Méthodes de statistique inférentielle.

A. Philippe

Laboratoire de mathématiques Jean Leray
Université de Nantes
Anne.Philippe@univ-nantes.fr

Version modifiée le 19 mai 2016

http://www.math.sciences.univ-nantes.fr/~philippe/

Plan du cours

Plan de la section

Quelques problèmes

1. Un fabricant souhaite vérifier la qualité des ampoules électriques produites par une nouvelle chaîne de production.
   Il faut donc évaluer la durée moyenne de fonctionnement des ampoules.

   Comment évaluer cette durée moyenne ?

   On ne peut pas tester toutes les ampoules !
2. Le responsable d'un parti politique souhaite estimer la proportion des militants favorables à la candidature de Mr X pour la prochaine élection présidentielle.

   Comment calculer la popularité d'un candidat au sein d'une population ?

   Interroger tous les militants est trop coûteux.

Population & Échantillon

Définition

La population : l'ensemble de tous les éléments considérés dans une étude.

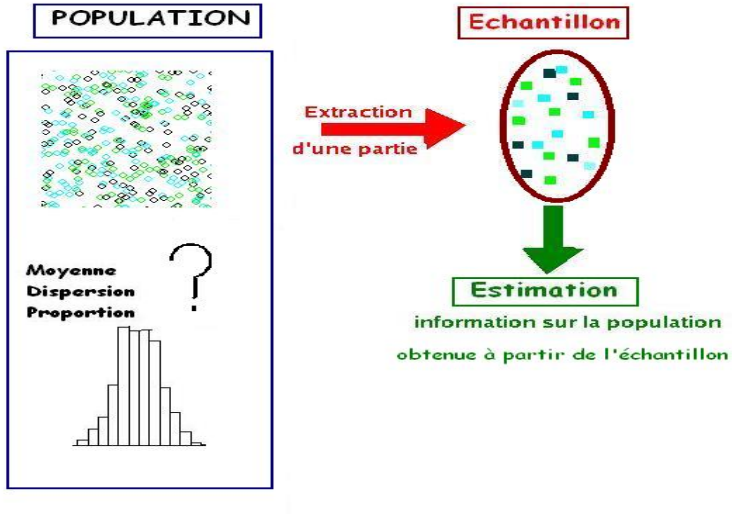
Définition

L'échantillon est un sous ensemble fini de la population.
La taille de l'échantillon est le nombre d'éléments sélectionnés pour constituer l'échantillon.

Le but de l'inférence statistique.

Tirer des conclusions concernant certaines caractéristiques de la population à partir des informations contenues dans l'échantillon.

Pour résumer



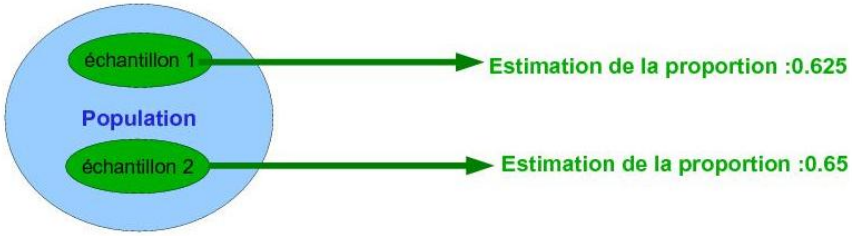
Retour aux exemples

1. Le fabricant d'ampoules.
   Il prélève un échantillon constitué de 130 ampoules.
   Pour chaque ampoule, il mesure la durée de fonctionnement.
   La moyenne de l'échantillon vaut 36 000 heures.
   Une estimation pour la population est 36 000 heures.
2. Le responsable du parti.
   Il constitue un échantillon de taille 400. Parmi les personnes sélectionnées, 250 sont favorables au candidat proposé.
   Une estimation de la proportion de la population favorable à Mr X est $250/400 = 0.625$

Quelle est la qualité de ces deux estimations ?

Erreur d'échantillonnage

Elle résulte de l'utilisation d'un sous ensemble de la population (l'échantillon) et non de la population toute entière.
Exemple : le responsable du parti (suite). deux échantillons différents vont fournir des estimations différentes.



Quelle est la précision des estimations réalisées ?

Plan de la section

Un exemple de loi discrète : la loi Binomiale

Un hôtel possède 50 chambres. Au printemps le taux de remplissage est de 75%.
On note X le nombre de chambres occupées un jour donné. C'est une variable aléatoire.
$X \in \{0, \dots, 50\}$ prend un nombre fini de valeurs,
c'est une variable aléatoire discrète.
La loi de X est la loi binomiale de paramètre $n = 50$ et $p = 0.75$.
c'est à dire, pour tout $k \in \{0, \dots, 50\}$, on a

$$P(X = k) = C_{50}^{k} p^{k} (1-p)^{50-k}$$

La probabilité que l'hôtel soit complet vaut

$$P(X = 50) = C_{50}^{50} 0.75^{50} (1 - 0.75)^{0} = 0.75^{50}$$

Plus généralement

- Une variable aléatoire discrète prend un nombre au plus dénombrable de valeurs. L'ensemble des valeurs prises par X peut donc s'écrire de la forme $\{x_i, i \in E\}$ où E est un sous ensemble de $\mathbb{N}$
- La loi de la variable aléatoire X est la suite des probabilités $p_k = P(X = x_k)$ pour tout $k \in E$

L'espérance (moyenne) de X :

$$\mathbb{E}(X) = \sum_{k \in E} p_k x_k$$

La variance de X :

$$\mathrm{var}(X) = \sum_{k \in E} p_k x_k^2 - \left(\sum_{k \in E} p_k x_k \right)^2$$

Un exemple de variable aléatoire non discrète

On note X le temps de vol entre Paris et Vilnius. C'est une variable aléatoire qui prend des valeurs comprises entre 135mn et 165mn.
La variable aléatoire X peut prendre toutes les valeurs de l'intervalle $[135, 165]$.
Cette variable aléatoire n'est donc pas une variable discrète.

Définition

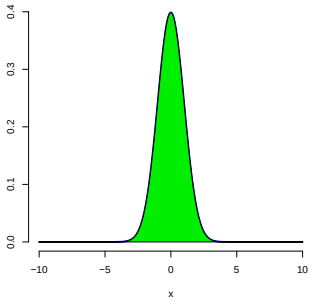
On dit que X est une variable aléatoire continue.

Définition

La loi d'une variable aléatoire continue est définie à partir d'une fonction f appelée densité qui vérifie les propriétés suivantes :

- f est positive
  pour tout $x \in \mathbb{R}$, $f(x) \geq 0$
- l'aire en dessous la courbe représentative de f vaut 1
  autrement dit

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$



Calcul des probabilités

L'aire comme mesure des probabilités

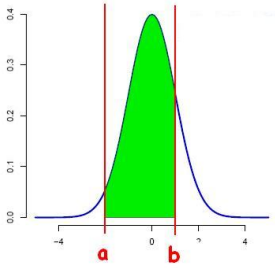
Soit X une variable aléatoire continue, f sa densité

Définition

La probabilité que X appartienne à l'intervalle $[a, b]$ $P(a \leq X \leq b)$ est égale à l'aire en dessous de la courbe représentative de la densité comprise entre $x = a$ et $x = b$

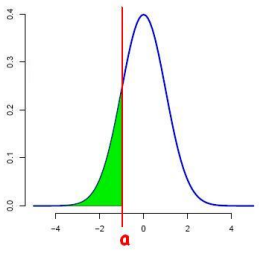
Autrement dit

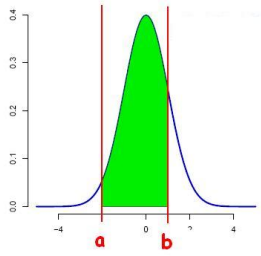
$$P(a \leq X \leq b) = \int_a^b f(t)\mathrm{d}t$$

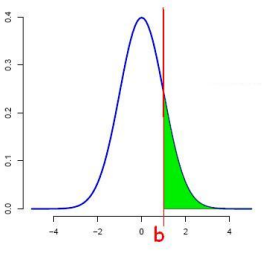


Illustration

1. La courbe en bleu représente la densité de la variable aléatoire
2. L'aire de la zone en vert représente
   - sur l'image de gauche : $P(X \leq a)$
   - sur l'image du milieu : $P(a \leq X \leq b)$
   - sur l'image de droite : $P(X \geq b)$







Définition

X une variable aléatoire continue.
La fonction de répartition de X (notée F) est définie par
$F(x) = P(X \leq x)$

Quelques propriétés

1. $P(X = x) = 0$
2. $P(X \leq x) = P(X < x)$
3. $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$
4. $P(X \geq b) = 1 - P(X \leq b) = 1 - F(b)$

Espérance/Variance

X une variable aléatoire continue de densité f
L'espérance de X s'écrit

$$\mathbb{E}(X) = \int xf(x)\,\mathrm{d}x$$

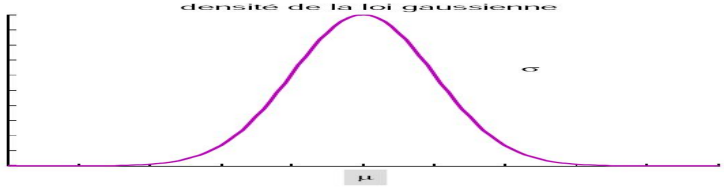
et la variance de X

$$\mathrm{var}(X) = \int x^2 f(x)\,\mathrm{d}x - \left(\int xf(x)\,\mathrm{d}x\right)^2$$

Définition de la loi normale ou gaussienne

La loi gaussienne est une loi continue qui dépend de deux paramètres $\mu \in \mathbb{R}$ et $\sigma > 0$. Sa densité est

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



Définition (Cas particulier)

On dit que la loi gaussienne est standard si $\mu = 0$ et $\sigma = 1$.
On note $F_{0,1}$ sa fonction de répartition.

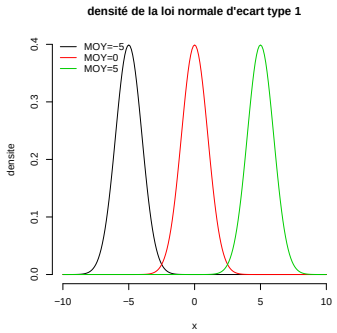
Le rôle des deux paramètres μ, σ

- μ est un paramètre de position
- σ un paramètre de dispersion

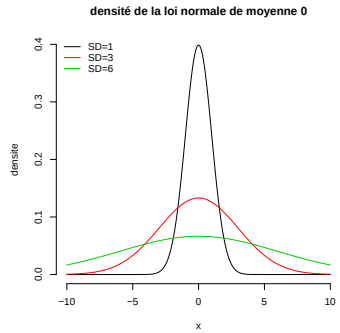
Propriétés

Soit X une variable aléatoire gaussienne.

- $\mathbb{E}(X) = \mu$, la moyenne
- $\mathrm{var}(X) = \sigma^2$, la variance
- σ est l'écart type de X



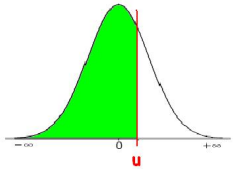
Densités de lois gaussiennes ayant la même variance mais des moyennes différentes



Densités de lois gaussiennes ayant la même moyenne mais des variances différentes

Table de la loi gaussienne standard

La table donne les valeurs de $F_{0,1}(u)$, $u \geq 0$ (aire en vert)



Prenons $u = 1.96 = 1.9 + 0.06$.

