Thống kê suy diễn

Đỗ Trọng Hợp
Khoa Khoa Học và Kỹ Thuật Thông Tin
Đại Học Công Nghệ Thông Tin TP. Hồ Chí Minh

Máy	Loại A	Loại B	d=B-A
1	132	140	8
2	90	108	18
3	101	112	11
4	143	140	-3
5	107	118	11
6	66	64	-2
7	100	98	-2
8	115	125	10
9	88	96	8
10	123	136	13

- Doanh số loại B nhiều hơn A có ý nghĩa thống kê với mức ý nghĩa α=0.05?
- Bộ phận kinh doanh tuyên bố B chỉ đầu tư sản phẩm B nếu B được mua nhiều hơn 10 lon/tuần. Với kết quả thí nghiệm này thì có cần đầu tư không. Mức ý nghĩa α=0.05
- Tính khoảng tin cậy của d=B-A
- Nhập data: d <- c(8 , 18 , 11 , -3 , 11 , -2 , -2 , 10 , 8 , 13)
- Hàm: mean(d), sd(d), var(d)
- Tính P ứng với t và bậc tự do df

  pt(t,df) (đuôi trái) hoặc pt(t,df,lower.tail = FALSE) (đuôi phải)

- Ví dụ: t=-2.262 ; df=9

  > pt(-2.262,9)   [1] 0.02500642

- Tính t ứng với α
  - One-tailed test : qt(α,df) (one-tailed test)
  - Ví dụ: (α=0.025): qt(0.025,9)   [1] -2.262157
  - Two-tailed test: qt(α/2,df) hoặc qt(c(α/2 , 1-α/2),df)
  - Ví dụ (α=0.05): qt(c(.025, .975),df=9)   [1] -2.262157 2.262157

Câu 1

Máy	d=B-A
1	8
2	18
3	11
4	-3
5	11
6	-2
7	-2
8	10
9	8
10	13

- Đặt μ là mean của toàn bộ quần thể B-A
- H0: μ=0 (doanh số hai loại thật sự bằng nhau)
- H1: μ>0

- n=10 , dbar=7.2
- s=7.16

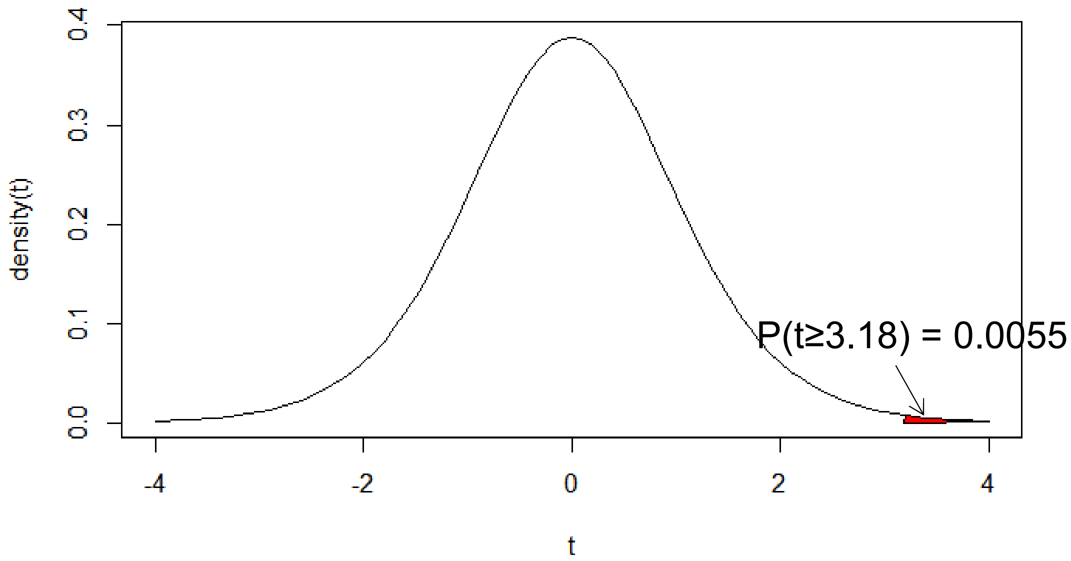
$$t = \frac{7.2-0}{7.16/\sqrt{10}}$$

- t=3.18
- P(dbar≥7.2)=P(t≥3.18)=0.0055 < α=0.05 → có thể bác bỏ H0
- Diễn giải: nếu thật sự H0 đúng thì xs P(dbar≥7.2) quá nhỏ (quá khó để xảy ra) → H0 không thể đúng được

```
> d <- c(8 , 18 , 11 , -3 , 11 , -2 , -2 , 10 , 8 , 13)
> dbar <- mean(d)
> s <- sd(d)
> n <- 10
> mu <- 0
> t <- (dbar-mu)/(s/sqrt(n))
> t
[1] 3.179222
> df <- n-1
> pt(t,df)
[1] 0.9944006      ← P(t≤3.18)
```

```
> T.values <- seq(-4,4,.1)
> x <- T.values
> y <- dt(T.values,9)
```

```
> plot(x , y , type = "l" , xlab = "t" , ylab = "density(t)")
> polygon(c(x[x>=3.18], 3.18), c(y[x>=3.18], 0), col="red")
```



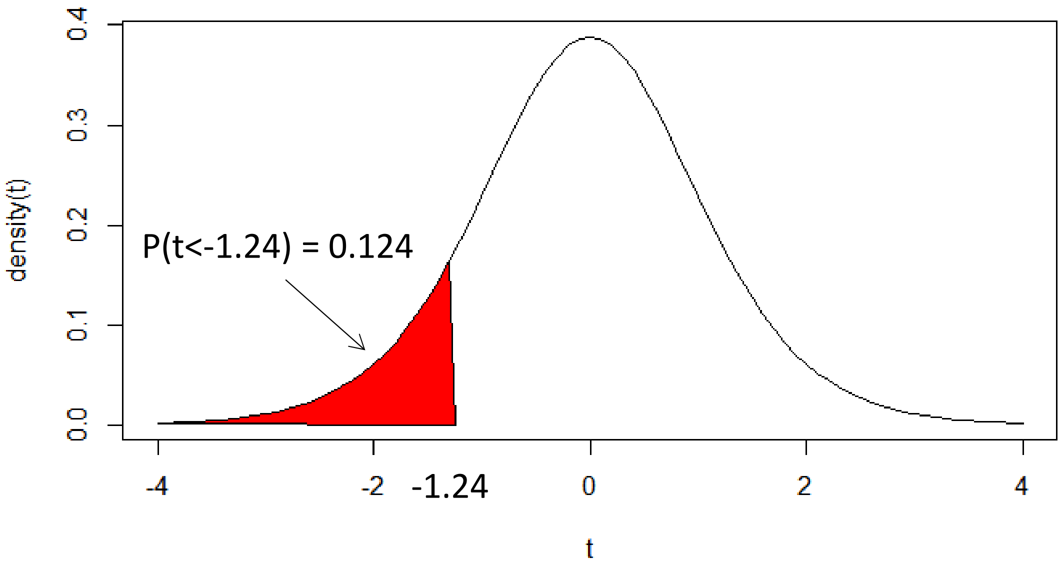

```
> pt(t,df,lower.tail = FALSE)
[1] 0.005599385 = P(t≥3.18) = P(dbar≥7.2) < α=0.05
```

Câu 2

H0: $\mu=10$

H1: $\mu<10$

```
> mu <- 10
> t <- (dbar-mu)/(s/sqrt(n))
> t
[1] -1.236364
> pt(t,df,lower.tail = TRUE)
[1] 0.1238078
> plot(x , y , type = "l" , xlab = "t" , ylab = "density(t)")
> polygon(c(x[x<=-1.24], -1.24), c(y[x<=-1.24], 0), col="rec
```



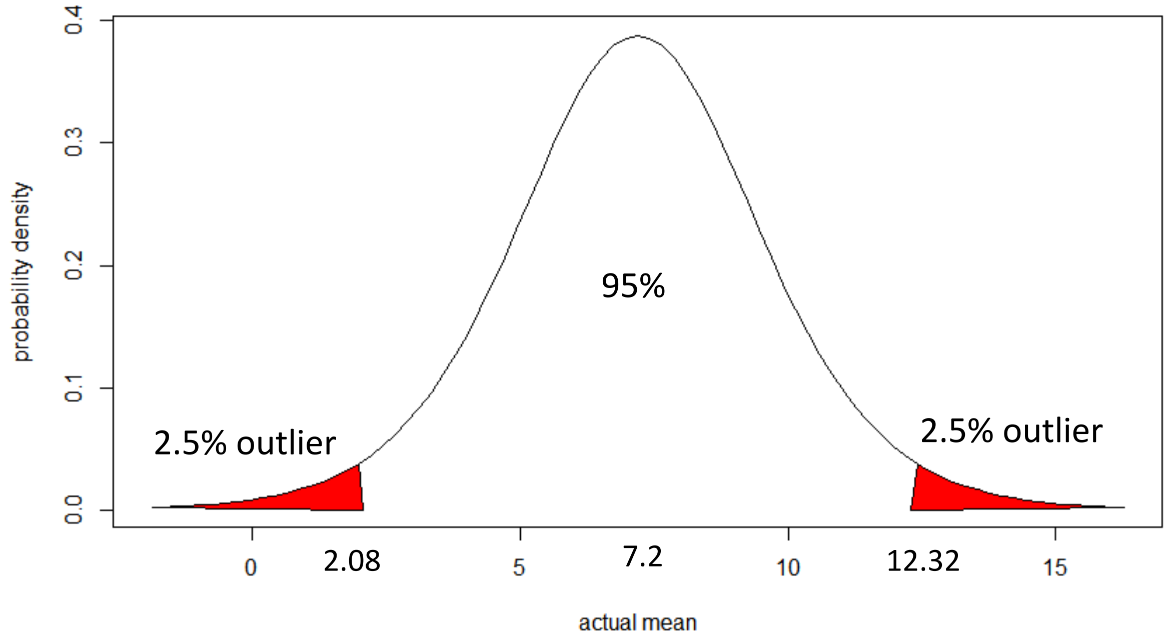
$P(dbar \le 7.2)=P(t \le -1.24)=0.124 > 0.05$

