Statistical Geophysics

Chapter 2

Descriptive Statistics

Descriptive Statistics

Setting the scene

Background

- Observing systems and computer models in geophysical sciences produce torrents of numerical data.
- One important application of statistical ideas is in making sense of a set of data.
- The goal is to extract insights about the processes underlying the generation of the numbers.
- Descriptive statistics is the discipline of quantitatively describing the main features of a collection of data (sample).
- More recently, a collection of summarisation techniques has been formulated under the heading of exploratory data analysis.

Elementary unit and population

Definition: Elementary unit

- Objects for which a statistical analysis is desired
- Symbol: ω

Definition: Population

- Aggregation of all elementary units defines a population
- Symbol: Ω
- $\omega_i \in \Omega$, $i = 1, \ldots, N$
- N is the size of the population

Elementary unit and population

Example: Households in Germany

- ω_i: a household in Germany
- Ω: all households in Germany
- Population size N: about 40.1 million (as of 2008)

Example: Fish in a lake

- ω_i: a fish in a lake
- Ω: all fish in a lake
- Population size: ?

Sample

Definition: Sample

- A sample is a subset of the elementary units, drawn from the population by means of a sampling method (e.g. random sample).

- Sampling theory is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population.

- Sample size: n ($n < N$)

- Statistical analysis of the sample allows us to draw conclusions about the population of interest (inferential statistics)

Variable and values of a variable

Definition: Variable or statistical variable

Properties, characteristics or attributes of an elementary unit

Definition: Variable values

The *different values* a variable can take. The values can be

- qualitative: variable values are not numbers, but may be coded by numerical values. Such variables are often called *categorical*.

- quantitative: variable values are numbers (numerical values)
  - discrete: finite or countable set of different values
  - continuous: uncountable set of different values
  - quasi-continuous: data are continuous but measured in a discrete way

Variable and values of a variable

Examples

- Gender: qualitative. Coding: 1=male, 2=female
- Hair colour: qualitative. Coding: 1=red, 2=brown, et cetera
- Temperature: quantitative, (quasi-)continuous
- Number of car accidents in 2012 in Germany: quantitative, discrete
- School grades: qualitative. Values: 1,2,3,4,5,6

Level of measurements

The level at which a variable is measured determines

- the choice of numerical summary measures to describe the main features of the data,
- what kind of graphical representations are useful for exploratory data analysis,
- which methods of statistical inference can be applied.

Measurement scales

Definition: Nominal scale

- Lowest level, *unordered set* of values
- Relation or operation: counting values, equality ($=$)
- Units cannot be ordered according to nominal values
- No arithmetic operations (addition, substraction, ratio) possible

Definition: Ordinal scale

- *Ordered set* of values
- Relation or operation: counting values, order ($<$)
- Units can be ordered according to ordinal values
- No arithmetic operations (addition, substraction, ratio) possible

Measurement scales

Definition: Metric scale

Interval scale

- All features of ordinal scale
- Differences of values are meaningful
- Zero value arbitrary

Ratio scale

- All features of interval scale
- Ratios of values are meaningful
- Zero value not arbitrary

Measurement scales

Examples: nominal scale

- Hair colour
- Gender

Examples: ordinal scale

- How often in a week do you eat carrots?
  Possible answers: 0 – 1 – 2 – 3 – more than 3 times
- School grades

Examples: metric scale

- Temperature in degrees Celsius (Fahrenheit): interval scale
- Temperature in degrees Kelvin: ratio scale
- Monthly income of a household: ratio scale

Descriptive Statistics

Frequency distributions

Absolute frequencies

- Let X be the variable of interest and suppose a sample of size n is given with observed values $x_1, x_2, \ldots, x_n$.
- Count the number of k *different* variable values ($k \leq n$): a_j ($j = 1, \ldots, k$).
- For each j ($j = 1, \ldots, k$): count the number n_j of elementary units with variable value a_j ($\sum_{j=1}^{k} n_j = n$).
- Frequency table of a_j and n_j for $j = 1, \ldots, k$.
- Graphical display: Bar chart. The x-axis gives the variable values a_j (ordered if scale is at least ordinal), the bars on the y-axis have length proportional to n_j.

