F20DL Data Mining and Machine Learning

Diana Bental
(with slide material from David Corne and Nick Taylor)

Lecture 6 Statistics

Fundamental Statistics Definitions

- A ***Population*** is the total collection of all items/individuals/events under consideration
- A ***Sample*** is that part of a population which has been observed or selected for analysis
  - E.g. **all students** is a population.
  - **Students at HWU** is a sample; **this class** is a sample, etc
- A ***Statistic*** is a measure which can be computed to describe a characteristic of the sample (e.g. the sample mean)
- The reason for doing this is almost always to estimate (i.e. make a good guess) things about that characteristic in the population

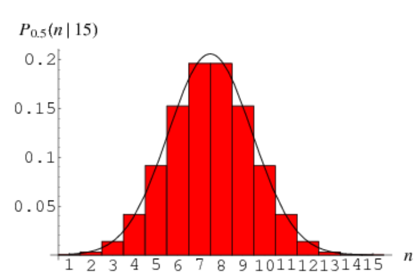
For example….

- This class is a sample from the **population** of students at HWU
- … it can also be considered as a sample of other **populations** – like what?
- One statistic of this sample is your mean weight. Suppose that is 65Kg. i.e. this is the sample mean
  - Is 65Kg a good estimate for the mean weight of the **population**?
- Another statistic: suppose 10% of you are married.
  - Is this a good estimate for the proportion that are married in the **population?**

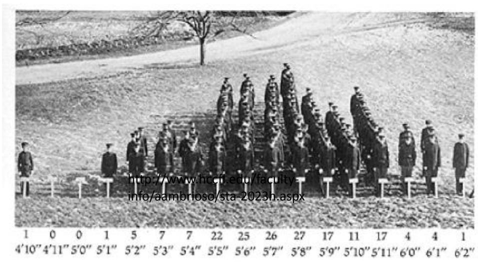
Some Simple Statistics

- The **Mean** (average) is the sum of the values in a sample divided by the number of values
- The **Median** is the midpoint of the values in a sample (50% above; 50% below) after they have been ordered (e.g. from the smallest to the largest)
- The **Mode** is the value that appears most frequently in a sample
- The **Range** is the difference between the smallest and largest values in a sample
- The **Variance** is a measure of the dispersion of the values in a sample – how closely the observations cluster around the mean of the sample
- The **Standard Deviation** is the square root of the variance of a sample.

Distributions / Histograms



A **Normal** (aka Gaussian) distribution (image from Mathworld)



http://www.hccfl.edu/faculty-info/aambrioso/sta-2023h.aspx

Figure 1.5
Differences in height in the same population: heights of conscripts over 60 years ago. (From A. Blakeslee, *Journal of Heredity*, vol. 5, 1914.)

http://www.hccfl.edu/faculty-info/aambrioso/sta-2023h.aspx

`Normal' or Gaussian distributions …

- … tend to be everywhere
- Given a typical numeric field in a typical dataset, it is common that most values are centred around a particular value (the mean), and the proportion with larger or smaller values tends to tail off.

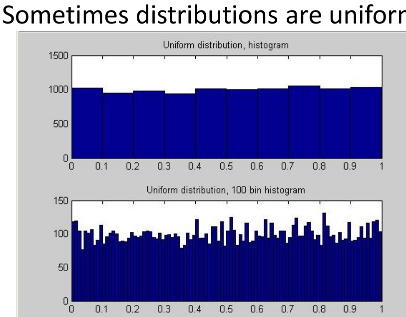
`Normal' or Gaussian distributions …

- … tend to be everywhere
- Given a typical numeric field in a typical dataset, it is common that most values are centred around a particular value (the mean), and the proportion with larger or smaller values tends to tail off.

