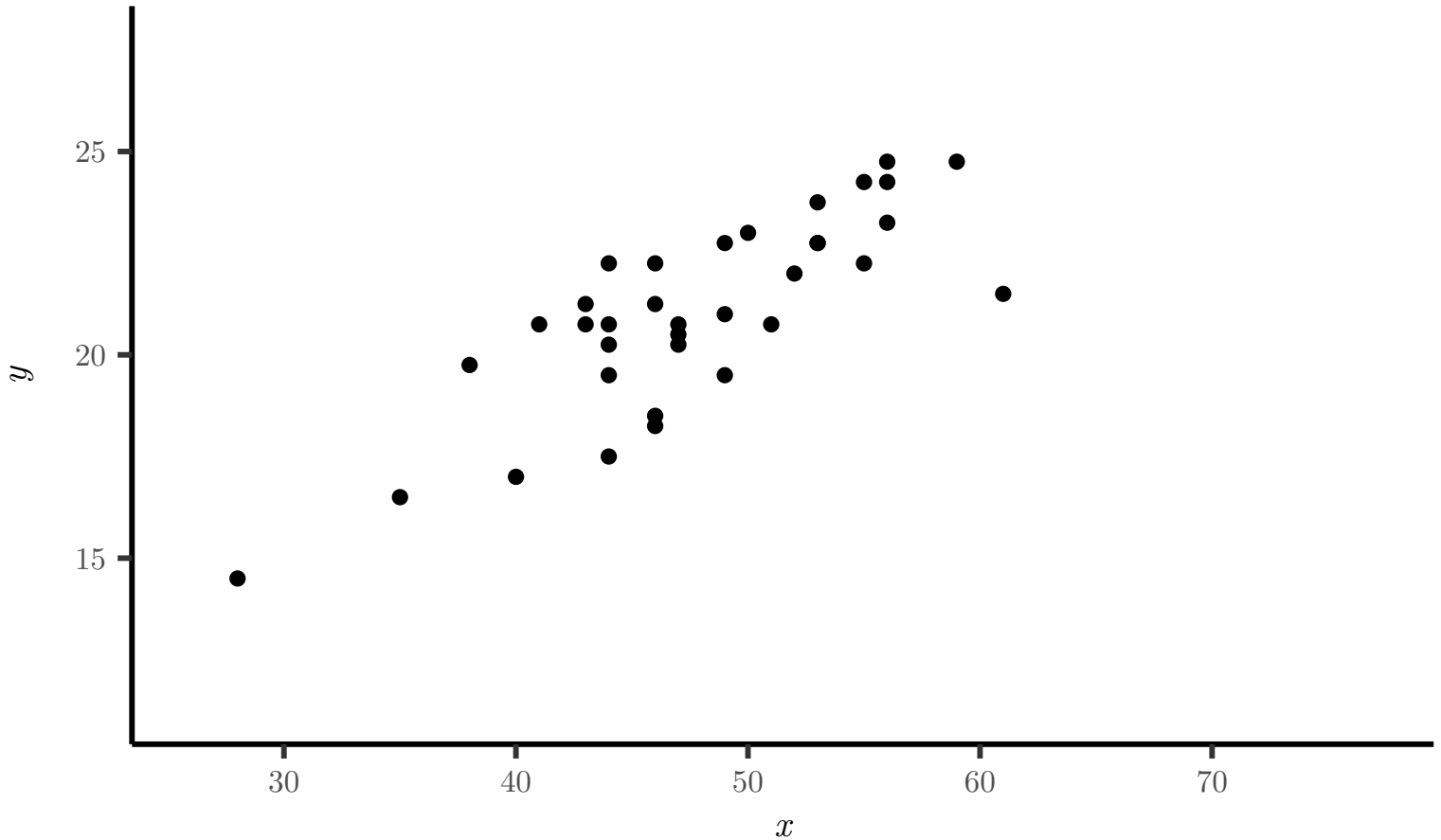
Les données

Individu	ht	circ
1	24.25	55
2	24.25	56
3	24.75	56
4	22	52
5	22.75	49
6	20.25	44
7	22.75	53
8	20.25	47
9	20.75	51
10	21.5	61

Au total $n = 35$ Eucalyptus.

Objectif

1. Expliquer la hauteur par la circonférence
2. Utiliser la hauteur & circonférence $\rightarrow$ volume



Représentation des mesures pour les $n = 35$ eucalyptus mesurés.

conclusion

Le modèle

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \cdots, n$$

1. x_i (circonférences) fixées
2. ε_i (erreur due à l'imprécision modèle & mesure) : aléatoire
   ε_i indépendants identiquement distribués $\mathcal{N}(0, \sigma^2)$
3. y_i (hauteur) : aléatoire
4. β_0, β_1 fixes et inconnus
   $\rightarrow$ β_0, β_1 à estimer...

