L3

Basic Statistics for Economists

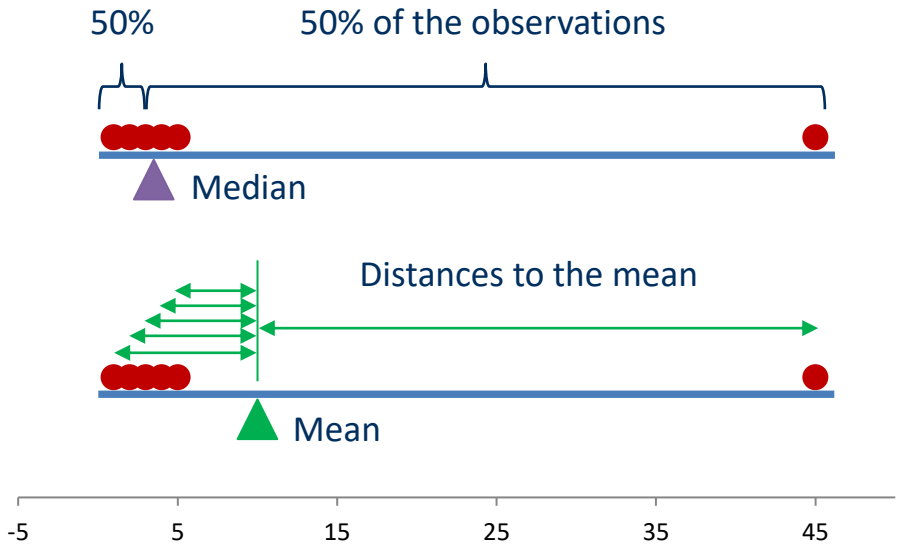
Spring 2020

Department of Statistics

Stockholm University

Short note on means and medians

- Dataset: $(\mathbf{1},\ \mathbf{2},\ \mathbf{3},\ \mathbf{4},\ \mathbf{5},\ \mathbf{45})$
- Median: $\boldsymbol{md} = \mathbf{3.5}$ Mean: $\bar{\boldsymbol{x}} = \mathbf{10}$



Equilibrium in **frequencies**

$50\% - 50\% = 0$

Equilibrium in **distances**

$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Stockholm University

Today – two things

- **Descriptive statistics for two or more variables**
  - Tables, frequency- and magnitude tables
  - Graphical, diagrams
  - Covariation, measuring relationship between two variables
- **Introduction to Probability Theory**
  - Short intro about Models
  - Random experiments (sv. *slumpförsök*)
  - Sample space, outcome space (*utfallsrum*)
  - Events, outcomes (*händelser*)
  - A little **Set Theory** (*mängdlära*) and Venn diagrams
  - Definitions and interpretations of probability
  - Intro to **Combinatorics**

Stockholm University

More than one variable

Why study more than one variable at the same time?

- **Analyzing relationships**
  - Is there evidence in the data to support the idea that different variables/phenomena affect each other?
  - Can we utilize the relationship between variables to predict the value of one variable given the others?

- Study **cross tabulations** and different **graphs/diagrams**
- Also, numerical measures of relationship
  - **covariances** and **correlations**

Stockholm University

Frequency tables

- ***Cross tabulations***, ”crossing” two or more variables
- Absolute frequencies or relative frequencies (%)
- However, … now we can choose:

  - ***simultaneous*** or ***joint*** relative frequencies

  - ***conditional*** relative frequencies (sv. *betingade*)

- With conditional relative frequencies it is often easier to do comparisons between groups and discover relation-ships and dependencies, e.g. if two groups differ in some way

Stockholm University

Two variables in one table

Joint frequency distribution

No. of employees per income class and department		Department			Total
		A	B	C	
Income	0 – 500’	18	4	2	24
	500’ –	2	1	-	3
	Total	20	5	2	27

Bivariate distribution

The marginals show the univariate distributions

Joint relative frequency distribution

% employees per income class and department		Department			Total
		A	B	C	
Income	0 – 500’	67 %	15 %	7 %	89 %
	500’ –	7 %	4 %	-	11 %
	Total	74 %	19 %	7 %	100 %

Joint distribution – every combination dep./income

Stockholm University

Conditional distributions (sv. *betingade*)

Rows

Distribution over department per income class

		Department			Total
		A	B	C	
Income	0 – 500’	75 %	17 %	8 %	100 %
	500’ –	67 %	33 %	-	100 %
	Total	74 %	19 %	7 %	100 %

Conditioning on income class

Columns

Distribution over income class per department

		Department			Total
		A	B	C	
Income	0 – 500’	90 %	80 %	100 %	89 %
	500’ –	10 %	20 %	-	11 %
	Total	100 %	100 %	100 %	100 %

Conditioning on department

Stockholm University

Combining categorical and numerical

- **Magnitude tables** (SCB)
- Cross two (or more) categorical ***background variables***
- In each cell/combination a third numerical ***response variable***
- The response variable can be displayed as a mean, a total, proportion etc. …

Stockholm University

Example: Magnitude tables

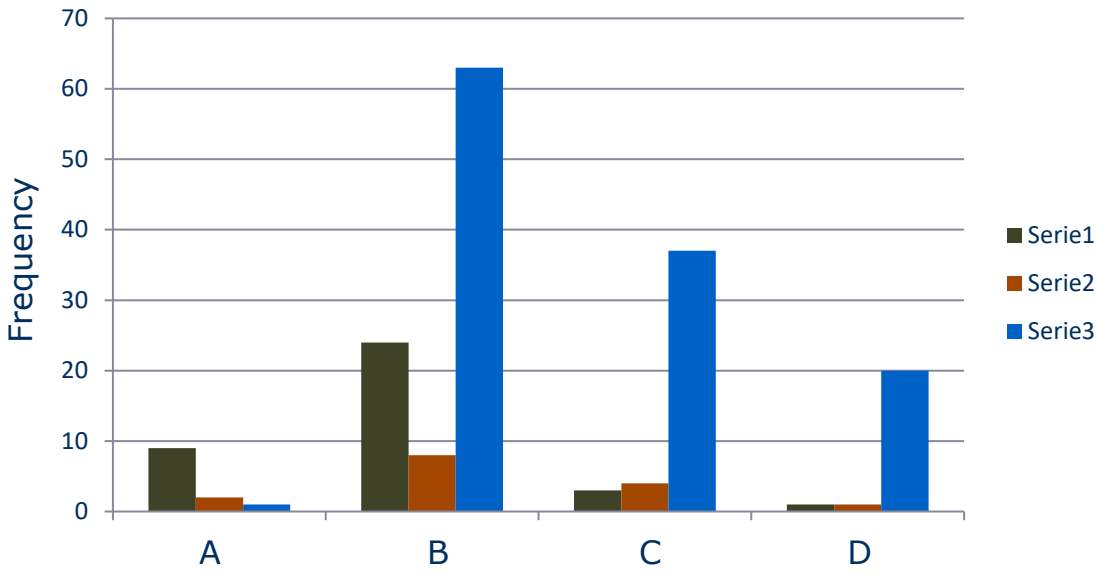
Rows for categorical variable 1, columns for categorical variable 2 and an ***aggregate*** (e.g. sum) of a **third** numerical variable in each cell/combination

