Nonparametric Tests

So far we made an assumption about a distribution and then tested what the parameters of that distribution could be

Instead, non-parametric tests are tests for the entire distribution
- i.e. whether X follows a normal distribution or not

Kolmogorov-Smirnov Test

- Suppose X_i are iid with cdf F
    - We want to test $H_0 : F = F_0$ or $H_1 : F \neq F_0$
- We introduce the empirical cdf:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq n)$$

    - $\hat{F}$ is an unbiased estimator of F
    - $nF(t) \sim \text{Binom}(n, F(t))$
        - By the CLT, this tells us that $\hat{F}$ converges to F asymptotically
- Intuition: reject the null if $|\hat{F}_n(t) - F_0(t)|$ is large for any t
- It turns out that for:

$$T = \sup_t \left| \hat{F}_n(t) - F_0(t) \right|$$

    - T is distributed according to the KS distribution
    - This is not dependent on F_0
    - It does depend on n

Kolmogorov Lilliefors Test

- Suppose we wanted to specifically test whether F is Gaussian or not
- We could create a cdf F_0 that is Gaussian and matches the mean and variance of the sample
- This would be ok, but T in this case would not actually have a KS distribution because we matched up the first two moments
- Instead, it follows a KL distribution and we could use those instead
    - Note: the KL rejection intervals are always smaller than the KS
    - So this means that the KS will retain the null in more cases than KL
    - So using the KS is ok but will give worse results

Permutation Test

- This is a 2-sample test where we have a set of samples X and Y and we want to test if $F_X = F_Y$ or not
    - I.e. test a simulation of the world with and without the Higgs boson and use that to determine if the Higgs boson exists or not
- If we computed the statistic $T = \sup_t \left| \hat{F}_X(t) - \hat{F}_Y(t) \right|$:
    - This would actually depend on the underlying distribution F and not be possible to calculate in most cases
- If we computed the statistic $\left|\overline{X} - \overline{Y}\right|$
    - If the means of the distribution are the same, but are fundamentally different, then we would be committing a large type II error
    - We also still don't know F

- The permutation test is based on the second idea and resolves the second issue, but **not** the first issue
  - We will actually compute $\left|\overline{X} - \overline{Y}\right|$
  - We will then attempt to compute the quantiles of this
  - This still fails if they happen to have the same mean
- Under the null hypothesis, all of the data is the same and it shouldn't matter if we shuffle them around
- Under the null hypothesis, we were equally likely to have seen any of the arrangements of the data
- There are a total of $2n!$ ways to rearrange the data, and then we could just split the samples into left and right halves and take the difference in sample means
- Therefore, if we just compute all possible permutations, we have $2n!$ samples for what the difference between left and right halves could be
  - We could compute the α th percentile for this to get the rejection region
  - In reality, we can't compute $2n!$ permutations, so we just do a random sample