Statistical Geophysics

Chapter 2

Descriptive Statistics 2

Descriptive Statistics 2

Numerical Summary Measures

Background

- The numerical summaries presented in this section can be subdivided into measures of location, spread and shape.

  1. Location refers to the central tendency of the data values.
  2. Spread denotes the degree of variation or dispersion around the center.
  3. Measures of shape tell you the amount and direction of departure from symmetry and how tall and sharp the central peak of the data is.

- Let X be the variable of interest. Suppose a sample of size n is given with observed values $x_1, \ldots, x_n$.

Mode

- The mode, x_{mod}, is the most frequently occurring value or category of X.
- The mode is the most important measure of location for categorical variables.
- The mode of the sample $\{1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17\}$ is 6.
- Given the list of data $\{1, 1, 2, 4, 4\}$ the mode is not unique - the data set may be said to be bimodal, while a set with more than two modes may be described as multimodal.

Arithmetic mean

- The arithmetic mean or average of a sample is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad ,$$

  for which it holds that $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.

- For frequency data with different observed values $a_1, \ldots, a_k$ and relative frequencies $f_1, \ldots, f_k$ the mean is

$$\bar{x} = \sum_{j=1}^{k} a_j f_j \quad .$$

- The mean is a meaningful measure for metric data.

- It is not a robust statistic, meaning that it is strongly affected by outliers.

Median

- The sorted, or ranked, data values from a particular sample are called the order statistics of that sample.
- Given $x_1, x_2, \ldots, x_n$ the order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ for this sample are the same numbers, sorted in ascending order.
- Equal proportions of the data fall above and below the median, x_{med}. Formally,

$$x_{med} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{if } n \text{ is even} \end{cases} .$$

- The median is a resistant measure of location and is meaningful for variables that possess at least an ordinal scale of measurement.

Quantiles

- A sample quantile, x_p, is a number having the same units as the data, which exceeds that proportion of the data given by the subscript p, with $0 < p < 1$.

- Computation:

$$x_p = \begin{cases} x_{(\lfloor np \rfloor + 1)} & \text{if } np \text{ is not an integer} \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right) & \text{if } np \text{ is an integer} \end{cases} ,$$

  where $\lfloor np \rfloor$ is the largest integer not greater than np.

- Commonly used quantiles: $x_{0.5} = x_{med}$; $x_{0.25}$: first (or lower) quartile; $x_{0.75}$: third (or upper) quartile.

Variance

- The empirical variance of $x_1, \ldots, x_n$ is

$$\tilde{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad .$$

- Since $\mathrm{E}(\tilde{s}^2) = \sigma^2 (n-1)/n$, an unbiased estimator for the population variance, σ^2, is the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad .$$

- The standard deviation, s, is obtained as $s = +\sqrt{s^2}$.

- Both s^2 and s are not resistant measures of dispersion.

Variance decomposition

- k groups $(x_{11}, x_{21}, \dots, x_{n_1,1}), \cdots, (x_{1k}, x_{2k}, \dots, x_{n_k,k})$ with

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \ , \quad (j = 1, \dots, k)$$

and

$$\tilde{s}^2_{n_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \ , \quad (j = 1, \dots, k) \ .$$

- Then

$$\tilde{s}^2_n = \frac{1}{n} \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 + \frac{1}{n} \sum_{j=1}^{k} n_j \tilde{s}^2_{n_j}$$

with $n = \sum_{j=1}^{k} n_j$ and $\bar{x} = \frac{1}{n} \sum_{j=1}^{k} n_j \bar{x}_j$.

Coefficient of variation

- The coefficient of variation is a normalized measure of dispersion of a frequency distribution.

- It is defined as

$$v = \frac{s}{\bar{x}}, \quad \bar{x} > 0 \ .$$

- The CV is independent of scale and can be used to compare different dispersions.

Range

- The range of a set of data is the difference between the largest and smallest values, $x_{(n)} - x_{(1)}$.

- It is the size of the smallest interval which contains all the data and provides an indication of statistical dispersion.

- The range can sometimes be misleading when there are extremely high or low values.

- Example: The range of the sample $\{8, 11, 5, 9, 7, 6, 3616\}$ is $3616 - 5 = 3611$.

Interquartile range

- The most common resistant measure of dispersion is the interquartile range (IQR).
- The IQR is defined as

$$\mathrm{IQR} = x_{0.75} - x_{0.25} \quad .$$

- The IQR is a good index of the spread in the central part of a data set, since it simply specifies the range of the central 50% of the data.

Median absolute deviation

- The IQR does not make use of a substantial fraction of the data.
- The median absolute deviation (MAD) is easiest to understand by imagining the transformation $y_i = |x_i - x_{0.5}|$.
- The MAD is then just the median of the transformed (y_i) values:

$$\text{MAD} = median(y_i) = median|x_i - x_{0.5}| \ .$$

- The MAD is analogous to computation of the standard deviation, but using operations that do not emphasize outlying data.

