

巨量資料管理學院碩士在職專班

統計分析

2022/11/11

陳光宏

假設檢定 (Hypothesis testing)

百分比的推論 (proportion)

$$p = \frac{x}{n} = \frac{\text{number of successes}}{\text{sample size}}$$

- 來自於二項式分布 (Binomial distribution)
- 根據中央極限定理，假設樣本數越大， p 的抽樣分布服從 $N(\pi_0, \frac{\pi_0(1-\pi_0)}{n})$
- 檢定統計量

$$z_{\text{calc}} = \frac{p - \pi_0}{\sigma_p} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$$

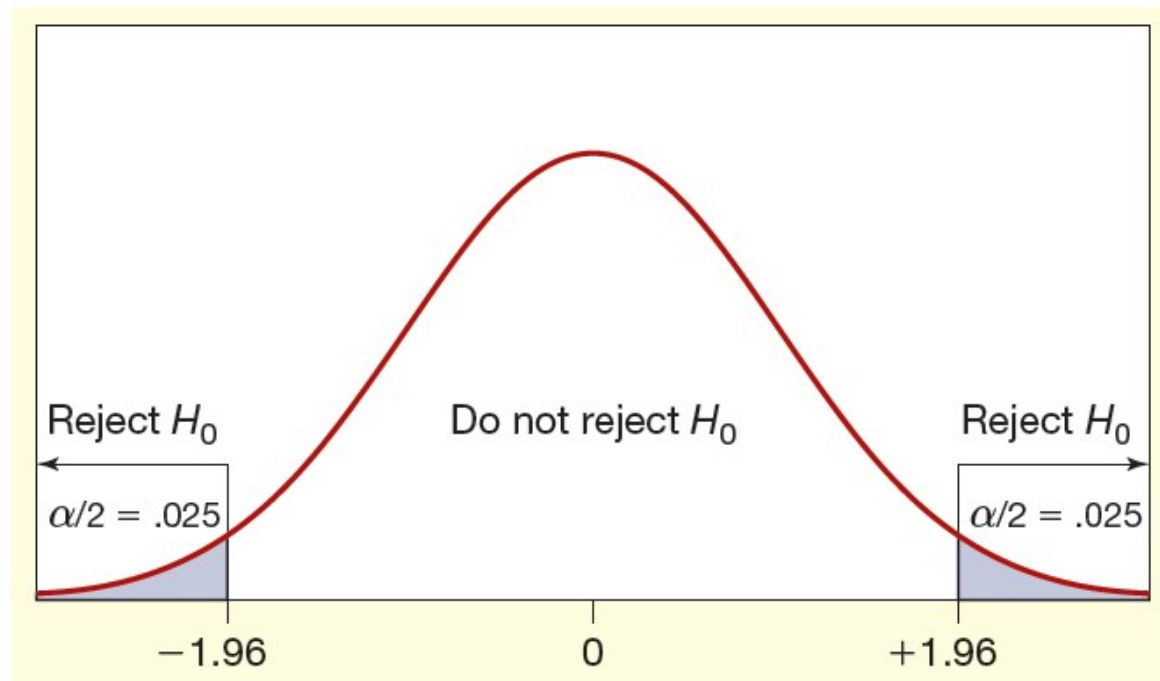
範例

- 零售商公司想要知道今年某商品退換貨的比例是否跟以前差不多
