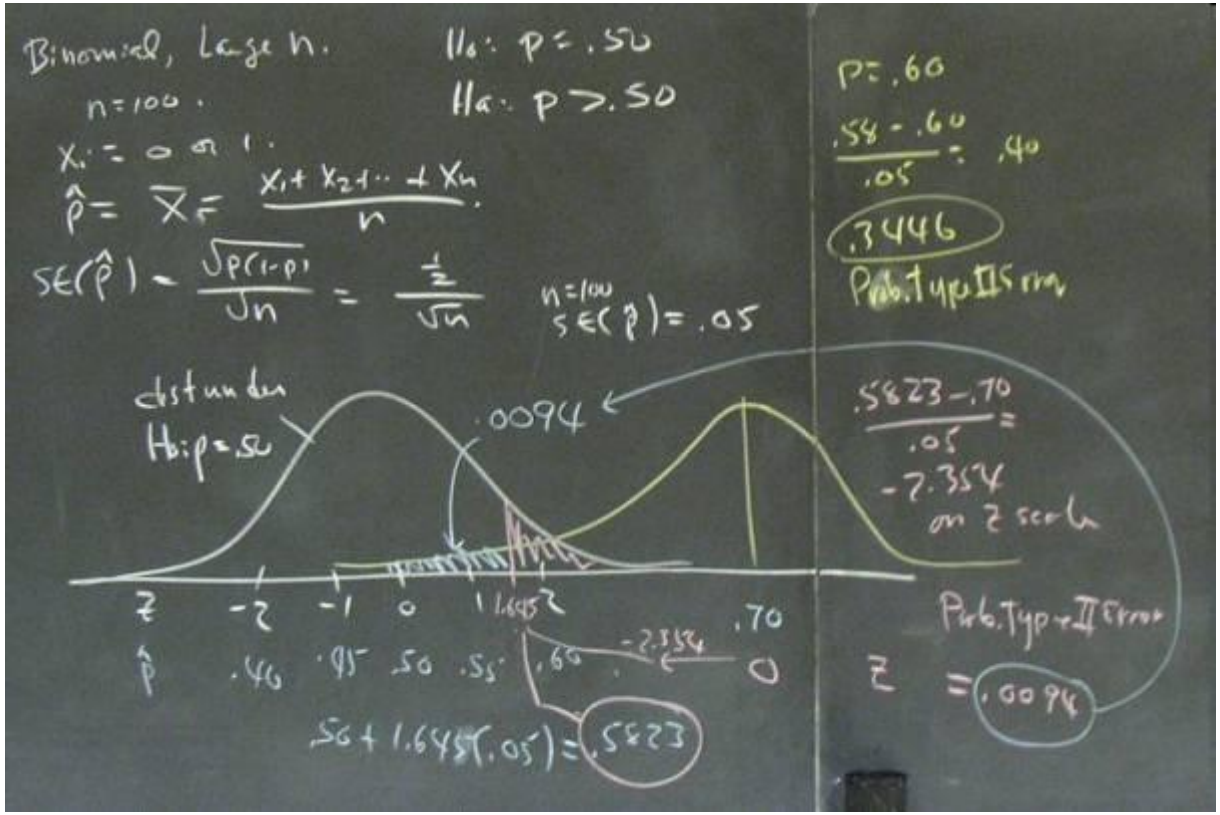
Analyse & traitement de données : mesurer, tester des hypothèses

Mise à jour du 20 octobre 2011

Rémi Bachelet

Dernière version des diapos disponible ici : [mesure et test d'hypothèses]



bachelet@bigfoot.com
École Centrale de Lille
Villeneuve d'Ascq - France

Source des images indiquées au-dessous ou en cliquant sur l'image

Image: source

Le cholera

  - Diarrhées brutales et très abondantes, crampes d'estomac, soif intense…
  - Les cas se déclenchent souvent à quelques heures d'intervalle
  - 50% de mortalité (de trois heures à trois jours).