Net turnover, million SEK		NACE Code			Total
		1	2	3	
No. of employees	0 – 19	64 209	12 518	4 414	81 141
	20 – 49	3 071	11 679	2 257	17 007
	50 – 99	1 391	9 677	1 994	13 062
	100 – 249	4 967	14 022	2 941	21 930
	250 –	26 916	83 009	2 999	112 924
	Total	100 554	130 905	14 605	246 064

Stockholm University

Graphical presentation multivariate data

- Grouped bars – absolute frequencies per group



Here, two categorical variables. Graph shows the frequency of each combination.

What about relative frequencies instead?

Categorical variable – nominal or ***ordinal***
Discrete numerical variable – ratio or ***interval***

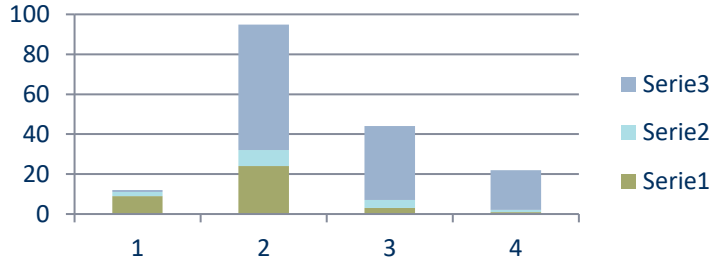
Stockholm University

Multivariate data, cont.

- Stacked bars (categorical data or discrete numerical)

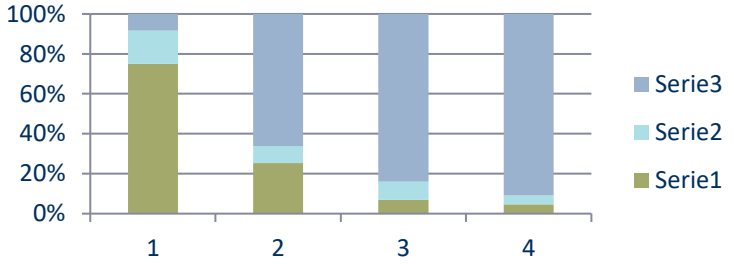
Absolute

frequencies per group (joint distribution)



Relative

frequencies per group **conditional** on groups

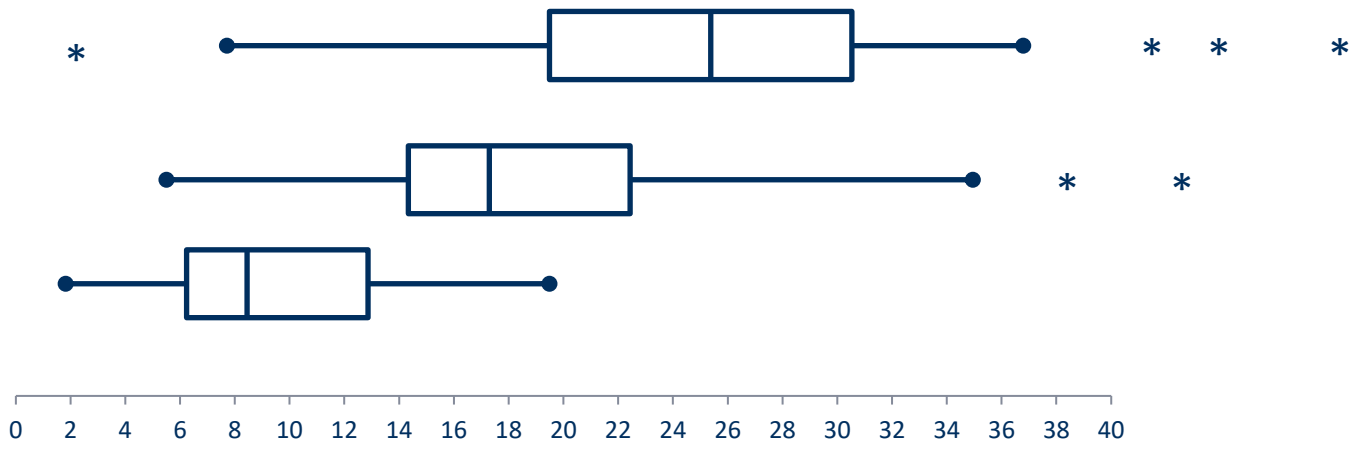


Categorical variable – nominal or ***ordinal***
Discrete numerical variable – ratio or ***interval***

Stockholm University

Numerical: comparing distributions

- Box plots again
- Three different groups



Continuous (or discrete) numerical variable – ratio or ***interval***

Stockholm University

Measuring relationship: Covariance

- Two numerical variables $\boldsymbol{x}$ and $\boldsymbol{y}$
- **Sample covariance** for $\boldsymbol{x}$ and $\boldsymbol{y}$:

Compare with the variance formulas

$$\boldsymbol{Cov(x,y)} = \boldsymbol{s_{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

- Covariance = measure of covariation, ***linear*** relationships
- May be negative or positive or zero
- Not entirely easy to interpret the value one gets
- Q: How do we calculate the **population covariance** σ_{xy}?

