SYS865 Inférence statistique avec programmation R

Ornwipa Thamsuwan

24 janvier 2024

Plan de la séance

Plan de la séance

- Récap: variables aléatoires
  - Espérance
  - Variance et covariance
  - Indépendance
- Échantillonage
  - Méthodes d'échantillonage
  - Taille d'échantillon
- Début de l'inférence statistique
  - Intervalle de confiance

Récap et matière à réflexion

Récap et matière à réflexion

Lire des données

```
data <- read.csv("diabetes.csv")
```

Espérance

```
expectations <- sapply(data, mean)
```

Variance et covariance

```
covariances <- var(data)
```

Indépendance

Comment savoir si deux variables sont indépendantes l'une de l'autre ?

Indépendance

Comment savoir si deux variables sont indépendantes l'une de l'autre ?

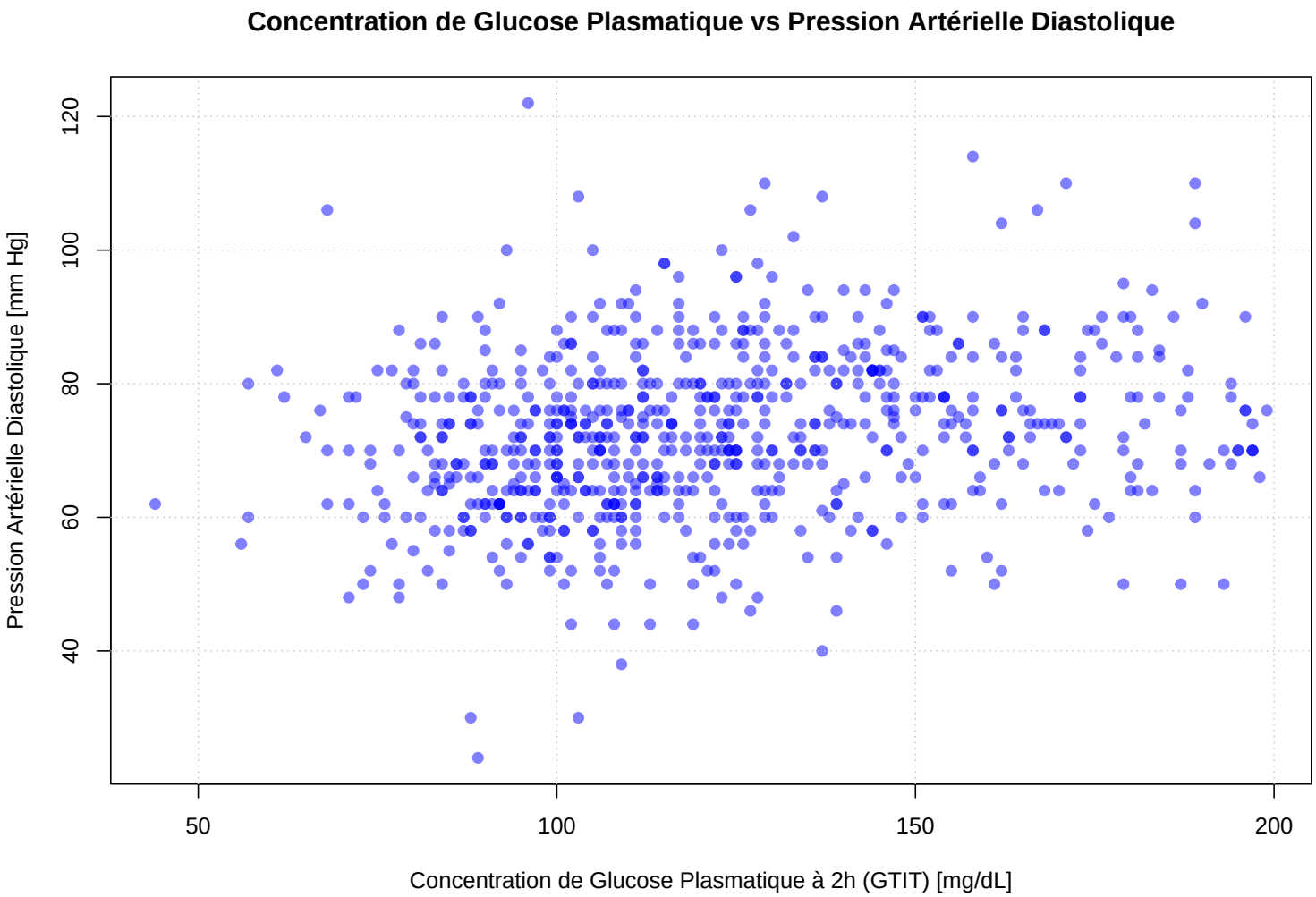
- Inspection visuelle par graphiques de dispersion ("Scatter plot" en anglais)

Indépendance

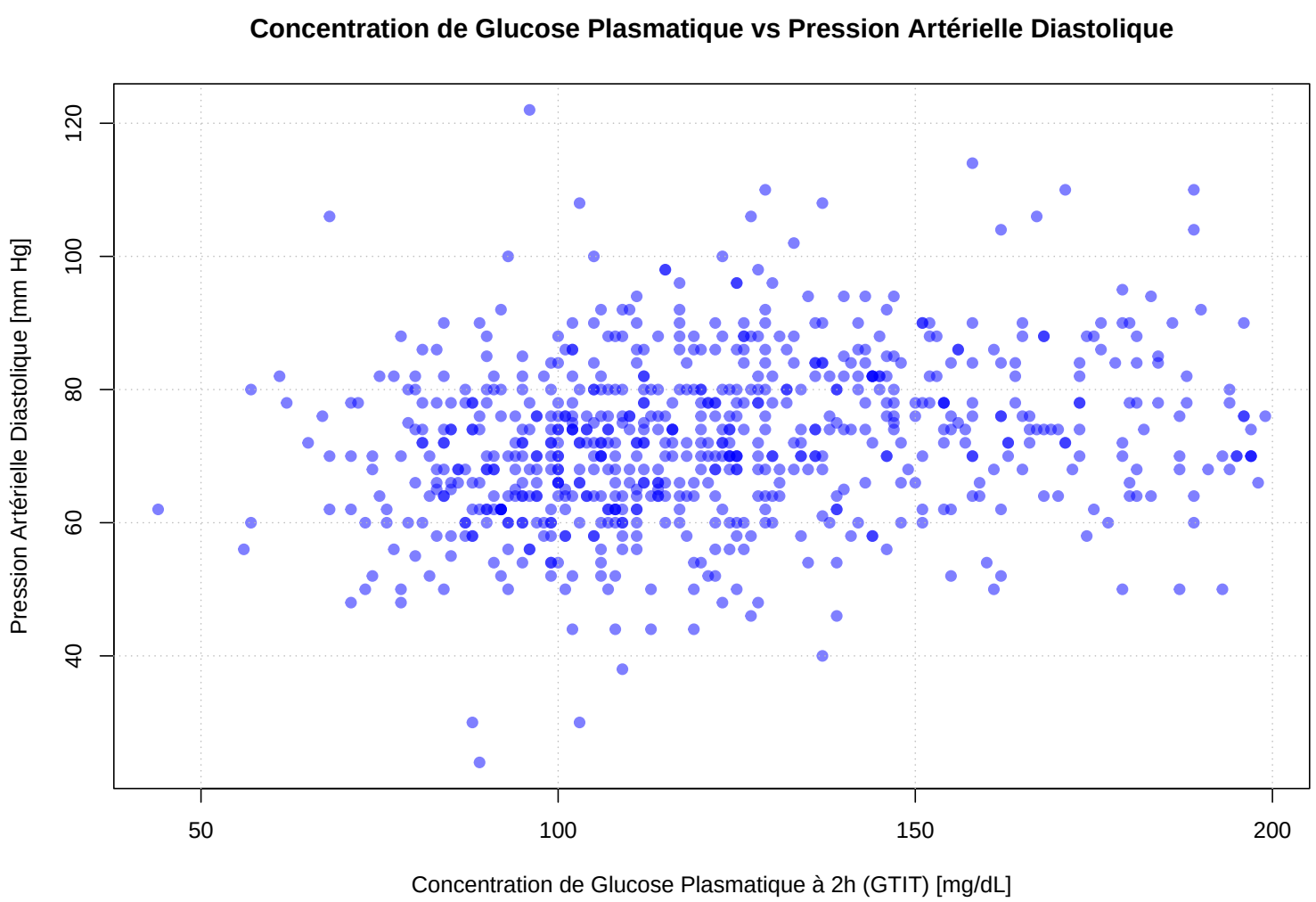
Comment savoir si deux variables sont indépendantes l'une de l'autre ?

