## ☆ Hypothesis testing (t-test)

problem ①: for linear regression model with p predictors.

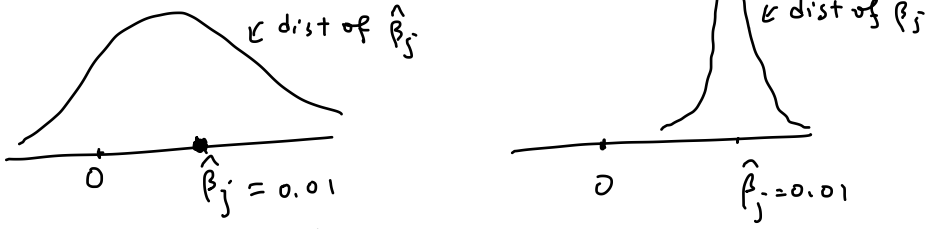
$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + e_i.$$

Want to test null hypothesis $H_0$ vs Alternative Hypothesis $H_A$:

$$H_0: \beta_j = \beta_j^* \quad \text{v.s} \quad H_A: \beta_j \neq \beta_j^*, \quad j \in \{0, 1, \cdots p\}$$

where $\beta_j^*$ is prespecified, such as $\beta_j^* = 0$ for many applications

E.g $H_0: \beta_j = 0$ v.s $H_A: \beta_j \neq 0$




§1. ☆ Review problem (from intro. to. Stats.)

(*) $Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} N(\beta_0, \sigma^2)$ ($\overset{iid}{\sim}$: independent and $\sim_R$)

Want to test $H_0: \beta_0 = 0$ v.s $H_A: \beta_0 \neq 0$

[Note that (*) is equivalent to the regression model with only intercept term in the mean function.

$i = 1, \cdots, n \quad Y_i = \beta_0 + e_i$

if $e_i \sim N(0, \sigma^2) \Rightarrow Y_i \sim N(\beta_0, \sigma^2)$

✔ Z-test (if $\sigma^2$ is known)

$$Z \text{ test statistic} = \frac{\bar{Y}}{\sigma/\sqrt{n}}$$

$\bar{Y}$ ← estimator of $\beta_0$

$\sigma/\sqrt{n}$ ← s.e. of this estimator.

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

$\sigma/\sqrt{n}$ is the standard error of the estimator $\bar{Y}$ (s.e)

$$\text{s.e.}(\bar{Y}) = \sqrt{\text{Var}(\bar{Y})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

(⇑ practice problem)

Recall that in Hypothesis testing, generally we want to control

$$\text{Type I error} = P(\text{Reject } H_0 \mid H_0 \text{ is true})$$

Smaller than ($\leq$) significant level $\alpha$

(typically $\alpha = 0.05$ in many applications)

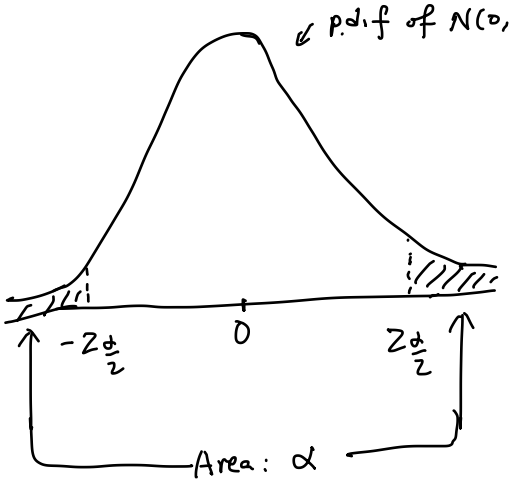
while having a good statistical power

= 1 − Type II error

with Type II error $= P(\text{Accept } H_0 \mid H_A \text{ is true})$

For z-test statistic, we have

When $H_0$ is true, $\quad z\text{-statistic} = \dfrac{\bar{Y}}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$.




Test Procedure:

① compute $z$-statistic from Data

② find a critical value $z_{\frac{\alpha}{2}}$ such that $P(\text{Rej } H_0 \text{ if } |Z| > z_{\frac{\alpha}{2}} \mid H_0 \text{ is true}) = \alpha$

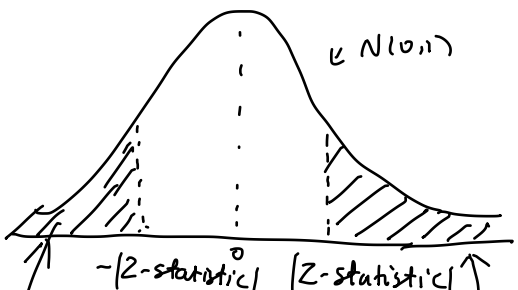
if $\alpha = 5\%$, $z_{\frac{\alpha}{2}} \approx 1.96$

③ Rej $H_0$ if $|z\text{-statistic}| > z_{\frac{\alpha}{2}}$

Accept $H_0$ otherwise.

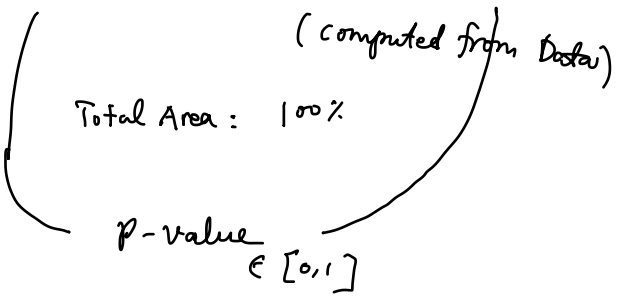
Or equivalently, we can use the p-value to perform the above test:

Replace ②&③ by ②′ calculate p-value $= P(N(0,1) > |z\text{-statistic}| \text{ or } < -|z\text{-statistic}|)$




N(0,1): dist of test statistic under $H_0$

$|z\text{-statistic}|$: test statistic




③' Rej $H_0$ if p-value $< \alpha$

Accept $H_0$ otherwise.

