IE 3301 - 004

ENGINEERING PROBABILITY

REAL-WORLD DATA: ANALYSIS

Part I

November 19, 2017

Joe Cloud

Student ID: 1000921236

University of Texas at Arlington

I, Joe Cloud, did not give or receive any assistance on this project, and the report submitted is wholly my own.

Contents

INTRODUCTION

The aim of this project, as stated in the project brief is to analyze real-world data using techniques covered in the course. For this portion of the project, this includes data collection of two independent data sets: one is a sample of a random variable that is suspected to be normally distributed, while the second is a measure of the time interval between events; then summarizing each dataset statistically.
For my project, I chose to collect a set of resistor values from a batch of 100 1kΩ ± 5% resistors, for my first set. For the second set, I compiled login logs for users from a compute cluster and ran preprocessing scripts. The result was a dataset of the login time (in seconds) interval between user logins.

MOTIVATION

DATA

Set One

Data collection began with the process of measuring the resistance of each resistor and recording the value. The measurements were conducted with a calibrated 4.5 digit digital multimeter (see appendix 4 for equipment information), each measurement was entered into a spreadsheet and later outputed as CSV file for further processing by data analysis scripts.
Before data analysis was performed, 0.33Ω was subtracted from each value, this done to account for the resistance added by the multimeter probes. Although this minor offset does not impact the distribution of measurements for the purposes of the project, it does provide us with truer results. The values are offset by approximately 0.03%.

In an ideal world, the resistors would measure identically to the 1kΩ label. In reality, the cost of highly accurate manufacturing processes is expensive, tolerances guarantee a range of possible [random] values. This dataset was created to explore the distribution of these values, within the specified, yet continuous range.

Set Two

The data collection for Set Two revolved around login access to a compute cluster and may be difficult to produce outside the context I outlined above. Though, it is possible for the reader to perform similar analysis with data from their own machine using their own login log. On a Unix-like operating system this can be performed with the use of the `last` command. One of the benefits of computer-triggered collection is that it is much less susceptable to timing errors as opposed to that of a 'human polling' based approach.
I included the source code necessary to collect the data in appendix REPLACEME.

TESTING

REFERENCES

Weiss, S., 1989. Tissue Destruction by Neutrophils, *New England Journal of Medicine,* 320, pp. 365-376.