MACHINE LEARNING FROM DATA
Fall 2018

Exploratory data analysis

Index

1. What is exploratory data analysis

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments [1].
The goal of EDA is to explore the data to reveal patterns and features that will help the analyst better understand, analyze and model the data.
EDA is a collection of techniques for revealing information about the data and methods for visualizing them to see what they can tell us about the underlying process that generated it. In most situations, exploratory data analysis should precede confirmatory analysis (e.g., hypothesis testing, ANOVA, etc.) to ensure that the analysis is appropriate for the data set. [3]
Explanatory data analysis is majorly performed using the following methods:

- Univariate visualization: provides summary statistics for each feature in the dataset
- Multivariate visualization: is performed to understand interactions between features
- Dimensionality reduction: to reduce the number of variables under consideration by obtaining a set of principal variables
- Cluster analysis: to find natural grouping and structure in data

Here we will only discuss the first two groups of methods, dimensionality reduction and clustering will be addressed later in the course.

2. Histograms

A histogram is a way to graphically represent the frequency distribution of a dataset [2]. Histograms are a good way to

- Summarize a dataset to understand general characteristics of the distribution such as shape, spread or location
- Suggest possible probabilistic models
- Determine unusual behavior

A frequency histogram is obtained by creating a set of bins or intervals that cover the range of the dataset and counting the number of observations that fall into each bin. Usually bins do not overlap and have equal width.
A histogram may also be normalized to display relative frequencies. It then shows the proportion of observations that fall into each bin, with the sum of the values equaling 1.
Matlab functions for creating a histogram and normalized histogram plots of x:

```
histogram(x)
histogram(x,'Normalization','Probability')
```

3. Box-plots

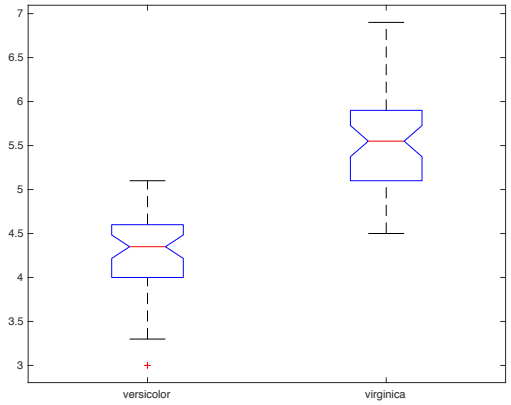
Box plots provide a visualization of summary statistics for sample data and contain the following features:

- The tops and bottoms of each "box" are the 25th and 75th percentiles of the samples, respectively. The distances between the tops and bottoms are the interquartile ranges. You can compute the value of the interquartile range using `iqr`.
- The line in the middle of each box is the sample median. If the median is not centered in the box, it shows sample skewness. You can compute the value of the median using the `median` function.

- The whiskers are lines extending above and below each box. Whiskers are drawn from the ends of the interquartile ranges to the furthest observations within the whisker length.
- Observations beyond the whisker length are marked as outliers. By default, an outlier is a value that is more than 1.5 times the interquartile range away from the top or bottom of the box, but this value can be adjusted with additional input arguments. Outliers are displayed with a red + sign.
- Notches display the variability of the median between samples. The width of a notch is computed so that box plots whose notches do not overlap (as above) have different medians at the 5% significance level. The significance level is based on a normal distribution assumption, but comparisons of medians are reasonably robust for other distributions. Comparing box-plot medians is like a visual hypothesis test, analogous to the *t* test used for means.

Example

```
load fisheriris
s1 = meas(51:100,3);
s2 = meas(101:150,3);
figure
boxplot([s1 s2],'notch','on',...
        'labels',{'versicolor','virginica'})
```



In the example, the notches of the two box plots do not overlap, which indicates that the median petal length of the versicolor and virginica irises are significantly different at the 5% significance level. The median line in the versicolor plot does not appear to be centered inside the box, which indicates that the sample is slightly skewed. Additionally, the versicolor data contains one outlier value, while the virginica data does not contain any outliers.

4. Measures of central tendency

Measures of central tendency locate a distribution of data along an appropriate scale.

The following functions calculate the measures of central tendency:

`geomean` (geometric mean), `harmmean` (harmonic mean), `mean` (arithmetic average), `median` (50th percentile), `mode` (most frequent value), `trimmean` (trimmed mean)

The average (`mean`) is a simple and popular estimate of location. If the data sample comes from a normal distribution, then the sample mean is also optimal (minimum variance unbiased estimator). Unfortunately, outliers, data entry errors, or glitches exist in almost all real data. The sample mean is sensitive to these problems. One bad data value can move the average away from the center of the rest of the data by an arbitrarily large distance.

