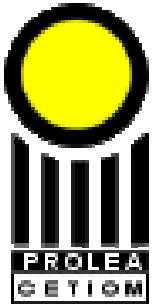
Séminaire *AppliBugs*

Paris, juin 2007

Une approche bayésienne pour évaluer la précision d'un test de diagnostic du sclérotinia du colza

D. Makowski	INRA
J-B. Denis	INRA
L. Ruck	CETIOM
A. Penaud	CETIOM



Qu’est-ce que le colza ?

Qu’est-ce que le sclérotinia ?

- **Maladie du colza induite par un champignon:**

  ***Sclerotinia sclerotiorum*.**

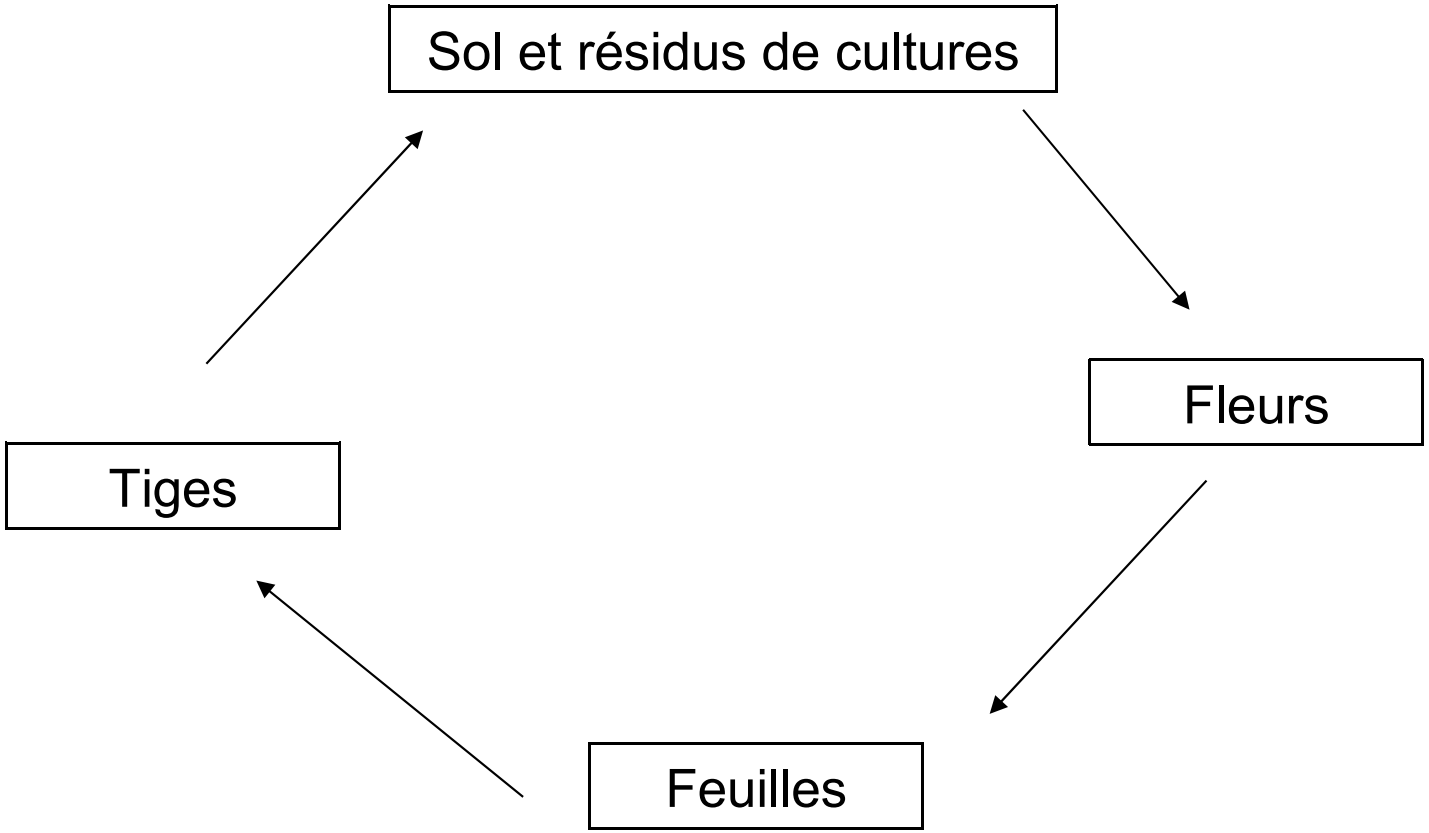
- **Incidence très variable entre sites, entre années.**

