**Intern Weekly Diary**

1. What is the main problem you are trying to solve this week?
2. What technical tools are you using?  E.g.  R and RShiny packages, Python libraries
3. How did you use these tools to try to solve the problem?
4. What web resources or reading materials have you been accessing for your work?

Week 1

1. Complete mandatory training (IG, H&S, Fire, DSE), read Starter Guide to CAS to familiarise myself with the available data, review Paul’s Simulacrum test suite code (SQL)
2. Oracle SQL Developer, Notepad
3. I have used this SQL software (and Notepad) to review and experiment with the existing codebase provided by Paul, as well as generally re-familiarise myself with using SQL
4. SVN repository: CAS general documents, Starter Guide to CAS;  
   SQL Tutorial: <https://www.w3schools.com/sql/default.asp>

Week 2

1. Mandatory ‘Healthy Working’ training; researching probability distributions and statistical hypothesis testing; reviewing, modifying, and running Paul’s Simulacrum test suite code (SQL, Python), and discussing preliminary results; discussing underlying assumptions and limitations of hypothesis tests to be highlighted in methodology and interpretation of results; started developing intern presentation.
2. Oracle SQL Developer, Python
3. Script development
4. Web resources and reading materials
   1. SQL Tutorials:
      1. <https://www.w3schools.com/sql/default.asp>
      2. <https://mode.com/sql-tutorial/sql-performance-tuning/>
      3. <https://jaxenter.com/10-sql-tricks-that-you-didnt-think-were-possible-125934.html>
   2. Probability distributions:
      1. [https://en.wikipedia.org/wiki/Binomial\_distribution](https://en.wikipedia.org/wiki/Binomial_distribution#Normal_approximation)
      2. <https://en.wikipedia.org/wiki/Categorical_distribution>
      3. <https://en.wikipedia.org/wiki/Multinomial_distribution>
      4. <https://en.wikipedia.org/wiki/Dirichlet_distribution>
      5. <https://en.wikipedia.org/wiki/Conjugate_prior>
      6. <https://en.wikipedia.org/wiki/Bayesian_network>
      7. <https://en.wikipedia.org/wiki/Gibbs_sampling#Variations_and_extensions>
      8. <https://en.wikipedia.org/wiki/Dirichlet-multinomial_distribution>
   3. Statistical Hypothesis Testing - Theory
      1. <https://en.wikipedia.org/wiki/Statistical_hypothesis_testing>
      2. <https://en.wikipedia.org/wiki/Test_statistic>
      3. <https://en.wikipedia.org/wiki/Type_I_and_type_II_errors>
      4. <https://en.wikipedia.org/wiki/Family-wise_error_rate>
      5. <https://en.wikipedia.org/wiki/False_discovery_rate>
      6. <https://en.wikipedia.org/wiki/Binomial_test>
      7. <https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test>
      8. <https://en.wikipedia.org/wiki/Multinomial_test>
      9. <https://en.wikipedia.org/wiki/Likelihood-ratio_test>
      10. <https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test>
      11. <https://en.wikipedia.org/wiki/Lilliefors_test>
      12. <https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test>
      13. <https://stats.stackexchange.com/questions/96921/comparing-two-distributions>
      14. <https://stats.stackexchange.com/questions/103207/statistical-description-of-how-two-distributions-are-different>
      15. <https://stats.stackexchange.com/questions/174024/can-you-use-the-kolmogorov-smirnov-test-to-directly-test-for-equivalence-of-two>
      16. <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm>
      17. <https://towardsdatascience.com/kolmogorov-smirnov-test-84c92fb4158d>
   4. Statistical Hypothesis Testing in practice
      1. <https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Tests_for_Two_Proportions.pdf>
      2. <https://www.coursera.org/lecture/biostatistics-2/two-sample-binomial-tests-comparing-2-binomial-proportions-eIL8m>
      3. <https://www.stat.washington.edu/peter/342/non_param.pdf>
      4. <https://www.stat.berkeley.edu/~stark/SticiGui/Text/chiSquare.htm>
      5. <https://www.biostat.wisc.edu/~kbroman/teaching/labstat/fourth/notes02.pdf>
      6. <http://www.r-tutor.com/elementary-statistics/goodness-fit/multinomial-goodness-fit>
      7. <https://stats.stackexchange.com/questions/43479/test-for-multinomial-distribution>
      8. <https://rstudio-pubs-static.s3.amazonaws.com/300209_c94d97787e7e45a28a19de02ee0e4636.html>
      9. <https://www.youtube.com/watch?v=OGctoeNVn5A>
      10. <https://www.youtube.com/watch?v=y6ezWg5ce0g>
   5. Machine Learning
      1. <https://link.springer.com/chapter/10.1007/978-3-030-21642-9_42>

