Eléments de statistique
Leçons 5 et 6 - Tests d'hypothèses

Louis Wehenkel

Département d'Electricité, Electronique et Informatique - Université de Liège
B24/II.93 - L.Wehenkel@ulg.ac.be

MATH0487-1 : 3BacIng, 3BacInf - 19/11/2013

Find slides: http://montefiore.ulg.ac.be/∼lwh/Stats/

Objectifs des tests d'hypothèses

- On souhaite exploiter un dataset pour prendre une décision.
  - P.ex., on souhaite décider si oui ou non il faut se lancer dans la production d'un nouveau médicament. Le dataset, nous dit pour un certain nombre de patients, si le médicament a eu un effet bénéfique. On connait, par ailleurs, le taux de succès du meilleur médicament concurrent. Si le taux de succès de notre nouveau médicament est meilleur, nous déciderions de le mettre en production. La question est donc de prendre cette décision, sachant que le dataset ne nous fournit qu'une information partielle.
- De façon plus abstraite
  - On souhaite choisir entre deux alternatives, appelées 'Hypothèses', en exploitant le dataset et les autres données du problème;
  - i.e., on souhaite savoir si la première des deux hypothèses (appelée "hypothèse nulle", et indiquée par H_0), doit être rejetée, au vu des données, en faveur de la seconde hypothèse (appelée "hypothèse alternative", et notée H_1).
  - (On comprendra plus tard la raison du choix de terminologie "hypothèse nulle" et "hypothèse alternative".)

Exemple de problème de test d'hypothèse

- On dispose d'un échantillon de n patients qui ont été traités à l'aide d'un médicament, et on constate qu'après trois jours de traitement, 43% se sont remis de leur maladie.
- On sait par ailleurs que dans un large échantillon de patients traités au moyen d'un placebo, 37% se sont remis dans le même temps.
- Doit on conclure que le traitement est inefficace, ou bien au contraire accepter l'hypothèse alternative qu'il y a bien un effet.
- Si on souhaite pouvoir rejeter H_0 avec un risque de 1% de se tromper malgré tout, quelle taille d'échantillon devrions nous considérer, en supposant que l'effet est bien de $+6\%$?

Exemple de problème de test d'hypothèse

- On dispose d'un échantillon de 20 copies corrigées pour l'examen de 2012. La moyenne des 20 cotes pour le projet 1 y vaut 13.65 et celle du projet 3 y vaut 14.40, les deux écarts-types étant de 6.70 et de 7.42.
- On souhaite, voir si à partir de cela on peut déjà raisonnablement conclure que le projet 1 est moins bien fait en moyenne sur les 182 copies rendues, que le projet 3.
- On formule l'hypothèse H_0 que la moyenne de la population du projet 1 est supérieure ou égale à celle du projet 3, et l'hypothèse alternative H_1 que celle de du projet 3 est supérieure à celle du projet 1.
- On choisit un risque de première espèce $\alpha = 0.05$, de rejeter l'hypothèse nulle alors qu'elle est vraie.

Plan du cours

- Démarche de la théorie de la décision bayesienne
  - le robot rationnel dans un monde idéal, qui optimise sa fonction de coût en prenant la décision en moyenne la plus intéressante
- Motivation de la formulation classique
  - le statisticien dans le monde réel, qui craint d'être ridicule, et donc se demande quand il doit s'abstenir.
- Neyman-Pearson rationale (risques I et II)
  - comment décider au mieux quand il faut s'abstenir, dans le monde réel.
- Construction de quelques tests "classiques".
- Synthèse et remarques
  - (p.ex. tests alternatifs, test multiples, ...)

Théorie de la décision bayesienne

- L'objectif est de choisir une action (i.e. prendre une décision) qui minimise une certaine fonction de coût.
- Pour prendre la décision on dispose d'observations, en générales partielles, sur la situation courante.
- Le coût des décisions dépend de la situation dans laquelle on se trouve.
- Si on connaissait de manière exacte la situation, on choisirait l'action de coût minimal.
- Si on n'a qu'une idée vague de la situation, exprimée par une distribution de probabilité, on choisira l'action qui minimise l'espérance du coût.

Formellement (dans le cas discret)

- Soit $\mathcal{S} \in S = \{s_1, \ldots, s_I\}$ une variable aléatoire discrète qui désigne la situation, $A = \{a_1, \ldots, a_k\}$ l'ensemble (fini) d'actions possibles, et $c(s, a)$ le coût associé à l'action a si on est dans la situation s.
- Stratégies optimales, en fonction de l'information disponible:
  - Si nous connaissons la situation, au moment de décider: $a^*(s) = \arg\min_{a \in A} c(s, a)$.
  - Si nous disposons seulement d'une loi a priori $P(s)$, pour $\mathcal{S}$: $a^* = \arg\min_{a \in A} \sum_{s \in S}^{I} P(s)c(s, a)$.
  - Si nous disposons d'une observation (use mesure) m, et que nous connaissons $f(m|s)$ et $P(s)$: $a^*(m) = \arg\min_{a \in A} \sum_{s \in S}^{I} P(s|m)c(s, a)$.
  - Notons que $\arg\min_{a \in A} \sum_{s \in S}^{I} P(s|m)c(s, a) = \arg\min_{a \in A} \sum_{s \in S}^{I} P(s)f(m|s)c(s, a)$.

Exemple (avec terminologie des tests d'hypothèses)

- Supposons qu'il n'y a que deux situations envisagées appelées "hypothèses" H_0 et H_1, de probabilité a priori $P(H_0)$ et $P(H_1) = 1 - P(H_0)$.
  - Exemple: H_0 signifie qu'un certain médicament est inefficace, H_1 signifie que ce même médicament a un effet bénéfique.
- Remarque: si nous sommes a priori agnostiques quant à l'effet du médicament, nous postulerions $P(H_0) = P(H_1) = 0.5$

Exemple (suite)

- Supposons que nous puissions faire une expérience qui nous révèle la valeur d'une variable aléatoire qui aurait une distribution différente sous l'hypothèse H_0 et sous l'hypothèse H_1.
  - Exemple: pour un patient malade traité au moyen du médicament, nous pouvons faire un test sanguin et en dériver un score de santé appelé $\mathcal{X}$.
- H_0: si le médicament est inefficace, $\mathcal{X}$ fluctue autour de $\mu_0 = 1$ avec un $\sigma_0 = \sqrt{40}$.
- H_1: si le médicament est efficace, $\mathcal{X}$ fluctue autour de $\mu_1 = 6$, avec $\sigma_1 = \sqrt{10}$
- Nous faisons l'expérience sur $n = 10$ patients malades traités moyen du médicament, en collectant un dataset $D_n = (x_1, \ldots, x_n)$ dont nous extrayons la statistique m_x.

