A Power Calculation

- Suppose, for illustation, that we are interested in testing the hypothesis

$$H_0\text{: } \theta_1 = \theta_2 \text{ vs. } H_A\text{: } \theta_1 \neq \theta_2$$

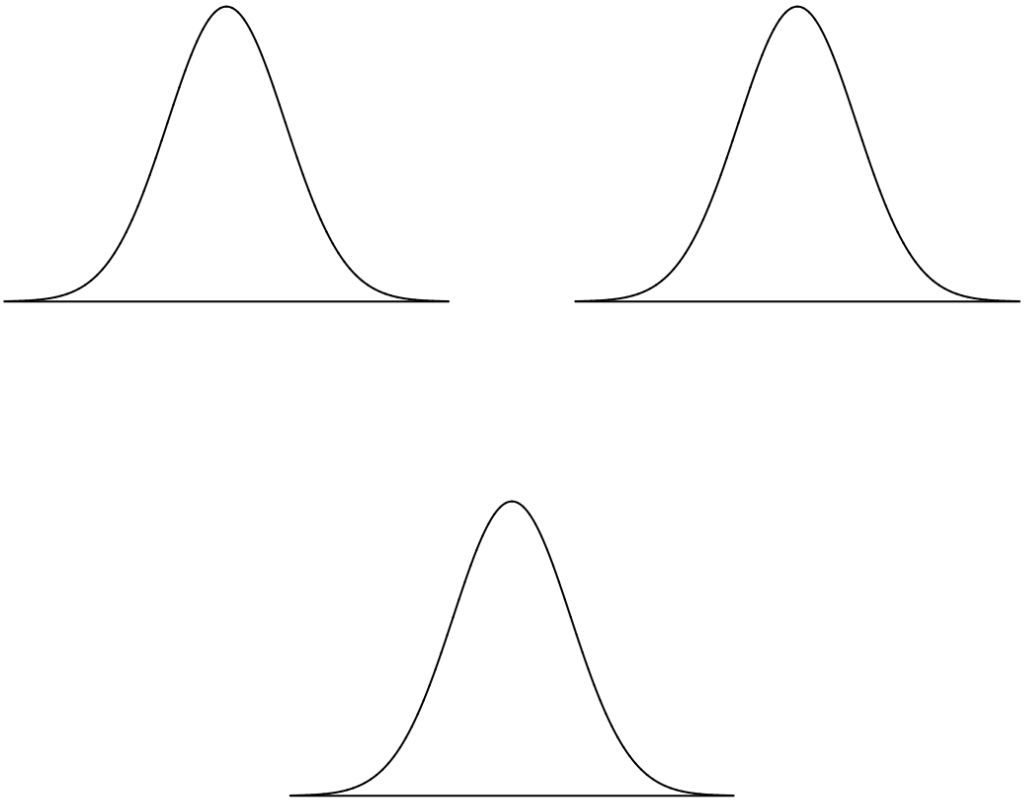
- Suppose, also for illustration, that the test statistic associated with this test has the form

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\text{Var}[Y_1]}{n} + \frac{\text{Var}[Y_2]}{n}}} \sim \text{N}(0, 1)$$

- It will be useful to define the notion of a **rejection region** $\mathcal{R}$: all values of the observed test statistic t that would lead to the rejection of H_0:

$$\mathcal{R} = \{t \mid H_0 \text{ is rejected}\}$$

  - If $t \in \mathcal{R}$, we reject H_0
  - If $t \in \mathcal{R}^c$, we do not reject H_0



- Defining Type I and Type II error rates in terms of a rejection region is also useful:

$$\begin{aligned} \alpha = \Pr(\text{Type I Error}) &= \Pr(\text{Reject } H_0 \mid H_0 \text{ is true}) \\ &= \Pr(T \in \mathcal{R} \mid H_0 \text{ is true}) \end{aligned}$$

