Sommerkurs Statistik

A. Hirner

Recommended Reading

- Lecture Notes
- Anderson, D. R., Sweeney, D. J., Williams, T. A. (2008) . Statistics for Business and Economics. Ohio: Cengage Learning.
- Ross, Sheldon (2010) . A first course in probability. New Jersey: Pearson.

I. The Basics

- Creating a Data Matrix
- Scales of Measurement
- Summary Tables
- Rules for Sums and Products

Creating a Data Matrix (1/1)

$n \times k$ raw data matrix

var1	var2	var3	var4
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

Technical Terms

- Rectangular, spreadsheet-like data structure, **n observations** and **k variables**.
- One column corresponds to one **variable (data field)**, one row corresponds to one **observation** (e.g. client, company, product, ...)
- Variables can have different levels of measurement (see later)

Important

This data structure is very common in statistics (Excel, SPSS, SAS, R, Python...). Reshaping your data is usually a time-consuming process. Stick to this data structure right from the beginning and avoid variables that contain text entries!

Scales of Measurement (1/5)

Nominal Scale

The data for a variable consists of labels. Numeric codes for the labels can be used. Calculations (except for counting the values) usually do not make sense at all.

Nominal Scale: Examples

- sex
- hair color
- country of origin
- product category

Scales of Measurement (2/5)

Ordinal Scale

Data can be meaningfully ranked.

Ordinal Scale: Examples

- *Standard & Poor's* rating (AAA, AA, A, BBB, $\cdots$)
- items in consumer research (strongly agree, agree, disagree, strongly disagree)
- satisfaction with a service (excellent, good, poor)

Scales of Measurement (3/5)

Interval Scale

Data can be ranked and the interval between values is expressed in terms of a fixed unit of measure. The zero is chosen arbitrarily and dealing with ratios is not possible.

Interval Scale: Examples

- temperature
- date

Scales of Measurement (4/5)

Ratio Scale

Ratio data is the highest level of data measurements. Ratio scales have an absolute zero.

Ratio Scale: Examples

- distance
- height
- weight

Scales of Measurement (5/5)

Interval Scale and Ratio Scale: What's the Difference?

- If individual A weighs 50 kilos and another individual - say, B - weights 100 kilos, B weighs twice as much as A (ratio scale).
- If the temperature outside today is $+2^{\circ} C$ and yesterday it was $+1^{\circ} C$ it is impossible to say that it is twice as warm today since the zero is arbitrarily chosen (in physics, zero degrees on a celsius scale is defined as the freezing point of water but it could be anything else as well...)

Summary Tables (1/1)

Example:

N=12 companies, X is the number of employees. What does the last column in the summary table tell us? For which level of measurement does it make sense to calculate the last column?

Raw Data

observation	X
1	1
2	1
3	4
4	3
5	1
6	2
7	1
8	1
9	3
10	2
11	1
12	2

Summary Table

x_j	f_j	r_j	cumulative r_j
1	6	50,00 %	50,00 %
2	3	25,00 %	75,00 %
3	2	16,67 %	91,67 %
4	1	8,33 %	100,00 %

Notation

- f_j is the **absolute frequency count of value** x_j and
- r_j the **relative frequency of value** x_j.

Rules for Sums and Products (1/3)

Instead of using uppercase letters (X, Y, Z) we can use double indices

$$x_{i,j} \text{ with } i \in \{1, \cdots, n\} \text{ and } j \in \{1, \cdots, k\}$$

where i is the row index and j the column index.

$n \times k$ raw data matrix

	$X_{.1}$	$X_{.2}$	$\cdots$	$X_{.k}$
observation 1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,k}$
observation 2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,k}$
$\cdots$	.	.	.	.
$\cdots$	.	.	.	.
$\cdots$	.	.	.	.
observation n	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,k}$

We will now introduce a new, handy notation for sums and products.

Rules for Sums and Products (2/3)

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

$$\sum_{i=1}^{n} x_i \cdot y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + \cdots + x_n \cdot y_n$$

$$\sum_{i=1}^{n} \alpha = n \cdot \alpha$$

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

$$\sum_{i=1}^{n} (x_i + \alpha) = (x_1 + \alpha) + (x_2 + \alpha) + \cdots + (x_n + \alpha) = \sum_{i=1}^{n} x_i + n \cdot \alpha$$

$$\sum_{i=1}^{n} c \cdot x_i = c \cdot \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} = \sum_{j=1}^{k} \sum_{i=1}^{n} x_{ij}$$

BUT:

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i \right)^2$$

Rules for Sums and Products (3/3)

Sometimes you will come across the product sign:

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdots x_n$$

$$\prod_{i=1}^{n} \alpha \cdot x_i = \alpha \cdot x_1 \cdot \alpha \cdot x_2 \cdots \alpha \cdot x_n = \alpha^n \cdot \prod_{i=1}^{n} x_i$$

$$\prod_{i=1}^{n} i = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n = n!$$

$$\prod_{i=1}^{n} x_i y_i = \prod_{i=1}^{n} x_i \cdot \prod_{i=1}^{n} y_i$$

Note that this is simply an abbreviation for the multiplication of some values!

