Expliquer une probabilité

Régression logistique avec les données Titanic

Aurélien Nicosia

À la fin de la matinée

Vous saurez expliquer une probabilité, interpréter un rapport de cotes et évaluer un modèle sans confondre classement, calibration et décision.

Une matinée, trois productions

1. Contester une explication simple
2. Recoder sans changer les prédictions
3. Évaluer et choisir un seuil

> Chaque mission ajoute une preuve et une limite à la fiche finale.

Une probabilité de 0,70 signifie-t-elle que la personne survivra?

Votez, puis précisez ce qui manque pour répondre.

Un jeu facile à lire, mais limité

`titanic_train` contient 891 passagers et 12 variables.

- survie enregistrée;
- classe;
- sexe;
- âge;
- tarif;
- proches à bord.

> Il s'agit d'un échantillon d'apprentissage, pas d'un manifeste complet.

Une ligne, une réponse, un événement

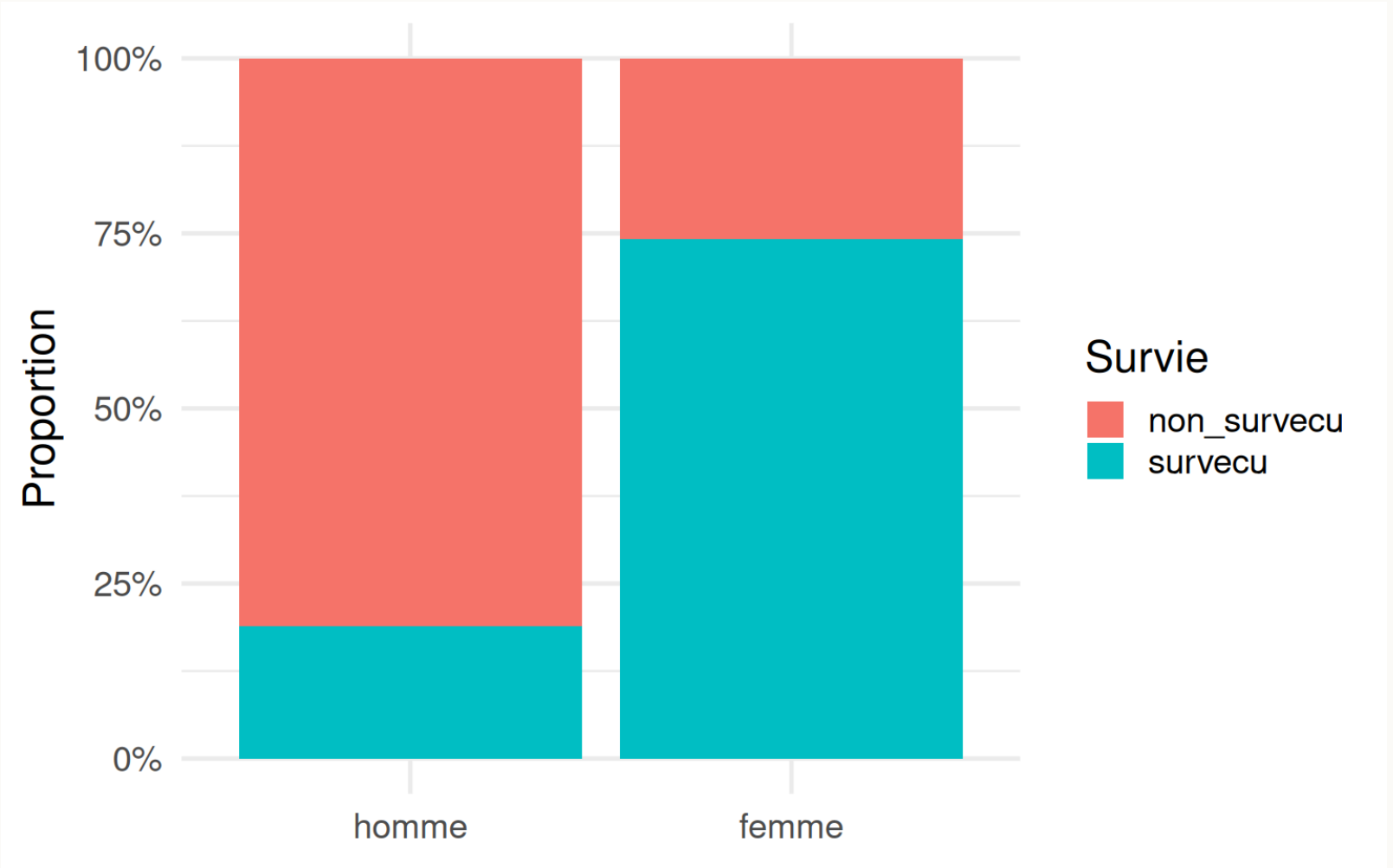
Élément	Définition
Unité	un passager de l’échantillon
Réponse	`survie`, codée 0 ou 1
Événement positif	`survecu`
Population immédiate	les 891 lignes observées

