DS Visualisation and Analysis

Abbey Waldron

This Week - Distributions and Research

1. Show last week's work to the class
2. Sampling
3. Applied statistics, means, modes, variances, errors *etc.*
4. Parameter measurement
5. Start work on problems

Random Student Generator

Q1

Make a cubic spline interpolation of the total cases data. Use your spline to predict the total number of cases on outbreak day 158.

Q1 Answer

261

Q2

Test out a few different fits to the total cases data, including the following. Show your results on plots and give your best fit parameters:

- Linear
- Quadratic
- Exponential

Q3

What do you think is a good function to fit to the total cases data? Justify your answer from theory.

Q4

Fit your solution to the data, show the fit on a plot and quote the fitted parameters and goodness of fit.

Q5

Predict how many cases of ebola there were in total on the first of March, 2015 using the 2014 data. Quote an error on your prediction (think…).

Q5 Answer

24k

Q6

*Optional: repeat the above procedures for the total number of deaths. What do you notice about the similarities/differences between the two?

This week

This week we will work on applying some basic concepts that you will have met in statistics class to real World problems.

Today's Experiment

Very often, we want to make a measurement of some quantity.

Let's suppose we want to know the average number of caffeinated drinks per day consumed by pupils in the 4th year of the Hogeschool, and we use this class as our sample…

Conduct Experiment

Sample Mean

How do we calculate the sample mean? Hopefully you remember this one from Statistics class:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \quad (1)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i \quad (2)$$

Where n of course is the number in this class and the x_i are your caffeinated drink numbers in this case.

Error on the Mean

What is the error on our measurement of the number of caffeinated drinks for the year?

If we would ask different classes, we would get different answers.

$$\text{Error on the Mean} = \frac{1}{\sqrt{n}}\sigma$$

So when we take a large enough sample we recover the true population mean.

Random Errors

Probably not all of you drink exactly the class mean number of caffeinated drinks per day!

This is not an error with our experimental procedure (me asking you) but rather a real feature of the caffeine addiction distribution.

Random Errors

Estimating the population variance:

$$\sigma^2 \approx \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3}$$

Systematic Errors

Should we expect this value to hold if we considered all students at the Hogeschool?

Now we get into the difficult part. Your measurements will have some uncertainty due to the finite size of your data set, but it may also have some inherent error due to how you have conducted your experiment.

Systematic Errors

In general the best strategy is to design your experiment to minimise such systematic errors - in this case we should use a better sampling method.

Quantification of them in general is beyond the scope of this course, one great option if possible if to perform your measurement using a different experiment and see how well the results agree.

Week 5 Problems

1. A more realistic research challenge
2. Describing distributions and finding outliers

Backup Slides