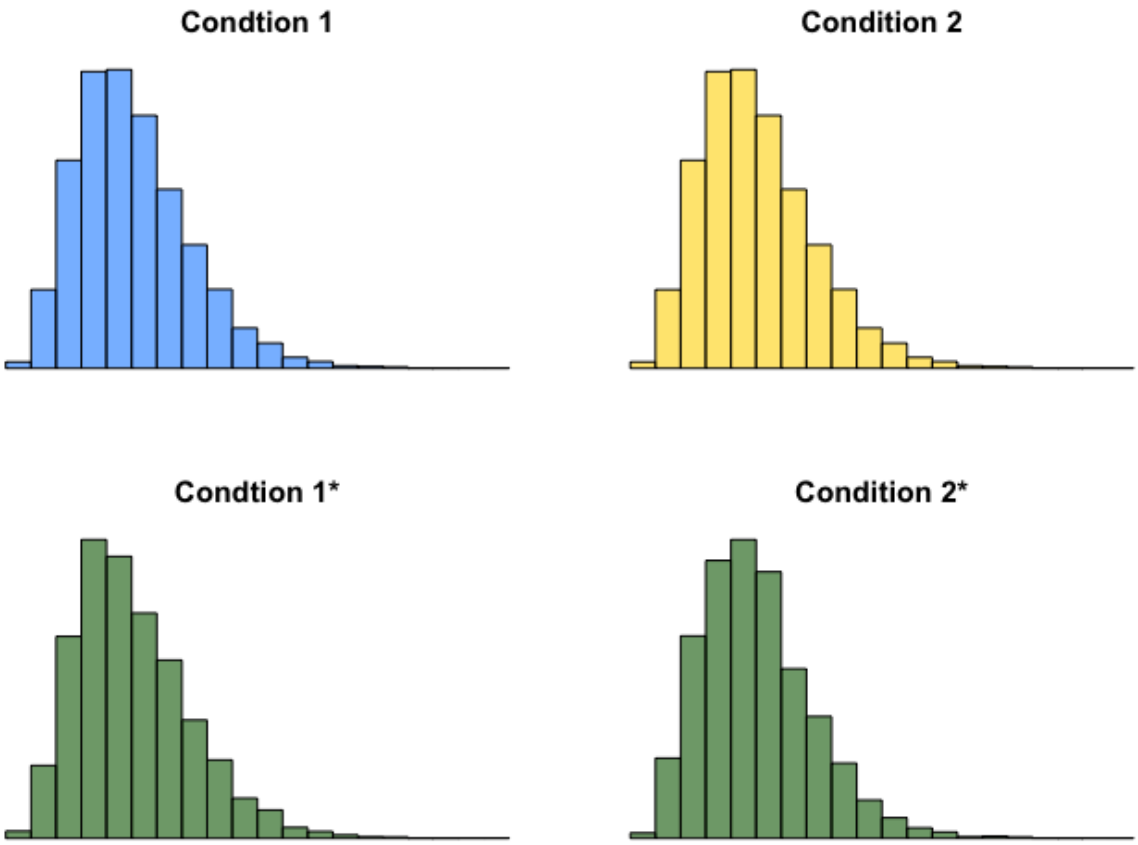
$$\begin{aligned} \beta = \Pr(\text{Type II Error}) &= \Pr(\text{Do Not Reject } H_0 \mid H_0 \text{ is false}) \\ &= \Pr(T \in \mathcal{R}^c \mid H_0 \text{ is false}) \end{aligned}$$

