COPENHAGEN BUSINESS ACADEMY





Basic statistics and machine learning

Jens Egholm Pedersen
<jeep@cphbusiness.dk>

Learning how to learn

- A word on metacognition
  - What does that mean?
- Dunning-Kruger effect
  - Stupid people think they are smart
  - Why is that a bad thing?
- Continuous feedback
  - Where would you like to be when this course ends?
  - Keep evaluating yourself
  - … and be honest!

See also: Dunning-Kruger effect, Metacognition improves your grade!

About these lectures 1/2

- Lectures: Practical part
  - Giving you hands-on experience
  - Resolving immediate problems
  - You need your computer for the practical part
- Lectures: Theoretical part
  - Answering the ‘why’ question
  - Putting things in context (do not underestimate this)
  - You have 1 (one) job
    - You learn best by writing things down. By hand!
  - You do not need your computer for the theoretical part (!)

About these lectures 2/2

- Exploit what we prepared for you
  - Bloom's Taxonomy
  - Lecture = Comprehending
  - Lecture + Preparation = Analyzing
    - Please read the literature. Please?
  - Lecture + Preparation + Exercises = Evaluating
- When studying for the exam use 'see also'
  - **Not part of the curriculum!**

See also: Something to read, Bloom's taxonomy

Goal of this block

- Have a basic understanding and knowledge of various terms, models and tests in statistics.
- Compute basic statistics on data using the Python's scientific stack and the Sklearn library.
- Develop an informed guess of when to choose a certain model to answer a concrete type of question and apply technology appropriately.

See also: BI plan

Goal for today

- Introduction to Scikit learn (sklearn)
- Introduction to statistics
  - Populations
  - Normal distributions
  - Standard deviations
- Introduction to machine learning
  - Prediction
  - Training versus testing
- Linear regression

See also: BI plan

Pandas and sklearn

- `http://pandas.pydata.org/`
- `http://scikit-learn.org/stable/`

Statistics

- What is statistics to you?

- "Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data." - Wikipedia

- For us, statistics means inference
  - Reasoning new knowledge from existing evidence

See also: Statistics

Why statistics?

- Statistics can help us answer questions from data
  - Can I make money on this?
  - Should I smoke this cigarette?
  - Should I buy this house?

- Data is growing. Fast
  - By 2050 there will be around 5200 GB per person

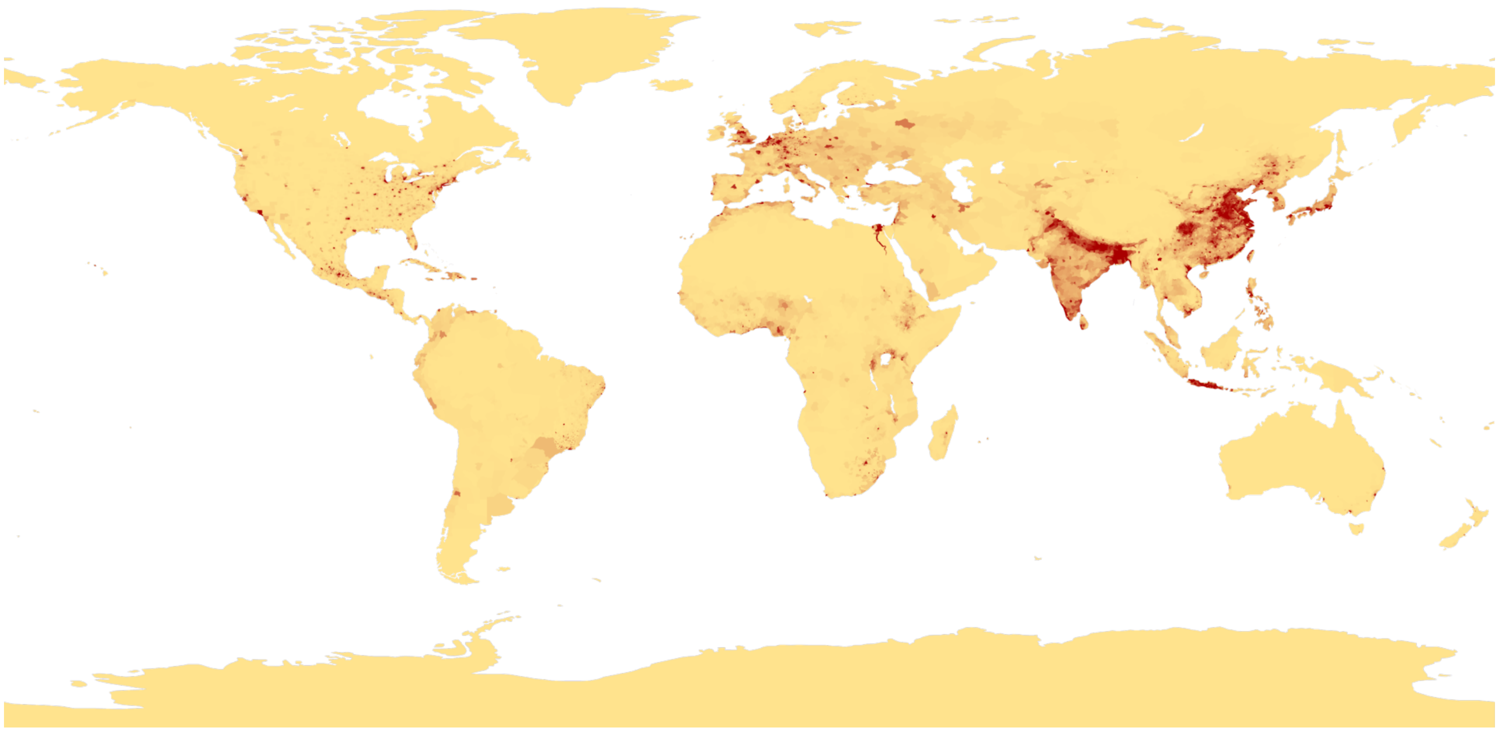
See also: Data is the new oil of the Digital Economy