Skewness and kurtosis

- Skewness and kurtosis measures are often used to describe shape characteristics of a distribution.
- Skewness tells you whether the distribution is symmetric or skewed to one side.
- If the bulk of the data is at the left (right) and the right (left) tail is longer, we say that the distribution is skewed right (left) or positively (negatively) skewed.
- The height and sharpness of the peak relative to the rest of the data are measured by the kurtosis. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak.

Skewness and kurtosis II

- The moment coefficients of skewness, g_1, and kurtosis, g_2, are typically defined as

$$g_1 = \frac{m_3}{m_2^{3/2}} \quad \text{and} \quad g_2 = \frac{m_4}{m_2^2} - 3 \ ,$$

  where the rth sample central moment of a sample of size n is defined as

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r \ .$$

- The sample central moments are not unbiased estimates of the population central moments.

Skewness and kurtosis III

- To remove the bias in g_1 and g_2 corrections need to be applied.

- The sample skewness, G_1, and kurtosis, G_2, are defined as

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \quad \text{and} \quad G_2 = \frac{n-1}{(n-2)(n-3)} [(n+1)g_2 + 6] \ .$$

- $G_1 = 0$ for symmetric distributions; $G_1 > 0$ $(G_1 < 0)$ for distributions that are right-skewed (left-skewed).

- $G_2 = 0$ for mesokurtic distributions; $G_2 > 0$ $(G_2 < 0)$ for distributions that are leptokurtic (platykurtic).

Statistical Geophysics

Chapter 2

Boxplots

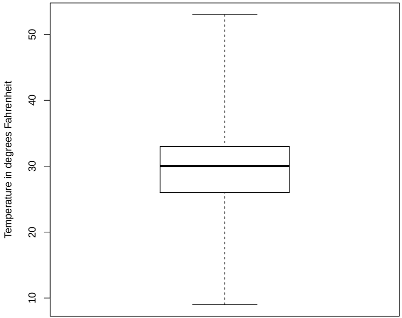
Graphical summary of location measures

- The boxplot, or box-and-whisker plot, is a very widely used graphical tool.

- It is a simple plot of five numbers: the minimum, $x_{(1)}$, the lower quartile, $x_{0.25}$, the median, $x_{0.5}$, the upper quartile, $x_{0.75}$, and the maximum, $x_{(n)}$.

- Using these five numbers, the boxplot presents a sketch of the distribution of the underlying data.

Boxplot: Example II



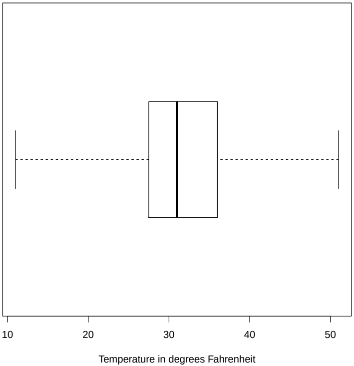


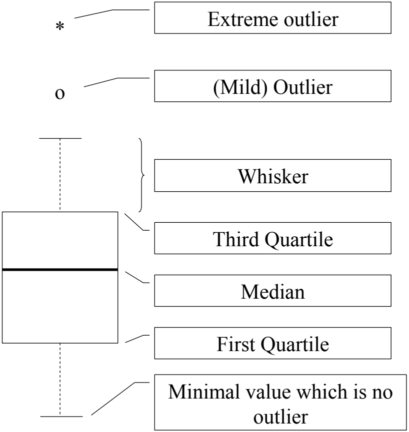
Figure: Boxplot for the January 1987 Ithaca (left) and Canandaigua (right) maximum temperature data ($n = 31$)

Boxplot: modified version

- The following quantities (called fences) can be used for identifying extreme values in the tails of the distribution:
  - lower inner fence: $x_{0.25} - 1.5 \times \mathrm{IQR}$;
  - upper inner fence: $x_{0.75} + 1.5 \times \mathrm{IQR}$;
  - lower outer fence: $x_{0.25} - 3 \times \mathrm{IQR}$;
  - upper outer fence: $x_{0.75} + 3 \times \mathrm{IQR}$.

- Outlier detection criteria: A point beyond an inner fence on either side is considered a mild outlier. A point beyond an outer fence is considered an extreme outlier.

Design of a boxplot



Boxplot: Example II

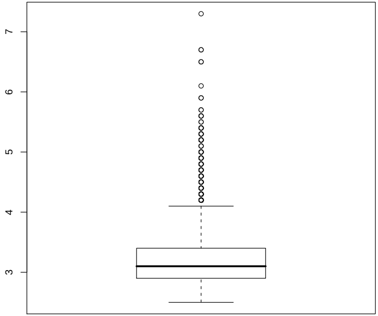


Figure: Boxplot for the earthquake magnitudes in South Carolina, 1987-1996 ($n = 4843$).

Boxplots for variables by group

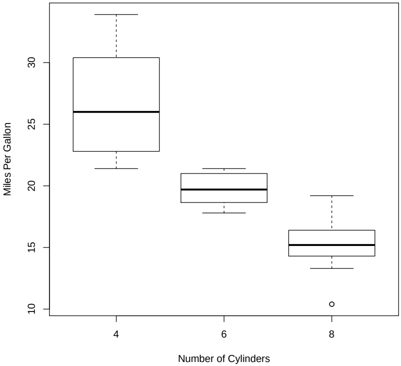


Figure: Boxplot of miles per gallon by car cylinder for car mileage data ($n = 32$).