II. Descriptive Statistics

- Measures of Central Tendency
- Measures of Dispersion
- Five-Number Summaries and Boxplots
- Linear Transformations

Measures of Central Tendency 1/4

A measure of central tendency is a central or typical value for a probability distribution.

Arithmetic Mean.
The **arithmetic mean (or average)** is defined as the sum of all values divided by the sample size. Note that the mean is sensitive to the presence of outliers. There are three equivalent ways to calculate the mean:

- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (using raw data).
- $\overline{x} = \frac{1}{n}\sum_{j=1}^{m} x_j \cdot f_j$ (using absolute frequencies).
- $\overline{x} = \sum_{j=1}^{m} x_j \cdot r_j$ (using relative frequencies).

Measures of Central Tendency 2/4

Median. The **median** of a set of data is a value that divides the bottom 50% of the data from the top 50%.

- The median is the value in the middle when the data are arranged in ascending order (odd number of observations) or
- the average of the two values in the middle (even number of observations, data arranged in ascending order).

The median is **not** sensitive to the presence of outliers. Sometimes we divide data into four parts, with each part containing 25% of the observations. Q_1 (the first quartile) is the median of the lower 50% of the data points, Q_3 (the third quartile) is the median of the upper 50% of our data. (Q_2 = median).

Measures of Central Tendency 3/4

Mode. The **mode** is the value in a data set that occurs the most often.

- If two or more such values exist, we say the data set is **bimodal** or **multimodal**.
- The mode can also be used to describe the distribution of a nominal variable.

Measures of Central Tendency 4/4

Geometric Mean. The geometric mean is more appropriate than the arithmetic mean for describing proportional growth. The x_i-values are **factors**, not percentages.

$$\overline{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Example: Average annual return.

The price of a certain share went up by 3% in the first year and by 5% in the second year. In the third year it dropped by 4%. Average annual return?

Year 1	+ 3%	1.03
Year 2	+ 5%	1.05
Year 3	- 4%	0.96

$x_G = \sqrt[3]{1.03 \cdot 1.05 \cdot 0.96} \approx 1.012587$.
Average return approximately +1.2587 % per year.

Measures of Dispersion 1/4

Dispersion or variability is the extent to which a distribution is stretched or squeezed.

Range. The range is the difference between the largest and the smallest value.

Measures of Dispersion 2/4

Inter Quartile Range.

$$\mathrm{IQR} = Q_3 - Q_1$$

The inter quartile range is the difference between third and first quartile.

Measures of Dispersion 3/4

Variance.

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

There are three equivalent ways to calculate the variance:

- $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ (by definition).
- $s^2 = \frac{1}{n}\sum_{j=1}^{m} f_j \cdot (x_j - \overline{x})^2$ (using absolute frequencies).
- $s^2 = \sum_{j=1}^{m} r_j \cdot (x_j - \overline{x})^2$ (using relative frequencies).

Measures of Dispersion 4/4

Standard Deviation. The standard deviation is the square root of the variance.

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

Five-Number Summaries and Boxplots (1/1)

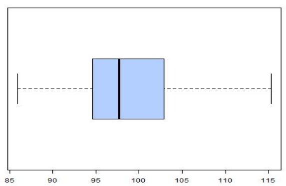
Five-Number Summary

In a so-called *five-number summary* the following five numbers are used to summarize your univariate data:

- smallest value
- first Quartile (Q_1)
- median (Q_2)
- third Quartile (Q_3)
- largest value

Boxplot

A proverb says *a picture is worth a thousand words*. This is why we sometimes use a **boxplot** to visualize these five numbers:



Linear Transformations (1/8)

If X is a random variable (a column in our data matrix) and if α and β are any constants ($\beta \neq 0$), then

$$Y = \alpha + \beta \cdot X$$

is a linear transformation of X and results in a new random variable Y.

Example 1

You collected the height of n=5 subjects in meters. If you want a new variable Y (a new column) - that contains the values in centimeters, each value has to be multiplied by 100.

$$Y = 100 \cdot X$$

This is a linear transformation (with $\alpha = 0$ and $\beta = 100$).

