Statistiques et Probabilités

Rappel

I-Statistique descriptive (suite)

1. **Complément sur la régression-Corrélation**
2. **Exemples**

II-Statistique mathématique

1. **Différences entre les stats descriptives et mathématiques**
2. **Théorie des grands nombres**
3. **Définition de la probabilité**
4. **Fonction de répartition**
5. **Loi de Laplace Gauss ou Loi Normale**
6. **Propriétés de la courbe de Gauss**
7. **Exemples**

III-Conclusion

Juste un petit rappel souligné par le prof ;l'histogramme p10 de la ronéo de biostat2 est bien faux…

Toujours sur la même ronéo p9 une petite erreur s'est glissée ; il faut lire $p \leq 0.05$ *il y avait un 0 de trop…*

Pas de questions de cours à l'examen sur toute la partie théorique de proba..C'est juste pour arriver à la courbe de Gauss…

Ne vous inquiétez pas il y a beaucoup de théorie dans ce cours mais en conclusion vous verrez…c'est Gauss qu'on doit retenir !!!

IMPORTANT : Attention aux petites et grandes lettres (x et X..)La différence est dans les détails suivants

I-Statistique descriptive

1. Complément sur la Régression-Corrélation

On a déjà vu précédemment la distribution **bivariée** i.e. *y=f(x).*

Mais il peut exister également des distributions **multivariées** i.e. $y=f(x_1,x_2,..,x_n)$ *avec n variables.* En fait, la variable ne dépend pas que d'une seule autre variable mais de **plusieurs** .

C'est le cas par exemple lorsqu'on trouve à la variable y=f(x), $r=0.2$ $p\leq 0.05$;on a là un r(coefficient de corrélation) qui est loin de 0.7 c.à.d. que la variable y ne s'explique par la seule variable x…On a donc besoin d'autres variables dont va dépendre y et on obtient au final un R global>0.7 .

n.b. : *Le prof ne pense pas que cela tombe en LCA..Trop compliqué..*

Mais comprenons ce que fait le logiciel face à des données :

Chaque variable possède un coéfficient de corrélation r_1 pour $y=f(x_1)$,r_2 pour $y=f(x_2)$,…r_n pour $y=f(x_n)$; le logiciel cherche alors un **maximum de vraisemblance** à partir de ces plusieurs variables voire même jusqu'à en éliminer pour ne retenir que les plus **pertinentes** dans le but d'avoir un **R global signant une corrélation** entre ces différentes variables.

2. Exemples :

Etude à plusieurs variables pour expliquer les résistances bactériologiques..

Pr Albertini travaillant sur le BNP dont un article sortira sur la revue Diabètes&Metabolism montre que celui-ci ne dépend pas seulement de paramètres cardiaques purs mais également de pression pulsée, de micro-albuminurie..L'âge intervenant peu..

II-Statistique mathématique

1. Différences

On a vu depuis le début des **statistiquess descriptives**..ce à quoi on est confronté dans nos études :les biostatistiques.

Ce sont des **stats liées à l'observation**.

A une variable stat X de valeur x_i est associée n_i ou $f_i = n_i/N$ avec $\sum f_i=1$

X : xi---------n_i

Rappel : Quand on a une variable continue on prend le centre de classe c_i.

Maintenant on va se placer dans le monde du **hasard**..On n'est pas sûr de ce qu'on va avoir…Ce sont les **statistiques mathématiques.** Ce sont des

statistiques posées à posteriori grâce à toutes les observations faites antérieurement.

A une variable aléatoire de réalisation x_i est associée une probabilité p_i avec $0 \leq p_i \leq 1$. A chaque événement dont on ne connaît pas le résultat, on associe une probabilité p_i. Si chaque évènement est indépendant des autres alors,

$f_i \longrightarrow p_i$ (grâce à la théorie des grands nombres)

(Ex : Mesure du cholestérol chez x personnes, on a x événements possibles)

Exemple : Le sexe ratio

La probabilité d'être mâle à la naissance (établie sur 100000 naissances en France) est de p(mâle)=0.52

Cela ne sert plus aujourd'hui car grâce aux échographies, on sait d'avance.

De plus, cela peut sembler surprenant que plus de mâles naissent sachant que dans notre population vieillissante il y a plus de femmes mais en fait les hommes tiennent plus de conduites à risque ils meurent donc plus jeunes…Mais à âge équivalent les hommes sont en meilleure santé que les femmes !

