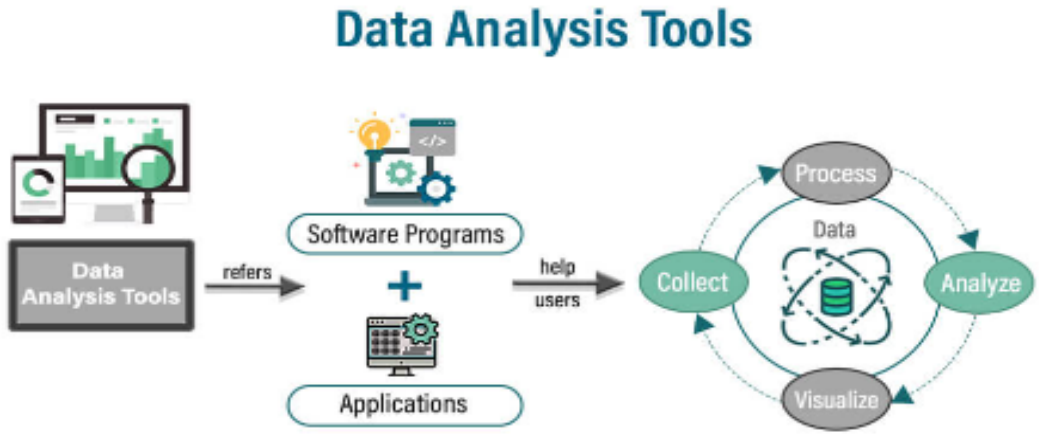
เอกสารประกอบการสอน

747-341

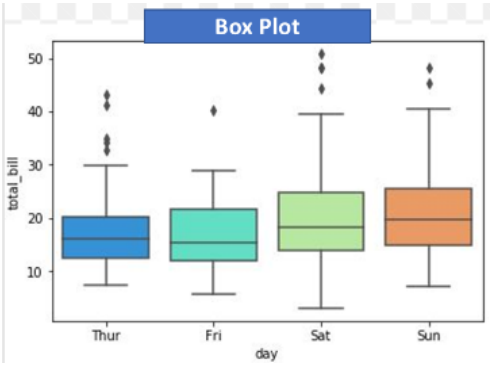
Data analytics and data visualization module



Asst. Prof. Dr. Arinda Ma-a-lee

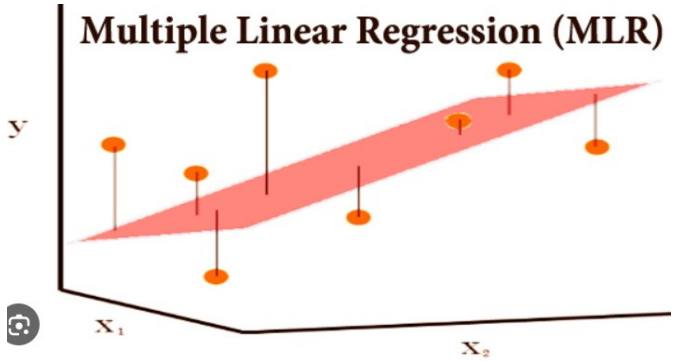
สถิติและการวิเคราะห์ข้อมูล

- การวิเคราะห์ข้อมูลเบื้องต้น 1 ตัวแปร
    - ตัวแปรต่อเนื่อง
    - ตัวแปรกลุ่ม



-การวิเคราะห์ข้อมูลรายคู่ระหว่างตัวแปรอิสะกับตัวแปรตาม

-การวิเคราะห์ข้อมูลด้วยตัวแบบทางสถิติ



สถิติและการวิเคราะห์ข้อมูล

ประเภท		ตัวแปรตาม		
		ทวิภาค (2 กลุ่ม)	ประเภท (3 กลุ่มขึ้นไป)	ปริมาณ (ต่อเนื่อง)
ตัวแปรอิสระ	ทวิภาค (2 กลุ่ม)	-2×2 tables -chi-squared -logistic regression	-2×c tables -chi-squared -multinomial regression	-two sample t tests -linear regression
	ประเภท (3 กลุ่มขึ้นไป)	-r×2 tables -chi-squared -logistic regression	-r×c tables -chi-squared -Multinomial regression	-analysis of variance (ANOVA) -linear regression
	ปริมาณ (ต่อเนื่อง)	-logistic regression	-multinomial regression	-linear regression

การวิเคราะห์ข้อมูลและการสรุปข้อมูล 1 ตัวแปร

- การสรุปข้อมูล เป็นการสรุปค่าทางสถิติเบื้องต้น
- สามารถสื่อสารได้เข้าใจง่าย
- เพื่อประกอบการตัดสินใจเลือกวิธีการจัดการข้อมูลได้อย่างเหมาะสมก่อนการวิเคราะห์ขั้นถัดไป
- เพื่อการเลือกสถิติที่เหมาะสมในการวิเคราะห์ข้อมูลเชิงอ้างอิงต่อไป

การวิเคราะห์ข้อมูลและการสรุปข้อมูล 1 ตัวแปร

ตัวอย่างข้อมูล

- ใช้ข้อมูลตัวอย่างจาก library(MASS) เป็นข้อมูลน้ำหนักของทารกแรกคลอด ที่เก็บรวบรวมจากศูนย์การแพทย์ Baystate (Baystate Medical Center) ในเมือง Springfield รัฐแมสซาชูเซตส์ จากหญิงตั้งครรภ์จำนวน 189 ราย ข้อมูลนี้ถูกเก็บในชื่อ “Bweight.csv” รายละเอียดข้อมูลแสดงดังตารางที่ 1

ตารางที่ 1 รายละเอียดตัวแปรของข้อมูล ชื่อ “birthwt”

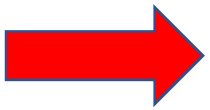
ตัวแปร	คำอธิบายตัวแปร	คำอธิบายค่าตัวเลข
low	น้ำหนักแรกคลอดต่ำกว่า 2,500 กรัม	0 คือ >= 2,500 , 1 คือ < 2,500
age	อายุของมารดาเป็นปี	
lwt	น้ำหนักของมารดาที่มีประจำเดือนครั้งสุดท้าย (ปอนด์)	
race	เชื้อชาติของมารดา	1=ผิวขาว, 2=ผิวดำ, 3=อื่นๆ
smoke	การสูบบุหรี่ขณะตั้งครรภ์	0= ไม่สูบ, 1= สูบ
ptl	จำนวนของการคลอดก่อนกำหนด	
ht	ประวัติการเป็นโรคความดันโลหิตสูง	0=ไม่มี, 1=มี
ui	การกลั้นปัสสาวะไม่อยู่	0=ไม่มี, 1=มี
ftv	จำนวนครั้งในการพบแพทย์ของไตรมาสแรก	
bwt	น้ำหนักแรกคลอดเป็นกรัม	

วัตถุประสงค์ของการศึกษา

- เพื่อศึกษาปัจจัยเสี่ยงที่มีความสัมพันธ์ต่อน้ำหนักแรกคลอดของทารก(กรัม)

ตัวแปร

ตัวแปรต้น



ตัวแปรตาม

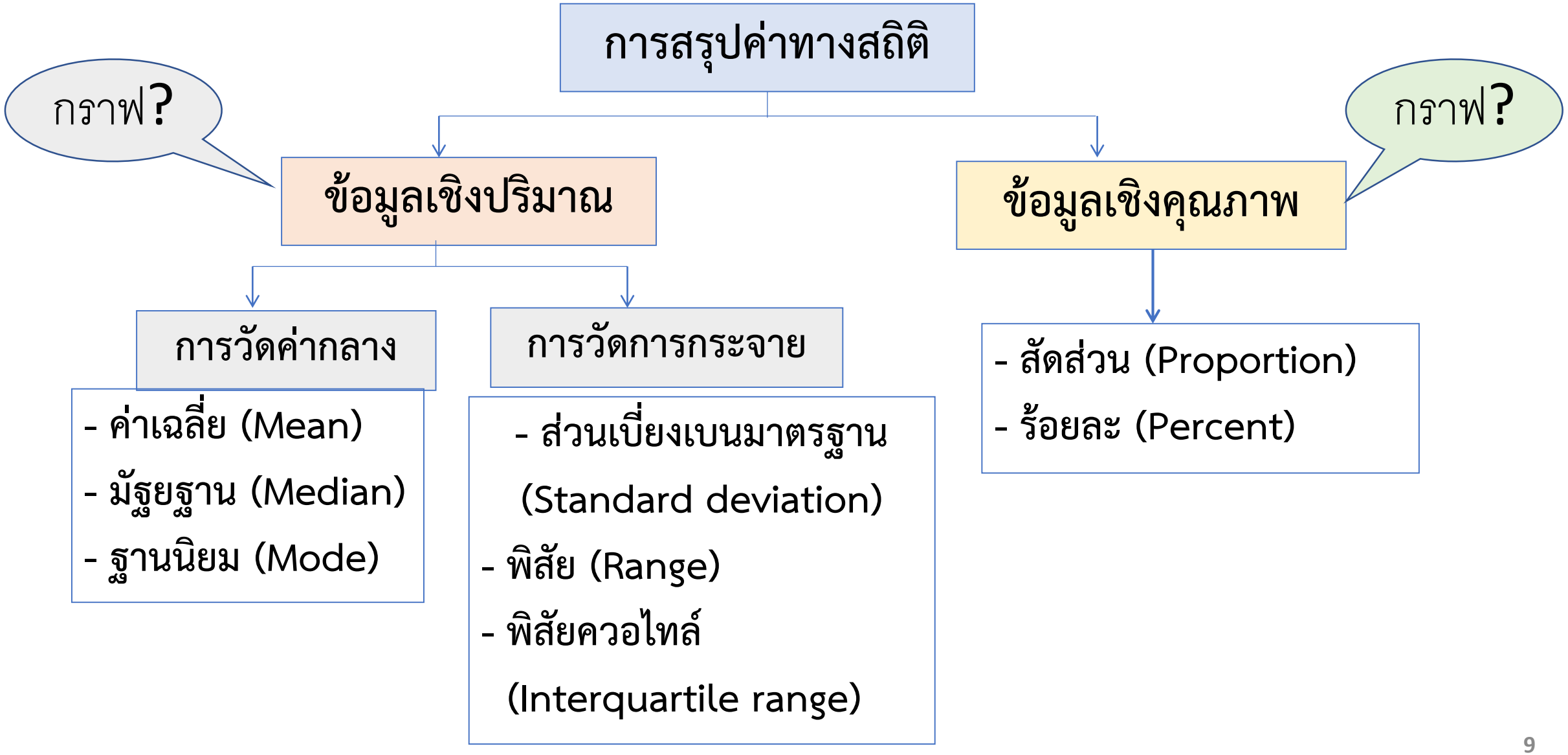
Data structure

Variables

Filter

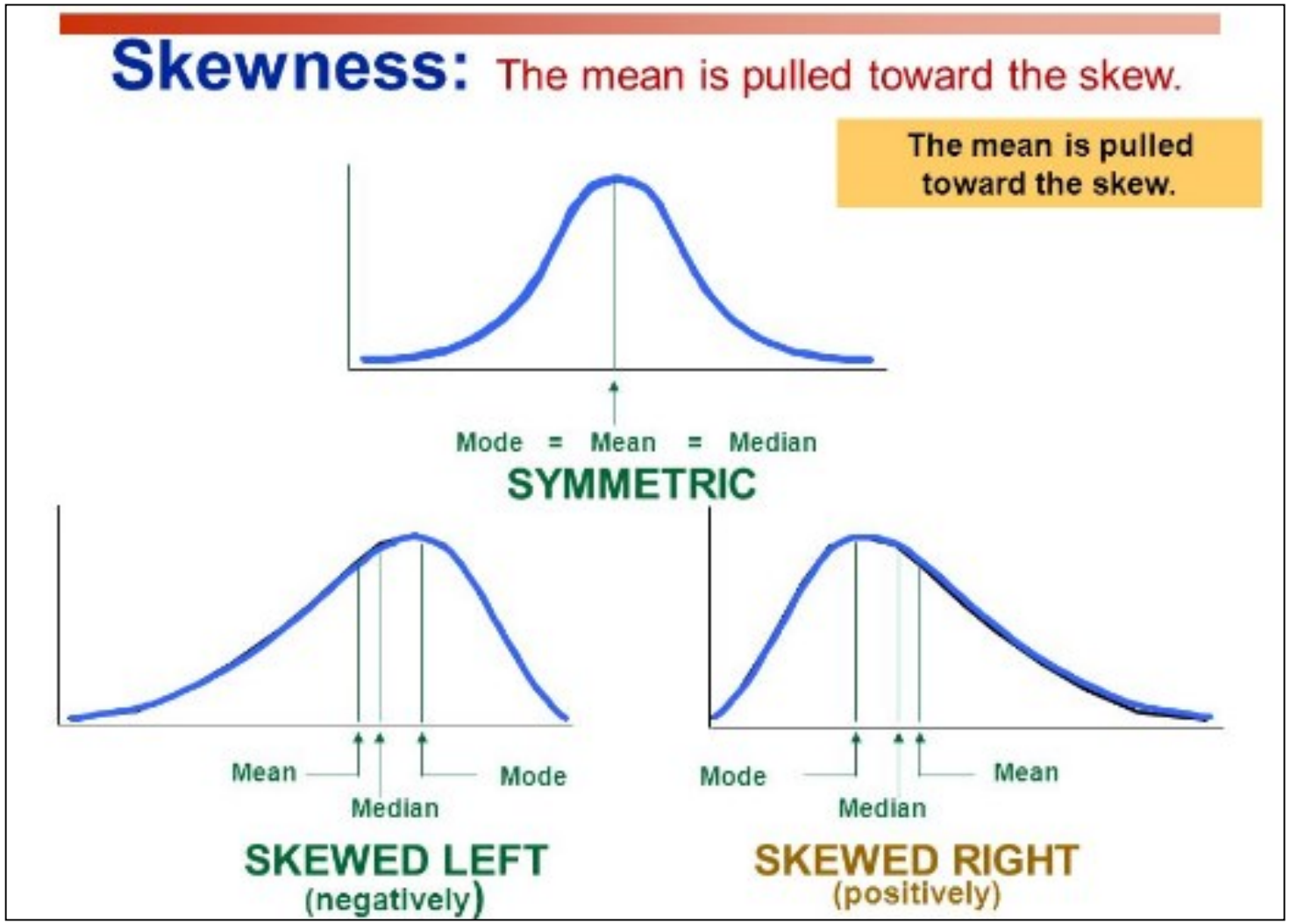
	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
1	1	0	19	182	2	0	0	0	1	0	2523
2	2	0	33	155	3	0	0	0	0	3	2551
3	3	0	20	105	1	1	0	0	0	1	2557
4	4	0	21	108	1	1	0	0	1	2	2594
5	5	0	18	107	1	1	0	0	1	0	2600
6	6	0	21	124	3	0	0	0	0	0	2622
7	7	0	22	118	1	0	0	0	0	1	2637
8	8	0	17	103	3	0	0	0	0	1	2637
9	9	0	29	123	1	1	0	0	0	1	2663

การวิเคราะห์ข้อมูล 1 ตัวแปร



การวิเคราะห์ข้อมูล 1 ตัวแปร: ข้อมูลเชิงปริมาณ

การวัดค่ากลางของข้อมูล



ข้อมูลเชิงปริมาณ

การวัดค่ากลางของข้อมูล ต่อ

- ตัวอย่างที่ 1 ผลการวิเคราะห์สถิติเบื้องต้นของชุดข้อมูล “birthwt” แสดงดัง ตารางที่ 2

ตารางที่ 2 แสดงค่าสถิติเบื้องต้นของข้อมูล

Id	Variables	Min	Q1	Median	Mean	Q3	Max
1	age	14	19	23	23.4	26	45
2	lwt	80	110	121	129.8	140	250
3	ptl	0	0	0	0.19	0	3
4	bwt	709	2,414	2,977	2,945	3,487	4,990

