# Problem Set 2: Basic Statistics

## Introduction

In this Problem Set, we will practice doing basic statistical analyses and visualizations in R with a health dataset. Download the dataset associated with this assignment, "f10IntercrossMissing.csv" from the Problem Set 2 folder on GitHub.

Crosses of inbred mouse lines are commonly used to study traits associated with disease in humans. The "f10IntercrossMissing.csv" file contains data for 1302 mice from the 10th generation of an advanced intercross between the Large and Small inbred strains of mice. Eight variables are reported related to body size and weight. This is part or a much larger study that includes genetic data, growth over time and collection of other disease related variables such as glucose and insulin response. These variables are important for studies identifying genetic variants related to growth, maternal effects, obesity, environmental interactions and diseases such as diabetes and other metabolic related diseases.

You will submit an R script file with code that replicates your results from beginning to end. Someone else running your R script file should be able to reproduce your results exactly. In addition, you will submit clear answers to each question or instruction from this assignment. It can either be in a separate file or written as comments interspersed in your R script.

## Instructions

Follow the instructions below. Be sure to include the answers to any questions posed in your report. For the purpose of reproducibility, start your R analysis by making a new R script file to which we can save our code. It's recommended that you write commands into your script file and then run the code from your script in R, rather than typing commands into the R console and then copy/pasting them into your script, for the sake of reproducibility. When we evaluate and grade your R script, we should be able to run it and arrive at the same answers as you did, so make sure that you check that your R script is able to do so from a clean environment (i.e. starting from scratch, with no data loaded into your R environment).

Start your script file off by setting your seed to the number 1389 (or any other number) with the following command, so our analyses can be more easily reproduced:

```
set.seed(1389)
```

Then, we'll start, as usual, by loading the libraries that we'll want to use. Looking ahead in this assignment, we'll need at least the packages `e1071`, `MASS`, `car`, `mice` and `mvoutlier`. If you don't have these packages, you'll need to install them.

```
library(e1071)
library(MASS)
library(car)
library(mice)
library(mvoutlier)
```

If you want to use the `tidyverse` package or any other packages, load them here as well. We will not be using `tidyverse` in the following example code, but feel free to if you feel comfortable doing so.

Next, we'll need code that loads in our data. Remember that you can use the function `file.choose()` in place of a file location to pull up an interactive prompt to help you find your files. Here, we demonstrate this assuming that the data file is located in our current working directory—your data file is most likely not located in your current working directory unless you've moved it there or changed your working directory to the folding containing your data file.

```
# replace "f10IntercrossMissing.csv", including the quotes, with file.choose() if desired.
dat <- read.csv("f10IntercrossMissing.csv", header = TRUE)
head(dat) # look at the first few rows to check
```

## Data Quality Control

**First, assess and report the percentage of missing data for each variable.** If you don't remember how to do this, make sure you've seen the lecture material for week 5 and reviewed the accompanying R code. While you're reviewing the R code accompanying the lecture, we recommend that you run it side-by-side with the lecture material to get a better feel for it. You may scavenge bits of code from the lecture code to use in this assignment

**Document and discard any variables with greater than 40% missing data.** Use multivariate imputation using the mice package to impute missing data points for the remaining variables.

For any variable with excess skewness (in this case any value between -1 & 1 is acceptable), **anchor the variable to 1** (see R script from lecture) using the function supplied below. You can paste the code below into R which will make the anchor1 function available for use.

```
#function of shift minimum to 1 (i.e. anchor to 1)
anchor1=function(y){
    y1=y-min(y)+1
    y1
}
```

On the "anchored variables, use the `boxcox` function from the `MASS` package to identify and **report the "optimal" lambda for each variable that needs transformation.**

### Advanced (optional, extra credit)

Use the `bcPower` function from the `car` package to transform the skewed variables with the optimal lambda previously found. **Create a new data.frame with all of the variables after imputation but replacing the variables with skew with their box cox transformations.** Use the uni.plot function (with its default setting) in the mvoutlier package to assess multivariate outliers. Report how many outliers it suggests there are.

**Report the most extreme outlier distance values** (hint: look at the mahalanobis distances in the $md vector of the output) and report which sample it belongs to using the row number as the ID for that sample.

### References

Fawcett GL, Jarvis JP, Roseman CC, Wang B, Wolf JB, Cheverud JM. 2010. Fine-mapping of Obesity-related Quantitative Trait Loci in an F9/10 Advanced Intercross Line. Obesity (Silver Spring) 18.

Cheverud JM, Lawson HA, Fawcett GL, Wang B, Pletscher LS, R Fox A, Maxwell TJ, Ehrich TH, Kenney-Hunt JP, Wolf JB, Semenkovich CF. 2011. Diet-dependent genetic and genomic imprinting effects on obesity in mice. Obesity (Silver Spring) 19: 160–170.

## The report

Submit your code as an R script along with clear answers to the course 2GW site. The answers can be in a separate file (pdf or word) or clearly noted in the script.

## Due date

Day 7, Week 5