Review of Statistics [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

June 9, 2021

[1] This section is based on Stock and Watson (2020), Chapter 3.

Estimators

Suppose you want to understand the distribution of X in the population.

- When a statistic $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is a function of an i.i.d. sample, then the distribution is determined by the population distribution is F and the sample size is n.
- We call the distribution of $\hat{\theta}$ the **sample distribution**.

The goal of an estimator $\hat{\theta}$ is to learn about the parameter θ, we evaluate the

- The exact bias and variance.
- The distribution under normality.
- The asymptotic distribution as $n \to \infty$.

Goodness of Estimators

Let $\hat{\theta}$ be an estimator of θ. Then

- The bias of $bias(\hat{\theta})$ is $\hat{\theta} - \theta$.
  - We say an estimator is **unbiased** if the bias is 0.
- The **mean squared error** of an estimator $\hat{\theta}$ for θ is

$$mse(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

  - The mean squared error is $mse(\hat{\theta}) = \text{var}(\hat{\theta}) + (bias(\hat{\theta}))^2$.

Best Unbiased Estimator

Definition 1 (Best Linear Unbiased Estimator (BLUE))

If $\sigma^2 < \infty$ the sample mean $\bar{X}_n$ has the lowest variance among all linear unbiased estimators of μ.

Hypothesis

- A point hypothesis is the statement that θ equals a specific value θ_0.
- A common example is θ measures the effect the proposed policy. A typical question is whether $\theta = 0$, which can be written as $\theta_0 = 0$.
- The **null hypothesis**, written as $H_0 : \theta = \theta_0$, is the restriction $\theta = \theta_0$.
- The **alternative hypothesis**, written as $H_A : \theta \neq \theta_0$, is the set $\{\theta \in \Theta : \theta \neq \theta_0\}$.
  - **One-sided** hypothesis: $H_A : \theta > \theta_0$.
  - **Two-sided** hypothesis: $H_A : \theta \neq \theta_0$.

Acceptance and Rejection

- A hypothesis test is a decision based on data. We can either **fail to reject** the null hypothesis or **reject** the alternative hypothesis.
- An alternative way to express a decision rule is to construct a real-valued function of the data called a **test statistics**

$$T = T(X_1, \ldots, X_n)$$

together with a **critical region** C.
- A hypothesis can be expressed as
  - Accept H_0 if $T \in C$.
  - Reject H_0 if $T \notin C$.

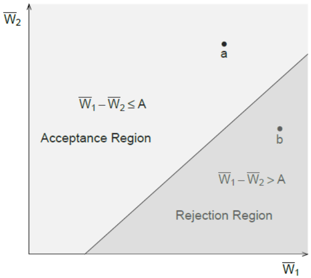
Note: "Accept" H_0 does not mean H_0 is true.

Example - Hypothesis Testing I

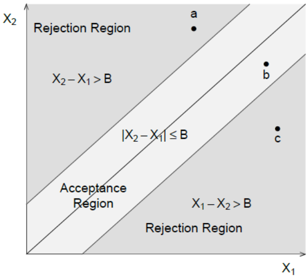
Consider the following examples:

- $2n$ adults who were raised in similar settings, n attended early childhood education. Let $\bar{W}_1$ be the average wage in the early childhood education group, and let $\bar{W}_2$ be the average wage in the remaining sample. Null hypothesis $H_0 : \bar{W}_1 > \bar{W}_2$.
- You ride each bus once and record the time it takes to travel from home to the university. Let X_1 and X_2 be the two recorded travel times. You adopt the following decision rule: If the absolute difference in travel times is greater than B minutes you will reject the hypothesis that the average travel times are the same, otherwise you will accept the hypothesis.

Example - Hypothesis Testing II



(a) Early Childhood Education Example



(b) Bus Travel Example

Type I and Type II error

- A false rejection of the null hypothesis is a **Type I error**.
- A false acceptance of the alternative hypothesis is a **Type II error**.

	Accept H_0	Reject H_0
H_0 true	Correct Decision	Type I Error
H_1 true	Type II Error	Correct Decision

Size and Power

The **power function** of a hypothesis test is the probability of rejection

$$\pi(F) = \mathbb{P}(\text{Reject } H_0 | F) = \mathbb{P}(T \in C | F).$$

- The **size** of a hypothesis test is the probability of a Type I error.
- The **power** of a hypothesis test is the complement of the probability of the Type II error.

P-value

Suppose we use a test which has the form: " Reject H_0 when $T > c$", how to report the results?
A simple choice is to report the "**p-value**", which is

$$p = 1 - G_0(T),$$

where $G_0(\cdot)$ is the null sampling distribution.
If $G_0(c) = \alpha$, the decision is identical to "Reject H_0 if $p < \alpha$".
Reporting p-values is especially useful when T has complicated or unusual distribution.

