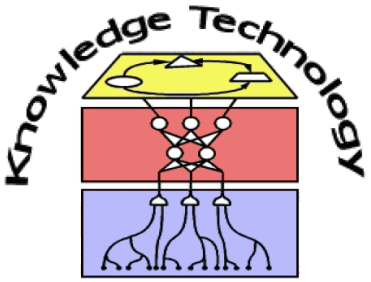
Research Methods

Empirical Analysis, Data and Variability

Dr. Sven Magg, Prof. Dr. Stefan Wermter



http://www.informatik.uni-hamburg.de/WTM/

Plan for today!

1. What is “empirical analysis”?
2. Types of empirical studies
3. What are data and data scales?
4. What to measure?
5. Variability, and how to quantify it!
6. Explaining variability

A word about MatLab

- All the examples are generated with GNU Octave
- Octave = MatLab – License - Toolboxes
- Octave has (for our purpose) the same syntax and functionality
- I'm not teaching you MatLab/Octave!

- Data files will appear in the MinCommSy soon
- Scripts, containing all commands to create the examples and graphs used will be made available

What is “empirical” analysis?

- Aristotle: Two types of arguments:
  - Dialectical: based on logical deduction
  - Empirical: based on practical considerations.
- “empirical” = derived from experiment and observation rather than theory

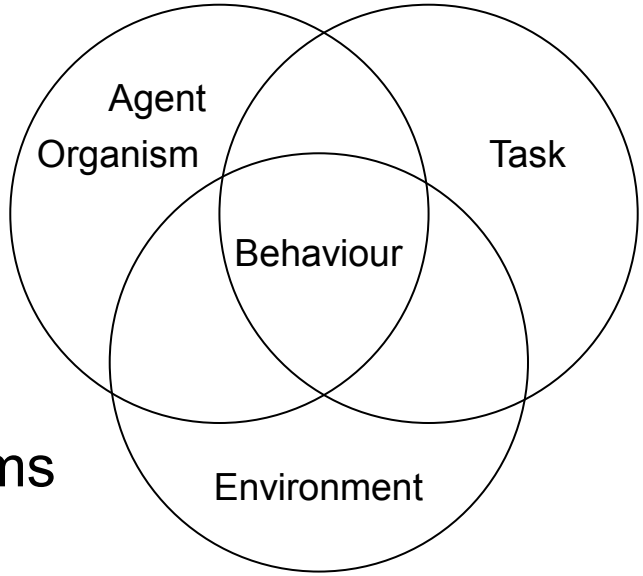
empirical = exploratory + experimental

- Exploratory techniques:
  - visualisation, summarisation, exploration, modelling
- Confirmatory techniques:
  - Hypothesis testing, predictions

What are our test subjects?

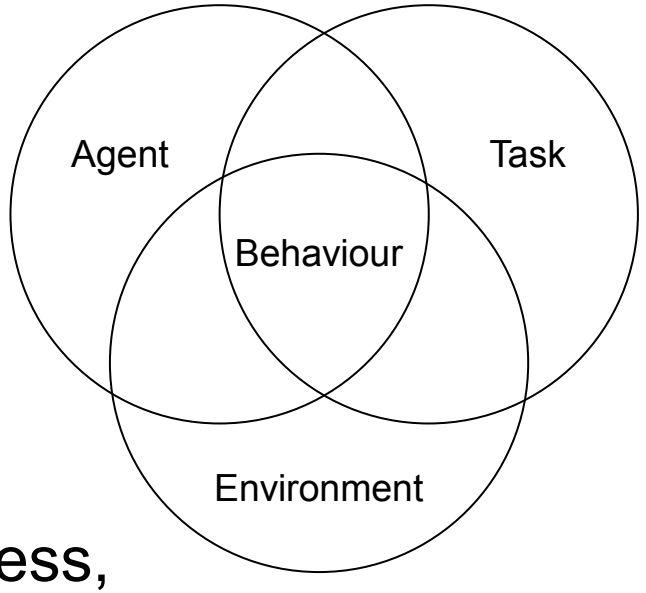
- We are computer scientists, so we have
  - Computer programmes
  - Artificial agents, robots
  - Humans that interact with those systems

- Artificial Systems
  - perform a task in an environment
  - are often too complex to predict the behaviour accurately
  - are therefore similar to real organisms

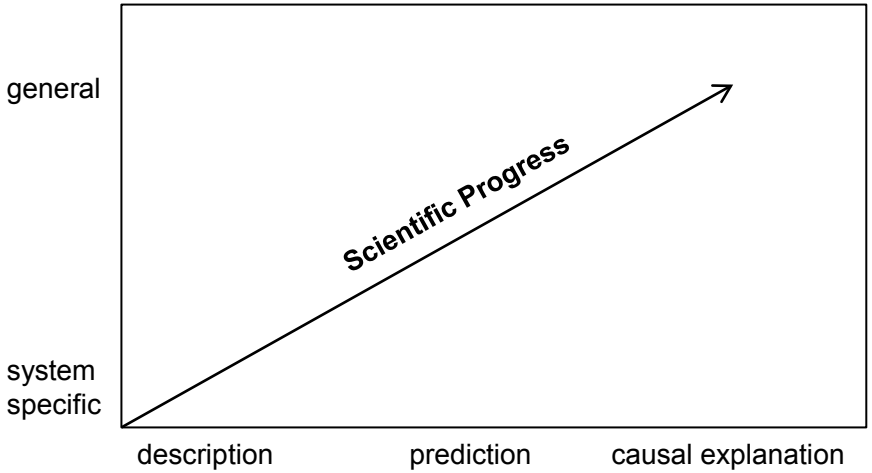


Research questions

- General form:
  How does a change in one given the other two affect behaviour?
- Behaviour is what you can measure!