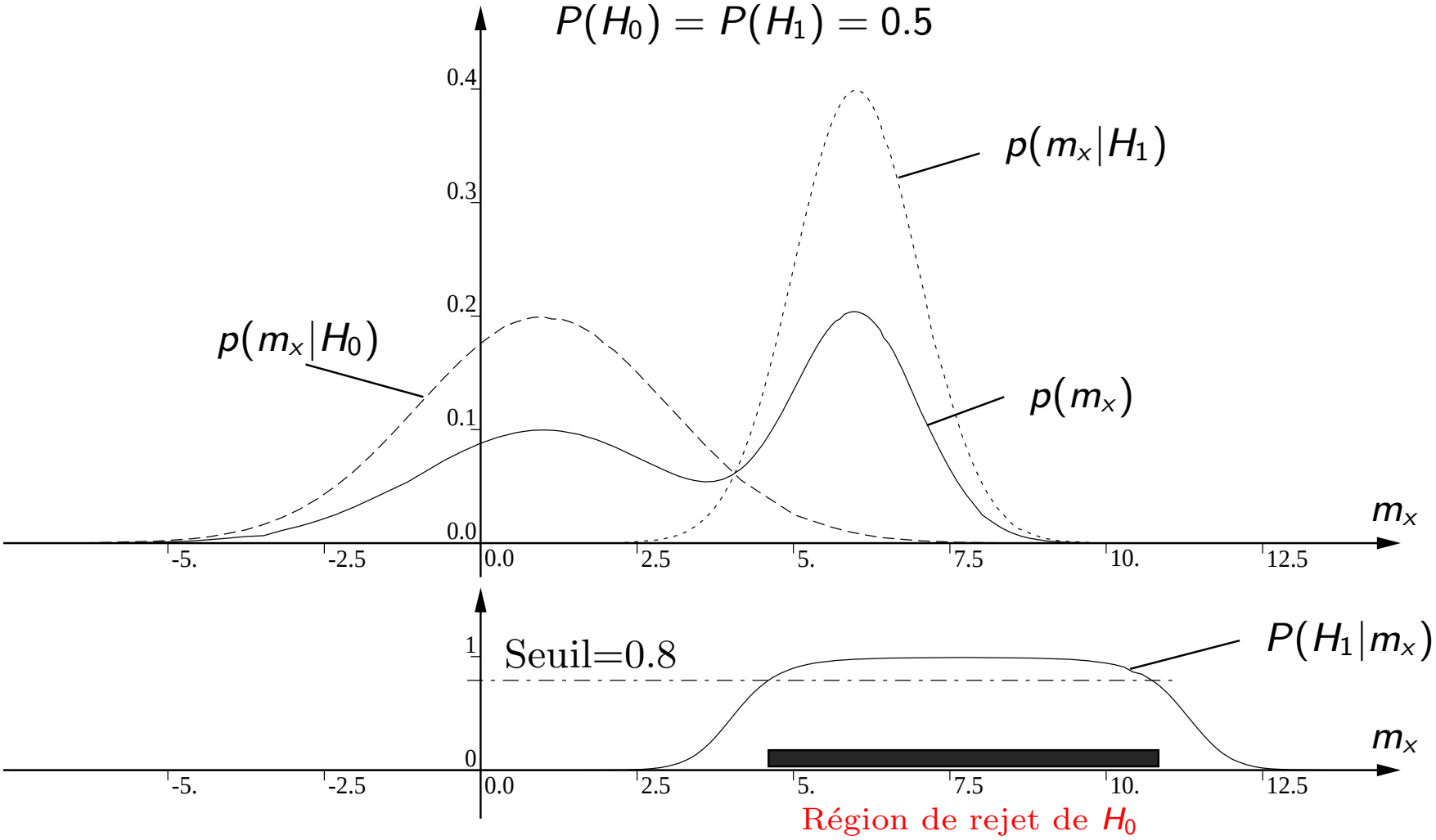
Exemple (suite)

- Sur base de la valeur de m_x nous souhaitons "décider" si le médicament est efficace et peut donc être produit en masse.
- On peut déduire de ce qui précède que m_x fluctuera ($n = 10$)
  - sous l'hypothèse H_0 autour de $\mu_0 = 1$, avec un $\sigma_{m_x,0} = \sqrt{40}/\sqrt{n} = 2$
  - sous l'hypothèse H_1, autour de $\mu_1 = 6$, avec un $\sigma_{m_x,1} = \sqrt{10}/\sqrt{n} = 1$

Exemple (suite)

- Supposons, que les aspects économiques soient les suivants:
  - L'opportunité de vente est de $\lambda_{ov} = 100 MEuros$, et le coût production et de mise en vente est de $\lambda_{cp} = 80 MEuros$.
  - Si nous décidons (action a_1) de produire et mettre en vente le médicament, alors qu'il n'est pas efficace, nous perdons $y = 80 MEuros$
  - Si nous décidons de produire le médicament, alors qu'il est efficace, nous gagnons $\lambda_{ov} - \lambda_{cp} = 20 MEuros$
  - Si nous ne produisons pas le médicament (action a_0), nous ne gagnons rien et nous ne perdons rien sous les deux hypothèses.
- Nous choisirons donc a_1 (i.e. rejet de H_0) seulement si $(\lambda_{ov} - \lambda_{cp})P(H_1|m_x) - \lambda_{cp}P(H_0|m_x) > 0$, en d'autres mots si $(\lambda_{ov} - \lambda_{cp})P(H_1|m_x) - \lambda_{cp}(1 - P(H_1|m_x)) > 0$, i.e. si $P(H_1|m_x) > \lambda_{cp}/\lambda_{ov} = 0.8$, ou encore si $P(H_0|m_x) \leq 0.2$.

Exemple (graphiquement, hypothèse de lois normales)



Exemple (discussion)

- On voit que l'hypothèse nulle H_0 (correspondant à un médicament inefficace) est rejetée pour un intervalle de valeurs de m_x environ de $[4.5; 11]$: on appelle cette région la "région de rejet de H_0, ou région critique", et les seuils 4.5 et 11 sont appelées les "valeurs critiques" de la statistique m_x utilisée pour construire le test.
- On définit deux types de probabilité d'erreur:
  - Erreur de première espèce: probabilité α de rejeter H_0 alors qu'elle est vraie.
  - Erreur de seconde espèce: probabilité β d'accepter H_0 alors que c'est H_1 qui est vraie.
- $1 - \beta$, la probabilité de rejeter H_0 alors qu'elle est fausse, est appelée la puissance du test.

Exemple (discussion et remarques)

Dans notre exemple, les valeurs de α et de β découlent des données du problème (modèle probabiliste et coûts associés aux décisions sous les deux hypothèses).

On peut se convaincre que:

- si λ_{ov} augmente (et λ_{cp} reste constant), la région de rejet de H_0 devient plus grande, et donc α augmente et β diminue;
- si $P(H_0)$ augmente (et donc $P(H_1)$ diminue, la région de rejet de H_0 devient plus petite, et donc α diminue et β augmente;
- si l'effet du médicament (en cas de guérison) est plus faible (i.e. si l'écart entre μ_1 et μ_0 est plus faible), à la fois α et β augmentent.

Exemple (discussion et remarques)

Dans notre procédure de décision, nous avons a priori choisi d'extraire du dataset $\mathbf{D}_n$ la statistique m_x, et puis nous avons construit notre test uniquement sur base de cette statistique.

- Deux datasets de même moyenne conduisent donc à la même décision, même si leurs variances s_x^2 sont fort différentes.
- On peut donc légitimement se demander si nous n'aurions pas mieux fait de baser tout notre raisonnement sur le couple de statistiques (m_x, s_x^2), voir en exploitant l'entièreté du dataset $\mathbf{D}_n$.

Réponses dans ce qui suit.

Démarche Bayesienne: principe général

- On se donne une loi a priori sur $\theta \in \Theta$, notée $f_\theta(\cdot)$
- Les deux hypothèses H_0 et H_1 correspondent à deux sous-ensembles disjoints Θ_0 et Θ_1 tels que $\Theta = \Theta_0 \cup \Theta_1$. Donc H_0 est vraie ssi $\theta \in \Theta_0$, sinon H_1 est vraie.
- On dispose d'un dataset $\mathbf{D}_n = (x_1, \ldots, x_n)$ i.i.d. selon $f_\mathcal{X}(\cdot; \theta)$. On en déduit

$$f_{\theta|\mathbf{D}_n}(\theta) = \frac{f_\theta(\theta)\mathcal{L}(x_1, \ldots, x_n; \theta)}{\int_{\theta\in\Theta} f_\theta(\theta)\mathcal{L}(x_1, \ldots, x_n; \theta)\mathrm{d}\theta}$$

(cf cours sur l'estimation bayesienne), puis

$$P(H_0|\mathbf{D}_n) = \int_{\theta\in\Theta_0} f_{\theta|\mathbf{D}_n}(\theta)\mathrm{d}\theta \quad \text{et} \quad P(H_1|\mathbf{D}_n) = 1 - P(H_0|\mathbf{D}_n).$$

- On choisit H_0 ou H_1 à partir de ces probabilités, sur base d'un seuil calculé en fonction des coûts.

(NB: la généralisation à K hypothèses mutuellement exclusives est immédiate.)

Méthode de Neyman-Pearson (motivation)

- L'approche bayesienne repose sur un modèle qui comporte les deux aspects suivants qui en pratique sont difficiles (souvent impossibles) à spécifier de manière crédible:
  - Une densité a priori f_θ sur Θ (ou bien une loi P_θ si Θ est discret).
  - Les coûts associés aux différentes décisions possibles et hypothèses.
- L'approche de Neyman-Pearson propose une démarche rationnelle lorsque ces données sont absentes
  - Elle fixe a priori la valeur α du risque de première espèce.
  - Etant donnée une valeur de α, elle propose alors de choisir entre H_0 et H_1 de façon à minimiser le risque β.
  - La méthode permet à la fois de choisir une bonne statistique et la bonne valeur critique correspondante, qui pour le risque α donné, maximisent conjointement la puissance $1-\beta$ du test.
- Nous allons expliquer les idées principales de cette méthode dans le cas le plus simple, c'est-à-dire lorsque $\Theta = \{\theta_0, \theta_1\}$ (les deux hypothèses confrontées sont alors dites "simples").