Week 3

1. Goals for the week:
   1. Collect and join tables of values counts grouped by field for both real and simulated datasets
   2. Identify outlier fields or categories which perform particularly poorly across the board to inform caveats about Simulacrum capabilities
   3. Organise (as chair) Analysts East Team Meeting for August 2019
   4. Communicate with SACT team at Wellington House, London about collaborative work on running existing report-generation code on the simulated SACT datasets
2. Oracle SQL Developer, Python (sqlalchemy, getpass, pandas), Jupyter Notebook, [Oracle Instant Client](https://oracle.github.io/odpi/doc/installation.html#windows)
3. Script development, securely connecting Python to the SQL server to extract CAS data. Writing SQL code to join datasets: tumour tables and lookups in AV2017; joining summary tables. Documenting existing work.
4. Web resources and reading materials
   1. <https://www.linkedin.com/pulse/connecting-sql-database-python-patrick-j-ryan>
   2. <https://docs.sqlalchemy.org/en/13/core/engines.html>
   3. <https://pyformat.info/>
   4. <https://thispointer.com/python-how-to-copy-a-dictionary-shallow-copy-vs-deep-copy/>
   5. <https://stackoverflow.com/questions/9546324/adding-directory-to-path-environment-variable-in-windows>
   6. <https://cancerstats.ndrs.nhs.uk/sact/regimen2018>
   7. <https://docs.oracle.com/database/121/SQLRF/functions131.htm#SQLRF00684>
   8. <https://medium.com/@rukavina.andrei/how-to-write-a-python-script-header-51d3cec13731>
   9. <https://realpython.com/documenting-python-code/>
   10. <https://docs.python-guide.org/writing/structure/>
   11. <https://stackoverflow.com/questions/15109548/set-pythonpath-before-import-statements/15109660>
   12. <https://stackoverflow.com/questions/3108285/in-python-script-how-do-i-set-pythonpath>

Week 4

1. Goals for the week:
   1. Writing code to statistically compare values counts grouped by categories in data fields across a pair of datasets (e.g. real vs. simulated datasets) – automating and adapting a binomial-model one-sample z-test based on Paul’s previous work.
   2. Developing and implementing alternative testing methods: Pearson’s Chi-squared, Likelihood-Ratio, binomial-model pooled two-sample z-test; assessing and comparing the results.
   3. Identify outlier fields or categories which perform particularly poorly across the board to inform caveats about Simulacrum capabilities – (e.g. Change in Laterality coding, simulating ages 94+)
   4. Communicating test results graphically using data visualization techniques
   5. Meeting with SACT team at Wellington House, London for collaborative project to produce data visualization reports on the simulated SACT datasets to enable their publication as well as for comparison with the real SACT reports (limited release)
2. Notepad++, Oracle SQL Developer, Python (sqlalchemy, getpass, pandas, matplotlib, ipywidgets), Jupyter Notebook, [Oracle Instant Client](https://oracle.github.io/odpi/doc/installation.html#windows)
3. Notepad++, Oracle SQL Developer for SQL script development and documentation; Python (pandas, matplotlib, ipywidgets), Jupyter Notebook for automating database requests and interactively plotting the results
4. Web resources and reading materials
   1. <https://en.wikipedia.org/wiki/G-test>
   2. <https://en.wikipedia.org/wiki/Binomial_distribution#Normal_approximation>

