

Evaluation de l'interaction

Michel Beaudouin-Lafon

Université Paris-Saclay

mbl@lri.fr

<http://ex-situ.lri.fr>

Plan

Du moins formel vers le plus formel

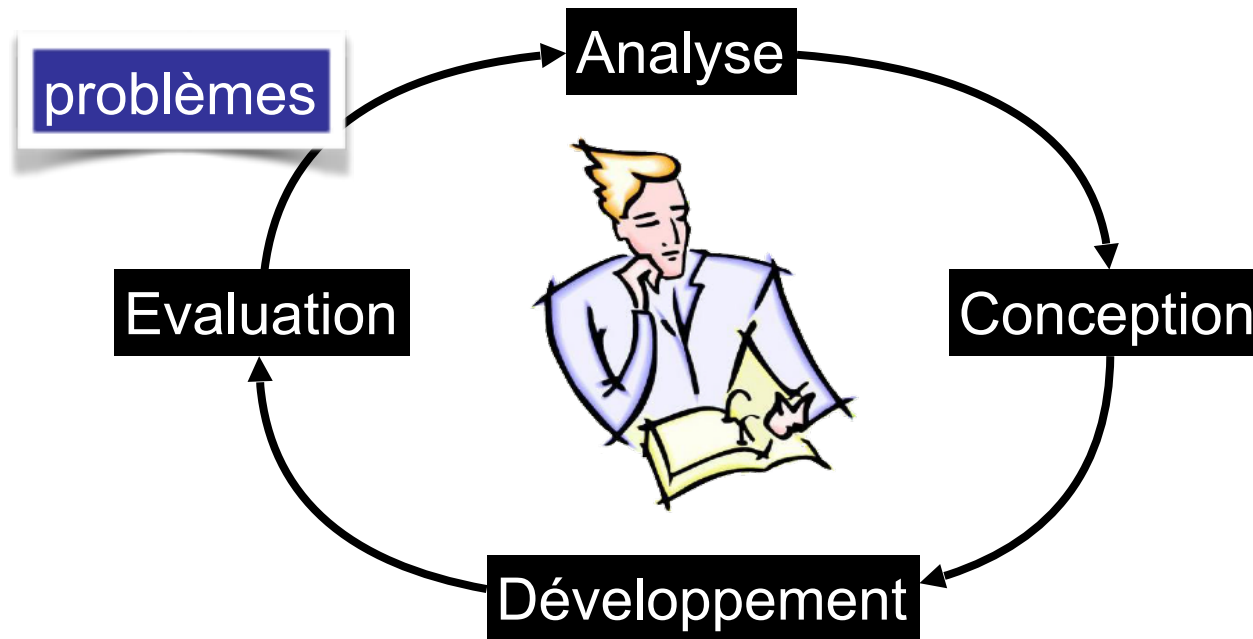
« Walkthrough » - revue de conception

GOMS KLM

Test A/B

Expérimentation contrôlée

Cycle de conception d'un système interactif



Walkthrough (revue de conception)

Walkthrough (revue de conception)

Evaluation pas à pas d'une interface pour identifier des problèmes

But : identifier un maximum de problèmes

La revue peut concerner :

- Une table d'interaction

- Un prototype non fonctionnel

- Un logiciel fonctionnel

- Le code du logiciel

Types de problèmes recherchés :

- Bugs

- Performance

- Utilisabilité : simplicité, cohérence, etc.

- Fonctionnalité manquante

Walkthrough : procédure

Petit groupe (4-8 personnes)

Définir l'objet évalué (design, prototype, logiciel, ...)

Etablir une liste de procédures à passer en revue

Centrée sur l'usage : tâches à effectuer ou scénarios d'usage

Centrée sur les fonctionnalités : liste de fonctions

Etablir une liste de critères d'évaluation

Pour chaque procédure :

Une personne la présente, puis chacun fait part des problèmes

Pas de jugement : on note tout

Pas de discussion : on ne cherche pas de solution (ni d'excuse)

On peut passer les critères un par un,

ou bien chaque personne est en charge d'un ou deux critères,

ou bien tout le monde évalue tous les critères

Walkthrough : recommandations

Types de commentaires :

Se concentrer sur l'**objet évalué**, pas sur les auteurs

Faire des commentaires **constructifs**, pas destructifs

Faire des commentaires **spécifiques**, pas généraux

Se concentrer sur les **problèmes**,
puis les **questions** et enfin les **suggestions**

Exemples :

“Le text est trop petit pour être lisible”

“On ne voit pas comment changer ce réglage”

“Il faut quatre étapes pour cette opération”

Trouver le maximum de problèmes,
ne pas discuter des solutions possibles, ni des problèmes