`Normal' or Gaussian distributions …

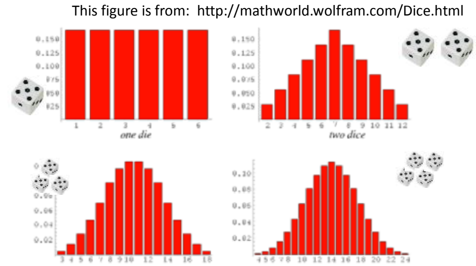
- Heights, weights, times (e.g. for 100m sprint, for lengths of software projects), measurements (e.g. length of a petal, waist measurement, coursework marks, level of protein A in a blood sample, …) all tend to be Normally distributed. **Why**??

Sometimes distributions are uniform



Uniform distributions. Every possible value tends to be equally likely

This figure is from: http://mathworld.wolfram.com/Dice.html



One die: uniform distribution of possible totals:
But look what happens as soon as the value is **a sum of things**;
The more things, the more Gaussian the distribution.
Are measurements (etc.) usually the sum of many factors?

Probability Distributions

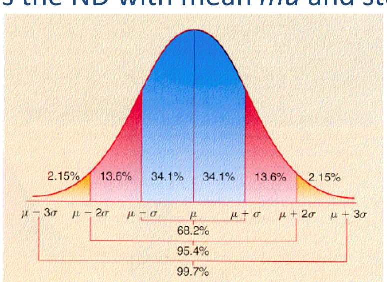
- If a population (e.g. field of a dataset) is expected to match a standard probability distribution then a wealth of statistical knowledge and results can be brought to bear on its analysis
- statistics of a sample provide info about a sample but…
- if we can assume that our statistic is normally distributed in the population, then our sample statistic provides info about the population

The power of assumptions…

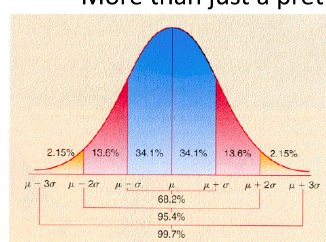
- You are a random sample of 30 (ish) HWU/Riccarton students. Suppose:
  - The mean height of this **sample** is 1.685cm
  - There are 5,000 students in the population
- With no more information, what can we say about the mean height of the **population** of 5,000 students?

A closer look at the **normal** distribution

This is the ND with mean *mu* and std *sigma*



More than just a pretty bell shape

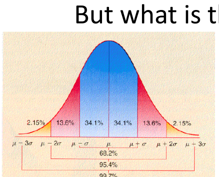


Suppose the standard deviation of your sample is 0.12

Theory tells us that if a population is Normal, the sample std is a fairly good guess at the population std

So, the sample STD is a good estimate for the population STD
So we can say, for example, that ~95% of the population of 5000 students (4750 students) will be within 0.24m of the population mean

But what is the population mean?



Mean of our **sample** was 1.685

The *Standard Error of the Mean is*

pop std / sqrt(sample size)

which we can approximate by:

sample std / sqrt(sample size)

… in our case this 0.12/5.5 = 0.022

This 'standard error' (SE) is actually the standard deviation of the distribution of sample means We can use this it to build a confidence interval for the actual population mean. **Basically, we can be 95% sure that the pop mean is within 2 SEs of the sample mean …**

The power of assumptions…

- You are a random sample of 30 (ish) HWU/Riccarton students. Suppose:
  - The mean height of this **sample** is 1.685cm
  - There are 5,000 students in the population
- With no more information, what can we say about the mean height of the **population** of 5,000 students?

If we assume the *population is normally distributed*
…. our sample std (0.12) is a good estimate of the pop std
….. so, means of samples of size 30 will generally have their own std, of 0.022 (calculated on last slide)
… so, we can be 95% confident that the pop mean is between 1.641 and 1.729 (2 SEs either side of the sample mean)

This is a good time to mention:
Z-normalisation
(converting measurements to z-scores)

Given any collection of numbers (e.g. the values of a particular field in a dataset) we can work out the mean and the standard deviation.

Z-Normalisation

- We looked at min-max normalisation
  - Scale the values so the lowest value is 0 and the highest is 1
  - And the remainder fall in between
- Z-score normalisation means converting the numbers into *units of standard deviation*.

