IN4080 – 2022 FALL
NATURAL LANGUAGE PROCESSING

Jan Tore Lønning

Looking at data

Data

- "Data is the new oil"
- We generate enormous amounts around the world every day
- The commodity of Google, Facebook, … and the gang
- "Data Science":
  - Used in various scientific fields to extract knowledge from data
  - Master's program at UiO
  - UiO is establishing a center for DS
- Language data is the raw material of modern NLP

https://pixabay.com/no/illustrations/skjerm-bin%C3%A6re-bin%C3%A6rt-system-1307227/

Data

- Advise in "data science", machine learning and data-driven NLP: Start by taking a look at your data
  - (But tuck away your test data first)
- General form:
  - A set of observations (data points, objects, experiments)
  - To each object some associated attributes
    - Called variables in statistics
    - Features in machine learning
    - (Attributes in OO-programming)

Example data set: email spam

	spam	chars	lines breaks	'dollar' occurs. numbers	'winner' occurs?	format	number
1	no	21,705	551	0	no	html	small
2	no	7,011	183	0	no	html	big
3	yes	631	28	0	no	text	none
4	no	2,454	61	0	no	text	small
5	no	41,623	1088	9	no	html	small
…							
50	no	15,829	242	0	no	html	small

- Data are typically represented in a table
- Each column one attribute
- Each row an observation (n-tuple, vector)
- (cf. Data base)

There are more variables (attributes) in the data set

Example data set: email spam

	spam	chars	lines breaks	'dollar' occurs. numbers	'winner' occurs?	format	number
1	no	21,705	551	0	no	html	small
2	no	7,011	183	0	no	html	big
3	yes	631	28	0	no	text	none
4	no	2,454	61	0	no	text	small
5	no	41,623	1088	9	no	html	small
…							
50	no	15,829	242	0	no	html	small

50 observations, rows

7 variables, columns

4 categorical variables

3 numeric variables

Some words of warning

- This is how data sets often are presented in texts on
  - Statistics
  - Machine learning
- But we know that there is a lot of work before this
  1. Preprocessing text
  2. Selecting attributes (variables, features)
  3. Extracting the attributes

Text as a data set

	token	POS
1	He	PRON
2	looked	VERB
3	at	ADP
4	the	DET
5	lined	VERB
6	face	NOUN
7	with	ADP
8	vague	ADJ
9	interest	NOUN
10	.	.
11	He	PRON
12	smiled	VERB
13	.	.

- Two attributes
  - Token type (‘He’, ‘looked’, …)
  - POS (part of speech)
    - = classes of words
    - we will see a lot to them

Types of (statistical) variables (attributes, features)

<table>
<tr><th colspan="3">All variables</th></tr>
<tr><td rowspan="2">Categorical</td><td colspan="2">Numerical (quantitative)</td></tr>
<tr><td>Discrete</td><td>Continuous</td></tr>
</table>

- Binary variables are both
  - Categorical (two categories)
  - Numerical, {0, 1}
- We will see ways to represent
  - A categorical variable in terms of numerical variables
  - and the other way around
- Machine learning, difference btw.
  - Categorical (classification)
  - Numeric (regression)
- Statistics, difference btw.
  - Discrete
  - Continuous

Categorical variables

- Categorical:
  - Person: Name
  - Word: Part of Speech (POS)
    - {Verb, Noun, Adj, …}
  - Noun: Gender
    - {Mask, Fem, Neut}
- Binary/Boolean:
  - Email: spam?
  - Person: 18 ys. or older?
  - Sequence of words: Grammatical English sentence?

Numeric variables

- Discrete
  - Person: Years of age, Weight in kilos, Height in centimeters
  - Sentence: Number of words
  - Word: length
  - Text: number of occurrences of *great*, (42)
- Continuous
  - Person: Height with decimals
  - Program execution: Time
  - Occurrences of a word in a text: Relative frequency (18.666…%)

Frequencies of categorical variables

Frequencies

- Given a set of observations O
  - Which each has a variable, *f*, which takes values from a set V
- To each v in V, we can define
  - The absolute frequency of v in O:
    - the number of elements x in O such that x.f = v
      - (requires O finite)
  - The relative frequency of v in O:
    - The absolute frequency/the number of elements in O

Universal POS tagset (NLTK)

Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

Distribution of universal POS in Brown

- Brown corpus:
  - ca1.1 mill. words
- For each word occurrence:
  - attribute: simplified tag
  - 12 different tags
- Frequency(absolute)
  - for each of the 12 values:
  - the number of occurrences in Brown
- Frequency (relative)
  - the relative number
    - Same graph pattern
    - Different scale

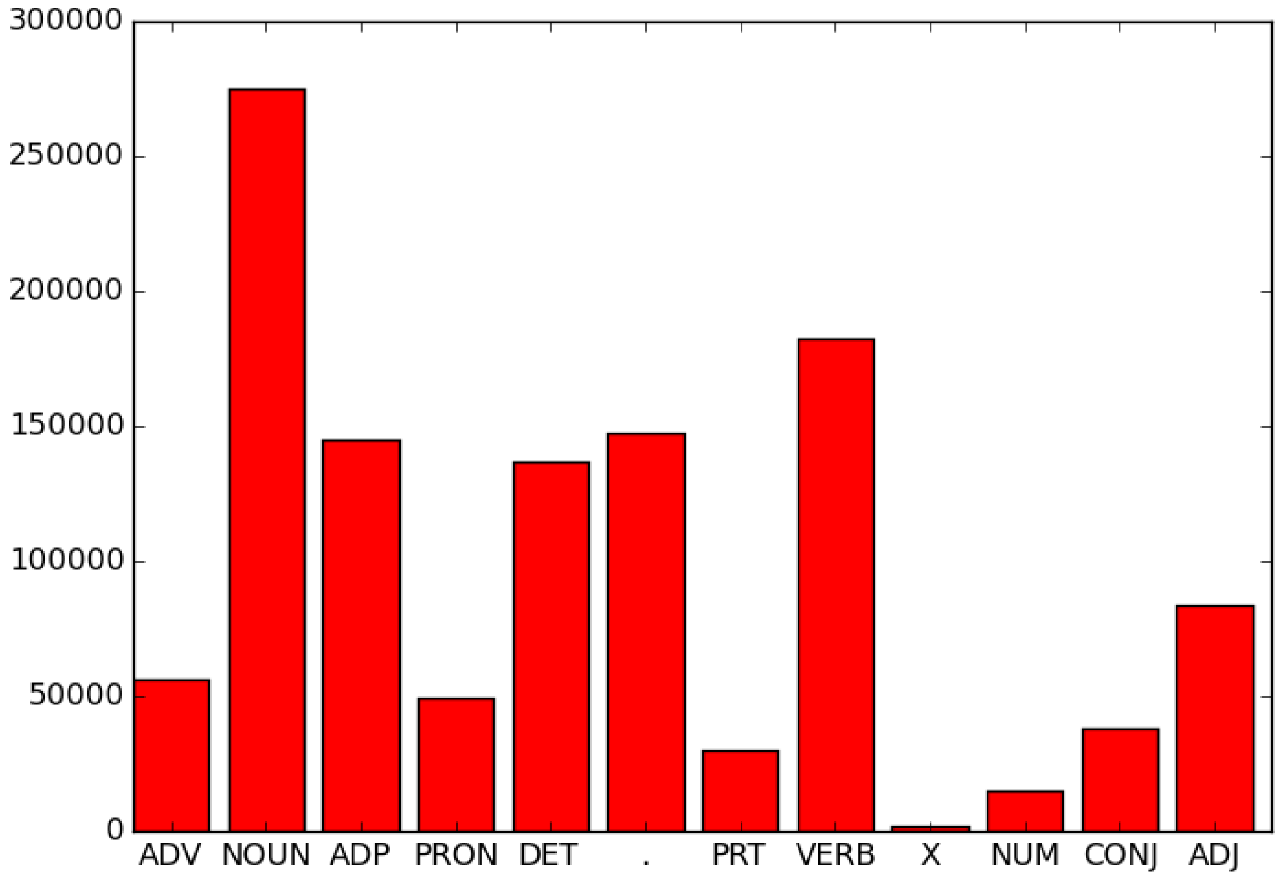
Cat	Freq
ADV	56 239
NOUN	275 244
ADP	144 766
NUM	14 874
DET	137 019
.	147 565
PRT	29 829
VERB	182 750
X	1 700
CONJ	38 151
PRON	49 334
ADJ	83 721

Frequency table
Normally the Cat will be one row (not column) and the frequencies another row

(Numbers from 2015)

Distribution of universal POS in Brown

Cat	Freq
ADV	56 239
NOUN	275 244
ADP	144 766
NUM	14 874
DET	137 019
.	147 565
PRT	29 829
VERB	182 750
X	1 700
CONJ	38 151
PRON	49 334
ADJ	83 721



