Chapter 1 Exploratory Data Analysis

Terminology Differences O

Terminology for rectangular data can be confusing. Statisticians

and data scientists use different terms for the same thing. For a statistician,

*predictor variables* are used in a model to predict a

*response* or *dependent variable*. For a data scientist, *features* are used

to predict a *target*. One synonym is particularly confusing: computer

scientists will use the term *sample* for a single row; a *sample*

to a statistician means a collection of rows.

Graphs in Statistics O

In computer science and information technology, the term *graph*

typically refers to a depiction of the connections among entities,

and to the underlying data structure. In statistics, *graph* is used to

refer to a variety of plots and *visualizations*, not just of connections

among entities, and the term applies only to the visualization, not

to the data structure.

Metrics and Estimates B

Statisticians often use the term estimate for a value calculated from

the data at hand, to draw a distinction between what we see from

the data and the theoretical true or exact state of affairs. Data scientists

and business analysts are more likely to refer to such a value as

a metric. The difference reflects the approach of statistics versus

that of data science: accounting for uncertainty lies at the heart of

the discipline of statistics, whereas concrete business or organizational

objectives are the focus of data science. Hence, statisticians

estimate, and data scientists measure.

Anomaly Detection B

In contrast to typical data analysis, where outliers are sometimes

informative and sometimes a nuisance, in *anomaly detection* the

points of interest are the outliers, and the greater mass of data

serves primarily to define the “normal” against which anomalies

are measured.

Other Robust Metrics for Location H

Statisticians have developed a plethora of other estimators for location,

primarily with the goal of developing an estimator more

robust than the mean and also more efficient (i.e., better able to

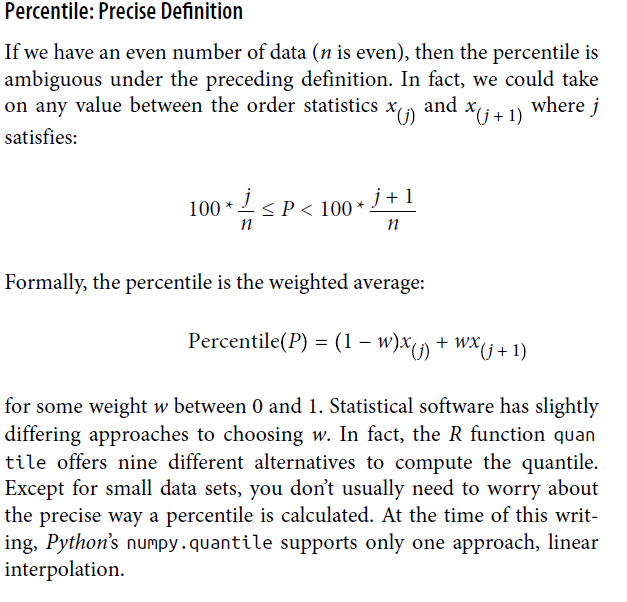
discern small location differences between data sets). While these

methods are potentially useful for small data sets, they are not

likely to provide added benefit for large or even moderately sized

data sets.

Estimates

H

Statistical Moments H

In statistical theory, location and variability are referred to as the

first and second *moments* of a distribution. The third and fourth

moments are called *skewness* and *kurtosis*. Skewness refers to

whether the data is skewed to larger or smaller values, and kurtosis

indicates the propensity of the data to have extreme values. Generally,

metrics are not used to measure skewness and kurtosis;

instead, these are discovered through visual displays such as Figures

1-2 and 1-3.

Density Estimation H

Density estimation is a rich topic with a long history in statistical

literature. In fact, over 20 *R* packages have been published that

offer functions for density estimation. [Deng-Wickham-2011] give

a comprehensive review of *R* packages, with a particular recommendation

for ASH or KernSmooth. The density estimation methods

in pandas and scikit-learn also offer good implementations. For

many data science problems, there is no need to worry about the

various types of density estimates; it suffices to use the base

functions

Numerical Data as Categorical Data B

In “Frequency Tables and Histograms” on page 22, we looked at

frequency tables based on binning the data. This implicitly converts

the numeric data to an ordered factor. In this sense, histograms and

bar charts are similar, except that the categories on the x-axis in the

bar chart are not ordered. Converting numeric data to categorical

data is an important and widely used step in data analysis since it

reduces the complexity (and size) of the data. This aids in the discovery

of relationships between features, particularly at the initial

stages of an analysis.

Other Correlation Estimates H

Statisticians long ago proposed other types of correlation coefficients,

such as *Spearman’s rho* or *Kendall’s tau*. These are correlation

coefficients based on the rank of the data. Since they work

with ranks rather than values, these estimates are robust to outliers

and can handle certain types of nonlinearities. However, data scientists

can generally stick to Pearson’s correlation coefficient, and its

robust alternatives, for exploratory analysis. The appeal of rankbased

estimates is mostly for smaller data sets and specific hypothesis

tests.