2. Théorie des Grands Nombres :

Rappel : Plus les observations sont nombreuses, plus l'histogramme tend à se lisser et on obtient soit une courbe de Gauss soit une courbe penchée à gauche ou à droite

Soit la répétition d'une épreuve (une épreuve peut également être une observation).

Exemple, le jeté d'une pièce de monnaie :

p (face)=p(f)

} 2 évènements complémentaires indépendants p(f)+p(p)=1

p (pile)=p(p)

- Si j'effectue 10 jetés et que j'ai 6 face et 4 pile, j'obtiens p(f)=0.6 et p(p)=0.4
- Si je lance 100 fois j'obtiens p(f)=0.56 et p(p)=0.44
- Pour 1000 jetés j'ai p(f)=0.48 et p(p)=0.52
- …..
- A 10^6 jetés j'obtiens 5.10^5 pile et 5.10^5 face…

Donc la théorie des grands nombres est que la probabilité est égale à la fréquence lorsque le nombre d'épreuves tend vers l'infini.

$$\mathbf{P_i = f_i \text{ quand } n \longrightarrow \infty}$$

3. Définition d'une probabilité

La probabilité est le nombre des cas favorables à la réalisation d'un évènement (n) sur le nombre de cas possibles (N)

$$p = \frac{n}{N}$$

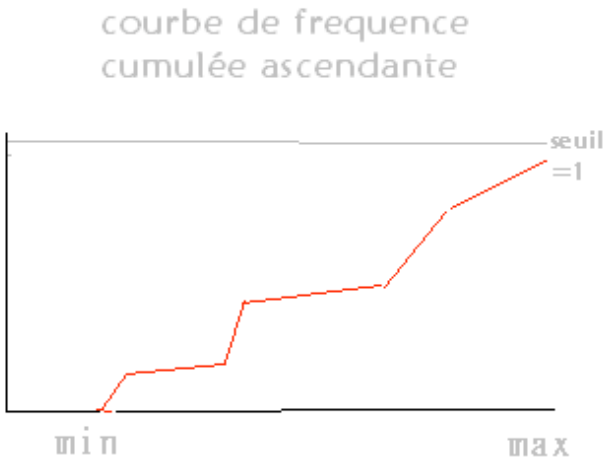
Tout ceci est très absolu et ne sert qu'à la compréhension des stats..Pas de questions là-dessus..

4. Fonction de répartition

Revenons à une variable aléatoire continue (≠ variable stat car on ne connaît pas sa réalisation).

Soit X dont la réalisation x se trouve dans un intervalle de valeurs]-∞ ; +∞ [.On introduit alors une **fonction de répartition** concernant cette variable ainsi qu'une courbe associée.

Rappel ou remarque sur la courbe de fréquence cumulée ascendante par analogie avec celle de la fonction de répartition ; il s'agit d'une courbe brisée ≠ continue ;en abscisse on a un min et un max et en ordonnée on a $\sum f_i=1$



La courbe associée à la fonction de répartition de la variable aléatoire continue a un aspect qui ne correspond qu'aux variables aléatoires continues. Il existe un point d'inflexion indiquant un changement de concavité

Schéma voir à la fin du cours

Pour obtenir une probabilité, cette fonction de répartition s'écrit

$$F(x)=p(X<x).$$

Autrement dit, si on a 2 réalisations **a** et **b,** la probabilité pour que x appartienne à [a ; b] est : $p(a\leq X<b)=p(X<b) - p(X<a) = F(b) - F(a)$.

On peut aussi vouloir trouver autre chose. Quelque chose d'assez important ; il s'agit de la probabilité infinitésimale sur un intervalle dx (très petit):

$$p(x\leq X<x+dx)= F(x+ dx) - F(x)$$

$$\lim_{dx\to 0} \frac{F(x+dx) - F(x)}{dx} = F'(x) = f(x)$$

F'(x) est la dérivée de la fonction de répartition qui graphiquement se trouve être la **courbe de Gauss ;** elle représente la **densité de probabilité.**

Schéma à voir à la fin du cours

On a vu que la **moyenne** $\bar{x} = \sum_i f_i x_i$ ou $\bar{x} = \sum_i f_i c_i$ avec les variables continues

- Avec les probas pour une variable <u>discontinue</u>, on a $\bar{x} = \sum_i p_i c_i$ avec $\sum p_i=1$

