

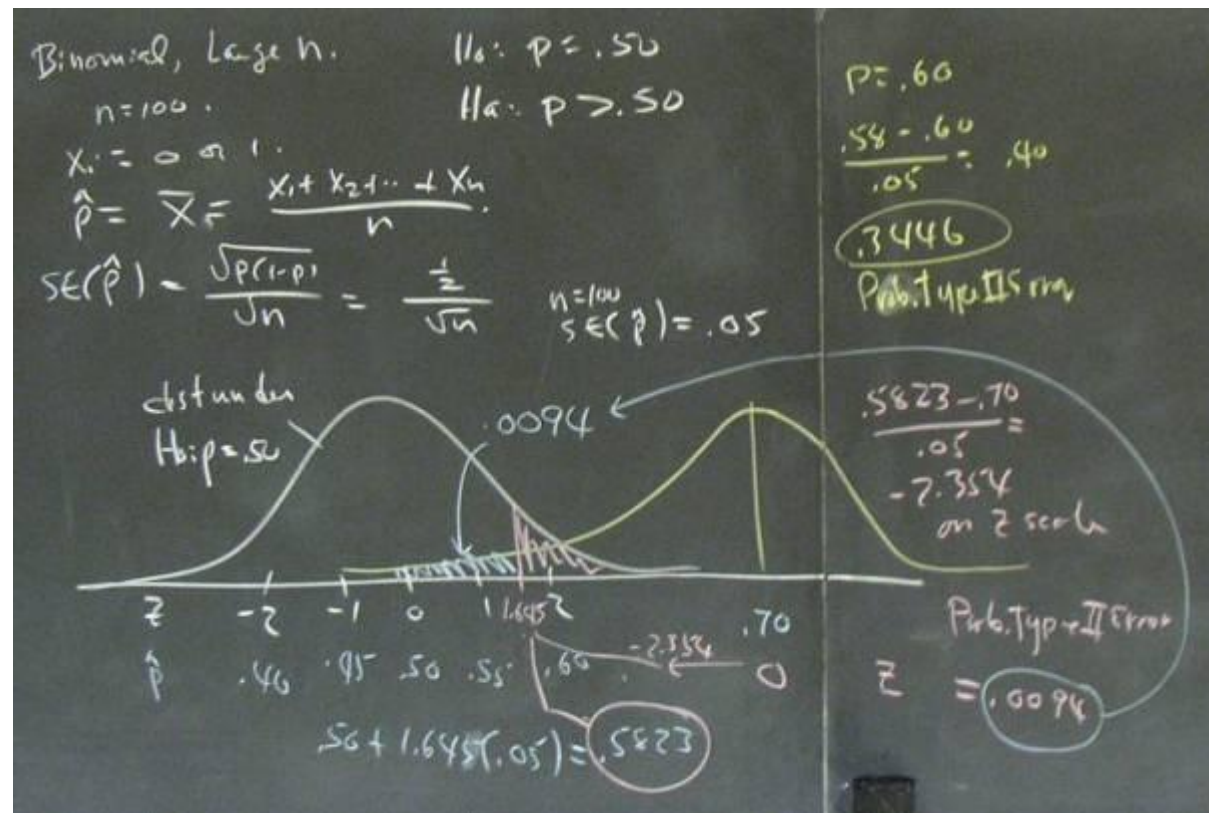


Analyse & traitement de données : mesurer, tester des hypothèses

Mise à jour du 20 octobre 2011

Rémi Bachelet

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



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



bachelet@bigfoot.com

École Centrale de Lille

Villeneuve d'Ascq - France

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*)*



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]]

