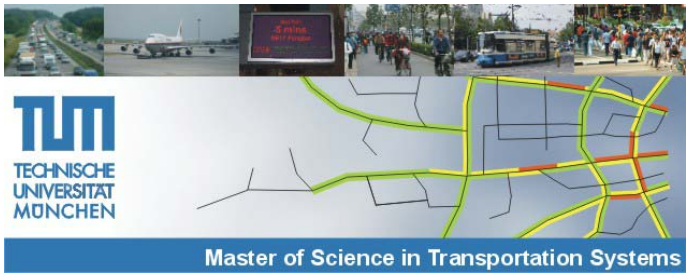
M.Sc. in ‚Transportation Systems'



Applied Statistics in Transport
Descriptive Analyses

Regine Gerike
Technische Universität München, mobil.TUM
regine.gerike@tum.de

Munich, 08/11/2011, 15/11/2011

Plan for Today's Lecture: Descriptive Statistics

- Measures of central tendency (Mean, Median, Midrange, Mode)
- Measures of Dispersion (Range, Mean Absolute Deviation, Variance and Standard Deviation, Coefficient of Variation)
- Measures of Position (Percentile, Quartiles)
- Measures of Shape (Skewness)
- Exploratory Data Analysis/Grapical Output/Visualisation

Descriptive Measures

Definition:

- Descriptive measure = single number that provides information about a sample.

Types of descriptive measures:

- Measures of central tendency: Where is the middle of the data? What values occurs most often?
- Measures of dispersion: How spread are the values? (wide/small range)
- Measures of position: Which values of data were exceeded by e.g. 95% of the data?
- Measures of shape: Is the distribution of values in the sample symmetric?, If it is not symmetric: What is the level of non-symmetricity/unbalance (skewness) in the data?

Measures of Central Tendency: The Arithmetic Mean

The Mean has two related meanings in statistics:

- The mean for a data set (arithmetic, geometric, harmonic mean)
- The expected value of a random variable (called population mean)

- For a real-valued random variable X, the mean is the expectation of X (see distributions).
- For a data set, the mean is the sum of the observations divided by the number of observations. It describes the central location of the data.

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum x_i f_i}{\sum f_i} = \frac{\sum x_i}{n}$$

Measures of Central Tendency: The Geometric Mean, The Harmonic Mean

The Geometric Mean is the nth root of the product of all growth factors x_i:

$$\overline{x_g} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

The Harmonic Mean is the reciprocal value of the arithmetic mean of the reciprocal values of the numbers $x_1,...,x_n$:

$$\overline{x}_h = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum\frac{1}{x}}$$

The Harmonic mean is used when the values to be averaged are ratios of two values, e.g. speed (distance/time), density (mass/weight per volume)

Measures of Central Tendency: The Median

- The median is described as the number separating the higher half of a sample (or a population, probability distribution) from the lower half.
- The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one.
- The median is not unique in case of an even number of observations. It can be determined then by taking the mean of the two middle values.

$$x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[n]}$$

$$m = \begin{cases} x_{\left[\frac{n+1}{2}\right]} \; when \; n = 2k+1 \; (n \; odd) \\ (x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n+1}{2}\right]})/2 \; when \; n = 2k \; (n \; even) \end{cases}$$

Measures of Central Tendency: The Mid-Range

- The mid-range of a set of statistical data values is the arithmetic mean of the maximum and minimum values in a data set:

- The mid-range is highly sensitive to outliers, very non-robust statistic and rarely used in statistical analysis.

$$mr = \frac{x_1 + x_n}{2}$$

Measures of Central Tendency: The Mode

- The mode is the value that occurs the most frequently in a data set.
- The mode may be very different from the mean and the median for strongly skewed data.
- The mode is not necessarily unique.

Descriptive Measures for Different Scales

Scale	Nominal	Ordinal	Cardinal
Measure	Mode	Mode Median	Mode Median Arithmetic Mean

Measures of Dispersion

- Dispersion is the measure of the spread of data values around the mean $\overline{x}$
- Controls the variability of the data
- The measures of dispersion should increase when variability increases

Measures of Dispersion: The Range

- Simplest measure of dispersion, distance between the lowest and the highest value
- The range is strongly influenced by outliers.
- Consider sample x_1, ..., x_n, ordered sample: $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$

$$rn = x_{[n]} - x_{[1]}$$

Measures of Dispersion: Mean Absolute Deviation

- In statistics, the absolute deviation of an element of a data set is the absolute difference between that element and a given point (typically the median or the mean): $D = |x_i - \overline{x}|$
- The average absolute deviation of a data set is the average of the absolute deviations.
- Mean absolute deviation – average absolute deviation related to the mean
- Consider sample $x_1, \ldots, x_n$
- Mean:

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Mean absolute deviation: $$md = \frac{|x_1 - \overline{x}| + \cdots + |x_n - \overline{x}|}{n} = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

Measures of Dispersion: The Variance and the Standard Deviation

- Variance: Average of the squared distances of the elements of the sample from the mean.
- Consider sample $x_1, \ldots, x_n$
- Mean:
$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$
- Sample Variance: $$s^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$
- Standard Deviation:
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

$$s^2 = \frac{1}{n-1}\left[\sum_{1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}\right]$$

Measures of Dispersion: The Coefficient of Variation

- The coefficient of variation (CV) is defined as the ratio of the standard deviation s to the mean $\overline{x}$.
- Consider sample x_1, ..., x_n
- Mean:

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Standard Deviation: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$

- Coefficient of Variation:

$$CV = \frac{s}{\overline{x}} * 100$$

Measures of Position: Percentile / Quartiles

- A percentile is the value of a variable below which a certain percent of observations fall.
- Consider sample $x_1, \ldots, x_n$, ordered sample: $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$

- Percentile: $$x(p) = \begin{cases} \frac{1}{2}\left(x_{[l]} + x_{[l+1]}\right), & where\ l = \frac{p * n}{100} \in \mathbb{N} \\ x_{[k]}, & where\ k = [l] + 1, if\ l \notin \mathbb{N} \end{cases}$$

- Quartiles: Determined in the same was as percentiles for p = 25, 50, 75.

$$Q_1 = x(25) first\ quartile$$
$$Q_2 = x(50)\ second\ quartile$$
$$Q_3 = x(75)\ third\ quartile$$

Measures of Shape: Skewness

- The skewness is a measure of the asymmetry of the data.
- Pearson coefficient of skewness:

$$sk = \frac{3(\overline{x} - m)}{s}$$

Thank you for your attention.

Regine Gerike
Technische Universität München
mobil.TUM
Office: 1753
Tel +49.89.289.28575
regine.gerike@tum.de