- ผลการวิเคราะห์สำหรับตัวแปร bwt พบว่า ค่าเฉลี่ยของน้ำหนักแรกคลอดของเด็กทารก คือ 2,945 กรัม โดยมีค่าน้ำหนักต่ำสุดและสูงสุด คือ 709 และ 4,990 กรัม ตามลำดับ

ข้อมูลเชิงปริมาณ

การวัดการกระจายของข้อมูล

การวัดการกระจาย	**ความหมาย**	**สูตร**
พิสัย (Range)	ค่าสูงสุดลบด้วยค่าต่ำสุดของข้อมูล	Max-Min
พิสัยควอร์ไทล์ (Interquartile range: IQR)	ความแตกต่างระหว่าง Q3 และ Q1	Q3-Q1
ส่วนเบี่ยงเบนมาตรฐาน (Standard deviation: SD)	ค่าของการเบี่ยงเบนไปจากค่าเฉลี่ย	$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

ข้อมูลเชิงปริมาณ

การวัดการกระจายของข้อมูล ต่อ

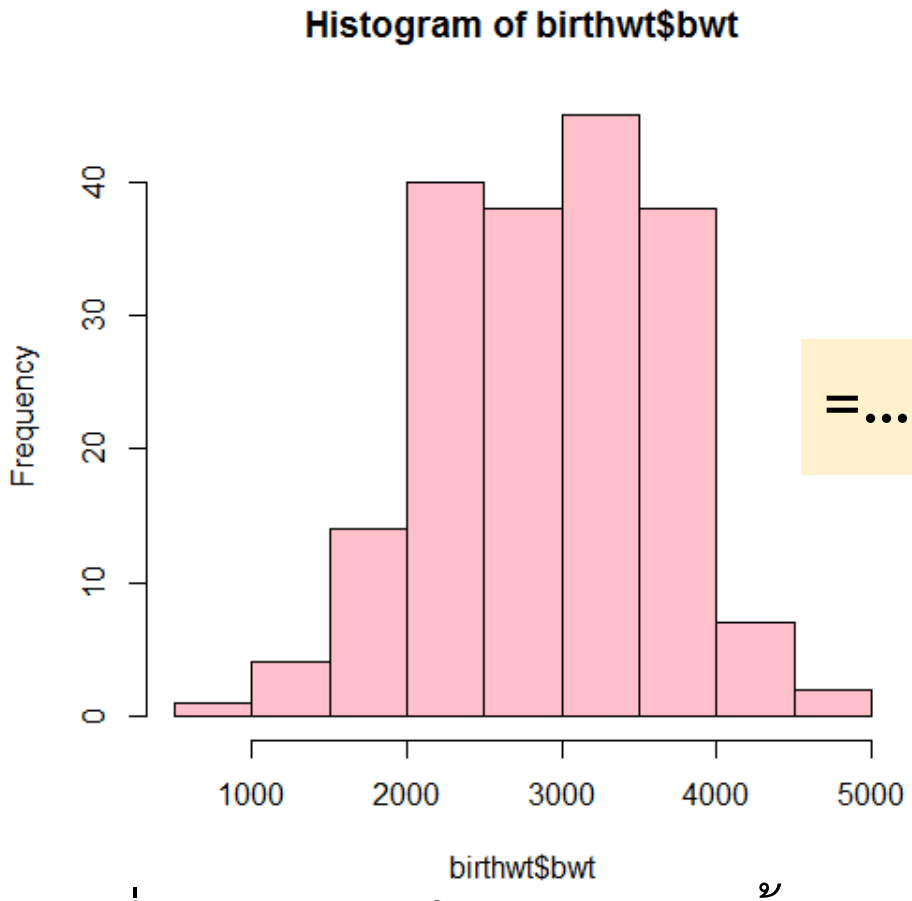
- ตัวอย่างที่ 3 จากผลการวิเคราะห์สถิติเบื้องต้นของ ตัวอย่างที่ 1 จงหาค่าการกระจายของตัวแปรต่อไปนี้

Id	Variables	Range	Interquartile range (IQR)	Standard deviation (SD)
1	age			5.3
2	lwt			30.58
3	ptl			0.49
4	bwt			0.24

ข้อมูลเชิงปริมาณ

การนำเสนอข้อมูลด้วยกราฟของตัวแปรเดียว

- ตัวอย่างที่ 4 การนำเสนอข้อมูลด้วยกราฟฮิสโตแกรม ของตัวแปร bwt



สถิติเบื้องต้น

=.....?

- จากกราฟ พบว่า

รูปที่ 1 กราฟฮิสโตแกรมของน้ำหนักแรกคลอดของทารก

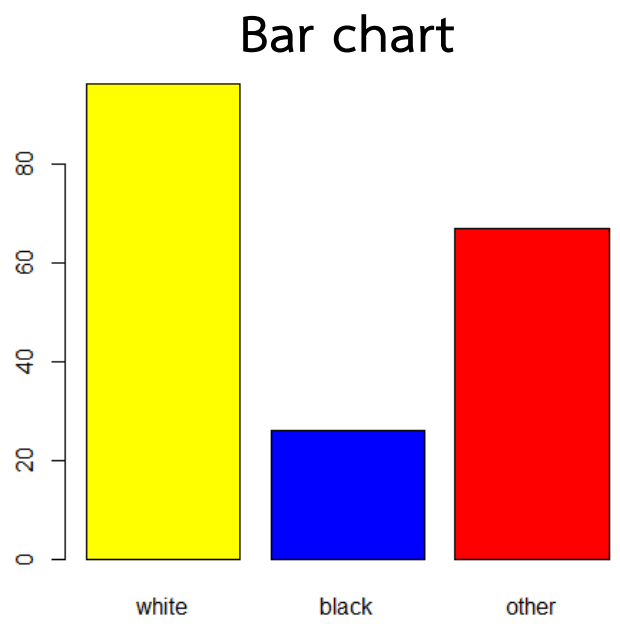
การวิเคราะห์ข้อมูล 1 ตัวแปร: ข้อมูลเชิงคุณภาพ

การนำเสนอข้อมูลด้วยตารางแจกแจงความถี่เดียวแสดงค่าสถิติร้อยละ

- การสรุปค่าทางสถิติของข้อมูลเชิงคุณภาพ คือ การนับจำนวนความถี่ของแต่ละตัวเลือกของตัวแปรกลุ่ม แล้วนำมาแสดงในรูปของค่าร้อยละ
- ตัวอย่างที่ 5 การสรุปค่าทางสถิติเบื้องต้นของข้อมูลกลุ่มแสดงดัง ตารางที่ 3

ตารางที่ 5 เชื้อชาติของกลุ่มตัวอย่างมารดาที่คลอดบุตร

เชื้อชาติ	จำนวน	ร้อยละ
ผิวขาว	96	
ผิวดำ	26	
ผิวอื่นๆ	27	

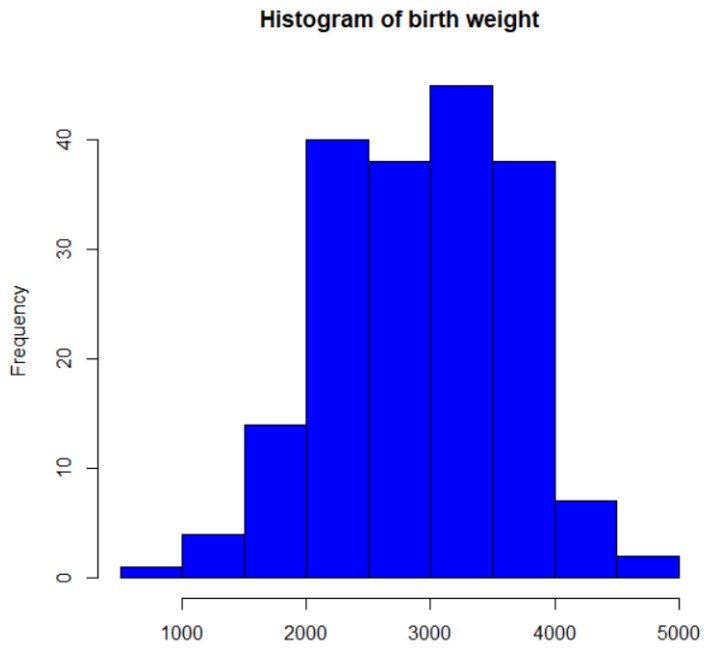


การแสดงผลการวิเคราะห์ข้อมูล 1 ตัวแปร (ตัวแปรตาม)

แสดงการแจกแจงตัวแปรตาม (ต่อเนื่อง)

```
> summ(wt$bwt)
 obs. mean     median  s.d.    min.   max.
 189  2944.587 2977    729.214 709    4990
```

```
> hist(wt$bwt, col="blue",main="Histogram of birth weight")
```



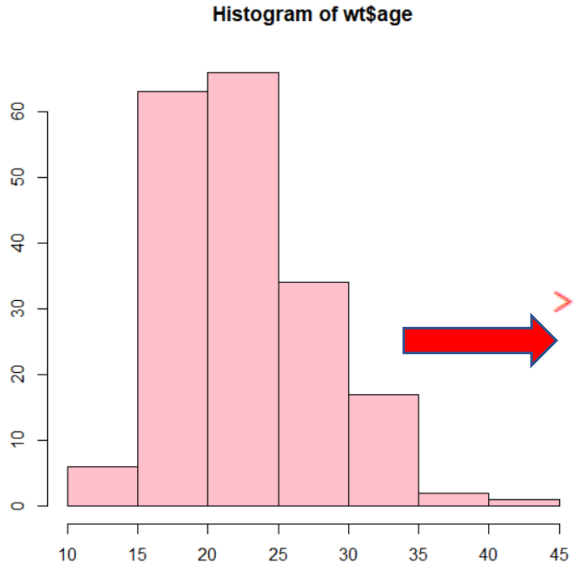

```
> shapiro.test(wt$bwt) # p-value >0.05

        Shapiro-Wilk normality test

data:  wt$bwt
W = 0.99244, p-value = 0.4353
```

- ตัวแปรตามมีการแจกแจงแบบปกติหรือไม่?
- .....
  .....

การแสดงผลการวิเคราะห์ข้อมูล 1 ตัวแปร (ตัวแปรอิสระ)

แสดงการแจกแจงตัวแปรอิสระ age

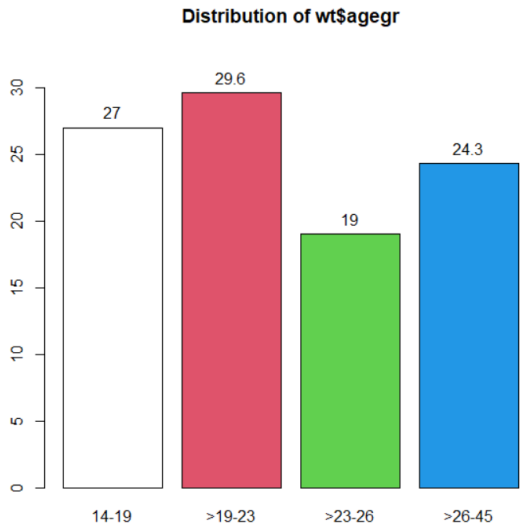
```
> summ(wt$age)
 obs. mean   median  s.d.   min.   max.
 189  23.238 23      5.299  14     45
> hist(wt$age, col="pink")
> shapiro.test(wt$age)

        Shapiro-Wilk normality test

data:  wt$age
W = 0.95977, p-value = 3.189e-05
```




```
> quantile(wt$age) # quantile
0%  25%  50%  75% 100%
14   19   23   26   45
```

```
> wt$agegr <- cut(wt$age,c(13,19,23,26,45),label=c("14-19",">19-23",">23-26",">26-45"))
> tab1(wt$agegr,bar.value="percent")
```

```
 wt$agegr :
         Frequency Percent Cum. percent
14-19           51    27.0          27.0
>19-23          56    29.6          56.6
>23-26          36    19.0          75.7
>26-45          46    24.3         100.0
  Total        189   100.0         100.0
```

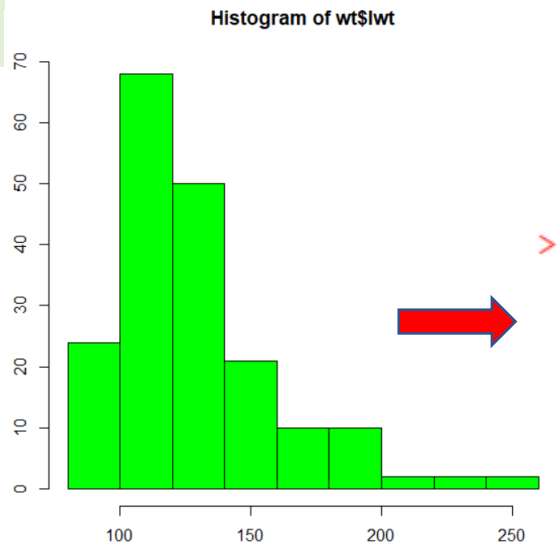


- แม่ของทารกแรกเกิดส่วนใหญ่มีอายุมากกว่า 19-23 ปี รองลงมาคือ อายุ14-19 ปี คิดเป็นร้อยละ 29.6 และ 27.0 ตามลำดับ

การแสดงผลการวิเคราะห์ข้อมูล 1 ตัวแปร (ตัวแปรอิสระ)

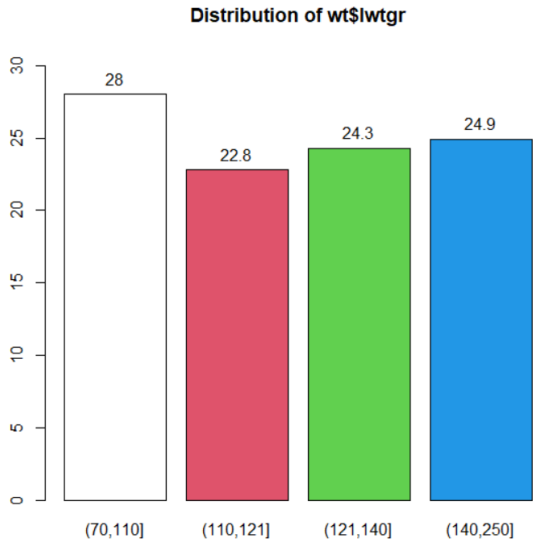
แสดงการแจกแจงตัวแปรอิสระ lwt

```
> summ(wt$lwt)
 obs. mean    median  s.d.    min.    max.
 189  129.815 121     30.579 80      250
> hist(wt$lwt, col="green")
> shapiro.test(wt$lwt)

        Shapiro-Wilk normality test

data:  wt$lwt
W = 0.89331, p-value = 2.242e-10
```




```
> quantile(wt$lwt) # 0%
  0%  25%  50%  75% 100%
  80  110  121  140  250
```

```
> wt$lwtgr <- cut(wt$lwt, c(70,110,121,140,250))
> tabl(wt$lwtgr,bar.value="percent")
wt$lwtgr :
          Frequency Percent Cum. percent
(70,110]         53    28.0          28.0
(110,121]        43    22.8          50.8
(121,140]        46    24.3          75.1
(140,250]        47    24.9         100.0
  Total         189   100.0         100.0
```

