Biostatistiques

Principe d'un test statistique

Professeur Jean-Luc BOSSON

Objectifs pédagogiques

- Comprendre le principe général d'un test d'hypothèse
- Comprendre les règles de bon usage des tests statistiques
- Connaître les notions de
  - Risque de première espèce
  - Risque de deuxième espèce
    - Puissance d'un test
  - Calcul du nombre de sujets

Objectifs des statistiques

- Estimer les paramètres d'une population inconnue
  - De tendance centrale
  - De dispersion
- Comparer des paramètres observés sur plusieurs échantillons (populations)
  - Échantillon vs population
  - Plusieurs échantillons
- Etablir des modèles prédictifs

Généralités sur les statistiques

- 2 possibilités pour décrire / comparer des populations:
  - Tester des populations entières mais c'est exceptionnel
  - Tester des échantillons et extrapoler à leurs populations source
    - une statistique = paramètres de l 'échantillon (moyenne, écart-type, …)
    - inférence statistique = porter une conclusion sur l'ensemble de la population source

Les tests statistiques sur échantillons n'ont d'intérêt que rapportés à leur population source

- Les tests statistiques paramétriques sont liés à des **lois de probabilités**
  - associer une probabilité de survenue à tout événement

Les tests statistiques en pratique

- Quelques questions:
  - dans une usine de produits chimiques, le Volume Globulaire Moyen de 30 ouvriers a été testé (92,5µm^3) et comparé à celui de 30 employés de bureau (94,7µm^3)
  - Avec un traitement A on observe 77% de guérison. Avec un traitement B 68 %

  $\Rightarrow$ Ces différences sont-elles réelles ou dues au hasard ?
- Les tests statistiques substituent à une solution empirique un risque d 'erreur
  - alpha est le risque qu'on accepte de prendre de dire que la différence n'est pas due au hasard alors qu'elle est due au hasard
  - choix, a priori d 'un risque « raisonnable » alpha = 5%

Les statistiques en pratique

- Autre mode de résolution pour savoir si deux paramètres différent réellement : l'approche par intervalles de confiance (IC)
  - Calcul de précision d 'une statistique
  - « si je multipliais les expériences donc les échantillons, 95 fois sur 100 le paramètre mesuré serait compris entre X et entre Y
  - C'est l'intervalle de confiance à 95% (IC95%) du paramètre mesuré (moyenne, pourcentage, risque relatif, odds ratio….)

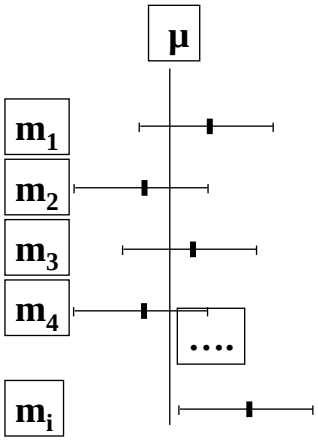
Les statistiques en pratique

- Pour savoir si deux paramètres différent réellement on peut utiliser l'approche par intervalles de confiance. Exemple :
  - Proportion de guérison avec le traitement A
    - 40% Intervalle de confiance à 95 % IC95% = [30;50]
  - Proportion de guérison avec le TTT B
    - 15% IC95% [6;21]
  - Dans le meilleur des cas B guéris 21 % des patients
    - estimation la plus haute de l'intervalle de confiance
  - Dans le pire des cas A guéris 30 % des patients
    - estimation la plus basse de l'intervalle de confiance
- A semble meilleur que B
  - risque alpha = 5% puisque les IC sont estimés à 95 %

Rappel IC variable quantitative

- Pour la moyenne on peut donner la précision de l'estimation de la moyenne par l'intervalle de confiance de la moyenne
  - moyenne $\pm$ 1,96 *((écarts type) / Racine(n)) n étant le nombre de cas
- C'est l'intervalle de confiance à 95 % de la moyenne
  - si l'on retirait au sort 100 fois un échantillon de même taille 95 moyennes estimées sur 100 seraient comprises dans cet intervalle
  - Ex : la moyenne des patients suspects d'embolie pulmonaire peut être estimée à 63 ans $\pm$ 1 an
  - Ce n'est pas la dispersion des valeurs individuelles mais la précision de la mesure

Estimation IC 95%



Attention !
μ reste constant
C'est l'intervalle de confiance qui varie autour de μ pour chaque échantillon.

Si l'on a choisi un risque de 5%, en moyenne, l'estimation obtenue dans 1 échantillon sur 20 ne contiendra pas la vraie valeur μ.