Bar chart

To better understand our data we may use graphics.
For frequency distributions, the bar chart is the most useful

Frequencies

- Frequencies can be defined for all types of value sets V (binary, categorical, numerical) as long as there are only finitely many observations or V is countable,
- But doesn't make much sense for continuous values or for numerical data with very varied values:
  - The frequencies are 0 or 1 for many (all) values

More than one categorical feature

Two features, example NLTK, sec. 2.1

	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

- Example of a contingency table (directly from NLTK)
- Observations, O, all occurrences of the five modals in Brown
- For each observation, two parameters
  - f1, which modal, V1 = {can, could, may, might, must, will}
  - f2, genre, V2={news, religion, hobbies, sci-fi, romance, humor}

Two features, example NLTK, sec. 2.1

	can	could	may	might	must	will	Total
news	93	86	66	38	50	389	722
religion	82	59	78	12	54	71	356
hobbies	268	58	131	22	83	264	826
science_fiction	16	49	4	12	8	16	105
romance	74	193	11	51	45	43	417
humor	16	30	8	8	9	13	84
Total	549	475	298	143	249	796	2510

- Example of complete contingency table
  - Added the sums for each row and column

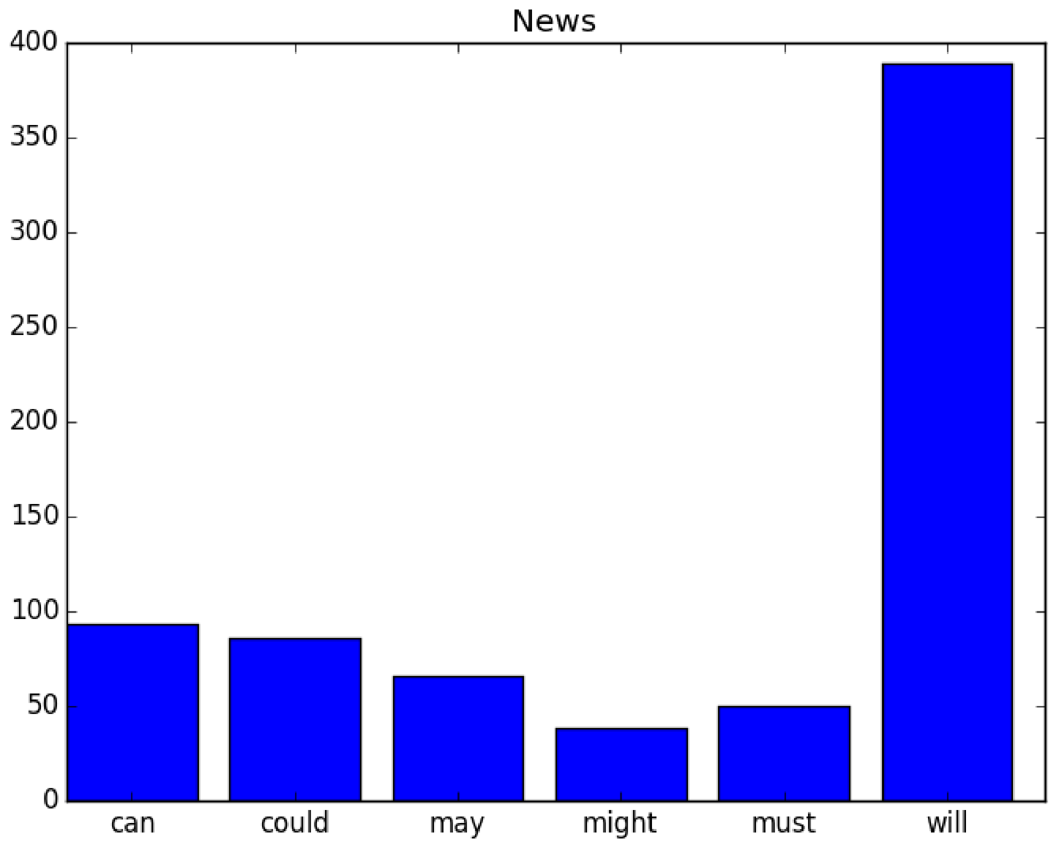
Two features, example NLTK, sec. 2.1

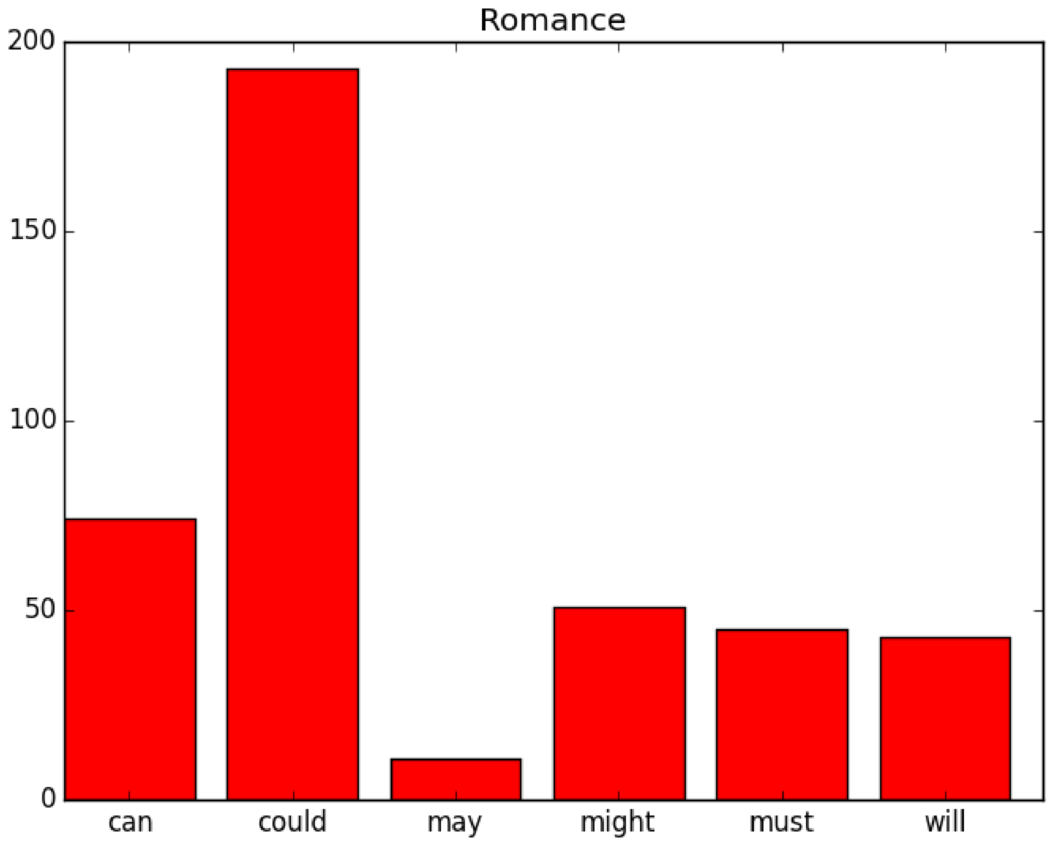
	can	could	may	might	must	will	Total
news	93	86	66	38	50	389	722
religion	82	59	78	12	54	71	356
hobbies	268	58	131	22	83	264	826
science_fiction	16	49	4	12	8	16	105
romance	74	193	11	51	45	43	417
humor	16	30	8	8	9	13	84
Total	549	475	298	143	249	796	2510

- Each row and each column is a frequency distribution
- We can calculate the relative frequency for each row
  - E.g. news: 93/722, 86/722, 66/722, etc.
- We can make a chart for each row and inspect the differences

