Quantifying Uncertainty

1. Standard Error. Fill in the math.
2. **Sharp Null Sampling Distribution** (distribution of estimates that we would receive by chance if there were no effect) and **Randomization Inference** (permutation test; proportion of these simulations that are more-extreme than the treatment effect observed in the experiment). Groups may differ by chance, even if the treatment has no effect.
   1. Estimating Treatment Effects. How large of an “effect” estimate would we reach by chance?
      1. Distribution of estimates one would reach if treatment had no effect. How likely is this estimate to have just arisen by chance?
      2. **Sharp Null Hypothesis**: for every unit, there is no effect. It is testing a hypothesis about our sample (not a population) in our experiment. Re-randomizing within our sample. How often did I get a randomization under the sharp null where the estimate was larger than my actual estimate?
         1. This is a sampling distribution (distribution of all the possible randomization combinations of our sample).
         2. p-values (if the treatment had no effect, how likely is it that the data would generate a difference this extreme just by chance) and hypothesis tests. Convention is to reject if less than 0.05.
            1. When there’s a really big treatment effect, the probability of observing that treatment effect, even if the sharp null hypothesis were true, is really unlikely.
      3. Randomization Inference will be symmetric. Treatment and control will be symmetric.
   2. **Power**: detecting non-zero treatment effects; the probability that a particular experiment design/measurement/test, will reject the null hypothesis in a world where it *should* reject the null hypothesis.
      1. Power increases with size of the effect (larger effects are easier to detect), square root of the sample size (), precision of the measurement, reduction of variance within groups
      2. Concentrate the tests. Decrease the sample size and give a higher “dosage” to the treatment group. Increases statistical power by exposing a smaller number of people to a larger dose of treatment.
      3. Power decreases with larger amounts of variation in the measured outcomes and standard deviation of the outcome
         1. More diverse populations create more differences in baseline differences; relative to the effect size, this “mutes” the ability to measure an effect
         2. More “noise” in the measurement raises the “floor” of what one must detect to look different from that noise; precise measurements are preferred to imprecise measurements
         3. Ratio of the true treatment effect to the standard error of the estimated effect:
   3. Confidence Intervals