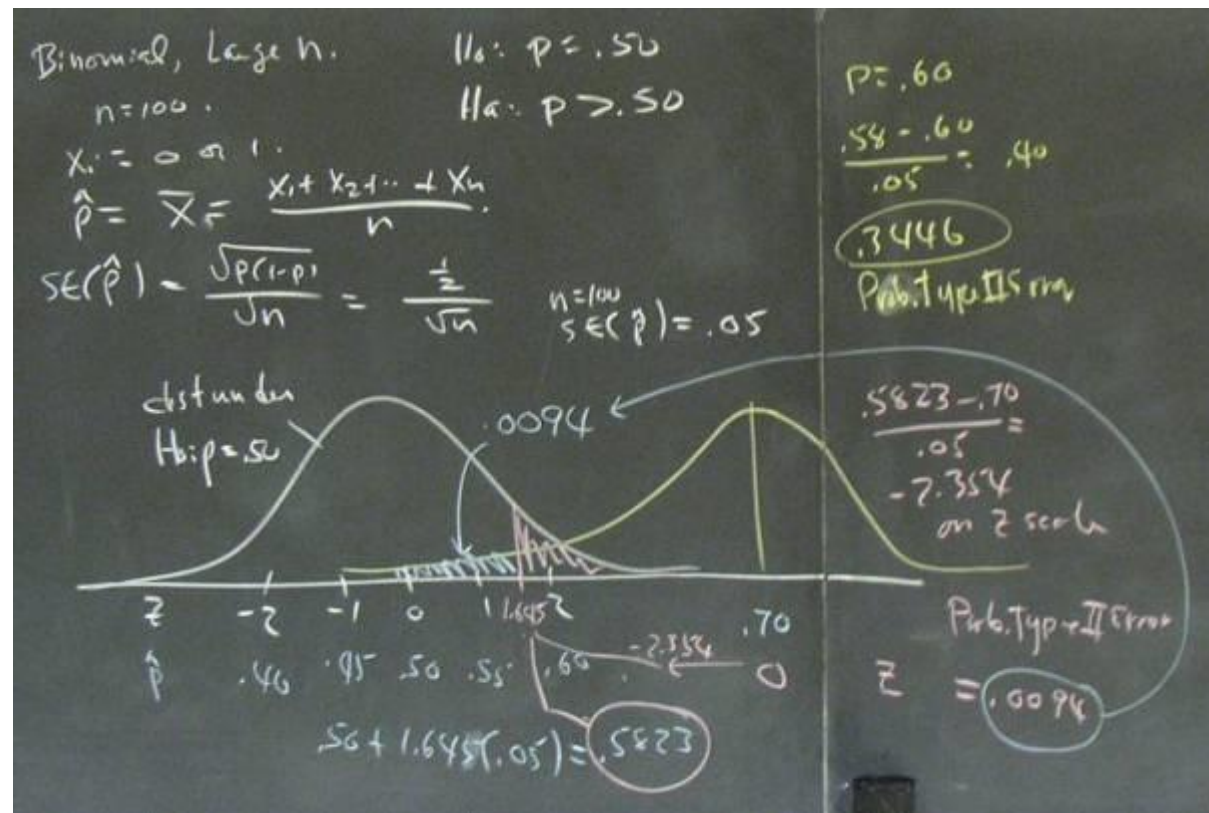


Analyse & traitement de données : mesurer, tester des hypothèses

Mise à jour du 26 août 2020

Rémi Bachelet

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



Cours distribué sous licence
Creative Commons,
selon les conditions suivantes :



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 maîtrisés, 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]]