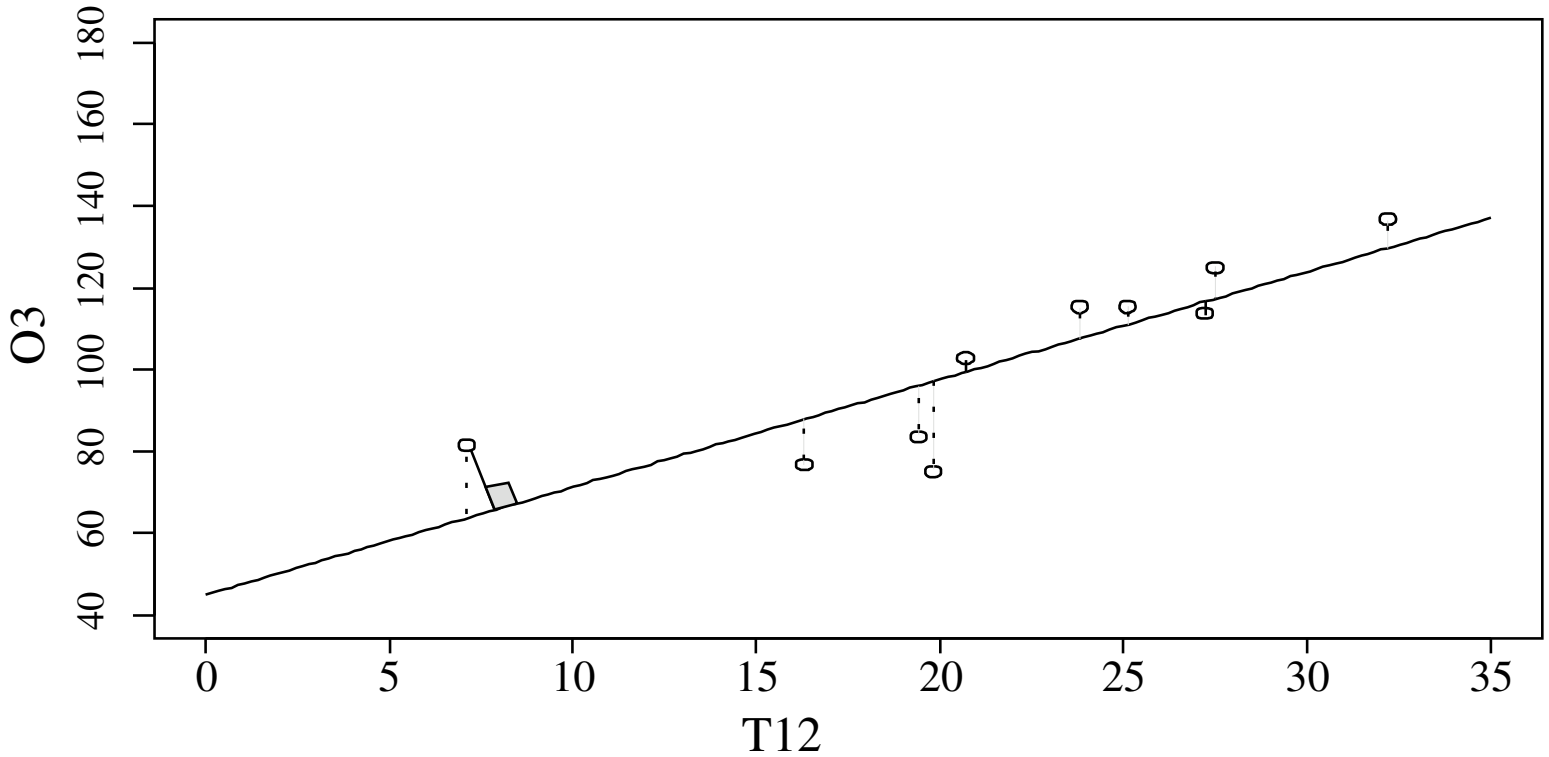
Le coût

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Objectif

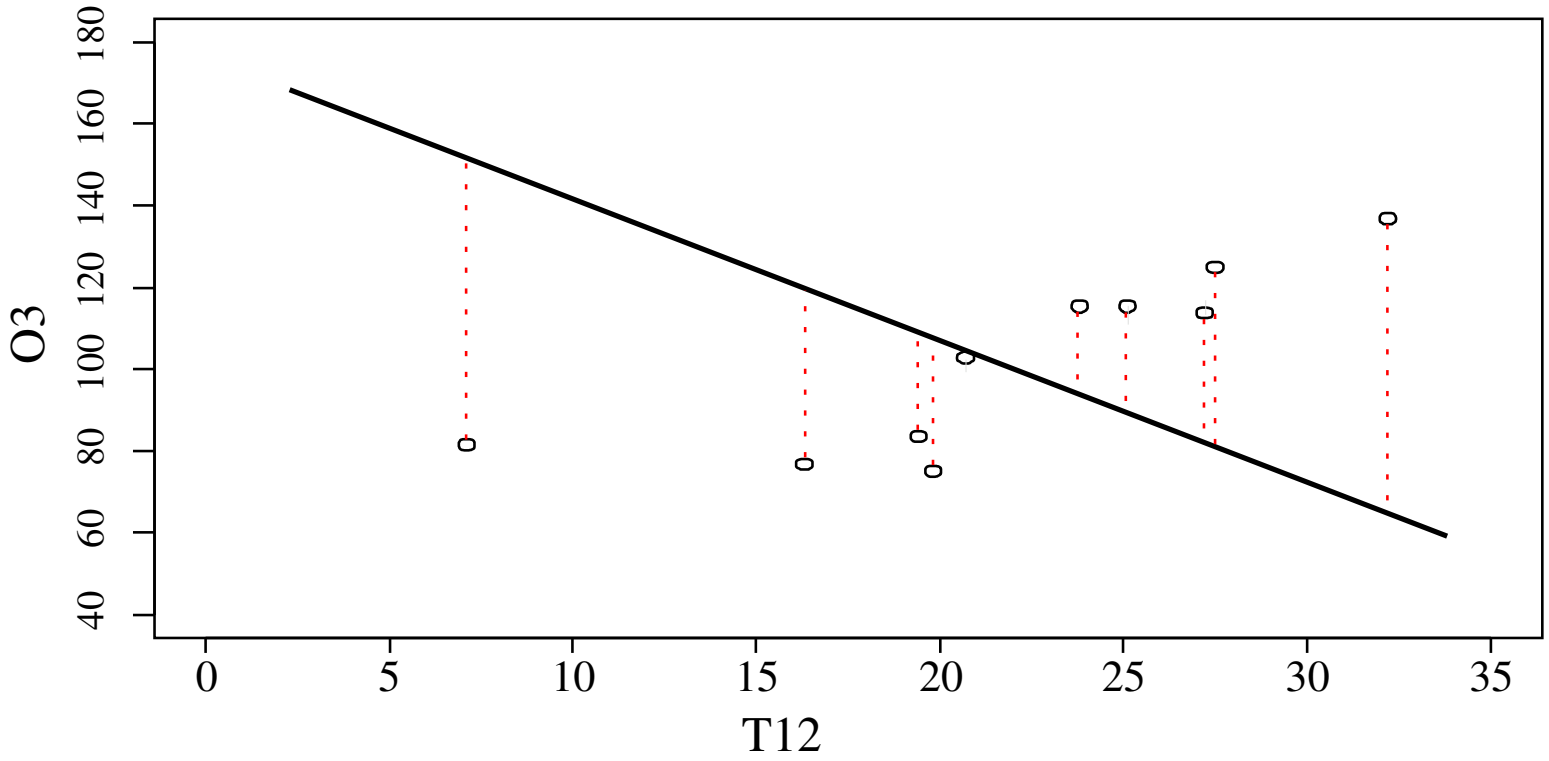
Minimiser le coût

Exemple de droite minimisant le coût



Distances à la droite : coût absolu (pointillés) et distance d'un point à une droite.

Exemple de droite non trouvée



Cette droite ne minimise pas le coût (et nous intéresse pas)

Le calcul des estimateurs

On cherche les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ qui minimisent le coût :

$$S(\hat{\beta}_0, \hat{\beta}_1) \quad \text{minimum}$$

Que valent $\hat{\beta}_0, \hat{\beta}_1$?

