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
- Tests d'hypothèse
  - Types d'erreur
  - Test sur la moyenne d'un échantillon
  - Test sur la moyenne des deux échantillons
  - 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.

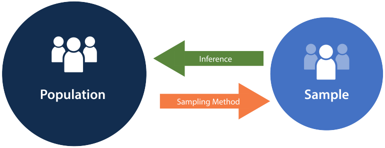


Figure 2: Inférence statistique

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.

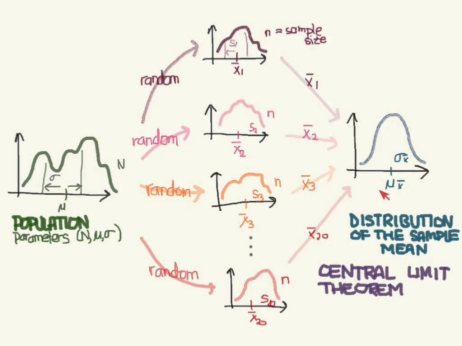


Figure 3: Théorème Central Limite

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.

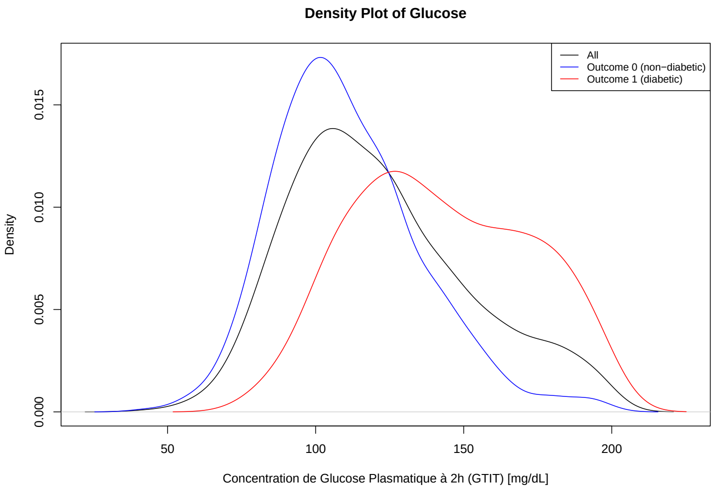
Lorsque σ est Connue

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

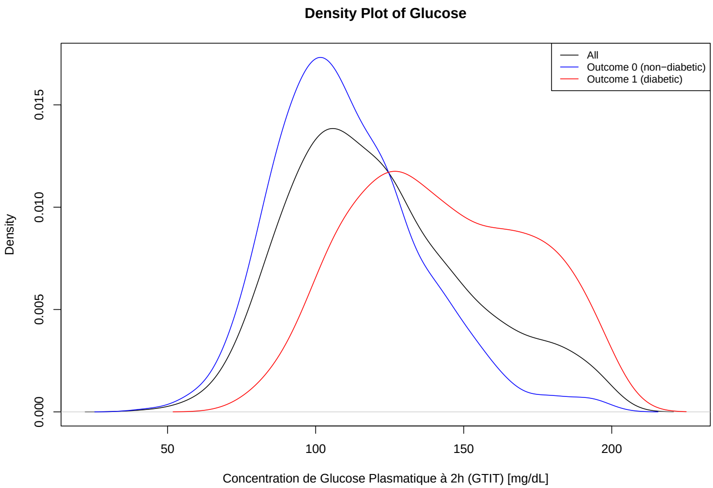
Lorsque σ est Inconnue

- 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)



