CUHK STAT1013: Data Science Toolbox Spring 2023

A/B Test: Test Statistic

Test statistic: construction and distribution

Lecturer: Ben Dai

1 Test statistic: construction

Next, we try to find out the distribution of a test statistic. Recall our goal is to check if there is any contradiction between the test statistic $T(X_1,\cdots,X_n)$ under H_0 and its empirical estimator $t_* = T(x_1,\cdots,x_n)$.

A question raises: how to construct a test statistic? In fact, there is not a fixed procedure to construct a test statistic, but there are some guidelines or motivation.

- The test statistic is initially motivated by the hypothesis, e.g. the population mean.
- We aim to construct the test statistic so that its distribution should be as simple as possible. For example, $N(0,1)$.

In A/B tests, most test statistics we will see have the following form:

$$T(X_1,\cdots,X_n) = \frac{\widehat{\theta}_n - \theta_0}{\mathrm{sd}(\widehat{\theta}_n)}, \quad \Big(\frac{\text{empirical estimate} - \text{hypothesized value under } H_0}{\text{sd of estimate under } H_0}\Big) \tag{1}$$

Recall the example,

$$H_0 : \delta = r_B - r_A = 0, \quad vs. \quad H_1 : \delta = r_B - r_A \neq 0$$

The test statistics we will consider is:

$$T = \frac{\widehat{\delta}_n - \delta}{\mathrm{sd}(\widehat{\delta}_n)} = \frac{\widehat{r}_B - \widehat{r}_A - 0}{\mathrm{sd}(\widehat{r}_B - \widehat{r}_A)}. \tag{2}$$

Thus, we tend to investigate the distribution of the test statistics, alternative, the distribution of estimators.

2 Test statistic: sampling distribution

In general, it is almost impossible to study the entire population (or population parameters such as the population mean or the population standard deviation). For example, denote

- θ: unknown but fixed **population** parameter;
  - For instance, r_A and r_B are population parameters. Note that the hypothesis testing wants a conclusion at population level.
- $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$: an empirical estimate based on a random sample.

Quality of an estimator (optinal). For any given unknown population parameter there are many possible empirical estimates. For example, if we want to estimate an unknown population mean μ, we should use the sample mean $\bar{X}$ or the sample median $\tilde{X}$ (or anything else for that matter). So how do we measure the "quality" of an estimate?

Definition 2.1 (Unbiased estimator (optinal))**.** An estimator $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$ is an unbiased estimator of the parameter θ if $\mathbb{E}(\widehat{\theta}) = \theta$.

Definition 2.2 (BUE(optinal))**.** An unbiased estimator with the smallest variance is called the best unbiased estimator (BUE).

Example 2.3 (optinal)**.** *CLT Show that*

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$$

are unbiased estimators of μ and σ^2, respectively.

Recall that our goal is to find the distribution of the test statistic T in (2). Fundamentally, the sampling distribution of $\widehat{\mu}$. Let's itemize the results.

2.1 Distribution of T with known σ

Theorem 2.4 (Distribution of $\widehat{\mu}$ with known σ)**.** *Suppose $X_1, \cdots, X_n \overset{iid}{\sim} X$ with mean μ and variance σ^2, with sufficient data ($n \geq 30$), then by CLT*

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \ \dot{\sim}\ N(\mu, \frac{\sigma}{\sqrt{n}}).$$

Alternatively,

$$T := \frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \ \dot{\sim}\ N(0,1).$$

Example 2.5. *A factory manufactures mask packages that have a number of masks that is approximately normally distributed with a standard deviation of 1. If a sample of 100 packages has an average 246 masks, i.e, $\bar{\mu}_n = 246$. If my hypothesis is $\mu = 250$, what's the probability of observing a sample mean smaller or equal than $\bar{\mu}_n$.*

$$\mathbb{P}(\widehat{\mu} \leq \bar{\mu}_n) = \mathbb{P}\big(\frac{\widehat{\mu} - \mu}{\sigma/\sqrt{n}} \leq \frac{246-250}{1/\sqrt{100}}\big) = \mathbb{P}(Z \leq -40) = \texttt{norm().cdf(-40)} \sim 7.62 \times 10^{-24}.$$

We tend to believe that the hypothesis is wrong, since it is almost impossible to observe the sample mean like this if our hypothesis is correct.