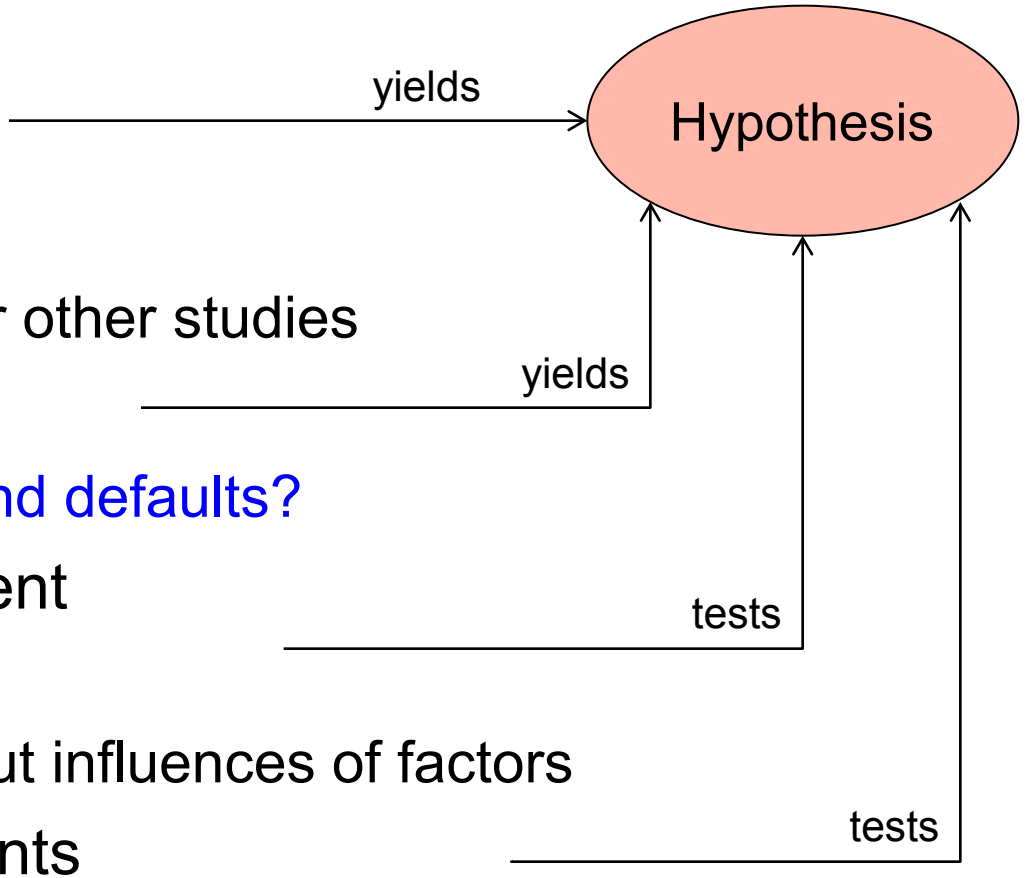
- Answering these questions is a process, from
  - description to prediction and causal explanation
  - system specific to general



Types of studies

We have a system,….

- Exploratory Study
  - what can/does it do?
  - Yields hypotheses for other studies
- Assessment Study
  - where are its limits and defaults?
- Manipulation Experiment
  - what happens if….?
  - Test hypotheses about influences of factors
- Observation Experiments
  - how correct is my model of what should happen?



Data & Samples

Factors

Physical

Social

Variables

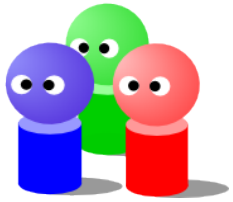
Individuals

Name	Hair Colour	Height	Weight	Income	Education
Hans	brown	192.5	high	1024	BSc
Jenny	blond	168.37	medium	750	MSc

Data & Samples

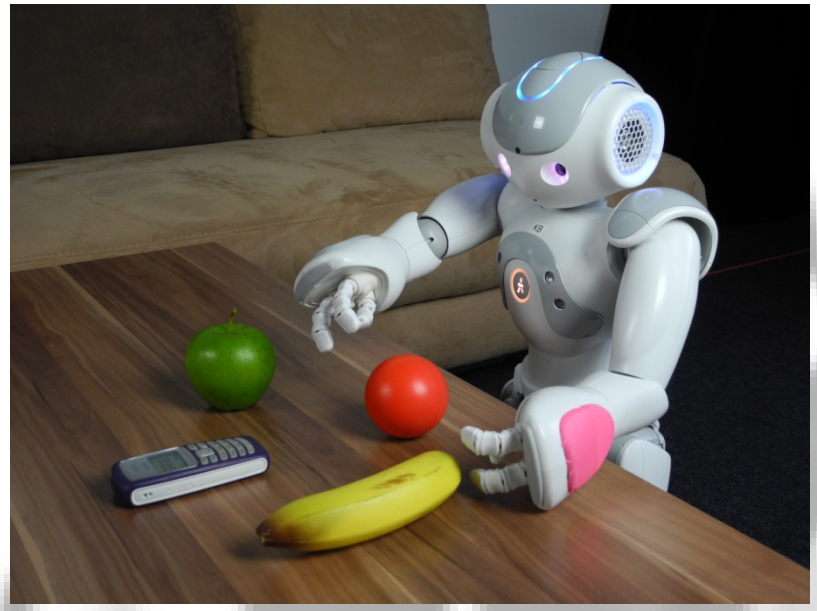
Data:	Collection of measurements
Individual:	Data of one measurement
Factor:	Abstract characteristics of an individual, represented by a set of variables
Variable:	Label for a function that maps from individuals to data scales. Variables produce measurements of factors
Sample:	A collection of individuals

Group Task!



