Psy1004 – Section 13:

Puissance d'un test

Plan du cours:

- Varia
- 13.1: Efficacité d'un test
- 13.2: Puissance d'un test
- 13.3: Comment rendre un test plus puissant
- 13.4: Méthode de Cohen
- 13.5: Exemples
- 13.6: Synthèse des différents tests statistiques
- Période de questions

Disponible sur: http://mapageweb.umontreal.ca/cousined/home/course/PSY1004

Varia

- L'examen a lieu au prochain cours.
- Pour l'examen:
  - L'examen débute à 13h00 et se termine à 16h00.
  - Une feuille 8 ½ x 5 ½ avec des notes personnelles est permise.
  - Ainsi que la calculatrice.

Varia

13.1: Efficacité d'un test (1/4)

- Quel est le but d'un test statistique?
  - Rendre une décision à savoir si une hypothèse semble vraie ou semble fausse:

		Dans la réalité, H_0 est:	
basé sur un échantillon:		VRAI	FAUX
Décision:	non-rejet H_0	**aucun effet sig.**	erreur β
	rejet de H_0	erreur α	**effet significatif**

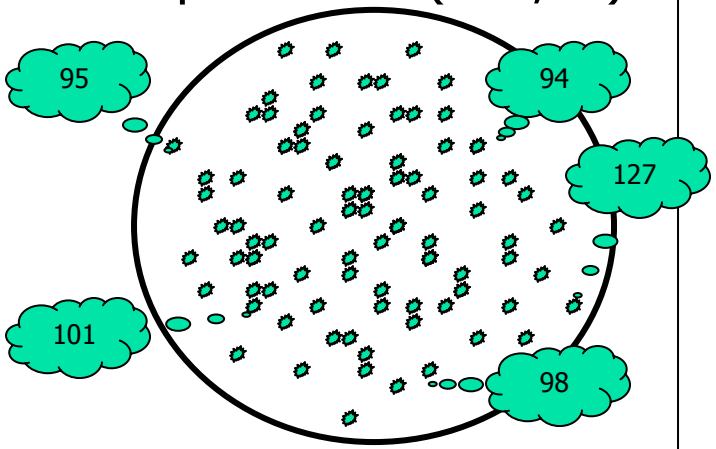
Sera publié!

- Quelle est la procédure de décision?
  - Mesurer une statistique de la population
  - Connaître comment varie cette statistique (les postulats) dans le but de fixer un seuil de décision
  - Rejet de H_0 si une formule impliquant la statistique est supérieure au seuil de décision:
    - on confronte la réalité avec nos attentes;
    - La dernière étape fait le lien entre l'empirique et le théorique.

13.1: Efficacité d'un test (2/4)

Par exemple, pour un test sur une moyenne:

Population $\mathcal{N}(100,10)$



La population contient 1 million de nombres, on va en échantillonner 25.

Question: "Est-ce que la moyenne est de 100?"

Procedure:

a) H_0: $\mu = 100$

b) $\alpha = .05$

c) test t

d) échantillonner et rejet de H_0 si $\frac{|\overline{\mathbf{X}} - \mu|}{SE_{\overline{\mathbf{X}}}} > s(\alpha/2)$

Tests:

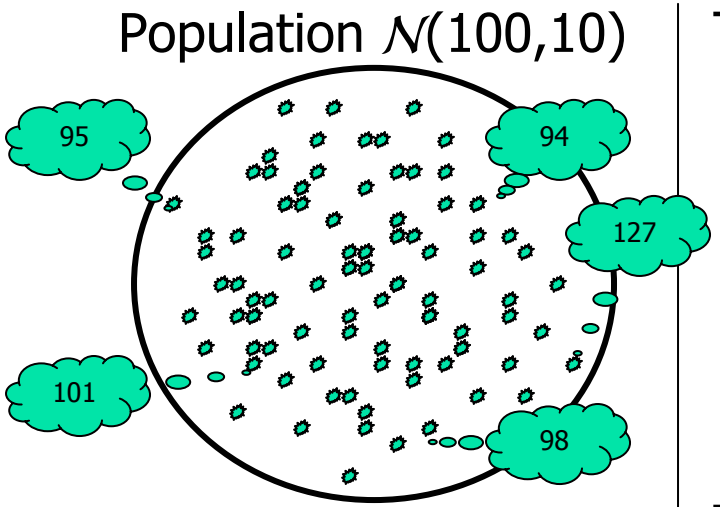
1. $\overline{\mathbf{X}} = 103.948$, calcul=1.89, non rejet
2. $\overline{\mathbf{X}} = 98.9632$, calcul=0.68, non rejet

...