2.2 Distribution of T with unknown σ

Theorem 2.6 (Distribution of $\widehat{\mu}$ with unknown σ)**.** *When σ is unknown we estimate it by using a sample standard deviation $\widehat{\sigma}$, then*

$$T := \frac{\widehat{\mu} - \mu}{\widehat{\sigma}/\sqrt{n}} \overset{\cdot}{\sim} t(n-1),$$

where $t(n-1)$ is a t-distribution with $n-1$ degrees of freedom, c.f. Source.

Note that $\lim_{n\to\infty} t(n-1) \overset{d}{=} N(0,1)$, that is when sample size is sufficiently large, the t-distribution will be a standard normal distribution.

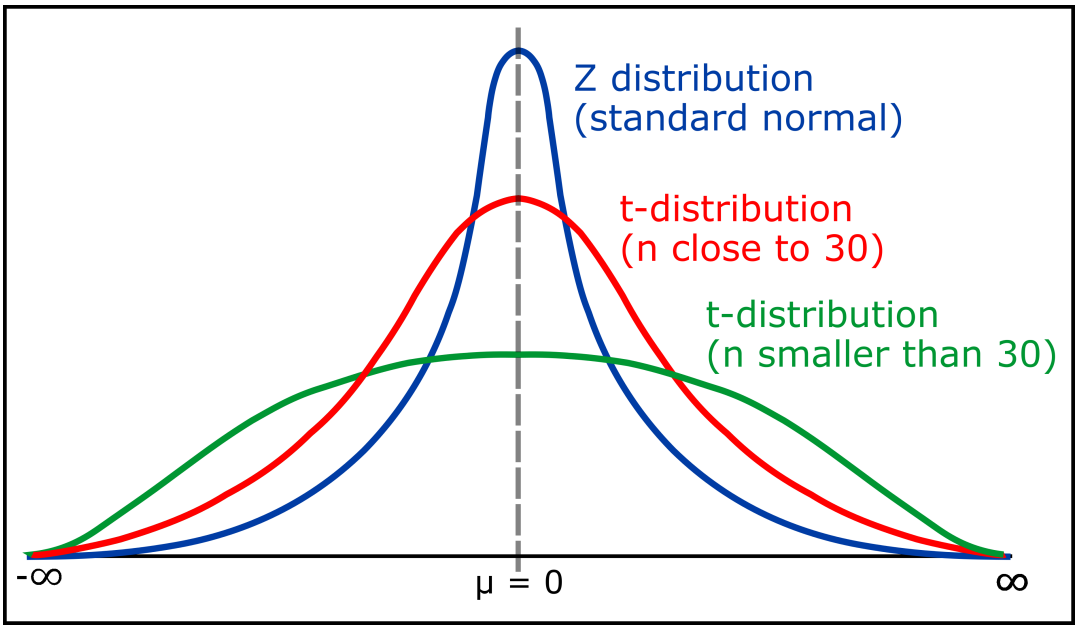

Figure 1: Source

3 p-value

The p-value is the probability that test statistic T (under H_0) takes a value as or more extreme than the one we observed t_*, or the probability of obtaining a sample statistic. Depending on alternative hypothesis H_1, the p-value is calculated differently.

