# Assignment: Pearson's chi-square test

## Instructions

Answer the following questions. Show your working by including output from your R session. All four questions are of equal mark value.

## Question 1.

62 students in a certain statistics class are asked which is their favourite colour among Red, Orange, Blue, and Green. The results are as follows:

```
o <- c(red = 10, orange = 16, yellow = 24, green = 12)
```

- State a sensible null hypothesis
- State the observed values and the expected values under your null.
- Calculate the Pearson's chi-square statistic `sum((o-e)^2/e)` and show your working; state what its approximate null distribution is
- State the precise definition of p-value and explain what "more extreme" means in this context.
- Calculate the pvalue using `pchisq()` and your result from above, and interpret.
- Excellence question: using random simulation, or otherwise, estimate the probability that all four colours have different numbers of students.

## Question 2

A call centre logs the number of support calls received each day for a year and tabulates the results:

```
o <- c(c0=8, c1=12, c2=36, c3=54, c4=67,c5=66, c6=41, c7=37, c8=23, c9=10, c10=11)
o
```

```
##  c0  c1  c2  c3  c4  c5  c6  c7  c8  c9 c10
##   8  12  36  54  67  66  41  37  23  10  11
```

This means that on 8 days there were zero calls, on 12 days there was one call, on 36 days there were two call, and so on up to 11 days when there were 10 callouts. We wish to test the hypothesis that the number of calls on any day has a Poisson distribution.

- Give a plausible reason why the Poisson distribution might be appropriate
- Verify that the dataset contains 365 observations. Calculate the number of calls in the year and then calculate the average number of calls per day.
- Use your estimated value of the mean number of calls per day as $\lambda$ in the Poisson distribution to calculate the probability of having $0, 1, 2, \ldots, \geqslant 10$ calls on any day. Remember that the final probability is "10 calls *or more*" so ensure that your probabilities sum to one.
- Use R to calculate the expected number of calls with $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, \geqslant 10$ in the year.
- Use R to calculate the Badness-of-fit $B = \sum \frac{(o-e)^2}{e}$

- Calculate a $p$-value for your B and interpret (note that the degrees of freedom is now $11-1-1=9$: there are 11 categories, minus one because the total is known, minus another one because the expectation uses an estimated value for $\lambda$.
- excellence question: Someone observes that the number of days with no calls at all is quite high. Formulate a sensible null hypothesis and test it. Interpret and give a plausible reason for your finding.

## Question 3.

This question requires you to use R for random sampling.

Suppose we have a bag with one red, one green, and one blue ball in a bag. We make 12 draws with replacement and tally the number of red, blue, and green balls in our sample.

- How many red, green, and blue balls do you expect to observe?
- For an observation of `c(r=5, g=3, b=4)`, calculate the chi-square statistic. Justify your number of degrees of freedom.
- Using `sample()`, or otherwise, use R to generate a sequence of random draws.
- Generate some sequences of random draws. For these, calculate: the chi-square statistic $B$ and the $p$-value and present histograms or tables [you might find `tabulate(...,nbins=3)` useful]
- Excellence question: compare your empirical distribution of $B$ above with its asymptotic distribution.

## Question 4

Give an example of Pearson's chi-square test from your daily life. Have at least three categories. State your observation *clearly*, state your null and expected values *clearly*. Calculate and state the value of the chi-square statistic $B$. State and justify the null distribution of $B$ (including number of degrees of freedom), give a $p$-value. Interpret your findings.