvature? In Chapters 9 and 11, other ways to answer this question will be presented, but many judgements on linearity are made solely on the basis of plots. To aid in this judgement, a "smooth" will be superimposed on the data.

The human eye is an excellent judge of the range of data on a scatterplot, but has a difficult time accurately judging the center – the pattern of how y varies with x. This results in two difficulties with judging linearity on a scatterplot as evident in figure **??**. Outliers such as the two lowest sand concentrations may fool the observer into believing a linear model may not fit. Alternatively, true changes in slope are often difficult to discern from only a scatter of data. To aid in seeing central patterns without being strongly influenced by outliers, a resistant center line can be fit to the data whose direction and slope varies locally in response to the data themselves. Many methods are available for constructing this type of center line – probably the most familiar is the (non-resistant) moving average. All such methods may be called a "middle smooth", as they smooth out variations in the data into a coherent pattern through the middle. We discuss computation of smooths in Chapter 10. For now, we will merely illustrate their use as aids to graphical data analysis. The smoothing procedure we prefer is called LOWESS, or LOcally WEighted Scatterplot Smoothing (**?**; **?**).

Figure **??** presents the Lees Ferry sediment data of figure **??**, with a superimposed middle smooth. Note the nonlinearity now evident by the curving smooth on the left-hand side of the plot. The rate of sand transport slows above 6600 ($e^{8.8}$) cfs. This curvature is easier to see with the superimposed smooth. It is important to remember that no model, such as a linear or quadratic function, is assumed prior to computing a smooth. The smoothed pattern is totally derived by the pattern of the data, and may take on any shape. As such, smooths are an exploratory tool for discerning the form of relationship between y and x. Seeing the pattern of figure **??**, a quadratic term might be added, a piecewise linear fit used, or a transformation stronger than logs used prior to performing a linear regression of concentration versus discharge (see Chapter 9).

Middle smooths should be regularly used when analyzing data on scatterplots, and when presenting those data to others. As no model form is assumed by them, they let the data describe the pattern of dependence of y on x. Smooths are especially useful when large amounts of data are to be plotted, and several