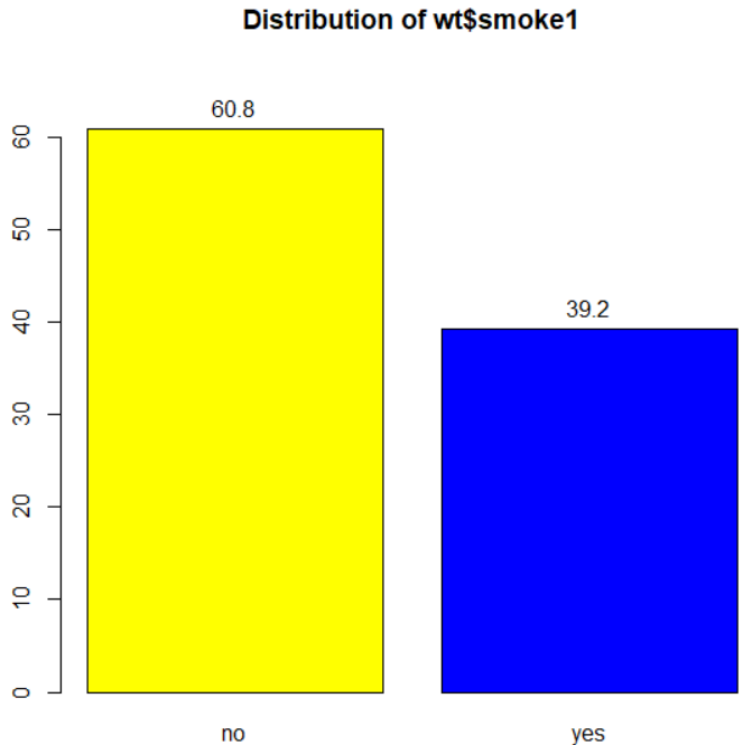


อธิบายผล

การแสดงผลการวิเคราะห์ข้อมูล 1 ตัวแปร (ตัวแปรอิสระ)

แสดงการแจกแจงตัวแปรอิสระ smoke

```
> wt$smoke1 <- factor(wt$smoke)
> levels(wt$smoke1) <- c("no","yes")
> tab1(wt$smoke1, col=c("yellow","blue"),bar.value="percent")
wt$smoke1 :
        Frequency Percent Cum. percent
no            115    60.8         60.8
yes            74    39.2        100.0
  Total       189   100.0        100.0
```



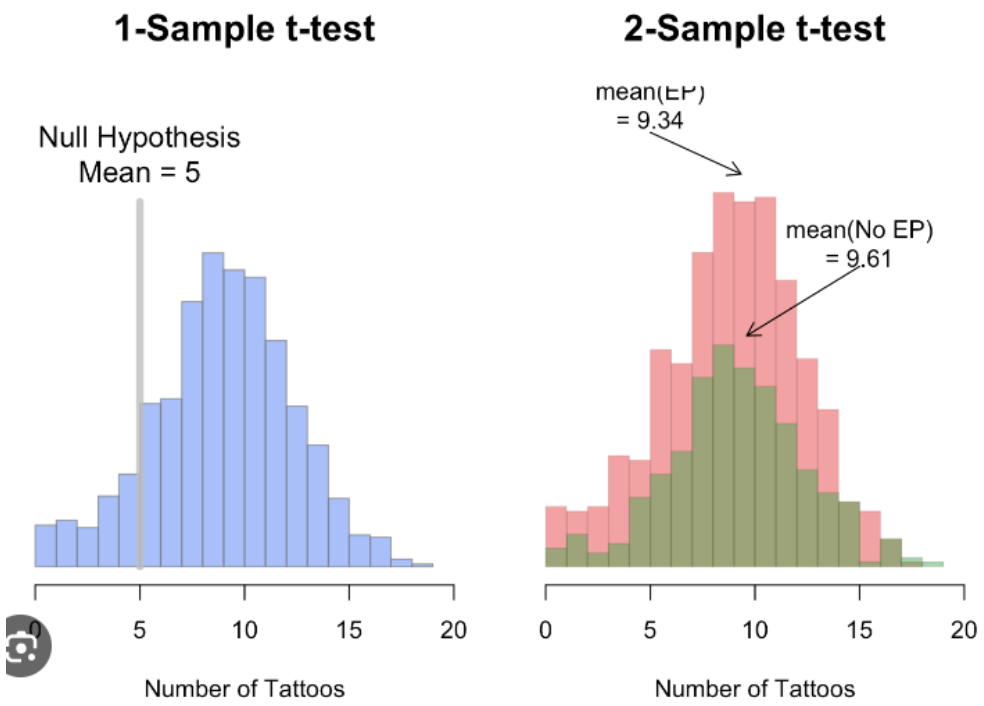
- แม่ของทารกแรกเกิดส่วนใหญ่ไม่สูบบุหรี่ คิดเป็นร้อยละ 60.8 ส่วนแม่ที่สูบบุหรี่ คิดเป็นร้อยละ 39.2

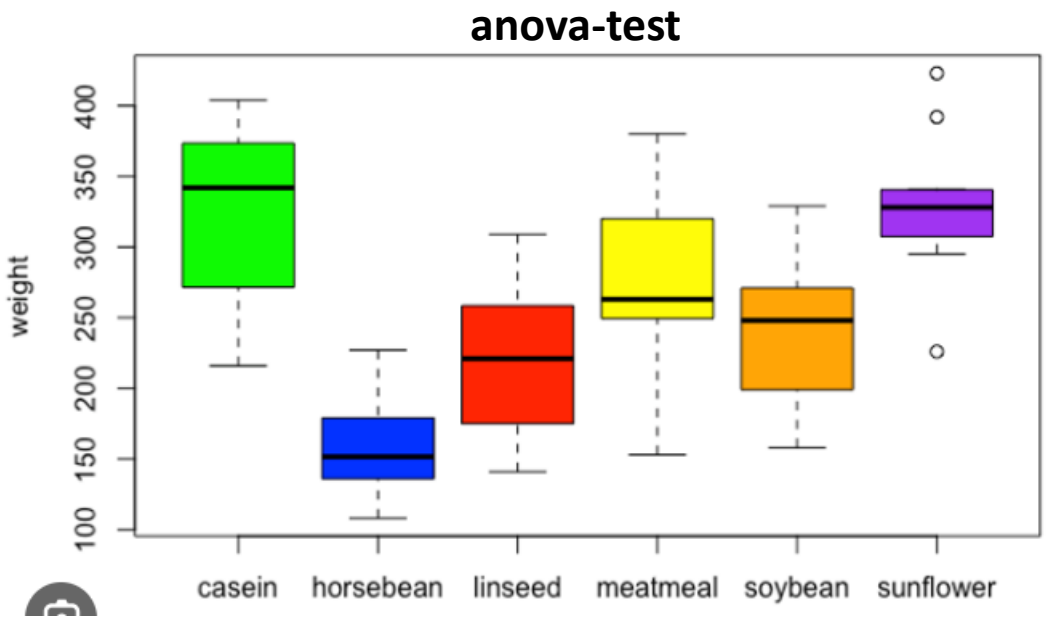
การแสดงผลการวิเคราะห์ข้อมูล 1 ตัวแปร

แสดงการแจกแจงตัวแปรอิสระที่เหลือเป็นอย่างไร???

ตัวแปรอิสระ	ประเภท	การจัดการตัวแปร	สถานะตัวแปรสุดท้าย	สถิติที่ใช้	กราฟ
ptl					
ftv					
race					
ht					
ui					

การวิเคราะห์ข้อมูลและการสรุปข้อมูล 2 ตัวแปร

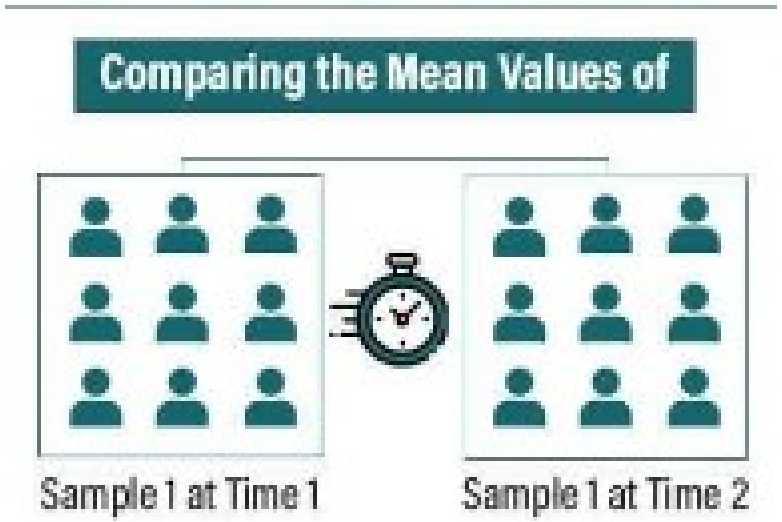




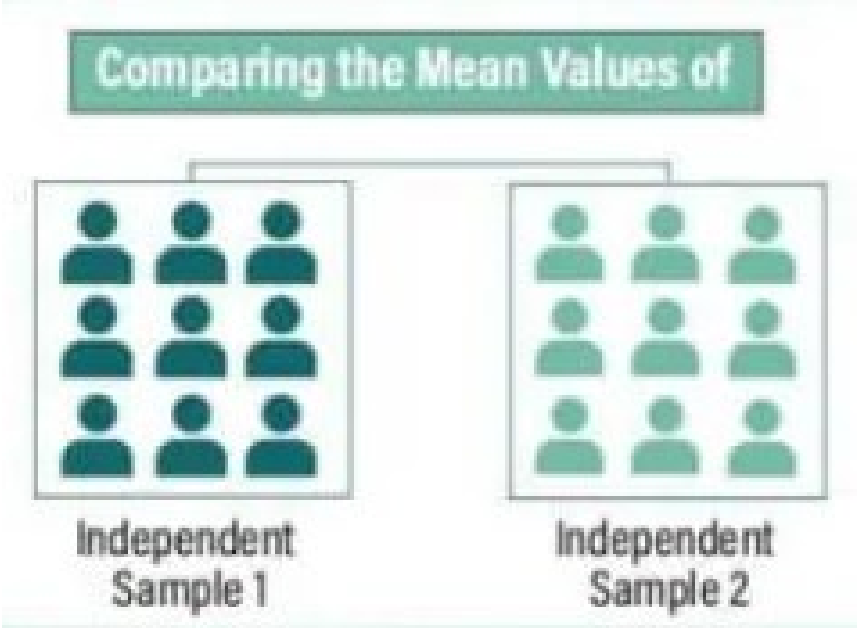
การวิเคราะห์ข้อมูลและการสรุปข้อมูล 2 ตัวแปร

Two sample t-test

Paired t-test



Independent two-sample t-test



Paired t-test and paired data

รูปแบบ Research question

“Does the population mean change after the subjects have been given some treatment”

“อัตราการเต้นของชีพจรเปลี่ยนแปลงหรือไม่หลังจากออกกำลังกาย”

•สถิติสำหรับการทดสอบสมมติฐานนี้ คือ t-test เพื่อทดสอบผลต่างของข้อมูล นั่นคือ $d = y_1 - y_2$

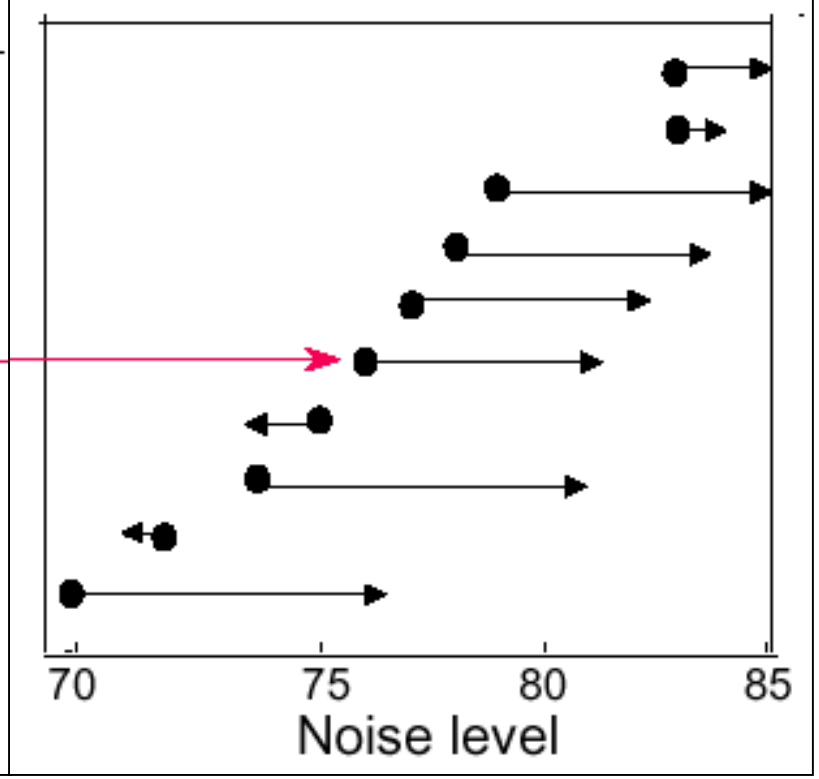
** เมื่อ y คือตัวแปรที่เราสนใจศึกษา

ตัวอย่าง: Pair data

Research Question: Did the noise level change in suburbs around Sydney Airport after the new runway was built?

	before	*after*	*increase*
Peterspam	83	85	2
Leichsoft	83	84	1
Askineville	79	85	6
Magnelle	78	84	6
St Joes	77	83	6
Balminor	76	82	6
Drummin	75	74	-1
Enless	74	81	7
Hunters Dale	72	71	-1
Annanhill	70	77	7

Noise (dBAs) from 10 suburbs



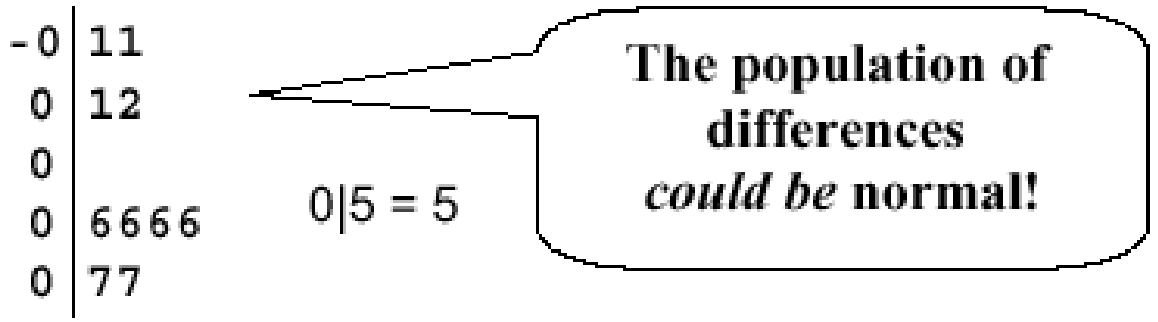
ตัวอย่าง:(ต่อ)

Preliminary Exploration:

Target population:หมู่บ้านรอบ ๆ สนามบินซิดนีย์

Sample: ตัวอย่างสุ่ม และสามารถเป็นตัวแทนประชากรได้

Variable of interest: ผลต่างระดับเสียง (d)



Paired t.test

Numerical Summary:

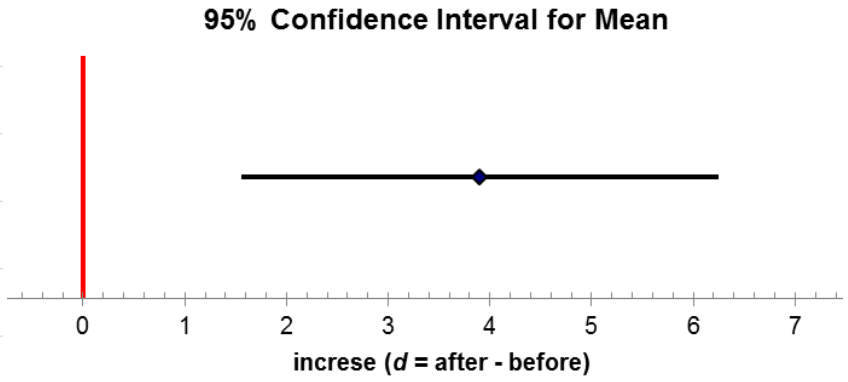
$$n = 10, \bar{d} = 3.9, s_d = 3.2812$$

	before	after	increse (d=after-before)
Peterspam	83	85	2
Leichsoft	83	84	1
Askineville	79	85	6
Magnelle	78	84	6
St. Joes	77	83	6
Balminor	76	82	6
Drummin	75	74	-1
Enless	74	81	7
Hunters Dale	72	71	-1
Annanhill	70	77	7

$$t_{0.05,9} = 2.262$$

$$\bar{d} \pm \left(t_{\alpha,\nu} s / \sqrt{n} \right)$$

$$3.9 \pm \left(2.262 \times 3.2812 / \sqrt{10} \right)$$



increse (d = after-before)

