Probabilités, théorie de l'information et QCM bayésiens

Rémi Peyre [*]

Les Probas du vendredi
21 novembre 2025

[*]. Institut Élie Cartan de Lorraine (Nancy)

Plan

Plan

Approche combinatoire

Probabilité vue comme proportion.

- Univers Ω composé d'éventualités ω dont aucune n'est privilégiée ;
- $\mathbb{P}(A) \coloneqq |A| \,/\, |\Omega|$.

Approche combinatoire

Probabilité vue comme proportion.

- Univers Ω composé d'éventualités ω dont aucune n'est privilégiée ;
- $\mathbb{P}(A) \coloneqq |A| \,/\, |\Omega|$.

⊕

- Facile à comprendre.
- Les règles de calcul s'expliquent naturellement.

Approche combinatoire

Probabilité vue comme proportion.

- Univers Ω composé d'éventualités ω dont aucune n'est privilégiée ;
- $\mathbb{P}(A) \coloneqq |A| \,/\, |\Omega|$.

⊕
- Facile à comprendre.
- Les règles de calcul s'expliquent naturellement.

⊖
- Passage au cadre continu bancal.
- Pas facile à connecter à l'usage quotidien : les éventualités, ce seraient des variantes équiprobables du futur !? Mais ici, que signifierait « équiprobables » au juste ?...

Approche combinatoire

Probabilité vue comme proportion.

- Univers Ω composé d'éventualités ω dont aucune n'est privilégiée ;
- $\mathbb{P}(A) \coloneqq |A| \,/\, |\Omega|$.

⊕
- Facile à comprendre.
- Les règles de calcul s'expliquent naturellement.

⊖
- Passage au cadre continu bancal.
- Pas facile à connecter à l'usage quotidien : les éventualités, ce seraient des variantes équiprobables du futur !? Mais ici, que signifierait « équiprobables » au juste ?...

$\rightsquigarrow$ Intérêt d'une approche par théorie de la mesure.

Approche combinatoire

Probabilité vue comme proportion.

- Univers Ω composé d'éventualités ω dont aucune n'est privilégiée ;
- $\mathbb{P}(A) \coloneqq |A| \,/\, |\Omega|$.

⊕
- Facile à comprendre.
- Les règles de calcul s'expliquent naturellement.

⊖
- Passage au cadre continu bancal.
- Pas facile à connecter à l'usage quotidien : les éventualités, ce seraient des variantes équiprobables du futur !? Mais ici, que signifierait « équiprobables » au juste ?...

$\rightsquigarrow$ Intérêt d'une approche par théorie de la mesure.

Mais quelle interprétation de « probabilité » dans un tel cadre ?

Approche fréquentiste

Probabilité vue comme fréquence asymptotique.
Pour $A \subseteq \Omega$, on peut évaluer $\mathbb{P}(A)$ en répétant l'expérience indépendamment à l'infini :

$$\mathbb{P}^{\otimes\mathbb{N}}\left(\left\{(\omega_n)_{n\in\mathbb{N}} \;\middle|\; \frac{\left|\{i \in [0,n[\mid \omega_i \in A\}\right|}{n} \overset{n\to\infty}{\to} \mathbb{P}(A)\right\}\right) = 1. \quad \text{(LGN)}$$

Approche fréquentiste

Probabilité vue comme fréquence asymptotique.
Pour $A \subseteq \Omega$, on peut évaluer $\mathbb{P}(A)$ en répétant l'expérience indépendamment à l'infini :

$$\mathbb{P}^{\otimes\mathbb{N}}\left(\left\{(\omega_n)_{n\in\mathbb{N}} \;\middle|\; \frac{\left|\{i \in [0,n[\mid \omega_i \in A\}\right|}{n} \overset{n\to\infty}{\to} \mathbb{P}(A)\right\}\right) = 1. \quad \text{(LGN)}$$

⊕

- La probabilité est une notion objective.
- Les règles de calcul s'expliquent naturellement.

Approche fréquentiste

Probabilité vue comme fréquence asymptotique.
Pour $A \subseteq \Omega$, on peut évaluer $\mathbb{P}(A)$ en répétant l'expérience indépendamment à l'infini :

$$\mathbb{P}^{\otimes\mathbb{N}}\left(\left\{(\omega_n)_{n\in\mathbb{N}} \;\middle|\; \frac{\left|\{i \in [0,n[\mid \omega_i \in A\}\right|}{n} \overset{n\to\infty}{\to} \mathbb{P}(A)\right\}\right) = 1. \quad \text{(LGN)}$$

⊕
- La probabilité est une notion objective.
- Les règles de calcul s'expliquent naturellement.

⊖
- Cohérence interne peu intuitive : la LGN forte est un théorème assez difficile !
- Besoin d'une théorie σ-additive de la mesure, plus délicate à défendre.
- Évaluation de $\mathbb{P}(A)$ pas implémentable en pratique, surtout pour des occurrences uniques !
- Conditionnement par rapport à un évènement de probabilité nulle ?!

Approche épistémique

Une probabilité exprime un niveau de croyance (alias « crédence ») en quelque chose : pour quels enjeux serais-je prêt à "parier" sur le fait que l'évènement se réalise ?

Approche épistémique

Une probabilité exprime un niveau de croyance (alias « crédence ») en quelque chose : pour quels enjeux serais-je prêt à "parier" sur le fait que l'évènement se réalise ?

Étalonnage à partir de l'approche combinatoire : si je sais juste que $\omega \in \Omega$, sans privilégier aucune valeur dans Ω, alors ma crédence en $\{\omega \in A\}$ est $|A| \,/\, |\Omega|$.

Approche épistémique

Une probabilité exprime un niveau de croyance (alias « crédence ») en quelque chose : pour quels enjeux serais-je prêt à "parier" sur le fait que l'évènement se réalise ?

Étalonnage à partir de l'approche combinatoire : si je sais juste que $\omega \in \Omega$, sans privilégier aucune valeur dans Ω, alors ma crédence en $\{\omega \in A\}$ est $|A| \,/\, |\Omega|$.

⊕

- C'est l'usage pratique du mot « probabilité ».
- Paradigme bayésien inhérent au concept : si l'information change, les probabilités changent aussi !
- Approche naturellement continue, mais sans besoin de σ-additivité.

Approche épistémique

Une probabilité exprime un niveau de croyance (alias « crédence ») en quelque chose : pour quels enjeux serais-je prêt à "parier" sur le fait que l'évènement se réalise ?

Étalonnage à partir de l'approche combinatoire : si je sais juste que $\omega \in \Omega$, sans privilégier aucune valeur dans Ω, alors ma crédence en $\{\omega \in A\}$ est $|A| \,/\, |\Omega|$.

⊕
- C'est l'usage pratique du mot « probabilité ».
- Paradigme bayésien inhérent au concept : si l'information change, les probabilités changent aussi !
- Approche naturellement continue, mais sans besoin de σ-additivité.

⊖
- Les probabilités sont subjectives, mais pas arbitraires : c'est subtil... (Confer priore en stat. bayésienne).
- Explication des règles de calcul guère évidente !

L'approche épistémique est *forcément* probabiliste

Soient :

- $\mathscr{C}r$ ensemble des crédences, totalement ordonné de ✗ à ✓ ;
- $\mathscr{F}$ algèbre booléenne d'évènements, de minimum et maximum notés resp. `FAUX` et `VRAI`) ;
- Notation $cr_A(B)$ pour désigner la crédence en B sachant A.

L'approche épistémique est *forcément* probabiliste

Soient :

- $\mathscr{C}r$ ensemble des crédences, totalement ordonné de ✗ à ✓ ;
- $\mathscr{F}$ algèbre booléenne d'évènements, de minimum et maximum notés resp. `FAUX` et `VRAI`) ;
- Notation $cr_A(B)$ pour désigner la crédence en B sachant A.