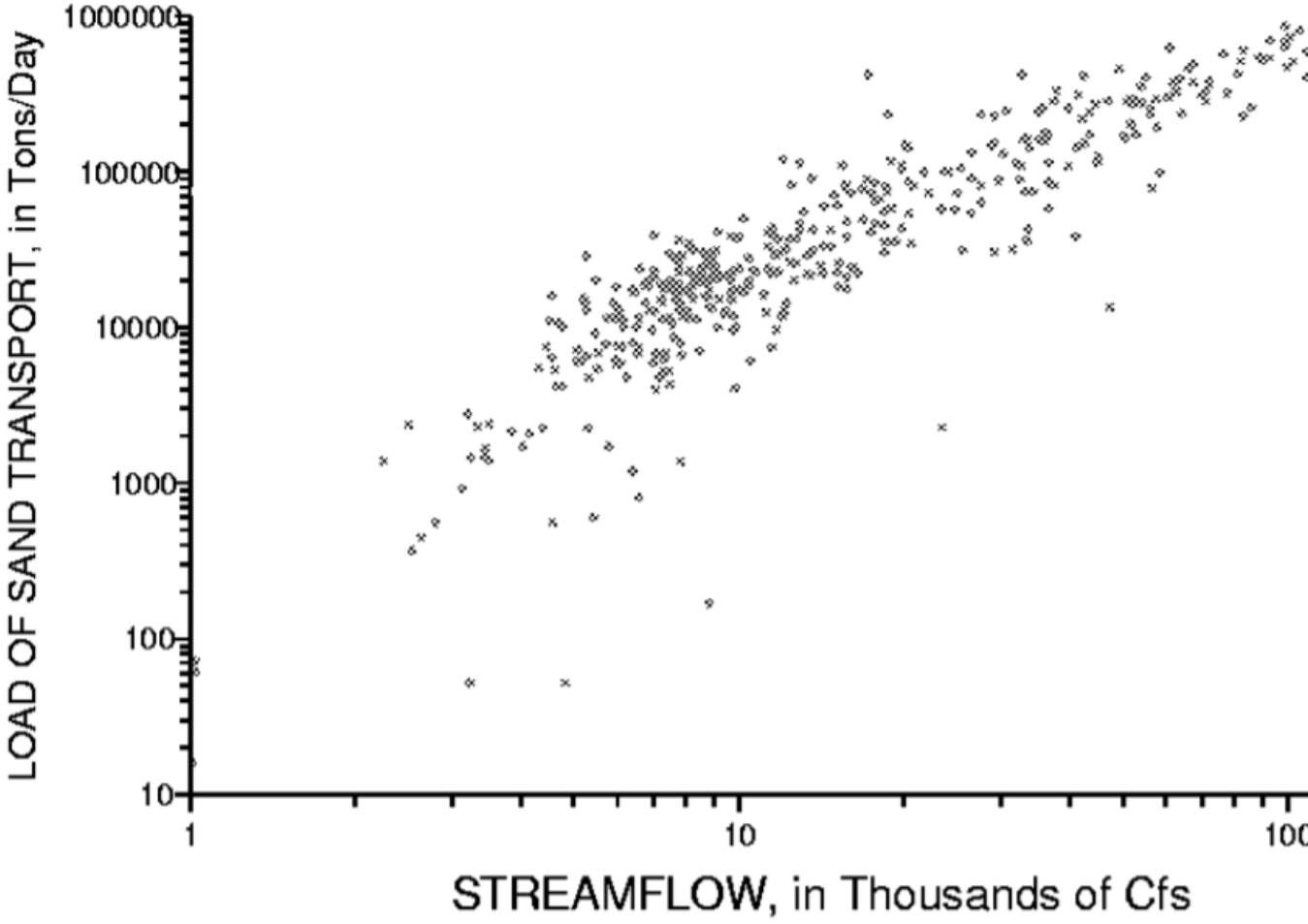


Figure 2.22: Suspended sand transport at Lees Ferry, Arizona, 1949-1952

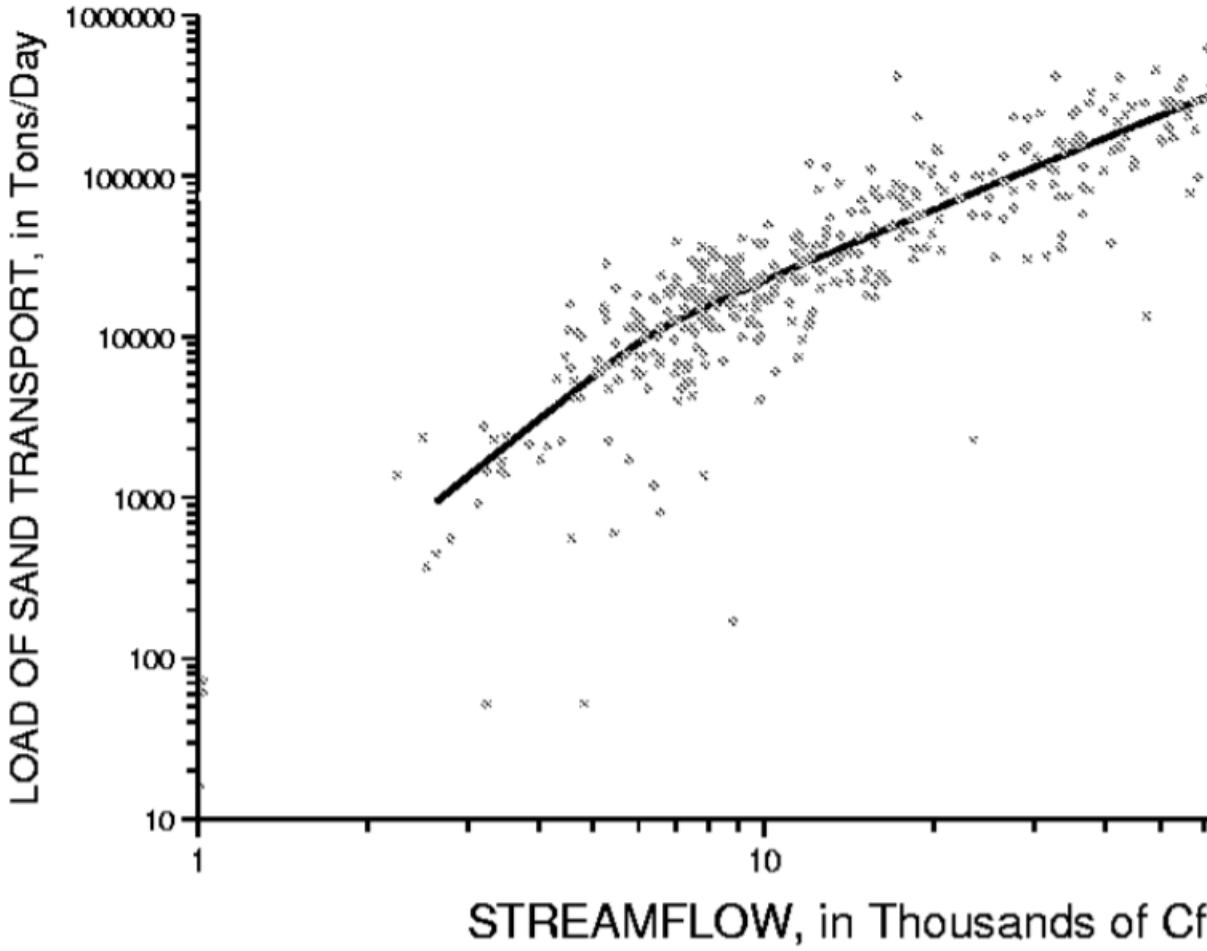
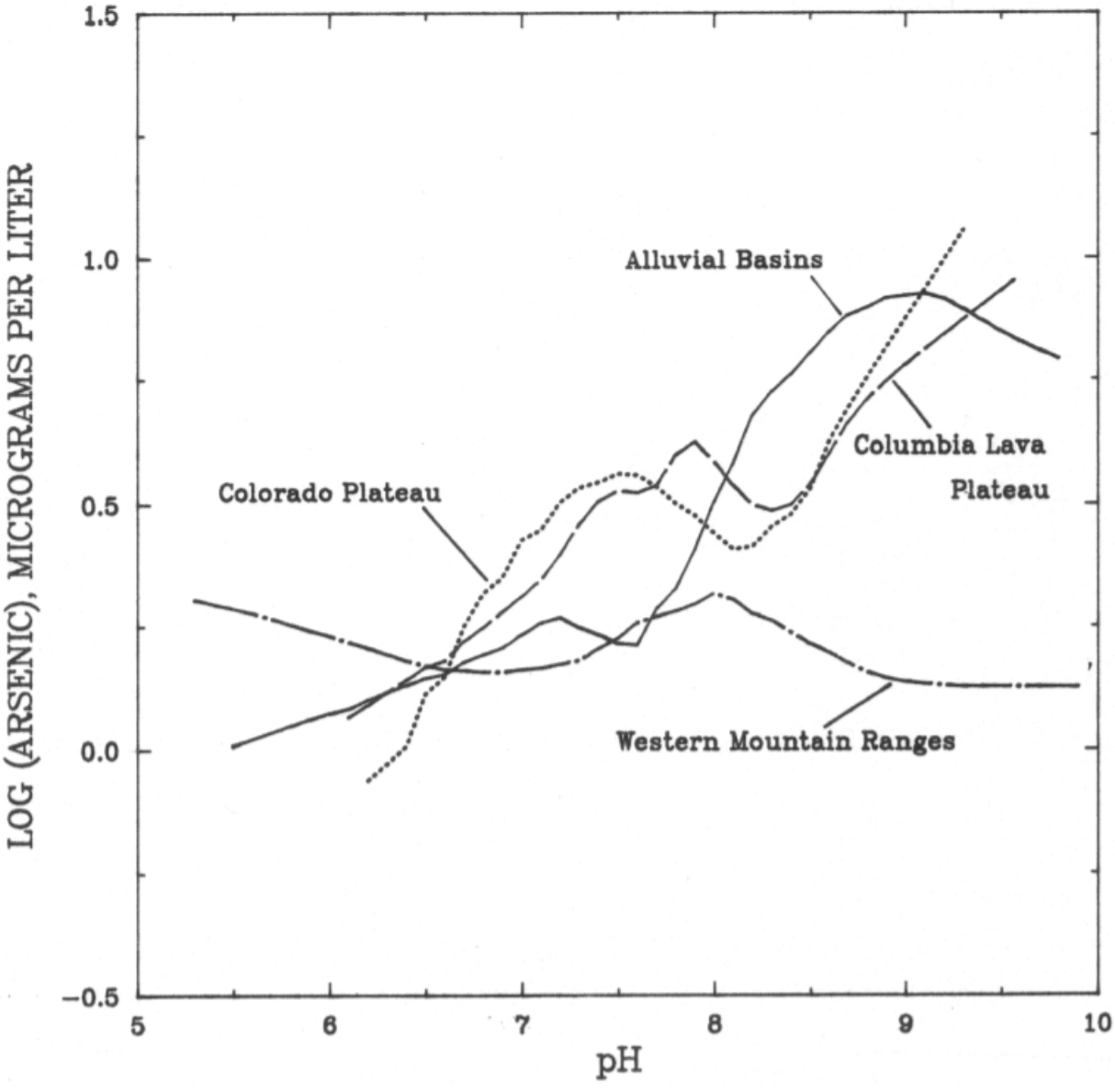


Figure 2.23: Data of figure **??** with superimposed lowess smooth

groups of data are placed on the same plot. For example, **?** depicted the dependence of log of arsenic concentration on pH for thousands of groundwater samples throughout the western United States (figure **??**). By using middle smooths, data from one physiographic province was seen to differ from the other three provinces in its relationship between pH and arsenic.
\begin{figure}



{

}

\caption{Dependance of log(As) on pH for 4 areas in the western U.S. (**?**)}
\end{figure} ### Evaluating Differences in Location on a Scatterplot

Figure **??** is a scatterplot of conductance versus pH for samples collected at low-flow in small streams within the coal mining region of Ohio (data from **?**). Each stream was classified by the type of land it was draining – unmined land, lands mined and later reclaimed, and lands mined and then abandoned without reclamation. These three types of upstream lands are plotted with different symbols in figure **??**.