4

10



Robot behaviour measured over time

In this setting, what are good examples for:

- **Data?**
- **Individual?**
- **Factors?**
- **Variables?**
- **Sample?**

Scales of data

Name	Hair Colour	Height	Weight	Income	Education	Exam
Hans	brown	192.5	high	1024	BSc	Fail
Jenny	blond	168.37	medium	750	MSc	Pass

- Categorical data: Individual is assigned to a category
- Ordinal data: Data points can be ranked
  - distances between points are still unknown
- Interval data: Distances between points are meaningful
- Ratio data: Interval data with true & meaningful zero point
  - Ratios of values along the scale are meaningful

What about temperature in Celsius, Fahrenheit, and Kelvin?

Scales of data

- Other distinctions or terms:
- Nominal data: Measurement is a name/tag
- Continuous vs. discrete
  - Interval/Ratio data often are continuous data
  - Continuous data can be measured to any level of precision e.g. age, distance, time
  - Discrete data can take only given values (e.g. 5-point scale)

Scales of data

- Depending on scale, different computations can be performed:

	Categorial	Ordinal	Interval	Ratio
Frequency Distribution, Mode	YES	YES	YES	YES
Median, Percentiles	NO	YES	YES	YES
Addition, Subtraction, Mean, Standard Deviation, Standard Error	NO	NO	YES	YES
Ratio, Coefficient of variation	NO	NO	NO	YES

- Which scale to choose for which variable?
- Scales provide representations of data, i.e. reality

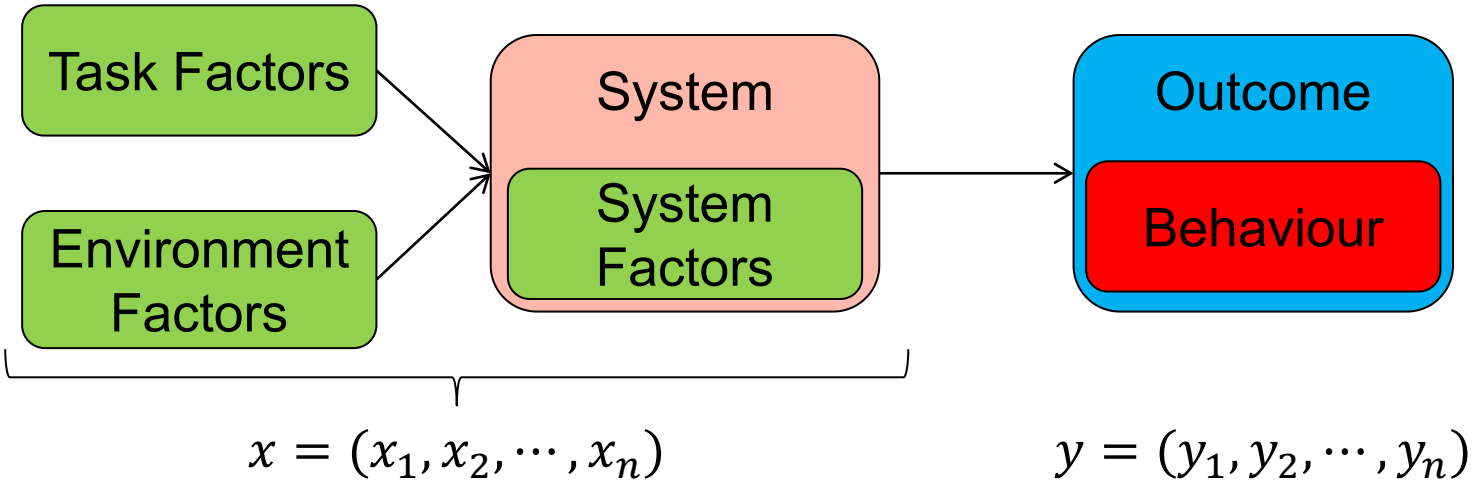
$\Rightarrow$ Representation of reality depends on your purpose!

Scales of data

- Can I transform data from one scale to the other?
  - Yes: Ratio/Interval → Ordinal → Categorical
  - e.g. Replace each datum by rank:
    (45,7,38,0) → (4,2,3,1)
  - Transformation loses information!
- Measurement theory
  - Which inferences are protected by transformations?
  - e.g. monotonic transformations preserve order but not magnitude

