Epidémiologie

Plan

1. Introduction
2. Enquête de cohorte
3. Enquête cas-témoins
4. Mesures de risques
5. Mesures d'association
6. Biais de sélection
7. Biais de classement
8. Biais de confusion
9. Stratégie d'analyse
10. Puissance
11. Modèles multivariés
12. Régression logistique

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

Epidémiologie

9. Protocole et stratégie d'analyse

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

9. Protocole et stratégie d'analyse

I. Objectifs de l'enquête

II. Choix du type d'enquête

III. Modalités de réalisation

IV. Puissance et nombre de sujets nécessaires

V. Démarche de l'analyse statistique

VI. Interprétation des résultats

I. Objectifs de l'enquête

- Enquêtes descriptives
  - Estimer la fréquence d'une maladie
  - Décrire les tendances temporelles et spatiales
  - Générer des hypothèse
- Enquêtes analytiques
  - Identifier des facteurs de risque
  - Identifier des facteurs causaux
  - Identifier les facteurs expliquant un comportement
- Enquêtes expérimentales
  - Démontrer l'efficacité d'un traitement ou d'une intervention

II. Choix du type d'enquête

- Enquêtes descriptives
  - Surveillance épidémiologique à partir de données d'enregistrement continu
  - générer des hypothèses
- Enquêtes analytiques
  - Enquêtes observationnelles (Enquêtes de cohorte, Cas-témoins et transversale)
  - Causalité non démontrable (faisceau d'arguments)
- Enquêtes quasi-expérimentales
  - Enquêtes "avant-après" ou "ici-ailleurs"
  - Niveau de preuve plus faible qu'une enquête expérimentale
- Enquêtes expérimentales
  - Essai contrôlé, randomisé, en double aveugle
  - Causalité démontrée

Enquête de cohorte ou enquête cas-témoins

	Cohorte	Cas-témoins
Adapté aux exposition rares	Oui	Non
Adapté aux maladies rares	Non	Oui
Choix des groupes de référence	Non exposé relativement facile à trouver	Témoins difficile à trouver Biais de sélection
Rapidité, coût	Long à cause du suivi (plus rapide si cohorte rétrospective)	Rapide si période d'inclusion courte (plus long si cas incidents)
Recueil de l'exposition	Chronologie entre E et M connue (suivi longitudinal)	Recueil de E toujours rétrospectif Biais de classement (différentiel ou non)
Perdu de vue	Oui (+ si cohorte rétrospective) Biais de sélection	Non (pas de suivi)
Constitution d'une base de données pour des objectifs multiples	Oui	Non
Causalité	Niveau de preuve supérieur à une enquête cas-témoins car les risques de biais sont moins importants	Niveau de preuve plus faible qu'une enquête de cohorte car les risques de biais sont plus importants

Enquête de cohorte ou enquête cas-témoins

- Quantités estimables dans une enquête de cohorte
  - Prévalence de l'exposition si construite à partir d'une enquête transversale
  - Risque relatif et odds ratio
  - Le suivi longitudinal permet d'estimer
    - Le risque de la maladie et le taux d'incidence
    - Analyse de survie (modèle de Cox, …)

- Quantités estimables dans une enquête cas-témoins
  - Prévalence de l'exposition chez les cas
  - Prévalence de l'exposition chez les témoins → généralisation possible à la population source
  - Odds ratio mais pas le risque relatif (car pas d'estimation de la prévalence de la maladie)
  - Pas d'estimation du risque de la maladie et du taux d'incidence (pas de suivi)

III. Modalités de réalisation

1. Définition des populations cibles et sources
2. Mode d'échantillonnage
3. Recueil des données
4. Modalités pratiques

➡ Chacun de ces points est détaillé dans les chapitres concernant les enquêtes cas-témoins et les enquêtes de cohorte

IV. Puissance et nombre de sujets nécessaires

- La puissance d'une enquête pour mettre en évidence une association entre E et M
  - Dépend du nombre de sujets exploitables pour l'analyse statistique
    - Nombre de sujets à calculer au moment de la mise en place de l'enquête
    - Choisir une population source (et un protocole) permettant de minimiser le risque de non participation, de données manquantes et de perdus de vue
  - Dépend du choix des groupes comparés
    - Choisir des groupes les plus contrastés possible (grande différence entre les groupes)
    - Choisir des groupes homogènes (faible variation au sein d'un groupe)
    - La puissance meilleure si les effectifs sont équilibrés (préférer les enquêtes de cohorte et cas-témoins)
    - La puissance peut être meilleure en cas d'échantillon apparié
  - Dépend du choix des instruments de mesure de l'exposition et de la maladie
    - Instruments de mesure standardisés, objectifs et précis
    - Instruments de mesure avec une bonne sensibilité et une bonne spécificité

      ➡ limiter les biais de classement non différentiel (qui ramène l'estimation de l'OR et du RR vers 1)

V. Démarche de l'analyse statistique

- Lien entre l'exposition (E) est la maladie (M)
  - E significativement associée à M ?
    - Estimation des mesures d'association (Odds ratio, risque relatif, …)
    - Intervalle de confiance
    - Test statistique (p-value)
  - E fortement associée à M ?
    - Grande valeur de la mesure d'association
  - E est la cause de M ?
    - Pas de preuve scientifique dans les enquêtes observationnelles
    - Faisceau d'argument, critère de présomption causale de Bradford Hill
  - Contribution de E au taux d'incidence de la maladie
    - Suppose que E est la cause de M
    - Risque attribuable

Démarche de l'analyse statistique

1. Analyses descriptives : contrôle de cohérence et de la qualité des données

- Préparation d'un fichier propre, rechercher les données aberrantes
- Vérifier la comparabilité des groupes comparés
  - E+/E- ou M+/M- issus de la même population
  - Contrôler la qualité de l'appariement
- Evaluer une déformation possible par rapport à la sélection initiale
  - Non réponse (totale ou partielle), perdus de vue
- Décrire les expositions
  - Répartition, niveau, durée, type
- Evaluer la mortalité et/ou la morbidité (taux d'incidences)
  - Uniquement dans les enquêtes de cohorte
- Comparer les résultats obtenus avec ceux de la littérature (recherche de relations connues)