- Les explications :
  1. Théorie des miasmes (dominante à l'époque) : les vecteurs des maladies contagieuses sont les gaz qui s'échappent des corps en décomposition
  2. Théorie des germes : les maladies sont causées par des micro-organismes (subsistants essentiellement dans les eaux)
- Le 31 août 1854, une épidémie de cholera frappe le quartier de Soho.

- En trois jours, 127 personnes habitant près de Broad Street meurent. Les trois quart des autres résidents fuient le quartier
- John Snow enquête … et collecte des données

*On proceeding to the spot, I found that nearly **all the deaths had taken place within a short distance of the [Broad Street] pump**. There were only ten deaths in houses situated decidedly nearer to another street-pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pumps which were nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street... (...)*

I had an interview with the Board of Guardians of St James's parish, on the evening of the 7th inst [Sept 7], and represented the above circumstances to them. In consequence of what I said, the handle of the pump was removed on the following day.

John Snow

(le puit de pompe était creusé à 1 m d'une fosse d'aisance, dans laquelle avait été jetée la couche d'un bébé infecté par le Vibrio cholerae)

Image: source

Relevé des victimes du choléra John Snow, 1854

Mesurer et tester des hypothèses

1. La causalité
   - Comment tester une théorie ?
2. Définir des construits, mesurer des variables
   - Variables nominales, ordinales, métriques…
   - Variables qualitatives : l'analyse de contenu

Qu'est-ce que la causalité ?

Selon John Stuart Mill (1806-1873), trois critères permettent d'inférer la causalité :

i. La covariation,
   - Cause et effet sont corrélés

ii. La précédence temporelle
   - La cause précède l'effet

iii. L'élimination d'explications alternatives.
   - Pas de troisième variable

Comment tester une théorie ?

Une théorie propose des **construits** qui permettent de formuler des hypothèses

1. Définir rigoureusement les construits
   - 1/ Concept => 2/dimensions => 3/composantes
   - « Température de la terre » => t° eau; t° air, t° sol => t° (x, y, z, t)
2. .. puis mesurer des variables pour estimer les composantes
   - Variables métriques (sc physiques), mais aussi nominales, ordinales (sc humaines)
3. Tester mathématiquement les hypothèses

Les variables métriques sont de divers types

- Continues ou discrètes
  - Poids, taille (métrique continu)
  - Image scanner, capacité à grimper sur une échelle jusqu'à un certain barreau (métrique discret)

- On peut faire énormément de calculs, surtout avec les variables continues : ACP

Les variables nominales

Elles ne peuvent faire l'objet d'un classement par ordre croissant…

par exemple :

- Lieu de naissance, plat préféré
- Sexe (dichotomique)

La plupart des calculs à partir de variables nominales sont impossibles, car il n'ont pas de sens.

- Calculer une « moyenne » entre des marques de voitures ?
- Mais, on peut parfois les convertir en variables métriques
  - destinations de vacances => distance (km)
  - marques de voitures => prix moyen
  - vote à une élection => échelle droite <=> gauche.

Variables Ordinales

- Elles sont ordonnées, mais pas métriques
  - Réponse sur une échelle d'estime de soi
  - .. une échelle du type de celles proposées par Rensis Likert (1903 - 1981)

    « J'ai confiance en moi »,

    cochez la case correspondant à votre opinion

    □ tout à fait d'accord □ plutôt d'accord □ plutôt pas d'accord □ pas d'accord du tout

- Problème : pour les traiter.. faut-il les considérer comme ..
  1. … des variables métriques (tout à fait = 1, plutôt = 2 …)
  2. ..ou des variables nominales ?

- Effets pervers
  - En numérisant un Likert (pas du tout d'accord = 1, assez d'accord =2..) on est tenté de faire des calculs : moyenne écart-type ..
  - Or, ces chiffres n'ont en fait que **peu de sens**, il impliquent notamment un postulat caché sur les « distances » entre les réponses
    - passer de « pas du tout d'accord » à « assez d'accord » est-il identique à passer de « assez d'accord » à « plutôt d'accord » ?

