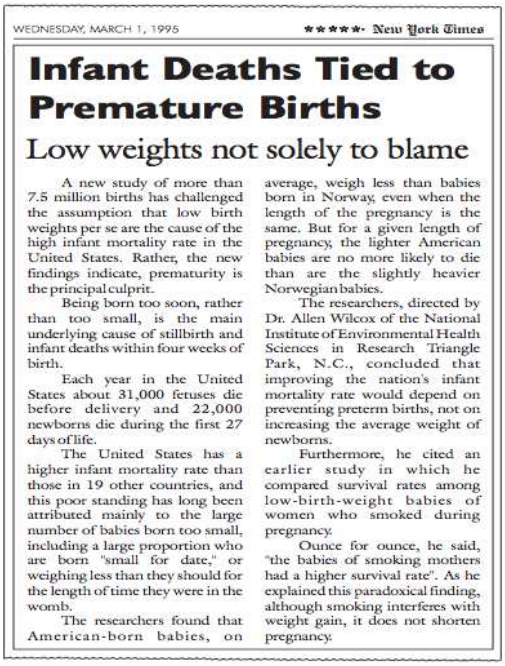
Traitement de données

GLMA 512

Grossesses et cigarettes : impact sur la santé du nouveau né ?
Statistiques descriptives



WEDNESDAY, MARCH 1, 1995 ★★★★★ New York Times

Infant Deaths Tied to Premature Births

Low weights not solely to blame

A new study of more than 7.5 million births has challenged the assumption that low birth weights per se are the cause of the high infant mortality rate in the United States. Rather, the new findings indicate, prematurity is the principal culprit.

Being born too soon, rather than too small, is the main underlying cause of stillbirth and infant deaths within four weeks of birth.

Each year in the United States about 31,000 fetuses die before delivery and 22,000 newborns die during the first 27 days of life.

The United States has a higher infant mortality rate than those in 19 other countries, and this poor standing has long been attributed mainly to the large number of babies born too small, including a large proportion who are born "small for date," or weighing less than they should for the length of time they were in the womb.

The researchers found that American-born babies, on average, weigh less than babies born in Norway, even when the length of the pregnancy is the same. But for a given length of pregnancy, the lighter American babies are no more likely to die than are the slightly heavier Norwegian babies.

The researchers, directed by Dr. Allen Wilcox of the National Institute of Environmental Health Sciences in Research Triangle Park, N.C., concluded that improving the nation's infant mortality rate would depend on preventing preterm births, not on increasing the average weight of newborns.

Furthermore, he cited an earlier study in which he compared survival rates among low-birth-weight babies of women who smoked during pregnancy.

Ounce for ounce, he said, "the babies of smoking mothers had a higher survival rate". As he explained this paradoxical finding, although smoking interferes with weight gain, it does not shorten pregnancy.

Mise en garde !

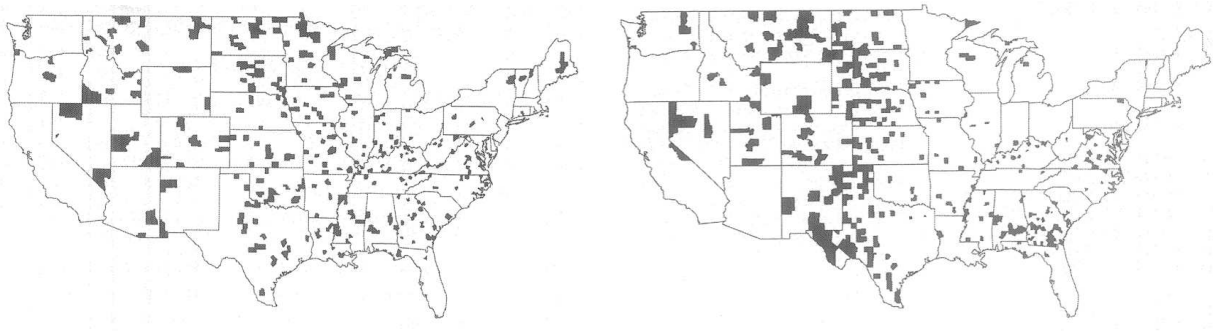


Figure 1: *Les contés des USA ayant les 10% plus forts taux de mortalité suite à un cancer du rein/urêtre pour les hommes sur la période 1980–1989. Données standardisées par rapport à l'âge. Droite : idem mais les 10 plus faibles !*

0.1 Problématique

Fumer pendant la grossesse nuit à la santé de votre enfant.