Absolute frequencies: Example

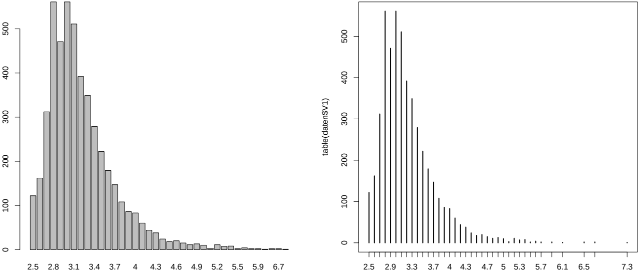


Figure: Earthquake magnitudes in South Carolina, 1987-1996 ($n = 4843$).

Absolute frequencies: Example II

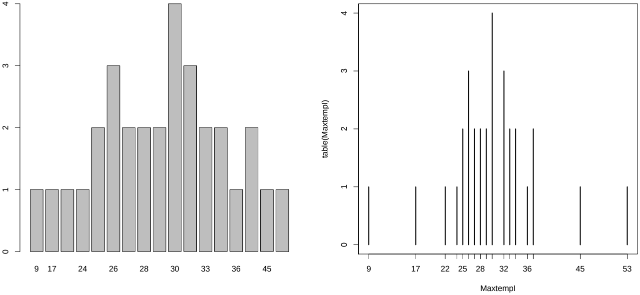


Figure: January 1987 Ithaca maximum temperature data ($n = 31$).

Relative frequencies

- Given the absolute frequencies divide each n_j by the sample size n: $f_j = n_j/n$ for $j = 1, \ldots, k$ ($\sum_{j=1}^{k} f_j = 1$).

- Frequency table of a_j, n_j and f_j for $j = 1, \ldots, k$.

- Graphical display: Bar chart. The x-axis gives the variable values a_j (ordered if scale is at least ordinal), the bars on the y-axis have length proportional to f_j.

Relative frequencies: Example

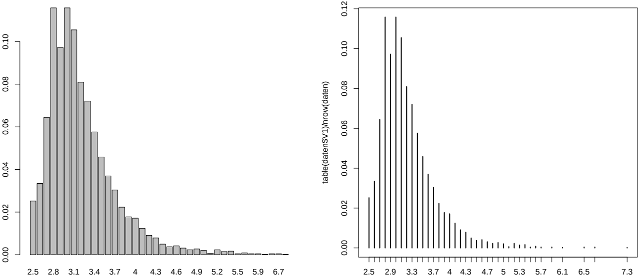


Figure: Earthquake magnitudes in South Carolina, 1987-1996 ($n = 4843$).

Relative frequencies: Example II

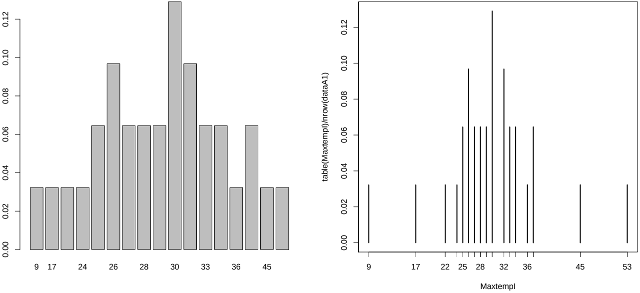


Figure: January 1987 Ithaca maximum temperature data ($n = 31$).

Metric variables

- Bar charts are not useful if $k \approx n$.
- If $k \approx n$ it may be worth defining classes or intervals.
- Count how many values fall within the range of each interval.
- Example: $[72, 86]$, $(86, 100]$, $(100, 114]$, $(114, 128]$.
- Graphical displays:
  1. Histogram or
  2. Kernel density estimate ('smooth histogram')

Histograms

- The range of the data is divided into class intervals or bins.
- The number of values falling into each interval is counted.
- The histogram consists of a series of rectangles whose
  - widths are defined by the class limits implied by the bin width, and whose
  - height depend on the number of values in each bin.
- Usually the widths of the bins are chosen to be equal. In this case the heights of the histogram bars are proportional to the number of counts (absolute or relative frequencies).
- If the histogram bins are chosen to have unequal widths, it is the areas of the histogram bars that are proportional to the number of counts.

