Data Analysis:
Data Description and Summarisation

Technological University Dublin Tallaght Campus

Department of Computing

Why describe and summarise data? Because from looking at them in raw form, such as in the picture, we cannot easily (or at all) tell what is going on.

1	Date	Distance_	Average_	Calories	Tyre_MM	Bike_Type	App	Weight	Rain	Wind_Dir	Wind_Spe	Temp
2	07/05/2017	43.6	21.03	1926	35	Hybrid	Map my Ride	100	0	70	4	16.3
3	14/05/2017	43.69	20.85	1920	35	Hybrid	Map my Ride	100	6.4	210	14	14.5
4	22/05/2017	6.33	19.29	277	35	Hybrid	Map my Ride	98.5	0	210	12	16.4
5	22/05/2017	6.42	17.77	263	35	Hybrid	Map my Ride	98.5	0	210	12	16.4
6	23/05/2017	6.3	23.44	310	35	Hybrid	Map my Ride	98.5	0	240	12	18.2
7	23/05/2017	6.46	16.83	228	35	Hybrid	Map my Ride	98.5	0	240	12	18.2
8	24/05/2017	6.1	21.59	287	35	Hybrid	Map my Ride	98.5	0	290	5	19.2
9	24/05/2017	6.46	21.5	303	35	Hybrid	Map my Ride	98.5	0	290	5	19.2
10	26/05/2017	6.29	18.58	262	35	Hybrid	Map my Ride	100	1.6	150	13	24
11	26/05/2017	6.33	23.41	305	35	Hybrid	Map my Ride	100	1.6	150	13	24
12	28/05/2017	66.52	20.57	2980	35	Hybrid	Map my Ride	100	2.2	20	3	14.3
13	30/05/2017	30.28	20.55	1355	35	Hybrid	Map my Ride	100	0.6	230	14	17.3
14	02/06/2017	6.3	22.26	301	35	Hybrid	Map my Ride	100	0.8	240	9	15.6
15	02/06/2017	6.38	17.15	238	35	Hybrid	Map my Ride	100	0.8	240	9	15.6
16	05/06/2017	30.62	20.42	1393	35	Hybrid	Map my Ride	100	5.6	240	12	14.1
17	07/06/2017	6.07	24.19	304	35	Hybrid	Map my Ride	99	16.1	250	10	13.8
18	07/06/2017	6.89	20.91	318	35	Hybrid	Map my Ride	99	16.1	250	10	13.8
19	09/06/2017	6.52	16.53	216	35	Hybrid	Map my Ride	99	9.2	230	10	15.5
20	09/06/2017	6.26	23.52	309	35	Hybrid	Map my Ride	99	9.2	230	10	15.5
21	13/06/2017	52.2	20.92	2412	35	Hybrid	Map my Ride	99	0	210	10	16.4
22	17/06/2017	43.89	21.81	2079	35	Hybrid	Map my Ride	99	0	230	8	21.9
23	19/06/2017	30.05	22.93	1464	35	Hybrid	Map my Ride	99	0	290	4	22.9
24	23/06/2017	2.94	23.19	142	35	Hybrid	Map my Ride	99	1.1	240	12	15.6
25	25/06/2017	1.63	10.11	76	35	Hybrid	Map my Ride	99	0	290	7	13
26	27/06/2017	29.87	22.33	1415	35	Hybrid	Map my Ride	99	0	200	8	20.2
27	02/07/2017	31.85	21.98	1495	35	Hybrid	Map my Ride	99	0.4	210	10	15.2
28	09/07/2017	43.81	22.53	2052	35	Hybrid	Map my Ride	99	0.4	240	10	19.4
29	22/07/2017	66.98	20.15	2931	35	Hybrid	Map my Ride	99	3.1	50	6	17.5
30	29/07/2017	43.61	20.45	1927	35	Hybrid	Map my Ride	99	0	230	13	16.3

... and another 300 lines

The statistical data characterisation concepts described in this presentation all refer to a single data table column i.e. to a single 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	I
63	I
64	
65	III
66	I
67	IIII
68	~~IIII~~
69	~~IIII~~ I
70	IIII
71	II
72	II
73	
74	I

[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.

Parameters and statistics

A quantity that describes a variable may:

- pertain to the whole *population*, in which case it is called a **parameter**
- be calculated for a *sample*, in which case it is called a **statistic**

Measures of central tendency

One of the most important ways of summarising a variable is finding some kind of centre around which its values are grouped. There are three measures of central tendency:

- **mode** - the most commonly occurring value among those observed (defined for variables with *discrete* values)
- **median** - the value that is surpassed by exactly half of the observed values (defined for variables with values that *can be ordered*)
- **mean** - the average value (defined for *numeric* variables)

HOWTO

Determining the **mode** for a set of values

Find the most commonly occurring value.

Example 1 (one mode):

`3, 4, 5, 6, 7, 7, 7, 8, 8, 9` $\longrightarrow$ The mode is **7**.

Example 2 (bimodal data):

`3, 4, 5, 6, 7, 7, 7, 8, 8, 8` $\longrightarrow$ The modes are **7 and 8**.

Note: More than two modes i.e. modes of *multimodal* data are reported similarly.

HOWTO

Determining the **median** for a set of values

The **median** is the middle value in the set, when the set is ordered.

Example 1 (odd number of values):

`3, 4, 5, 6, 7,` **7**, `8, 8, 8, 9, 9` $\longrightarrow$ The median is **7**.

Example 2 (even number of values):

`3, 4, 5, 6, 7,` **7, 8**, `8, 8, 9, 9, 10` $\longrightarrow$ The median is $\frac{7+8}{2} = \mathbf{7.5}$.

HOWTO

Determining the **mean** for a set of values

The **mean** is the average of the values. For a variable called x with n values it is calculated as:

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

Example:

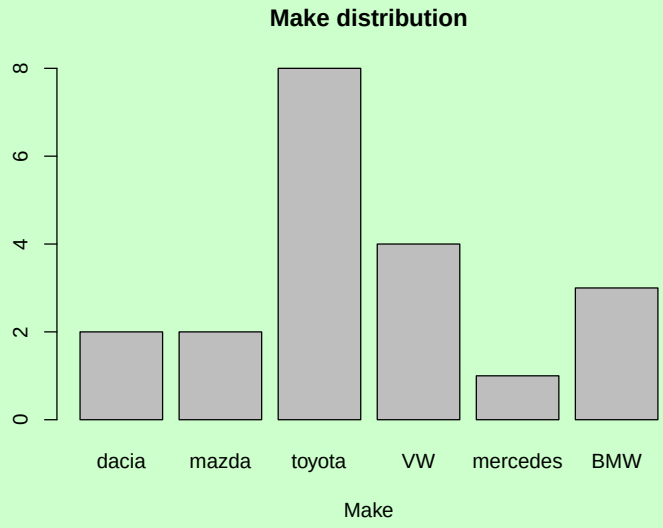
`3, 4, 5, 6, 7, 7, 8, 8, 8, 9, 9` $\longrightarrow$ The mean is $\frac{3+4+5+6+7+7+8+8+8+9+9}{11} = \mathbf{6.73}$.

Distribution visualisation

The distribution of a set of values can be visualised in several ways, each suitable for specific types of data.

Bar chart

A bar chart shows how many time each value occurs in a set. It is used with **categorical variables** and **discrete numeric variables of limited range**.



Frequency histogram

A frequency histogram is useful for ordered variables with many values. Indeed, it is most often **numeric** data that is presented in this way.

A frequency histogram groups the values into ranges and gives an idea of the relative frequencies of the ranges.