- **Pertes de rendement pouvant être importante**

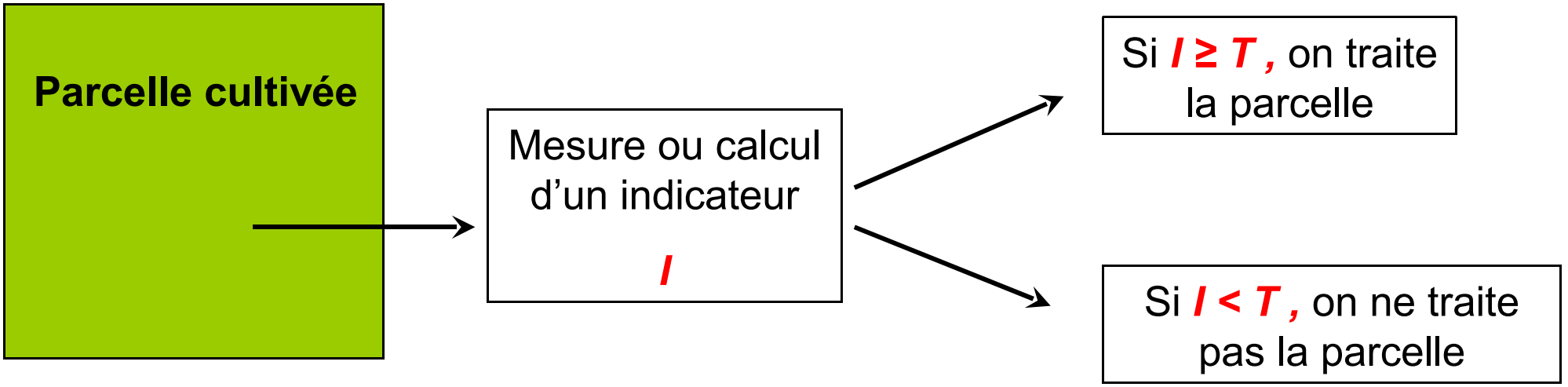
  si *% plantes malades à la récolte* > 10%.

- **Traitement fongicide efficace mais souvent inutile.**

Un cycle simplifié



Principe des tests pour décider de traiter une culture atteinte par une maladie



Intérêt de ces tests

Éviter les traitements systématiques pour

- **faire des économies**
- **limiter l'apparition des résistances des maladies aux traitements**
- **réduire les risques de pollution de l'eau et des sols.**

Tests basés sur un nombre d'organes malades

Exemple 1: Nombre de ***pieds de blé atteint par le piétin verse*** sur un échantillon de 40 pieds.

Exemple 2: Nombre de ***fleurs de colza atteintes par le sclérotinia*** sur un échantillon de 40 fleurs.

Turkington et al., 1991.

Taverne et al. 2003.

Test basé sur un nombre de fleurs de colza malades

Perte de rendement +

Perte de rendement -

Prélèvement de *n* fleurs.

Incubation.

***y* fleurs malades.**

Si $\frac{y}{n} \geq T$**, traitement.**

Questions pratiques

1. Précision de l'indicateur / fréquences des erreurs ?

3. Combien de fleurs ?

3. Quel seuil de décision ?

Les deux types d'erreur du test

Erreur 1:

« Le test recommande *de ne pas traiter la parcelle* alors qu'un *traitement était nécessaire* ».

On a $I \geq T$ (traitement non recommandé)

alors que $D = 1$ (traitement nécessaire)

Erreur 2:

« Le test recommande *de traiter la parcelle* alors qu'un *traitement était inutile* ».

On a $I < T$ (traitement recommandé)

alors que $D = 0$ (traitement inutile)

L'analyse ROC pour un test de diagnostic « agronomique »

$m_{\overline{D}}$ parcelles expérimentales avec *D*=0 (traitement inutile).

m_D parcelles expérimentales avec *D*=1 (traitement nécessaire).

(i). Calcule de la valeur de l'indicateur pour chaque parcelle.

