The data analysis cycle

map1_desc.pdf

map1_knowhps-converted-to.pdf

map1_updown_arrs-eps-converted-to.pdf

map1_ext_prep

map1_mouseclicks-converted-to.pdf

Data Analysis:
Data Preparation and Description

Technological University Dublin Tallaght Campus

Department of Computing

Data

- Data can be structured or unstructured
- Unstructured data is often transformed into structured data or can at least be abstracted as such
- In this module, the term 'data' will be used as meaning 'structured data', unless indicated otherwise
- Structured data can be shown in a table

Generalized data table

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{\ldots}$	$\boldsymbol{x}_{\mathrm{p}}$
i_1	x_{11}	x_{21}	$\ldots$	$x_{\mathrm{p}1}$
i_2	x_{12}	x_{22}	$\ldots$	$x_{\mathrm{p}2}$
.	.	.		.
.	.	.		.
.	.	.		.
i_{n}	$x_{1\mathrm{n}}$	$x_{2\mathrm{n}}$	$\ldots$	x_{pn}

Data table: terminology in statistics (S), databases (DB) and machine learning (ML)

S: Variables ('independent', 'dependent')
DB: Columns, attributes
ML: Features (some of which are target ones)

S: Observations
DB: Rows,
records,
examples
ML: Instances

The terms are used across discipline boundaries as well!

	x_1	x_2	$\dots$	x_p
i_1	x_{11}	x_{21}	$\dots$	x_{p1}
i_2	x_{12}	x_{22}	$\dots$	x_{p2}
.	.	.		.
.	.	.		.
.	.	.		.
i_n	x_{1n}	x_{2n}	$\dots$	x_{pn}

Observed values, values

'Independent' vs. 'dependent': Let's say we are looking at happiness. Respondents are asked to record how they feel every day for a year, because we are studying seasonal variations and the effect of sunlight on the feeling of happiness. The independent variables here are 'amount of sunlight' (which we can obtain from meteorological records) and 'time of year'. The dependent variable is 'happiness'. However, the two 'independent' variables are not really independent of each other (in the summer there is more sunlight). *'Independent' in this context simply indicates the relationship with the 'dependent' variable*, which, by the same token, may be found not to depend on the 'independent' variables at all.

Variable types

- **discrete** (e.g. natural numbers) or **continuous variables** (e.g. real numbers)
- **categorical** or **numeric**
- classification according to scale:
  - **nominal scale**: categorical variable, no ordering among the possibe values (e.g. 'telecomms industry', 'pharmaceutical industry' etc.)
  - **ordinal scale**: categorical variable, values that can be ordered but without quantification of differences (e.g. low, medium and high)
  - **interval scale**: numeric variable, intervals between values can be compared but not ratios (e.g. temperature values $5°$C, $10°$C and $15°$C)
  - **ratio scale**: numeric variable, both intervals and ratios between values can be calculated (e.g. prices €5, €10 and €15)
- dichotomous: the variable can take only two values and if the values are 0 and 1 the variable is referred to as *binary* (e.g. 'customer has defaulted on their mortgage' or 'has not defaulted')
- variables not used in analysis (e.g. customer id)

Data table: pets for adoption

Name	Species	Breed	Gender	Age	Vaccinated
Axel	Cat	Common	M	around 4	NO
Baz	Dog	Lurcher	M	2	YES
Clodagh	Cat	Persian	F	10	YES
Dan	Cockatoo	N/A	M	5	?
Emma	Dog	Staffie	F	1.5	YES

Data preparation

- Integration of data from multiple sources
  - removal of duplicate entries
  - unit conversion
  - normalisation

HOWTO

Min-max normalisation

Sometimes it is necessary to 'shift and rescale' a variable's data to a different range from the one they are originally provided in. This is typically done when a method requires normalised data, e.g. z-score for a hypothesis test, or when joining data sets that have the same variable specified on different scales (e.g. different units of measurement). The operation is called **min-max normalisation**. The formula to apply for each value is:

$$x_{i(NEW)} = \frac{x_{i(OLD)} - x_{min(OLD)}}{x_{max(OLD)} - x_{min(OLD)}}(x_{max(NEW)} - x_{min(NEW)}) + x_{min(NEW)}$$

Example:

Let's say we have the set of values 2, 4, and 7 ($x_1 = 2$, $x_2 = 4$ and $x_3 = 7$) and are required to normalise it into the range 0-1, the new values would be:

$$x_{1(NEW)} = \frac{2-2}{7-2}(1-0) + 0 = 0, \quad x_{2(NEW)} = \frac{4-2}{7-2}(1-0) + 0 = 0.4, \quad x_{3(NEW)} = \frac{7-2}{7-2}(1-0) + 0 = 1$$

The new set is 0, 0.4 and 1.