The median and trimmed mean are two measures that are resistant (robust) to outliers. The median is the 50th percentile of the sample, which will only change slightly if you add a large perturbation to any value. The idea

behind the trimmed mean is to ignore a small percentage of the highest and lowest values of a sample when determining the center of the sample.

The geometric mean and harmonic mean, like the average, are not robust to outliers. They are useful when the sample is distributed lognormal or heavily skewed.

Example: compute and compare measures of location for sample data that contains one outlier.

```
% Generate sample data that contains one outlier.
x = [ones(1,6),100]
% Compute the geometric mean, harmonic mean, mean, median, and trimmed mean
% for the sample data.
locate = [geomean(x) harmmean(x) mean(x) median(x) trimmean(x,25)]

x =      1     1     1     1     1     1   100
locate = 1.9307    1.1647   15.1429    1.0000    1.0000
```

In the example the mean (`mean`) is far from any data value because of the influence of the outlier. The geometric mean (`geomean`) and the harmonic mean (`harmmean`) are influenced by the outlier, but not as significantly. The median (`median`) and trimmed mean (`trimmean`) ignore the outlier value and describe the location of the rest of the data values.

5. Measures of dispersion

The purpose of measures of dispersion is to find out how spread out the data values are on the number line. Another term for these statistics is measures of spread.

The following functions calculate the measures of dispersion: `iqr` (interquantile range), `mad` (mean absolute deviation), `moment` (central moment of all orders), `range` (range), `std` (standard deviation), `var` (variance)

The `range` (the difference between the maximum and minimum values) is the simplest measure of spread. But if there is an outlier in the data, it will be the minimum or maximum value. Thus, the range is not robust to outliers.

The `standard deviation` and the `variance` are popular measures of spread that are optimal for normally distributed samples. Neither the standard deviation nor the variance is robust to outliers. A data value that is separate from the body of the data can increase the value of the statistics by an arbitrarily large amount.

The mean absolute deviation (`mad`) is also sensitive to outliers. But the MAD does not move quite as much as the standard deviation or variance in response to bad data.

The interquartile range (`iqr`) is the difference between the 75th and 25th percentile of the data. Since only the middle 50% of the data affects this measure, it is robust to outliers.

Example: compute and compare measures of dispersion for sample data that contains one outlier.

```
#Generate sample data that contains one outlier value.
x = [ones(1,6),100]
#Compute the interquartile range, mean absolute deviation,
#range, and standard deviation of the sample data.
stats = [iqr(x),mad(x),range(x),std(x)]
```

The interquartile range (iqr) is the difference between the 75th and 25th percentile of the sample data, and is robust to outliers. The range (range) is the difference between the maximum and minimum values in the data, and is strongly influenced by the presence of an outlier. Both the mean absolute deviation (mad) and the standard deviation (std) are sensitive to outliers. However, the mean absolute deviation is less sensitive than the standard deviation.

Skewness

Skewness is a measure of the **asymmetry** of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If skewness is positive, the data are spread out more to the right. The skewness of the normal distribution (or any perfectly symmetric distribution) is zero.

The skewness of a distribution is defined as

$$Sk(x) = \frac{\mu_3}{\mu_2\sqrt{\mu_2}} = \frac{E\left[\left(x-\mu\right)^3\right]}{\sigma^3}$$

where μ is the mean of x and σ is the standard deviation of x. The Matlab function `skewness` computes a sample version of this population value.

Kurtosis

Kurtosis is a measure of the "**tailedness**" of a distribution, or how outlier-prone a distribution is. For this measure, higher kurtosis is the result of infrequent extreme deviations (or outliers), as opposed to frequent modestly sized deviations.

The kurtosis of the normal distribution is 3. Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3.

The kurtosis of a distribution is defined as

$$K(x) = \frac{\mu_4}{\left(\mu_2\right)^2} = \frac{E\left[\left(x-\mu\right)^4\right]}{\sigma^4}$$

where μ is the mean of x and σ is the standard deviation of x. The Matlab function `kurtosis` computes a sample version of this population value. It is common to compute $K(x)-3$, so that the normal distribution has a value of 0.

6. Scatter plots

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. Matlab function `scatter(x,y)` creates a scatter plot with circles at the locations specified by the vectors x and y.