Simple z-normalisation example

values
2.8
17.6
4.1
12.7
3.5
11.8
12.2
11.1
15.8
19.6

Simple z-normalisation example

values
2.8
17.6
4.1
12.7
3.5
11.8
12.2
11.1
15.8
19.6

Mean: 11.12
STD: 5.93

Simple z-normalisation example

Mean: 11.12
STD: 5.93

subtract mean, so that these are centred around zero

values	Mean subtracted
2.8	-8.32
17.6	6.48
4.1	-7.02
12.7	1.58
3.5	-7.62
11.8	0.68
12.2	1.08
11.1	-0.02
15.8	4.68
19.6	8.48

Simple z-normalisation example

Mean: 11.12
STD: 5.93

subtract mean, so that these are centred around zero

Divide each value by the std; we now see how usual or unusual each value is

values	Mean subtracted	In *Z* units
2.8	-8.32	-1.403
17.6	6.48	1.092
4.1	-7.02	-1.18
12.7	1.58	0.27
3.5	-7.62	-1.28
11.8	0.68	0.11
12.2	1.08	0.18
11.1	-0.02	-0.003
15.8	4.68	0.79
19.6	8.48	1.43

A bit more basic statistics: Correlation and Regression

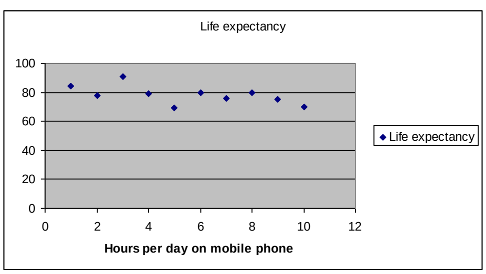
- Correlation
  - Understanding whether two fields of the data are related
  - Can you predict one from the other?
  - Or is there some underlying cause that affects both?
- Basic Regression
  - Very, very often used
  - Given that there is a correlation between A and B (e.g. hours of study and performance in exams; height and weight ; radon levels and cancer, etc … this is used to *predict* B from A.
  - Linear Regression (predict value of B from value of A)
  - But be careful…..

Correlation

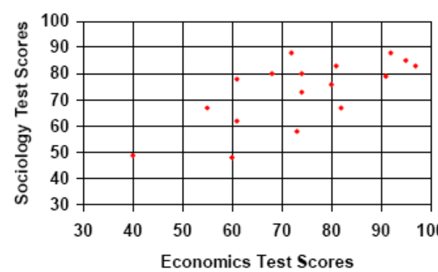
Are these two things correlated?

Phone use (hrs)	Life expectancy
1	84
2	78
3	91
4	79
5	69
6	80
7	76
8	80
9	75
10	70

Correlation

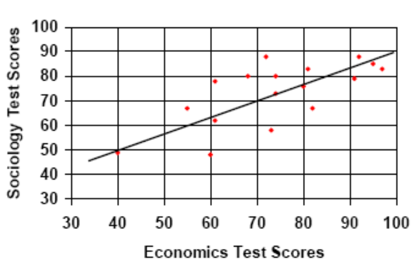


What about these (web credit)



David Corne, and Nick Taylor, Heriot-Watt University - dwcorne@gmail.com
These slides and related resources:
http://www.macs.hw.ac.uk/~dwcorne/Teac

What about these (web credit)



Correlation Measures

- It is easy to calculate a number that tells you how well two things are correlated. The most common is "Pearson's **R**"
- The **r** measure is:

  r = 1 for perfectly positively correlated data (as A increases, B increases, and the line exactly fits the points)

  r = -1 for perfectly negative correlation (as A increases, B decreases, and the line exactly fits the points)

Correlation Measures

r = 0 No correlation – there seems to be not the slightest hint of any relationship between A and B

- More general and usual values of *r*:
  - if **r >= 0.9** (r <= -0.9) -- a `strong' correlation
  - else if **r >= 0.65** (r <= -0.65) -- a moderate correlation
  - else if **r >= 0.2** (r <= -0.2) -- a weak correlation,