Introducing statistics

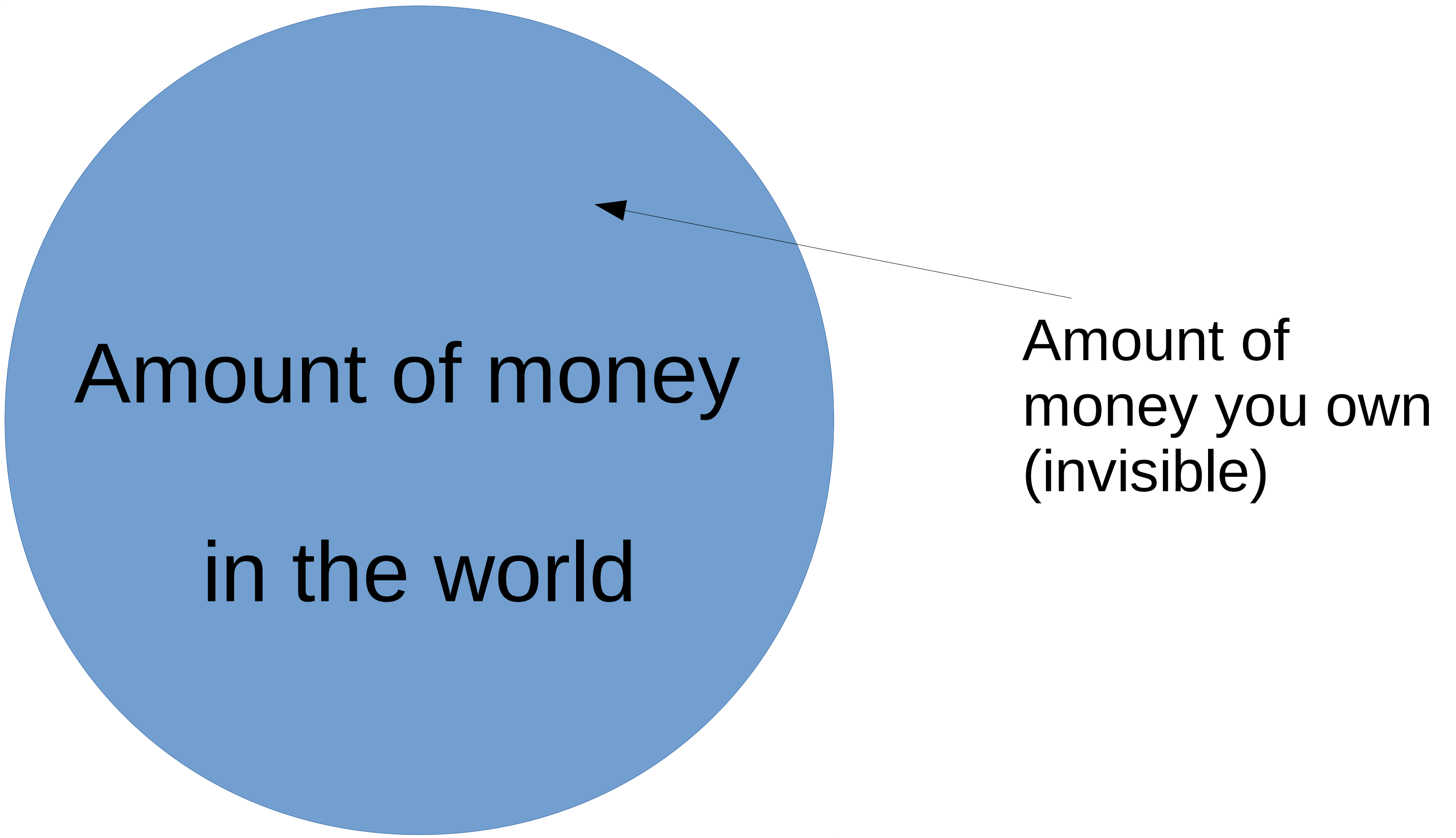
- Descriptive statistics
  - Summarizing the information of a population
- Inferential statistics
  - Predict behaviour based on a population