Neyman-Pearson (plan de la suite)

Voici la démarche qui sera développée en détails dans les slides qui suivent:

1. On se donne $f_{\mathcal{X}}(\cdot;\theta)$, les deux hypothèses correspondant à deux valeurs ponctuelles de θ, soit θ_0, soit θ_1.
2. On dispose d'un échantillon $\mathbf{D}_n = (x_1, \ldots, x_n) \in \mathbb{R}^n$ i.i.d. selon soit $f_{\mathcal{X}}(\cdot;\theta_0)$, soit $f_{\mathcal{X}}(\cdot;\theta_1)$.
3. La question est de construire une région de $R_0 \subset \mathbb{R}^n$ de rejet, i.e. si $\mathbf{D}_n \in R_0$ on rejettera l'hypothèse nulle (celle qui dit que $\theta = \theta_0$) et on adopte l'hypothèse alternative H_1 (qui dit que $\theta = \theta_1$).
4. On souhaite que $P(\mathbf{D}_n \in R_0|\theta_0) = \alpha$ et que sous cette contrainte on maximise la puissance $1-\beta$, i.e. on souhaite que que $P(\mathbf{D}_n \in R_0|\theta_1)$ soit maximale.
5. On démontre ensuite que la région critique R_0^* optimale est nécessairement de la forme

$$R_0^* = \left\{ \mathbf{D}_n \in \mathbb{R}^n \,\middle|\, \frac{\mathcal{L}(\mathbf{D}_n;\theta_1)}{\mathcal{L}(\mathbf{D}_n;\theta_0)} > k_\alpha \right\}$$

   où k_α est juste choisi pour que $P(\mathbf{D}_n \in R_0^*|\theta_0) = \alpha$ (de façon à satisfaire notre contrainte de départ).
6. On voit que la région critique optimale est construite uniquement sur base du "rapport de vraisemblance du dataset sous les deux hypothèses". Donc, si nous pouvons construire une statistique T sur $\mathbf{D}_n$, telle que

$$\mathcal{L}(\mathbf{D}_n;\theta) = g(T(\mathbf{D}_n);\theta)h(\mathbf{D}_n), \forall\theta$$

   alors le test optimal peut aussi être calculé uniquement en fonction de cette statistique T. Une statistique qui vérifie cette propriété est appelée statistique 'exhaustive' (ou aussi 'suffisante').
7. P.ex. si $f_{\mathcal{X}}(\cdot;\theta)$ est gaussienne de moyenne θ et de variance connue, m_x est une statistique exhaustive. Dans ce cas, formuler le test uniquement sur base de m_x est une bonne idée; et en plus à α fixé, il n'y aura en général pas de choix multiples pour définir k_α. P.ex., pour autant que $\theta_1 > \theta_0$, le seuil de décision (valeur critique) ne dépendra pas de la valeur exacte de θ_1. On aurait donc la même région critique optimale quelque soit θ_1, mais la puissance $1-\beta$ dépendra de façon forte de la valeur de θ_1.

Neyman-Pearson (position du problème)

- On dispose d'un dataset $\mathbf{D}_n \in \mathbb{R}^n$.
- L'hypothèse nulle H_0 est que ce dataset est i.i.d. selon $f_{\mathcal{X}}(\cdot;\theta_0)$
- L'hypothèse alternative H_1 est qu'il est i.i.d. selon $f_{\mathcal{X}}(\cdot;\theta_1)$
- On fixe le risque de première espèce α.
- 1. On souhaite déterminer une région R_0 de rejet de l'hypothèse nulle telle que
$$P(\mathbf{D}_n \in R_0|\theta_0) = \alpha.$$
- 2. Parmi toutes les région de rejet qui satisfont cette contrainte, on souhaite choisir celle qui maximise la puissance $1-\beta$ du test, i.e. telle que
$$P(\mathbf{D}_n \in R_0|\theta_1) = 1-\beta$$
soit maximale.

Neyman-Pearson (solution)

- **Théorème de Neyman-Pearson:** la région de rejet R_0 optimale est définie par l'ensemble des points $(x_1, \ldots, x_n) \in \mathbb{R}^n$ tels que

$$\frac{\mathcal{L}(x_1, \ldots, x_n; \theta_1)}{\mathcal{L}(x_1, \ldots, x_n; \theta_0)} > k_\alpha,$$

où la constante k_α est telle que $P((x_1, \ldots, x_n) \in R_0 | \theta_0) = \alpha$.

- Nota Bene: on a

$$P(\mathbf{D}_n \in R_0 | \theta_0) = \alpha = \int_{R_0} \mathcal{L}(x_1, \ldots, x_n; \theta_0) \mathrm{d}x_1 \cdots \mathrm{d}x_n,$$

et on a

$$P(\mathbf{D}_n \in R_0 | \theta_1) = 1 - \beta = \int_{R_0} \mathcal{L}(x_1, \ldots, x_n; \theta_1) \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

Neyman-Pearson (démonstration)

Dans ce qui suit, nous remplaçons $\mathbf{D}_n = (x_1, \ldots, x_n)$ par la notation compacte $\mathbf{x}$.

- Montrons tout d'abord, que lorsque $f_{\mathcal{X}}(\cdot;\theta)$ est une densité bornée, il existe toujours une constante k telle que

$$P\left(\frac{\mathcal{L}(\mathbf{x};\theta_1)}{\mathcal{L}(\mathbf{x};\theta_0)} > k \middle| H_0\right) = \alpha.$$

- En effet, lorsque $k = 0$, cette probabilité vaut 1. D'autre part, cette probabilité décroit monotonément et continument vers zéro, lorsque $k \to \infty$.
  Par conséquent, il doit donc exister une valeur finie de k, appelée k_α qui satisfait l'égalité, $\forall \alpha \in]0;1[$.
- Désignons alors par R_0, le sous-ensemble de $\mathbb{R}^n$ suivant,

$$R_0 \triangleq \left\{ \mathbf{x} \in \mathbb{R}^n \middle| \frac{\mathcal{L}(\mathbf{x};\theta_1)}{\mathcal{L}(\mathbf{x};\theta_1)} > k_\alpha \right\},$$

et soit R une autre partie de $\mathbb{R}^n$, telle que $P(\mathbf{x} \in R|\theta_0) = \alpha$. Montrons que $P(\mathbf{x} \in R_0|\theta_1) > P(\mathbf{x} \in R|\theta_1)$.

Neyman-Pearson (démonstration, suite)

- Considérons les trois sous-ensembles suivants de $\mathbb{R}^n$:
  - $R_0 \cap R$: la partie commune, et les différences $R_0 \setminus R$ et $R \setminus R_0$.
  - Puisque R et R_0 sont toutes les deux de probabilité α sous H_0, on a que

$$P(\mathbf{x} \in R_0 \setminus R|\theta_0) = P(\mathbf{x} \in R \setminus R_0|\theta_0) = \alpha - P(\mathbf{x} \in R_0 \cap R|\theta_0).$$

    et $P(\mathbf{x} \in R_0 \cap R|\theta_0) < \alpha$ (sinon R et R_0 seraient "identiques").
  - Notons que pour toute partie Q de $\mathbb{R}^n$, on a

$$P(\mathbf{x} \in Q|\theta_0) = \int_Q \mathcal{L}(\mathbf{x};\theta_0)\mathrm{d}\mathbf{x}.$$

$$P(\mathbf{x} \in Q|\theta_1) = \int_Q \left(\frac{\mathcal{L}(\mathbf{x};\theta_1)}{\mathcal{L}(\mathbf{x};\theta_0)}\right) \mathcal{L}(\mathbf{x};\theta_0)\mathrm{d}\mathbf{x}.$$

- Montrons alors, pour conclure, que

$$P(\mathbf{x} \in R \setminus R_0|\theta_1) < P(\mathbf{x} \in R_0 \setminus R|\theta_1).$$

Neyman-Pearson (démonstration, suite)