Supposons :

- $B \mapsto cr_A(B)$ croissante ; $cr_A(\texttt{FAUX}) =$ ✗ ; $cr_A(\texttt{VRAI}) =$ ✓.
- Fonction $compl : \mathscr{C}r \to \mathscr{C}r$, décroissante, telle que

$$cr_A(\neg B) = compl(cr_A(B)).$$

- Fonction $chaine : \mathscr{C}r \times \mathscr{C}r \to \mathscr{C}r$, croissante en chaque variable, telle que

$$cr_A(B \wedge C) = chaine(cr_A(B), cr_{A \wedge B}(C)).$$

- "Principe de la chose certaine" :

$$(cr_{A \wedge B}(C) \geqslant \gamma \text{ et } cr_{A \wedge (\neg B)}(C) \geqslant \gamma) \implies cr_A(C) \geqslant \gamma.$$

L'approche épistémique est *forcément* probabiliste

Soient :

- $\mathscr{C}r$ ensemble des crédences, totalement ordonné de ✗ à ✓ ;
- $\mathscr{F}$ algèbre booléenne d'évènements, de minimum et maximum notés resp. `FAUX` et `VRAI`) ;
- Notation $cr_A(B)$ pour désigner la crédence en B sachant A.

Supposons :

- $B \mapsto cr_A(B)$ croissante ; $cr_A(\texttt{FAUX}) =$ ✗ ; $cr_A(\texttt{VRAI}) =$ ✓.
- Fonction $compl : \mathscr{C}r \to \mathscr{C}r$, décroissante, telle que

$$cr_A(\neg B) = compl(cr_A(B)).$$

- Fonction $chaine : \mathscr{C}r \times \mathscr{C}r \to \mathscr{C}r$, croissante en chaque variable, telle que

$$cr_A(B \wedge C) = chaine(cr_A(B), cr_{A \wedge B}(C)).$$

- "Principe de la chose certaine" :

$$(cr_{A \wedge B}(C) \geqslant \gamma \text{ et } cr_{A \wedge (\neg B)}(C) \geqslant \gamma) \implies cr_A(C) \geqslant \gamma.$$

- $\forall A \in \mathscr{F} \quad \forall \gamma \in \mathscr{C}r \quad \exists B \in \mathscr{F} \quad cr_A(B) = \gamma.$

L'approche épistémique est *forcément* probabiliste

Théorème. Sous les hypothèses ci-devant, il existe $\pi : \mathscr{C}r \to [0,1]$ qui convertit les crédences en probabilités, au sens où :

- $\pi(\bullet)$ est croissante ; $\pi(\text{✗}) = 0$; $\pi(\text{✓}) = 1$;
- $\pi(compl(\gamma)) = 1 - \pi(\gamma)$;
- $\pi\Big(chaine(\gamma, \delta)\Big) = \pi(\gamma)\pi(\delta)$.

En outre, cette fonction est unique.

Remarque. $\pi(\bullet)$ n'est pas forcément surjective, ni injective.

L'approche épistémique est *forcément* probabiliste

Théorème. Sous les hypothèses ci-devant, il existe $\pi : \mathscr{C}r \to [0,1]$ qui convertit les crédences en probabilités, au sens où :

- $\pi(\bullet)$ est croissante ; $\pi(\text{✗}) = 0$; $\pi(\text{✓}) = 1$;
- $\pi(compl(\gamma)) = 1 - \pi(\gamma)$;
- $\pi\Big(chaine(\gamma, \delta)\Big) = \pi(\gamma)\pi(\delta)$.

En outre, cette fonction est unique.

Remarque. $\pi(\bullet)$ n'est pas forcément surjective, ni injective.

Remarque. En philosophie, on parle d'épistémologie bayésienne lorsque la notion « connaissance » se comprend en termes des niveaux de confiance quantifiables.

Plan

Principe de base

Une question ; plusieurs items proposés ; un seul est correct.

Exemple

Quel a été le prénom le plus donné en France sur l'ensemble de la période 1946-2024 ?

A. Jean
B. Léa
C. Marie
D. Nathalie
E. Nicolas
F. Thomas

Au lieu de cocher un item, on indique notre crédence pour chaque item [†].

[†]. Le total sera automatiquement normalisé à 1.

Principe de base

Une question ; plusieurs items proposés ; un seul est correct.

Exemple

Quel a été le prénom le plus donné en France sur l'ensemble de la période 1946-2024 ?

A.	Jean	22 %
B.	Léa	7 %
C.	Marie	26 %
D.	Nathalie	4 %
E.	Nicolas	30 %
F.	Thomas	11 %

Au lieu de cocher un item, on indique notre crédence pour chaque item [†].

[†]. Le total sera automatiquement normalisé à 1.

Principe de base

Une question ; plusieurs items proposés ; un seul est correct.

Exemple

Quel a été le prénom le plus donné en France sur l'ensemble de la période 1946-2024 ?

A.	Jean	22 %
B.	Léa	7 %
C.	Marie	26 %
D.	Nathalie	4 %
E.	Nicolas	30 %
F.	Thomas	11 %

Au lieu de cocher un item, on indique notre crédence pour chaque item [†].

Score dépendant de nos crédences et de l'identité de l'item correct. (Sera sommé sur l'ensemble du quiz).

[†]. Le total sera automatiquement normalisé à 1.

Quel barème ? Théorie de l'information

- Apprendre que A se réalise représente d'autant plus d'information qu'on y attribuait une crédence faible.
- On voudrait que la quantité d'information d'apprendre A puis B s'ajoute ; or au niveau de la crédence, on multiplie $\rightsquigarrow$ $info = -\log(crédence)$ [‡].

[‡]. Peut aussi être justifié par un argument de compression de données.

Quel barème ? Théorie de l'information

- Apprendre que A se réalise représente d'autant plus d'information qu'on y attribuait une crédence faible.
- On voudrait que la quantité d'information d'apprendre A puis B s'ajoute ; or au niveau de la crédence, on multiplie $\rightsquigarrow$ $info = -\log(crédence)$ [‡].

Info apportée sur l'item correct par le répondant =

Info que la révélation de l'item correct apporterait à un ignorant

– Info que la révélation de l'item correct apporte au répondant.

Formellement, c'est $\mathrm{H}_{\mathrm{Sh}}(\delta_\star \parallel U) - \mathrm{H}_{\mathrm{Sh}}(\delta_\star \parallel Cr)$.

[‡]. Peut aussi être justifié par un argument de compression de données.

Quel barème ? Théorie de l'information

- Apprendre que A se réalise représente d'autant plus d'information qu'on y attribuait une crédence faible.
- On voudrait que la quantité d'information d'apprendre A puis B s'ajoute ; or au niveau de la crédence, on multiplie $\rightsquigarrow$ $info = -\log(crédence)$ [‡].

Info apportée sur l'item correct par le répondant =

Info que la révélation de l'item correct apporterait à un ignorant

– Info que la révélation de l'item correct apporte au répondant.

Formellement, c'est $\mathrm{H}_{\mathrm{Sh}}(\delta_\star \parallel U) - \mathrm{H}_{\mathrm{Sh}}(\delta_\star \parallel Cr)$.

Concrètement, pour k items, l'item correct ayant reçu la crédence $q_\star$, le score vaut

$$-\log(1/k) - (-\log(q_\star)) = \log(kq_\star).$$

[‡]. Peut aussi être justifié par un argument de compression de données.

Quel barème ? Théorie de l'information

$$score = \log(kq_\star).$$

Quel barème ? Théorie de l'information

$$score = \log(kq_\star).$$

Unités de mesure : $\log 2 = 1$ bit ; $\log 10^{1/10} = 1$ dban.