Stockholm University

Correlation

- Measure of the ***linear*** relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$
- Sample correlation coefficient for $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\boldsymbol{Corr(x,y) = r_{xy} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{S_{xy}}{\sqrt{S_x^2 \cdot S_y^2}}}$$

- Population correlation: $\boldsymbol{\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}}$ *("rho", or sv. "rhå")*
- Actually: *product moment correlation coefficient*
- **Important! Correlation is always in the interval [-1,1]**

Stockholm University

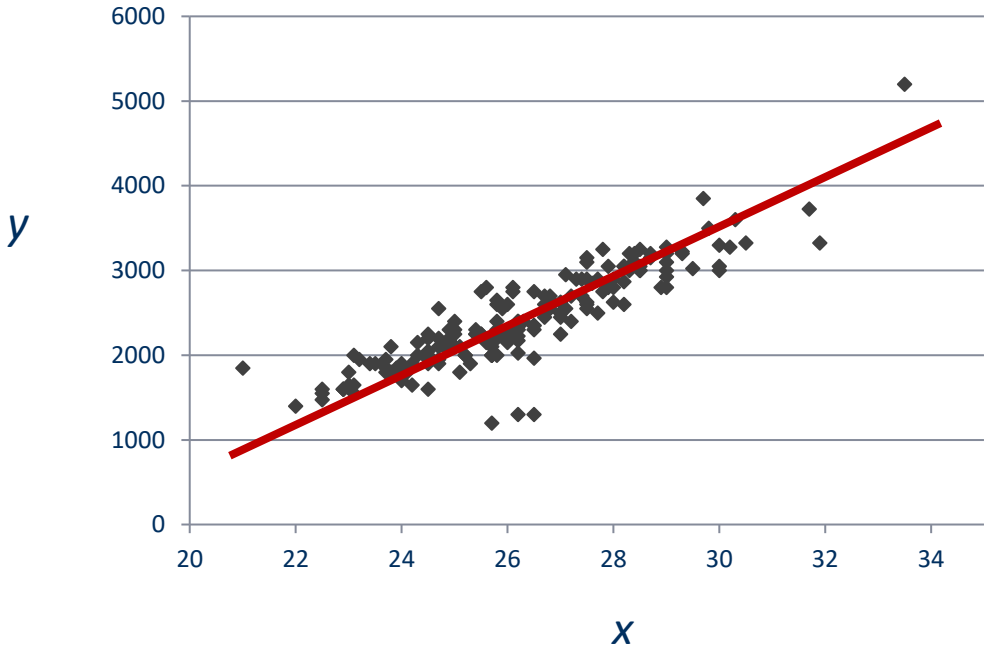
Covariances and correlations

- ***Covariance*** can in principal take on any value (-∞,∞)
  - Large absolute values suggest strong ***linear*** relationship
  - Scale dependent (e.g. different value if SEK, Euro or USD)
  - Difficult to compare covariances directly
- ***Correlation*** re-scales the covariance to **[-1,1]**
  - Scale independent, easier to compare values directly
  - Absolute values close to 1, i.e. -1 or +1, suggest strong ***linear*** relationships
  - Values close to zero suggest no linear relationship

Stockholm University

Scatter plots

- Sv. *punktplottar, spridningsdiagram*
- Two numerical continuous or discrete variables: $\boldsymbol{x}$ and $\boldsymbol{y}$
- Each pair $(\boldsymbol{x_i}, \boldsymbol{y_i})$, $i = 1, \ldots, n$, is represented by a dot:

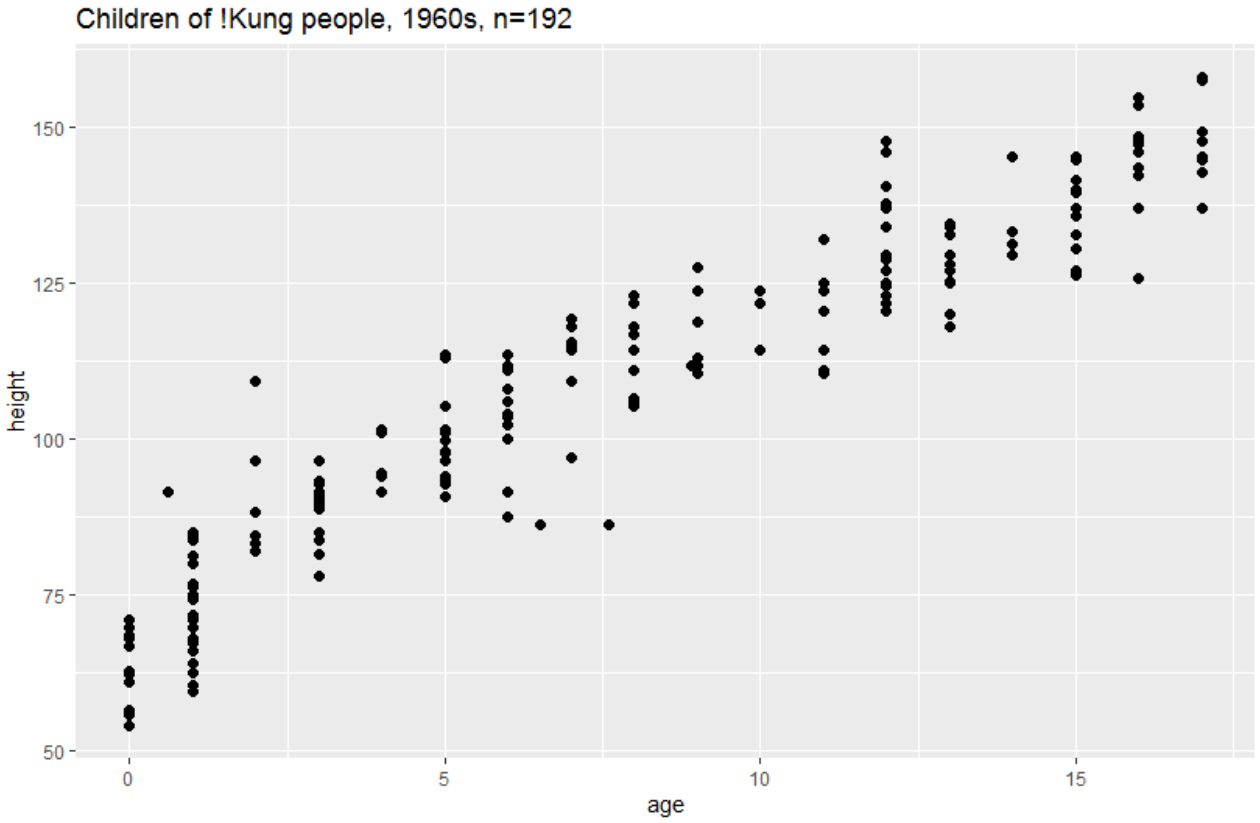


Linear relationship if the imposed line is straight, not curved

If it's discrete numerical, data points (x_i, y_i) are likely to coincide with each other