- ☐ Objectif : Comparaison du poids à la naissance selon le statut fumeur de la mère.
- ☐ Données : 1236 accouchements aux alentours de San Francisco (1960–1967) – pas de jumeaux, triplés, . . . et ayant vécu au moins 28 jours

Une partie des données

Table 1: *Partie du tableau de données*

Poids Naissance	120	113	128	123	108	136	138	132
Statut Fumeur	0	0	1	0	1	0	0	0

- Les poids sont en onces. 1g = 0.035 once.
- Le statut fumeur vaut 1 si la mère fumait pendant la grossesse et 0 sinon.
- Le tableau en entier à donc 1236 (+1) colonnes.
- Visualisation impossible $\Longrightarrow$ besoin de résumer les données par quelques valeurs numériques, graphiques utiles
- C'est ce que nous allons voir dans ce cours.

Tableau des fréquences

- Une étude de ces données a montré que le taux de mortalité infantile chez les mères fumeuses était plus faible !
- Cette conclusion est basée sur le tableau suivant

Classe de poids	Non fumeur	Fumeur
< 1500	792	565
1500–2000	406	346
2000–2500	78	27
2500–3000	11.6	6.1
3000–3500	2.2	4.5
≥ 3500	3.8	2.6

Table 2: *Taux de mortalité infantile (pour 1000 naissances) en fonction du poids (g) à la naissance différencié selon le statut fumeur. Tableau croisé.*

- Des critiques sur ce tableau ? Age de la mère ou autres facteurs...
- Une étude tenant compte de ces facteurs donne la même conclusion.

- Une autre étude préconise de travailler sur les poids à la naissance standardisés

$$\text{poids standardisé} = \frac{\text{poids} - \text{moyenne}}{\text{écart type}}.$$

- Cette standardisation est faite séparément pour les fumeurs et pour les non fumeurs.
- Quel est l'intérêt ?
- Comparer ce qui est comparable ! Exemple : si les bébés de mères fumeuses ont toujours un poids plus faible.
- Ainsi on comparera le taux de mortalité d'un bébé pesant 2680g (fumeur) à celui pesant 3000g (non fumeur) car ces valeurs sont exactement moyenne $- 1 \times$ écart type dans les deux cas.

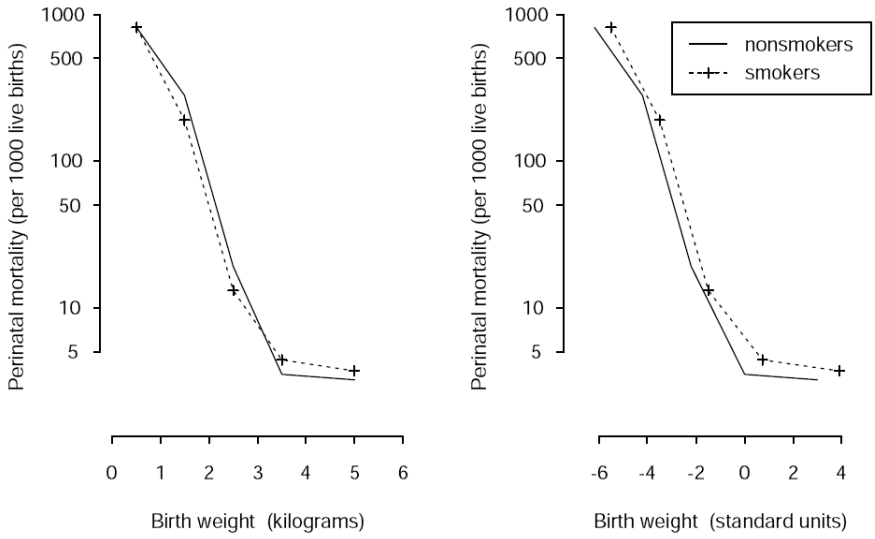


Figure 2: *Taux de mortailité (pour 1000) en fonction du poids et du poids standardisé*

- □ Il semblerait maintenant que les bébés de mères fumant aient un taux de mortalité plus élevé.
- □ Moralité : Faire attention aux effets "cachés".