Variable	Size	Mean	StDev	StErr	95% C.I.	
increse (d = after-before)	10	3.9000	3.2813	1.0376	1.5527	6.2473

Hypothesis Testing

H H_0: $\mu_d = 0$ นั่นคือ

A จากกราฟ stem & leaf เราคาดว่าผลต่างมาจากประชากรผลต่างที่มีการแจกแจงแบบปกติ

T ใช้ t-test ในการทดสอบ

$$t = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}} = \frac{3.90 - 0}{3.28 / \sqrt{10}} = \frac{3.90}{1.0376} = 3.76$$

t-test on population mean

p-value	*t-value*	μ_0
0.0045	3.7586	0.0000

ค่า ν = 10-1 = 9

P 0.002 < p-value < 0.005, ค่า p-value < 0.05 และค่า 0 อยู่ในช่วงความเชื่อมั่น เราปฏิเสธ H_0

C สรุปได้ว่าหลังจากสร้าง runway ใหม่ ค่าเฉลี่ยของระดับเสียงเพิ่มขึ้นอยู่ระหว่าง 1.55 และ 6.25 dBAs

Example for Paired t-test by using R

แฟ้มนี้ประกอบด้วยข้อมูลที่เกี่ยวกับการเต้นของชีพจรของเด็กนักเรียนซึ่งประกอบด้วยตัวแปรดังตาราง

“pulseRan.csv”

Variables	Description
ID	ลำดับที่ของนักเรียน
PULSE1_R	จำนวนครั้งการเต้นของชีพจรก่อนวิ่งอยู่กับที่
PULSE2_R	จำนวนครั้งการเต้นของชีพจรหลังจากวิ่งอยู่กับที่
PULSE1_NR	จำนวนครั้งการเต้นของชีพจรครั้งแรกของกลุ่มที่ไม่วิ่ง
PULSE2_NR	จำนวนครั้งการเต้นของชีพจรครั้งที่สองของกลุ่มไม่วิ่ง

ถ้าเราต้องการทราบว่าการวิ่งอยู่กับที่ทำให้การเต้นของชีพจรเปลี่ยนแปลงหรือไม่**?**

Example for Paired t-test by using R

ผลการวิเคราะห์ข้อมูล

```
# Open library epiDisplay
library(epiDisplay)
# Set working directory
setwd("D:\\2566-1\\747-341\\data & command")
pul <- read.csv("pulseRan.csv") # read data into R
dir()
#descriptive
des(pul)
summ(pul)
View(pul)
```

Data structure

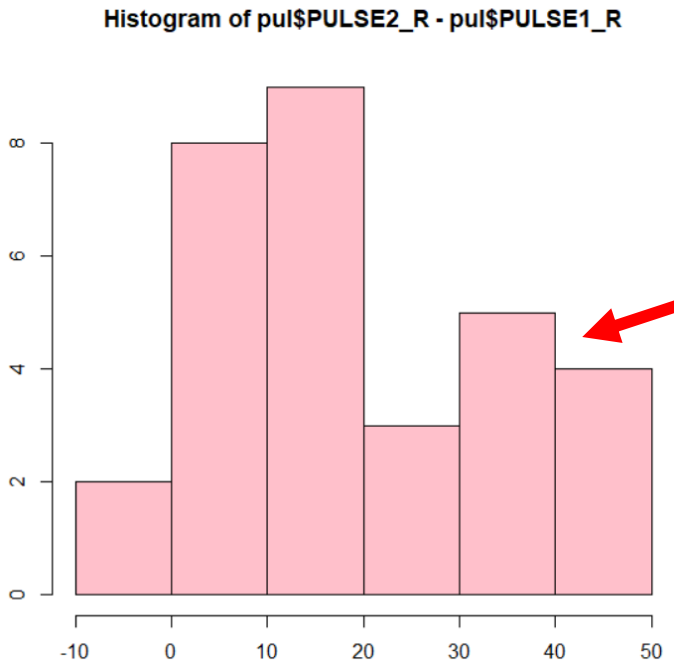
	ID	PULSE1_R	PULSE2_R	PULSE1_NR	PULSE2_NR
1	1	100	115	90	88
2	2	64	88	66	72
3	3	68	72	72	68
4	4	74	84	62	66
5	5	74	76	54	56
6	6	88	110	72	74
7	7	66	82	80	74

Example for Paired t-test by using R

ผลการวิเคราะห์ข้อมูล

```
#paired t-test
mdiff <- pul$PULSE2_R-pul$PULSE1_R # mean difference
summ(mdiff)
shapiro.test(mdiff)
```

```
hist(mdiff,col="pink")
```




```
        Shapiro-Wilk normality test

data:  mdiff
W = 0.93505, p-value = 0.06025
```

Example for Paired t-test by using R

ผลการวิเคราะห์ข้อมูล

```
# to test mean difference based on paired t.test
t.test(pul$PULSE2_R,pul$PULSE1_R,paired=TRUE,var.equal=TRUE)
```

```
        Paired t-test

data:  pul$PULSE2_R and pul$PULSE1_R
t = 6.8917, df = 30, p-value = 1.19e-07
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 13.57385 25.00680
sample estimates:
mean difference
       19.29032
```

Hypothesis Testing

H H_0:

ข้อความ คือ

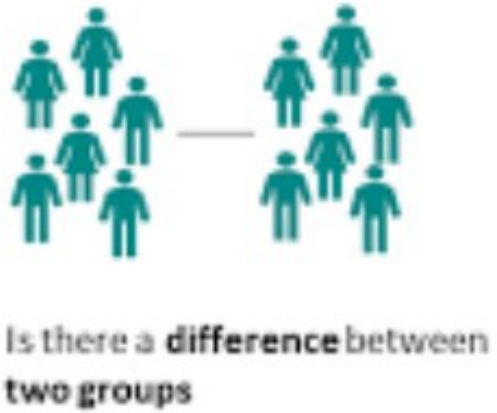
A

T

P

C

Independent two-sample t-test



A new research question?

“หลังจากสร้าง runway ใหม่แล้วระดับเสียงที่หมู่บ้าน ทางเหนือของสนามบินและทางใต้ของสนามบินซิดนีย์ มีความแตกต่างกัน หรือไม่”

ข้อมูล: วัดระดับเสียง (dbAs) จากการสุ่มตัวอย่างหมู่บ้านทางเหนือของ สนามบินซิดนีย์ 12 หมู่บ้าน และทางใต้ของสนามบินซิดนีย์ 10 หมู่บ้าน

Independent two-sample t-test

ตัวอย่างข้อมูล

เหนือ		ใต้	
หมู่บ้าน	ระดับเสียง	หมู่บ้าน	ระดับเสียง
Dove Lake	85	Queensford	82
Mundle	84	Ascot	83
Rebora	79	Oldtown	85
Lee Way	70	Bluefern	87
Thimble	75	Cunella	80
Walk	82	St. Georgina	85
Stepping	83	Bookwood	79
Straight Straight	74	Goldvania	81
Wommanly	76	Minty	80
Marswood	83	Atticus	81
Murraturra	76		
Pickly	72		

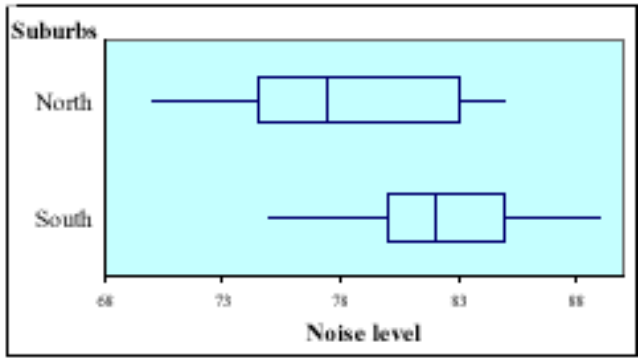
ตัวอย่าง:(ต่อ)

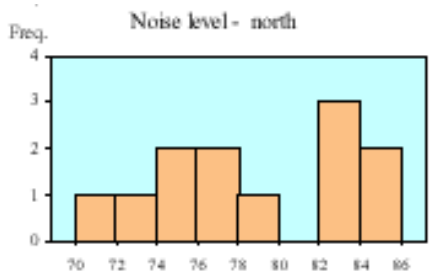
Preliminary Exploration:

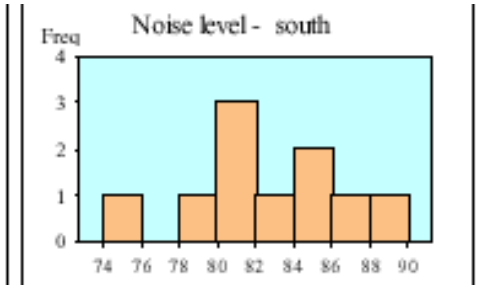
Target population: หมู่บ้านรอบ ๆ สนามบินซิดนีย์

Sample: ตัวอย่างสุ่ม 12 หมู่บ้านทางเหนือ 10 หมู่บ้านทางใต้ และสามารถเป็นตัวแทนประชากรได้

Variable of interest: ให้ y_1 แทนระดับเสียง (เหนือ)
y_2 แทนระดับเสียง (ใต้)







Numerical Summary:

Comparison:	Noise level			
Suburbs	*Size*	*Mean*	*SE*	*StDev*
North	12	78.250	1.467	5.083
South	10	82.400	1.327	4.195

ตัวอย่าง: (ต่อ) Confidence interval

เราจะหาช่วงความเชื่อมั่นสำหรับผลต่างของค่าเฉลี่ยของประชากร สองกลุ่ม

นั่นคือ ช่วงความเชื่อมั่นของ $\mu_1 - \mu_2$

<u>ข้อตกลง</u>

- กลุ่มตัวอย่างทั้งสองเป็นอิสระต่อกันและมาจากประชากรที่มีการแจกแจงแบบปกติ
- ประชากรทั้งสองกลุ่มมีค่าส่วนเบี่ยงเบนมาตรฐานเหมือนกัน $(\sigma_1 = \sigma_2)$

Confidence interval for differences

ช่วงความเชื่อมั่นสำหรับผลต่าง $\mu_1 - \mu_2$ คำนวณได้จาก

$$(\bar{y}_1 - \bar{y}_2) \pm t_{crit} \times se_{(\bar{y}_1 - \bar{y}_2)}$$

$$se_{(\bar{y}_1 - \bar{y}_2)} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

S_p คือ ค่าส่วนเบี่ยงเบนมาตรฐานร่วมของสองกลุ่ม (Pooled standard deviation)

ตัวอย่าง: (ต่อ)

คำนวณหาช่วงความเชื่อมั่น 95% สำหรับผลต่างของค่าเฉลี่ยของ กลุ่มประชากร (ตัวอย่างระดับเสียง ณ สนามบิน) จากข้อมูลการวิเคราะห์เบื้องต้น

$$\bar{y}_1 - \bar{y}_2 = (78.25 - 82.40) = -4.15$$

t_{crit} มีค่า df= $n_1 + n_2 - 2$, นั่นคือ df = 20; เปิดตาราง t ได้ค่า t_{crit} = 2.086 ; s_p = 4.704

95% Confidence interval

$$= (78.25 - 82.40) \pm \left[2.086 \times 4.705 \sqrt{\frac{1}{12} + \frac{1}{10}} \right]$$

$$= (-8.350,\ 0.052)$$

Hypothesis testing

ค่าเฉลี่ยของสองกลุ่ม
ไม่มีความแตกต่างกัน

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

A จาก histogram และ box plot; ข้อมูลตัวอย่างมาจากประชากรที่มีการแจกแจงแบบปกติ และทั้งสองกลุ่มมีค่าส่วนเบี่ยงเบนมาตรฐานไม่แตกต่างกัน

T $$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} ; \quad = \frac{78.25 - 82.4}{4.704\sqrt{\frac{1}{12} + \frac{1}{10}}}$$

$= -2.06$ ด้วยค่า $\nu = 12 + 10 - 2 = 20$

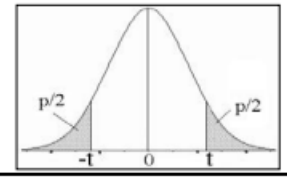
P P-value: $0.05 < \text{p-value} < 0.1$ เพราะฉะนั้นเราไม่ปฏิเสธ H_0

C ระดับเสียงโดยเฉลี่ยที่วัดได้จากหมู่บ้านทางเหนือและทางใต้ของสนามบินไม่แตกต่างกัน

Independent two-sample t-test

พื้นที่ใต้โค้งการแจกแจงที (Student's t distribution)

สะสมจาก $-\infty$ ถึง -t และ t ถึง ∞



ν	0.0005	0.001	0.002	0.005	0.01	0.02	0.05	0.1	0.2	0.5
1	1273.24	636.62	318.31	127.32	63.66	31.82	12.71	6.31	3.08	1.00
2	44.70	31.60	22.33	14.09	9.92	6.96	4.30	2.92	1.89	0.82
3	16.33	12.92	10.21	7.45	5.84	4.54	3.18	2.35	1.64	0.76
4	10.31	8.61	7.17	5.60	4.60	3.75	2.78	2.13	1.53	0.74
5	7.976	6.869	5.893	4.773	4.032	3.365	2.571	2.015	1.476	0.727
6	6.788	5.959	5.208	4.317	3.707	3.143	2.447	1.943	1.440	0.718
7	6.082	5.408	4.785	4.029	3.499	2.998	2.365	1.895	1.415	0.711
8	5.617	5.041	4.501	3.833	3.355	2.896	2.306	1.860	1.397	0.706
9	5.291	4.781	4.297	3.690	3.250	2.821	2.262	1.833	1.383	0.703
10	5.049	4.587	4.144	3.581	3.169	2.764	2.228	1.812	1.372	0.700
11	4.863	4.437	4.025	3.497	3.106	2.718	2.201	1.796	1.363	0.697
12	4.716	4.318	3.930	3.428	3.055	2.681	2.179	1.782	1.356	0.695
13	4.597	4.221	3.852	3.372	3.012	2.650	2.160	1.771	1.350	0.694
14	4.499	4.140	3.787	3.326	2.977	2.624	2.145	1.761	1.345	0.692
15	4.417	4.073	3.733	3.286	2.947	2.602	2.131	1.753	1.341	0.691
16	4.346	4.015	3.686	3.252	2.921	2.583	2.120	1.746	1.337	0.690
17	4.286	3.965	3.646	3.222	2.898	2.567	2.110	1.740	1.333	0.689
18	4.233	3.922	3.610	3.197	2.878	2.552	2.101	1.734	1.330	0.688
19	4.187	3.883	3.579	3.174	2.861	2.539	2.093	1.729	1.328	0.688
20	4.146	3.850	3.552	3.153	2.845	2.528	2.086	1.725	1.325	0.687
21	4.110	3.819	3.527	3.135	2.831	2.518	2.080	1.721	1.323	0.686
22	4.077	3.792	3.505	3.119	2.819	2.508	2.074	1.717	1.321	0.686
23	4.047	3.768	3.485	3.104	2.807	2.500	2.069	1.714	1.319	0.685
24	4.021	3.745	3.467	3.091	2.797	2.492	2.064	1.711	1.318	0.685
25	3.996	3.725	3.450	3.078	2.787	2.485	2.060	1.708	1.316	0.684
26	3.974	3.707	3.435	3.067	2.779	2.479	2.056	1.706	1.315	0.684
27	3.954	3.690	3.421	3.057	2.771	2.473	2.052	1.703	1.314	0.684
28	3.935	3.674	3.408	3.047	2.763	2.467	2.048	1.701	1.313	0.683
29	3.918	3.659	3.396	3.038	2.756	2.462	2.045	1.699	1.311	0.683
30	3.902	3.646	3.385	3.030	2.750	2.457	2.042	1.697	1.310	0.683
35	3.836	3.591	3.340	2.996	2.724	2.438	2.030	1.690	1.306	0.682
40	3.788	3.551	3.307	2.971	2.704	2.423	2.021	1.684	1.303	0.681
45	3.752	3.520	3.281	2.952	2.690	2.412	2.014	1.679	1.301	0.680
50	3.723	3.496	3.261	2.937	2.678	2.403	2.009	1.676	1.299	0.679
60	3.681	3.460	3.232	2.915	2.660	2.390	2.000	1.671	1.296	0.679