→ không thể bác bỏ H0

Diễn giải: Nếu H0 thật sự đúng thì xs để xảy ra trường hợp dbar = 7.2 hoặc thấp hơn là 0.12 (không quá nhỏ để có thể nói là rất khó xảy ra) → không bác bỏ được H0. Trường hợp ngược lại, nếu p-value < 0.05 có nghĩa là nếu H0 đúng thì xs để xảy ra dbar=7.2 hoặc thấp hơn là rất khó → bác bỏ H0 (chấp nhận H1: $\mu<10$)

Câu 3

```
> t025=abs(qt(0.025,9))
> t025
[1] 2.262157
> lbound=dbar-t025*s/sqrt(n)
> lbound
[1] 2.076881
> ubound=dbar+t025*s/sqrt(n)
> ubound
[1] 12.32312
```




```
> m <- dbar+T.values*s/sqrt(n)
> plot(m , y , type = "l" , xlab = "actual mean" ,
"probability density")
> polygon(c(m[m<=2.08], 2.08), c(y[m<=2.08], 0), col="red")
> polygon(c(m[m>=12.32], 12.32), c(y[m>=12.32], 0),
col="red")
```

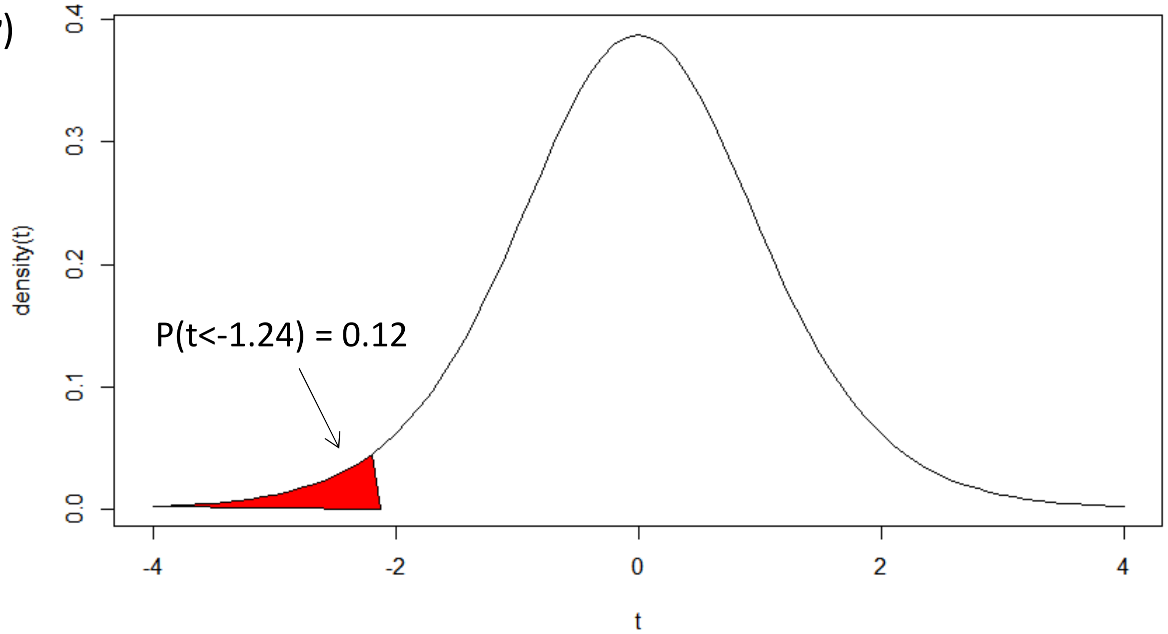
One-tailed test

- Nhà đầu tư (thích mạo hiểm) sẽ xem xét việc đầu tư sản phẩm B nếu chênh lệch (B-A) nhiều hơn **12** lon/tuần. Với kết quả thí nghiệm này thì có cần xem xét đầu tư không. Mức ý nghĩa α=0.05

- H0: μ=12
- H1: μ<12

(Nếu H1 đúng thì không đầu tư)

```
> mu <- 12
> t <- (dbar-mu)/(s/sqrt(n))
> t
[1] -2.119481
> pt(t,df,lower.tail = TRUE)
[1] 0.03154536
```



P(dbar≤7.2)=P(t≤-2.12)=0.03 < 0.05

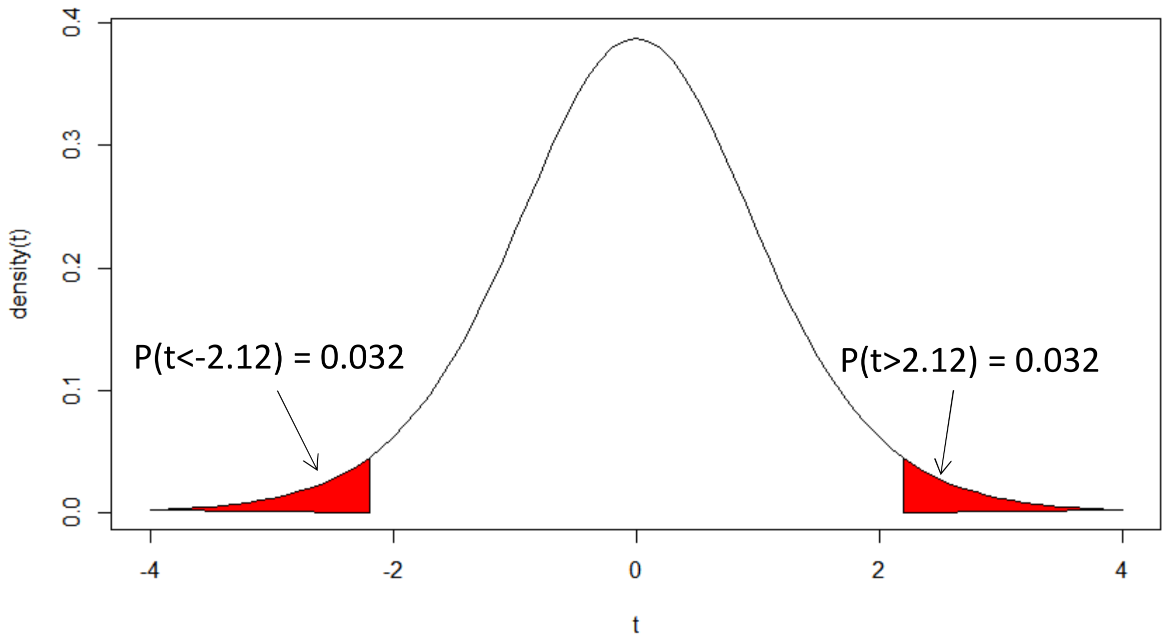
→ bác bỏ H0 (chấp nhận H1: mức chênh lệch (B-A) phải ít hơn 12 → không cần xem xét đầu tư)

Two-tailed test

- Bộ phận phát triển sản phẩm B tuyên bố mức chênh lệch doanh số (B-A) trung bình là **12** lon/tuần. Với kết quả thí nghiệm này thì có thể bác bỏ tuyên bố đấy không. Mức ý nghĩa α=0.05

- H0: μ=12
- **H1: μ≠12**

```
> mu <- 12
> t <- (dbar-mu)/(s/sqrt(n))
> t
[1] -2.119481
> pt(t,df,lower.tail = TRUE)
[1] 0.03154536
```



P(dbar≤7.2 | dbar≥16.8)=P(t≤-2.12 | t≥2.12)=0.06 > 0.05

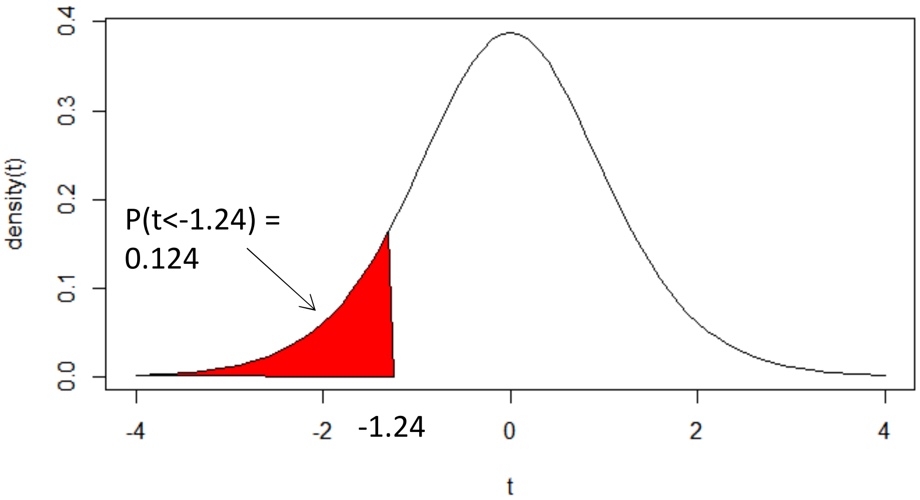
→ Không thể bác bỏ H0 (chấp nhận H0)

Lưu ý: có thể dùng khoảng tin cậy 2 bên để kiểm tra giả thuyết này.