0.2 Théorie

Histogramme

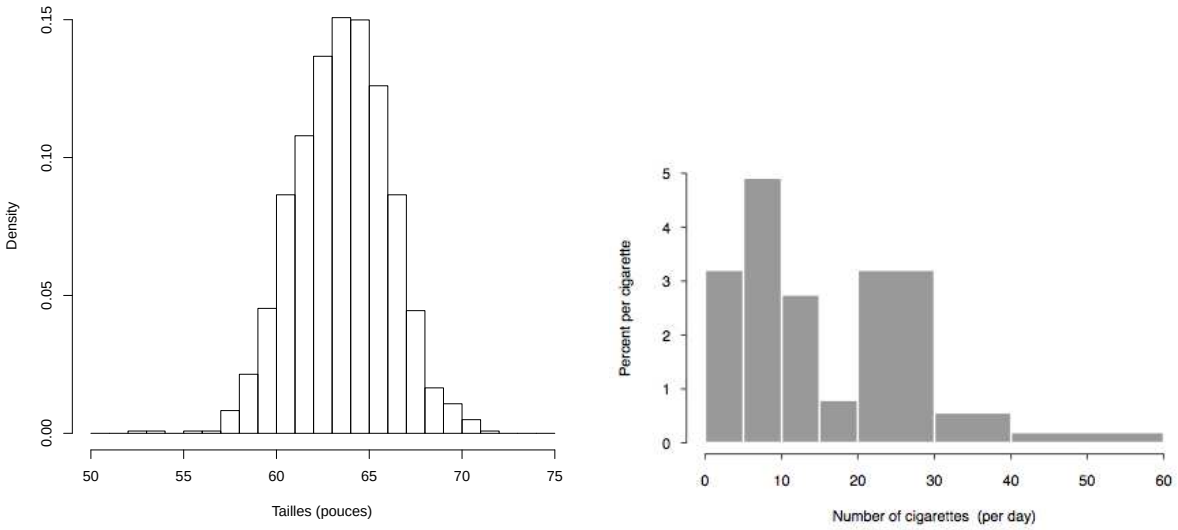


Figure 3: *Histogramme de la taille (pouces) issu des 1214 mères (droite) et du nombre de cigarettes fumées par jour issu des 384 mères fumeuses.*

- C'est quoi ? Juste une représentation graphique des observations
- Utilité :
  - Forme de la distribution : unimodalité, symétrie, étendue
  - Détection des valeurs aberrantes (outliers).
  - Que peut on dire sur les deux histogrammes précédents ?
- Construction :

$$\text{hauteur}_k = \frac{\text{effectif de la classe } k}{\text{effectif total} \times \text{largeur de la classe } k}.$$

Par construction l'aire de cet histogramme est égale à 1.

Exemple : Construction

Nb de cig.	Nb. de fumeurs (%)
0–5	16
5–10	25
10–15	14
15–20	4
20–30	32
30–40	5
40–60	4
Total	100

Table 3: *Distribution du nombre de cigarettes fumées par jour pour les 484 mères fumeuses.*

- Notion de classe. Le nombre 5 appartient à quelle classe ?
  A la deuxième !
- Hauteur du rectangle pour certaines classes

$$h_1 = \frac{16}{5 \times 100} = 0.032, \qquad h_5 = \frac{32}{10 \times 100} = 0.032.$$

Résumés numériques : Mesures de position

- Il est souvent utile de "résumer" les données par un nombre limité de valeurs numériques. On parle alors de statistiques.
- Une statistique de position mesure le centre de la distribution.
- La moyenne (empirique) est un exemple

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad x_i \text{ observations.}$$

- La médiane en est une autre plus robuste aux valeurs aberrantes

médiane = valeur qui sépare les données en deux parties de mêmes effectifs.