- Inspection visuelle par graphiques de dispersion ("Scatter plot" en anglais)
- Test de hypothèse
  - Test χ^2
  - Test de corrélation
  - Regression linéaire
  - Regression logistique

Graphique de dispersion

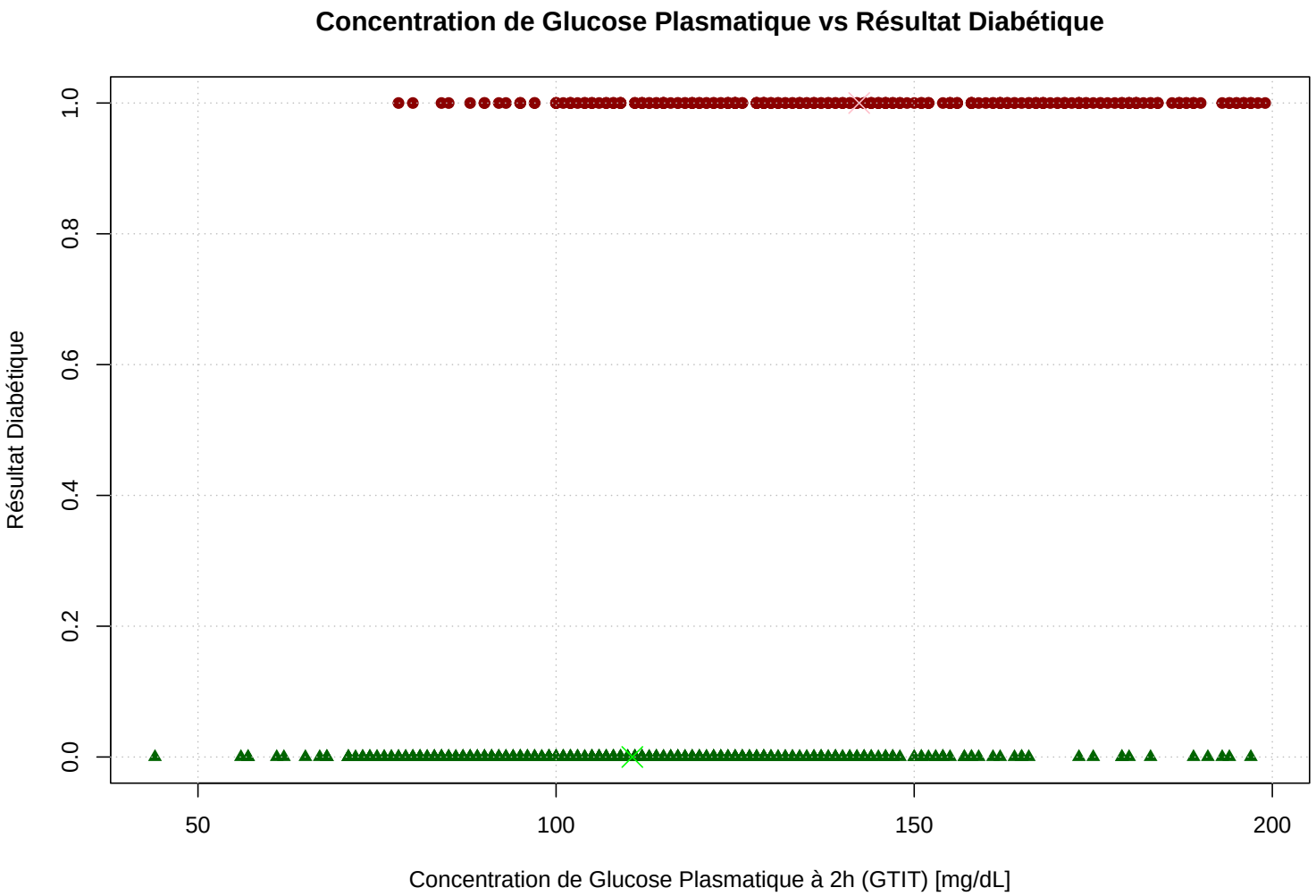


Graphique de dispersion

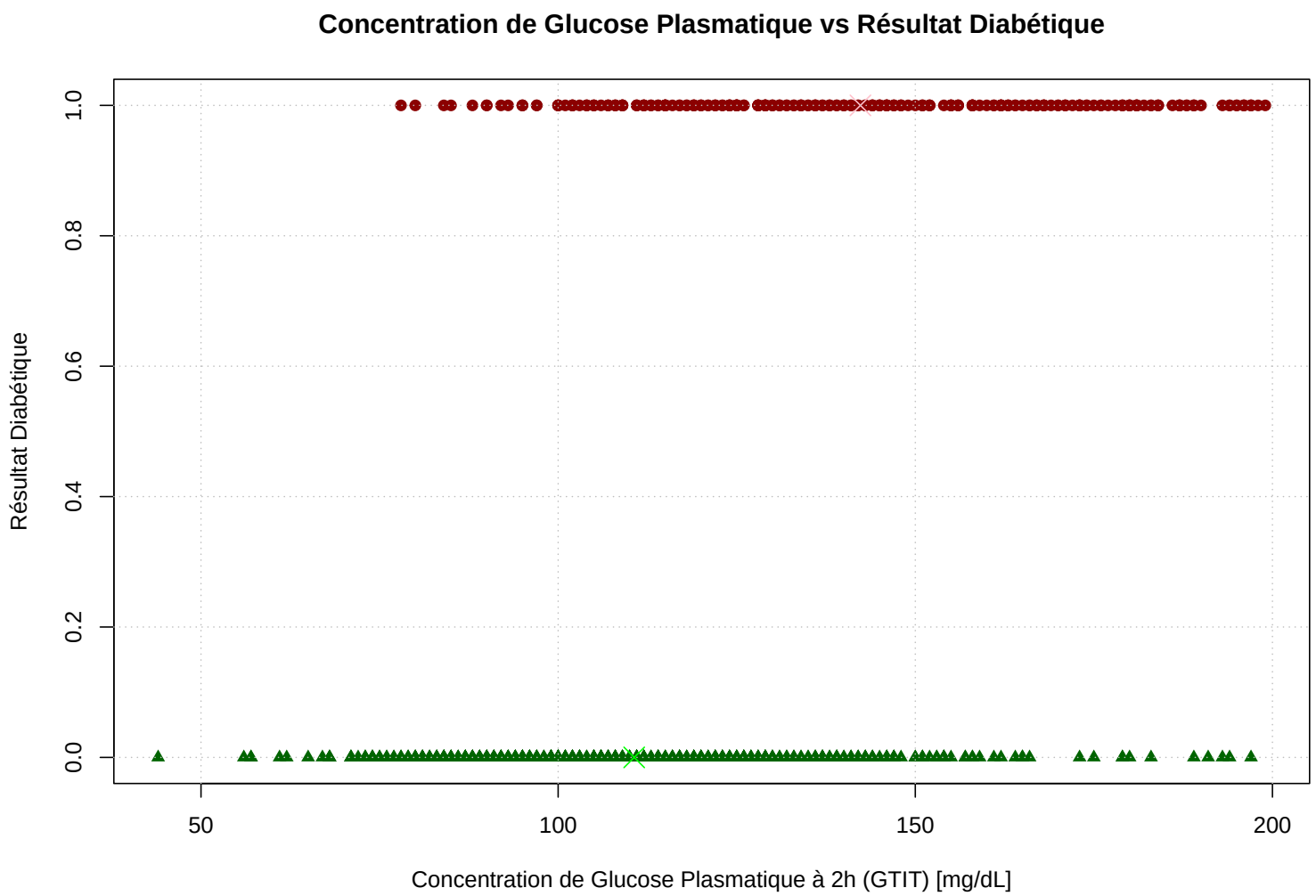


À votre avis, la covariance entre le glucose et la pression artérielle est positive, négative ou proche de zéro ?