Kết quả t-test vs khoảng tin cậy

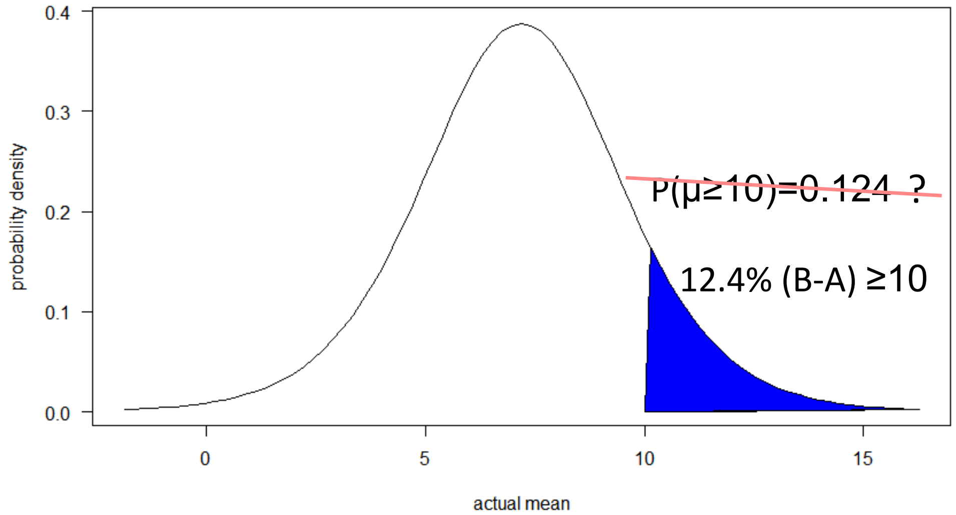
μ là trung bình thật sự của quần thể (B-A)

H0: $\mu=10$
H1: $\mu<10$



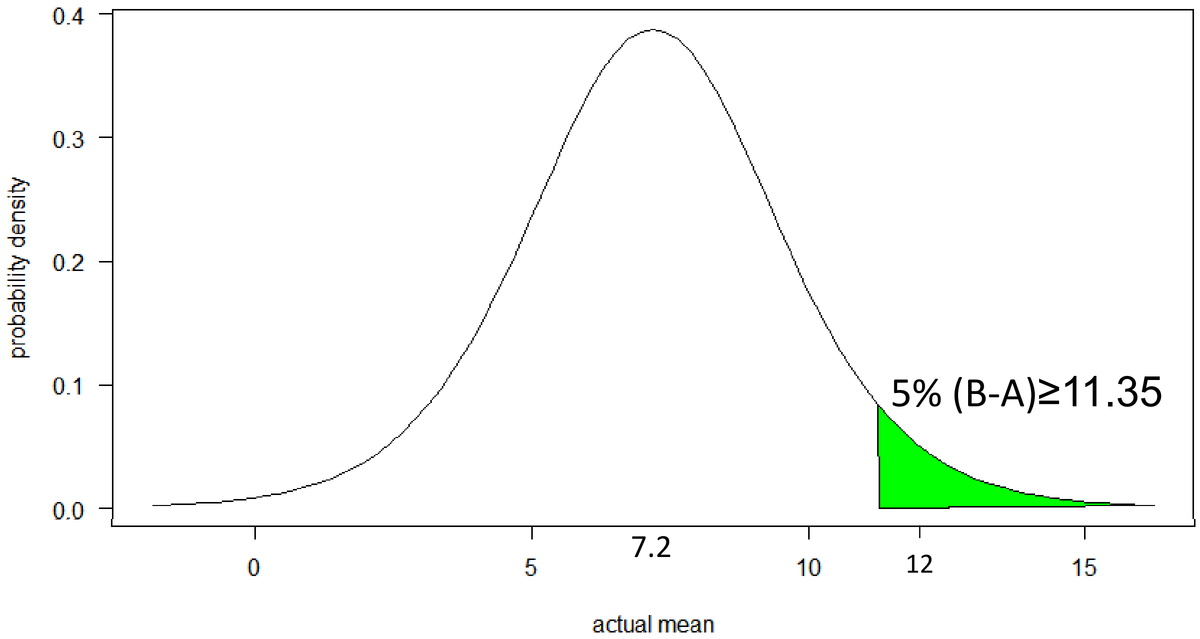
kết quả t-test

Lưu ý: ($\mu \geq 10$) không phải là biến ngẫu nhiên



khoảng tin cậy

Ý nghĩa khoảng tin cậy



One sample t test = paired t test

n=10

Máy	d=B-A
1	8
2	18
3	11
4	-3
5	11
6	-2
7	-2
8	10
9	8
10	13

One sample test

n=10

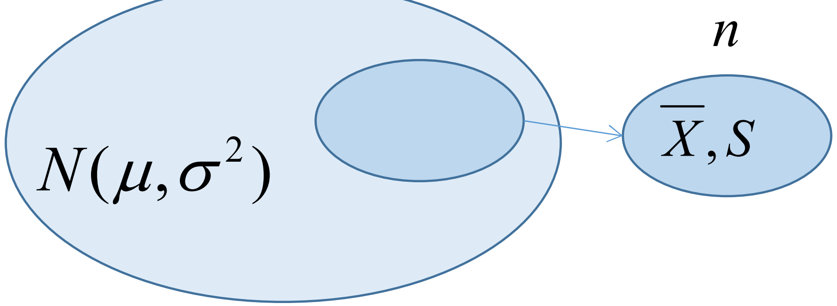
Máy	Loại A	Loại B	d=B-A
1	132	140	8
2	90	108	18
3	101	112	11
4	143	140	-3
5	107	118	11
6	66	64	-2
7	100	98	-2
8	115	125	10
9	88	96	8
10	123	136	13

Paired test

$$t = \frac{dbar - \mu}{s/\sqrt{n}}$$

Các experiment unit loại A và B được bắt cặp (cùng máy bán hàng, cùng vị trí)

Ý nghĩa công thức tính t



- Mean và variance của $\bar{X}$

mean	variance	standard deviation*
$\mu_{\bar{X}} = \mu$	$\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n}$	$\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

?