See also: Free book on statistics (pdf)

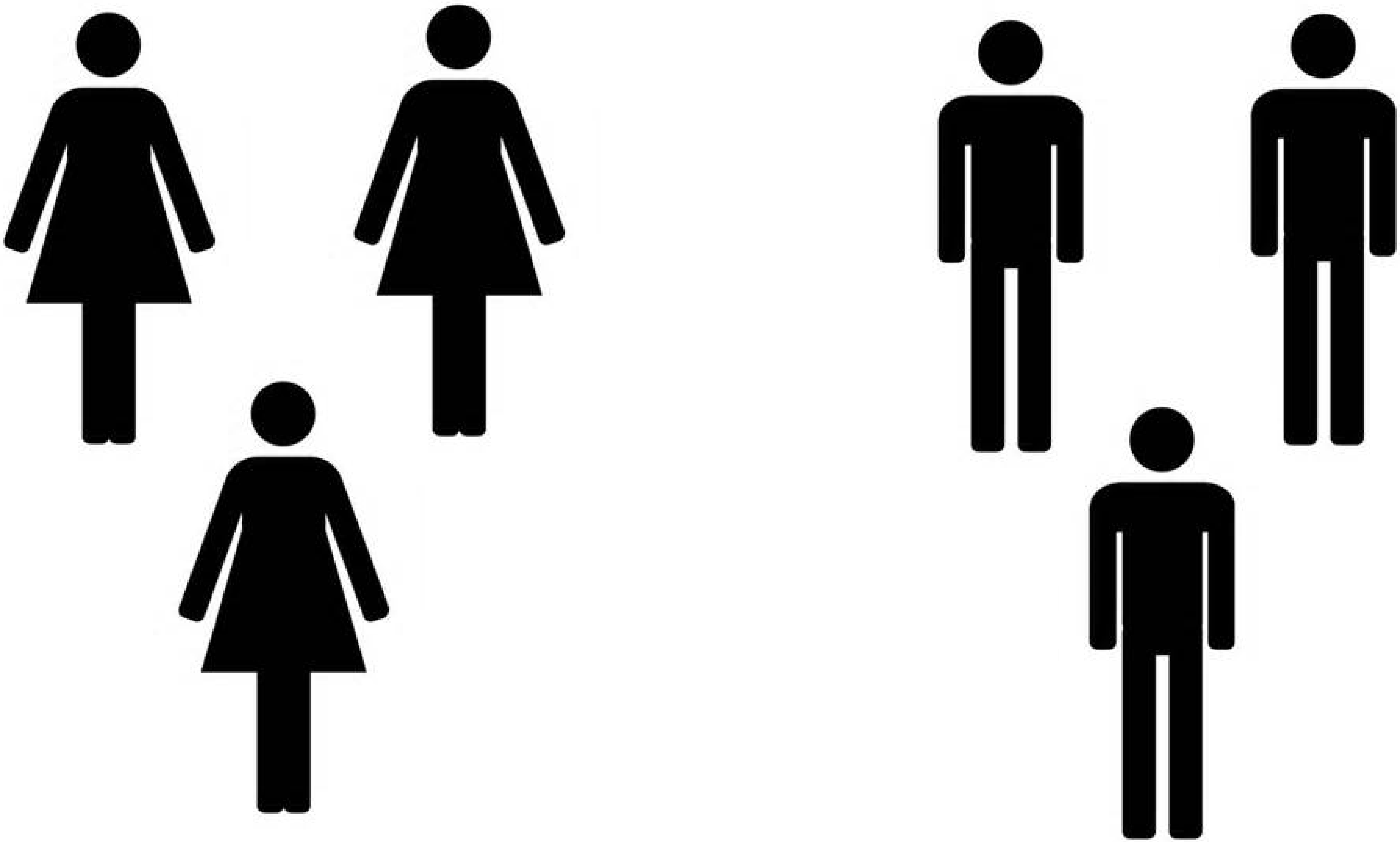
Populations, example 1



Populations, example 2



Populations, example 3



Defining populations

- Sample = subset of population

- Complete sample
  - Entire population
- Representative sample
  - Represents your population

- You have been hired to find out who will win the municipality election in Copenhagen
  - Who will you ask?

Representative samples

- Example: The Literary Digest mailed out millions of mock ballots for the 1936 presidential campaign
  - The results that poured in during the months leading up to the [1936 presidential] election showed a landslide victory for Republican Alf Landon. In more than two million ballots it had received, the incumbent, Roosevelt, had polled only about 40 percent of the votes.
  - Within a week it was apparent that both their results and their methods were erroneous. Roosevelt was re-elected by an even greater margin than in 1932.
- The mailing lists the editors used were from directories of automobile owners and telephone subscribers.
- People prosperous enough to own cars have always tended to be somewhat more Republican than those who do not, and this was particularly true in [the] heart of the Depression.
  - The Digest's experience conclusively proved that no matter how massive the sample, it will produce unreliable results if the methodology is flawed.