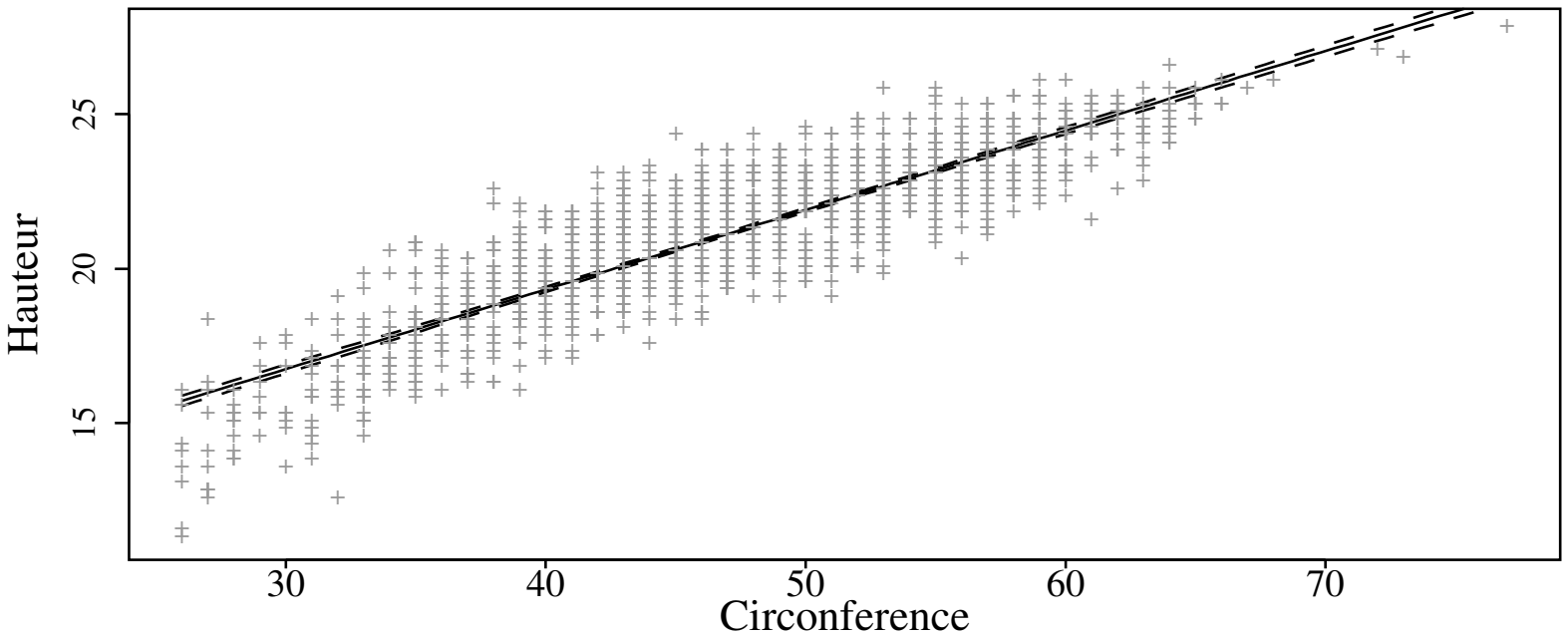
Le résultat

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{1}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}, \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2 \tag{3}$$

L'ajustement linéaire



La prévision d'une hauteur

1. On me donne une circonférence x^*
2. J'estime $\hat{\beta}_0$ et $\hat{\beta}_1$
3. Je redonne la prévision $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$

Que manque t-il ?

Quelle bonne question, que proposez vous comme réponse ?

Que manque t-il ?

Quelle bonne question, que proposez vous comme réponse ?

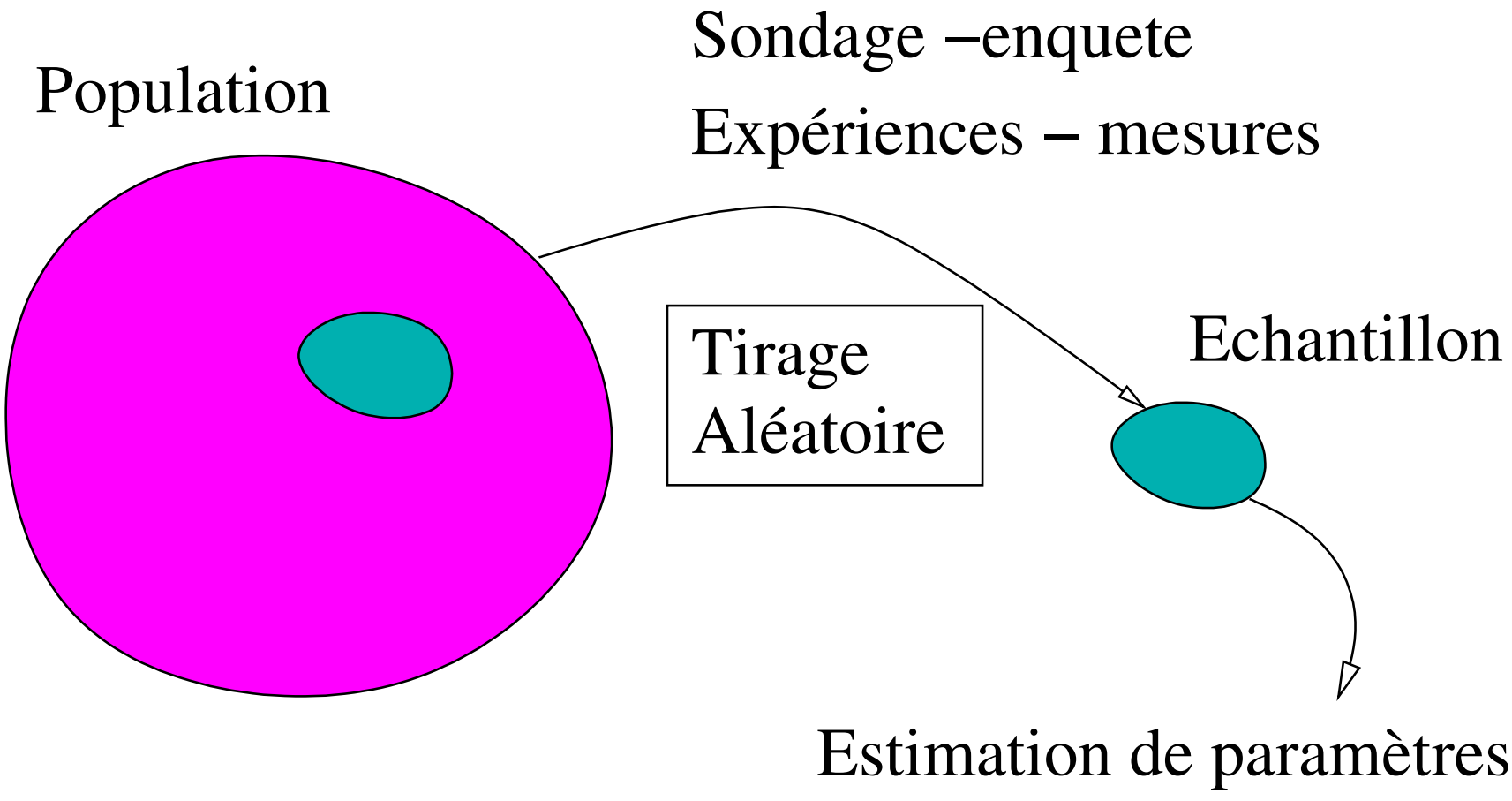
1. Variabilité/Précision de l'estimation

Que manque t-il ?