Example continued

	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13



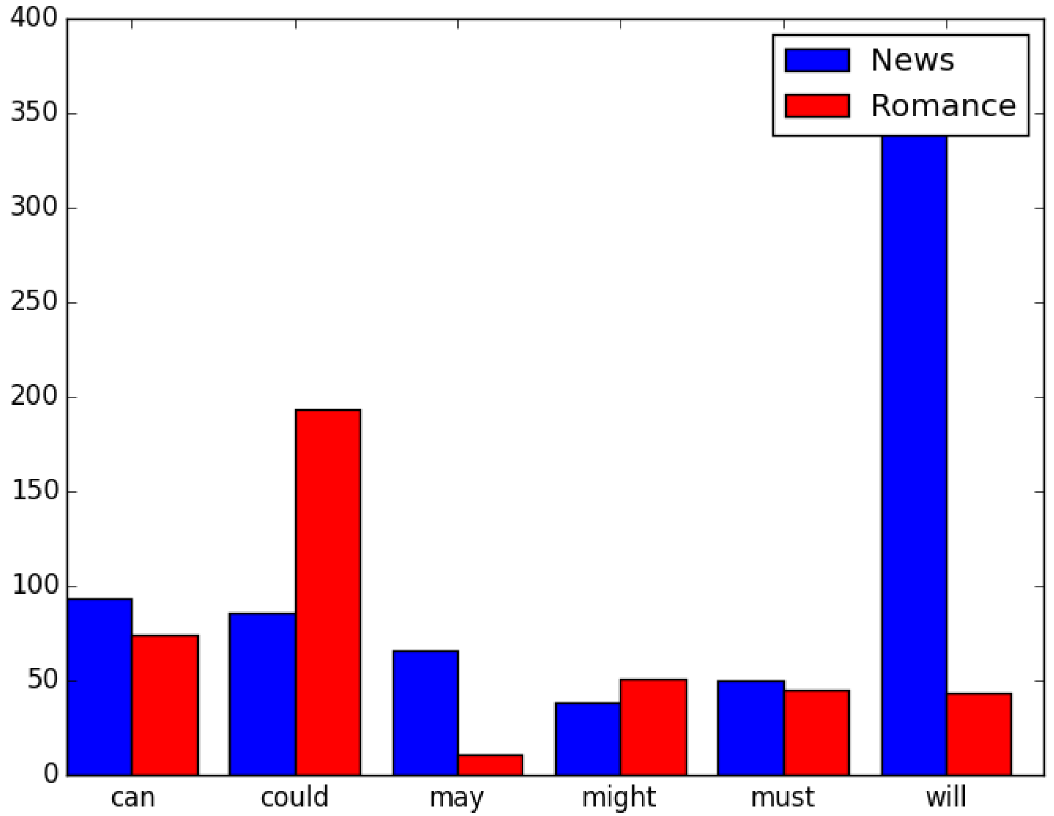


We see the same differences in pattern, the same shapes, whether we use absolute or relative frequencies

Example continued

	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

- Or we could color code to display two dimensions in the same chart
  - (In this chart it would have been more enlightening to use relative frequencies)

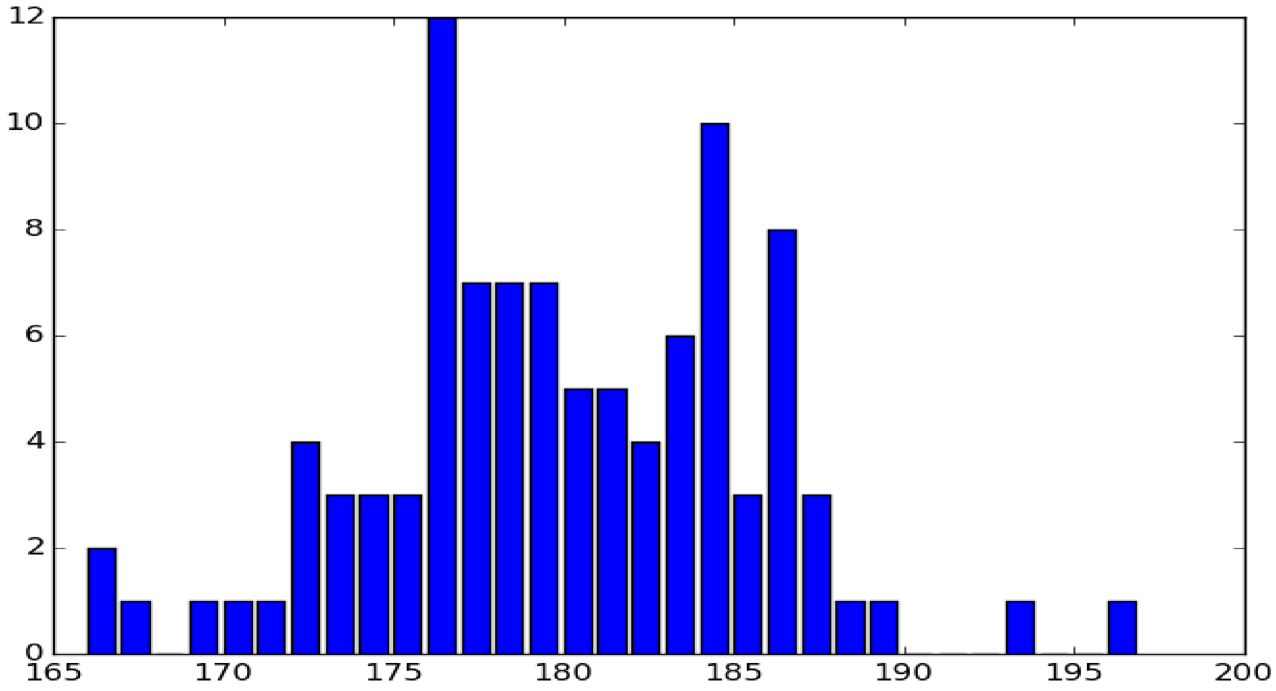


Numerical data

Numerical values

Ex 1

173 172 173 183 177 177 186 180 178 187 179 181 184 172 180 180 171 176 186 175 176 181 176 177 178 176 174 186 172 175 186 183 185 184 176 179 175 193 181 178 177 183 196 187 184 179 182 184 181 176 185 180 176 176 176 167 178 182 176 186 179 176 166 186 169 186 183 178 186 184 179 177 174 176 184 174 177 178 173 182 182 184 185 172 179 179 189 178 170 183 166 188 187 184 184 177 181 180 183 184



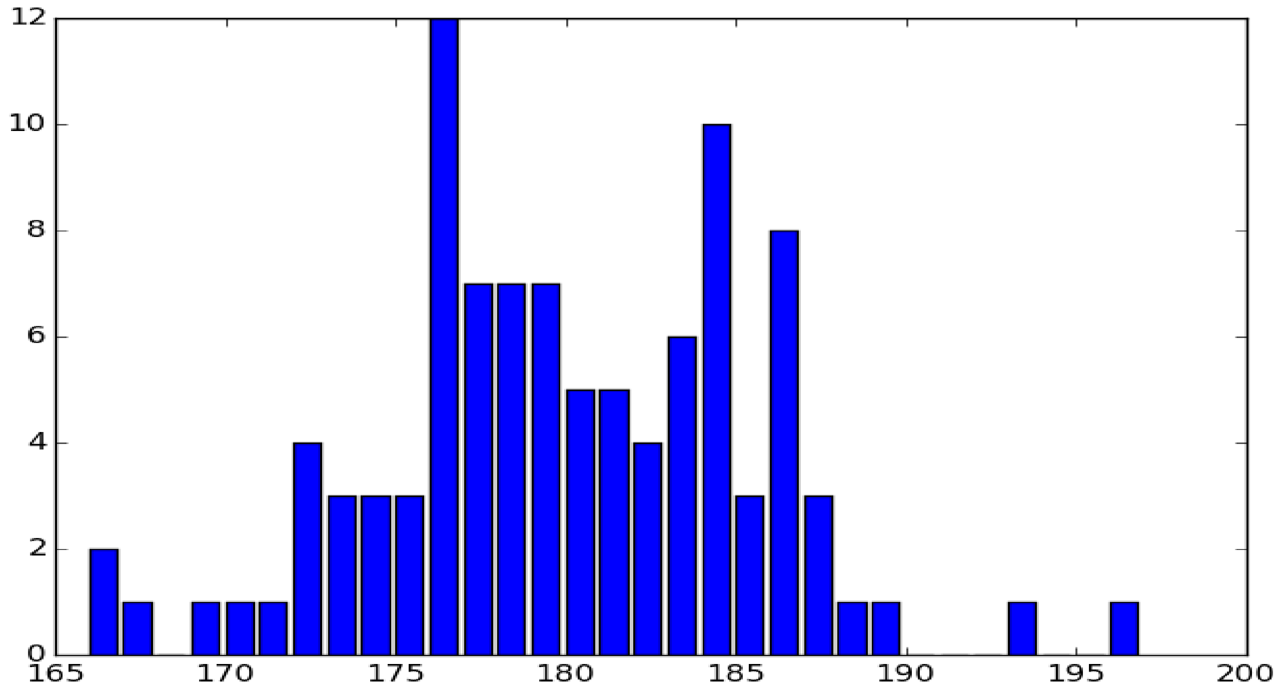
- With finally many different values, we may use
  - Table
  - Bar chart