5. Pour une variable aléatoire continue, on a une **Esperance mathématique E(x)**

$E(x) = \bar{x} = \int_{-\infty}^{+\infty} x f(x) dx$ avec $\int_{-\infty}^{+\infty} f(x) dx = 1$ (on retrouve $\sum p_i=1$ MAIS pour des variables continues)

La densité de probabilité est donnée par la loi de Laplace-Gauss ou Loi-normale ; elle dépend de la moyenne $\bar{x}$ et de l'écart type σ, $\mathcal{N}(\bar{x},\sigma)$

Rappel : $V(x) = \left[\int_{-\infty}^{+\infty} x^2 f(x) dx\right] - \bar{x}$ avec $\bar{x} = \int_{-\infty}^{+\infty} x f(x) dx$ et $\sigma = \sqrt{V(x)}$

Dans la loi normale, on précise ce que l'on a été obtenu à partir de l'observation : on à alors la densité de probabilité. Les probabilités sont obtenues par intégration de cette densité de probabilité dont voici l'expression

$$\mathcal{N}(\bar{x},\sigma) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp\left(-\frac{1}{2}\frac{x_i - \bar{x}}{\sigma}\right)^2$$

Mais cette loi normale est trop compliquée on fait alors un changement de variable pour en simplifier l'écriture et afin d'en établir une table qui prendra une valeur universelle.

On pose alors une nouvelle variable aléatoire $T = \frac{X - \bar{x}}{\sigma}$ dont les réalisations sont t_i.

On a alors une expression simplifiée de la densité de probabilié:

$$\mathcal{N}(0;1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\ t_i\right)^2$$

On a alors graphiquement la **Loi Normale centrée réduite** car centrée sur 0 et réduite car l'écart type ne vaut qu'1.

Schéma voir à la fin du cours

La probabilité de T, $\prod(T<t)$ est la surface (ou aire) sous la courbe de Gauss.

Du fait de la symétrie de la courbe de GAUSS on a

$$\prod(T\geq t)= 1 - \prod(T<t) \text{ et } \prod(T<-t) = \prod(T\geq t$$

Pour obtenir une probabilité, on intègre de l'infini jusque la valeur étudiée.

Il existe une table universelle donnant $\prod(T<t$.*Le prof a dit qu'il n'était pas nécessaire que nous l'ayons (car on la trouve dans tous les bouquins de statisriques)…*

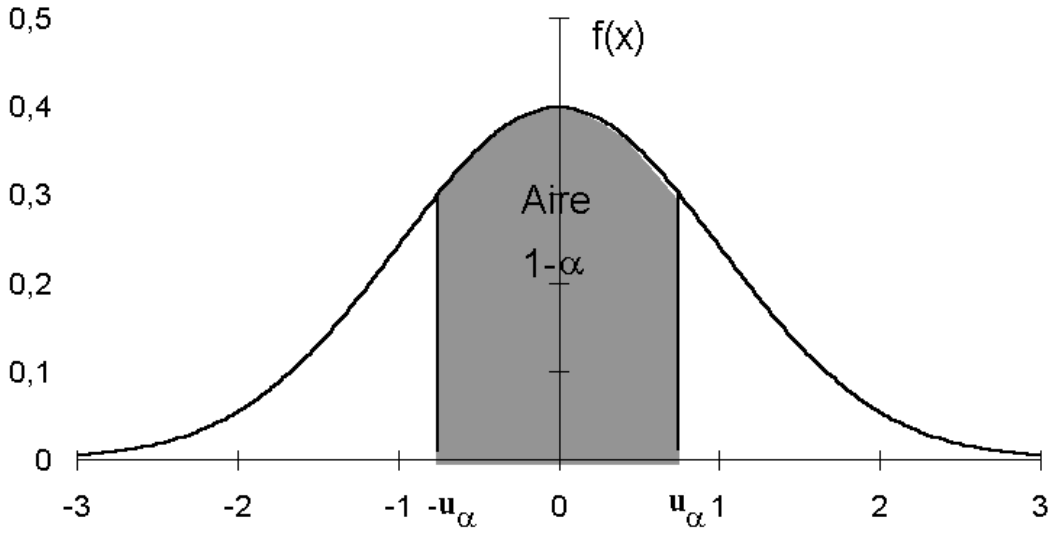
Mais elle ne donne que des valeurs positives on applique alors $\prod(T<-t) = \prod(T\geq t)$ et