Quelle bonne question, que proposez vous comme réponse ?

1. Variabilité/Précision de l'estimation
2. Le modèle est-il bon ?

Echantillon et population



Inférence : les enjeux

But A partir de l'échantillon dresser des conclusions valables pour la population

Moyen tirer profit de l'aléa en utilisant des modèles

La précision (la variabilité)

Question

Pourquoi ma prévision est variable ?

Réponse

La précision (la variabilité)

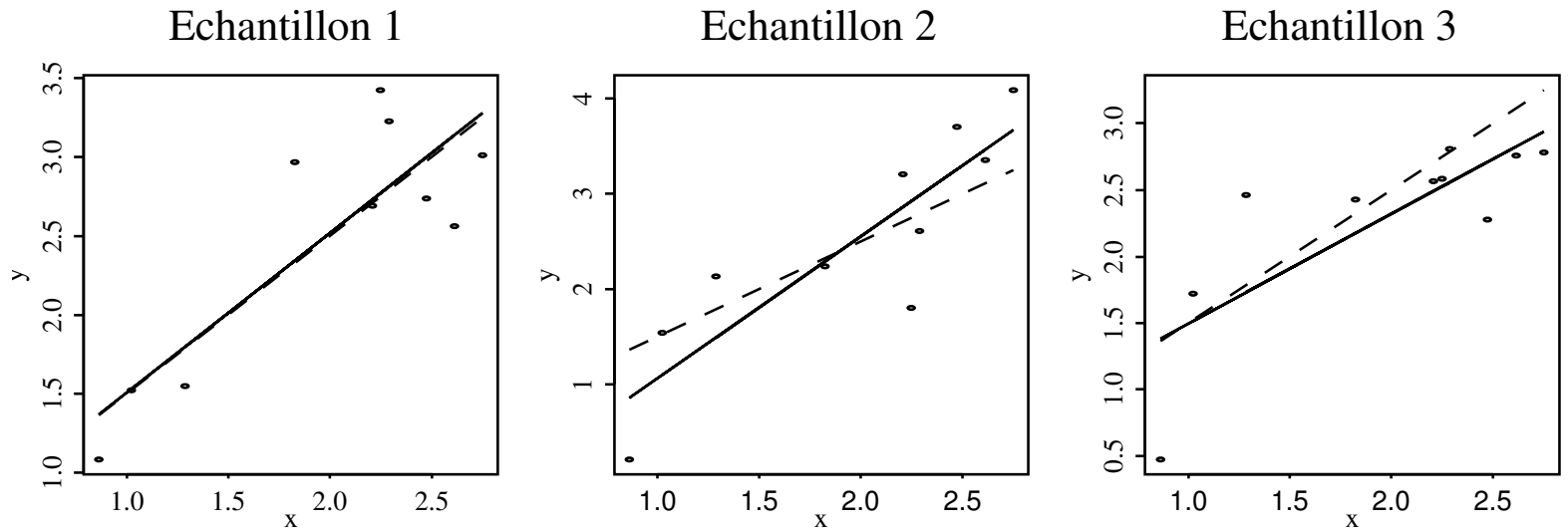
Question

Pourquoi ma prévision est variable ?

Réponse

$\hat{\beta}_0$ et $\hat{\beta}_1$ sont aléatoires

La précision (la variabilité) : illustration



Estimation

$\hat{\beta}_2 \approx 1.01$
$\hat{\beta}_1 \approx 0.499$

Estimation

$\hat{\beta}_2 \approx 1.49$
$\hat{\beta}_1 \approx -0.424$

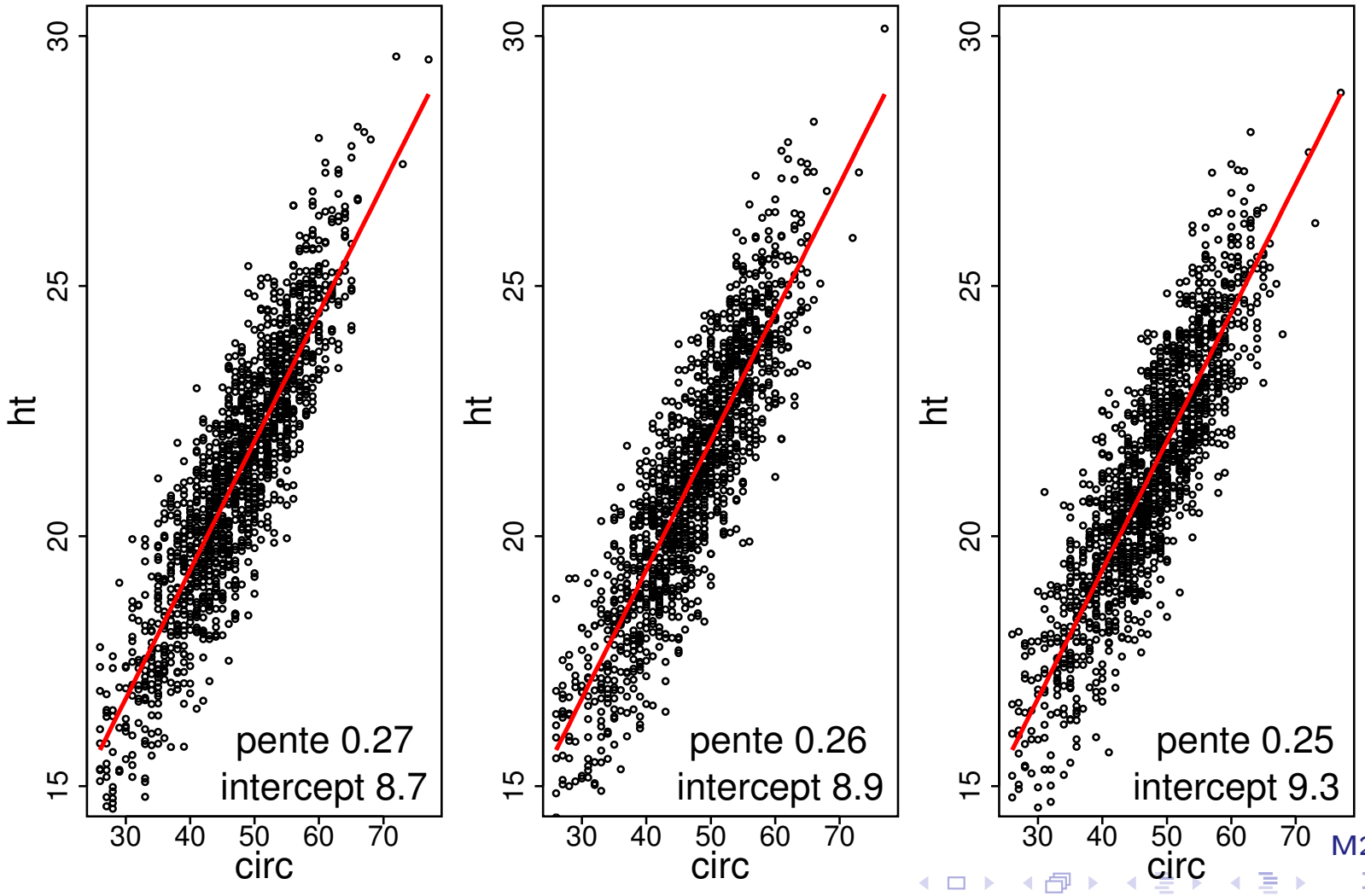
Estimation

$\hat{\beta}_2 \approx 0.825$
$\hat{\beta}_1 \approx 0.669$

Valeurs des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ pour différents échantillons

