Data Science

Data Exploration and Visualisation

Alexandra Posekany

Summerschool 2023

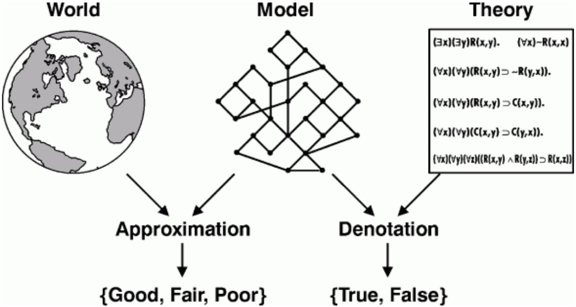
Statistics - better than its reputation?

Some Citations to take away

> There are three kinds of lies:
> lies, damned lies, and statistics.
>
> [Mark Twain (referring to Benjamin Disraeli)]

> And thirdly, the code is more what you'd call
> "guidelines" than actual rules.
>
> [Barbossa (Pirates of the Carribbean)]

Building Models


World
Model
Theory
$(\exists x)(\exists y)R(x,y).$
$(\forall x)\sim R(x,x)$
$(\forall x)(\forall y)(R(x,y) \supset \sim R(y,x)).$
$(\forall x)(\forall y)(R(x,y) \supset C(x,y)).$
$(\forall x)(\forall y)(C(x,y) \supset C(y,x)).$
$(\forall x)(\forall y)(\forall z)((R(x,y) \wedge R(y,z)) \supset R(x,z))$
Approximation
Denotation
{Good, Fair, Poor}
{True, False}

Moral of the story

Essentially, all models are wrong,
but some are useful.

[George Box]

Models are always a simplification of the real world, translating structures and processes into mathematical formulations in order to make them computationally tractable. In addition to the simplification error, a stochastic error inhibits all measurements.

Goals of Statistics in Data Science

We divide quantiative methods in Statistics according to their main purposes:

1. **Descriptive** Analysis: Organising and summarising data

exploratory data analysis, data visualisation, summary statistics, sample estimators

2. **Inferential** Statistics: Analysing the data in order to answer research questions, explain relationships between variables or perform forecasts

testing hypotheses, building a model for explanation and prediction (forecasting)

An Overview of Measurement Scales

Measurement: Mapping of observable phenomena to numbers.

- **categorial** variables; discrete categories
  - **nominal**; categories without any ordering

    examples are: male/female; different products in marketing; colours; etc.
  - **ordinal**; ordered categories

    Credit Ratings (AAA – D); Quality categories; grades
- **metric**; real valued measurements
  - **interval**; ordering matters, differences matter $(-\infty, \infty)$

    examples are: temperatures, companies' sales, years

    only differences, but NOT multiples can be interpreted

    -10°C cannot be interpreted as twice the temperature of -5°C

    The 2000 AD is not twice the year 1000 AD. What would 5000 BC then mean?
  - **ratio**; ordering and ratios matter $(0, \infty)$, absolute zero

    examples are: incomes, heights, lengths, expenditures, ...

    Here, multiples and differences can be interpreted.

    2m is twice the length of 1m. 2m are 1m more than 1m.

Measurement Scales: Examples

Classification into data types

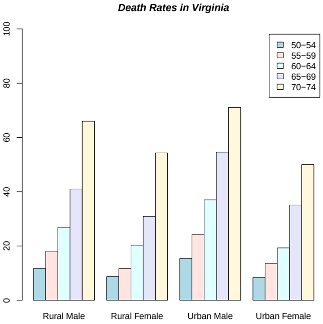
Item	**possible values**	**n/o/m**	**d/c**
Germ types	Bacteria (=1) Fungi (=2), Viruses (=3), etc.	nominal	discrete
Credit Ratings	AAA (=1), AA (=2), . . . , D (=18)	ordinal	discrete
logarithmic concentrations (pH)	1.40, 6.32 EUR, 9.6 EUR, . . .	metric (ratio)	contin.
Years	1999, 2012, . . .	metric (interval)	discrete
cell counts	20,500; 4,746; . . .	metric (ratio)	discrete
Temperatures	37.0°C, 38.1°C, . . .	metric (interval)	cont.