Démarche de l'analyse statistique

2. Première analyses statistiques : analyses univariées et stratifiées

- Comparaison (interne) des groupes E+/E- ou M+/M-
  - Estimation des risques relatifs (uniquement cohorte) ou des odds ratios
  - Analyses stratifiées
  - Recherche des interactions et estimation des RR (ou OR) ajustés par la méthode de Mantel - Haenszel

- Comparaison externe (uniquement dans les enquêtes de cohorte)
  - Standardisation des mesures de risques (taux d'incidence, prévalence, …)
    - SMR : standardisation indirecte (basée sur la mesure de risque de la population de référence)
    - CMF : standardisation directe (basée sur la structure d'âge de la population de référence)

Démarche de l'analyse statistique

3. Analyses approfondies de la relation dose-effet

- Type d'exposition, délais depuis le début ou l'arrêt de l'exposition, dose totale d'exposition
- Etude conjointe de la durée, du niveau et de la dose d'exposition
- Analyse en fonction de sous-catégories
- Recherche d'effet seuil, évaluation de la période d'induction

4. Analyses statistiques multivariées

- Dans les enquêtes de cohorte (analyse des données longitudinales)
  - Analyse de survie (Modèle de Cox), modèle mixte, modèle multi-états, …
  - Modèle de Poisson (si données groupées)
- Dans les enquêtes cas-témoins
  - Régression logistique
  - Autres méthodes de classification supervisée (SVM, CART, forêt aléatoire, …)

VI. Interprétation des résultats

- Résultats significatifs
  - Discussion sur les biais possibles
    - Biais de sélection
    - Biais de classement différentiels
    - Facteurs de confusion pas ou mal pris en compte
  - Arguments en faveur de la causalité
    - **Critères internes à l'étude**
      - force de l'association
      - Relation dose-effet
      - Pas d'ambiguïté sur la chronologie
      - Spécificité de l'association
    - **Critères externes à l'étude**
      - Constance des résultats dans la littérature
      - Plausibilité biologique (mécanismes explicatifs)
      - Cohérence des résultats avec les hypothèses de départ

Interprétation des résultats

- Résultats non significatifs
  - Ajustement sur un facteur intermédiaire (suppression de l'association)
  - Sur-appariement
  - Discussion sur les biais
    - Biais de sélection, biais de classement différentiel
    - Facteur de confusion non ou mal pris en compte
  - Manque de puissance
    - Nombre de sujets plus faible que prévu (non réponse, perdus de vues)
    - Biais de classement non différentiel
    - Evaluer la puissance a posteriori
      - Si la puissance est ≥ 80% ➡ on peut conclure à l'absence d'association
      - Si la puissance est < 80% ➡ il y a peut être une différence mais on ne la voit pas

Epidémiologie

10. Puissance d'un test statistique

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

10. Puissance d'un test statistique

I. Précision de l'estimation et nombre de sujets

- Estimation d'un pourcentage
- Estimation d'une moyenne

II. Rappels sur la puissance

III. Puissance et nombre de sujets nécessaire

- Comparaison de moyennes
- Comparaison de pourcentages, d'un OR et d'un RR à la valeur 1
- Comparaison d'un SMR à la valeur 1

IV. Puissance dans une enquête

I. Précision de l'estimation et nombre de sujets

- Contexte
  - Estimation d'une moyenne ou d'un pourcentage
  - Calcul d'un intervalle de confiance avec un risque d'erreur α fixé

- Objectif
  - Nombre de sujets nécessaire pour estimer le paramètre avec une précision i fixée
  - Soit $\hat{\theta}$ un estimateur et $IC_{\alpha}(\theta) = [\theta_{inf} ; \theta_{sup}] = [\hat{\theta} - i ; \hat{\theta} + i]$ son intervalle de confiance
  - Déterminer le nombre de sujets nécessaire pour obtenir la précision i souhaitée

Estimation d'un pourcentage

- Rappels
  - Considérons un échantillon de taille n
  - Soit P le vrai pourcentage et $\hat{P}$ un estimateur de P obtenu sur l'échantillon
  - Intervalle de confiance de P de niveau α

$$IC_{\alpha}(P) = \left[\hat{P} - i\,;\hat{P} + i\right] = \left[\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\,;\hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right]$$

- Contexte
  - On souhaite déterminer n pour estimer P avec une certaine précision i fixée
  - On se base sur une idée a priori $\tilde{P}$ (littérature) de la valeur de P attendue dans la population

$$n = \frac{\left[z_{\frac{\alpha}{2}}\right]^2}{i^2} \times \tilde{P}(1-\tilde{P})$$

Estimation d'une moyenne

- Rappels
  - Considérons un échantillon de taille n
  - Soit M la vraie moyenne et $\widehat{M}$ un estimateur de M obtenu sur l'échantillon
  - Intervalle de confiance de M de niveau α

$$IC_{\alpha}(M) = \left[\widehat{M} - i\,;\widehat{M} + i\right] = \left[\widehat{M} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}^2}{n}}\,;\widehat{M} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}^2}{n}}\right]$$

- Contexte
  - On souhaite déterminer n pour estimer M avec une certaine précision i fixée
  - On se base sur une idée a priori $\tilde{\sigma}^2$ (littérature) de la valeur de σ^2 attendue dans la population

$$n = \frac{\left[z_{\frac{\alpha}{2}}\right]^2}{i^2} \times \tilde{\sigma}^2$$

Exemples

- Estimation du pourcentage P de fumeur d'une population
  - Données a priori : le pourcentage de fumeur serait de 30% dans la population
  - Nombre de sujets à prévoir dans l'échantillon pour estimer P avec une précision de $\pm 3\%$ au risque $\alpha = 5\%$

$$n = \frac{[1.96]^2}{0.03^2} \times 0.3(1 - 0.3) = 897 \qquad (i = \pm 5\% \implies n = 323)$$

- Estimation du poids de naissance M moyen dans une population
  - Données a priori : le poids moyen serait de $3500g$ et l'écart-type de $500g$
  - Nombre de sujets à prévoir dans l'échantillon pour estimer M avec une précision de $\pm 50g$ au risque $\alpha = 5\%$

$$n = \frac{[1.96]^2}{50^2} \times 500^2 = 384 \qquad (\tilde{\sigma}^2 = 250g \implies n = 92)$$