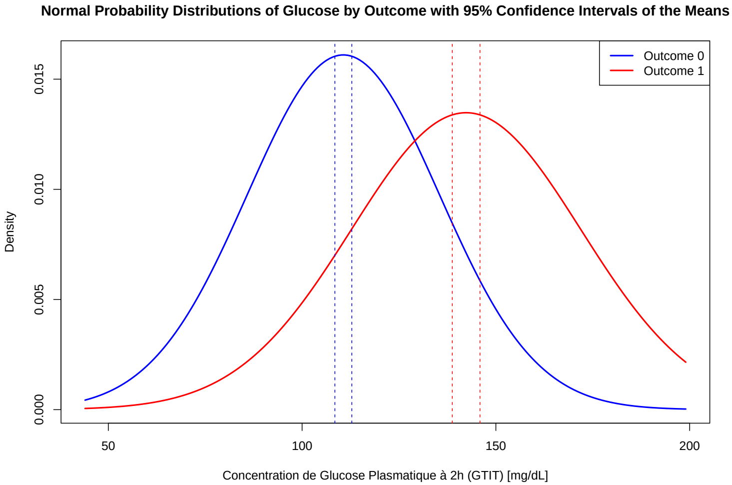
Intervalle de confiance (dernier TP)



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

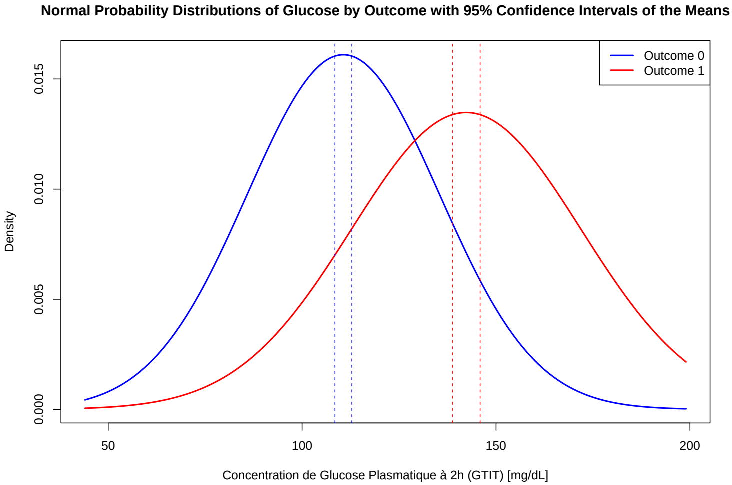
Intervalle de confiance (dernier TP)

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 sur la moyenne d'un échantillon
  - Test sur la moyenne des deux échantillons
  - 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

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 ?

Alors, voici un exemple ...

Figure 4: Erreur de 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

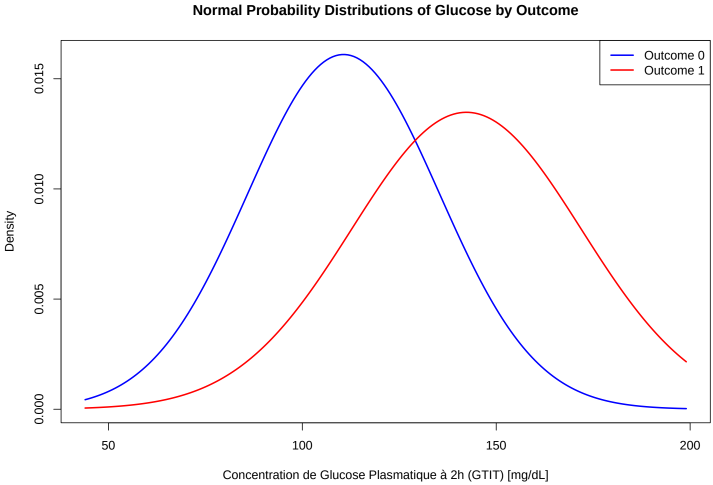
Beta β représente la probabilité de commettre une erreur de type II dans un test d'hypothèse. C'est la probabilité de ne pas rejeter une hypothèse nulle fausse.

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.