  as for categorical data
- We will of course put the values in order

Numerical values

Ex 1

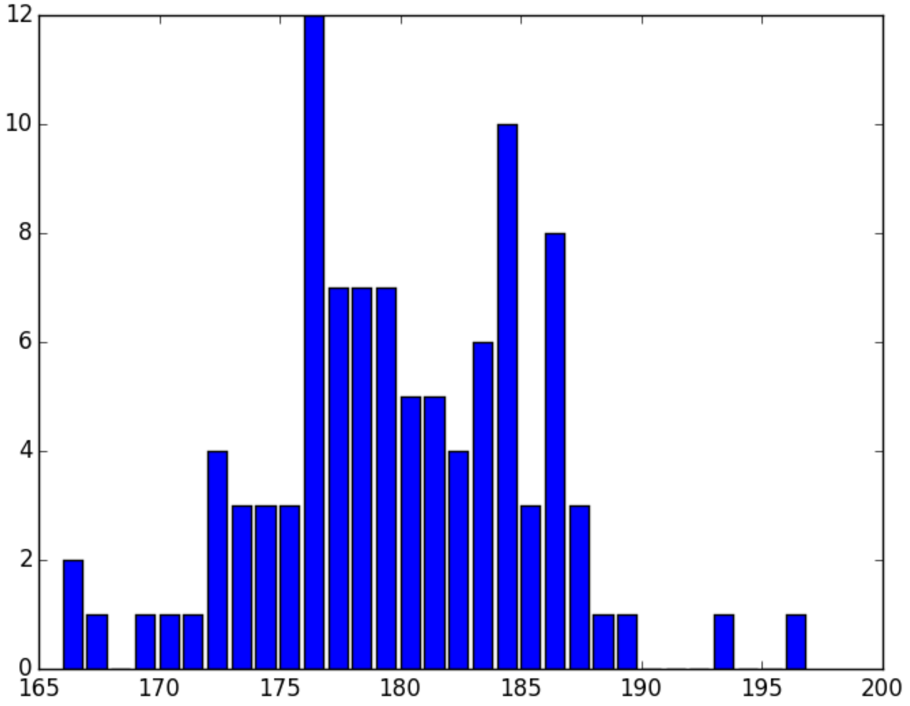
173 172 173 183 177 177 186 180 178 187 179 181 184 172 180 180 171 176 186 175 176 181 176 177 178 176 174 186 172 175 186 183 185 184 176 179 175 193 181 178 177 183 196 187 184 179 182 184 181 176 185 180 176 176 176 167 178 182 176 186 179 176 166 186 169 186 183 178 186 184 179 177 174 176 184 174 177 178 173 182 182 184 185 172 179 179 189 178 170 183 166 188 187 184 184 177 181 180 183 184



We may ask more questions:

- Max?
  - 196
- Min?
  - 166
- Middle, average?

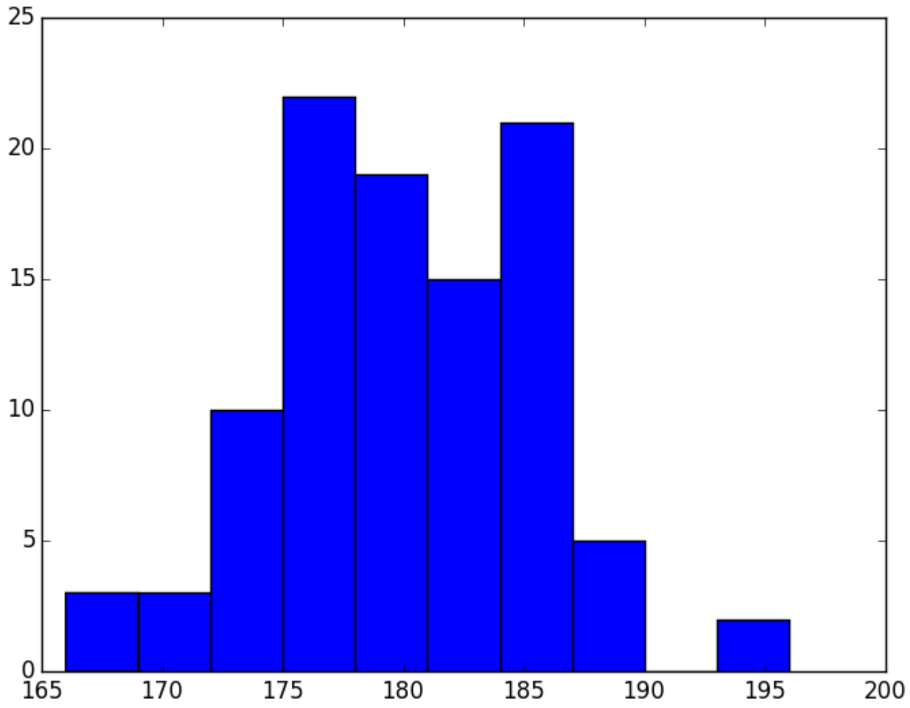
3 ways to define "middle", "average"



- Median (in the example: 179)
  - equally many above and below,
  - Formally, order $x_1, x_2, \ldots, x_n$, then
    - the median is $x_{(n/2)}$ if n is even and
    - $(x_{(n-1)/2} + x_{(n+1)/2})/2$ if n is odd.
- Mean: ex: 179.54
  - $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Mode, the most frequent one, ex: 176

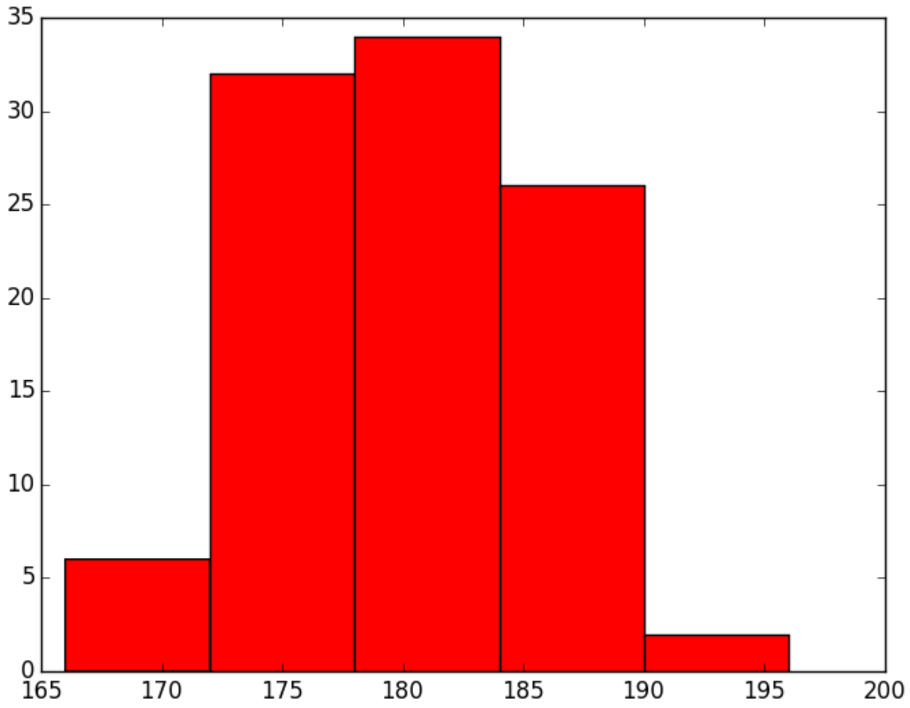
Histogram for numerical data

- Split the set of values into equally sized intervals
- For each interval, ask how many individuals take a value in it
- Over the interval, draw a rectangle with height proportional to this frequency
- The y-axis may be tagged with
  - Absolute frequencies
  - Relative frequencies, or
  - Densities (= absolute frequencies/elements in the interval)