Original source: [MSD]

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:

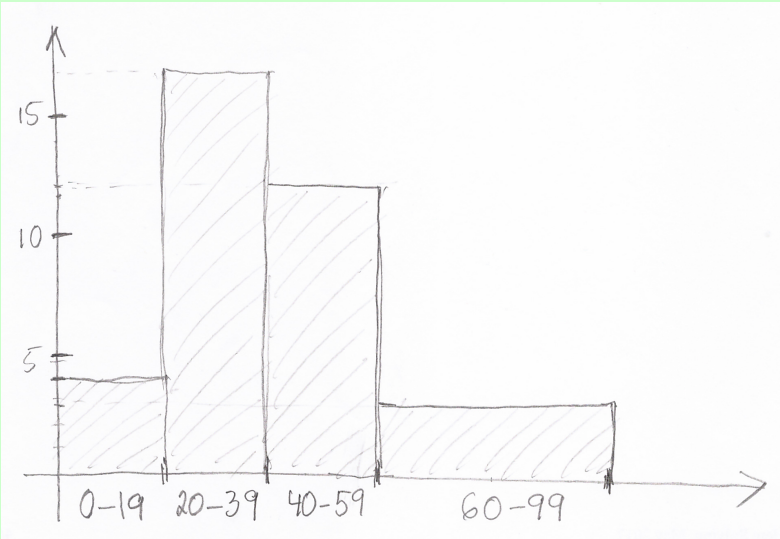
The data to be displayed in a histogram is given in the following table:

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.

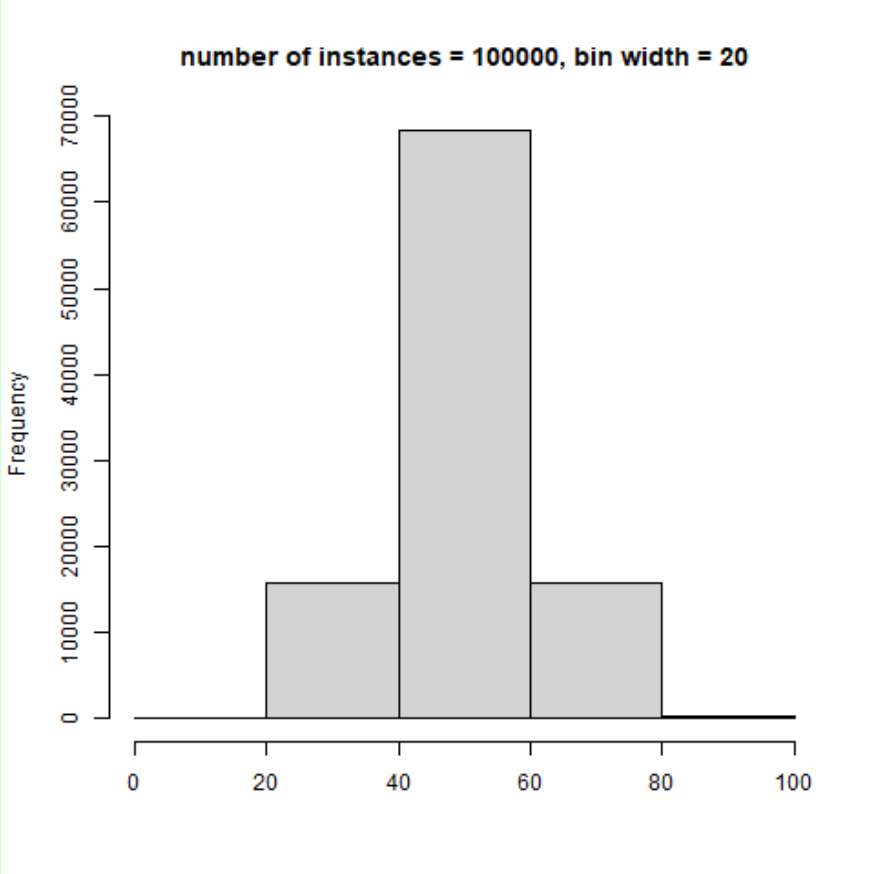


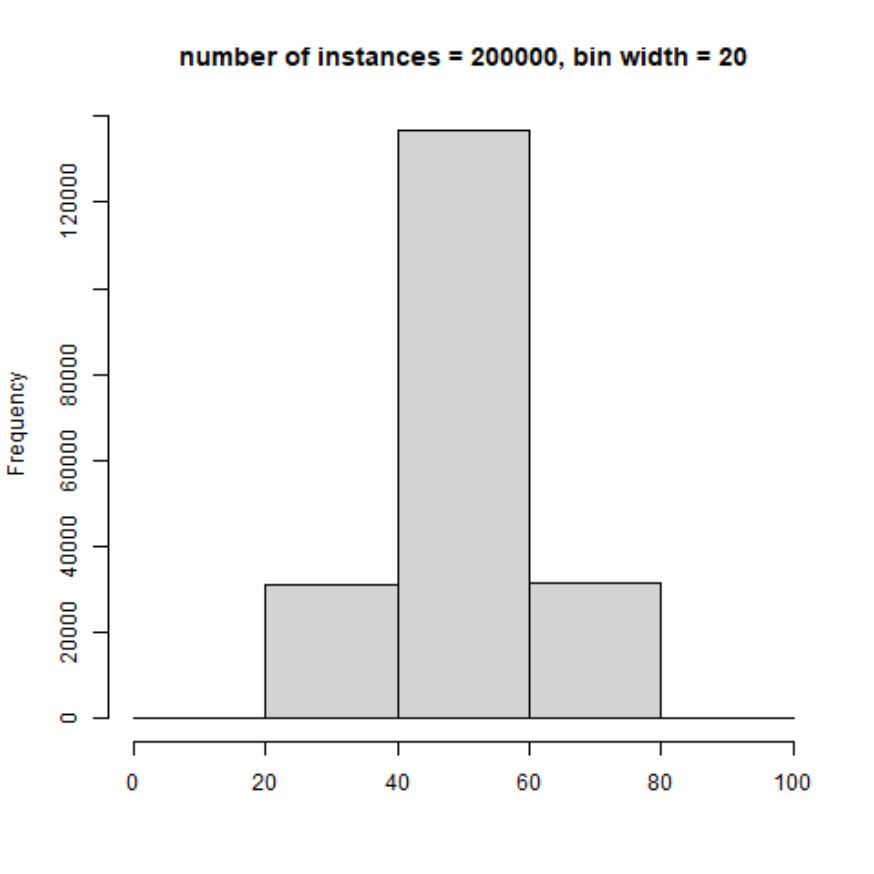
Probability density function

A probability density function (PDF) shows the idealised (based on an infinite population) distribution of a continuous variable. It could be viewed as a histogram with infinitely small bins and values normalised to a population of 1. The values on the y-axis are in probability density units and the area under the entire curve equals 1.

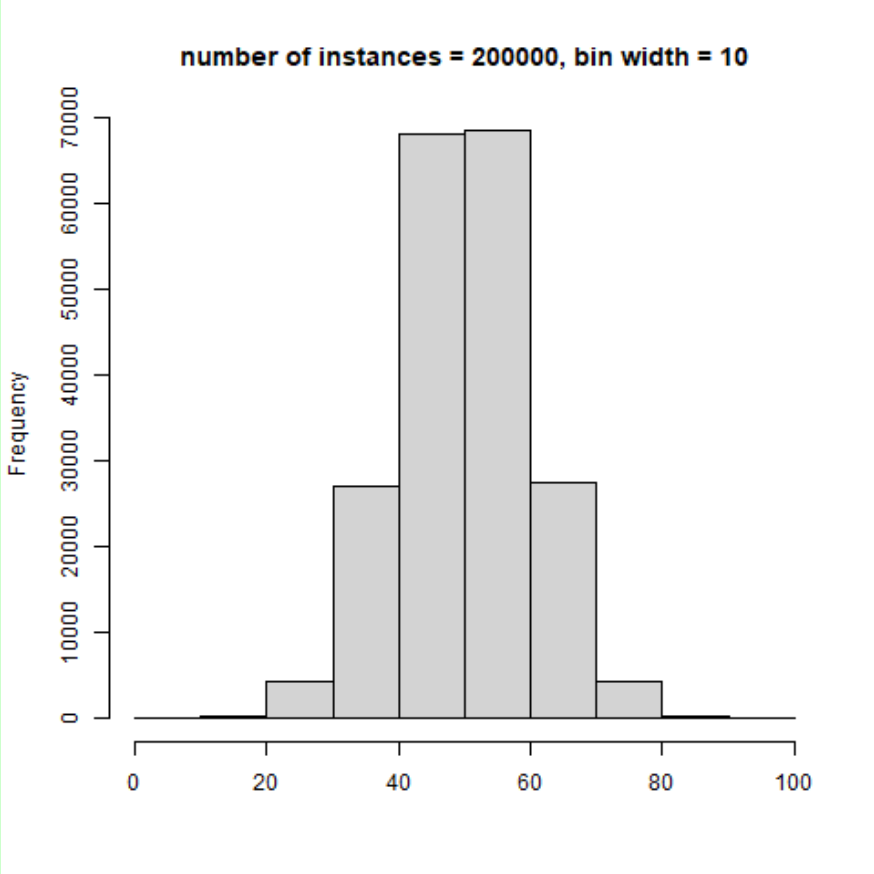
The following pictures show how a histogram 'turns into' a PDF with the population and number of bins increasing.