Stockholm University

Example, height and age among children



$$s_{xy} = 130\ years \cdot cm \qquad r_{xy} = 0.94$$

Stockholm University

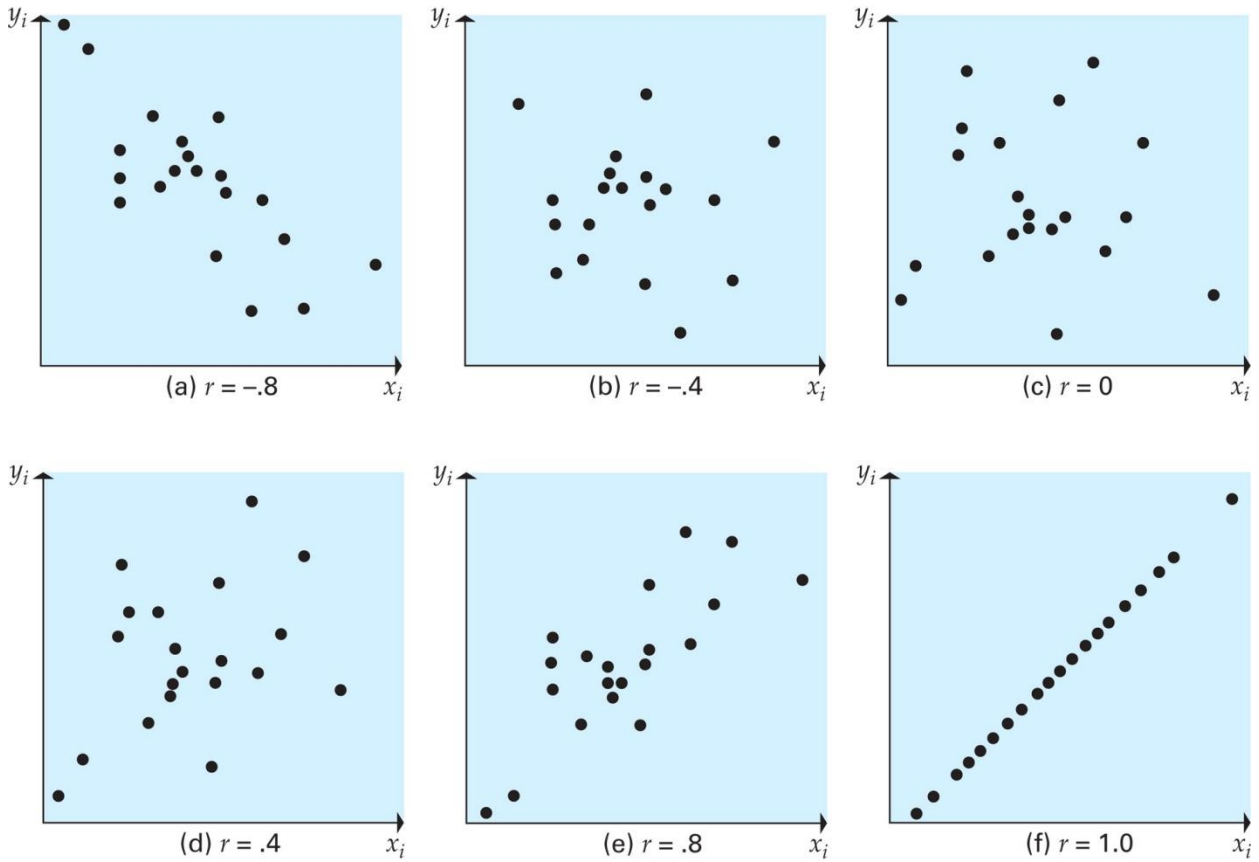
Scatter plots, cont.

What are we looking for?

- ***Strong*** or ***weak*** relationships
- ***Positive*** or ***negative*** relationships
- ***Linear*** or ***non-linear*** relationships
- Also, ***deviant*** and ***extreme*** values/observations