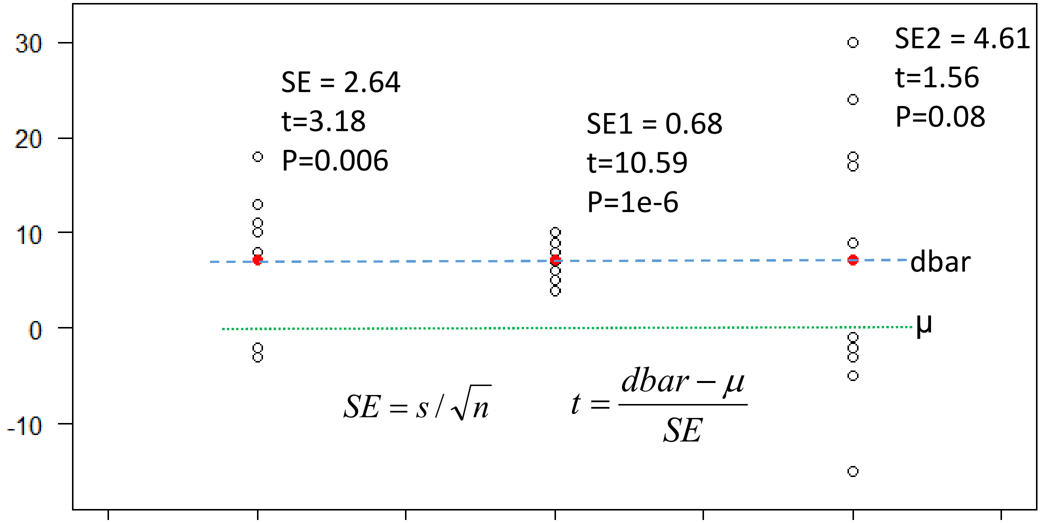
Signal

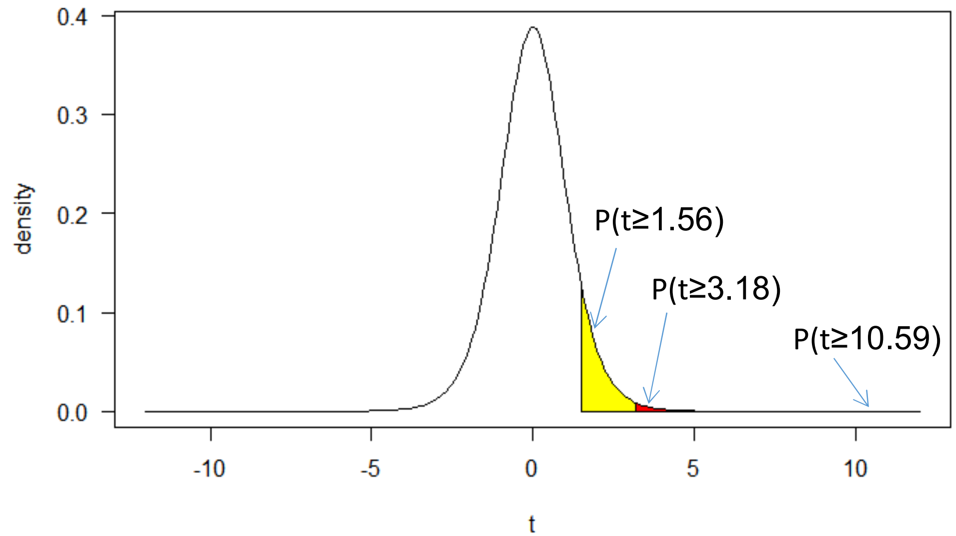
$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

Noise

- Standard Error (SE) của $\bar{X}$

$$SE = s / \sqrt{n}$$





Máy	d=B-A
1	8
2	18
3	11
4	-3
5	11
6	-2
7	-2
8	10
9	8
10	13

Máy	d=B-A
1	8
2	10
3	7
4	5
5	8
6	5
7	6
8	10
9	4
10	9

Máy	d=B-A
1	-15
2	18
3	9
4	-1
5	17
6	-5
7	-2
8	30
9	-3
10	24

- H0: $\mu=0$
- dbar = dbar1 = dbar2 = 7.2
- SE = 2.64 ; t=3.18
- $P(t \geq 3.18)=0.006$

- SE1=0.68 ; t=10.59
- $P(t \geq 10.59)=0.000001$

- SE2=4.61 ; t=1.56
- $P(t \geq 1.56)=0.08$

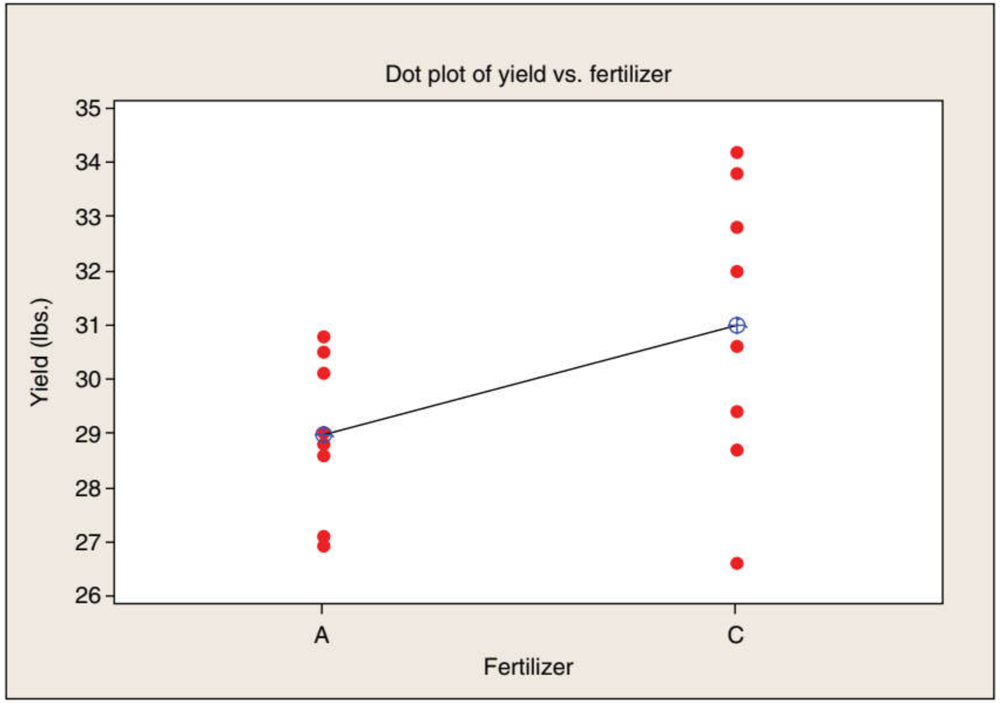
Two-sample t test

- Thí nghiệm 2 loại phân bón lên năng suất cà chua

Position	*Fertilizer*	*Yield*
1	A	30.5
2	A	28.8
3	C	32.0
4	A	29.0
5	A	27.1
6	A	30.1
7	C	26.6
8	C	34.2
9	C	28.7
10	C	32.8
11	C	30.6
12	A	30.8
13	A	26.9
14	C	32.8
15	C	29.4
16	A	28.8

Yield in Pounds.

Two-sample t test



H0: C = A

H1: C > A

2 tập mẫu thí nghiệm của 2 loại phân bón hoàn toàn tách biệt
các experiment unit loại A và C không được bắt cặp (khác vị trí trồng, khác cây)
→ cần dùng two-sample t test

Two-sample t test (giả sử 2 mẫu có **cùng phương sai**)

- Nhắc lại one-sample t test

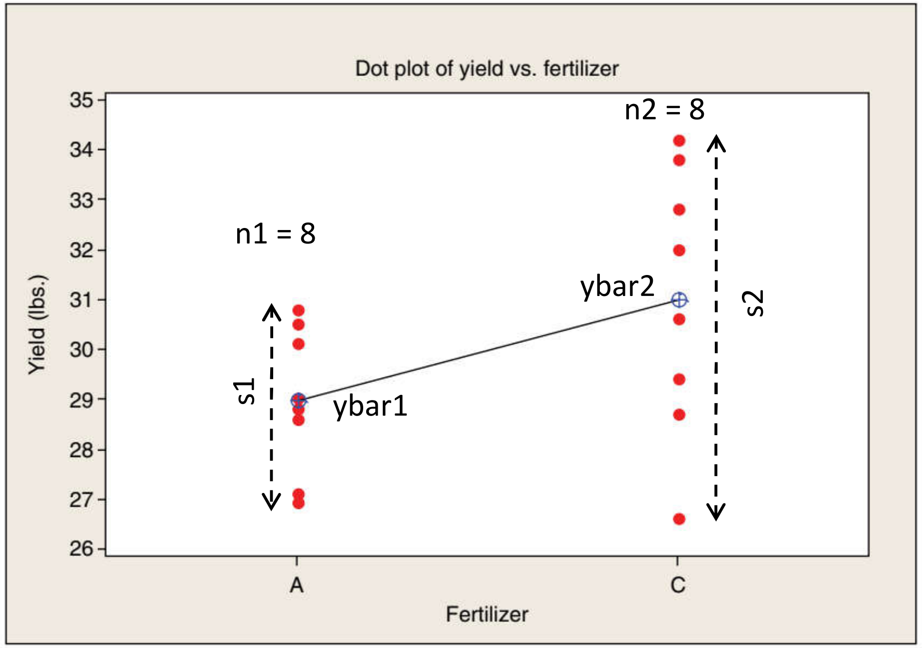
$$t = \frac{dbar - \mu}{s/\sqrt{n}}$$

- Two-sample t test
  - Giả sử $\sigma_1 = \sigma_2 = \sigma$
  - μ_1 và μ_2 là trung bình thật sự của 2 nhóm
  - s_p^2 là phương sai gộp (pooled variance) của 2 nhóm
  - $df = n_1 + n_2 - 2$

$$s_p = \sqrt{\left[\frac{\left((n_1-1)s_1^2 + (n_2-1)s_2^2\right)}{(n_1+n_2-2)}\right]},$$

$$t = \frac{(ybar_1 - ybar_2) - (\mu_1 - \mu_2)}{s_p\sqrt{(1/n_1 + 1/n_2)}}$$

$$SE = s_p\sqrt{(1/n_1 + 1/n_2)}$$



  - SE là standard deviation (hay standard error) của $(ybar_1 - ybar_2)$

Two-sample t test (giả sử 2 mẫu có **cùng phương sai**)