1, 4, 5, 8, 12, médiane = 5
1, 4, 5, 8, 12, 15, médiane = (5 + 8) / 2 = 6.5

Résumés numériques : Mesures de dispersion

- Elles mesurent la dispersion de la distribution, i.e., sa variabilité.
- L'écart-type est un exemple

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

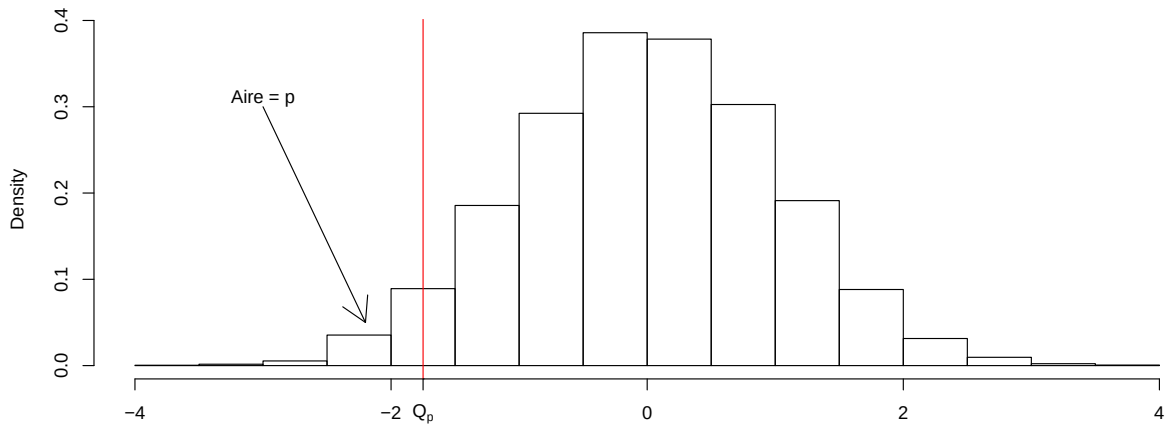
- L'écart inter-quartile en est une autre

$$IQR = Q_3 - Q_1,$$

où Q_1 est le nombre tel que $1/4$ de l'effectif lui est inférieur et Q_3 lui est supérieur.

Comment calculer les quantiles (approchés) ?

- A partir des données, il n'existe pas une unique manière de calculer les quantiles — cf. `?quantile` dans R.
- On va donc voir une manière de les calculer (approximativement) à partir d'un histogramme.
- Q_p est défini comme le nombre tel que l'aire de l'histogramme avant ce point soit égale à p.



Exemple : Calculs de Q_1 et Q_3.

Classe	Effectif	h_k (%)	Aire$_k$ (%)	Cum. Aires (%)
0–5	3	2.4	12	12
5–10	10	8	40	52
10–15	7	5.6	28	80
15–25	5	2	20	100
Total	25	—	100	100

- $Q_1 = Q_{0.25}$ se situe quelque part entre 5 et 10 puisque à $x = 5$ l'aire à gauche fait 12.
- Pour arriver à 25 il faut donc rajouter 13 à partir du deuxième rectangle qui a pour hauteur 8 donc

$$\Delta \times 8 = 13 \Longrightarrow \Delta = 13/8 = 1.63, \qquad \text{et donc} \quad Q_1 = 5 + \Delta = 6.63.$$

- A vous de faire pour Q_3 la réponse est $Q_3 = 14.11$.