- Le théorème de la moyenne nous dit que pour toute partie Q de $\mathbb{R}^n$ $\exists \tilde{\mathbf{x}} \in Q$ tel que
$$\int_Q \left(\frac{\mathcal{L}(\mathbf{x};\theta_1)}{\mathcal{L}(\mathbf{x};\theta_0)}\right)\mathcal{L}(\mathbf{x};\theta_0)\mathrm{d}\mathbf{x} = \left(\frac{\mathcal{L}(\tilde{\mathbf{x}};\theta_1)}{\mathcal{L}(\tilde{\mathbf{x}};\theta_0)}\right)\int_Q \mathcal{L}(\mathbf{x};\theta_0)\mathrm{d}\mathbf{x}.$$
- Par conséquent,
$$P(\mathbf{x} \in R \setminus R_0|\theta_1) = \left(\frac{\mathcal{L}(\tilde{\mathbf{x}}_1;\theta_1)}{\mathcal{L}(\tilde{\mathbf{x}}_1;\theta_0)}\right)P(\mathbf{x} \in R \setminus R_0|\theta_0), \text{ avec } \tilde{\mathbf{x}}_1 \in R \setminus R_0$$
et
$$P(\mathbf{x} \in R_0 \setminus R|\theta_1) = \left(\frac{\mathcal{L}(\tilde{\mathbf{x}}_2;\theta_1)}{\mathcal{L}(\tilde{\mathbf{x}}_2;\theta_0)}\right)P(\mathbf{x} \in R_0 \setminus R|\theta_0), \text{ avec } \tilde{\mathbf{x}}_2 \in R_0 \setminus R.$$
- Comme
$$\left(\frac{\mathcal{L}(\tilde{\mathbf{x}}_2;\theta_1)}{\mathcal{L}(\tilde{\mathbf{x}}_2;\theta_0)}\right) > k_\alpha, \text{ alors que } \left(\frac{\mathcal{L}(\tilde{\mathbf{x}}_1;\theta_1)}{\mathcal{L}(\tilde{\mathbf{x}}_1;\theta_0)}\right) \leq k_\alpha,$$
et que $P(\mathbf{x} \in R \setminus R_0|\theta_0) = P(\mathbf{x} \in R_0 \setminus R|\theta_0) > 0$

on a forcément que

$$\boxed{P(\mathbf{x} \in R \setminus R_0|\theta_1) < P(\mathbf{x} \in R_0 \setminus R|\theta_1).}$$

CQFD

Neyman-Pearson (discussion)

- Le théorème de Neyman-Pearson nous indique qu'une approche optimale pour confronter deux hypothèses simples, est de "seuiller" le rapport de vraisemblance

$$\frac{\mathcal{L}(\mathbf{D}_n;\theta_1)}{\mathcal{L}(\mathbf{D}_n;\theta_0)} > k_\alpha.$$

- La valeur de k_α croît lorsque α décroit (en même temps que $1-\beta$). On montre aussi que pour ce test optimal on a toujours $1-\beta > \alpha$.
- S'il existe une statistique T telle que $\forall\theta$

$$\mathcal{L}(\mathbf{D}_n;\theta) = g(T(\mathbf{D}_n);\theta)h(\mathbf{D}_n) \quad \text{(factorisation)}$$

alors

$$\frac{\mathcal{L}(\mathbf{D}_n;\theta_1)}{\mathcal{L}(\mathbf{D}_n;\theta_0)} = \frac{g(T(\mathbf{D}_n);\theta_1)}{g(T(\mathbf{D}_n);\theta_0)},$$

autrement dit, le test optimal peut alors s'écrire uniquement en fonction de cette statistique (dite exhaustive).

Statistiques exhaustives pour un paramètre θ

- Si une statistique T est telle que $(\mathbf{D}_n \perp \theta | T)$, i.e. telle que $F_{\mathcal{X}^n}(\mathbf{D}_n | T(\mathbf{D}_n) = t, \theta) = F_{\mathcal{X}^n}(\mathbf{D}_n | T(\mathbf{D}_n) = t), \forall t, \theta$, alors la vraisemblance se factorise en

$$\mathcal{L}(\mathbf{D}_n; \theta) = g(T(\mathbf{D}_n); \theta) h(\mathbf{D}_n), \forall \theta.$$

  Remarque: $g(\cdot; \theta)$ est la densité (proba) de T sachant θ, et $h(\cdot)$ est la densité (proba) de $\mathbf{D}_n$ sachant T.

- Dans ce cas, on a

$$\hat{\theta}_{ML} = \arg\max_{\theta} g(T(\mathbf{D}_n); \theta)$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} f_{\theta}(\theta) g(T(\mathbf{D}_n); \theta)$$

  et donc les estimateurs au maximum de vraisemblance et bayesiens ne dépendent de l'échantillon $\mathbf{D}_n$ qu'au travers de la statistique T.

- De même, la région de rejet optimale d'un test d'hypothèse sur θ peut alors être calculée uniquement à partir de la statistique T.

Statistique exhaustive: exemple important

- Soit θ le paramètre μ d'une v.a. $\mathcal{X} \sim \mathcal{N}(\mu; \sigma^2)$ dont on suppose connue la variance σ^2. Soit $\mathbf{D}_n$ i.i.d. selon $\mathcal{X}$.
- Alors $m_x(\mathbf{D}_n)$ est une statistique exhaustive pour μ, et sa densité est

$$g(m_x(\mathbf{D}_n); \mu) = \frac{1}{\sigma\sqrt{\frac{2\pi}{n}}} \exp\left(-\frac{1}{2}\left(\frac{m_x(\mathbf{D}_n) - \mu}{\sigma/\sqrt{n}}\right)^2\right).$$

  La démonstration de cela est un **HOMEWORK**. *Suggestion*: montrer que $\mathcal{L}(\mathbf{D}_n; \mu)/g(m_x(\mathbf{D}_n); \mu)$ ne dépend plus de μ.

- Par conséquent, tester de façon optimale (avec une puissance maximale) $H_0 : \mu = \mu_0$ contre $H_1 : \mu = \mu_1$ revient tester si

$$\frac{g(m_x(\mathbf{D}_n); \mu_1)}{g(m_x(\mathbf{D}_n); \mu_0)} > k_\alpha$$

  ou k_α est ajusté pour que $P(\mathbf{D}_n \in R_0 | \mu_0) = \alpha$.

Statistique exhaustive: exemple important (suite)

- Tester si $\frac{g(m_x(\mathbf{D}_n);\mu_1)}{g(m_x(\mathbf{D}_n);\mu_0)} > k_\alpha$ revient (calcul simple) à tester si