Example: การวิเคราะห์ข้อมูล 2 ตัวแปร by using independent two-sample t.test

ตัวแปรอิสระ

age
lwt
race
smoke
ptl
ht
ui
ftv

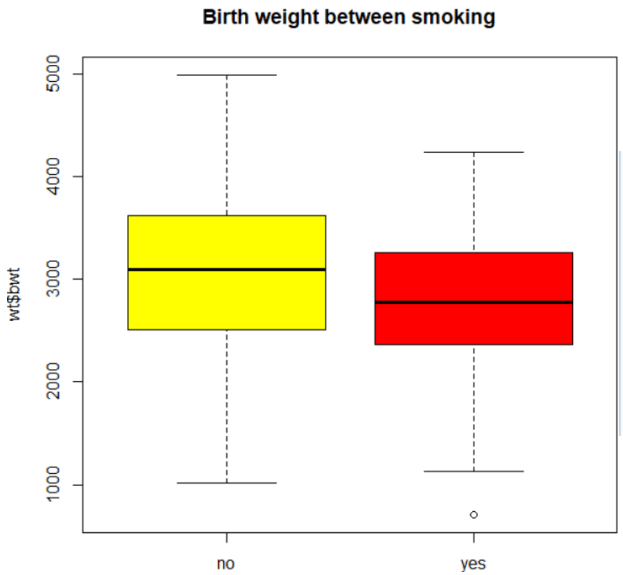
ตัวแปร

ตัวแปรตาม

bwt

ความสัมพันธ์ระหว่างน้ำหนักของทารกแรกเกิดกับการสูบบุหรี่

```
> boxplot(wt$bwt~wt$smoke1, col=c("yellow","red"),main="Birth weight between smoking")
```




```
> summ(wt$bwt,by=wt$smoke1)
For wt$smoke1 = no
 obs. mean     median  s.d.    min.  max.
 115  3055.696 3100    752.657 1021  4990

For wt$smoke1 = yes
 obs. mean     median  s.d.    min.  max.
 74   2771.919 2775.5  659.635 709   4238
```

```
> t.test(wt$bwt~wt$smoke1, var.equal=T)

        Two Sample t-test

data:  wt$bwt by wt$smoke1
t = 2.6529, df = 187, p-value = 0.008667
alternative hypothesis: true difference
95 percent confidence interval:
  72.75612 494.79735
sample estimates:
 mean in group no mean in group yes
         3055.696          2771.919
```

H $H_0 : \mu_1 = \mu_2$

ค่าเฉลี่ยของน้ำหนักของกลุ่มที่แม่สูบและไม่สูบบุหรี่ไม่แตกต่างกัน

A ข้อมูลตัวอย่างมาจากประชากรที่มีการแจกแจงแบบปกติ และทั้งสองกลุ่มมีค่าส่วนเบี่ยงเบนมาตรฐานไม่แตกต่างกัน

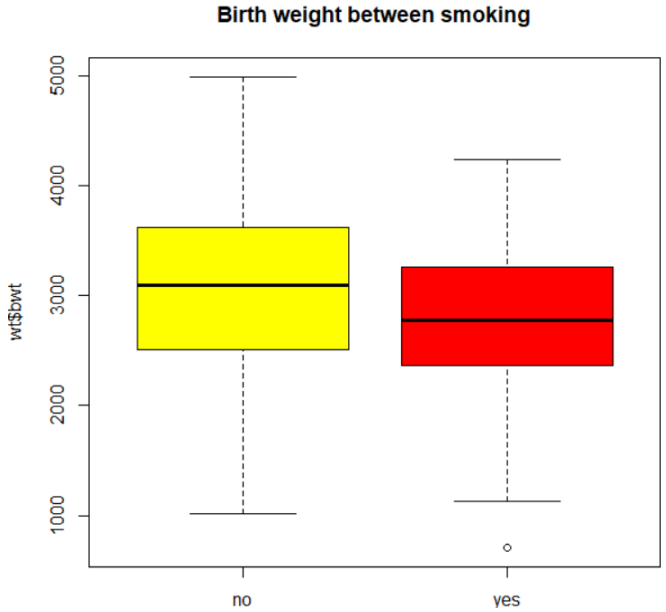
T T=2.65, df=187

P P=0.009; reject H_0

C ค่าเฉลี่ยของน้ำหนักของทารกแรกคลอดของกลุ่มที่แม่สูบและไม่สูบบุหรี่แตกต่างกัน

ความสัมพันธ์ระหว่างน้ำหนักของทารกแรกเกิดกับการสูบบุหรี่

```
> boxplot(wt$bwt~wt$smoke1, col=c("yellow","red"),main="Birth weight between smoking")
```




```
> t.test(wt$bwt~wt$smoke1, var.equal=T)

        Two Sample t-test

data:  wt$bwt by wt$smoke1
t = 2.6529, df = 187, p-value = 0.008667
alternative hypothesis: true difference
95 percent confidence interval:
  72.75612 494.79735
sample estimates:
 mean in group no mean in group yes
         3055.696          2771.919
```

H $H_0 : \mu_1 = \mu_2$

A

T

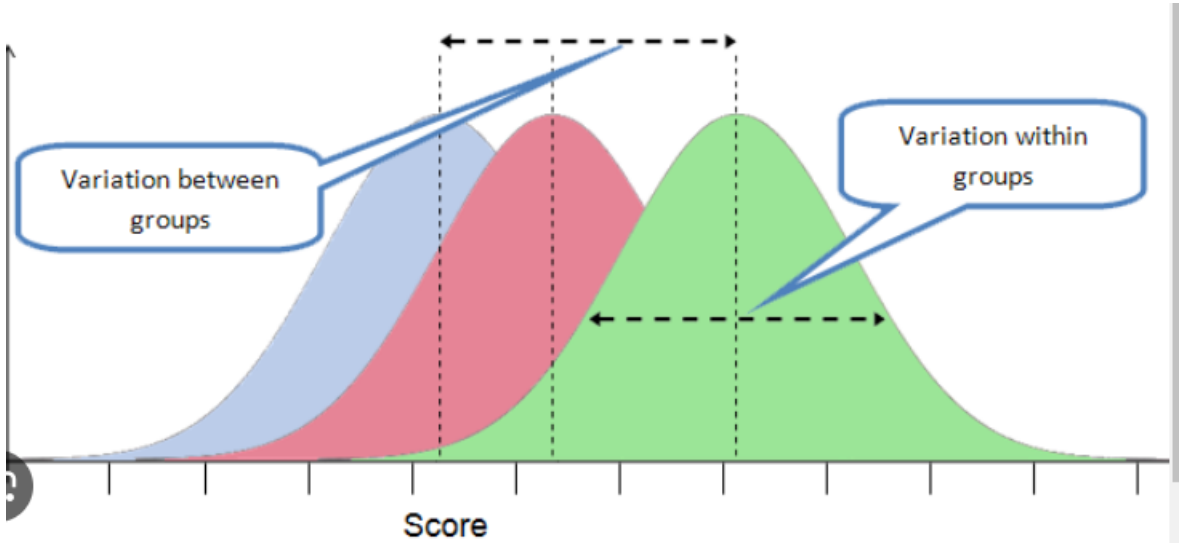
P

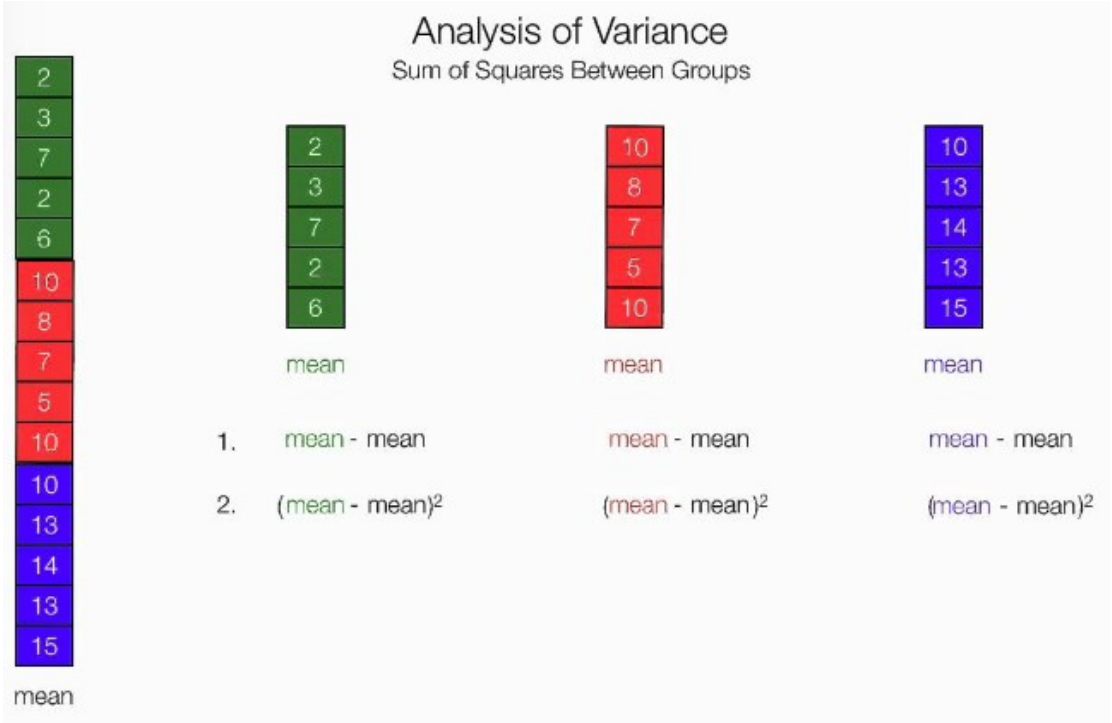
C

Comparison of more than two populations means

Analysis of Variance (ANOVA)

การวิเคราะห์ความแปรปรวน





Analysis of Variance (ANOVA)

A new research question

“หลังจากสร้าง runway ใหม่แล้วระดับเสียงที่หมู่บ้าน ทางเหนือ ทางใต้ ทางตะวันออก และทางตะวันตก ของสนามบินซิดนีย์ มีความแตกต่างกัน หรือไม่”

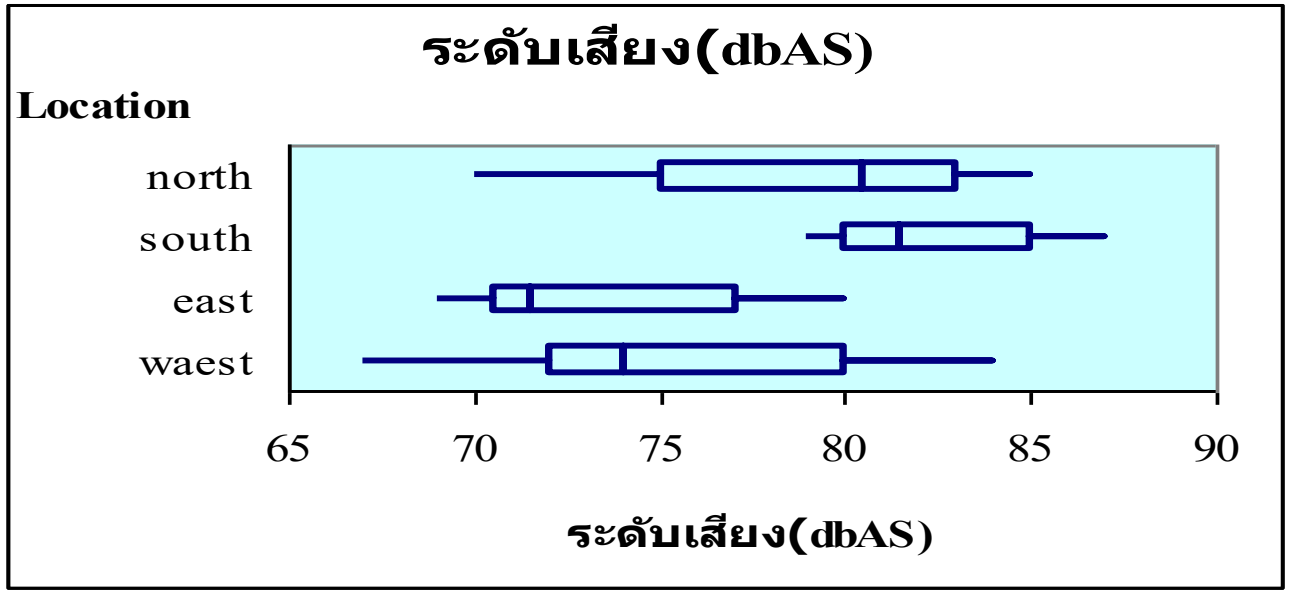
ข้อมูล: วัดระดับเสียง (dbAs) จากการสุ่มตัวอย่างหมู่บ้านทางเหนือของ สนามบินซิดนีย์ 10 หมู่บ้าน ทางใต้ของสนามบินซิดนีย์ 10 หมู่บ้าน ทางตะวันออกของสนามบินซิดนีย์ 10 หมู่บ้าน และทางตะวันตกของ สนามบินซิดนีย์ 10 หมู่บ้าน

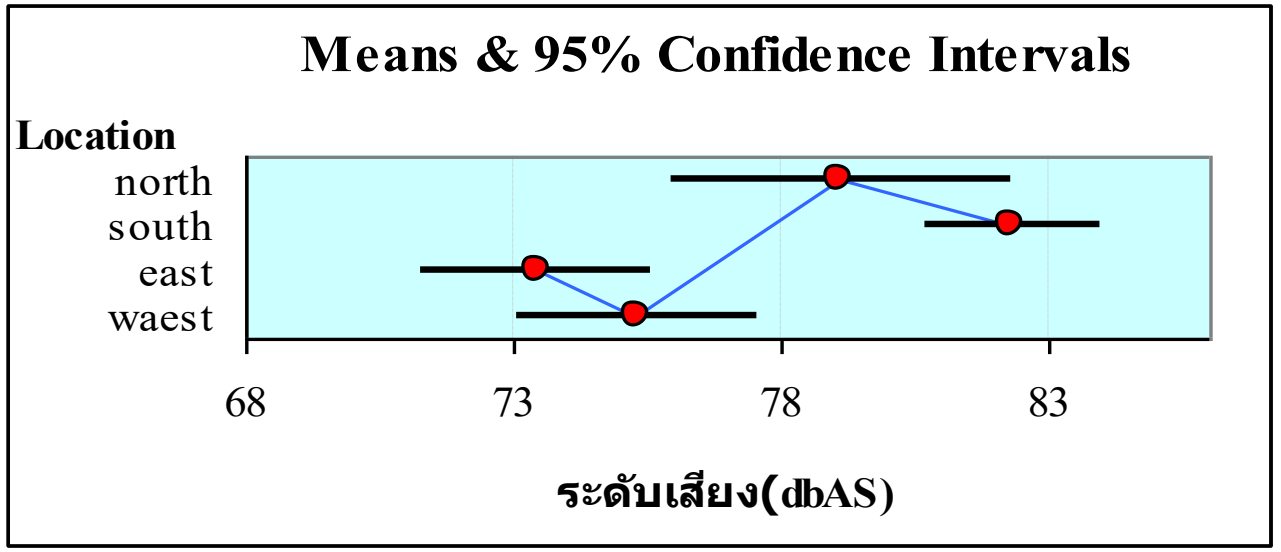
Analysis of Variance (ANOVA)