Boxplot

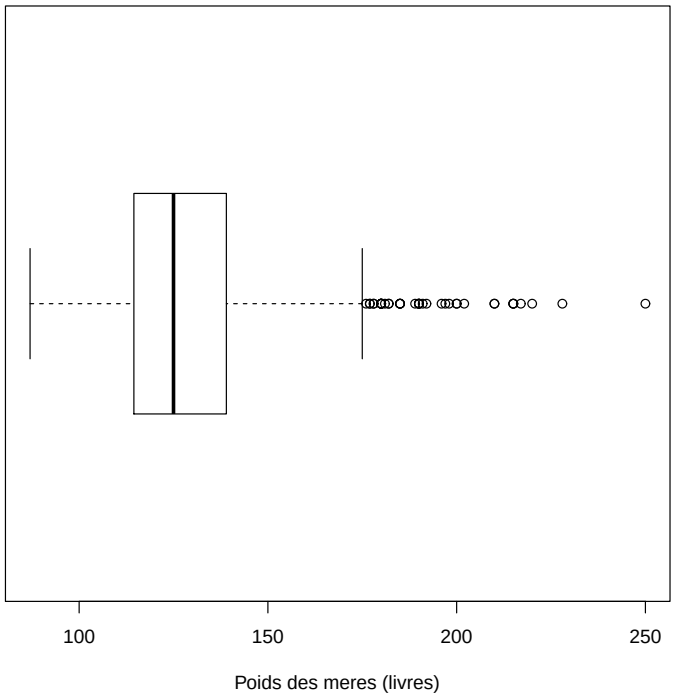


Figure 4: *Boxplot du poids des 1200 mères.*

- C'est quoi ? C'est une autre représentation graphique d'un jeu de données.
- Utilité :
  - Pareil que l'histogramme.
  - Il contient plus d'information que les résumés numériques mais moins qu'un histogramme.
  - Mais peut-être utile pour visualiser un grand nombre de variables (car compact).
- Construction :
  - Il faut calculer la médiane, Q_1 et Q_3
  - Et faire quelques incantations magiques...