To see the three locations more clearly, a smooth can be constructed for each group which encloses either 50 or 75 percent of the data. This type of smooth is called a polar smooth (**?**), and its computation is detailed in Chapter 10. Briefly, the data are transformed into polar coordinates, a middle or similar smooth computed, and the smooth is re-transformed back into the original units. In figure **??**. a polar smooth enclosing 75 percent of the data in each of the types of upstream land is plotted. These smooths are again not limited to a prior shape or form, such as that of an ellipse. Their shapes are determined from the data.

Polar smooths can be a great aid in exploratory data analysis. For example, the irregular pattern for the polar smooth of data from abandoned lands in figure **??** suggests that two separate subgroups are present, one with higher pH than the other. Using different symbols for data from each of the two geologic units underlying these streams shows indeed that the basins underlain by a limestone unit have generally higher pH than those underlain by a sandstone. Therefore the type of geologic unit should be included in any analysis or model of the behavior of chemical constituents for these data.

Polar smooths are especially helpful when there is a large amount of data to be plotted on a scatterplot. In such situations, the use of different symbols for distinguishing between groups will be ineffective, as the plot will be too crowded to see patterns in the locations of symbols. Indeed, in some locations it will not be possible to distinguish which symbol is plotted. Plots presenting small data points and the polar smooths as in figure **??**, or even just the polar smooths themselves, will provide far greater visual differentiation between groups.

2.3.2 Evaluating Differences in Spread

In addition to understanding where the middle of data lie on a scatterplot, it is often of interest to know something about the spread of the data as well. Homoscedasticity (constant variance) is a crucial assumption of ordinary least-squares regression, as we will see later. Changes in variance also invalidate parametric hypothesis test procedures such as analysis of variance. From a more exploratory point of view, changes in variance may be as important or more important than changes in central value. Differences

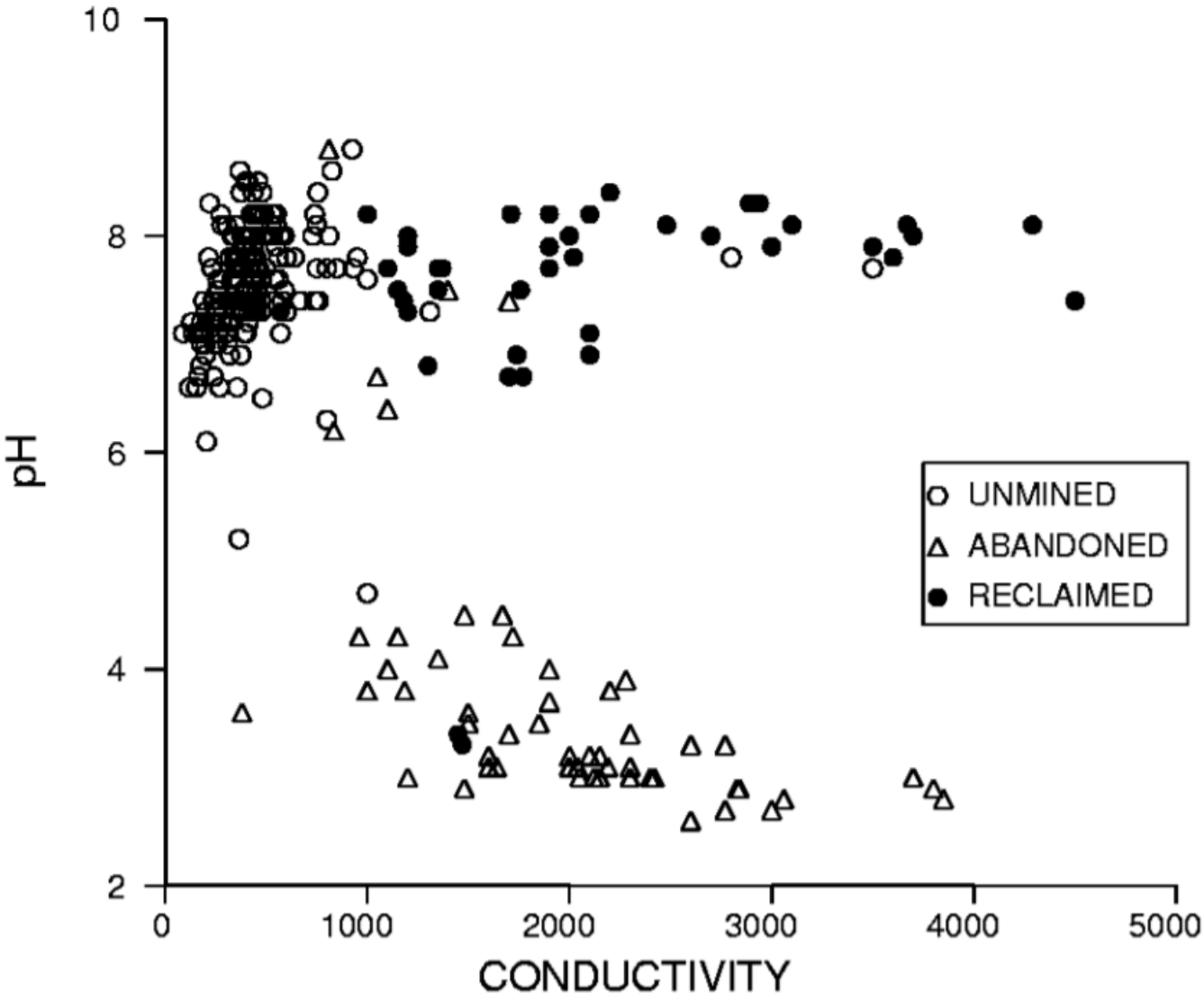


Figure 2.24: Scatterplot of water-quality draining three types of upstream land use