En sciences humaines, les variables mesurées sont rarement quantitatives au départ

- Affirmation
  - Opinion, réponse sur une échelle d'estime de soi
- Comportement
  - Rencontrer quelqu'un, éviter de faire quelque chose
- …. Voire discours sur un comportement
  - Par exemple « utilisation d'un préservatif »
  - Cf. biodata dans le cours sur la conception de questionnaires

Autres types de variables

- « Classez par ordre de préférence »
  - Premier choix, réponses multiples ..
  - ➢ Données dures à exploiter !
- Graphes
  - Par exemple réseau relationnel / sociogramme
  - ➢ Conversion du graphe en matrice et analyse structurale
- Variables textuelles
  - Texte brut ou transcription d'un entretien
  - ➢ Analyse de contenu, voir ci-après

⇒ Erreur très fréquente : collecter des données et ne pas être capable de les exploiter ensuite !

1. Savoir-faire : logiciels maitrisés, éviter de croire que « plus on utilise de mathématiques, meilleur c'est »
2. Méthodologie : rigoureuse et … comprise par le client
3. Temps .. et coût..(3* la durée d'un entretien pour le taper et autant pour l'analyser).

L'analyse de contenu

- Elle se fait « avec le cerveau » !
  1. Construire un tableau des concepts
  2. Faire une carte cognitive / conceptuelle

Création d'une carte conceptuelle +

Ou d'une mind map ….

Logiciels d'aide à la fabrication de cartes conceptuelles :

- Freeplane
- Visual Understanding Environment (VUE)

[[Guide - Réaliser une carte conceptuelle]]