Quel barème ? Théorie de l'information

$$score = \log(kq_\star).$$

Unités de mesure : $\log 2 = 1\,\text{bit}$; $\log 10^{1/10} = 1\,\text{dban}$.

Dans mon exemple, $k = 6$ et $q_\star = 26\ \% \rightsquigarrow$ Je score $+0{,}64\,\text{bit}$, soit $+1{,}93\ \text{dban}$.

Quel barème ? Théorie de l'information

$$score = \log(kq_\star).$$

Unités de mesure : $\log 2 = 1\,\text{bit}$; $\log 10^{1/10} = 1\,\text{dban}$.

Dans mon exemple, $k = 6$ et $q_\star = 26\ \% \rightsquigarrow$ Je score $+0{,}64\,\text{bit}$, soit $+1{,}93\ \text{dban}$.

Remarque.

- Score possiblement négatif si $q_\star < 1/k$...
- ... Voire infiniment négatif si $q_\star = 0$!

Quel barème ? Incitation à l'honnêteté

Autre critère pour concevoir un bon barème :

- Que le barème incite à l'évaluation honnête des crédences :

$$\arg\max_{\vec{q}} \sum_i \big(p_i \times score(\vec{q}, i)\big) = \vec{p}.$$

Quel barème ? Incitation à l'honnêteté

Autre critère pour concevoir un bon barème :

- Que le barème incite à l'évaluation honnête des crédences :

$$\arg\max_{\vec{q}} \sum_i \big(p_i \times score(\vec{q}, i)\big) = \vec{p}.$$

- Que tous les items soient traités de la même façon ;
- Qu'une réponse complètement ignorante donne un score nul ;

Quel barème ? Incitation à l'honnêteté

Autre critère pour concevoir un bon barème :

- Que le barème incite à l'évaluation honnête des crédences :

$$\arg\max_{\vec{q}} \sum_i \big(p_i \times score(\vec{q}, i)\big) = \vec{p}.$$

- Que tous les items soient traités de la même façon ;
- Qu'une réponse complètement ignorante donne un score nul ;
- Que $score(\vec{q}, i)$ ne dépende que de q_i, pas du détail des $q_{j\neq i}$.

Quel barème ? Incitation à l'honnêteté

Autre critère pour concevoir un bon barème :

- Que le barème incite à l'évaluation honnête des crédences :

$$\arg\max_{\vec{q}} \sum_i \big(p_i \times score(\vec{q}, i)\big) = \vec{p}.$$

- Que tous les items soient traités de la même façon ;
- Qu'une réponse complètement ignorante donne un score nul ;
- Que $score(\vec{q}, i)$ ne dépende que de q_i, pas du détail des $q_{j\neq i}$.

Théorème. Le barème informationnel vérifie ces propriétés ; et pour $k \geqslant 3$, c'est le seul à le faire (modulo facteur constant).

Remarque. Les différentes questions peuvent être pondérées différemment.

Pertinence pédagogique

- Récompense la capacité à certifier certaines connaissances : notant $I(\vec{p})$ l'espérance de score pour des crédences $\vec{p}$, la fonction $\vec{p} \mapsto I(\vec{p})$ est convexe.

[‡].

Pertinence pédagogique

- Récompense la capacité à certifier certaines connaissances : notant $I(\vec{p})$ l'espérance de score pour des crédences $\vec{p}$, la fonction $\vec{p} \mapsto I(\vec{p})$ est convexe.
- Barème continu $\rightsquigarrow$ Part d'aléa minorée par rapport à un QCM classique.
- Récompense la capacité à exclure certaines possibilités, bien plus nettement qu'un QCM classique.
- Tout à fait pertinent de proposer, dans une même question, des items incorrects de difficultés variées.

[‡].

Pertinence pédagogique

- Récompense la capacité à certifier certaines connaissances : notant $I(\vec{p})$ l'espérance de score pour des crédences $\vec{p}$, la fonction $\vec{p} \mapsto I(\vec{p})$ est convexe.
- Barème continu $\rightsquigarrow$ Part d'aléa minorée par rapport à un QCM classique.
- Récompense la capacité à exclure certaines possibilités, bien plus nettement qu'un QCM classique.
- Tout à fait pertinent de proposer, dans une même question, des items incorrects de difficultés variées.
- Pas d'intérêt à bluffer $\rightsquigarrow$ Incite les étudiants à prendre conscience des limites de leurs connaissances pour évaluer correctement leurs crédences.

[‡].

Pertinence pédagogique

- Récompense la capacité à certifier certaines connaissances : notant $I(\vec{p})$ l'espérance de score pour des crédences $\vec{p}$, la fonction $\vec{p} \mapsto I(\vec{p})$ est convexe.
- Barème continu $\rightsquigarrow$ Part d'aléa minorée par rapport à un QCM classique.
- Récompense la capacité à exclure certaines possibilités, bien plus nettement qu'un QCM classique.
- Tout à fait pertinent de proposer, dans une même question, des items incorrects de difficultés variées.
- Pas d'intérêt à bluffer $\rightsquigarrow$ Incite les étudiants à prendre conscience des limites de leurs connaissances pour évaluer correctement leurs crédences.

$\rightsquigarrow$ Évaluation mieux alignée avec les objectifs pédagogiques des enseignants [‡] ! ◡̈

[‡]. Sous réserve que l'évaluation par QCM fasse sens.

Pertinence pédagogique

- Récompense la capacité à certifier certaines connaissances : notant $I(\vec{p})$ l'espérance de score pour des crédences $\vec{p}$, la fonction $\vec{p} \mapsto I(\vec{p})$ est convexe.
- Barème continu $\rightsquigarrow$ Part d'aléa minorée par rapport à un QCM classique.
- Récompense la capacité à exclure certaines possibilités, bien plus nettement qu'un QCM classique.
- Tout à fait pertinent de proposer, dans une même question, des items incorrects de difficultés variées.
- Pas d'intérêt à bluffer $\rightsquigarrow$ Incite les étudiants à prendre conscience des limites de leurs connaissances pour évaluer correctement leurs crédences.

$\rightsquigarrow$ Évaluation mieux alignée avec les objectifs pédagogiques des enseignants [‡] ! ◡̈

Remarque. Méthode utilisée pour entrainer les IA génératives... ◡̈ (« Entropie croisée »).

[‡]. Sous réserve que l'évaluation par QCM fasse sens.

Le problème de la calibration

Quid si on a des crédences $\vec{p}$ en la bonne réponse, mais que la conversion en probabilités $\vec{q}$ est mal faite ? Espérance du score :

$$\sum_i p_i \log(k q_i) =: S(\vec{p} \parallel \vec{q})$$

Remarque. $S(\vec{p} \parallel \vec{p}) = \mathrm{H}_{\mathrm{Sh}}(U) - \mathrm{H}_{\mathrm{Sh}}(\vec{p})$; $S(\vec{p} \parallel \vec{q}) = \mathrm{H}_{\mathrm{Sh}}(U) - \mathrm{H}_{\mathrm{Sh}}(\vec{p} \parallel \vec{q}) = S(\vec{p} \parallel \vec{p}) - \mathrm{D}_{\mathrm{KL}}(\vec{p} \parallel \vec{q})$.

Le problème de la calibration

Quid si on a des crédences $\vec{p}$ en la bonne réponse, mais que la conversion en probabilités $\vec{q}$ est mal faite ? Espérance du score :

$$\sum_i p_i \log(kq_i) =: S(\vec{p} \parallel \vec{q})$$

Remarque. $S(\vec{p} \parallel \vec{p}) = \mathrm{H_{Sh}}(U) - \mathrm{H_{Sh}}(\vec{p})$; $S(\vec{p} \parallel \vec{q}) = \mathrm{H_{Sh}}(U) - \mathrm{H_{Sh}}(\vec{p} \parallel \vec{q}) = S(\vec{p} \parallel \vec{p}) - \mathrm{D_{KL}}(\vec{p} \parallel \vec{q})$.