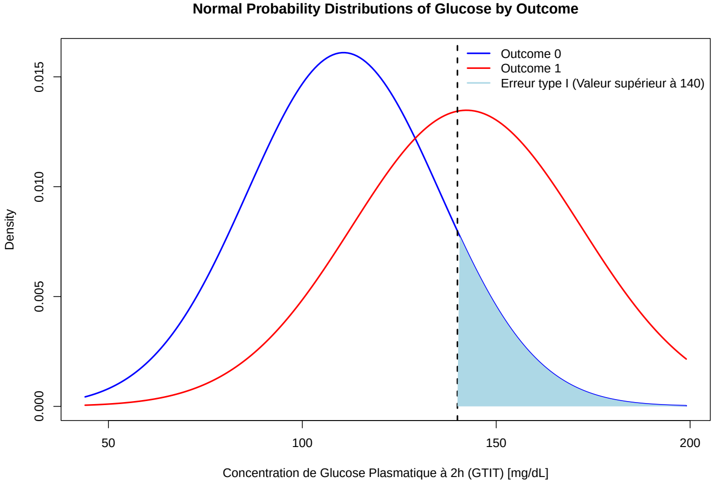
Types d'erreur : mise en application

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.

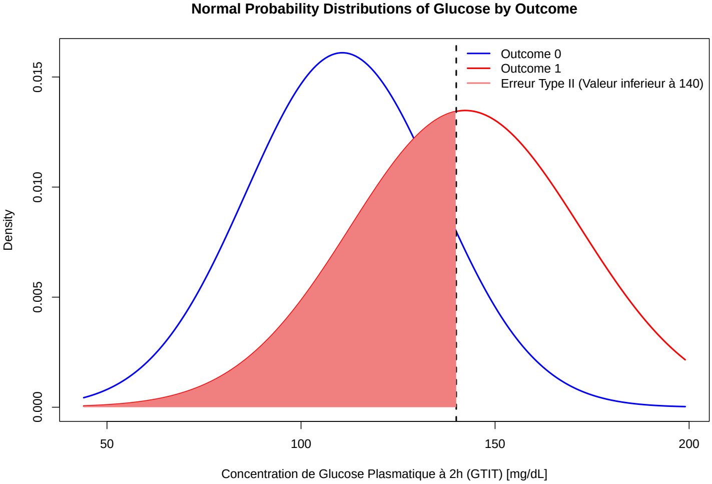


Tableau de contingence

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 : 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.

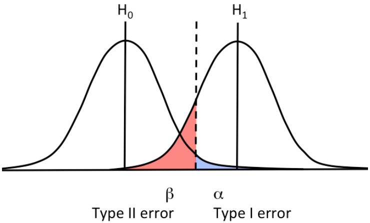


Figure 5: Équilibre entre α et β

Tests d'hypothèse

Tests d'hypothèse

Les tests d'hypothèses sont utilisés pour déterminer s'il existe suffisamment de preuves pour soutenir une croyance ou une théorie particulière sur un paramètre de population.

Tests d'hypothèse

Les tests d'hypothèses sont utilisés pour déterminer s'il existe suffisamment de preuves pour soutenir une croyance ou une théorie particulière sur un paramètre de population.

Exemples dans le génie mécanique :

- Résistance des matériaux : Tester si un matériau composite possède une résistance à la traction supérieure à celle d'un matériau standard.
- Efficacité thermique : Comparer l'efficacité thermique de deux liquides de refroidissement moteur différents.
- Aérodynamique : Évaluer l'impact d'un nouveau design d'aile sur le rapport portance/traînée d'un aéronef.
- Analyse des Vibrations : Analyser si une nouvelle technique d'amortissement des vibrations réduit plus efficacement les vibrations qu'une méthode actuelle.

Hypothèse nulle (H_0) et alternative (H_1)

L'**hypothèse nulle** est une déclaration indiquant qu'il n'y a aucun effet ou aucune différence dans un contexte particulier. C'est une position par défaut suggérant que toute différence ou signification observée dans un ensemble de données est purement due au hasard.

Hypothèse nulle (H_0) et alternative (H_1)

L'**hypothèse nulle** est une déclaration indiquant qu'il n'y a aucun effet ou aucune différence dans un contexte particulier. C'est une position par défaut suggérant que toute différence ou signification observée dans un ensemble de données est purement due au hasard.