Computing p-value

- Suppose we are interested in testing the null hypothesis in $H_0 : \mathbb{E}(X) = \mu$ with the alternative hypothesis $H_A : \mathbb{E}(X) \neq \mu$.
  - Two-sided test.
- We observe the realization of $X_1, \ldots, X_n$ as $x_1, \ldots, x_n$.
- Note that $\bar{X}$ is a function of $X_1, \ldots, X_n$, which are i.i.d., therefore is a random variable.
  - Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - and $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
- Under H_0, the distribution of $\frac{\bar{X} - \mathbb{E}(X)}{\sigma_{\bar{X}}} \sim N(0, 1)$(CLT).
- $p = 1 - \mathbb{P}\left(\left| \frac{\bar{X} - \mathbb{E}(X)}{\sigma_{\bar{X}}} \right| < \left| \frac{\bar{x} - \mathbb{E}(X)}{\sigma_{\bar{X}}} \right| \right)$.

Issue: $\sigma_{\bar{X}}$ unknown.

Sample Variance

If the following assumptions hold:

1. $X_1, \ldots, X_n$ are i.i.d.
2. $\mathbb{E}(X_i) < \infty$.

The sample variance is computed

$$\bar{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- μ is unknown, need to be estimated.
- $\mathbb{E}((X - \bar{X})^2) \to \frac{n-1}{n}\sigma$.
- The sample variance is a consistent estimator of the population variance.

t-statistic

The standardized sample average can be constructed using

$$t = \frac{\bar{X} - \mu}{\sqrt{\bar{s}^2}}.$$

With the sample of $x_1, \ldots, x_n$, we can compute the sample t-statistic t^{sample}.
The p-value is given by

$$p\text{-value} = 2\Phi(-|t^{sample}|).$$

Significance Level

When construct hypothesis test, can fix a significance level.

- α-significance test means the tolerance to make Type I error is α.
- α is referred to as the **size** of the test.

Suppose the two-sided test has the **significance level** of α, the rule is "Reject H_0 if $|t^{sample}| > 1 - \Phi^{-1}(\alpha/2)$".

- $\alpha = 1\%$, $1 - \Phi^{-1}(\alpha/2) = 2.58$.
- $\alpha = 5\%$, $1 - \Phi^{-1}(\alpha/2) = 1.96$.
- $\alpha = 10\%$, $1 - \Phi^{-1}(\alpha/2) = 1.64$.

Confidence Interval I

We are interested in learning a parameter of interest θ from i.i.d. random sample of $X_1, \ldots, X_n$.

- With random sampling error, it's impossible to learn the exact value of the parameter of interest.
- Construct a **confidence set**: the parameter of interested has $1 - \alpha$ probability to fall into the confidence set.
- The **coverage probability** of the interval estimator is the probability that the random interval contains the true parameter.
    - An $1 - \alpha$ **asymptotic confidence interval** for a parameter has the **asymptotic coverage probability** $1 - \alpha$.

Confidence Interval II

A normal-based $1-\alpha$ confidence interval is

$$CI = [\hat{\theta} - Z_{1-\alpha/2} s(\hat{\theta}), \hat{\theta} + Z_{1-\alpha/2} s(\hat{\theta})],$$

where $\hat{\theta}$ is the estimator for θ and $se(\hat{\theta})$ is the estimated standard deviation. $Z_{1-\alpha/2}$ is the $1-\alpha/2$-quantile of a normal distribution.

Test for Difference Between Two Groups I

Suppose we observe the i.i.d sample $W_1, \ldots, W_{n_1}, \ldots, W_n$.

- Sample $W_1, \ldots, W_{n_1}$ are the monthly wage of graduates with master's degree, let μ_1 denote the population mean and σ_1^2 the population variance of group 1.
- Sample $W_{n_1+1}, \ldots, W_n$ are the monthly wage of graduates with bachelor's degree, let μ_2 denote the population mean and σ_2^2 the population variance of group 2.
- Let $n_2 = n - n_1$.
- $H_0 : \mu_1 - \mu_2 > d_0$, $H_1 : \mu_1 - \mu_2 \leq d_0$, with significance level of α.

Test for Difference Between Two Groups II

- The parameter of interest is $\theta = \mu_1 - \mu_2$.
- Let $\bar{W}_1$ and $\bar{W}_2$ be the estimated sample mean and s_1^2 and s_2^2 be the estimated sample variance for group 1 and group 2.
- The standard error of $\hat{\theta} = \bar{W}_1 - \bar{W}_2$ is $se(\hat{\theta}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
- We construct the t-statistic as $t = \frac{\hat{\theta} - d_0}{se(\hat{\theta})}$.
- We reject H_0 if $t > Z_{1-\alpha}$.

References I

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, volume 4. Pearson New York.