Pour des énoncés à deux items : En ligne, la crédence réelle (en l'item privilégié) ; en colonne, la crédence déclarée. (Scores en dban).

	50 %	60 %	70 %	80 %	90 %	100 %
50 %						
60 %						
70 %						
80 %						
90 %						
100 %						

Le problème de la calibration

Quid si on a des crédences $\vec{p}$ en la bonne réponse, mais que la conversion en probabilités $\vec{q}$ est mal faite ? Espérance du score :

$$\sum_i p_i \log(k q_i) =: S(\vec{p} \parallel \vec{q})$$

Remarque. $S(\vec{p} \parallel \vec{p}) = \mathrm{H_{Sh}}(U) - \mathrm{H_{Sh}}(\vec{p})$; $S(\vec{p} \parallel \vec{q}) = \mathrm{H_{Sh}}(U) - \mathrm{H_{Sh}}(\vec{p} \parallel \vec{q}) = S(\vec{p} \parallel \vec{p}) - \mathrm{D_{KL}}(\vec{p} \parallel \vec{q})$.

Pour des énoncés à deux items : En ligne, la crédence réelle (en l'item privilégié) ; en colonne, la crédence déclarée. (Scores en dban).

	50 %	60 %	70 %	80 %	90 %	100 %
50 %	0	−0,09	−0,38	−0,97	−2,22	$-\infty$
60 %	0	+0,09	−0,01	−0,37	−1,26	$-\infty$
70 %	0	+0,26	+0,36	+0,24	−0,31	$-\infty$
80 %	0	+0,44	+0,73	+0,84	+0,64	$-\infty$
90 %	0	+0,62	+1,09	+1,44	+1,60	$-\infty$
100 %	0	+0,79	+1,46	+2,04	+2,55	+3,01

Le problème de la calibration

	50 %	60 %	70 %	80 %	90 %	100 %
50 %	0	−0,09	−0,38	−0,97	−2,22	$-\infty$
60 %	0	+0,09	−0,01	−0,37	−1,26	$-\infty$
70 %	0	+0,26	+0,36	+0,24	−0,31	$-\infty$
80 %	0	+0,44	+0,73	+0,84	+0,64	$-\infty$
90 %	0	+0,62	+1,09	+1,44	+1,60	$-\infty$
100 %	0	+0,79	+1,46	+2,04	+2,55	+3,01

- On n'évalue pas que la connaissance, mais aussi la méta-connaissance : le répondant sait-il s'il sait ?

Le problème de la calibration

	50 %	60 %	70 %	80 %	90 %	100 %
50 %	0	−0,09	−0,38	−0,97	−2,22	$-\infty$
60 %	0	+0,09	−0,01	−0,37	−1,26	$-\infty$
70 %	0	+0,26	+0,36	+0,24	−0,31	$-\infty$
80 %	0	+0,44	+0,73	+0,84	+0,64	$-\infty$
90 %	0	+0,62	+1,09	+1,44	+1,60	$-\infty$
100 %	0	+0,79	+1,46	+2,04	+2,55	+3,01

- On n'évalue pas que la connaissance, mais aussi la méta-connaissance : le répondant sait-il s'il sait ?
- En soi, c'est plutôt souhaitable pédagogiquement...

Le problème de la calibration

	50 %	60 %	70 %	80 %	90 %	100 %
50 %	0	−0,09	−0,38	−0,97	−2,22	$-\infty$
60 %	0	+0,09	−0,01	−0,37	−1,26	$-\infty$
70 %	0	+0,26	+0,36	+0,24	−0,31	$-\infty$
80 %	0	+0,44	+0,73	+0,84	+0,64	$-\infty$
90 %	0	+0,62	+1,09	+1,44	+1,60	$-\infty$
100 %	0	+0,79	+1,46	+2,04	+2,55	+3,01

- On n'évalue pas que la connaissance, mais aussi la méta-connaissance : le répondant sait-il s'il sait ?
- En soi, c'est plutôt souhaitable pédagogiquement...
- ... Mais cela devient malsain si les erreurs de calibration deviennent prépondérantes dans le score !

Plan

Plan

Test de calibration

Si les réponses sont correctement calibrées, à une question donnée, le répondant s'attend à une espérance

$$I(\vec{q}) := \sum_i q_i \log(k q_i)$$

et à une variance

$$V(\vec{q}) := \sum_i q_i \big(\log(k q_i) - I(\vec{q})\big)^2.$$

D'où espérance et variances pour le score global (modulo indépendance), avec approximation normale (si questions nombreuses).

Test de calibration

Si les réponses sont correctement calibrées, à une question donnée, le répondant s'attend à une espérance

$$I(\vec{q}) \coloneqq \sum_i q_i \log(k q_i)$$

et à une variance

$$V(\vec{q}) \coloneqq \sum_i q_i \big(\log(k q_i) - I(\vec{q})\big)^2.$$

D'où espérance et variances pour le score global (modulo indépendance), avec approximation normale (si questions nombreuses).

$\rightsquigarrow$ Test d'hypothèse nulle par comparaison du score réel avec la distribution attendue (z-score).

Test de calibration

Si les réponses sont correctement calibrées, à une question donnée, le répondant s'attend à une espérance

$$I(\vec{q}) \coloneqq \sum_i q_i \log(k q_i)$$

et à une variance

$$V(\vec{q}) \coloneqq \sum_i q_i \big(\log(k q_i) - I(\vec{q})\big)^2.$$

D'où espérance et variances pour le score global (modulo indépendance), avec approximation normale (si questions nombreuses).

$\rightsquigarrow$ Test d'hypothèse nulle par comparaison du score réel avec la distribution attendue (z-score).

- $z < -\sqrt{3}$: Score est pire que prévu, on a pris "trop de risques" ;
- $z > \sqrt{3}$: Le score est meilleur que prévu, on été "trop prudent".

Test de calibration

Si les réponses sont correctement calibrées, à une question donnée, le répondant s'attend à une espérance

$$I(\vec{q}) \coloneqq \sum_i q_i \log(k q_i)$$

et à une variance

$$V(\vec{q}) \coloneqq \sum_i q_i \big(\log(k q_i) - I(\vec{q})\big)^2.$$

D'où espérance et variances pour le score global (modulo indépendance), avec approximation normale (si questions nombreuses).

$\rightsquigarrow$ Test d'hypothèse nulle par comparaison du score réel avec la distribution attendue (z-score).

- $z < -\sqrt{3}$: Score est pire que prévu, on a pris "trop de risques" ;
- $z > \sqrt{3}$: Le score est meilleur que prévu, on été "trop prudent".

⚠ p-valeur $\neq$ taille d'effet...

Paramètre d'excès de confiance

Modèle : On passe des crédences réelles $\vec{p}$ aux crédences déclarées $\vec{q}$ par

$$\frac{q_j}{q_i} = \left(\frac{p_j}{p_i}\right)^{\varepsilon},$$

avec $\varepsilon \in \mathbb{R}_+^*$ paramètre d'« excès de confiance », indépendant de la question.

Indicateur de taille d'effet :

- $\varepsilon = 1$: Calibration correcte ;
- $\varepsilon < 1$: Réponses trop prudentes (et à quel point) ;
- $\varepsilon > 1$: Réponses trop risquées (et à quel point).