✓ t-test. (when $\sigma^2$ is unknown)

$$t\text{-statistic} = \frac{\bar{Y}}{\hat{\sigma}/\sqrt{n}}$$

($\bar{Y}$: estimator of $\beta_0$; $\hat{\sigma}/\sqrt{n}$: estimated s.e.)

where $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$ ← sample variance of $\{Y_1 \cdots Y_n\}$.

Note that $\hat{\sigma}/\sqrt{n}$ is an estimated standard error of $\bar{Y}$.

Under $H_0$ ($H_0$ is true),

$$t\text{-statistic} = \frac{\bar{Y}}{\hat{\sigma}/\sqrt{n}} \overset{H_0}{\sim} t\text{-distribution}$$

with degrees of freedom (d.f.) $n-1$.

when $n \to \infty$ t-dist with $n-1$ d.f. $\longrightarrow N(0,1)$.

t-test $\approx$ Z-test.

## §2. Regression Setting.

mean function: $E[Y_i|x_i] = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_j x_{i,j} + \cdots + \beta_p x_{i,p}$

(problem ①): $H_0: \beta_j = \beta_j^*$ v.s $H_A: \beta_j \neq \beta_j^*$.

$H_0$: mean function $E[Y_i|x_i] = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_j^* x_{i,j} + \cdots + \beta_p x_{i,p}$

t-test Test Statistic:

$$t\text{-statistic} = \frac{\hat{\beta}_j - \beta_j^*}{s.e(\hat{\beta}_j)}$$

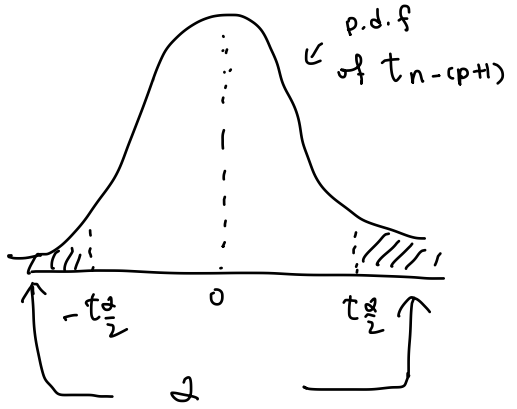
$\varepsilon_{n\times 1} \sim N(0, \sigma^2 I)$

When $H_0$ is true and statistical errors $(\varepsilon_1 \cdots \varepsilon_n)$ are normally distributed.

the above t-statistic $\overset{H_0}{\sim}$ t-distribution with d.f. $n-(p+1)$

✓ Test Procedure: ① compute t-statistic from Data

② Find the critical value $t_{\frac{\alpha}{2}}$, such that




$$P(\text{Rej } H_0 \text{ if } |t\text{-stat}| > t_{\frac{\alpha}{2}} \mid H_0 \text{ is true}) \leq \alpha \quad (\text{e.g } \alpha = 5\%)$$

$$t_{\frac{\alpha}{2}} \approx 2$$

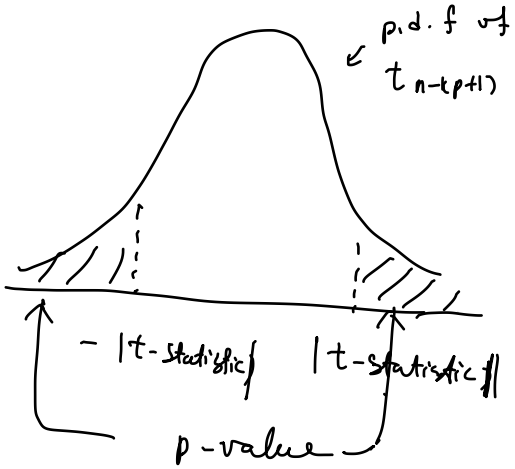
③ Rej Ho if $|t\text{-statistic}| > t_{\frac{\alpha}{2}}$

Accept Ho otherwise.

$$-t_{\frac{\alpha}{2}} < \frac{\hat{\beta}_j - \beta_j^*}{s.e(\hat{\beta}_j)} < t_{\frac{\alpha}{2}}.$$

Equivalently to ② & ③, we can use the following ②' ③'

②' we compute the p-value of the t-statistic




$$p\left(\begin{array}{l} t\text{-distribution with } n-(p+1) \text{ d.f.} \\ > |t\text{-statistic}| \text{ or } < -|t\text{-statistic}| \end{array}\right)$$

③' Reject Ho if p-value $< \alpha$

Accept otherwise.

✓ ② & ③ are also equivalent to:

Accept Ho: $\beta_j = \beta_j^*$ if $\beta_j^*$ (prespecified value) falls in the interval

$$\beta_j^* \in \left[\hat{\beta}_j - t_{\frac{\alpha}{2}} \cdot s.e(\hat{\beta}_j),\ \hat{\beta}_j + t_{\frac{\alpha}{2}} \cdot s.e.(\hat{\beta}_j)\right],$$

($t_{\frac{\alpha}{2}}$: critical value in ②)

and Rej Ho otherwise.

The above interval is the $(1-\alpha)$-level Confidence Interval of $\beta_j$ $(j=0,1,\cdots,p)$.

A $(1-\alpha)$ level Confidence Interval is defined as a range of values such that with $(1-\alpha)$ probability, the range will contain the true unknown value of $\beta_j$.