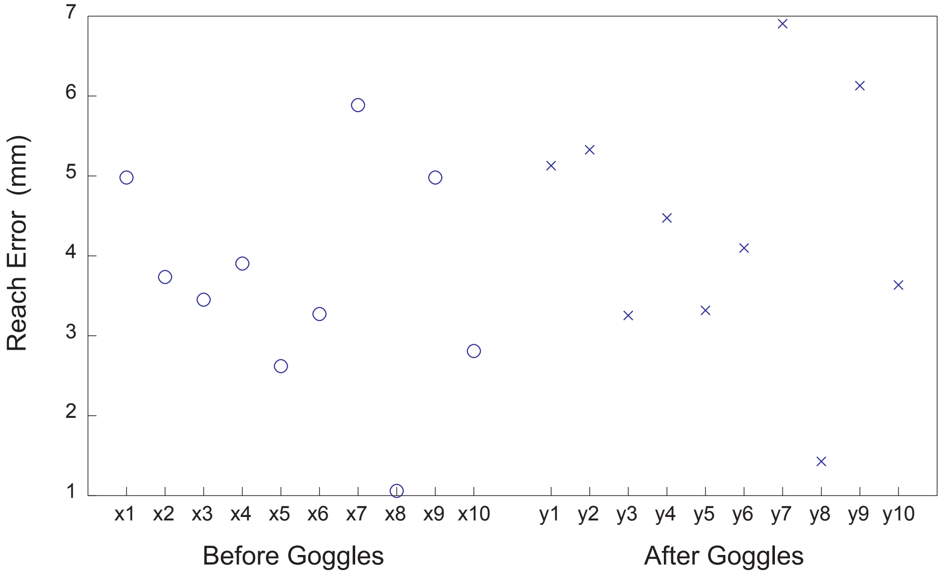
Problem Set: Resampling Methods for Statistical Estimation and Hypothesis Testing



The following questions all pertain to the dataset shown above. The variables x and y represent errors made by a particular subject when reaching with his hand to visual targets. The x variables are for reaches made before the subject wore prism goggles, and the y variables are for reaches made after the subject wore the goggles. The dataset consists of N=10 reaches with each hand, $X = \{x_i\}_{i=1}^{10}$ and $Y = \{y_i\}_{i=1}^{10}$. The data should be obtained using the supplied function GetSample.m:

```
[X,Y] = GetSample;
```

Although GetSample can take arguments, for now you should use the form above (no input arguments).

We are interested in whether the prism goggles affected reaching performance, in other words whether the means of the two sample groups, μ_x and μ_y, are the same or different.

1.a. Consider the statistic $d = \bar{x} - \bar{y}$, which is a measure of adaptation in this example. As described in class, use the Bootstrap to estimate the standard deviation of d and its 95% Confidence Interval. Use the "percentile interval", as defined in the notes.

Is there evidence of adaptation, i.e. do the group means look significantly different?

1.b. Now you are told that the data are actually paired. This means that x_i and y_i were sampled together (e.g. at the same time, or from the same animal, etc). In this case, it means that there were 10 different reach targets, and x_i and y_i were both made to the ith target.

Design an appropriate Bootstrap resampling scheme for this paired-sample case. The idea is to resample with replacement from the dataset while preserving the fact that the data were collected in a paired fashion. Use this scheme to compute the standard deviation of d and the 95% Confidence Interval, as above.

Does the extra knowledge that the data were paired change your interpretation?

2.a. Using the method of Section 2.1 in the Notes, design and implement a Permutation Test to determine whether X and Y have different means. For now, ignore the fact that the data are paired.

Is there evidence of significant adaptation?

2.b. Now devise and implement a Permutation Test that takes into account the fact that the data are paired. One again, the key is to randomize the data in a fashion that “breaks” the difference between X and Y (the whole idea of two-sample Permutation Tests) yet that preserves the fact that each x_i and y_i are paired.

Does the extra knowledge that the data were paired change the results of your test?

3. *Power Analysis.* Here we will compare the statistical power of the two permutation tests (unpaired and paired) when applied to paired datasets.

The power of a hypothesis test is defined as

$$\beta = P(\text{rejecting } H_0|H_1),$$

the probability of rejecting the Null Hypothesis given that alternative really is true. If one has access to many (> 100) datasets drawn from the same distribution, the power of a test can be determined simply by applying the test to each dataset and counting the percent for which H_0 was rejected.

Of course β will depend on many factors including: the confidence level α, the effect size, the sample size, the statistic you are testing, the shape of the data distribution, etc. Here we will focus on two factors, the effect size and the sample size.

New datasets drawn from the same distribution as $\{X, Y\}$ above can be obtained with the command,

```
[X,Y] = GetSample(N,D);
```

Here N is the sample size (i.e. the number of $\{x_i, y_i\}$ pairs), and D is the effect size (i.e. the “true” difference in error before and after prisms).

Conduct a power analysis for both the unpaired and paired Permutation Tests from Question 2 using a range of sample sizes and effect sizes. The following values worked well for me:

$$\begin{aligned} N &\in [10, 20] \\ D &\in [0.12, 0.25, 0.5, 1, 2]. \end{aligned}$$

Using 200 datasets for each of the 10 (N, D) pairs and a (perhaps slightly small) value of R=500 permutation resamplings for each test, the entire analysis took 420 seconds to run on my old laptop. Use the value $\alpha = .05$, i.e. reject H_0 at the 95% confidence level.

For each test (unpaired, paired), make a plot of the power β as a function of D, with separate lines for each value of N. Which test is more powerful? How much smaller of an effect size can the more powerful test detect?

Comments on the datasets: If you are curious about the datasets, type `help GetSample` or `type GetSample` in Matlab. In practice, one does not typically have a source of infinite data at hand (if we did, we wouldn't need statistics). So how can you run a power analysis in practice? In a case like this, you might have a good idea of the variance in the data (from pilot studies). You can then create simulated datasets by adding known effect sizes D to simulated noise, just as is done in GetSample.m. In other words, if we had a rough estimate of Sx and Sd in this example (see the Matlab file), then we could perform the exact analysis done here. This approach is very useful for planning experiments and for grant applications!