SYS865 Inférence statistique avec programmation R

Ornwipa Thamsuwan

31 janvier 2024

Plan de la séance

Plan de la séance

- Récap
  - Échantillonage
  - Théorème Central Limite
  - Intervalle de confiance
- Test d'hypothèse
  - Types d'erreur
  - Test de moyenne
  - Test de variance
  - Test de deux populations
  - Test nonparamétrique

Récap et matière à réflexion

Récap et matière à réflexion

Réponses anonymes

Go to wooclap.com

Enter the event code FFBQSE

Figure 1: Lien à l'activité sur Wooclap

Échantillonage

Characteristiques de l'échantillonage probabilistique

1. **Sélection aléatoire** : Les individus sont choisis de manière aléatoire, ce qui assure l'impartialité dans la sélection.

Échantillonage

Characteristiques de l'échantillonage probabilistique

1. **Sélection aléatoire** : Les individus sont choisis de manière aléatoire, ce qui assure l'impartialité dans la sélection.

2. **Probabilité égale ou connue** : Chaque membre de la population a une chance égale ou connue d'être inclus dans l'échantillon. Ça permet d'avoir une représentation équitable de la population.

Échantillonage

Characteristiques de l'échantillonage probabilistique

1. **Sélection aléatoire** : Les individus sont choisis de manière aléatoire, ce qui assure l'impartialité dans la sélection.

2. **Probabilité égale ou connue** : Chaque membre de la population a une chance égale ou connue d'être inclus dans l'échantillon. Ça permet d'avoir une représentation équitable de la population.

3. **Représentativité** : L'échantillon a de fortes chances d'être représentatif de la population globale. Cela rend possible de généraliser les résultats de l'échantillon à l'ensemble de la population.

Échantillonage (suite)

Characteristiques de l'échantillonage probabilistique

4. **Inférence statistique** : Ces méthodes permettent de calculer des erreurs d'échantillonnage, des intervalles de confiance et de réaliser des tests de significativité. Cela offre la possibilité de tirer des conclusions statistiques sur la population à partir de l'échantillon.

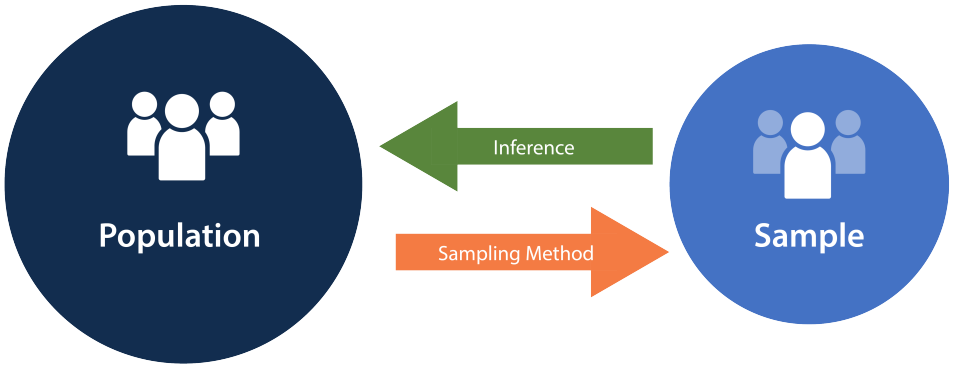


Figure 2: Inférence statistique

Théorème Central Limite

À mesure que l'échantillon s'agrandit, la distribution de la moyenne de cet échantillon $\overline{X}_n$ se rapproche d'une distribution normale, indépendamment de la forme de la distribution de la population.