$\prod(T<-t) = \prod(T\geq t)$ car $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathbf{exp}\left(-\frac{1}{2}\ \mathbf{t}\right)^2 \mathbf{dt} = 1$

6. Propriétés de la Courbe de Gauss+++

La courbe de Gauss est remarquable car :

1. elle est **centrée sur la moyenne**
2. la **demi largeur à la mi-hauteur est égale à l'écart type**
3. **elle permet de déduire l'intervalle de confiance (**permettant de donner une échelle de NORMALITE)



Lorsque x appartient à $\bar{x} \pm \sigma$, la probabilité est réalisée dans 68%

Lorsque x appartient à $\bar{x} \pm 2\sigma$, la probabilité est réalisée à 95.4%.

Pour obtenir le FAMEUX intervalle de confiance de 95% définissant la normalité, il faut que

$$x = \bar{x} \pm 1{,}96\sigma$$

Ce qui correspond à un risque de 5% ou 0.05

Le fameux p≤0.05 que l'on rencontre dans les études statistiques.

Donc pour être NORMAL au sens biologique il faut qu'une mesure x appartienne à l'intervalle de confiance à 95% (x est appelée estimation ponctuelle)

La normalité c'est d'être dans les 95% de la population.

Exemple :TAUX DE POTASSIUM NORMAL ENTRE 3.5 ET 4.5 pour 95% de la population (saine ?) possède un taux compris dans cet intervalle ; les autres risquent des pathologies (torsades de pointe, tachycardie ventriculaire..) où sont des exceptions.

7.Exemples

Loi normale à partir d'observations $\mathcal{N}(5 ;2)$ i.e. moyenne=5 et écart type=2

- P(X<9) x=9 chgt de variable t=*,9-5-2.=2* puis on trouve dans la table universelle p(2) qui est de 0.9772 d'où p(X<9)=0.9772
- P(X≥8.36) x=8.36 chgt de variable t=*,8.36-5-2.=1.68* □ '□ù p(1.68)=0.9535 ainsi p(X<8.36)=0.9535 et p(X≥8.36)=1-0.9535=0.0465

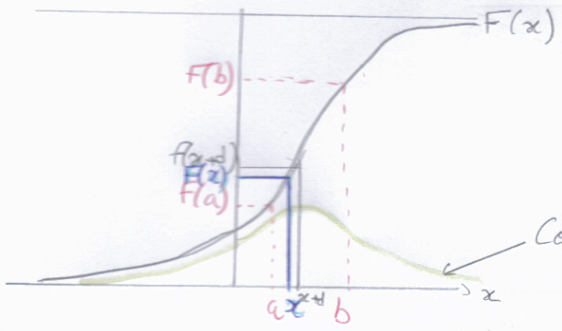
Conclusion :

Beaucoup de théorie ….pour finalement devoir retenir **l'importance de la courbe de Gauss ;ses propriétés ; son utilisation(intervalle de confiance !!)**

Comprendre également la différence entre statistique descriptive et statistique mathématique ; la théorie des grands nombres qui tend à lisser un histogramme en courbe

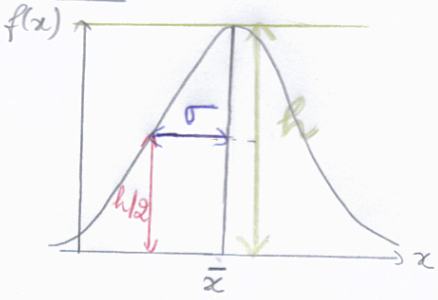
Si à l'examen, si on nous propose une courbe de Gauss, on mesure très facilement (avec un décimètre) la moyenne (en abscisse) et l'écart type (à la demi-largeur à mi hauteur)

Fonction de Repartition



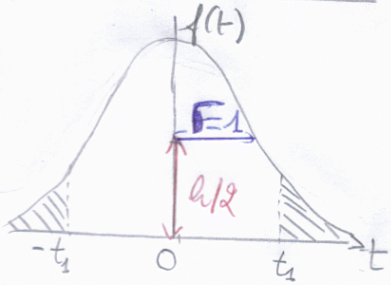
$F(a) = p(X < a)$

Courbe de Gauss



h : hauteur de la courbe gauss

Loi Normale Centrée Reduite



$\mathcal{N}(0;1)$ $f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$