Walkthrough : après la session

Rassembler les problèmes par catégories

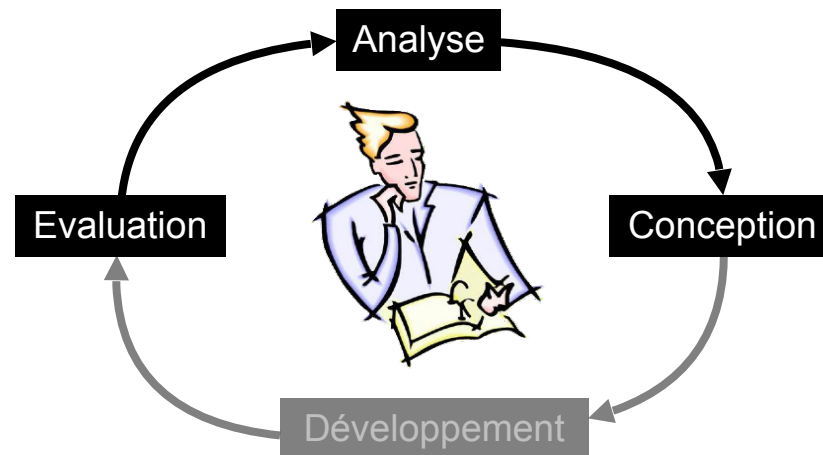
Selon les tâches / fonctionnalités examinées

Selon les critères utilisés

Repasser les problèmes en revue

Les classer par importance (du point de vue de l'utilisateur)

Ces listes classées vont servir à la prochaine étape d'analyse/conception



GOMS KLM

GOMS KLM : évaluation prédictive

GOMS = **G**oals **O**perators **M**ethods **S**election rules

Ensemble de modèles pour décrire et évaluer l'interaction

Buts : ce que veut faire l'utilisateur

Opérateurs : actions de l'utilisateur reconnues par le système

Méthodes : séquences de sous-buts et d'opérateurs pour atteindre un but

Règles de sélection : règles de l'utilisateur pour choisir une méthode

KLM = Keystroke-Level Model

Description détaillée des actions de l'utilisateur

À chaque opérateur est associé un temps d'exécution

Evaluation prédictive du temps pour effectuer une tâche

Comparaison de différentes interfaces ou différentes méthodes

GOMS KLM : exemple

But :

effacer un fichier

Opérateurs :

P : pointer = déplacer le curseur sur une cible (1,1s)

B : appuyer ou relâche le bouton de la souris (0,1s)

K : taper au clavier (0,5s pour une lettre au hasard,
0,1s à 0,3s pour une personne qui tape couramment du texte)

H : déplacer la main entre le clavier et la souris (0,4s)

M : activité mentale de l'utilisateur (1,35s)

Méthodes :

Cliquer-tirer l'icône du fichier vers l'icône de la poubelle

Sélectionner l'icône du fichier et taper Command-Delete

GOMS KLM : exemple

Evaluation de la méthode 1 :

Identifier l'icône du fichier : M (1,35s)

Pointer l'icône : P (1,1s)

Appuyer sur le bouton de la souris : B (0,1s)

Déplacer l'icône jusqu'à la poubelle : P (1,1s)

Relâcher le bouton de la souris : B (0,1s)

Revenir à la position initiale : P (1,1s)

Total = 4,85s

Evaluation de la méthode 2 :

Identifier l'icône du fichier : M (1,35s)

Pointer l'icône et la sélectionner (clic) : P B B (1,1s + 0,1s + 0,1s)

Déplacer la main de la souris vers le clavier : H (0,4s)

Taper Command-Delete : K K (0,5s + 0,5s)

Ramener la main sur la souris : H (0,4s)

Total = 4,45s

GOMS KLM

Les prédictions de KLM ne sont pas très précises :

Difficulté à placer l'opérateur M

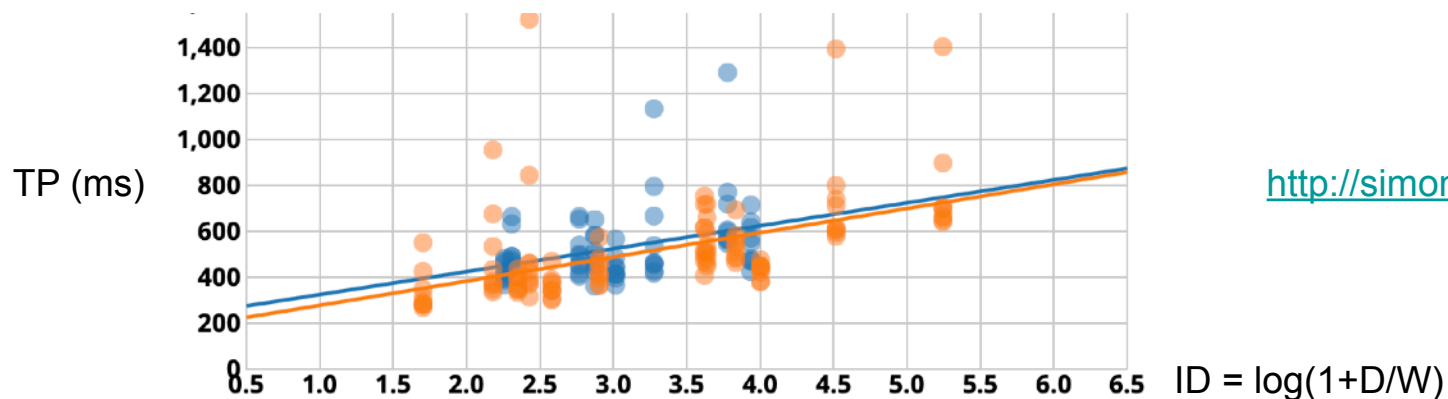
dépend de l'expertise de l'utilisateur

Evaluation grossière du temps de pointage ($P = 1,1s$)

peut être affiné avec la Loi de Fitts

$$TP = 0,2 + 0,1 \log_2(1 + D/W)$$

où D est la distance à la cible et W sa taille ("width")



Les comparaisons sont cependant en général valides

Test A/B

Test A/B

Comparer deux versions d'une interface

Utilisé à grande échelle par Google, Microsoft, Facebook

Exemple : la couleur de fond des publicités affecte le nombre de clics

Principe :

Réaliser deux versions d'une interface (souvent un site web)

Etablir une mesure de performance binaire

cliquer ou non sur une publicité

convertir ou non une visite en achat

etc.