II. Rappels sur la puissance

- Principe d'un test d'hypothèse (statistique)
  - Consiste à rejeter ou à ne pas rejeter une hypothèse H_0 à partir d'un échantillon
  - Deux risques d'erreur

		Décision	
		Rejet de H_0	Non rejet de H_0 ($\neq$ accepter H_1)
Vérité	H_0 vraie	Erreur de 1ère espèce $\alpha = P(rejet\ de\ H_0 \mid H_0\ vraie)$	Pas d'erreur $1-\alpha = P(non\ rejet\ de\ H_0 \mid H_0\ vraie)$
	H_1 vraie	Pas d'erreur $1-\beta = P(rejet\ de\ H_0 \mid H_1\ vraie)$	Erreur de 2ème espèce $\beta = P(non\ rejet\ de\ H_0 \mid H_1\ vraie)$

- On contrôle le risque d'erreur de 1ère espèce α (en général fixé à 5%) : erreur la plus grave
- Ex: Justice $\rightarrow$ condamner un innocent ou relâcher un coupable
  - H_0 : accusé est innocent et H_1 : accusé coupable
  - $\alpha =$ "condamner un innocent" et $\beta =$ "relâcher un coupable"

Rappels sur la puissance

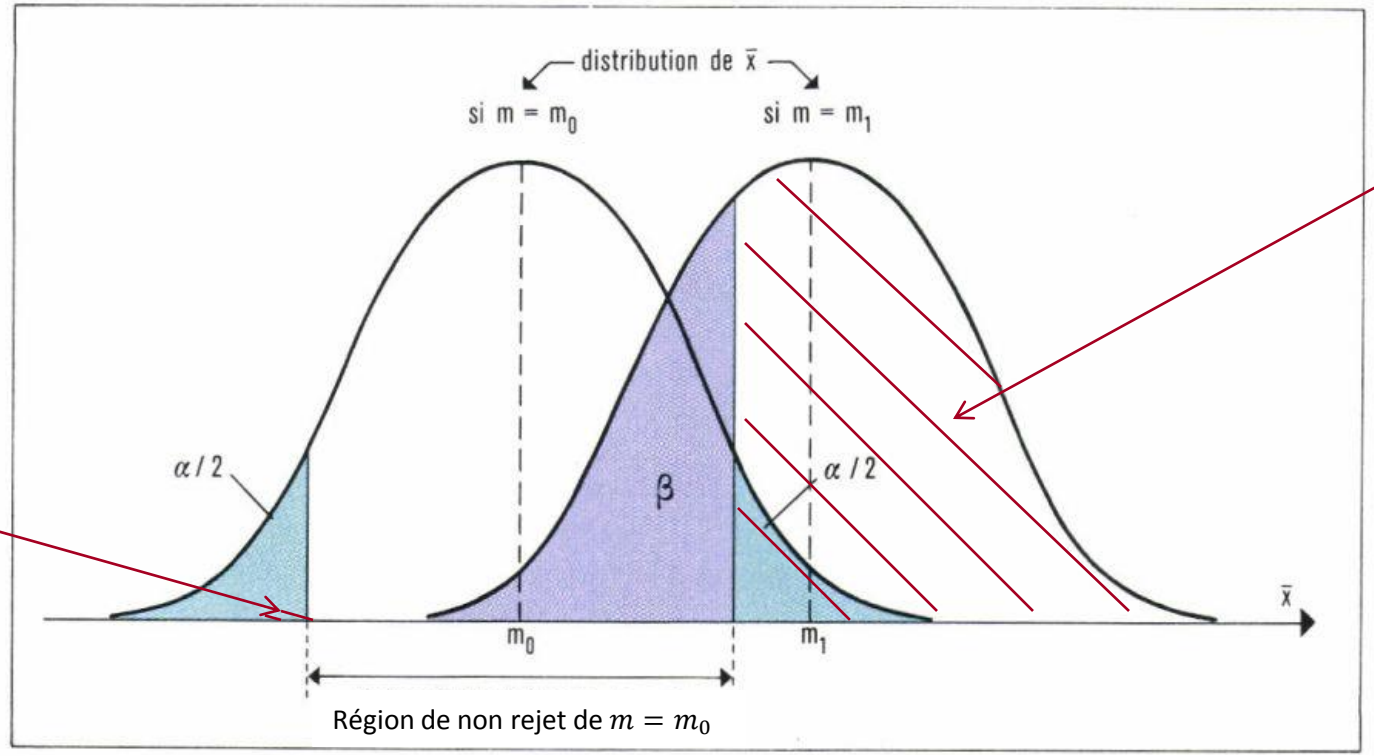
- Erreur $\alpha = P(rejet\ de\ H_0 | H_0\ vraie)$ $\rightarrow$ erreur possible quand on rejette H_0
- Erreur $\beta = P(non\ rejet\ de\ H_0 | H_1\ vraie)$ $\rightarrow$ erreur possible quand on ne rejette pas H_0
- Puissance $= 1 - \beta = P(rejet\ de\ H_0 | H_1\ vraie)$
- La puissance mesure la capacité d'un test à mettre en évidence une différence qui existe réellement
- Exemples :
  - Justice : on veut une puissance $(1 - \beta)$ suffisante pour ne pas relâcher un coupable (erreur β)
  - Comparaison de l'effet de deux traitements : éviter de passer à coté d'une différence qui pourrait permettre des progrès thérapeutique

Rappels sur la puissance

- Calcul de la puissance pour la comparaison d'une moyenne à une moyenne théorique
  - Hypothèses
    - $X_1, \dots, X_n$ un échantillon indépendant et de même loi que X de moyenne m et de variance σ^2
    - X_i normale ou $n \geq 30$ $\Rightarrow$ la moyenne estimée $\bar{X}$ est une variable normale : $\bar{X} \sim N\left(m, \frac{\sigma^2}{n}\right)$
    - σ^2 connue (pour simplifier) et identique sous H_0 et H_1
  - Test bilatéral $\begin{cases} H_0 : m = m_0 \\ H_1 : m \neq m_0 \end{cases}$ (m la vraie moyenne et m_0 la moyenne théorique)
  - Sous H_0, $m = m_0$ $\rightarrow$ la statistique de student $Z = \frac{\bar{X} - m_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$
  - Sous H_1, on pose $m = m_1$ $\rightarrow$ la statistique de student $Z = \frac{\bar{X} - m_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N\left(\frac{m_1 - m_0}{\sqrt{\frac{\sigma^2}{n}}}, 1\right)$
  - Pour calculer la puissance on doit spécifier l'hypothèse H_1