- Never ever ever ever (ever) forget this

Source: Unrepresentative samples

Representative samples

- Generally
  - If X% of a sample of people have Y
  - It does NOT mean that X% of people have Y
- To conclude on your results, always consider whether your data is representative
- Never ever ever ever (ever) forget this
- How do you measure “representativity”

Measuring variation

- To obtain a representative sample:
  - Same variation in the sample as in the population
  - Example: 0.01% thinks rape is ok in population
    - Not representative: 50% thinks rape is ok in sample
    - Representative: 0.01% thinks rape is ok

- How do you measure variation?
  - Mean
  - Deviation from mean / variability

Python tools

- http://pandas.pydata.org/
- http://scikit-learn.org/stable/

Practice: Sampling

https://github.com/datsoftlyngby/soft2017fall-business-intelligence-teaching-material/

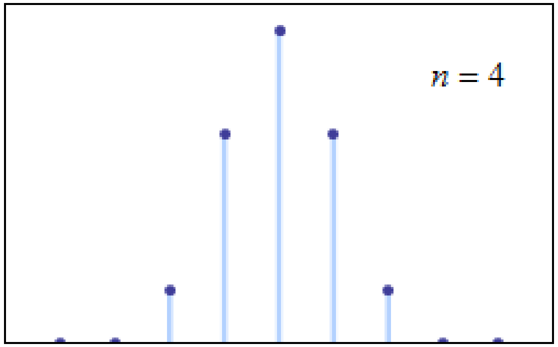
Unknown population

- Let's assume you don't know the full population
  - You only know what is the maximum and minimum values
- What would you think is the mean of the population?
- What would you think is the mean of a sample of that population?
- What is the probability that the mean == center?

Central limit theorem

- The sum of many random variables will approximate a bell curve
  - For instance the mean of many random populations

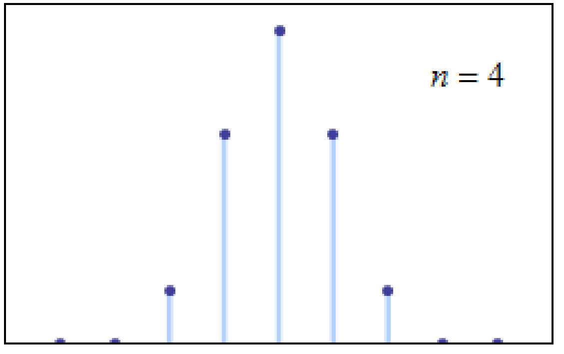
- Why is this important?



Central limit theorem

- The sum of many random variables will approximate a bell curve
  - For instance the mean of many random populations

- Your sample will *probably* be centered around the mean.



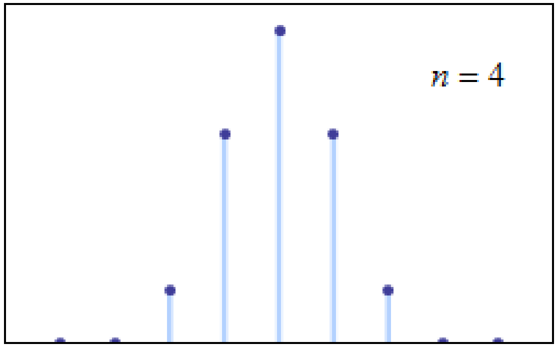
Practice: Mean error

https://github.com/datsoftlyngby/soft2017fall-business-intelligence-teaching-material/

Central limit theorem

- The sum of many random variables will approximate a bell curve
  - For instance the mean of many random populations

- Your sample will *probably* be centered around the mean.