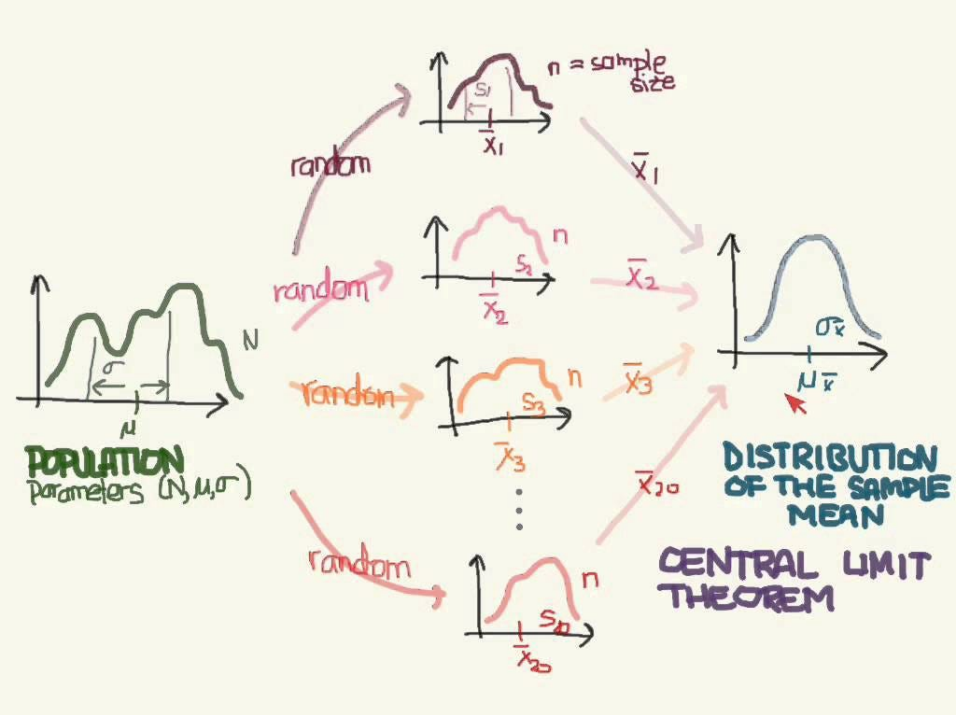


Figure 3: Théorème Central Limite

Théorème Central Limite (suite)

Le Théorème Central Limite peut être résumé par l'équation suivante :

$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Où :

- $\overline{X}_n$ est la moyenne de l'échantillon d'un ensemble de n variables aléatoires **indépendantes et identiquement distribuées**.
- $N\left(\mu, \frac{\sigma^2}{n}\right)$ indique que $\overline{X}_n$ suit approximativement une distribution normale avec une moyenne μ (la moyenne de la population) et une variance $\frac{\sigma^2}{n}$ (la variance de la population divisée par la taille de l'échantillon n).

Intervalle de confiance

Un IC est une plage de valeurs statistiques utilisée pour estimer la fiabilité d'une estimation d'un paramètre de population, comme la moyenne. Il est exprimé avec un niveau de confiance, indiquant la probabilité que cet intervalle contienne le vrai paramètre de la population.