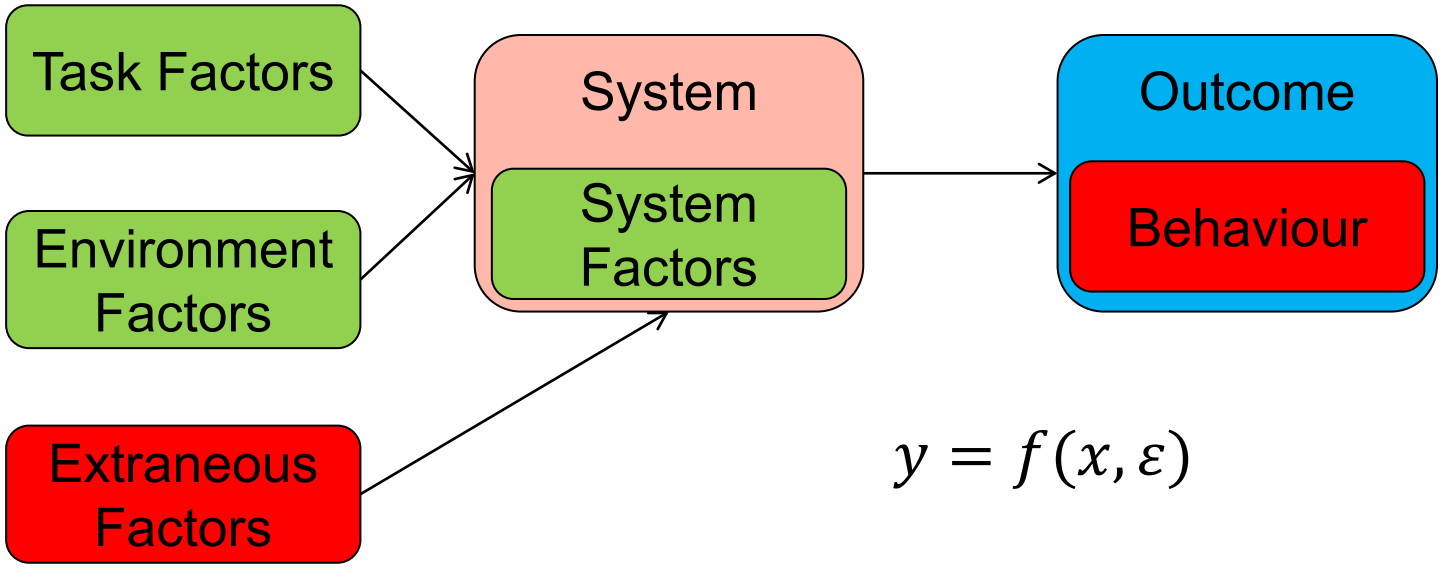
- Again, think which relationships are important for your purpose!

What to measure?



- x_n: Independent (manipulated, controlled, predictor, explanatory) variables, risk factor, feature
  - If x_n does not change within experiment: often “parameter”
- y_n: Dependent (measured, response) variables
- $y = f(x)$: Model of system to explain causal relationships

What to measure?

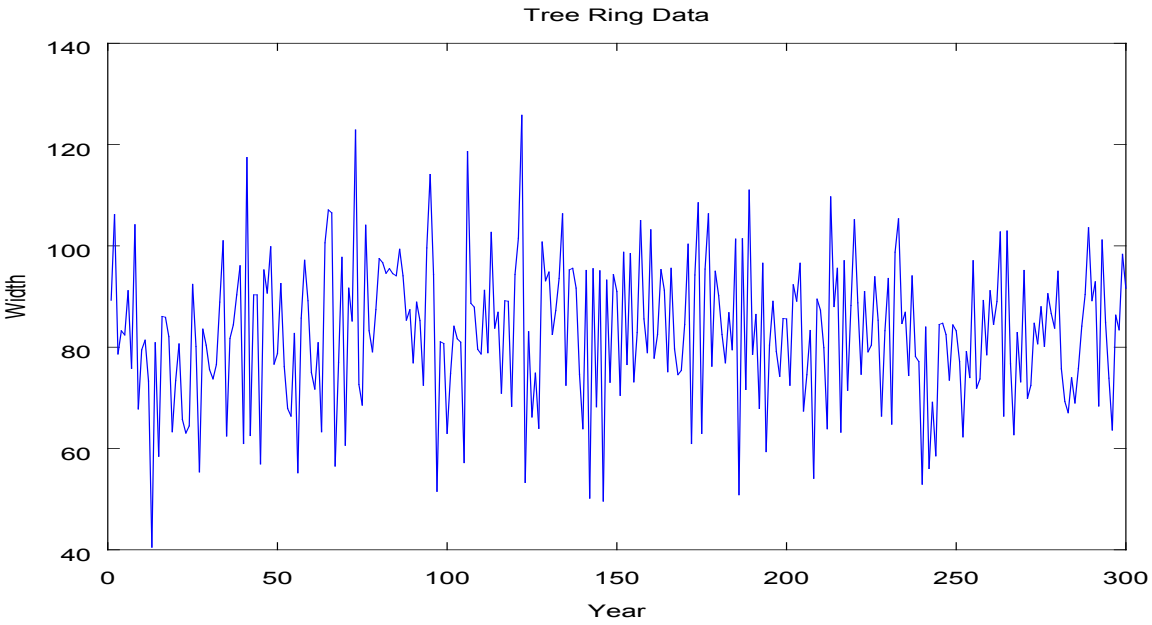


- Extraneous, non-measured variables
  - Assumption that there is only a small and constant effect
- Problem: Confounding, hidden, lurking factors
  - Correlates with both dependent and independent factors

What to measure?

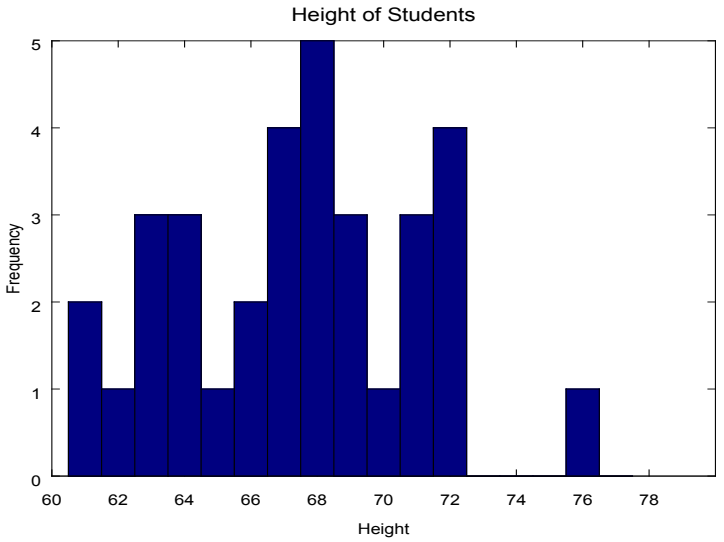
- Several criteria for measurements:
- Validity
  - Measured value represents what we think it does
  - Measurement is not affected by other hidden factors
  - e.g. does a 1-5 point happiness scale measure happiness?
- Reliability
  - Measurement produces same result in same condition
- Measurement error
  - Difference between recorded and real value
  - Validity and reliability high $\Rightarrow$ reduced error