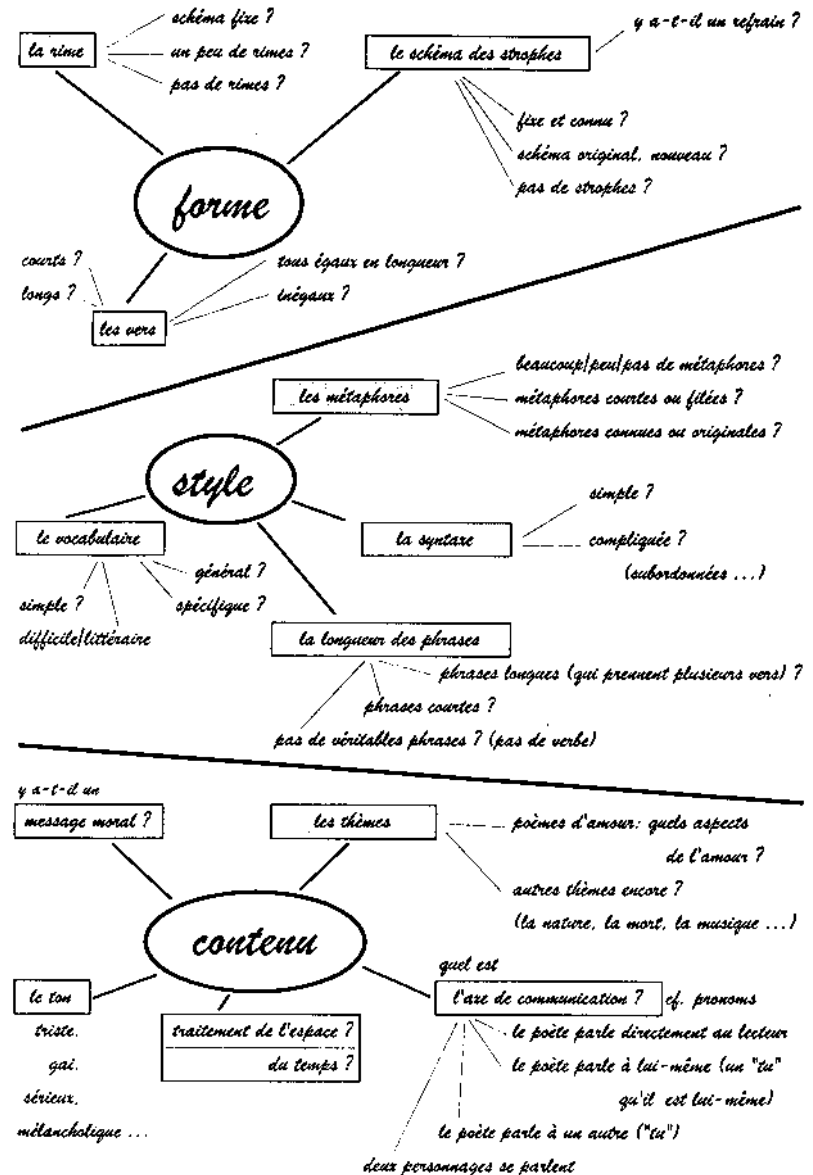


Image: [source](#)

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 H_0 , en fonction de données.