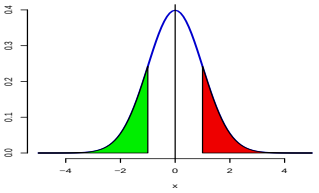
$u = u_1 + u_2$ (u_1 \ u_2)	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.5039	0.5079	0.5119	0.5159	0.5199	0.5239	0.5279	0.5318	0.5358
0.1	0.5398	0.5437	0.5477	0.5517	0.5556	0.5596	0.5635	0.5674	0.5714	0.5753
0.2	0.5792	0.5831	0.587	0.5909	0.5948	0.5987	0.6025	0.6064	0.6102	0.614
0.3	0.6179	0.6217	0.6255	0.6293	0.633	0.6368	0.6405	0.6443	0.648	0.6517
0.4	0.6554	0.659	0.6627	0.6664	0.67	0.6736	0.6772	0.6808	0.6843	0.6879
0.5	0.6914	0.6949	0.6984	0.7019	0.7054	0.7088	0.7122	0.7156	0.719	0.7224
0.6	0.7257	0.729	0.7323	0.7356	0.7389	0.7421	0.7453	0.7485	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7703	0.7733	0.7763	0.7793	0.7823	0.7852
0.8	0.7881	0.791	0.7938	0.7967	0.7995	0.8023	0.8051	0.8078	0.8105	0.8132
0.9	0.8159	0.8185	0.8212	0.8238	0.8263	0.8289	0.8314	0.8339	0.8364	0.8389
1	0.8413	0.8437	0.8461	0.8484	0.8508	0.8531	0.8554	0.8576	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8707	0.8728	0.8749	0.8769	0.8789	0.8809	0.8829
1.2	0.8849	0.8868	0.8887	0.8906	0.8925	0.8943	0.8961	0.8979	0.8997	0.9014
1.3	0.9031	0.9049	0.9065	0.9082	0.9098	0.9114	0.913	0.9146	0.9162	0.9177
1.4	0.9192	0.9207	0.9221	0.9236	0.925	0.9264	0.9278	0.9292	0.9305	0.9318
1.5	0.9331	0.9344	0.9357	0.9369	0.9382	0.9394	0.9406	0.9417	0.9429	0.944
1.6	0.9452	0.9463	0.9473	0.9484	0.9494	0.9505	0.9515	0.9525	0.9535	0.9544
1.7	0.9554	0.9563	0.9572	0.9581	0.959	0.9599	0.9607	0.9616	0.9624	0.9632
1.8	0.964	0.9648	0.9656	0.9663	0.9671	0.9678	0.9685	0.9692	0.9699	0.9706
1.9	0.9712	0.9719	0.9725	0.9731	0.9738	0.9744	0.975	0.9755	0.9761	0.9767
2	0.9772	0.9777	0.9783	0.9788	0.9793	0.9798	0.9803	0.9807	0.9812	0.9816
2.1	0.9821	0.9825	0.9829	0.9834	0.9838	0.9842	0.9846	0.9849	0.9853	0.9857

On a $u_1 = 1.9$ et $u_2 = .06$ d'où $F_{0,1}(1.96) = 0.975$.

Propriétés de la loi gaussienne standard

Soit X une variable aléatoire gaussienne standard.

- Pour tout x, on a $P(X \leq -x) = P(X \geq x)$



- $P(X \leq -x) = 1 - P(X \leq x)$ autrement dit $F_{0,1}(-x) = 1 - F_{0,1}(x)$.
- $P(-x \leq X \leq x) = F_{0,1}(x) - F_{0,1}(-x) = 2F_{0,1}(x) - 1$

Applications

Soit X une variable aléatoire gaussienne standard.

1. En utilisant la table : $P(X \leq 1.96) = F_{0,1}(1.96) = 0.975$
2. Calcul de $P(X \leq -1.96)$. Cette valeur n'est pas dans la table.

$$\begin{aligned} P(X \leq -1.96) &= F_{0,1}(-1.96) = 1 - F_{0,1}(1.96) \\ &= 1 - 0.975 = 0.025 \end{aligned}$$

3. Calcul de $P(-x \leq X \leq x)$ pour $x = 1, 2, 3$

$$\begin{aligned} P(-x \leq X \leq x) &= F_{0,1}(x) - F_{0,1}(-x) \\ &= 2F_{0,1}(x) - 1 \\ &= \begin{cases} 0.68 & x = 1 \\ 0.95 & x = 2 \\ 0.99 & x = 3 \end{cases} \end{aligned}$$

Lien entre les lois gaussiennes

1. Si la loi de X est la loi gaussienne de moyenne μ et d'écart type σ alors la loi de $Y = \frac{X-\mu}{\sigma}$ est la loi gaussienne de moyenne 0 et d'écart type 1
2. Si la loi de Y est la loi gaussienne de moyenne 0 et d'écart type 1 alors la loi de $X = \sigma Y + \mu$ est la loi gaussienne de moyenne μ et d'écart type σ

Calcul pour la loi gaussienne (μ, σ)

Soit X est une variable gaussienne de moyenne μ et d'écart type σ. Pour calculer $P(X \leq x)$, on se ramène à une loi gaussienne standard. On pose

$$Y = \frac{X - \mu}{\sigma} \qquad \Leftrightarrow \qquad X = \sigma Y + \mu$$

$$\begin{aligned} P(X \leq x) &= P(\sigma Y + \mu \leq x) \\ &= P(Y \leq \frac{x - \mu}{\sigma}) \end{aligned}$$

Comme la loi de Y est la loi gaussienne standard, le dernier terme est donné par la table de la loi gaussienne.

$$P(X \leq x) = F_{0,1}\left(\frac{x - \mu}{\sigma}\right)$$

Exemple

Si la loi de X est gaussienne de moyenne 4 et d'écart type 2. On pose $Y = \frac{X-4}{2}$

$$\begin{aligned} P(X \leq 6.5) &= P(2Y + 4 \leq 6.5) \\ &= P(Y \leq \frac{6.5 - 4}{2}) \\ &= P(Y \leq 1.25) = 0.8943 \end{aligned}$$

Plan de la section

La situation

Le directeur du personnel du groupe $\alpha\beta$ a été chargé de développer le profil de 2500 responsables de sociétés appartenant au groupe $\alpha\beta$.
Les caractéristiques à étudier sont

- le salaire moyen annuel et sa dispersion
- la participation au programme de formation en gestion mis en place par la société.

On a donc trois paramètres à calculer

- la moyenne μ et l'écart type σ du salaire annuel pour la population
- la proportion p de la population ayant suivi la formation

Deux méthodes

- Le recensement. On doit interroger 2500 personnes. Le coût de la collecte est très élevé, il nécessite un entretien avec chaque responsable.
- L'estimation. On estime les trois paramètres à partir d'un échantillon de taille $n << 2500$. Il faut alors
  1. Construire un échantillon de taille n
  2. Calculer des estimateurs des trois paramètres
  3. Évaluer la qualité des estimateurs.

On construit un échantillon constitué de 30 responsables de sociétés du groupe.
Pour chaque personne de l'échantillon, on collecte deux informations

- son salaire. On note $S_1, \ldots, S_{30}$ les salaires
- s'il a participé au programme de formation que l'on code par 1 pour oui et 0 pour non. On note $F_1, \ldots, F_{30}$ les réponses

les données collectées

	S	F		S	F		S	F
1	50427.82	1	11	53714.13	1	21	54276.3	1
2	47770.71	1	12	56641.81	1	22	58389.2	1
3	51686.39	1	13	45535.32	0	23	48762.44	0
4	44520.07	1	14	55626.63	1	24	48916.25	0
5	47976.9	0	15	54898.44	0	25	51026.77	1
6	59979.41	1	16	49246.59	0	26	50999.26	1
7	47022.2	1	17	57261.6	1	27	55811.3	1
8	44252.88	1	18	52876.62	0	28	48622.47	1
9	51641.93	1	19	49841.11	1	29	47226.59	0
10	51206.19	1	20	54256.2	0	30	53419.27	1

S = salaire
F = formation (0 :non, 1 :oui)

Caractéristiques de l'échantillon

1. moyenne de l'échantillon : $\bar{x} = 51461.09$
2. écart type de l'échantillon : $S = 4091.18$
3. proportion de l'échantillon ayant suivi le programme de formation : $\bar{p} = .7$

$x_1, \ldots, x_n$ un échantillon de taille n.

- sa moyenne : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- sa variance : $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- son écart type $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Recensement

Après un recensement de la population entière, on obtient

1. moyenne de la population $\mu = 51800$ $\bar{x} = 51461.09$
2. écart type de la population $\sigma = 4000$ $S = 4091.18$
3. proportion de la population ayant suivi le programme de formation $p = .67$ $\bar{p} = .7$

Les valeurs calculées sur l'échantillon ne correspondent pas exactement aux valeurs de la population.

Erreur d'échantillonnage

Évaluation des erreurs

- Erreur absolue : $EA = |\text{estimation} - \text{vraie valeur}|$
- Erreur relative : $ER = \dfrac{EA}{\text{vraie valeur}}$

ici

1. sur la moyenne : $EA = |\bar{x} - \mu| = 338.90$ et $ER = \dfrac{|\bar{x} - \mu|}{\mu} < 0.01\%$
2. Sur l'écart type : $EA = 91.18$ et $ER = 2.2\%$
3. sur la proportion : $EA = .03$ et $ER = 5\%$

Définition d'un échantillon

On suppose que l'on dispose d'un échantillon aléatoire de taille n issu d'une population.

L'échantillon satisfait les conditions suivantes

1. Tous les individus sont sélectionnés dans la même population
2. Les individus sont sélectionnés de façon indépendante.

Estimation d'une moyenne

Soit X une caractéristique/variable de la population. On note

- μ sa moyenne dans la population
- σ son écart type.

Question

Comment estimer le paramètre μ ?
Quelle est la précision de l'estimation ?

Les données

On dispose des valeurs de la variable X pour les n individus sélectionnés dans l'échantillon :

$$x_1, \ldots, x_n$$

Construction de l'estimateur de μ

On estime la moyenne de la population par la moyenne de l'échantillon

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

$\bar{x}$ est une estimation ponctuelle de μ

Remarque

$\bar{x}$ est une variable aléatoire.
À chaque répétition du processus d'échantillonnage, il est vraisemblable d'obtenir une valeur différente pour la moyenne $\bar{x}$.

On peut donc calculer la loi de $\bar{x}$, sa moyenne, sa variance etc

Propriétés de l'estimateur $\bar{x}$

1. La moyenne de $\bar{x}$ est égale à la moyenne de la population μ.

$$\mathbb{E}(\bar{x}) = \mu$$

2. La variance de $\bar{x}$:

$$\mathrm{var}(\bar{x}) = \frac{\sigma^2}{n}$$

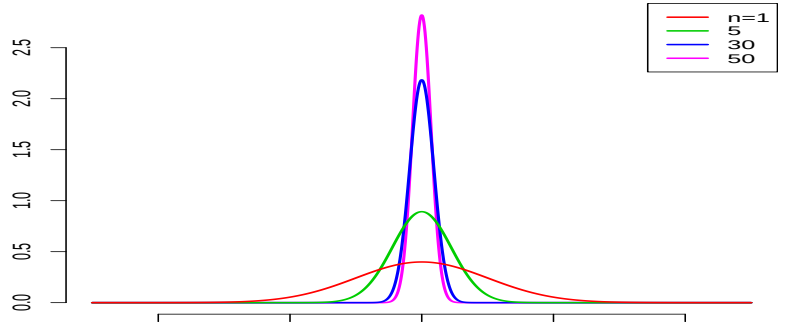
où σ^2 est la variance de la population.

3. L'écart type de $\bar{x}$:

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

- l'écart type décroît vers zéro quand la taille de l'échantillon tend vers l'infini.
- la moyenne reste inchangée quelque soit la taille de l'échantillon n

Graphique Évolution de la loi de $\bar{x}$ en fonction de la taille de l'échantillon.
La population est gaussienne de moyenne $\mu = 10$ et d'écart type $\sigma = 1$



Loi de $\bar{x}$: cas gaussien

Lorsque la distribution de la population est gaussienne alors la loi de $\bar{x}$ est aussi une loi gaussienne

	Population	$\bar{x}$
loi	gaussienne	gaussienne
moyenne	μ	μ
variance	σ^2	$\frac{\sigma^2}{n}$
écart type	σ	$\frac{\sigma}{\sqrt{n}}$

Loi de $\bar{x}$: le cas des grands échantillons

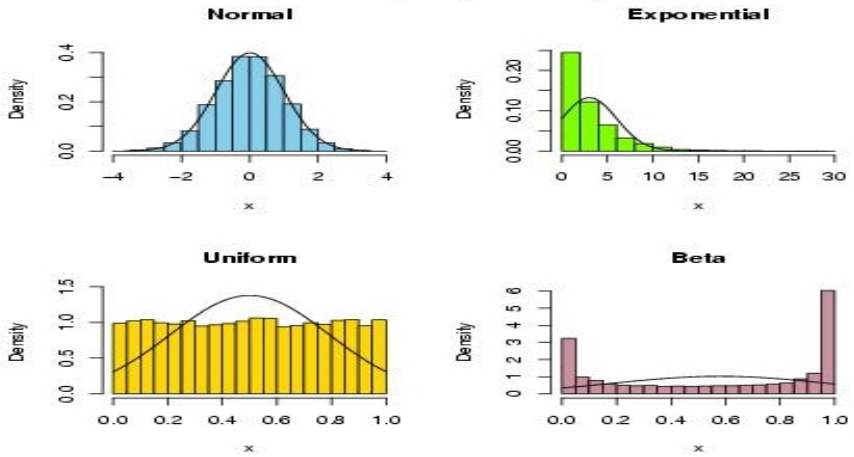
Le théorème central limite donne la loi de $\bar{x}$ pour les grands échantillons quelque soit la loi de la population.