Test A/B : un exemple

Prenons l'exemple de l'effet de la couleur sur le taux de clic d'une publicité

Publicité

Publicité

1000 utilisateurs voient la publicité sur fond jaune

12 cliquent dessus, soit 1,2%

1000 utilisateurs voient la publicité sur fond bleu

15 cliquent dessus, soit 1,5%

La différence est-elle suffisante pour dire que le fond bleu est plus efficace ?

=> Test d'hypothèse nulle

Hypothèse nulle = il n'y a pas de différence

On essaie de montrer que cette hypothèse est fausse

Test d'hypothèse nulle

“Null Hypothesis Significance Testing” ou NHST

On teste un échantillon de la population,
et on veut inférer un résultat sur l'ensemble de la population

On veut savoir si la différence observée entre deux conditions est **statistiquement significative**, c'est-à-dire si l'on a une bonne chance qu'elle ne soit pas l'effet du hasard (dans le choix de l'échantillon)

En pratique on se donne un seuil de confiance de 95% :
on effectue un test statistique, destiné à rejeter l'hypothèse nulle :
s'il est positif, on est sûr à 95% que la différence existe bien dans la population
s'il est négatif, on ne peut rien dire : il y a peut-être une différence, mais le test n'a pas permis de la déceler

Test du Chi-2 : observations

Observations	Jaune	Bleu	Total
Clic	12	15	
Pas de clic			
Total	1000	1000	

Test du Chi-2 : observations

Observations	Jaune	Bleu	Total
Clic	12	15	27
Pas de clic	988	985	1973
Total	1000	1000	2000

Test du Chi-2 : hypothèse nulle

Observations	Jaune	Bleu	Total
Clic	12	15	27
Pas de clic	988	985	1973
Total	1000	1000	2000
Hypothèse nulle	Jaune	Bleu	Total
Clic			27
Pas de clic			1973
Total	1000	1000	2000

Test du Chi-2 : hypothèse nulle

Observations	Jaune	Bleu	Total
Clic	12	15	27
Pas de clic	988	985	1973
Total	1000	1000	2000
Hypothèse nulle	Jaune	Bleu	Total
Clic	13,5		27
Pas de clic			1973
Total	1000	1000	2000

Calculer les proportions dans l'hypothèse où il n'y a pas de différence

$$\text{Clic bleu} = \text{total clic} * \text{total bleu} / \text{total}$$

Test du Chi-2 : hypothèse nulle

Observations	Jaune	Bleu	Total
Clic	12	15	27
Pas de clic	988	985	1973
Total	1000	1000	2000
Hypothèse nulle	Jaune	Bleu	Total
Clic	13,5	13,5	27
Pas de clic	986,5	986,5	1973
Total	1000	1000	2000

Calculer les proportions dans l'hypothèse où il n'y a pas de différence

Clic jaune = total clic * total jaune / total

Clic bleu = total clic * total bleu / total

Pas de clic jaune = total pas de clic * total jaune / total

Pas de clic bleu = total pas de clic * total bleu / total

Test du Chi-2 : calcul du Chi-2

Observations	Jaune	Bleu	Total
Clic	12	15	27
Pas de clic	988	985	1973
Total	1000	1000	2000
Hypothèse nulle	Jaune	Bleu	Total
Clic	13,5	13,5	27
Pas de clic	986,5	986,5	1973
Total	1000	1000	2000
Chi-2	Jaune	Bleu	Total
Clic	0,167		
Pas de clic			
Total			

$$\text{Chi}2 = (\text{valeur observée} - \text{valeur attendue})^2 / \text{valeur attendue}$$

Test du Chi-2 : résultat

Observations	Jaune	Bleu	Total
Clic	12	15	27
Pas de clic	988	985	1973
Total	1000	1000	2000
Hypothèse nulle	Jaune	Bleu	Total
Clic	13,5	13,5	27
Pas de clic	986,5	986,5	1973
Total	1000	1000	2000
Chi-2	Jaune	Bleu	Total
Clic	0,167	0,167	0,333
Pas de clic	0,002	0,002	0,005
Total	0,169	0,169	0,338

Test : si $\text{Chi-2} > 3,84$ alors 95% de chance que la différence soit significative

Test du Chi-2 : résultat avec plus grand échantillon

Observations	Jaune	Bleu	Total
Clic	120	150	270
Pas de clic	9880	9850	19730
Total	10000	10000	20000
Hypothèse nulle	Jaune	Bleu	Total
Clic	135	135	270
Pas de clic	9865	9865	19730
Total	10000	10000	20000
Chi-2	Jaune	Bleu	Total
Clic	1,67	1,67	3,33
Pas de clic	0,02	0,02	0,05
Total	1,69	1,69	3,38