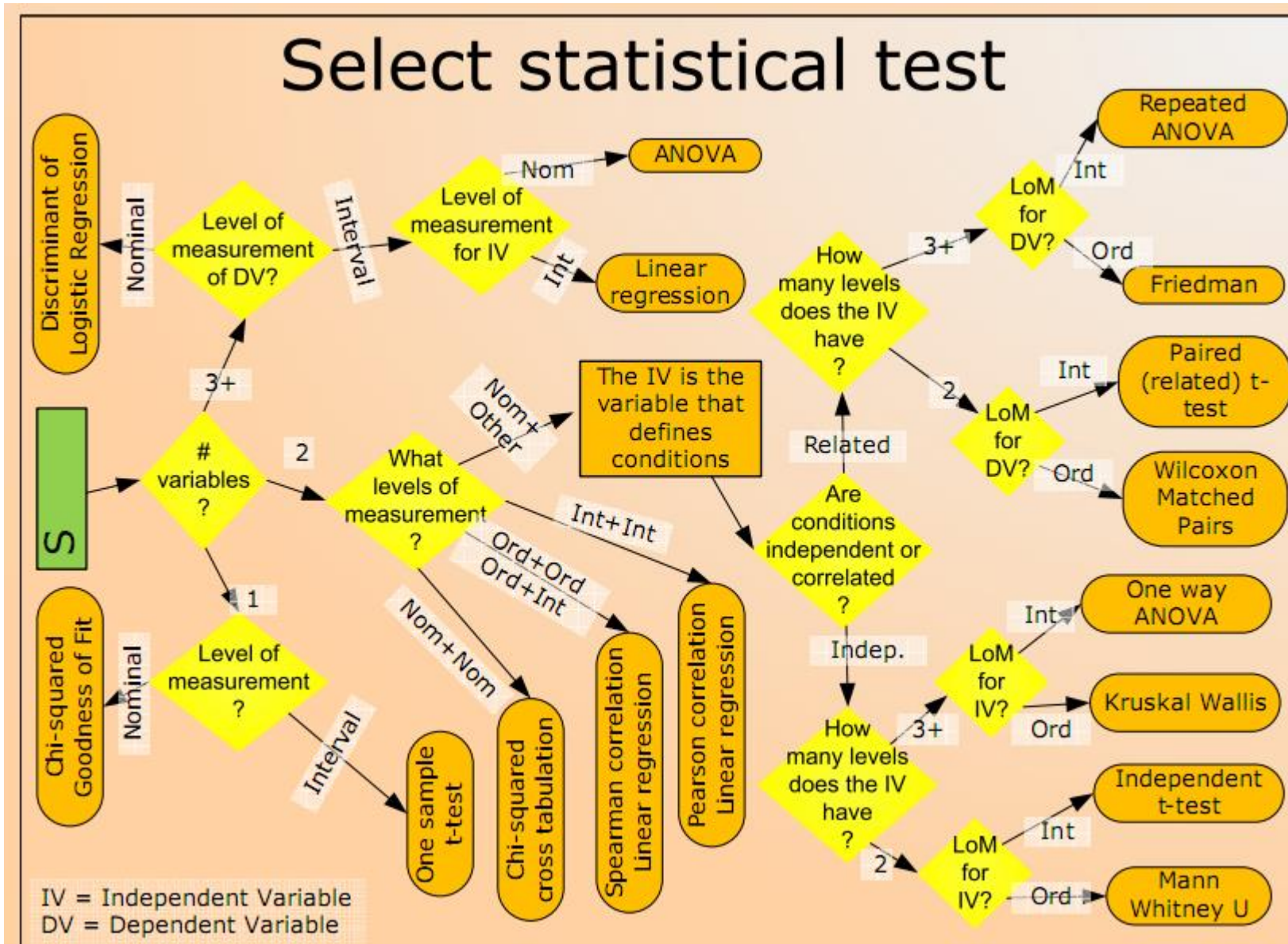
- On cherche à tester si un paramètre a une valeur donnée.
 - L'hypothèse nulle H_0 est par exemple « patient déclaré séropositif au VIH » et l'hypothèse contraire = H_1 « patient déclaré séronégatif ».
- Il y a deux façons de se tromper lors d'un test statistique :
 1. **Rejeter à tort H_0 .** risque de **première espèce** $\alpha =$ faux positif : accepter une hypothèse alors qu'elle était fautive (test positif à tort).
 2. **Accepter à tort H_0 :** risque de **deuxième espèce** $\beta =$ 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 H_0 (et de l'hypothèse alternative H_1).
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 H_0 sera probable.
 - Calcul de la **probabilité**, en supposant que H_0 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 H_0 est vraie, l'expérimentateur se trompera et la rejettera).
4. Si la p-value est plus grande que 5% on accepte l'hypothèse H_0 . 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



Empirical Research Methods Poster



Check for the latest version at: <http://researchmethods.poster.net>

Define research question

Can you answer your question? Can you answer it in a way that is interesting? Can you answer it in a way that is relevant? Can you answer it in a way that is novel? Can you answer it in a way that is important? Can you answer it in a way that is useful? Can you answer it in a way that is interesting? Can you answer it in a way that is relevant? Can you answer it in a way that is novel? Can you answer it in a way that is important? Can you answer it in a way that is useful?