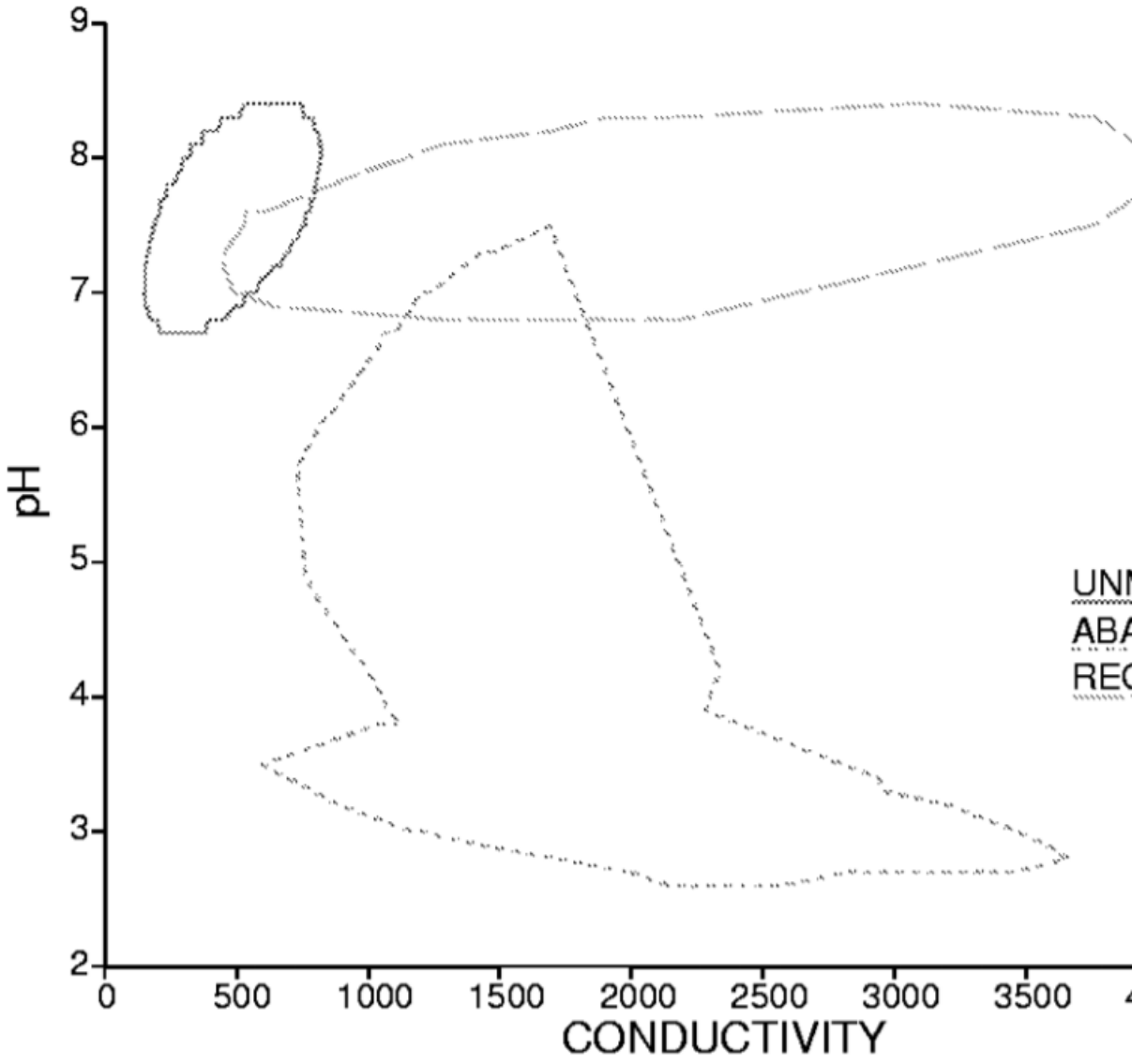


Figure 2.25: Polar smooths for the three groups of data in figure **??**

between estimation methods for flood quantiles, or between methods of laboratory analysis of some chemical constituent, are often differences in repeatability of the results and not of method bias. Graphs again can aid in judging differences in data variability, and are often used for this purpose.

A major problem with judgements of changing spread on a scatterplot is again that the eye is sensitive to seeing the range of data. The presence of a few unusual values may therefore incorrectly trigger a perception of changing spread. This is especially a problem when the density of data changes across a scatterplot, a common occurrence. Assuming the distribution of data to be identical across a scatterplot, and that no changes in variablility or spread actually occur, areas where data are more dense are more likely to contain outlying values on the plot, and the range of values is likely to be larger. This leads to a perception that the spread has changed.

One graphical means of determining changes in spread has been given by **?**. First, a middle smooth is computed, as in figure **??**. The absolute values of differences $|d_i|$ between each data point and the smooth at its value of x is a measure of spread.

$$|d_i| = |y_i - l_i| \quad \text{where } l_i \text{ is the value for the lowess smooth at } x_i \tag{2.4}$$

By graphing these absolute differences $|d_i|$ versus x_i, changes in spread will show as changes in absolute differences. A middle smooth of these differences should also be added to make the pattern more clear. This is done in figure **??**, a plot of the absolute differences between sand concentration and its lowess smooth for the Lees Ferry data of figure **??**. Note that there is a slight decrease in $|d_i|$, indicating a small decrease of variability or spread in concentration with increasing discharge at that site.

2.4 Graphs for Multivariate Data

Boxplots effectively illustrate the characteristics of data for a single variable, and accentuate outliers for further inspection. Scatterplots effectively illustrate the relationships between two variables, and accentuate points which appear unusual in their x-y relationship. Yet there are numerous situations where relationships between more than two variables should be considered simultaneously. Similarities and differences between groups of observations based on 3 or more variables are frequently of interest. Also of interest is the detection of outliers for data with multiple variables. Graphical methods again can provide insight into these relationships. They supplement and enhance the understanding provided by formal hypothesis test procedures. Two multivariate graphical methods already are widely used in water-quality studies – Stiff and Piper diagrams. These and other graphical methods are outlined in the following sections. For more detailed discussions on multivariate graphical methods, see **?**, or the textbook by **?**.