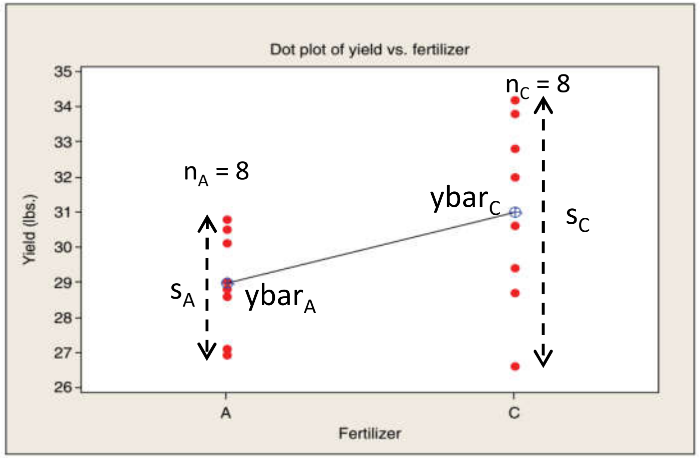
- H0: C=A (μ_C-μ_A=0)
- **H1: C>A (μ_C-μ_A>0)**
- $ybar_C$=30.9 ; $ybar_A$=29
- $ybar_C$ - $ybar_A$ = 1.9
- s_A = 1.45; s_C = 2.53
- df=n_A + Nc - 2 = 14

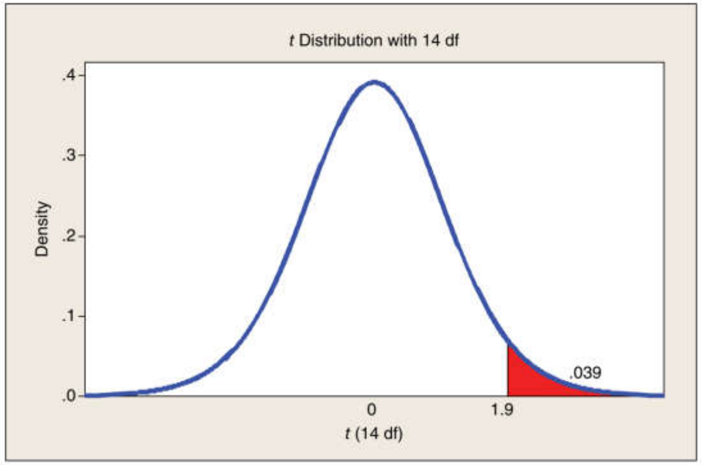
$$s_p = \sqrt{\left[\frac{\left((n_A - 1)s_A^2 + (n_C - 1)s_C^2\right)}{(n_A + n_C - 2)}\right]}.$$

- sp =sqrt((7*2.53^2+7*1.45^2)/14) = 2.06

$$t = \frac{(ybar_1 - ybar_2) - (\mu_1 - \mu_2)}{s_p\sqrt{(1/n_1 + 1/n_2)}} = \frac{ybar_C - ybar_A}{s_p\sqrt{(1/n_A + 1/n_C)}}.$$

- t = 1.83, df = 14, p-value = 0.044
- P(t≥1.83) = P($ybar_C$ - $ybar_A$ ≥ 1.9) = 0.044 → bác bỏ H0





Two-sample t test (giả sử 2 mẫu **khác phương sai**)

- Giả sử $\sigma_1 \neq \sigma_2$
- $Var(ybar_1 - ybar_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$
- $t = \dfrac{(ybar_1 - ybar_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}}$ có phân phối xấp xỉ phân phối t với một "effective df" v
- $\nu \approx \dfrac{\left(\dfrac{s_1^2}{N_1} + \dfrac{s_2^2}{N_2}\right)^2}{\dfrac{s_1^4}{N_1^2\nu_1} + \dfrac{s_2^4}{N_2^2\nu_2}}$ ← Welch–Satterthwaite equation
- t = 1.83, df = 11.156, p-value = 0.0473
- $P(t \geq 1.83) = P(ybar_C - ybar_A \geq 1.9) = 0.047$ → bác bỏ H0

Two-sample t test in R

```
> A <- c(30.5,28.8,29,27.1,30.1,30.8,26.9,28.8)
> C <- c(32,26.6,34.2,28.7,32.8,30.6,32.8,29.4)
```

```
> t.test(C,A,var.equal=TRUE,alternative="greater")

        Two Sample t-test

data:  C and A
t = 1.8267, df = 14, p-value = 0.04457
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.0675831       Inf
sample estimates:
mean of x mean of y
  30.8875   29.0000
```

```
> t.test(C,A,var.equal=FALSE,alternative="greater")

        Welch Two Sample t-test

data:  C and A
t = 1.8267, df = 11.156, p-value = 0.0473
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.03422726       Inf
sample estimates:
mean of x mean of y
  30.8875   29.0000
```

Khoảng tin cậy

- Khoảng tin cậy trong one-sample t test (nhắc lại)

$$dbar - t_{.025}s/\sqrt{n} < \mu < dbar + t_{.025}s/\sqrt{n}.$$

- Khoảng tin cậy trong two-sample t test (giả sử 2 mẫu **cùng variance**)

$$t = \frac{(ybar_1 - ybar_2) - (\mu_1 - \mu_2)}{s_p\sqrt{(1/n_1 + 1/n_2)}} \Rightarrow ybar_C - ybar_A \pm t_{.025}(df)\, s_p\sqrt{(1/n_A + 1/n_C)} \qquad s_p = \sqrt{\left[\frac{((n_A - 1)s_A^2 + (n_C - 1)s_C^2)}{(n_A + n_C - 2)}\right]}.$$

- Khoảng tin cậy trong two-sample t test (giả sử 2 mẫu **khác variance**)

$$t = \frac{ybar_C - ybar_A - (\mu_C - \mu_A)}{\sqrt{(s_C^2/n_C + s_A^2/n_A)}} \Rightarrow ybar_C - ybar_A \pm t_{.025}(df)\sqrt{(s_C^2/n_C + s_A^2/n_A)}.$$

Khoảng tin cậy 95% 2 bên (two-sided CI)

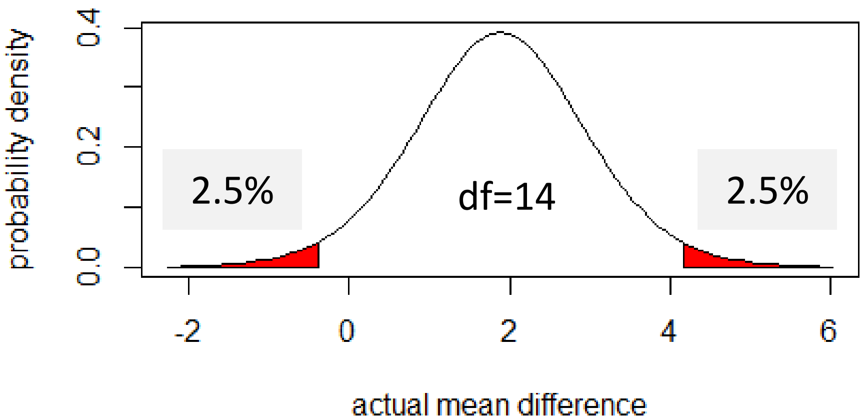
- Khoảng tin cậy trong two-sample t test (giả sử 2 mẫu **cùng variance**)

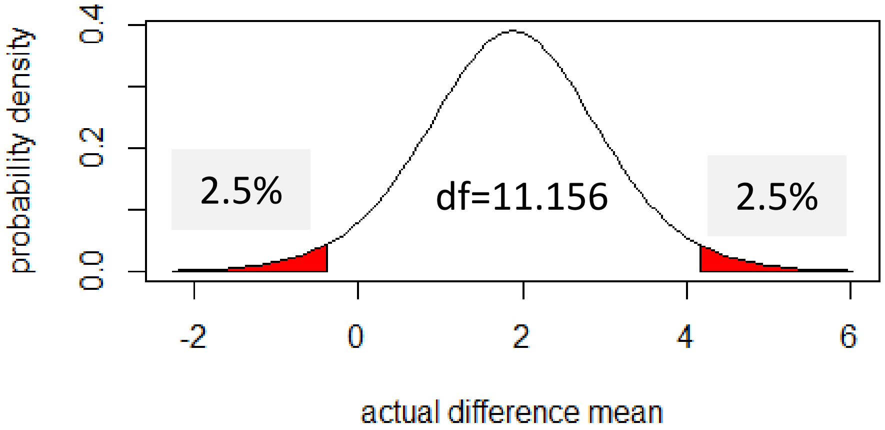
$$t = \frac{(ybar_1 - ybar_2) - (\mu_1 - \mu_2)}{s_p \sqrt{(1/n_1 + 1/n_2)}} \Rightarrow ybar_C - ybar_A \pm t_{.025}(df) s_p \sqrt{(1/n_A + 1/n_C)} \qquad s_p = \sqrt{\left[\frac{((n_A - 1)s_A^2 + (n_C - 1)s_C^2)}{(n_A + n_C - 2)}\right]}.$$

- Khoảng tin cậy trong two-sample t test (giả sử 2 mẫu **khác variance**)

$$t = \frac{ybar_C - ybar_A - (\mu_C - \mu_A)}{\sqrt{(s_C^2/n_C + s_A^2/n_A)}} \Rightarrow ybar_C - ybar_A \pm t_{.025}(df)\sqrt{(s_C^2 / n_C + s_A^2 / n_A)}.$$

```
> A <- c(30.5,28.8,29,27.1,30.1,30.8,26.9,28.8)
> C <- c(32,26.6,34.2,28.7,32.8,30.6,32.8,29.4)
```