- 假設過去幾年退換貨的比例約13%
- 今年該商品售出250件，22件有退換貨記錄

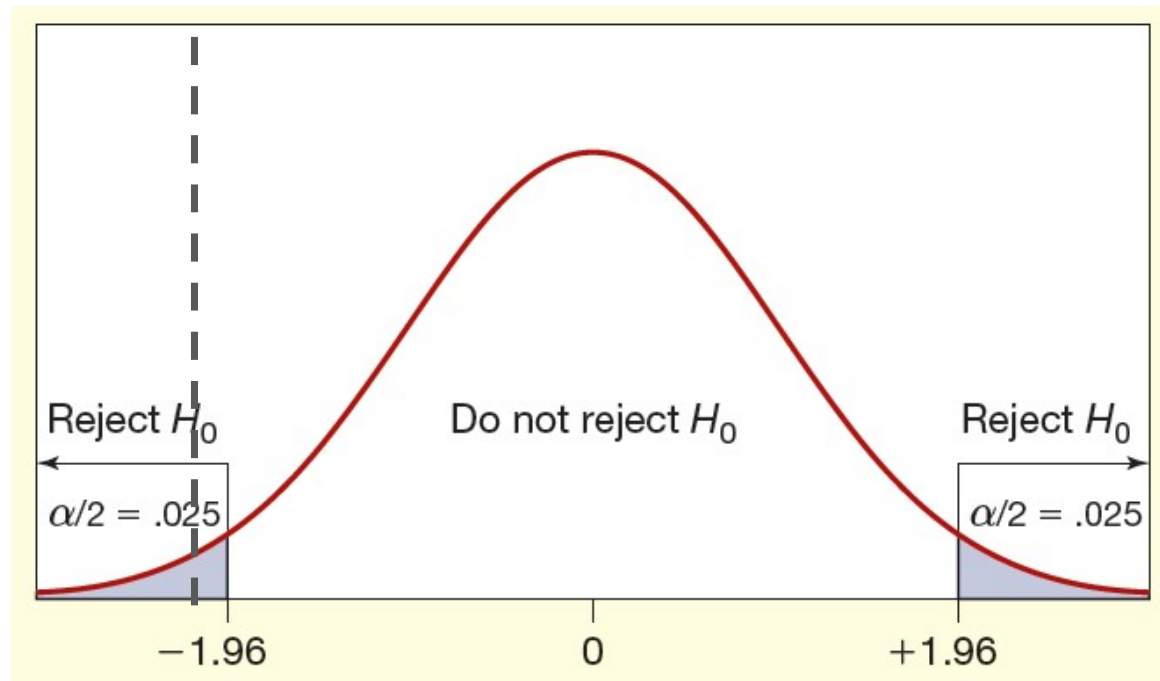
範例解答

$H_0: \pi = .13$ (return rate is the same as the historical rate)

$H_1: \pi \neq .13$ (return rate is different from the historical rate)



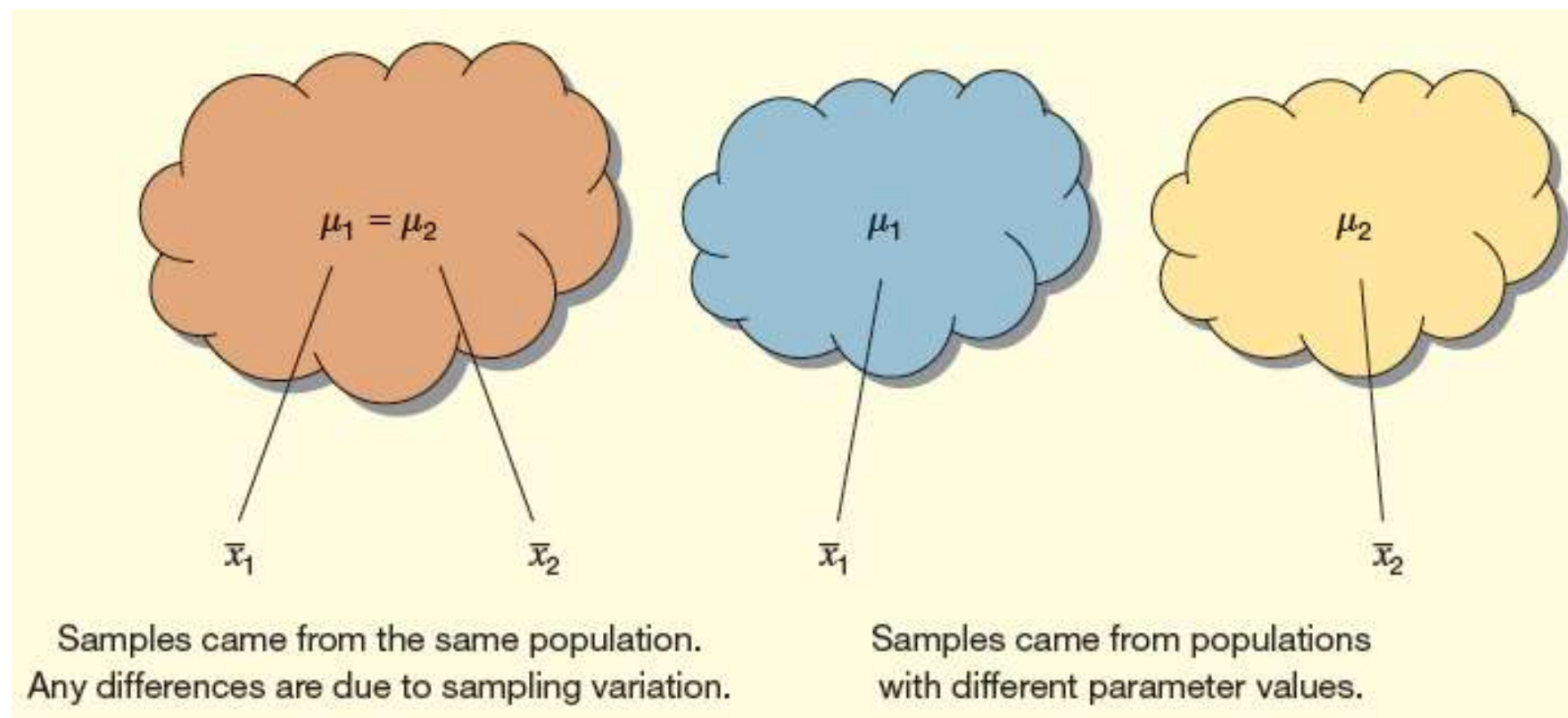
$$z_{\text{calc}} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{.088 - .13}{\sqrt{\frac{.13(1 - .13)}{250}}} = \frac{-.042}{.02127} = -1.975$$