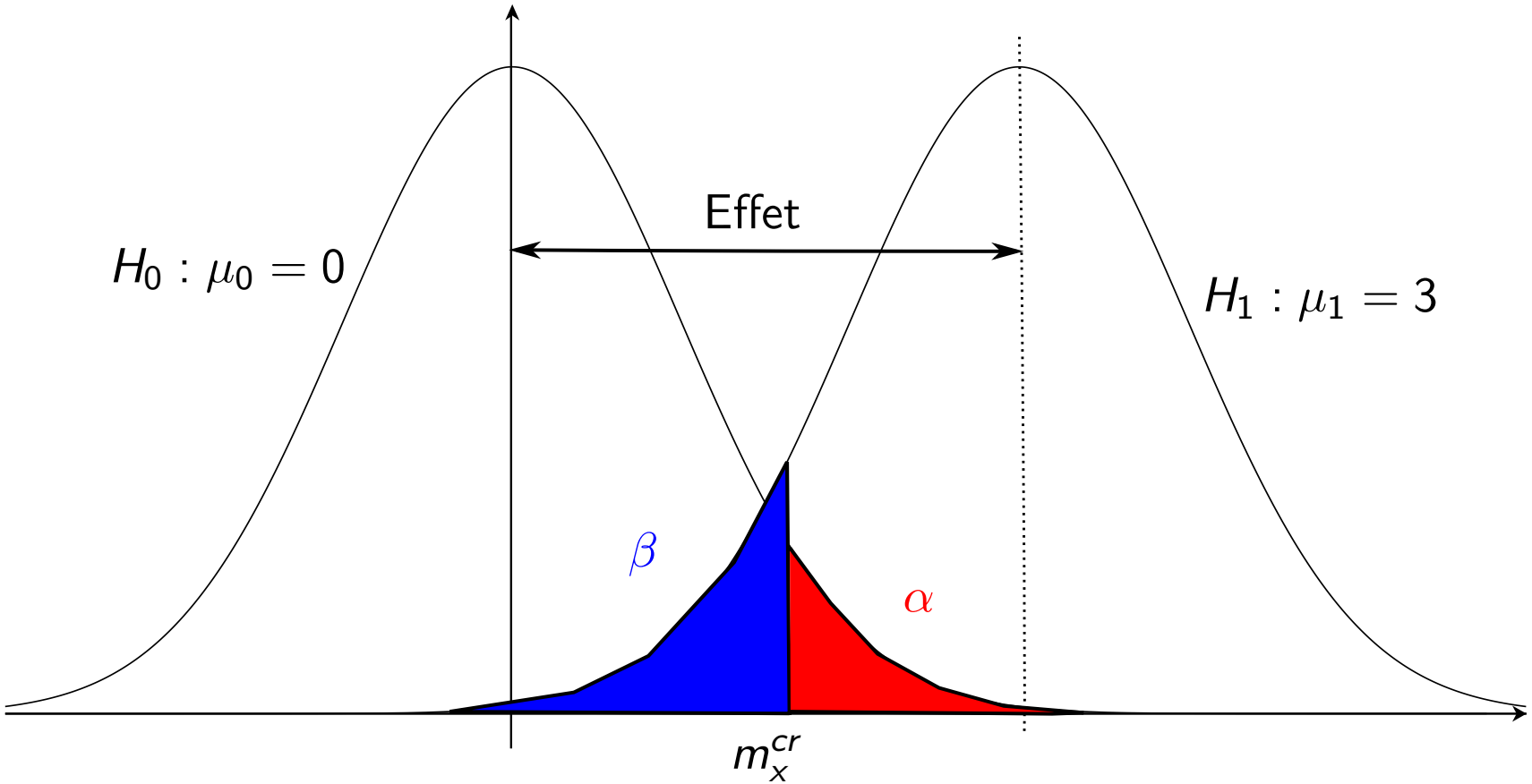
$$(\mu_1 - \mu_0)(2m_x - \mu_0 - \mu_1) > k'_\alpha \quad \text{c'est-à-dire}$$

  - si $\mu_1 > \mu_0$: tester si $m_x > k^+_\alpha$
  - si $\mu_1 < \mu_0$: tester si $m_x < k^-_\alpha$.

- Remarque: les constantes k^+_α et k^-_α ne dépendent que de μ_0, σ, n, α et pas de μ_1, car fixées par la condition $P(\mathbf{D}_n \in R_0|\mu_0) = \alpha$. On a
  - si $\mu_1 > \mu_0$: $P(\mathbf{D}_n \in R_0|\mu_0) = P(m_x > k^+_\alpha|\mu_0) = \alpha$
  - si $\mu_1 < \mu_0$: $P(\mathbf{D}_n \in R_0|\mu_0) = P(m_x < k^-_\alpha|\mu_0) = \alpha$.
- Autrement dit, le test peut être réalisé seulement en spécifiant si H_1 correspond à $\mu_1 > \mu_0$ ou bien $\mu_1 < \mu_0$, sans spécifier la valeur exacte de μ_1.
- Mais la puissance $1 - \beta$ du test dépend bien de la valeur de μ_1

Graphiquement...

Ici $\sigma^2 = 1$ et est supposé connu, et $\mu_1 > \mu_0$.



Si l'effet $|\mu_1 - \mu_0|$ diminue et TACRE, alors β augmente.

Si α diminue et TACRE, alors β augmente.

Statistiques exhaustives pour différents cas classiques

- Loi de Bernoulli: $T = f_1$ est exhaustif pour $\theta = p_1$.
- Loi normale $\mathcal{N}(\mu; \sigma^2)$:
  - σ est connu: $T = m_x$ est exhaustif pour $\theta = \mu$.
  - μ est connu: $T = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ est exhaustif pour $\theta = \sigma^2$.
  - μ et σ^2 inconnus: $T = (m_x, s_x^2)$ est exhaustif pour $\theta = (\mu, \sigma^2)$.
- Loi exponentielle $\frac{1}{\theta}\exp\left(\frac{-x}{\theta}\right)$: $T = m_x$ est exhaustif pour θ.
- Selon une procédure analogue à celle de notre "Exemple Important", on peut exploiter ces statistiques exhaustives pour construire différents test de puissance maximale applicables dans différentes situations pratiques.
- Le théorème central limite peut aussi être invoqué pour élargir les domaines d'application de ces tests, ou bien pour les formuler de manière plus simple. (CF. suite du cours et séances de répétitions.)

Neyman-Pearson (exégèse et pièges)

- "P-VALUE" :
  - Jusqu'ici nous avons fixé α a priori, puis la procédure de test en est déduite, puis $\mathbf{D}_n$ est utilisé par cette procédure pour rejeter ou non H_0.
  - Inversement, étant donnée une procédure de test et un échantillon $\mathbf{D}_n$, on peut déterminer la valeur minimale de α qui aurait conduit à rejeter l'hypothèse nulle, et mesurer ainsi "la force avec laquelle" l'échantillon semble contredire cette hypothèse.
  - Cette valeur minimale de α est appelée dans la littérature "p-value" (ou "valeur-p"). Plus elle est faible et plus l'expérience semble intéressante.
  - Une pratique courante consiste alors à formuler un grand nombre d'hypothèses nulles, puis à calculer leurs "p-values", puis à trier ces hypothèses par ordre croissant de ces dernières, puis à diagnostiquer que les hypothèses en tête de plotton sont les plus intéressantes à rejeter.
- "EFFECT SIZE" :
  - Elle mesure l'écart entre l'hypothèse nulle et les alternatives; elle est souvent négligée (car inconnue) en pratique.
  - Mais, plus l'EFFET est faible, et moins bonne sera la puissance du test.

BAD SCIENCE (...dans les médias...)

Résumé des différents type d'erreurs

	H_0 vraie	H_1 vraie
Acceptation de H_0	$1-\alpha$	β
Rejet de H_0	α	$1-\beta$

Supposons qu'on teste 1000 hypothèses H_0, dont 900 son vraies, avec $\alpha = 0.05$ et $1-\beta = 0.8$. On a alors le décompte suivant

	H_0 vraie	H_1 vraie
Acceptation de H_0	855	20
Rejet de H_0	45	80

Le revues scientifiques ne publient généralement que les résultats "étonnants", c'est-à-dire ceux qui correspondent au rejet de H_0. Or, parmi ces cas, on compte dans notre exemple $45/125 = 36\%$ d'erreurs. Cependant, les résultats négatifs (acceptation de H_0) sont bien plus fiables (2.3% d'erreurs).

Cliquer sur ce lien pour lire "The Economist" à ce sujet !

Tests basés sur les moyennes d'échantillons

- Traiter le cas du test bilatéral.
- Relâcher l'hypothèse qu'on connaît σ^2 et qu'il est le même sous les deux hypothèses.
- Relâcher l'hypothèse de lois normales, pour des échantillons suffisamment grands.
- Etendre au cas ou l'on souhaite comparer deux moyennes, issues de deux échantillons différents.
- Etendre au cas ou l'on souhaite comparer les moyennes de deux v.a. mesurées dans un même échantillon.

Tests sur une moyenne: panorama

- On dispose d'un dataset $\mathbf{D}_n$ supposé i.i.d. selon une v.a. $\mathcal{X} \in \mathbb{R}$.
- L'hypothèse nulle H_0 dit $\mu_{\mathcal{X}} = \mu_0$ (la valeur de μ_0 étant donnée).
- Plusieurs cas se présentent pour ce qui est de l'hypothèse alternative, selon le problème pratique:
  - Test unilatéral à droite: $H_1 : \mu_{\mathcal{X}} > \mu_0$.
  - Test unilatéral à gauche: $H_1 : \mu_{\mathcal{X}} < \mu_0$.
  - Test bilatéral: $H_1 : \mu_{\mathcal{X}} \neq \mu_0$.
- Plusieurs cas se présentent, selon la nature des autres hypothèses sur la v.a. $\mathcal{X}$, et sur la taille de l'échantillon:
  - Loi inconnue sous forme paramétrique, mais $n > 30$ et donc théorème central limite applicable à m_x.
  - Loi connue sous forme paramétrique (p.ex. Bernoulli, Gauss, Poisson, etc.), permettant de construire des statistiques exhaustives valables pour des petits échantillons.

