Pivigo statistics challenges

## Counting statistics

Q1) Examine the income of users in the data set 'Training\_predictions.txt'. Calculate a) the mean, b) median and c) modal income of a user.

***A1) a) 88200 b) 85000 c) 78000***

Q2) Find a) the standard deviation and b) variance of the age population age.

***A2) a) 37400 b) 1.4\*10^9***

## Simple probability

Q1) What is the probability of rolling at least one 6 when rolling three dice?

***A1) 0.421***

Q2) In a group of 50 marbles, 15 are clear and the remaining 35 has either red, blue or both. a) If 20 are red and 25 are blue, how many are red and blue?

What is the probability of drawing; b) a red marble, c) a blue marble and d) a red and blue marble, e) red OR blue or multi-coloured marble at random from a bag?

***A2 a) 10, b) 0.29, c) 0.36, d) 0.14, e) 0.79***

Q3) A fracking company is conducting exploratory drilling in three sites. Each site has 20% probability of proving profitable and each site it independent of the others. If successful, the payoff from a single site is $4 million, from two sites it is $3 million (each) and from all three sites it is $2.5 million each (due to excess supply lowering prices). Find the expected payoff.

***A3) $2.17 million***

## Counting probability

Q1) A media firm identify that 85% of users of their app are using it to read sports related news articles. Out of a random sample of 100 users, find the probability that a) exactly three quarters of the sample are reading sports articles?

***A1) 0.00311***

Q2) Calculate the probability being dealt a straight flush (5 consecutive cards of the same suit) off the top of a shuffled deck of cards?

***A2) 1.53907716933e-05***

Q3) Each Pivigo team is composed of 4 people drawn randomly from a pool of 30 women and 60 men. What is the probability that the teams are composed of ; a) two women and two men, b) four men, c) at least 1 women.

***A3) a) 0.3010, b) 0.1908, c) 0.8092***

## Conditional probability and classification

Q1) You are playing a game whereby you draw two cards at random from a deck of 52 playing cards with the aim of drawing matching suits. If you draw a spade, what is the probability that you will draw a second spade?

***A1) 0.235***

Q2) The fraction of S2DS students born outside of the UK and currently working in the UK is 7/20. The total fraction born outside of the UK is 2/3. Find the probability that a student is currently working in the UK, given that they were born outside of the UK?

***A2) 0.53***

Q3) What is the probability that the total of two dice will be greater than 7, given that the first die is a 4?

***A3) 2/3***

Q4) You are playing a game whereby you draw two cards at random from a deck of 52 playing cards with the aim of drawing matching suits. a) If you draw a spade, what is the probability that you will draw a second spade? b) What is the probability drawing two cards of the same suit?

***A4) a) 0.235, b) 0.018***

Q5) A pharmaceutical lab is testing a new viral detection drug. They find a false positive rate of 0.2% and a false negative rate of 0.9%. The virus is estimated to be present in 1 in 50,000 people in the population. Calculate a) the instances of the drug failing to detect the virus and b) the instances of the drug incorrectly identifying the virus (hint: in format of 1 per X number of people).

***A5 a) 1:5,600,000, b) 1:500***

Q6) 0.04% of the population was inflected with swine flu. An experimental test has a 99.75% chance of returning a positive result if a subject has the virus. a) Calculate the probability that a randomly selected person tests positive. b) What is the probability that a person is infected, given that they tested positive?

## *A6) a) 0.0030, b) 0.138*

## Normal distributions

Q1). For the following standard deviation intervals (σ) about the mean in a Gaussian (Normal) distribution, find the probability that a sample lies in that sample. A) ≤+/- 1 σ, B) ≤+/-3 σ, C) ≥+/-5 σ?

***A1) a) 0.683, b) 0.997, c) 0.00573.***

Q2). Likewise, find the following confidence intervals in terms of σ. A) 50%, B) 95%, C) 99%.

***A2) a) 0.674 σ, b) 1.64 σ, c) 2.58 σ.***

Q3) The S2DS 2016 cohort contained 90 people. The ages are distributed normally and have a mean age of 30 with a standard deviation of 5. A student is selected at random. What is the probability that; a) their age is greater than 35, b) their age is less than 22, c) their age is between 22 and 35, d) Approximately how many students are older than 35?