Théorème

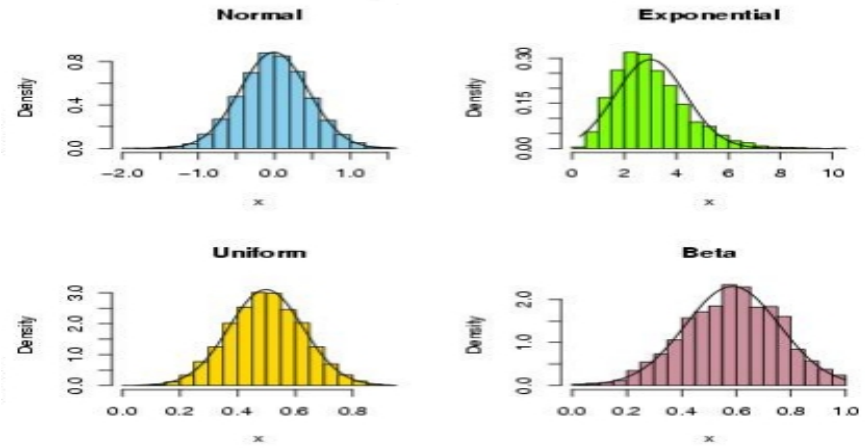
On suppose que la loi de la population est de moyenne μ et d'écart type σ.
Lorsque la taille de l'échantillon n est assez grande, la loi de $\bar{x}$ peut être approchée par une loi gaussienne de moyenne μ et d'écart type $\frac{\sigma}{\sqrt{n}}$.

Illustration du TCL

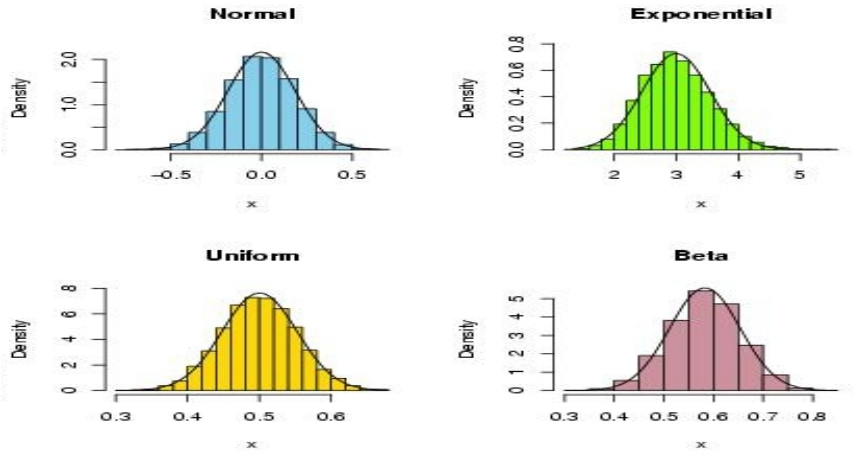
Loi de la population.



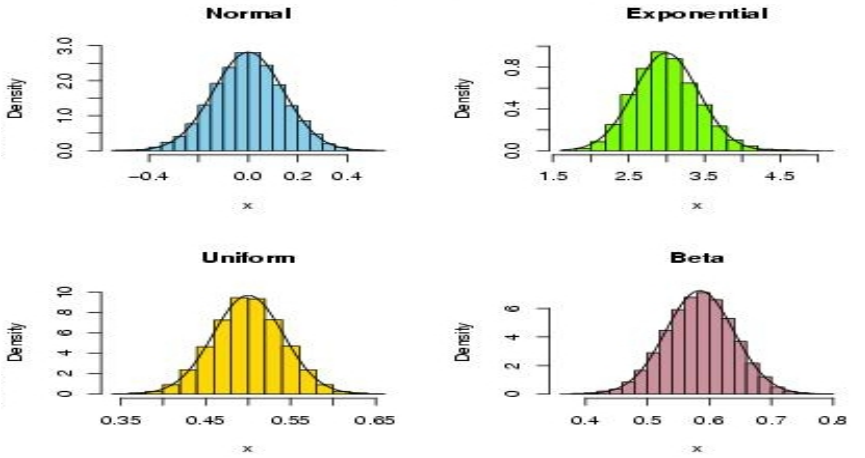
Loi de $\bar{x}$ pour des échantillons de taille $n = 5$



Loi de $\bar{x}$ pour des échantillons de taille $n = 30$



Loi de $\bar{x}$ pour des échantillons de taille $n = 50$



En pratique

On peut approcher la loi de $\bar{x}$ par une loi gaussienne pour des grands échantillons. On admet souvent comme limite $n > 30$.

Remarque

Si la loi de la population est gaussienne alors la loi de $\bar{x}$ est gaussienne quelque soit la taille de l'échantillon.

Remarque

La loi d'échantillonnage révèle la façon dont les valeurs de $\bar{x}$ sont distribuées autour de μ. Nous allons utiliser cette loi

- *pour contrôler l'erreur d'estimation*
- *pour construire une estimation par intervalle.*

Erreur d'estimation : conclusions probabilistes

La connaissance de la loi de $\bar{x}$ permet de tirer des conclusions probabilistes sur l'erreur $|\bar{x} - \mu|$ (même si μ est inconnu)
Les situations étudiées sont les suivantes

- les grands échantillons
  - σ connu
  - σ inconnu
- les petits échantillons pour des populations gaussiennes
  - σ connu
  - σ inconnu

Cas des grands échantillons $n > 30$

D'après le théorème central limite la loi de $\bar{x}$ peut être approchée par une loi gaussienne de moyenne μ et d'écart type $\sigma/\sqrt{n}$.
$\Rightarrow$ la loi de $\sqrt{n}\dfrac{\bar{x} - \mu}{\sigma}$ peut être approchée par une loi gaussienne standard.

Soit Z une variable gaussienne standard. D'après la table de la loi gaussienne, on sait que $P(Z \in [-1,96\ ;\ 1.96]) = 0.95$

En effet
$P(Z \in [-a\ ;\ a]) = 2F_{0,1}(a) - 1 = 0.95$ et $F_{0,1}(1.96) = 0.975$

Par conséquent

$$P\left(\sqrt{n}\frac{\bar{x} - \mu}{\sigma} \in [-1,96\ ;\ 1.96]\right) = 0.95$$

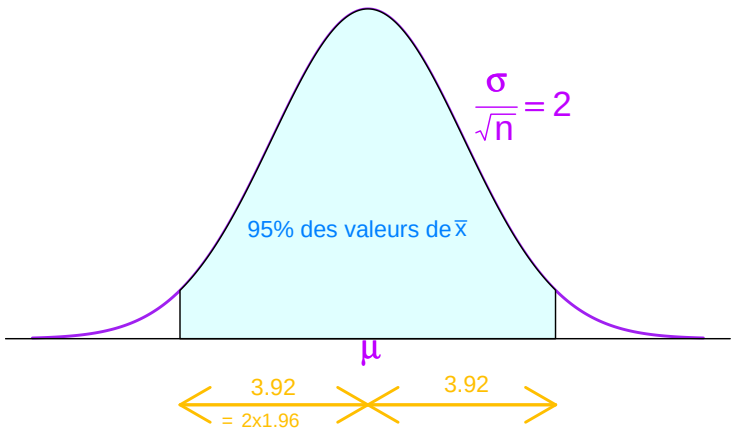
c'est à dire

$$P\left(\bar{x} - \mu \in \left[-1,96\frac{\sigma}{\sqrt{n}}; 1.96\frac{\sigma}{\sqrt{n}}\right]\right) = 0.95$$

Conclusion probabiliste sur l'erreur

95% des valeurs de $\bar{x}$ génèrent une erreur absolue inférieure à $1,96\dfrac{\sigma}{\sqrt{n}}$

Illustration : distribution de la loi de $\bar{x}$

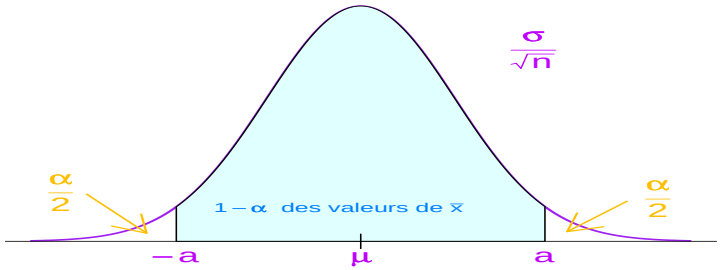


Généralisation

1. On fixe $\alpha \in]0,1[$, $1-\alpha$ est de niveau de confiance.
2. On construit a (qui dépend de α) tel que

$$P(\bar{x} - \mu \in [-a\ ;\ a]) = 1 - \alpha$$

$\bar{x}$ génère une erreur absolue inférieure à a avec une probabilité de $1-\alpha$.

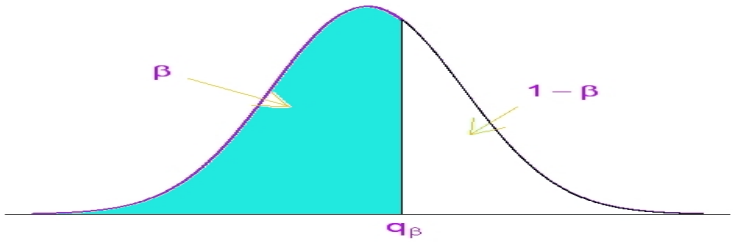


Quantile de la loi gaussienne standard.

Définition

Soit X une variable gaussienne standard.
Le quantile d'ordre β de la loi gaussienne standard est le réel $q(\beta)$ tel que

$$P(X \leq q(\beta)) = \beta \qquad \Longleftrightarrow \qquad F_{0,1}(q(\beta)) = \beta$$



Erreur d'estimation : n grand σ connu

Théorème

Hypothèses

- *la taille de l'échantillon est assez grande (n>30)*
- *la variance de la population σ^2 est connue*

Soit α fixé. On a

$$P\left(\bar{x} - \mu \in \left[-q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}\ ;\ q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha$$

$\bar{x}$ génère une erreur absolue inférieure à $q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$ avec une probabilité de $1-\alpha$.

le calcul ...

On remarque que

$$\bar{x} - \mu \in \left[-q(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \; ; \; q(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \right]$$

$$\Updownarrow$$

$$\frac{\sqrt{n}}{\sigma}(\bar{x} - \mu) \in [-q(1-\alpha/2) \; ; \; q(1-\alpha/2)]$$

Comme la loi de $\frac{\sqrt{n}}{\sigma}(\bar{x} - \mu)$ peut être approchée par la loi gaussienne standard, on a

$$\begin{aligned} P &= P\left(\bar{x} - \mu \in \left[-q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}; q(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \right] \right) \\ &= F_{0,1}(q(1-\alpha/2)) - F_{0,1}(-q(1-\alpha/2)) \\ &= 2F_{0,1}(q(1-\alpha/2)) - 1 = 2(1-\alpha/2) - 1 = 1 - \alpha \end{aligned}$$

Grands échantillons, σ est inconnu

Les intervalles dépendent de l'écart type de la population σ qui généralement est inconnu.
On estime l'écart type de la population par celui de l'échantillon

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Remarque

S^2 est un estimateur ponctuel de la variance de la population σ^2

Théorème

Quand n est assez grand, la loi de $\frac{\sqrt{n}}{S}(\bar{x} - \mu)$ peut être approchée par la loi gaussienne standard.

Erreur d'estimation : *n* grand σ inconnu

Théorème

Hypothèses

- *la taille de l'échantillon est assez grande (n>30)*
- *la variance de la population σ^2 est inconnue*

Soit α fixé. On a

$$P\left(\bar{x} - \mu \in \left[-q(1-\alpha/2)\frac{S}{\sqrt{n}}; q(1-\alpha/2)\frac{S}{\sqrt{n}} \right] \right) = 1 - \alpha$$

$\bar{x}$ génère une erreur absolue inférieure à $q(1-\alpha/2)\frac{S}{\sqrt{n}}$ avec une probabilité de $1-\alpha$.

Cas des petits échantillons gaussiens

Si la loi de la population est gaussienne alors la loi de $\frac{\sqrt{n}}{\sigma}(\bar{x} - \mu)$ est la loi gaussienne standard

Théorème

Hypothèses

- *la population est gaussienne*
- *la variance de la population σ^2 est connue*

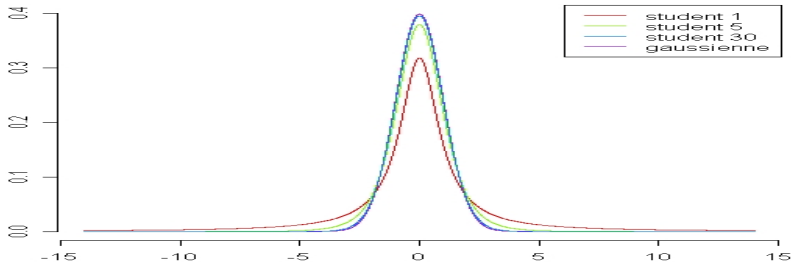
Soit α fixé. On a

$$P\left(\bar{x} - \mu \in \left[-q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}; q(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \right] \right) = 1 - \alpha$$

$\bar{x}$ génère une erreur absolue inférieure à $q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$ avec une probabilité de $1-\alpha$.