Tests sur une moyenne: cas pratiques importants

- Cas où $\mathcal{X} \sim \mathcal{N}(\mu; \sigma^2)$:
  - Lorsque σ^2 est connu:
    - Sous H_0, $m_x \sim \mathcal{N}(\mu_0; \sigma^2/n)$.
    - Unilatéral à droite: rejet de H_0 si $m_x > \mu_0 + u_\alpha \frac{\sigma}{\sqrt{n}}$.
    - Unilatéral à gauche: rejet de H_0 si $m_x < \mu_0 - u_\alpha \frac{\sigma}{\sqrt{n}}$.
    - Bilatéral: rejet de H_0 si $m_x > \mu_0 + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ ou $m_x < \mu_0 - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.
  - Lorsque σ^2 est inconnu:
    - Sous H_0, $T = \frac{m_x - \mu_0}{s_{n-1}/\sqrt{n}} \sim$ Student à $n-1$ ddl.
    - Unilatéral à droite: rejet de H_0 si $m_x > \mu_0 + t_\alpha \frac{s_{n-1}}{\sqrt{n}}$.
    - Unilatéral à gauche: rejet de H_0 si $m_x < \mu_0 - t_\alpha \frac{s_{n-1}}{\sqrt{n}}$.
    - Bilatéral: rejet de H_0 si $m_x > \mu_0 + t_{\alpha/2} \frac{s_{n-1}}{\sqrt{n}}$ ou $m_x < \mu_0 - t_{\alpha/2} \frac{s_{n-1}}{\sqrt{n}}$.
  - Cas générique avec $n > 30$, pour autant que $\sigma < \infty$: on peut appliquer les mêmes formules.

Tests bilatéraux sur une moyenne, vs intervalle de confiance

Remarque:

- Les deux variantes du test **bilatéral**, consistent à ne pas rejeter l'hypothèse nulle, si respectivement:
  - lorsque σ^2 est connu:

$$\mu_0 \in [m_x - u_{\alpha/2}\frac{\sigma}{\sqrt{n}}; m_x + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}],$$

  - et lorsque σ^2 est inconnu:

$$\mu_0 \in [m_x - t_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}}; m_x + t_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}}].$$

- Cela est donc équivalent à construire un intervalle de confiance pour μ à partir de $\mathbf{D}_n$, puis à vérifier si μ_0 appartient à cet intervalle de confiance.

Comparaison de 2 moyennes (échantillons indépendants)

- On dispose de deux datasets $\mathbf{D}_{n_1}^1$ et $\mathbf{D}_{n_2}^2$ obtenus indépendamment, et chacun supposé i.i.d., respectivement selon $\mathcal{X}_1$ et $\mathcal{X}_2$.
- On souhaite tester l'hypothèse que $\mu_{\mathcal{X}_1} = \mu_{\mathcal{X}_2}$ de façon unilatérale (à gauche, ou à droite) ou bien de façon bilatérale.
  - Par exemple, on suppose que sous l'hypothèse nulle $\mathcal{X}_1 \sim \mathcal{N}(\mu; \sigma^2)$ et $\mathcal{X}_2 \sim \mathcal{N}(\mu; \sigma^2)$, sans spécifier μ.
  - On a alors que sous H_0, la variable $\Delta m_x = m_{x_1} - m_{x_2} \sim \mathcal{N}\left(0; \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$.
  - On peut construire les trois tests sur base de Δm_x:
    - $\Delta m_x > u_\alpha \left(\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$,
    - $\Delta m_x < -u_\alpha \left(\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$,
    - $|\Delta m_x| > u_{\alpha/2} \left(\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$.
  - Voir ouvrages de référence, pour extensions à des cas plus généraux.

Comparaison de 2 moyennes (échantillons appariés)

- On dispose d'un seul dataset $\mathbf{D}_n$ reprenant des couples de valeurs de $\mathcal{X}_1$ et $\mathcal{X}_2$, et supposés i.i.d. selon leur loi conjointe $f_{\mathcal{X}_1,\mathcal{X}_2}$.
- On souhaite tester l'hypothèse que $\mu_{\mathcal{X}_1} = \mu_{\mathcal{X}_2}$ de façon unilatérale (à gauche, ou à droite) ou bien de façon bi-latérale.
  - Par exemple, si on suppose que sous l'hypothèse nulle $\mathcal{X}_1$ et $\mathcal{X}_2$ son conjointement gaussiennes, on a que $\Delta\mathcal{X} = \mathcal{X}_1 - \mathcal{X}_2 \sim \mathcal{N}(0; \sigma^2_{\Delta\mathcal{X}})$.
  - On peut alors construire un test sur base de $m_{\Delta x}$ et $s^2_{\Delta x}$, en posant que la grandeur $T_{\Delta x} = \frac{m_{\Delta x}}{s_{\Delta x}/\sqrt{n-1}}$ suit une loi de Student à $n-1$ ddl.
  - Voir ouvrages de référence, pour extensions à des cas plus généraux.
  - Nous ne voyons pas les tests sur les variances, qui ont une portée limitée au cas gaussien.

Tests sur une proportion

- On dispose d'un dataset $\mathbf{D}_n$ supposé i.i.d. selon une v.a. de Bernoulli $\mathcal{X} \in \{0, 1\}$.
- L'hypothèse nulle H_0 dit $p_1 = p_{1,0}$ (la valeur de $p_{1,0}$ étant donnée).
- Plusieurs cas se présentent pour ce qui est de l'hypothèse alternative, selon le problème pratique:
  - Test unilatéral à droite: $H_1 : p_1 > p_{1,0}$.
  - Test unilatéral à gauche: $H_1 : p_1 < p_{1,0}$.
  - Test bilatéral: $H_1 : p_1 \neq p_{1,0}$.
- Dès lors que $\min\{np_{1,0}, n(1 - p_{1,0})\} \geq 5$, on peut se contenter de l'approximation gaussienne pour construire un test optimal sur base de la statistique exhaustive f_1.
  - On a $f_1 \sim \mathcal{N}\left(p_{1,0}, \sqrt{\frac{p_{1,0}(1-p_{1,0})}{n}}\right)$, et les tests s'en déduisent selon le raisonnement déjà utilisé pour la moyenne...

Tests sur des proportions (commentaires)

- Toujours sous l'hypothèse gaussienne (échantillons suffisamment grands), on construit en suivant les raisonnements déjà utilisés pour les tests de comparaison de moyennes, des tests de comparaison de proportions pour le cas des échantillons indépendants et pour le cas des échantillons appariés.
- Un test de comparaison de proportions permet aussi de formuler un test d'indépendance, cf suite.
- Un test de comparaison de proportions permet aussi de formuler des test d'ajustement, cf suite.

Tests d'ajustement et d'indépendance

- Expliquer l'idée générale.
- Illsutrer les applications, notamment aux tests de "normalité" et aux test "d'indépendance".
- Revenir sur "Kolmogorov-Smirnov".

Tests d'ajustement (idée générale)

- On dispose d'un échantillon $\mathbf{D}_n$ reportant les valeurs d'une v.a. discrète ou continue, en supposant que l'échantillon est i.i.d. selon une certaine distribution parente.
- On veut ici seulement tester l'hypothèse H_0 que $\mathbf{D}_n$ est issu d'une certaine loi, et cela de façon aussi générique que possible.
- Par exemple:
  - $\mathcal{X}$ est discrète, et peut prendre k modalités: on souhaite tester l'hypothèse que ces modalités sont de probabilités données $p_1, \ldots, p_k$ (extension du test sur une proportion).
  - $\mathcal{X}$ est continue, on souhaite tester l'hypothèse qu'elle suit une loi donnée $f_{\mathcal{X}}$ (extension du test sur une moyenne et/ou sur une variance, par rapport au cas gaussien).
- Remarque: l'approche de Neyman-Pearson répond en principe à cette question, si H_1 est simple; ici nous illustrons quelques tests souvent utilisés en pratique.

Test d'ajustement du χ^2 (fondement; cas discret)

- Soit $\mathcal{X}$ une v.a. discrète à k modalités, respectivement de probabilités, $p_1, \ldots, p_k$, soit $\mathbf{D}_n$ un dataset i.i.d. selon $\mathcal{X}$, et désignons par n_i le nombre fois que la modalité i de $\mathcal{X}$ y est observée.
- On démontre que, lorsque tous les $np_i > 5$, alors la grandeur suivante

$$D^2 \triangleq \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = n \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}$$

  suit (approximativement) une loi en χ^2_{k-1} ($k-1$ ddl).
- On peut donc tester l'hypothèse H_0 que $\mathcal{X}$ suit la loi $(p_1, \ldots, p_k)$ en calculant la valeur critique $D^2_{cr}(\alpha, k-1)$ au niveau α, associée à la loi χ^2_{k-1}, et en rejetant l'hypothèse nulle si $D^2 > D^2_{cr}(\alpha, k-1)$.