Echantillon **10 fois plus grand** (même taux de clic) : résultat **non significatif**

Test du Chi-2 : résultat avec très grand échantillon

Observations	Jaune	Bleu	Total
Clic	1200	1500	2700
Pas de clic	98800	98500	197300
Total	100000	100000	200000
Hypothèse nulle	Jaune	Bleu	Total
Clic	1350	1350	2700
Pas de clic	98650	98650	197300
Total	100000	100000	200000
Chi-2	Jaune	Bleu	Total
Clic	16,67	16,67	33,33
Pas de clic	0,23	0,23	0,46
Total	16,89	16,89	33,79

Echantillon **100** fois plus grand (même taux de clic) : résultat **significatif**

Test A/B : estimer le nombre d'observations

Lorsque les différences sont faibles, il faut un grand nombre d'observations pour obtenir un résultat statistiquement significatif

Comment estimer le nombre d'observations nécessaires ?

=> estimer la différence minimale que l'on souhaite observer
par exemple : une augmentation de 25% du taux de clic

=> Utiliser un calculateur, par exemple G*Power, ou
<https://www.abtasty.com/sample-size-calculator/>

Classic Sample Size Calculator

FAQ

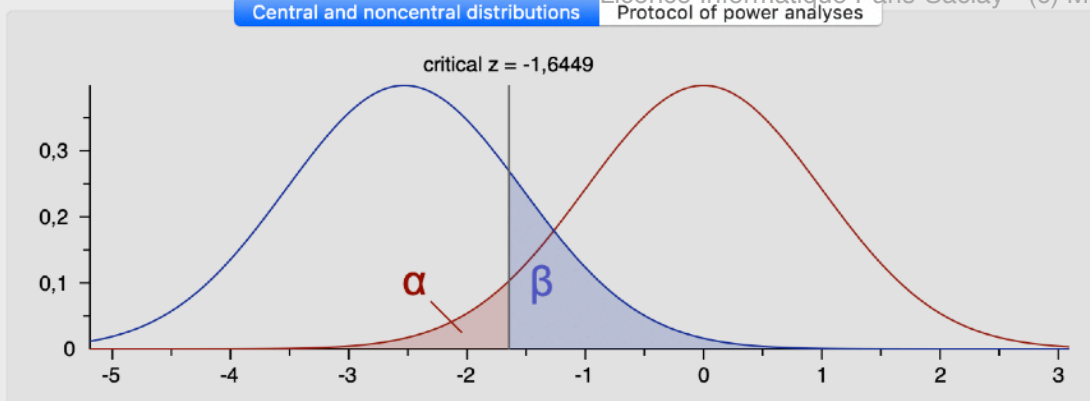
Calculate the minimum sample size as well as the ideal duration of your A/B tests based on your audience, conversions and other factors like the Minimum Detectable Effect.

How many users do you need?

Conversion Rate [?]	<input type="text" value="1.2"/> %
Minimum Detectable Effect [?]	<input type="text" value="25"/> %
Statistical Significance [?]	<input type="text" value="95"/> %

Required number of tested visitors per variation

21,428



Test family

Exact

Statistical test

Proportions: Inequality, two independent groups (Fisher's exact test)

Type of power analysis

A priori: Compute required sample size - given α , power, and effect size

Input parameters

Determine

Tail(s)

One

Proportion p_1 0,012Proportion p_2 0,015 α err prob 0,05Power ($1-\beta$ err prob) 0,8Allocation ratio N_2/N_1 1

Output parameters

Sample size group 1 18958

Sample size group 2 18958

Total sample size 37916

Actual power 0,8000183

Actual α 0,0500000

Options

X-Y plot for a range of values

Calculate

Test A/B : conclusion

Test le plus simple :

2 conditions, en général interface actuelle et amélioration espérée

Mesure binaire : clic ou pas clic

Tests plus avancés :

Plus de 2 conditions, pour tester plusieurs alternatives

Mesures ordinales :

exemple : réponses à un questionnaire de type

pas du tout d'accord, plutôt pas d'accord, neutre, plutôt d'accord, tout à fait d'accord

Mesures numériques :

exemples : temps passé sur une page Web

=> différents types de tests statistiques

Nécessitent souvent un grand nombre d'observations

Expérimentation contrôlée

Expérimentation contrôlée

Test effectué en laboratoire, avec un nombre faible de participants (~30)

Même principe de test d'hypothèse nulle que le test A/B

MAIS

- Souvent avec une combinaison de variables

- Souvent avec un plan d'expérience intra-participants :

 - chaque participant teste les différentes possibilités

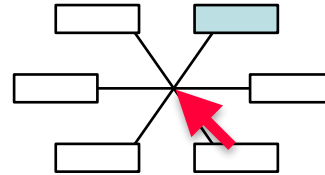
 - => permet des tests plus robustes avec moins de participants

- Analyses statistiques plus approfondies

Exemple : comparer menu linéaire et menu radial

Exemple : comparaison de menus

Variable principale : type de menu



Autres **variables** :

taille du menu : 4, 6, 8, 12 items

item à sélectionner : 1er, 2nd, ..., dernier

Hypothèse nulle :

Il n'y a pas de différence entre les deux types de menus quels que soient le nombre d'items et le rang de l'item sélectionné

