# Introduction to NonParametric Statistics

*Derek Sonderegger*

*7/19/2018*

www.dereksonderegger.github.io/talks

### Linear Models in general

- Have several model assumptions
  - Independence of error terms
  - Constant variance of error terms
  - Normally distributed error terms
- "All models are wrong, but some are useful" - George Box
- Model assumptions must be *approximately* met for the results to be useful.
- Often use transformations on y-variable to address variance and normality violations.
  - $\log(Y)$
  - $\sqrt{(}Y)$
  - *rank transformation*
  - *sign transformation*

### Rank Transformation

- Sort all of the data, smallest-to-largest, and call the order number the *rank*.
- Smallest value has rank 1, second smallest has rank 2, etc, until the largest value has rank $n$.
- If there are ties, give an average rank.

### Wilcoxen Rank Sum

- Let
  - $n_i$ be the number of observations in group $i$
  - $R_{ij}$ the the rank of the $j$th observation in group $i$
  - $n = \sum n_i$
- For each group, calculate the sum of the ranks

$$R_i = \sum_{j=1}^{n_i} R_{ij}$$

- Note, for two groups: $R_1 + R_2 = n(n+1)/2$
- Under the null hypotheses,

$$R_1 \approx R_2 \approx \frac{n(n+1)}{4}$$

- Let $W = R_1 - R_2$

### Simulating the Sampling Distribution under $H_0$

```
SamplingDist <- mosaic::do(10000) *
  Tents %>%
    mutate( Rank.Sim = mosaic::shuffle(Rank) ) %>%
    group_by(Type) %>%
    summarise( R = sum( Rank.Sim ) ) %>%
    summarise( W=diff(R) )
```

### Conclusions

- Nonpparametric tests are more widely applicable than a standard t-test.
- By using rank or sign transformed responses, we cannot make confidence intervals on the scale that is scientifically useful.
- If the usual requirements are met for a t-test, then the standard approach is more powerful and should preferentially be used.