Problem Set 7

Submit your solutions as a single PDF file via Canvas by **8am Tuesday February 28th**.

- If writing up problems by hand, please use a pen and not a pencil, as it is difficult to read scanned submission of pencil work. Typed solutions are preferred.

- For problems that require coding, Colab notebooks will be provided. Please copy and save the shared notebook and edit your own copy, which you should then submit by including a clickable link in your submitted homework. Prior to submission make sure that you code runs from beginning to end without any error reports.

Problem 1 (20 points)
   P-values are random variables

   In general, a hypothesis test involves a random variable T (also known as a test statistic), which has a (cumulative) distribution function $F(t)$ under the null hypothesis (e.g. T-distribution in T-tests). The p-value, $P(T)$, being a function of the test statistic, is also a random variable. Assuming the test statistic is a continuous random variable (so that F is an invertible function), show that the p-value for a one-sided test, $P(T) = 1 - F(T)$, is uniformly distributed between 0 and 1.

Problem 2 (40 points)

The Bonferroni correction is a conservative method for dealing with the multiple comparisons problem. The method works by controlling the familywise error rate (FWER), which is the probability of rejecting at least one null hypothesis when it is in fact true.

(a) (8 points) Suppose you are carrying out n independent hypothesis tests, for each test you will reject the null hypothesis if the p-value is less than α, where $0 < \alpha < 1$. Furthermore, suppose there are n_0 true null hypotheses, whose value is presumably unknown. Write down an expression for the FWER in terms of only n_0 and α.

(b) (8 points) Use your result from part (a), explain why making a large number of comparisons without adjusting the significance level α may be problematic (hint: n_0 is likely to be large when n is large).

(c) (8 points) Show that the FWER $\leq n_0\alpha$

(d) (8 points) Show that we can ensure FWER $\leq \alpha$ by setting the significance level of each individual hypothesis to be $\frac{\alpha}{n}$.

(e) (8 points) Compare the probability of not rejecting at least one null hypothesis when it is in fact false (false negative) when the significance level is set to α versus α/n, which significance level produces a higher number of false negatives?

Problem 3   (40 points)

In this problem you will be using various statistical tests to look for differential expression between cells in different cell types i.e. testing null versus alternative hypotheses that gene expression is the same or different between groups of cells. This will involve comparing mean expression values between different cell populations across genes, determining gene candidates with 'significant' differences in expression, and gauging how accurate or trustworthy such results are.

The Problem 3 notebook is here.

Your edited version of the notebook *must be submitted* for this problem. Reminder to check that your notebook runs all the way through with the the `Runtime` → `Restart` and `Runtime` → `Run All` commands.