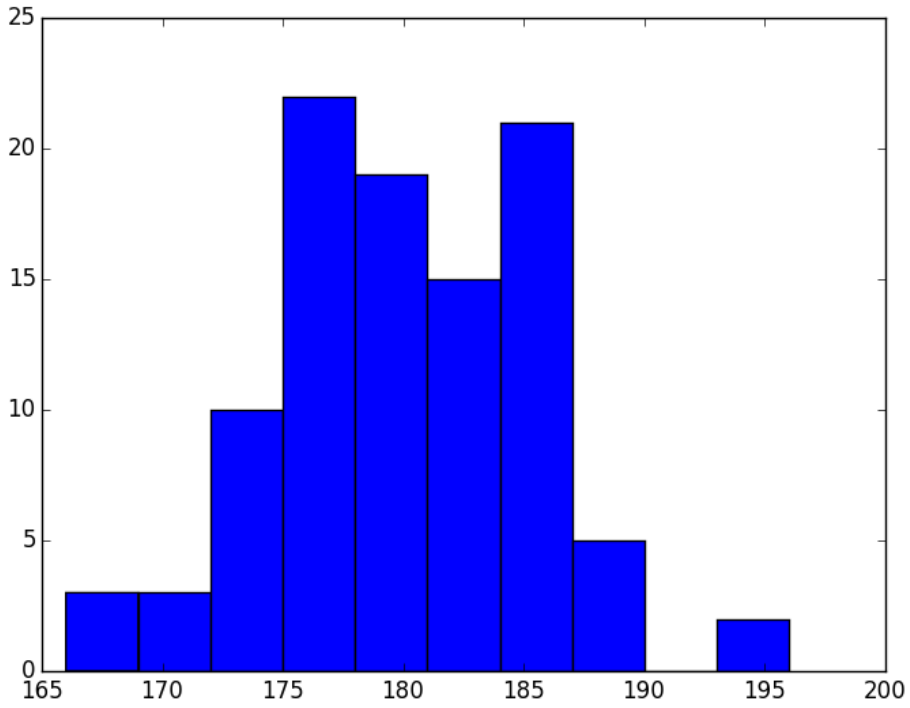


Ex 1: 10 bins

Histogram for numerical data



Ex 1: 5 bins

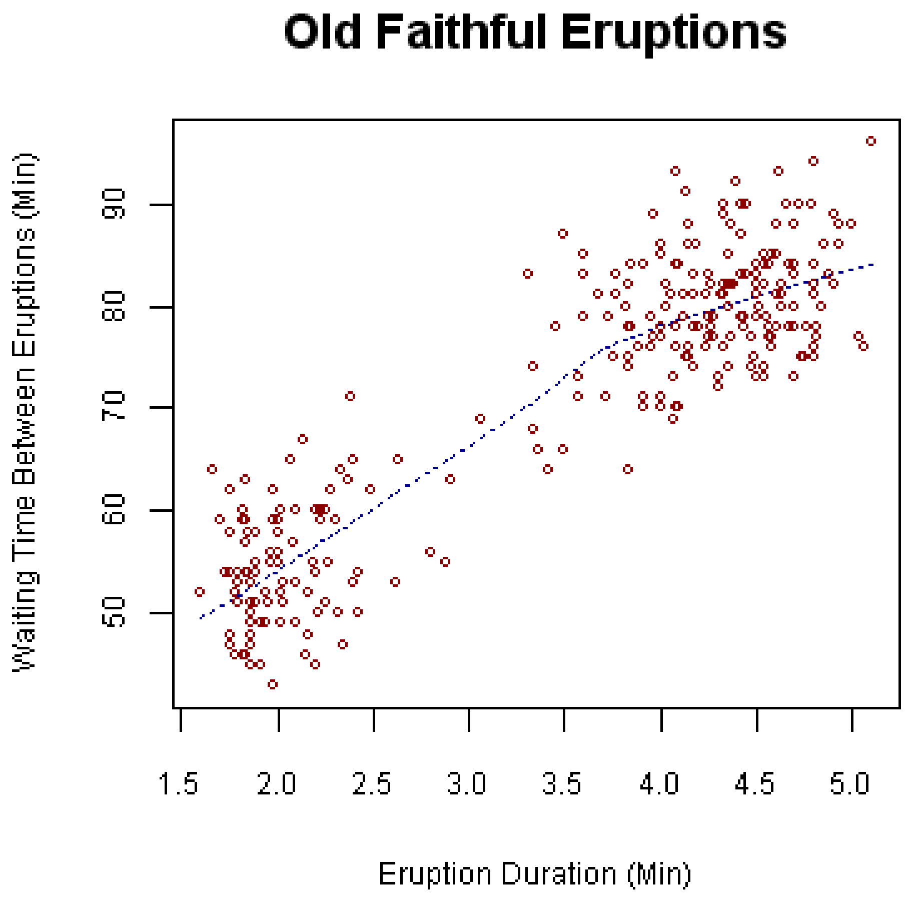


Ex 1: 10 bins

More than one numerical feature

Scatter plot

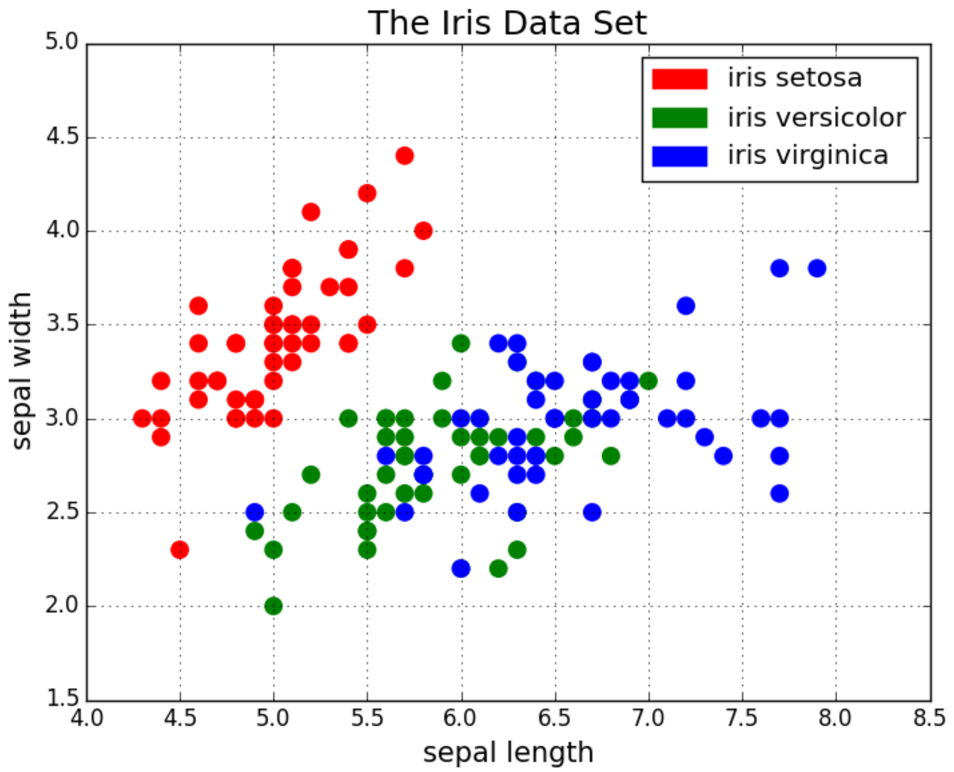
- When the objects have two numerical attributes, we may plot the pairs for each object in a coordinate system.
- Called a scatter plot
- A good way to visualize numerical data



https://en.wikipedia.org/wiki/Scatter_plot

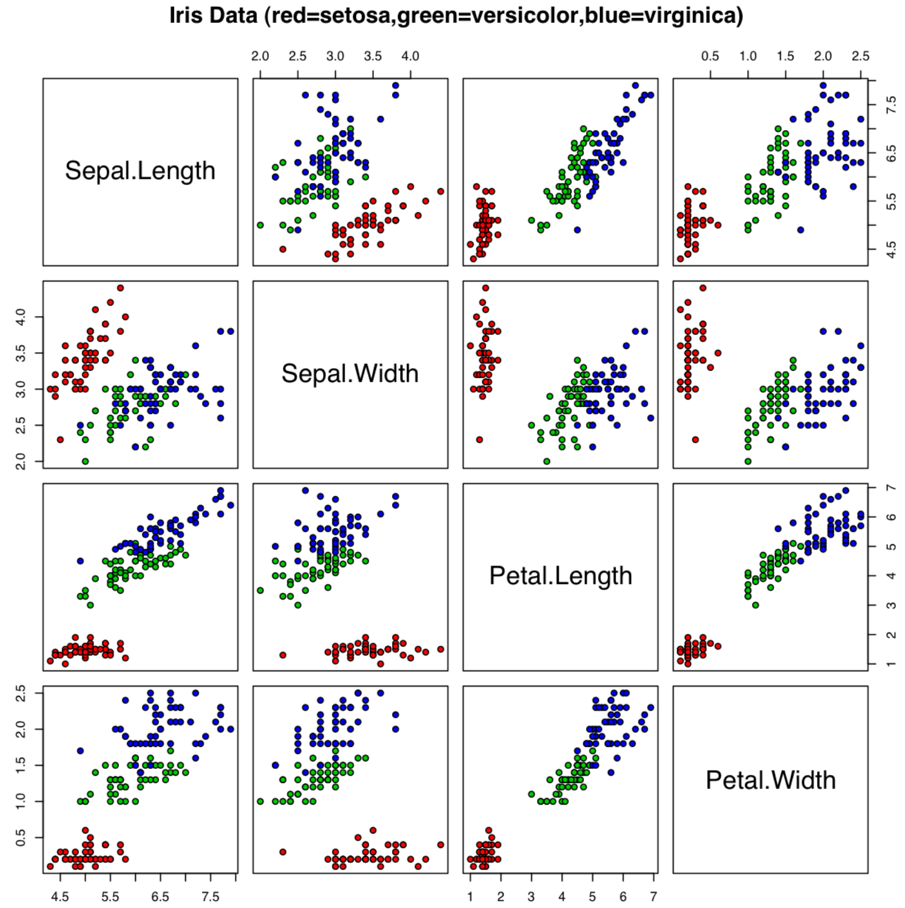
Scatter plot too

- Scatter plot with:
  - 2 numerical features
  - one categorical feature
- Use different colors – or symbols – to indicate categorical feature
- Common in machine learning to explain algorithms