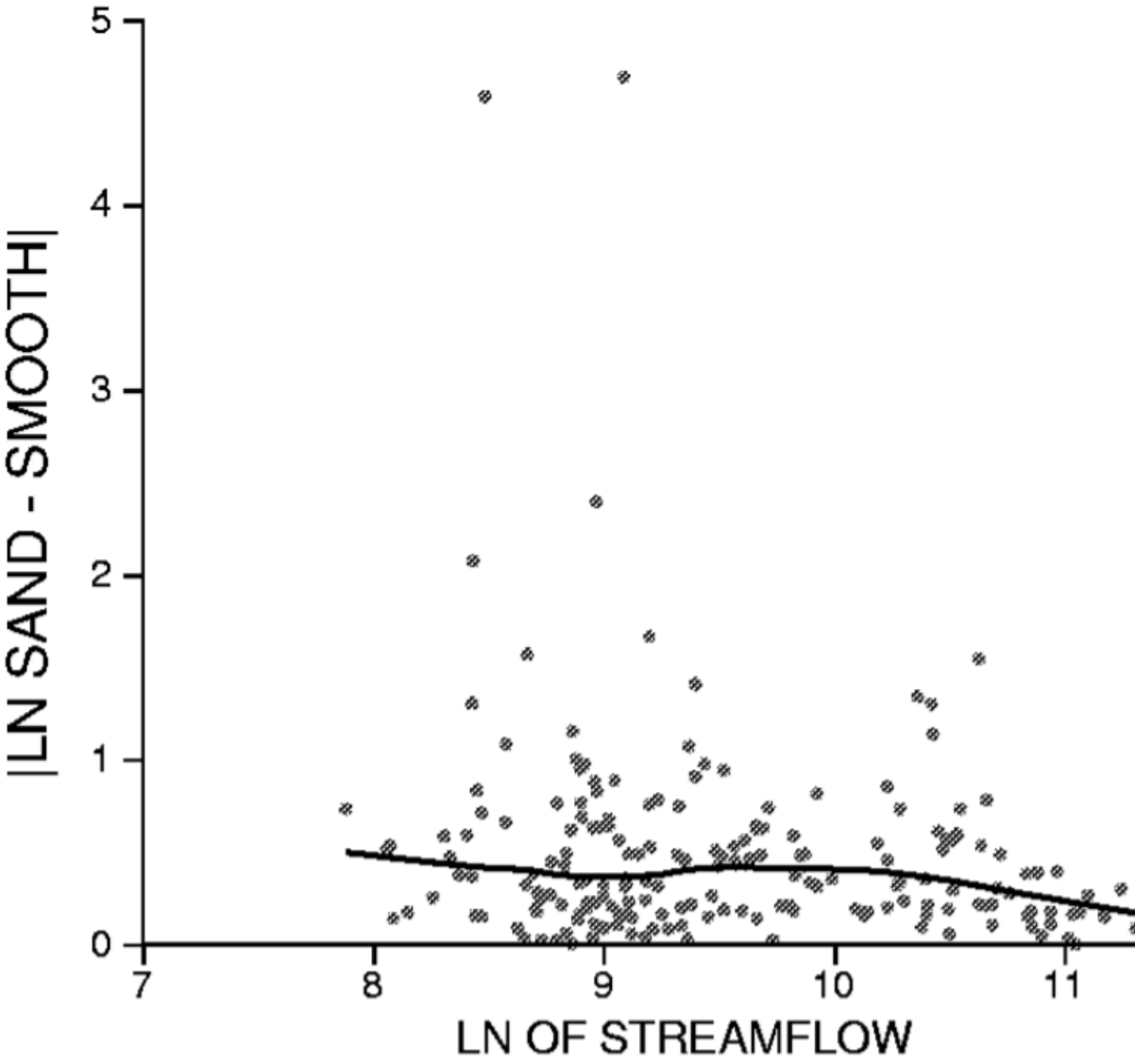
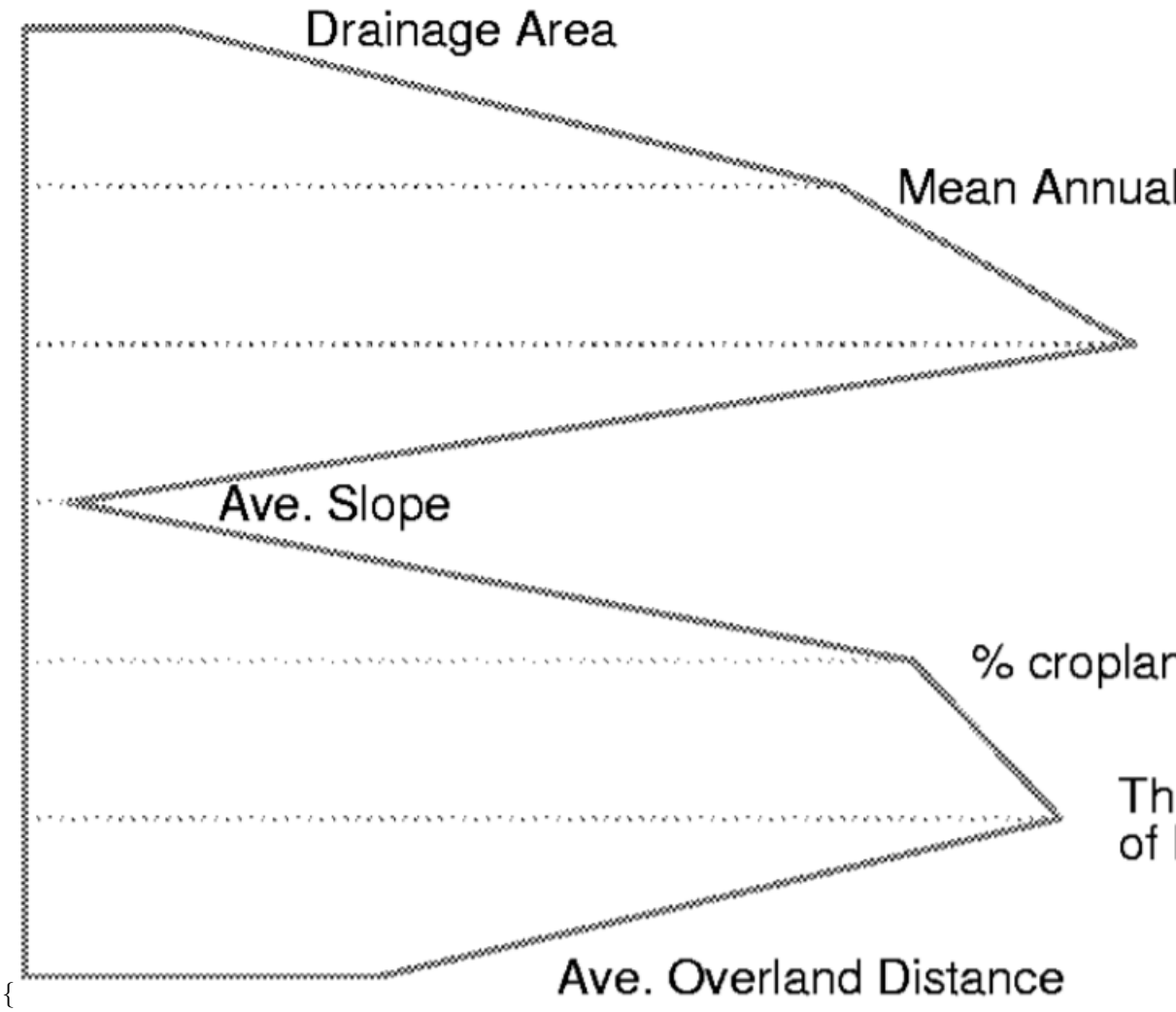


Figure 2.26: Absolute residuals show wheter the spread changes with changing x – sediment concentrations at Lees Ferry, Arizona

2.4.1 Profile Plots

Profile plots are a class of graphical methods which assign each variable to a separate and parallel axis. One observation is represented by a series of points, one per axis, which are connected by a straight line forming the profile. Each axis is scaled independently, based on the range of values in the entire data set. Comparisons between observations are made by comparing profiles.