Measuring variation

- How do you measure variation?
  - Mean
  - Deviation from mean / variability

- Standard deviation

$$s = \sqrt{\frac{\sum of\ squared\ deviations}{samplesize - 1}}$$

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sigma$$

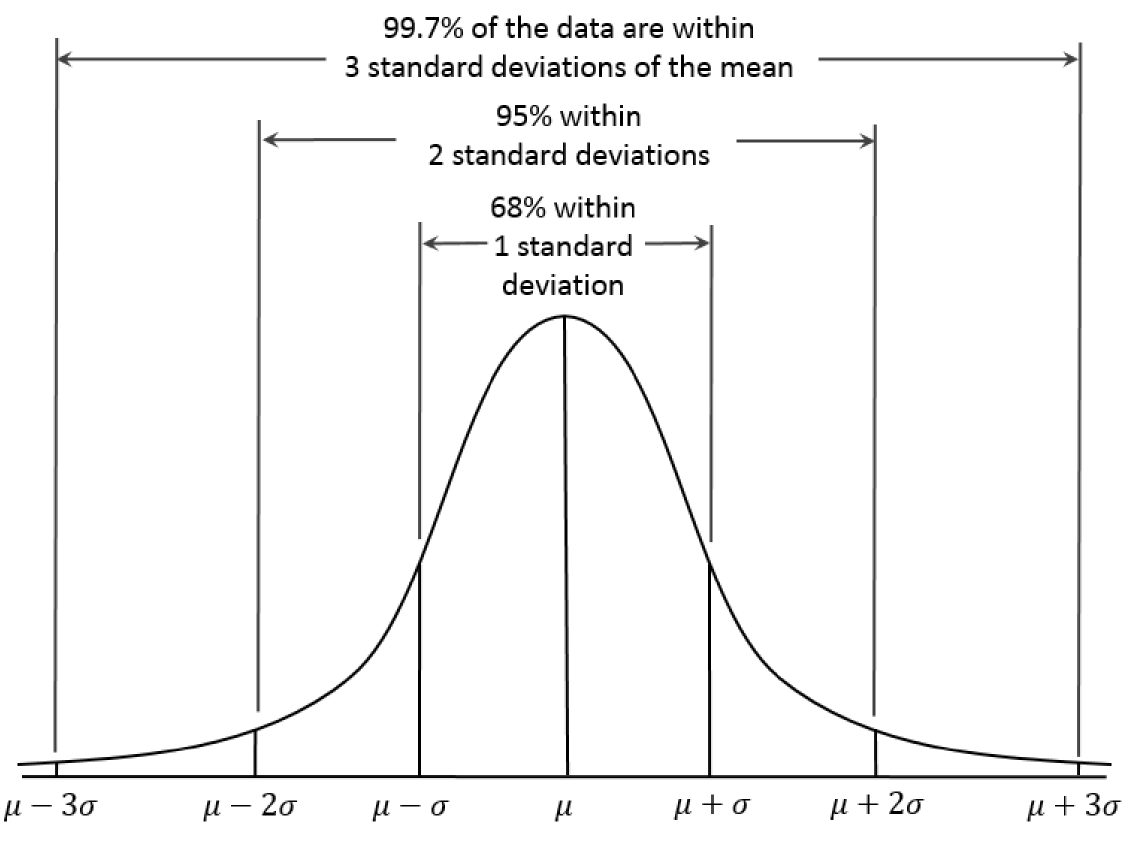
Probability distribution

- What is the “variation” between the population and the sample?
- The probability that your sample == population
- What is that probability?
- Also called a ‘probability density function’. Why?

See also: Probability distribution

Probability distribution

- We are looking for the probability that our sample is equal to the population

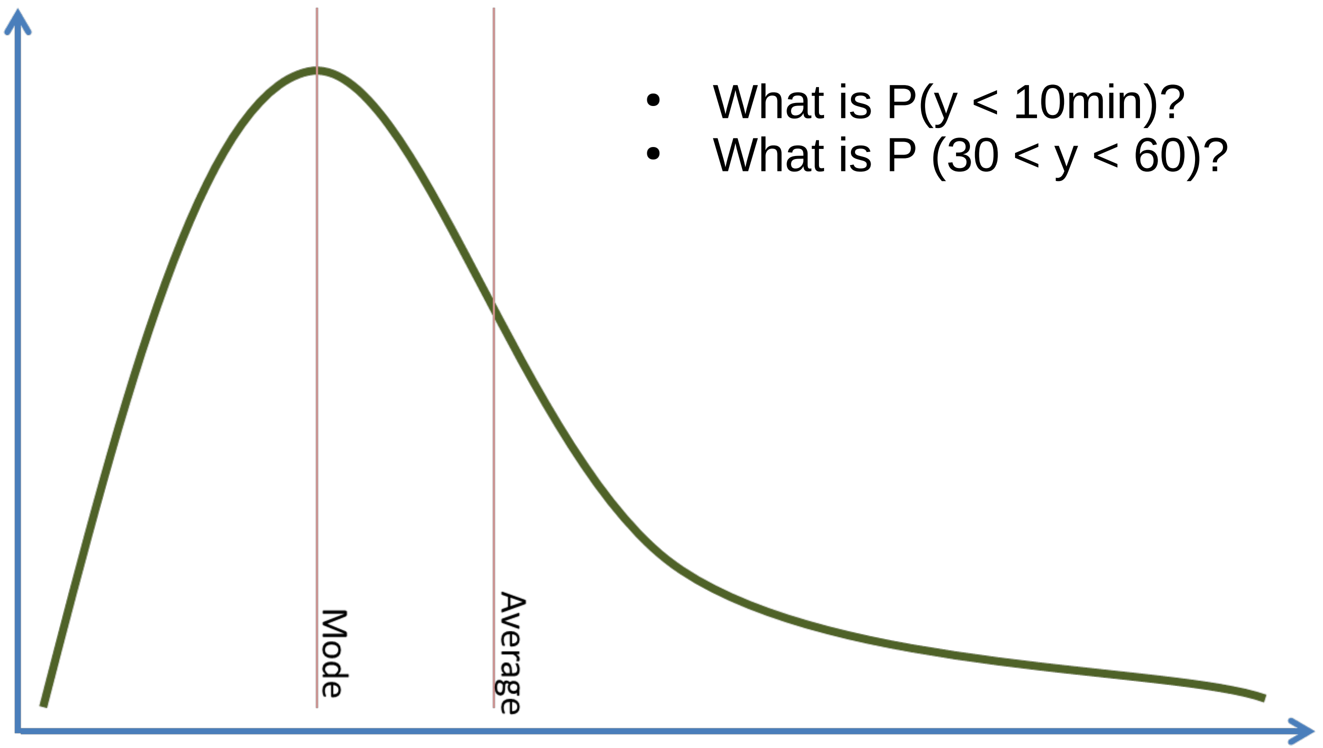


Probability distribution

- We are looking for the probability that our sample is equal to the population
- A probability distribution over a sample **provides a likelihood that a value would equal that sample**