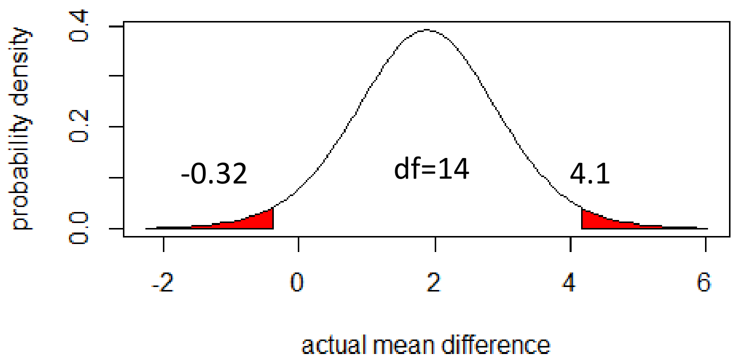


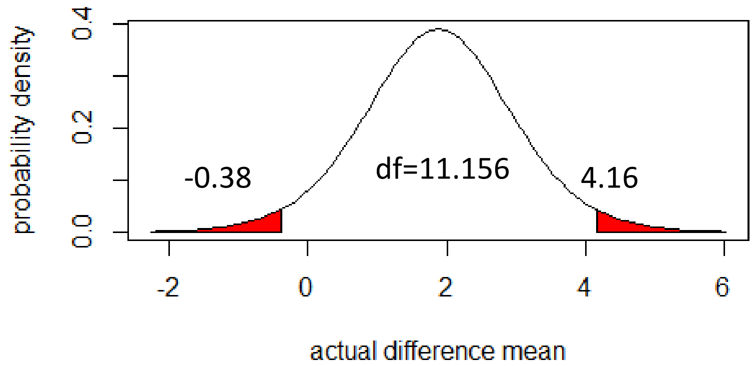


Diễn giải: chênh lệch (C-A) là từ -0.32 đến 4.1 với mức ý nghĩa 95%

Khoảng tin cậy 95% 2 bên (two-sided CI)

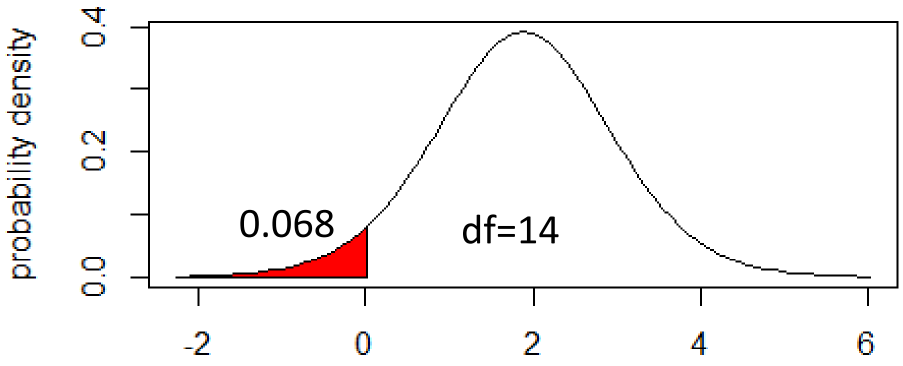
- Nhà sản xuất tuyên bố chênh lệch C-A là 4lbs. Có thể bác bỏ tuyên bố này hay không?

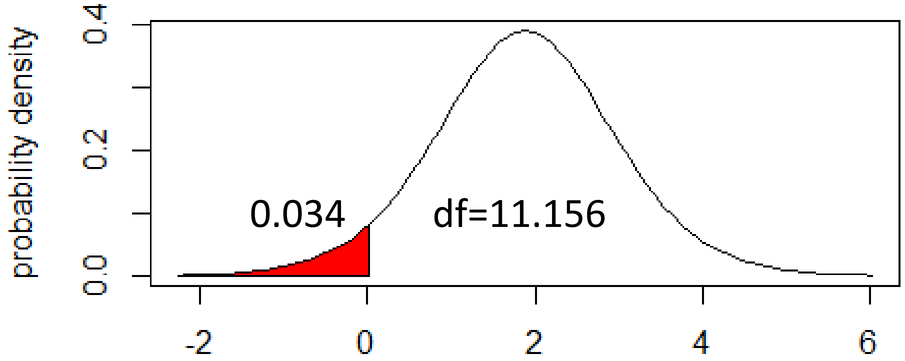




- H0: μC-μA=4. Thực tế cho phép mức chênh lệch này nhỏ hơn hoặc lớn hơn → kiểm tra 2 đầu
- t=(mean(C)-mean(A)-4)/(sp*sqrt(1/nc+1/na)) = -2.044471
- p(t≤-2.04)=0.03
- p(t≥2.04)=0.03
- p-value=0.06 > 0.05 → không bác bỏ (kiểm tra 2 đầu)

Diễn giải ý nghĩa của khoảng tin cậy





Diễn giải: chênh lệch (C-A) **thấp nhất** là từ 0.068 với mức tin cậy 95%

~~Xác suất P(µC-µA≥0.068)=95%~~ ((µC-µA) không phải biến ngẫu nhiên)

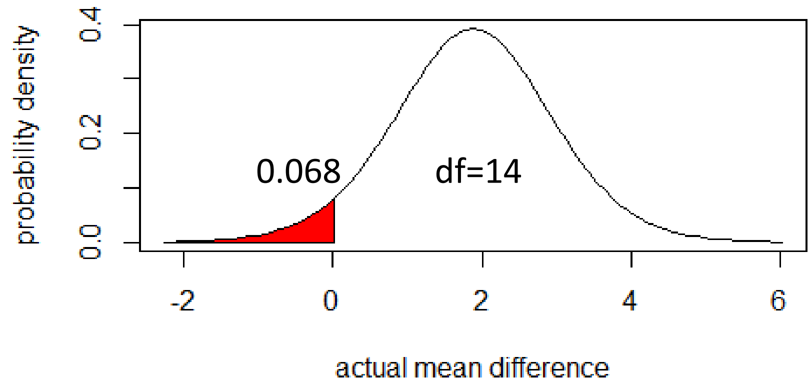
Nếu lặp lại thí nghiệm, xs P(ybarC-ybarA≥0.068) là 95%.

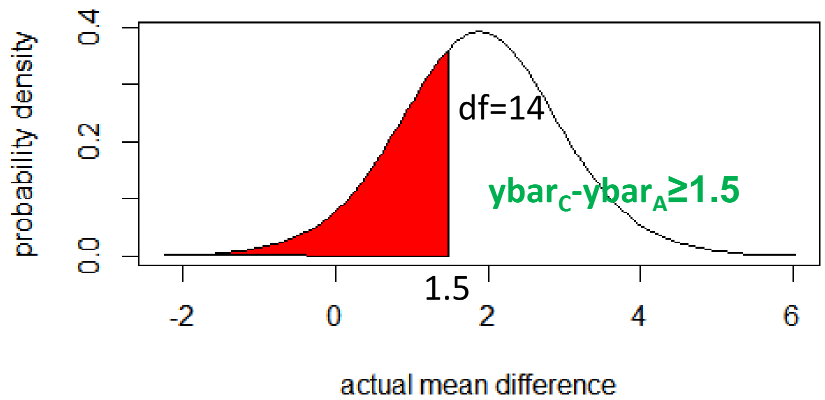
Nếu lặp lại thí nghiệm 100 lần, trung bình có 95 lần giá trị (ybarC-ybarA)≥0.068

Lưu ý: P(ybarC-ybarA ≥ x) ≠ P(ybarC-ybarA ≥ x | H0 đúng)

Ước lượng kích thước mẫu

- Nhà tư vấn tài chính tuyên bố: Nếu C làm tăng sản lượng cà chua ít nhất 1.5lbs thì ta nên chuyển sang dùng C vì lúc đó mới có lời.
- Thí nghiệm hiện tại tuyên bố C làm tăng ít nhất là 0.068lbs (với mức ý nghĩa 5%)
- Hiện tại thì xs lời khi chuyển sang C là khoảng **64.3%** (không đủ cao nên không thể đưa ra quyết định chuyển)
- Cần phải thí nghiệm trên bao nhiêu mẫu để đưa ra tuyên bố C làm tăng sản lượng ít nhất 1.5lbs ở mức ý nghĩa 5%?
- Lưu ý: ta giả sử lượng tăng trung bình ở thí nghiệm tương lai vẫn như thí nghiệm hiện tại.