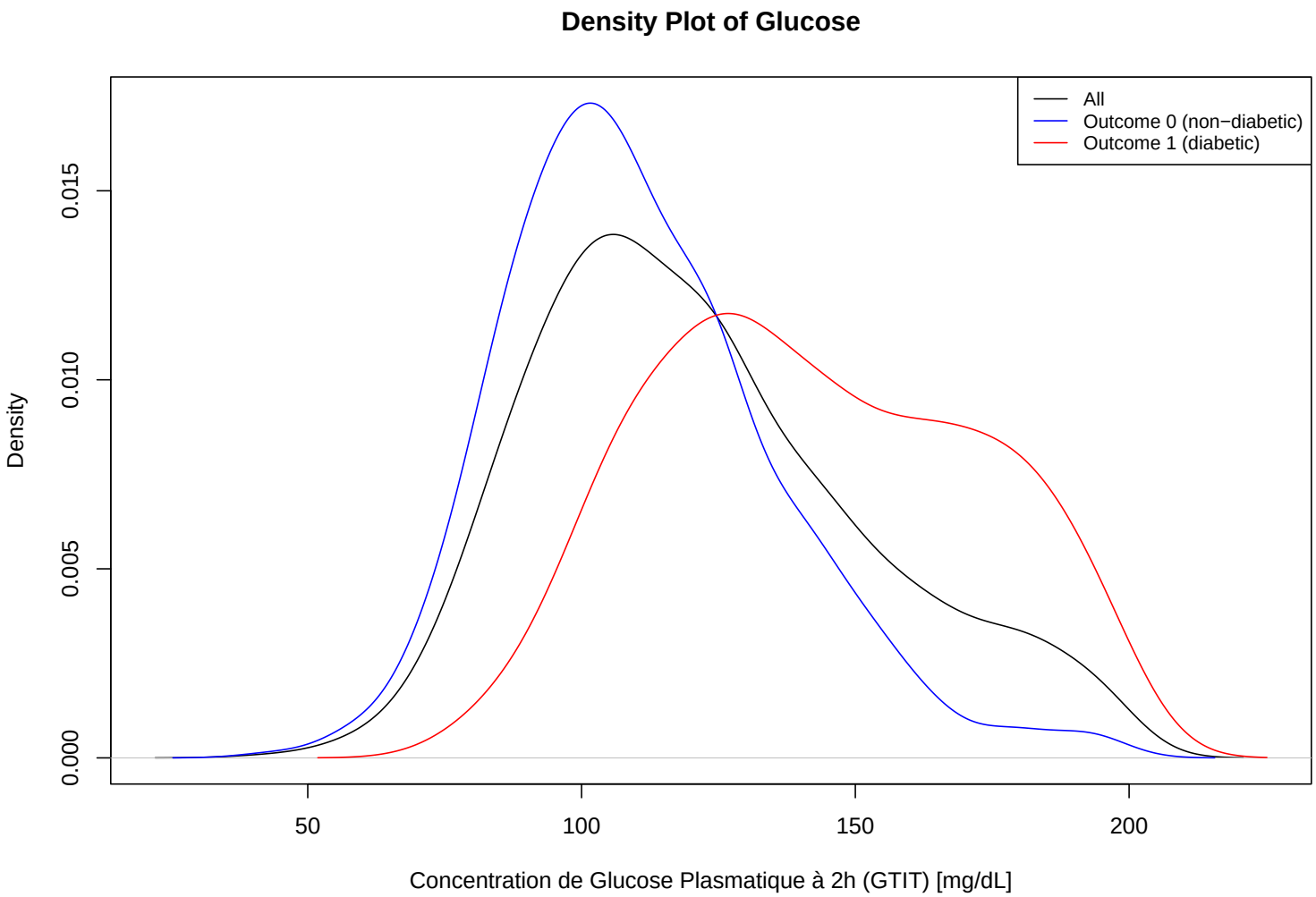
Lorsque σ est Connue

- Formule : $CI = \overline{x} \pm z \times \frac{\sigma}{\sqrt{n}}$
- z : Score Z de la distribution normale, correspondant au niveau de confiance souhaité.

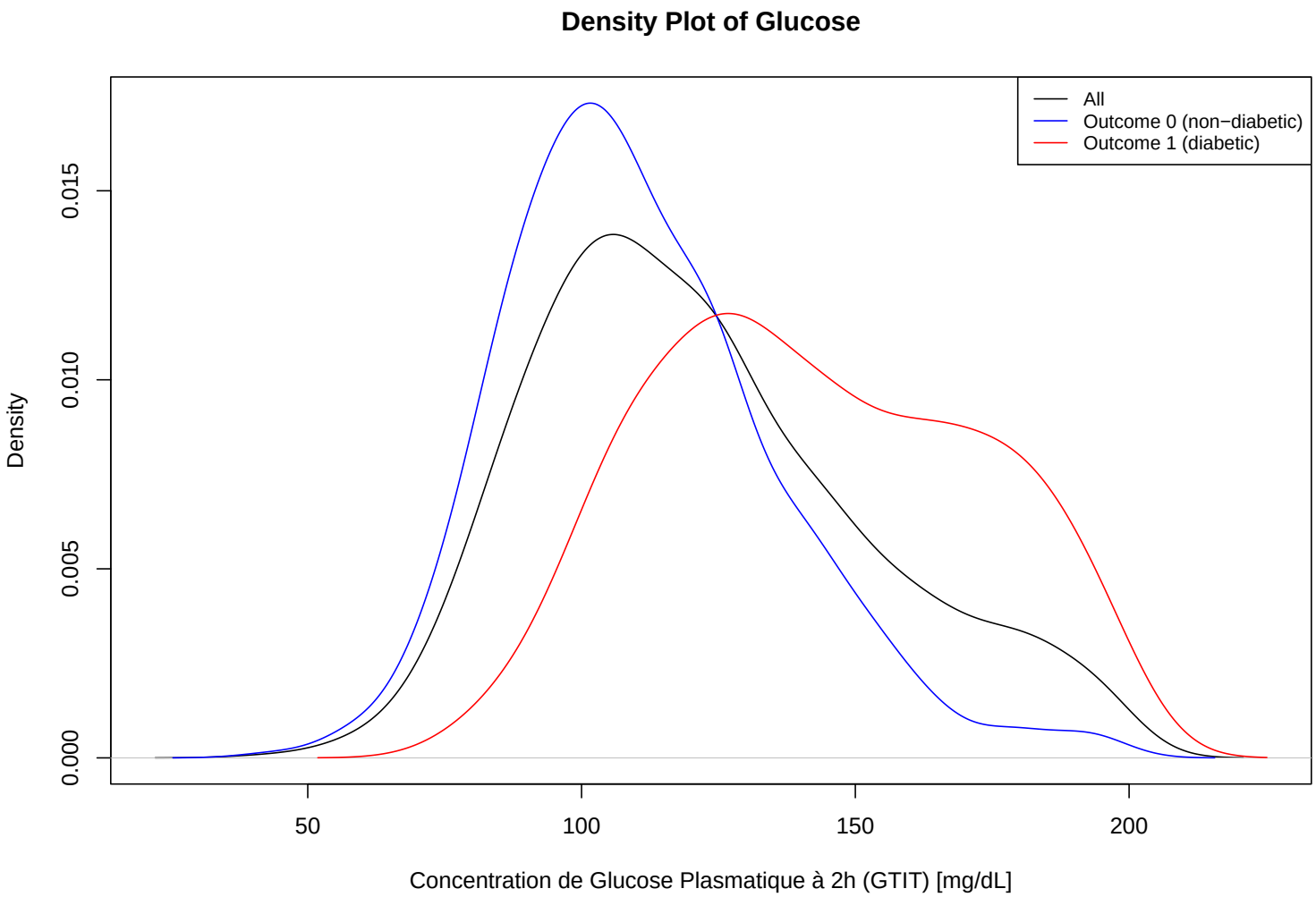
Lorsque σ est Inconnue

- Formule : $CI = \overline{x} \pm t \times \frac{s}{\sqrt{n}}$
- t : Score t de la distribution t, variant selon la taille de l'échantillon.