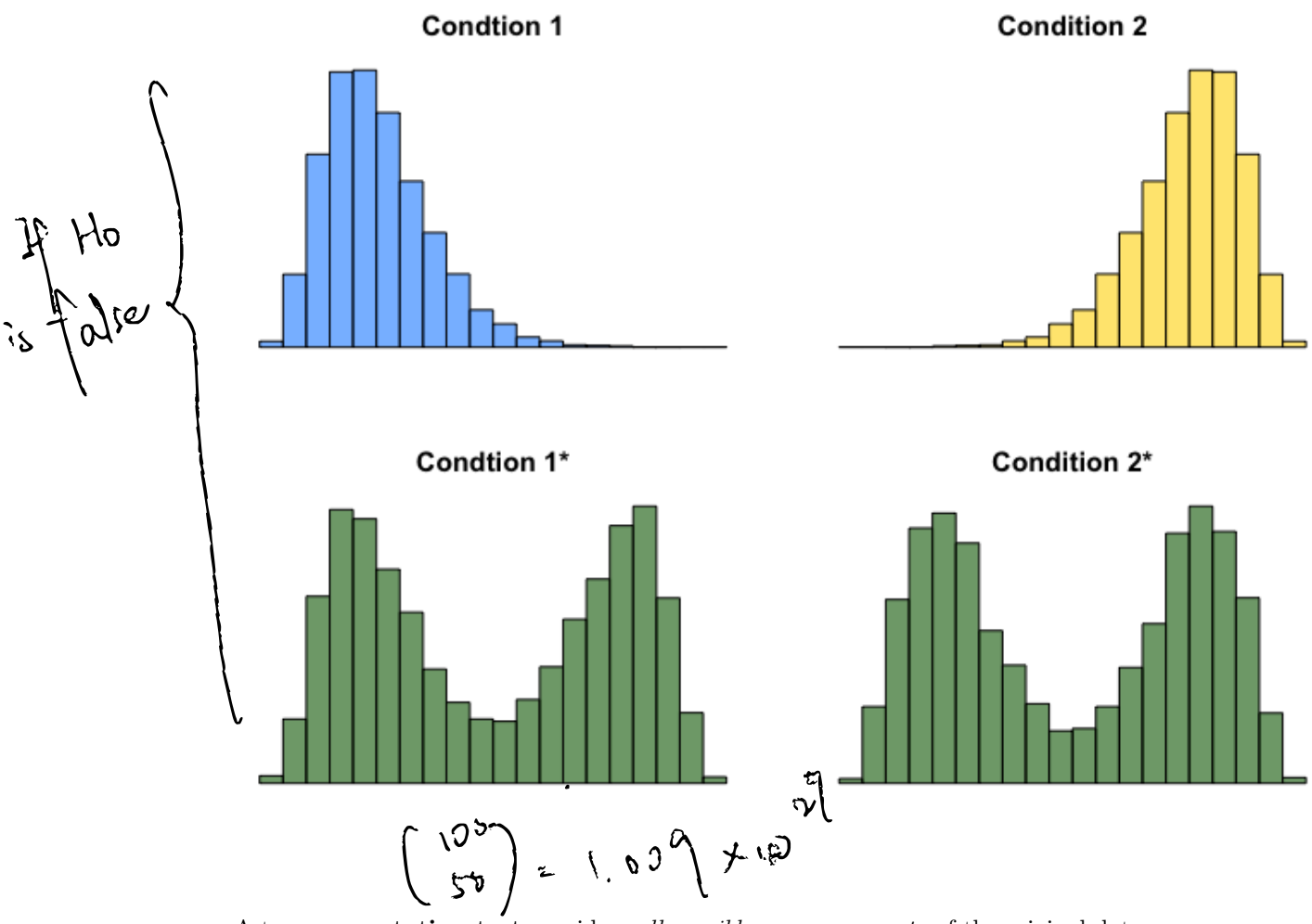
Permutation and Randomization Tests

- All of the previous tests have made some kind of distributional assumption for the response measurements
- It would be preferable to have a test that does not rely on *any* assumptions (about what kind of distribution)
- This is precisely the purpose of permutation and randomization tests.
  - These tests are *nonparametric* and rely on resampling.
  - The motivation is that if $H_0 : \theta_1 = \theta_2$ is true, any random rearrangement of the data is *equally likely to have been observed*. (i.e. 2 samples are from the same population)
  - With n_1 and n_2 units in each condition, there are

$$\binom{n_1 + n_2}{n_1} = \binom{n_1 + n_2}{n_2}$$

arrangements of the $n_1 + n_2$ observations into two groups of size n_1 and n_2 respectively