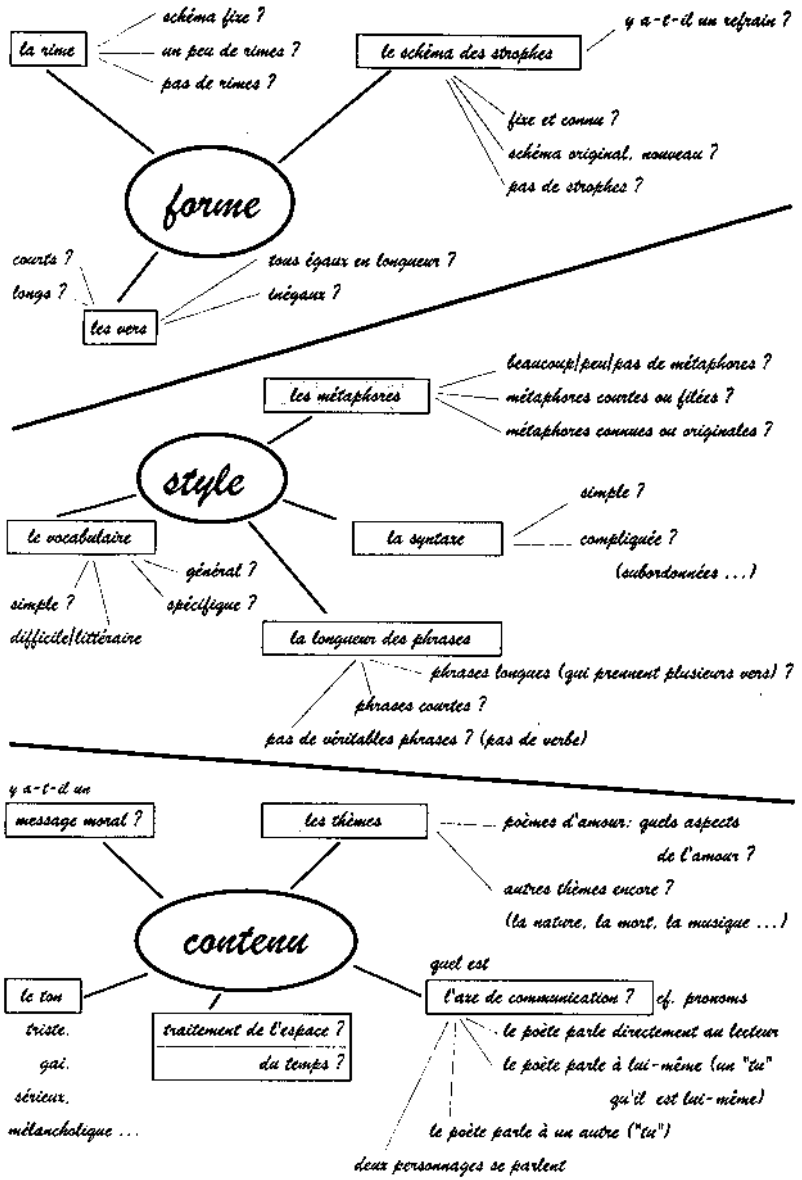


Image: source

13

Test d'hypothèses

1. L'hypothèse nulle
   - Risques de première et deuxième espèces
2. Choisir parmi les tests statistiques
   - Variables nominales, ordinales, métriques…
3. Panorama des méthodes de recherche

Test d'hypothèse

Une démarche consistant à **rejeter** une hypothèse statistique, appelée *H0*, en fonction de données.

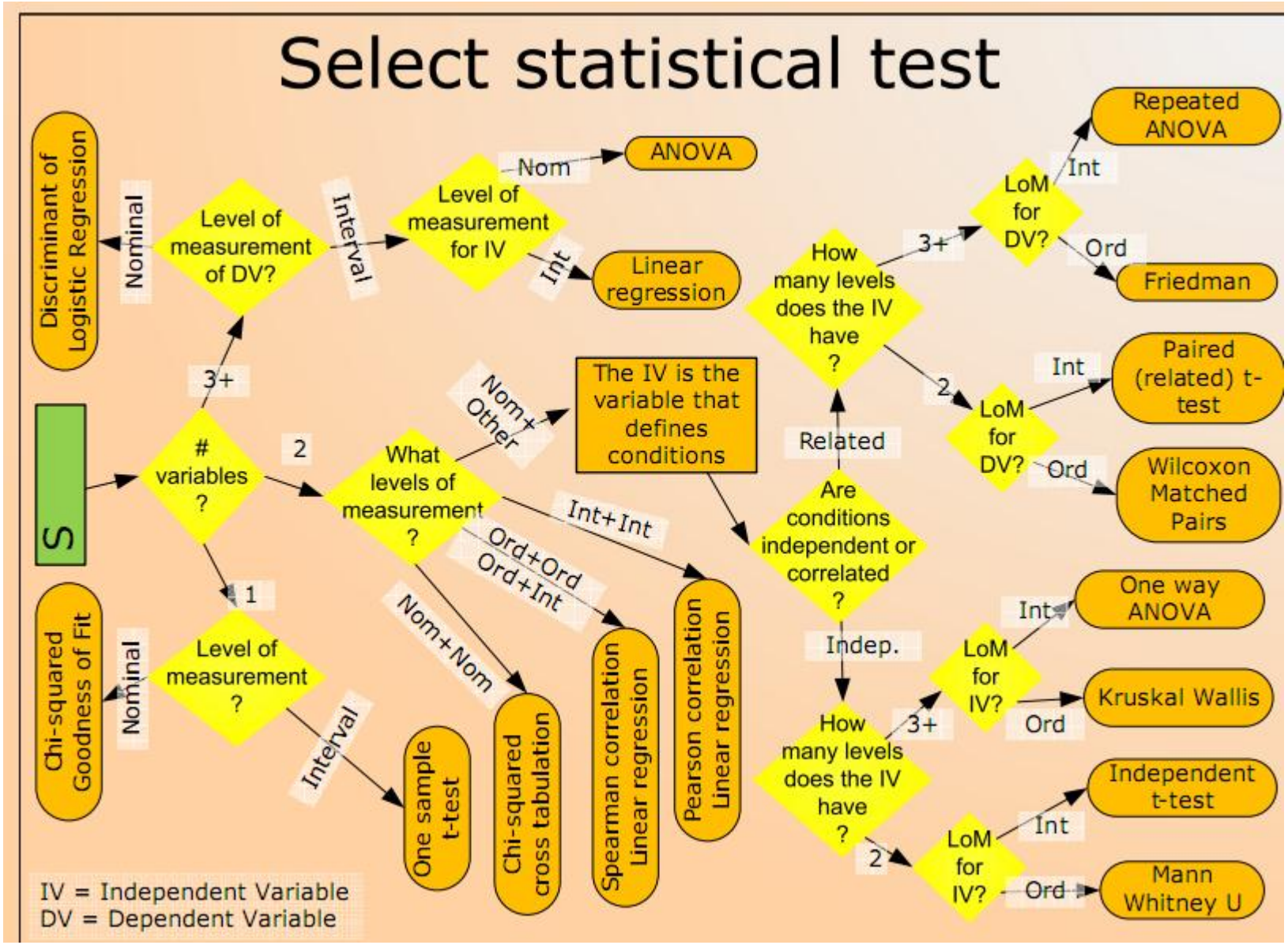
- On cherche à tester si un paramètre a une valeur donnée.
  - L'hypothèse nulle H0 est par exemple « patient déclaré séropositif au VIH » et l'hypothèse contraire = H1 « patient déclaré séronégatif ».
- Il y a deux façons de se tromper lors d'un test statistique :
  1. **Rejeter à tort** H0. risque de **première espèce** α = faux positif : accepter une hypothèse alors qu'elle était fausse (test positif à tort).
  2. **Accepter à tort** H0 : risque de **deuxième espèce** β =faux négatif : rejeter une hypothèse alors qu'en fait elle était vraie (test négatif à tort).