More attributes

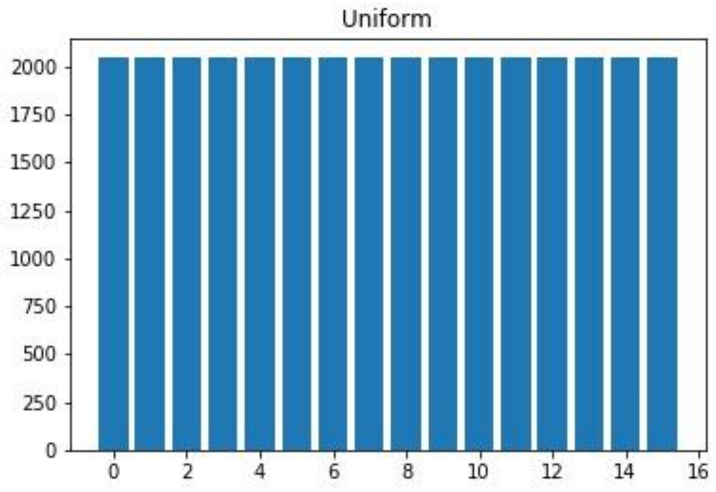
- A scatterplot only shows to numeric attributes
- With more attributes, we may use more plots
- (But there is a limit to informative they are with, say, 100 attributes).



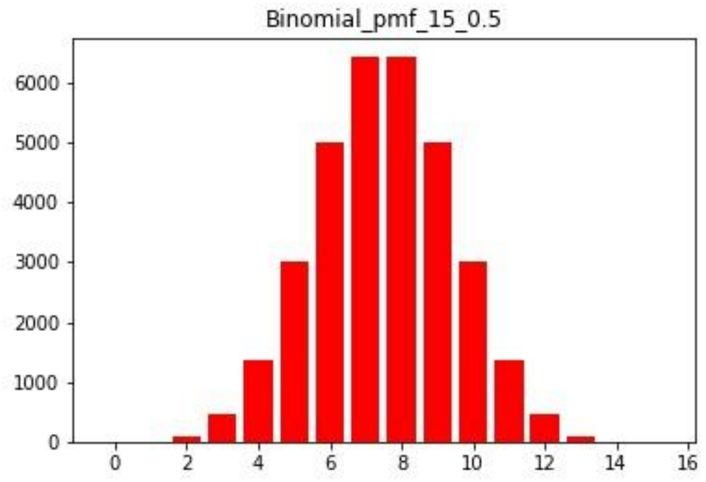
Dispersion

Dispersion

- Median or mean does not say everything
- Nor does max, mean or range (=max-min)
- Example:
  - Two sets
  - The same median=mean=4, min:0, max:8



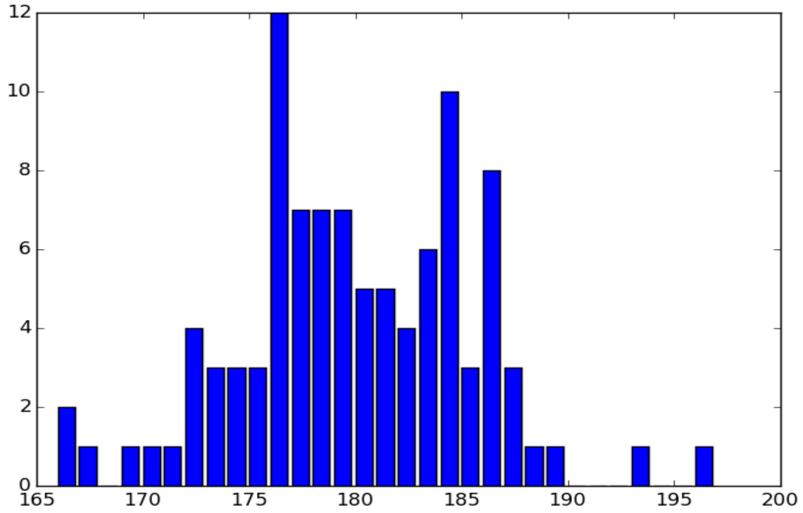
Ex 2: Uniform

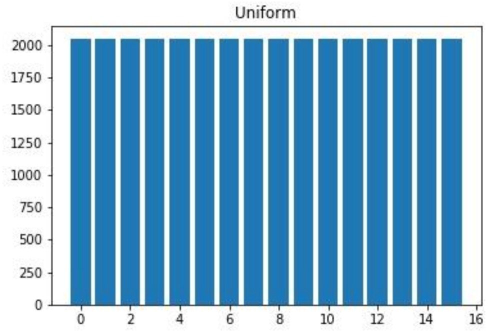


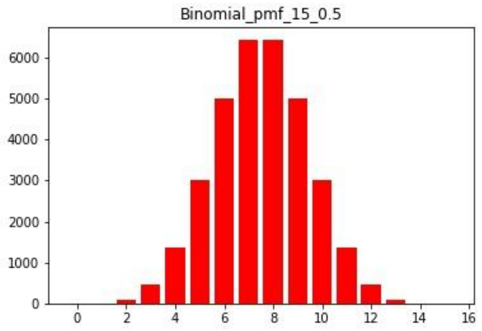
Ex 3: Binomial

Median, quartile, percentile (approach 1)

- The *n*-percentile *p*:
  - *n* percent of the objects are below *p*
  - (100–*n*) percent are above *p*
  - (where 0<*n*<100)
- Median is the 50-percentile
- Quartiles are the 25-, 50-, 75-percentiles
  - Split the objects into 4 equally big bins
  - Example 1: 176, 179, 184
  - Example 2: 3.75, 7.5, 11.25
  - Example 3: 6, 7.5, 9

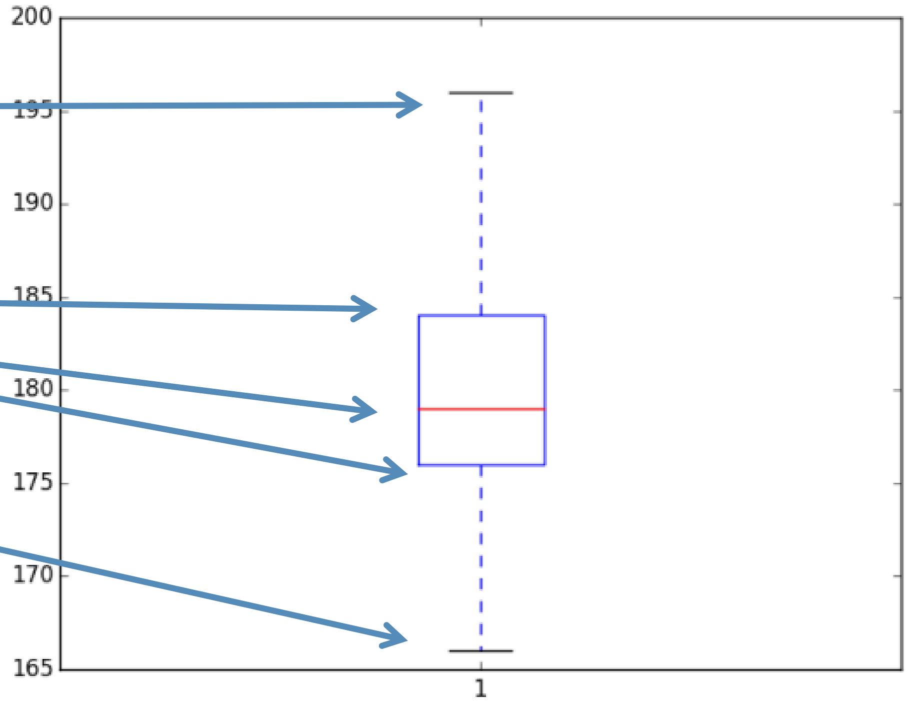


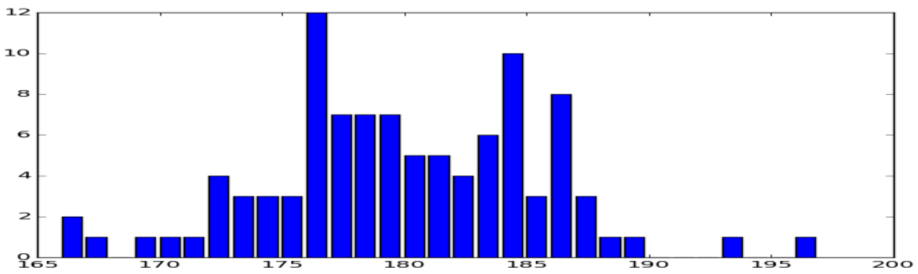




Boxplot

- Example 1:
  - Max 196
  - Quartiles:
  - 176, 179, 184
  - Min 166
- Also good for continuous data
- (The exact definition for the "end points" may vary when "outliers")