`gplotmatrix(x,y,group)` creates a matrix of scatter plots. Each individual set of axes in the resulting figure contains a scatter plot of a column of x against a column of y. All plots are grouped by the grouping variable `group`. x and y are matrices with the same number of rows. If x has *p* columns and y has *q* columns, the figure contains a *p*-by-*q* matrix of scatter plots.

If x is a NxD matrix corresponding to a dataset with N subjects and D features per subject in a classification problem with C classes, `gplotmatrix(x,x,C)` can be used to create scatter plots of all pairs of features grouped by class. It is useful to visually analyze the separability of the classes.

7. Distribution plots

Distribution plots visually assess the distribution of sample data by comparing the empirical distribution of the data with the theoretical values expected from a specified distribution. Use distribution plots in addition to more formal hypothesis tests to determine whether the sample data comes from a specified distribution.

Matlab Statistics and Machine Learning Toolbox™ offers several distribution plot options:

- *Normal probability plots* assess whether sample data comes from a normal distribution (`normplot`).

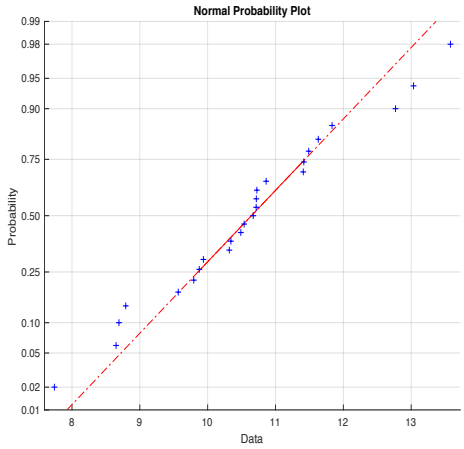
- *Quantile-quantile (q-q) plots* assess whether two sets of sample data come from the same distribution family, and is robust with respect to differences in location and scale (`qqplot`).
- *Cumulative distribution plots* display the empirical cumulative distribution function (cdf) of the sample data for visual comparison to the theoretical cdf of a specified distribution (`cdfplot`, `ecdf`, `stairs`)
- You can create distribution plots for distributions other than normal, or explore the distribution of censored data, using `probplot`.

Normal probability plots:

Normal probability plots are used to assess whether data comes from a normal distribution. Many statistical procedures make the assumption that an underlying distribution is normal, so normal probability plots can provide some assurance that the assumption is justified, or else provide a warning of problems with the assumption. An analysis of normality typically combines normal probability plots with hypothesis tests for normality.

Example: generate a data sample of 25 random numbers from a normal distribution with mu = 10 and sigma = 1, and create a normal probability plot of the data.

```
rng default;  % For reproducibility
x = normrnd(10,1,25,1);
normplot(x)
```



The plus signs plot the empirical probability versus the data value for each point in the data. A solid line connects the 25th and 75th percentiles in the data, and a dashed line extends it to the ends of the data. The *y*-axis values are probabilities from zero to one, but the scale is not linear. The distance between tick marks on the *y*-axis matches the distance between the quantiles of a normal distribution. The quantiles are close together near the median (probability = 0.5) and stretch out symmetrically as you move away from the median.

In a normal probability plot, if all the data points fall near the line, an assumption of normality is reasonable. Otherwise, the points will curve away from the line, and an assumption of normality is not justified.

Quantile-quantile plots

Quantile-quantile plots are used to determine whether two samples come from the same distribution family. They are scatter plots of quantiles computed from each sample, with a line drawn between the first and third quartiles. If the data falls near the line, it is reasonable to assume that the two samples come from the same distribution. The method is robust with respect to changes in the location and scale of either distribution.

To create a quantile-quantile plot, use the `qqplot` function.

