# (PART) Statistical analysis

# Hypothesis testing

การทดสอบสมมติฐาน คือ การใช้วิธีทางสถิติในตอบคำถามสมมติฐานที่ต้องการ โดยมีขั้นตอน คือ

## สร้างสมมติฐาน (Construct the hypothesis)

* สมมติฐานว่าง (Null hypothesis; ) คือสมมติฐานที่ต้องการทดสอบ ซึ่งเปรียบเทียบได้กับสิ่งที่ทุกคนมีความเชื่อกันอยู่แล้ว (default belief) หรือไม่ทราบแน่ชัดว่าเป็นอย่างไร ซึ่งจะการแนวปฏิเสธไว้ก่อน คือ **ไม่มีความแตกต่างกัน**
* สมมติฐานทางเลือก (Alternative hypothesis; ) คือสมมติฐานที่คาดหวังว่าจะเป็นการค้นพบใหม่

## ทำการวิเคราะห์ค่าทางสถิติ (Construct the test statistics)

โดยเครื่องมือจะมีหลายรูปแบบตามลักษณะของข้อมูล เช่น t-test, Chi-square เป็นต้น ซึ่งผลลัพธ์จากการหาค่าทางสถิติจะเป็นไปตามการทดสอบนั้นๆ เช่น t-test: , Chi-square: , F-test: เป็นต้น

## หา p-value

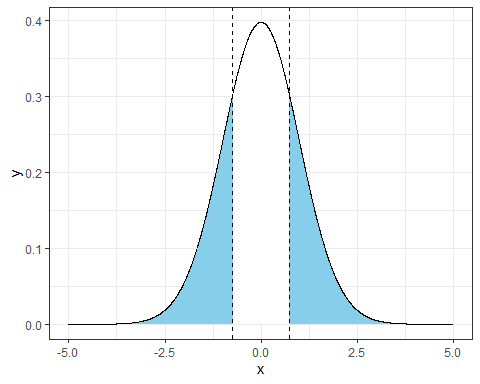
คือ โอกาสที่สามารถสังเกตค่าทางสถิติที่มากกว่าค่าการวิเคราะห์ทางสถิติที่โดยที่ยังอยู่ภายใต้ ได้ โดยทางปฏิบัติแล้วคือ การสร้างแบบจำลองข้อมูลภายใต้ ขึ้นมาแล้วแล้วดูว่า ที่ค่าสถิตินั้น มีโอกาสเกิดกี่ %

ยกตัวอย่าง เมื่อวิเคราะห์ t-test จะทำการสร้าง t-distribution ขึ้นมา (ตัวอย่างของการวิเคราะห์ อยู่[บทถัดไป](#sec-t-test) ) ภายใต้ ว่า mean = 10

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

t\_dist <- dt(seq(-5,5,0.01), df = 99) # t-distribution มี mean = 0 sd = 1  
  
tval <- 0.7345 # t-value วิธีคำนวณอยู่ในบท t-test  
  
density\_df <- data.frame(x = seq(-5,5,0.01), y = t\_dist) |>  
 mutate(area = ifelse(between(x,-tval,tval), TRUE,FALSE))  
  
ggplot(density\_df, aes(x,y)) +   
 geom\_area(fill = "skyblue", color = "black") +  
 geom\_area(data = filter(density\_df, area), fill = "white", color = "black") +  
 geom\_vline(xintercept = c(-tval,tval), linetype = "dashed") +  
 theme\_bw()



* ที่ t < -0.7345 หรือ t > 0.7345 (พื้นที่ใต้กราฟสีฟ้า) หมายถึงโอกาสที่ค่านั้น**ไม่ได้**เกิดจากความบังเอิญภายใต้ ที่ยังเป็นจริง (extreme value)
* ที่ -0.7345 < t < 0.7345 (พื้นที่ใต้กราฟสีขาว) หมายถึงโอกาสที่ค่านั้นเกิดจากความบังเอิญภายใต้ ที่ยังเป็นจริง

density\_df |>   
 filter(!area) |>   
 summarize(AUC = sum(y)) |> pull()

## [1] 46.40728

p <- 2\*pt(0.7345, 99, lower.tail = FALSE)   
p # p-value

## [1] 0.4643803

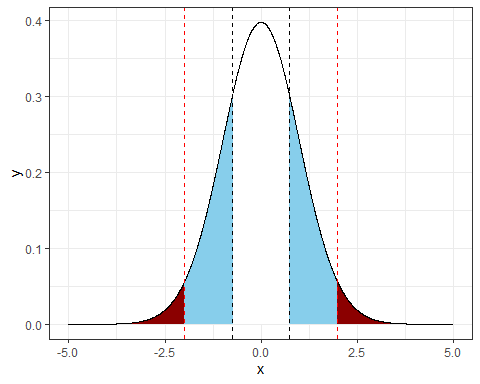
จะเห็นว่าค่า AUC = p-value = ~46.4% (ต่างกันเล็กน้อยจากความละเอียดของการพล็อต)

## เปรียบเทียบ p-value กับ ค่าวิกฤติ (critical value) ที่ยอมรับได้

จุดนี้จะเป็นการตัดว่า ท่านสามารถยอมรับความบังเอิญนี้ที่กี่ % โดยทั่วไปมักใช้ที่น้อยกว่า 5% (0.05) หรือ 1% (0.01) ซึ่งแบ่งเป็น

* Left-tailed คือ โอกาสที่ critical value
* Right-tailed คือ โอกาสที่ critical value
* Two-tailed คือ โอกาสที่ critical value นิยมใช้วิเคราะห์มากที่สุด

crit <- qt(1-0.05/2, df = 99) # ตัดที่ <= p 0.05, two-tailed  
  
density\_df <- density\_df |>   
 mutate(crits = ifelse(between(x,-crit,crit), TRUE,FALSE))  
  
ggplot(density\_df, aes(x,y)) +   
 geom\_area(fill = "darkred", color = "black") +  
 geom\_area(data = filter(density\_df, crits), fill = "skyblue", color = "black") +   
 geom\_area(data = filter(density\_df, area), fill = "white", color = "black") +  
 geom\_vline(xintercept = c(-tval,tval), linetype = "dashed") +  
 geom\_vline(xintercept = c(-crit,crit), linetype = "dashed", color = "red") +  
 theme\_bw()



จะเห็นว่า โอกาสที่ความแตกต่างนั้นเกิดจากความบังเอิญภายใต้ นั้นอยู่ที่ 46% ซึ่งมากกว่า 5% ที่ต้องการ จึงไม่สามารถ reject null hypothesis ได้ เรียกอีกแบบหนึ่งว่า ไม่ได้แตกต่างอย่างมีนัยสำคัญ (จนสามารถ reject ได้)

ตรงส่วนพื้นที่ระหว่าง critical value (ระหว่างเส้นประสีแดง) คือ พิสัยของความแตกต่างที่สามารถรับได้ ซึ่งจะถูกนำไปใช้ในการคำนวณ**ค่าความเชื่อมั่น (confidence interval)** ต่อไป

## Trade-off ของ p-value threshold

การตั้งค่าวิกฤตินั้นมี trade-off ระหว่าง false positive และ false negative ที่ต้องพิจารณา

| Decision/Truth |  |  |
| --- | --- | --- |
| **Reject** | Type I error | Correct |
| **Accept** | Correct | Type II error |

* Type I error = false positive = = critical value
* Type II error = false negative = = 1 - power

power = true negative