As an example, assume that sediment loads are to be regionalized. That is, mean annual loads are to be predicted at ungaged sites based on basin characteristics (physical and climatic conditions) at those sites. Of interest may be the interrelationships between sites based on their basin characteristics, as well as which characteristics are associated with high or low annual values. Profile plots such as the one of site basin characteristics in figure **??** would effectively illustrate those relationships. \begin{figure}

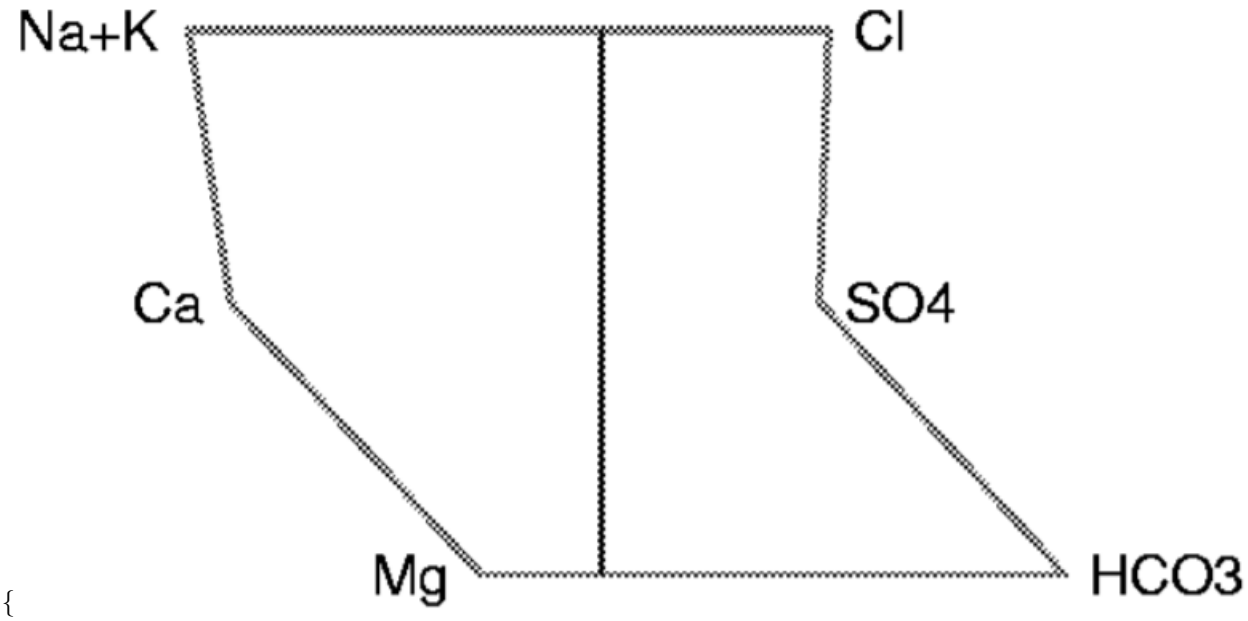


{

}

\caption{Profile plot of selected basin characteristics, Cow Creek near Lyons, Kansas (data from **?**)} \end{figure} #### Stiff diagrams

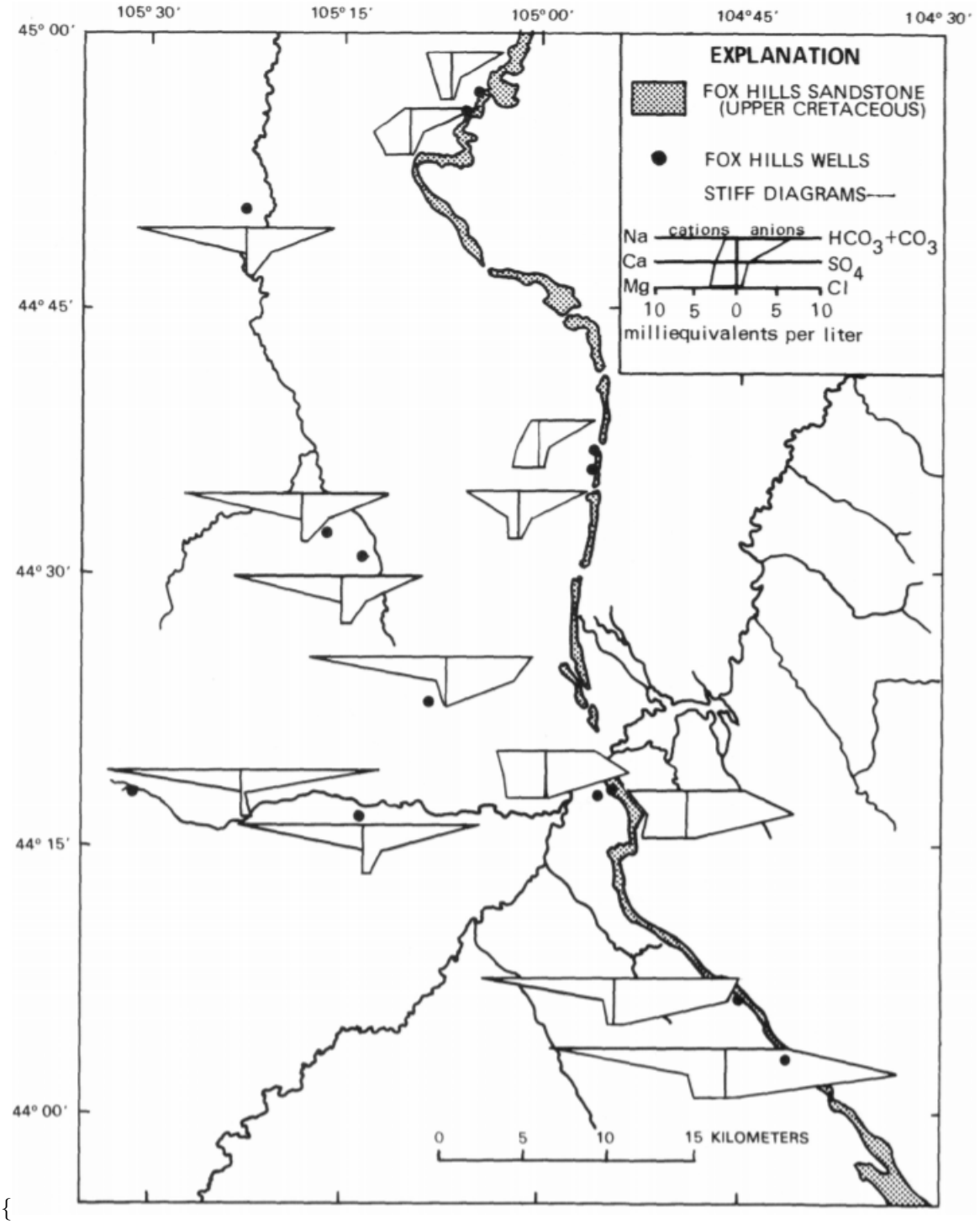
Stiff diagrams (**?**) are the most familiar application of profile plots in water resources. In a Stiff diagram, the milliequivalents of major water-quality constituents are plotted for a single sample, with the cation profile plotted to the left of the center line, and anion profile to the right (figure **??**). Comparisons between several samples based on multiple water-quality constituents is then easily done by comparing shapes of the Stiff diagrams. Figure **??** shows one such comparison for 14 groundwater samples from the Fox Hills Sandstone in Wyoming (**?**). \begin{figure}



{

}

\caption{Stiff diagram for a groundwater sample from the Columbia River Basalt aquifer, Oregon (data from **?**)} \end{figure} \begin{figure}