NOTES:

- When working with samples, the minimum and maximum values need to be those known for the population, rather than for the sample. Let's say we have data for exam marks given in percentages: 45, 67, 88 and 91. We know that the minimum and maximum in this case are 0 and 100 and we use those, rather than 45 and 91, for $x_{min(OLD)}$ and $x_{max(OLD)}$.
- The same technique can be used in any case where the scaling factor and a pair of corrsespondent source and target rangeg are known, even if no minimum and maximum don't exist. For example, converting temperatures from degrees celsius to farenheit involves a new 'minimum' of 32 and an old to new range mapping of 100 to 180: $temperature(°F) = \frac{temperature(°C)}{100} \times 180 + 32$

- Cleaning
  - resolution of ambiguities and errors (e.g abbreviations vs. full names or non-numeric values where numeric is expected, such as 'about 4')
  - removal or correction of invalid outliers, caused by:
    * erroneous entry (typing error)
    * a mistake in measurement (e.g. a traffic counter reports a flow of 30 cars per second, which indicates a fault)
    * mixing of different measurement units (e.g. an adult's weight recorded as 55lbs may in fact be in kilograms)
  - adjustments for time-dependent effects, such as inflation
  - removal of columns that are not relevant to the analysis (e.g. customer id or calibration information)
  - handling missing data:
    * removal of observations

    - imputation (ranging from simple replacement with the mean to complex methods such as multiple imputation)
- recording the steps of the cleaning process, so that if needed the information is available during future analyses

The statistical data characterization concepts described in the remainer of this presentation all refer to a single data table column/variable/attribute

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\cdots$	$\boldsymbol{x}_{\mathsf{p}}$
i_1	x_{11}	x_{21}	$\cdots$	$x_{\mathsf{p}1}$
i_2	x_{12}	x_{22}	$\cdots$	$x_{\mathsf{p}2}$
.	.	.		.
.	.	.		.
.	.	.		.
i_{n}	$x_{1\mathsf{n}}$	$x_{2\mathsf{n}}$	$\cdots$	$x_{\mathsf{p}\mathsf{n}}$

Tally charts and frequency distributions

For datasets collected and/or processed manually

Score	Tallies
62	|
63	|
64	
65	|||
66	|
67	||||
68	~~||||~~
69	~~||||~~ |
70	||||
71	||
72	||
73	
74	|

[LEFT] Tally chart of the scores made in their final round by the 30 leading golfers in the 1992 Scottish Open

0	4, 7, 1, 0, 1, 6, 0, 1, 0
1	2, 7, 0, 3, 0, 1, 4
2	2, 6
3	9
4	
5	8

[ABOVE] Results from a low-scoring cricket match, shown in a stem and leaf diagram. The stems are chosen so that there are up to about 10 of them, for clarity.

Central Tendency

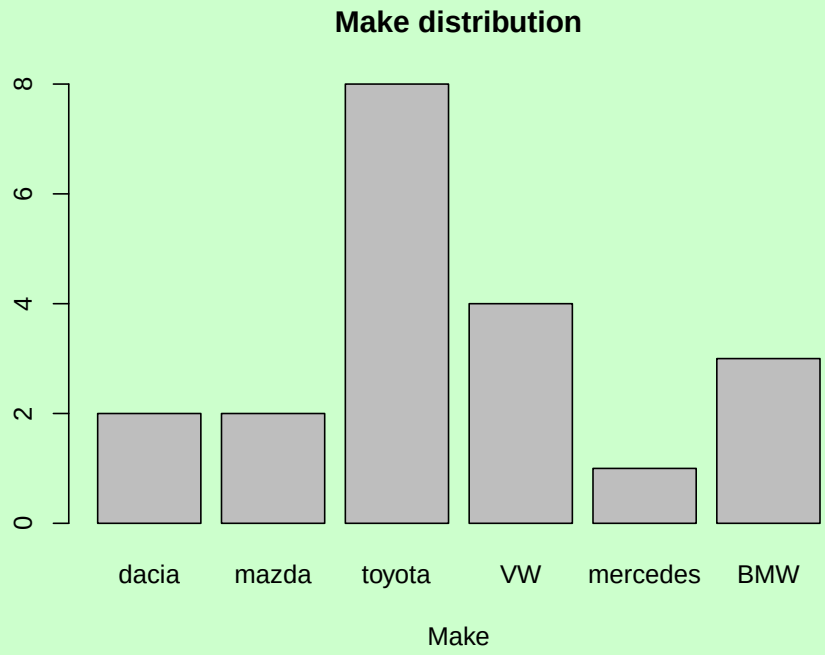
- One of the most important way of summarising a variable is finding the centre of the values associated with it
- There are three measures of central tendency: the *mode*, the *median* and the *mean*.
  - The **mode** is the value that is most commonly reported for the variable. For example, if the following data were reported: `3, 4, 5, 6, 7, 7, 7, 8, 8, 9` then the mode would be `7`, as this is the most commonly occurring value in the set. If the set included an additional `8`, then the mode would be either `7, 8` or `7.5`. Always used with variables on a nominal scale, to which the other two measures of central tendency do not apply, but can be and is also used with variables measured on the other types of scale.
  - The **median** is the middle value in the set, when the set is ordered. For example, in the set `3, 4, 5, 6, 7,` **`7,`** `8, 8, 8, 9, 9` the median is `7`, as this is the value with the same number of values on either side of it. If the number of values is even, the median is calculated as the average of the two middle values. For example, in the set `3, 4, 5, 6, 7,` **`7, 8,`** `8, 8, 9, 9, 9` the median is `(7+8)/2=7.5`.