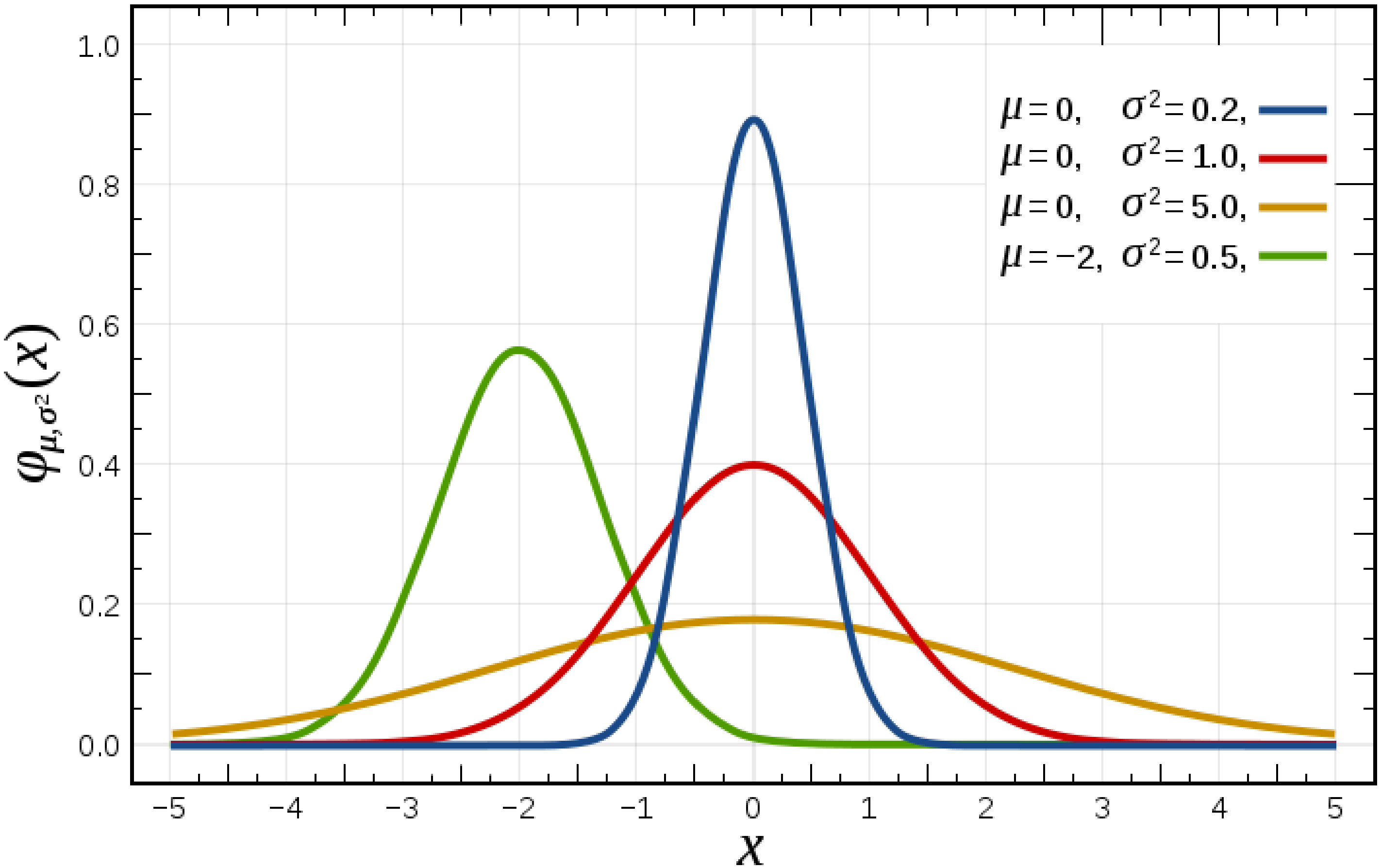
Example: Commuting

- How long does it take to commute to work?
  - Let's plot it



- What is P(y < 10min)?
- What is P (30 < y < 60)?