Stockholm University

Strong/weak – positive/negative

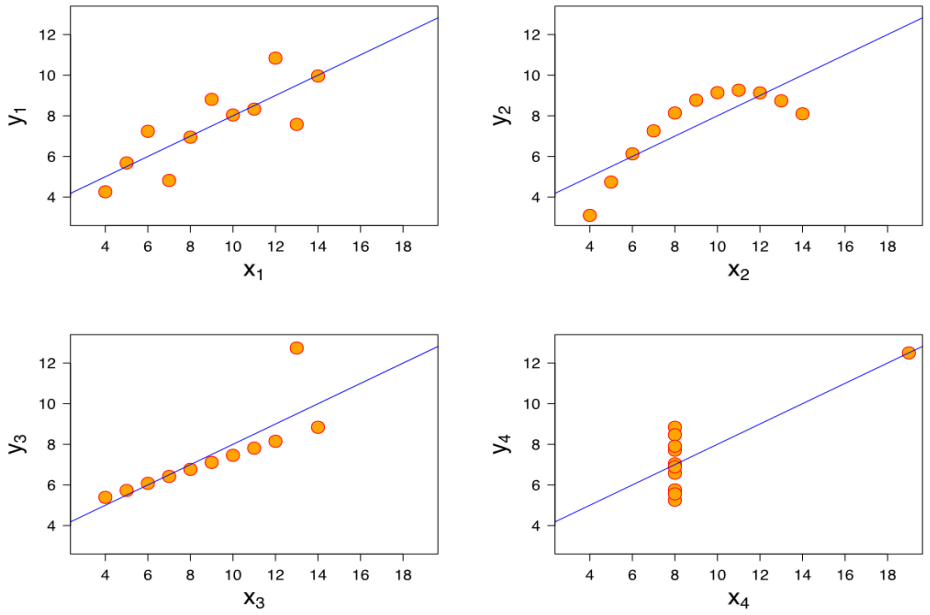


Copyright ©2013 Pearson Education, publishing as Prentice Hall

Stockholm University

Linear/non-linear – deviant points

- Anscomb's data set

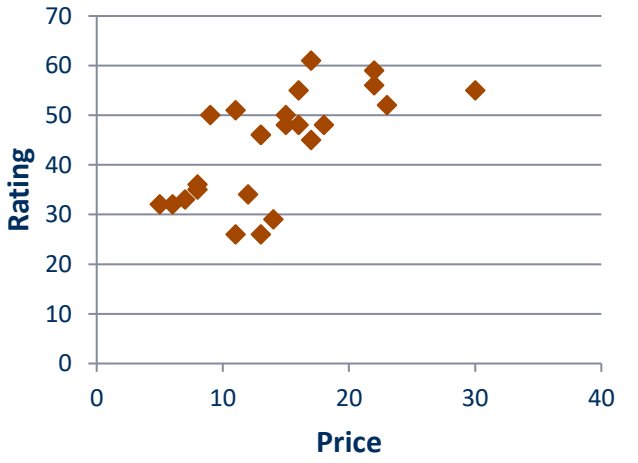


- Four pairs of variables all of which yield the same correlation: **0,816**
- Only one shows truly linear relationship (which?)
- One could be perfectly linear but for one extreme point (which?)

Stockholm University

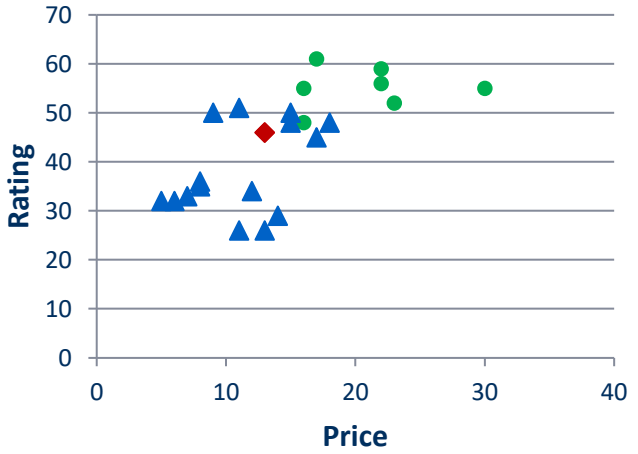
Example

ID	Type	Rating	PricePerLoad
1	Liquid	61	17
2	Liquid	59	22
3	Liquid	56	22
4	Liquid	55	16
5	Liquid	55	30
6	Liquid	52	23
7	Powder	51	11
8	Powder	50	15
9	Powder	50	9
10	Liquid	48	16
11	Powder	48	15
12	Powder	48	18
13	Gel	46	13
14	Gel	46	13
15	Powder	45	17
16	Powder	36	8
17	Powder	35	8
18	Powder	34	12
19	Powder	33	7
20	Powder	32	6
21	Powder	32	5
22	Powder	29	14
23	Powder	26	11
24	Powder	26	13



$s_{xy} = 43,2$

$r_{xy} = 0,67$



$r_{xy} = 0,42$

$r_{xy} = 0,07$

$r_{xy} =$ not def.

Stockholm University

Answers: $\bar{x} = 10$ $s_x^2 = 16$
$\bar{y} = 8$ $s_y^2 = 4$
$s_{xy} = 6$ $r_{xy} = 0{,}75$

DIY Exercise

- Calculate means and variances for *x* and *y* respectively
- Calculate the covariance and correlation between *x* and *y*
- Construct a scatter plot, displaying the data
- What is the relationship? How does it relate to the correlation?

i	1	2	3	4	5	6	7	8	9	10	Σ
x_i	10	6	14	10	12	16	8	14	6	4	
y_i	8	6	7	10	9	10	8	11	5	6	
x_i^2											
y_i^2											
$x_i y_i$											

Stockholm University

Summary

We have reviewed how to present empirical data, both one variable at a time and several together (**univariate** and **multivariate**)

- Tables
- Graphics, types of diagrams, charts
- Summarizing numerical measures
  - Location, variability and now linear correlation (relationship)

Now we will move towards ***models***, mathematical constructs that "explain" how the data was generated, how they relate and so on

- Specifically **probability models**, i.e. theoretical models that we trust will provide an approximate description of reality

Stockholm University

Models

- A model is a **simplified description** of something real based on **assumptions**
  - only relevant aspects are included in the model
  - to test the durability of a material, the color may not matter
  - we approximate (“almost right”)
- We denote the relevant aspects of reality with ***symbols***
  - we turn it into “mathematics”
  - e.g. $\boldsymbol{d = \frac{1}{2} \cdot g \cdot t^2}$
  - $\boldsymbol{d}$, $\boldsymbol{g}$ and $\boldsymbol{t}$ are symbols that represent the values of the different variables, i.e. phenomena in the real world

Stockholm University

Stochastic models – random models

In a **deterministic model**, there is no room for exemption, everything is precisely prescribed in the model (e.g. Newton) and all is perfectly predictable (if the model is "true").

Ex. Ohm's law (physics): $\boldsymbol{R = U/I}$

A **stochastic model** is characterized in that it contains a **random** component. We do not know exactly what the outcome will be but we might be able to say something about what could happen and how often.

Ex. $\boldsymbol{X}$ = "no. of deaths caused by horse kicks to the head"

Stockholm University

Random experiment

- A process or procedure that results in one of at least two possible outcomes (sv. *slumpförsök*)
- However, we cannot be certain which it will be

Sample space, outcome space

- An enumeration, listing or description of all possible **elementary outcomes** (**basic outcomes**) of a random experiment
  - numerical or categorical outcomes?
  - discrete or continuous?
  - limited or unlimited?
- Denoted in NCT by S (capital S)

Stockholm University

Events (sv. *händelser*)

- A ***subset*** of S
- May be a single basic outcome or a collection (set) of basic outcomes or even S itself
- Ex: Rolling a die (dice)
  - Sample space $S = \{1,2,3,4,5,6\}$ where $1,2,\ldots,6$ are basic outcomes
  - Some possible ***events***:
    - ”less than 3” $= \{1,2\}$
    - ”odd number” $= \{1,3,5\}$
    - ”equal to 4” $= \{4\}$
    - ”not 4” $= \{1,2,3,5,6\} = S - \{4\}$
    - ”larger than zero” $= \{1,2,3,4,5,6\} = S$

Stockholm University

Examples

Random experiment = ”toss two dice”

1. Define $\boldsymbol{Y}$ = “sum of the two outcomes”

   Describe the sample space

   Is every possible basic outcome equally probable?