Why (Explanatory research)
How (Descriptive research)
When (Explanatory research)
Where (Explanatory research)
What (Explanatory research)

Level of access
Level of control
Level of access

Research question examples:

- What are the key success factors of agile/agile frameworks?
- How do agile frameworks impact on the success of agile projects?
- How do agile frameworks impact on the success of agile projects?
- How do agile frameworks impact on the success of agile projects?

Consider threats to the research

Threats to the research are related to operationalization and measurement issues.

- Operationalization issues** - The validity of the operationalization
- Measurement issues** - Reliability, validity, sensitivity (see below)

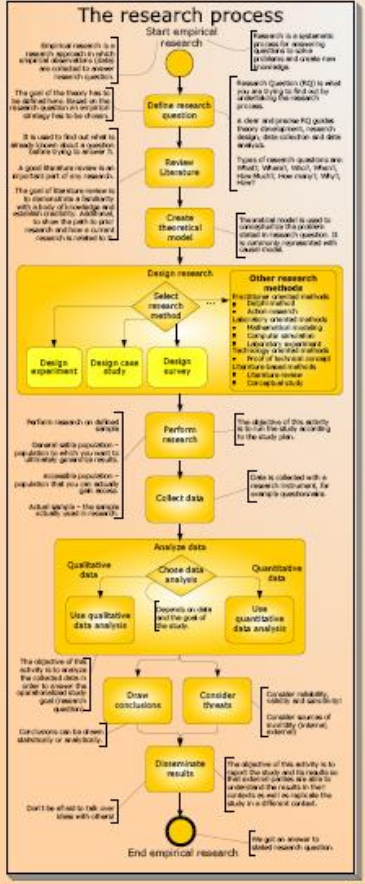
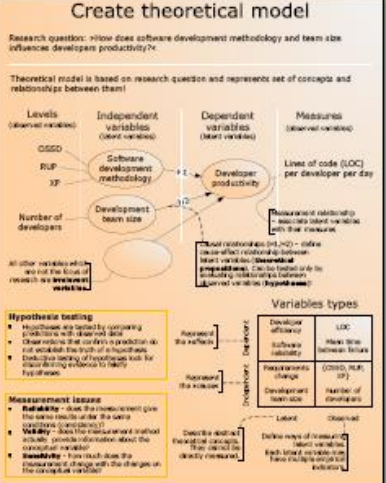
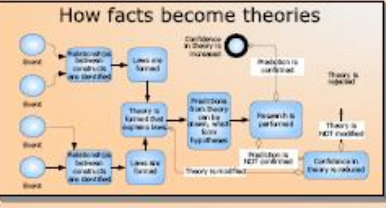
Reliability threats - refers to the extent where the research can be repeated with the same results.

- Stability reliability** - Does the measurement vary over time?
- Internal reliability** - Does the measurement give the same answer when asked in different ways?
- Inter-rater reliability** - Does the measurement give the same answer when asked by different raters?
- Test-retest reliability** - Does the measurement give the same answer when asked at different times?

Validity threats - refers to the extent where the research can be repeated with the same results.

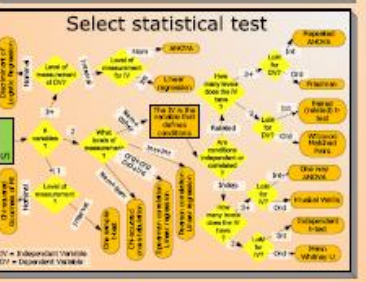
- Construct validity** - Are the measurement tools used to measure what they are intended to measure?
- Internal validity** - Does the measurement give the same answer when asked in different ways?
- External validity** - Does the measurement give the same answer when asked in different contexts?