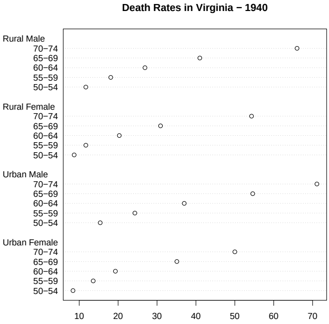
Data Exploration and Visualisation of Frequentist statistics

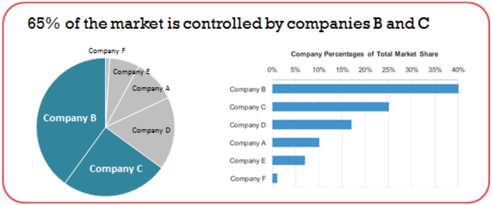
Frequency plots

- **bar charts** visualise absolute or relative frequencies of categories (recommendable in 90% of cases)
  human eye and brain discern lengths better than angles
  ‘‘‘r barplot(table(xcat)) ‘‘‘
- **Cleveland dot charts** visualise absolute or relative frequencies of categories
  alternative to bar charts
- **pie charts** visualise relative frequencies
  only when visualising majorities (in order to make a pie chart fully interpretable, always add all percentages)

Frequency Plots





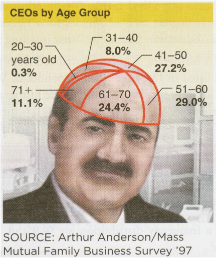


[1]

[1] http://speakingppt.com/2013/03/18/why-tufte-is-flat-out-wrong-about-pie-charts/

When pie charts are useful

When pie Charts are evil - Meet the worst pie chart EVER



CEOs by Age Group

SOURCE: Arthur Anderson/Mass Mutual Family Business Survey '97

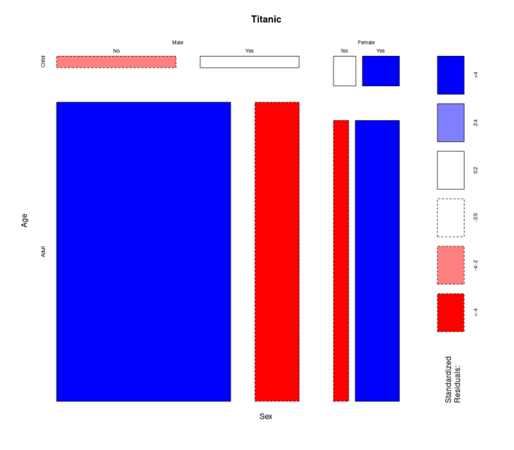
Please note how 24.4% and 27.2% cover different areas and what this visually suggests about their respective relations.

The right one is the best way to use pie charts properly.

Visualising contingency tables

Mosaic plots

Visualises contingency tables as blocks in matrix where the area corresponds to the absolute frequency of the category, defined by co-occurrence of two or more events.



Example: Skin color and death sentence

The Data and scenario

The following example may appear not politically correct at first sight, but it is a famous example for misinterpretations of tabulated data.

The following Data come from the *New York Times Magazine*, March 11, 1979. Originally, they were published concerning the frequency of death sentences in Florida. They led to nationwide discussions and questioning of statistics.

Watch out not to repeat such mistakes in your own research!

Case no.	Skin color of accused	Death sentence
1	b	0
2	b	0
3	w	0
4	b	1
⋮	⋮	⋮
4764	w	0

The true story?

	black skin	white skin	Σ
Death sentence	59	72	131
No death sentence	2448	2185	4633
Σ	2507	2257	4764
Proportion in %	2.4	3.3	2.8

This table was published and produced an uproar amongst those looking for racial discrimination. According to overall summaries more caucasian accused had been sentenced to death than African Americans. To learn why one should tread carefully when dealing with conditional probabilities will we extend the example by additional information.

The true story!