Intervalle de confiance (dernier TP)



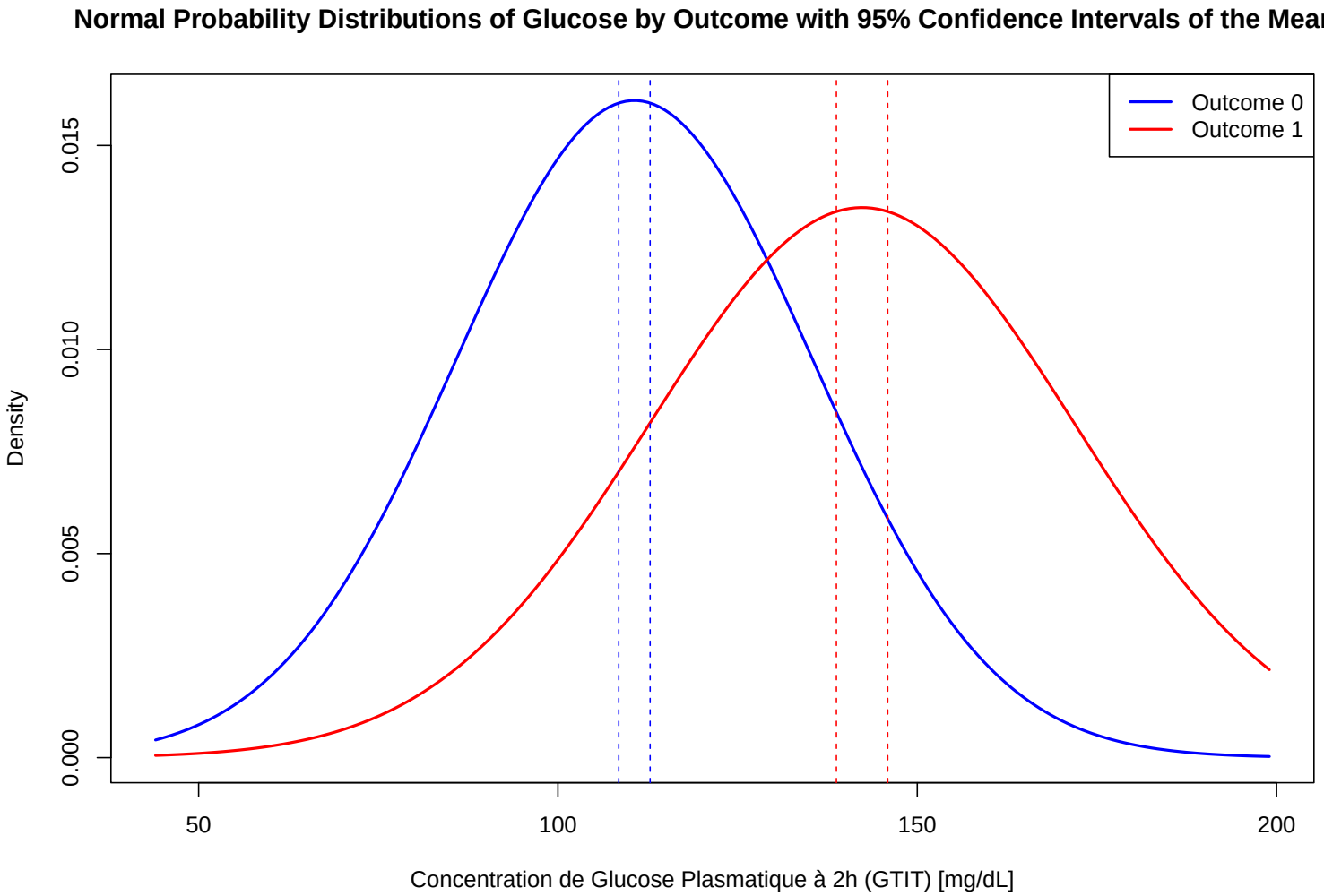
Intervalle de confiance (dernier TP)



Les personnes diabétiques et non diabétiques ont-elles des niveaux différents du glucose plasmatique ?