La catégorie majoritaire atteint déjà 61,6 %

```
       verite   n proportion
1 non_survecu 549  0.6161616
2     survecu 342  0.3838384
```

Une exactitude doit être comparée à une règle simple.

Les proportions diffèrent selon le sexe enregistré



La classe structure aussi les proportions

Décrire n’est pas expliquer causalement

Les différences observées peuvent refléter:

- des règles et pratiques historiques;
- des caractéristiques liées entre elles;
- des variables absentes;
- la construction de l’échantillon;
- des valeurs manquantes.

> Le modèle décrit des associations conditionnelles dans les données.

L’âge manque pour 177 passagers

```
  lignes age_manquant lignes_completes
1    891          177              714
```

L’analyse en cas complets est simple, mais elle change l’échantillon.

Une droite peut sortir de l’intervalle

Une régression linéaire sur une réponse 0 ou 1:

1. peut prédire moins de 0 ou plus de 1;
2. impose une variance constante;
3. ignore la saturation près de 0 et de 1.

> Nous gardons un prédicteur linéaire, mais nous changeons d’échelle.

La réponse suit une loi de Bernoulli

$$Y_i \mid X_i \sim \text{Bernoulli}(p_i)$$

Pour chaque profil X_i :

- Y_i vaut 0 ou 1;
- p_i est une probabilité conditionnelle;
- la variance dépend de p_i .

Probabilité, cote et log-cote

$$\text{cote} = \frac{p}{1-p} \qquad \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Probabilité	Cote	Log-cote
0,20	0,25	-1,39
0,50	1,00	0
0,80	4,00	1,39