Let us now also look at the skin-color of the *victim* and construct a three way cross tabulation:

Skin-color of victim	black		white	
Skin-color of accused	b	w	b	w
Death sentence	11	0	48	72
No death sentence	2209	111	239	2074
Sum	2220	111	287	2146
Proportion in %	0.5	0.0	20.1	3.5

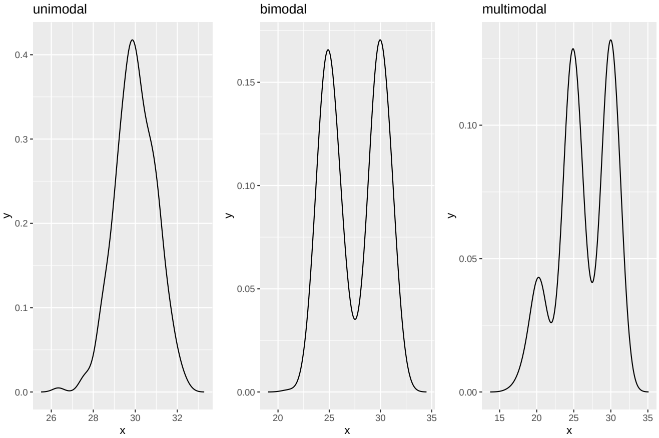
This table changes the whole picture the previous table showed us. An African American accused of murder of a white person was sentenced to death more than 20 % of the time, not a single white accused of murder of an African American victim was sentenced to death. However, as most victims and accused share the same skin colour these extreme effects are covered by the majority of sentences.

This effect is called **Simpson's paradox** which happens when subgroups are extremely unbalanced and show opposing effects to the majority.

Characteristics of data

1. **Modality.** Do the data have a single "center" or do they consist of several different parts with different characteristics? The number of "peaks" determines the modality.
2. **Center.** The location of the middle of the data illustrated by a representative or average value. This makes sense for unimodal data or for multimodal data in each part separately.
3. **Variation.** A measure of the variability of measurements with respect to its central location value. This makes sense for unimodal data or for multimodal data in each part separately.
4. **Distribution.** How the spread of the data is shaped and behaves. Possibly, we consider whether data are symmetric or asymmetric. In addition we consider "peakedness" or in other words the weight of the tails which is the amount of data further away from the center than would be expected for Gaussian data.
5. **Outliers.** Values of the sample showing a behaviour which differs from the rest of the sample, often but not always are those values located very far away from the vast majority of the other sample values. Sometimes, values far from the center have close neighbours which have a similar behaviour which are not necessarily outliers.

Modality



unimodal
0.4
0.3
0.2
0.1
0.0
y
26
28
30
32
x
bimodal
0.15
0.10
0.05
0.00
y
20
25
30
35
x
multimodal
0.10
0.05
0.00
y
15
20
25
30
35
x

Rank and Order Statistics

Most important order statistics:

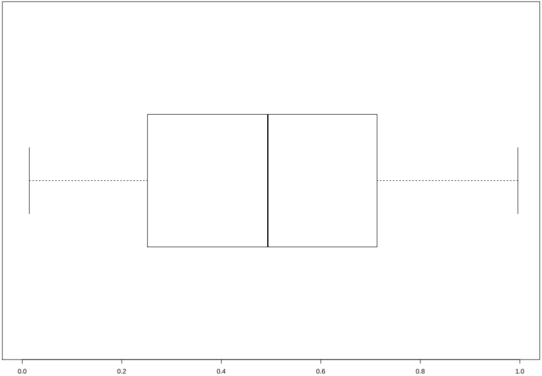
- minimum (rank=1),
- maximum (rank=sample size);
- range=(maximum-minimum);
- quantiles
  - The best known quantile is the **median**, which is the 50%-quantile.
  - The **quartiles** refer to the 25% and 75% which in addition to the median spread the data in quarters. A robust measure for variation is based on the quartile, the inter-quartile range.