2. Define $(\boldsymbol{X_1}, \boldsymbol{X_2})$ = ”dots on die no. 1 and die no. 2”

   Describe the sample space

   Is every possible basic outcome equally probable?

3. Define $\boldsymbol{Z}$ = “the age of a randomly chosen person in this room”

Stockholm University

Elementary Set Theory

- Let $\boldsymbol{O_1, O_2, \ldots}$ denote the ***basic elements of S (outcomes)***
- Let capital letters $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{S}$, … denote ***sets*** of elements
  - braces { } are used to list the elements in a set
  - ex. $A = \{1,2\}$, $B = \{4\}$, $C = \{O_1, \ldots, O_N\}$
- O_k is a **member of/lies in** A, denoted $\boldsymbol{O_k \in A}$
  - ex. $1 \in \{1,2\}$ but $3 \notin \{1,2\}$
- A is a ***subset*** of B, is denoted $\boldsymbol{A \subset B}$
  - Subset is denoted $\subseteq$, strict subset is denoted $\subset$
  - ex. $\{1,2\} \subset \{1,2,3,4,5,6\}$, $\{1,2\} \subseteq \{1,2,3,4,5,6\}$, $\{1,2\} \subseteq \{1,2\}$
  - ex. $\{5,6\} \not\subset \{1,2,3,4,5\}$

Stockholm University

Elementary Set Theory, cont.

Assume that $S = \{1,2,3,4,5,6\}$ and that $A = \{1,2\}$, $B = \{2,3,4\}$ and $C = \{3\}$

- ***Complement*** is all that is not a member of the set and is denoted $\overline{A}$, (alternative notations include A', A^c, A^*, $\complement A$, not A)
  - ex. $\bar{A} = \{3,4,5,6\}$

- ***Union***, denoted ∪: all members in A or B or both A and B
  - ex. $A \cup B = \{1,2,3,4\}$ $B \cup C = B$

- ***Intersection***, denoted ∩: all which are members in both *A* and *B*
  - ex. $A \cap B = \{2\}$ $B \cap C = C$

Stockholm University

Elementary Set Theory, cont.

- The ***empty set*** is the set containing nothing, denoted $\emptyset$.

- Two sets are ***disjoint*** or ***mutually exclusive*** (sv. *disjunkta, oförenliga*) if their intersection is empty, if they have nothing in common, they are ***non-overlapping***

  - ex. $A = \{1,2\}$ and $C = \{3\}$
    $A \cup C = \{1,2,3\}$
    $A \cap C = \emptyset$

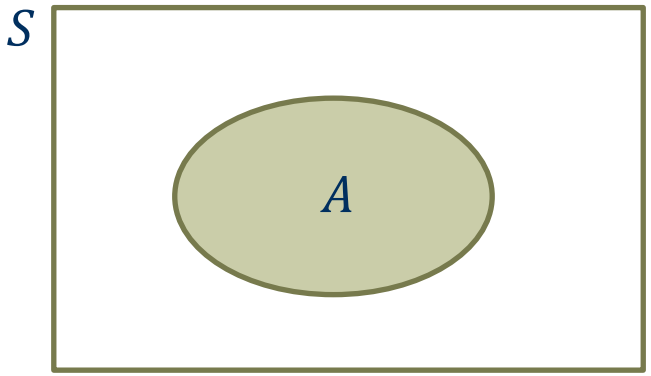
- *What is the complement of S? What is $S \cup \emptyset$? $S \cap \emptyset$?*

Stockholm University

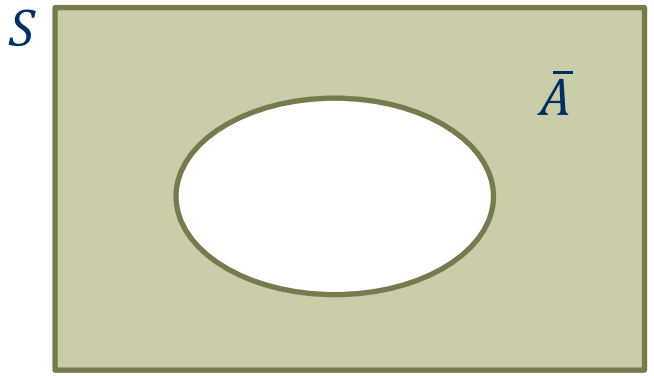
Venn diagrams

- A way to illustrate sets and their relationships to other sets
- The shaded areas denote the set in focus

$A \subseteq S$



$\bar{A} \subseteq S$



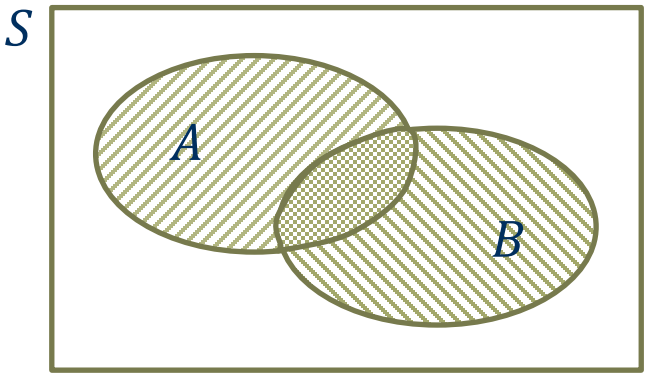
$A \cup \bar{A} = S$ $A \cap \bar{A} = \emptyset$

Stockholm University

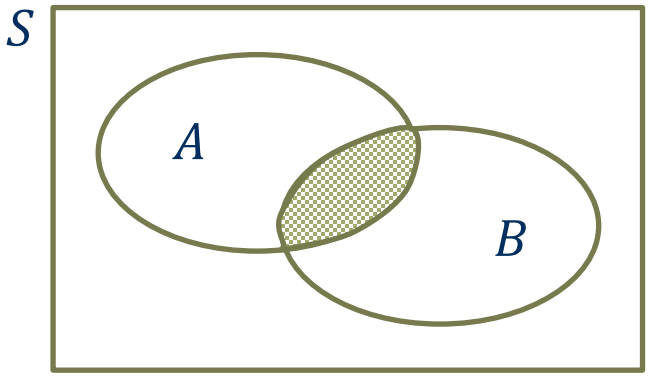
Venn diagrams, cont.