Histogram: Example

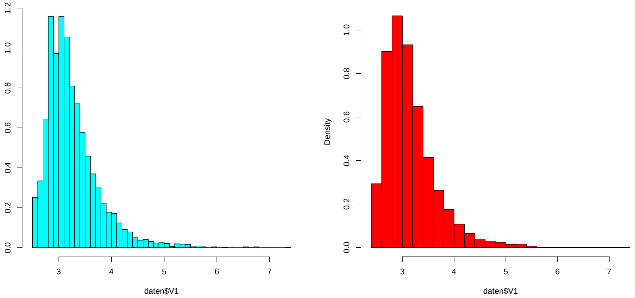


Figure: Histograms of the earthquake magnitudes in South Carolina, 1987-1996.

Histogram: Example II

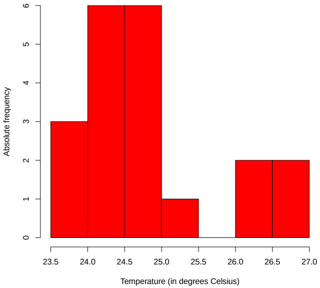


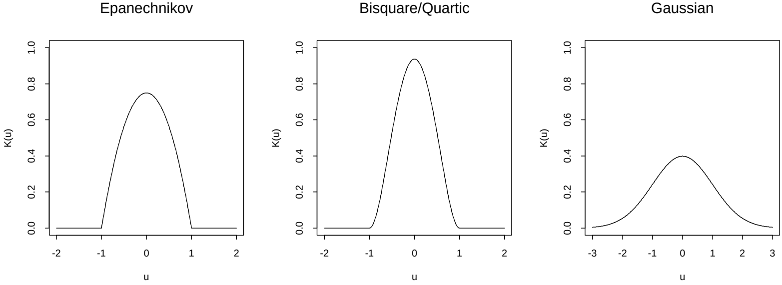
Figure: Histogram of the June temperature data in Guayaquil, Ecuador (1951-1970).

Kernel density smoothing

- An alternative to the histogram that produces a smooth result, is kernel density smoothing.
- It produces the kernel density estimate, which is a nonparametric alternative to the fitting of a parametric pdf.
- It is easiest to understand kernel density smoothing as an extension to histograms.
- Characteristic shapes (kernels) are used that are generally smoother than rectangles.
- A kernel is a non-negative, real-valued, integrable function K satisfying $\int_{-\infty}^{+\infty} K(u)\mathrm{d}u = 1$ and $K(u) = K(-u)$.

Some commonly used kernels

- Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2)$ for $-1 < u < 1,\ 0$ elsewhere
- Bisquare/Quartic: $K(u) = \frac{15}{16}(1 - u^2)^2$ for $-1 < u < 1,\ 0$ elsewhere
- Gaussian: $K(u) = \frac{1}{\sqrt{2\pi}}\ \exp\left(-\frac{1}{2}u^2\right)$ for $u \in \mathbb{R}$



Kernel density estimate

- For data $x_1, \ldots, x_n$, the kernel density estimate of $f(x_0)$ at a given value x_0 is defined as

$$\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right) \quad .$$

- $f(x_0)$ is meant to be the true, unknown population density of X at x_0.

- The bandwidth parameter $h > 0$ controls the amount of smoothness of the kernel density estimate.

Kernel density smoothing: Example

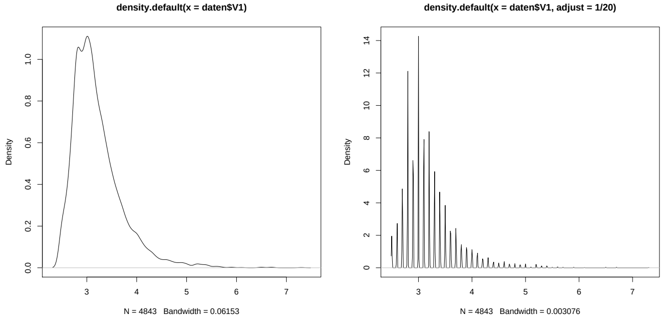


Figure: Kernel density estimates for the earthquake magnitudes in South Carolina, 1987-1996.