$$IQR = x_{0.75} - x_{0.25}$$

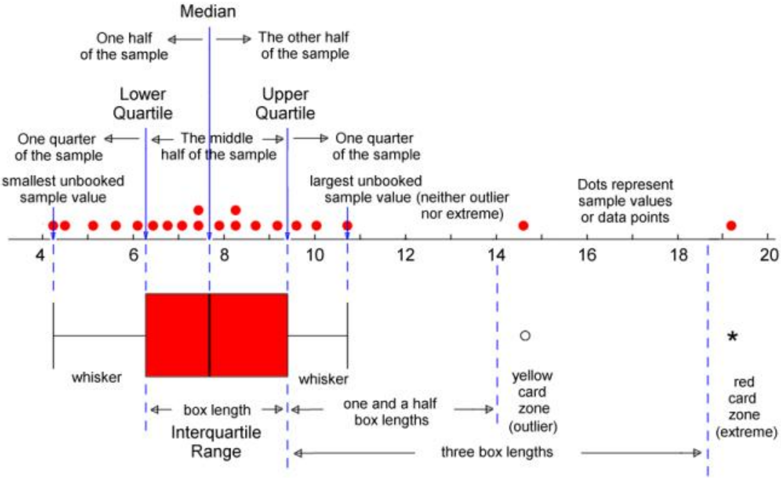
Five Number Summary = Box of the Boxplot

The **boxplot** is the visualisation of the most important quantiles, the quartiles which divide the data into four parts with an equal number of observations contained.

Minimum	1st Quartile	Median	3rd Quartile	Maximum
x_0	$x_{0.25}$	$x_{0.5}$	$x_{0.75}$	x_1
0.0142	0.2565	0.4936	0.7110	0.9960

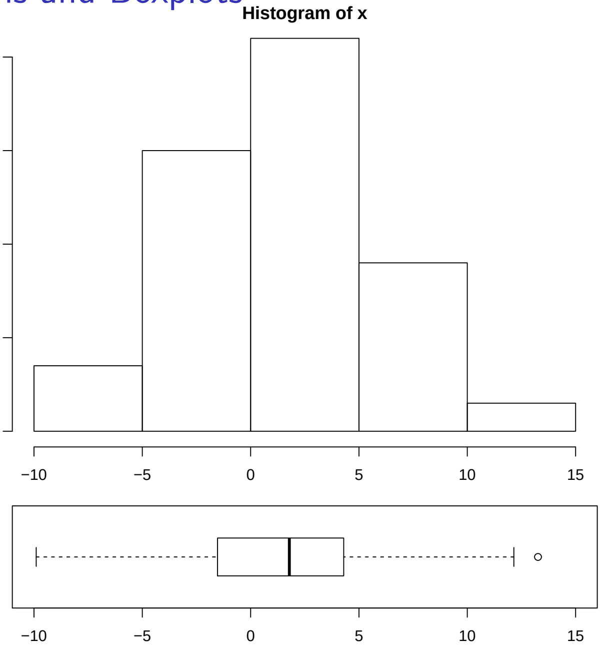


Extended Boxplot


Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
The middle half of the sample
One quarter of the sample
smallest unbooked sample value
largest unbooked sample value (neither outlier nor extreme)
Dots represent sample values or data points
4
6
8
10
12
14
16
18
20
whisker
whisker
box length
Interquartile Range
one and a half box lengths
yellow card zone (outlier)
red card zone (extreme)
three box lengths