Loi de Student

Soit $\nu \in \mathbb{R}^+$. La loi de Student à ν degrés de liberté est une loi continue dont la densité est de la forme

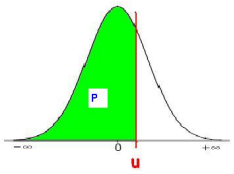


Proposition

Quand le degré de liberté ν est grand, on peut approcher la loi de Student par une loi gaussienne standard

Fonction de répartition des lois de Student

Soit X une variable distribuée suivant la loi de Student à ν degrés de liberté.
$P = P(X \leq u)$ (aire en vert)



si $\nu = 8$ alors $P(X < 1.859) = 0.95$.

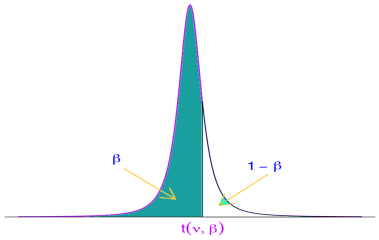
e

ν \ P	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	0.975	0.99
1	0.158	0.324	0.509	0.726	1	1.376	1.962	3.077	6.313	12.706	31.82
2	0.142	0.288	0.444	0.617	0.816	1.06	1.386	1.885	2.919	4.302	6.964
3	0.136	0.276	0.424	0.584	0.764	0.978	1.249	1.637	2.353	3.182	4.54
4	0.133	0.27	0.414	0.568	0.74	0.94	1.189	1.533	2.131	2.776	3.746
5	0.132	0.267	0.408	0.559	0.726	0.919	1.155	1.475	2.015	2.57	3.364
6	0.131	0.264	0.404	0.553	0.717	0.905	1.134	1.439	1.943	2.446	3.142
7	0.13	0.263	0.401	0.549	0.711	0.896	1.119	1.414	1.894	2.364	2.997
8	0.129	0.261	0.399	0.545	0.706	0.888	1.108	1.396	1.859	2.306	2.896
9	0.129	0.26	0.397	0.543	0.702	0.883	1.099	1.383	1.833	2.262	2.821
10	0.128	0.26	0.396	0.541	0.699	0.879	1.093	1.372	1.812	2.228	2.763
11	0.128	0.259	0.395	0.539	0.697	0.875	1.087	1.363	1.795	2.2	2.718
12	0.128	0.259	0.394	0.538	0.695	0.872	1.083	1.356	1.782	2.178	2.68
13	0.128	0.258	0.393	0.537	0.693	0.87	1.079	1.35	1.77	2.16	2.65
14	0.127	0.258	0.393	0.536	0.692	0.868	1.076	1.345	1.761	2.144	2.624
15	0.127	0.257	0.392	0.535	0.691	0.866	1.073	1.34	1.753	2.131	2.602
16	0.127	0.257	0.392	0.535	0.69	0.864	1.071	1.336	1.745	2.119	2.583
17	0.127	0.257	0.391	0.534	0.689	0.863	1.069	1.333	1.739	2.109	2.566
18	0.127	0.257	0.391	0.533	0.688	0.862	1.067	1.33	1.734	2.1	2.552
19	0.127	0.256	0.391	0.533	0.687	0.86	1.065	1.327	1.729	2.093	2.539
20	0.127	0.256	0.39	0.532	0.686	0.859	1.064	1.325	1.724	2.085	2.527

Quantiles de la loi de Student

On note $t(\nu, \beta)$ le quantile d'ordre β de la loi de Student à ν degrés de liberté.

$$P(X \leq t(\nu, \beta)) = \beta$$



Fixons $\beta = 0.975$

ν	1	2	3	20	30	40	500
$t(\nu, 0.975)$	12.706	4.302	3.182	2.085	2.041	2.022	1.960

Pour la loi gaussienne standard, on a $q(0.975) = 1.96$.

Petits échantillons gaussiens, σ inconnu

Important : On commence par corriger l'estimateur de la variance
On pose

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}S^2$$

Définition

S_c^2 est la variance modifiée/corrigée de l'échantillon. C'est un estimateur ponctuel de la variance de la population

Théorème

La loi de $\frac{\sqrt{n}}{S_c}(\bar{x} - \mu)$ est une loi de Student à $n-1$ degrés de liberté.

Erreur d'estimation : population gaussienne, σ inconnu

Théorème

Hypothèses

- *la population est gaussienne*
- *la variance de la population σ^2 est inconnue*

Soit α fixé. On a

$$P\left(\bar{x} - \mu \in \left[-t(n-1, 1-\alpha/2)\frac{S_c}{\sqrt{n}} \; ; \; t(n-1, 1-\alpha/2)\frac{S_c}{\sqrt{n}}\right]\right) = 1 - \alpha$$

$\bar{x}$ génère une erreur absolue inférieure à $t(n-1, 1-\alpha/2)\frac{S_c}{\sqrt{n}}$ avec une probabilité de $1-\alpha$

Estimation par intervalle

À partir de l'échantillon, on souhaite construire un intervalle qui vérifie la propriété suivante :

il y a une probabilité $1-\alpha$ que l'intervalle contienne la moyenne de la population.

Définitions

1. *$1-\alpha$ est le coefficient de confiance.*
2. *L'intervalle obtenu est appelé intervalle de confiance de niveau $1-\alpha$.*

Cas des grands échantillons

Estimation par intervalle de la moyenne d'une population

Hypothèses

- la taille de l'échantillon est assez grande (n>30)
- la variance de la population σ^2 est connue

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}}q(1-\alpha/2) \; ; \; \bar{x} + \frac{\sigma}{\sqrt{n}}q(1-\alpha/2)\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ

il y a une probabilité $1-\alpha$ que l'intervalle de confiance contienne la moyenne de la population.

le calcul

Il y a une probabilité $1-\alpha$ que la valeur de $\bar{x}$ génère une erreur inférieure à $\frac{\sigma}{\sqrt{n}}q(1-\alpha/2)$ d'où

$$P(|\bar{x}-\mu| \leq \frac{\sigma}{\sqrt{n}}q(1-\alpha/2)) = 1-\alpha$$

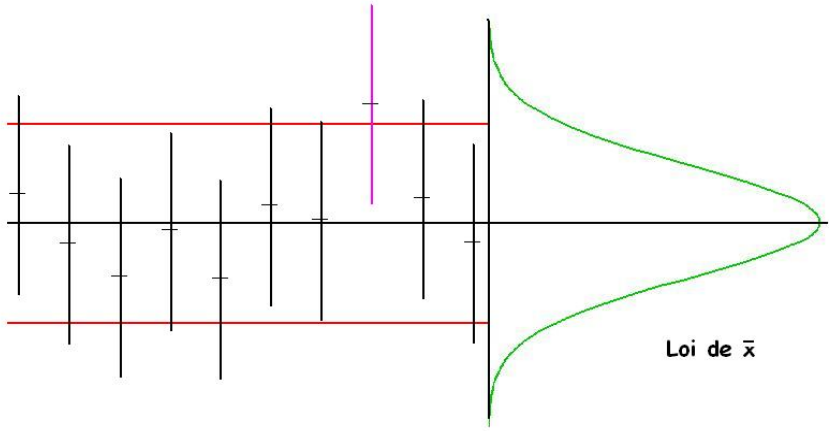
Ensuite, il suffit de remarquer que

$$|\bar{x}-\mu| \leq \frac{\sigma}{\sqrt{n}}q(1-\alpha/2)$$

$$\Updownarrow$$

$$\mu \in \left[\bar{x} - \frac{\sigma}{\sqrt{n}}q(1-\alpha/2)\ ;\ \bar{x} + \frac{\sigma}{\sqrt{n}}q(1-\alpha/2)\right]$$

La courbe en vert est la densité de la loi de $\bar{x}$.
On construit 10 intervalles de confiance de niveau 95% à partir de 10 échantillons différents.



L'intervalle en rose ne contient pas la vraie valeur de la moyenne.

Cas des grands échantillons, σ inconnu

On estime σ par l'écart type de l'échantillon S

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Estimation par intervalle de la moyenne d'une population

Hypothèses

- la taille de l'échantillon est assez grande (n>30)
- la variance de la population σ^2 est inconnue

$$\left[\bar{x} - \frac{S}{\sqrt{n}}q(1-\alpha/2)\ ;\ \bar{x} + \frac{S}{\sqrt{n}}q(1-\alpha/2)\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ

Petits échantillons gaussiens, σ connu

On retrouve le résultat des grands échantillons.

Estimation par intervalle de la moyenne d'une population

Hypothèses

- la population est gaussienne
- la variance de la population σ^2 est connue

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}}q(1-\alpha/2)\ ;\ \bar{x} + \frac{\sigma}{\sqrt{n}}q(1-\alpha/2)\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ

Petits échantillons gaussiens, σ inconnu

On utilise l'écart type corrigé de l'échantillon S_c pour estimer σ

$$S_c = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Estimation par intervalle de la moyenne d'une population

Hypothèses

- la population est gaussienne
- la variance de la population σ^2 est inconnue

$$\left[\bar{x} - \frac{S_c}{\sqrt{n}}t(n-1, 1-\alpha/2)\ ;\ \bar{x} + \frac{S_c}{\sqrt{n}}t(n-1, 1-\alpha/2)\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ.

Retour à l'exemple du groupe $\alpha\beta$

On suppose que la population est gaussienne.
Situation 1 On dispose d'un échantillon de taille 30 et la variance de la population est connue.
Avec une probabilité de 95%, l'erreur est inférieure à

$$1.96\sigma\frac{1}{\sqrt{n}} = 1.96 * 4000/\sqrt{30} = 1431.382$$

L'intervalle de confiance au niveau 95% est
$[51461.09 - 1431.38\ ;\ 51461.09 + 1431.38] = [50029.7\ ;\ 52892.4]$

Remarque

Sur l'échantillon sélectionné, nous avions $EA = |\bar{x} - \mu| = 338.90$ *après recensement. Le cas observé appartient aux* 95% *des cas favorables.*

Situation 2 On suppose que la loi des salaires est gaussienne. La variance de la population est inconnue.

- Calcul de la variance modifiée $S_c^2 = S^2 30/29$. D'où $S_c = \sqrt{S^2 30/29} = 4161.12$
- Dans la table de la loi de Student , on trouve $t(29, 0.975) = 2.04$

Avec une probabilité de 95%, l'erreur est inférieure à
$2.04 * 4161.1/\sqrt{30} = 1553.78$
L'intervalle de confiance au niveau 95% est

$$[51461.09 - 1553.78\ ;\ 51461.09 + 1553.78] = [49907.31\ ;\ 53014.87]$$

Pour résumer

Les intervalles de confiance sur la moyenne de la population

	petits échantillons loi gaussienne	grands échantillons quelle que soit la loi
σ connu	$\left[\bar{x} \pm \frac{\sigma}{\sqrt{n}}q(1-\alpha/2)\right]$	$\left[\bar{x} \pm \frac{\sigma}{\sqrt{n}}q(1-\alpha/2)\right]$
σ inconnu	$\left[\bar{x} \pm \frac{S_c}{\sqrt{n}}t(n-1, 1-\alpha/2)\right]$	$\left[\bar{x} \pm \frac{S}{\sqrt{n}}q(1-\alpha/2)\right]$

Notations :

- $[a \pm b]$ est l'intervalle $[a-b; a+b]$
- $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ et $S_c = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $q(\beta)$ est le quantile d'ordre β de la loi gaussienne standard et $t(\nu, \beta)$ celui de la loi de Student à ν degrés de liberté

Construction de l'estimateur

On souhaite estimer la variance de la population.
1er estimateur : On estime la variance de la population par la variance de l'échantillon

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Remarque (estimation biaisée)

$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$ *on dit que l'estimateur a un biais.*

2ème estimateur : On améliore l'estimateur S^2 en prenant la variance modifiée

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Le biais est corrigé, on a $\mathbb{E}(S_c^2) = \sigma^2$

Propriétés de S_c^2

- La moyenne de S_c^2 est égale à la variance de la population μ

$$\mathbb{E}(S_c^2) = \sigma^2$$

- La variance de S_c^2 converge vers zéro pour des variables L^4. De plus si l'échantillon est gaussien, on a

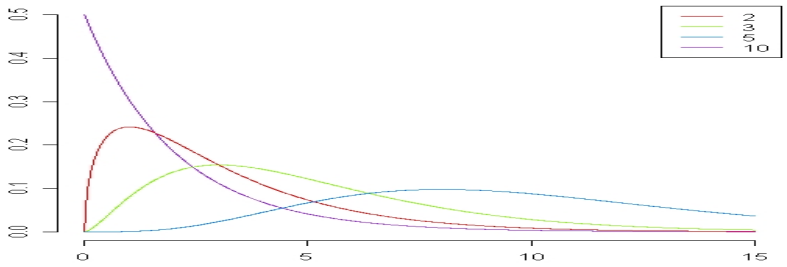
$$\mathrm{var}(S_c^2) = \sigma^4 \frac{2}{n-1}$$

Comparaison des deux estimateurs

Quand la taille de l'échantillon est grande, les deux estimateurs sont équivalents.