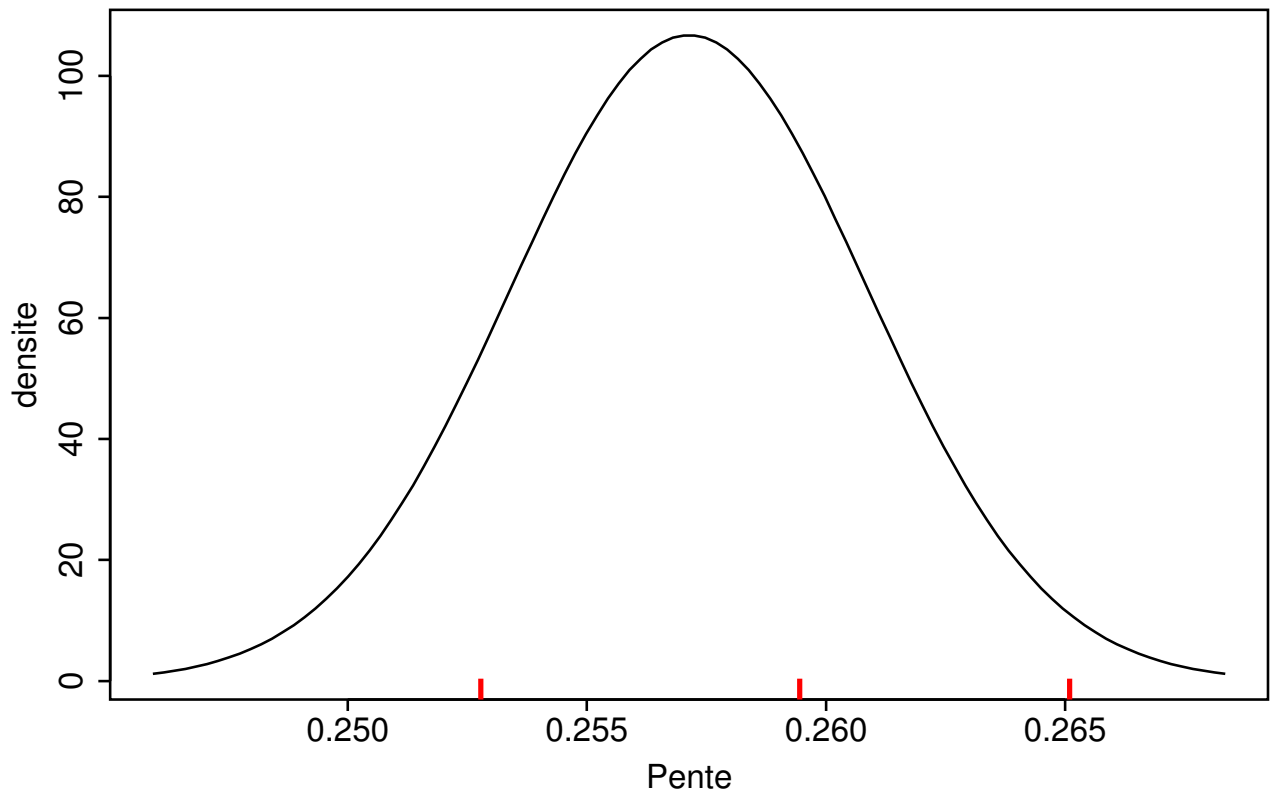
La précision (la variabilité) : illustration 2

3 expériences avec $n = 1429$ mesures :



Loi du paramètre pente

Avec une infinité d'expériences on a :



La précision (la variabilité) : suite

- Espérance, variance de $\hat{\beta}_0$ et $\hat{\beta}_1$
- IC de β_0 et β_1
- IC de y^*

La variabilité de $\hat{\beta}_1$

- Espérance, variance de $\hat{\beta}_1$:

$$\begin{aligned} \mathbb{E}(\hat{\beta}_1) &= \beta_1 \\ \mathbb{V}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

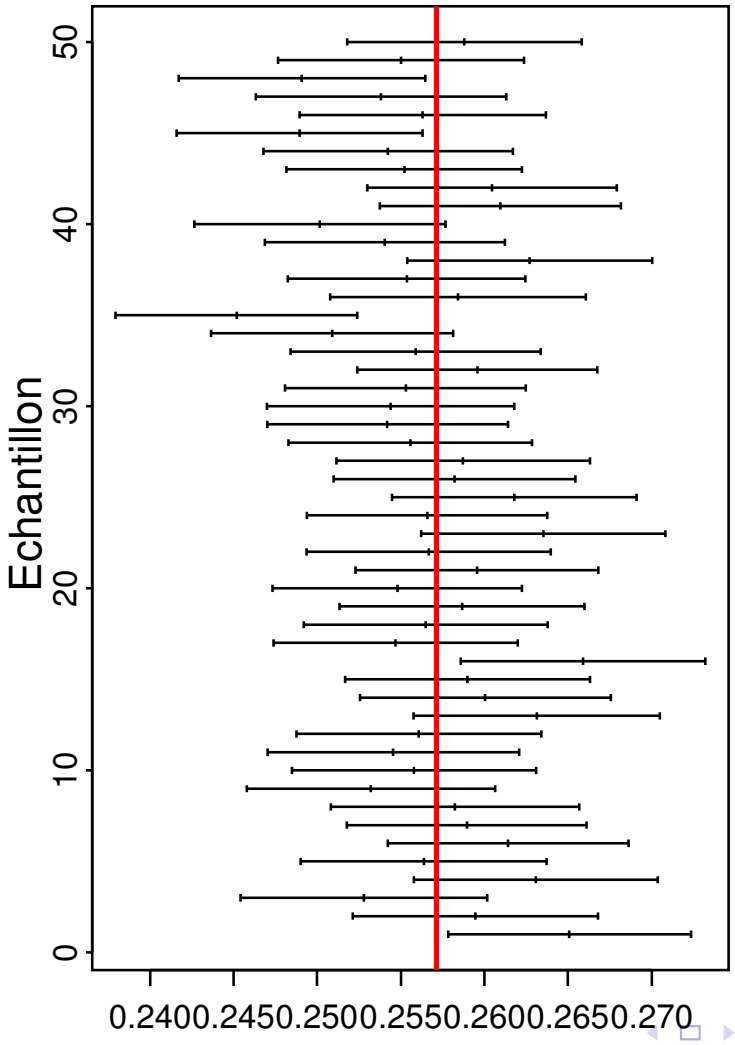
- IC à 95% pour β_1 :

$$\hat{\beta}_1 - t_{0.975}(n-2)\sqrt{\hat{\mathbb{V}}(\hat{\beta}_1)} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.975}(n-2)\sqrt{\hat{\mathbb{V}}(\hat{\beta}_1)}$$

avec un niveau de ...