Example:

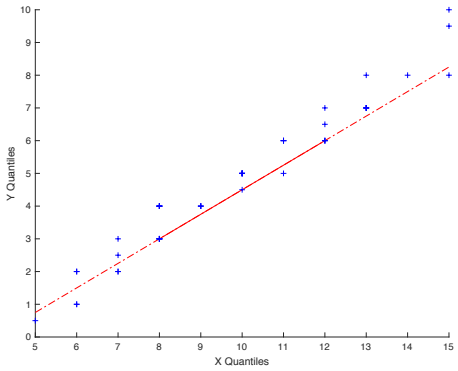
The following example generates two data samples containing random numbers from Poisson distributions with different parameter values, and creates a quantile-quantile plot. The data in x is from a Poisson distribution with lambda = 10, and the data in y is from a Poisson distribution with lambda = 5.

```
x = poissrnd(10,50,1);
y = poissrnd(5,100,1);
qqplot(x,y);
```

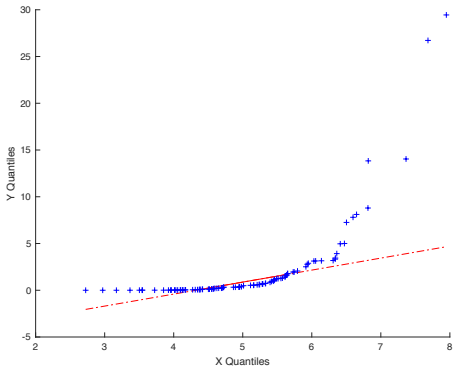
Even though the parameters and sample sizes are different, the approximate linear relationship suggests that the two samples may come from the same distribution family. As with normal probability plots, hypothesis tests can provide additional justification for such an assumption. For statistical procedures that depend on the two samples coming from the same distribution, however, a linear quantile-quantile plot is often sufficient.

The following example shows what happens when the underlying distributions are not the same. Here, x contains 100 random numbers generated from a normal distribution with mu = 5 and sigma = 1, while y contains 100 random numbers generated from a Weibull distribution with A = 2 and B = 0.5.

```
x = normrnd(5,1,100,1);
y = wblrnd(2,0.5,100,1);
qqplot(x,y);
```



Example1



Example2

Cumulative distribution plots

An empirical cumulative distribution function (cdf) plot shows the proportion of data less than each *x* value, as a function of *x*. The scale on the *y*-axis is linear; in particular, it is not scaled to any particular distribution. Empirical cdf plots are used to compare data cdfs to cdfs for particular distributions.

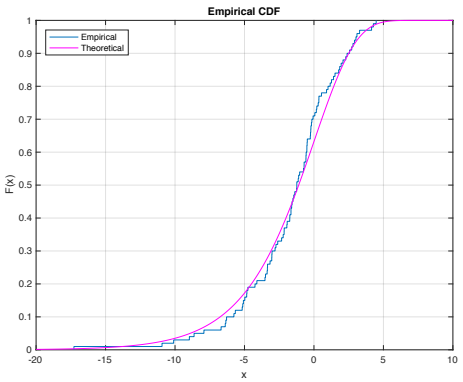
To create an empirical cdf plot, use the `cdfplot` function (or `ecdf` and `stairs`).

Example:

The following example compares the empirical cdf for a sample from an extreme value distribution with a plot of the cdf for the sampling distribution. In practice, the sampling distribution would be unknown, and would be chosen to match the empirical cdf.

```
y = evrnd(0,3,100,1);
cdfplot(y)
hold on
x = -20:0.1:10;
f = evcdf(x,0,3);
plot(x,f,'m')
```

```
legend('Empirical','Theoretical','Location','NW')
```



Other probability plots

A probability plot, like the normal probability plot, is just an empirical cdf plot scaled to a particular distribution. The *y*-axis values are probabilities from zero to one, but the scale is not linear. The distance between tick marks is the distance between quantiles of the distribution. In the plot, a line is drawn between the first and third quartiles in the data. If the data falls near the line, it is reasonable to choose the distribution as a model for the data.

To create probability plots for different distributions, use the `probplot` function.

Example:

The following example assesses two samples, one from a Weibull distribution with A = 3 and B = 3, and one from a Rayleigh distribution with B = 3, to see if either distribution may have come from a Weibull population.

```
x1 = wblrnd(3,3,100,1);
x2 = raylrnd(3,100,1);
probplot('weibull',[x1 x2])
legend('Weibull Sample','Rayleigh Sample','Location','NW')
```

8. z-scores

For a random variable *x* with mean μ and standard deviation σ, the *z*-score of a value *x* is

$$z = \frac{(x - \mu)}{\sigma}$$

For sample data with mean $\bar{X}$and standard deviation *S*, the *z*-score of a data point *x* is

$$z = \frac{(x - \bar{X})}{S}$$

z-scores measure the distance of a data point from the mean in terms of the standard deviation. This is also called *standardization* of data. The standardized data set has mean 0 and standard deviation 1, and retains the shape properties of the original data set (same skewness and kurtosis).