Example 2

You have measured last months' daily temperature in in Celsius and for some reason you want a change in units from Celsius to Fahrenheit. You find the formula

$$F = \frac{9 \cdot C}{5} + 32$$

This is a linear transformation (with $\alpha = 32$ and $\beta = \frac{9}{5}$)

Linear Transformations (2/8): Effects on $\overline{x}$

Effects of Linear Transformations: Arithmetic Mean

Suppose we have a linear transformation of our values in column **X** which results in a new column **Y**. Remember that every value in X is transformed using the formula

$$Y = \alpha + \beta \cdot X$$

How does the mean change?

$$\overline{Y} = \alpha + \beta \cdot \overline{X}$$

The mean undergoes the same transformation as every single data point.

Linear Transformations (3/8): Effects on s_x^2

Effects of Linear Transformations: Variance

Suppose we have a linear transformation of our values in column **X** which results in a new column **Y**. Remember that every value in X is transformed using the formula

$$Y = \alpha + \beta \cdot X$$

How does the variance change?

$$s_y^2 = \beta^2 \cdot s_x^2$$

The variance of the values in the new column Y is the variance of the values in X times the squared constant β. The additive constant α does not have an effect.

Linear Transformations (4/8): Examples

Example 1

You observe the height in cm (**X**) of 12 subjects and get $\overline{x} = 171$ and $s_x^2 = 225$. If you calculate a new column (**Y**) that contains the height in meters, you will get

$\overline{y} =$ ______ and $s_y^2 =$ ______

Example 2

You observe 100 values of a numeric variable (**X**) and get $\overline{x} = 412$ and $s_x^2 = 2500$. Which values for $\overline{y}$ and s_y^2 do you get after applying the transformation $Y = 3 \cdot X + 10$?

$\overline{y} =$ ______ and $s_y^2 =$ ______

Example 3

You observe 100 values of a numeric variable (**X**) and get $\overline{x} = 1000$ and $s_x^2 = 15$. Which values for $\overline{y}$ and s_y^2 do you get after applying the transformation $Y = 1000X + 100$?

$\overline{y} =$ ______ and $s_y^2 =$ ______

Linear Transformations (5/8): Examples (continued)

Example 1

You observe the height in cm (**X**) of 12 subjects and get $\overline{x} = 171$ and $s_x^2 = 225$. If you calculate a new column (**Y**) that contains the height in meters, you will get

$\overline{y} =$ 1.71 and $s_y^2 =$ 0.0225

Example 2

You observe 100 values of a numeric variable (**X**) and get $\overline{x} = 412$ and $s_x^2 = 2500$. Which values for $\overline{y}$ and s_y^2 do you get after applying the transformation $Y = 3 \cdot X + 10$?

$\overline{y} =$ 1246 and $s_y^2 =$ 22500

Example 3

You observe 100 values of a numeric variable (**X**) and get $\overline{x} = 1000$ and $s_x^2 = 15$. Which values for $\overline{y}$ and s_y^2 do you get after applying the transformation $Y = 1000X + 100$?

$\overline{y} =$ 1000100 and $s_y^2 =$ 15000000

Linear Transformations (6/8): Centering

Centering a variable simply means subtracting
the arithmetic mean ($\overline{x}$) from every value.

Centering.
$$Y = X - \overline{x}$$

Note that this is again a linear transformation (with $\alpha = -\overline{x}$ and $\beta = 1$.)

A centered variable has zero mean. The variance remains unchanged:

- $\overline{Y} = 0$
- $s_Y^2 = s_X^2$

Since we subtracted the mean from every data point, we can now easily see which values are greater than the sample mean (positive values of Y) and which are smaller than the sample mean (negative values of Y)!

Linear Transformations (7/8): z-Scores

Sometimes we are also interested in the relative location of our values within a data set. By using both the mean and standard deviation, we can determine the relative location of any observation and compare the values even across different groups. Note that the **z-score** is often called the **standardized value**. In other words, the standard score is the signed number of standard deviations by which the value of an observation or data point differs from the mean value.

Finding a z-Score

$$z = \frac{x - \overline{x}}{s_x}$$

Where...

- x is the value of interest
- $\overline{x}$ is the sample mean and
- s_x is the standard deviation.

We will make use of this transformation more often later!

Linear Transformations (8/8): z-Scores (continued)

Application: Comparisons

Suppose you have two different high school tests, namely the SAT and ACT that measure the same ability. SAT results have a **mean of 1500 points** and a **standard deviation of 300 points** whereas ACT results have a **mean of 21 points** and a **standard deviation of 5 points**. Suppose that **Alice** scored 1800 on the SAT, and **Bob** scored 24 on the ACT. Who performed better?

Solution: Comparisons

At first we calculate Alice's z-score:

$$z_A = \frac{1800 - 1500}{300} = 1.$$

Bob's z-score is

$$z_B = \frac{24 - 21}{5} = 0.6$$

Since

$$z_A > z_B,$$

Alice's performance is better than Bob's.