- A true **permutation test** considers *all possible rearrangements* of the original data
  - The test statistic t is calculated on the original data and on every one of its rearrangements
  - This collection of test statistic values generate the empirical null distribution

- A **randomization test** is carried out similarly, except that we do not consider all possible rearrangements
  - We just consider a large number N of them

Randomization Test Algorithm

1. Collect response observations in each condition.

2. Calculate the test statistic t on the original data.

3. Pool all of the observations together and randomly sample (without replacement) n_1 observations which will be assigned to "Condition 1" and the remaining n_2 observations are assigned to "Condition 2". Repeat this N times.

preserve n_1 and n_2

4. Calculate the test statistic $t_k^\star$ on each of the "shuffled" datasets, $k = 1, 2, \ldots, N$.

k indexing number of resamples

5. Compare t to $\{t_1^\star, t_2^\star, \ldots, t_N^\star\}$, the empirical null distribution and calculate the p-value:

$$\text{p-value} = \frac{\#\text{ of } t^\star\text{'s that are } \textit{at least as extreme as } t}{N} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{I}[t_k^* \text{ at least as extreme as } t]$$

Example: Pokemon Go

- Suppose that Niantic is experimenting with two different promotions within Pokémon Go:
  - Condition 1: Give users nothing
  - Condition 2: Give users 200 free Pokécoins
  - Condition 3: Give users a 50% discount on Shop purchases
- In a small pilot experiment $n_1 = n_2 = n_3 = 100$ users are randomized to each condition
- For each user, the amount of real money (in USD) they spend in the 30 days following the experiment is recorded
- The data summaries are:
  - $\overline{y}_1 = \$10.74$, $Q_1(0.5) = \$9$
  - $\overline{y}_2 = \$9.53$, $Q_2(0.5) = \$8$
  - $\overline{y}_3 = \$13.41$, $Q_3(0.5) = \$10$

3 Experiments with More than Two Conditions

3.1 Anatomy of an A/B/m Test OR A/B/C/D ... testing

- We now consider the design and analysis of an experiment consisting of more than two experimental conditions – or what many data scientists broadly refer to as "A/B/m Testing".

  - Canonical A/B/m test:

Figure 1: Button-Colour Experiment

- Other, more tangible, examples:

  - Netflix
  - Etsy

- Typically the goal of such an experiment is to decide which condition is optimal with respect to some metric of interest θ. This could be a

  - mean
  - proportion
  - variance
  - quantile
  - technically any statistic that can be calculated from sample data

- From a design standpoint, such an experiment is *very* similar to a two-condition experiment

  1. Choose a metric of interest θ which addresses the question you are trying to answer
  2. Determine the response variable y that must be measured on each unit in order to estimate $\widehat{\theta}$
  3. Choose the design factor x and the m levels you will experiment with.
  4. Choose $n_1, n_2, \ldots, n_m$ and assign units to conditions at random
  5. Collect the data and estimate the metric of interest in each condition:

$$\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_m$$

- Determining which condition is optimal typically involves a series of pairwise comparisons

g-test, randomization tests

- But it is useful to begin such an investigation with a *gatekeeper* test which serves to determine whether there is *any* difference between the m experimental conditions. Formally, such a question is phrased as the following statistical hypothesis.

$$H_0\text{: } \theta_1 = \theta_2 = \cdots = \theta_m \text{ versus } H_A\text{: } \theta_j \neq \theta_k \text{ for some } j \neq k \tag{1}$$

3.2 Comparing Multiple Means with an F-test

- We assume that our response variable follows a normal distribution and we assume that the mean of the distribution depends on the condition in which the measurements were taken, and that the variance is the same across all conditions.