Déroulement d'un test

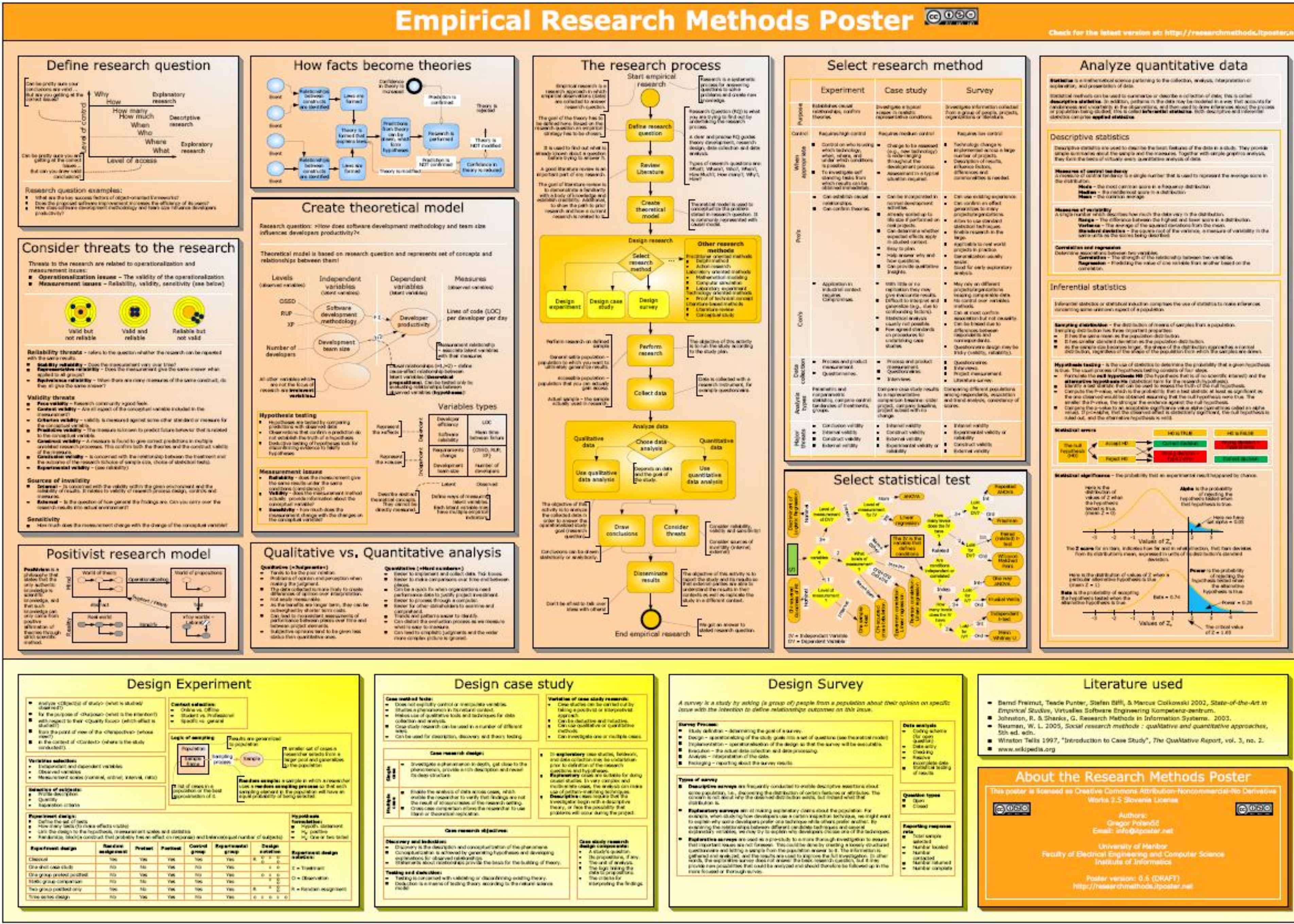
1. Énoncé de l'hypothèse nulle H0 (et de l'hypothèse alternative H1).
2. Calcul d'une variable de décision
   - = une mesure de la distance entre les deux échantillons (test d'homogénéité), ou entre l'échantillon et la loi statistique (test de conformité).
   - Plus cette distance sera grande et moins l'hypothèse nulle H0 sera probable.
   - Calcul de la **probabilité**, en supposant que H0 est vraie, d'obtenir une valeur de la variable de décision au moins aussi grande que la valeur de la statistique que l'on a obtenue avec notre échantillon. Cette probabilité est appelée la p-value.
3. Conclusion du test, en fonction d'un risque seuil.
   - Souvent, un risque de 5% est considéré comme acceptable (c'est-à-dire que dans 5% des cas quand H0 est vraie, l'expérimentateur se trompera et la rejettera).
4. Si la p-value est plus grande que 5% on accepte l'hypothèse H0. Si la p-value est plus petite que 5% on la rejette.

Ici (et souvent) le seul risque α est utilisé comme critère de décision et on étudie un test unilatéral.

Choisir parmi les tests statistiques d'hypothèses



source



Empirical Research Methods Poster
Define research question
How facts become theories
The research process
Select research method
Analyze quantitative data
Consider threats to the research
Create theoretical model
Descriptive statistics
Inferential statistics
Select statistical test
Positivist research model
Qualitative vs. Quantitative analysis
Design Experiment
Design case study
Design Survey
Literature used
About the Research Methods Poster
Experiment
Case study
Survey
Start empirical research
Define research question
Review literature
Create theoretical model
Design research
Select research method
Design experiment
Design case study
Design survey
Perform research
Collect data
Analyze data
Qualitative data
Quantitative data
Choose data analysis
Use qualitative data analysis
Use quantitative data analysis
Draw conclusions
Consider threats
Disseminate results
End empirical research
Valid but not reliable
Valid and reliable
Reliable but not valid
Variables types

Questions ?