Intervalles de confiance

Sur 50 expériences de $n = 1429$ données on a :



La variabilité de $\hat{\beta}_0$

- Espérance, variance de $\hat{\beta}_0$:

$$\begin{aligned} \mathbb{E}(\hat{\beta}_0) &= \beta_0 \\ \mathbb{V}(\hat{\beta}_0) &= \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

- IC à 95% pour β_0 :

$$\hat{\alpha} - t_{0.975}(n-2)\sqrt{\hat{\mathbb{V}}(\hat{\beta}_0)} \leq \beta_0 \leq \hat{\beta}_0 + t_{0.975}(n-2)\sqrt{\hat{\mathbb{V}}(\hat{\beta}_0)}$$

avec un niveau de ...

Variabilité des coefficients

1. Diminuer σ^2
2. Augmenter $\sum_{i=1}^{n} (x_i - \bar{x})^2$
   - Augmenter n
   - Augmenter la dispersion
3. Pour $\hat{\beta}_0$: diminuer $\sum_{i=1}^{n} x_i^2$

Variabilité des coefficients

```
> modele1 <- lm(ht~1+circ,data=eucalyptus)
> summary(modele1)

Call:
lm(formula = ht ~ 1+circ, data = eucalyptus)

Residuals:
     Min       1Q   Median       3Q      Max
-4.76589 -0.78016  0.05567  0.82708  3.69129

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.037476   0.179802   50.26   <2e-16 ***
circ        0.257138   0.003738   68.79   <2e-16 ***
---
```

Modélisation avec R

```
> modele1 <- lm(ht~circ,data=eucalyptus)
> summary(modele1)

Call:
lm(formula = ht ~ circ, data = eucalyptus)

Residuals:
     Min       1Q   Median       3Q      Max
-4.76589 -0.78016  0.05567  0.82708  3.69129

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.037476   0.179802   50.26   <2e-16 ***
circ        0.257138   0.003738   68.79   <2e-16 ***
---
```

IC : coefficients et prevision

```
> confint(modele1,level=0.95)
                2.5 %    97.5 %
(Intercept) 8.6847719 9.3901795
circ        0.2498055 0.2644702

> aprevoir=data.frame(circ=c(40,57))
> predict(modele1,newdata=aprevoir,level=0.95,interval="prec
       fit      lwr      upr
1 19.32299 16.96920 21.67678
2 23.69433 21.34010 26.04857
```

Variabilité d'une prévision

Une prévision

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

L'objectif (inconnu)

$$y^* = \beta_0 + \beta_1 x^* + \varepsilon^*$$

Notre but : avoir une différence minimum.
Espérance, variance

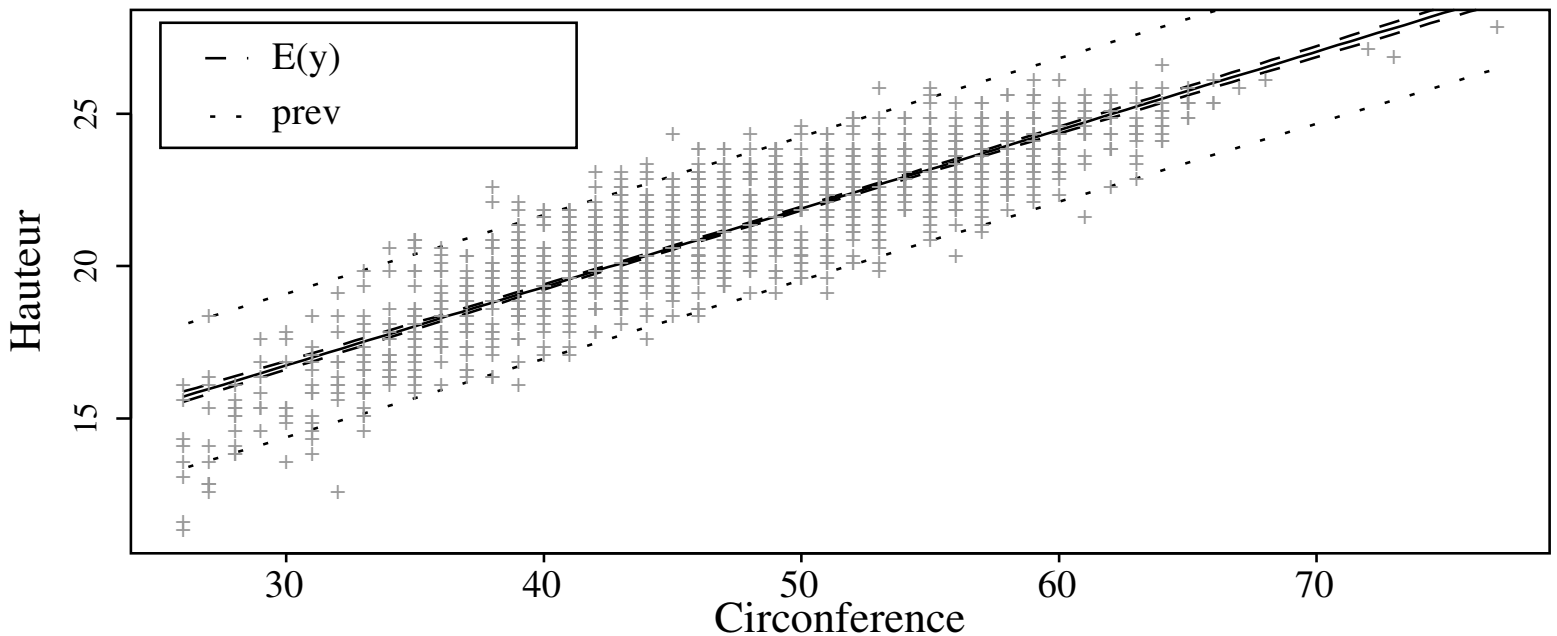
1. $\mathbb{E}(\hat{y}^* - y^*) = 0$
2. $\mathbb{V}(\hat{y}^* - y^*) = \sigma^2 \left(1 + \frac{1}{n} + \frac{\sum_{i=1}^{n} (x^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)$

IC pour une prévision

IC à 95 % pour y^* :

$$\hat{y}^* - t_{0.975}(n-2)\sqrt{\hat{\mathbb{V}}(\hat{y}^* - y^*)} \leq y^* \leq \hat{y}^* + t_{0.975}(n-2)\sqrt{\hat{\mathbb{V}}(\hat{y}^* - y^*)}$$

La précision (la variabilité) : prise en compte



Le modèle est-il bon ?