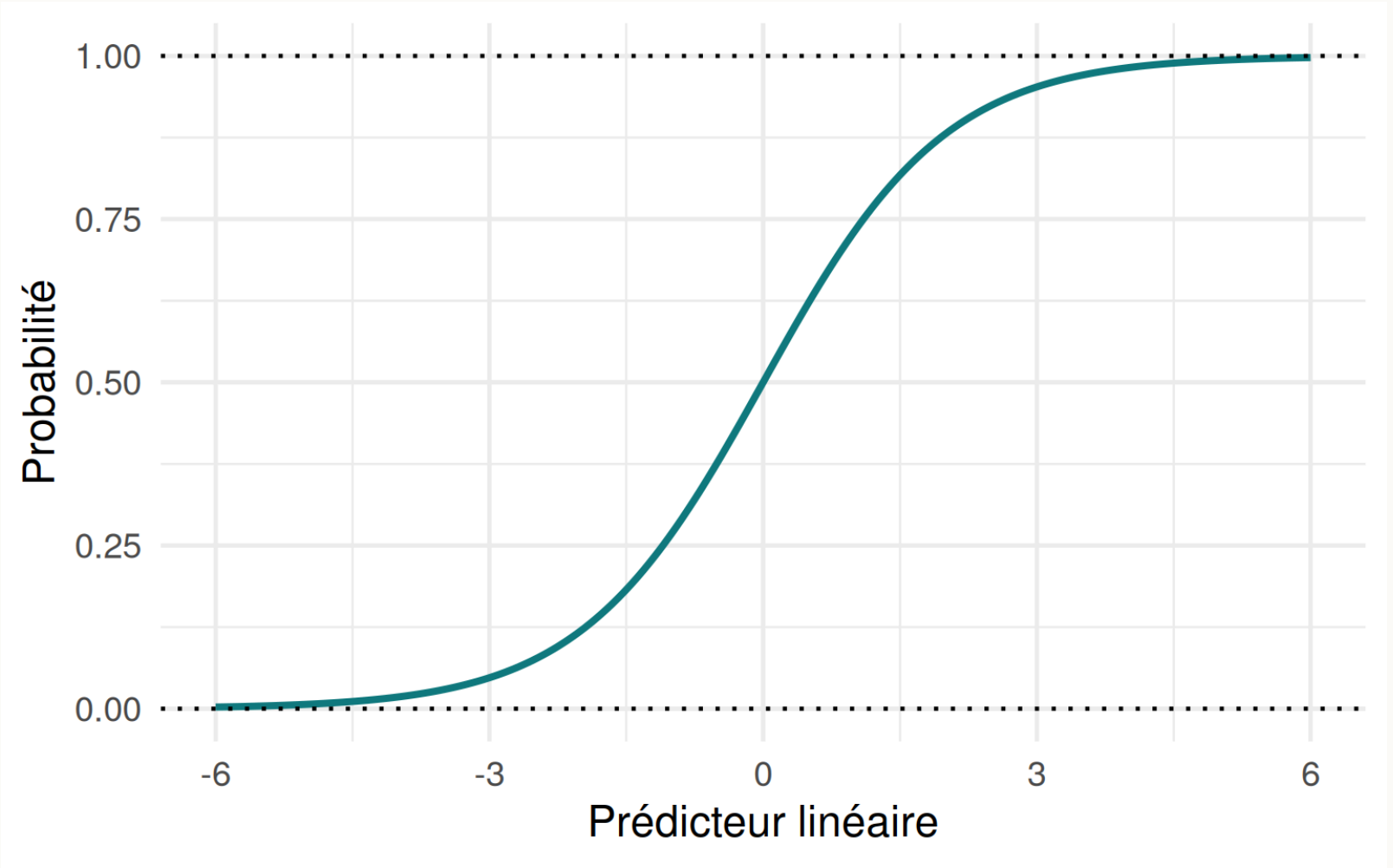
Une cote de 3 n’est pas une probabilité de 3

$$p = \frac{\text{cote}}{1 + \text{cote}}$$

Une cote de 3 correspond à une probabilité de 0,75.

Une cote de 0,5 correspond à une probabilité d’environ 0,33.

La fonction logistique borne les prédictions



Le modèle est linéaire sur la log-cote

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

Puis:

$$p_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}.$$

La vraisemblance remplace les moindres carrés

`glm()` choisit les coefficients qui rendent les réponses observées les plus plausibles sous le modèle.

- survivant avec probabilité élevée: contribution favorable;
- non-survivant avec probabilité faible: contribution favorable;
- prédiction confiante et fausse: forte pénalité.

Les catégories de référence organisent la comparaison

Dans le modèle commun:

- `homme` est la référence pour `sexe`;
- `3e` est la référence pour `classe`;
- l'âge s'interprète par année;
- le tarif s'interprète par unité monétaire.

> Un coefficient catégoriel n'existe pas sans catégorie de référence.

Le jeu test reste fermé

```
# A tibble: 2 × 3
  ensemble         n proportion_survecu
  <chr>        <int>              <dbl>
1 Entraînement   535              0.406
2 Test           179              0.408
```

Le modèle apprend sur 535 lignes et sera évalué sur 179 lignes réservées.

Mission 1: la classe suffit-elle?

En 25 minutes, produisez:

1. un graphique à l'intérieur des classes;
2. un tableau sur les âges manquants;
3. deux formules candidates;
4. un verdict de quatre phrases.

Ouvrir la mission 1

Débrief: persistance n’est pas causalité

$$\text{survie} \sim \text{sexe} + \text{classe} + \text{âge} + \text{tarif}$$

Nous cherchons des associations conditionnelles lisibles, pas le modèle historique définitif.

L’écart associé au sexe enregistré persiste dans chaque classe, mais son ampleur varie. L’interaction `sexe * classe` devient une hypothèse à évaluer, pas une conclusion automatique.