- Two sets, A and B
- The shaded areas denote the set in focus

$A \cup B$



$A \cap B$

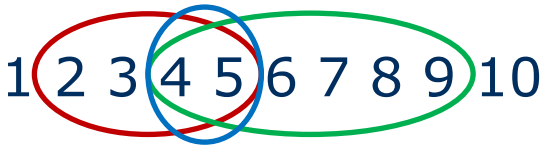


- DIY: $A \cup \bar{B}$, $\bar{A} \cup B$, $\bar{A} \cup \bar{B}$, and $A \cap \bar{B}$, $\bar{A} \cap B$, $\bar{A} \cap \bar{B}$

Stockholm University

Or like this …

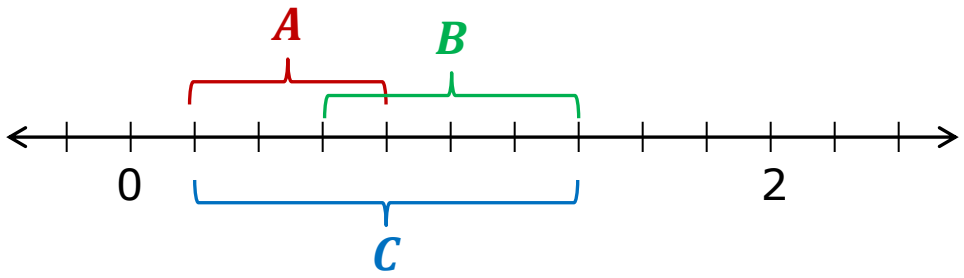
- List all set members and encircle the subsets (discrete sets):



$\boldsymbol{A} = \{2,3,4,5\}$ $\boldsymbol{B} = \{4,5,6,7,8,9\}$

$\boldsymbol{C} = A \cap B = \{4,5\}$

- Line of real numbers, mark subsets/intervals (continuous sets):



$\boldsymbol{A} = (0.2\,,0.8)$

$\boldsymbol{B} = (0.6\,,1{,}4)$

$\boldsymbol{C} = A \cup B = (0.2\,,1.4)$

$A \cap B = (0.6\,,0.8)$

Stockholm University

DIY Exercise – answers on Athena

Let $S = \{1,2,3,4,5,6\}$ and describe the relationship between them:

1. $A = \{1,2\}$ $B = \{3,4,5\}$ Disjoint, non-overlapping $A \cap B = \emptyset$
2. $A = \{1,2,3\}$ $B = \{3,4,5\}$ Overlapping, $A \cap B \neq \emptyset$
3. $A = \{1,2,3,4\}$ $B = \{1,2\}$ B is a strict subset of A, $B \subset A$

For each of the 3 cases above, what is the resulting set of

a) $A \cup B$ c) $\bar{A} \cup B$ e) $\bar{A} \cup \bar{B}$

b) $A \cap B$ d) $\bar{A} \cap B$ f) $\bar{A} \cap \bar{B}$

List the basic outcomes and encircle subsets!

Stockholm University

Probability

Once an ***experiment*** and the ***sample space*** $\boldsymbol{S}$ is defined we also need to know how probable any given event (subset) is.

- Let $\boldsymbol{O_k}$ denote a basic outcome (sv. *elementärt utfall*)
- $\boldsymbol{O_k}$ is in $\boldsymbol{S}$ and we write this as $\boldsymbol{O_k} \in \boldsymbol{S}$

- Let $\boldsymbol{P}(\boldsymbol{O_k})$ denote the ***probability*** that the outcome is $\boldsymbol{O_k}$
- The probability $\boldsymbol{P}(\boldsymbol{O_k})$ is a ***number,*** a numerical measure
- $\boldsymbol{P}$ for *probability*; $\boldsymbol{P}(\boldsymbol{O_k})$ can be viewed as a function

Stockholm University

Probability and event

- Let $\boldsymbol{A}$ be any event, i.e. $\boldsymbol{A} \subseteq \boldsymbol{S}$
  - ex. $A = \{1,2\} \subseteq \{1,2,3,4,5,6\} = S$

- The probability of $\boldsymbol{A}$: $\boldsymbol{P(A)}$ = **sum of** $\boldsymbol{P(O_k)}$ where $\boldsymbol{O_k} \in \boldsymbol{A}$
  - ex. $P(\text{"1 or 2"}) = P(\text{"1"}) + P(\text{"2"})$

Note! ”1” and ”2” are ***disjoint*** basic outcomes

An (almost) complete ***stochastic model*** is defined if

1. The sample space $\boldsymbol{S}$ is well-defined
2. For all $\boldsymbol{A} \subseteq \boldsymbol{S}$ we can provide $\boldsymbol{P(A)}$, including $\boldsymbol{P(S)}$ and $\boldsymbol{P(\emptyset)}$

Stockholm University

But what is a probability?

- Probabilities $P(A)$ are numbers
- Should for all practical reasons represent how "probable" an event is
- Our intuition tells us:
  - if the probability is zero, it should never happen?
  - if the probability is 100%, it should always happen?
- Interpretations of probabilities can vary:

  Classical, ***Frequentist*** or ***Subjectively***

Stockholm University

Classical interpretation

If possible, we can gauge the "size" of a subset $\boldsymbol{A}$ and compare it to the "size" of $\boldsymbol{S}$

- Assume that $\boldsymbol{S}$ can be described as set of basic outcomes $\boldsymbol{O_k}$, ***all of which are equally probable***
- Count the number of elements/members in $\boldsymbol{A}$ (size of $\boldsymbol{A}$)
- Compare to the total number of basic outcomes (size of $\boldsymbol{S}$)

