

Randomization Test for Comparing Two Treatment Means

Alternative to two sample t-test for experiments

- Two treatments with means μ_1 (Treatment 1) and μ_2 (Treatment 2)
e.g., μ_1 =the mean lean percentage for all hogs on regular/control diet,
 μ_2 =the mean lean percentage for all hogs on new diet
- In the experiment, there are a total of $n_1 + n_2$ experimental units (EUs) (e.g., 100 hogs)
- Randomly assign n_1 EUs to Treatment 1 & n_2 EUs to Treatment 2
e.g., $n_1 = 50$ hogs to control diet, $n_2 = 50$ to new diet
- Compute averages \bar{Y}_1, \bar{Y}_2 for treatment groups & observe a difference of $\bar{Y}_2 - \bar{Y}_1 = d$, say
e.g., $\bar{Y}_2 - \bar{Y}_1 = 52.3 - 50.8 = 1.5 = d$ lean percentage points
- We wish to conduct a hypothesis test

Hypotheses: $H_0: \mu_1 = \mu_2$ (or $\mu_2 - \mu_1 = 0$) *null hypothesis*
 $H_a: \mu_2 > \mu_1$ (or $\mu_2 - \mu_1 > 0$) *alternatives*
 $H_a: \mu_2 < \mu_1$ (or $\mu_2 - \mu_1 < 0$)
 $H_a: \mu_2 \neq \mu_1$ (or $\mu_2 - \mu_1 \neq 0$)

- We need to assess if the observed difference d is more consistent with $H_0: \mu_1 = \mu_2$ or some alternative H_a that implies a difference in the treatments means/effects (e.g., $H_a: \mu_2 > \mu_1$).

Under $H_0: \mu_1 = \mu_2$ we would expect the difference in sample averages $\bar{Y}_2 - \bar{Y}_1$ to be around *zero*, but samples vary and our observed difference d from our experiment *might not be zero* just due to chance variation.

Note also if:

$H_a: \mu_2 > \mu_1$ is really true, “large” *positive* values of $\bar{Y}_2 - \bar{Y}_1$ should be expected
 $H_a: \mu_2 < \mu_1$ is really true, “large” *negative* values of $\bar{Y}_2 - \bar{Y}_1$ should be expected
 $H_a: \mu_2 \neq \mu_1$ is really true, “large” *absolute* values of $|\bar{Y}_2 - \bar{Y}_1|$ should be expected

- To see if the observed difference d from our data supports $H_0: \mu_1 = \mu_2$ or one of the selected alternatives, do the following steps of a **Randomization Test**:

1. collect all $n_1 + n_2$ observed values in the experiment together into one pooled group
2. from the pooled group of $n_1 + n_2$ observed values, list all the possible ways that n_1 of these values could be split into a “Treatment 1” group and the remaining n_2 values could be placed into a “Treatment 2” group. Suppose there are N total possible splits into two groups.

3. For *each* possible arrangement/split into two groups, compute the sample averages \bar{Y}_1 and \bar{Y}_2 for our artificially created “Treatment 1” and “Treatment 2” groups and then compute the difference $\bar{Y}_2 - \bar{Y}_1$

Note: In place of Step 2-3, if it’s hard to list all possible splits into two groups, then for some large number of repetitions, randomly divide the $n_1 + n_2$ observed values into a “Treatment 1” group with n_1 observations and a “Treatment 2” group with n_2 observations. Let N denote the number of times we conducted random splits of our $n_1 + n_2$ observed values into two groups. For each random split performed, compute the difference $\bar{Y}_2 - \bar{Y}_1$.

4. To assess how “unusual” our original observed difference d is (i.e., does the difference d support H_0 or H_a ?), compute a p-value (a proportion) as follows.

$$\text{For } H_a: \mu_2 > \mu_1, \text{ p-value} = \frac{\# \text{ of times a difference } \bar{Y}_2 - \bar{Y}_1 \text{ was greater than or equal to our } d}{N}$$

$$\text{For } H_a: \mu_2 < \mu_1, \text{ p-value} = \frac{\# \text{ of times a difference } \bar{Y}_2 - \bar{Y}_1 \text{ was less than or equal to our } d}{N}$$

$$\text{For } H_a: \mu_2 \neq \mu_1 \text{ p-value} = \frac{\# \text{ of times an absolute diff } |\bar{Y}_2 - \bar{Y}_1| \text{ was greater than or equal to } |d|}{N}$$

The test is based on the idea that, if there is no difference in treatments, then the two treatment groups made in the original experiment were artificial labeling. We would have observed the same values for the EUs no matter in which group the EUs were placed. The p-value tells us how unusual our observed average difference d is compared to the background of all the average differences possible just by randomly dividing the observations into two treatment groups. If our difference d is unusual (as measured by the p-value), then we could claim that our difference d is *not* likely due to just chance randomization but really represents a difference caused by the treatments.

Hog Example: wish to test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_2 > \mu_1$

want to show hogs on treatment diet (“Treatment 2”) have higher mean lean percentage than the control hogs (“Treatment 1”)

From the $N = 10,000$ random divisions of 100 observed hog percentages into two groups of 50 hog percentages, 2 out of 10,000 splits produced an average difference $\bar{Y}_2 - \bar{Y}_1$ greater than or equal to our observed difference of $d = 1.5$.

Hence, p-value = $\frac{2}{10000} = 0.0002$. Such a difference of $d = 1.5$ is unlikely due to just chance.

Note the two sample t-test (different from randomization test) performed on the same data gives almost identical results (p-value < 0.0005).

Can use the two sample t-test to approximate p-values from randomization tests for experiments.