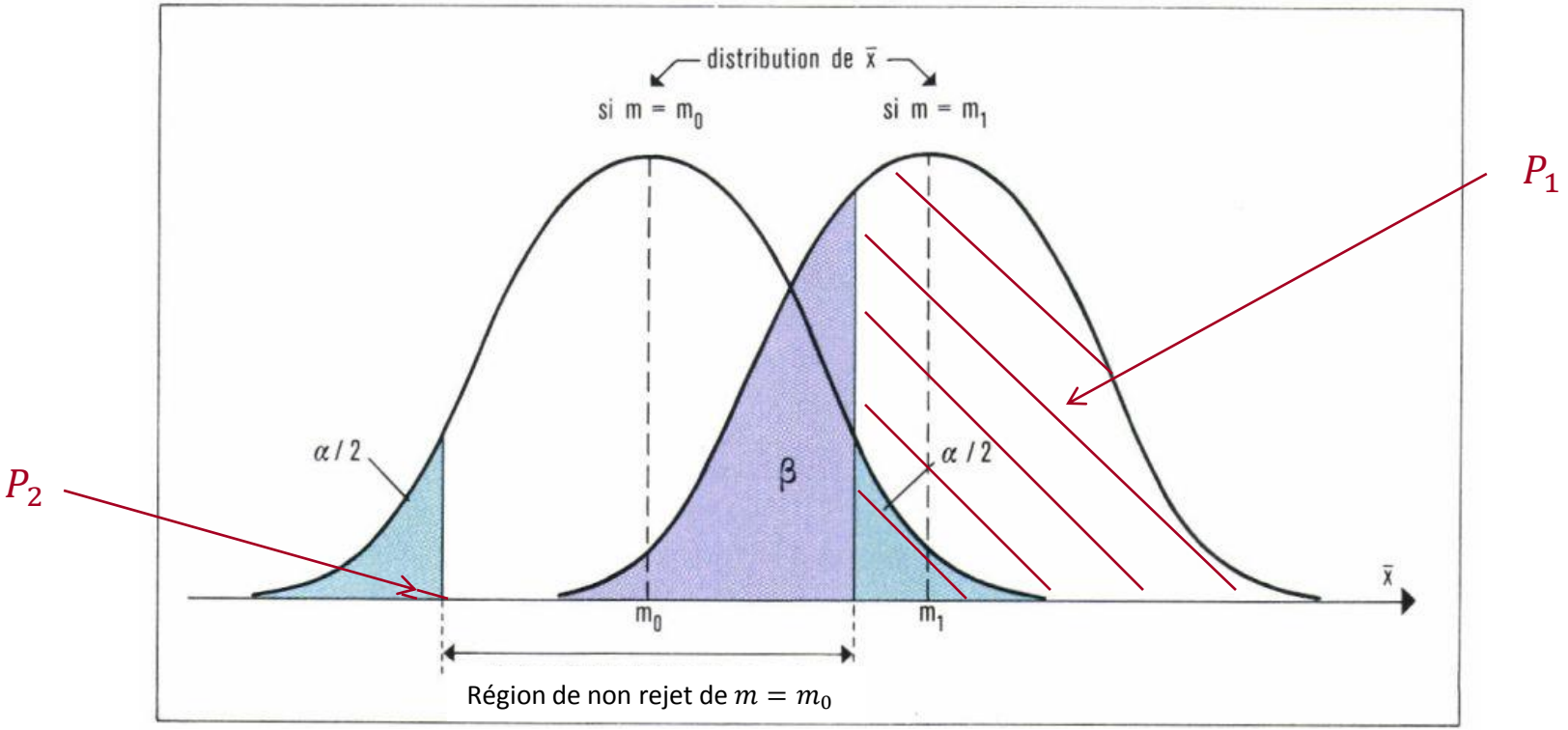
Rappels sur la puissance

- $\alpha = P(rejet\ de\ H_0 | H_0\ vraie) = P\left(|Z| \geq z_{\frac{\alpha}{2}} | H_0\ vraie\right)$
- $\beta = P(non\ rejet\ de\ H_0 | H_1\ vraie) = P\left(|Z| < z_{\frac{\alpha}{2}} | H_1\ vraie\right)$
- $1 - \beta = P(rejet\ de\ H_0 | H_1\ vraie) = P\left(|Z| \geq z_{\frac{\alpha}{2}} | H_1\ vraie\right) = P_1 + P_2$



