Descriptive statistics

SET07106 Mathematics for Software Engineering

School of Computing
Edinburgh Napier University
Module Leader: Uta Priss

2010

Outline

Percentages

Central tendency

Other measures

Graphical representations

Percentages

$3\% = \frac{3}{100} = 0.03$
$4\% = \frac{4}{100} = 0.04$
$40\% = \frac{40}{100} = \frac{2}{5} = 0.4$

Investing £20 for a year with a 3% interest rate?

Percentages

$3\% = \frac{3}{100} = 0.03$
$4\% = \frac{4}{100} = 0.04$
$40\% = \frac{40}{100} = \frac{2}{5} = 0.4$

Investing £20 for a year with a 3% interest rate?

$20 \times \frac{3}{100} = 20 \times 0.03 = 0.6$

The result is £20 + £0.6 = £20.60

Exercise: which ones belong together?

This medicine ...

- has a 90% survival rate
- cures 75% of the patients
- has deadly side-effects in 0.2% of patients
- has a 50% probability of curing a patient
- is 200% safe
- has a 10% success rate
- has an 11% success rate

- helps 1 in 10 patients
- kills 1 in 10 patients
- helps every 9th patient
- nonsense
- is useless half of the time
- kills 1 in 500 patients
- useless for every 4th patient

Example: statistics and politics

Country A has an unemployment rate of 8%.

Country B has an unemployment rate of 10%.

Which country has a higher unemployment rate?
Which country has more unemployed people?

How to count

- At what age does employment start?
- What is the age of retirement?
- Are people on welfare counted as unemployed?
- Are people in training courses considered unemployed?
- What about young people who can't find a job and go to College instead?
- What is the employment status of people on parental leave?
- What about part-time or seasonal workers?

Measures of central tendency

- Mode: the most frequent value in the data set
- Median: middle value separating the higher half from the lower half
- Arithmetic mean: sum of all values divided by the number of values
- Geometric mean: the *n*th root of the product of the values where *n* is the number of values

Nominal (or categorical) data

Labels or names without any particular ordering.

Examples: names, phone numbers, number plates

The **mode** can be calculated:
Example: the most commonly used first names for boys and girls in a particular year.

No other measure can be calculated:
Example: the average of phone numbers does not makes sense.

Ordinal data

Rank orders

Examples:
rank order in a competition (first $\leq$ second $\leq$ third);
bed sizes (single $\leq$ double $\leq$ queen $\leq$ king);
coffee sizes (small $\leq$ medium $\leq$ large)

The **mode** and **median** can be calculated.

Measures for ordinal data

Example: a store sells the following beds in one day:
10 single, 5 double, 6 queen, 1 king.

The **mode** is:

Measures for ordinal data

Example: a store sells the following beds in one day:
10 single, 5 double, 6 queen, 1 king.

The **mode** is: single

22 beds:
$s \leq s \leq s \leq s \leq s \leq s \leq s \leq s \leq s \leq s < d \leq d \leq d \leq d \leq d < q \leq q \leq q \leq q \leq q \leq q < k$
The **median** is:

Measures for ordinal data

Example: a store sells the following beds in one day:
10 single, 5 double, 6 queen, 1 king.

The **mode** is: single

22 beds:
$s \leq s \leq s \leq s \leq s \leq s \leq s \leq s \leq s \leq s < d \leq d \leq d \leq d \leq d <$
$q \leq q \leq q \leq q \leq q \leq q < k$
The **median** is: double (the 12th value in the list)

Even if these are converted into numbers ($1 =$ single ... $4 =$ king), it would not make sense to calculate the **mean** (1.91). What is single.91?

Exercise

Calculate the mode and median for this example.
Can the mean be calculated?

4 children	1 family
3 children	2 families
2 children	10 families
1 children	17 families
0 children	10 families

Exercise

Calculate the mode and median for this example.
Can the mean be calculated?

4 children	1 family
3 children	2 families
2 children	10 families
1 children	17 families
0 children	10 families

Note: if the distances between the ordinal values are approximately the same, the mean may be calculated, although it depends on the application domain.

Interval data

Numerical data. Distances can be calculated.
The 0 point can be anywhere on the scale.

Examples: temperature in Celsius

Intervals can be calculated (for example: $10^\circ\mathrm{C} - 5^\circ\mathrm{C} = 5^\circ\mathrm{C}$)
But these values cannot be multiplied! ($0^\circ\mathrm{C} \times 5^\circ\mathrm{C} = 0^\circ\mathrm{C}$)

Measures for interval data

The **mode**, **median** and **arithmetic mean (average)** can be calculated.

