Introduction to Geostatistics

2. Variable types, descriptive statistics

Edzer J. Pebesma

edzer.pebesma@uni-muenster.de
Institute for Geoinformatics (**ifgi**)
University of Münster

summer semester 2007/8,
April 14, 2008

Types of observation variables

We can distinguish four types: Nominal, Ordinal, Interval and Ratio variables.

- Nominal: can be named
- Ordinal: can be ordered (and named)
- Interval: can be subtracted (and ordered and named)
- Ratio: can be divided (and subtracted, ordered and named)

Nominal variables

Nominal variables

- can only be *separated*, but not (uniquely) ranked
- can be coded as numbers (0,1,2,...), but many numerical operations do not make sense


```
> soil.char = c("Sand", "Sand", "Clay", "Sand", "Peat")
> soil.f = factor(soil.char)
> soil.f

[1] Sand Sand Clay Sand Peat
Levels: Clay Peat Sand

> table(soil.f)

soil.f
Clay Peat Sand
   1    1    3

> as.numeric(soil.f)

[1] 3 3 1 3 2

> table(soil.f)/length(soil.f)

soil.f
Clay Peat Sand
 0.2  0.2  0.6
```

Nominal variables

Descriptive statistics: frequencies, proportions.

Nominal variables

Binomial variables: yes/no, TRUE/FALSE, 1/0.
Every nominal variable with p classes can be encoded in a set of $p - 1$ binomial variables

soil	IsSand?	IsClay?
Sand	TRUE	FALSE
Sand	TRUE	FALSE
Clay	FALSE	TRUE
Peat	FALSE	FALSE
Clay	FALSE	TRUE

measures of (central) tendency

Organized by variable kind:

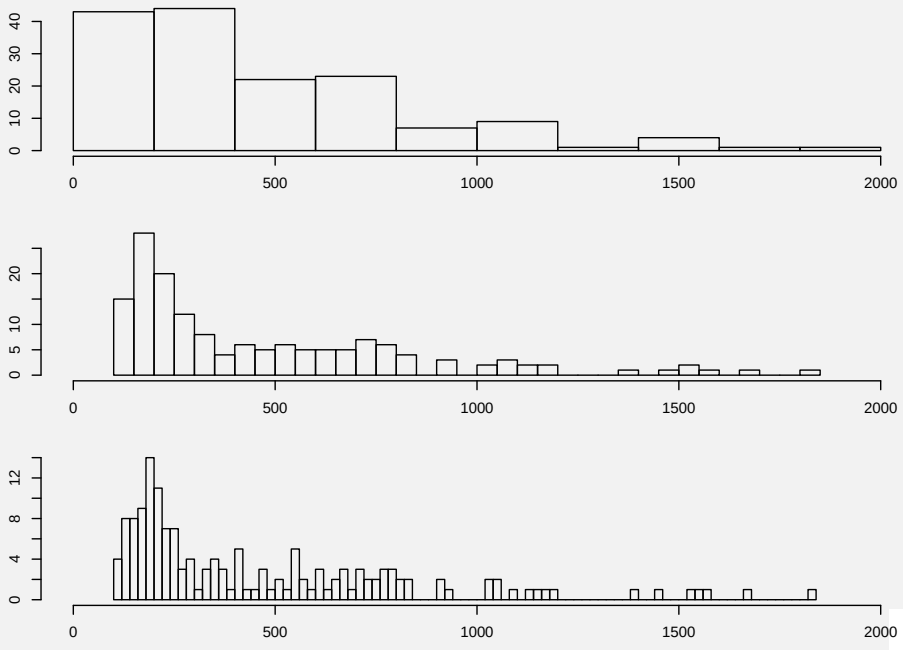
- Nominal: measure of tendency is the mode, the class with the largest frequency, or dominant class
- Ordinal: central tendency: the median, the value above (below) which 50% of the data lie
- Interval, ratio: the mean, sum of the observations divided by n

Organized by measure:

- Mean: relevant for interval/ratio, disputable for ordinal
- Median: relevant for interval/ratio and ordinal
- Mode: relevant for data for which frequencies make sense

ifgi

Mode for a continuous distributions?



ifgi

Quantiles, percentiles, fractions

Quantiles or percentiles generalize the idea of the median for ordinal, interval or ratio data. If the median is the value *below* which 50% of the data lie, the p-percentile is the value *below* which p % of the data lie. The q-quantile is the value *below* which a fraction of q lies. They are expressed in units of measurement. Fractions invert this reasoning. Given a threshold, we can find the number (frequency), or fraction of observations below this value. These are unitless.

```
> load("students.RData")
> attach(students)
> quantile(Length, c(0.25, 0.75))
```

```
25% 75%
170 185
```

```
> mean(Length < 180)
```

```
[1] 0.4651163
```

Measures of spread

The first statistic one usualy considers is a measure of central tendency, as a *typical value*. The second one is a measure of *spread*, or variability.
For ordinal variables, this can e.g. be the *range* (min, max), or the inter-quartile range:

```
> quantile(Length, c(0, 1))
```

```
  0% 100%
 150  220
```

```
> quantile(Length, c(0.25, 0.75))
```

```
25% 75%
170 185
```

For interval/ratio variables, a typical value is the variance, or its square root, the standard deviation.

Mean, variance, standard deviation

Let the n observations be written as x_i, $i = 1, ..., n$. Then, the mean is computed as

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

The variance is then computed

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

and the standard deviation is $s = \sqrt{s^2}$.

```
> var(Length)
[1] 134.5747
> sqrt(var(Length))
[1] 11.60063
```

Why divide by $n - 1$? Consider the case where $n = 1$...

Variability and variance

- Variance, standard deviation, inter-quartile range are *measures* of variability, meaning they can be used to express (measure) variability quantitatively.
- Variability itself is the generic *concept* of something that varies, and is non-quantitative.