– Finally, the **mean** is the average of the values and is calculated as:

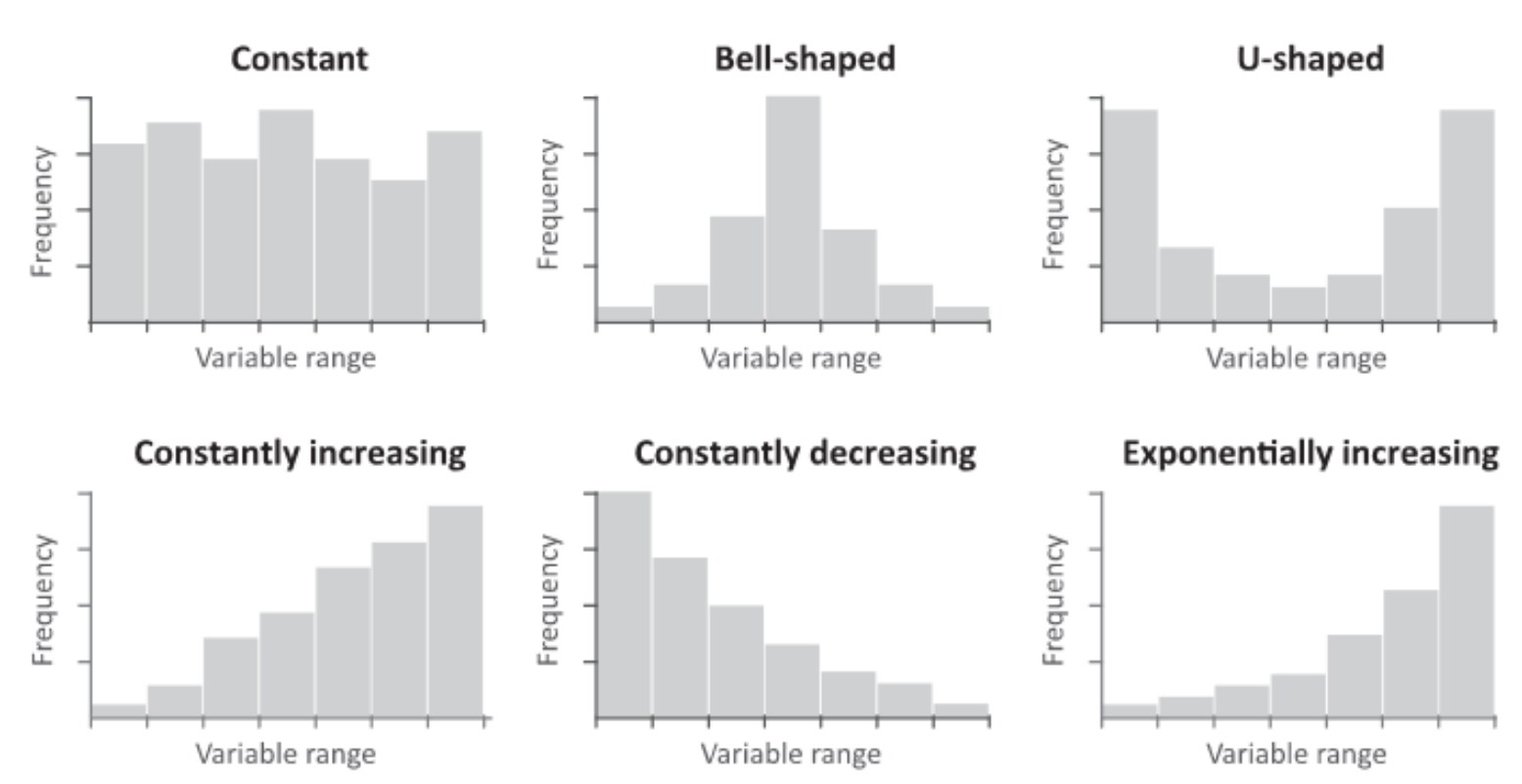
$$\bar{x} = \frac{\sum_{\mathrm{i}=1}^{n} x_{\mathrm{i}}}{\mathrm{n}}$$

Distribution

- Mode, median and mean provide a good initial idea of the data values but nothing about where in the range of possible values the set at hand actually is.
- A good place to start is a visual summary of the distribution, by way of bar charts and frequency histograms
  - A bar chart simply displays the 'counts' for the different values of a variable:



– A frequency histogram is useful for ordered variables with many values. It groups the values into ranges and gives an idea of the relative frequencies of the ranges.

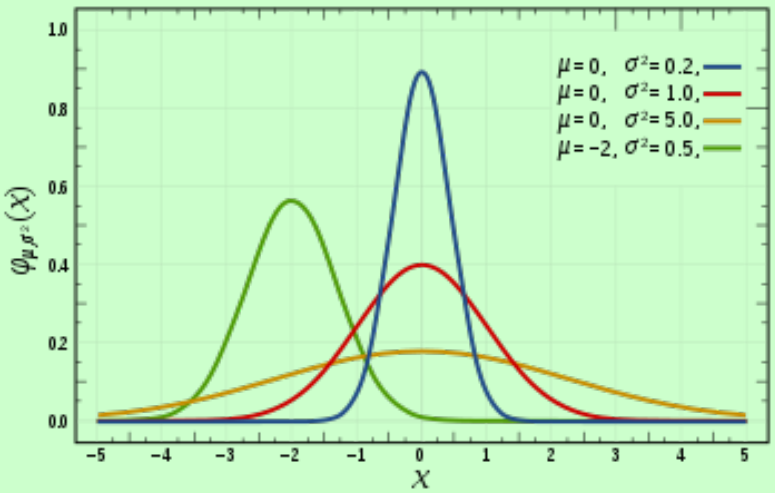


Source: [MSD]

– A probability density function (pdf) is a smooth function representing data in the same way as a histogram but based on a very large sample or theoretically derived. A very common probability density function is that for the *normal distribution*, which can be expressed with the formula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ the mean and σ the standard deviation.