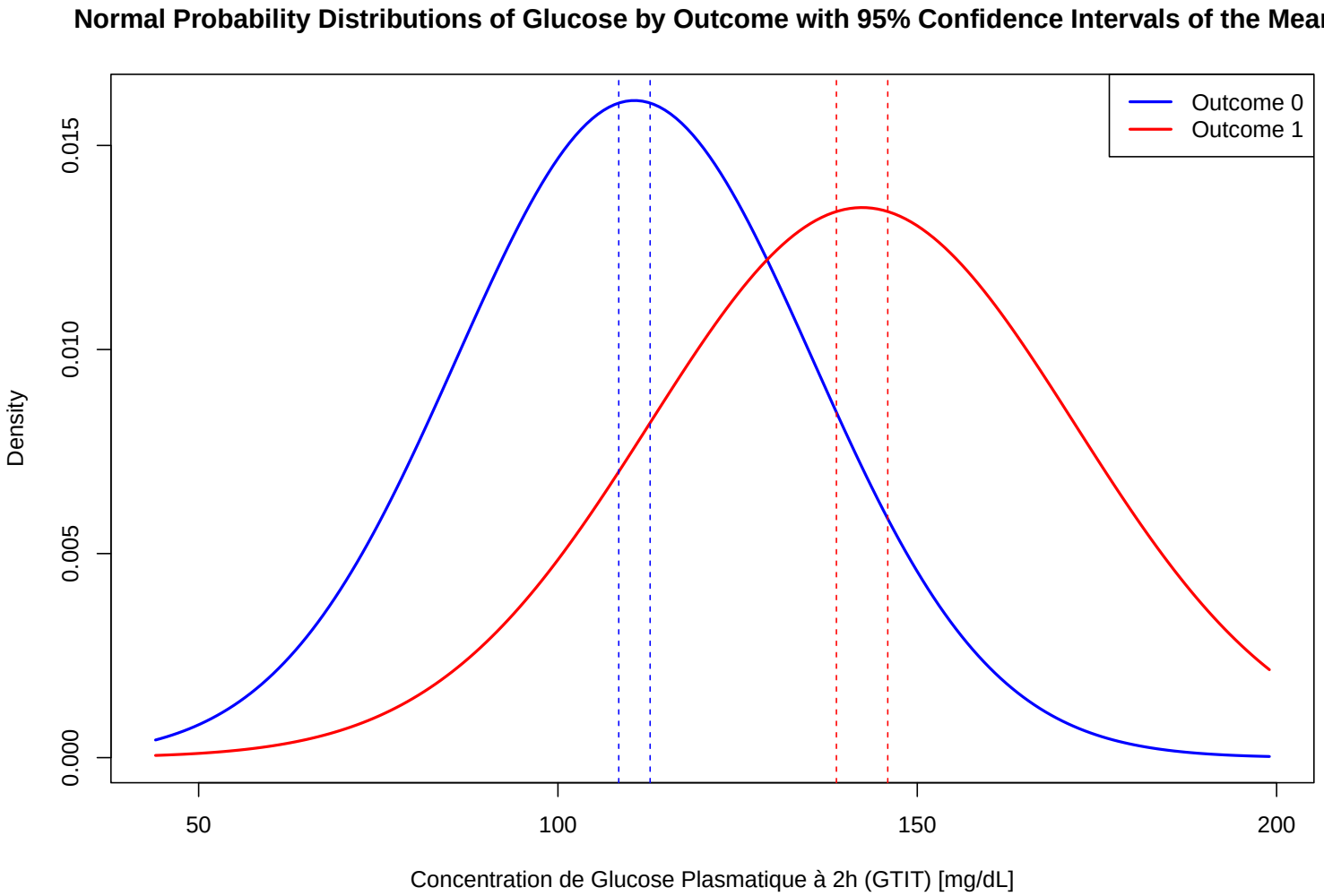
Intervalle de confiance (dernier TP)

En appliquant le Théorème Central Limite . . .



Intervalle de confiance (dernier TP)

En appliquant le Théorème Central Limite . . .



Les deux moyennes sont-elles différentes ?

Plan de la séance

- Récap
  - Échantillonage
  - Théorème Central Limite
  - Intervalle de confiance
- Test d'hypothèse
  - Types d'erreur
  - Test de moyenne
  - Test de variance
  - Test de deux populations
  - Test nonparamétrique

Types d'erreur

Types d'erreur

Erreur de type I (faux positif) : l'enquêteur rejette une *hypothèse nulle* qui est réellement vraie dans la population.

Erreur de type II (faux négatif) : l'investigateur ne parvient pas à rejeter une *hypothèse nulle* qui est en réalité fausse dans la population.

Types d'erreur

Erreur de type I (faux positif) : l'enquêteur rejette une *hypothèse nulle* qui est réellement vraie dans la population.

Erreur de type II (faux négatif) : l'investigateur ne parvient pas à rejeter une *hypothèse nulle* qui est en réalité fausse dans la population.

. . . mais quelle est l'hypothèse nulle ?

Types d'erreur (suite)

Alors, voici un exemple . . .

Figure 4: Erreur type I et II

Erreur type I

Alpha α représente le seuil de probabilité de commettre une erreur de type I dans un test d'hypothèse. C'est la probabilité maximale acceptable de rejeter à tort l'hypothèse nulle.

Communément fixé à 0,05 (5 %), un α de 0,05 signifie qu'il y a 5 % de chances de rejeter l'hypothèse nulle alors qu'elle est en réalité vraie.

Réduire α diminue les chances d'une erreur de type I, mais augmente le risque d'une erreur de type II.