***A3) a) 0.1587, b) 0.0548, c) 0.7865, d) 14.***

Q4) The leader of the 'British Democratic Party' has a mean public satisfaction rating of -8% with a standard deviation of 12 (assume a normal distribution). Find a) the probability that a person chosen at random will have ranked them greater than +15 and b), if the survey size is 1053, calculate the number of individuals whose rankings are considered 'outliers' (+/-3sigma). c) What were the upper and lower bounds of their scores? Note: satisfaction rankings run from -100 to +100.

***A4) a) 0.0276 b) 2, c) 28 and -44***

## Other distributions

Q1) You are playing the game perudo. They are 6 players and a total of 36 dice in play. The player to your left calls 13 sixes (recall that rolls of one are a wildcard that can equal any other chosen roll). What is the probability that there are exactly 13 sixes in play?

***A1) 0.13***

Q2) A recruiter takes commission from on average 5 placements per month. Find the probability that a) they place at least one person per month, and b) the probability that they we fail to place anyone in any given week. The manager want's to improve sales and sets a target of 6 or more placements per month. c) calculate the probability of meeting this. (Hint: use Poisson's law and assume 4 weeks in a month).

***A2) a) 0.993, b) 0.287, c) 0.384***

Q3) The recruitment agency makes contact with 85 needy data scientists every month but only 7% are successfully placed with a client. A) what is the mean and standard deviation of placements per month?, B) what is the probability of you successfully finding a job with this agency within a month? C) 3 other S2DS graduates join this agency. What is the probability that at least one of you find a job within a month?

***A3) a) 6, 2, b) 0.0126, c) 0.244 (???)***

## Monte-Carlo

Q1) You are given the following function:

y = log(x)+x^(1/2)-3

And an input variable of x=5. a) calculate y. The standard deviation on x = 0.25. b) calculate the standard deviation on y numerically using a monte-carlo method. c) Compare this answer to that analytical solution found using the propagation of error.

A1) a) 0.846, b) 0.1060, c) 0.1059

## Hypothesis testing

Q1) We hypothesise that users older than 90 years are inherently wealthier. We propose a null hypothesis that states that "the 95% (or greater) confidence that the incomes of 90+ year old users IS consistent with being drawn randomly from the main sample".

Select a subsample of data for users over 90 years old. a) find it's mean and b) calculate its 'one sample Z test statistic'. C) find the 95% confidence interval of the 90 year old sample. D) Is the null hypothesis 'True' or 'False'.

***A1) a) 88500 b) 1.32 c) 465 d) True***

Q2) We hypothesise that customers of this company are more wealthy than the national average. We propose a null hypothesis that states that "there is a 95% or greater confidence that customers could have been drawn randomly from the national population".

The mean national income is 26500. a) Calculate the "student's" t-stat value and b) find the 95% confidence interval of the customer database (hint: consider the degrees of freedom to be inf.). C) is the null hypothesis true or false?

***A2) a) -16.5, b) 7331, c) False***

Q3) Taking the standard deviation on the mean national income to be 6000 and assuming the distribution is Gaussian; use the two sample KS-test to compare the distributions of incomes national, and in the company. Find a) the KS-statistic and b) the probability that each distribution is drawn from the same sample (p-value) - hint: as the population of the UK is 65 million but our data base is only 2 million you will need to normalise).

***A3) a) 0.36 b) 5.9%***

## Validation

Q1) Using the 'Modelling\_predictions.txt' data set. Find the a) mean square error (MSE) b) mean absolute error (MAE) c) symmetric mean absolute percentage error (SMAPE) d) co-efficient of determination (R^2) and e) Chi^2 value of the both the ORF and RRF Customer Value models, compared to the training set labelled Customer Values. D) Which is the best model?

***A2) a) MSE: ORF = 432 RRF = 6400 , b) MAE: ORF = 9.24 RRF = 39.8, c)SMAPE: ORF = 10.5 RRF = 43.9, d) R2: ORF = 0.944 RRF = 0.166, e) Chi2: ORF = 4110000 RRF = 158000000, D) ORF.***

# Model testing – move to ML???

Q1] Using the data set 'time\_series\_1.csv'. Fit a unweighted linear model to data and show that the coefficient of determination (R^2) is 97.4%.

Q2] Using the data set 'time\_series\_1.csv'. Fit a weighted linear model to data and show that the coefficient of determination (R^2) is 97.3%.

Q3] Using the data set 'time\_series\_2.csv'. Show that a weighted polynomial model, of order 3, is a better model than a weighted linear model, using Chi^2.

Q4] Use Chi^2 to find the polynomial order of the model that best fits the data? What one word describes why this value is not suitable for producing a predictive model?

***A4]. 9. Overfitting.***