Source: Wikipedia

HOWTO

Drawing a **histogram** from a frequency table manually

The x-axis should show the ranges, while the area of the box above a range should be proportional to the frequency i.e. the number of values in that range. This proportionality is achieved when the height of the boxes corresponds to a *frequency density* i.e. the number of values per some 'unit of range', which can be chosen arbitrarily.

Example:

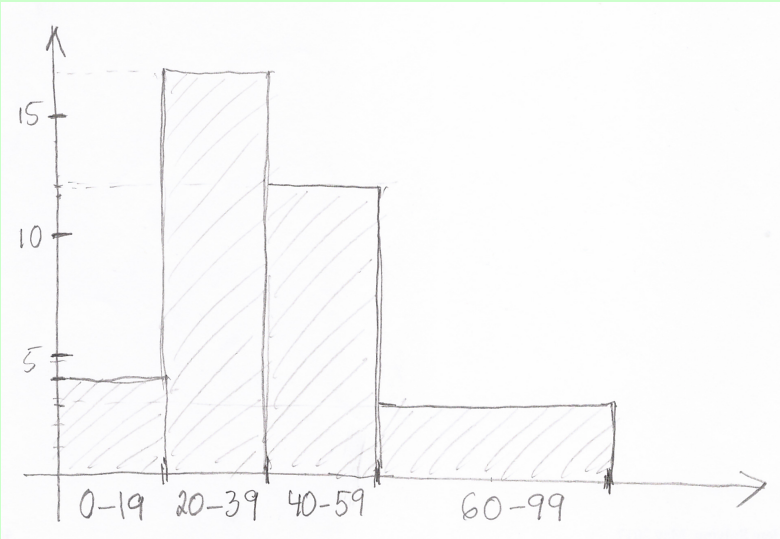
The data to be displayed in a histogram is given in the following table:

Range of values	0-19	20-39	40-59	60-99
Frequency	4	17	12	6

The frequency density value shown in a histogram for a range can be expressed as:

$$FD = \frac{number\ of\ values\ in\ range}{\dfrac{width\ of\ range}{width\ of\ 'unit\ of\ range'}}$$

For a range that has the same width as the 'unit of range' the frequency density is equal to the frequency. If we choose 20 for the 'unit of range', 3 out of the 4 ranges in the table will have 'unit of range' as their width and in those cases frequency density can be read directly from the table. For the fourth range, which has a width of 40, the frequency density value to be displayed in the histogram is $FD_4 = \dfrac{6}{\dfrac{40}{20}} = 3$. A sketch of the histogram derived in this way is shown below.



HOWTO

Drawing a **histogram** from a frequency table with R

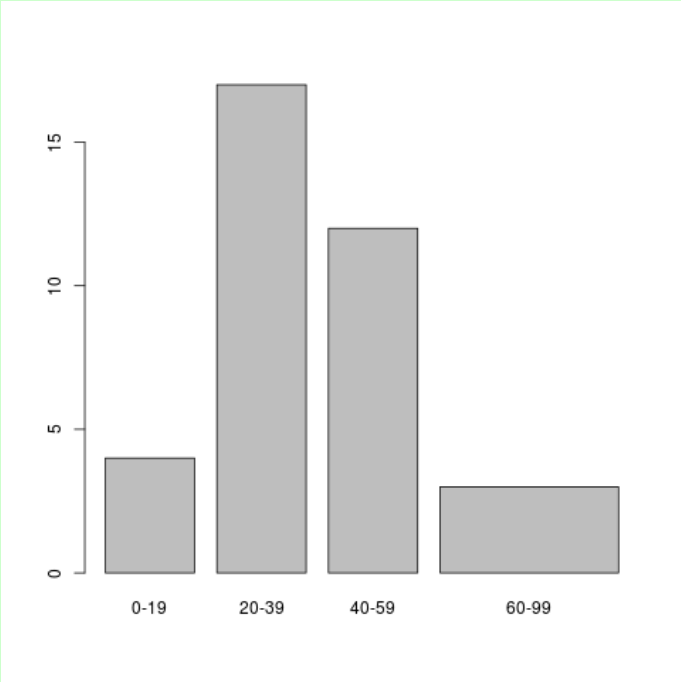
Frequency table data can be displayed in a boxplot, which is equivalent to the histogram for that data.

Example:

Using the same frequency table as above, we call the barplot function in R:

```
> barplot(height=c(4,17,12,3),width=c(1,1,1,2),names=c("0-19","20-39","40-59","60-99"))
```

The height vector contains the calculated frequency densities, the width vector contains the widths of the ranges in 'units of range' and the names vector contains the range labels.