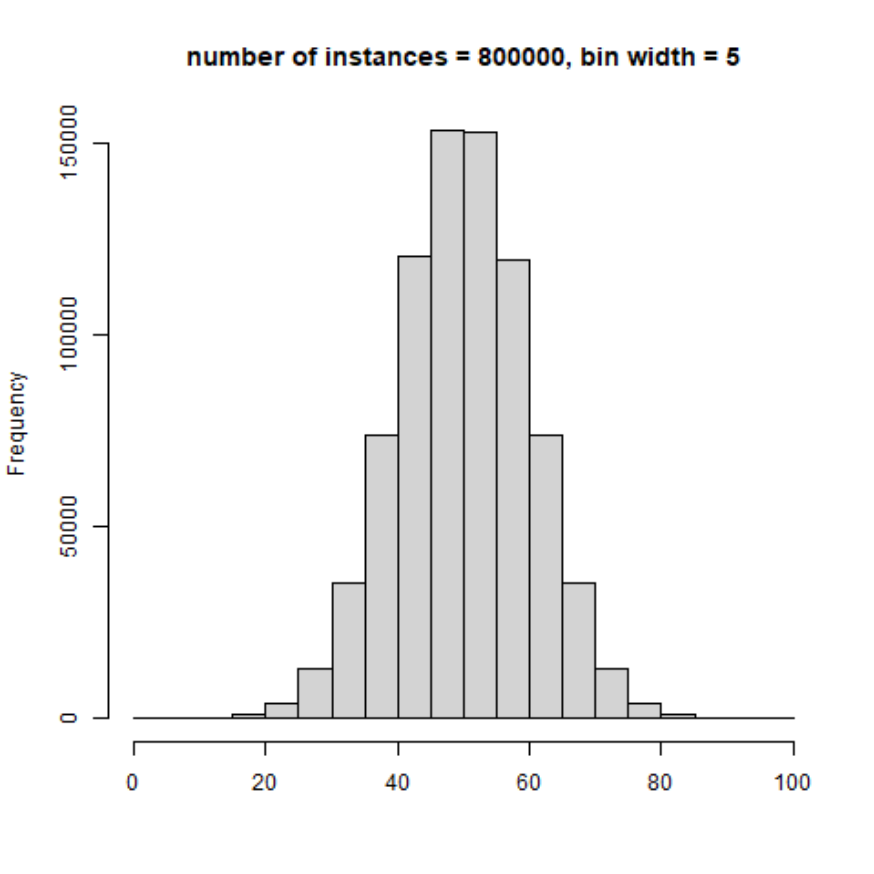
Two different population sizes



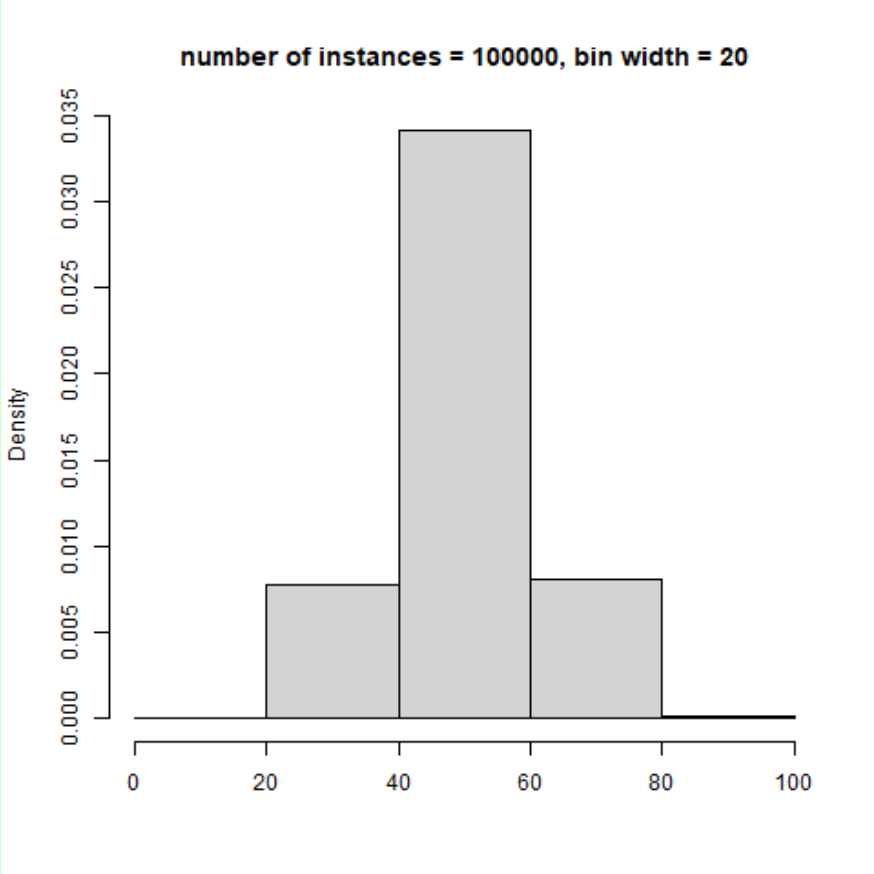


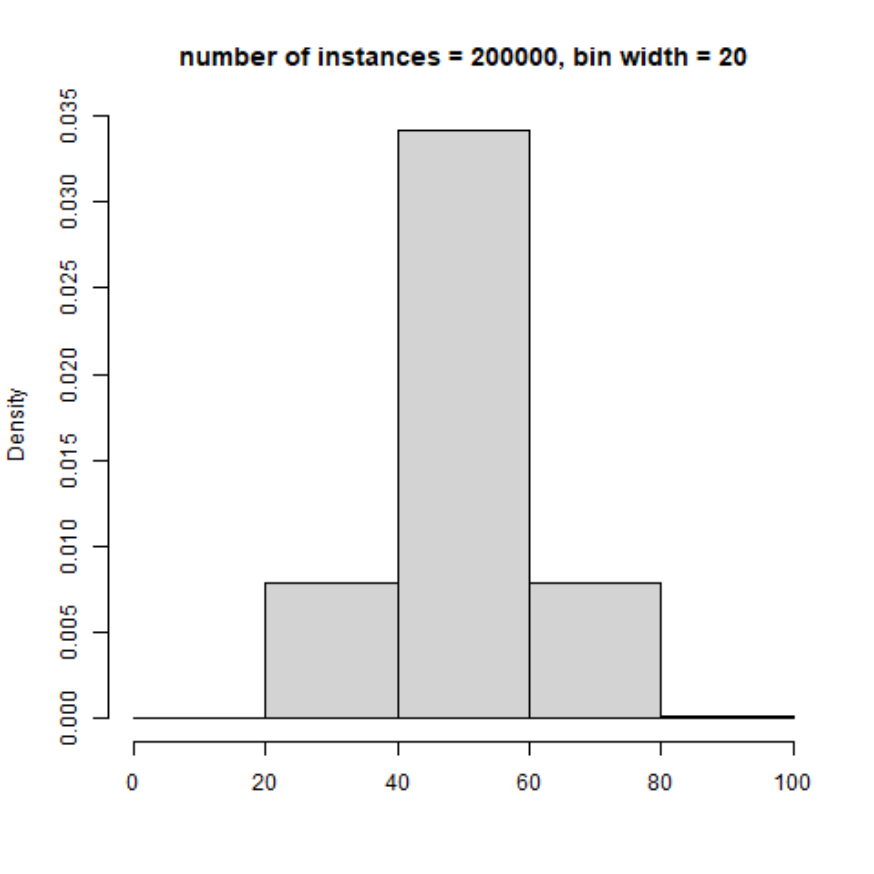
Further changes to population and bin width



