Module 5F: Unsupervised Learning

A smattering of options: PCA, permutations, bootstrap

Randomization/Permutation (a resampling method):

- Asks: **are two variables independent?**
- **Assumptions:** random sampling, distribution of variables have approximately same shape

- Versatile
  - Variables can be any combination of numerical or categorical
  - We don't need a null hypothesis *because we build it ourselves*. A randomization test generates a ***null distribution*** for the association between two variables.
  - **MWU test is a type of permutation tests** – but you lose power when you use ranks instead of the actual data

- Basis: **Permutation**
  - Sampling without replacement
  - Method:
    1. Create data set
       - Response variable of a test statistic measuring association **randomly assigned to Explanatory variable**
         - **You are effectively exchanging labels**
       - **All data points are used exactly once**
    2. Calculate measure of association for randomized sample
    3. Repeat randomization many times
       - A NULL distribution

Pretty much gives you a p-value and not much else!

Randomization example:

The following is a very small data set of birth weights (in kg) of either singleton or individuals who were born with a twin. Create a legitimate randomized data set:

Singleton: 3.5, 2.7, 2.6, 4.4

Twin: 3.4, 4.2, 1.7