L'**hypothèse alternative** est ce que l'on souhaite prouver. C'est une déclaration qui indique une différence ou un effet. Cette hypothèse est acceptée uniquement lorsque les données fournissent suffisamment de preuves pour rejeter l'hypothèse nulle.

Démarches dans les tests d'hypothèses

Dans les tests d'hypothèses, on commence par supposer que l'hypothèse nulle est vraie. Ensuite, en fonction des données de l'échantillon, on teste cette hypothèse. Si les preuves sont suffisamment fortes contre H_0, on la rejette en faveur de H_1. La décision est souvent prise en utilisant les valeurs p et des niveaux de signification prédéfinis (comme $\alpha = 0,05$).

Test sur la moyenne d'un échantillon

Test sur la moyenne d'un échantillon

1. Formulation des hypothèses

H_0 postule que la moyenne de la population (μ) est égale à une valeur spécifique (μ_0).

- Formellement, $H_0 : \mu = \mu_0$

H_1 ou H_a suggère que la moyenne de la population diffère de cette valeur.

- $H_0 : \mu \neq \mu_0$ (test bilatéral), $H_0 : \mu > \mu_0$ (test unilatéral droit), ou $H_0 : \mu < \mu_0$ (test unilatéral gauche)

Test sur la moyenne d'un échantillon

1. Formulation des hypothèses

H_0 postule que la moyenne de la population (μ) est égale à une valeur spécifique (μ_0).

- Formellement, $H_0 : \mu = \mu_0$

H_1 ou H_a suggère que la moyenne de la population diffère de cette valeur.

- $H_0 : \mu \neq \mu_0$ (test bilatéral), $H_0 : \mu > \mu_0$ (test unilatéral droit), ou $H_0 : \mu < \mu_0$ (test unilatéral gauche)

2. Collecte de données et calculs statistiques

À partir des données collectées, calculer

- la moyenne de l'échantillon ($\overline{x}$),
- l'écart-type de l'échantillon (s) et
- la taille de l'échantillon (n).

Test sur la moyenne d'un échantillon

3. Choix du test statistique et calculs

Utiliser un test (t) de Student pour comparer la moyenne de l'échantillon à la valeur spécifiée (μ_0).

- $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$.

Test sur la moyenne d'un échantillon

3. Choix du test statistique et calculs

Utiliser un test (t) de Student pour comparer la moyenne de l'échantillon à la valeur spécifiée (μ_0).

- $t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$.

4. Décision statistique et interprétation

Déterminer une **valeur critique** par le niveau de signification α (souvent 0,05).

Test sur la moyenne d'un échantillon

3. Choix du test statistique et calculs

Utiliser un test (t) de Student pour comparer la moyenne de l'échantillon à la valeur spécifiée (μ_0).

- $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$.

4. Décision statistique et interprétation

Déterminer une **valeur critique** par le niveau de signification α (souvent 0,05).
Si la valeur t est *en dehors des limites* de la valeur critique, rejeter H_0.
Rejeter H_0 indique que la moyenne de la population diffère significativement de μ_0.

Test sur la moyenne d'un échantillon

3. Choix du test statistique et calculs

Utiliser un test (t) de Student pour comparer la moyenne de l'échantillon à la valeur spécifiée (μ_0).

- $t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$.

4. Décision statistique et interprétation

Déterminer une **valeur critique** par le niveau de signification α (souvent 0,05).
Si la valeur t est *en dehors des limites* de la valeur critique, rejeter H_0.
Rejeter H_0 indique que la moyenne de la population diffère significativement de μ_0.
Si H_0 n'est pas rejetée, il n'y a pas suffisamment de preuves pour affirmer que la moyenne de la population est différente de μ_0.

Valeur critique

La **valeur critique** est un seuil utilisé pour déterminer si la statistique de test calculée (comme une valeur t ou z) indique un résultat statistiquement significatif, à partir de laquelle on décide de rejeter ou non l'hypothèse nulle.