Calculating **r**

- You will remember the Sample standard deviation, when you have a sample of *n* different values whose mean is μ

Sample std is square root of $$\frac{1}{(n-1)} \cdot \sum_{x \in Sample} (x-\mu)^2$$

Calculating **r**

If we have pairs of (x,y) values, Pearson's **r** is:

$$\frac{1}{(n-1)} \cdot \sum_{(x,y) \in Sample} \frac{(x-\mu_x)}{std_x} \cdot \frac{(y-\mu_y)}{std_y}$$

Interpretation of this should be obvious (?)

Correlation (Pearson's **R**) and covariance

As we just saw, this is Pearson's **R**

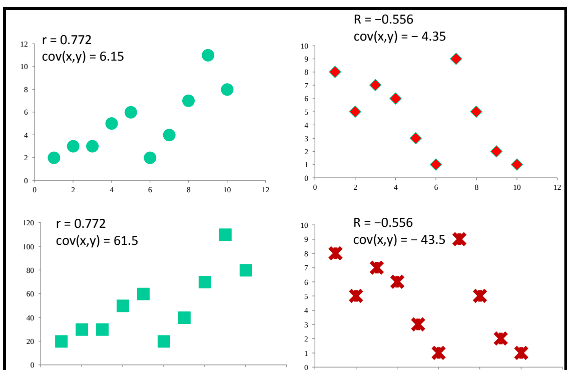
$$\frac{1}{(n-1)} \cdot \sum_{(x,y) \in Sample} \frac{(x-\mu_x)}{std_x} \cdot \frac{(y-\mu_y)}{std_y}$$

Correlation (Pearson's R) and covariance

And *this* is the covariance between x and y

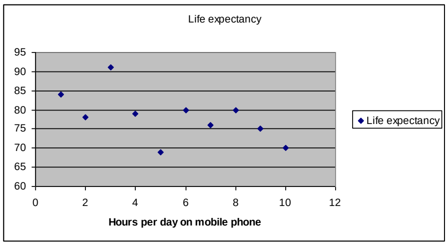
$$\frac{1}{(n-1)} \cdot \sum_{(x,y) \in Sample} (x-\mu_x)\ (y-\mu_y)$$

often called cov(x,y)

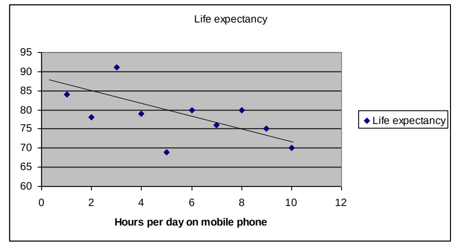


David Corne, and Nick Taylor, Heriot-Watt University - dwcorne@gmail.com
These slides and related resources:

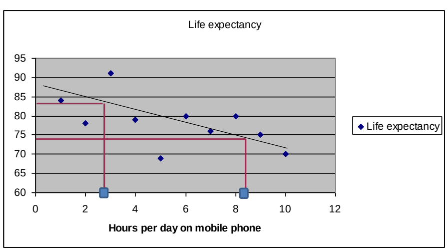
So, we calculate *r*, we think there is a correlation, what next?



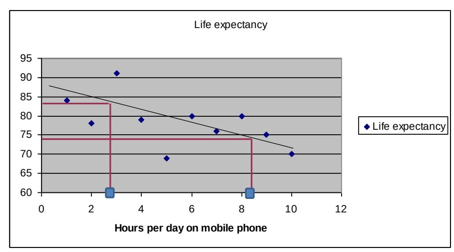
We find the 'best' line through the data



We can now make predictions! We have just done 'regression'



So, how do we calculate the best line?