Graphique de dispersion



Graphique de dispersion



Remarquez la différence dans la moyenne et dans la plage en comparant le cas des diabétiques et des non-diabétiques ?

Récap et matière à réflexion (suite)

Pouvons-nous utiliser les données fournies pour répondre à ces questions ?

Les données sont-elles représentatives de la population ?

Récap et matière à réflexion (suite)

Pouvons-nous utiliser les données fournies pour répondre à ces questions ?

Les données sont-elles représentatives de la population ?

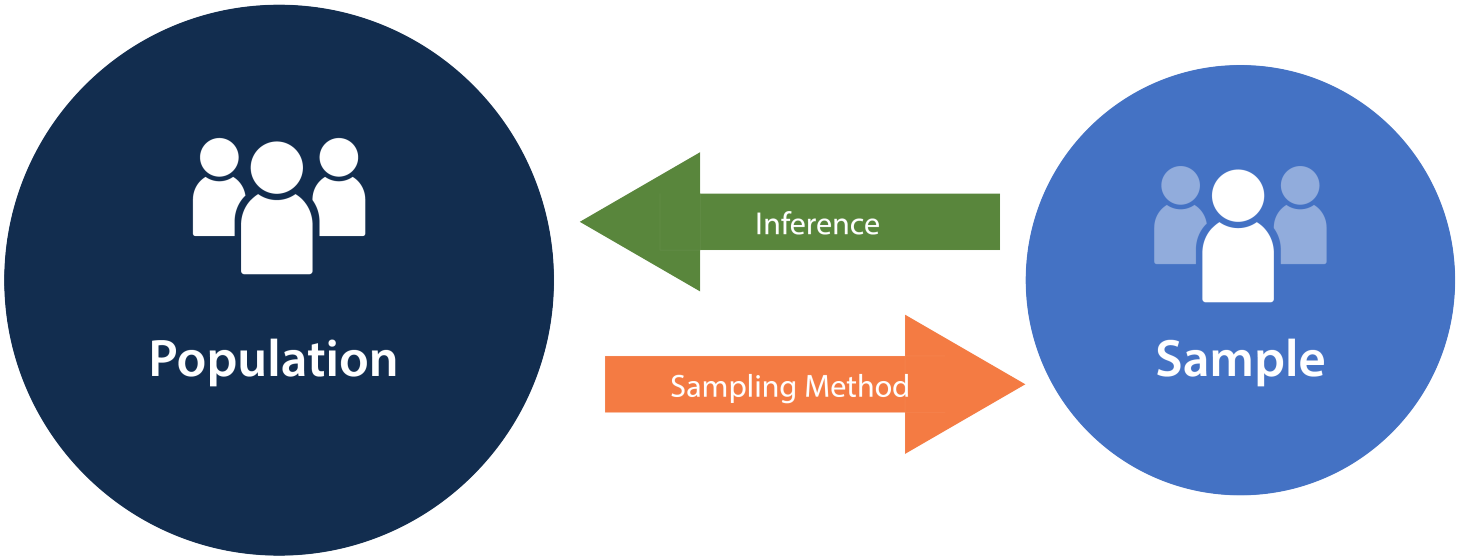


Figure 1: Relation entre population et échantillon

Méthodes d'échantillonage

Méthodes d'échantillonage

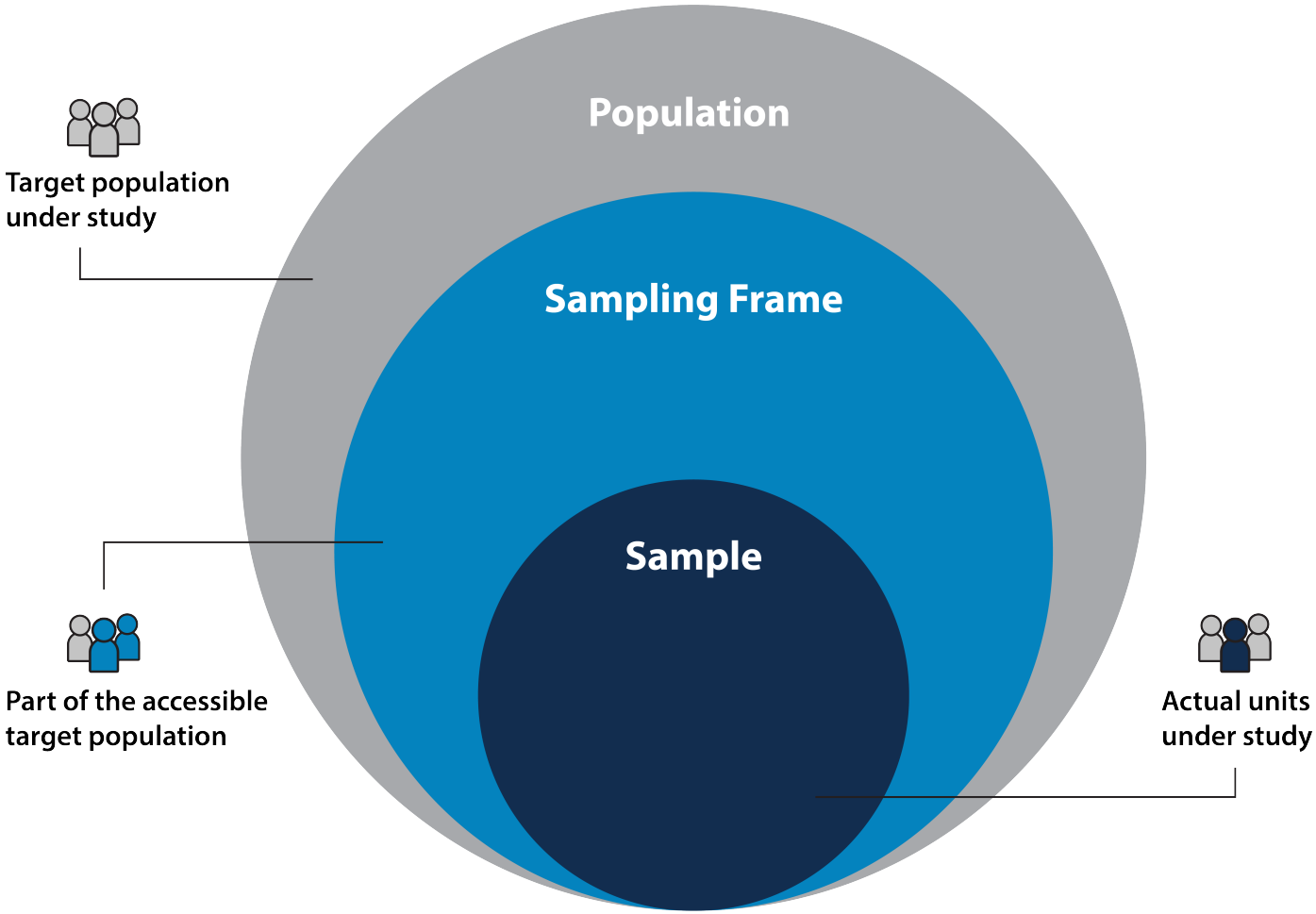


Figure 2: Échantillonage

Définitions et exemples

Recherche sur la complexité des problèmes de statistiques dispensés par les professeurs aux différentes spécialisations à l'ÉTS :

Définitions et exemples

Recherche sur la complexité des problèmes de statistiques dispensés par les professeurs aux différentes spécialisations à l'ÉTS :

Population : Un ensemble complet d'éléments (personnes, objets ou sujets) ayant des caractéristiques spécifiques que vous souhaitez étudier et sur lesquelles vous souhaitez faire des inférences.