Exemple : comparaison de menus

24 conditions : 2 types de menus X 4 tailles X 3 rangs (début/milieu/fin)
Chaque participant verra chaque condition 5 fois => $5 * 24 = 120$ essais

Pour chaque essai on **mesure**

le temps d'exécution

le succès / échec

Pour éviter les effets d'ordre, chaque participant verra les conditions dans un ordre différent => au minimum 24 ordres, donc 24 participants

Parfois on complète les mesures objectives avec un questionnaire

Exemple : comparaison de menus

On collecte une table de données

Participant	Menu	Taille	Item	Essai	Temps	Succès
P1	Radial	4	début	1	0,45	oui
P1	Radial	4	début	2	0,34	oui
P1	Radial	4	fin	1	0,20	non
P1	Radial	4	milieu	1	0,39	oui
P1	Radial	4	début	3	0,29	oui
P1	Linéaire	12	milieu	1	1,23	oui
P1	Linéaire	12	milieu	2	1,18	non
P1	Linéaire	12	fin	3	2,11	non
P1	Linéaire	12	début	1	0,98	oui
P1	Linéaire	12	début	1	1,03	oui
...

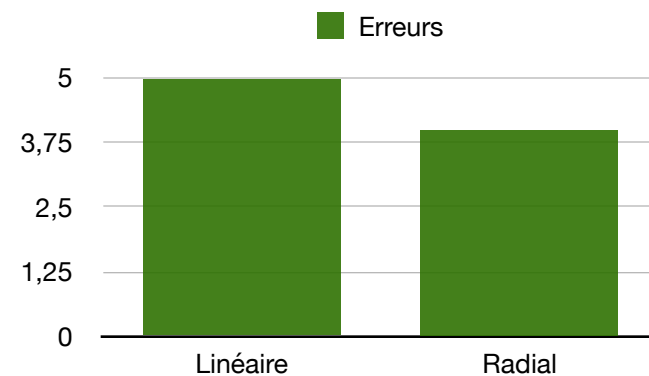
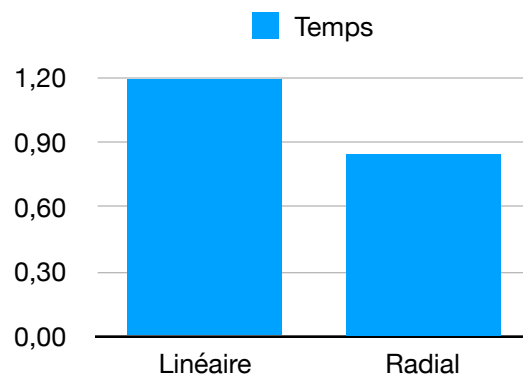
Exemple : comparaison de menus

Analyser les mesures en fonctions de variables

Analyse de variance (ANOVA - ANalysis Of VAriance)

Permet de détecter des effets principaux :

le type de menu affecte le temps de sélection et le taux d'erreur



Exemple : comparaison de menus

Analyser les mesures en fonctions de variables

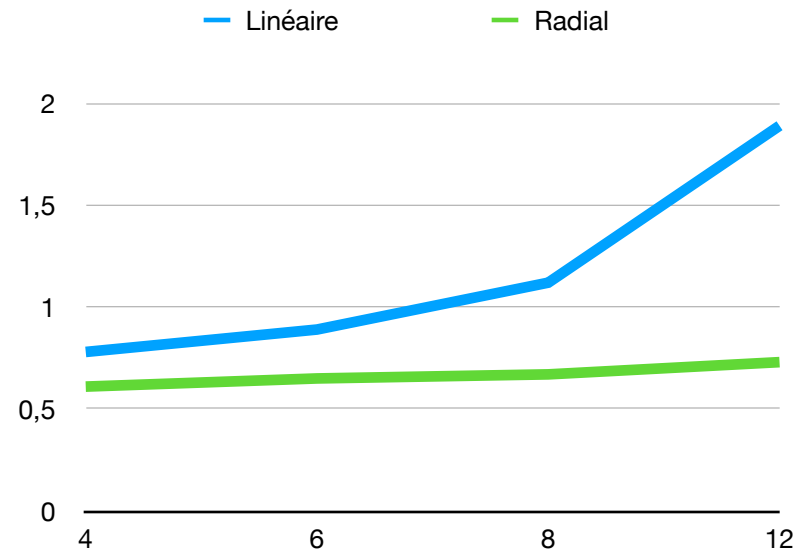
Analyse de variance (ANOVA - ANalysis Of VAriance)

Permet de détecter des effets principaux :

le type de menu affecte le temps de sélection et le taux d'erreur

Et des interactions :

le temps moyen de sélection augmente avec la taille du menu
pour les menus linéaires mais pas les menus radiaux



Conclusion

L'évaluation des interfaces fait appel à des méthodes empiriques :

- Analyse qualitative (walkthrough)

- Analyse quantitative (test A/B, expérimentations contrôlées)

Les méthodes prédictives (GOMS KLM) sont utiles mais insuffisantes :

- Prédit uniquement les temps d'exécution

- Prédictions approximatives

- Fastidieux à utiliser systématiquement

L'évaluation est indispensable dans le cycle de conception itératif

Complément : t-test

Un peu de statistiques : lancer de pièce

Supposons que l'on a une pièce que l'on tire à pile ou face 100 fois
Combien de fois va-t-on tirer pile ?