HOWTO

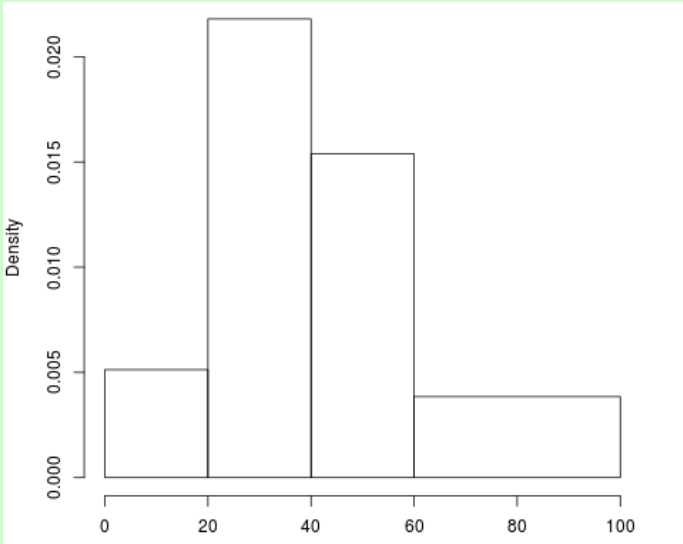
Drawing a **histogram** from a dataset with R

A function is available in R for drawing a histogram directly from a vector containing some attribute's values.

Example:

Let's 'make up' some data that fits the previous example. We feed the data into the R hist() function, with the resulting diagram equivalent to the one obtained from the frequency table.

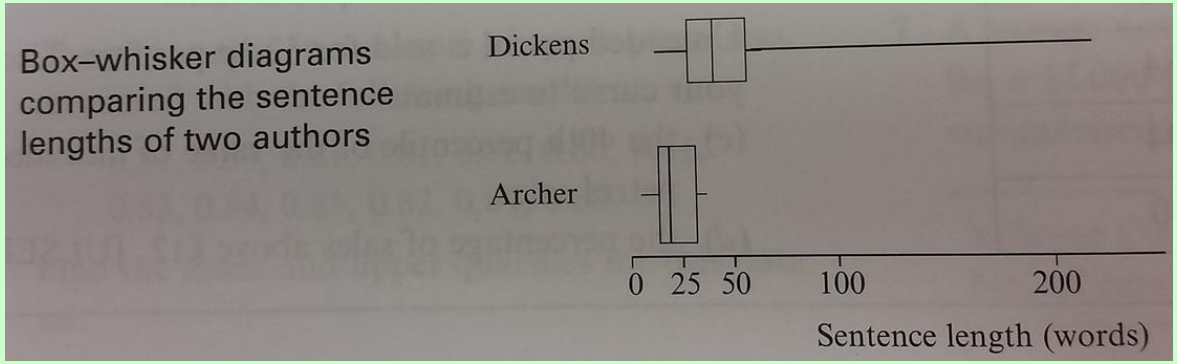
```
> hist(c(3,5,7,17,21,21,23,25,25,28,30,31,32,33,33,33,34,36,37,38,39,41,44,44,45,47,49,
50,50,52,55,56,57,62,66,70,76,84,91),breaks=c(0,20,40,60,100))
```



NOTE: The values on the y-axis are chosen so that the area under the entire histogram is equal to 1 (the 'unit of range' is 1/39, where 39 is the number of values in the set). However, the proportions remain the same as in the histogram created from the frequency table in the previous HOWTO panel.

- Distribution is also expressed through the measures of *range* (the difference between the highest and the lowest value) and *inter-quartile range* (the difference between the two values that lie (a) below the highest quarter of values and (b) above the lowest quarter of values, respectively)
  - Box and whisker diagrams show the positions of the quartiles and the median in relation to the entire range. The vertical sides of the box are at the lower and upper quartile. The vertical line splitting the box is at the median. The leftmost and rightmost tips of the horizontal lines (the whiskers) are at the lowest and highest value in the range.

    Example:



Source: [US]

The box and whisker diagrams in this example show that Dickens' sentences are a lot longer (some over 200 words long!) than Archer's but also that Dickens produced a greater range of sentence lengths than Archer.

HOWTO

Calculating the **range** of a dataset manually

1. order the data values by size
   Example: `2 3 4 4 5 5 5 7 9 10 10 11 12 12 15`
2. the *range* is the difference between the largest and the smallest value

   Example: $range = 15 - 2 = 13$

Finding the **range** of a dataset with R

Use the following R command:

```
> max(DATASET) - min(DATASET)
```

Example:

```
> dataset <- c(2, 3, 4, 4, 5, 5, 5, 7, 9, 10, 10, 11, 12, 12, 15)
> max(dataset) - min(dataset)
[1] 13
```

The value given for the range by R is 13.

HOWTO

Calculating the **inter-quartile range (IQR)** manually

1. order the data values by size

   Example 1 (odd number of data values): `2 3 4 4 5 5 5 7 9 10 10 11 12 12 15`

   Example 2 (even number of data values): `2 3 4 4 5 5 5 7 9 10 10 11 12 12 15 16`

2. identify the lower and upper half of the dataset (in the case of a dataset with an odd number of values, the median value is discarded)

   Example 1: `(2 3 4 4 5 5 5)` **`7`** `(9 10 10 11 12 12 15)`

   Example 2: `(2 3 4 4 5 5 5 7) (9 10 10 11 12 12 15 16)`

3. the *first quartile* or the *lower quartile* or *Q1* is equal to the median of the lower half of the dataset

   Example 1: `2 3 4` **`4`** `5 5 5,` $Q1 = 4$

   Example 2: 2 3 4 **4** **5** `5 5 7,` $Q1 = \frac{4+5}{2} = 4.5$

4. the *third quartile* or the *upper quartile* or *Q3* is equal to the median of the lower half of the dataset

   Example 1: `9 10 10` **`11`** `12 12 15,` $Q3 = 11$

   Example 2: `9 10 10` **`11 12`** `12 15 16,` $Q3 = \frac{11+12}{2} = 11.5$

5. the *inter-quartile range* or *IQR* is the difference between the upper and lower quartiles

   Example 1: $IQR = Q3 - Q1 = 11 - 4 = 7$

   Example 2: $IQR = Q3 - Q1 = 11.5 - 4.5 = 7$

NOTE: The median of the data set is equal to the *second quartile* or *middle quartile* of the set.