Erreur type II

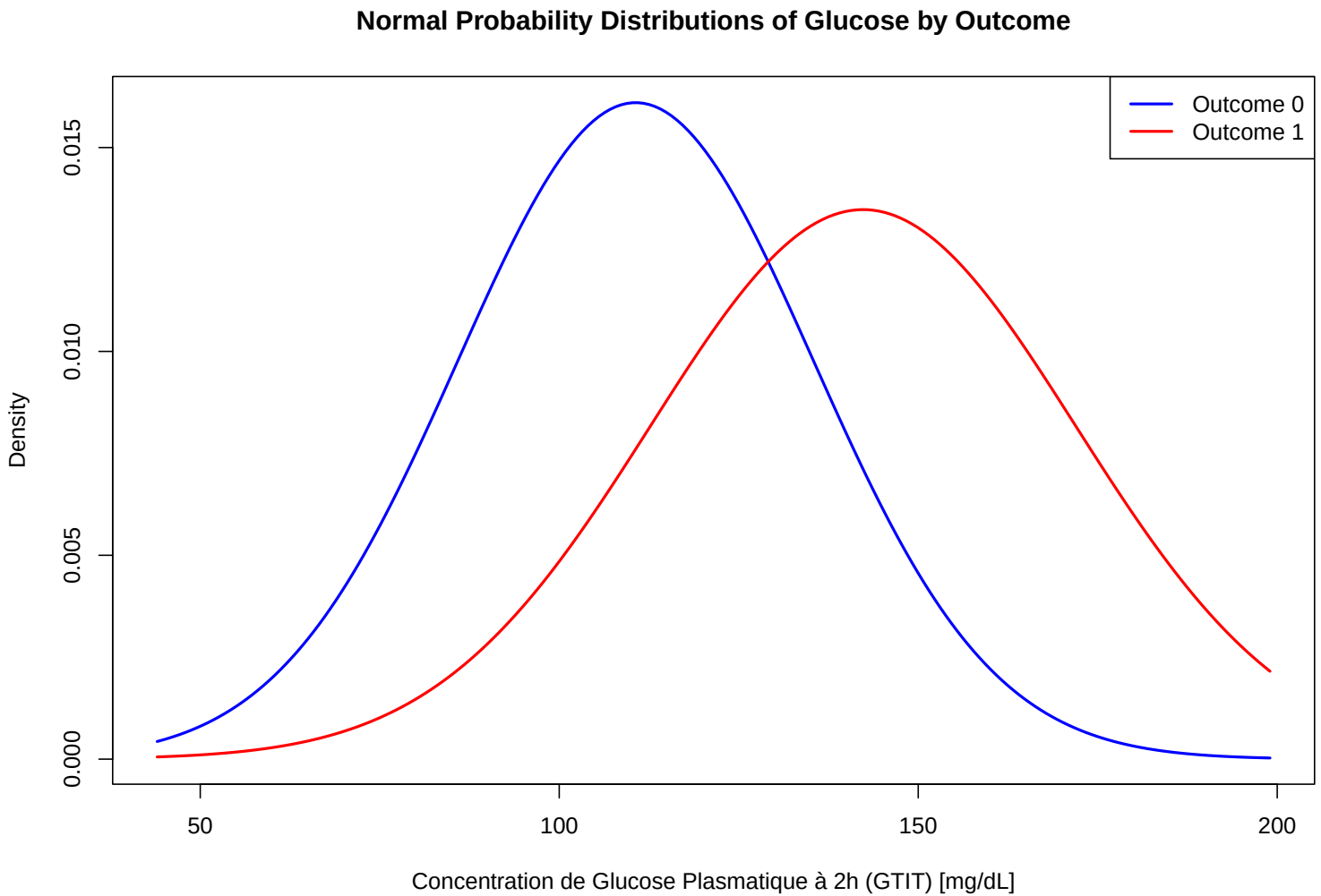
Beta β représente la probabilité de commettre une erreur de type II dans un test d'hypothèse. C'est la probabilité de ne pas rejeter une hypothèse nulle fausse.

La puissance d'un test, qui est $1 - \beta$, indique la capacité du test à rejeter correctement une fausse hypothèse nulle.

Réduire β (augmentant ainsi la puissance) nécessite souvent d'augmenter la taille de l'échantillon ou la taille de l'effet.