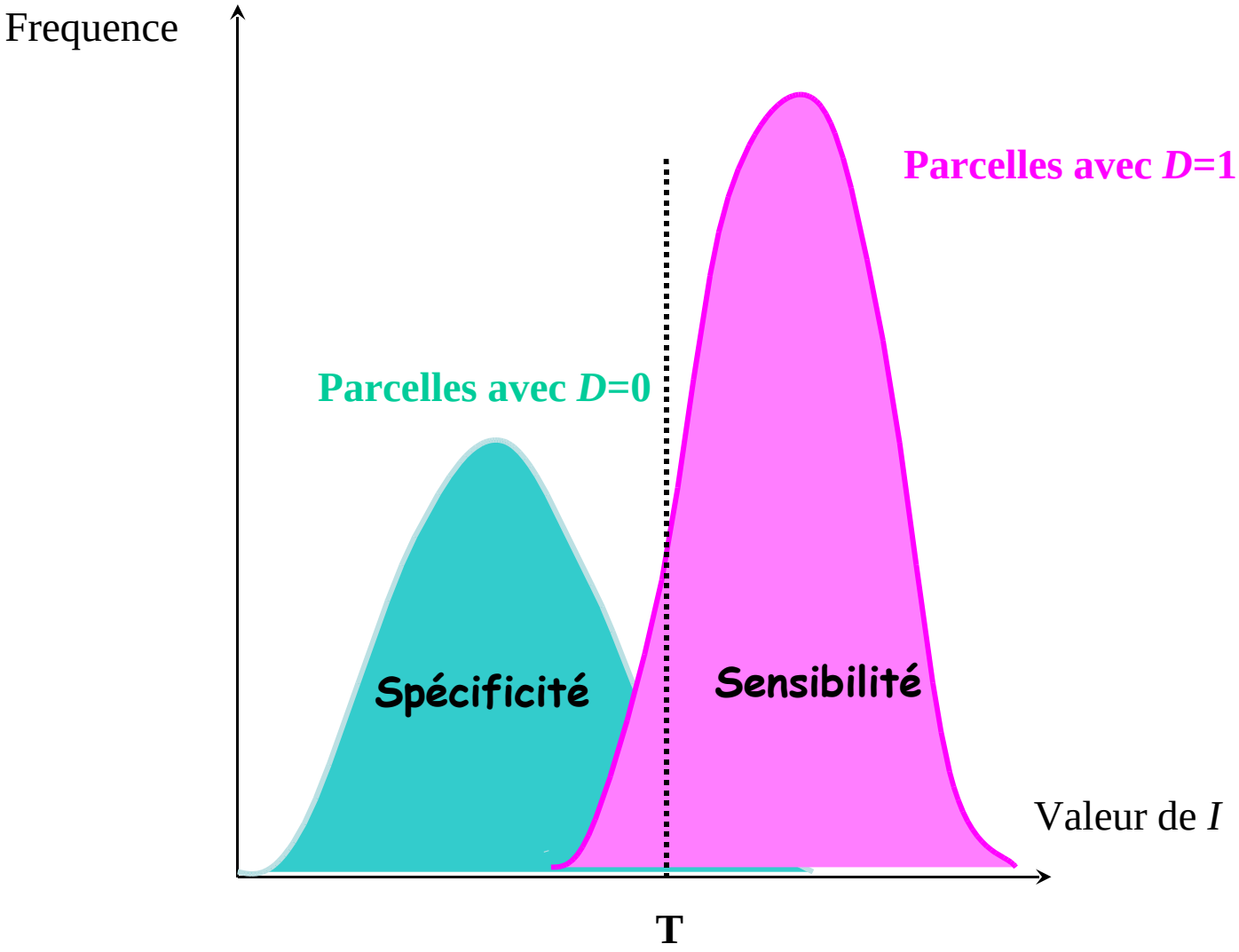
(ii). Définition d'un seuil de décision, *T* .