Variability



Variability over time within one tree

Variability between individuals



Variability

- Variability of a datum is the result of a complex causal relationship between factors

Science is about identifying and measuring factors that explain variability!

$$y = f(x, \varepsilon)$$

- Explained variability due to measured/controlled factors (x)
- Unexplained variability due to unmeasured factors (ε)

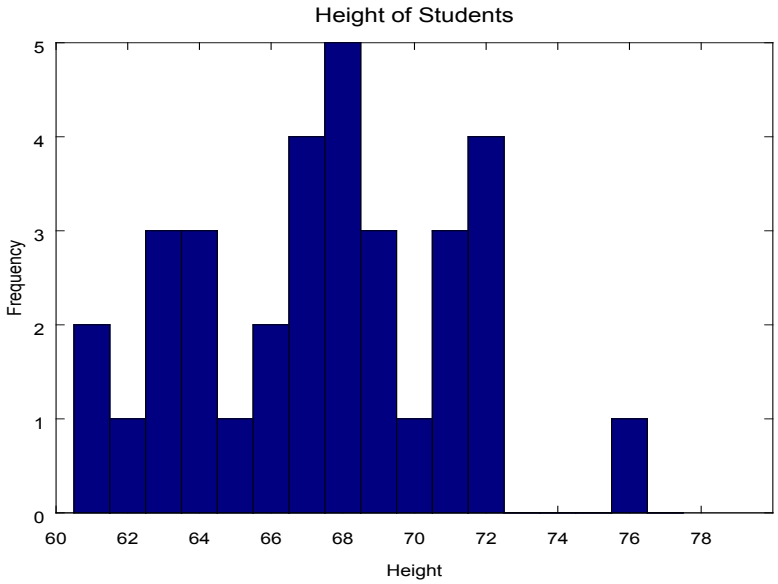
Variability

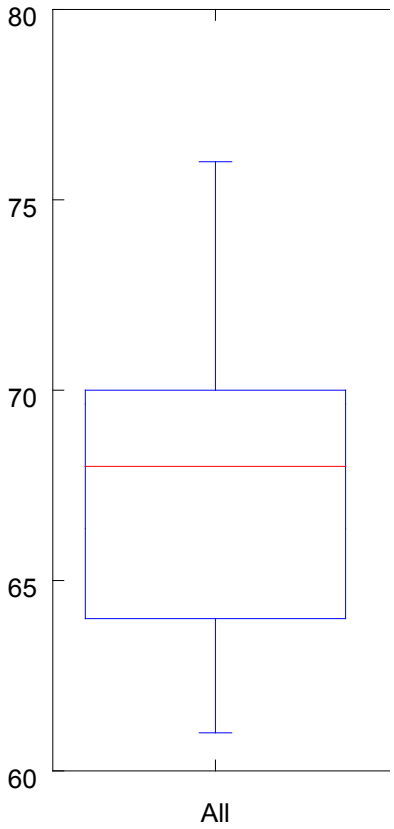
- A scientists task is, to
  - find factors that influence variability in y
  - explain how they combine (f)
  - reduce unexplained variability

- There usually always is some unexplained variability
  - due to measurement errors
  - due to extraneous variables too difficult to measure