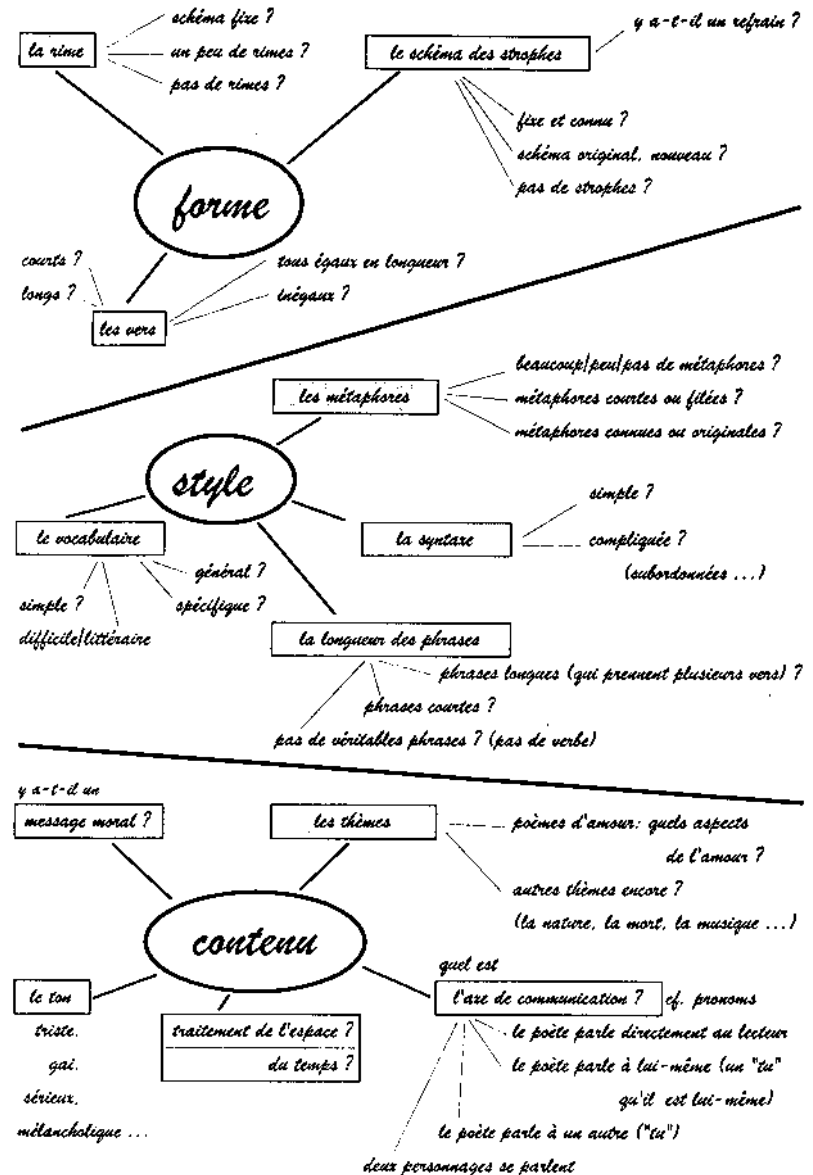


Image: [source](#)

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.

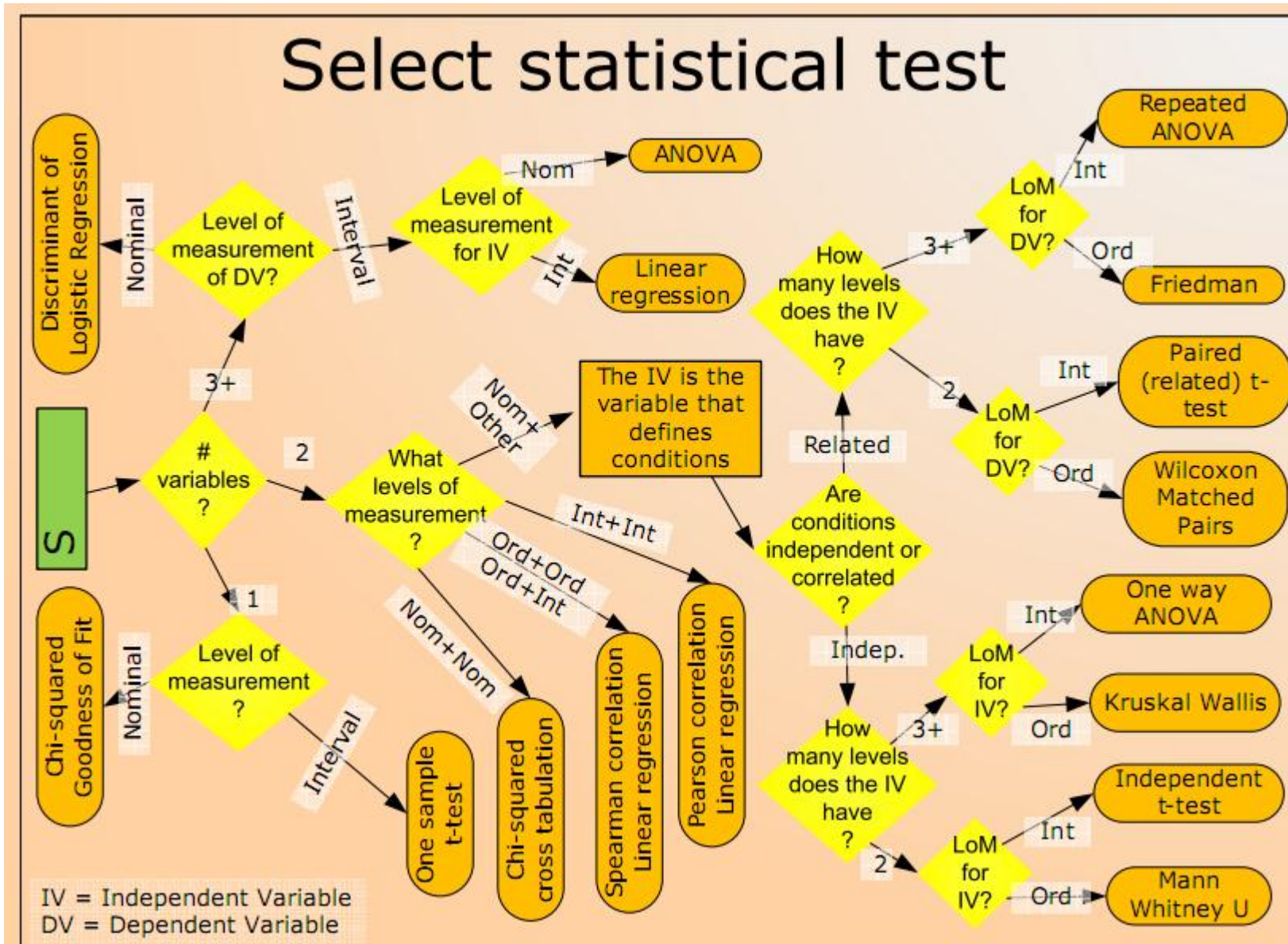
- On cherche à tester si un paramètre a une valeur donnée.
 - L'hypothèse nulle H_0 est par exemple « patient déclaré séropositif au VIH » et l'hypothèse contraire = H_1 « patient déclaré séronégatif ».
- Il y a deux façons de se tromper lors d'un test statistique :
 1. **Rejeter à tort H_0 .** risque de **première espèce** α = faux positif : accepter une hypothèse alors qu'elle était fautive (test positif à tort).
 2. **Accepter à tort H_0 :** 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 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