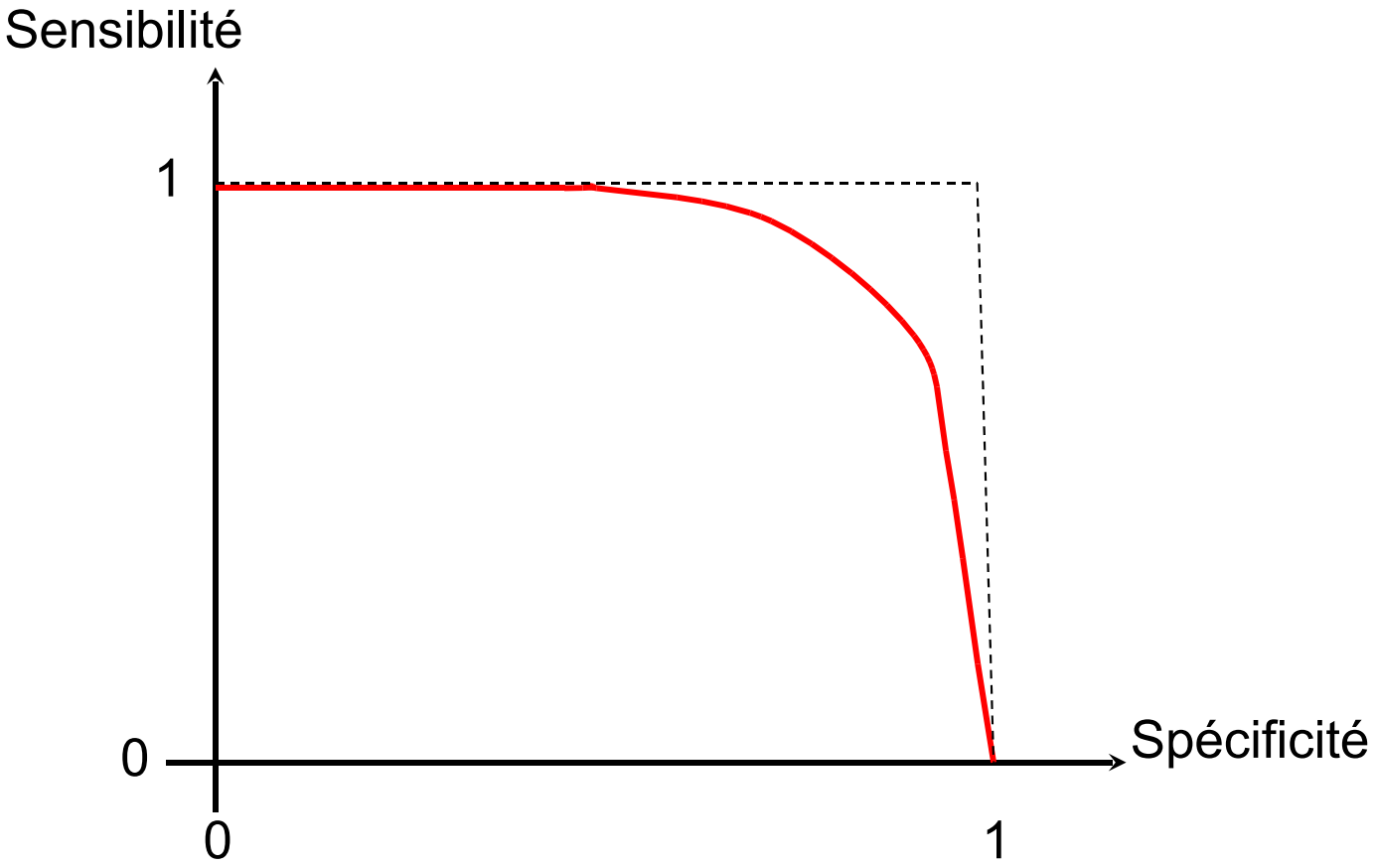
(iii). **Sensibilité** = *Prob*(*I* ≥ *T* | *D=1*)

(iv). **Spécificité** = *Prob*(*I* < *T* | *D=0*)

(v). **Courbe ROC**: Sensibilité(*S*) *versus* 1 - Spécificité(*S*)

(vi). Estimation de l'aire sous la courbe ROC (**ASC**) .


Frequence
Parcelles avec D=1
Parcelles avec D=0
Spécificité
Sensibilité
Valeur de I
T



Aire sous la courbe = $P\left(I_D > I_{\overline{D}}\right)$

Hanley et McNeil, 1982

Comment estimer la sensibilité, la spécificité, l'aire sous la courbe ?

Approche non paramétrique.

+ peu d'hypothèse sur la distribution des mesures

+ facile à appliquer

- peu adaptée pour étudier l'effet de covariables sur la précision du test.

Modélisation.

+ permet d'étudier l'effet de covariables sur la précision du test.

- nécessite des hypothèses sur la distribution des mesures.

Pepe, 1998
O'Malley et al., 2001
Farragi et Reiser, 2002

Un modèle pour étudier la relation entre la taille de l'échantillon de fleurs et la précision du test

Niveau 1: intra parcelle

$$y_{Di} \mid \theta_{Di} \sim Bin(n_i, \theta_{Di}) \qquad i=1, \ldots, m_D,$$

$$y_{\overline{D}j} \mid \theta_{\overline{D}j} \sim Bin(n_j, \theta_{\overline{D}j}) \qquad j=1, \ldots, m_{\overline{D}}$$

Niveau 2: inter parcelles

$$\text{logit}(\theta_{Di}) \mid \mu_D, \sigma_D^2 \sim N(\mu_D, \sigma_D^2) \text{ iid} \qquad i=1, \ldots, m_D,$$