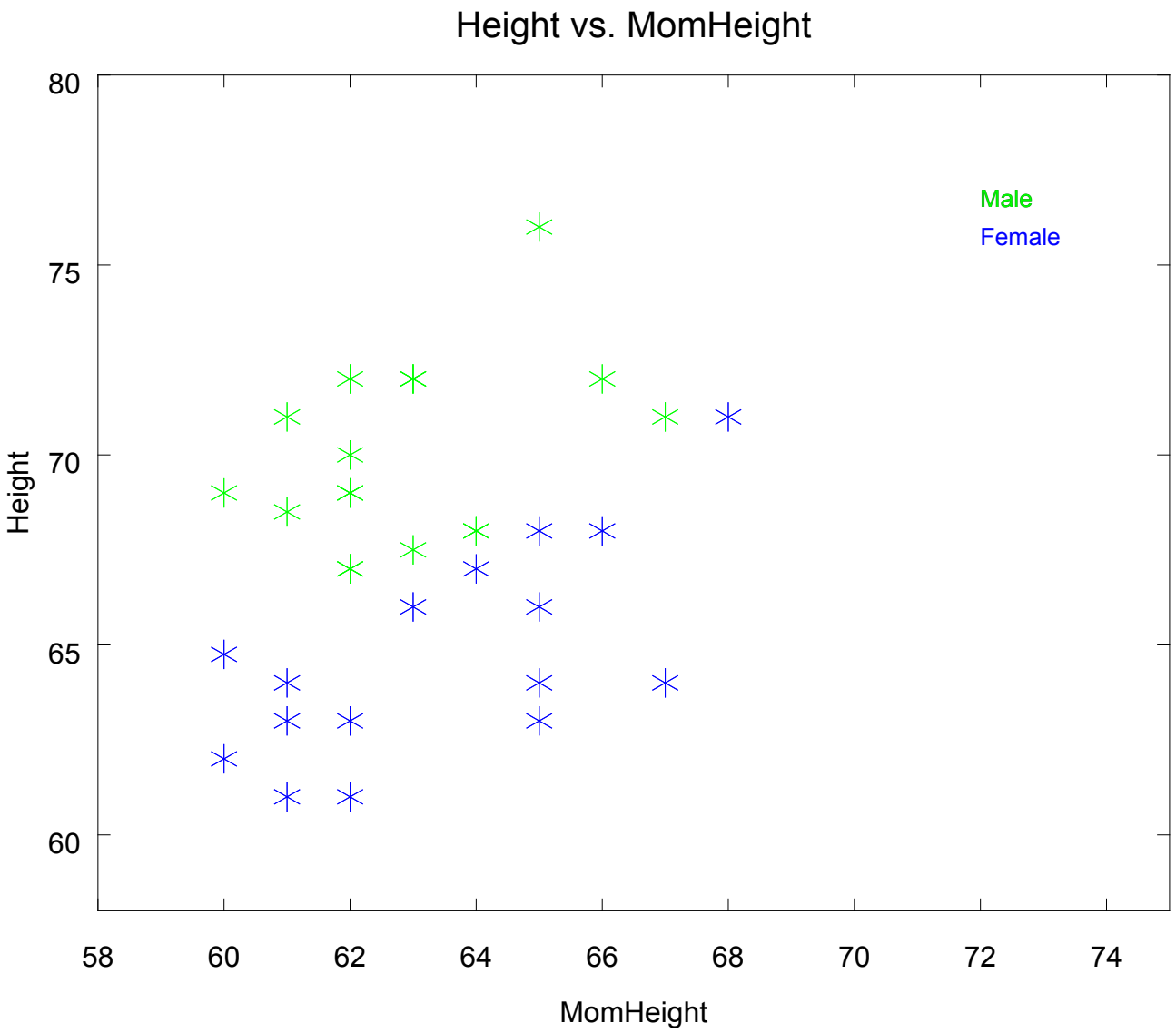
- Try to explain variability *"well enough"*

Variability Example

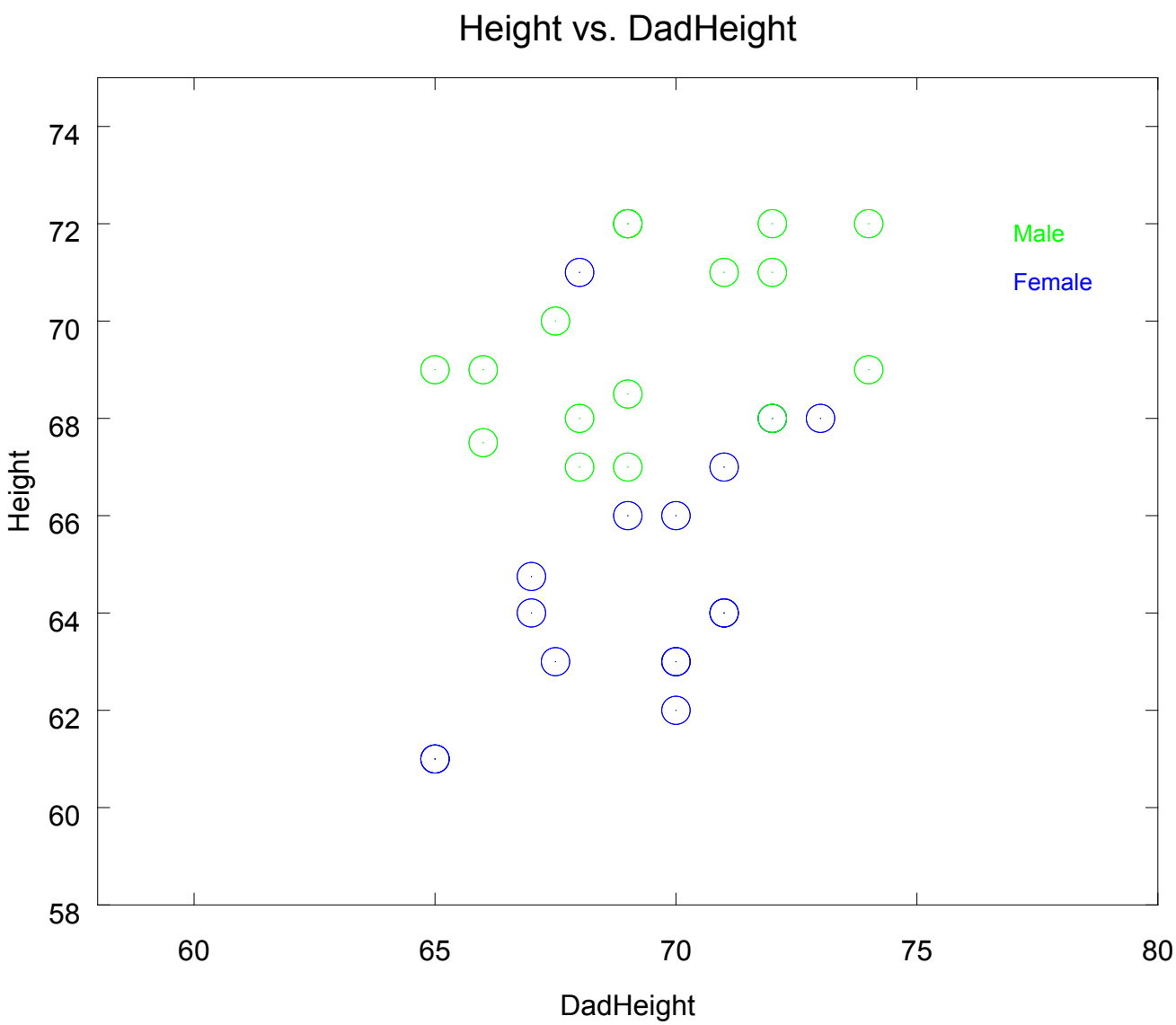




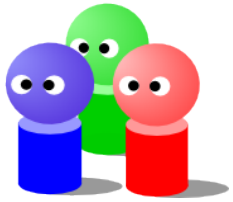
Variability Example



Variability Example



Group Task!