ข้อมูล

ระดับเสียง (dbAs)			
เหนือ	ใต้	ตะวันออก	ตะวันตก
85	82	76	81
84	83	78	80
79	85	75	72
70	87	72	75
75	80	80	67
82	85	71	71
83	79	69	73
74	81	78	84
76	80	71	77
83	81	70	73

Analysis of Variance (ANOVA)





Analysis of Variance (ANOVA)

ANOVA การวิเคราะห์ความแปรปรวน

- จำนวนประชากรที่ต้องการทดสอบมีมากกว่า 2 กลุ่ม
- ทดสอบเกี่ยวกับปัจจัยที่ต้องการศึกษา เรียกว่า ทรีทเมนต์ (Treatments)
- เพื่อดูว่าผลที่เกิดจากทรีทเมนต์แตกต่างกันหรือไม่โดยใช้ หลักการวิเคราะห์ความแปรปรวน

หลักการวิเคราะห์ความแปรปรวน

การแบ่งความแปรปรวนที่เกิดขึ้นทั้งหมดในการศึกษา หรือการทดลอง (Total Variation) ออกเป็น

- ความแปรปรวนที่เกิดขึ้นภายในประชากรแต่ละกลุ่มหรือ ภายในกลุ่มเดียวกัน

- ความแปรปรวนที่เกิดขึ้นเนื่องจากทรีทเมนต์ที่แตกต่างกัน หรือเป็นความแปรปรวนระหว่างกลุ่ม

ตัวอย่างปัญหา

การศึกษาวิธีการสอนภาษาอังกฤษที่แตกต่างกัน 4 วิธี โดยวิธีสอนแต่ละวิธีกับเด็กที่มีพื้นฐานเหมือน ๆ กัน เช่น ระดับสติปัญญา อายุ ฯลฯ หลังจากสอนจบ ทำการทดสอบแล้วนำคะแนนเฉลี่ยของนักเรียนด้วยวิธีสอนแต่ละวิธีมาเปรียบเทียบกัน

วิธีการสอนที่แตกต่างกัน คือ ทรีทเมนต์ ซึ่งมี 4 ทรีทเมนต์

การศึกษาในลักษณะนี้ เป็นการจำแนกข้อมูลโดยอาศัยเพียงปัจจัยเดียว เรียกว่า การจำแนกแบบทางเดียว หรือ
วิเคราะห์แบบทางเดียว (One-way ANOVA)

Analysis of Variance (ANOVA)

เป็นการวิเคราะห์ข้อมูลซึ่งเป็นผลมาจากปัจจัยที่นำมาศึกษาเพียงอย่างเดียว แต่ละปัจจัยจะมีจำนวนข้อมูลเท่ากันหรือไม่เท่ากันก็ได้ โดยแยกความแปรปรวนออกได้ดังนี้

ความแปรปรวนรวม = ความแปรปรวนที่เกิดขึ้นระหว่างทรีทเมนต์ + ความคลาดเคลื่อน

ความแปรปรวนรวม หาจากค่าเบี่ยงเบนของข้อมูลแต่ละค่าจากค่าเฉลี่ยรวม นำค่ามายกกำลังสองแล้วรวมกัน และหารด้วยค่าองศาแห่งความเป็นอิสระ ซึ่งเรียกว่าผลรวมกำลังสองทั้งหมด (Total Sum Square) ซึ่งเขียนย่อเป็น SST

Analysis of Variance (ANOVA)

ความแปรปรวนที่เกิดขึ้นระหว่างทรีทเมนต์ หรือผลรวมกำลังสองระหว่างกลุ่ม เป็นค่าที่ได้จากส่วนเบี่ยงเบนของคะแนนเฉลี่ยในแต่ละกลุ่มกับค่าเฉลี่ยรวม ซึ่งเรียกว่า Treatment Sum Square หรือเขียนย่อว่า SSTr

ความคลาดเคลื่อนหรือผลรวมกำลังสองภายในกลุ่ม เป็นค่าที่ได้จากส่วนเบี่ยงเบนของข้อมูลแต่ละตัวกับค่าเฉลี่ยของแต่ละกลุ่ม เรียกว่า Within group Sum Square หรือ Error Sum Square หรือเขียนย่อว่า SSE

ลักษณะของข้อมูล

สมมติการศึกษา k ทรีทเมนต์ โดยแต่ละทรีทเมนต์มีข้อมูล n_i

(i = 1, 2, ...k และ j = 1, 2 , ... n_i)

ทรีทเมนต์				
1	2	...	k	
x_{11}	x_{21}	...	x_{k1}	
x_{12}	x_{22}	...	x_{k2}	
.	.	.	.	
.	.	.	.	
.	.	.	.	
x_{1n}	x_{2n}	...	x_{kn}	
$T_{1.}$	$T_{2.}$	...	$T_{k.}$	$T_{..}$
$\bar{x}_{1.}$	$\bar{x}_{2.}$	...	$\bar{x}_{k.}$	$\bar{x}_{..}$

ยอดรวมแต่ละทรีทเมนต์

ค่าเฉลี่ย ($\bar{x}_{1.}$)

ลักษณะของข้อมูล (ต่อ)

y_{ij} คือ ค่าของข้อมูลหน่วยที่ j ในทรีทเมนต์ที่ i

$T_{i.}$ คือ ผลรวมของข้อมูลในทรีทเมนต์ที่ i $= \sum_{j=1}^{n_i} y_{ij}$

$\bar{y}_{i.}$ คือ ค่าเฉลี่ยของข้อมูลในแต่ละทรีทเมนต์ $= \frac{T_{i.}}{n_i}$

$T_{..}$ คือ ผลรวมของข้อมูลทั้งหมด $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$

$\bar{y}_{..}$ คือ ค่าเฉลี่ยของข้อมูลทั้งหมด $= \frac{T_{..}}{\sum_{i=1}^{k} n_i}$

การคำนวณ

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2..}{kn}$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

SST=SSTr+SSE

SSTr ในกรณีที่จำนวนข้อมูลในแต่ละทรีทเมนต์เท่ากัน

$$SSTr = n\sum_{i=1}^{n}(\bar{y}_{i.} - \bar{y}..)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} T_{i.}^2 - \frac{T^2..}{kn}$$

ในกรณีที่จำนวนข้อมูลในแต่ละทรีทเมนต์ไม่เท่ากัน

$$SSTr = \sum_{i=1}^{n_i} n_i(\bar{y}_{i.} - \bar{y}..)^2$$

$$= \sum_{i=1}^{k} \frac{T_{i.}^2}{n_i} - \frac{T^2..}{kn}$$

การคำนวณ (ต่อ)

ในการคำนวณค่าของ SST, SSTr และ SSE จะมีค่าองศาแห่งความเป็นอิสระ (degree of freedom) มาเกี่ยวข้องซึ่งคำนวณจาก

SST มีค่า degree of freedom คือ $kn - 1$

SSTr มีค่า degree of freedom คือ $k - 1$

SSE มีค่า degree of freedom คือ $k(n - 1)$

ค่าผลรวมกำลังสอง SST=SSTr+SSE

ค่า degree of freedom kn-1= (k-1) + k(n-1)

การคำนวณ (ต่อ)

นำค่า degree of freedom ไปหารผลรวมกำลังสองของแต่ละค่า จะได้ค่าต่าง ๆ ดังนี้

$\frac{SST}{kn-1}$ = ค่าเฉลี่ยกำลังสองของยอดรวม หรือ Mean square of total เขียนย่อ MST

$\frac{SSTr}{k-1}$ = ค่าเฉลี่ยกำลังสองของทรีทเมนต์ หรือ Mean square of treatment เขียนย่อ MSTr

$\frac{SSE}{k(n-1)}$ = ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน หรือ Mean square of error เขียนย่อ MSE

การคำนวณ (ต่อ)

ค่าทดสอบทางสถิติคือ

$$F = \frac{SSTr/k-1}{SSE/k(n-1)}$$

หรือ

$$= \frac{MSTr}{MSE}$$

โดยมีค่า degree of freedom เท่ากับ $df_1 = (k - 1)$ และ $df_2 = k(n - 1)$

Analysis of Variance (ANOVA)

ตารางวิเคราะห์ความแปรปรวน

SOV	df	SS	MS	F
Treatment	k-1	SSTr	MSTr	F
Error	K(n-1)	SSE	MSE	
Total	kn-1	SST		

โดยศึกษาที่ระดับนัยสำคัญ 0.05

การทดสอบสมมติฐาน

การวิเคราะห์ความแปรปรวนแบบทางเดียวมีวัตถุประสงค์เพื่อทดสอบว่าค่าเฉลี่ยที่ได้จากแต่ละทรีทเมนต์มีความแตกต่างกันหรือไม่

สมมติฐานหลักทางสถิติคือ

H H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

A ข้อตกลงของการทดสอบ

1. การแจกแจงของผลที่ได้จากแต่ละหน่วยทดลองในแต่ละทรีทเมนต์จะมีการแจกแจงแบบปกติ
2. ความแปรปรวนในแต่ละทรีทเมนต์มีค่าเท่ากันหมด

การตัดสินใจ

- P-value ?

- ตาราง F หาค่า วิกฤต

โดยการนำค่า F ที่คำนวณได้พร้อมกับค่า degree of freedom มาเทียบกับค่า F วิกฤตที่เปิดได้จากตาราง

ถ้าค่า F ที่คำนวณได้มีค่า มากกว่า ค่า F วิกฤต เราจะปฏิเสธ H_0

Percentage Point of the F distu

$F_{0.05,\ \nu_1, \nu_2}$

Denominator Degrees of Freedom	Numerator Degree								
	1	2	3	4	5	6	7	8	9
1	161.4	199.5	213.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.88	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65

การตัดสินใจ (ต่อ)

T เช่น คำนวณค่า $F = 8.47$ ด้วยค่า $df_1 = 2$ และ $df_2 = 12$

เปิดตาราง F ที่ระดับนัยสำคัญ **0.05** ด้วยค่า $df_1 = 2$ และ $df_2 = 12$

P $$F_{0.05(2,\ 12)} = 3.89$$

F คำนวณ > F วิกฤต เพราะฉะนั้น ปฏิเสธ H_0

C สรุปได้ว่า มีอย่างน้อยหนึ่งทรีทเมนต์ที่มีค่าเฉลี่ยแตกต่างจากทรีทเมนต์อื่น ๆ

Example: การวิเคราะห์ข้อมูล 2 ตัวแปร by using anova test

ตัวแปร

ตัวแปรอิสระ

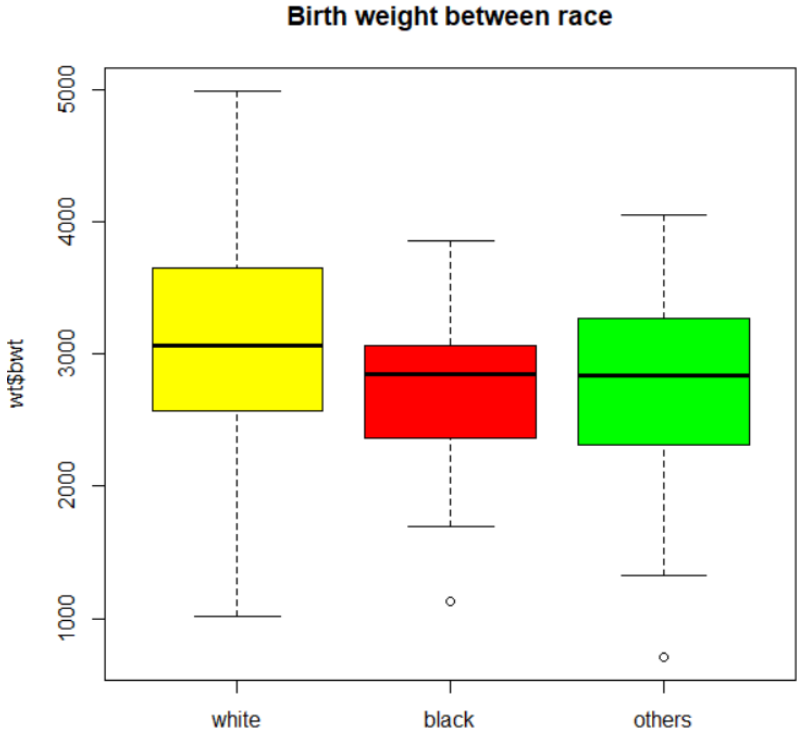
age
lwt
race
smoke
ptl
ht
ui
ftv

ตัวแปรตาม

bwt

ความสัมพันธ์ระหว่างน้ำหนักของทารกแรกเกิดกับสีผิว

```
> boxplot(wt$bwt~wt$race1, col=c("yellow","red","green"), main="Birth weight between race")
```



H $H_0 : \mu_1 = \mu_2 = \mu_3$

โดยเฉลี่ยแล้วน้ำหนักของทารกแรกคลอดในแต่ละกลุ่มไม่แตกต่างกัน

A กลุ่มตัวอย่างมาจากประชากรที่มีการแจกแจงแบบปกติและความแปรปรวนในแต่ละกลุ่มค่าเท่ากัน

T F= 4.91, df=2/df=186

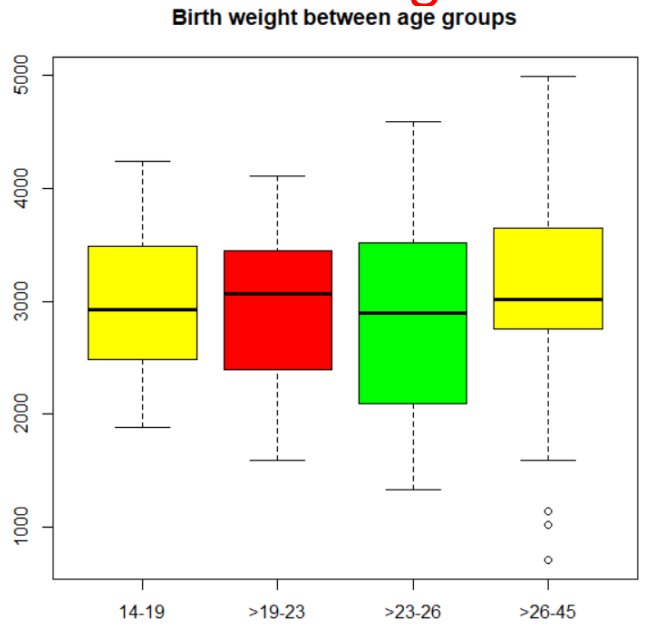
P P=0.008

C โดยเฉลี่ยแล้วน้ำหนักของทารกแรกคลอดมีอย่างน้อย 1 คู่ แตกต่างกัน

```
> oneway.test(wt$bwt~wt$race1, var.equal=T)

        One-way analysis of means

data:  wt$bwt and wt$race1
F = 4.9125, num df = 2, denom df = 186, p-value = 0.008336
```

ความสัมพันธ์ระหว่างน้ำหนักของทารกแรกเกิดกับกลุ่มอายุ

```
# campare mean, 3 groups oneway.test
summ(wt$bwt,by=wt$agegr)
boxplot(wt$bwt~wt$agegr, col=c("yellow","red","green"),
main="Birth weight between age groups")
```