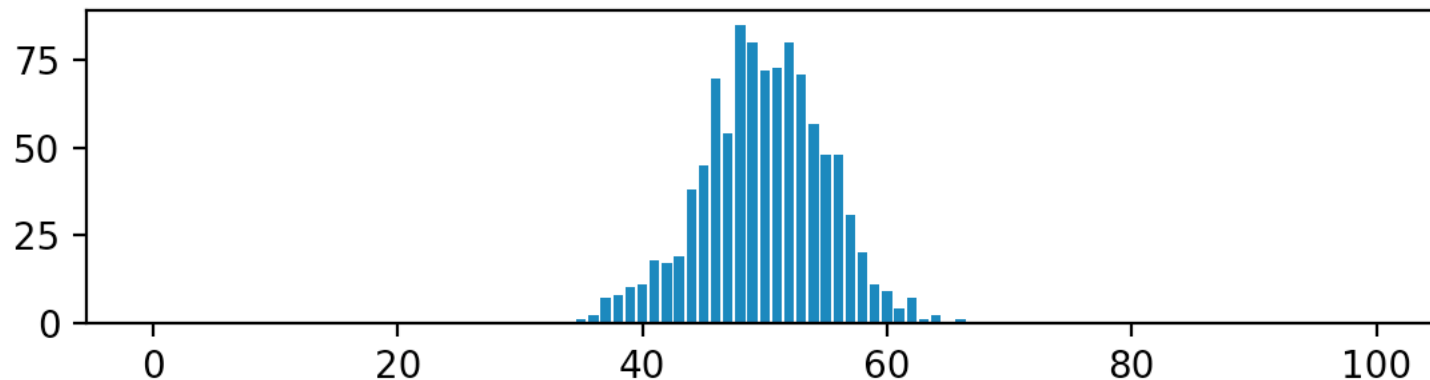
Cela dépend...

Les chances d'avoir 100 pile et 0 face sont très, très, très faibles

Les chances d'avoir 50 pile et 50 face sont assez élevées

Mais on peut aussi avoir 51 pile et 49 face, etc.

Si l'on fait un grand nombre de séries de 100 lancers et que l'on compte le nombre de pile, on obtient une courbe comme celle-ci (1000 séries) :



Un peu de statistiques : lancer de pièce

Supposons que l'on a une pièce que l'on tire à pile ou face 100 fois
Combien de fois va-t-on tirer pile ?

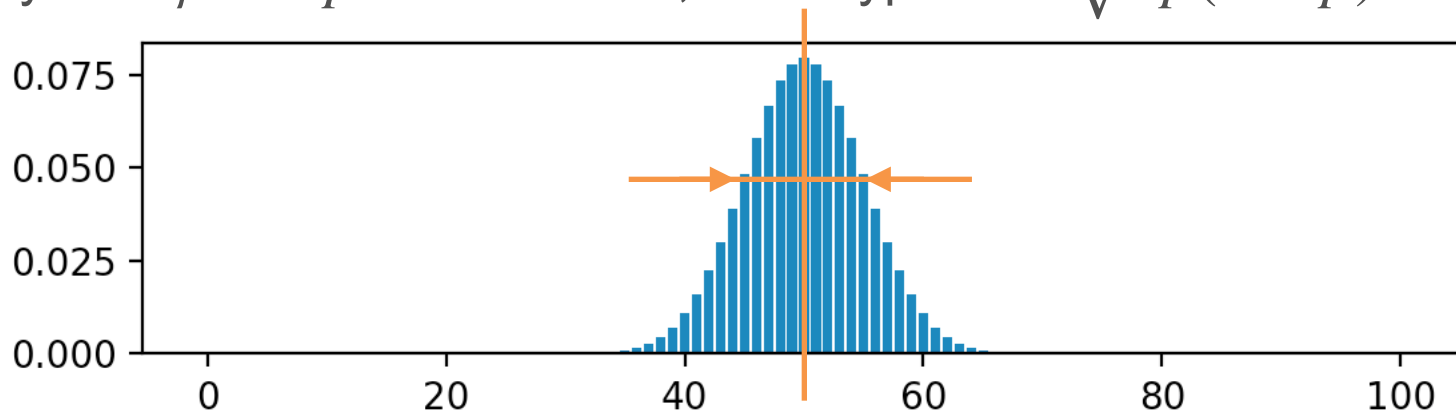
Cela dépend...

Les chances d'avoir 100 pile et 0 face sont très, très, très faibles

Les chances d'avoir 50 pile et 50 face sont assez élevées

Mais on peut aussi avoir 51 pile et 49 face, etc.