2 5

1. What happens to variability if we manage to control all important factors?
2. Using the graphs (scatter and box), how can we see whether a factor had no influence?

Measuring Variability

Raw data

63	63	66	62	63	64	66	61	68	68	64,8	67	64	64	71	61	72
72	71	68	68,5	69	67	69	67	76	72	68	71	70	69	67,5	72	

Frequency data

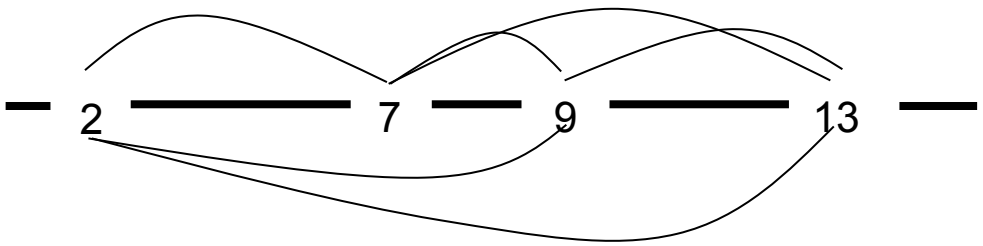
61	62	63	64	64.75	66	67	67.5	68	68.5	69	70	71	72	76
2	1	3	3	1	2	3	1	4	1	3	1	3	4	1

- We need to measure variability to explain it!
- The range is [61,76]
  - Gives us the minimum and maximum

- What we are interested in: Difference between all pairs of data points

Average Squared Difference

- Data: (2,9,13,7)



	2	7	9	13
2	0	-5	-7	-11
7	5	0	-2	-6
9	7	2	0	-4
13	11	6	4	0

Average Squared Difference

$$D^2 = \frac{\sum_{i,j=1, i \neq j}^{N} (x_i - x_j)^2}{N^2 - N}$$

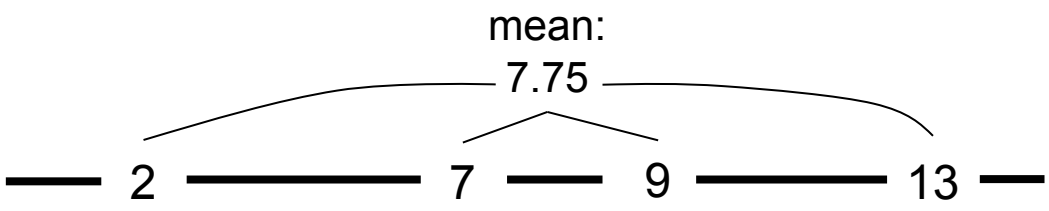
$$D^2 = \frac{(2-7)^2 + (2-9)^2 + \cdots + (13-9)^2}{4^2 - 4}$$

$$= 41.833$$

Average difference:

$$D = 6.467$$

Deviance from the mean



Mean

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

- Mean as a model for our data
- Estimate the accuracy of this model by sum of squared errors (SS):

Sum of Squared Errors

$$SS = \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$SS = (-5{,}75)^2 + \cdots + (-5{,}25)^2 = 62.75$$

Downside: SS is dependent on N

To get average error, divide by N

Sample Variance

Sample Variance

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

Standard Deviation

$$s = \sqrt{s^2}$$

$$s^2 = \frac{((2-7.75)^2 + \cdots + (13-7.75)^2)}{3}$$

$$= 20.916$$

Standard Deviation:

$$s = \sqrt{s^2} = 4.5735$$

- Sum of Squared Errors, sample variance, and standard deviation all measure how well our model "Mean" fits the data

$$outcome_i = model + error_i$$

$$Deviation = \sum (observed - model)^2$$

ASD vs. sample variance

$$D^2 = \frac{\sum_{i,j=1,i\neq j}^{N}(x_i - x_j)^2}{N^2 - N} = 41.833$$

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1} = 20.916$$

- $D^2 = 2s^2$
- The sample variance is proportional to the average squared difference between data points in our sample
- Sample variance represents the variability between data points
- $\mathrm{D} = s\sqrt{2}$
- Standard deviation times $\sqrt{2}$ is the average difference between two data points

Sample vs. population

- A sample is taken from a (maybe unknown) population
- How well does the sample represent the population?
- Reality:
  - $N \ll |population|$

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} \qquad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