Test d'ajustement du χ^2 (le mystère des ddl)

- Le théorème central limite nous dit que les k grandeurs $(f_i - p_i)/\sqrt{p_i(1-p_i)/n}$ finiront par suivre une loi normale centrée et réduite, si n est suffisamment grand.
- Si ces k grandeurs étaient indépendantes, la loi de leur somme de carrés serait une loi χ^2_k (k ddl).
- Cependant, ces variables non sont pas indépendantes, puisqu'elles sont liées par une équation linéaire, à savoir que $\sum_{i=1}^{k} f_i = 1$.
- On démontre que D^2 suit en réalité une loi en χ^2_{k-1} ($k-1$ ddl).
- Dans le résultat précédent, le nombre de degrés de liberté est réduit de k à $k-1$ car les valeurs comparées sont liées par une seule équation linéaire, à savoir $\sum_{i=1}^{k} f_i = 1$. Aussi les facteurs $1/(1-p_i)$ ont été corrigés, de façon à ce que $E\{D^2\} = k-1$.

Test d'ajustement du χ^2 (variables continues)

- On souhaite vérifier si une v.a. continue suit une certaine loi $f_{\mathcal{X}}$, disons la loi normale.
- On peut alors bâtir sur le test χ^2 de la façon suivante:
  - On divise l'intervalle de variation de $\mathcal{X}$ en k parties, dont on calcule les probabilités p_i selon la loi $f_{\mathcal{X}}$.
  - On exploite l'échantillon, pour en déduire les f_i correspondant aux observations tombant dans le i-ème intervalle.
  - On applique le test χ^2, comme ci-dessus, aux f_i et p_i, en fixant la valeur de α a priori.
- En pratique, on peut se poser la question de savoir comment choisir le nombre k et la définition des k intervalles, chaque choix donnant lieu à un test différent.
- Si la spécification loi sous H_0 nécessite l'estimation de certains paramètres (p.ex. μ et σ) le nombre de degrés de liberté doit être réduit par autant.

Test d'indépendance du χ^2

- Soient deux v.a. discrètes $\mathcal{X}$ et $\mathcal{Y}$.
  - La première $(\mathcal{X})$ a k modalités, et la seconde $(\mathcal{Y})$ en a l.
  - Désignons par $p_{i,j}$ la probabilité d'observer $[\mathcal{X} = i \wedge \mathcal{Y} = j]$.
  - Désignons aussi par $p_{i,.} = \sum_j p_{i,j}$ la probabilité d'observer $[\mathcal{X} = i]$, et par $p_{.,j} = \sum_i p_{i,j}$ celle d'observer $[\mathcal{Y} = j]$.
- Nous voulons tester $H_0 : \mathcal{X} \perp \mathcal{Y}$, i.e. $\forall i, j : p_{i,j} = p_{i,.} p_{.,j}$, sur base d'un échantillon $\mathbf{D}_n$ d'observations conjointes de $\mathcal{X}$ et $\mathcal{Y}$.
- Définissons de façon analogue les fréquences relatives $f_{i,j}$, $f_{i,.}$ et $f_{.,j}$ calculées à partir de $\mathbf{D}_n$.
  - On montre que , si $np_{i,j} > 5, \forall i, j$, la statistique suivante

$$D^2_{\mathcal{X} \perp \mathcal{Y}} = n \sum_i \sum_j \frac{(f_{i,j} - f_{i,.} f_{.,j})^2}{f_{i,.} f_{.,j}}$$

suit $\approx$ une loi $\chi^2_{(k-1)(l-1)}$ (à $(k-1)(l-1)$ ddl) sous H_0.

Test d'ajustement de Kolmogorov-Smirnov

- On souhaite tester H_0 qui dit que la fonction de répartition de $\mathcal{X}$ est une certaine fonction de répartition donnée, disons $F^0_{\mathcal{X}}$.
- Sous H_0, on sait que la distance de Kolmogorov-Smirnov entre la fonction de répartition empirique $\hat{F}_x(\mathbf{D}_n)$ et $F^0_{\mathcal{X}}$ suit une distribution indépendante de la forme de $F^0_{\mathcal{X}}$.
- Le test consiste alors à vérifier si

$$D_n^{KS} = \sup_y |\hat{F}_x(\mathbf{D}_n, y) - F^0_{\mathcal{X}}(y)| > k_{\alpha,n}.$$

- Les valeurs critiques sont disponibles dans des tables. Par exemple, si $n > 80$, on a

$$k_{\alpha=0.05,n} = \frac{1.3581}{\sqrt{n}} \quad \text{et} \quad k_{\alpha=0.01,n} = \frac{1.6276}{\sqrt{n}}.$$

Test Kolmogorov-Smirnov (discussion, extensions)

- Le test K-S de base suppose que $F^0_{\mathcal{X}}$ est entièrement spécifiée sous l'hypothèse nulle.
- Il existe cependant des variantes du test, permettant de traiter le cas où certains paramètres de $F^0_{\mathcal{X}}$ doivent être estimés à partir de $\mathbf{D}_n$. Par exemple, si on veut tester l'hypothèse que $\mathcal{X}$ est gaussienne, sans postuler a priori les valeurs de μ et σ.
- Une autre variante du test permet de vérifier si deux fonctions de répartitions sont identiques, à partir de deux échantillons i.i.d., sur base de la distance de K-S des deux fonctions de répartitions empiriques correspondantes.
- Cela permet par exemple de vérifier si une variable $\mathcal{X}$ continue est indépendante d'une variable $\mathcal{Y}$ de Bernoulli, en comparant les deux fonctions de répartition empiriques conditionnelles de $\mathcal{X}$ obtenues à partir des deux sous-échantillons correspondant aux deux valeurs possibles de $\mathcal{Y}$.

Problématique des tests multiples (exemple)

- Exemple: on veut savoir si une certaine maladie est liée à des facteurs génétiques. H_0 prétend que cela n'est pas le cas.
  - Pour ce faire, on récolte des échantillons biologiques sur un ensemble de n personnes, dont une partie est atteinte de cette maladie, et les autres non. Ensuite on mesure l'ADN des ces personnes, à l'aide d'un certain nombre (disons m) de sondes qui regardent à des endroits spécifiques sur le génome, si la personne présente une mutation ou non.
  - Nous pouvons analyser ces données, en construisant un test d'indépendance (p.ex. du χ^2, au niveau α) entre la variable de Bernoulli qui indique si une personne est malade, et chacune des variables de Bernoulli qui indique si elle présente une mutation au niveau d'une des m sondes.
- Si on applique un risque α pour chacun de ces tests,
  - alors la probabilité de trouver un test 'rejetant l'hypothèse d'indépendance alors qu'elle est vraie' parmi les m test effectués, peut-être aussi grande que $\overline{\alpha} = 1 - (1 - \alpha)^m$.
  - Numériquement, si $\alpha = 0.01$, $m = 100$, on trouve $\overline{\alpha} = 0.63$. Autrement dit, on a plus d'une chance sur deux de rejeter H_0, alors qu'elle est vraie.

Problématique des tests multiples (discussion)

- La valeur de $\overline{\alpha} = 1 - (1 - \alpha)^m$ calculée ci-dessous correspond au pire des cas, c'est-à-dire lorsque les m mesures génétiques sont indépendantes.
- Si ces mesures étaient parfaitement dépendantes, i.e. si l'une permet de prédire les valeurs de toutes les autres, alors on aurait au contraire que $\overline{\alpha} = \alpha$, puisque tous les m tests seraient alors soit rejetés soit acceptés simultanément.
- En pratique, dans une situation de test multiples on est généralement dans une situation intermédiaire, à savoir que les différentes variables testées sont partiellement corrélées, ce qui rend plus compliqué le contrôle du risque global de première espèce $\overline{\alpha}$.
- Le problème des tests multiples est accru par les développements technologiques récents, qui permettent de collecter de très grandes bases de données, et induisent le besoin de les analyser de façon rigoureuse.
- P.ex. en génomique, les études classiques faites depuis quelques années considèrent un nombre de variables de plus en plus important (actuellement $m = 1,000,000$ est le standard !).
- Depuis les années 1990, une littérature scientifique croissante s'intéresse à ce problème des test multiples, et de nouvelles solutions dont l'étude dépasse le cadre de ce cours introductif ont été proposées.