Cette distribution est modélisée par une loi binomiale de probabilité $p = 0.5$
moyenne $\mu = np = n/2 = 50$, écart-type $\sigma = \sqrt{np(1-p)} = \sqrt{n/2} = 5$

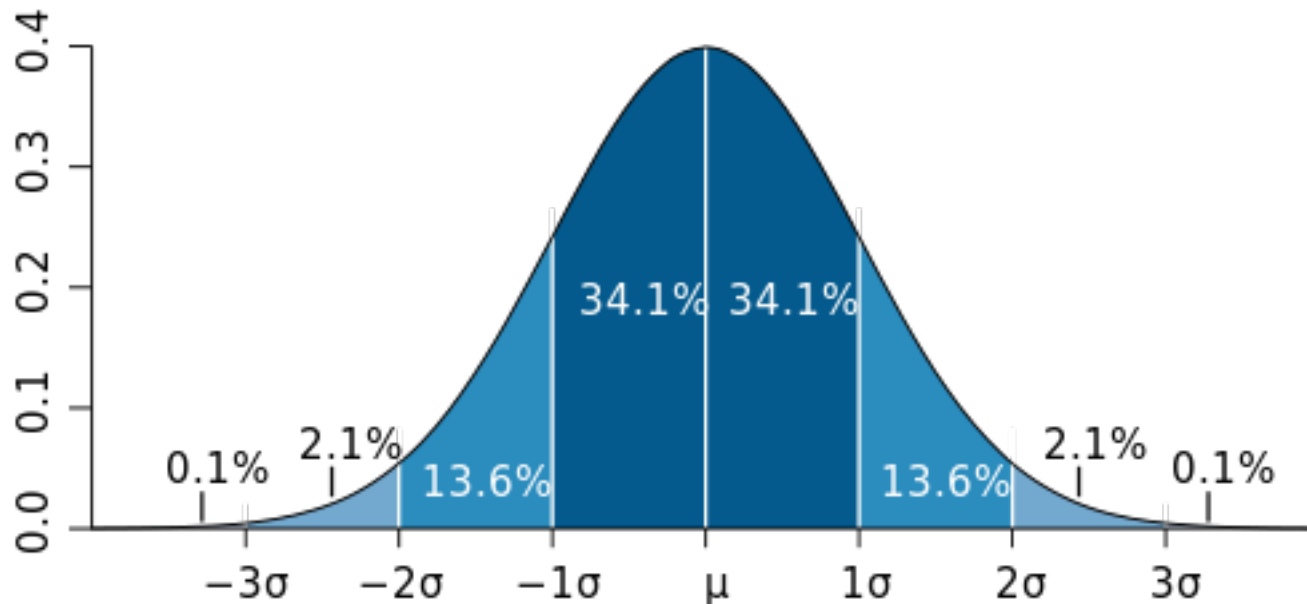


Un peu de statistiques : la loi normale

Pour un grand nombre de lancers, cette courbe approche la **loi normale**

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ = moyenne, σ = écart-type

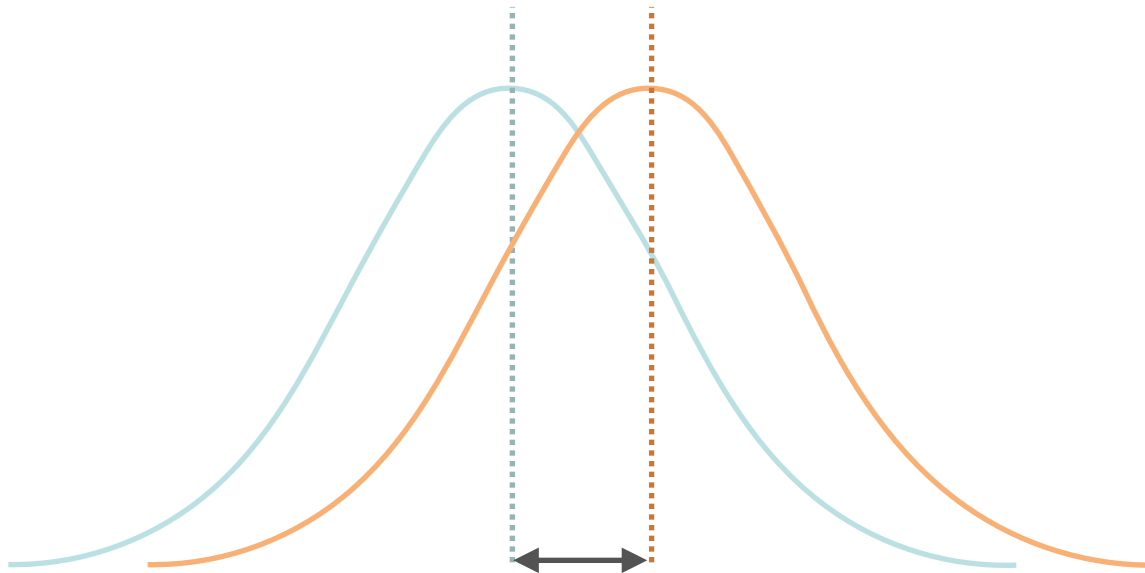


Un peu de statistiques : détecter une fausse pièce

Comment savoir si la pièce que l'on lance est bien équilibrée ?

Comparer les distributions

Quelle différence est suffisante pour conclure ?