- $H_1 : \mu < 0$

$$p = \mathbb{P}\big(T \leq t_n | H_0\big)$$

- $H_1 : \mu > 0$

$$p = \mathbb{P}\big(T \geq t_n | H_0\big).$$

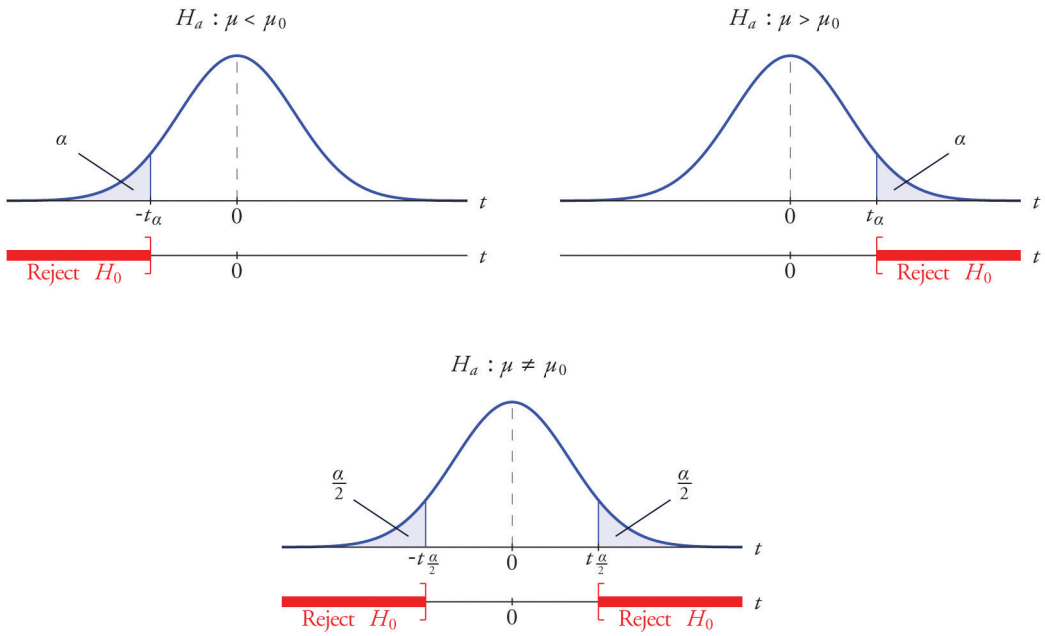

Figure 2: Distribution of the Standardized Test Statistic and the Rejection Region. [Source]

- $H_1 : \mu \neq 0$

$$p = \mathbb{P}\big(|T| \geq |t_n||H_0\big).$$

4 Summary

Finally, we summarize the construction and sampling distribution of our one-sample and two-sample test statistics.

4.1 One-sample (paired) test statistic

In this case, each homogeneous experimental unit receives both population conditions; as a result, each experimental unit has a pair of observations, one for each population.

For example, if we run a testing on a new treatment based on 50 individuals, the measures before and after going on the treatment form the information for our two samples. In this content, the two populations are "before" and "after", and the experimental unit is the individual. Obviously, the observations in a pair have something in common and not independent.

Given a paired dataset $\mathscr{D} = \{X_i, Y_i\}_{i=1}^n$, to determine if the treatment is effective, we consider the differences $d_i = Y_i - X_i$ in paired observations.

- Null Hypothesis:

$$H_0 : \delta = \mu_B - \mu_A = d_0$$

- Test statistic:

$$T = \frac{\text{empirical estimate} - \text{hypothesized value under } H_0}{\text{sd of estimate under } H_0},$$

  - Empirical estimate:

$$\delta \leftarrow \widehat{\delta} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i).$$

  - Hypothesized value under H_0: $\delta = 0$.
  - sd of estimate under H_0:

$$\mathbf{Var}(\widehat{\delta}) = \sigma_d^2/n,$$

    where σ_δ is the standard deviation of $d = X - Y$.

Therefore, we summarize as following two cases:

CASE 1. One Sample test statistic (difference between the means) with σ_δ is known.

Theorem 2.4 yields that

$$T = \frac{\widehat{\delta} - d_0}{\sigma_\delta/\sqrt{n}} \;\dot{\sim}\; N(0,1).$$

CASE 2. One Sample test statistic (difference between the means) with σ_δ is unknown.