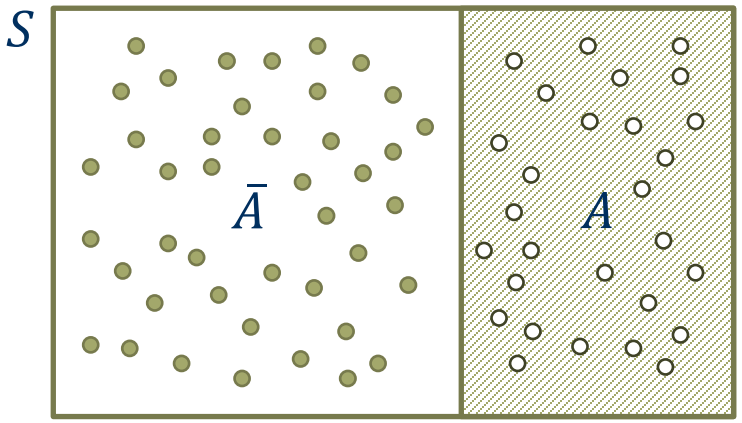
$$\frac{\#(A)}{\#(S)} = \frac{\text{number of members in } A}{\text{number of members in } S} = P(A) \qquad \frac{size(A)}{size(S)} = P(A)$$

Stockholm University

Classical interpretation – example

- 70 lottery tickets of which 28 are wins
- Draw one randomly
- Venn diagram

Every dot represents a ticket, dark are no-wins, white are wins



$$\#(A) = size(A) = 28$$

$$\#(S) = size(S) = 70$$

$$\boldsymbol{P(A) = 28/70 = 0,4}$$

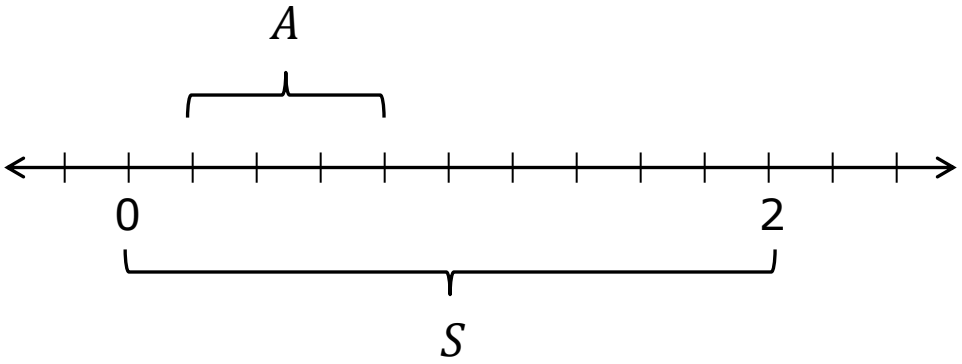
Stockholm University

Variant with a continuous set

- Let $S = (0,2)$ i.e. an **interval**

  $S = \{x; 0 < x < 2\}$, all x such that $0 < x < 2$

- Assume that partial intervals that are equal in size, are equally probable
- Let $A = \{x; 0{,}2 < x < 0{,}8\}$
- Select a random point in the interval S



$size(A) = 0{,}6$

$size(S) = 2$

$\boldsymbol{P(A) = 0{,}6/2 = 0{,}3}$

Stockholm University

Combinatorics – counting the outcomes

Q. Menu with ***3 starters***, ***4 main courses*** and ***2 deserts***.

How many different 3-course dinners can you order?

A: $3 \cdot 4 \cdot 2 = \mathbf{24}$ Illustrate with a tree diagram!

In combinatorics, the number of possibilities often explode!

Ex. *Students are to be arranged in pairs. How many different configurations of 15 pairs are there in a group of 30 students?*

Answer: 6 190 283 353 629 375

Stockholm University

Principal of Multiplication

- An experiment has $\boldsymbol{m_1}$ possible outcomes
- Another experiment has $\boldsymbol{m_2}$ possible outcomes
- We carry out the first, then the second experiment
  - *the two experiments may be different or the same*

- In total there are

$$\boldsymbol{m_1 \times m_2}$$

possible combinations

Note! m_1 and m_2 may be different numbers or equal; depends on the experiments!

Note! Assumes that m_2 isn't affected by the outcome of experiment 1

Stockholm University

Combinatorics, cont.

Example

A bag with numbered balls $\mathbf{1}, \ldots, \boldsymbol{n}$

- Randomly draw one ball and write up the number

  How many possible outcomes?

- Randomly draw one ball again and write up the number

  How many possible outcomes?

- ***How many possible outcomes in total? First and second draw considered together?***

Stockholm University

Repeated draws, draw *x* times from *n*

- **Without replacement, *wor*** (sv. *utan återläggning)*
  - A ball that is drawn is put to the side and cannot be drawn again
  - How many possible outcomes, combinations?

    $n \cdot (n-1) \cdot \ldots \cdot (n-x+1)$ Multiplication principal

- **With replacement, *wr*** (sv. *med återläggning)*
  - A ball that is drawn is put back into the bag and may be drawn again, repeatedly
  - How many possible outcomes, combinations?

    $n \cdot n \cdot \ldots \cdot n = n^x$ Multiplication principal

Stockholm University

Combinatorics, cont.

Does the ***order*** in which the balls are drawn matter?

Assume that $x = 2$ letters are to be chosen (**wor**) from the set of $n = 4$ letters **A**, **B**, **C**, **D**

- Order matters:
  - the outcomes **A**,**B** och **B**,**A** are considered as **different**
  - the formulas on the preceding page – order matters
- Order doesn't matter:
  - the outcomes **A**,**B** och **B**,**A** are treated as **equal**
  - in many cases the ordering is not essential to the problem

Stockholm University

Factorial

- From the multiplication principal it should be clear that the number of ways we can arrange x different objects is

$$x \cdot (x-1) \cdot (x-2) \cdots 3 \cdot 2 \cdot 1 = x!$$

↑ 1st draw ↑ 2nd draw ↑ 3rd draw … ↑ last draw

- $x!$ pronounced ***x factorial*** (sv. *x fakultet*)

- Ex: 0! = 1 1! = 1 2! = 2 3! = 6 4! = 24

  10! = 3 628 800 20! = 2 432 902 008 176 640 000

Stockholm University

Summary

Understand the concepts, start with the Quiz F1-F4 today! :

- Random experiments – generating random data
- Sample space – all that is possible
- Events – a part of the sample space (basic outcomes)
- Set theory – structured way to describe
- Probability – what it is (really)
  - Classical interpretation ⇒ combinatorics, art of counting

Next time we will continue exploring combinatorics, set a foundation for probability theory and models and learn how to calculate probabilities.

Stockholm University