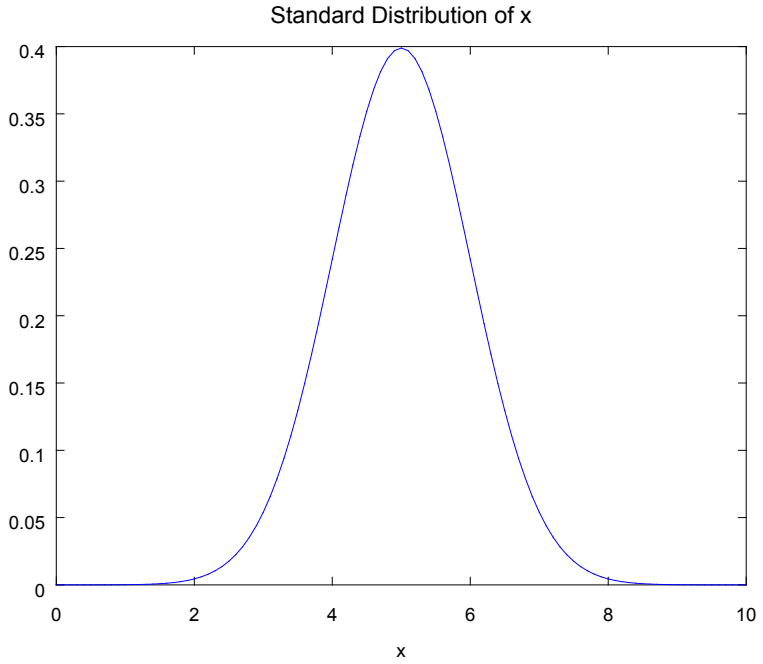
  - Sample variance ≠ population variance
- Sample variance is an estimate of population variance
- N-1 because we lose one degree of freedom by keeping the mean constant (assuming it is the population mean)

Central Limit Theorem

- If we draw N samples from a population, we get N sample means due to sampling variation
- Histogram of all sample means: Sampling distribution
- How does a sampling distribution look?

Central Limit Theorem

The sampling distribution of the mean of samples of size N approaches a normal distribution with increasing N



Standard error

- Central Limit Theorem
  - It holds irrespective of the shape of the population
  - If samples are drawn from a population with mean μ and standard deviation σ, the mean of the sampling distribution is μ and its standard deviation $\sigma/\sqrt{N}$
  - N > 30 $\Rightarrow$ sample distribution of the mean $\bar{x}$ is normal

- Standard error of the mean (SE):
  - Standard deviation of the sample mean $SE = \sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$
  - SE of small samples systematically under-estimate the true standard error of the population $\frac{\mu}{\sqrt{N}}$

Another interpretation

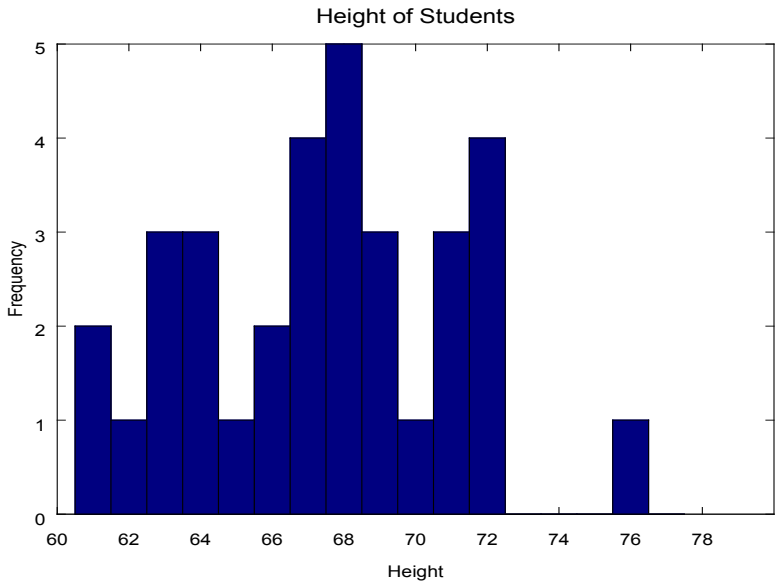
- You have a population of unique elements $E = e_1, e_2, \ldots, e_N$ and associated probabilities p_i. Then mean and population variance are:

$$\mu = \sum_{i=1}^{N} p_i e_i \qquad \sigma = \sum_{i=1}^{N} p_i (e_i - \mu)^2$$

- For $e = (2, 7, 9, 13)$ and $\mathrm{p} = (.25, .25, .25, .25)$
  - Mean: 7.75
  - Population variance: 15.6875
- Related to accuracy of predictions, the error you make when always guessing the mean in a guessing game!

Explaining variance

- If the variance in y is lower when I know x than when I don't know x, x explains that difference
- Example again with Heights:



- D^2= 26,05
- D= 5.1
- Variance: $\sigma^2 = 13.025$
- Standard Deviation $\sigma = 3.609$

Explaining variance

- How big is the variance if we know the gender?
- Variance of full population: $\sigma^2{}_{All} = 13.02525$
- $\sigma^2{}_{males} = 5.378893$
- $\sigma^2{}_{females} = 7.18335$

- Now we can calculate the weighted variance when we know gender:

$$\sigma^2{}_{Pop|gender} = \frac{\sigma^2{}_{males} * |males| + \sigma^2{}_{females} * |females|}{|males| + |females|}$$

$$= \frac{5.378893 * 17 + 7.18335 * 16}{33} = 6.253781$$

Explaining variance

- How big is the variance if we know the gender?
- Variance of full population: $\sigma^2_{All} = 13.02525$
- $\sigma^2_{males} = 5.378893$
- $\sigma^2_{females} = 7.18335$
- $\sigma^2_{All|gender} = 6.253781$

- Variance is reduced to 6.254 when you know gender
- Variance explained by gender is

$$(13.025 - 6.254)/13.025 = 52\%$$

What have we learned?

1. Each data point has a complex causal story associated with it and is influenced by many different factors
2. We try to identify those factors and map them onto data scales via variables
3. Each variable has some variability which we can measure by its variance
4. The job of science is to explain variability in data by finding models that fit “well enough” and reduce variability
5. We play a guessing game, using $y = f(x, \varepsilon)$ to guess y and the standard deviation represents our average error