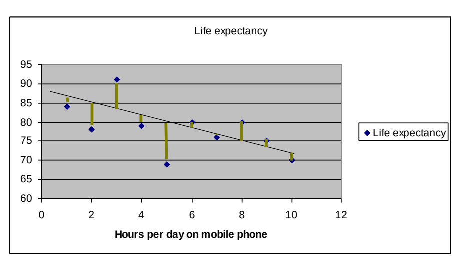


Best line means …

- We want the "closest" line, which *minimizes* the error. I.e. it minimizes the degree to which the actual points stray from the line . For any given point, its distance from the line is called a *residual.*
- We *assume* that there really is a **linear relationship** (e.g. for a top long distance runner, time = miles x 5 minutes), but we expect that there is also random `**noise**' which moves points away from the line (e.g. different weather conditions, different physical form, when collecting the different data points). We want this noise (deviations from line) to be as small as possible.
- We also want the **noise** to be unrelated to A (as in predicting B from A) – in other words, we want the correlation between A and the noise to be 0.

David Corne, and Nick Taylor, Heriot-Watt University - dwcorne@gmail.com
These slides and related resources:

Noise/Residuals: the residuals



- We want the line, which *minimizes* the sum of the (squared) residuals.
- And we want the residuals themselves to have zero correlation with the variables

- When there is only one *x*, it is called univariate regression, and closely related to the *correlation* between *x* and *y*
- In this case the mathematics turns out to be equivalent to simply working out the correlation, plus a bit more

Calculating the regression line in univariate regression

- First, recall that any **line** through (x,y) points can be represented by: $y = mx + c$

where *m* is the gradient, and *c* is where it crosses the y-axis at x=0

Calculating the regression line

$$y = mx + c$$

To get *m*, we can work out Pearson's *r*, and then we calculate:

$$m = r\left(\frac{std_y}{std_x}\right)$$

to get *c*, we just do this:

$$c = \mu_y - m\mu_x$$

Job done

Now we can:

- Try to gain some insight from the slope *m*
  - e.g. "since *m* = −4, this suggests that every extra hour of watching TV leads to a reduction in IQ of 4 points
  - Or "since *m* = 0.05, it seems that an extra hour per day of mobile phone use leads to a 5% increase in likelihood to develop brain tumours".

BUT….

- Be careful –
- It is easy to calculate the regression line, but always remember what the value of ***r*** actually is, since this gives an idea of how accurate is the assumption that the relationship is really linear.
- So, the regression line might suggest:
  - "… extra hour per day of mobile phone use leads to a 5% increase in likelihood to develop brain tumours"

but if ***r*** ~ 0.3 (say) we can't say we are very confident about this. Either not enough data, or the relationship is different from linear

And … what about a dataset with more than one non-class field?

The general solution to *multiple regression*

Suppose you have a numeric dataset, e.g.

```
3.1  3.2  4.5  4.1  …  2.1  1.8    5.1
4.1  3.9  2.8  2.4  …  2.0  1.5    3.2
…
6.0  7.4  8.0  8.2  …  7.1  6.2    9.5
```

The general solution to *multiple regression*

Suppose you have a numeric dataset, e.g.

X **y**

```
3.1  3.2  4.5  4.1  …  2.1  1.8    5.1
4.1  3.9  2.8  2.4  …  2.0  1.5    3.2
…
6.0  7.4  8.0  8.2  …  7.1  6.2    9.5
```

The general solution to *multiple regression*

Suppose you have a numeric dataset, e.g.

X **y**

```
3.1  3.2  4.5  4.1  …  2.1  1.8    5.1
4.1  3.9  2.8  2.4  …  2.0  1.5    3.2
…
6.0  7.4  8.0  8.2  …  7.1  6.2    9.5
```

A linear *classifier* or *regressor* is:

$\mathbf{y}_n = \beta_1\mathbf{x}_{n1} + \beta_2\mathbf{x}_{n2} + \ldots \beta_m\mathbf{x}_{nm}$ - So, how do you get the β values?

The general solution to *multiple regression*

By straightforward linear algebra. Treat the data as matrices and vectors …

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

The general solution to *multiple regression*

By straightforward linear algebra. Treat the data as matrices and vectors …

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

And the solution is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

And that's general multivariate linear regression.