- *Tous les enseignants de l'ÉTS qui dispensent des cours en statistiques aux différentes spécialisations.*

Définitions et exemples

Recherche sur la complexité des problèmes de statistiques dispensés par les professeurs aux différentes spécialisations à l'ÉTS :

Population : Un ensemble complet d'éléments (personnes, objets ou sujets) ayant des caractéristiques spécifiques que vous souhaitez étudier et sur lesquelles vous souhaitez faire des inférences.

- *Tous les enseignants de l'ÉTS qui dispensent des cours en statistiques aux différentes spécialisations.*

Cadre d'échantillonnage : le matériel source ou la liste complète à partir de laquelle un échantillon est tiré. C'est une compilation exhaustive de tous les éléments de votre population.

- *Le registre de l'ÉTS qui liste tous les enseignants des cours en statistiques.*

Définitions et exemples

Échantillon : Un sous-ensemble d'une population. Il s'agit de l'ensemble spécifique d'éléments à partir desquels vous collecterez des données.

- *Sous-ensemble des enseignants de l'ÉTS que vous sélectionnez pour votre étude.*

Définitions et exemples

Échantillon : Un sous-ensemble d'une population. Il s'agit de l'ensemble spécifique d'éléments à partir desquels vous collecterez des données.

- *Sous-ensemble des enseignants de l'ÉTS que vous sélectionnez pour votre étude.*

Taille de l'échantillon : le nombre de membres de la population enquêtés, mesurés ou observés.

La taille de l'échantillon détermine la quantité de données, ce qui influence davantage la précision de votre étude et la fiabilité de vos résultats.

Échantillonnage probabiliste

Échantillonnage aléatoire simple : Chaque membre de la population a une chance égale d'être sélectionné.

Échantillonnage probabiliste

Échantillonnage aléatoire simple : Chaque membre de la population a une chance égale d'être sélectionné.

Échantillonnage systématique : On sélectionne des membres de la population à intervalles réguliers, par exemple, choisir chaque 10ème personne de la liste.

Échantillonnage probabiliste

Échantillonnage aléatoire simple : Chaque membre de la population a une chance égale d'être sélectionné.

Échantillonnage systématique : On sélectionne des membres de la population à intervalles réguliers, par exemple, choisir chaque 10ème personne de la liste.

Échantillonnage stratifié : La population est divisée en sous-groupes (strates) qui partagent des caractéristiques similaires. Un échantillon aléatoire est ensuite prélevé dans chacune de ces strates. Cette méthode garantit une représentation de chaque sous-groupe.

Échantillonnage probabiliste

Échantillonnage aléatoire simple : Chaque membre de la population a une chance égale d'être sélectionné.

Échantillonnage systématique : On sélectionne des membres de la population à intervalles réguliers, par exemple, choisir chaque 10ème personne de la liste.

Échantillonnage stratifié : La population est divisée en sous-groupes (strates) qui partagent des caractéristiques similaires. Un échantillon aléatoire est ensuite prélevé dans chacune de ces strates. Cette méthode garantit une représentation de chaque sous-groupe.

Échantillonnage par grappes : La population est divisée en grappes, généralement basées sur des zones géographiques, et un échantillon aléatoire de ces grappes est choisi. Tous les individus des grappes sélectionnées sont dans l'échantillon.

Échantillonnage probabiliste (suite)

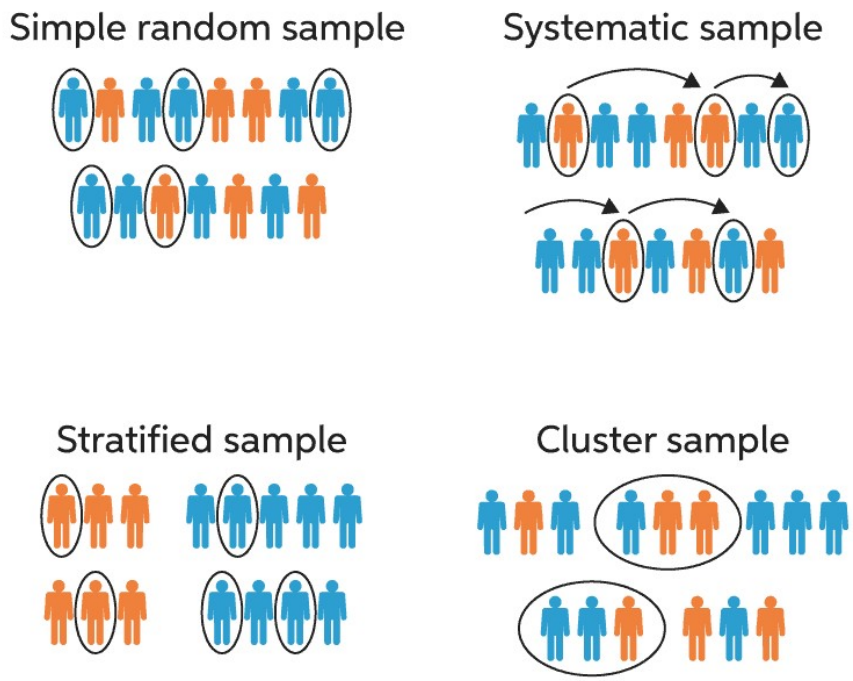


Figure 3: Échantillonage probabiliste

Échantillonnage non-probabiliste

Échantillonnage de convenance : Les participants choisis sont les plus faciles à atteindre. Ce n'est pas un échantillon aléatoire et est souvent utilisé pour les tests pilotes.

Échantillonnage non-probabiliste

Échantillonnage de convenance : Les participants choisis sont les plus faciles à atteindre. Ce n'est pas un échantillon aléatoire et est souvent utilisé pour les tests pilotes.

Échantillonnage intentionnel : Les participants sont sélectionnés en fonction du but de l'étude et du jugement du chercheur.

Échantillonnage non-probabiliste

Échantillonnage de convenance : Les participants choisis sont les plus faciles à atteindre. Ce n'est pas un échantillon aléatoire et est souvent utilisé pour les tests pilotes.

Échantillonnage intentionnel : Les participants sont sélectionnés en fonction du but de l'étude et du jugement du chercheur.

Échantillonnage à réponse volontaire : C'est les sujets qui choisissent de participer, souvent en réponse à une invitation générale.

Échantillonnage non-probabiliste

Échantillonnage de convenance : Les participants choisis sont les plus faciles à atteindre. Ce n'est pas un échantillon aléatoire et est souvent utilisé pour les tests pilotes.

Échantillonnage intentionnel : Les participants sont sélectionnés en fonction du but de l'étude et du jugement du chercheur.

Échantillonnage à réponse volontaire : C'est les sujets qui choisissent de participer, souvent en réponse à une invitation générale.

Échantillonnage boule de neige : Les sujets actuels recrutent de futurs sujets parmi leurs connaissances. Cela est particulièrement utile pour atteindre des populations difficiles d'accès.