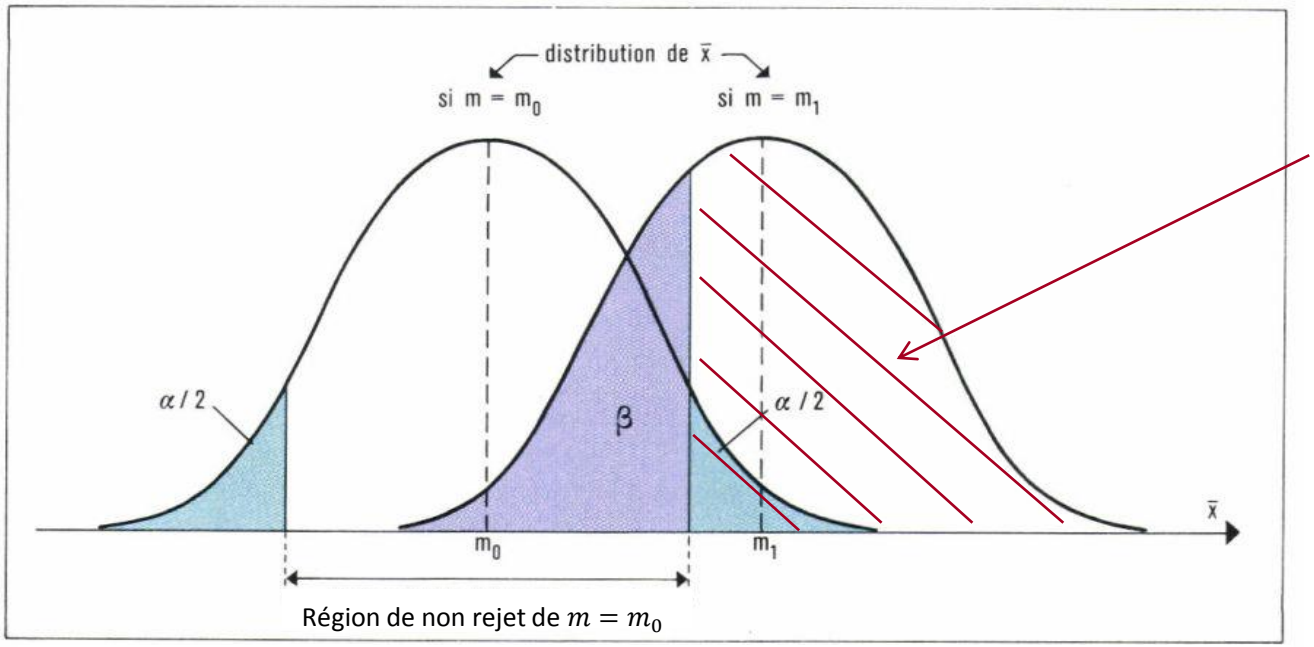
Rappels sur la puissance

- On suppose que P_2 est négligeable
- $1-\beta \approx P_1 = P\left(Z \geq z_{\frac{\alpha}{2}} \,|H_1\ vraie\right) = P\left(N(0,1) \geq z_{\frac{\alpha}{2}} - \frac{|m_1-m_0|}{\sqrt{\frac{\sigma^2}{n}}}\right)$ ➡ $z_{1-\beta} = z_{\frac{\alpha}{2}} - \frac{|m_1-m_0|}{\sqrt{\frac{\sigma^2}{n}}}$



Rappels sur la puissance

- $z_{1-\beta} = z_{\frac{\alpha}{2}} - \frac{|m_1 - m_0|}{\sqrt{\frac{\sigma^2}{n}}}$ ➡ $1-\beta$ augmente quand $z_{1-\beta}$ diminue
  - $1-\beta$ augmente quand $m_1 - m_0$ augmente
  - $1-\beta$ augmente quand σ^2 diminue
  - $1-\beta$ augmente quand n augmente
  - $1-\beta$ augmente quand $z_{\frac{\alpha}{2}}$ diminue (*i.e.* quand $\frac{\alpha}{2}$ augmente) ➡ compromis entre les 2 erreurs



III. Puissance et nombre de sujets nécessaire

Comparaison d'une moyenne à une moyenne théorique

	Test bilatéral : $H_0 : m = m_0$ $H_1 : m \neq m_0$	**Test unilatéral : $H_0 : m = m_0$ $H_1 : m > m_0$**
Puissance (Table de la loi normale Table 1 du livre "rose")	$z_{1-\beta} = z_{\frac{\alpha}{2}} - \frac{\lvert m_1 - m_0 \rvert}{\sqrt{\frac{\sigma^2}{n}}}$	$z_{1-\beta} = z_{\alpha} - \frac{\lvert m_1 - m_0 \rvert}{\sqrt{\frac{\sigma^2}{n}}}$
Puissance (Table 4a et 4b)	$\phi = \frac{\lvert m_1 - m_0 \rvert}{\sqrt{\frac{\sigma^2}{n}}} = z_{\frac{\alpha}{2}} - z_{1-\beta}$	$\phi = \frac{\lvert m_1 - m_0 \rvert}{\sqrt{\frac{\sigma^2}{n}}} = z_{\alpha} - z_{1-\beta}$
Nombre de sujets nécessaire	$n = \frac{\sigma^2}{(m_1 - m_0)^2} \times \left(z_{\alpha/2} - z_{1-\beta}\right)^2$	$n = \frac{\sigma^2}{(m_1 - m_0)^2} \times \left(z_{\alpha} - z_{1-\beta}\right)^2$

- Pour $\alpha = 5\%$: $z_{\frac{\alpha}{2}} = 1.96$ et $z_{\alpha} = 1.64$

Exemple

- Comparaison de l'effet antalgique de 2 médicaments A et B données successivement aux mêmes malades. On mesure le nombre d'heures sans douleur x_A et x_B
- Echantillon apparié ➡ pour chaque patient on calcule $d = x_A - x_B$
- Données a priori : la variance de $d = 12$
- Objectif : calculer le nombre de sujets nécessaire pour détecter une différence de 0.7h avec une puissance de 80%
- $H_0 : m_d = 0$ et $H_1 : m_d \neq 0$ avec $\alpha = 5\%$
- $$n = \frac{\sigma^2}{(m_1 - m_0)^2} \times \left(z_{\alpha/2} - z_{1-\beta}\right)^2 = \frac{12^2}{(0.7)^2} \times \left(1.96 - (-0.842)\right)^2 = 193$$
- Il faut sélectionner 193 sujets qui recevront chacun 2 traitements

Puissance : comparaison de deux moyennes

- On suppose que les deux populations de vraie moyenne m_1 et m_2 ont la même variance σ^2
- Sous H_1, $Z = \frac{\bar{X}-\bar{Y}}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \underset{H_1}{\rightarrow} N\left(\frac{m_1-m_2}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, 1\right)$

	Test bilatéral : $H_0 : m_1 = m_2$ $H_1 : m_1 \neq m_2$	**Test unilatéral : $H_0 : m_1 = m_2$ $H_1 : m_1 > m_2$**
Puissance (Table de la loi normale Table 1 du livre "rose")	$z_{1-\beta} = z_{\frac{\alpha}{2}} - \frac{\lvert m_1 - m_2 \rvert}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	$z_{1-\beta} = z_{\alpha} - \frac{\lvert m_1 - m_2 \rvert}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
Puissance (Table 4a et 4b)	$\phi = \frac{\lvert m_1 - m_2 \rvert}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = z_{\frac{\alpha}{2}} - z_{1-\beta}$	$\phi = \frac{\lvert m_1 - m_2 \rvert}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = z_{\alpha} - z_{1-\beta}$
Nombre de sujets nécessaire $n_1 = n_2 = n$	$n = \frac{2\sigma^2}{(m_1-m_2)^2} \times \left(z_{\alpha/2} - z_{1-\beta}\right)^2$	$n = \frac{2\sigma^2}{(m_1-m_2)^2} \times \left(z_{\alpha} - z_{1-\beta}\right)^2$
Nombre de sujets nécessaire $n_2 = k \times n_1$	$n_1 = \frac{k+1}{k}\frac{\sigma^2}{(m_1-m_2)^2} \times \left(z_{\alpha/2} - z_{1-\beta}\right)^2$ $n_2 = k \times n_1$	$n_1 = \frac{k+1}{k}\frac{\sigma^2}{(m_1-m_2)^2} \times \left(z_{\alpha} - z_{1-\beta}\right)^2$ $n_2 = k \times n_1$

