# README

## About CountSMART

Longitudinal count data are often collected in a variety of health domains. This repository contains code to estimate sample size needed to compare dynamic treatment regimens using longitudinal count outcomes from a Sequential Multiple Assignment Randomized Trial (SMART). A particular focus of this repository is on longitudinal count data having overdispersion.

A pair of dynamic treatment regimens embedded in a planned SMART (aka. 'EDTRs') can be compared using differences in end-of-study means, or more generally, differences in a weighted average of means across various time points, which we denote as $\Delta_Q$; $Q$ is simply shorthand for 'quantity', e.g.,$\Delta_{EOS}$ denotes the quantity *difference in end-of-study means*.

CountSMART is about a Monte Carlo simulation-based approach developed to estimate sample size required to attain power of $1-\eta$ to the test of the null $H_0 : \Delta_Q = 0$ against the alternative $H_a : \Delta_Q \neq 0$ at type-I error rate $\alpha$.

## About this repository

This repository contains code implementing CountSMART methodology and simulation studies examining the validity of the approach.

### 1. Setting up this repository

#### 1.1 Packages used in the project

1. The collection of packages and their version numbers used for this repository are recorded in the renv.lock file. The package, renv, can facilitate installation of these packages in the machine of end-users of this repository. See renv package documentation here for more details: https://rstudio.github.io/renv/articles/renv.html

#### 1.2 Tell R where to pull code from from and where to push data to

1. Create a new R file named 'paths.R' and save this file within the root directory of the repository (usually where the .Rproj file is located).
2. Within 'paths.R', set the value of the following variables below by replacing the three dots '. . . ' with the appropriate directory.

- path.output_data = ". . . /output"
- path.code = ". . . /code"
- path.plots = ". . . /plots"

Note that 'paths.R' is included in the '.gitignore' file, preventing any user-specific directories from being displayed in the repository. Also, since 'paths.R' is included in the '.gitignore' file, a new 'paths.R' file would need to be created by each end-user of the repository.

## 2. The `code` folder

### 2.1 Collection of functions for input-checking, simulation, and data analysis

| File Name | Brief Description |
|---|---|
| input-utils.R | Contains a function for checking validity of time-specific means and proportion of zeros provided as inputs to the sample size estimation procedure. |
| datagen-utils.R | Collection of functions to generate potential outcomes and observed outcomes. |
| analysis-utils.R | Collection of functions to 'analyze' data from a SMART. |

### 2.2 Collection of functions for executing calculations

| File Name | Brief Description |
|---|---|
| calc-covmat.R | Calculate estimated covariance matrix. |
| calc-corr-params-curve.R | Implement simulation to estimate relationship between $\rho$ and $\tau_{MAX}$ and the relationship between $\rho$ and $\tau_{MIN}$. |
| calc-truth-beta.R | Calculate true value of parameters in a model for the mean trajectory of dynamic treatment regimens embedded in a SMART, implied by inputs provided to Monte Carlo simulation. |
| calc-truth-contrasts.R | Calculate true value of $\Delta_Q$ in a model for the mean trajectory of dynamic treatment regimens embedded in a SMART, implied by inputs provided to Monte Carlo simulation. |
| plot-truth-deltaQ.R | Wrapper for calc-truth-beta.R and calc-truth-contrasts.R. Visualize true mean trajectory of each dynamic treatment regimen embedded in a SMART, implied by inputs provided to Monte Carlo simulation. |
| geemMod.R | Modification of the `geem.R` script from the R package `geeM`: setting the additional argument `fullmat=TRUE` allows custom specification of working correlation matrix for each participant-time. |

## 3. The `output` folder

**Results using an autoregressive structure**

| File Name | Brief Description |
|---|---|
| create-scenarios-ar.R | A script to create simulation study scenarios. |
| calculate-dispersion-param.R | A script to calculate the value of the negative binomial dispersion parameter in the different simulation scenarios. |
| simulation-study-pipeline-ar.R | A script to document and run steps in the simulation study pipeline. |
| sim_size_test | A directory containing a collection of scripts to execute simulation studies concerning empirical type-I error rate. Results of simulation studies are also provided here (e.g., `power.csv` file). |
| sim_vary_effect | A directory containing a collection of scripts to execute simulation studies investigating how power changes as specific choices of $\Delta_Q$ are increased across a grid of total sample sizes N=100, 150, 200, ..., 550. Results of simulation studies are also provided here (e.g., `power.csv` file). |

| File Name | Brief Description |
|---|---|
| sim_vary_n4 | A directory containing a collection of scripts to execute simulation studies investigating whether power is sensitive to a violation in our working assumption on the number of individuals who would not respond to either first-stage intervention option. Results of simulation studies are also provided here (e.g., `power.csv` file). |
| sim_vary_eta | A directory containing a collection of scripts to execute simulation studies investigating whether power is sensitive to the actual value of $\eta$, given fixed value of $\rho$ and N. Results of simulation studies are also provided here (e.g., `power.csv` file). |

**Results using an exchangeable structure**

| File Name | Brief Description |
|---|---|
| create-scenarios-exch.R | A script to create simulation study scenarios. |
| calculate-dispersion-param.R | A script to calculate the value of the negative binomial dispersion parameter in the different simulation scenarios. |
| simulation-study-pipeline-exch.R | A script to document and run steps in the simulation study pipeline. |
| sim_vary_effect | A directory containing a collection of scripts to execute simulation studies investigating how power changes as specific choices of $\Delta_Q$ are increased across a grid of total sample sizes N=100, 150, 200, ..., 550. Results of simulation studies are also provided here (e.g., `power.csv` file). |

## 4. The `plots` folder

**Plot results using an autoregressive structure**

| File Name | Brief Description |
|---|---|
| data-viz-pipeline-ar.R | A script to document and run steps in the data visualization pipeline. |
| plot-sim-size-test.R | A script to plot results in sim_size_test |
| plot-sim-vary-effect.R | A script to plot results in sim_vary_effect |
| plot-sim-vary-n4.R | A script to plot results in sim_vary_n4 |
| plot-sim-vary-eta.R | A script to plot results in sim_vary_eta |
| corviz_sim_vary_effect | A directory containing visualization of empirical correlation matrices corresponding to each scenario considered. These results accompany those in the directory sim_vary_effect. Values of parameters identical to those used to obtain results in the directory sim_vary_effect were used to calculate the values displayed, except that N was fixed to 1000. |
| corviz_sim_vary_eta | A directory containing visualization of empirical correlation matrices corresponding to each scenario considered. Values of parameters identical to those used to obtain results in the directory sim_vary_eta were used to calculate the values displayed, except that N was fixed to 1000. |

**Plot results using an exchangeable structure**

| File Name | Brief Description |
|---|---|
| data-viz-pipeline-exch.R | A script to document and run steps in the data visualization pipeline. |
| plot-sim-vary-effect.R | A script to plot results in sim_vary_effect |