Loi du χ^2

Soit $\nu \in \mathbb{R}^+$. La loi du χ^2 à ν degrés de liberté est une loi continue. La densité est de la forme

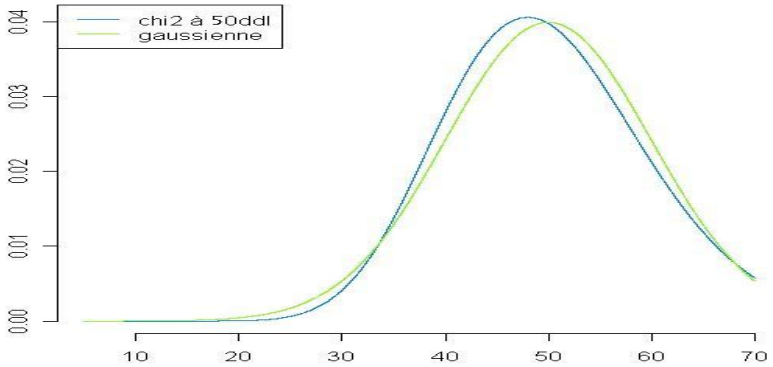


Remarque

La densité est nulle sur $\mathbb{R}^-$ *donc* $P(X < 0) = 0$ *et* $P(X \geq 0) = 1$

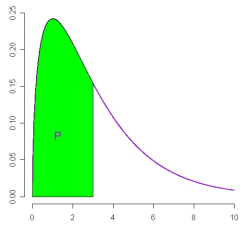
Proposition

Quand le degré de liberté ν est grand, on peut approcher la loi du χ^2 par la loi gaussienne de moyenne ν et d'écart type $\sqrt{2\nu}$



Fonction de répartition des lois du χ^2

Soit X une variable distribuée suivant la loi du χ^2 à ν degrés de liberté.
$P = P(X \leq u)$



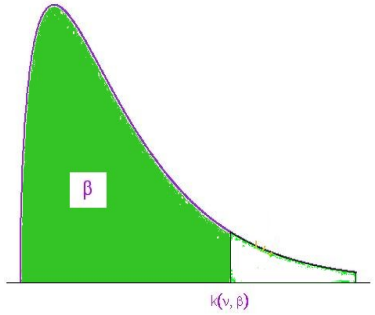
si $\nu = 5$ alors $P(X < 11.07) = 0.95$.

ν \ p	0.01	0.025	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.99	0.995
1	0.0002	0.001	0.0039	0.01	0.1	0.45	1.32	2.7	3.84	5.02	6.63	7.87
2	0.02	0.05	0.1	0.21	0.57	1.38	2.77	4.6	5.99	7.37	9.21	10.59
3	0.11	0.21	0.35	0.58	1.21	2.36	4.1	6.25	7.81	9.34	11.34	12.83
4	0.29	0.48	0.71	1.06	1.92	3.35	5.38	7.77	9.48	11.14	13.27	14.86
5	0.55	0.83	1.14	1.61	2.67	4.35	6.62	9.23	11.07	12.83	15.08	16.74
6	0.87	1.23	1.63	2.2	3.45	5.34	7.84	10.64	12.59	14.44	16.81	18.54
7	1.23	1.68	2.16	2.83	4.25	6.34	9.03	12.01	14.06	16.01	18.47	20.27
8	1.64	2.17	2.73	3.48	5.07	7.34	10.21	13.36	15.5	17.53	20.09	21.95
9	2.08	2.7	3.32	4.16	5.89	8.34	11.38	14.68	16.91	19.02	21.66	23.58
10	2.55	3.24	3.94	4.86	6.73	9.34	12.54	15.98	18.3	20.48	23.2	25.18
11	3.05	3.81	4.57	5.57	7.58	10.34	13.7	17.27	19.67	21.92	24.72	26.75
12	3.57	4.4	5.22	6.3	8.43	11.34	14.84	18.54	21.02	23.33	26.21	28.29
13	4.1	5	5.89	7.04	9.29	12.33	15.98	19.81	22.36	24.73	27.68	29.81
14	4.66	5.62	6.57	7.78	10.16	13.33	17.11	21.06	23.68	26.11	29.14	31.31
15	5.22	6.26	7.26	8.54	11.03	14.33	18.24	22.3	24.99	27.48	30.57	32.8
16	5.81	6.9	7.96	9.31	11.91	15.33	19.36	23.54	26.29	28.84	31.99	34.26
17	6.4	7.56	8.67	10.08	12.79	16.33	20.48	24.76	27.58	30.19	33.4	35.71
18	7.01	8.23	9.39	10.86	13.67	17.33	21.6	25.98	28.86	31.52	34.8	37.15
19	7.63	8.9	10.11	11.65	14.56	18.33	22.71	27.2	30.14	32.85	36.19	38.58
20	8.26	9.59	10.85	12.44	15.45	19.33	23.82	28.41	31.41	34.16	37.56	39.99

Quantiles de la loi du χ^2

On note $k(\nu, \beta)$ le quantile d'ordre β de la loi du χ^2 à ν degrés de liberté.

$$P(X \leq k(\nu, \beta)) = \beta$$



Fixons $\beta = 0.975$

ν	1	3	5	10	20	500
$k(\nu, 0.975)$	5.02	9.35	12.83	20.48	34.17	563.85

Pour la loi gaussienne de moyenne 500 et d'écart type $\sqrt{1000}$, le quantile supérieur d'ordre $\beta = 0.975$ vaut 561.97

Loi de l'estimateur S_c^2

Théorème

Si la population est gaussienne alors la loi de $\frac{n-1}{\sigma^2} S_c^2$ est la loi du χ^2 à $n-1$ degrés de liberté.

Grands échantillons gaussien

Quand la taille de la population est assez grande ($n > 30$), on peut approcher la loi de $\frac{n-1}{\sigma^2} S_c^2$ par la loi gaussienne de moyenne $n-1$ et d'écart type $\sqrt{2n-2}$.

Autrement dit on peut approcher la loi de $\left(\frac{S_c^2}{\sigma^2} - 1\right) \frac{\sqrt{n-1}}{\sqrt{2}}$ par la loi gaussienne standard

Intervalle de confiance pour la variance

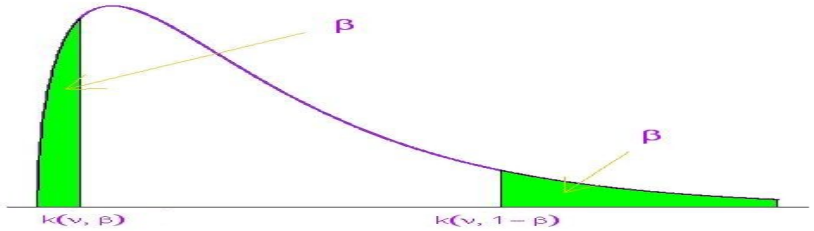
Estimation par intervalle de la variance d'une population

Hypothèses

- la population est gaussienne

$$\left[\frac{(n-1)S_c^2}{k(n-1,1-\alpha/2)} ; \frac{(n-1)S_c^2}{k(n-1,\alpha/2)}\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour la variance σ^2



Approximation gaussienne

Estimation par intervalle de la variance d'une population gaussienne

Quand la taille de l'échantillon est assez grande $n > 30$,

$$\left[\frac{S_c^2}{1+\frac{q(1-\alpha/2)\sqrt{2}}{\sqrt{n-1}}} ; \frac{S_c^2}{1-\frac{q(1-\alpha/2)\sqrt{2}}{\sqrt{n-1}}}\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour la variance σ^2

Construction de l'estimateur

On étudie une caractéristique X qui prend deux modalités $\{0,1\}$.
Soit p la proportion de la population qui possède la modalité 1
On veut estimer p à partir de notre échantillon.

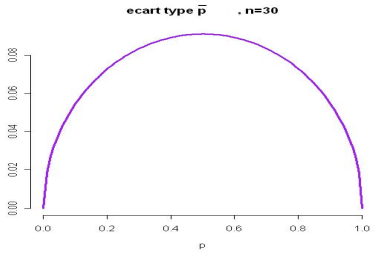
Construction de l'estimateur

On note $\bar{p}$ la proportion de l'échantillon qui possède la modalité 1.
C'est un estimateur ponctuel de p

Propriétés de la loi de $\bar{p}$

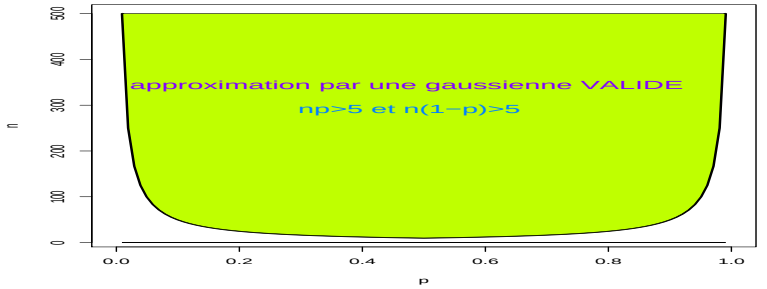
1. La moyenne de la variable $\bar{p}$ est égale à la proportion p dans la population.
2. L'écart type de $\bar{p}$ vaut $\sqrt{\frac{p(1-p)}{n}}$.

Le graphique suivant représente l'écart type en fonction de p.



Loi d'échantillonnage de $\bar{p}$

Quand la taille de l'échantillon est assez grande, on peut approcher la loi de $\bar{p}$ par une loi gaussienne de moyenne p et d'écart type $\sqrt{\frac{p(1-p)}{n}}$.
On peut considérer que n est grand si $np \geq 5$ et $n(1-p) \geq 5$.



Précision de l'estimation : grands échantillons

Soit α fixé. On a

$$P\left(\bar{p} - p \in \left[-q(1-\alpha/2)\sqrt{\frac{p(1-p)}{n}}\ ;\ q(1-\alpha/2)\sqrt{\frac{p(1-p)}{n}}\right]\right) = 1-\alpha$$

$\bar{p}$ génère une erreur absolue inférieure à $q(1-\alpha/2)\sqrt{\frac{p(1-p)}{n}}$ avec une probabilité de $1-\alpha$.

Remarque

L'erreur dépend de p qui est inconnu.

Estimation par intervalle : grands échantillons

On estime l'écart type de la loi de $\bar{p}$ par $\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

Théorème

Pour n assez grand, la loi de

$$\sqrt{\frac{n}{\bar{p}(1-\bar{p})}}(\bar{p} - p)$$

peut être approchée par la loi gaussienne standard.

Intervalle de confiance

Estimation par intervalle de la proportion p

Hypothèse

- la taille de l'échantillon est assez grande $np \geq 5$ et $n(1-p) \geq 5$. en pratique on vérifie si $\bar{p}n \geq 5$ et $n(1-\bar{p}) \geq 5$

$$\left[\bar{p} - q(1-\alpha/2)\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \; ; \; \bar{p} + q(1-\alpha/2)\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour la proportion p

Retour à l'exemple du groupe $\alpha\beta$

L'estimation de p : $\bar{p} = .7$ et la taille de l'échantillon est $n = 30$. On a bien $\bar{p}n = 21 \geq 5$ et $n(1-\bar{p}) = 9 \geq 5$
On peut utiliser l'approximation par une gaussienne

1. Avec une probabilité de 95%, l'erreur sur l'estimation de p est inférieure à

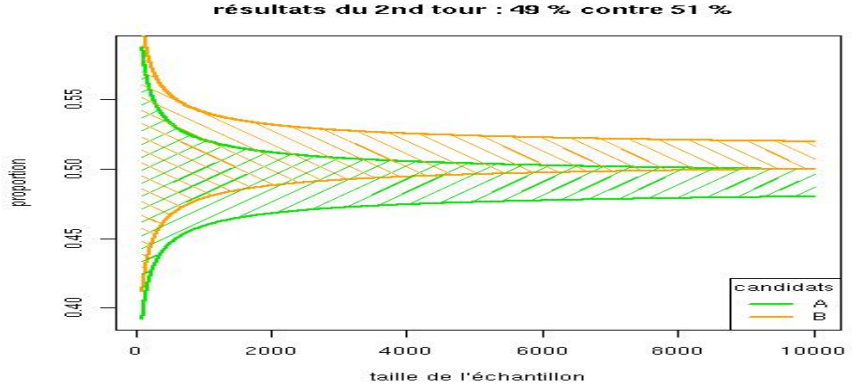
$$1.96\frac{\sqrt{\bar{p}(1-\bar{p})}}{\sqrt{n}} = 1.96 * \sqrt{0.3 * 0.7}/\sqrt{30} = 0.16$$