Exemple

- Comparaison les poids de naissance des nouveau nés selon que la mère a consommé ou non du tabac pendant la grossesse

- Données a priori : Ecart-type du poids de naissance est de 500g

- Objectif : calculer la puissance pour détecter une différence de 100g avec un échantillon de 300 femmes fumeuses et 300 non fumeuses

- $H_0 : m_F = m_{NF}$ et $H_1 : m_F \neq m_{NF}$ avec $\alpha = 5\%$

- $$\phi = \frac{|m_F - m_{NF}|}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{100}{\sqrt{500^2\left(\frac{1}{300}+\frac{1}{300}\right)}} = 2.449$$

  Table 4a ➡ $1 - \beta = 69\%$

- $$z_{1-\beta} = z_{\frac{\alpha}{2}} - \frac{|m_F - m_{NF}|}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = 1.96 - \frac{100}{\sqrt{500^2\left(\frac{1}{300}+\frac{1}{300}\right)}} = -0.489$$

  Table 1 ➡ $1 - \beta = 69\%$

Puissance : comparaison d'un pourcentage

- Hypothèses
  - $X_1, \dots, X_n$ un échantillon indépendant et de même loi que $X \sim B(p)$
  - $np \geq 5$ et $n(1-p) \geq 5$ $\Rightarrow$ la moyenne estimée $\bar{X}$ vérifie $\bar{X} \to N\left(p, \frac{p(1-p)}{n}\right)$
- Problème : la variance dépend de p et n'est pas identique sous H_0 et H_1
- Solution : on utilise la transformation $V = \text{arcsinus}\left(\sqrt{p}\right)$
  - La distribution de V est approximativement une loi normale
  - $Var(V)$ est indépendante de p $\to$ $Var(V) = \frac{1}{4n}$
- Il faut se placer en mode "radian" pour calculer l'arcsinus
- Comparaison de 2 pourcentages : la variance de la statistique de test est identique sous H_0 et H_1

Puissance : comparaison d'un pourcentage

- Soit p_0 (resp. p_1) le pourcentage théorique sous H_0(resp. H_1)

	Test bilatéral : $H_0 : p = p_0$ $H_1 : p \neq p_0$	**Test unilatéral : $H_0 : p = p_0$ $H_1 : p > p_0$**
Puissance (Table de la loi normale Table 1 du livre "rose" Table 5 : valeur de Arcsin	$z_{1-\beta} = z_{\frac{\alpha}{2}} - \frac{\lvert \mathrm{Arcsin}(\sqrt{p_1}) - \mathrm{Arcsin}(\sqrt{p_0}) \rvert}{\sqrt{\frac{1}{4n}}}$	$z_{1-\beta} = z_{\alpha} - \frac{\lvert \mathrm{Arcsin}(\sqrt{p_1}) - \mathrm{Arcsin}(\sqrt{p_0}) \rvert}{\sqrt{\frac{1}{4n}}}$
Puissance (Table 4a et 4b)	$\phi = \frac{\lvert \mathrm{Arcsin}(\sqrt{p_1}) - \mathrm{Arcsin}(\sqrt{p_0}) \rvert}{\sqrt{\frac{1}{4n}}} = z_{\frac{\alpha}{2}} - z_{1-\beta}$	$\phi = \frac{\lvert \mathrm{Arcsin}(\sqrt{p_1}) - \mathrm{Arcsin}(\sqrt{p_0}) \rvert}{\sqrt{\frac{1}{4n}}} = z_{\alpha} - z_{1-\beta}$
Nombre de sujets nécessaire	$n = \frac{\left(z_{\alpha/2} - z_{1-\beta}\right)^2}{4\left(\mathrm{Arcsin}(\sqrt{p_1}) - \mathrm{Arcsin}(\sqrt{p_0})\right)^2}$	$n = \frac{\left(z_{\alpha} - z_{1-\beta}\right)^2}{4\left(\mathrm{Arcsin}(\sqrt{p_1}) - \mathrm{Arcsin}(\sqrt{p_0})\right)^2}$

Puissance : comparaison de deux pourcentages

- Soit p_1 (resp. p_2) le vrai pourcentage dans la population 1 (resp. population 2)
- Sous H_1, $Z = \frac{\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2})}{\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}}} \underset{H_1}{\rightarrow} N\left(\frac{(\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2}))}{\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}}}, 1\right)$

	Test bilatéral : $H_0 : p_1 = p_2$ $H_1 : p_1 \neq p_2$	**Test unilatéral : $H_0 : p_1 = p_2$ $H_1 : p_1 > p_2$**
Puissance (Table de la loi normale Table 1 du livre "rose" Table 5 : valeur de Arcsin	$z_{1-\beta} = z_{\frac{\alpha}{2}} - \frac{\lvert\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2})\rvert}{\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}}}$	$z_{1-\beta} = z_{\alpha} - \frac{\lvert\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2})\rvert}{\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}}}$
Puissance (Table 4a et 4b)	$\phi = \frac{\lvert\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2})\rvert}{\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}}} = z_{\frac{\alpha}{2}} - z_{1-\beta}$	$\phi = \frac{\lvert\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2})\rvert}{\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}}} = z_{\alpha} - z_{1-\beta}$
Nombre de sujets nécessaire $n_1 = n_2 = n$	$n = \frac{(z_{\alpha/2} - z_{1-\beta})^2}{2(\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2}))^2}$	$n = \frac{(z_{\alpha} - z_{1-\beta})^2}{2(\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2}))^2}$
Nombre de sujets nécessaire $n_2 = k \times n_1$	$n_1 = \frac{k+1}{k} \frac{(z_{\alpha/2} - z_{1-\beta})^2}{4(\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2}))^2}$ $n_2 = k \times n_1$	$n_1 = \frac{k+1}{k} \frac{(z_{\alpha} - z_{1-\beta})^2}{4(\text{Arcsin}(\sqrt{p_1}) - \text{Arcsin}(\sqrt{p_2}))^2}$ $n_2 = k \times n_1$