Un appel suffit pour ajuster

```
modele_logistique <- glm(
  survie ~ sexe + classe + age + tarif,
  data = entrainement,
  family = binomial()
)
```

`family = binomial()` utilise le lien logit par défaut.

L’exponentielle donne les rapports de cotes

terme	rapport de cotes	borne basse	borne haute
sexefemme	11.46	7.18	18.29
classe2e	3.81	2.20	6.60
classe1re	12.15	5.70	25.86
age	0.96	0.95	0.98
tarif	1.00	1.00	1.01

« Identiques » doit être plausible

Pour `sexefemme`:

> À classe, âge et tarif identiques, les cotes de survie enregistrée des femmes sont estimées à environ 11.5 fois celles des hommes.

Ce nombre n’est pas une différence de probabilités.

Le tarif est répété pour plusieurs personnes partageant un numéro de billet. Il ne mesure pas automatiquement la richesse personnelle. Comme la classe et le tarif sont liés, leur comparaison conditionnelle repose sur la zone où leurs valeurs se chevauchent.

Dix années donnent une unité plus lisible

```
# A tibble: 2 × 2
  comparaison rapport_cotes
  <chr>               <dbl>
1 1 année             0.963
2 10 années           0.688
```

L’unité d’interprétation doit être substantiellement compréhensible.

Le même rapport de cotes produit des écarts différents

```
# A tibble: 2 × 2
  profil           probabilite
  <chr>                  <dbl>
1 Homme, 3e classe      0.0862
2 Femme, 3e classe      0.520
```

Significativité, importance et prédiction

Élément	Ce qu'il renseigne
Rapport de cotes	taille et direction conditionnelles
Intervalle	incertitude d'estimation
Valeur p	compatibilité avec l'absence d'association
Jeu test	performance hors échantillon
Calibration	qualité des probabilités

> Aucun élément ne répond seul à toutes les questions.

Mission 2: changer le codage

En 25 minutes, produisez:

1. deux paramétrisations comparées;
2. une vérification des probabilités;
3. deux profils inédits;
4. un mémo corrigé.

Ouvrir la mission 2

Pause

Revenez avec une phrase:

> Un coefficient peut changer sans que les prédictions…

Les probabilités sont évaluées sur le test

```
predictions_test <- test |>
  mutate(
    probabilite_survie = predict(
      modele_logistique,
      newdata = test,
      type = "response"
    )
  )
```

Une probabilité élevée n'est pas une certitude. Sa qualité doit être vérifiée.

Le seuil fabrique une classe

$$\widehat{Y} = \begin{cases} \text{survecu}, & \hat{p} \geq s, \\ \text{non\_survecu}, & \hat{p} < s. \end{cases}$$

Changer s ne change pas les probabilités, seulement les classes.