$Y_{ij} \sim N(\mu_j, \sigma^2)$ for $i = 1, \ldots, n_j$, $j = 1, \ldots, m$

- The "gatekeeper" test for means is tested using an F-test

$H_0: \mu_1 = \cdots = \mu_m$ $H_A: \mu_j \neq \mu_k$ for some j and k

- In particular, we use the F-test for overall significance in an *appropriately defined linear regression model*:
  - The *appropriately defined linear regression model* in this situation is one in which the response variable depends on $m - 1$ indicator variables:

$$x_{ij} = \begin{cases} 1 & \text{if unit } i \text{ is in condition } j \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, 2, \ldots, m-1$.

  - For a particular unit i, we adopt the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{m-1} x_{i,m-1} + \varepsilon_i$$

$i = 1, \ldots, N = \sum_{j=1}^{m} n_j$

have the diagnostic plots/tests for residuals is relevant.

– In this model the β's are unknown parameters and may be interpreted in the context of the following expectations:

$$\mathrm{E}[Y_i|x_{i1} = x_{i2} = \cdots = x_{i,m-1} = 0] = \beta_0 = \mu_m$$

$$\mathrm{E}[Y_i|x_{ij} = 1] = \beta_0 + \beta_j = \mu_j \ (j \neq m)$$

– Based on these assumptions, H_0 in (1) is true if and only if $\beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0$. Thus testing (1) is equivalent to testing

$$H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ vs. } H_A\text{: } \beta_j \neq 0 \text{ for some } j$$

– This hypothesis corresponds, as noted, to the F-test for overall significance in the model.

- In regression parlance, the test statistic is defined to be the ratio of the regression mean squares (MSR) to the mean squared error (MSE) in a standard regression-based analysis of variance (ANOVA):

$$t = \frac{MSR}{MSE}$$

- In our setting we can more intuitively think of the test statistic as comparing the response variability between conditions to the response variability within conditions:

$$\bar{y}_{\cdot\cdot} = \frac{\sum_{j=1}^{m} n_j \bar{y}_{\cdot j}}{\sum_{j=1}^{m} n_j}$$

$$SS_C = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 \quad \text{variation between conditions}$$

$$SS_E = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (\bar{y}_{\cdot j} - y_{ij})^2 \quad \text{variation within conditions}$$

$$SS_T = SS_C + SS_E$$

- The null distribution for this test is $F_{(m-1,N-m)}$

- The p-value for this test is calculated by

$$\text{p-value} = P(T \geq t)$$

  where $T \sim F_{(m-1,N-m)}$

- **Example: Candy Crush Boosters**
  - Candy Crush is experimenting with three different versions of in-game "boosters": the lollipop hammer, the jelly fish, and the color bomb.

Figure 2: Candy Crush Experiment

  - Users are randomized to one of these three conditions ($n_1 = 121$, $n_2 = 135$, $n_3 = 117$) and they receive (for free) 5 boosters corresponding to their condition. Interest lies in evaluating the effect of these different boosters on the length of time a user plays the game.
  - Let μ_j represent the average length of game play (in minutes) associated with booster condition $j = 1, 2, 3$. While interest lies in finding the condition associated with the longest average length of game play, here we first rule out the possibility that booster type does not influence the length of game play (i.e., $\mu_1 = \mu_2 = \mu_3$).
  - In order to do this we fit the linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

    where x_1 and x_2 are indicator variables indicating whether a particular value of the response was observed in the jelly fish or color bomb conditions, respectively. The lollipop hammer is therefore the reference condition.

Optional Exercises:

- Calculations: 2, 7
- Proofs: 1, 5, 6, 9, 10, 14, 17, 18
- R Analysis: 2, 5, 6, 8, 13(g), 17 (not g,h), 22(h), 23(a-f)