Exemple

- Comparaison des taux d'échec de grossesse selon la consommation ou non de café
- Données a priori : pourcentage d'échec de l'ordre de 10% sans consommation de café
- Objectif : calculer la puissance pour détecter une augmentation de 10% chez les consommatrices avec un échantillon de 100 femmes consommatrices et 100 non consommatrices
- $H_0 : p_C = p_{NC}$ et $H_1 : p_C \neq p_{NC}$ avec $\alpha = 5\%$
- $p_{NC} = 0.1$ ➡ $\text{Arcsin}\left(\sqrt{p_{NC}}\right) = 0.322$ (Table 5)
- $p_C = 0.2$ ➡ $\text{Arcsin}\left(\sqrt{p_C}\right) = 0.464$ (Table 5)
- $$\phi = \frac{\left|\text{Arcsin}\left(\sqrt{p_C}\right) - \text{Arcsin}\left(\sqrt{p_{NC}}\right)\right|}{\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}}} = \frac{0.464 - 0.322}{\sqrt{\frac{1}{4\times100} + \frac{1}{4\times100}}} = 2.008$$

Table 4a ➡ $1 - \beta = 52\%$

Exemple

- Comparaison des taux d'échec de grossesse selon la consommation ou non de café
- Données a priori : pourcentage d'échec de l'ordre de 10% sans consommation de café
- Objectif : calculer le nombre de sujets nécessaire pour détecter une fréquence d'échec de 20% avec la consommation de café et une puissance de 80% $(z_{1-\beta} = 0.842)$
- $H_0 : p_C = p_{NC}$ et $H_1 : p_C \neq p_{NC}$ avec $\alpha = 5\%$
- $p_{NC} = 0.1$ ➡ $\text{Arcsin}\left(\sqrt{p_{NC}}\right) = 0.322$ (Table 5)
- $p_C = 0.2$ ➡ $\text{Arcsin}\left(\sqrt{p_C}\right) = 0.464$ (Table 5)
- $$n = \frac{\left(z_{\alpha/2} - z_{1-\beta}\right)^2}{2\left(\text{Arcsin}\left(\sqrt{p_C}\right) - \text{Arcsin}\left(\sqrt{p_{NC}}\right)\right)^2} = \frac{\left(1.96 - (-0.842)\right)^2}{2 \times (0.464 - 0.322)^2} = 195$$
- Il faut sélectionner 195 femmes par groupe

Comparaison d'un OR ou d'un RR à la valeur 1

- On se ramène à la comparaison de deux pourcentages ($\Leftrightarrow OR = 1$ ou $RR = 1$)

- Dans une enquête transversale ou une enquête de cohorte

  - n_{E+} nombre de patients exposés et n_{E-} nombre de non exposés

$$P_0 = P(M+|E-)$$

$$P_1 = P(M+|E+) = RR \times P_0$$

$$P_1 = P(M+|E+) = \frac{OR \times P_0}{1+(OR-1)P_0}$$

$$\left.\begin{array}{l} H_0 : RR = 1 \\ H_0 : OR = 1 \end{array}\right\} \longrightarrow z_{1-\beta} = z_{\frac{\alpha}{2}} - \frac{\left|\mathrm{Arcsin}(\sqrt{P_1}) - \mathrm{Arcsin}(\sqrt{P_0})\right|}{\sqrt{\frac{1}{4n_{E+}} + \frac{1}{4n_{E-}}}}$$

Comparaison d'un OR ou d'un RR à la valeur 1

- Dans une enquête cas-témoins
  - n_{M+} nombre de patients malades et n_{M-} nombre de non malades
  - La fréquence de la maladie n'est pas estimable (formule de l'OR avec les fréquence d'exposition)
  - $P_{E_0} = P(E+|M-)$
  - $P_{E_1} = P(E+|M+) = \frac{OR \times P_{E_0}}{1+(OR-1)P_{E_0}}$

$$H_0 : OR = 1 \longrightarrow z_{1-\beta} = z_{\frac{\alpha}{2}} - \frac{\left|\text{Arcsin}(\sqrt{P_{E_1}}) - \text{Arcsin}(\sqrt{P_{E_0}})\right|}{\sqrt{\frac{1}{4n_{M+}} + \frac{1}{4n_{M-}}}}$$

- Tables du livre "rose"
  - Table 4a : Puissance $1-\beta$ en fonction des valeurs de ϕ pour $\alpha = 0.05$, test bilatéral
  - Table 4b : Puissance $1-\beta$ en fonction des valeurs de ϕ pour $\alpha = 0.05$, test unilatéral
  - Table 6a : Nombre de sujets pour une puissance de 80% en fonction de l'OR et de P_0 (ou de P_{E_0} dans une enquête cas-témoins) pour $\alpha = 0.05$
  - Table 6b : Valeur de l'OR garantissant une puissance de 80% en fonction de P_0 et du nombre d'exposés et de non exposés (ou de P_{E_0} et du nombre de témoins et de cas dans une enquête cas-témoins) pour $\alpha = 0.05$ et $n_1 = n_2$

Exemple

- Evaluer l'effet de l'exposition à des solvants chez des patients atteints d'un cancer et des témoins
- Données a priori : pourcentage de l'exposition autour de 20% en population générale (témoins)
- Objectif : calculer la puissance pour mettre en évidence un $OR = 2$ avec un échantillon de 100 cas et de 100 témoins
- $H_0 : OR = 1$ et $H_1 : OR \neq 1$ avec $\alpha = 5\%$
- $p_{E0} = 0.2 \rightarrow \text{Arcsin}\left(\sqrt{0.2}\right) = 0.464$
- $p_{E1} = \frac{OR \times P_{E_0}}{1+(OR-1)P_{E_0}} = \frac{2\times 0.2}{1+(2-1)\times 0.2} = 0.33 \rightarrow \text{Arcsin}\left(\sqrt{0.33}\right) = 0.612$
- $\phi = \frac{\left|\text{Arcsin}\left(\sqrt{p_{E1}}\right)-\text{Arcsin}\left(\sqrt{p_{E0}}\right)\right|}{\sqrt{\frac{1}{4n_{M+}}+\frac{1}{4n_{M-}}}} = \frac{0.612-0.464}{\sqrt{\frac{1}{4\times 100}+\frac{1}{4\times 100}}} = 2.093$
- Table 4a $\rightarrow$ $1 - \beta = 55\%$