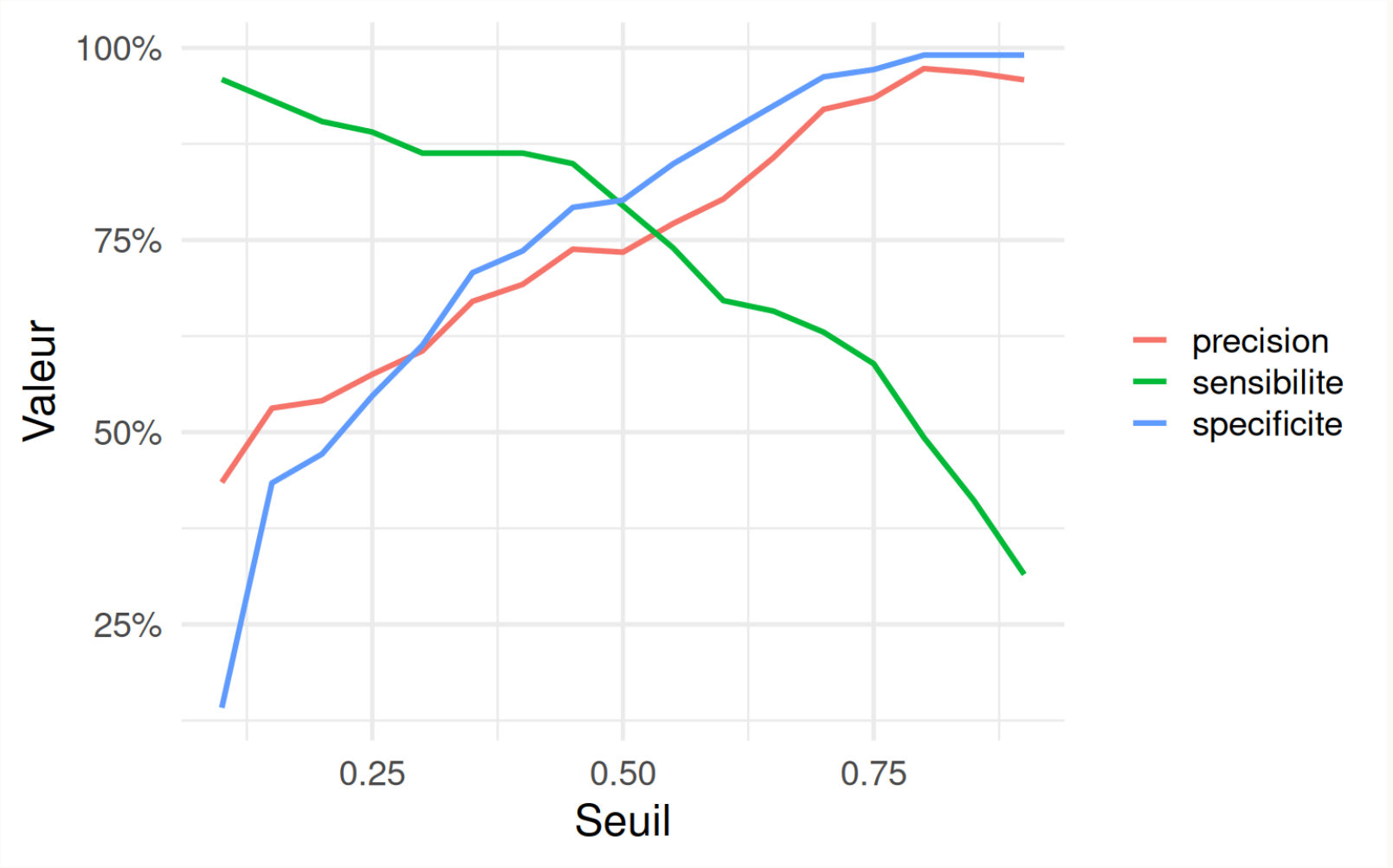
La matrice nomme quatre résultats

```
             Truth
Prediction    non_survecu survecu
  non_survecu          85      15
  survecu              21      58
```

Chaque mesure change de dénominateur

Mesure	Point de départ
Exactitude	toutes les observations
Sensibilité	survivants enregistrés
Spécificité	non-survivants enregistrés
Précision	prédictions positives