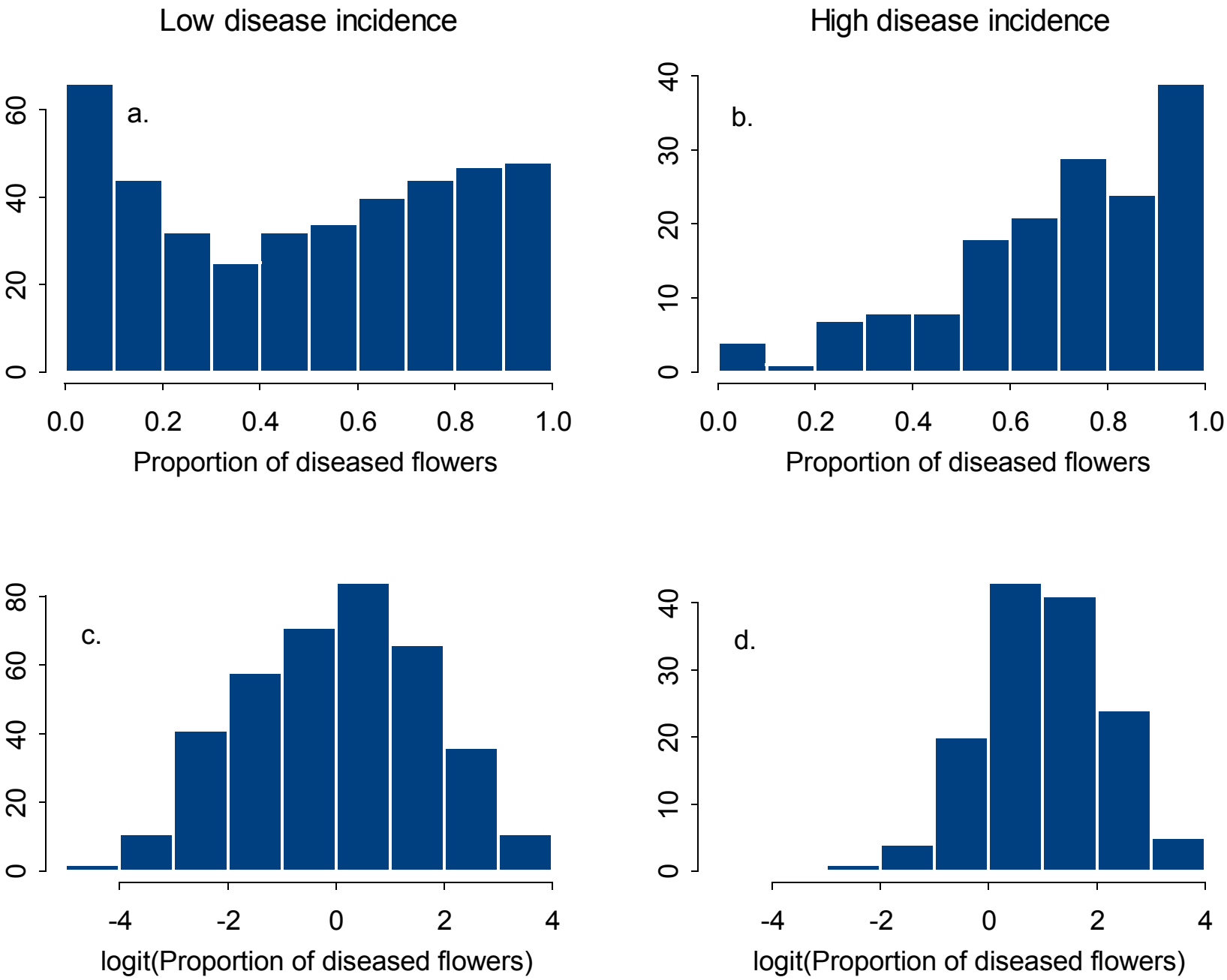
$$\text{logit}(\theta_{\overline{D}j}) \mid \mu_{\overline{D}}, \sigma_{\overline{D}}^2 \sim N(\mu_{\overline{D}}, \sigma_{\overline{D}}^2) \text{ iid} \qquad j=1, \ldots, m_{\overline{D}}.$$

Niveau 3: a priori

$N(0, 10^6)$ pour μ_D et $\mu_{\overline{D}}$, et gamma(0.001, 0.001) pour $1/\sigma_D^2$ et $1/\sigma_{\overline{D}}^2$.

Données expérimentales

Year	Number of plots with low disease incidence	Number of plots with high disease incidence	Total number
2002	30	14	44
2003	40	4	44
2004	65	19	84
2005	202	46	248
2006	75	76	151
All years	412	159	571


Low disease incidence
a.
0
20
40
60
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of diseased flowers
High disease incidence
b.
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of diseased flowers
c.
0
20
40
60
80
-4
-2
0
2
4
logit(Proportion of diseased flowers)
d.
0
10
20
30
40
-4
-2
0
2
4
logit(Proportion of diseased flowers)

Le programme WinBUGS

```
model
# y[i]= number of diseased flowers in the ith plot.
# n[i]= number of collected flowers in the ith plot.
# theta[i]=probability of a diseased flower in the ith plot.
# d[i]=binary variable indicating the status of plot i
# (0=low disease incidence at harvest,
#  1=high disease incidence at harvest)

{
### MODEL ###
## Likelihood ##
     for (i in 1:N) {
          y[i] ~ dbin(theta[i],n[i])
          logit(theta[i])<-mu[i]+eps[i]
          eps[i]~dnorm(0,prec[d[i]+1])
          mu[i]<-muLow*(1-d[i])+muHigh*d[i]
     }
     for (k in 1:2) {var[k]<-1/prec[k]}
## Prior ##
     muLow~dnorm(0.0,1.0E-6)
     muHigh~dnorm(0.0,1.0E-6)
     for (k in 1:2) {prec[k]~dgamma(0.001,0.001)}

……

}
```

```
list(
y=c(5 , 7 , 66 , 10 , 2 , 37 , 14 , 5 , 34 , 40 , 5 , 20 , 58 , 11 , ……..),
n=c(80 , 80 , 80 , 80 , 80 ,40, 40, 40,…….),
d=c(1 , 0 , 1 , 0 , 0 , 0 , 0 , 0 , 1 , 0 , 0 , 0 , 1 , 0 , 1 , 0 , 0 , 0 , 1 , 1 , ……),
N=571)
```

```
list(alpha=c(0,0),prec=c(0.5,0
```