Example: lowest temperatures in Edinburgh, Jan 4 - 10, 2010:

Jan 4	$-8^{\circ}\mathrm{C}$
Jan 5	$-10^{\circ}\mathrm{C}$
Jan 6	$-10^{\circ}\mathrm{C}$
Jan 7	$-11^{\circ}\mathrm{C}$
Jan 8	$-11^{\circ}\mathrm{C}$
Jan 9	$-6^{\circ}\mathrm{C}$
Jan 10	$1^{\circ}\mathrm{C}$

Mode:

Measures for interval data

The **mode**, **median** and **arithmetic mean (average)** can be calculated.

Example: lowest temperatures in Edinburgh, Jan 4 - 10, 2010:

Jan 4	$-8°\mathrm{C}$
Jan 5	$-10°\mathrm{C}$
Jan 6	$-10°\mathrm{C}$
Jan 7	$-11°\mathrm{C}$
Jan 8	$-11°\mathrm{C}$
Jan 9	$-6°\mathrm{C}$
Jan 10	$1°\mathrm{C}$

Mode: -10, -11 (bimodal)
Median:

Measures for interval data

The **mode**, **median** and **arithmetic mean (average)** can be calculated.

Example: lowest temperatures in Edinburgh, Jan 4 - 10, 2010:

Jan 4	$-8°\mathrm{C}$
Jan 5	$-10°\mathrm{C}$
Jan 6	$-10°\mathrm{C}$
Jan 7	$-11°\mathrm{C}$
Jan 8	$-11°\mathrm{C}$
Jan 9	$-6°\mathrm{C}$
Jan 10	$1°\mathrm{C}$

Mode: -10, -11 (bimodal)
Median: -10 (the 4th value: -11,-11,-10,-10,-8,-6,1)
Arithmetic mean (average):

Measures for interval data

The **mode**, **median** and **arithmetic mean (average)** can be calculated.

Example: lowest temperatures in Edinburgh, Jan 4 - 10, 2010:

Jan 4	$-8°\mathrm{C}$
Jan 5	$-10°\mathrm{C}$
Jan 6	$-10°\mathrm{C}$
Jan 7	$-11°\mathrm{C}$
Jan 8	$-11°\mathrm{C}$
Jan 9	$-6°\mathrm{C}$
Jan 10	$1°\mathrm{C}$

Mode: -10, -11 (bimodal)
Median: -10 (the 4th value: -11,-11,-10,-10,-8,-6,1)
Arithmetic mean (average): -7.86 $= \frac{-11-11-10-10-8-6+1}{7}$

Ratio data

Same as interval data, but the lowest measurements start at 0. There are no negative values.

Examples: height, length, temperature in Kelvin

All the previous measures can be calculated (mode, median, average), but also the **geometric mean**.

Geometric mean

The geometric mean is often used for data that grows exponentially.

Example: a chain letter that is sent to 15 people

1 iteration	15
2 iteration	225
3 iteration	3375
4 iteration	50625
5 iteration	759375

Arithmetic mean (average):

Geometric mean

The geometric mean is often used for data that grows exponentially.

Example: a chain letter that is sent to 15 people

1 iteration	15
2 iteration	225
3 iteration	3375
4 iteration	50625
5 iteration	759375

Arithmetic mean (average): 162723
Geometric mean:

Geometric mean

The geometric mean is often used for data that grows exponentially.

Example: a chain letter that is sent to 15 people

1 iteration	15
2 iteration	225
3 iteration	3375
4 iteration	50625
5 iteration	759375

Arithmetic mean (average): 162723
Geometric mean:
$3375 = \sqrt[5]{15 \times 15^2 \times 15^3 \times 15^4 \times 15^5} = \sqrt[5]{15^{15}} = 15^3$

Which value is more "in the middle"?

