Introduction et problématique

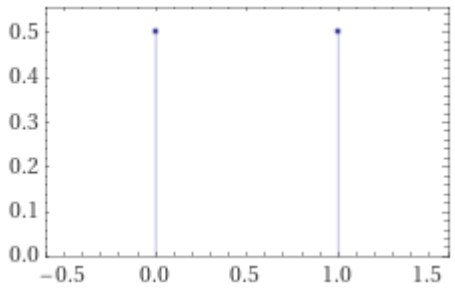
Nous allons maintenant nous pencher sur l'interprétation de résultats statistiques. Commençons en regardant un exemple très simple: supposons qu'exactement la moitié des gens préfèrent la couleur rouge au bleu, et l'autre moitié préfèrent le bleu. La préférence de chaque personne est une variable statistique: c'est une propriété de chaque individu de la population, inconnue (avant de demander) mais existante. Si nous ferions un recensement de la population au complet, on obtiendrait une valeur exacte: la moitié de la population préfère le bleu et l'autre moitié le rouge.

Cependant, en pratique, il est rarement possible de faire l'étude d'une population au complet (et, en général, ce n'est pas nécessaire). Penchons-nous donc sur ce qui se passe lorsqu'on prend un échantillon de la population. Si nous prenons une personne au hasard et lui demandons sa préférence, la valeur de la variable n'existe en fait pas car l'individu auquel on demandera n'est pas fixe. Il y a simplement une probabilité: 50% qu'elle dise bleu, 50% rouge. Il s'agit d'une variable aléatoire. Notre but sera de voir comment travailler dans ce cas-ci.

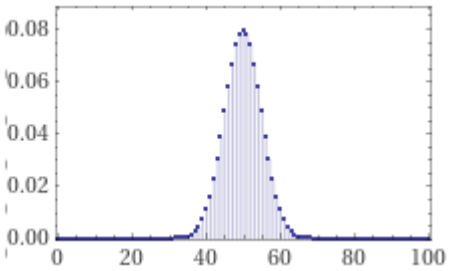
Notions de probabilités

Un évènement, en probabilités, est un résultat possible de toute éxpérience qui implique une variable aléatoire. La probabilité de cette évènement est une valeur entre 0 et 1 donnant la proportion de fois que cet évènement se produirait si l'expérience serait répétée à l'infini. Par exemple, dans notre exemple, nous pourrions appeler B l'évènement "le répondant préfère le bleu" et R l'évènement "le répondant préfère le rouge". Comme on a vu, ces deux évènements ont une probabilité de 0,5, ce que nous écririons $P(B) = 0.5$, $P(R) = 0.5$.

Nous allons toutefois préférer travailler avec des valeurs numériques, alors nous allons prendre X comme étant le nombre de fois que quelqu'un dit préférer le bleu. La distribution d'une variable X est une fonction qui associe à chaque valeur possible sa probabilité. Par exemple, en demandant à une personne, on aura la distribution:



Naturellement, demander à une seule personne ne nous permet pas de conclure grande chose. Si nous demanderions à 100 personnes, la distribution deviendrait:



On remarque deux choses ici. Premièrement, quoique notre distribution n'est définie que pour des valeurs entières de x, elle ressemble de plus en plus au graphique d'une fonction continue plus typique, comme celles vues dans d'autres cours de mathématiques. Deuxièmement, sa forme rappele celle assez connue de la "cloche de Gauss". C'est ce que nous verrons maintenant.

Distributions normales

Définition 1. *Soit X une variable aléatoire pouvant prendre des valeurs dans un intervalle de nombres réels. Sa fonction de distribution, f sera une fonction ayant les propriétés suivantes:*

1. *$f(x) \geq 0$ pour tout x.*
2. *L'aire entre l'axe des x et la courbe du graphique de f est de 1.*
3. *Pour deux nombres a et b, $P(X \in [a, b])$ sera égale à l'aire entre l'axe des x et la courbe du graphique de f entre a et b.*

Un exemple particulièrement intéressant pour nous sera la loi normale.

Définition 2. *On dit que X est une variable normale de paramètres μ et σ^2, ce que l'on note $X \sim N(\mu; \sigma^2)$, si sa fonction de distribution est:*

$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

Remarque. *La fonction de distribution de la loi normale est la fameuse cloche de Gauss, ou "bell curve". La formule de la loi normale est bien moins importante qu'on ne pourraît le croire car il est très difficile de travailler avec.*

En général, nous ne travaillerons pas directement avec les fonctions de distribution et la notion n'est inclue ici qu'à des fins de complétude.

À remarquer qu'on voit parfois la notation $N(\mu;\sigma)$ plutôt que $N(\mu;\sigma^2)$.

Définition 3. *La variable normale centrée réduite, que l'on notera de façon canonique par Z, est la fonction de distribution $N(0,1)$. Des valeurs choisies de l'aire sous la courbe de Z - c'est-à-dire de $P(Z \in [0,z])$ pour certaines valeurs de z - vous sont fournies dans le fichier table_normale.pdf.*

Les résultats qui suivent vous expliquent comment utiliser le fichier fourni pour trouver des probabilités pour une variable ayant une distribution normale

Théorème 1. *Soit $a,b \in \mathbb{R}$, $a<b$.*

i) si $a,b>0$, $P(Z \in [a,b]) = P(Z \in [0,b]) - P(Z \in [0,a])$

ii) si $a<0, b>0$, $P(Z \in [a,b]) = P(Z \in [0,-a]) + P(Z \in [0,b])$

iii) si $a>0$, $P(Z>a) = 0,5 - P(Z \in [0,a])$, $P(Z<a) = 0,5 + P(Z \in [0,a])$

Les cas i) avec $a,b<0$ et iii) avec $a<0$ se font par symmétrie de la loi normale.