Valeur critique

La **valeur critique** est un seuil utilisé pour déterminer si la statistique de test calculée (comme une valeur t ou z) indique un résultat statistiquement significatif, à partir de laquelle on décide de rejeter ou non l'hypothèse nulle.

La valeur critique est déterminée en fonction du niveau de signification α (souvent 0,05) et, pour un test t, du nombre de degrés de liberté (df ou "degree of freedom" en anglais).

Valeur critique

La **valeur critique** est un seuil utilisé pour déterminer si la statistique de test calculée (comme une valeur t ou z) indique un résultat statistiquement significatif, à partir de laquelle on décide de rejeter ou non l'hypothèse nulle.

La valeur critique est déterminée en fonction du niveau de signification α (souvent 0,05) et, pour un test t, du nombre de degrés de liberté (*df* ou "degree of freedom" en anglais).

```
qt(1 - 0.05/2, df = 29)
```

```
## [1] 2.04523
```

Valeur critique

La **valeur critique** est un seuil utilisé pour déterminer si la statistique de test calculée (comme une valeur t ou z) indique un résultat statistiquement significatif, à partir de laquelle on décide de rejeter ou non l'hypothèse nulle.

La valeur critique est déterminée en fonction du niveau de signification α (souvent 0,05) et, pour un test t, du nombre de degrés de liberté (*df* ou "degree of freedom" en anglais).

```
qt(1 - 0.05/2, df = 29)
```

```
## [1] 2.04523
```

Pour un test z, la valeur critique dépend de la distribution normale standard, $N(0, 1)$.

Valeur critique

La **valeur critique** est un seuil utilisé pour déterminer si la statistique de test calculée (comme une valeur t ou z) indique un résultat statistiquement significatif, à partir de laquelle on décide de rejeter ou non l'hypothèse nulle.

La valeur critique est déterminée en fonction du niveau de signification α (souvent 0,05) et, pour un test t, du nombre de degrés de liberté (*df* ou "degree of freedom" en anglais).

```
qt(1 - 0.05/2, df = 29)
```

```
## [1] 2.04523
```

Pour un test z, la valeur critique dépend de la distribution normale standard, $N(0, 1)$.

```
qnorm(0.975)
```

```
## [1] 1.959964
```

Tests d'hypothèse avec la distribution t

Test bilatéral (à deux queues)

- **Valeur critique** : $t_{\alpha/2,\text{df}}$ et $-t_{\alpha/2,\text{df}}$
- Rejeter de H_0 si $t_{\text{test}} > t_{\alpha/2,\text{df}}$ ou $t_{\text{test}} < -t_{\alpha/2,\text{df}}$

Test unilatéral gauche (à queue gauche)

- **Valeur critique** : $t_{\alpha,\text{df}}$
- Rejeter de H_0 si $t_{\text{test}} < -t_{\alpha,\text{df}}$

Test unilatéral droit (à queue droite)

- **Valeur critique** : $t_{\alpha,\text{df}}$
- Rejeter de H_0 si $t_{\text{test}} > t_{\alpha,\text{df}}$

Tests d'hypothèse avec la distribution t

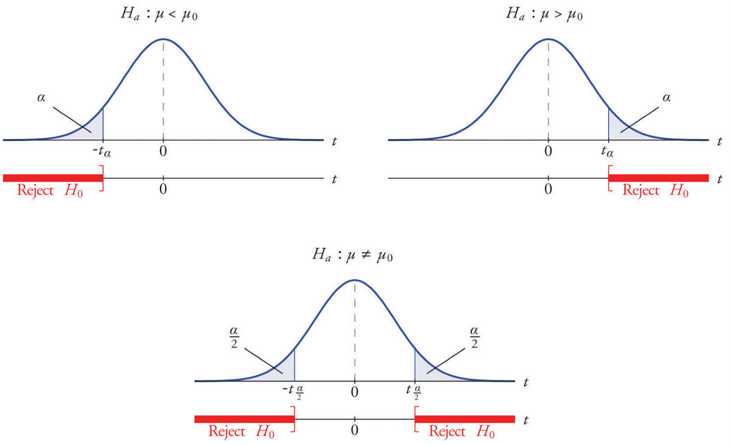


Figure 6: Rejeter H_0

Exemples avec R

À partir de la base des données “Pima Indian Diabetes” . . .