You can use *z*-scores to put data on the same scale before further analysis. This lets you to compare two or more data sets with different units.

9. Random number generator

The Random Number Generation user interface generates random samples from specified probability distributions, and displays the samples as histograms. Use the interface to explore the effects of changing parameters and sample size on the distributions.

Run the user interface by typing `randtool` at the command line. Start by selecting a distribution, then enter the desired sample size.

You can also

- Use the controls at the bottom of the window to set parameter values for the distribution and to change their upper and lower bounds.
- Draw another sample from the same distribution, with the same size and parameters.
- Export the current sample to your workspace. A dialog box enables you to provide a name for the sample.

10.Confidence interval

A confidence interval is an estimated range of values with a specified probability of containing the true population value of a parameter. Upper and lower bounds for confidence intervals are computed from the sample estimate of the parameter and the known (or assumed) sampling distribution of the estimator. A typical assumption is that estimates will be normally distributed with repeated sampling (as dictated by the Central Limit Theorem). Wider confidence intervals correspond to poor estimates (smaller samples); narrow intervals correspond to better estimates (larger samples).
Confidence interval for the mean:

Let's assume we have samples randomly selected from a random process with unknown parameters (mean and variance). We want to estimate the actual population mean $\boldsymbol{\mu_o}$, but we can only get the sample mean $\bar{x}$, a point estimate (a single number). A confidence interval provides additional information about variability.

A confidence interval gives a range estimate of values: it takes into consideration variation in sample statistics from sample to sample, it is based on all the observations from one sample and gives information about closeness to unknown population parameters; it is stated in terms of a level of confidence-

Confidence Level quantifies the level of confidence that the interval will contain the unknown parameter ($\boldsymbol{\mu_o}$) a percentage (less than 100%).

The *level of significance*, or "α" is the chance we take that the true population parameter is not contained in the confidence interval. Therefore, a 95% confidence interval would have an "α" of 5%

If the population standard deviation **σ** is unknown, we can substitute the sample standard deviation (**s**). This introduces extra uncertainty, since '**s**' is variable from sample to sample.

Given a sample, we compute the statistic *t* , which follows a *t*-student distribution with *n*-1 degrees of freedom

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}; \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

t follows a *t*-student distribution with *n*-1 degrees of freedom $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
For a level of significance α = 0.05

Let $t_{a/2}$ be the value such that

$$\frac{\alpha}{2} = \Pr\left\{t \le -t_{\alpha/2}\right\} = \Pr\left\{+t_{\alpha/2} \le t\right\}$$

Then

$$P\left(-t_{\alpha/2} \le t \le t_{\alpha/2}\right) = 1-\alpha$$
$$P\left(-t_{\alpha/2} \le \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \le t_{\alpha/2}\right) = 1-\alpha$$
$$P\left(\overline{x}-t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu_0 \le \overline{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

So we can state that the true mean value m_0 is within this interval

$$\left(\overline{x}-t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

with (1-α)% = 95% confidence.

11. Hypothesis testing

Hypothesis testing is a common method of drawing inferences about a population based on statistical evidence from a sample.

As an example, suppose someone says that at a certain time in the state of Massachusetts the average price of a gallon of regular unleaded gas was $1.15. How could you determine the truth of the statement? You could try to find prices at every gas station in the state at the time. That approach would be definitive, but it could be time-consuming, costly, or even impossible.

A simpler approach would be to find prices at a small number of randomly selected gas stations around the state, and then compute the sample average. Sample averages differ from one another due to chance variability in the selection process. Suppose your sample average comes out to be $1.18. Is the $0.03 difference an artifact of random sampling or significant evidence that the average price of a gallon of gas was in fact greater than $1.15? Hypothesis testing is a statistical method for making such decisions.

All hypothesis tests share the same basic terminology and structure:

A *null hypothesis* is an assertion about a population that you would like to test. It is "null" in the sense that it often represents a status quo belief, such as the absence of a characteristic or the lack of an effect. It may be formalized by asserting that a population parameter, or a combination of population parameters, has a certain value. In the example given, the null hypothesis would be that the average price of gas across the state was $1.15. This is written H0: μ = 1.15.

An *alternative hypothesis* is a contrasting assertion about the population that can be tested against the null hypothesis. In the example given, possible alternative hypotheses are:

H1: μ ≠ 1.15 — State average was different from $1.15 (two-tailed test)

H1: μ > 1.15 — State average was greater than $1.15 (right-tail test)

H1: μ < 1.15 — State average was less than $1.15 (left-tail test)

To conduct a hypothesis test, a random sample from the population is collected and a relevant test statistic is computed to summarize the sample. This statistic varies with the type of test, but its distribution under the null hypothesis must be known (or assumed).