8. $\overline{\mathbf{X}} = 95.7251$, calcul=2.22, REJET!

ERREUR!
erreur de type I
ou erreur alpha

13.1: Efficacité d'un test (3/4)



La population contient 1 million de nombres, on va en échantillonner 25.

Tests:

1. $\bar{x}$ = 103.948, calcul=1.89, non rejet
2. $\bar{x}$ = 98.9632, calcul=0.68, non rejet

...

8. $\bar{x}$ = 95.7251, calcul=2.22, REJET!

...

10000. $\bar{x}$= 101.91, calcul=1.05, non rejet

Total:

sur 10 000 essais, j'ai obtenu 515 rejets (soit 515 décisions erronées), soit un pourcentage de 5.15 %

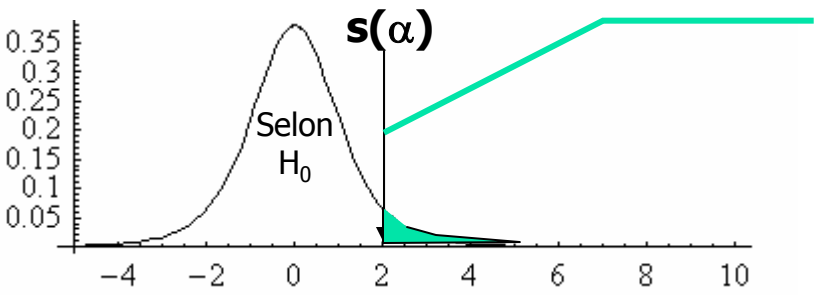
Ce pourcentage est très proche du seuil α.

Est-ce un hasard?

13.1: Efficacité d'un test (4/4)

Non.

- Si l'hypothèse est vraie, elle me dit exactement tout ce dont j'ai besoin pour savoir la probabilité d'obtenir une moyenne extrême.
  - Selon le théorème central limite, une moyenne se distribue comme une distribution *t* de Student, ce qui permet de fixer un seuil de décision.



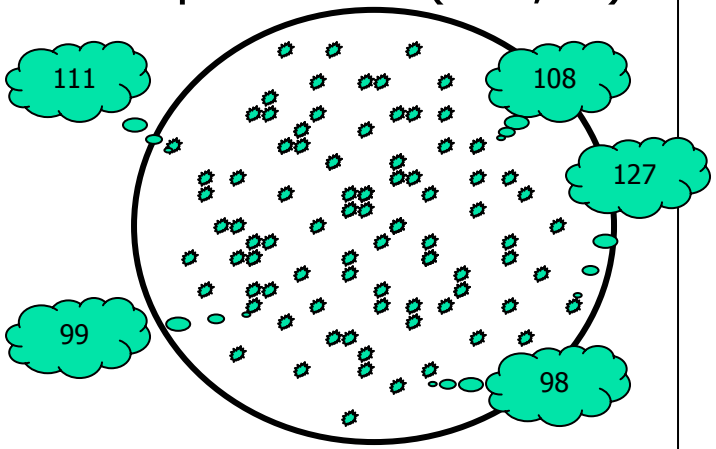
En principe, notre calcul doit donner une valeur proche de zéro.
La probabilité d'elle soit plus grande qu'une valeur x est calculable puisqu'on connaît sa distribution.
Dans notre cas, on veut que la probabilité qu'elle soit plus grande que s(5%) soit de 5%. Donc, s(5%) = 2.015 (si unicaudal).

- Le seuil α est la probabilité de faire une erreur alpha.
- La probabilité de faire l'erreur a résulte d'un choix éclairé.

13.2: Puissance d'un test (1/4)

Quand est-il si H_0 n'est pas vrai?

Population $\mathcal{N}(110,10)$



La population contient 1 million de nombres, on va en échantillonner 25.

Question: "Est-ce que la moyenne est de 100?"