\end{itemize}

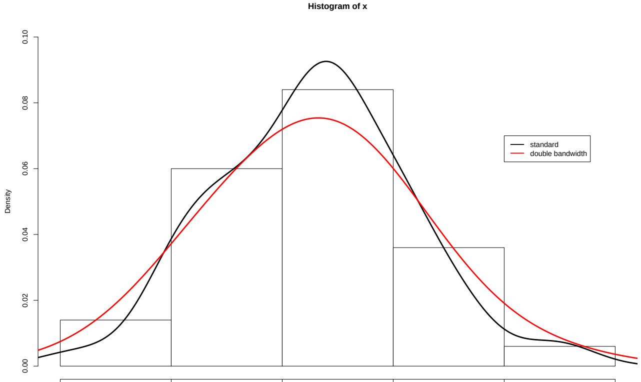
Histograms and Boxplots



Kernel density estimation

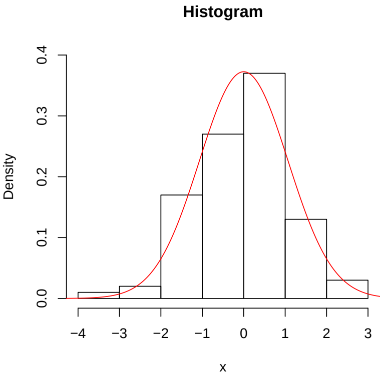
Kernel density estimator
provide a smooth estimate of the density (≠ histogram is discretised) with a smoothing parameter h and kernel K_h

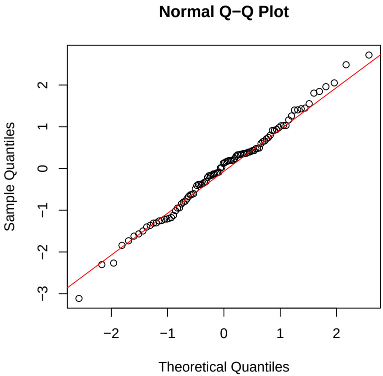
$$\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$



Example: Assessing normality

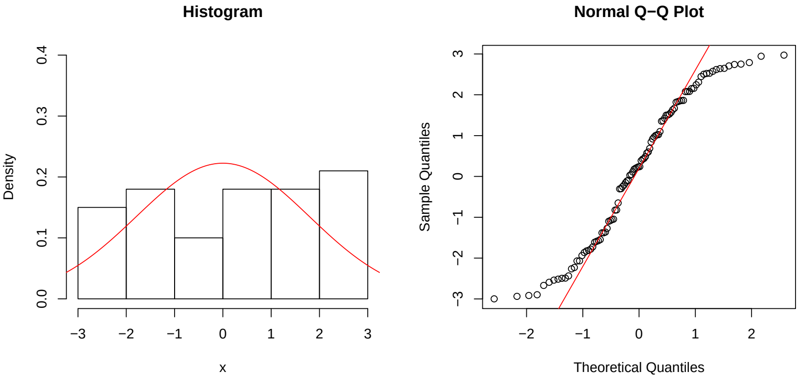
100 draws from a standard normal distribution:





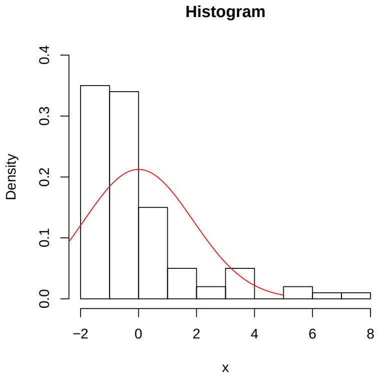
Example: Assessing normality

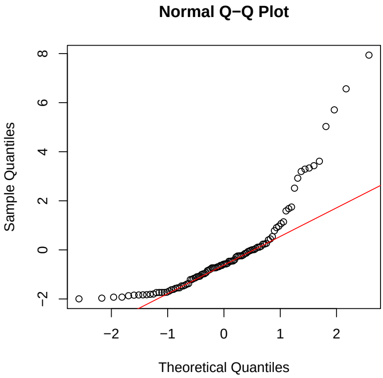
100 draws from a uniform distribution on [-3,3]:



Example: Assessing normality

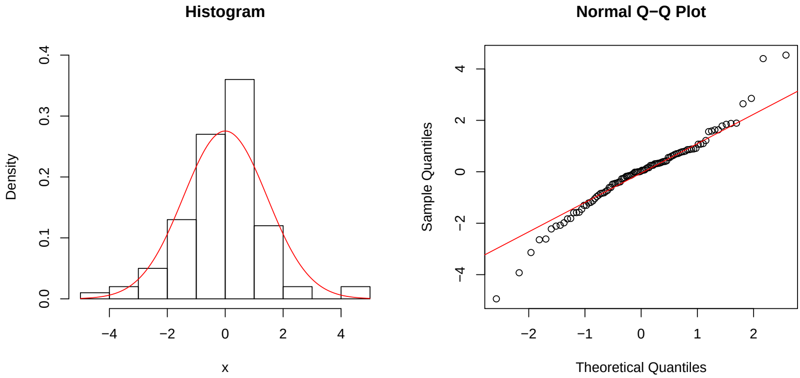
100 draws from an exponential distribution:


Histogram


Normal Q−Q Plot

Example: Assessing normality

100 draws from a t distribution with 2 degrees of freedom:



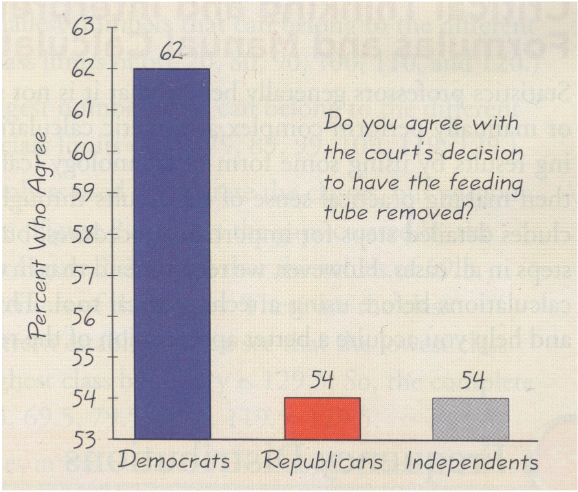
The guide for every PhD student how NOT to present your data

- Display as little information as possible, your readers, advisors and co-authors will thank you.
- Obscure what you do show (with chart junk). If one extra-large legend is not enough, add a superfluous label or two hiding your graphs.
- Use pseudo-3d and color gratuitously - the human eye is so easy to distract.
- Make a pie chart (preferably in color and 3d) - please refer to the worst pie chart ever for how to do this properly.
- Use a poorly chosen scales to trick the hman eye into interpreting multiples in the completely wrong way.

Spoiler Warnings: Please take none of this seriously - it's sarcasm. [2]

[2] www.biostat.wisc.edu/ kbroman/

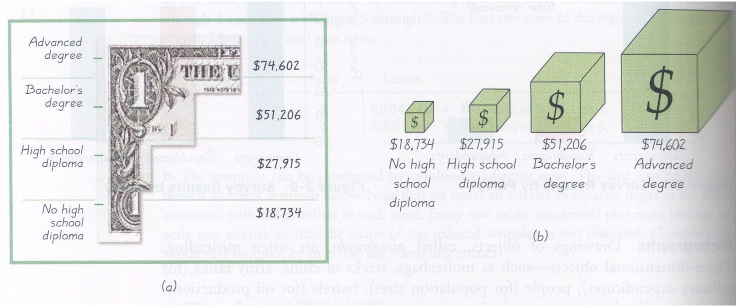
How not to present your data - Visualised



from Triola: Essentials of Statistics (2011)

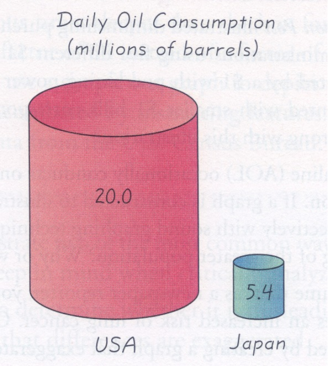
Enjoy the difference between 62% und 54% and the visual difference when the eye interprets lengths as multiples, if you cut your axis at 53%.

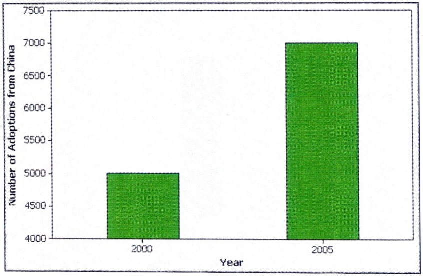
How not to present your data - Visualised



from Triola: Essentials of Statistics (2011)

How not to present your data - Visualised





from Triola: Essentials of Statistics (2011)