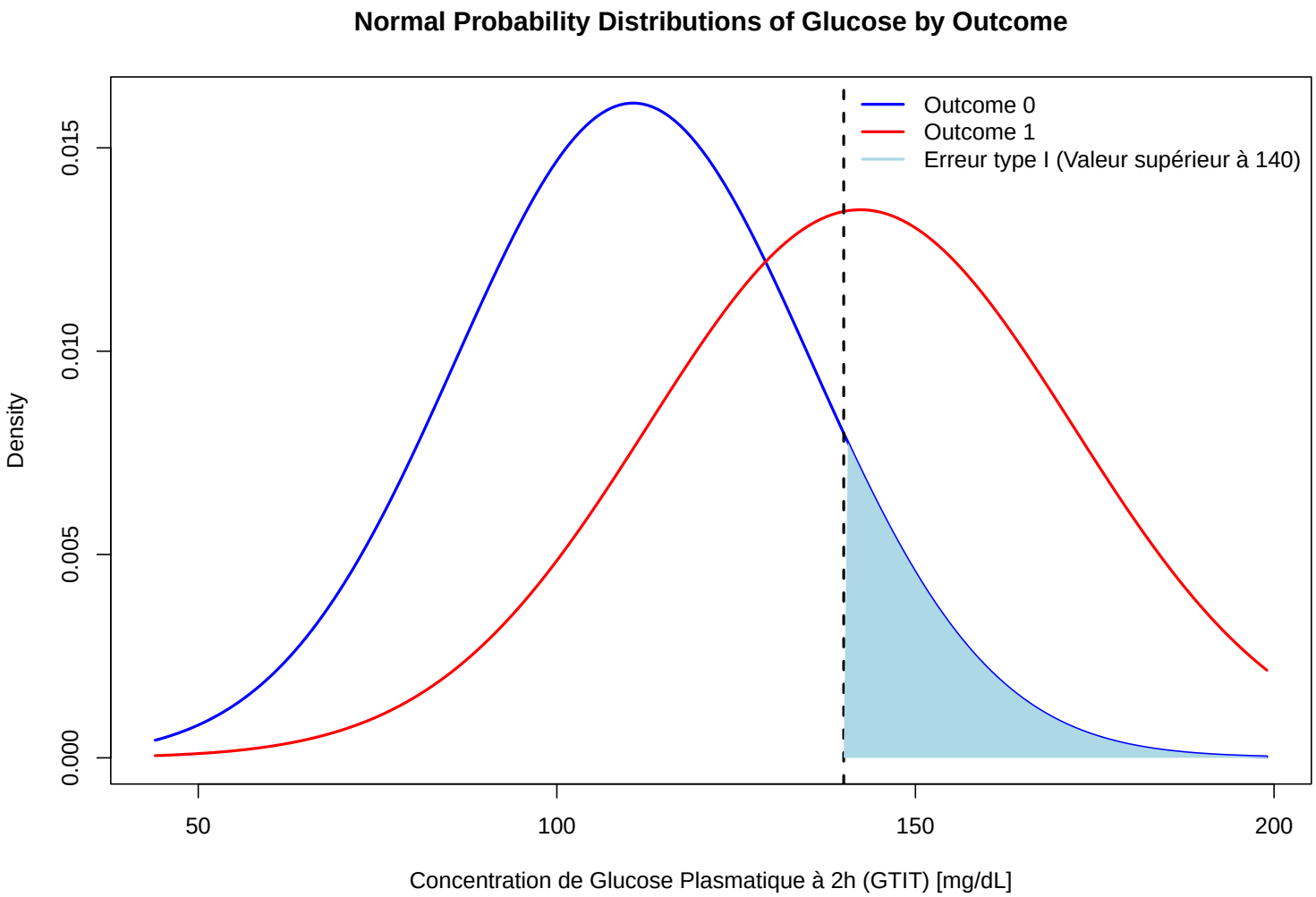
Types d'erreur (suite)

En supposant que le paramètre "Glucose" est normalement distribué ...



Erreur type I

Une concentration de glucose plasmatique (à 2h) normale est inférieure à 140 mg/dL.



Erreur type II

Une concentration de glucose plasmatique (à 2h) normale est inférieure à 140 mg/dL.