Valeurs estimées des paramètres

Parameter	Mean	Standard deviation	Quantile 0.025	Quantile 0.975
μ_D	1.325	0.146	1.049	1.623
$\mu_{\overline{D}}$	0.022	0.106	-0.1874	0.224
σ_D^2	2.807	0.418	2.103	3.749
$\sigma_{\overline{D}}^2$	4.209	0.357	3.574	4.184

Analyse ROC

$$Se(T) = P\left(\frac{y_D^*}{n} \geq T \right)$$

$$Sp(T) = P\left(\frac{y_{\overline{D}}^*}{n} < T \right)$$

$$A = P\left(\frac{y_D^*}{n} > \frac{y_{\overline{D}}^*}{n} \right) = P\left(y_D^* > y_{\overline{D}}^* \right)$$

$$A_t = P\left(\theta_D^* > \theta_{\overline{D}}^* \right)$$

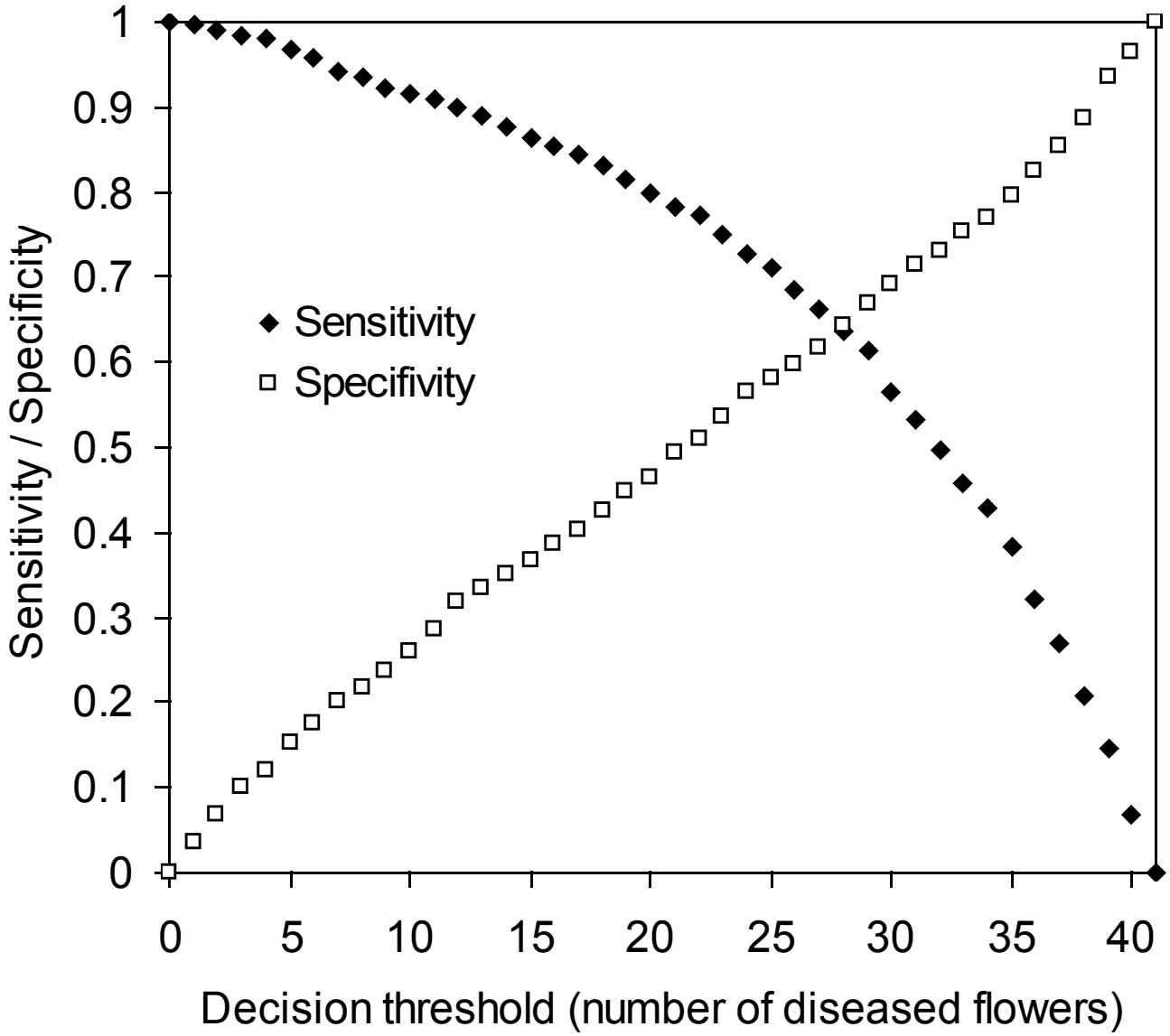
Le programme WinBUGS (suite)