Paramètre d'excès de confiance

Modèle : On passe des crédences réelles $\vec{p}$ aux crédences déclarées $\vec{q}$ par

$$\frac{q_j}{q_i} = \left(\frac{p_j}{p_i}\right)^{\varepsilon},$$

avec $\varepsilon \in \mathbb{R}_+^*$ paramètre d'« excès de confiance », indépendant de la question.

Indicateur de taille d'effet :

- $\varepsilon = 1$: Calibration correcte ;
- $\varepsilon < 1$: Réponses trop prudentes (et à quel point) ;
- $\varepsilon > 1$: Réponses trop risquées (et à quel point).

Notation : $\vec{q} = \mathrm{ampl}(\varepsilon, \vec{p})$. Bijection réciproque : $\vec{p} = \mathrm{ampl}(1/\varepsilon, \vec{q})$.

Paramètre d'excès de confiance

Modèle : On passe des crédences réelles $\vec{p}$ aux crédences déclarées $\vec{q}$ par

$$\frac{q_j}{q_i} = \left(\frac{p_j}{p_i}\right)^{\varepsilon},$$

avec $\varepsilon \in \mathbb{R}_+^*$ paramètre d'« excès de confiance », indépendant de la question.

Indicateur de taille d'effet :

- $\varepsilon = 1$: Calibration correcte ;
- $\varepsilon < 1$: Réponses trop prudentes (et à quel point) ;
- $\varepsilon > 1$: Réponses trop risquées (et à quel point).

Notation : $\vec{q} = \mathrm{ampl}(\varepsilon, \vec{p})$. Bijection réciproque : $\vec{p} = \mathrm{ampl}(1/\varepsilon, \vec{q})$.

Definition. Le *résultat ε-rectifié* s'obtient ainsi :

- Pour chaque question n, remplacer la réponse $\vec{q}^{(n)}$ par $\mathrm{ampl}(1/\varepsilon, \vec{q}^{(n)})$;
- Regarder le résultat global du quiz dans ce cas. ♡

Paramètre d'excès de confiance

Modèle : On passe des crédences réelles $\vec{p}$ aux crédences déclarées $\vec{q}$ par

$$\frac{q_j}{q_i} = \left(\frac{p_j}{p_i}\right)^{\varepsilon},$$

avec $\varepsilon \in \mathbb{R}_+^*$ paramètre d'« excès de confiance », indépendant de la question.

Indicateur de taille d'effet :

- $\varepsilon = 1$: Calibration correcte ;
- $\varepsilon < 1$: Réponses trop prudentes (et à quel point) ;
- $\varepsilon > 1$: Réponses trop risquées (et à quel point).

Notation : $\vec{q} = \mathrm{ampl}(\varepsilon, \vec{p})$. Bijection réciproque : $\vec{p} = \mathrm{ampl}(1/\varepsilon, \vec{q})$.

Definition. Le *résultat ε-rectifié* s'obtient ainsi :

- Pour chaque question n, remplacer la réponse $\vec{q}^{(n)}$ par $\mathrm{ampl}(1/\varepsilon, \vec{q}^{(n)})$;
- Regarder le résultat global du quiz dans ce cas. ♡

$\rightsquigarrow$ Estimateur $\hat{\varepsilon}$ de l'excès de confiance par maximisation du résultat ε-rectifié.

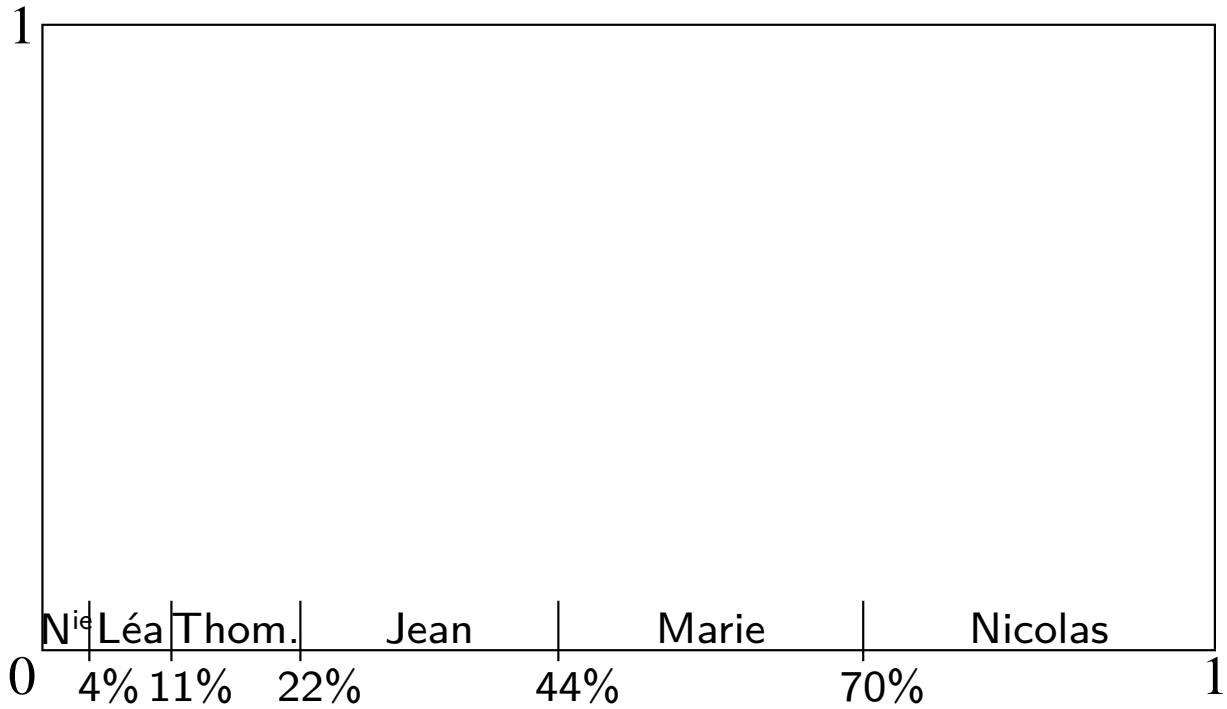
Courbe de calibration

Definition. La *courbe de calibration* d'un répondant à un quiz est définie ainsi :

Courbe de calibration

Definition. La *courbe de calibration* d'un répondant à un quiz est définie ainsi :

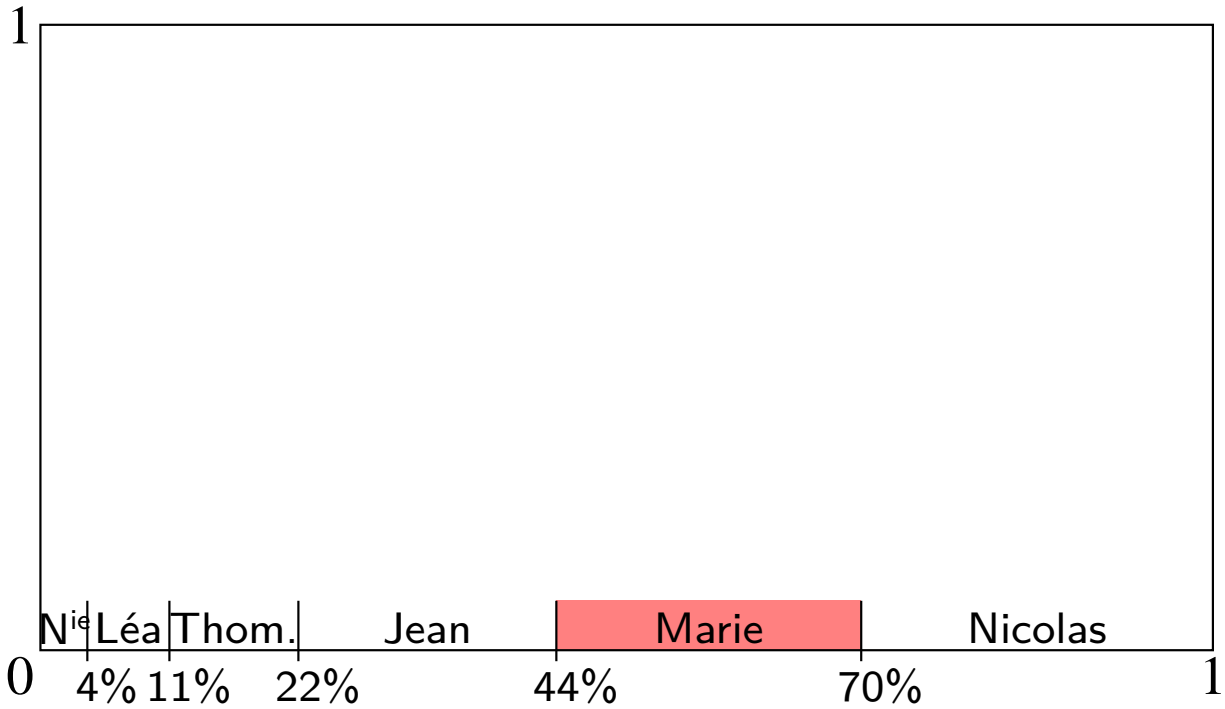
- Une réponse étant donnée, on classe les items par crédence croissante.