Table 6a : Il faut 170 témoins et 170 cas pour mettre en évidence un $OR = 2$ avec une puissance de 80%

Table 6b : Avec 100 témoins et 100 cas et une puissance de 80% on peut mettre en évidence un $OR > 2.43$

Comparaison d'un SMR à la valeur 1

- Soit M le nombre de cas observés et E le nombre de cas attendus
- Test unilatéral : $H_0 : SMR = 1$

  $H_1 : SMR > 1$
- Sous H_1, $Z = 2\left(\sqrt{M} - \sqrt{E}\right) \underset{H_1}{\rightarrow} N\left(2\sqrt{E} \times \left(\sqrt{SMR} - 1\right), 1\right)$
- $z_{1-\beta} = z_{\alpha} - \phi$ avec $\phi = 2\sqrt{E} \times \left(\left|\sqrt{SMR} - 1\right|\right)$
- Tables du livre "rose"
  - Table 4a : Puissance $1 - \beta$ en fonction des valeurs de ϕ pour $\alpha = 0.05$, test bilatéral
  - Table 4b : Puissance $1 - \beta$ en fonction des valeurs de ϕ pour $\alpha = 0.05$, test unilatéral
  - Table 7a : Valeur de la puissance selon E et la valeur du SMR pour $\alpha = 0.05$, test unilatéral
  - Table 7b : Valeur du SMR qu'on peut mettre en évidence en fonction de la puissance et de E, $\alpha = 0.05$, test unilatéral

Exemple

- Comparaison du nombre de cas de cancers dans une cohorte de 400 hommes d'une entreprise du Bas-Rhin au nombre attendu si le taux d'incidence est égal à celui de la population du département
- Nombre de cas attendu $E = 20.71$ (Nombre Personnes/année = 2216)
- Objectif : calculer la puissance pour mettre en évidence un SMR de 2
- $H_0 : SMR = 1$et $H_1 : SMR > 1$ avec $\alpha = 5\%$
- $$\phi = 2\sqrt{E} \times \left(\left|\sqrt{SMR} - 1\right|\right) = 2\sqrt{20.71} \times \left(\left|\sqrt{2} - 1\right|\right) = 3.770$$
- Table 4b ➡ $1 - \beta = 98\%$
- Table 7b ➡ Pour $E = 20$ et $1 - \beta = 80\%$ on peut mettre en évidence un $SMR = 1.67$

IV. Puissance dans une enquête

- La puissance d'une enquête dépend de
  - La différence des valeurs comparées (puissance augmente quand la différence augmente)
  - De la variabilité de la variable étudiée (puissance augmente quand la variance diminue)
    - Cas des moyennes : σ^2 petite
    - Cas des pourcentages : p_1 et p_2 le plus écartés possibles et éloignés de 0.5
  - Du nombre de sujets inclus (puissance augmente quand la taille des échantillons augmente)
  - De l'erreur de 1ère espèce α (puissance augmente quand α augmente)
    - En général α est fixé à 5%

Optimiser la puissance au niveau du protocole

- Choix des populations comparées

  - Ecart entre les populations le plus grand possible (différence importante)

    Ex : Non fumeur et gros fumeur

  - Choisir des populations sensibles pour mieux observer les effets (différence importante)

    Ex : personnes âgées, jeunes, femmes enceintes

  - Choisir des populations homogènes (pour avoir une petite variance)

    Ex : Comparer le personnels d'un atelier (plutôt que le personnel d'une usine) à des témoins

    L'exposition est moins homogène dans une usine que dans un atelier

  - Attention : – ne pas générer un biais de sélection

    – ne répond pas toujours à l'objectif initial

Optimiser la puissance au niveau du protocole

- Choix des paramètres de santé et d'exposition
  - Choisir des paramètres les plus spécifiques possibles

    Ex : Effet des champs électromagnétiques → leucémies plutôt que tous les cancers

    Ex : Etude d'un cancer spécifique à l'amiante → amiante plutôt que toutes les poussières

  - Limiter les erreurs et imprécisions de mesure, définitions précises et standardisées

    ⟶ Limiter les biais de classement non différentiels (perte de puissance)

- Choix du mode d'échantillonnage
  - Puissance meilleure dans enquête cas-témoins ou exposés/non exposés que dans une enquête transversale
  - A nombre de sujets égal, la puissance est meilleure quand les effectifs des groupes comparés sont équilibrés
  - Appariement et stratification peuvent apporter un gain de puissance lors des tests de comparaison
  - Choisir des tailles d'échantillons le plus importantes possible
  - Anticiper et estimer la proportion de non participation, de non réponse, de perdus de vue

Interprétation des résultats

- Résultats significatifs : rejet de H_0 ⟶ On peut faire l'erreur de 1ère espèce α
- Résultats non significatifs: non rejet de H_0 ⟶ On peut faire l'erreur de 2ème espèce β
  - Calculer la puissance a posteriori (à partir des données réellement disponibles)
  - Soit il n'y a pas de différence (H_0 vraie, il n'y a pas de différence)
    - Décision possible si la puissance a priori et à posteriori est $\geq 80\%$
  - Soit il y a un manque de puissance (H_1 vraie mais on ne voit pas la différence)
    - Décision possible si la puissance a priori et à posteriori est $< 80\%$
    - Calcul de la différence minimale détectable
      - Ex : On trouve que la différence détectable pour une puissance de 70% est $\Delta = 5$

        On peut conclure que "la vraie différence est vraisemblablement inférieure à 5" avec un risque d'erreur de 30% (plutôt que "on n'a pas mis en évidence de différence")

Interprétation des résultats

- Remarque importante
  - Quand la différence n'est pas significative, on ne conclut pas "qu'on accepte H_0"
  - En effet, le risque d'erreur encouru est inconnu car il dépend de " la valeur sous H_1" qui est inconnue
  - " Image du microscope " :
    - Si on voit une différence c'est qu'il y en a une
    - Si on ne voit rien, c'est peut être que la différence est trop petite et qu'on ne l'a pas vue