1. Proposer un autre modèle et comparer
   $\rightarrow$ Test : F ou T

Le modèle est-il bon ?

1. Proposer un autre modèle et comparer
   $\rightarrow$ Test : F ou T
2. Analyser la qualité globale : R^2
   $0 \leq R^2 \leq 1$

Le modèle est-il bon ?

1. Proposer un autre modèle et comparer
   $\rightarrow$ Test : F ou T
2. Analyser la qualité globale : R^2
   $0 \leq R^2 \leq 1$
3. Analyser les résidus :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Residu=Valeur obervée-Valeur prévue

Qualité du modèle par R^2

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Variabilité expliquée / Variabilité totale

- compris entre 0 (pas de lien linéaire) et 1 (lien linéaire)
- comparer les R^2 de modèles avec le même nombre de paramètres.

Problème

Question

Est-ce qu'il existe un lien (causal) entre Circonférence et Hauteur ?

Méthode

- Si le modèle 0

$$y_i = \beta_0 + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma) \quad i = 1, \cdots, n$$

est bon alors pas de lien ($\beta_1 = 0$)

- Si le modèle 1

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma) \quad i = 1, \cdots, n$$

est bon, alors lien (linéaire) ($\beta_1 \neq 0$)

Conclusion

Choix de modèles : méthode naïve

- Estimation β_0, β_1
- Selon la valeur de $\hat{\beta}_1$:
  - $\hat{\beta}_1 = 0$ alors modèle 0
  - $\hat{\beta}_1 \neq 0$ alors modèle 1

Problème : $\hat{\beta}_1$ variable

Question formalisée

Question (Modèle 0 = Modèle 1) $\Leftrightarrow$ ($\beta_1 = 0$)

ou

(Modèle 0 < Modèle 1) $\Leftrightarrow$ ($\beta_1 \neq 0$) ?

Formalisation pour résoudre le pb

1. **Modèle :** Régression linéaire simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma) \quad i = 1, \cdots, n$$

2. **Deux hypothèses :**
   - $\mathcal{H}_0 : \beta_1 = 0$
   - $\mathcal{H}_1 : \beta_1 \neq 0$

But Choisir entre $\mathcal{H}_0$ et $\mathcal{H}_1$ en faisant le moins d'erreur !

Erreur(s) ?

Deux types d'erreurs

1. Erreur de première espèce α : Se tromper alors que $\mathcal{H}_0$ est vrai. Cette erreur est choisie par l'utilisateur (en général 5% ou 1%)
2. Erreur de seconde espèce β : Se tromper alors que $\mathcal{H}_0$ est vrai

Test T de lien linéaire

1. **Modèle :** Régression linéaire simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma) \quad i = 1, \cdots, n$$

2. Choix d'**une erreur de première espèce** $\alpha = 5\%$
3. **Deux hypothèses :**
   - $\mathcal{H}_0 : \beta_1 = 0$
   - $\mathcal{H}_1 : \beta_1 \neq 0$
4. Statistique de test (et loi sous $\mathcal{H}_0$)

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\mathbb{V}}(\hat{\beta}_1)}} \overset{\mathcal{H}_0}{\sim} \mathcal{T}(n-2)$$

5. Observation de la stat. de test : T_{obs}
   $\rightarrow$ probabilité critique (p-value)
6. Conclusion

Tests T avec R

```
> modele1 <- lm(ht~circ,data=eucalyptus)
> summary(modele1)

Call:
lm(formula = ht ~ circ, data = eucalyptus)

Residuals:
     Min       1Q   Median       3Q      Max
-4.76589 -0.78016  0.05567  0.82708  3.69129

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.037476   0.179802   50.26   <2e-16 ***
circ        0.257138   0.003738   68.79   <2e-16 ***
---
```

Choix de modèles et question(s)

Question

Est-ce qu'il existe un lien (causal) entre Circonférence et Hauteur ?

Réponse

- Si le modèle 0

$$y_i = \beta_0 + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma) \quad i = 1, \cdots, n$$

est bon alors pas de lien

- Si le modèle 1

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma) \quad i = 1, \cdots, n$$

est bon, alors lien (linéaire)

Conclusion

Choix de modèles : test F

- Choix d'**une erreur de première espèce** $\alpha = 5\%$
- **Modèle :** Régression linéaire simple

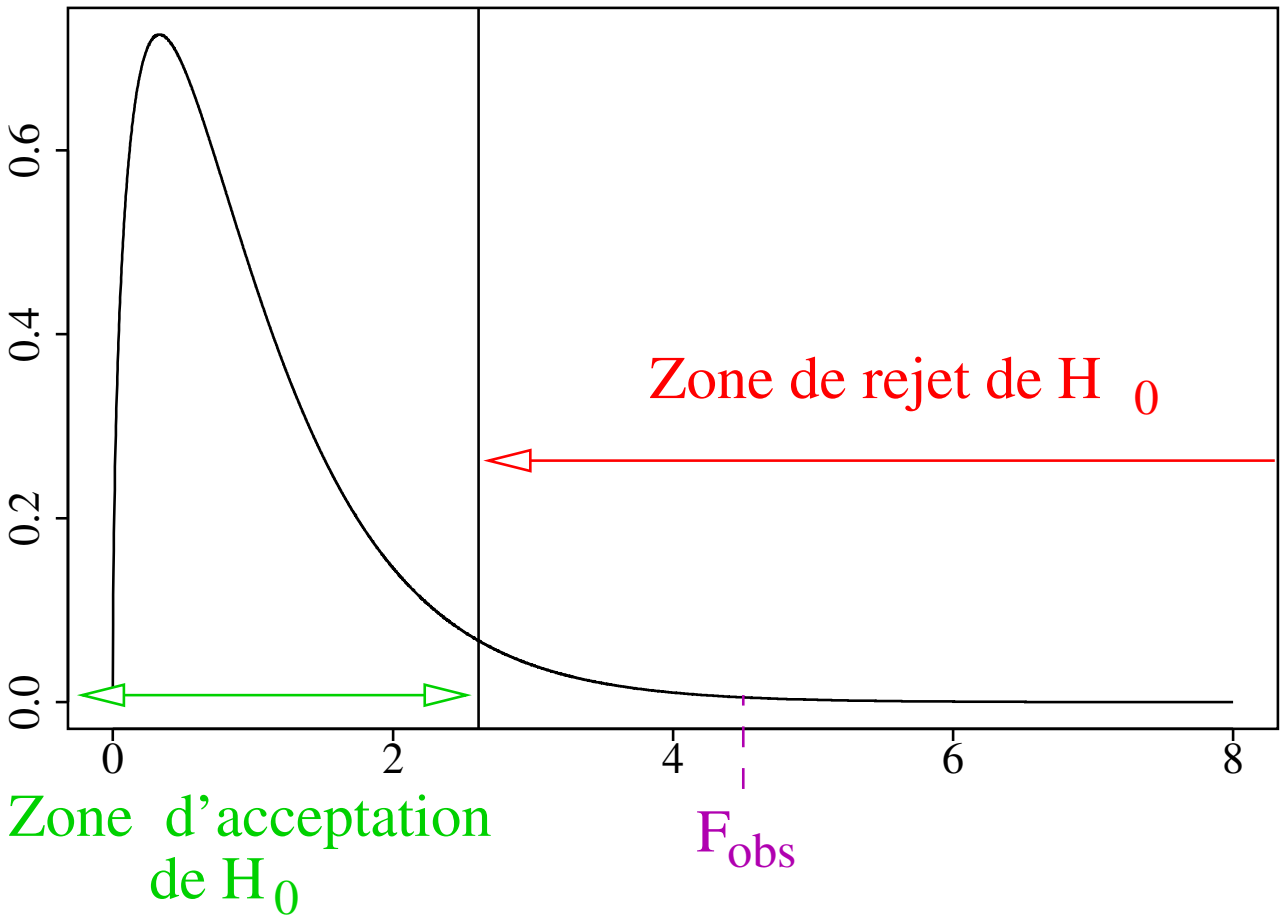
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma) \quad i = 1, \cdots, n$$