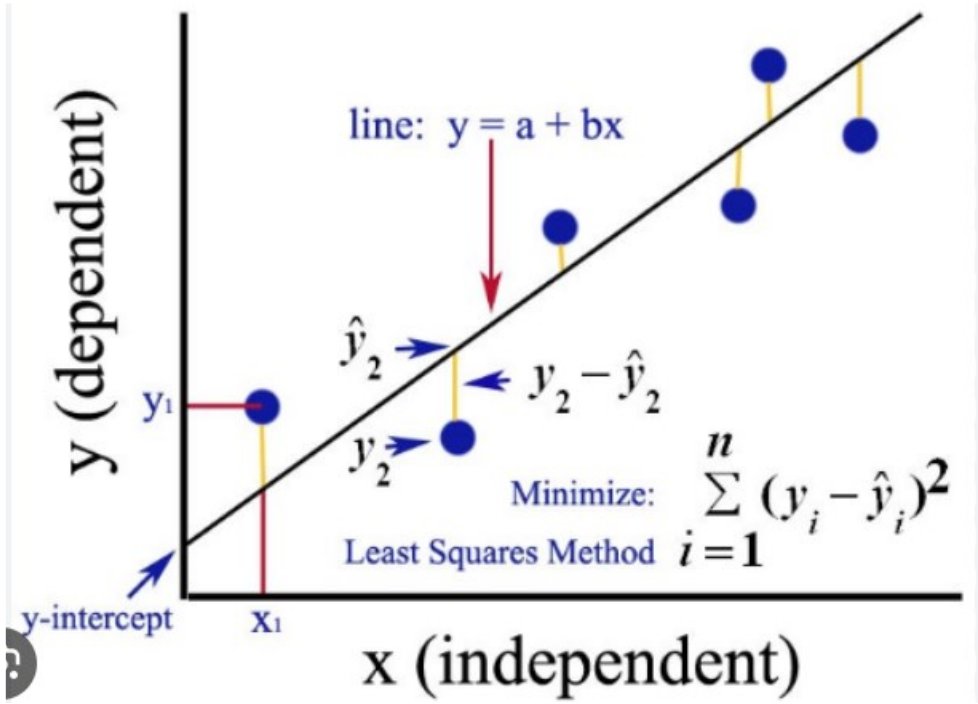
H0 จงแสดงการทดสอบสมมุติฐานของการทดสอบนี้

HATPC

```
oneway.test(wt$bwt~wt$agegr, var.equal=T)
```

```
        One-way analysis of means

data:  wt$bwt and wt$agegr
F = 0.5787, num df = 3, denom df = 185, p-value = 0.6297
```

การวิเคราะห์ข้อมูลและการสรุปข้อมูลโดยใช้ simple linear regression



Simple linear regression

การประเมินสมการความสัมพันธ์ (Assessing Relations)

- การทดสอบสมมติฐานเกี่ยวกับ slope
- การทำนายค่าโดยใช้สมการ regression
- ภาวะสารูปดี (Goodness-of-fit) ของ สมการ regression
- สัมประสิทธิ์สหสัมพันธ์ (correlation coefficient, r)

Simple linear regression

ทบทวน

Relations คือ ความสัมพันธ์ระหว่างตัวแปร continuous กับ ตัวแปร continuous ขั้นตอนการหาความสัมพันธ์เป็นดังนี้

- ระบุตัวแปรตาม Y และตัวแปรอิสระ X
- สร้างกราฟ scatter plot ระหว่าง Y และ X
- สร้างสมการถดถอย (Least Squares Regression Line, LRS)

  โดยใช้สูตร $\hat{y} = a + b\,x$
- วาดเส้นตรงบนกราฟ $b = S_{XY}/S_{XX}\ ,\ a = \bar{y} - b\,\bar{x},$
- ตรวจสอบข้อตกลงของโมเดลเชิงเส้น

Simple linear regression

ความรู้เบื้องต้น

- ความสัมพันธ์ของข้อมูลประชากรมีสมการเป็น $\hat{y} = \alpha + \beta x$

  และ α เรียก β ค่าพารามิเตอร์ของประชากร (ไม่ทราบค่า)

- ความสัมพันธ์ของข้อมูลตัวอย่างไม่คงที่ เปลี่ยนไปตามตัวอย่างที่สุ่มได้

  ค่าประมาณของ α และ β คือ a และ b ตามลำดับ ซึ่งค่าจะเปลี่ยนไปตามตัวอย่างที่สุ่มได้

- ค่า slope b เป็นค่าที่เราสนใจ เพราะเป็นค่าที่บอกว่าตัวแปรตาม Y และตัวแปรอิสระ X มีความสัมพันธ์กันหรือไม่ และ b คือค่าประมาณของการเปลี่ยนแปลงของตัวแปรตาม (Y) โดยเฉลี่ยต่อการเปลี่ยนแปลงของตัวแปรอิสระ (X) 1 หน่วย

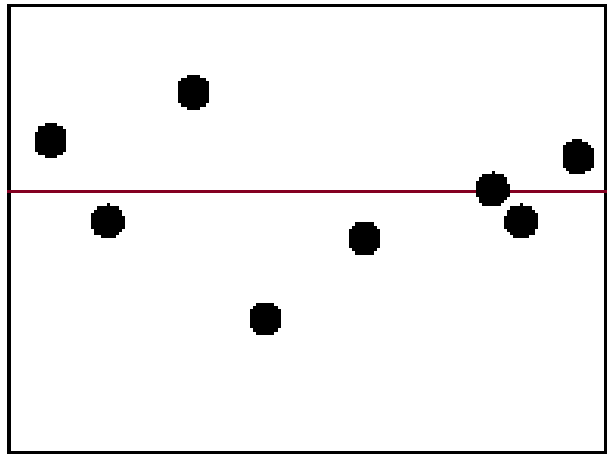
Simple linear regression

ความสัมพันธ์ระหว่าง X กับ Y

ถ้าไม่มีความสัมพันธ์ระหว่าง X กับ Y แล้ว

$\beta = 0$ ดังนั้น $\hat{y} = \alpha = \mu_y$

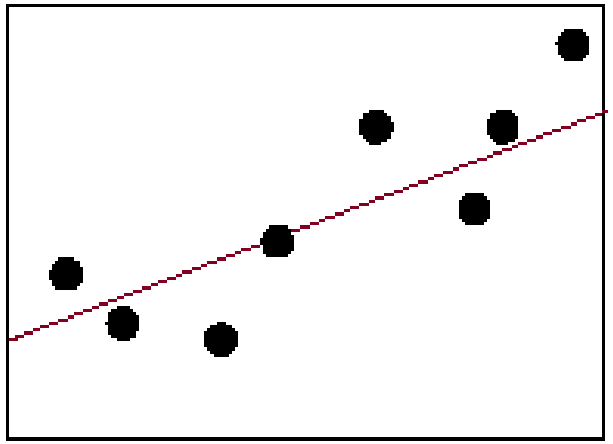
ค่า Y จะเหมือนเดิมไม่ว่าค่า X จะเปลี่ยนไป เป็นเท่าไร



ถ้ามีความสัมพันธ์ระหว่าง X กับ Y แล้ว

$\beta \neq 0$ ดังนั้น $\hat{y} = \alpha + \beta x$

ค่า Y จะเปลี่ยนค่าไป เมื่อค่า X จะเปลี่ยนไป



Simple linear regression

ข้อมูลประชากร

ความสัมพันธ์ระหว่าง X กับ Y มีจริงหรือไม่

การตรวจสอบว่าความสัมพันธ์ระหว่าง X กับ Y มีจริงหรือไม่ ทำได้โดยการทดสอบสมมติฐานเกี่ยวกับ slope หรืออีกนัยหนึ่งคือการทดสอบสมมติฐานเกี่ยวกับ β

โดยที่ตัวประมาณของ β คือ slope b จากข้อมูลกลุ่มตัวอย่าง

การทดสอบสมมติฐานเกี่ยวกับ β (Hypothesis Test for $\beta = 0$)

ขั้นตอนการวิเคราะห์ข้อมูลและทดสอบสมมติฐาน β เป็นดังนี้

ประชากรเป้าหมาย:	ระบุกลุ่มประชากรเป้าหมาย
ตัวอย่าง:	ระบุกลุ่มตัวอย่าง
ตัวแปรที่สนใจ:	ระบุตัวแปรตาม (Y) และ ตัวแปรอิสระ (X)
	เขียนกราฟ Scatter Plot และอธิบายกราฟ
ค่าสรุปเกี่ยวกับสมการ:	ค่า b และ ค่า SE(b)
ช่วงความเชื่อมั่น:	สร้างและเขียนกราฟช่วงความเชื่อมั่น 95% ของ β

การทดสอบสมมติฐาน (ต่อ)

Simple linear regression

H สมมติฐานหลัก (Null hypothesis)

$H_0 : \beta = 0$ หมายถึงไม่มีความสัมพันธ์ระหว่าง x และ y

A ตรวจสอบข้อตกลงของโมเดลเชิงเส้น (Check the assumptions of the linear model)

T สถิติในการทดสอบ (Test statistic) คือ $t = \frac{b - \beta}{SE(b)}$

เมื่อ SE(b) เป็นค่าประมาณ standard error ของ b คำนวณจาก

$SE(b) = \frac{s}{s_x \sqrt{n-1}}$, S เป็นส่วนเบี่ยงเบนมาตรฐานของ residuals

S_x เป็นส่วนเบี่ยงเบนมาตรฐานของ x

การทดสอบสมมติฐาน (ต่อ)

P

หาค่า p-value หาจากตาราง t, degree of freedom เป็น n-2

การตัดสินใจ ถ้า p-value น้อยกว่า 0.05 หรืออีกนัยหนึ่ง 0 ไม่อยู่ในช่วงความเชื่อมั่น ดังนั้น ปฏิเสธสมมติฐานหลัก

ถ้าปฏิเสธสมมติฐานหลัก (ปฏิเสธ $H_0 : \beta = 0$)

C

สรุปว่า มีความสัมพันธ์เชิงเส้นระหว่าง x กับ y ในกลุ่มประชากรเป้าหมาย

ถ้าไม่ปฏิเสธสมมติฐานหลัก (ไม่ปฏิเสธ $H_0 : \beta = 0$)

สรุปว่า ไม่มีความสัมพันธ์เชิงเส้นระหว่าง x กับ y ในกลุ่มประชากรเป้าหมาย

Simple linear regression

ตัวอย่าง

ตัวอย่างการทดสอบสมมติฐาน เกี่ยวกับข้อมูลราคารถ

ประชากรเป้าหมาย: รถยนต์ฮอนด้ามือสอง

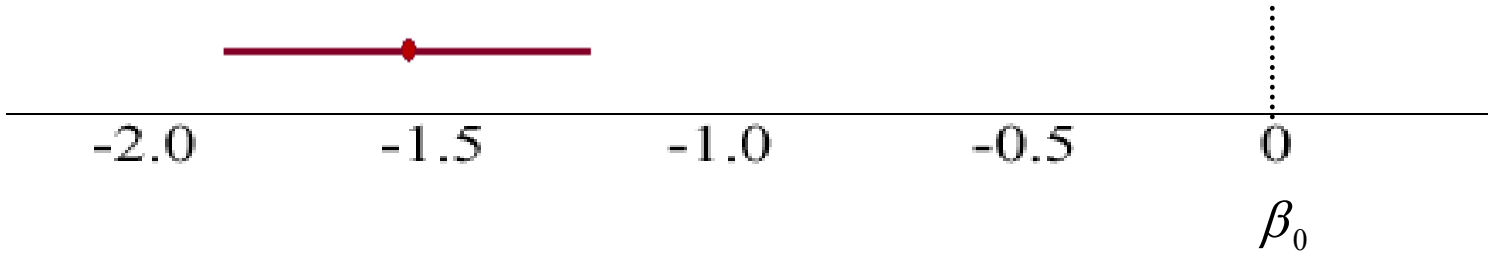
ตัวอย่าง: รถยนต์ฮอนด้ามือสองที่ถูกสุ่มมาเป็นกลุ่มตัวอย่าง จำนวน 10 คัน

ตัวแปรที่สนใจ: ตัวแปรตาม (Y)=ราคารถ และ ตัวแปรอิสระ (X) = อายุรถ

ค่าสรุปเกี่ยวกับสมการ: b= -1.6082 และ SE(b)= 0.116

ช่วงความเชื่อมั่น: ช่วงความเชื่อมั่น 95% ของ β คือ

$-1.6082 \pm 2.306 \times 0.1160 = (-1.8757, -1.3407)$



ตัวอย่าง (ต่อ)

Simple linear regression

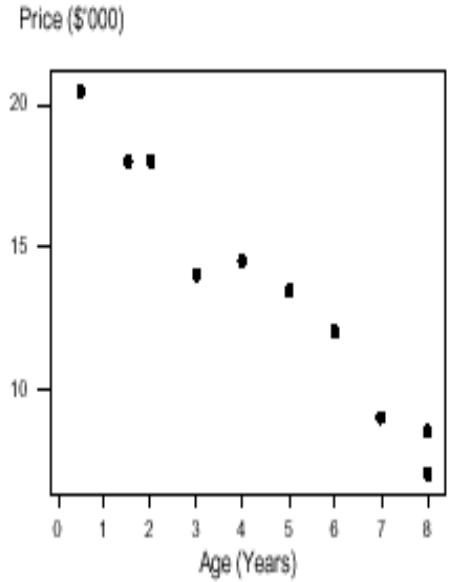
H

สมมติฐานหลัก (Null hypothesis) $H_0 : \beta = 0$

หรือไม่มีความสัมพันธ์ระหว่างอายุ และราคารถ

A

จากกราฟ Scatter Plot ข้อมูลสอดคล้องกับข้อตกลง



T

สถิติในการทดสอบ (Test statistic) คือ

$$t = \frac{b - \beta}{SE(b)}$$ =-1.6082/0.116=-13.86, df=n-2=8

P

จากตาราง t ได้ p-value<0.0005 เนื่องจาก p-value<0.05 และ 0 ไม่อยู่ในช่วงความเชื่อมั่น ดังนั้น ปฏิเสธ H_0

C

สรุปผล มีความสัมพันธ์เชิงเส้นระหว่างอายุและราคารถ ในขณะที่รถ ฮอนด้าอายุมากขึ้น ราคาจะลดลง

Example: การวิเคราะห์ข้อมูล 2 ตัวแปร by using simple linear regression

ตัวแปร

ตัวแปรอิสระ

age
lwt
race
smoke
ptl
ht
ui
ftv

ตัวแปรตาม

bwt

ตัวอย่าง “Bweight.csv”

```
> mod0 <- lm(bwt~race1, data=wt)
> summary(mod0)

Call:
lm(formula = bwt ~ race1, data = wt)

Residuals:
     Min       1Q   Median       3Q      Max
-2096.28  -502.72   -12.72   526.28  1887.28

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3102.72      72.92  42.548  < 2e-16 ***
race1black   -383.03     157.96  -2.425  0.01627 *
race1others  -297.44     113.74  -2.615  0.00965 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 714.5 on 186 degrees of freedom
Multiple R-squared:  0.05017,   Adjusted R-squared:  0.03996
F-statistic: 4.913 on 2 and 186 DF,  p-value: 0.008336
```

Hypothesis testing: example “Bweight.csv”