Define your own research question

Why? How? Exploratory research
How many? How often? When? What? Descriptive research
Where? Why? Exploratory research

Level of control
Level of access

Research question examples:
 • What are the success factors of organizational innovation?
 • How do the growth factors impact on the success of an owner?
 • How do different development technology will turn out for the developers?

How facts become theories

There is a fact
There is a description
There is an analysis
There is a pattern
There is a theory
There is a test
There is a refinement

The research process

Start empirical research

Define research question
Review literature
Create theoretical model
Design research
Perform research
Collect data
Analyze data
Draw conclusions
Disseminate results

End empirical research

Select research method

	Experiment	Case study	Survey
Purpose	Establish causal relationships, confirm theories	Develop a specific issue, explore a research phenomenon	Develop information collected from a group of people, conduct hypothesis testing
When to use	Requires high control	Requires medium control	Requires low control
Advantages	Control on what is being studied, which technology, which variables, which situations, which conditions, which procedures	Change to a research question, change the technology, change the development process, investigate the development process, investigate the situation	Technology change, investigate large number of subjects, investigate the influence factors, investigate the relationships and generalizations is easier
Disadvantages	Control on what is being studied, which technology, which variables, which situations, which conditions, which procedures	Can be investigated in natural development, Can control on what is being studied, Can control on what is being studied, Can control on what is being studied	Can be investigated in natural development, Can control on what is being studied, Can control on what is being studied

Analyze quantitative data

Statistical analysis is a mathematical science pertaining to the collection, analysis, interpretation, and presentation of data.

Descriptive statistics
 Descriptive statistics are used to summarize or describe a collection of data. This is called descriptive statistics. In military problems, the data may be ordered in a rank, and the ranks for each rank, and the mean, and the standard deviation, and the variance, and the standard error of the mean, and the standard error of the estimate, and the standard error of the regression coefficient, and the standard error of the correlation coefficient, and the standard error of the regression coefficient, and the standard error of the correlation coefficient.

Inferential statistics
 Inferential statistics are used to make inferences about a population based on a sample of data. This is called inferential statistics. Inferential statistics are used to make inferences about a population based on a sample of data. This is called inferential statistics. Inferential statistics are used to make inferences about a population based on a sample of data. This is called inferential statistics.

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 to which the research can be repeated with the same results.
 Validity threats – Does the measurement truly measure what it is intended to measure?
 Sensitivity – How much does the measurement change with the change of the conceptual variable?