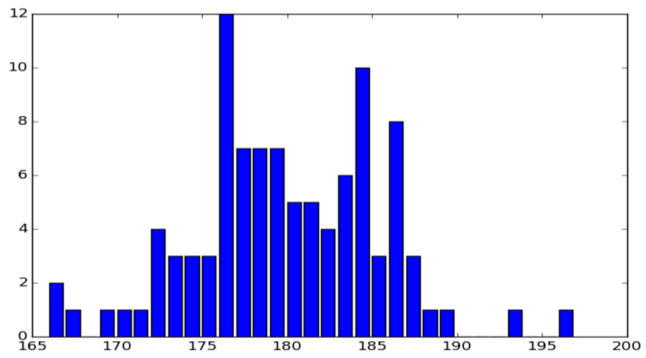
Variance (approach 2)

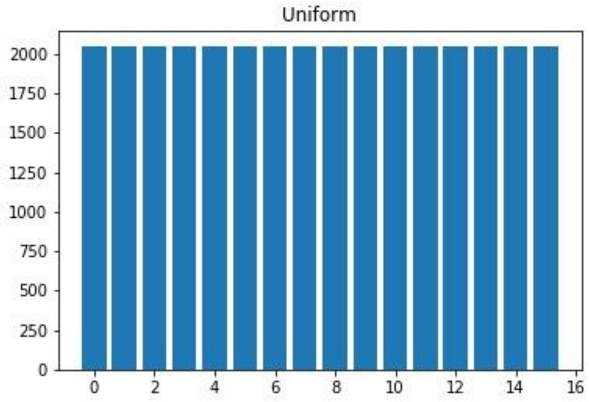
- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Variance: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Idea:
  - Measure how far each point is from the mean
  - Take the average
  - Square – otherwise the average would be 0
- Standard deviation: square root of the variance
  - "Correct dimension and magnitude"

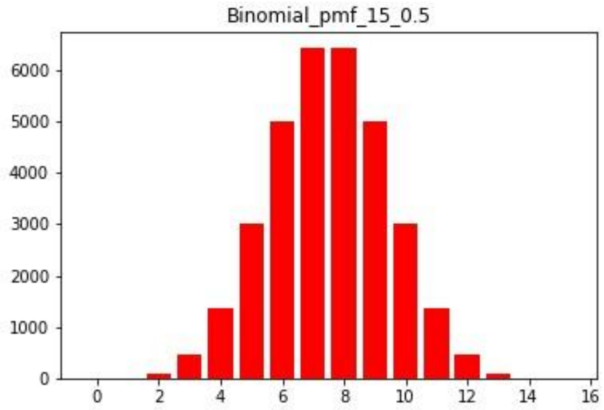
Beware:
For some statistical purposes one divides by (n-1) instead of n.

The examples

EX	Min	25%	Median	75%	Max	Mean	Vari.	s.d
1	166	176	179	184	196	179.54	30.33	5.5
2	0	3.75	7.5	11.25	15	7.5	21.21	4.61
3	0	6	7.5	9	15	7.5	3.75	1.94