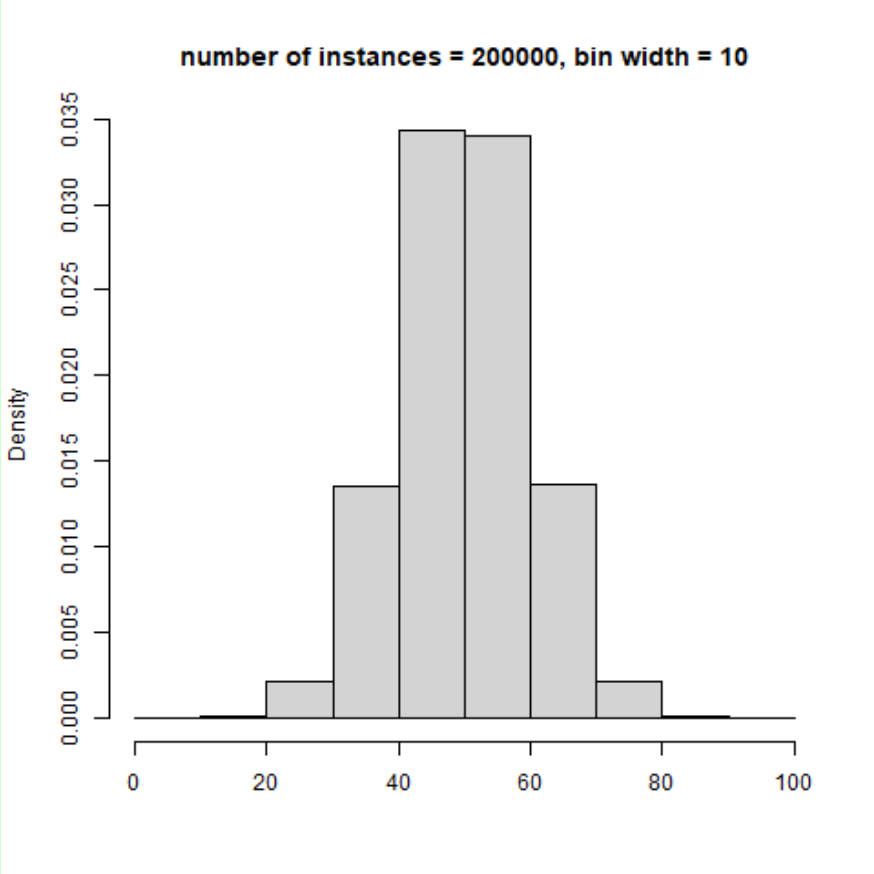


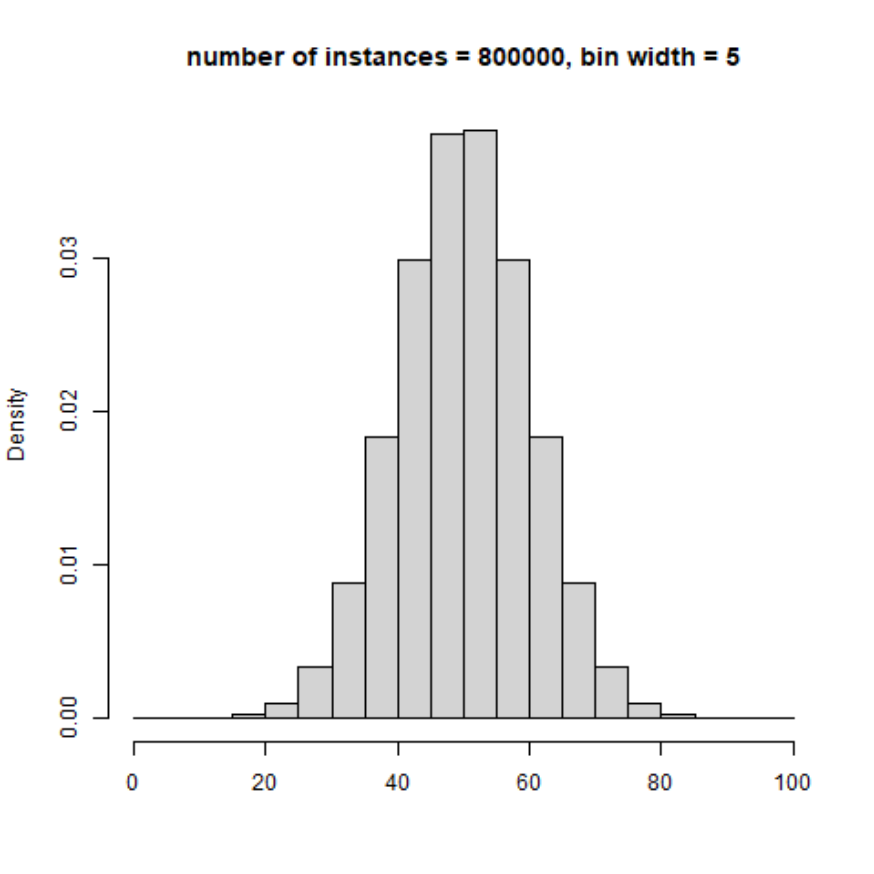
Using density rather than frequency makes the diagrams comparable regardless of population and bin size