The *p-value* of a test is the probability, under the null hypothesis, of obtaining a value of the test statistic as extreme or more extreme than the value computed from the sample.

The *significance level* of a test is a threshold of probability α agreed to before the test is conducted. A typical value of α is 0.05. If the p-value of a test is less than α, the test rejects the null hypothesis. If the p-value is greater than α, there is insufficient evidence to reject the null hypothesis. Note that lack of evidence for rejecting the

null hypothesis is not evidence for accepting the null hypothesis. Also note that substantive "significance" of an alternative cannot be inferred from the statistical significance of a test.

The significance level α can be interpreted as the probability of rejecting the null hypothesis when it is actually true—a type I error. The distribution of the test statistic under the null hypothesis determines the probability α of a type I error. Even if the null hypothesis is not rejected, it may still be false—a type II error. The distribution of the test statistic under the alternative hypothesis determines the probability β of a type II error. Type II errors are often due to small sample sizes. The power of a test, 1 – β, is the probability of correctly rejecting a false null hypothesis.

Results of hypothesis tests are often communicated with a confidence interval. If the null hypothesis asserts the value of a population parameter, the test rejects the null hypothesis when the hypothesized value lies outside the computed confidence interval for the parameter.

Different hypothesis tests make different assumptions about the distribution of the random variable being sampled in the data. These assumptions must be considered when choosing a test and when interpreting the results.

One of the tests that we are going to use is the *chi-square goodness-of-fit test*, that determines if a data sample comes from a specified probability distribution, with parameters estimated from the data.

The test groups the data into bins, calculating the observed and expected counts for those bins, and computing the chi-square test statistic

$$\chi^2 = \sum_{i=1}^{N} (O_i - E_i)^2 / E_i$$

where O_i are the observed counts and E_i are the expected counts based on the hypothesized distribution. The test statistic has an approximate chi-square distribution when the counts are sufficiently large.

`chi2gof(x)` returns a test decision for the null hypothesis that the data in vector x comes from a normal distribution with a mean and variance estimated from x, using the chi-square goodness-of-fit test. The alternative hypothesis is that the data does not come from such a distribution. The result h is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise.

12. Example of exploratory analysis of data

This example shows how to explore the distribution of data using descriptive statistics.

1. *Generate sample data*

Generate a vector containing randomly-generated sample data.

```
rng default  % For reproducibility
x = [normrnd(4,1,1,100),normrnd(6,0.5,1,200)];
```

2. *Plot a histogram.*

Plot a histogram of the sample data with a normal density fit. This provides a visual comparison of the sample data and a normal distribution fitted to the data.

```
histfit(x)
```

The distribution of the data appears to be left skewed. A normal distribution does not look like a good fit for this sample data.

3. Obtain a normal probability plot.

Obtain a normal probability plot. This plot provides another way to visually compare the sample data to a normal distribution fitted to the data.

```
probplot('normal',x)
```

The probability plot also shows the deviation of data from normality.

4. Compute the quantiles.

Compute the quantiles of the sample data.

```
p = 0:0.25:1;
y = quantile(x,p);
z = [p;y]
```

5. Create a box plot to visualize the statistics.

The box plot shows the 0.25, 0.5, and 0.75 quantiles. The long lower tail and plus signs show the lack of symmetry in the sample data values.

```
boxplot(x)
```

6. Compute descriptive statistics.

Compute the mean and median of the data.

```
y = [mean(x),median(x)]
```

The mean and median values seem close to each other, but a mean smaller than the median usually indicates that the data is left skewed.

Compute the skewness and kurtosis of the data.

```
y = [skewness(x),kurtosis(x)]
```

A negative skewness value means the data is left skewed. The data has a larger peakedness than a normal distribution because the kurtosis value is greater than 3.

7. Compute z-scores.

Identify possible outliers by computing the z-scores and finding the values that are greater than 3 or less than -3.

```
Z = zscore(x);
find(abs(Z)>3);
```

Based on the z-scores, the 3rd and 35th observations might be outliers.

13. References

[1] https://en.wikipedia.org/wiki/Exploratory_data_analysis
[2] Computational statistics handbook with Matlab, W. Martinez, A. Martinez, 2002.
[3] Matlab Documentation