Sensitivity - How much does the measured change with the change of the conceptual variable?



Select research method

	Experiment	Case study	Survey
Purpose	Establish causal relationships, control confounding factors	Investigate a causal process in a real world situation	Investigate a causal process in a real world situation
When to use	Requires high control	Requires moderate control	Requires low control
When to avoid	Control is not possible when technology, when users, and when interactions are involved	Change to a research method is required during the development process	Technology changes rapidly, large number of projects, distribution of results, influence factors, differences and confounding factors is needed
Pros	Control over what is measured, when, where, and how interactions are involved	Can be implemented in a real world situation	Can be implemented in a real world situation
Cons	Application in industrial context is difficult	Requires a lot of resources	Requires a lot of resources



Analyze quantitative data

Statistical analysis is a method of analyzing data to describe, summarize, and present data.

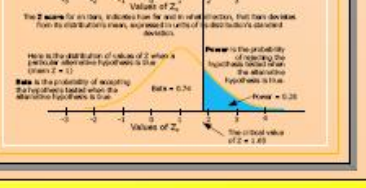
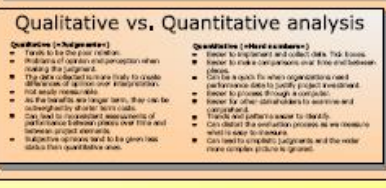
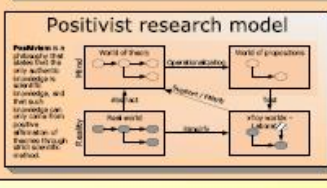
Descriptive statistics: Describe the data in terms of its central tendency, dispersion, and shape.

- Measures of central tendency:** Mean, Median, Mode.
- Measures of dispersion:** Range, Variance, Standard deviation.

Inferential statistics: Inferential statistics are used to make inferences about a population based on a sample.

- Parametric tests:** t-test, ANOVA, F-test, etc.
- Non-parametric tests:** Chi-square, Fisher's exact test, etc.

Statistical significance: The probability that an experimental result occurred by chance.



Design Experiment

	Researcher	Participant	Researcher	Participant	Researcher	Participant	Design
Control	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Random assignment	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over what is measured	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over when, where, and how interactions are involved	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the development process	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the distribution of results	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the influence factors	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the differences and confounding factors	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E

Design Case Study

	Researcher	Participant	Researcher	Participant	Researcher	Participant	Design
Control	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Random assignment	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over what is measured	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over when, where, and how interactions are involved	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the development process	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the distribution of results	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the influence factors	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the differences and confounding factors	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E

Design Survey

	Researcher	Participant	Researcher	Participant	Researcher	Participant	Design
Control	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Random assignment	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over what is measured	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over when, where, and how interactions are involved	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the development process	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the distribution of results	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the influence factors	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the differences and confounding factors	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E

Literature used

	Researcher	Participant	Researcher	Participant	Researcher	Participant	Design
Control	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Random assignment	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over what is measured	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over when, where, and how interactions are involved	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the development process	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the distribution of results	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the influence factors	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E
Control over the differences and confounding factors	Yes	Yes	Yes	Yes	Yes	Yes	A, B, C, D, E

About the Research Methods Poster

This poster is licensed under Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License.

Authors: Gregor Podszus, Email: podszus@uni-leipzig.de

University of Member: Faculty of Electrical Engineering and Computer Science, Institute of Informatics

Poster version: 6.0 (DRAFT)
<http://researchmethods.poster.net>

Questions ?

Autres cours :

1. Explorer ou vérifier ? Deux catégories d'approches
2. Éventails des démarches de recueil de données
3. Conception de questionnaires
4. Techniques d'entretien et reformulation
5. L'Analyse Factorielle des Correspondances pour les nuls
6. Validité et Fiabilité des données