Ước lượng kích thước mẫu

- Chặn dưới 95% của $ybar_C$-$ybar_A$ (giả sử cùng phương sai)

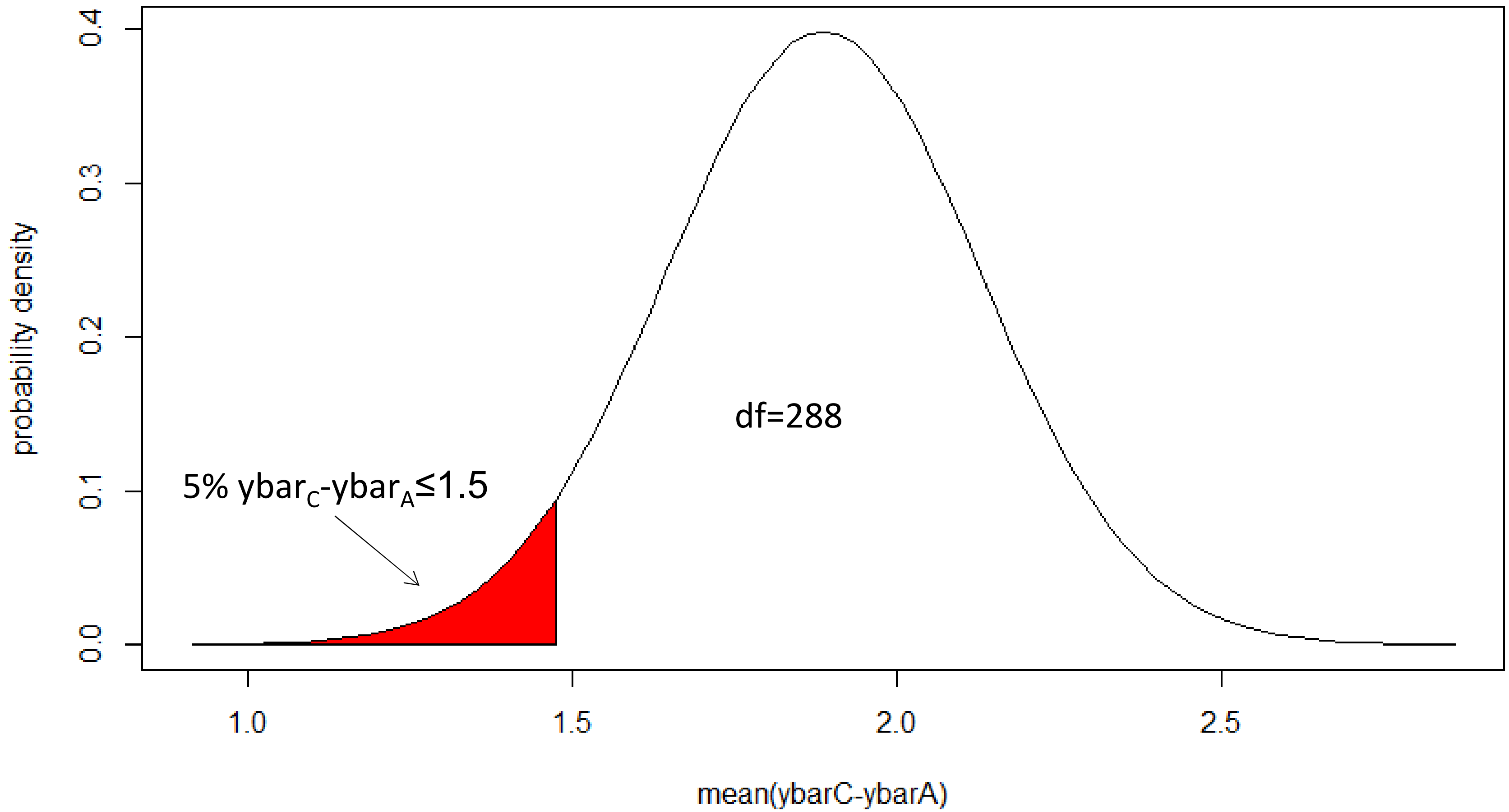
$$ybar_C - ybar_A - t_{.05,n_C+n_A-2} \times s_p \times \sqrt{1/n_C + 1/n_A}$$

$$ybar_C - ybar_A - t_{.05,n_C+n_A-2} \times s_p \times \sqrt{1/n_C + 1/n_A} \geq 1.5 \quad \leftrightarrow \quad 1.9 - 1.76 \times 2.07 \times \sqrt{1/n_C + 1/n_A} \geq 1.5$$

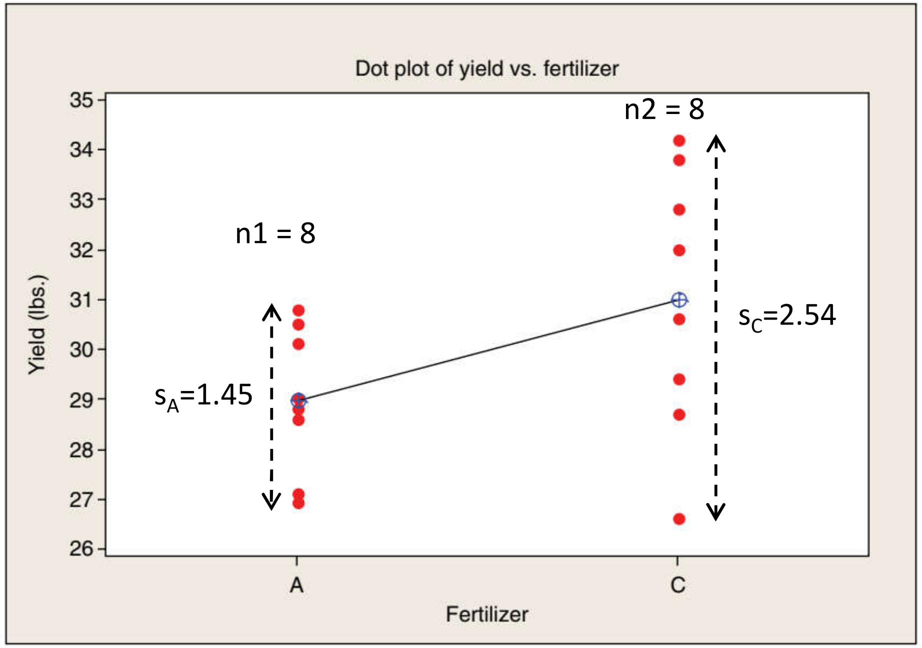
- Với s_p=2.067 ; $t_{.05,14}$=1.76
- Tổng số nA+nC nhỏ nhất khi nA=nC=n $\rightarrow \sqrt{\frac{2}{n}} \leq 0.11 \rightarrow n \geq 164$
- Thay n=164 tính lại $t_{.05,326}$=1.65 $\longrightarrow$ n=145

Chú ý: 145 chỉ là con số ước lượng vì ta chưa biết các thông số $ybar_C$-$ybar_A$ và sp của thí nghiệm tương lai

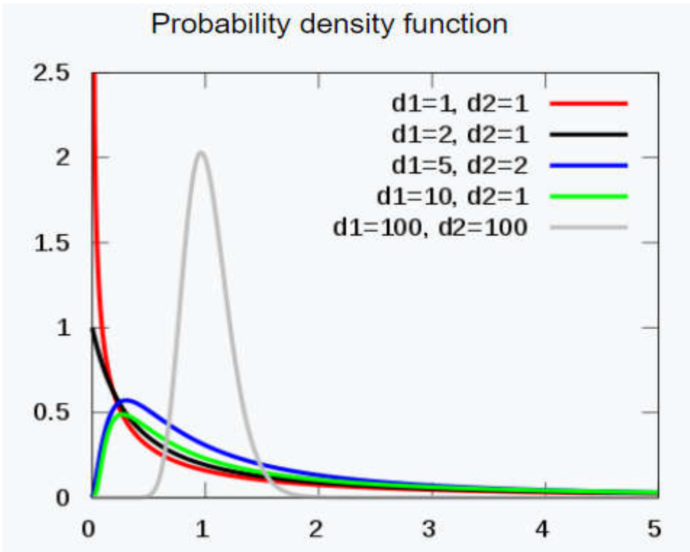
Khoảng tin cậy 95% 1 bên khi n1=n2=145



F-test of equality of variances



- H0: $\sigma_C = \sigma_A$; H1: $\sigma_C > \sigma_A$
- Tính $F_{ex} = \dfrac{s_C^2}{s_A^2}$
- Tính $P(F>F_{ex})$