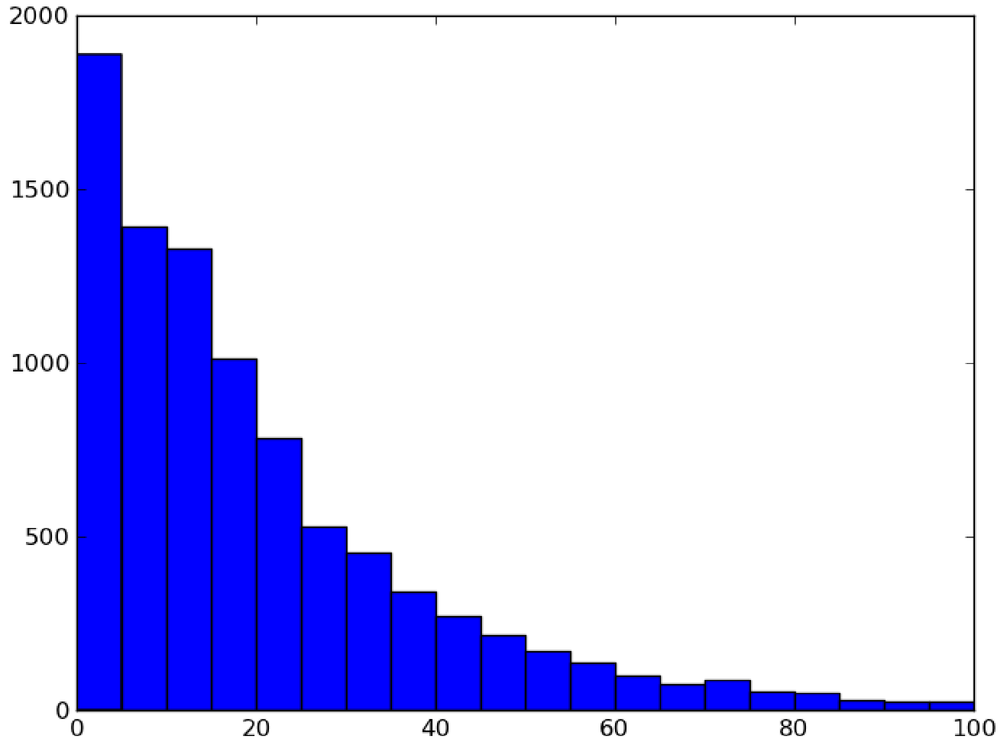






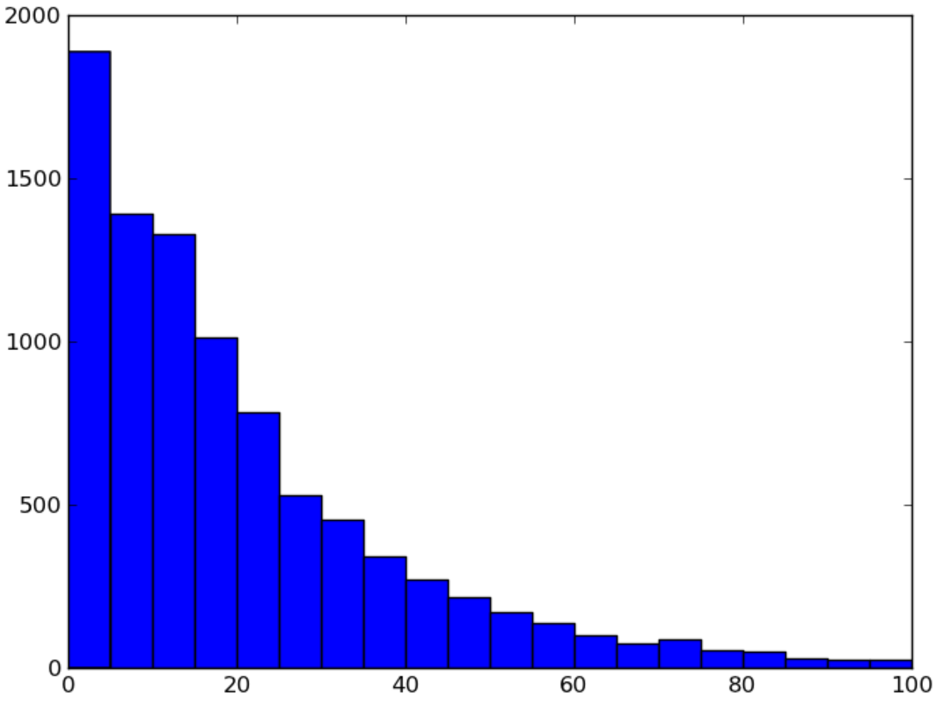
Example: sentence length

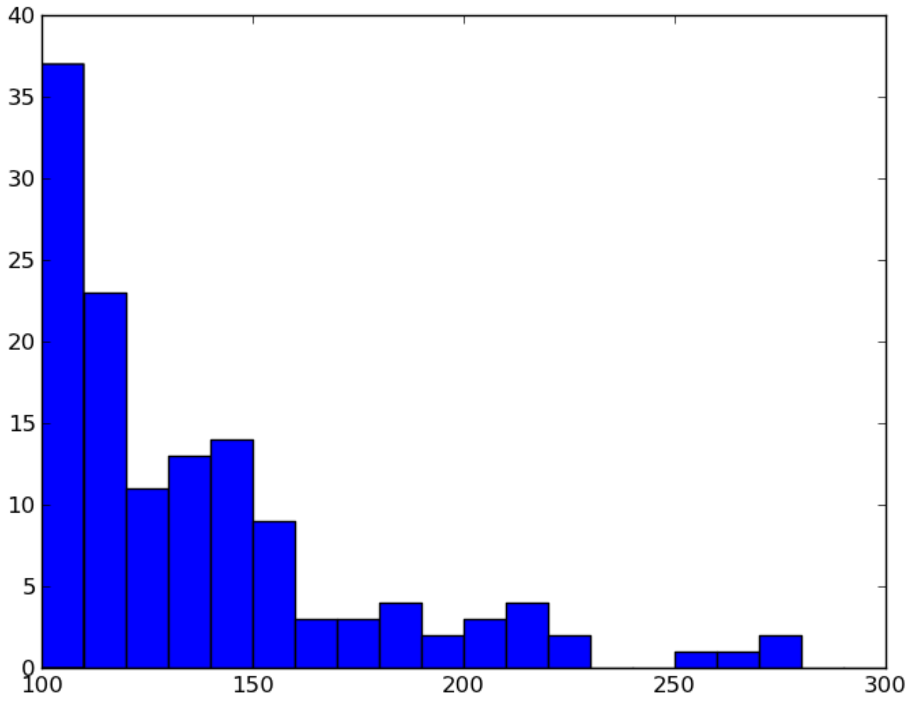
- NLTK: austen-emma.txt
- Number of sentences: 9111
- Length:
  - Min: 1
  - Max: 274
  - Mean: 21.3
  - Median: 14
  - Q1-Q2-Q3: 6-14-29
  - Std.dev.: 23.86



+…274

Example cntd.: the whole picture

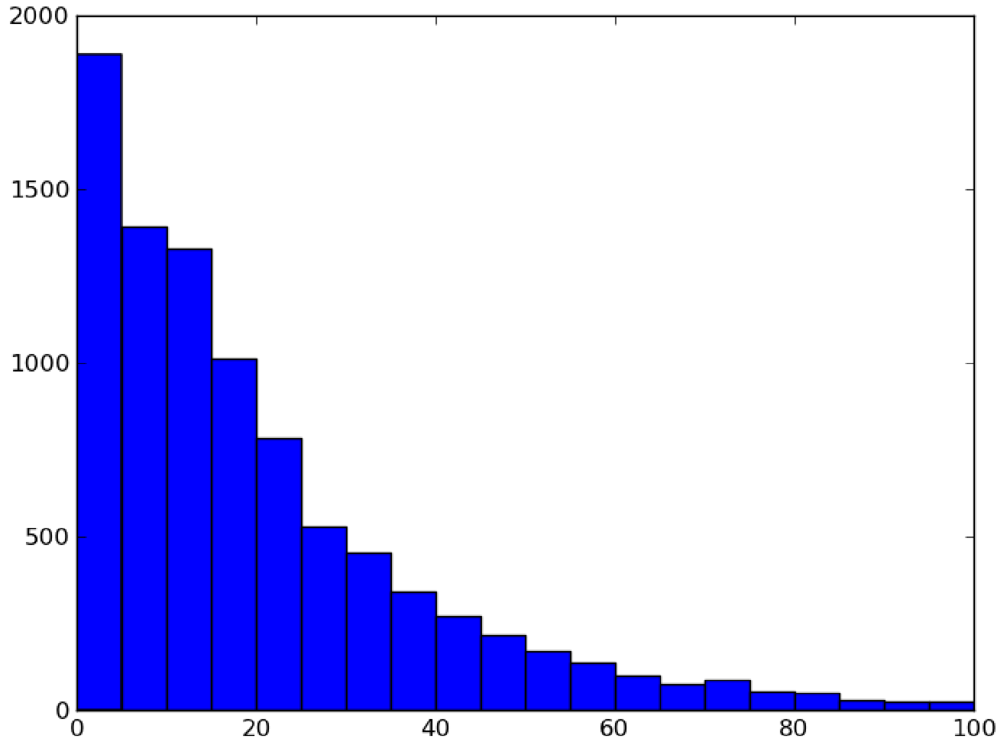




Observe: Different scales on the y-axes

Example: sentence length

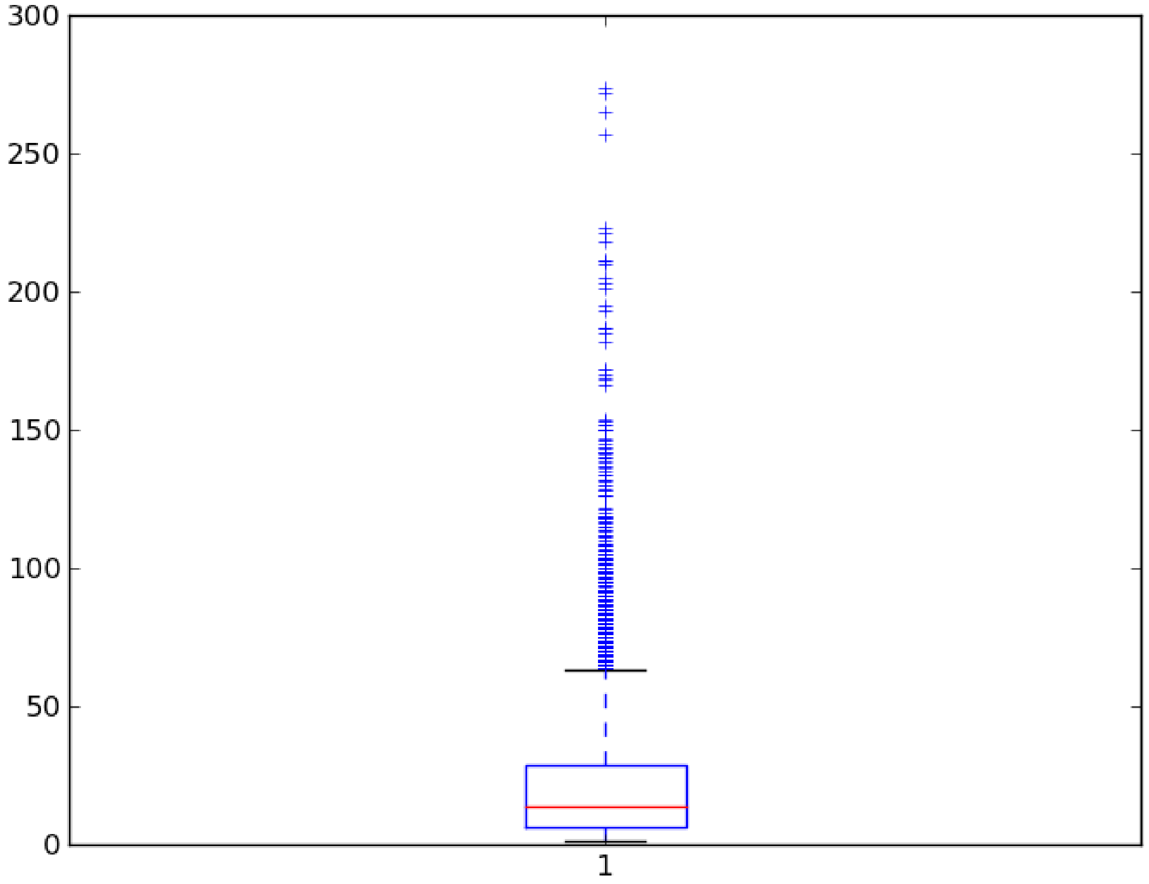
- NLTK: austen-emma.txt
- Number of sentences: 9111
- Length:
  - Min: 1
  - Max: 274
  - Mean: 21.3
  - Median: 14
  - Q1-Q2-Q3: 6-14-29
  - Std.dev.: 23.86



+…274

Example: sentence length

- NLTK: austen-emma.txt
- Number of sentences: 9111
- Length:
  - Min: 1
  - Max: 274
  - Mean: 21.3
  - Median: 14
  - Q1-Q2-Q3: 6-14-29
  - Std.dev.: 23.86



Boxplot with outliers

Take home

- Statistical variables:
  - Categorical
  - Numerical
    - Discrete
    - Continuous
- Frequencies
- Median
  - Quartiles, percentiles
- Mean
  - Variance
  - Standard deviation
- Tables
  - Contingency table
- Bar chart
- Histogram
- Scatter plot
- Boxplot