Procedure:

a) H_0: $\mu = 100$

b) $\alpha = .05$

c) test t

d) échantillonner et rejet de H_0 si $\frac{|\overline{\mathbf{X}} - \mu|}{SE_{\overline{\mathbf{X}}}} > s(\alpha/2)$

Tests:

1. $\overline{\mathbf{X}} = 117.598$, calcul=9.21, rejet
2. $\overline{\mathbf{X}} = 115.652$, calcul=7.65, rejet

...

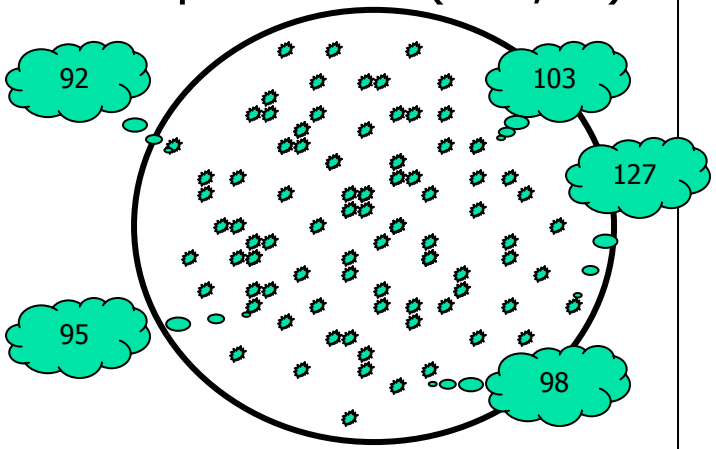
854. $\overline{\mathbf{X}} = 104.442$, calcul=2.04, NON REJET!

ERREUR!
erreur de type II
ou erreur beta

13.2: Puissance d'un test (2/4)

Quand est-il si H_0 n'est pas vrai, mais de très peu?

Population $\mathcal{N}(105,10)$



La population contient 1 million de nombres, on va en échantillonner 25.

erreur beta très commune: souvent, je conserve H_0 de façon erronée.

Question: "Est-ce que la moyenne est de 100?"

Procedure:

a) H_0: $\mu = 100$
b) $\alpha = .05$
c) test *t*
d) échantillonner et rejet de H_0 si $\frac{|\overline{\mathbf{X}} - \mu|}{SE_{\overline{\mathbf{X}}}} > s(\alpha/2)$

Tests:

1. $\overline{\mathbf{X}} = 107.395$, calcul=3.62, rejet
2. $\overline{\mathbf{X}} = 105.768$, calcul=2.64, rejet
3. $\overline{\mathbf{X}} = 102.415$, calcul=1.16, NON REJET!

...

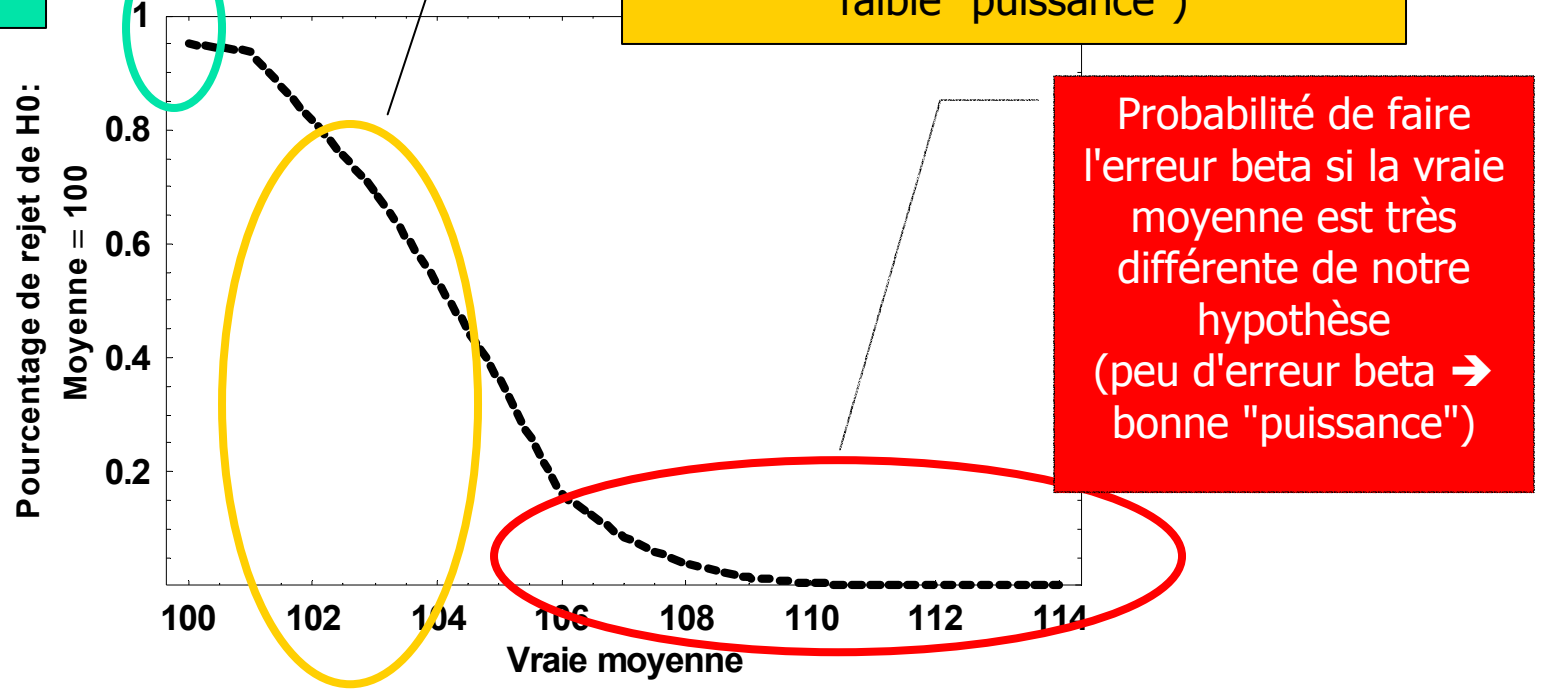
13.2: Puissance d'un test (3/4)