Intervalle de confiance d'une proportion

- Un médecin observe 20 cas de guérison parmi 50 cas de cancer.
- Quel est l'intervalle de confiance qui a 95 chances sur 100 de contenir la vraie valeur du pourcentage de guérison ?

$$\pi : [\ 0.40 \pm 1.96 \times \sqrt{\frac{0.40 \times 0.60}{50}}\ ]$$

$$\pi : [\ 0.40 \pm 1.96 \times 0.07\ ]$$

où *1.96* est la valeur de la variable N(0,1) correspondant au risque choisi : 5 %

Estimation ponctuelle sur 50 cas : **40 % (pourcentage observé)**
Intervalle de confiance : **26 % à 54 %**

Résumé IMPORTANT intervalle de confiance

- Intervalle de confiance
  - Construit autour de la moyenne observée sur l' échantillon
  - Construit en utilisant la variance observée
  - Définit l'intervalle « raisonnable » dans lequel la moyenne vraie (théorique) peut se situer au risque 1-α
  - Si 2 IC95% sont disjoints alors le test statistique correspond est significatif au seuil 5%

Principe d'un test statistique

- Pour comparer deux paramètres (deux moyennes par exemple) on va se ramener à une valeur qui suit une loi de distribution connue
  - Ex : N (0,1) pour la différence de deux moyennes
- On va ensuite regarder sur une table de cette loi de distribution, si la valeur observée est une valeur
  - « étonnante » ie peu probable pour cette loi de distribution
  - « banale »

Objectif d'un test statistique

- **Un test permet de porter une affirmation en contrôlant le risque d'erreur**
- Il donne une réponse à la question :
  - La différence observée entre mes deux paramètres peut-elle être due aux fluctuations d'échantillonnage ?
  - Les deux paramètres sont-ils deux estimations d'une même population théorique ?

Etapes d'un test statistique

- **Poser les hypothèses**
  - Hypothèse nulle H0 : les moyennes sont égales
  - Hypothèse alternative H1 : les moyennes sont différentes
- Chercher à rejetter l'hypothèse nulle pour accepter avec un risque d'erreur connu l'hypothèse alternative
  - Si H0 est vrai alors ma statistique de test suit une loi de distribution connue
- Définir le risque d'erreur acceptable c'est le risque alpha de première espèce
  - C'est le risque de rejetter H0 alors que H0 est vrai (conclure à tort à une différence)
- Calculer la p-valeur observée (table de la loi normale par exemple)
- Comparer la p-valeur observée au risque alpha choisi (0,05)
- Si p-valeur observée est inférieur au risque alpha on conclu à l'existence d'une différence statistiquement significative

la p value observée

- C'est un **calcul précis** ex : p = 0,0275 que l'on va comparer au seuil usuel de 0,05
- Cela ne **reflète pas directement l'importance de l'effet clinique**
  - Delta important cliniquement mais peu de sujets p>>0,05 et pourtant …
  - Delta sans signification clinique mais beaucoup de sujets p<<0,001 et pourtant …
- C'est un calcul **a posteriori** Ex : jeu de cartes et pari
- C'est une **variable aléatoire**

la p value observée

- C'est une **variable fragile**
  - choix du test
  - conditions d'applications
  - tests multiples, on ne contrôle plus le risque d'erreur
    - ex : taille, poids et périmètre crânien des nouveaux nés
    - On a en fait une probabilité de : **$1 – (1- \alpha^n)$** d'avoir au moins un test significatif n = nb de tests
    - au seuil $\alpha = 0,05$ soit ici 15%
    - Ref : DATA TORTURING NEJM

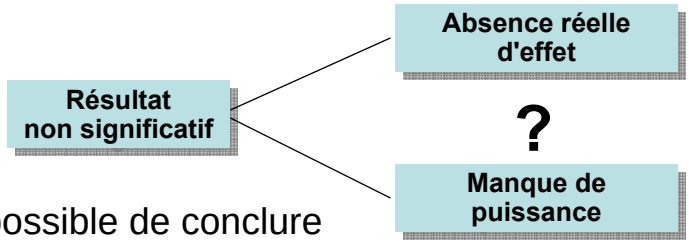
Interpréter un test significatif

- Si la p-valeur calculée est inférieure au seuil alpha choisi alors
  - On rejette H0 hypothèse d'égalité des paramètres
  - On accepte H1 donc on conclu à l'existence d'une différence significative avec un risque de se tromper inférieur à alpha
- **Donc un test significatif permet de porter une conclusion et de commencer l'interprétation causale**