Kernel density smoothing: Example II

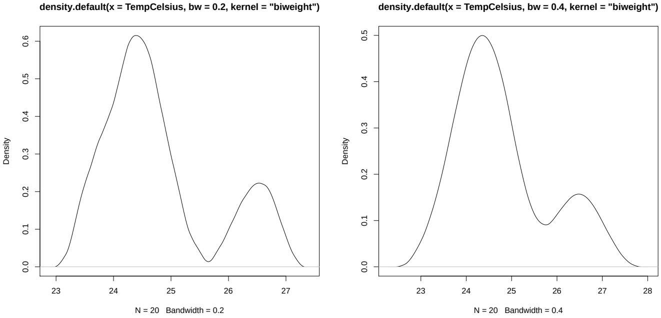


Figure: Kernel density estimates for the June temperature data in Guayaquil, Ecuador (1951-1970) for two different choices of *h*.

Empirical cumulative distribution function (ECDF)

- Sort the different observed values in ascending order:

$$a_{(1)} < a_{(2)} < \cdots < a_{(k)}$$

- Compute relative frequencies $f_{a_{(j)}}$ $(j = 1, \ldots, k)$.
- Compute cumulative relative frequencies:

$$f_{a_{(1)}}, f_{a_{(1)}} + f_{a_{(2)}}, \ldots, f_{a_{(1)}} + f_{a_{(2)}} + \cdots + f_{a_{(k)}}$$

- The ECDF is the step function defined as

$$F_n(x) = \sum_{a_{(j)} \leq x} f_{a_{(j)}}$$

ECDF: Example

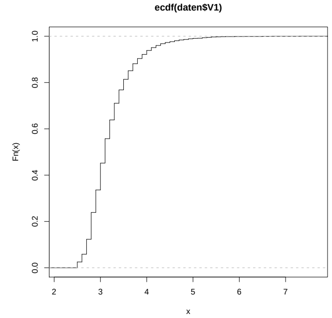


Figure: ECDF for the earthquake magnitudes in South Carolina, 1987-1996 ($n = 4843$).

ECDF: Example II

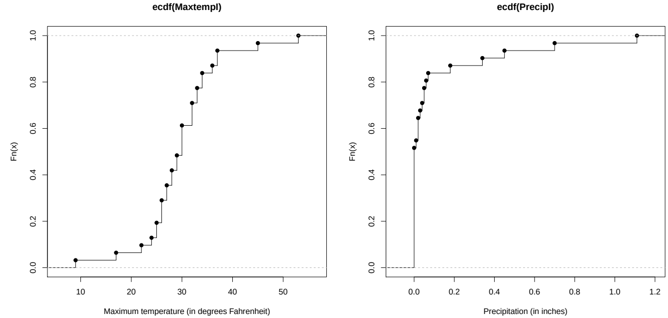


Figure: ECDF for the January 1987 Ithaca maximum temperatures (left) and precipitation data ($n = 31$).

Stem-and-leaf display

- A stem-and-leaf plot provides the analyst with an initial exposure to the individual data values.
- In its simplest form, the stem-and-leaf display groups the data values according to their all-but-least significant digits.
- These values are written in either ascending or descending order to the left of a vertical bar, constituting the “stems”.
- The least significant digit for each data value is then written to the right of the vertical bar, on the same line as the more significant digits with which it belongs. These least significant values constitute the “leaves”.

Stem-and-leaf display: Example

```
 The decimal point is 1 digit(s) to the right of the |
  0 | 9
  1 |
  1 | 7
  2 | 24
  2 | 55666778899
  3 | 00002223344
  3 | 677
  4 |
  4 | 5
  5 | 3
```

Stem-and-leaf plot for the January 1987 Ithaca maximum temperatures. Separate stems are used for least-significant digits from 0 to 4 and from 5 to 9.

Stem-and-leaf display: Example II

```
The decimal point is 1 digit(s) to the left of the |
   0 | 0000000000000001222345567
   1 | 8
   2 |
   3 | 4
   4 | 5
   5 |
   6 |
   7 | 0
   8 |
   9 |
  10 |
  11 | 1
```

Stem-and-leaf plot for the January 1987 Ithaca precipitation data.