Summary 1

	nominal	ordinal	interval	ratio
Defined categories	x	x	x	x
Some ordering		x	x	x
Differences can be calculated			x	x
The lowest point is 0				x

Summary 2

	nominal	ordinal	interval	ratio
mode	x	x	x	x
median		x	x	x
arithmetic mean		?	x	x
geometric mean				x

Other statistical measures

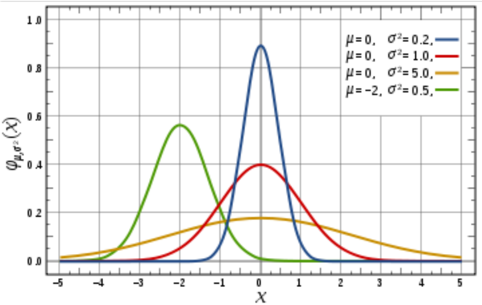
Dispersion:

- Variance, standard deviation: how close the data is to the mean.

Association:

- Correlation: whether two sets of data are related or dependent on each other.

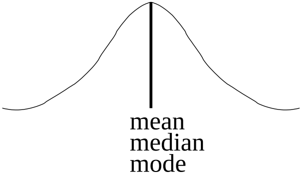
Normal distribution (Gaussian distribution)



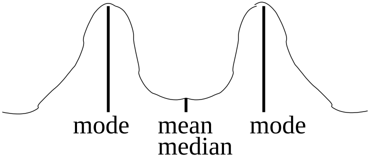
Mean $=$ mode $=$ median $= \mu$
Variance $= \sigma^2$
Standard deviation $= \sigma$

(Source: `http://commons.wikimedia.org/wiki/File:Normal_Distribution_PDF.svg`)

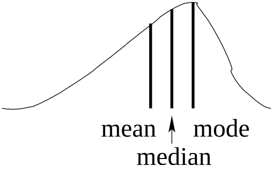
Normal distribution



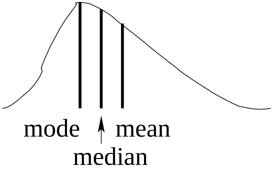
Bimodal



Skewed to left



Skewed to right



Standard deviation

The standard deviation measures how close or far the values are from the mean.

Ideally, the standard deviation should be close to 1.

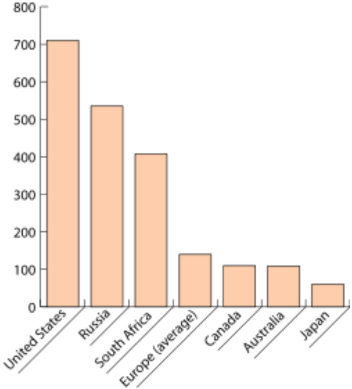
Unlike the variance (which is the square of the standard deviation), the standard deviation is expressed in the same units as the data.

Example

In the School of Computing the following measures are calculated for student marks:

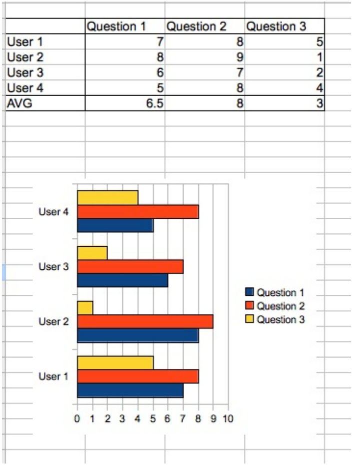
- **Mean**: in order to compare one module with instances of the module in previous years or other modules.
- **Standard deviation**: this can be high if, for example, several students don't complete all parts of an assessment, marks are capped for students who submit late or there is something wrong with the marking scheme.
- A **correlation** measure that compares student marks across modules. If the same students get much better (worse) results in one module than another the module results may be scaled down (up).

Bar chart



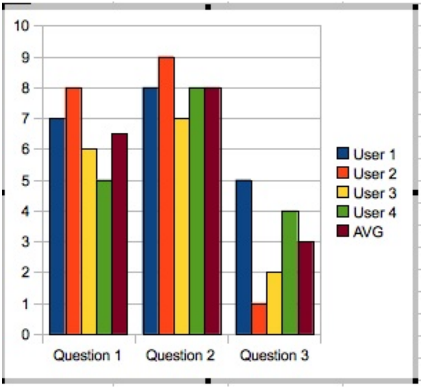
Bar charts represent frequencies of nominal/categorical data. (Source: `http://commons.wikimedia.org/wiki/File:Incarceration_Rates_Worldwide_ZP.svg`)

What is wrong with this bar chart?