Nous voulons tester si la population étudiée a un IMC moyen considéré comme normal, défini entre 18,5 et 24,9, avec un niveau de signification α de 0,05.

Exemples avec R

À partir de la base des données “Pima Indian Diabetes” ...

Nous voulons tester si la population étudiée a un IMC moyen considéré comme normal, défini entre 18,5 et 24,9, avec un niveau de signification α de 0,05.

1. Formulation des hypothèses

H_0 : La moyenne de l'IMC de la population est normale (c'est-à-dire, située dans la plage 18.5 - 24.9).

- **Formellement**: $H_0 : 18.5 \leq \mu \leq 24.9$

H_1 : La moyenne de l'IMC de la population n'est pas normale (c'est-à-dire, en dehors de la plage 18.5 - 24.9).

- **Formellement**: $H_1 : \mu < 18.5$ ou $\mu > 24.9$

Exemples avec R (suite)

2. Collecte de données et calculs statistiques

```
filtered_bmi <- data$BMI[data$BMI>0 & !is.na(data$BMI)]
mean_bmi <- mean(filtered_bmi)
sd_bmi <- sd(filtered_bmi)
n <- length(filtered_bmi)
```

Exemples avec R (suite)

2. Collecte de données et calculs statistiques

```
filtered_bmi <- data$BMI[data$BMI>0 & !is.na(data$BMI)]
mean_bmi <- mean(filtered_bmi)
sd_bmi <- sd(filtered_bmi)
n <- length(filtered_bmi)
```

3. Choix du test statistique et calculs

```
mu_0_lower <- 18.5
mu_0_upper <- 24.9
tstat_upper <- (mean_bmi-mu_0_lower) / (sd_bmi/sqrt(n))
tstat_lower <- (mean_bmi-mu_0_upper) / (sd_bmi/sqrt(n))
cat(tstat_lower, tstat_upper)
```

```
## 30.02652 55.45432
```

Exemples avec R (suite)

4. Décision statistique et interprétation

Déterminer les valeurs critiques $t_{-\alpha/2}$ et $t_{\alpha/2}$

Exemples avec R (suite)

4. Décision statistique et interprétation

Déterminer les valeurs critiques $t_{-\alpha/2}$ et $t_{\alpha/2}$

```
alpha <- 0.05
tcrit_lower <- qt(alpha/2, df = n-1)
tcrit_upper <- qt(1 - alpha/2, df = n-1)
cat(tcrit_lower, tcrit_upper)
```

```
## -1.963107 1.963107
```

Exemples avec R (suite)

4. Décision statistique et interprétation

Déterminer les valeurs critiques $t_{-\alpha/2}$ et $t_{\alpha/2}$

```
alpha <- 0.05
tcrit_lower <- qt(alpha/2, df = n-1)
tcrit_upper <- qt(1 - alpha/2, df = n-1)
cat(tcrit_lower, tcrit_upper)
```

```
## -1.963107 1.963107
```

Les statistiques t sont supérieures aux valeurs critiques. Donc, rejeter H_0 et conclure que l'IMC moyen n'est pas normal.

Test sur la moyenne des deux échantillons

Quelques exercices stimulants . . .

Réponses anonymes

Go to wooclap.com

Enter the event code FFBQSE

Figure 7: Lien à l'activité sur Wooclap

Test t pour deux échantillons indépendants

1. Hypothèses

H_0 : La moyenne de la première population (μ_1) est égale à la moyenne de la seconde population (μ_2).

- Formellement, H_0: $\mu_1 = \mu_2$

H_1 : Les moyennes des deux populations sont différentes.