- Les erreurs beta peuvent être rares ou très fréquentes, et on a peu de contrôle, car si notre hypothèse est fausse, on ne connaît pas la vraie réalité.

Probabilité de faire l'erreur alpha

Probabilité de faire l'erreur beta si la vraie moyenne est proche de notre hypothèse (beaucoup d'erreur beta ➔ faible "puissance")

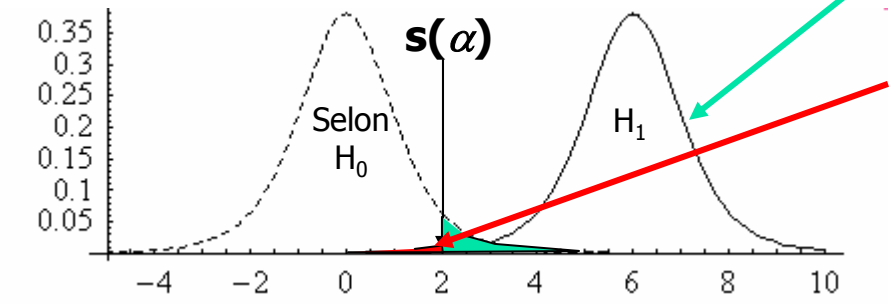
Probabilité de faire l'erreur beta si la vraie moyenne est très différente de notre hypothèse (peu d'erreur beta ➔ bonne "puissance")



13.2: Puissance d'un test (4/4)

Autrement dit:

- Si H_1, la vérité, est vraiment différente de H_0:

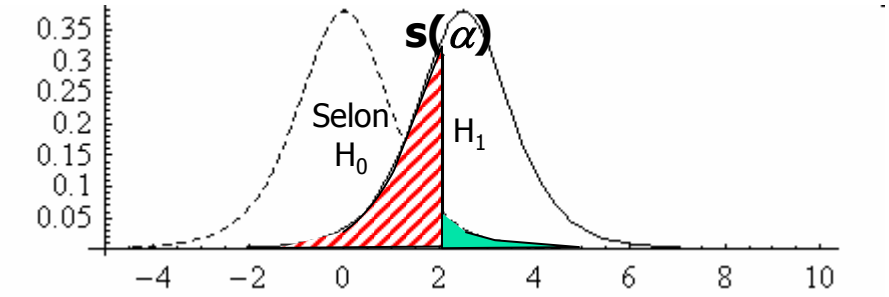


Vraie distribution de notre statistique élaborée, à notre insu...

La probabilité de faire une erreur β est la probabilité que notre moyenne soit très basse étant donnée qu'il y a un effet.