Courbe de calibration

Definition. La *courbe de calibration* d'un répondant à un quiz est définie ainsi :

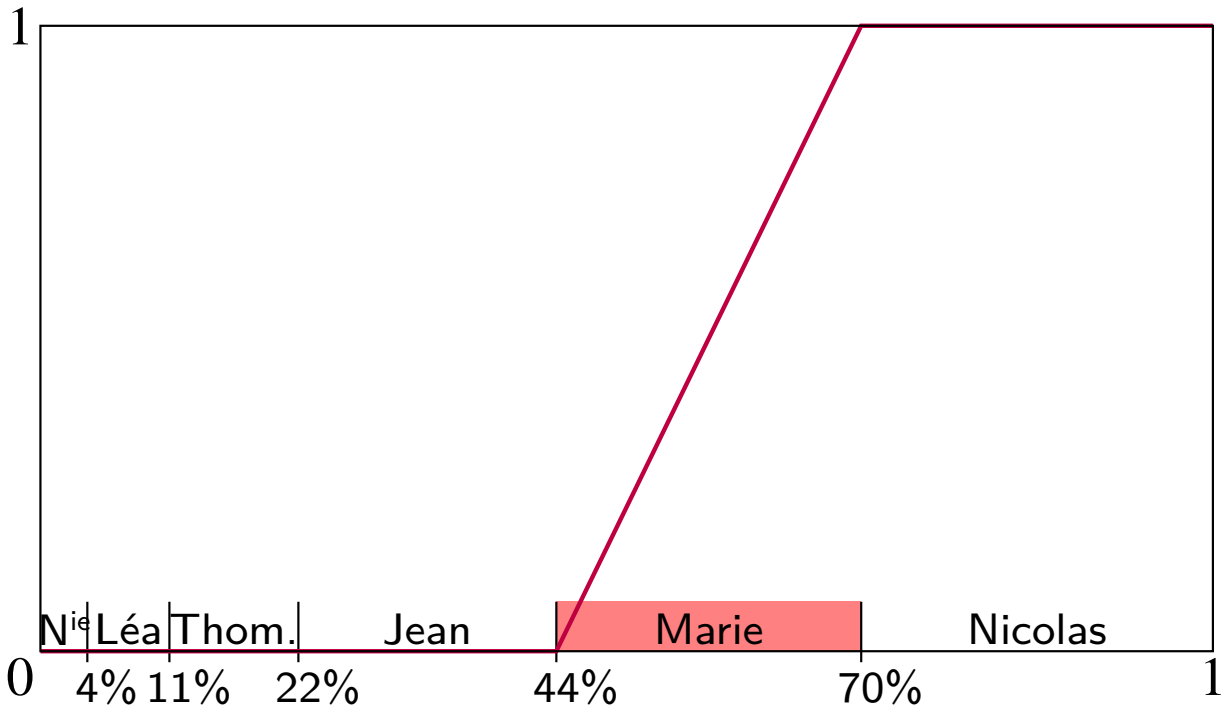
- Une réponse étant donnée, on classe les items par crédence croissante.
- À la vraie réponse, on associe un point de loi uniforme dans l'intervalle correspondant.



Courbe de calibration

Definition. La *courbe de calibration* d'un répondant à un quiz est définie ainsi :

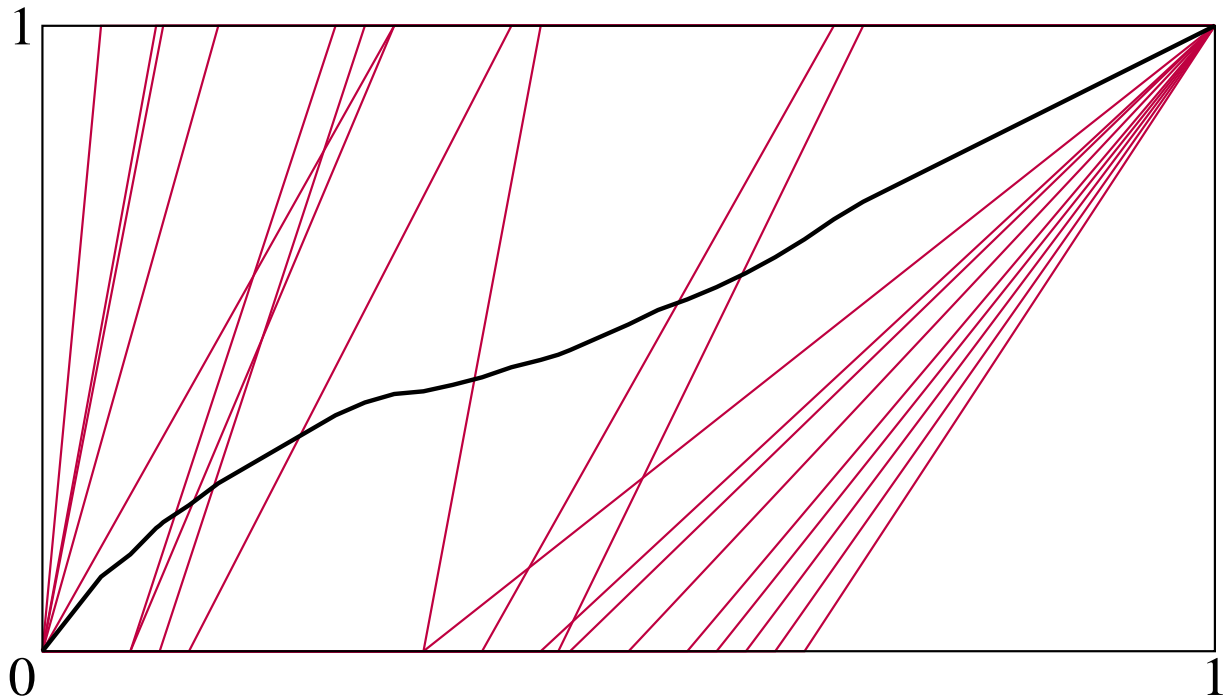
- Une réponse étant donnée, on classe les items par crédence croissante.
- À la vraie réponse, on associe un point de loi uniforme dans l'intervalle correspondant.
- On considère alors la fonction de répartition de cette loi...