105° 30′
105° 15′
105° 00′
104° 45′
104° 30′
45° 00′
44° 45′
44° 30′
44° 15′
44° 00′
EXPLANATION
FOX HILLS SANDSTONE (UPPER CRETACEOUS)
FOX HILLS WELLS
STIFF DIAGRAMS
Na
Ca
Mg
cations
anions
HCO_3+CO_3
SO_4
Cl
10 5 0 5 10
milliequivalents per liter
0 5 10 15 KILOMETERS

}

\caption{Stiff diagrams to display areal differences in water quality (from **?**)}
\end{figure} ### Star Plots

A second method of displaying multiple axes is to have them radiate from a central point, rather than aligned parallel as in a profile plot. Again, one observation would be represented by a point on each axis, and these points are connected by line segments. The resulting figures resemble a star pattern, and are often called star plots. Angles between rays of the star are $360^\circ/k$, where k is the number of axes to be plotted. To provide the greatest visual discrimination between observations, rays measuring related characteristics should be grouped together. Unusual observations will stand out as a star looking quite different than the other data, perhaps having an unusually long or short ray. In figure **??**, the basalt water-quality data graphed using a Stiff diagram in figure **??** is displayed as a star plot. Note that the cations are grouped together on the top half of the star, with anions along the bottom.

2.4.1.1 Kite diagrams

A simplified 4-axis star diagram, the "kite diagram", has been used for displaying water-quality compositions, especially to portray compositions of samples located on a map (**?**). Cations are plotted on the two vertical axes, and anions on the two horizontal axes. The primary advantage of this plot is its simplicity. Its major disadvantage is also its simplicity, in that the use of only four axes may hide important characteristics of the data. One might need to know whether calcium or magnesium were present in large amounts, for example, but that could not be determined from the kite diagram. There is no reason why a larger number of axes could not be employed to give more detail, making the plot a true star diagram. Compare for example the basalt data plotted as a star diagram in figure **??** and as a kite diagram in figure **??**.

One innovative use of the kite diagram was made by **?**. They plotted the quartiles of all observations taken from each of several formations, and at different depth ranges, in order to compare water quality between formations and depths (figure **??**). The kite plots in this case are somewhat like multivariate boxplots. There is no reason why the other multivariate plots described here could not also present percentile values for a group of observations rather than descriptions of individual values, and be used to compare among groups of data. \begin{figure}