   Après le recensement, nous avions une erreur absolue de : $EA = .03$
2. L'intervalle de confiance au niveau 95% est

$$[0.7 - 0.16, 0.7 + 0.16] = [0.54, 0.86]$$

Le second tour d'une élection présidentielle

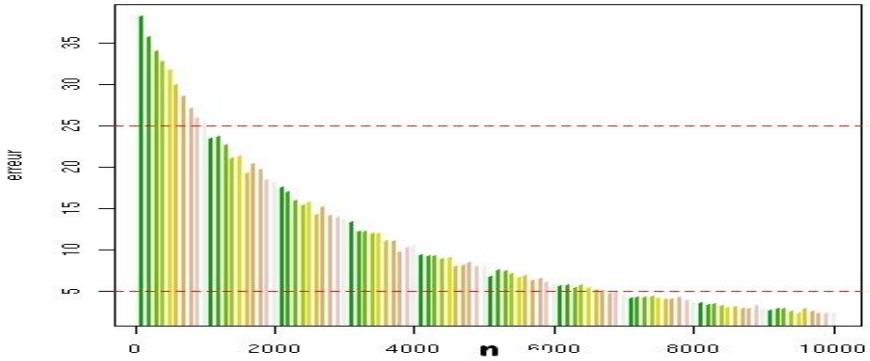
A et B sont les deux candidats présents au second tour. Les résultats du second tour sont B 51% et A 49%
Les régions de confiance pour les deux proportions en fonction de la taille de l'échantillon



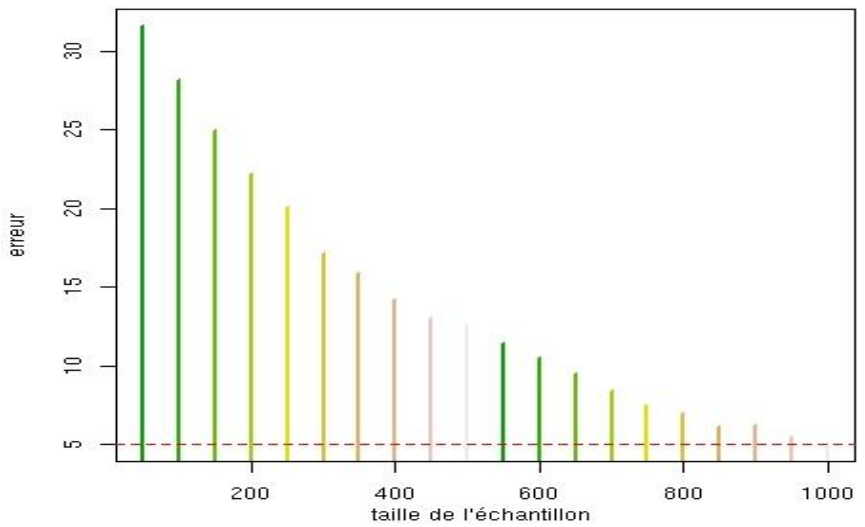
Incertitude sur le candidat vainqueur

Quelle est la précision des sondages ?

On réalise de nombreux sondages sur des échantillons de taille n afin d'évaluer le pourcentage de sondages qui ne donnent pas le bon candidat vainqueur. Ce graphique représente ce pourcentage en fonction de n.



un autre résultat : 52,5% contre 47.5%



La bonne démarche

La démarche statistique pour estimer une caractéristique/un paramètre de la population (moyenne, variance, proportion, etc.) est la suivante

1. On constitue un échantillon de taille n
2. On récolte les observations $x_1, \ldots, x_n$
3. On calcule l'estimateur du paramètre d'intérêt.
4. Avant d'évaluer la qualité de l'estimateur, on doit répondre aux questions suivantes :
   1. Dispose-t-on d'un grand échantillon ?
   2. La population est-elle gaussienne ?
5. On fixe un niveau de confiance $1 - \alpha$
6. On calcule l'erreur d'estimation et/ou un intervalle de confiance

Plan de la section

Un test statistique

Dans la première partie du cours un échantillon est utilisé pour estimer les paramètres d'une caractéristique de la population, par exemple

- une moyenne
- une variance
- une proportion

Nous poursuivons l'inférence statistique par la description des tests statistiques.

> *Un test statistique est utilisé pour déterminer si une assertion sur une caractéristique de la population doit être rejetée.*

Le contrôle de qualité.

Dans une des entreprises du groupe $\alpha\beta$, on procède à l'assemblage de 10 composants électroniques sur une plate-forme.
La qualité de soudure sur la plate-forme ne satisfait pas les critères de qualité établis pour ce produit.

l'avis de l'ingénieur

Un ingénieur a émis l'hypothèse que le problème serait dû à des défauts de placage sur les plates-formes.

Question

La proportion de plates-formes défectueuses dans les stocks de l'entreprise est-elle supérieure à celle annoncée par le fournisseur ?

Principe général

Étape 1 On commence par formuler une première hypothèse sur une caractéristique de la population.
Cette hypothèse, notée H_0, est appelée l'hypothèse nulle.

Étape 2 On définit ensuite une seconde hypothèse qui contredit l'hypothèse nulle H_0. Cette hypothèse, notée H_a, est appelée l'hypothèse alternative.

Étape 3 On utilise les données issues d'un échantillon pour tester les deux hypothèses en compétition H_0 et H_a.

Illustration

Situation : Une société de transport annonce que la durée moyenne μ du trajet entre Paris et Lille a été réduite de 5 minutes, la durée moyenne du trajet serait de 58mn au lieu de 1h03. Une association d'usagers conteste cette annonce.

Les hypothèses On confronte les deux hypothèses suivantes :

- H_0 : l'affirmation de l'association d'usagers $\mu = 63mn$
- H_a : l'affirmation de la société de transport $\mu = 58mn$

On dispose d'un échantillon de taille $n = 35$ dont la moyenne des durées de trajet vaut $\bar{x} = 59.1mn$ et l'écart type $S = 5.1mn$.

La différence entre $\bar{x}$ et 63 peut-elle être attribuée aux fluctuations de l'échantillonnage ou doit-elle être attribuée à une réduction réelle de la durée du trajet ?

Quelle décision peut-on prendre ?

Remarques

- Quelle est la probabilité de commettre une erreur si H_0 est vraie ?
- Quelle est la probabilité de commettre une erreur si H_a est vraie ?

la société de transport (suite) la loi de $\bar{x}$

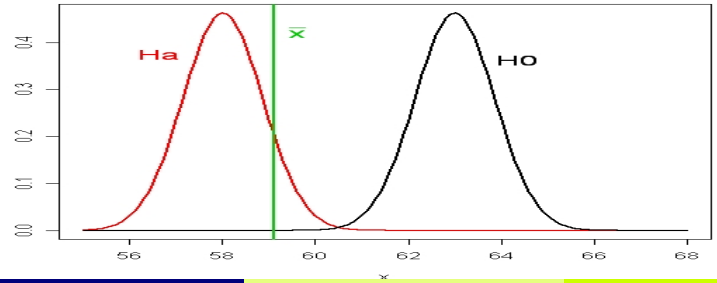
L'hypothèse H_0 est vraie

la loi de $\bar{x}$ peut être approchée par la loi gaussienne de moyenne 63 et d'écart type $\frac{5.1}{\sqrt{35}} \approx 0.86$

L'hypothèse H_a est vraie

la loi de $\bar{x}$ peut être approchée par la loi gaussienne de moyenne 58 et d'écart type $\frac{5.1}{\sqrt{35}}$

Représentation de la loi de $\bar{x}$

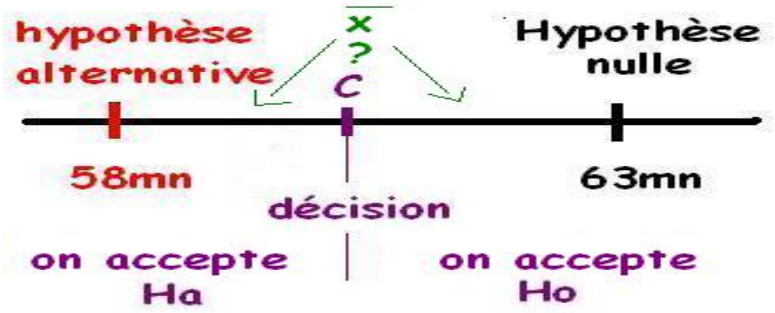


la société de transport (suite)

Supposons que l'hypothèse H_0 soit vraie.
On calcule la probabilité d'observer une valeur inférieure à 59.1.
On a

$$\begin{aligned} P_0(\bar{x} \leq 59.1) &= P\left(\frac{\bar{x} - 63}{0.86} \leq \frac{59.1 - 63}{0.86}\right) \\ &= F_{0,1}(-4.53) \\ &= 1 - F_{0,1}(4.53) \\ &\approx 310^{-6} \end{aligned}$$

la société de transport (suite)



Comment choisir la limite c ?

On fixe $\alpha = 5\%$, la probabilité de commettre une erreur quand H_0 est vraie, autrement dit α est la probabilité que $\bar{x} < c$ quand H_0 est vraie.

la société de transport (suite)

Autrement dit on cherche la valeur c telle que

1. la loi de $\bar{x}$ peut être approchée par la loi gaussienne de moyenne 63 et d'écart type $\dfrac{5.1}{\sqrt{35}}$
2. $P_0(\bar{x} < c) = 0.05$

$$P_0(\bar{x} < c) = P\left(\frac{\bar{x} - 63}{0.86} < \frac{c - 63}{0.86}\right)$$
$$= F_{0,1}\left(\frac{c - 63}{0.86}\right) = 0.05$$

d'où

$$F_{0,1}\left(-\frac{c - 63}{0.86}\right) = 0.95$$

et

$$-\frac{c - 63}{0.86} = 1.64 \quad \Rightarrow \quad c = 61.58$$

la société de transport (fin)

La décision

On a observé $\bar{x} = 59.1$. Comme $\bar{x} < c = 61.58$, on décide de rejeter l'hypothèse nulle (on accepte la réduction de la durée du trajet) pour le test de seuil $\alpha = 5\%$.

Un autre type d'erreur

On calcule la probabilité de décider H_0 alors que H_a est vraie

La loi de $\bar{x}$ peut être approchée par la loi gaussienne de moyenne 58 et d'écart type $\dfrac{5.1}{\sqrt{35}}$

$$P_1(\bar{x} > 61.58) = P\left(\frac{\bar{x} - 58}{0.86} > \frac{61.58 - 58}{0.86}\right)$$
$$= 1 - F_{0,1}\left(\frac{61.58 - 58}{0.86}\right) = 10^{-5}$$

Décision et erreur

On teste les hypothèses H_0 contre H_a

		État de la population	
		H_0 est vraie	H_a est vraie
Décision	Accepter H_0	Décision correcte	Erreur de 2nde espèce
	Rejeter H_0	Erreur de 1ère espèce	Décision correcte

Notations :

- α est la probabilité de commettre une erreur de première espèce
- β est la probabilité de commettre une erreur de seconde espèce

La démarche

1. On fixe la probabilité d'erreur de première espèce α
   $\rightsquigarrow$ c'est le risque de rejeter H_0 (accepter H_a) alors que H_0 est vraie.
2. On construit une région R_0 telle que
   - si $\bar{x} \in R_o$ alors on rejette l'hypothèse nulle H_0 (on accepte H_a)
   - la probabilité de $\bar{x} \in R_o$ est égale à α quand H_0 est vraie

Définition

On dit que la décision est prise au niveau α

Remarque

La probabilité d'erreur de seconde espèce β n'est pas fixée par le statisticien qui met en œuvre le test.
Pour de nombreux tests, il n'est pas possible de calculer la valeur de β.

Les décisions

La décision est prise à partir d'un échantillon de taille n.
On calcule la moyenne de l'échantillon $\bar{x}$.

- Si $\bar{x} \in R_o$ alors on décide de rejeter H_0 (d'accepter H_a).
  Le risque de commettre une erreur est inférieur ou égal à α.
- Si $\bar{x} \notin R_o$ alors on décide d'accepter H_0.

Remarque

Lorsque β est inconnu, on utilise plutôt l'expression "on ne peut pas rejeter H_0" plutôt que "on accepte H_0".
Utiliser cette expression permet de différer tout jugement et toute action.

Tester les hypothèses de recherche

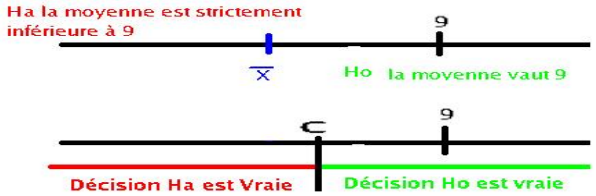
Situation : Les voitures de type XYZ consomment en moyenne, 9 litres d'essence tous les 100 kilomètres. Des chercheurs ont développé un nouveau moteur pour ce modèle.