HOWTO

Finding the **inter-quartile range (IQR)** with R

```
> IQR(DATASET)
```

Example:

```
> IQR(c(2, 3, 4, 4, 5, 5, 5, 7, 9, 10, 10, 11, 12, 12, 15))

 [1] 6
```

NOTE: The value returned by R (6) differs from that calculated manually (7). This is because there are several methods for determining the IQR. All these are valid and the difference in results is not significant in the case of large datasets, which are the type that are encountered in real-life statistics.

- The **variance** describes the spread of the data and measures how much the values of a variable differ from the mean. When calculated with data across an entire population, the following formula is used:

$$\sigma^2 = \frac{\sum_{\mathrm{i}=1}^{n}(x_{\mathrm{i}} - \bar{x})^2}{\mathrm{n}}$$

where σ^2 is the variance, n is the number of data values in the population, x_i is the ith value and $\bar{x}$ is the mean value for the population. In the case that *sample data* is used to

estimate the variance of a population, the formula is:

$$S^2 = \frac{\sum_{\mathrm{i}=1}^{n}(x_{\mathrm{i}} - \bar{x})^2}{\mathrm{n} - 1}$$

where S^2 is the exstimated variance, n is the number of values in the sample, x_i is the ith value and $\bar{x}$ is the mean value for the sample. Dividing by n when using sample data would consistently underestimate the population variance, essentially because the contribution of the difference between the estimated and population mean is always positive (owing to the squaring in the formula).

- The **standard deviation** is the square root of the variance. Both the population and sample standard deviation are calculated as the square root of the respective variance:

$$\sigma = \sqrt{\frac{\sum_{\mathrm{i}=1}^{n}(x_{\mathrm{i}} - \bar{x})^2}{\mathrm{n}}} \qquad S = \sqrt{\frac{\sum_{\mathrm{i}=1}^{n}(x_{\mathrm{i}} - \bar{x})^2}{\mathrm{n} - 1}}$$

Frequency distribution shape

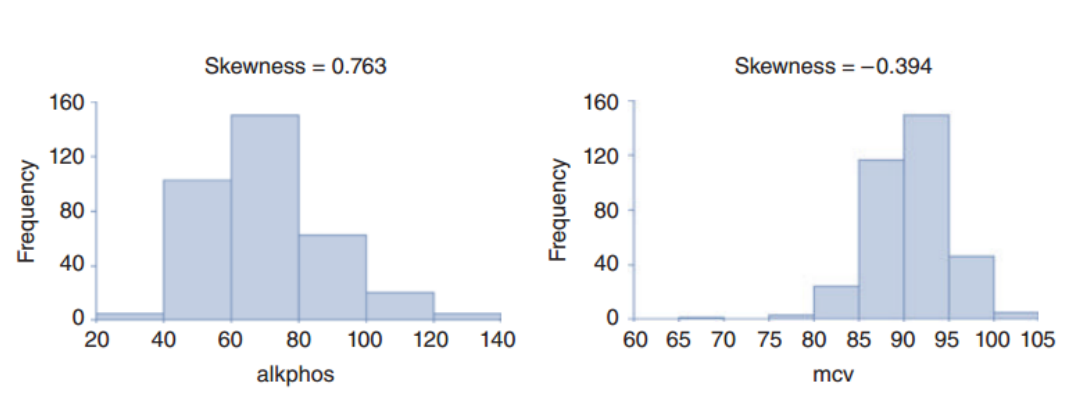
The shape is represented with two measures, **skewness** and **kurtosis**.

- **skewness** is a measure of imbalance in how the values of a distribution are placed to the left and the right of the mean. it can be calculated using the following formula:

$$skewness = \left(\frac{\sqrt{n \times (n-1)}}{n-2}\right) \times \frac{{}^{1}\!/_{n} \times \sum_{i=1}^{n}(x_i - \bar{x})^3}{\left({}^{1}\!/_{n} \times \sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

Source: [MSD]

two distributions, one with a positive and one with a negative skewness:



Source: [MSD]

there are many other ways of measuring skewness, such as the **quartile coefficient of**