- Formellement, H_1: $\mu_1 \neq \mu_2$

Test t pour deux échantillons indépendants

1. Hypothèses

H_0 : La moyenne de la première population (μ_1) est égale à la moyenne de la seconde population (μ_2).

- Formellement, H_0: $\mu_1 = \mu_2$

H_1 : Les moyennes des deux populations sont différentes.

- Formellement, H_1: $\mu_1 \neq \mu_2$

2. Conditions du test – Important!

- Les échantillons sont indépendants.
- Les échantillons sont suffisamment grands ou proviennent de distributions normalement distribuées.
- Les variances des deux populations sont supposées égales.

Test t pour deux échantillons indépendants

3. Statistique de test

La statistique de test t est calculée comme suit :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

où $\overline{x}_1$ et $\overline{x}_2$ sont les moyennes des échantillons, n_1 et n_2 les tailles des échantillons, et s_p^2 la variance combinée.

Test t pour deux échantillons indépendants

4. Décision et interprétation

Comparer la valeur de t calculée à une valeur critique de la distribution t.

- Les degrés de liberté sont $\mathrm{df} = n_1 + n_2 - 2$.
- Le niveau de signification habituel : $\alpha = 0.05$.
- Rejeter H0 si $|t| > t_{\text{critique}}$.

Test t pour deux échantillons indépendants

4. Décision et interprétation

Comparer la valeur de t calculée à une valeur critique de la distribution t.

- Les degrés de liberté sont $\text{df} = n_1 + n_2 - 2$.
- Le niveau de signification habituel : $\alpha = 0.05$.
- Rejeter H0 si $|t| > t_{\text{critique}}$.

Rejeter H_0 indique une différence statistiquement significative entre les moyennes des deux populations.
Si H_0 n'est pas rejetée, cela suggère l'absence de preuve suffisante d'une différence.

Test t pour deux échantillons appariés

Pour des échantillons appariés, la méthode est similaire mais basée sur les différences appariées entre les deux ensembles de données.

Test t pour deux échantillons appariés

Pour des échantillons appariés, la méthode est similaire mais basée sur les différences appariées entre les deux ensembles de données.

Quelques exemples :

- Mesurer la capacité de charge d'une batterie *avant* et *après* un nombre de cycles de charge pour déterminer la perte de capacité au fil du temps
- Comparer la force d'un joint soudé *avant* et *après* un processus de vieillissement accéléré pour simuler l'effet à long terme de l'utilisation
- Évaluer la préférence ou le confort d'utilisateurs avec deux designs différents d'un produit pour déterminer lequel est ergonomiquement supérieur (A|B testing)

Test t pour deux échantillons appariés

Hypothèse

$$H_0 : \mu_{\text{diff}} = 0$$

$$H_1 : \mu_{\text{diff}} \neq 0$$

où μ_{diff} est la moyenne des différences entre les paires.

Test t pour deux échantillons appariés

Hypothèse

$$H_0 : \mu_{\text{diff}} = 0$$

$$H_1 : \mu_{\text{diff}} \neq 0$$

où μ_{diff} est la moyenne des différences entre les paires.

Statistique de test

$$t = \frac{\overline{d}}{s_d/\sqrt{n}}$$

où $\overline{d}$ est la moyenne des différences entre les deux mesures $d_i = x_{i1} - x_{i2}$ pour chaque paire i, s_d est l'écart-type des différences et n est le nombre de paires.

Exemples avec R

À partir de la base des données “Pima Indian Diabetes” ...

Nous voulons tester si les IMC des deux groupes (diabétiques et non diabétiques) sont égaux.

Exemples avec R

À partir de la base des données “Pima Indian Diabetes” ...

Nous voulons tester si les IMC des deux groupes (diabétiques et non diabétiques) sont égaux.

Les deux groupes sont-ils indépendants ou appariés ?

Exemples avec R

À partir de la base des données “Pima Indian Diabetes” . . .