Hypothèses : Les chercheurs veulent prouver que le nouveau moteur est plus économique.
On note μ la consommation moyenne en litres pour 100 kilomètres.
L'hypothèse de recherche est $\mu < 9$
Les hypothèses appropriées sont

$$\mathrm{H}_0 : \mu = 9 \text{ et } \mathrm{H}_a : \mu < 9$$

Construction du test sur la consommation

- On mesure la consommation sur un échantillon de 100 voitures équipées du nouveau moteur. On calcule la moyenne $\bar{x}$

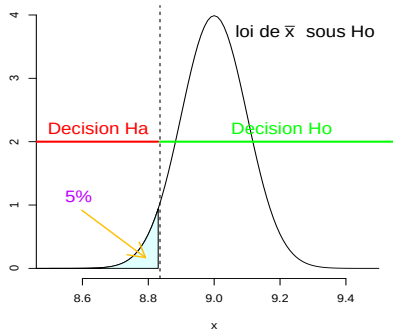


Si $\bar{x} \leq C$
alors on accepte $\mathrm{H_a}$
sinon on accepte $\mathrm{H_0}$

- Comment fixer la limite C ?
  1. On fixe l'erreur de première espèce $\alpha = 0.05$
  2. On cherche la valeur de C telle que si $\mathrm{H_0}$ est vraie $[\mu = 9]$, on a

$$P(\text{accepter } \mathrm{H_a}) = P(\bar{x} < C) = 0.05$$

On dispose d'un grand échantillon $n = 100 > 30$ et $\sigma = 1$ est connu.
Si $\mathrm{H_0}$ est vraie alors la loi de $Z = \dfrac{\bar{x} - 9}{1/\sqrt{100}}$ peut être approchée par une loi gaussienne standard



On cherche C telle que

$$P(\bar{x} < C) = P(Z \leq \frac{C - 9}{1/\sqrt{100}}) = 0.05$$

Dans la table, on lit $\dfrac{C - 9}{1/\sqrt{100}} = -1.64$ donc $C = 8.83$

Si $\bar{x} < 8.83$ alors on rejette l'hypothèse nulle (on accepte l'hypothèse alternative) au niveau 5%

- Sur l'échantillon constitué par les ingénieurs, la moyenne des consommations est égale à $\bar{x} = 8.5$.
- Les résultats de l'échantillon indiquent que l'on rejette $\mathrm{H_0}$ et donc que l'on accepte $\mathrm{H_a}$ au niveau 5%
- Les ingénieurs ont le support statistique nécessaire pour affirmer que le nouveau moteur est plus économique.
  La production pourra alors commencer.

Tester la validité d'une assertion

Situation : Un producteur de tiges filetées prétend que la longueur moyenne μ des tiges est d'un mètre.
Un échantillon de tiges est constitué et leur longueur est mesurée pour tester l'affirmation du fabricant.

Hypothèses : On accorde le bénéfice du doute au producteur et son assertion correspond à $\mathrm{H_0}$.
On formule les hypothèses

$$\mathrm{H_0} : \mu = 1 \text{ et } \mathrm{H_a} : \mu \neq 1$$

Construction du test sur la qualité des pièces

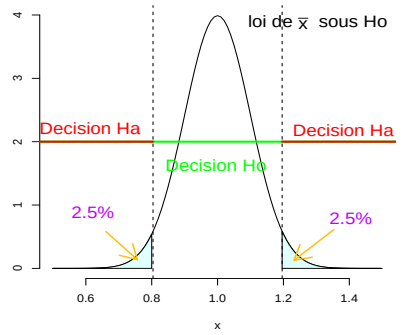
- On mesure la longueur de 100 tiges. On calcule la moyenne $\bar{x}$

Si $|\bar{x} - 1| \geq C$
alors on accepte $\mathrm{H_a}$
sinon on accepte $\mathrm{H_0}$

- Comment fixer la limite C ?
  1. On fixe l'erreur de première espèce $\alpha = 0.05$
  2. On cherche la valeur de C telle que si $\mathrm{H_0}$ est vraie $[\mu = 1]$ alors

$$P(\text{accepter } \mathrm{H_a}) = P(|\bar{x} - 1| > C) = 0.05$$

On dispose d'un grand échantillon $n = 100 > 30$ et $\sigma = 1$ est connu.
Si $\mathrm{H_0}$ est vraie alors la loi de $Z = \dfrac{\bar{x} - 1}{1/\sqrt{100}}$ peut être approchée par une loi gaussienne standard



On cherche C telle que

$$P(|\bar{x} - 1| > C) = P(|Z| \geq \frac{C}{1/\sqrt{100}}) = 0.05$$

Dans la table, on lit $\dfrac{C}{1/\sqrt{100}} = 1.96$ donc $C = 0.19$

Si $\bar{x} < 0.81$ ou $\bar{x} > 1.19$ alors on rejette l'hypothèse nulle (autrement dit on accepte l'hypothèse alternative) au niveau 5%.

- Sur l'échantillon de tiges, la longueur moyenne des tiges est $\bar{x} = 1.1$.
- Les données de l'échantillon ne permettent pas de rejeter $\mathrm{H_0}$. On accepte $\mathrm{H_0}$.
- On ne peut pas contester l'affirmation du fabricant.

Les différentes hypothèses sur la moyenne de la population

Hypothèse nulle $\mathrm{H_0}$

- la moyenne est égale à μ_0 — $\mathrm{H_0} : \mu = \mu_0$
- la moyenne est supérieure ou égale à μ_0 — $\mathrm{H_0} : \mu \geq \mu_0$
- la moyenne est inférieure ou égale à μ_0 — $\mathrm{H_0} : \mu \leq \mu_0$

Hypothèse alternative $\mathrm{H_a}$

- la moyenne est différente de μ_0 — $\mathrm{H_a} : \mu \neq \mu_0$
- la moyenne est strictement supérieure à μ_0 — $\mathrm{H_a} : \mu > \mu_0$
- la moyenne est strictement inférieure à μ_0 — $\mathrm{H_a} : \mu < \mu_0$

Remarque

L'égalité doit toujours apparaître dans l'hypothèse nulle $\mathrm{H_0}$.

Test sur la moyenne : n grand, σ connu

Hypothèse nulle H_0	Hypothèse alternative H_a	H_a est acceptée H_0 est rejetée
$\mu = \mu_0$ $\mu \leq \mu_0$	$\mu > \mu_0$	$\bar{x} > \mu_0 + q(1-\alpha)\frac{\sigma}{\sqrt{n}}$
$\mu = \mu_0$ $\mu \geq \mu_0$	$\mu < \mu_0$	$\bar{x} < \mu_0 - q(1-\alpha)\frac{\sigma}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\bar{x} > \mu_0 + q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$ ou bien $\bar{x} < \mu_0 - q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$

Test sur la moyenne : n grand, σ inconnu

Hypothèse nulle H_0	Hypothèse alternative H_a	H_a est acceptée H_0 est rejetée
$\mu = \mu_0$ $\mu \leq \mu_0$	$\mu > \mu_0$	$\bar{x} > \mu_0 + q(1-\alpha)\frac{S}{\sqrt{n}}$
$\mu = \mu_0$ $\mu \geq \mu_0$	$\mu < \mu_0$	$\bar{x} < \mu_0 - q(1-\alpha)\frac{S}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\bar{x} > \mu_0 + q(1-\alpha/2)\frac{S}{\sqrt{n}}$ ou bien $\bar{x} < \mu_0 - q(1-\alpha/2)\frac{S}{\sqrt{n}}$

Test sur la moyenne : cas gaussien, σ connu

Hypothèse nulle H_0	Hypothèse alternative H_a	H_a est acceptée H_0 est rejetée
$\mu = \mu_0$ $\mu \leq \mu_0$	$\mu > \mu_0$	$\bar{x} > \mu_0 + q(1-\alpha)\frac{\sigma}{\sqrt{n}}$
$\mu = \mu_0$ $\mu \geq \mu_0$	$\mu < \mu_0$	$\bar{x} < \mu_0 - q(1-\alpha)\frac{\sigma}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\bar{x} > \mu_0 + q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$ ou bien $\bar{x} < \mu_0 - q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$

Test sur la moyenne : cas gaussien, σ inconnu

Hypothèse nulle H_0	Hypothèse alternative H_a	H_a est acceptée H_0 est rejetée
$\mu = \mu_0$ $\mu \leq \mu_0$	$\mu > \mu_0$	$\bar{x} > \mu_0 + t(n-1, 1-\alpha)\frac{S_c}{\sqrt{n}}$
$\mu = \mu_0$ $\mu \geq \mu_0$	$\mu < \mu_0$	$\bar{x} < \mu_0 - t(n-1, 1-\alpha)\frac{S_c}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\bar{x} > \mu_0 + t(n-1, 1-\alpha/2)\frac{S_c}{\sqrt{n}}$ ou bien $\bar{x} < \mu_0 - t(n-1, 1-\alpha/2)\frac{S_c}{\sqrt{n}}$

Tests de comparaison

Problème On veut tester si deux échantillons ont la même moyenne.
Deux situations

1. les deux échantillons sont indépendants

Exemple

On veut comparer les salaires moyens des techniciens de deux entreprises.

2. les échantillons sont appariés

Exemple

Pour tester l'efficacité d'un médicament, on compare le taux de cholestérol avant et après le traitement sur un groupe de malades. Les échantillons ne sont pas indépendants car les mesures sont effectuées sur les mêmes individus.

Échantillons indépendants

Un grand magasin implante deux boutiques
- l'une est située dans le centre ville
- l'autre dans un centre commercial en banlieue

Le directeur des ventes remarque que les produits qui se vendent bien dans un des magasins ne se vendent pas forcément bien dans le second. Il attribue cette variation des ventes au fait que l'âge moyen des clients est différent entre les deux magasins.

	boutique	taille de l'échantillon	âge moyen	écart type
pop. 1	centre ville	$n_1 = 36$	$\bar{x}_1 = 40$ ans	$S_1 = 9$ ans
pop. 2	banlieue	$n_2 = 49$	$\bar{x}_2 = 35$ ans	$S_2 = 10$ ans

Plus généralement

On suppose que les deux populations sont indépendantes

Population 1	**Population 2**
moyenne μ_1	moyenne μ_2
écart type σ_1	écart type σ_2

La question

Les deux moyennes sont-elles égales ? $\mu_1 = \mu_2$?

On teste $\mu_1 = \mu_2$ contre $\mu_1 \neq \mu_2$

Les observations : on dispose de deux échantillons indépendants.

échantillon 1 extrait de la population 1	**échantillon 2 extrait de la population 2**
taille n_1 moyenne $\bar{x}_1$,	taille n_2, moyenne $\bar{x}_2$,
écart type S_1	écart type S_2

La procédure de test

- Le test $\mathrm{H}_0 : \mu_1 = \mu_2$ contre $\mathrm{H}_\mathrm{a} : \mu_1 \neq \mu_2$
- Hypothèses : on dispose de deux grands échantillons $n_1 > 30$ et $n_2 > 30$. Les deux échantillons sont indépendants. On suppose que σ_1 et σ_2 sont connus
- On pose

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Si $|Z| > q(1-\alpha/2)$
alors

on rejette l'hypothèse nulle H_0 *(donc on accepte* H_a*) au niveau* α .

sinon

on accepte H_0

Modification de la procédure de test

lorsque les variances sont inconnues

- Le test $\mathrm{H}_0 : \mu_1 = \mu_2$ contre $\mathrm{H}_\mathrm{a} : \mu_1 \neq \mu_2$
- Hypothèses : on dispose de deux grands échantillons $n_1 > 30$ et $n_2 > 30$. Les deux échantillons sont indépendants.
- On pose

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Si $|Z| > q(1-\alpha/2)$
alors

on rejette l'hypothèse nulle H_0 *(donc on accepte* H_a*) au niveau* α.

sinon

on accepte H_0

Retour à l'exemple des deux boutiques

- On calcule Z

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{40 - 35}{\sqrt{\frac{9^2}{36} + \frac{10^2}{49}}} = 2.41$$

- On fixe l'erreur de première espèce : $\alpha = 5\%$.
  On a
$$q(1-\alpha/2) = q(0.975) = 1.96$$
- On compare $|Z|$ et $q(0.975)$
- $|Z| = 2.41 > 1.96$ donc on accepte l'hypothèse alternative

  H_a *: l'âge moyen des deux populations est différent*

  au niveau 5%

Échantillons appariés

On dispose de deux méthodes pour réaliser une tâche sur une chaîne de production. On veut comparer les temps d'exécution de ces deux méthodes
On sélectionne un échantillon de $n = 40$ ouvriers qui vont exécuter cette tâche d'abord par la méthode 1 puis par la méthode 2. .
Pour chaque personne, on récolte deux temps d'exécution. Voici un extrait des données récoltées :

i	1	2	3	4	5	6	7	8	9	⋯
x_i	6.50	5.00	3.80	5.70	4.80	6.10	5.70	5.00	4.00	⋯
y_i	4.50	6.50	5.70	7.20	4.20	5.60	5.30	5.10	6.90	⋯

Etc

Remarque

On teste les deux méthodes sur le même groupe de la population pour diminuer les effets de l'échantillonnage.