Normal distribution



See also: Normal distributionNormal distribution

Recap

- Population != sample
  - Very. Very. Very (!) important
- What is the difference between the population and the sample?
  - Mean
  - Standard deviation (square root of variability)
- Probability distributions tells us how likely it is that our sample == population
- Descriptive statistics

Machine learning

- “The science (and art) of programming computers so they can learn from data”

- If your machine downloads an article from Wikipedia, is it smarter?
  - No. That is not machine learning

See also: Géron: Hands-on machine learning (book)

Machine learning

- “The science (and art) of programming computers so they can learn from data”

- Why?
  - Humans cannot process the vast amounts of data
  - Machines can test ideas (such as sample deviation) fast!
  - Machines can become much better than humans

- Example: >98% precision for handwriting recognition

See also: Géron: Hands-on machine learning (book)

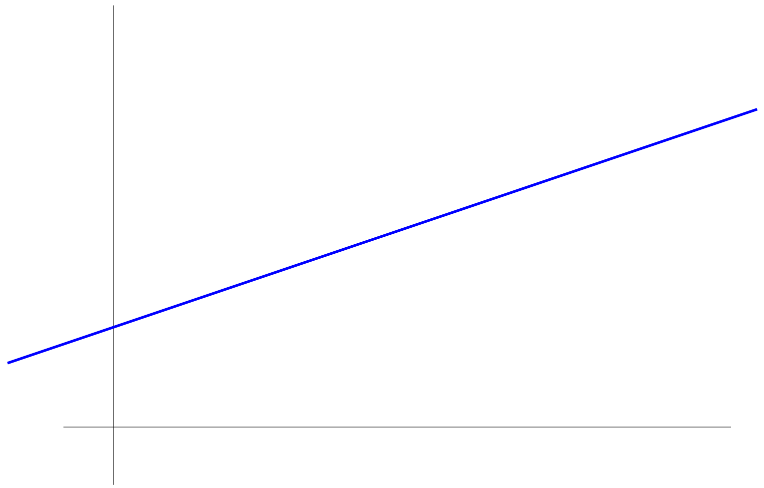
Types of machine learning

- Is it trained by a human
  - Supervised / unsupervised
- Can they learn on the fly?
  - Online learning / offline (batch) learning
- Do they include new data?
  - Instance-based / model-based
- Today: supervised, offline model-learning

See also: Géron: Hands-on machine learning (book)

Linear regression

- Inferential statistics
  - We want to predict something from data
- Supervised learning
  - We have data already
- Offline learning
  - We train it once then use it
- Model based
  - We use a model
  - $y = ax + b$