Interpréter un test non significatif

- Si la p-valeur calculée est supérieur au seuil alpha choisi alors
  - On ne rejette pas H0 hypothèse d'égalité des paramètres
  - Ne pas rejeter H0 c'est ne pas rejeter un modèle
    - Modèle d'égalité des paramètres
    - C'est un des modèles compatibles avec les données
  - Rejeter H0 c'est rejeter ce modèle
- **Donc un test non significatif ne permet pas de conclure formellement**
  - Ni à l'existence d'une différence puisque le test est non significatif
  - Ni à la preuve d'une absence de différence bien que cette hypothèse fasse partie des possibles

Différence non significative



- Impossible de conclure
- Ne prouve pas qu'il n'y a pas d'effet
  - Il n'y a peut être pas d'effet
  - l'effet est trop petit pour être décelable
  - ou les conditions de mesure de l'expérience ne permette pas de conclure (variance trop grande)

En pratique dans statview

Question : les patients avec une embolie pulmonaire sont-ils en moyenne plus âgés que les patients sans embolie pulmonaire

Hypothèses

- H0 : la moyenne des patients avec EP = la moyenne des patients sans EP
- H1 : ces deux moyennes sont différentes
- Si je rejette H0 alors je conclus à l'existence d'une différence

Résultats staview

1 Nb et moyennes

2 Conditions validité ?

Info. du groupe pour age
Variable "groupe" : EP

	nombre	Moy.	Variance	Dév Std	Erreur Std
NEGATIF	926	61,3	346,4	18,6	0,6
POSITIF	241	69,2	306,7	17,5	1,1

3 Interprétation test

Test-t séries non appariées pour age
Variable "groupe" : EP
Ecart théorique = 0

	Ecart moyen	DDL	t	p
NEGATIF, POSITIF	-7,9	1165	-5,9	<0,0001

Présentation des résultats

- La dimension clinique prime sur la p valeur
  - Les patients avec EP sont en moyenne plus âgés
    - 61 ans ± 18 ans
  - Que les patients sans EP
    - 69 ans ± 17 ans
  - Cette différence est statistiquement significative $p < 0,01$

En pratique dans R2web

Question : les patients avec un ATCD de MTE ont-ils plus souvent une embolie pulmonaire ?

Hypothèses

- H0 : le pourcentage d'embolie pulmonaire chez les patients **AVEC** ATCD de MTE = le pourcentage d'embolie pulmonaire chez les patients **SANS** ATCD de MTE
- H1 : ces deux fréquences sont différentes
- Si je rejette H0 alors je conclus à l'existence d'une différence
- Test bilatéral

Résultats R2Web

1 tableau de contingence

> Matrice des observations

	Non	Oui
NEGATIF	805	121
POSITIF	177	64

2 Conditions validité ?

> Matrice théorique

	Non	Oui
NEGATIF	779,205	146,795
POSITIF	202,795	38,205

34 % des patients avec ATCD ont une EP
18% des patients sans ATCD ont une EP

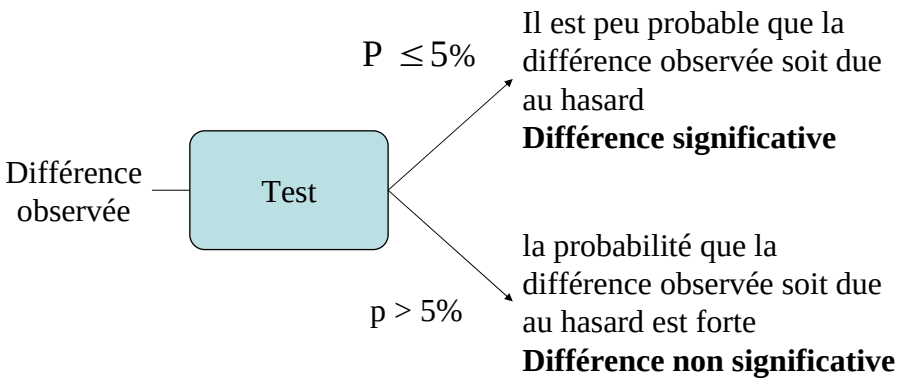
3 Interprétation test

Chi-deux = 25, P valeur observée <0,01

Présentation des résultats

- La dimension clinique prime sur la p valeur
  - Le pourcentage d'EP est plus élevé en cas d'ATCD de MTE
    - 64 ep sur 185 patients (34 %)
  - Que sans ATCD de MTE
    - 177 ep sur 982 patients (18 %)
  - Cette différence est statistiquement significative p < 0,01