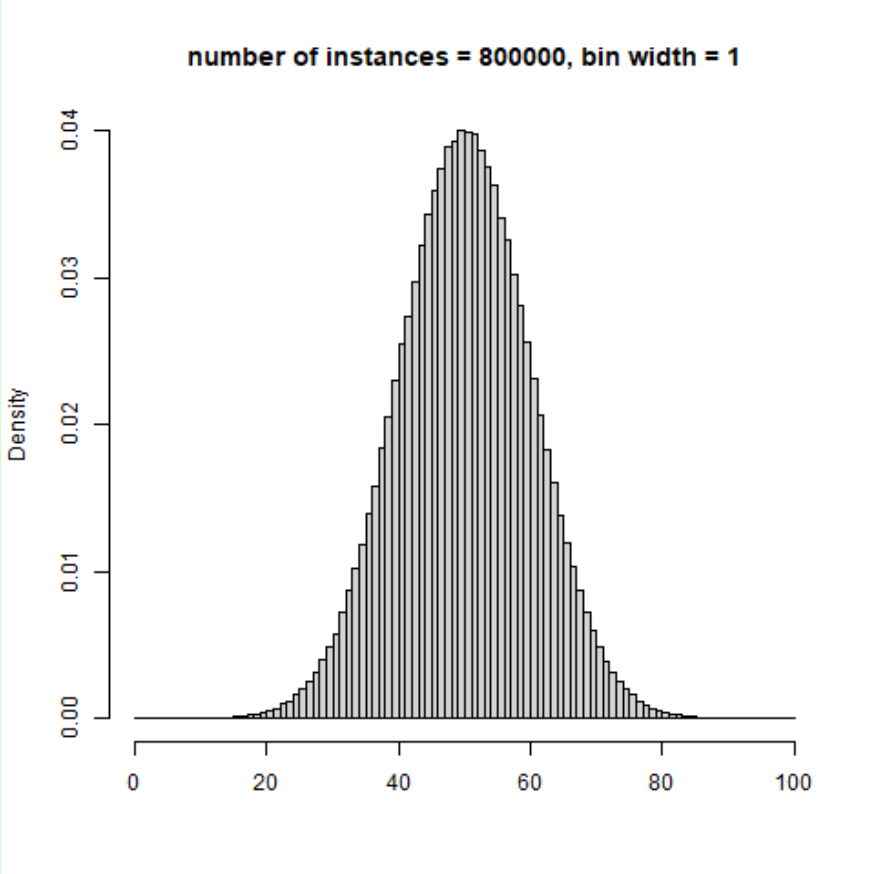


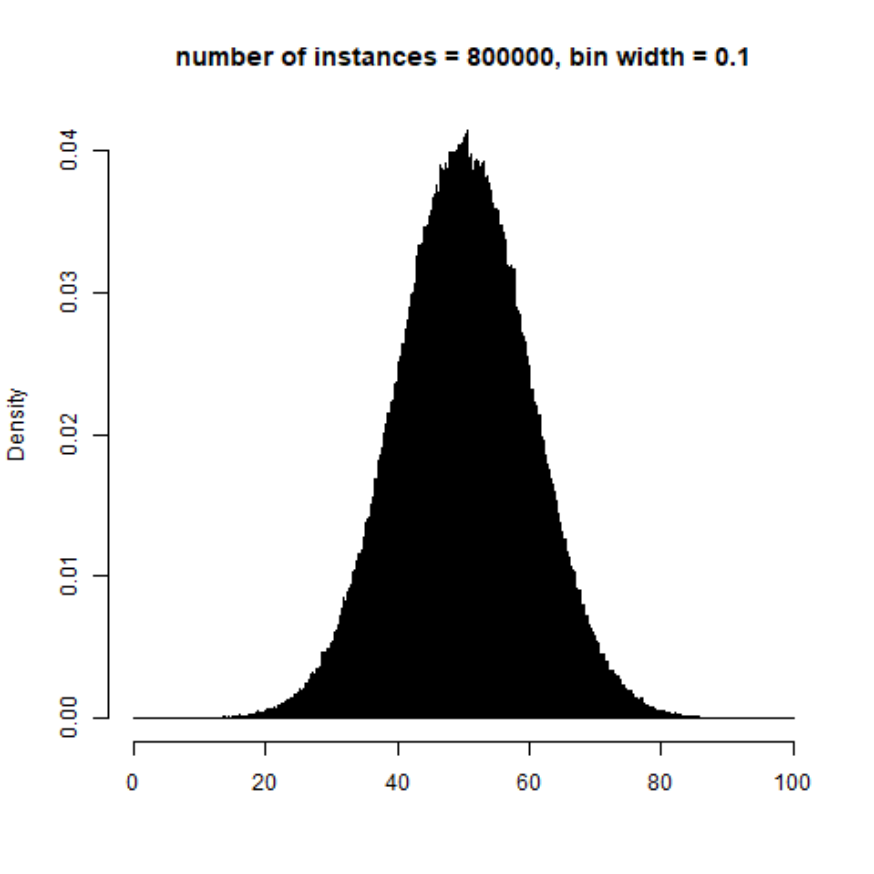
As the population gets bigger...



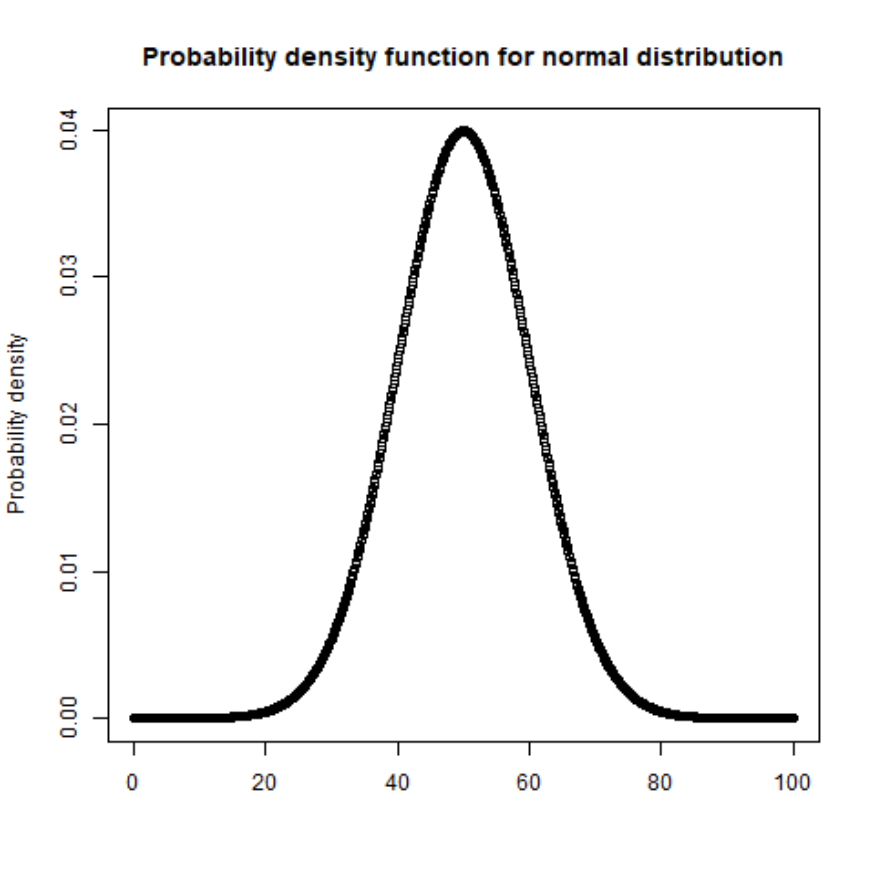


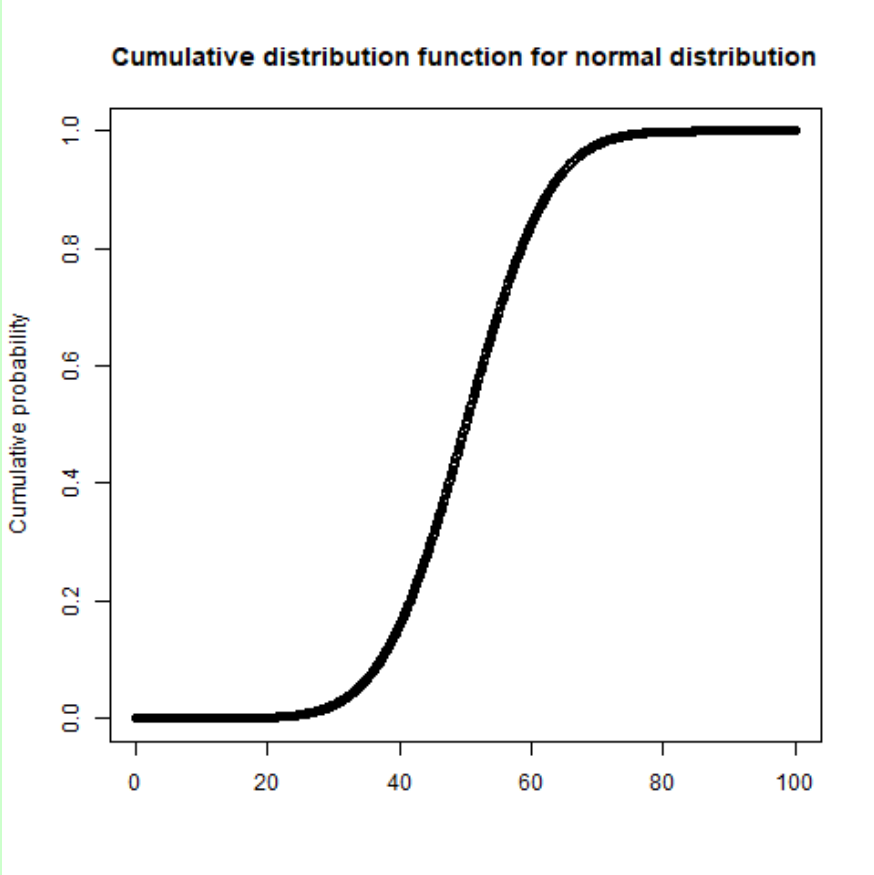
... and the bins smaller, we the distribution gets closer to the ideal