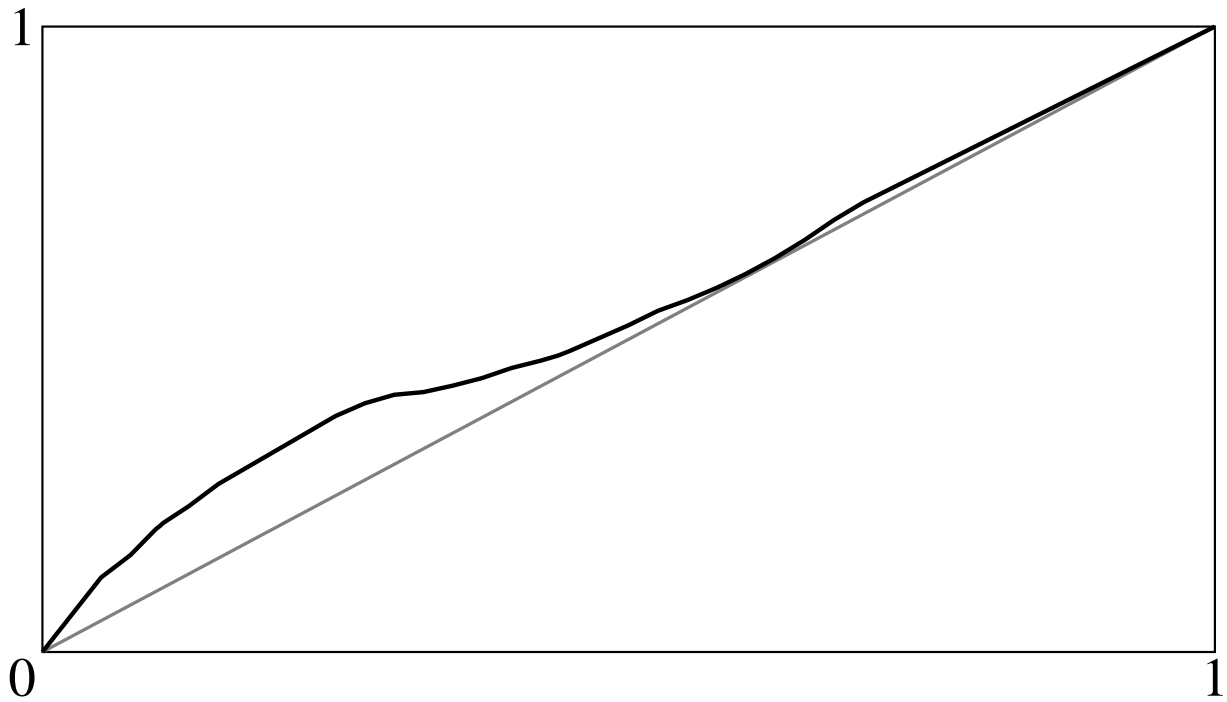
Courbe de calibration

Definition. La *courbe de calibration* d'un répondant à un quiz est définie ainsi :

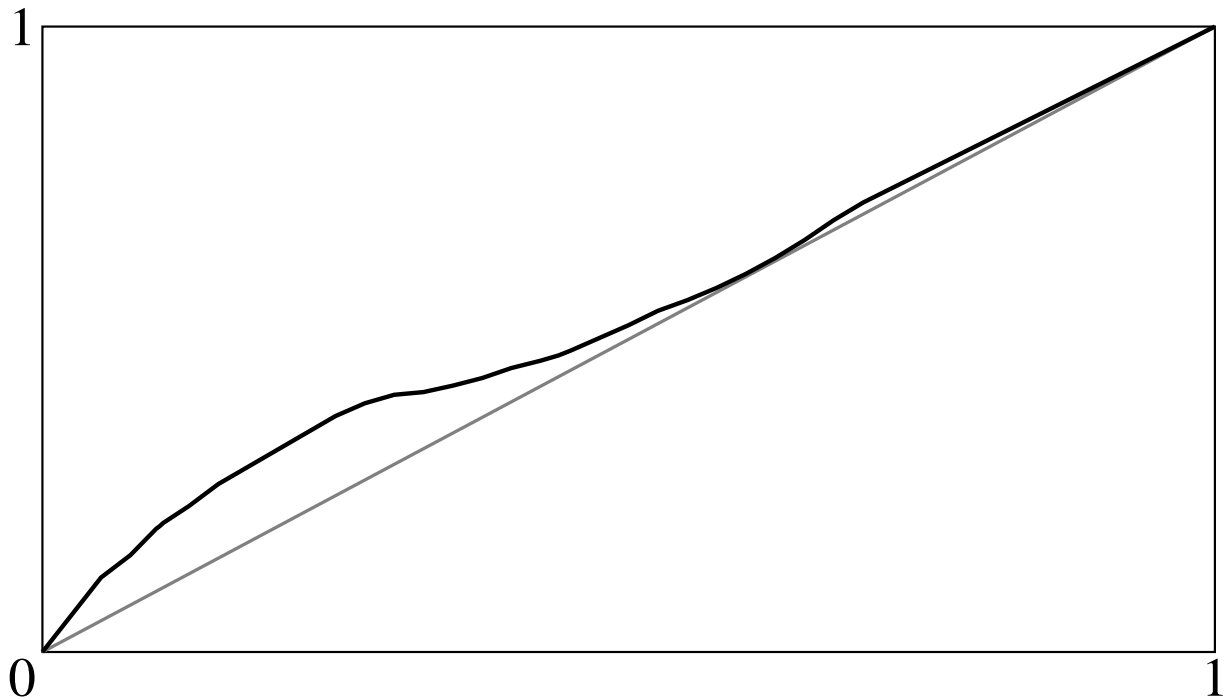
- Une réponse étant donnée, on classe les items par crédence croissante.
- À la vraie réponse, on associe un point de loi uniforme dans l'intervalle correspondant.
- On considère alors la fonction de répartition de cette loi...
- ... Qu'on moyenne sur l'ensemble des questions.



Courbe de calibration



Courbe de calibration



- Au-dessus de la diagonale : Trop risqué ;
- En-dessous de la diagonale : Trop prudent.

On peut en tirer des indicateurs de taille d'effet.

Score de connaissance pure

En utilisant les “véritables” crédences estimées via $\hat{\varepsilon}$ [§], resp. en ajustant les crédences fournies de sorte à aligner la courbe de calibration sur la diagonale [¶], on obtient un score de « connaissance pure », écartant d'éventuels défauts de méta-connaissance.

[§]. Ne fonctionne pas si certains items ont reçu une crédence de 0 : mais on peut prétraiter les données de sorte à limiter les prises de risque “objectivement déraisonnables”.

[¶]. Peut donner des choses bizarres... Il faut à minima convexfier la courbe des probabilités cumulées pour respecter l'ordre des crédences des différents items.

Score de connaissance pure

En utilisant les “véritables” crédences estimées via $\hat{\varepsilon}$ [§], resp. en ajustant les crédences fournies de sorte à aligner la courbe de calibration sur la diagonale [¶], on obtient un score de « connaissance pure », écartant d'éventuels défauts de méta-connaissance.

Remarque. Attention au biais de sur-ajustement : mais on peut l'évaluer par bootstrap.

[§]. Ne fonctionne pas si certains items ont reçu une crédence de 0 : mais on peut prétraiter les données de sorte à limiter les prises de risque “objectivement déraisonnables”.

[¶]. Peut donner des choses bizarres... Il faut à minima convexfier la courbe des probabilités cumulées pour respecter l'ordre des crédences des différents items.

Score de connaissance pure

En utilisant les “véritables” crédences estimées via $\hat{\varepsilon}$ [§], resp. en ajustant les crédences fournies de sorte à aligner la courbe de calibration sur la diagonale [¶], on obtient un score de « connaissance pure », écartant d'éventuels défauts de méta-connaissance.