Theorem 2.6 yields that

$$T = \frac{\widehat{\delta} - d_0}{\widehat{\sigma}_d/\sqrt{n}} \;\dot{\sim}\; t(n-1).$$

4.2 Two-samples test statistic

In this section, we consider the case that we have two independent datasets: $\mathscr{D}_A : \{X_i\}_{i=1}^n$ and $\mathscr{D}_B : \{Y_i\}_{i=1}^m$.

- Null Hypothesis:

$$H_0 : \delta = \mu_B - \mu_A = d_0$$

- Test statistic:

$$T = \frac{\text{empirical estimate} - \text{hypothesized value under } H_0}{\text{sd of estimate under } H_0},$$

  - Empirical estimate:

$$\delta \leftarrow \widehat{\delta} = \widehat{\mu}_B - \widehat{\mu}_A, \quad \mu_A \leftarrow \widehat{\mu}_A = \frac{1}{n}\sum_{i=1} X_i, \quad \mu_B \leftarrow \widehat{\mu}_B = \frac{1}{m}\sum_{i=1} Y_i.$$

  - Hypothesized value under H_0: $\delta = 0$.

– sd of estimate under H_0:

$$\mathbf{Var}(\widehat{\delta}) = \mathbf{Var}(\widehat{\mu}_B - \widehat{\mu}_A) = \mathbf{Var}(\widehat{\mu}_B) + \mathbf{Var}(\widehat{\mu}_A) = \sigma_A^2/n + \sigma_B^2/m.$$

Therefore, we summarize as following three cases:

CASE 3. Two Samples test statistic (difference between the means) with σ_A and σ_B are known.

$$T = \frac{\widehat{\mu}_B - \widehat{\mu}_A - d_0}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}} \;\dot{\sim}\; N(0,1).$$

CASE 4. Two Samples test statistic (difference between the means) with σ_A and σ_B are unknown but $\sigma_A = \sigma_B$.

In this case, the variance can be estimated by pooled variance[1] of two samples. The motivation is that we can pool $\mathscr{D}_A$ and $\mathscr{D}_B$ to estimate the common variance:

$$\widehat{\sigma}^2_{pool} = \frac{\sum_{i=1}^n (X_i - \widehat{\mu}_A)^2 + \sum_{i=1}^m (Y_i - \widehat{\mu}_B)^2}{n+m-2} = \frac{(n-1)\widehat{\sigma}_A^2 + (m-1)\widehat{\sigma}_B^2}{n+m-2}.$$

Thus, the test statistic is defined as:

$$T = \frac{\widehat{\mu}_B - \widehat{\mu}_A - d_0}{\sqrt{\frac{\widehat{\sigma}^2_{pool}}{n} + \frac{\widehat{\sigma}^2_{pool}}{m}}} \;\dot{\sim}\; t(n+m-2).$$

CASE 5 (Welch's t-test). Two Samples test statistic (difference between the means) with σ_A and σ_B are unknown and $\sigma_A \neq \sigma_B$.

In this case, the test statistic is defined as:

$$T = \frac{\widehat{\mu}_B - \widehat{\mu}_A - d_0}{\sqrt{\frac{\widehat{\sigma}_A^2}{n} + \frac{\widehat{\sigma}_B^2}{m}}} \;\dot{\sim}\; t\Big(\frac{\left(\frac{\widehat{\sigma}_A^2}{n} + \frac{\widehat{\sigma}_B^2}{m}\right)^2}{\frac{\left(\widehat{\sigma}_A^2/n\right)^2}{n-1} + \frac{\left(\widehat{\sigma}_B^2/m\right)^2}{m-1}} \Big) \;\dot{\sim}\; t\big(\min(n_1-1, n_2-1)\big).$$

Remark 4.1. This calculation for the degrees of freedom is cumbersome and is typically done by software. As an alternative, conservative option to using the exact degrees of freedom calculation can be made by choosing as $\min(n_1-1, n_2-1)$

[1] [Optional but recommended reading] https://en.wikipedia.org/wiki/Pooled_variance