Théorème 2. *Si $X \sim N\left(\mu;\sigma^2\right)$, alors $\frac{X-\mu}{\sigma} = Z$.*

Corollaire. *Soit $X \sim N(\mu;\sigma^2)$, $a,b \in \mathbb{R}$. Pour trouver $P(X \in [a,b])$:*

i) calculer $\frac{a-\mu}{\sigma}$ et $\frac{b-\mu}{\sigma}$

ii) calculer $P\left(Z \in \left[\frac{a-\mu}{\sigma}, \frac{b-\mu}{\sigma}\right]\right)$

Seuils de probabilité

Définition 4. *Nous aurons souvent besoin d'évaluer des seuils de probabilité pour une probabilité α, soit des valeurs z telles que $P(Z>z) \approx \alpha$. Nous appelerons cette valeur z_α.*

Pour évaluer z_α, nous allons d'abord calculer $0,5-\alpha$ puis chercher dans la table de la loi normale une valeur z_α telle que $P(0<Z<z_\alpha)$ soit aussi proche que possible de $0,5-\alpha$.

Nous aurons aussi souvent besoin d'évaluer des seuils de probabilité des deux côtés, soit des valeurs de z telles que $P(Z<-z \text{ ou } Z>z) \approx \alpha$. Par symmétrie de la loi normale, ceci revient à calculer $z_{\alpha/2}$.

Remarque. *Les seuils de probabilité les plus usuels seront:*

α	z_α	$z_{\alpha/2}$
0,1	*1,28*	*1,64 ou 1,65*
0,05	*1,64 ou 1,65*	*1,96*
0,01	*2,33*	*2,57 ou 2,58*

Théorème limite central

Le théorème suivant est fondamental pour l'étude des statistiques. Il en existe plusieurs formulations, nous donnons ici celle qui sera la plus facile à utiliser pour nous:

Théorème 3. *Soit X une variable statistique sur une population de taille N, de moyenne μ et variance σ^2.*

Soit $\overline{x}_n$ la moyenne d'un échantillon de taille n. Si $n > 30$, alors c'est une approximation acceptable de dire que $\overline{x}_n \sim N\left(\mu; \frac{\sigma^2}{n}\right)$.

Remarque. *Techniquement parlant, si il n'y a pas la possibilité que le même individu soit pris plusieurs fois dans l'échantillon - ce qui est généralement le cas - il serait plus correct de dire que $\overline{x}_n \sim N\left(\mu; \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right)$. Cependant, si n est très petit par rapport à N ou si nous ne connaissons pas la valeur du N, on peut utiliser l'approximation $\frac{N-n}{N-1} \approx 1$ et c'est ce que nous ferons pour ce cours.*

Maintenant, nous voudrions pouvoir appliquer nos résultats à des variables qualitatives. Pour ce faire, nous utiliserons l'astuce suivante:

Soit A une variable statistique qualitative et a une valeur qu'elle peut prendre. Nous allons définir une nouvelle variable quantitative X comme suit:

$$X = \begin{cases} 1 \text{ si } A = a \\ 0 \text{ sinon} \end{cases}$$

Théorème 4. *Supposons que la proportion d'individus pour lesquels $A = a$ est p, alors la moyenne de X sera p et la variance $p(1-p)$.*

Corollaire. *Supposons qu'une proportion p des individus d'une population ont une certaine caractéristique et soit $\overline{p}_n$ la proportion des individus d'un échantillon de taille n ayant cette caractéristique. Si $n > 30$, alors c'est une approximation acceptable de dire que $\overline{p}_n \sim N\left(p; \frac{p(1-p)}{n}\right)$.*

Exercices

1. Les résultats d'un examen standardisé ont suivi une distribution normale de moyenne 70. 95,44% des étudiant ont obtenu une note entre 60 et 80.

a) trouvez l'écart-tye et la variance

b) trouvez le pourcentage des étudiants ayant obtenu une note entre 68 et 72

c) trouvez le pourcentage des étudiants ayant obtenu une note de passage

d) trouvez le pourcentage des étudiants ayant obtenu plus de 85

2. Si on suppose que 10% de la population ont les chevex rouges et on fait une étude sur la couleur des cheveux de 100 individus, quelle serait la probabilité approximative que plus de 20 individus aient les cheveux rouges? Devrait-on changer notre supposition si on trouve que 21 individus de notre échantillon ont les cheveux rouges, et si oui comment?

3. Dans un collège, les résultats à un examen du ministère ont été distribués normalement avec une moyenne de 75 et une variance de 120. Un groupe contient 30 étudiants.

a) Quelle est la distribution de la moyenne des étudiants de ce groupe?

b) Quelle est la probabilité que la moyenne des résultats dans ce groupe soit entre 72 et 78?

c) Quel pourcentage des étudiants du collège ont coulé?

d) Quelle est la distribution du pourcentage des étudiants de ce groupe qui ont coulé?

Solutions

1.a) $\sigma = 5$, $\sigma^2 = 25$

b) $\approx 31,08\%$

c) $\approx 97,72\%$

d) $\approx 0,13\%$

2. $\approx 0,0004$, il serait raisonnable de conclure que le pourcentage est plus grand que 10%

3.a) $\overline{x} \sim N(75; 4)$

b) $\approx 0,8664$

c) $\approx 8,53\%$

d) $\overline{p} \sim N(0,0853; 0,0026)$