2.1.1. Test statistique : interprétation



Calcul du nombre de sujets nécessaire

- C’est le nombre de sujet nécessaire pour que la différence entre les 2 paramètres soit significative significative au risque alpha choisi (5%) si cette différence existe vraiment
  - 2 moyennes de tension artérielles selon TTT A ou B
  - 2 pourcentages d’une pathologie selon FDR présent ou absent (équivalent à RR ou OR = 1)

Calcul du nombre de sujets nécessaire

- Il dépend du delta entre les deux paramètres
  - Il diminue si ce delta est grand
  - Il augmente si delta est petit
- Il dépend de la variance estimée pour un paramètre continu
  - Il diminue si la variance est faible
  - Il augmente si la variance est grande
- Il dépend du risque alpha et de la puissance souhaitée pour l’expérience
  - Il augmente avec la puissance
  - Il augmente avec un risque alpha plus petit

Nombre de sujets nécessaire Exercice

- Allez sur le site http://www.u707.jussieu.fr/biostatgv/
  - Rubrique calcul du nombre de sujets nécessaire

  Comparer deux moyennes / deux groupes

  Simuler un essai TTT (puissance 80 % = 0.8, alpha 0.05) avec
  M1 = 135 mm de hg de tension artérielle systolique dans le groupe A
  M2 = 145 mm de hg dans le groupe B
  Combien de sujets
  - si la variance est faible écart type = 20 (mesure par holter sur 24 H)
  - si la variance est forte écart type = 60 (brassard à tension)

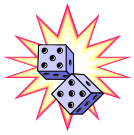
  Refaites les calculs avec une puissance de 0.9 puis 0.95

Un mauvaise usage des test : la répétition des tests

- Conclusion basée
  - non pas sur un seul test
  - mais sur plusieurs
- Conclusion à un effet à partir du moment où il existe au moins un test significatif parmi n tests
- Le risque α de la conclusion est bien supérieure à 5%
  - Inflation du risque alpha

Comparaisons multiples

Aux dés, la probabilité d 'obtenir un six est plus forte avec 3 dés qu 'avec un seul



Test 1 → **Test 2** → **Test 3** → **Test 4**

Risque de conclure à tort à une différence = 5%

Risque de conclure à tort à une différence = 5%

Risque de conclure à tort à une différence = 5%

Risque de conclure à tort à une différence = 5%

n	risque global
2	0.10
3	0.13
5	0.23
10	0.40

Globalement, le risque de conclure à tort à une différence lors de ces 4 comparaisons est bien plus important que 5%.

Inflation du risque alpha

Avoir un critère de jugement principal limite le nombre de tests statistiques

- Décès de toute cause
- Décès cardiovasculaire
- Mort subite
- Infarctus
- Accident vasculaire cérébraux
- Chirurgie

Pas de définition de critère principal

7 tests statistiques

Risque de conclure à tort à l 'efficacité du traitement = 30%

Critère principal

- Décès de toute cause

Critères secondaires

- Décès cardiovasculaire
- Mort subite
- Infarctus
- Accident vasculaire cérébraux
- Chirurgie

Définition a priori d 'un critère principal

Un seul test statistique

Risque de conclure à tort à l 'efficacité du traitement = 5%

Mauvais usage des tests : les analyses en sous-groupes a posteriori

Essai non significatif entre 2 TTT A et B	**p=0.92**	**NS**
I) J'essaye par tranche d'âge		
1 Age<75	0.92	NS
2 Age>75	0.95	NS
II) J'essaye selon le sexe		
3 Hommes	0.92	NS
4 Femmes	0.99	NS
III) Selon les ATCD		
5 Antécédents d'infarctus	0.87	NS
6 Pas d'antécédents d'infarctus	1.03	NS
IV) Selon la prise d'aspirine		
7 Prise d'aspirine	0.03	p<0.05
8 Pas d'aspirine	1.09	NS