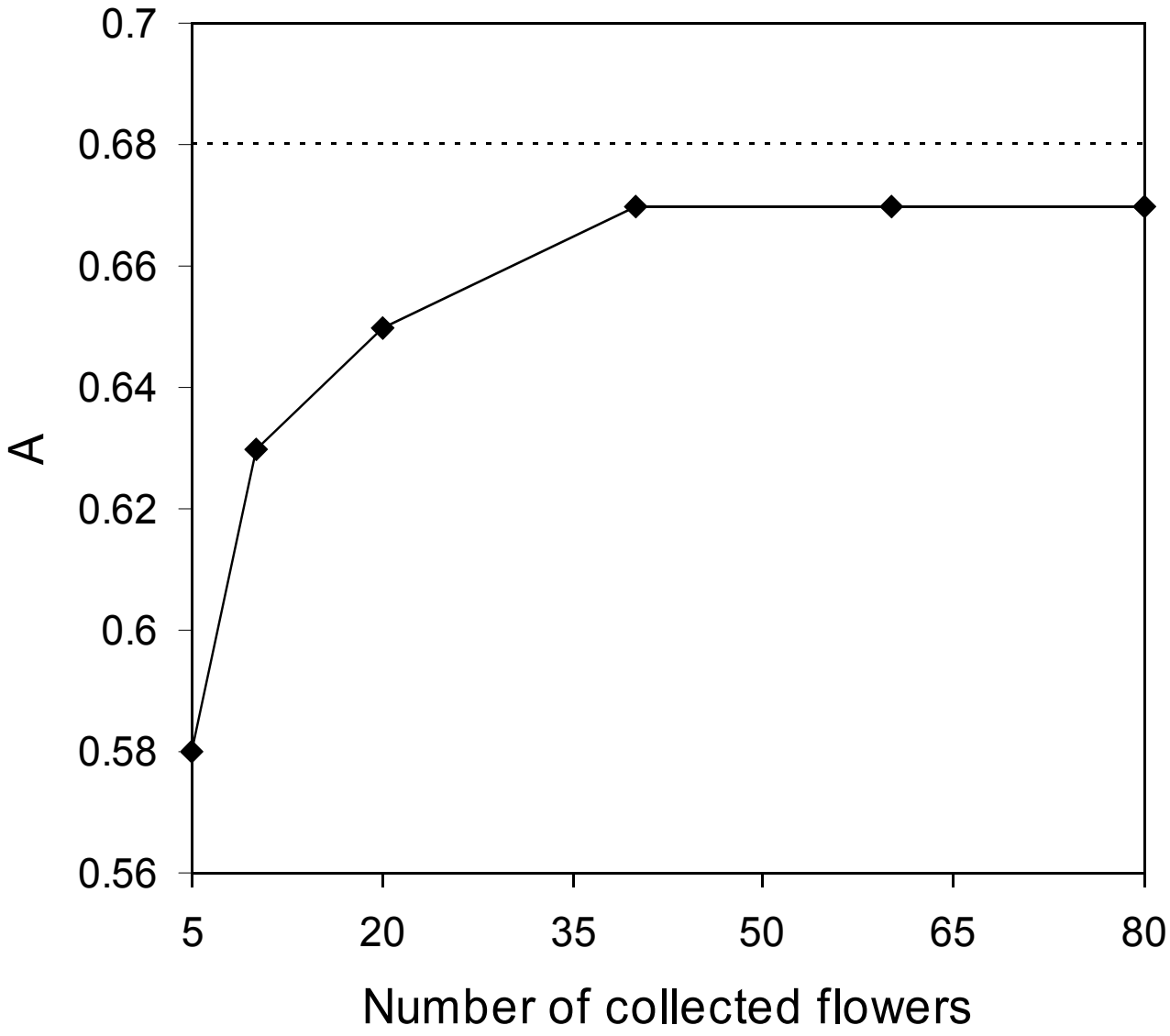
```
……
### ROC ANALYSIS ###
## Generation of values of Theta for d=1 et d=0
     LogitHigh~dnorm(muHigh,prec[2])
     LogitLow~dnorm(muLow,prec[1])
     ThetaHigh<-exp(LogitHigh)/(1+exp(LogitHigh))
     ThetaLow<-exp(LogitLow)/(1+exp(LogitLow))
## Generation of values of y when n=80
     YHigh80~dbin(ThetaHigh,80)
     YLow80~dbin(ThetaLow,80)
## A, Sensibility and Specificity when n=80
     A80<-step(YHigh80-YLow80-1)
     for (l in 1:80) {
          sensib80[l]<-step(YHigh80-l)
          specif80[l]<-step(l-1-YLow80)
     }

##..... the same for other values of n....


}
```