Monter le seuil change le compromis



Mission finale: évaluer et choisir

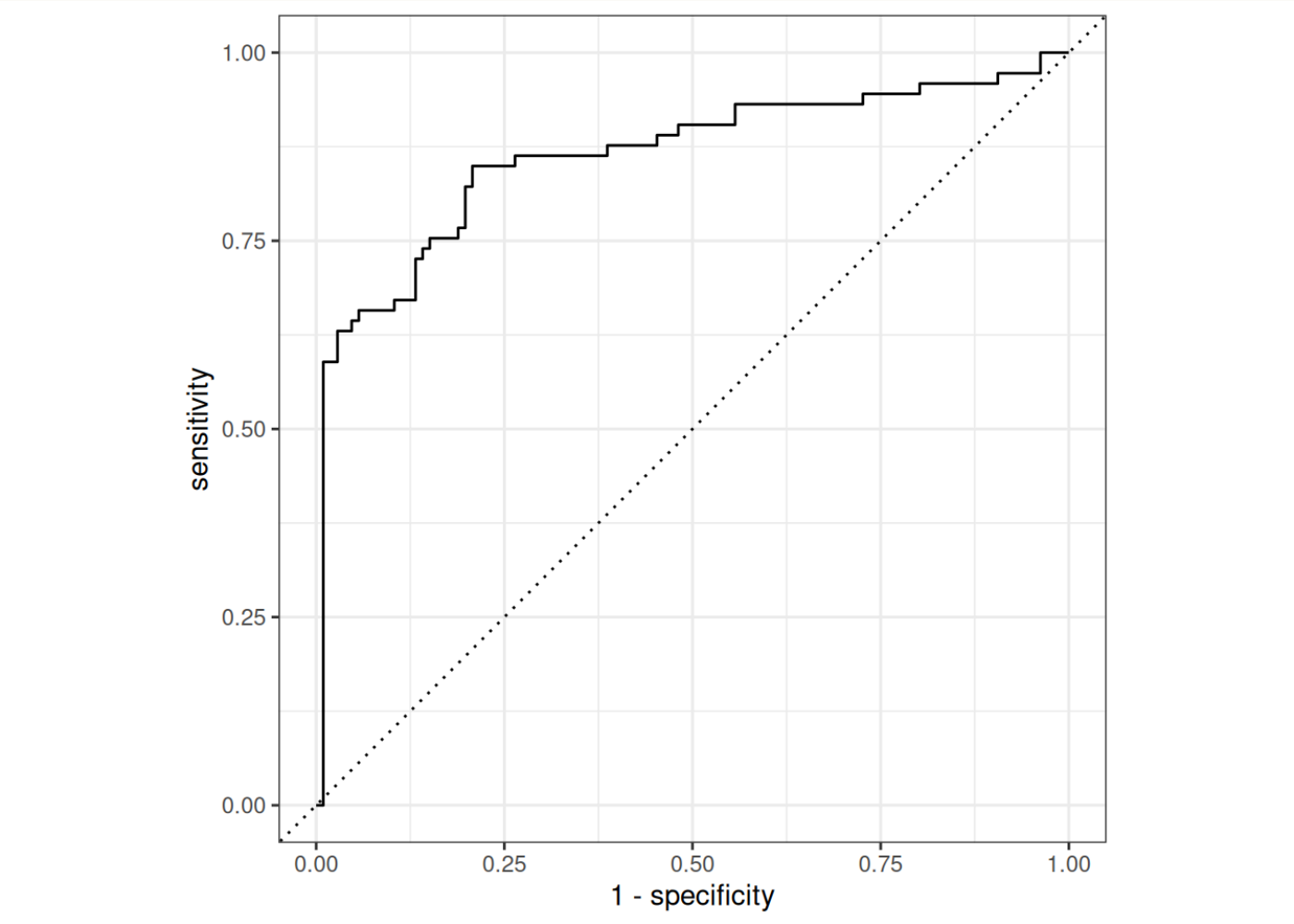
En 25 minutes:

1. évaluez le seuil 0,50;
2. comparez 0,30, 0,50 et 0,70;
3. examinez ROC et calibration;
4. consignez le seuil, le mandat et une limite.

Le Brier et la fiche en six phrases seront repris collectivement après la mission.

