DSCB 225
Winter, 2026
Problem Set 1: Choosing and Applying Statistical Tests in R

Overview

In this problem set, you will analyze a biological dataset using R. The goal is to demonstrate your ability to select appropriate statistical tests, implement them correctly in R, and interpret the results in a biological context.

All work must be completed in an R Notebook saved in the Code/ directory. Code that specifies file input and output must use relative file paths. Relative file paths describe how to get to the desired location from the current working directory (which is the directory that contains the R Notebook). For example, if the R Notebook is in the Code folder, a relative file path to the Data folder would be "../Data/" (the ../ means 'Go up one level').

Directory Structure

After cloning this repository from GitHub Classroom, you should see the following structure:

```
ProblemSet1/
|— Code/
|— Data/
   — differentiation_growth_data.csv
|— Results/
```

Dataset Description

The dataset contains simulated results of measurements from an experiment, performed in triplicate, on the effect that a drug has on cells with a GFP reporter that are cultured in a multiwell dish. The measurements are:

- Percent_Positive: The percent of cells in the well that are positive for a marker, NKX2.1
- Growth_Rate: The growth rate of the cells over a 24 hour period
- Colony_Morphology: A qualitative assessment colony morphology (monolayered or multilayered)
- Fluorescence_Intensity: A measurement of the total GFP fluorescence per well (one measurement per well)

Instructions

Code chunk 1: Setup Notebook and Load Libraries and Data (2 points)

- Create a new R Notebook called ProblemSet1_LastName.Rmd
  - Set up the first four lines like this:

```
---
title: "[First Name] [Last Name], Problem Set 1"
output: html_notebook
---
```

- Save the notebook in the Code/ directory
- Create a new chunk and add code to load the tidyverse and multcomp libraries and the data in 'differentiation_growth_data.csv' file into an object called 'CellDiff' (Don't forget to use a relative file path!)

- Display the first few rows of the data below the chunk

Code chunk 2 (3 points)

- Calculate the mean, median, and standard deviation of the values in the Percent_Positive column
- Use pipes (%>%), group_by(), and summarize() to compute means and standard deviations by condition and store the results in a new dataframe called 'CellDiff_summary'. The column headings of this new dataframe should be: 'Condition' 'Mean' and 'StdDev'
- Save the summary data as a .csv file in the Results folder (Don't forget to use a relative file path!)

Code chunk 3 (3 points)

- Use two methods to determine whether the data on the percent of NKX2.1$^+$ cells in the Control condition are normally distributed. (Hint: It's easier if you subset the CellDiff dataframe first)
- Repeat these methods in the same chunk to determine whether the percent of NKX2.1$^+$ cells in the Drug_10uM condition are normally distributed.
- In the space below the chunk, state your conclusion from this test and a brief justification

Code chunk 4 (3 points)

Subset the data so that it includes only the Control and Drug_10uM conditions and write code to answer the following questions. For each question, state the null hypothesis, alternative hypothesis, your conclusion from the test and a brief justification of this conclusion.

- Is the percent of NKX2.1$^+$ cells significantly different between these two conditions?
- Does drug treatment at 10 uM change the distribution of colony morphologies compared to control?

Code chunk 5 (3 points)

Write code that uses the full dataset to answer the following question:

- Is the percent of NKX2.1$^+$ cells significantly different between the Control condition and each of the drug treatment conditions?

Code chunk 6 (3 points)

Use the whole dataset for this chunk

- Generate a linear model of the growth rate as a function of the percent NKX2.1$^+$ values
- Generate a graph with the linear model superimposed on a scatterplot showing the relationship between the percent NKX2.1$^+$ and the growth rate. (Not required, but if you want to be fancy, you can use col = [put Conditions column reference here] to color the markers by condition. This is even prettier in ggplot!)
- Test whether there is a correlation between the NKX2.1$^+$ values and the growth rate. State your answers sto the following questions and provide a brief justification:
  - Is a linear model appropriate for this correlation?
  - Is there a strong correlation between these values?

Code chunk 7 (3 points)

Use only the data from the Control and Drug_10uM conditions for this chunk

- Generate a histogram of the GFP fluorescence intensity measurements. State whether you think they are normally distributed and write code to further test this conclusion.
- What would be the appropriate test to use to determine if there is a significant difference between the GFP fluorescence intensity measurements from the Control and Drug_10uM conditions?

Extra Credit (3 points)

Redo Code chunk 6 using data from individual conditions rather than the whole dataset. Are the outcomes different? Why or why not? Provide a justification for this conclusion from both a statistical perspective and a biological perspective.

Output and Submission

All code must run without errors. Any saved output must be written to the Results/ directory using relative paths. Push your completed R Notebook to GitHub Classroom before the deadline.