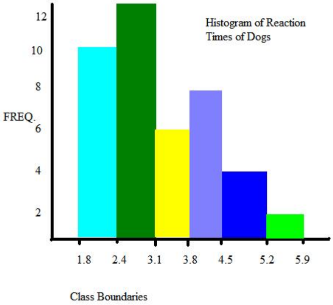


	Question 1	Question 2	Question 3
User 1	7	8	5
User 2	8	9	1
User 3	6	7	2
User 4	5	8	4
AVG	6.5	8	3

Improved version of the bar chart

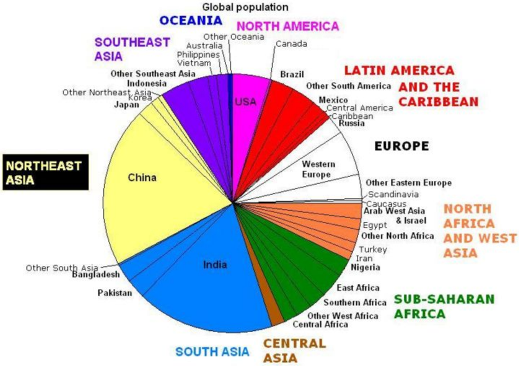


Histogram: reaction times of dogs



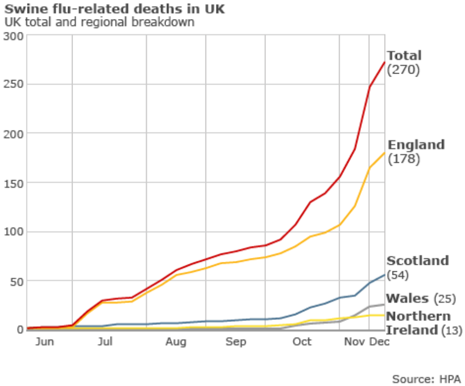
Histograms represent frequencies of interval data.
(Source: `http://commons.wikimedia.org/wiki/File:-8_Histogram.JPG`)

Pie chart: World population 2008



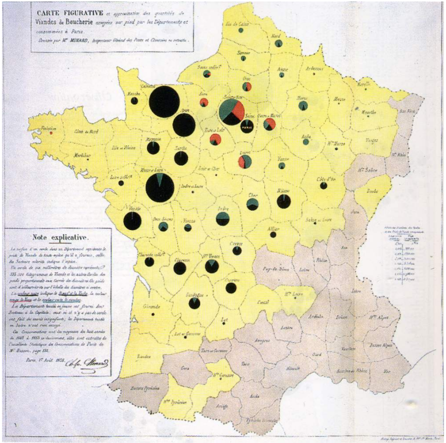
(Source: `http://commons.wikimedia.org/wiki/File:World_population_pie_chart.PNG`)

Line graph



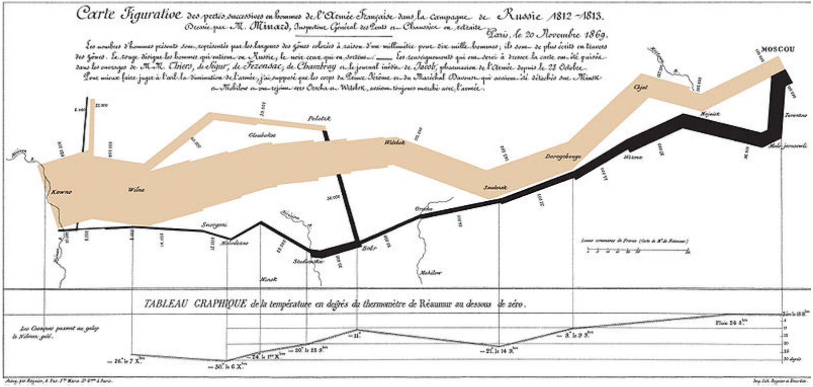
This line graph appeared on a BBC website on 3.12.09. What is odd about this graph?

Combining several aspects in one graph



Minard (1858): cattle sent from all around France for consumption in Paris

Graphical representation is an art!


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de MM. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Kowno
Wilna
Polotzk
Gloubokoe
Witebsk
Smolensk
Dorogobouge
Wixma
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Smorgoni
Molodetschno
Studienska
Minsk
Bobr
Orscha
Mohilow
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
Pluie 24 8.bre
Zéro le 18 8.bre
— 9° le 9 9.bre
— 21° le 14 9.bre
— 11°
— 20° le 28 9.bre
— 24° le 1.er X.bre
— 30° le 6 X.bre
— 26° le 7 X.bre

Minard (1869): Napoleon's Russian campaign of 1812