Ouvrir la mission finale

La courbe ROC parcourt tous les seuils



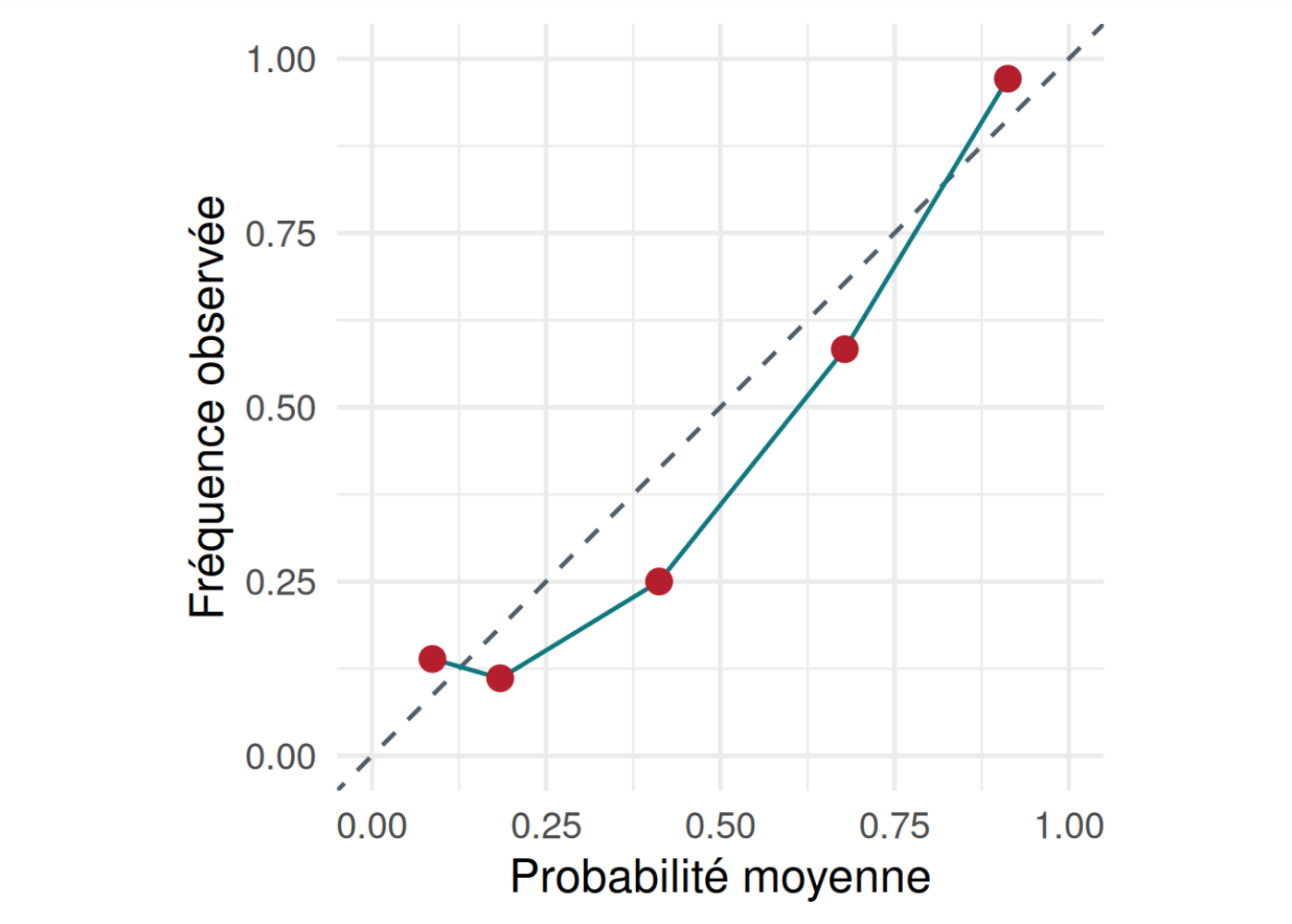
L’aire ROC mesure un classement

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.863
```

Elle compare le score d’un positif et celui d’un négatif choisis au hasard.

Elle ne garantit ni calibration, ni seuil utile, ni causalité.

La calibration vérifie les probabilités



Le Brier pénalise les probabilités fausses

```
  brier_modele brier_reference
1    0.1391384        0.241508
```

Le modèle réduit l’erreur quadratique par rapport à une probabilité constante.

> Discrimination et calibration restent deux propriétés distinctes.

Un appel réussi ne suffit pas

Vérifiez:

1. unité, événement et population;
2. indépendance suffisante des observations;
3. forme linéaire sur le logit;
4. absence de séparation complète;
5. stabilité des coefficients et cas influents;
6. performance sur des données réservées;
7. pertinence pour l’usage visé.

La fiche modèle en six phrases

1. Unité, échantillon et réponse
2. Prédicteurs et valeurs manquantes
3. Association conditionnelle
4. Discrimination et calibration
5. Seuil, mandat et erreur acceptée
6. Limite historique, causale ou de transfert

Sept questions avant de conclure

1. Quel événement est positif?
2. Quelle population est couverte?
3. Comment les valeurs manquantes sont-elles traitées?
4. Quelles sont les références?
5. Le test est-il resté fermé?
6. Classement et calibration sont-ils suffisants?
7. Quel seuil, quelle erreur et quelle limite?