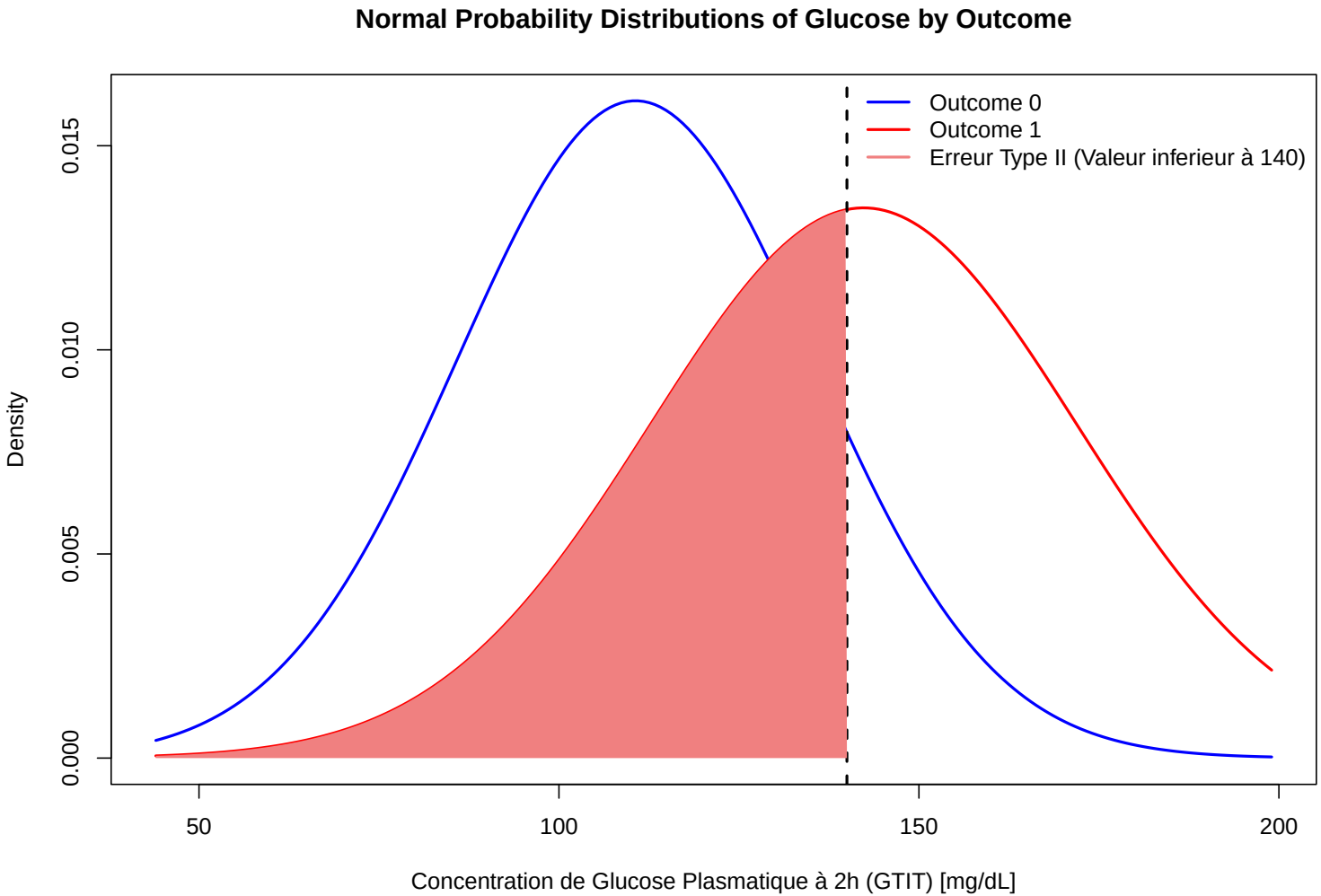


Tableau de contingence

Ou en utilisant directement les décomptes de données . . .

Tableau de contingence

Ou en utilisant directement les décomptes de données . . .

Discrétiser la variable 'Glucose'

```
data$GlucoseCtgr <- ifelse(data$Glucose < 140,
                           "Less than 140",
                           "140 and above")
```

Tableau de contingence

Ou en utilisant directement les décomptes de données . . .

Discrétiser la variable 'Glucose'

```
data$GlucoseCtgr <- ifelse(data$Glucose < 140,
                           "Less than 140",
                           "140 and above")
```

Créer un tableau de contingence avec les variables discrètes 'GlucoseCtgr' et 'Outcome'

```
contingency <- table(data$GlucoseCtgr, data$Outcome)
print(contingency)
```

```
##
##                    0   1
##    140 and above  62 135
##    Less than 140 438 133
```

Tableau de contingence (suite)

Quels sont les valeurs de α et β ?

Tableau de contingence (suite)

Quels sont les valeurs de α et β ?

- Erreur type I : 'Glucose' est '140 and above' et 'Outcome' est 0.
- Erreur type II : 'Glucose' est 'Less than 140' et 'Outcome' est 1.

Tableau de contingence (suite)

Quels sont les valeurs de α et β ?

- Erreur type I : 'Glucose' est '140 and above' et 'Outcome' est 0.
- Erreur type II : 'Glucose' est 'Less than 140' et 'Outcome' est 1.

La mauvaise manière ... à éviter!

```
contingency_prb <- prop.table(contingency)
print(contingency_prb)
```

```
##
##                          0          1
##   140 and above 0.08072917 0.17578125
##   Less than 140 0.57031250 0.17317708
```

Tableau de contingence (suite)

La bonne manière ...

```
total_negatives <- sum(contingency[, "0"])
false_positives <- contingency["140 and above", "0"]
alpha <- false_positives / total_negatives
cat("Alpha (Type I error rate):", alpha, "\n")
```

```
## Alpha (Type I error rate): 0.124
```

```
total_positives <- sum(contingency[, "1"])
false_negatives <- contingency["Less than 140", "1"]
beta <- false_negatives / total_positives
cat("Beta (Type II error rate):", beta, "\n")
```

```
## Beta (Type II error rate): 0.4962687
```

Types d'erreur (suite)

Conclusion

- α : Probabilité d'un faux positif (erreur de type I).
- β : Probabilité d'un faux négatif (erreur de type II).
- Équilibrer α et β est crucial dans les tests d'hypothèses, car la diminution de l'un augmente souvent l'autre. Le choix de α et β est influencé par le contexte de l'étude et l'importance relative des erreurs dans le scénario de recherche spécifique.

Test d'hypothèse