If the data are all numbers, you can get a *prediction* for your 'target' field by deriving the beta values with linear algebra.

Note that this is *regression*, meaning that you predict an actual real number value, such as wind-speed, IQ, weight, etc..., rather than *classification*, where you predict a class value, such as 'high', 'low', 'clever', ...

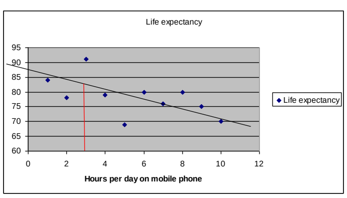
But ... easy to turn this into classification ... how?

Pros and Cons

- Arguably, there's no 'learning' involved, since the beta values are obtained by direct mathematical calculations on the data
- It's pretty fast to get the beta values too.
- HOWEVER
  - In many, many cases, <u>the 'linear' assumption is a poor one</u> – the predictions simply will not be as good as, for example, decision trees, neural networks, k-nearest-neighbour, etc ...
  - The maths/implementation involves getting the inverses of large matrices. This sometimes explodes.

Prediction

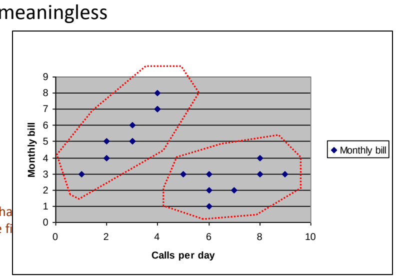
Commonly, we can of course use the regression line for prediction:



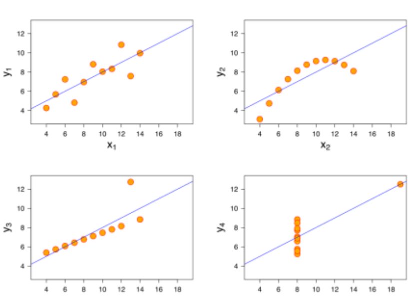
So, predicted life expectancy for 3 hrs per day is: 82 - fine

BUT

If correlation is not strong, the regression line may be meaningless



... or, outliers may be distorting the picture,
... or, a straight line is simply not an appropriate model for the data

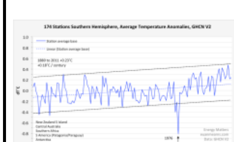


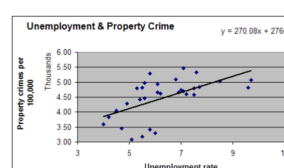
r is 0.816

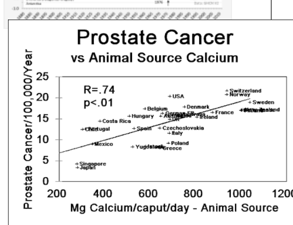
y = 3.00 + 0.500*x*

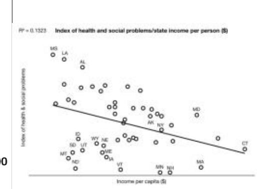
Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21.

Correlation and regression are used everywhere in science and industry.

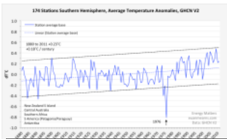


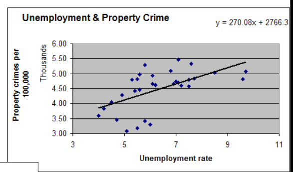


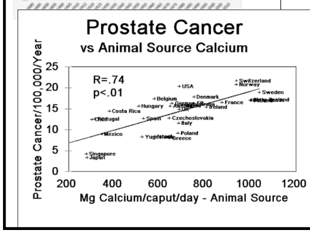


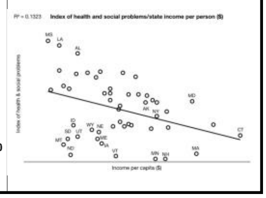


Correlation and regression are used everywhere in science and industry.









But see here … http://www.tylervigen.com/spurious-correlations