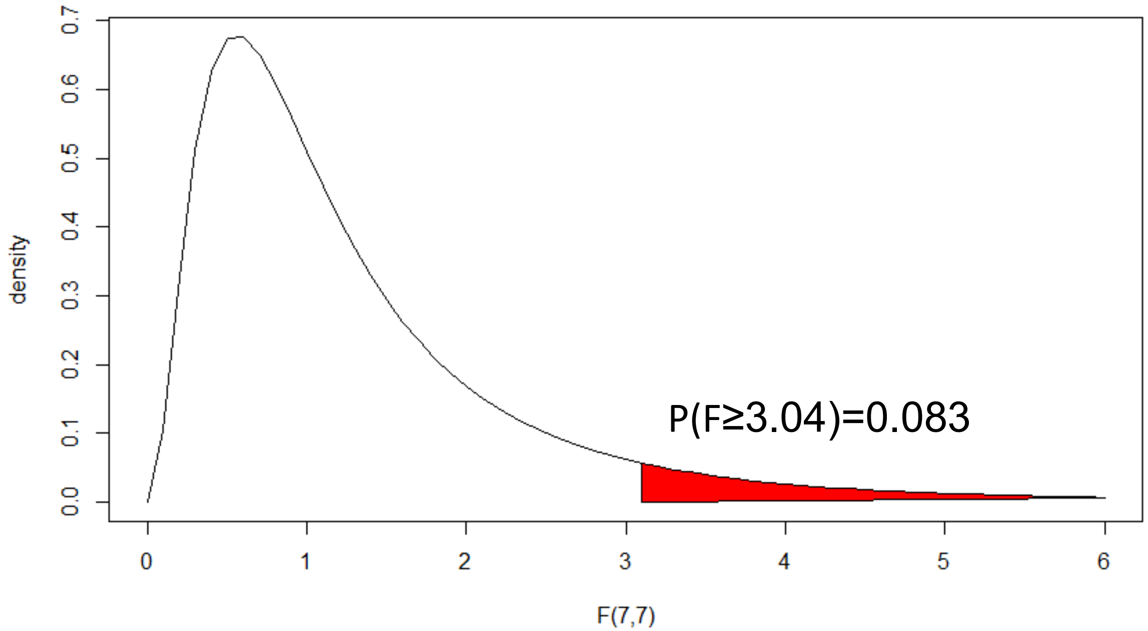


F-test of equality of variances

- H0: $\sigma_C = \sigma_A$; H1: $\sigma_C > \sigma_A$
- $F_{ex}=(2.54/1.45)^2=3.04$
- $df_C=7$; $df_A=7$

> pf(3.04,7,7,lower.tail = FALSE)=0.083

→ không thể bác bỏ H0



F-test of equality of variances

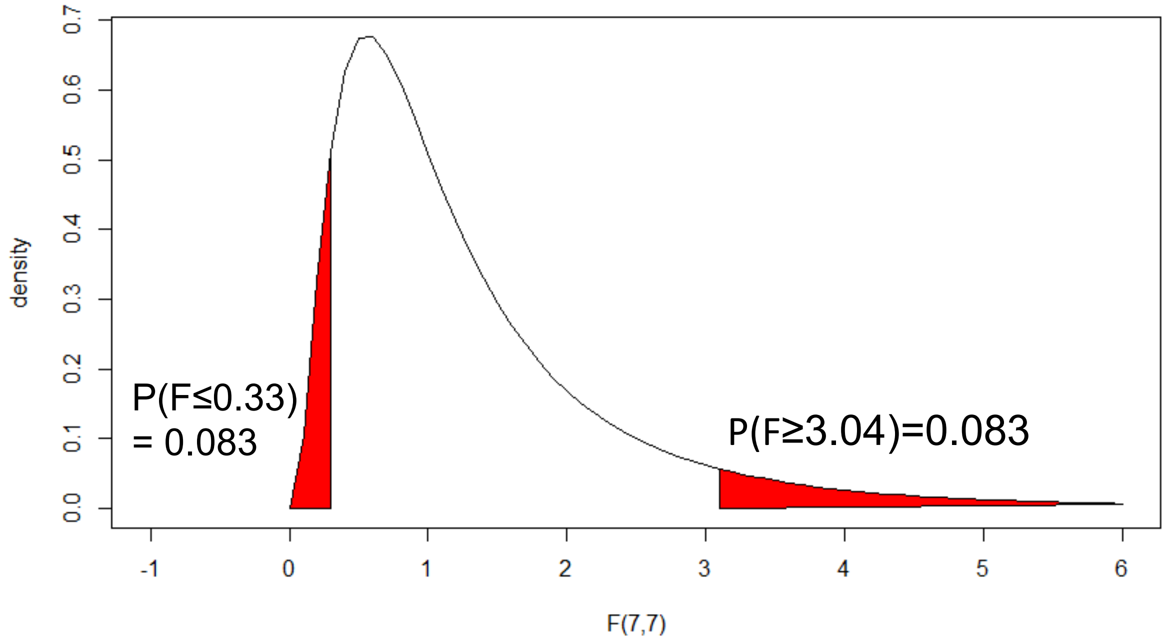
- H0: $\sigma_C = \sigma_A$; H1: $\sigma_C \neq \sigma_A$
- $F_{C>A} = (sC/sA)^2 = 3.04$
- $1/F_{C>A}=1/3.04=0.33$
- $df_C=7$; $df_A=7$

> pf(3.04,7,7,lower.tail = FALSE)=0.083

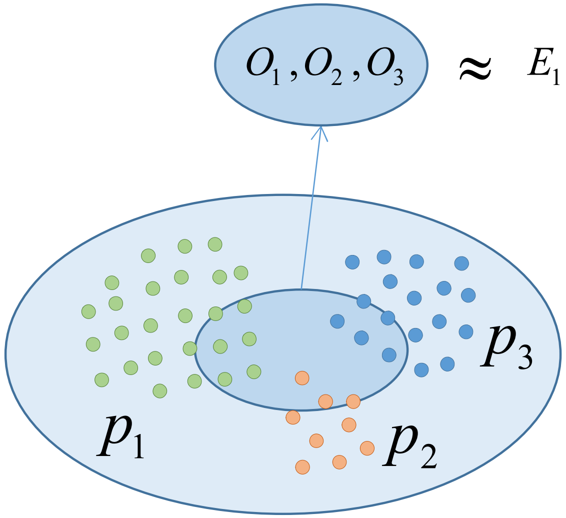
> pf(0.33,7,7,lower.tail = TRUE)=0.083

- P(F≤0.33 | F≥3.04)=0.166

→ không thể bác bỏ H0



Chi-square test



$\approx E_1, E_2, E_3$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}$$

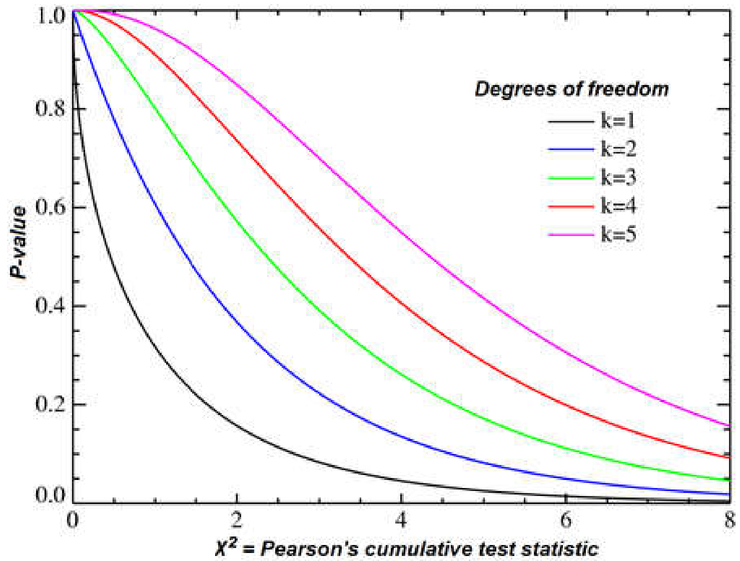
$$H_0 : f_i = p_i$$

Chọn mức ý nghĩa α

Tính p-value < α → bác bỏ H0

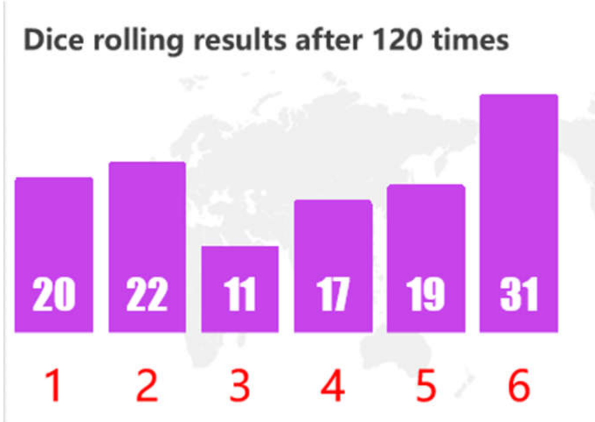
Tra bảng chi-square tìm $\chi_{\alpha;v}$

Nếu $\chi > \chi_{\alpha;v}$ ➡ bác bỏ H_0



Chi-square test

- Trắc nghiệm con xúc xắc có đều?
- Lý thuyết: xác suất mỗi mặt là 1/6 = Giả thuyết H0
- Kết quả như hình



- Tính $\chi^2=(20-20)^2/20+(22-20)^2/20+...+(31-20)^2/20=10.8$
- Bậc tự do v=k-1=6-1=5
- Tính $P(\chi 2 \geq 10.8)=0.055$ (> pchisq(10.8,5,lower.tail = FALSE))
- p-value > 0.05 → không bác bỏ H0
- $\chi^2_{.05;v=5} = 11.07$ (>qchisq(0.05,5,lower.tail = FALSE))
- $\chi<\chi^2_{.05;v=5}$ → không bác bỏ H0 (tức là công nhận xúc xắc đều)