1. A study was conducted to investigate the relative fitness costs of drug resistant strains of *Mycobacterium tuberculosis*. In an experimental setting, the researchers grew *M. tuberculosis* in the presence of the antibiotic rifampin until drug resistant strains appeared. Resistant strains were isolated and grown in competition with each strain’s parent to obtain a measure of the fitness cost of the drug resistance conferring mutation. All nine mutations realized in this fashion were Centers for Disease Control (CDC) clinical strains associated with Europe and the Americas. Relative fitness was also measured for four of these nine strains on a different genetic background (different from CDC) referred to as T85, and associated with East Asia strains of *M. tuberculosis*. The goal of this problem will be to investigate the effects of mutation (S531L, H526Y, H526D, S531W, H526R, S522L, Q531L, H526P, R529Q) and background (CDC, T85) on fitness cost of the mutations. The data can be found in the MatLab file `gagneux.mat` in four columns named `strain`, `mutation`, `fitness`, and `background` within the structure `gag` (e.g.: you can extract the mutations by typing `gag.mutation`). The 1st variable contains an identifier for the strain type and will not be used here; the 2nd variable contains a 5-character string identifier for the mutation type; the 3rd variable contains the measure of relative fitness; and the 4th variable contains a 3-character string identifier for the genetic background used.

(a) Perform some exploratory data analysis (EDA) on these data. Specifically, indicate any unusual mutations in terms of relative fitness, and construct a relevant graph comparing the relative fitness among groups. You might compute means and standard deviations for each mutation-background combination as a means of comparison. Present your initial findings in a well-written paragraph with no more than one plot.

(b) Using only the CDC background, run a one-way analysis of variance (ANOVA) to measure the effect of the 9 mutations on relative fitness for CDC. Report your ANOVA table and use this table and earlier EDA to address the following questions in a few complete sentences. [An ANOVA model can be run with the `anovan` function, where once you define vectors for the desired mutations and fitness values, you can type: `[p1,tab1,stats1,terms1] = anovan(fitness,mutation,1);` to run the model. An ANOVA table will automatically be printed and information stored in `tab1`. `p1` contains the factor p-values, `stats1` is a complex structure containing information to perform multiple comparisons among the mean relative fitness values of the 9 mutations, and `terms1` simply has a list of the model factors used. The "1" indicates that only first-order terms are needed in the model.]

i. Is there a significant effect of mutation on relative fitness for CDC strains?

ii. Which mutations are significantly different from one another in terms of their mean relative fitness?

iii. Which mutation is likely driving any differences among the mutations on relative fitness?

[For the 2nd question above, you will want to use the `multcompare` function to run a set of multiple comparisons among the mean relative fitness values of the 9 mutations. After running the ANOVA with the `anovan` function above, type: `[cis,means] = multcompare(stats1);`, where `stats1` was saved from the `anovan` command. This will use Tukey-based pairwise comparisons between mutation means and produces two types of output. First, it stores the confidence intervals (and center) for the difference between each pair of mutations in a matrix called `cis`, and stores the means and standard deviations for each mutation group in a matrix called `means`. Second, it produces a graph with confidence intervals for each mutation mean, and allows you to interactively select a mutation to see from which mutations it differs - very cool!]

(c) Construct a residual plot and normal quantile plot of the residuals resulting from this ANOVA model. The residuals are stored in the vector `stats1.resid`. [See the `meadowfoam.m` code on the course webpage to see how to construct a residual plot and normal quantile plot.] In a few clear sentences, explain clearly what these two plots indicate about the variance homogeneity and normality assumptions. Do they indicate any unusual values?
(d) Since this unusual mutation identified in part (b) was reported as being extremely rare and has never been observed in a clinical setting, it might be reasonable to exclude its effect as it does not seem to occur naturally. Rerun the 1-way ANOVA from part (b) without this unusual mutation, again reporting the ANOVA table, and address the effect that removal of this one mutation has on measuring the overall effect of mutation on relative fitness. Reconstruct a residual plot and normal quantile plot and interpret these in this case with 8 mutations. Use the multcompare function again and indicate between which pairs of mutations there are differences in mean relative fitness.

(e) Using only the four strains (S531L, H526Y, H526D, S531W) for which relative fitness was measured in both backgrounds (T85, CDC), run a two-way analysis of variance (ANOVA) to measure the effect of mutation and background on relative fitness. Include an interaction term in your model. Based on this ANOVA table, address the following questions with supported p-values in a well-written paragraph.

i. Is there an effect of mutation on relative fitness?
ii. Is there an effect of background on relative fitness?
iii. Does the effect of mutation on relative fitness depend on the genetic background?

(f) Construct an interaction plot to explore the nature of the significant mutation x background interaction present in the 2-way ANOVA model in the previous part. Describe this interaction in a few clear sentences. [In MatLab, you can construct an interaction plot using the interactionplot function. To obtain an interaction plot of the relative fitness means grouped both by mutation type and background, you can type: interactionplot(fitness,mutation,background)];

(g) Finally, download the paper from which these data originate [Gagneux et al., Science, 312, 30 June 2006], read the paper, and write a short synopsis of what the researchers found. Do their results agree with your analysis on their data? How does it differ?

(h) The question of whether the Gagneux analysis or that performed in this homework is correct boils down to whether their experimental design is a crossed design or a nested design. Read the attached pages on crossed and nested designs and argue for one type of analysis or the other. [It should be mentioned that Joran Elias (a former MEID student who completed his PhD in statistics) and I wrote a response to Science pointing out analysis errors in this paper, and although the errors were acknowledged (sort of) by the editors, they decided not to take any action to print a correction. I suppose I know where data analysis stands with respect to “science!”]

2. Complete the plot of beneficial calculation. The problem was as follows. Massive cancellations in subtraction do not always lead to catastrophic cancellation. Cancellation error does not always lead to inaccuracy. Straightforward implementation of

\[ f(x) = \frac{(x - 1)}{(e^{(x-1)} - 1)} \]  

is not accurate if x is close to 1. Surprisingly, the solution below achieves full working accuracy regardless of the size of x. Final computed answer more accurate than intermediate results through extreme cleverness in helpful cancellation! Error analysis can be very subtle and non-obvious.

function [z] = cancelCleverSequential(x)
    y = x - 1.0;
    z = exp(y);
    if z == 1.0
        return;
    else
        z = log(z) / (z - 1.0);
    end
• Write the function cancelClever(x) which takes a "vector" of x as input, and returns a "vector" of z.

• Write a function that naively computes \( f(x) = (x - 1)/(e^{(x-1)} - 1) \). This function should also be vectorized.

• Write a script to plot both functions on the same axes for \( 1 - 1 \times 10^{-7} \leq x \leq 1 + 1 \times 10^{-7} \). Label axes, create a legend, and generally make the plot suitable for submission as a homework.

3. Assemble the various components of a gauss elimination scheme presented in the notes into a function for solving a system of linear equation. Complete the partial pivoting scheme. Run the function against the test data in the slides and verify that the solution is correct. Turn in the run showing that it works, as well as a source code listing.

4. Verify or dispute the value of the \( \lambda \) appearing in the Kareiva (2000) paper. Provide plots of each age category of the salmon population as a function of the next 25 years under various assumptions about the survival of the salmon during in-river migration (\( s_z \)). Do you agree with the papers conclusions?