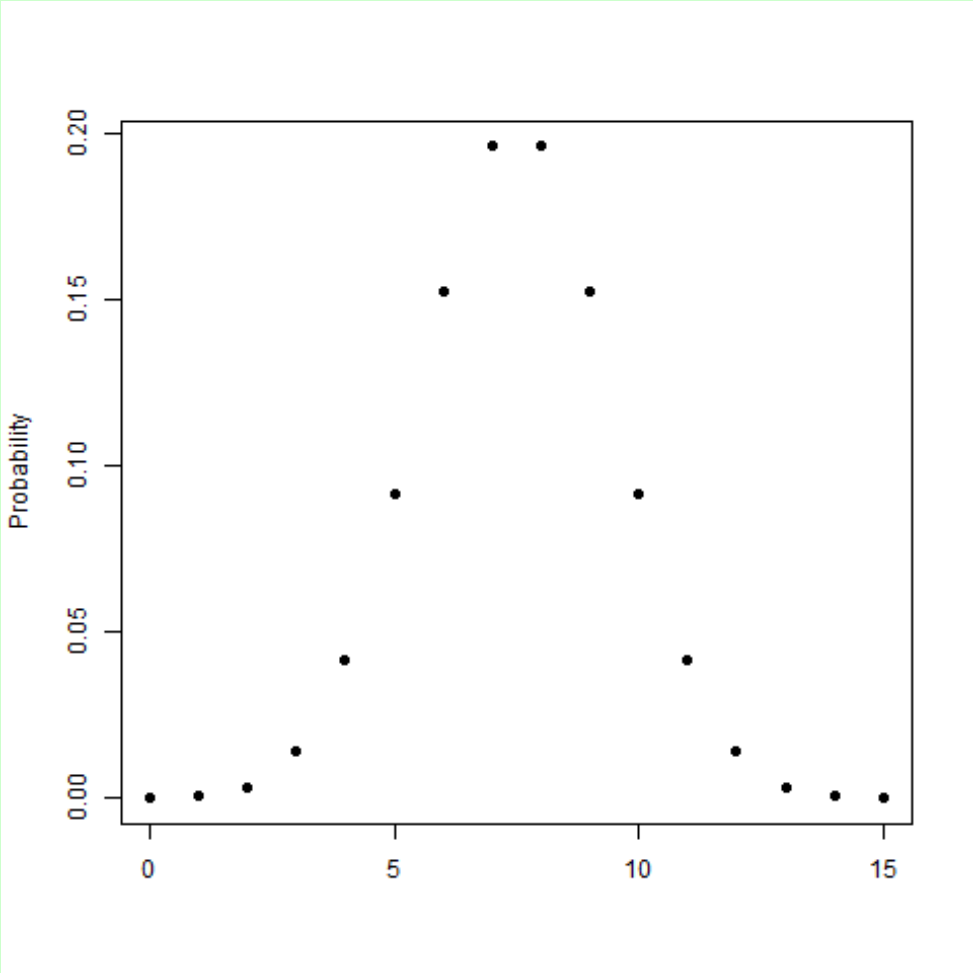
The picture on the left is the idealised distribution, with infinitesimally narrow 'bins' and the y-axis showing the **probability density function (PDF)**. The picture on the right shows the **cumulative distribution function**, which is the probability of values being smaller than the value on the x axis.





Discrete distributions

A probability density function represents the distribution of a continuous variable but a plot can also be constructed for **discrete variables**. In that case the y-axis shows the **probability** of occurring, for each discreet value. The picture shows the *binomial distribution*. An example of data that would exhibit this distribution are instances of counting heads in N subsequent coin throws (the picture corresponds to N=15): the number of heads can be between 0 and N and the probability of each of these outcomes is shown in the graph. Each outcome corresponds to one possible value for the data arising from the experiment.

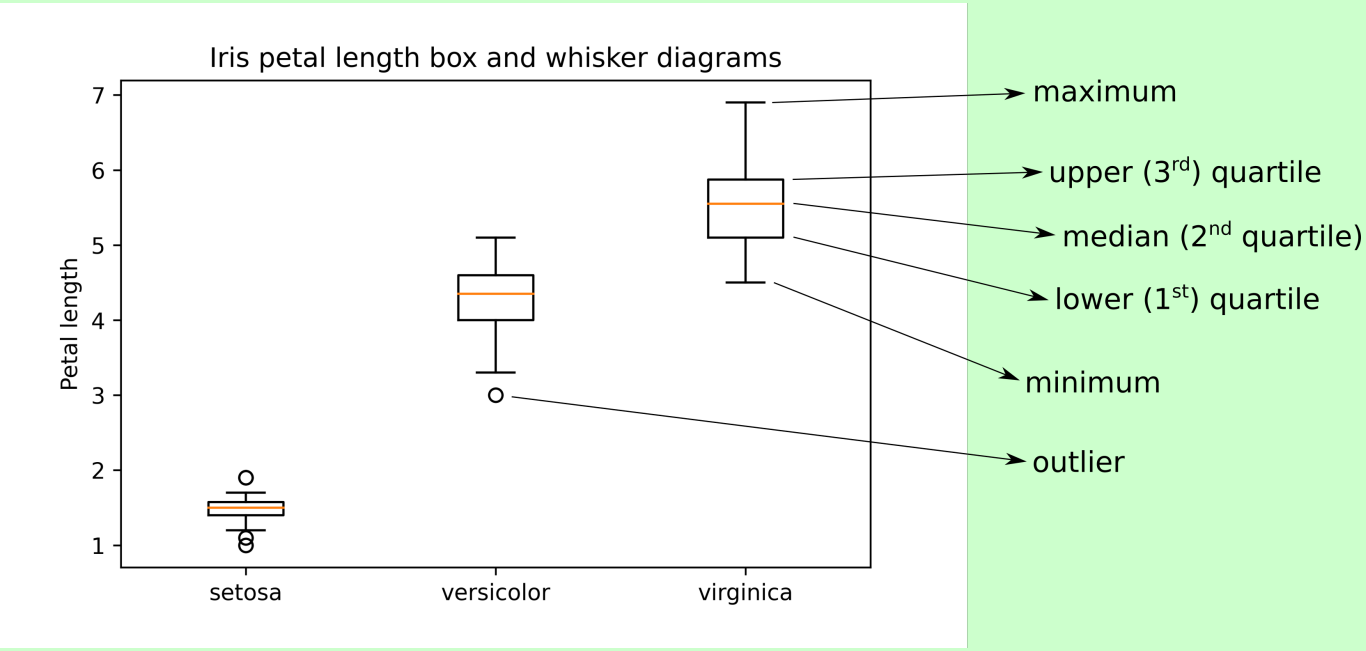


Box and whisker diagram

A box and whisker diagram gives a 5-point summary of a set of values. Several diagrams placed side-by-side are often used for comparison of different value sets.

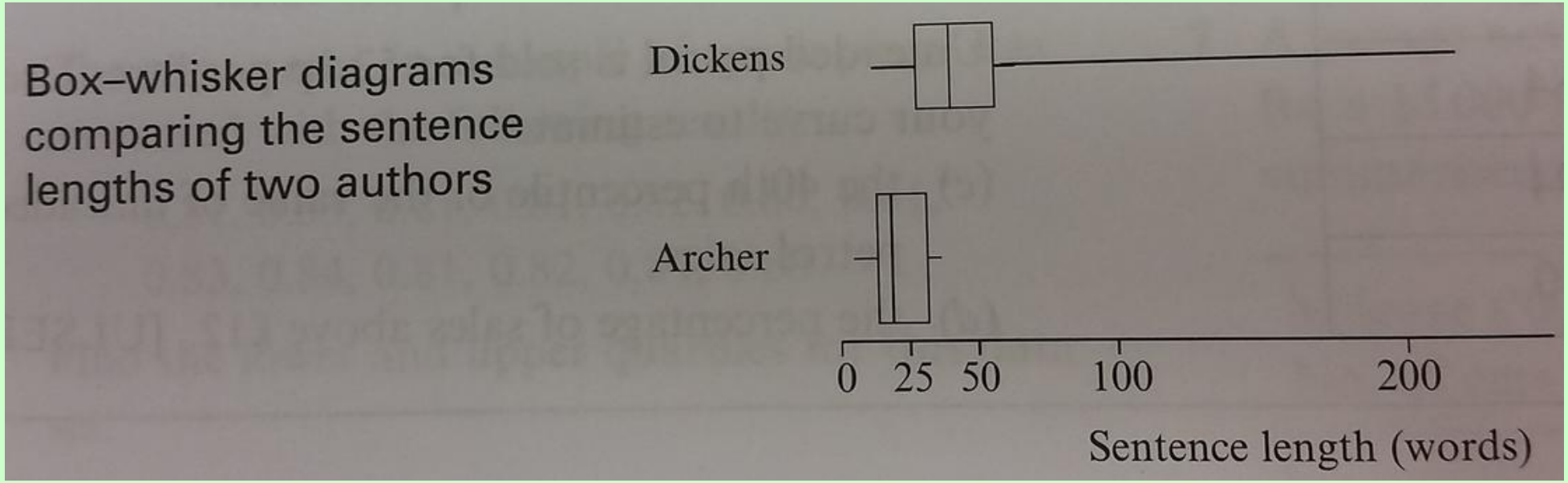
The points:

- minimum
- lower quartile (a value that is greater than 25% of the values in the set and smaller than 75% of values in the set)
- median
- upper quartile (a value that is greater than 75% of the values in the set and smaller than 25% of values in the set)
- maximum



Box and whisker diagram example:

The box and whisker diagrams in this example show that Dickens's 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.



Source: [US]

Measures of dispersion (spread)

How much numeric data is spread out can be expressed with several measures.

Range

The *range* is the difference between the highest and the lowest value in the set.

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: $range = 15 - 2 = 13$

Inter-Quartile Range (IQR)

The *inter-quartile range* is the difference between the upper quartile and the lower quartile i.e. the middle 50% of all the values.

Arbitrary percentiles

Sometimes it is useful to calculate arbitrary *percentiles*, for example the 97th percentile for some data.

There are several methods for calculating percentiles (there are 9 listed in the documentation for the R function `quantile`). The example on the following page uses the usual method for manual derivation of quartiles, however, with large samples, the choice of method makes little difference in most cases.

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.

Variance

The variance is the average squared distance of the values from the mean.

The unit of variance is the data unit squared. For example, if the data is in meters (m), then the variance will be in meters squared (m^2).

When calculated with data across an entire population, the following formula is used:

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

where σ^2 is the variance, n is the number of data values in the population, x_i is the ith value and μ 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 estimated 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.

Standard deviation (SD)

The standard deviation is the square root of the variance.

The standard deviation has the same unit as the data. For example, if the data is in litres (l), then the standard deviation will also be in litres (l).

The population and sample standard deviation are each calculated as the square root of the respective variance:

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

Note that σ^2 and σ are *parameters* (calculated for the population), while S^2 and S are *statistics* (calculated using a sample).

The denominator (n-1) in the sample-based formulae for variance and standard deviation

At first sight, the denominator $n-1$ in sample formulae seems like an arbitrary and unintuitive modification of the formulae for population parameters. In fact, the division by $n-1$ reflects the fact that the distances calculated against the sample mean, $\bar{x}$, are on average slightly smaller than they would be from the actual mean, μ. This is because the sample mean is constructed from the data values, reducing the randomness, expressed as **degrees of freedom**, by a factor of $\frac{n}{n-1}$.

The formulae for sample variance and standard deviation represent **unbiased estimates of the respective parameters**. This means that the **expected values** (average arising from a very large number of experiments) of the formulae are equal to the parameters.

Mean absolute deviation (MAD)

This measure is more intuitive for most people. However, because it is not easy to manipulate mathematically (in contrast to SD), it retains an inferior position in statistics. With computers now ubiquitous for data analysis, MAD is used more, especially in deep learning.

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

Distribution shape

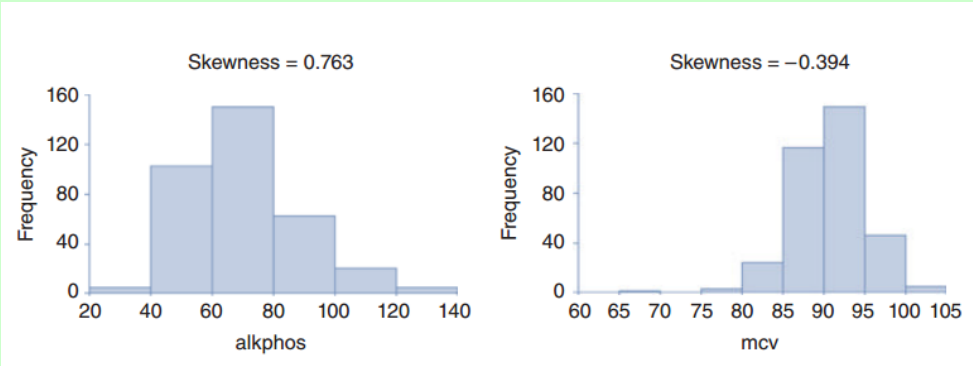
Skewness

Skewness is a measure of asymmetry of a distribution. There are various ways of calculating skewness, one of which is:

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

Source: [MSD]

Examples of distributions, one with a positive and one with a negative skewness:



Source: [MSD]

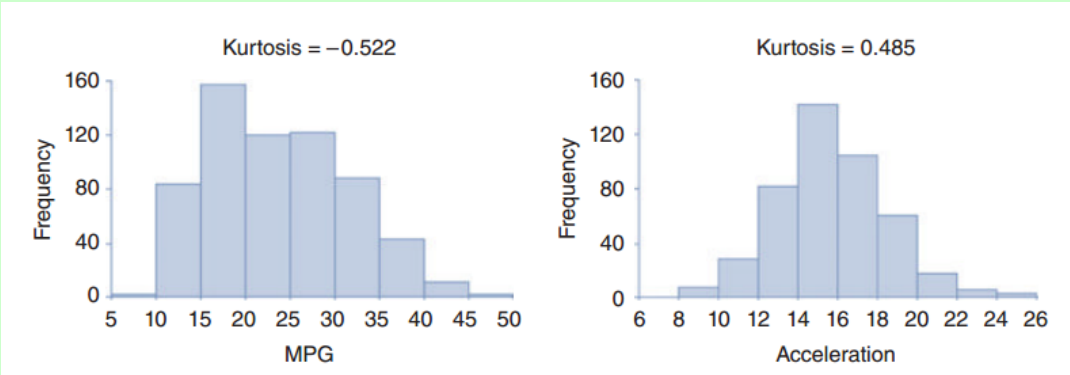
Kurtosis

Kurtosis is a measure of the 'tailedness' of a distribution i.e. of the extent and amount of outliers it includes:

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

Source: [MSD]

Two distributions, with different values for kurtosis: a distribution with a negative kurtosis is **platykurtic** while the distribution with a positive kurtosis is **leptokurtic**



Source: [MSD]

The normal distribution

The normal distribution occurs in many types of data in the world, for example:

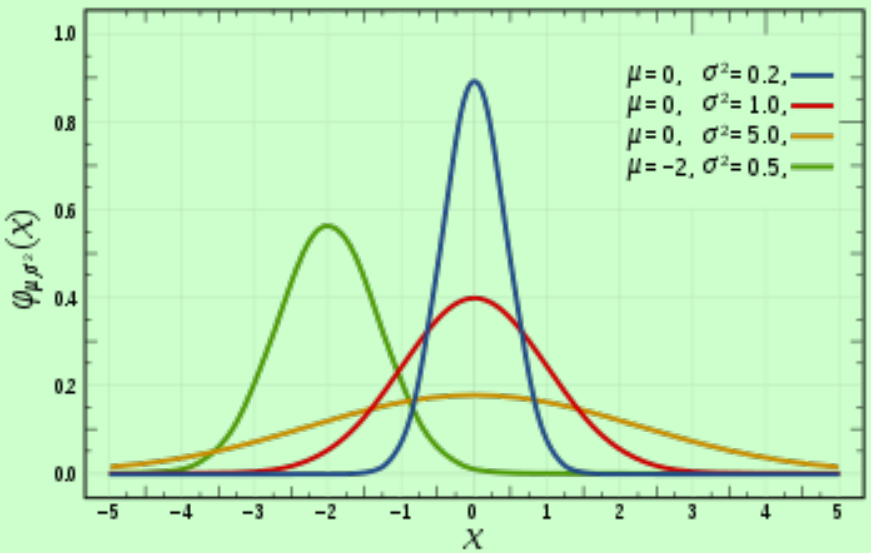
- measurement errors
- weather properties
- size in biology (this is often log-normal, with the exponential index of a quantity normally distributed)
- heights of children of the same age

Definition

The formula for the probability density function of the normal distribution is:

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

where μ is the population mean and σ the standard deviation.