Exploratory techniques for paired data

Scatterplots

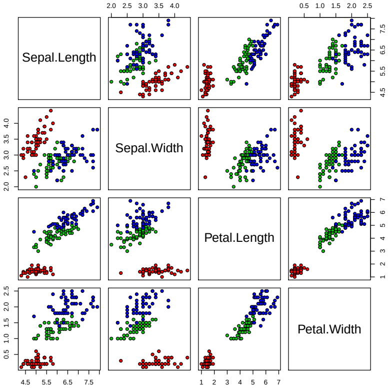


Figure: Scatterplot matrix of iris data ($n = 150$).

Pearson correlation

- Often an abbreviated, single valued measure of association between two variables is needed.

- The term correlation coefficient is used to mean the Pearson product-moment coefficient of linear correlation between two variables X and Y. Formally,

$$r_{XY} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \ ,$$

  where $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ is the sample covariance of X and Y.

- The heart of the Pearson correlation is the covariance between X and Y in the numerator. The denominator is in effect just a scaling constant.

Pearson correlation II

- $-1 \leq r_{XY} \leq 1$
- Interpretation:
  - $r_{XY} > 0$: positive linear correlation.
  - $r_{XY} < 0$: negative linear correlation.
  - $r_{XY} = 0$: no linear correlation.
- It is computationally easier to calculate

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} \quad .$$

Spearman rank correlation

- A robust measure of association is the Spearman rank correlation coefficient.

- The Spearman correlation is simply the Pearson correlation coefficient computed using the ranks of the data. Formally,

$$r_{SP} = \frac{\sum(\text{rank}(x_i) - \overline{\text{rank}}_X)(\text{rank}(y_i) - \overline{\text{rank}}_Y)}{\sqrt{\sum(\text{rank}(x_i) - \overline{\text{rank}}_X)^2 \sum(\text{rank}(y_i) - \overline{\text{rank}}_Y)^2}} \ ,$$

  where $\overline{\text{rank}}_X$ and $\overline{\text{rank}}_Y$ are the averages of the ranks of X and Y, respectively.

- The Spearman correlation can be used for variables that are measured on an ordinal scale.

Spearman rank correlation II

- In cases of ties (a particular data value appears more than once) all of these equal values are assigned their average rank.

- $-1 \leq r_{SP} \leq 1$

- Interpretation:
  - $r_{SP} > 0$: Y tends to increase when X increases.
  - $r_{SP} < 0$: Y tends to decrease when X increases.
  - $r_{SP} = 0$: No tendency for Y to either increase or decrease when X increases.

- If there are no ties, r_{SP} can be computed as

$$r_{SP} = 1 - \frac{6 \sum d_i^2}{(n^2 - 1)n} \,,$$

where d_i is the difference in ranks between the ith pair of data values.

Association between categorical variables

- Suppose two variables X and Y with observed tuples $(x_1, y_1), \dots, (x_n, y_n)$ are given.

- The k ($k \leq n$) different characteristics of X are denoted by $a_1, \dots, a_k$. The m ($m \leq n$) different characteristics of Y are denoted by $b_1, \dots, b_m$.

	b_1	$\dots$	b_m	$\sum$
a_1	n_{11}	$\dots$	n_{1m}	$n_{1.}$
a_2	n_{21}	$\dots$	n_{2m}	$n_{2.}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	n_{k1}	$\dots$	n_{km}	$n_{k.}$
$\sum$	$n_{.1}$	$\dots$	$n_{.m}$	n

Table: ($k \times m$)-contingency table of absolute frequencies for two categorical variables X and Y.

Association between categorical variables II

- The conditional frequency distribution of Y given $X = a_i$, $Y|X = a_i$, is

$$f_Y(b_1|a_i) = \frac{n_{i1}}{n_{i.}}, \ldots, f_Y(b_m|a_i) = \frac{n_{im}}{n_{i.}} \quad .$$

- The conditional frequency distribution of X given $Y = b_j$, $X|Y = b_j$, is

$$f_X(a_1|b_j) = \frac{n_{1j}}{n_{.j}}, \ldots, f_X(a_k|b_j) = \frac{n_{kj}}{n_{.j}} \quad .$$

- Postulate of empirical independence:

$$\frac{\tilde{n}_{ij}}{n_{i.}} = \frac{n_{.j}}{n} \Rightarrow \tilde{n}_{ij} = \frac{n_{i.}n_{.j}}{n} \quad ,$$

where $\tilde{n}_{ij}$ is the absolute frequency one would expect under the assumption of no association between X and Y.

Association between categorical variables III

- Association measure:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} \ , \qquad \chi^2 \in [0, \infty) \ .$$

- Contingency coefficient:

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}} \ ,$$

which can take values between 0 and $K_{max} = \sqrt{(M-1)/M}$ with $M = \min\{k, m\}$.

- The adjusted contingency coefficient is

$$K^* = \frac{K}{K_{max}} \ , \qquad K^* \in [0, 1] \ .$$