Week 5

1. Goals for the week:
   1. Continuing work on Simulacrum-SACT reporting project (writing, testing SQL code)
   2. Developing and refining tools for visualizing dataset comparisons and test results
   3. Preparing for and delivering a presentation at HDI’s Intern Presentation Day
2. Notepad++, Oracle SQL Developer, Python (sqlalchemy, getpass, pandas, matplotlib, ipywidgets), Jupyter Notebook, [Oracle Instant Client](https://oracle.github.io/odpi/doc/installation.html#windows), Microsoft PowerPoint
3. Script development and documentation, producing tables graphs of results
4. Web resources and reading materials
   1. <https://www.oracletutorial.com/oracle-basics/oracle-select-distinct/>
   2. <https://www.neowin.net/forum/topic/974024-sql-select-multiple-columns-with-only-one-distinct-column/>

Week 6

1. Goals for the week:
   1. Running tests to compare new version of Simulacrum against larger training cohort
   2. Implementing alternative testing methods: Pearson’s Chi-squared, Likelihood-Ratio; assessing and comparing the results.
   3. Contributing graphs of test results to Cong Chen for Simulcrum Methodology paper
   4. Learning about the Plotly graphing library for Python to improve plotting functionality and interactivity
   5. Organising, Chairing Analysts East Team Meeting, 22nd August 2019
   6. Produce code to apply tests grouped by pairs of variables
   7. Researching Kolmogorov-Smirnov test for continuous data (e.g. date fields, ages)
   8. Evaluate SACT reporting software usage for adaptation to simulated datasets
2. Oracle SQL Developer, Python (sqlalchemy, getpass, pandas, matplotlib, ipywidgets, scipy.stats, plotly), Jupyter Notebook, [Oracle Instant Client](https://oracle.github.io/odpi/doc/installation.html#windows)
3. Script development and documentation in Python (pandas), SQL
4. Web resources and reading materials
   1. <https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html>
   2. <https://en.wikipedia.org/wiki/Box_plot>
   3. <https://en.wikipedia.org/wiki/Outlier>
   4. <https://pandas.pydata.org/pandas-docs/stable/getting_started/comparison/comparison_with_sql.html?highlight=sql>
   5. <https://plot.ly/python/>
   6. <https://www.dataquest.io/blog/pandas-big-data/>

Week 7

1. Goals for the week:
   1. Improve plotting functionality and interactivity using the Plotly graphing library for Python
   2. Produce visualizations of results from testing univariate, bivariate distributions
   3. Identify outlier fields or categories which perform particularly poorly across the board to inform caveats about Simulacrum capabilities
   4. Implement tests for continuous data/large number of categories (Kolmogorov-Smirnov test)
   5. Adapt SACT reporting software (SAS) to simulated SACT datasets
2. Oracle SQL Developer, Python (sqlalchemy, getpass, pandas, matplotlib, ipywidgets, scipy.stats, plotly), Jupyter Notebook, [Oracle Instant Client](https://oracle.github.io/odpi/doc/installation.html#windows)
3. Using the tools:
   1. Plotly: Customizing grouped bar chart outputs to compare performance between Simulacrum versions, producing heatmaps of test results for pairs of categorical fields, plotting cumulative distribution functions for visualizing Kolmogorov-Smirnov test results.
   2. SQL: Introducing derived fields ‘DIAGNOSISMONTHBEST’ and ‘MONTH\_FIRST\_SURGERY’ to analyse time data as discrete variables.
4. Web resources and reading materials
   1. <https://www.jeannicholashould.com/tidy-data-in-python.html>