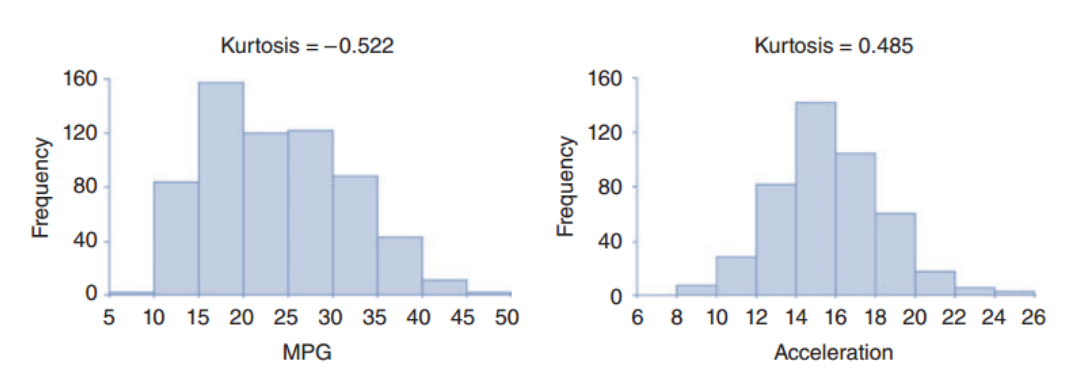
skewness or the **pearson's coefficient of skewness**

- **kurtosis** is a measure of of the peak of the distribution and how close it is to the mean. kurtosis formula:

$$kurtosis = \frac{n-1}{(n-2)\times(n-3)} \times \left((n+1) \times \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4 / n}{\left(\sum_{i=1}^{n} (x_i - x)^2 / n\right)^2} - 3 \right) + 6$$

Source: [MSD]

two distributions, with different values for kurtosis:



Source: [MSD]

Confidence interval

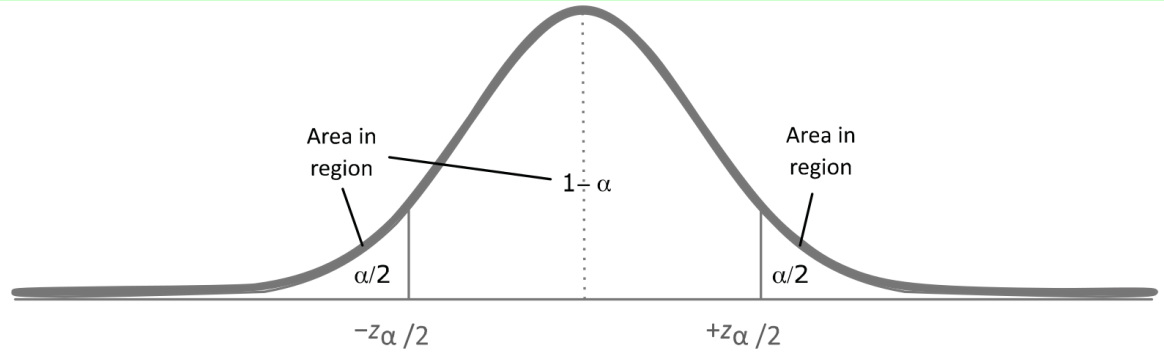
The confidence interval is a measure of how well a property calculated on a data sample represents the same property with respect to the entire population.

- A confidence interval is expressed as a range of values within which a statistic will be found in a large proportion of cases. This proportion is specified as a percentage, most often 95%. For example, a confidence interval may be:

$$\mathbf{x} = \mathbf{455.5} \pm \mathbf{5.4}, \text{ with a confidence level of } \mathbf{99}\%$$

- Confidence intervals are calculated using **sampling distributions**. These can be distributions of various statistics, e.g. the sample mean. Like population distributions, sampling distributions have:
  - a mean, called the **expected value**
  - a standard deviation, called the **standard error**
  - a distribution, called the **sampling distribution**

- For an example, we look at the confidence interval for the mean of a distribution.
  - The sampling distribution for a mean is a normal distribution as long as the samples contain more than 30 values, according to the *law of large numbers* or the *central limit theorem* (the sampling distribution for samples smaller than 30 is a *Student's t-distribution* and a confidence interval can be defined for that case as well, but it is outside of the scope of the module)
  - The standard error of the sampling distribution is $\frac{s}{\sqrt{n}}$, where s is the standard deviation of a sample and n is the size of the sample
  - The following picture shows a z-distribution (a normal distribution with standard deviation 1 and mean 0):



Source: [MSD]

If we define $\alpha = 1 - \frac{confidence\_level}{100}$, where $confidence\_level$ is the percentual value of the required or declared *confidence level* (e.g. 95%), then the values on the x-axis of the z-distribution below and above which the area covered by the distribution is equal to $\alpha/2$ define the normalized *confidence interval*. The two values, which we will name $-z_{\alpha/2}$ and $z_{\alpha/2}$, differ only in sign, since the z-distribution is symmetric with respect to the y-axis.

– The confidence interval is then:

$$\bar{x} \pm z_{\alpha/2}(\frac{s}{\sqrt{n}}), \text{ with a confidence of } confidence\_level$$

The meaning of this statement is that there is a $confidence\_level$ level of confidence that the actual mean of the population lies somewhere in the interval spanning the width $z_{\alpha/2}(\frac{s}{\sqrt{n}})$ to either side of the value of the sample mean, $\bar{x}$.