La probabilité de faire une erreur β est très faible (ici, de l'ordre de 1 ‰)

- Si H_1, la vérité, diffère seulement <u>légèrement</u> de H_0:



La probabilité de faire une erreur β est très grande (ici, de l'ordre de 30%), à peu prés une fois sur trois, je vais dire qu'il n'y a pas d'effet alors qu'il y a un effet.

En résumé

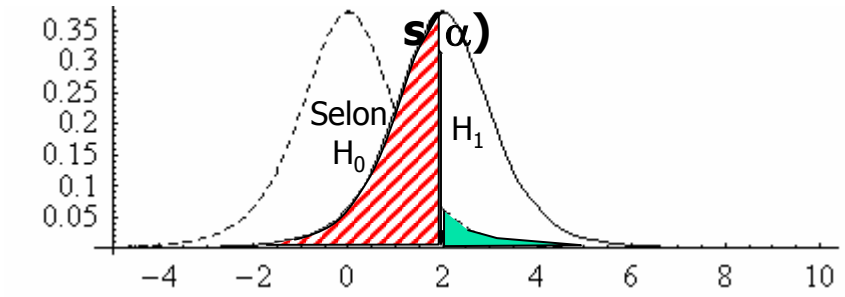
- L'erreur β est la probabilité de dire que H_0 est vraie (non rejet de H_0) alors qu'elle ne l'est pas.
  - L'erreur β est difficile à évaluer car on ne sait pas, quand effet il y a, quelle est son ampleur (faible, modéré, fort).
  - On appelle la puissance d'un test la probabilité de ne pas faire l'erreur β, c'est à dire, la probabilité de dire que H_1 est vraie (rejet de H_0) alors qu'elle est belle et bien vraie.
  - Autrement dit, $1 - \beta$ = puissance; si la probabilité de faire l'erreur β est de 20%, la puissance du test est de 80%.
- Quelle devrait être la puissance idéale d'un test? 99%, 95%, 90%
  i.e. la probabilité de l'erreur β 1%, 5%, 10%
  - On recommande en générale d'avoir une puissance de 80% (évidemment, 90%, 95% et même 99% sont mieux, mais…).
  - L'erreur α doit être contrôlé plus sérieusement que l'erreur β car dès qu'un effet est trouvé, il est publié dans les revues scientifiques.
  - L'erreur β peut être plus tolerée car si le chercheur a la conviction d'avoir raison, il peut toujours refaire l'expérience.

Anecdote

- Un chercheur (Cohen) a estimé la puissance de tous les tests statistiques publiés dans l'année 1962, puis aussi en 1988.
  - en supposant que tous les effets, s'il y en a, sont modéré,
  - en regardant le nombre de sujets utilisé *n*, il obtient:

➔que la puissance, en moyenne, était de 50%!

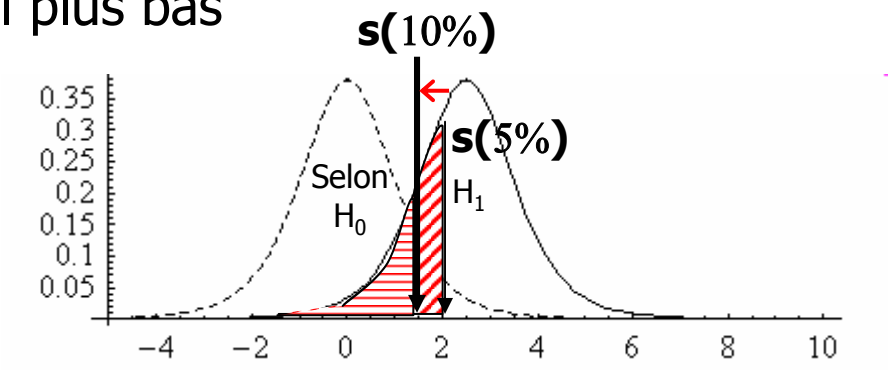
  - c.à.d. que lorsqu'il y a bel et bien un effet de la variable étudiée, il existe une chance sur deux que le chercheur décide que l'effet n'est pas significatif (sur la base du test statistique).