Échantillonnage non-probabiliste (suite)

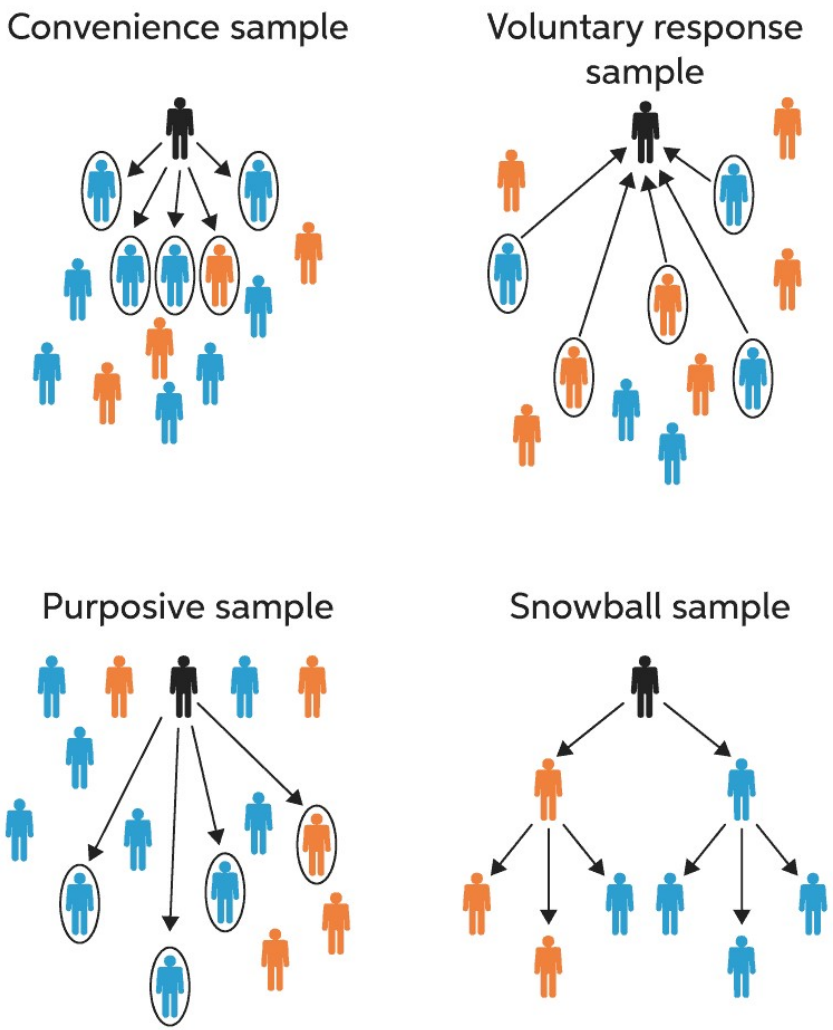


Figure 4: Échantillonage non-probabiliste

Discussion

En groupe de 3-4, discutez de quelle serait la situation dans laquelle chacune des méthodes d'échantillonnage serait utilisée ?

Vérification des connaissances

Pour une étude sur une maladie rare, un chercheur choisi délibérément les patients connus pour souffrir de cette maladie à partir des dossiers médicaux ou des établissements de santé spécialisés.

Vérification des connaissances

Pour une étude sur une maladie rare, un chercheur choisi délibérément les patients connus pour souffrir de cette maladie à partir des dossiers médicaux ou des établissements de santé spécialisés.

- Échantillonnage intentionnel

Vérification des connaissances

Pour une étude sur une maladie rare, un chercheur choisi délibérément les patients connus pour souffrir de cette maladie à partir des dossiers médicaux ou des établissements de santé spécialisés.

- Échantillonnage intentionnel

Pour évaluer l'efficacité d'une nouvelle campagne de santé, une organisation sélectionne au hasard cinq quartiers d'une ville. Ils enquêtent ensuite sur chaque foyer de ces quartiers.

Vérification des connaissances

Pour une étude sur une maladie rare, un chercheur choisi délibérément les patients connus pour souffrir de cette maladie à partir des dossiers médicaux ou des établissements de santé spécialisés.

- Échantillonnage intentionnel

Pour évaluer l'efficacité d'une nouvelle campagne de santé, une organisation sélectionne au hasard cinq quartiers d'une ville. Ils enquêtent ensuite sur chaque foyer de ces quartiers.

- Échantillonnage par grappes

Vérification des connaissances

Dans une étude sur la santé mentale des personnes sans abri, les premiers participants sans abri recommandent d'autres personnes sans abri qu'ils connaissent.

Vérification des connaissances

Dans une étude sur la santé mentale des personnes sans abri, les premiers participants sans abri recommandent d'autres personnes sans abri qu'ils connaissent.

- Échantillonnage boule de neige

Vérification des connaissances

Dans une étude sur la santé mentale des personnes sans abri, les premiers participants sans abri recommandent d'autres personnes sans abri qu'ils connaissent.

- Échantillonnage boule de neige

Une équipe de recherche sur la prévalence de l'hypertension divise la population en catégories ethniques, puis sélectionne au hasard un nombre proportionné d'individus dans chaque groupe pour garantir que tous soient représentés.

Vérification des connaissances

Dans une étude sur la santé mentale des personnes sans abri, les premiers participants sans abri recommandent d'autres personnes sans abri qu'ils connaissent.

- Échantillonnage boule de neige

Une équipe de recherche sur la prévalence de l'hypertension divise la population en catégories ethniques, puis sélectionne au hasard un nombre proportionné d'individus dans chaque groupe pour garantir que tous soient représentés.

- Échantillonnage stratifié

Attention!

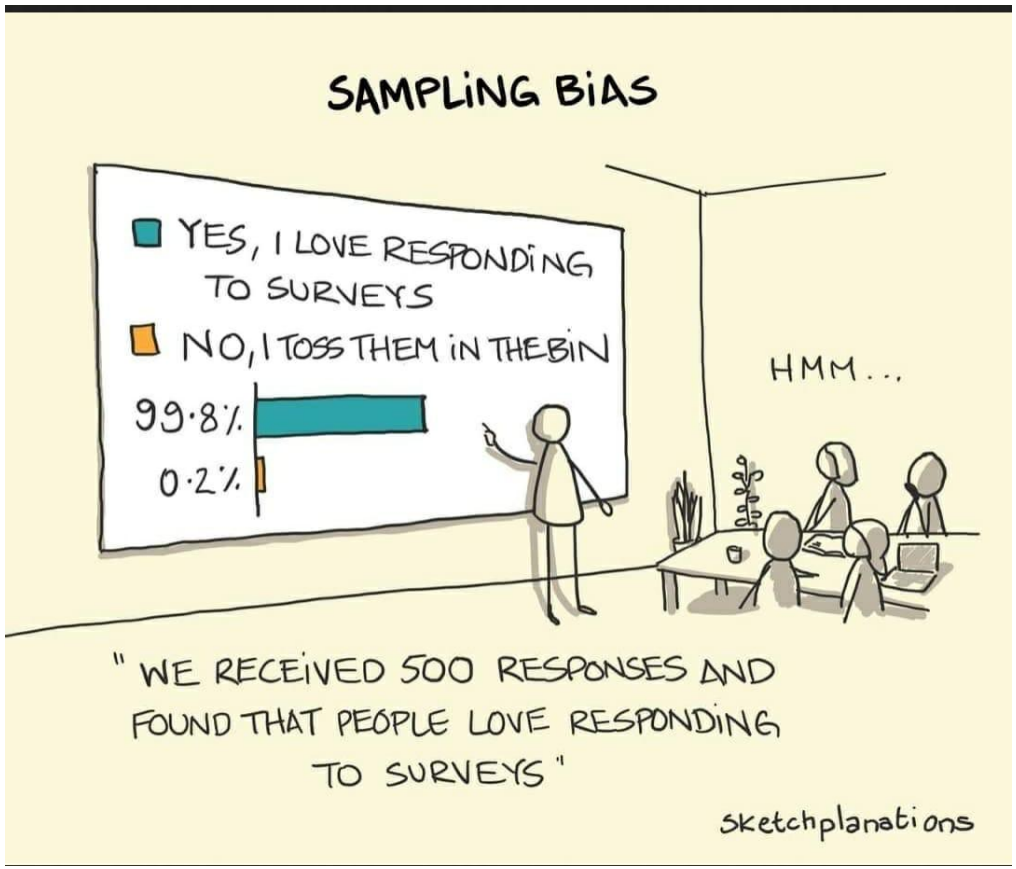


Figure 5: Biais de réponse

Attention!