– The value of $z_{\alpha/2}$ for a given confidence level percentage can be looked up from a z-distribution table

Upper-tail percentage points of the standard normal distribution

The table gives the values of z for which $P(Z > z) = p$, where the distribution of Z is $N(0, 1)$.

p	z	p	z	p	z	p	z	p	z
.50	0.000	.15	1.036	.025	1.960	.010	2.326	$.0^{3}4$	3.353
.45	0.126	.14	1.080	.024	1.977	.009	2.366	$.0^{3}3$	3.432
.40	0.253	.13	1.126	.023	1.995	.008	2.409	$.0^{3}2$	3.540
.35	0.385	.12	1.175	.022	2.014	.007	2.457	$.0^{3}1$	3.719
.30	0.524	.11	1.227	.021	2.034	.006	2.512	$.0^{4}5$	3.891
.25	0.674	.10	1.282	.020	2.054	.005	2.576	$.0^{4}1$	4.265
.24	0.706	.09	1.341	.019	2.075	.004	2.652	$.0^{5}5$	4.417
.23	0.739	.08	1.405	.018	2.097	.003	2.748	$.0^{5}1$	4.753
.22	0.772	.07	1.476	.017	2.120	.002	2.878	$.0^{6}5$	4.892
.21	0.806	.06	1.555	.016	2.144	.001	3.090	$.0^{6}1$	5.199
.20	0.842	.050	1.645	.015	2.170	$.0^{3}9$	3.121	$.0^{7}5$	5.327
.19	0.878	.045	1.695	.014	2.197	$.0^{3}8$	3.156	$.0^{7}1$	5.612
.18	0.915	.040	1.751	.013	2.226	$.0^{3}7$	3.195	$.0^{8}5$	5.731
.17	0.954	.035	1.812	.012	2.257	$.0^{3}6$	3.239	$.0^{8}1$	5.998
.16	0.994	.030	1.881	.011	2.290	$.0^{3}5$	3.291	$.0^{9}5$	6.109

Hypothesis tests

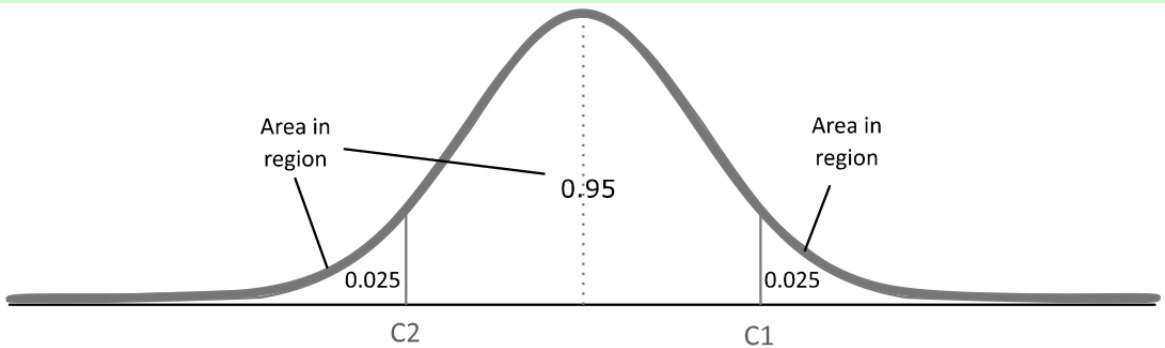
A hypothesis test is used to check if a sample of data supports a particular hypothesis made about the population from which the sample was drawn.

- A hypothesis test definition uses the same principle as confidence interval definition, i.e. works with an interval for a data statistic, which associated with a probability value
- In the case of hypothesis testing, the probability used is the *threshold of probability* or *level of significance* (while in the case of confidence intervals the probability was the *confidence level*)
- The interval in hypothesis testing is defined by between one or two critical values (while in the case of confidence intervals it was called the *confidence interval*)
- A hypothesis test is stated through a null hypothesis and an alternative hypothesis and the null hypothesis is either accepted or rejected at the end of the test
- For an example, let us look at the mean of a population once more. The hypotheses would be stated in this way:

$H_0 : \mu = 100$ null hypothesis

$H_a : \mu \neq 100$ alternative hypothesis

- Let's assume a sample of more than 30 observations, which means that the sampling distribution is normal (see the section on confidence intervals)
- The standard error of the sampling distribution is $\frac{s}{\sqrt{n}}$, where s is the standard deviation of a sample and n is the size of the sample
- Critical values are marked in a z-distribution (a normal distribution with standard deviation 1 and mean 0) so as to correspond to the test's level of significance. If the test is two-tailed, there are two critical values and if it single-tailed there is only one. In the picture below, critical values are shown for a two-tailed test with a level of significance of 5% (the areas under the z-distribution curve outside of the range C1 to C2 add up to 0.05):



Source: [MSD]

- The hypothesis test is accepted if the test value, calculated as follows, falls between C1 and C2:

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

  where $\bar{x}$ is the sample mean, μ_0 is the hypothesised mean, s is the standard deviation and n the size of the sample.
- The critical values can be looked up for a given level of significance in a z-distribution table

- Two types of errors can occur with respect to the outcome of a hypothesis test:
  - Type I error - when the null hypothesis is rejected even though it is true
  - Type II error - when the null hypothesis is accepted even though it is not true

References The pictures in this presentation were taken from the following books. The source for each picture is cited beside it.

[DSB] *Data Science for Business: What you need to know about data mining and data-analytic thinking*, by Foster Provost and Tom Fawcett, O'Reilly Media, 2013.

[MSD] *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, by Glenn J. Myatt and Wayne P. Johnson, John Wiley & Sons, 2014.

[US] Understanding Statistics, by Graham Upton and Ian Cook, Oxford University Press, 1996.