H

$H_0 : \beta = 0$

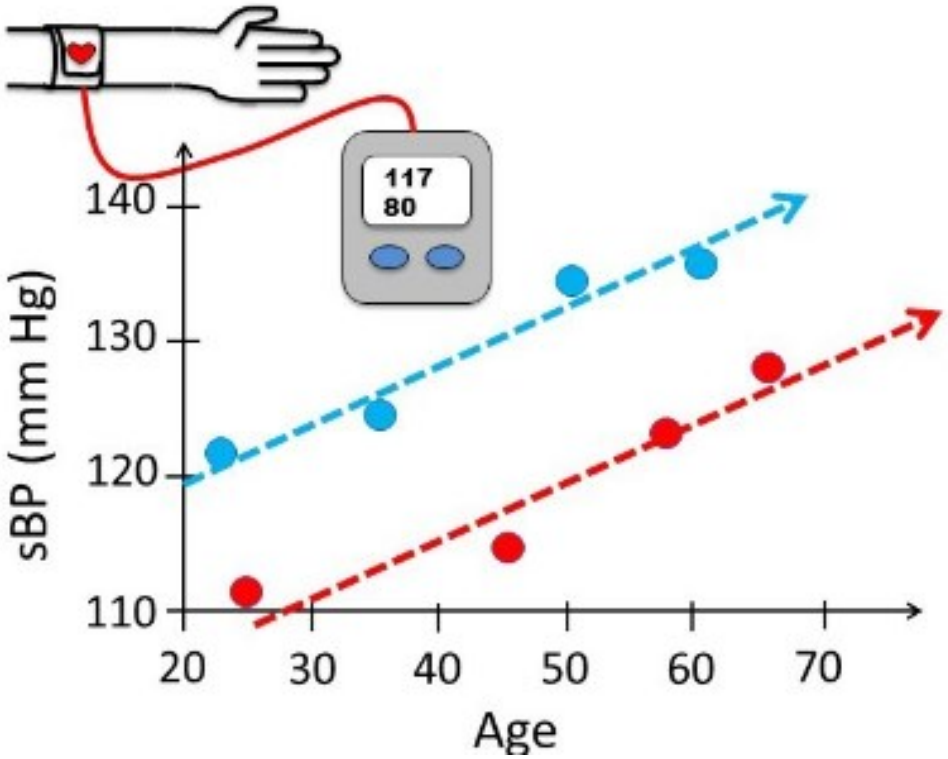
A

T

P

C

การวิเคราะห์ข้อมูลและการสรุปข้อมูลหลายตัวแปร
Multiple linear regression



$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Multiple linear regression (Full model)

```
> mod <- lm(bwt~ptl1+race1+smoke1+ht1+ui1+ftv1, data=wt)
> summary(mod)

Call:
lm(formula = bwt ~ ptl1 + race1 + smoke1 + ht1 + ui1 + ftv1,
    data = wt)

Residuals:
     Min       1Q   Median       3Q      Max
-1853.18  -429.60    36.71   498.71  1554.56

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3435.44      94.59  36.321  < 2e-16 ***
ptl11-3        -235.86     136.18  -1.732 0.084991 .
race1black     -415.49     146.34  -2.839 0.005041 **
race1others    -391.24     112.21  -3.487 0.000614 ***
smoke1yes      -345.99     106.53  -3.248 0.001388 **
ht1yes         -469.33     198.39  -2.366 0.019058 *
ui1yes         -526.17     137.92  -3.815 0.000187 ***
ftv11 time      -35.10     133.31  -0.263 0.792635
ftv12-6 times  -138.27     197.12  -0.701 0.483932
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 654.6 on 180 degrees of freedom
Multiple R-squared:  0.2285,    Adjusted R-squared:  0.1942
F-statistic: 6.664 on 8 and 180 DF,  p-value: 1.255e-07
```

H $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

A ตรวจสอบข้อตกลงของโมเดลเชิงเส้น

T

$$t = \frac{b - \beta}{SE(b)}$$

P

C

Multiple linear regression (Model selection)

```
> step(mod,direction = "backward")
Start:  AIC=2459.73
bwt ~ ptl1 + race1 + smoke1 + ht1 + ui1 + ftv1

         Df Sum of Sq      RSS    AIC
- ftv1    2    225935 77351286 2456.3
<none>                77125351 2459.7
- ptl1    1   1285310 78410662 2460.8
- ht1     1   2398045 79523396 2463.5
- smoke1  1   4519274 81644625 2468.5
- race1   2   6644715 83770066 2471.3
- ui1     1   6236292 83361644 2472.4

Step:  AIC=2456.28
bwt ~ ptl1 + race1 + smoke1 + ht1 + ui1

         Df Sum of Sq      RSS    AIC
<none>                77351286 2456.3
- ptl1    1   1260448 78611734 2457.3
- ht1     1   2386137 79737423 2460.0
- smoke1  1   4638557 81989843 2465.3
- race1   2   6678745 84030031 2467.9
- ui1     1   6176214 83527500 2468.8

Call:
lm(formula = bwt ~ ptl1 + race1 + smoke1 + ht1 + ui1, data = wt)

Coefficients:
(Intercept)      ptl1-3   race1black  race1others     smoke1yes       ht1yes
     3421.1       -232.3       -417.1       -390.1       -349.7       -465.5
     ui1yes
     -522.7
```

We exclude the variable above the <none>,
This means that if we remove the Weight, the AIC will be 228.66

Best model

Multiple linear regression (best model)

```
> mod1 <- lm(bwt~ptl1+race1+smoke1+ht1+ui1, data=wt)
> summary(mod1)

Call:
lm(formula = bwt ~ ptl1 + race1 + smoke1 + ht1 + ui1, data = wt)

Residuals:
     Min       1Q   Median       3Q      Max
-1877.36  -444.06    41.74   513.06  1568.94

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3421.06      88.71  38.563  < 2e-16 ***
ptl1-3       -232.32     134.90  -1.722 0.086746 .
race1black   -417.10     145.67  -2.863 0.004684 **
race1others  -390.12     111.31  -3.505 0.000575 ***
smoke1yes    -349.73     105.86  -3.304 0.001149 **
ht1yes       -465.47     196.45  -2.369 0.018861 *
ui1yes       -522.70     137.12  -3.812 0.000188 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 651.9 on 182 degrees of freedom
Multiple R-squared:  0.2263,    Adjusted R-squared:  0.2007
F-statistic:  8.87 on 6 and 182 DF,  p-value: 1.719e-08
```

Interpretation

แม่ที่สูบบุหรี่ระหว่างตั้งครรภ์มีผลให้น้ำหนักแรกคลอดของทารกโดยเฉลี่ยน้อยกว่า 349.73 หน่วยเมื่อเทียบกับแม่ที่ไม่สูบบุหรี่อย่างมีนัยสำคัญ

Model accuracy

Interpretation

- The estimated regression line equation can be written as

- สีผิว

- ประวัติความดันโลหิตสูง

- การกลั้นปัสสาวะ

Model accuracy

```
Residual standard error: 651.9 on 182 degrees of freedom
Multiple R-squared:  0.2263,    Adjusted R-squared:  0.2007
F-statistic:  8.87 on 6 and 182 DF,  p-value: 1.719e-08
```

Residual standard error (RSE)

- RSE is the residual variation, representing the average of the observation points around the fitted regression line.
- RSE provides an absolute measure of patterns in the data that can't be by the explained model. When comparing two models, the model with the small RSE is a good indication that this model fits the best data.

Model accuracy (cont.)

```
Residual standard error: 651.9 on 182 degrees of freedom
Multiple R-squared:  0.2263,    Adjusted R-squared:  0.2007
F-statistic:  8.87 on 6 and 182 DF,  p-value: 1.719e-08
```

R-squared and adjust R-squared

- The R-squared range 0 to 1 and represents the proportion of information (i.e. variation) in the data that can be explained by the model. The adjust R-squared adjusts for the degree of freedom.
- The R-squared measure, how well the model fits the data
- A high value of R-squared is a good indication. However, as the value of R-squared tends to increase when more predictors are added in the model.
- 22.63% ของความแปรปรวนของน้ำหนักทารกแรกคลอดสามารถอธิบายได้จากตัวแปรการคลอดก่อนกำหนด สีผิว ประวัติการสูบบุหรี่ ความดันโลหิตสูง การกลั้นปัสสาวะไม่อยู่

Model accuracy (cont.)

F-statistic:

- The F-statistic gives overall significance of the model. It assess whether at least one predictor variable has a non-zero coefficient.
- A large F-statistic will corresponds to a statistically significant p-value.

```
> anova(mod1)
Analysis of Variance Table

Response: bwt
           Df   Sum Sq Mean Sq F value    Pr(>F)
ptl1        1  4755731 4755731 11.1898 0.0009989 ***
race1       2  4718355 2359178  5.5509 0.0045702 **
smoke1      1  5337308 5337308 12.5582 0.0005013 ***
ht1         1  1630762 1630762  3.8370 0.0516603 .
ui1         1  6176214 6176214 14.5320 0.0001884 ***
Residuals 182 77351286  425007
```

We conclude that there is a linear association between ptl, race, hypertension, smoking, and urine.

MLR model assumptions

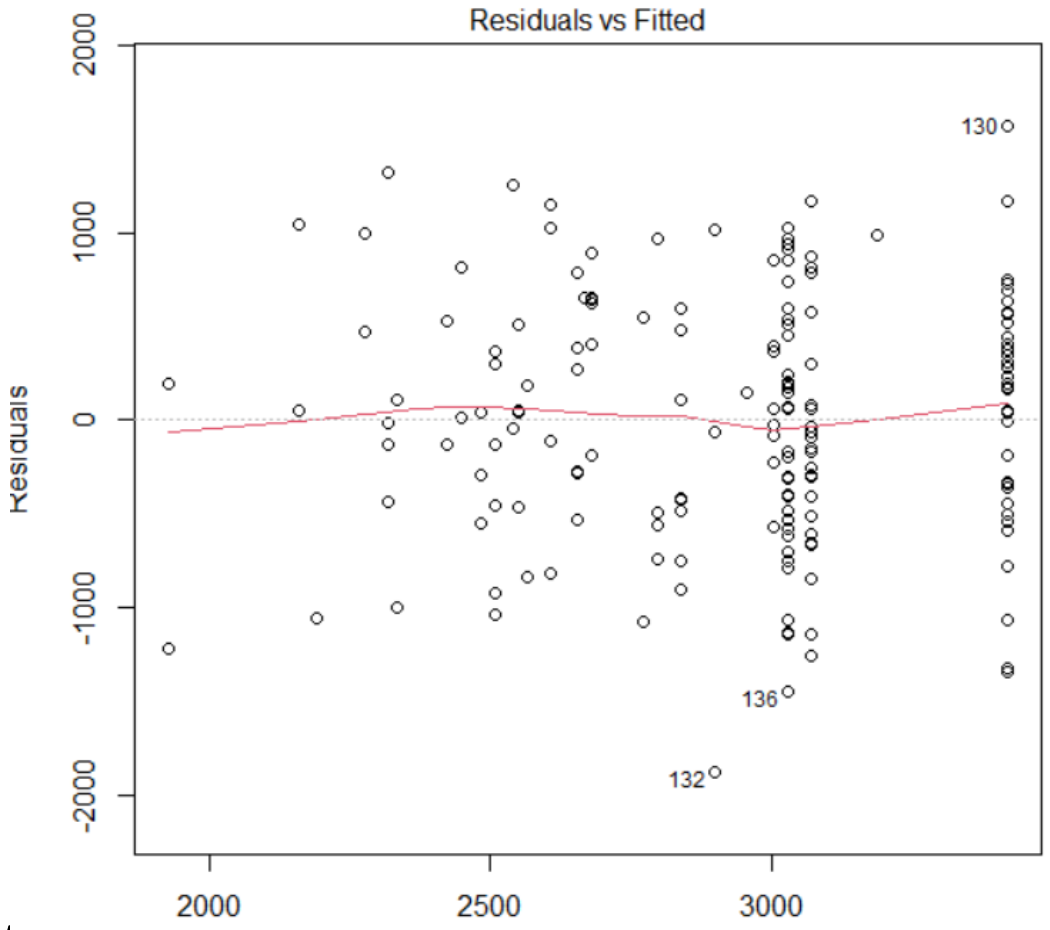
- The mean of the response, at each set of values of the predictors, is a linear function of the predictors.
- The errors are independent.
- The errors at each set of values of the predictors, are normally distributed.
- The errors at each set of values of the predictors have equal variances.

Model accuracy (cont.)

Summary:

- RSE closer to zero the better
- R-squared: higher the better
- F-statistic: higher the better

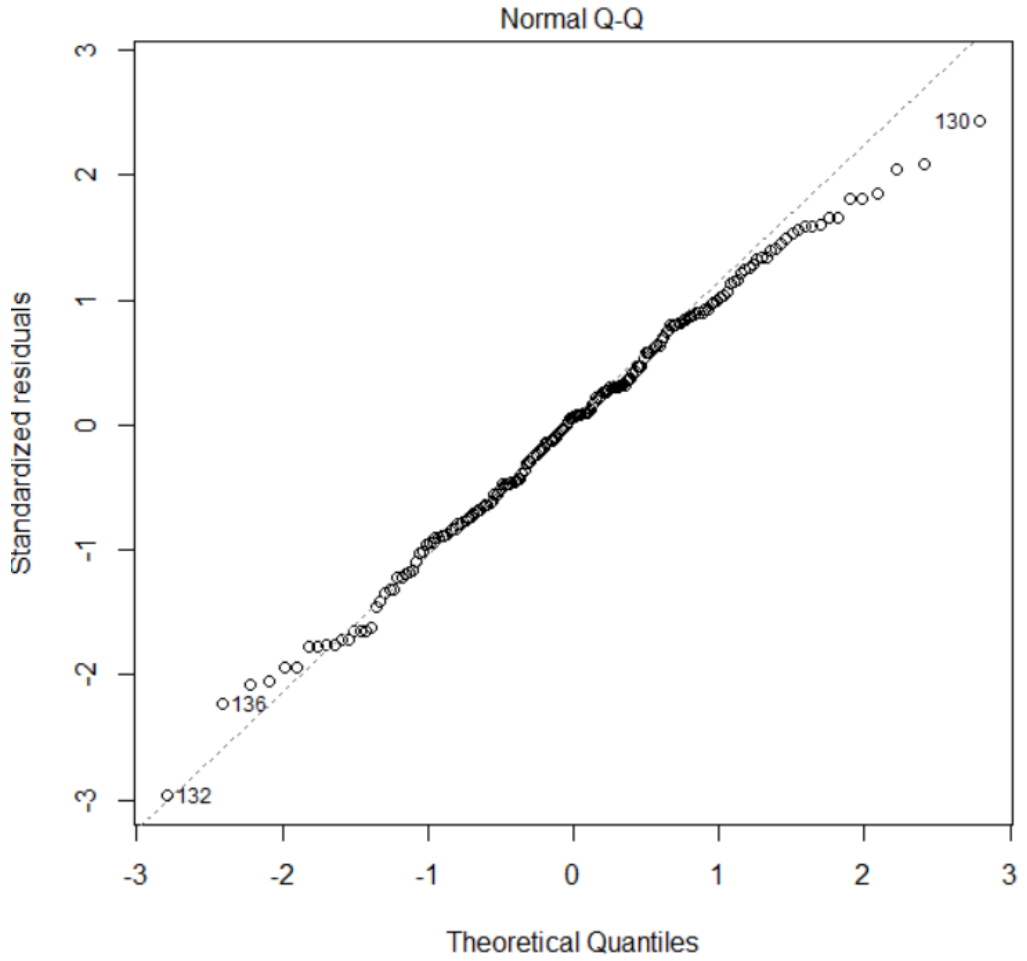
The linearity Assumptions



The plot of residuals versus fitted values is useful for assessing the assumption of linearity and homoscedasticity.
To assess the linearity: the residuals are not too far away from 0.
To assess the homoscedasticity: there is no pattern in the residual and that they are equally spread around the y=0 line.

We can see that:
- The average of the residual remains approximately 0.
- The variation of the residuals appears to be roughly constant.
- There are no excessively outlying points (except perhaps the observation with a residual of about 50).

The Normality Assumptions



The normality assumption is evaluated based on the residuals and can be evaluated using a q-q plot by comparing residuals to “ideal” normal observation. Observation lies well along the 45-degree line in the q-q plot.