13.3: Comment rendre un test plus puissant? (1/3)

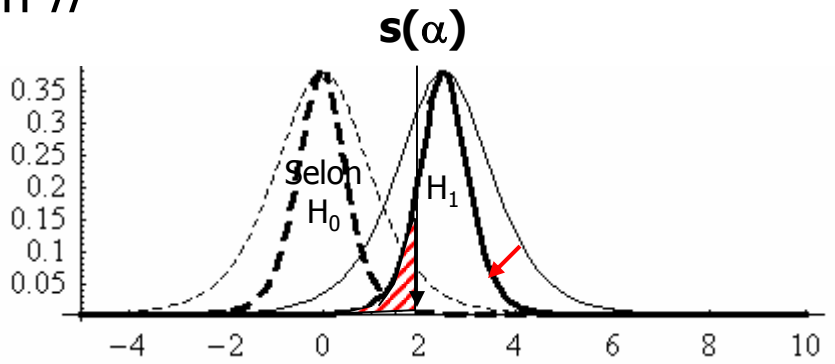
a) Changer le seuil α pour un seuil plus bas

En rapprochant α de H_0, je réduit le risque d'erreur β, mais! le test sera moins sévère.



b) Accroître le nombre d'observation *n*

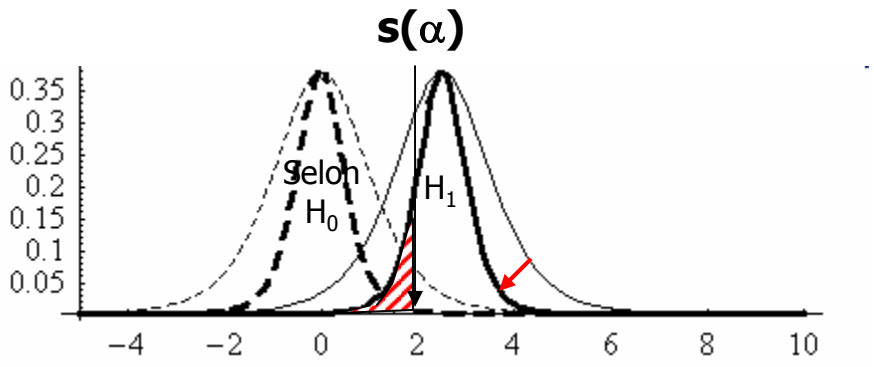
En utilisant un *n* plus grand, la distribution de l'hypothèse H_0 et de l'hypothèse H_1 devient moins variable, et donc, il y a moins de chevauchement entre les deux, mais! est plus coûteux.



13.3: Comment rendre un test plus puissant? (2/3)

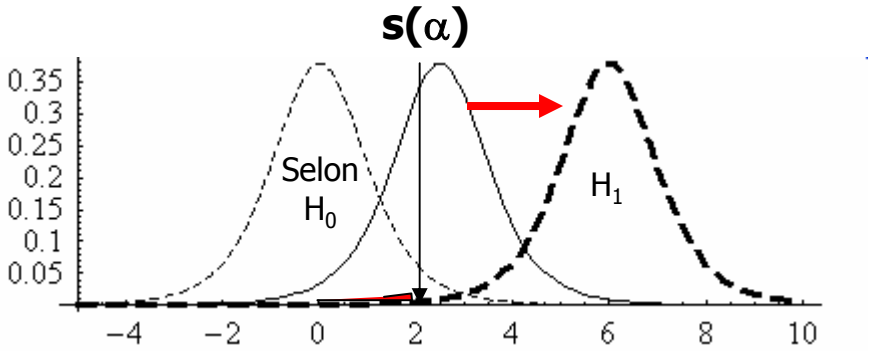
c) Réduire l'erreur expérimentale

En contrôlant tous les facteurs qui peuvent nuire à la concentration des sujets, en prenant le plus possible des sujets identiques (âge, heure, etc.), ou en entraînant les sujets pour qu'ils performent avec moins de variabilité.



d) Accroître la taille de l'effet

En étudiant des effets importants, (par exemple, en changeant la difficulté, la dose, etc.), on accroît la distance entre les distributions pour H_0 et H_1, donc moins de chevauchement, mais! parfois difficile (par exemple, dans l'étude sur le stress, la performance grimpe, puis redescend; il n'y a pas d'effet plus fort qu'au niveau 3).