$$2 \times P(Z < -1.975) = 2 \times .02413 = .04826$$

兩個獨立樣本的檢定

Two-sample t-test



假說建立

- 比較兩組有無差別
- 兩組平均值相減=0 (雙尾檢定)

Left-Tailed Test

$$H_0: \mu_1 - \mu_2 \geq D_0$$

$$H_1: \mu_1 - \mu_2 < D_0$$

Two-Tailed Test

$$H_0: \mu_1 - \mu_2 = D_0$$

$$H_1: \mu_1 - \mu_2 \neq D_0$$

Right-Tailed Test

$$H_0: \mu_1 - \mu_2 \leq D_0$$

$$H_1: \mu_1 - \mu_2 > D_0$$

檢定統計量

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Case 1: Known Variances

(10.1)

$$z_{\text{calc}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

檢定統計量

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Case 2: Unknown Variances Assumed Equal

(10.2)
$$t_{\text{calc}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \quad \text{where the pooled variance is}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad \text{and} \quad d.f. = n_1 + n_2 - 2$$

檢定統計量

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$


Case 3: Unknown Variances Assumed Unequal

$$(10.3) \quad t_{\text{calc}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{with } d.f. = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

練習

- 請參考
Data_w11.xlsx檔案
- 請比較兩個地區的平均價格是否有差異？

TABLE 10.2

Zocor Prices (30-Day Supply) in Two States  **Zocor**

Colorado Pharmacies		Texas Pharmacies	
City	Price (\$)	City	Price (\$)
Alamosa	125.05	Austin	145.32
Avon	137.56	Austin	131.19
Broomfield	142.50	Austin	151.65
Buena Vista	145.95	Austin	141.55
Colorado Springs	117.49	Austin	125.99
Colorado Springs	142.75	Dallas	126.29
Denver	121.99	Dallas	139.19
Denver	117.49	Dallas	156.00
Eaton	141.64	Dallas	137.56
Fort Collins	128.69	Houston	154.10
Gunnison	130.29	Houston	126.41
Pueblo	142.39	Houston	114.00
Pueblo	121.99	Houston	144.99
Pueblo	141.30		
Sterling	153.43		
Walsenburg	133.39		
$\bar{x}_1 = \$133.994$		$\bar{x}_2 = \$138.018$	
$s_1 = \$11.015$		$s_2 = \$12.663$	
$n_1 = 16$ pharmacies		$n_2 = 13$ pharmacies	

Source: Public Research Interest Group (www.pirg.org). Surveyed pharmacies were chosen from the telephone directory in 2004.