Mauvais usage des tests : les analyses en sous-groupes a posteriori

CA MARCHE !!!!
A et mieux que B si on prend de l'aspirine

en fait on a fait 8 tests avec chacun un risque d'erreur de 5%

Au total le risque d'erreur est >>> 5%

Il est de l'ordre de 50 %

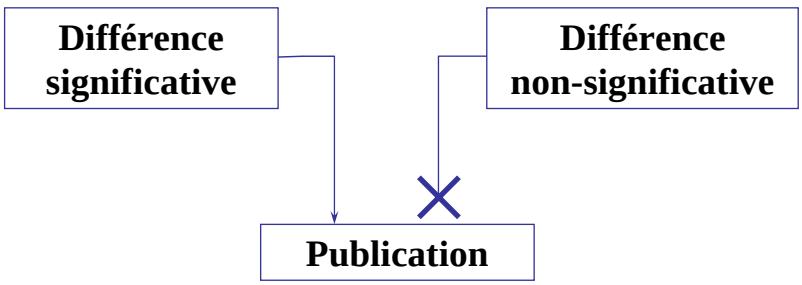
En multipliant les tests on perd le bénéfice des tests : conclure avec un risque d'erreur faible et connu

Analyses en sous groupes a posteriori

- Sont de nature exploratoire
- Ne donnent pas de démonstration
- Suggèrent des variations d'efficacité donc des idées de protocoles futurs

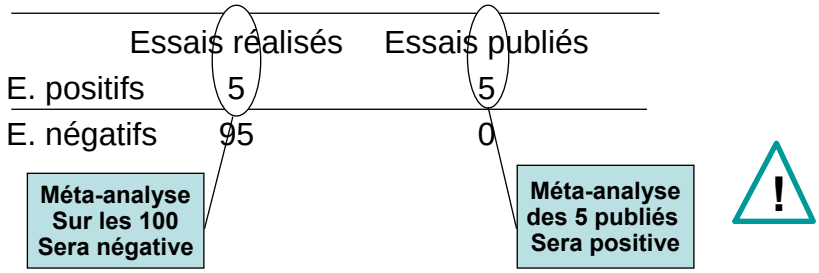
Biais de publication

Les essais statistiquement significatifs sont plus facilement publiés que les négatifs



Biais de publication

- Un essai peut être positif à tort (risque alpha)
- Exemple d'un traitement sans efficacité
  - On fait 100 simulations
- Risque alpha = 5%



Conclusion

- Un test statistique permet de porter une conclusion avec un risque d'erreur connu et contrôlé
  - Le risque alpha de première espèce usuel est 0,05
- Pour contrôler le risque d'erreur de conclure à tort il faut
  - Choisir le bon test
  - Respecter les conditions de validités des tests
  - Peu de tests, définis à l'avance dans un protocole
- La puissance d'un test augmente avec
  - Le nombre de sujets
  - Le contrôle des sources de variabilité (diminution de la variance)

Biostatistiques médicales > BIOSTAT01 > Principe d'un test statistique. - Questions du [10/11/2008] au [17/11/2008]

No	Diap	Intérêt	Question
1	6	1 Etud.	Quelle est la différence entre a et IC puisque les 2 mesurent le risque que la différence observée entre les paramètres soit réelle?
2	6	1 Etud.	Je ne comprends pas à quel niveau l'IC mesure si les paramètres diffèrent ou non.
3	9	1 Etud.	Pourriez vous réexpliquer le calcul sur l'intervalle de proportion qui est sur la diapo.
4	15	10 Etud.	Bonjour. Serais-t-il possible que vous repreniez plus en détail la notion de p value (= p valeur)lors du prochain cours svp?
5	16	1 Etud.	Qu'est ce que la référence DATO TORTURING NEJM?
6	21	2 Etud.	Pourquoi la dimension clinique prime sur la p valeur?
7	21	1 Etud.	Quel est le lien à faire entre la variance et les conditions de validité?
8	22	1 Etud.	Pourquoi la dimension clinique prime sur la p valeur?
9	26	2 Etud.	Quelle est la différence entre le delta et la variance?
10	27	2 Etud.	Pour le calcul du "nombre de sujets nécessaires" sur le site internet indiqué dans le diaporama,à quoi correspondent les termes "unilatéral" et "bilatéral" (il faut cocher l'un des deux pour faire les calculs)?
11	27	1 Etud.	Comment se fait le calcul du nombre de sujets nécessaires ?
12	30	5 Etud.	Pourquoi a-t-on une inflation du risque alpha?
13	30	2 Etud.	Comment on trouve un risque de 0,13 pour n=3? Idem pour les autres n
14	31	2 Etud.	Pourquoi est-ce qu'avoir un critère de jugement principal limite le nombre de tests?
15	32	2 Etud.	A quoi correspond NS?
16	36	2 Etud.	Qu'est-ce que la méta-analyse?

Med@TICE
Faculté de Médecine de Grenoble
PCEM2

Mentions légales

UNIVERSITE JOSEPH FOURIER
SCIENCES. TECHNOLOGIE. SANTÉ