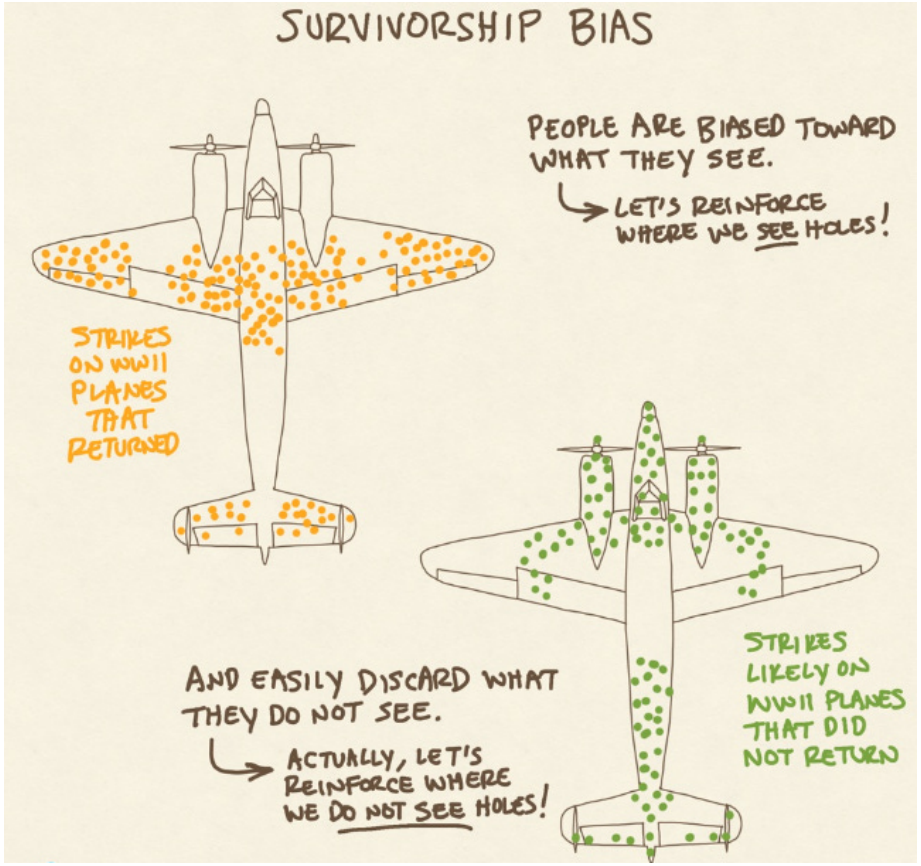


Figure 6: Biais de survie

Effet de la taille de l'échantillon sur l'erreur et l'inférence

Théorème Central Limite et Implications

Le Théorème Central Limite (TCL) stipule que, pour une taille d'échantillon suffisamment grande, la distribution des moyennes d'échantillons se rapprochera d'une distribution normale, indépendamment de la distribution originale de la population.

Théorème Central Limite et Implications

Le Théorème Central Limite (TCL) stipule que, pour une taille d'échantillon suffisamment grande, la distribution des moyennes d'échantillons se rapprochera d'une distribution normale, indépendamment de la distribution originale de la population.

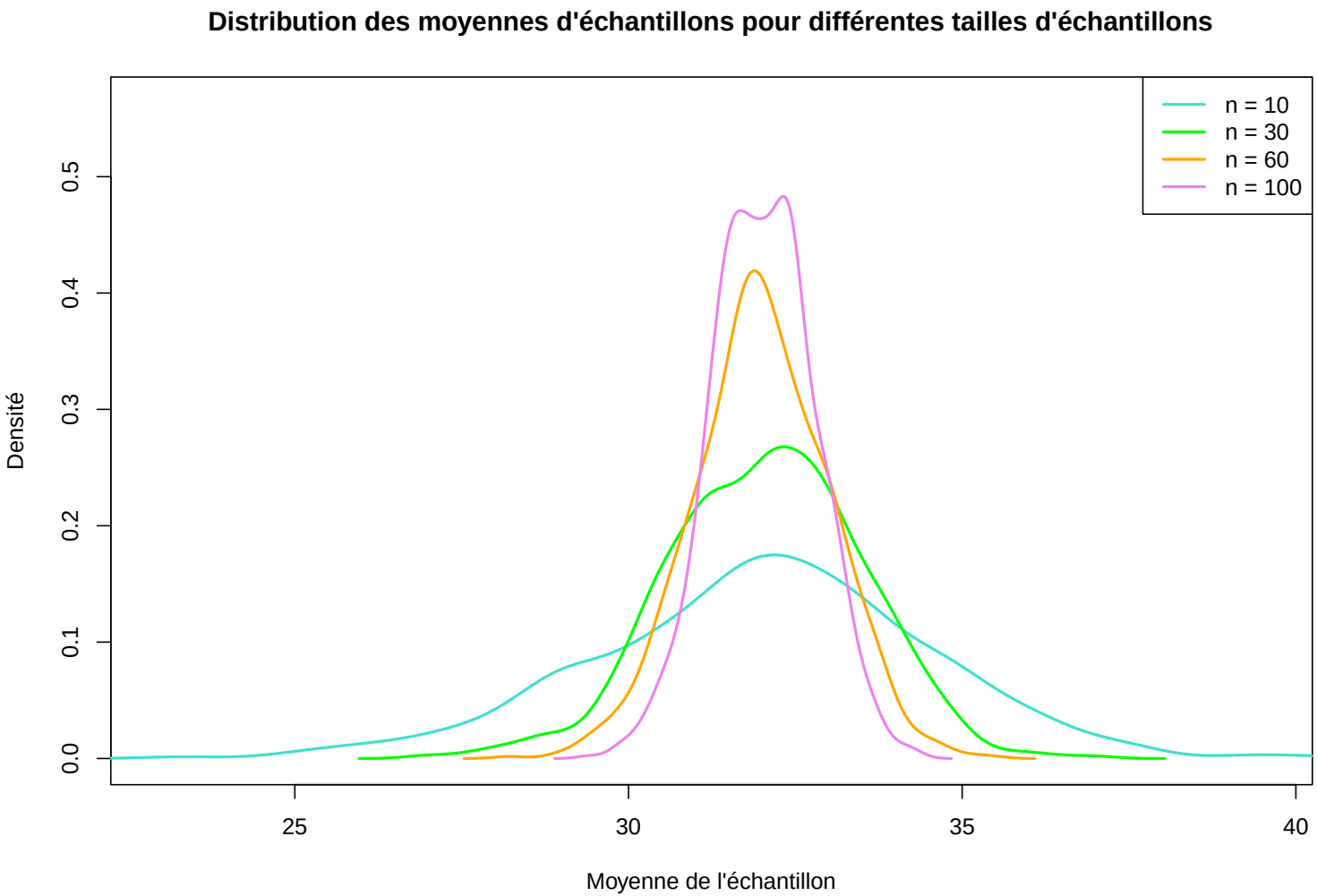
Définition

Si $X_1, X_2, X_3, ..., X_n$ sont des échantillons aléatoires pris d'une population avec une moyenne générale μ et une variance finie σ^2, la moyenne de l'échantillon $\overline{X} = \frac{1}{n}(X_1 + X_2 + X_3 + ... + X_n)$ sera approximativement distribuée normalement avec une moyenne μ et une variance $\frac{\sigma^2}{n}$, à mesure que n devient grand.

La distribution normale est notée $N(\mu, \frac{\sigma^2}{n})$.

Simulation avec des données

À partir de la base de données de “Pima Indian Diabetes Dataset”, les IMC (ou “BMI” en anglais) sont échantillonées.



Distribution Normale

À mesure que la taille de l'échantillon (n) augmente, la forme de la distribution de la moyenne de l'échantillon ($\overline{X}$) devient de plus en plus en cloche ("bell-shapred" en anglais) ou normale.

Le TCL justifie l'utilisation de la distribution normale dans l'inférence statistique et les tests d'hypothèses, même lorsque la population sous-jacente n'est pas normalement distribuée.

Réduction des Erreurs d'Échantillonnage

L'**erreur standard**, qui mesure la variabilité des moyennes d'échantillons $\overline{X}$, est donnée par $SE = \frac{\sigma}{\sqrt{n}}$.

Réduction des Erreurs d'Échantillonnage

L'**erreur standard**, qui mesure la variabilité des moyennes d'échantillons $\overline{X}$, est donnée par $SE = \frac{\sigma}{\sqrt{n}}$.