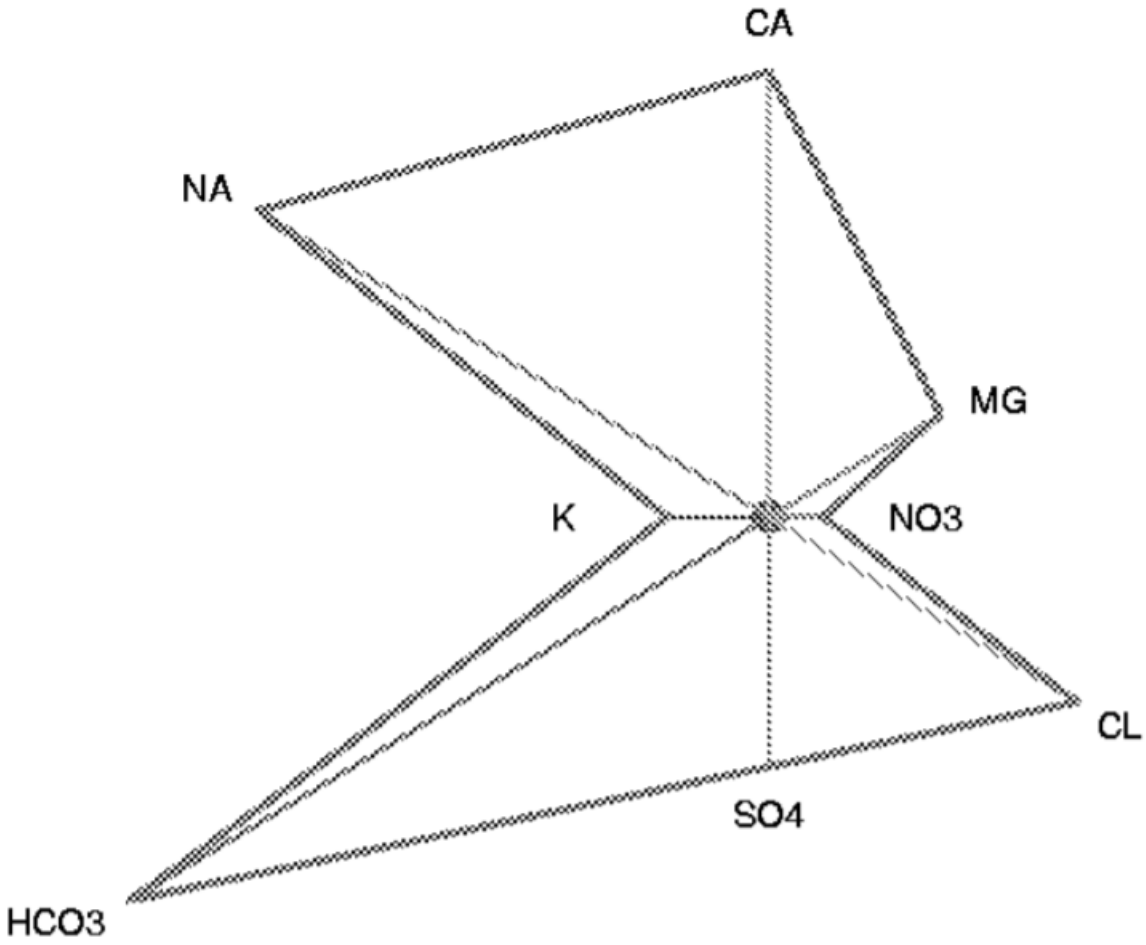


Figure 2.27: Star diagram of the basalt water-quality data

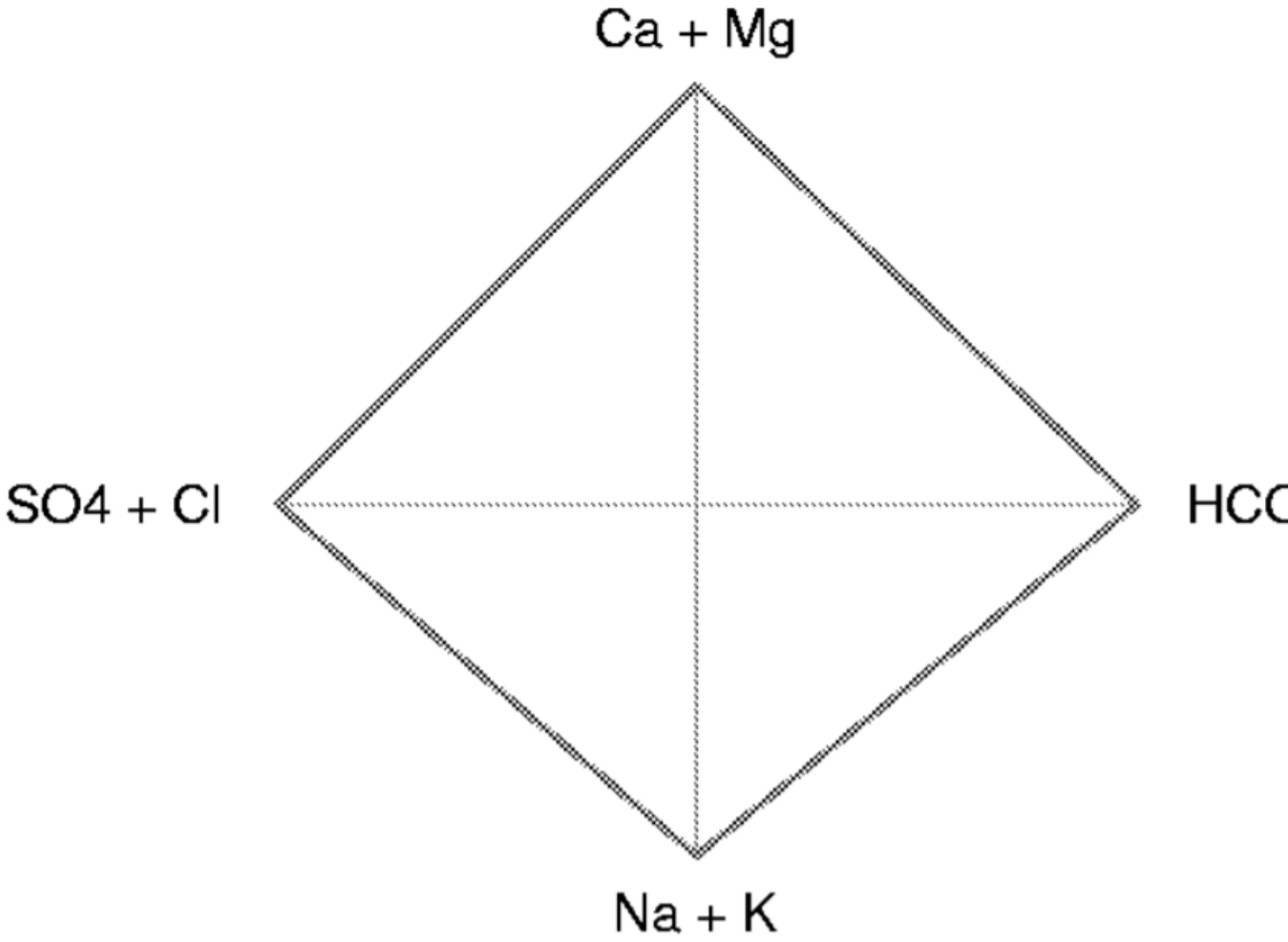
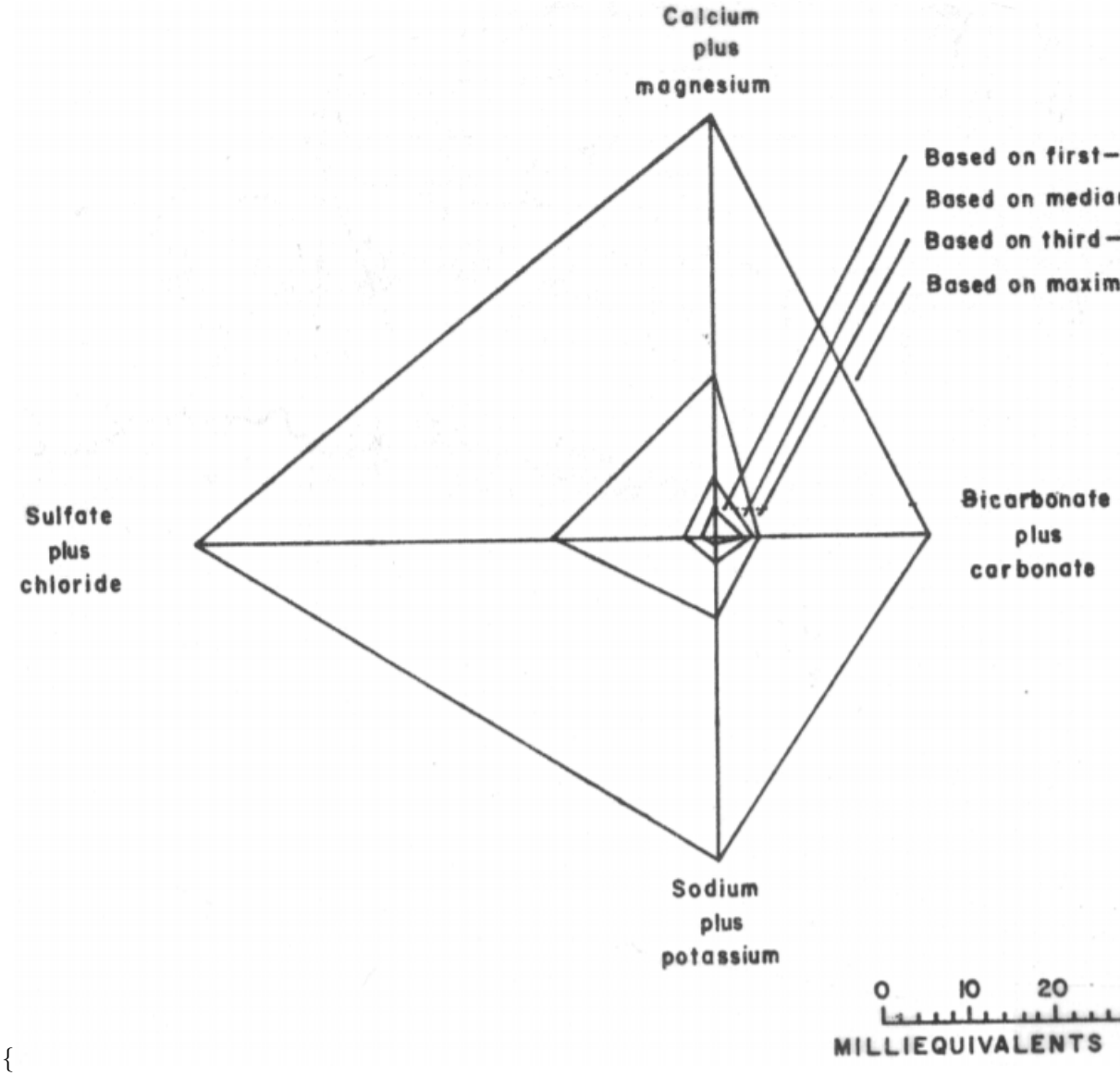


Figure 2.28: Kite diagram of the basalt water-quality data



{

}

\caption{Kite diagram of quartiles of composition from an alluvial formation in Montana (from **?**)} \end{figure} ### Trilinear Diagrams

Trilinear diagrams have been used within the geosciences since the early 1900's. When three variables for a single observation sum to 100 percent, they can be represented as one point on a triangular (trilinear) diagram. Figure **??** is one example – three major cation axes upon which is plotted the cation composition for the basalt data of figure **??**. Each of the three cation values,