Script pour estimation et analyse ROC

```
display(log)
set.seed(2)
check(c:/David/articles/ScleroBayes/Programmes/ModelRocBin.odc)
data(c:/David/articles/ScleroBayes/Programmes/DataRocBinAll.odc)
compile(1)
inits(1,c:/David/articles/ScleroBayes/Programmes/IniRocBin.odc)
gen.inits()
set(auc1)
set(auc5)
…
set(auc60)
set(auc80)
set(aucOpt)
set(alpha)
set(var)
set(sensib40)
set(specif40)
beg(5000)
thin.samples(40)
update(40000)
stats(*)
```

Résultats


Sensitivity / Specificity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Sensitivity
Specifivity
0
5
10
15
20
25
30
35
40
Decision threshold (number of diseased flowers)


0.7
0.68
0.66
0.64
0.62
0.6
0.58
0.56
A
5
20
35
50
65
80
Number of collected flowers

Simulations

Mean, standard deviation, minimum, maximum, and RMSE of 100 estimated values of *A* derived from 100 simulated datasets including 100, 300 or 500 experimental plots. It was assumed that *n*=40 flowers were collected in each plot. The true value of *A* is equal to 0.672.

Number of plots in each dataset	Min	Max	Mean	RMSE
100	0.543	0.874	0.670	0.059
300	0.581	0.759	0.669	0.038
500	0.608	0.740	0.673	0.028

Programme pour réaliser les simulations

- Fonction R pour générer 100 fichiers de données.

- Fonction R pour générer un script qui applique le programme WinBugs aux 100 fichiers.

- Lancement du script avec WinBugs.

```
### Generation d'un script pour les données simulées de Sclerotinia
#source("c:\\David\\articles\\ScleroBayes\\Programmes\\GenerationScript.txt")
setwd("c:/David/articles/ScleroBayes/Programmes")
Nsim<-100
NomFichier<-"ScriptScleroSimul.txt"
cat(" ",file=NomFichier)

for (j in 1:Nsim) {
Nom.j<-paste("data(c:/David/articles/ScleroBayes/Programmes/","DataScleroSim",j,".txt)",sep="")
cat("display(log)","\n",file=NomFichier,append=T)
cat("set.seed(1)","\n",append=T,file=NomFichier)
cat("check(c:/David/articles/ScleroBayes/Programmes/ModelRocBin.odc)","\n",append=T,file=NomFichier)
cat(Nom.j,"\n",append=T,file=NomFichier)
cat("compile(1)","\n",append=T,file=NomFichier)
cat("inits(1,c:/David/articles/ScleroBayes/Programmes/IniRocBin.odc)","\n",append=T,file=NomFichier)
cat("gen.inits()","\n",append=T,file=NomFichier)
cat("set(auc40)","\n",append=T,file=NomFichier)
cat("beg(5000)","\n",append=T,file=NomFichier)
cat("thin.samples(40)","\n",append=T,file=NomFichier)
cat("update(40000)","\n",append=T,file=NomFichier)
cat("stats(*)","\n",append=T,file=NomFichier)
cat(" ","\n",append=T,file=NomFichier)
}
cat("save(resultatSimul.txt)","\n",append=T,file=NomFichier)
```

```
…
display(log)
set.seed(1)
check(c:/David/articles/ScleroBayes/Programmes/ModelRocBin.odc)
data(c:/David/articles/ScleroBayes/Programmes/DataScleroSim100.txt)
compile(1)
inits(1,c:/David/articles/ScleroBayes/Programmes/IniRocBin.odc)
gen.inits()
set(auc40)
beg(5000)
thin.samples(40)
update(40000)
stats(*)

save(resultatSimul.txt)
```

Conclusions et perspectives

- Définition d'un modèle permettant d'évaluer des tests de diagnostic basés sur un nombre d'organes malades.

- Application à un test pour le sclérotinia.

- Les résultats montrent que le test est assez imprécis.

- Des échantillons de 40 fleurs sont suffisants.

- Le test ne peut pas être amélioré en augmentant le nombre de fleurs.

- Prise en compte de **l'incertitude** dans la **variable de référence**.