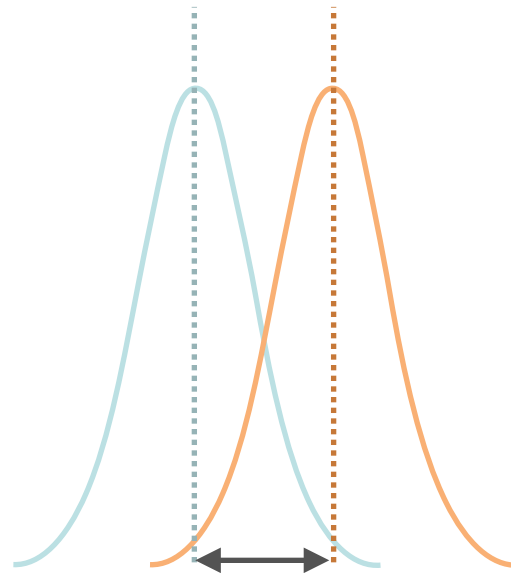
Cela dépend de la différence entre les moyennes

Un peu de statistiques : détecter une fausse pièce

Comment savoir si la pièce que l'on lance est bien équilibrée ?

Comparer les distributions

Quelle différence est suffisante pour conclure ?



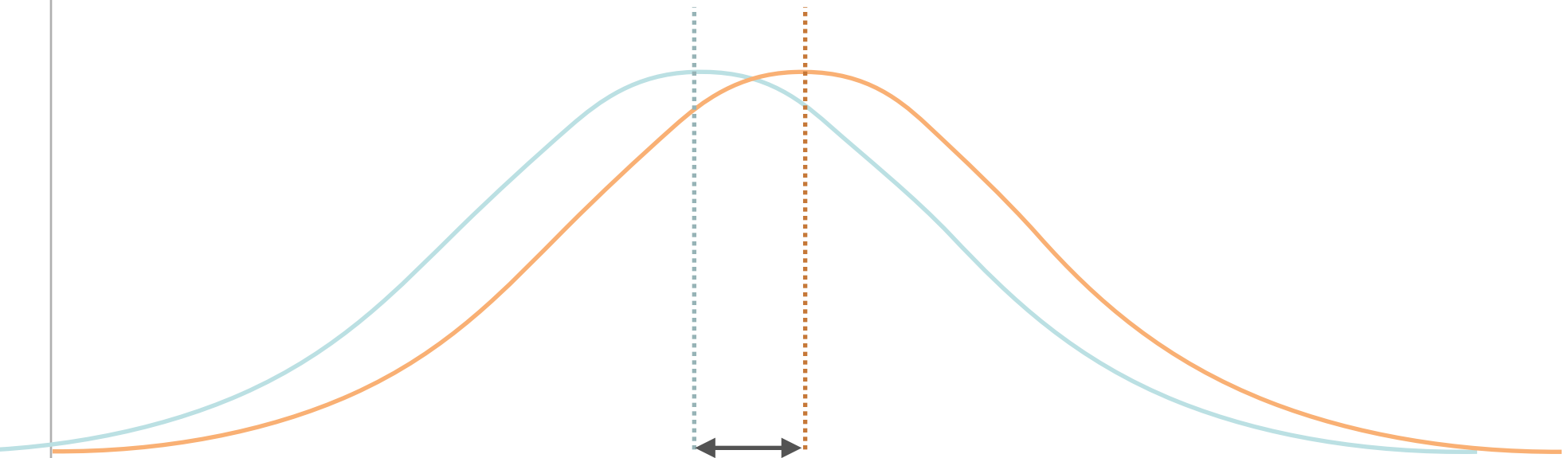
Mais cela dépend aussi de l'écart-type

Un peu de statistiques : détecter une fausse pièce

Comment savoir si la pièce que l'on lance est bien équilibrée ?

Comparer les distributions

Quelle différence est suffisante pour conclure ?



Mais cela dépend aussi de l'écart-type

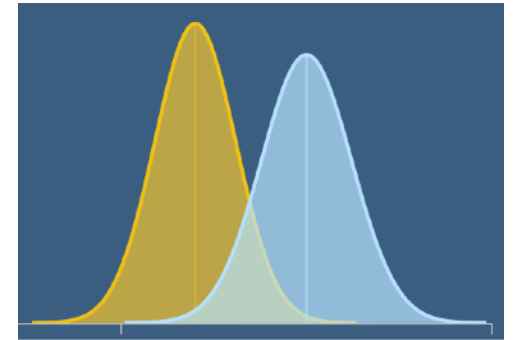
Un peu de statistiques : le t-test

On calcule le rapport signal / bruit, appelé t :

signal = différence entre les moyennes (ce qui nous intéresse)

bruit = variabilité (ce qui nous embête)

$$t = \frac{\mu_2 - \mu_1}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}$$



Comment interpréter t ?

Dans le cas qui nous intéresse, si $t > 2$ ou $t < -2$,

alors on est sûr à environ 95% que la différence est « réelle »

(On dit qu'elle est statistiquement significative)

Plus les écarts-types sont grands (plus les courbes sont plates),
plus la différence des moyennes doit être importante

Application

Collecter un ensemble de mesures d'une valeur numérique
exemple : temps passé à visiter une page, ou à effectuer une tâche

On dispose de n_1 mesures dans la condition de référence
et de n_2 mesures dans la condition expérimentale

Analyse des données :

On vérifie que les données sont distribuées selon une loi normale
(test de Kolmogorov-Smirnov)

On calcule la moyenne et l'écart-type de chaque série de mesures

$$\mu = \frac{1}{n} \sum m_i \quad \sigma^2 = \frac{1}{n-1} \sum (m_i - \mu)^2$$

On calcule la valeur de t

On conclut sur la validité (on non) de l'hypothèse nulle