Create theoretical model

Research question: How does software development methodology and team size influence developer productivity?

Theoretical model is based on research question and represents set of concepts and relationships between them!

Independent variables (color variables): OSSD, RUP, XP
 Dependent variables (start variables): Developer productivity
 Measures (observed variables): Lines of code (LOC) per developer per day

Design research

Design research is the process of designing a study to answer a research question.

Design approach: Design experiment, Design case study, Design survey

Perform research: Perform research on a population, Perform research on a specific group, Perform research on a specific individual

Collect data: Collect data with a research instrument, Collect data with a research instrument, Collect data with a research instrument

Select statistical test

Select statistical test based on the type of data and the research question.

Qualitative data: Use qualitative data analysis
 Quantitative data: Use quantitative data analysis

Qualitative data analysis: Content analysis, Grounded theory, Narrative analysis, etc.
 Quantitative data analysis: Descriptive statistics, Inferential statistics, etc.

Statistical tests

Statistical tests are used to test hypotheses about a population.

Parametric tests: t-test, ANOVA, F-test, etc.
Non-parametric tests: Mann-Whitney U-test, Wilcoxon signed-rank test, etc.

Statistical significance – the probability that an experimental result is due to chance.

Design Experiment

Design Experiment is a research method that involves manipulating one or more independent variables to observe the effect on a dependent variable.

Design Experiment
 • Analyze objectives of study (what is the study about?)
 • Determine the independent and dependent variables
 • Select the research instrument
 • Design the experiment
 • Conduct the experiment
 • Analyze the results

Qualitative vs. Quantitative analysis

Qualitative analysis is used to understand the meaning and nature of an event or situation.

Qualitative (subjective)
 • Focus on understanding the meaning and nature of an event or situation
 • Data is often unstructured and subjective
 • Analysis is often inductive and interpretive

Quantitative (objective)
 • Focus on understanding the meaning and nature of an event or situation
 • Data is often structured and objective
 • Analysis is often deductive and statistical

Design Case Study

Design Case Study is a research method that involves studying a single case or a small number of cases in depth.

Design Case Study
 • Analyze objectives of study (what is the study about?)
 • Determine the independent and dependent variables
 • Select the research instrument
 • Design the experiment
 • Conduct the experiment
 • Analyze the results

Design Survey

Design Survey is a research method that involves asking a group of people from a population about their opinions on a specific issue.

Design Survey
 • Analyze objectives of study (what is the study about?)
 • Determine the independent and dependent variables
 • Select the research instrument
 • Design the experiment
 • Conduct the experiment
 • Analyze the results

Literature used

Literature used in the research includes books, articles, and other sources.

Literature used
 • Bernd Freimut, Teade Purter, Stefan Billi, & Marcus Ciolekowski 2002, "State-of-the-Art in Empirical Studies, Virtualized Software Engineering Kompetenzzentrum."
 • Johnson, R., & Shavano, G. Research Methods in Information Systems, 2002.
 • Neumann, W. L., 2005, Social research methods - qualitative and quantitative approaches, 5th ed., Wiley.
 • Winston Teltz 1997, "Introduction to Case Study", The Qualitative Report, vol. 3, no. 2, www.elfed.org

Design Experiment

Design Experiment is a research method that involves manipulating one or more independent variables to observe the effect on a dependent variable.

Design Experiment
 • Analyze objectives of study (what is the study about?)
 • Determine the independent and dependent variables
 • Select the research instrument
 • Design the experiment
 • Conduct the experiment
 • Analyze the results

Design Case Study

Design Case Study is a research method that involves studying a single case or a small number of cases in depth.

Design Case Study
 • Analyze objectives of study (what is the study about?)
 • Determine the independent and dependent variables
 • Select the research instrument
 • Design the experiment
 • Conduct the experiment
 • Analyze the results

Design Survey

Design Survey is a research method that involves asking a group of people from a population about their opinions on a specific issue.

Design Survey
 • Analyze objectives of study (what is the study about?)
 • Determine the independent and dependent variables
 • Select the research instrument
 • Design the experiment
 • Conduct the experiment
 • Analyze the results

About the Research Methods Poster

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

About the Research Methods Poster
 • This poster is licensed under Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License.
 • Author: Gregor Polzella
 • Email: info@poster.net
 • University of Maribor, Faculty of Electrical Engineering and Computer Science, Institute of Informatics
 • Poster version: 6.6 (DRAFT)
 • <http://researchmethods.poster.net>

Questions ?