Plus généralement

Méthode 1	**Méthode 2**
moyenne μ_1	moyenne μ_2
écart type σ_1	écart type σ_2

On constitue un seul échantillon d'individus

L'échantillon 1 est constitué des résultats obtenus par la méthode 1	**L'échantillon 2 est constitué des résultats obtenus par la méthode 2**
taille n moyenne $\bar{x}_1$,	taille n, moyenne $\bar{x}_2$,
écart type S_1	écart type S_2

Définition

On dit que les échantillons sont appariés quand deux méthodes sont testées sur les mêmes individus

Construction du test

On note

- $x_1, \ldots, x_n$ l'échantillon obtenu pour la méthode 1
- $y_1, \ldots, y_n$ l'échantillon obtenu pour la méthode 2

On calcule les différences

$$d_1 = x_1 - y_1, \ \ldots \ , d_n = x_n - y_n$$

puis

- la moyenne des différences : $\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i$
- la variance : $S_d^2 = \frac{1}{n} \sum_{i=1}^{n} (d_i - \bar{d})^2$
- l'écart type $S_d = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (d_i - \bar{d})^2}$

Procédure de test

- Le test $\mathrm{H}_0 : \mu_1 = \mu_2$ contre $\mathrm{H}_\mathrm{a} : \mu_1 \neq \mu_2$
- Hypothèses : on suppose que les échantillons sont appariés et $n > 30$
- On pose

$$Z = \frac{\bar{d}}{\sqrt{\frac{S_d^2}{n}}}$$

- Si $|Z| > q(1 - \alpha/2)$
  alors

  on rejette l'hypothèse nulle et donc on accepte H_a *au niveau* α

  sinon

  on accepte H_0

Exemple (suite)

Sur l'échantillon de taille 40, on calcule

$$\bar{d} = -0.64$$
$$S_d = 1.413$$

puis $Z = -2.89$.
On compare $|Z|$ avec le quantile $q(1 - \alpha/2) = q(0.975) = 1.96$
Comme $|Z| > 1.96$, on rejette l'hypothèse H_0 au niveau 5%.
Autrement dit, on accepte l'hypothèse H_a :

les deux méthodes n'ont pas le même temps d'exécution

Test d'indépendance sur des tables de contingence

On teste l'indépendance entre deux variables.

Exemple

On dispose de trois types de bière : blanche / blonde / brune. Le groupe marketing se demande si les préférences des consommateurs sont différentes entre les hommes et les femmes

Les données :

	blanche	*blonde*	*brune*
homme	*20*	*40*	*20*
femme	*30*	*30*	*10*

Définition d'une table de contingence

On considère deux variables X et Y qui prennent un nombre fini de valeurs

- X prend les valeurs $A_1, \ldots, A_p$
- Y prend les valeurs $B_1, \ldots, B_q$

À partir d'un échantillon de taille n, on construit la table de contingence

$X \backslash Y$	B_1	B_2	$\cdots$	B_q
A_1	e(1,1)	e(1,2)	$\ldots$	e(1,q)
A_2	e(2,1)	e(2,2)	$\ldots$	e(2,q)
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_p	e(p,1)	e(p,2)	$\ldots$	e(p,q)

où $e(i,j)$ est égal au nombre d'individus dans l'échantillon qui possèdent les modalités A_i, B_j

Procédure de test

On teste H_0 : X et Y sont indépendantes contre H_a : X et Y ne sont pas indépendantes.
On note

- pour $i = 1 \ldots p$: ℓ_i le total de la ligne i
- pour $j = 1 \ldots q$: c_j le total de la colonne j

$X \backslash Y$	B_1	B_2	$\cdots$	B_q	
A_1	e(1,1)	e(1,2)	$\ldots$	e(1,q)	ℓ_1
A_2	e(2,1)	e(2,2)	$\ldots$	e(2,q)	ℓ_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_p	e(p,1)	e(p,2)	$\ldots$	e(p,q)	ℓ_p
	c_1	c_2	$\ldots$	c_q	n

On calcule

$$Q = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(e(i,j) - \frac{\ell_i c_j}{n}\right)^2}{\frac{\ell_i c_j}{n}}.$$

Si $Q > k((p-1)(q-1), 1-\alpha)$
alors

on rejette l'hypothèse nulle H_0 *(on accepte l'hypothèse alternative* H_a*) au niveau* α*. Les variables* X *et* Y *ne sont pas indépendantes*

sinon

on accepte l'hypothèse nulle H_0*, les variables sont indépendantes.*

[$k((p-1)(q-1), 1-\alpha)$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $(p-1)(q-1)$ degrés de liberté.]

Retour à l'exemple

	blanche	blonde	brune	
homme	20	40	20	80
femme	30	30	10	70
	50	70	30	150 = n

- On calcule $Q = 6.13$.
- On compare Q avec $k((2-1)(3-1), 0.95) = 5.99$
- Conclusion $Q = 6.13 > 5.99$ donc on rejette l'indépendance.
  Il existe un lien entre la préférence en matière de bière et le sexe du consommateur.

Plan de la section

La régression

On mesure deux variables continues (X, Y) sur n individus.
Les Observations : on observe donc n couples de points

$$(x_1, y_1), \ldots, (x_n, y_n)$$

Problème : Existe-t-il une liaison entre ces deux variables ?

Exemple (Une maison de vente par correspondance)

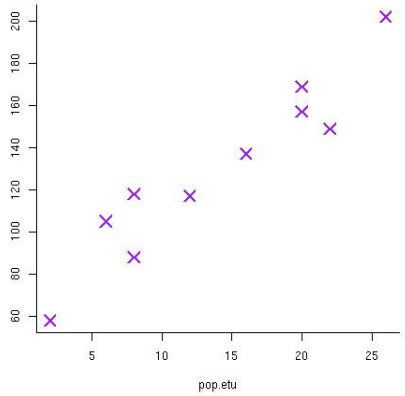
Existe-t-il un lien entre le poids du courrier reçu par une entreprise chaque matin et le nombre de commandes traitées pendant la journée.

Problème

- Tester l'existence d'une liaison entre ces deux variables
- Estimer la liaison, si elle existe.
- Utiliser cette liaison pour prévoir

Lien linéaire entre la proportion d'étudiants dans la clientèle d'un restaurant et les ventes de Pizza

	Prop. Etud. en %	Ventes en milliers euros
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202



Définition du coefficient de corrélation

Soit n couples $(x_1, y_1), \ldots, (x_n, y_n)$. La corrélation entre les variables X et Y est égale à

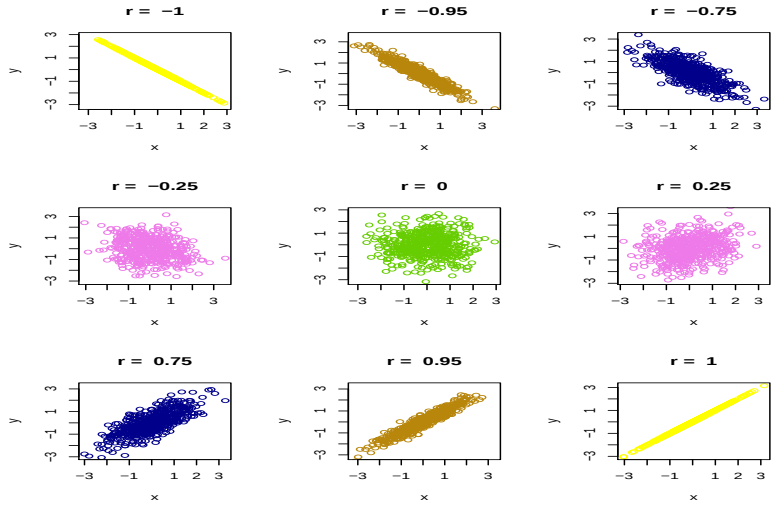
$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$$

où

- $\bar{x}$ représente la moyenne et S_x l'écart type de l'échantillon $x_1, \ldots, x_n$
- $\bar{y}$ représente la moyenne et S_y l'écart type de l'échantillon $y_1, \ldots, y_n$

1. r est un nombre entre -1 et 1.
2. $|r| = 1$ tous les points sont alignés
3. Une valeur de r proche de zéro indique que les variables ne sont pas linéairement liées

Illustration



En pratique

1. On calcule le coefficient de corrélation r
   1. Si r est proche de zéro les deux variables ne sont pas liées
   2. si $|r|$ est proche de 1, les variables sont liés.
      On cherche à déterminer si la nature du lien est linéaire ou d'une autre nature.
2. Un outil graphique. On représente le nuage de points (x_i, y_i) pour $i = 1, \ldots, n$
   - Si les points semblent dessiner une droite, alors le lien linéaire est confirmé.
   - On peut alors chercher la droite qui est la plus proche des points du nuage.

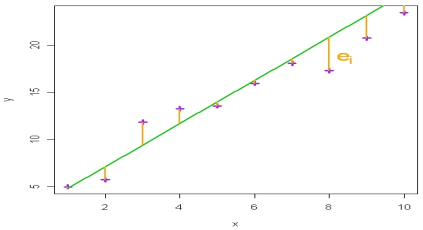
Modèle linéaire et méthode des moindres carrés

Estimation du lien linéaire entre X et Y c'est à dire $Y = aX + b + \epsilon$.

1. ϵ est une variable aléatoire appelée terme d'erreur
2. $y = ax + b$ est la droite de régression

On utilise les données $(x_1, y_1), \ldots, (x_n, y_n)$ pour estimer les coefficients de la droite (a, b).
On calcule la somme des carrés des erreurs $e_1, \ldots, e_n$

$$E_n(a, b) = \sum_{i=1}^{n} (e_i)^2$$



On cherche les coefficients a et b qui minimisent $E_n(a, b)$

Calcul de la droite de régression

- La pente est égale à

$$\hat{a} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x^2}$$

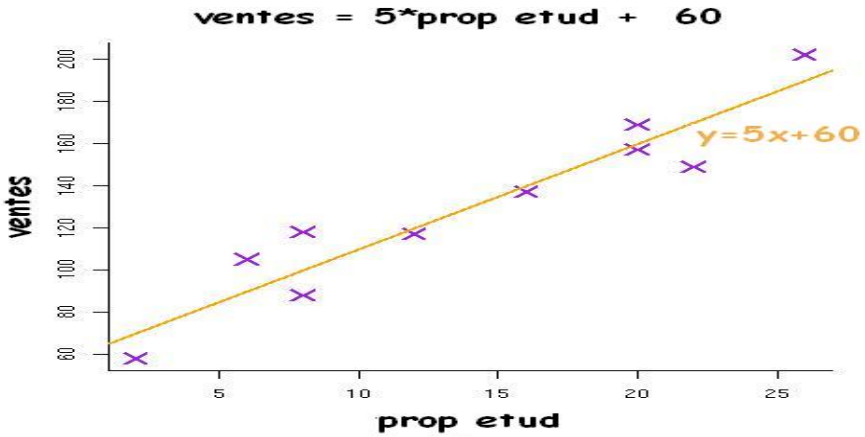
où

  - $\bar{x}$ représente la moyenne et S_x^2 la variance de l'échantillon $x_1, \ldots, x_n$
  - $\bar{y}$ représente la moyenne de l'échantillon $y_1, \ldots, y_n$
- L'ordonnée à l'origine est égale à

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

Suite de l'exemple sur les ventes de pizzas

La corrélation entre les deux variables vaut 0.95. l'ajustement linéaire est satisfaisant



Prévoir

S'il existe un lien linéaire entre X et Y, on peut prévoir la valeur prise par Y connaissant la valeur de X

Calcul de la prévision Si on connaît la valeur de X, $X = x_0$, on prévoit la valeur de la variable Y en prenant $\hat{a}x_0 + \hat{b}$.

Exemple

Un restaurateur sait que sa clientèle est composée de 10 % d'étudiants
Il peut prévoir ses ventes de pizzas en prenant
$\hat{a} \times 10 + \hat{b} = 5 \times 10 + 60 = 110$ *milliers d'euros*

Le bon usage du coefficient de corrélation

On dispose de 4 nuages de points

Dans les 4 cas, on a
$\bar{x} = 9$; $\bar{y} = 7.50$,
$S_x^2 = 10$; $S_y^2 = 3.75$
et $r = 0.816$.

données A		données B		données C		données D	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

On obtient donc la même droite $y = 0.5x + 3$ pour les 4 nuages de points.

Les nuages de points associés aux données