Remarque. Attention au biais de sur-ajustement : mais on peut l'évaluer par bootstrap.

$\rightsquigarrow$ Particulièrement utile lors de l'initiation à ce format d'épreuve, pour fournir aux répondants un retour d'expérience plus positif ☺

[§]. Ne fonctionne pas si certains items ont reçu une crédence de 0 : mais on peut prétraiter les données de sorte à limiter les prises de risque “objectivement déraisonnables”.
[¶]. Peut donner des choses bizarres... Il faut à minima convexfier la courbe des probabilités cumulées pour respecter l'ordre des crédences des différents items.

Plan

Pré-traitement par assurance

Score totaux typiques de l'ordre de $\lesssim 2\,\mathrm{dban}$ par question.

Supposons qu'à une question à deux items, on soit persuadé d'avoir la bonne réponse.

- Si je réponds « 98:2 » au lieu de « 100:0 », j'y perds $0{,}09\ \mathrm{dban}$...
- ... Mais mon score en cas d'erreur inattendue passe de $-\infty$ à $-14\ \mathrm{dban}$!

Pré-traitement par assurance

Score totaux typiques de l'ordre de $\lesssim 2\,\mathrm{dban}$ par question.

Supposons qu'à une question à deux items, on soit persuadé d'avoir la bonne réponse.

- Si je réponds « 98:2 » au lieu de « 100:0 », j'y perds $0{,}09\ \mathrm{dban}$...
- ... Mais mon score en cas d'erreur inattendue passe de $-\infty$ à $-14\ \mathrm{dban}$!

$\rightsquigarrow$ Il semble raisonnable de s'"assurer" contre l'éventualité d'une "catastrophe" inattendue. Si le répondant ne l'a pas fait : pourquoi pas pré-traiter les données pour implémenter une telle « assurance ».

Pré-traitement par assurance

Score totaux typiques de l'ordre de $\lesssim 2\,\mathrm{dban}$ par question.

Supposons qu'à une question à deux items, on soit persuadé d'avoir la bonne réponse.

- Si je réponds « 98:2 » au lieu de « 100:0 », j'y perds $0{,}09\ \mathrm{dban}$...
- ... Mais mon score en cas d'erreur inattendue passe de $-\infty$ à $-14\ \mathrm{dban}$!

$\rightsquigarrow$ Il semble raisonnable de s'"assurer" contre l'éventualité d'une "catastrophe" inattendue. Si le répondant ne l'a pas fait : pourquoi pas pré-traiter les données pour implémenter une telle « assurance ».

Definition. L'« assurance standardisée » optimise l'espérance du score en supposant les items corrects distribués ansi : avec proba $1/2$, un « malin génie » *choisit* un des items corrects aussi nuisiblement que possible ; et à part cela les items corrects suivent (indépendamment) nos crédences.

Pré-traitement par assurance

Score totaux typiques de l'ordre de $\lesssim 2\,\mathrm{dban}$ par question.

Supposons qu'à une question à deux items, on soit persuadé d'avoir la bonne réponse.

- Si je réponds « 98:2 » au lieu de « 100:0 », j'y perds $0{,}09\ \mathrm{dban}$...
- ... Mais mon score en cas d'erreur inattendue passe de $-\infty$ à $-14\ \mathrm{dban}$!

$\rightsquigarrow$ Il semble raisonnable de s'"assurer" contre l'éventualité d'une "catastrophe" inattendue. Si le répondant ne l'a pas fait : pourquoi pas pré-traiter les données pour implémenter une telle « assurance ».

Definition. L'« assurance standardisée » optimise l'espérance du score en supposant les items corrects distribués ansi : avec proba $1/2$, un « malin génie » *choisit* un des items corrects aussi nuisiblement que possible ; et à part cela les items corrects suivent (indépendamment) nos crédences.

Remarque. En pratique, grosso modo, l'assurance standardisée interdit aux ratios de crédences de dépasser 50.

Autres barèmes incitatifs

Peut-on inciter à l'évaluation honnête des crédences sans pour autant qu'une crédence nulle attribuée à tort ne ruine tout le quiz [‖] ?

[‖]. Le cas échéant, le score dépendra de la répartition des probabilités entre les mauvaises réponses.

Autres barèmes incitatifs

Peut-on inciter à l'évaluation honnête des crédences sans pour autant qu'une crédence nulle attribuée à tort ne ruine tout le quiz [‖] ?

Réponse : Oui !

Exemples (à transformation affine près) :

- $score(\vec{q}, i) \coloneqq \|\delta_i - \vec{q}\|_2^2$: « score quadratique ».
- $score(\vec{q}, i) \coloneqq q_i \,/\, \|\vec{q}\|_2$: « score cosinus ».

[‖]. Le cas échéant, le score dépendra de la répartition des probabilités entre les mauvaises réponses.

Famille de barèmes

Plus généralement :

Famille de barèmes

Plus généralement :

- Budget fixe à allouer : La récompense est $mise_\star^\alpha$ pour $\alpha \in]0,1[$. Il faut alors prendre $mise_i \propto crédence_i^{1/(1-\alpha)}$.

Famille de barèmes

Plus généralement :

- Budget fixe à allouer : La récompense est $mise_\star^\alpha$ pour $\alpha \in]0,1[$. Il faut alors prendre $mise_i \propto crédence_i^{1/(1-\alpha)}$.
- Même récompense, mais budget libre qui représente un cout $budget^\beta$ pour $\beta \in]\alpha, \infty[$ $\rightsquigarrow$ Budget total à allouer au mieux.

Famille de barèmes

Plus généralement :

- Budget fixe à allouer : La récompense est $mise_\star^\alpha$ pour $\alpha \in]0,1[$. Il faut alors prendre $mise_i \propto crédence_i^{1/(1-\alpha)}$.
- Même récompense, mais budget libre qui représente un cout $budget^\beta$ pour $\beta \in]\alpha, \infty[$ $\rightsquigarrow$ Budget total à allouer au mieux.

Cela donne la formule de score (à affinité près) :

$$score(\vec{p}, i) := \left(\frac{p_i^{r-1}}{\|\vec{\pi}\|_r^r} - \frac{\kappa - 1}{\kappa} \right) \|\vec{\pi}\|_r^\kappa$$

pour $r, \kappa \in]1, \infty[$ $\big(r \leftarrow 1/(1-\alpha), \kappa \leftarrow \beta \,/\, (\beta - \alpha)\big)$.

Famille de barèmes

Plus généralement :

- Budget fixe à allouer : La récompense est $mise_\star^\alpha$ pour $\alpha \in]0,1[$. Il faut alors prendre $mise_i \propto crédence_i^{1/(1-\alpha)}$.
- Même récompense, mais budget libre qui représente un cout $budget^\beta$ pour $\beta \in]\alpha, \infty[$ $\rightsquigarrow$ Budget total à allouer au mieux.

Cela donne la formule de score (à affinité près) :

$$score(\vec{p}, i) := \left(\frac{p_i^{r-1}}{\|\vec{\pi}\|_r^r} - \frac{\kappa - 1}{\kappa} \right) \|\vec{\pi}\|_r^\kappa$$

pour $r, \kappa \in]1, \infty[$ $\big(r \leftarrow 1/(1-\alpha), \kappa \leftarrow \beta \,/\, (\beta - \alpha)\big)$.

Cas particuliers

- $r \to 1, \kappa \in]1, \infty[$: barème informationnel (logarithmique).
- $r = 2, \kappa = 1$: barème cosinus.
- $r = 2, \kappa = 2$: barème quadratique.
- $r \to \infty, \kappa = 1$: barème classique !

La nature continue des paramètres permet d'approcher le barème logarithmique progressivement.

Plan

Pas de diapos pour cette partie...

FIN

Merci pour votre attention ! 🙂