13.3: Comment rendre un test plus puissant? (3/3)

- Quelle méthode préférer?
  - Accroître la taille de l'effet est la meilleure solution, mais pas toujours possible
  - Accroître le nombre d'observation est toujours une bonne solution, et presque toujours possible (mais coûteux)
  - Réduire l'erreur expérimental est souhaitable, mais il est difficile de prévoir toutes les causes possibles de variance entre les sujets…
  - Réduire le seuil α est non recommandé, car on troque une source possible d'erreur pour une autre…

13.4: Méthode de Cohen

Si vous choisissez d'accroître le nombre d'observation, Cohen a développer une méthode simple pour trouver le nombre de sujets tel que la puissance soit de .80.

- Si vous connaissez bien le domaine de recherche:
  - a) faites un estimé de ce que serait l'effet s'il existait;
  - b) regardez dans la Table 7a s'il s'agit d'un effet considéré comme "petit", "moyen", ou "grand".
- Si vous ne connaissez pas le domaine de recherche:
  - supposez qu'il s'agirait d'un effet "moyen", comme très souvent en psychologie.

- Selon votre seuil α et la taille de l'effet ("petit", "moyen", "grand"), lire dans la Table 7b le nombre de sujets requis.

13.5: Exemples (1/3)

- Soit une étude sur le Q.I. lors d'une période de deuil. On veut savoir si le Q.I. baisse.
  - H_0: Q.I. moyen = 100
  - H_1: Q.I. moyen < 100
    (selon des études, la maladie et le stress baisse le Q.I. de 5, donc, vous espéré un Q.I. d'à peu près 95)
  - En regardant la table 7a, première ligne, la taille de l'effet d serait d'environ 0.50 (car l'écart type d'un test de Q.I. est toujours de 10), soit un effet moyen.
  - En regardant la table 7b, avec α = 5% et effet moyen, il faudrait étudier 64 personnes en deuil.

13.5: Exemples (2/3)

- Il semble que le nombre de semaine depuis lequel la personne est en deuil ait une influence. Vous refaites donc l'expérience avec 4 durée de deuil différentes: depuis 1 semaine, depuis 2 semaines, depuis 3 semaines, et depuis 4 semaines.
  - H_0: $\mu_{1sem}=\mu_{2sem}=\mu_{3sem}=\mu_{4sem}$.
  - H_1: il existe au moins une paire où $\mu_{i\,sem} \neq \mu_{j\,sem}$
    (vous vous attendez, si effet il y a, à ce genre de résultats: 94, 96, 98, et 100 de Q.I. pour les niveaux dans l'ordre).

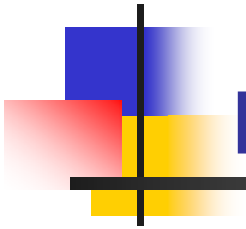
  - Pour utiliser la table 7a avec les ANOVA, il faut calculer
    - a) quelles serait l'erreur expérimentale, $\sigma_{S|A}$, facile ici puisque l'écart type lors d'un test de Q.I. est toujours de 10.
    - b) quelles serait l'écart type des prédictions (94, 96, 98, 100). Avec une calculatrice, on obtient: 2.58.
    - La valeur *f* serait donc: 2.58 / 10 = 0.258, soit un effet "large".
  - Dans la table 7b avec une ANOVA à 4 groupes, on trouve 18 sujets par groupe (soit 72 personnes en deuil, peut être difficile à trouver…).

13.5: Exemples (3/3)

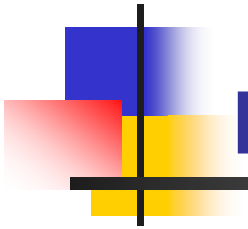
- Vous voulez plutôt connaître de combien le Q.I. baisse pour chaque semaine de convalescence après le deuil. Pour ce faire, vous faîtes une étude de corrélation et extrayez la pente de la droite de régression.
  - H_0: $r_{deuil,\ QI} = 0$
  - H_1: $r_{deuil,\ QI} \neq 0$
    (s'il y avait un effet, l'effet –la corrélation– serait autour de .20, selon ce que vous savez du domaine).
  - Dans la Table 7a, on voit qu'un *r* de .20 est considéré "moyen"
  - Dans la Table 7b, pour un *r* moyen, avec $\alpha = .05$, on trouve qu'il faut étudier environ 85 personnes au total pour établir si le *r* est significatif ou pas avec une puissance de .80.

13.6: Synthèse des différents tests statistiques

- Voir Tableau de synthèse PSY1004_1a13 Détails



Période de questions



Merci.