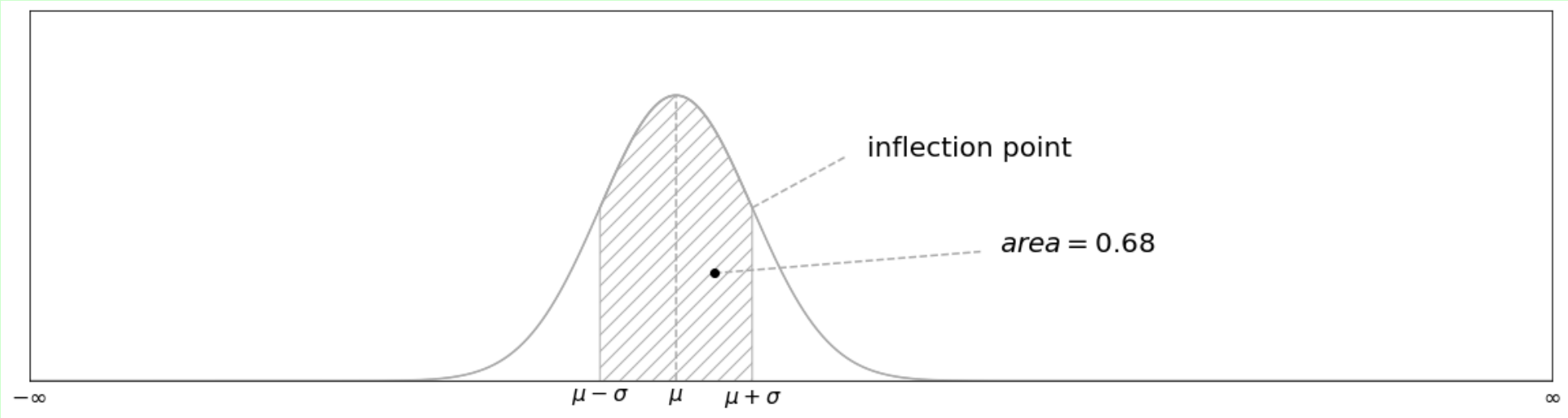
Source: Wikipedia

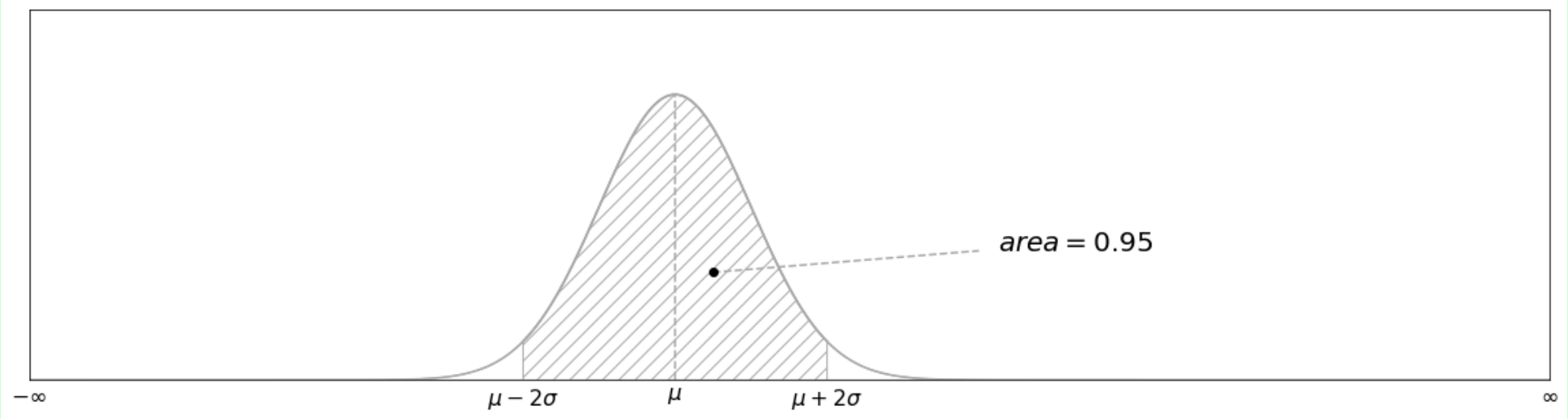
History

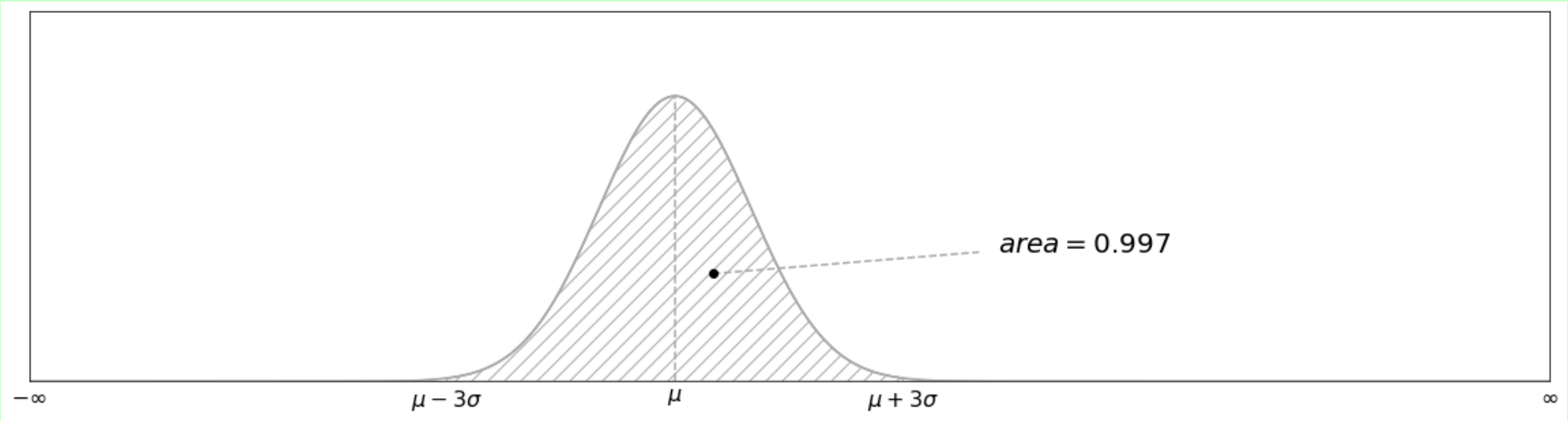
- Abraham de Moivre (1667 - 1754) - first derived the normal distribution as a **limit of the binomial distribution** when the sample size tends to infinity (an interesting article about this can be found **here**)
- Pierre-Simon Laplace (1749 - 1827) - proposed the **central limit theorem**, which states that by sampling data of any distribution with a big enough sample size, a normal distribution emerges in the sample means - this is probably the most intuitively convincing phenomenon for the 'naturalness' of the normal distribution
- Johann Carl Friedrich Gauss (1777 - 1855) - noticed that errors in measurement are always distributed in the same way i.e. normally, and derived the PDF for the normal distribution starting from that observation (more about de Moivre, Gauss and other contributors to work on the normal distribution **here**
- Claude Elwood Shannon (1916 - 2001) - the father of information theory; defined **information entropy**, which can be used to derive PDFs or discrete probability distributions through a process of entropy maximisation under given constraints

Properties

The *kurtosis* and *skewness* of the normal distribution are 0.




area = 0.95
−∞
μ − 2σ
μ
μ + 2σ
∞


area = 0.997
−∞
μ − 3σ
μ
μ + 3σ
∞

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

[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.