Exercice 1: test du caractère poissonnien d'un processus ponctuel

Eléments de théorie: on considère une famille $\mathcal{X}_t, \forall t \in [0; \infty]$ de variables de Bernoulli, représentant des événements qui peuvent se produire à un instant donné, tel que l'arrivée d'un appel téléphonique, l'arrivée d'une personne dans une file d'attente, l'arrivée d'un avion sur un aéroport, etc. Par définition, $\mathcal{X}_t = 1$ si un tel événement se produit à l'instant t. Les instants $t_1, t_2, \ldots,$ d'arrivée forment une suite aléatoire, qui représente un processus ponctuel de Poisson si les conditions suivantes sont vérifiées

- Les temps d'attente entre deux événements successifs $t_{i+1} - t_i$ sont des variables aléatoires indépendantes (processus sans mémoire).
- La loi du nombre d'événements se produisant dans un intervalle de temps $[t; t + T]$ ne dépend que de T (processus stationnaire).
- Deux événements ne peuvent arriver simultanément.

Exercice 1: test du caractère poissonnien d'un processus ponctuel

Propriétés: pour un processus ponctuel de Poisson on a:

- le temps d'attente Δt entre deux événements successifs suit une loi de densité $f(\Delta t) = \exp(-c\Delta t)c$ (on a $E\{\Delta t\} = 1/c$);
- le nombre N_T d'événements se produisant dans un intervalle de temps T suit une loi de Poisson de paramètre cT, i.e.

$$P(N_T = k) = \exp(-cT)\frac{(cT)^k}{k!}.$$

On a $E\{N_T\} = cT$; c est donc la fréquence moyenne d'arrivée des événements.

Exercice 1: test du caractère poissonnien d'un processus ponctuel

Enoncé:

- On souhaite étudier le processus d'arrivée des appels téléphoniques au secrétariat facultaire de la FSA. Notre hypothèse H_0 est que ce processus est un processus ponctuel de Poisson de paramètre c inconnu.
- Pour vérifier cette hypothèse, on fait une expérience, qui consiste à compter pour 50 périodes de temps de $T = 10$ minutes, le nombre N_T d'appels reçus. On relève le tableau suivant

k	0	1	2	3	4	5	6	> 6
n_k	3	7	15	12	7	5	1	–

où n_k représente ne nombre de périodes de 10 minutes pour lesquelles on a enregistré exactement k appels.

Exercice 1: test du caractère poissonnien d'un processus ponctuel

Solution (début):

- On va réaliser un test d'ajustement, pour vérifier si les valeurs observées

k	0	1	2	3	4	5	6	> 6
n_k	3	7	15	12	7	5	1	–

  suivent bien une loi de Poisson.

- On détermine d'abord à partir de l'échantillon une estimée de la valeur de cT, en calculant la moyenne des valeurs de k. On a

$$\widehat{cT} = \frac{0 \times 3 + 1 \times 7 + \cdots + 6 \times 1}{50} = 2.64$$

- On déduit ensuite, sous l'hypothèse H_0, le nombre attendu $n \times p_k$ de périodes de 10 minutes (sur les $n = 50$) où k appels sont reçus

$$n \times p_k = 50 \times \exp(-2.64)\frac{(2.64)^k}{k!}$$

Exercice 1: caractère poissonnien d'un processus

Solution (suite):

- On obtient les valeurs attendues sous H_0 suivantes

k	0	1	2	3	4	5	6	> 6
n_k	3	7	15	12	7	5	1	–
$n \times p_k$	3.57	9.42	12.43	10.94	7.22	3.81	1.68	0.93

- Afin de pouvoir faire le test χ^2, il faut d'abord regrouper des modalités, de façon à ce que les effectifs attendus soient tous supérieurs à 5. On peut ici regrouper les deux premières cases ($k \in \{0, 1\}$) et les trois dernières ($k = \{5, 6, > 6\}$), ce qui donne

k'	≤ 1	2	3	4	≥ 5
$n'_{k'}$	10	15	12	7	6
$n \times p'_{k'}$	12.99	12.43	10.94	7.22	6.42

Exercice 1: caractère poissonnien d'un processus

Solution (fin):

- A partir du tableau

k'	≤ 1	2	3	4	≥ 5
$n'_{k'}$	10	15	12	7	6
$n \times p'_{k'}$	12.99	12.43	10.94	7.22	6.42

on calcule

$$D^2 = \sum_{k'=1}^{5} \frac{(n'_{k'} - n \times p'_{k'})^2}{n \times p'_{k'}} = 1.35.$$

- Il faut ensuite déterminer le nombre de degrés de liberté. Ici nous avons estimé un seul paramètre à partir de l'échantillon ($\widehat{cT}$), par conséquent le nombre de ddl vaut $5 - 1 - 1 = 3$.
- En supposant que le risque de première espèce soit fixé à $\alpha = 0.05$, on trouve que

$$\chi^2_c = 7.81$$

(cf tables). Nous devons donc accepter H_0, puisque $D^2 \leq \chi^2_c$.

Exercice 2: test d'indépendance de deux variables discrètes

Enoncé:

- On souhaite savoir si les intentions de vote aux prochaines élections sont significativement différentes, selon la zone électorale. Il y a 3 zones électorales (notées z_i), et trois candidats (notés c_j).
- On fait un sondage, en tirant au hasard 310 électeurs parmi ceux des trois zones, et en leur demandant d'indiquer pour lequel des trois candidats il voteront. Le tableau suivant indique les résultats obtenus:

	z_1	z_2	z_3	*Total*
c_1	50	40	35	125
c_2	30	45	25	100
c_3	20	45	20	85
Total	100	130	80	310

Exercice 2: test d'indépendance de deux variables discrètes

Solution (début):

- Il s'agit de tester l'indépendance de la variable $\mathcal{Z}$ et de la variable $\mathcal{C}$ à partir du tableau des observations:

	z_1	z_2	z_3	*Total*
c_1	$n_{1,1} = 50$	$n_{1,2} = 40$	$n_{1,3} = 35$	$n_{1,.} = 125$
c_2	$n_{2,1} = 30$	$n_{2,2} = 45$	$n_{2,3} = 25$	$n_{2,.} = 100$
c_3	$n_{3,1} = 20$	$n_{3,2} = 45$	$n_{3,3} = 20$	$n_{3,.} = 85$
Total	$n_{.,1} = 100$	$n_{.,2} = 130$	$n_{.,3} = 80$	$n = 310$

- Pour ce faire, nous allons construire une version "attendue" des effectifs des cases de ce tableau sous l'hypothèse d'indépendance, i.e. que
$\forall i, j : n \times p_{i,j} = n \times \hat{p}_{i,.} \times \hat{p}_{.,j} = n \times \frac{n_{i,.}}{n} \times \frac{n_{.,j}}{n}$.

Exercice 2: test d'indépendance de deux variables discrètes

Solution (suite):

- Le valeurs attendues sous l'hypothèse d'indépendance sont donc (en rouge) vs les valeurs observées (en bleu):

	z_1	z_2	z_3	*Total*
c_1	40.32(50)	52.42(40)	32.26(35)	$n_{1,\cdot} = 125$
c_2	32.26(30)	41.94(45)	25.80(25)	$n_{2,\cdot} = 100$
c_3	27.42(20)	35.64(45)	21.94(20)	$n_{3,\cdot} = 85$
Total	$n_{\cdot,1} = 100$	$n_{\cdot,2} = 130$	$n_{\cdot,3} = 80$	$n = 310$

- On peut calculer $D^2 = \frac{(50-40.32)^2}{40.32} + \cdots + \frac{(20-21.94)^2}{21.04} = 10.55$
- La valeur critique pour le test χ^2 à $(3-1)\times(3-1)$ ddl, et au risque $\alpha = 0.05$ de première espèce vaut 9.49.
- On doit donc en principe rejeter l'hypothèse H_0 d'indépendance entre $\mathcal{Z}$ et $\mathcal{C}$..
  NB: avec $\alpha = 0.025$ on aurait cependant accepté H_0.