À mesure que la taille de l'échantillon (n) augmente, l'erreur standard SE diminue. Cela indique que des échantillons plus grands fournissent des estimations plus précises de la moyenne de la population, réduisant ainsi le risque d'erreur d'échantillonnage.

À noter que . . .

La taille d'échantillon "suffisamment grande" pour le TCL est généralement considérée comme étant 30 ou plus, mais cela peut varier en fonction de la population.

Intervalle de confiance

Intervalle de confiance

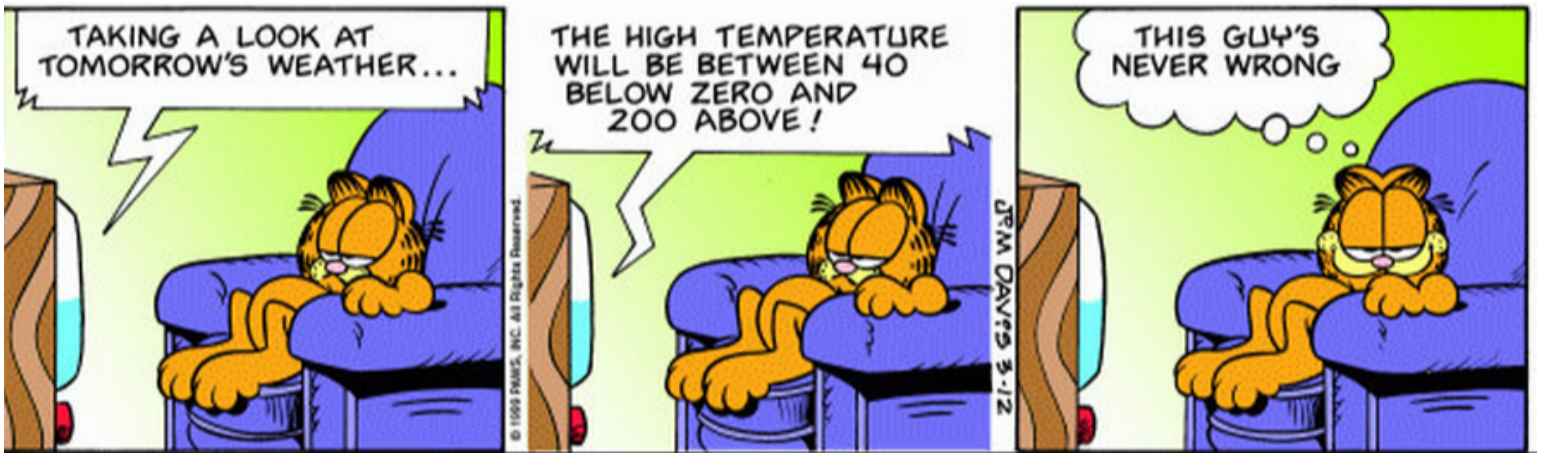


Figure 7: Intervalle de Confiance

Est-ce que ça donne une information importante ?

Définition de l'Intervalle de Confiance

Un intervalle de confiance est une plage de valeurs estimée à partir des données d'un échantillon, destinée à contenir un paramètre inconnu de la population (par exemple, moyenne de la population μ).

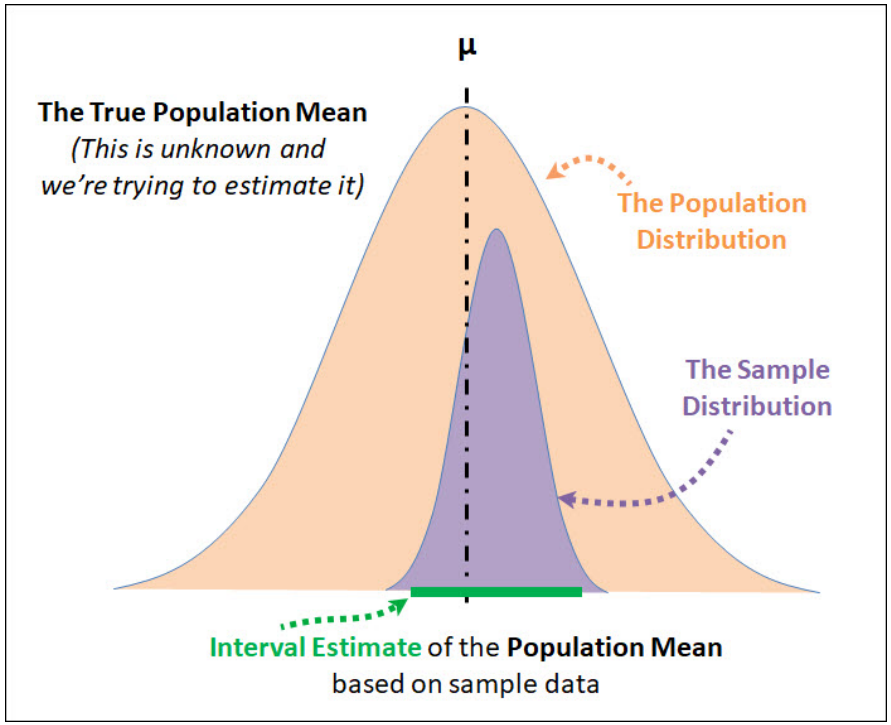


Figure 8: Estimation de l'IC

Composants d'un Intervalle de Confiance

- **Estimation Ponctuelle** est généralement la moyenne de l'échantillon ($\overline{x}$).
- **Niveau de Confiance**, exprimé en pourcentage (90 %, 95 %, 99 %, etc.), indique la **probabilité** que cet intervalle contienne le paramètre de la population si l'expérience est répétée plusieurs fois.
- **Marge d'Erreur (ME)** reflète l'incertitude autour de l'estimation ponctuelle et dépend de l'écart-type de la population σ et de la taille de l'échantillon n.

Composants d'un Intervalle de Confiance

- **Estimation Ponctuelle** est généralement la moyenne de l'échantillon ($\overline{x}$).
- **Niveau de Confiance**, exprimé en pourcentage (90 %, 95 %, 99 %, etc.), indique la **probabilité** que cet intervalle contienne le paramètre de la population si l'expérience est répétée plusieurs fois.
- **Marge d'Erreur (ME)** reflète l'incertitude autour de l'estimation ponctuelle et dépend de l'écart-type de la population σ et de la taille de l'échantillon n.

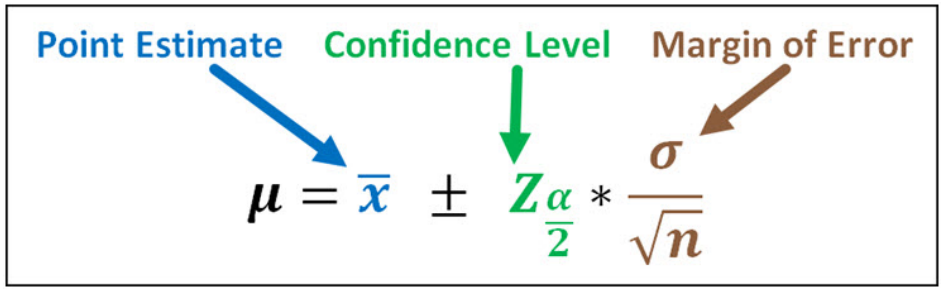


Figure 9: Composants d'un IC

Intervalle de Confiance pour la Moyenne

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.

Intervalle de Confiance pour la Moyenne

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.