Scikit learn

- Python statistics + machine learning library
- We will use this extensively

See also: Python SKlearn library

Building models

- Supervised learning
  - We instruct the model with data

- `model.fit()`
  - Trains the model to the data we feed it

- `model.predict()`
  - Predicts the outcome of the model

Linear regression

- Predicting y values based on x values
- Example: Murder rates based on income

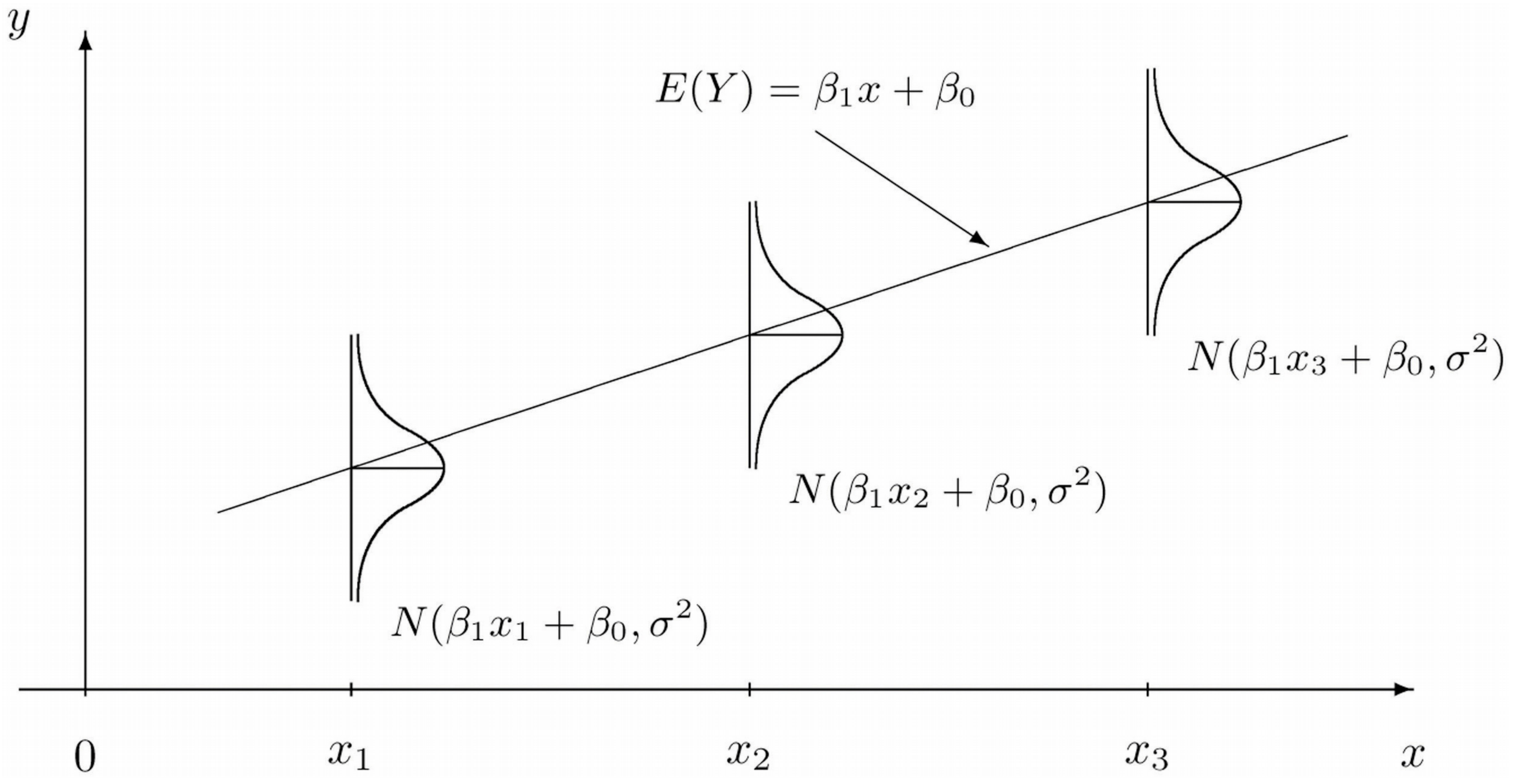
Linear regression

- What is the problem with this?

- Error between data and line
  - Error = distance between points to line

- We are actually not interested in $y = ax + b$
- We are interested in $E(y) = ax + b$
  - We want to make many models of $y = ax + b$ and find the best

Errors in the model



Typical error metrics

- Three metrics:
  - Mean Average Error (MAE)
  - Root Mean Square Error (RMSE)
  - Pearson's r

See also: sklearn on regression metrics

What is an error?

- A good model: expected = actual

$$y_i - \hat{y} = 0$$

- A bad model: expected != actual

$$y_i - \hat{y} = very\ large\ number$$

$$|y_i - \hat{y}| = very\ large\ number$$

See also: sklearn on regression metrics

What is an error?

- For each point, how much error do we add?
- Simplest error metric: Absolute error

$$AE = \sum |y - \hat{y}|$$

- What is the problem with this metric?

See also: sklearn on regression metrics

Mean Absolute Error (MAE)

- For each point, how much error do we add?
  - Controlled for the sample size!

$$MAE = \sum \frac{|y - \hat{y}|}{n}$$

- Sklearn:

```
from sklearn.metrics import mean_absolute_error
mean_absolute_error(y_true, y_pred)
```

See also: sklearn on regression metrics

(Root) Mean Squared Error

- For each point, how much error do we add?
  - Again, controlled for sample size

$$MSE = \sum \frac{\left(y - \hat{y}\right)^2}{n}$$

$$RMSE = \sqrt{\sum \frac{\left(y - \hat{y}\right)^2}{n}}$$

- Sklearn:

```
from sklearn.metrics import mean_squared_error
mean_squared_error(y_true, y_pred)
```

See also: sklearn on regression metrics

Pearson's r

- For each point, how much error do we add?

- All in all: we have two variables: x and y
  - So we can have two errors in our prediction:
    - Errors without x: $y - \bar{y}$
    - Errors with x: $y - \hat{y}$

- What are the differences?
  - Errors without x measures the errors without using the linear model

See also: sklearn on regression metrics

Pearson's r

- For each point, how much error do we add?
- Pearson's r: The coefficient of determination
  - In other words: The reduction in error from using the linear prediction equation instead of simple y values

$$r^2 \approx \frac{error_{linear}}{error_y}$$

$$r^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

See also: sklearn on regression metrics

Pearson's r

- For each point, how much error do we add?

$$r^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

- Sklearn:
  - `model.score(X, y)`
  - Example: Murder rates

See also: sklearn on regression metrics

Regression function

- We are interested in $E(y) = ax + b$
  – We want to make many models of $y = ax + b$ and find the best

- This is a **regression function**
  – A regression function describes how the mean of the prediction changes according to input
  – Or: it helps us to understand how the typical value of the dependent variable changes when the input variable changes

- We want to find the relationship between the input and the output
  – We have to try many different functions so we can find the best values for $y = ax + b$

Practice: Prediction

https://github.com/datsoftlyngby/soft2017fall-business-intelligence-teaching-material/

Training versus testing

- Machine learning models are trained – then tested
- Data for testing **cannot** be used for training
  - Why?
- For the next assignment:
  - 80% training 20% testing

See also: Introduction to Numpy

Recap

- Population != sample
  - Very. Very. Very (!) important
- What is the difference between the population and the sample?
  - Mean
  - Standard deviation (square root of variability)
- Probability distributions tells us how likely it is that our sample == population
- Machine learning
  - Training/testing split (80%/20%)

Next hand-in: Assignment 5

- Deadline: **13th of November 23:59:59**
- The hand-in (on Moodle) should be a link to a GitHub release containing a single file with the code and written text for the assignment parts

- This can either be a .ipynb, .py, .pdf or .md file
- The file must be clearly identifiable. Please name it accordingly. (for instance report.pdf or assignment5.ipynb)