- $\mathcal{H}_0$: modèle 0 vrai ; $\mathcal{H}_1$: modèle 1 vrai
- Statistique de test (et loi sous $\mathcal{H}_0$)

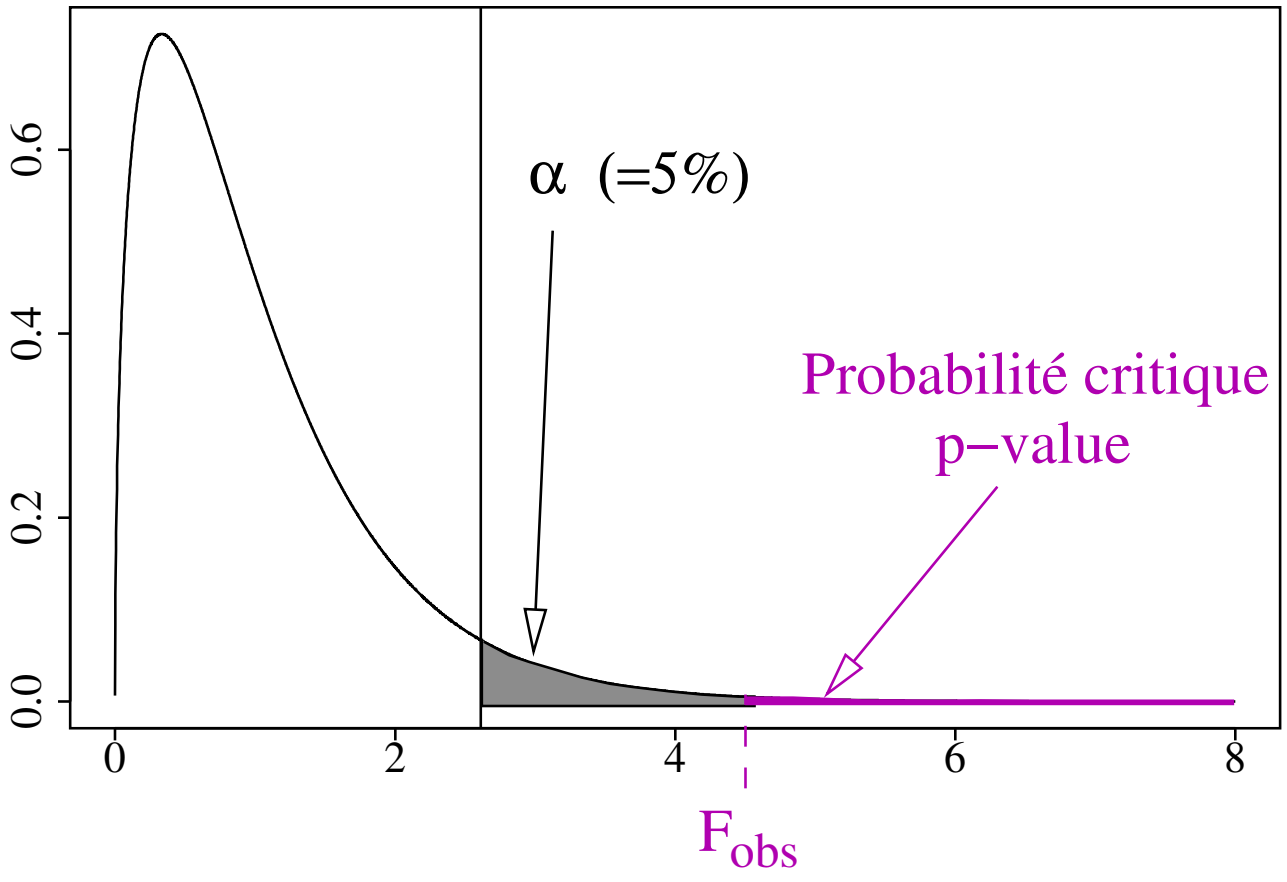
$$F = \frac{R_1^2 - R_0^2}{1 - R_1^2}(n-2) \overset{\mathcal{H}_0}{\sim} \mathcal{F}(1, n-2)$$

- Observation de la stat. de test : F_{obs}
  $\rightarrow$ probabilité critique (p-value)
- Conclusion

test F graphiquement



test *F* graphiquement, probabilité critique



Conclusion du test F

- Si probabilité critique (p-value) $> \alpha$ (erreur 1^{ere} espèce) $\rightarrow \mathcal{H}_0$ conservée
- Si probabilité critique (p-value) $< \alpha$ (erreur 1^{ere} espèce) $\rightarrow \mathcal{H}_0$ repoussée (au profit de $\mathcal{H}_1$)

Test F

On compare 2 modeles

```
> modele0 <- lm(ht~1,data=eucalyptus)
> modele1 <- lm(ht~1+circ,data=eucalyptus)
> anova(modele0,modele1)
Analysis of Variance Table

Model 1: ht ~ 1
Model 2: ht ~ 1 + circ
  Res.Df    RSS   Df Sum of Sq      F    Pr(>F)
1   1428 8857.4
2   1427 2052.1    1    6805.3 4732.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```