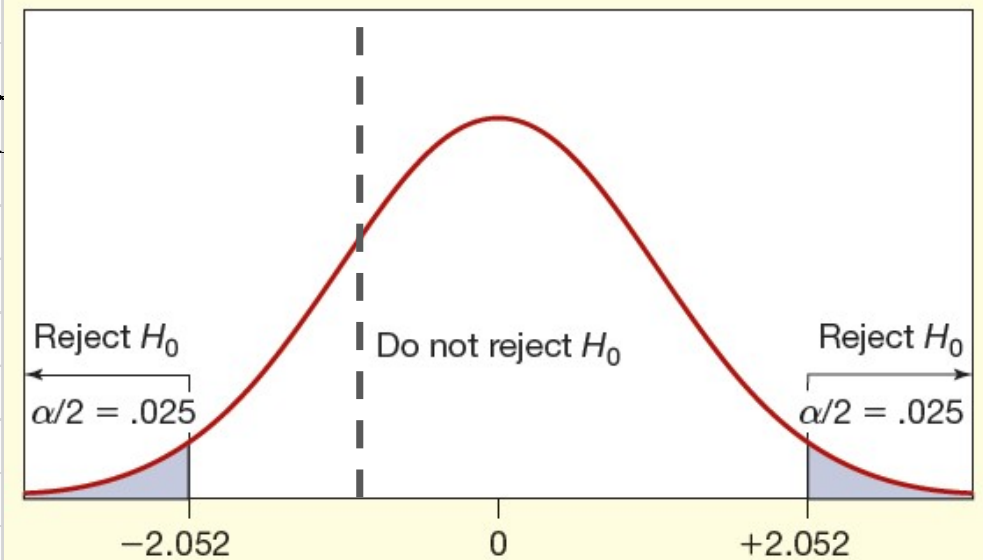
假設變異數相等

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

t 檢定：兩個母體平均數差的檢定，假設變異數相等

	變數 1	變數 2
平均數	133.99375	138.0184615
變異數	121.3293183	160.3542641
觀察值個數	16	13
Pooled 變異數	138.6737387	
假設的均數差	0	
自由度	27	
t 統計	-0.915314443	
P(T<=t) 單尾	0.184064721	
臨界值：單尾	1.703288446	
P(T<=t) 雙尾	0.368129443	
臨界值：雙尾	2.051830516	



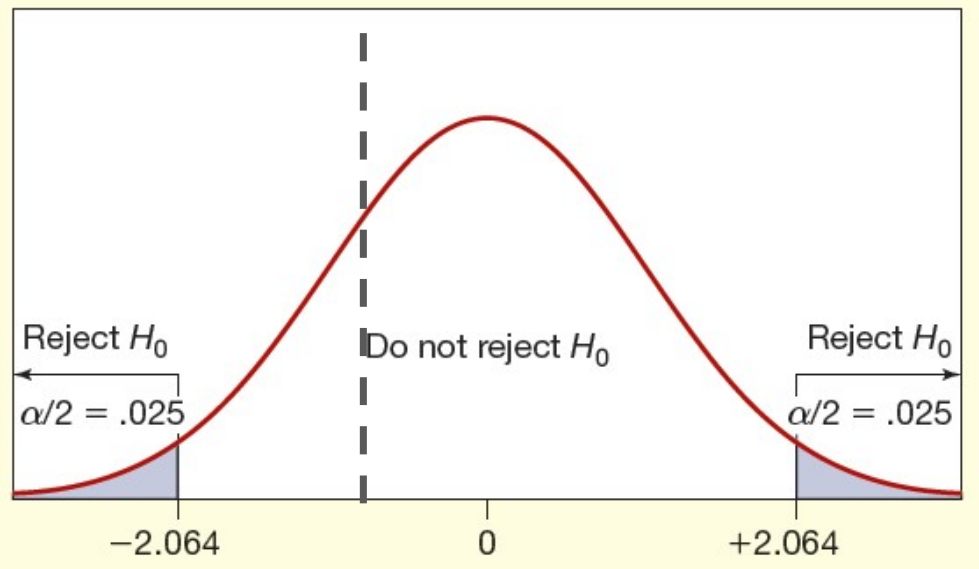
假設變異數不相等

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

t 檢定：兩個母體平均數差的檢定，假設變異數不相等

	變數 1	變數 2
平均數	133.99375	138.0184615
變異數	121.3293183	160.3542641
觀察值個數	16	13
假設的均數差	0	
自由度	24	
t 統計	-0.901802871	
P(T<=t) 單尾	0.188061598	
臨界值：單尾	1.71088208	
P(T<=t) 雙尾	0.376123196	
臨界值：雙尾	2.063898562	



討論

- 基本假設
 - 常態分布假設
 - 變異數相等假設
- 極端值的影響
- 遺漏值的影響
- 中央極限定理的重要性
 - 估計
 - 假設檢定

應用：A/B testing

- **情景:** 團隊想測試使用紅色按鈕還是藍色按鈕更能吸引顧客購買產品
- **A/B 測試:** Version A Website(紅色按鈕) vs. Version B Website(藍色按鈕)
- **主要成功指標 (Primary successful metrics):** 按鈕點擊率(Click Through Rate)

A/B Testing 是運用統計學的假設檢定(Hypothesis Testing) , 將兩個變量 (Control vs. Variant) 進行測試比較 , 以研究出哪一個變量效果更好。