Niveau de Confiance et Score z

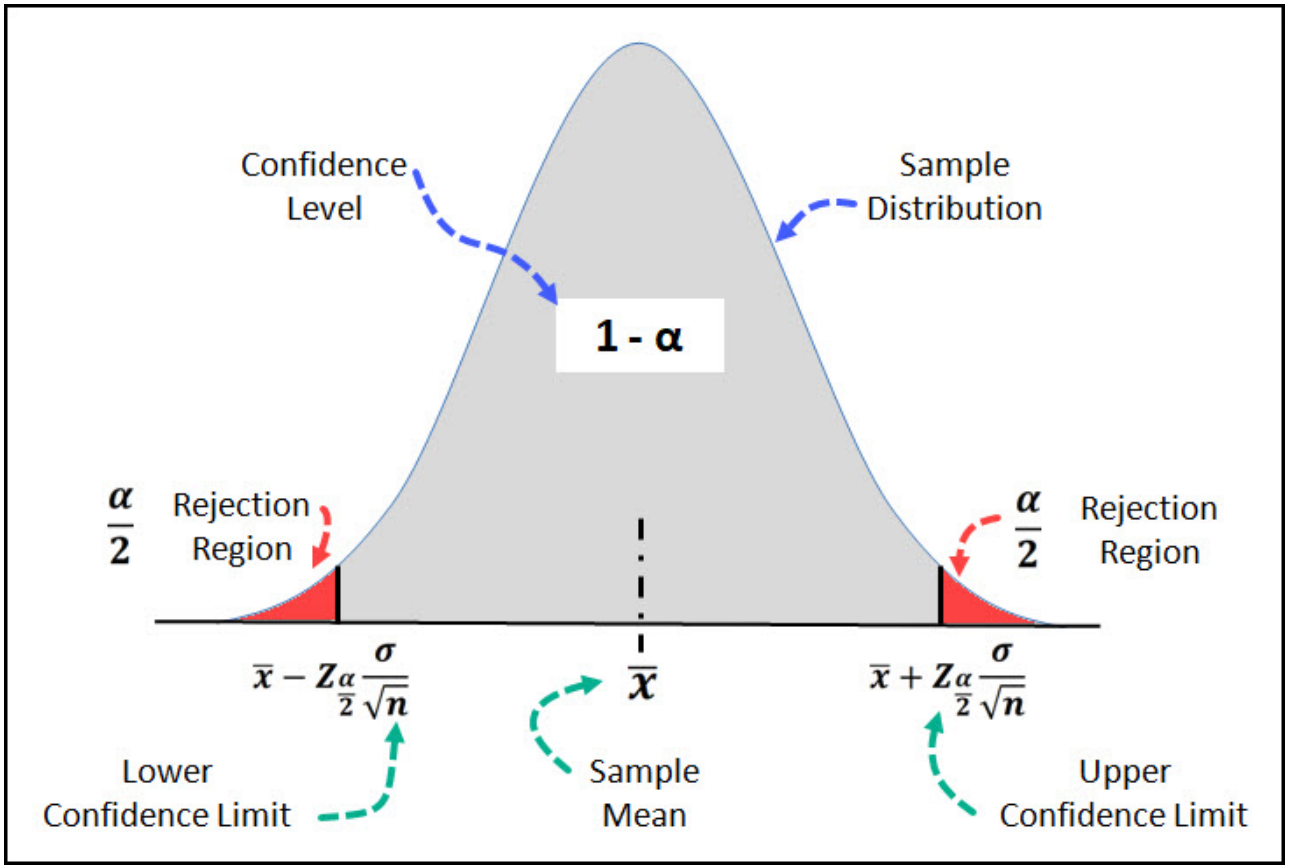


Figure 10: Niveau de Confiance ou $1 - \alpha$

- L'erreur α sera expliqué dans la prochaine séance sur les tests d'hypothèse.

Interprétation d'un Intervalle de Confiance

Un IC de 95 % signifie que si de nombreux échantillons sont pris et qu'un IC est construit à partir de chacun, environ 95 % de ces intervalles contiendront la vraie moyenne de la population. Cela ne signifie pas qu'il y a 95 % de probabilité que l'intervalle donné contienne la moyenne populationnelle.

Exemples avec R

IC du paramètre “Glucose” dans la base de données “Pima Indian Diabetes”

Exemples avec R

IC du paramètre “Glucose” dans la base de données “Pima Indian Diabetes”

- Calculer la moyenne et l’écart type

```
data <- subset(data, Glucose > 0)
mean_glucose <- mean(data$Glucose, na.rm = TRUE)
sd_glucose <- sd(data$Glucose, na.rm = TRUE)
```

Exemples avec R

IC du paramètre “Glucose” dans la base de données “Pima Indian Diabetes”

- Calculer la moyenne et l’écart type

```
data <- subset(data, Glucose > 0)
mean_glucose <- mean(data$Glucose, na.rm = TRUE)
sd_glucose <- sd(data$Glucose, na.rm = TRUE)
```

- Déterminer la taille de l’échantillon et l’erreur standard

```
n <- sum(!is.na(data$Glucose))
se <- sd_glucose / sqrt(n)
```

Exemples avec R

IC du paramètre “Glucose” dans la base de données “Pima Indian Diabetes”

- Calculer la moyenne et l’écart type

```
data <- subset(data, Glucose > 0)
mean_glucose <- mean(data$Glucose, na.rm = TRUE)
sd_glucose <- sd(data$Glucose, na.rm = TRUE)
```

- Déterminer la taille de l’échantillon et l’erreur standard

```
n <- sum(!is.na(data$Glucose))
se <- sd_glucose / sqrt(n)
```

- Préciser le niveau de confiance de 95 % avec le niveau de signification de 0,05

```
alpha <- 0.05
```

Exemples avec R (suite)

- Calculer la valeur critique à partir de la distribution t (t-score)

```
t_critical <- qt(1 - alpha/2, df = n - 1)
```

Exemples avec R (suite)

- Calculer la valeur critique à partir de la distribution t (t-score)

```
t_critical <- qt(1 - alpha/2, df = n - 1)
```

- Calculer la marge d'erreur et le IC

```
margin_error <- t_critical * se
ci_lower <- mean_glucose - margin_error
ci_upper <- mean_glucose + margin_error
```

Exemples avec R (suite)

- Calculer la valeur critique à partir de la distribution t (t-score)

```
t_critical <- qt(1 - alpha/2, df = n - 1)
```

- Calculer la marge d'erreur et le IC

```
margin_error <- t_critical * se
ci_lower <- mean_glucose - margin_error
ci_upper <- mean_glucose + margin_error
```

L'intervalle de confiance à 95 % pour le taux de glucose plasmatique moyen se situe entre 119.5 et 123.9 mg/dL.

Exemples avec R (suite)

Et les ICs à 95 % du taux de glucose plasmatique des personnes non diabétiques (Outcome = 0) vs diabétiques (Outcome = 1) ?

Exemples avec R (suite)

Et les ICs à 95 % du taux de glucose plasmatique des personnes non diabétiques (Outcome = 0) vs diabétiques (Outcome = 1) ?

- Séparer des données

```
data_outcome_0 <- subset(data, Outcome == 0)
data_outcome_1 <- subset(data, Outcome == 1)
```

Exemples avec R (suite)

Et les ICs à 95 % du taux de glucose plasmatique des personnes non diabétiques (Outcome = 0) vs diabétiques (Outcome = 1) ?

- Séparer des données

```
data_outcome_0 <- subset(data, Outcome == 0)
data_outcome_1 <- subset(data, Outcome == 1)
```

- Calculer les ICs

```
## Outcome 0: [ 108.4602 , 112.8275 ]

## Outcome 1: [ 138.7462 , 145.8929 ]
```

Exemples avec R (suite)

Et les ICs à 95 % du taux de glucose plasmatique des personnes non diabétiques (Outcome = 0) vs diabétiques (Outcome = 1) ?

- Séparer des données

```
data_outcome_0 <- subset(data, Outcome == 0)
data_outcome_1 <- subset(data, Outcome == 1)
```

- Calculer les ICs

```
## Outcome 0: [ 108.4602 , 112.8275 ]

## Outcome 1: [ 138.7462 , 145.8929 ]
```

Observez que ces deux intervalles de confiance ne se chevauchent pas.

Travaux pratiques

Travaux pratiques

Pour chacun des neuf paramètres dans la base de données sur les diabètes . . .

1. Calculer l'intervalle de confiance

En séparant les premiers huits paramètres selon le paramètre "Outcome" (0 ou 1) . . .

2. Calculer l'IC pour chaque groupe
3. Déterminer si les IC des deux groupes sont différents
4. (Bonus) Créer un graphique pour présenter la distribution de probabilité avec l'estimation ponctuelle et la marge d'erreur