Nous voulons tester si les IMC des deux groupes (diabétiques et non diabétiques) sont égaux.

Les deux groupes sont-ils indépendants ou appariés ?

Séparer des données en deux groupes basés sur ‘Outcome’

```
filtered_data <- subset(data, BMI > 0)
bmi_non_diabetic <-
  filtered_data$BMI[filtered_data$Outcome == 0]
bmi_diabetic <-
  filtered_data$BMI[filtered_data$Outcome == 1]
```

Exemples avec R (suite)

Vérifier la condition d'homogénéité des variances

```r
var.test(bmi_non_diabetic, bmi_diabetic)
```

```
##
##  F test to compare two variances
##
## data:  bmi_non_diabetic and bmi_diabetic
## F = 0.98367, num df = 490, denom df = 265, p-value = 0.870
## alternative hypothesis: true ratio of variances is not equ
## 95 percent confidence interval:
##  0.7932778 1.2115225
## sample estimates:
## ratio of variances
##          0.9836665
```

Exemples avec R (suite)

Effectuer le test t pour échantillons indépendants

```
t.test(bmi_non_diabetic, bmi_diabetic)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi_non_diabetic and bmi_diabetic
## t = -9.055, df = 539.79, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not eq
## 95 percent confidence interval:
##  -5.533527 -3.560659
## sample estimates:
## mean of x mean of y
##  30.85967  35.40677
```

Attention!

Les données ne sont pas normalement distribuées. Alors, il faut utiliser un test non paramétrique.

```
shapiro.test(bmi_non_diabetic)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bmi_non_diabetic
## W = 0.98054, p-value = 3.858e-06
```

```
shapiro.test(bmi_diabetic)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bmi_diabetic
## W = 0.94938, p-value = 5.738e-08
```

Test non paramétrique

Test de Wilcoxon

Le test de Wilcoxon, également connu sous le nom de test de rang signé de Wilcoxon, est un test non paramétrique utilisé en statistique pour comparer deux échantillons appariés ou pour tester des différences dans un seul échantillon apparié ou répété. Il est utilisé lorsque les conditions pour un test t paramétrique ne sont pas remplies, notamment lorsque les données ne suivent pas une distribution normale.

Test de Wilcoxon

Le test de Wilcoxon, également connu sous le nom de test de rang signé de Wilcoxon, est un test non paramétrique utilisé en statistique pour comparer deux échantillons appariés ou pour tester des différences dans un seul échantillon apparié ou répété. Il est utilisé lorsque les conditions pour un test t paramétrique ne sont pas remplies, notamment lorsque les données ne suivent pas une distribution normale.

- **Test de Rang Somme de Wilcoxon** : Utilisé pour comparer deux échantillons indépendants. Ce test est également connu sous le nom de test de Mann-Whitney.
- **Test de Wilcoxon pour Échantillons Appariés** : Utilisé pour comparer deux échantillons appariés ou mesures répétées sur un même groupe (par exemple, avant et après un traitement).

Exemples avec R

Avec les mêmes données ...

```
wilcox.test(bmi_non_diabetic, bmi_diabetic)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bmi_non_diabetic and bmi_diabetic
## W = 40875, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal t
```

Résumé des démarches

Résumé des démarches

Réponses anonymes

Go to wooclap.com

Enter the event code FFBQSE

Figure 8: Lien à l'activité sur Wooclap

Réponse correcte

- Formulation des hypothèses
- Collecte de données
- Calculs de la moyenne, de l'écart type et de la taille de l'échantillon
- Choix du test statistique et calculs
- Détermination des valeurs critiques par le niveau de signification
- Décision statistique de rejeter l'hypothèse nulle ou non
- Interprétation

Figure 9: Démarches de test d'hypothèse

Travaux pratiques

Travaux pratiques

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

1. Vérifier les conditions pour utiliser un test t
2. Effectuer le test t ou Wilcoxon pour comparer les valeurs entre les deux groupes (non diabétique et diabétique)
3. Interpréter les résultats