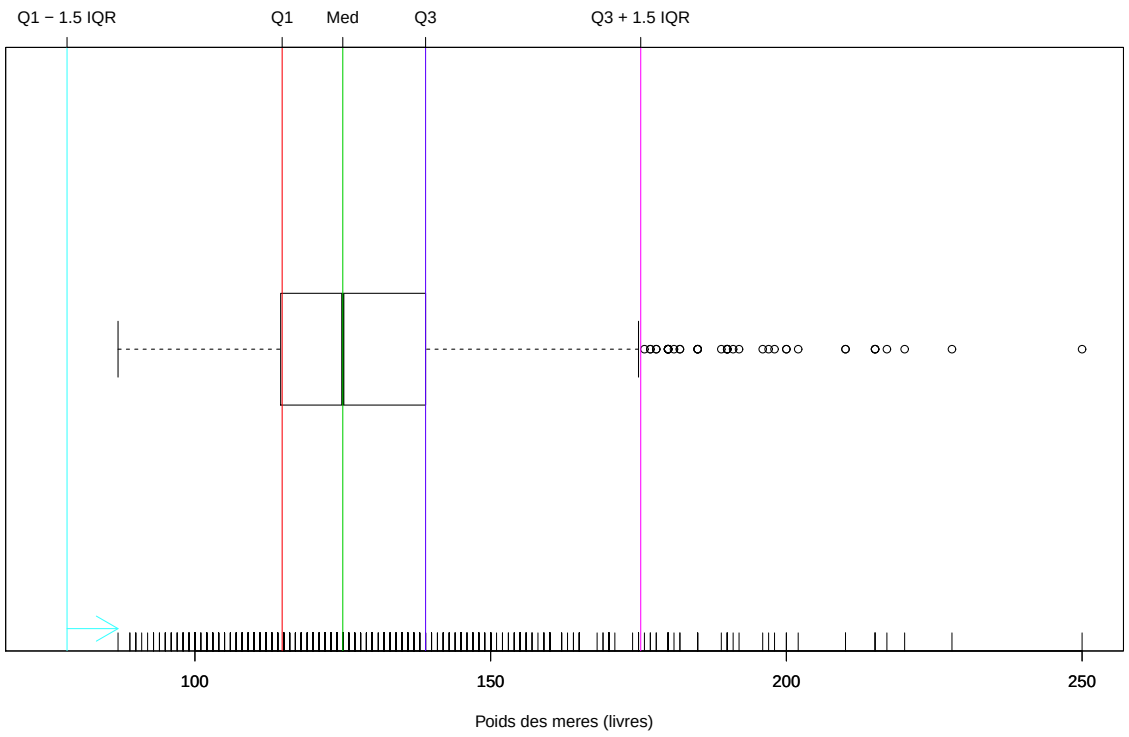


Figure 5: *Boxplot du poids des 1200 mères — avec les détails.*

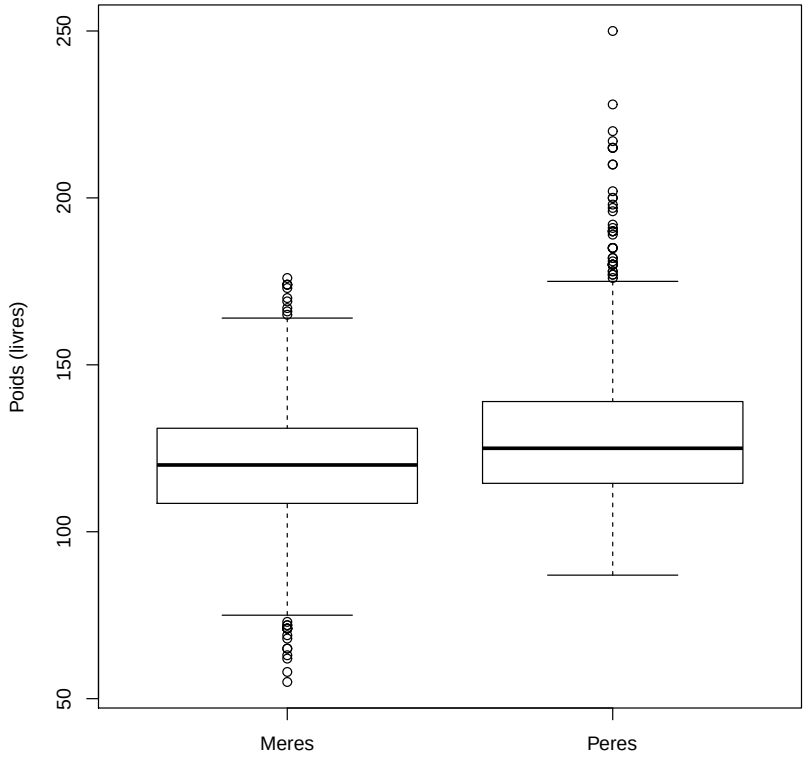


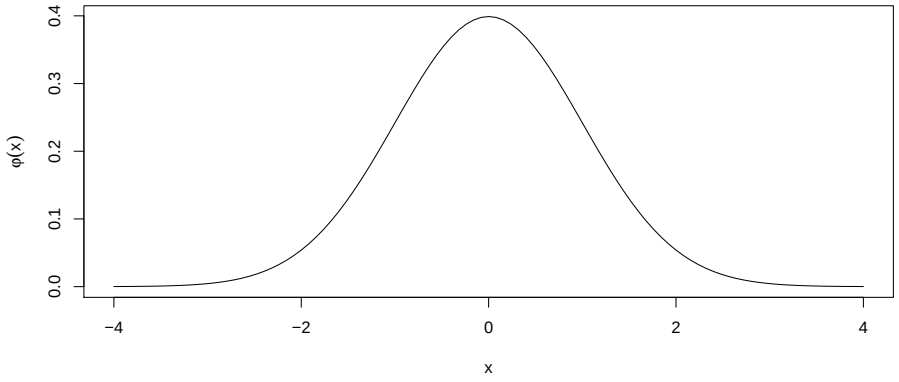
Figure 6: *Comparaison de la distribution des poids des mères et pères à l'aide de boxplot.*

0.3 Loi normale

Loi Normale

- La loi Normale joue un rôle central en statistique
- De nombreuses données suivent approximativement cette loi
- Des théorèmes nous disent que certaines variables aléatoires suivent approximativement cette loi dès lors que n est grand.
- La densité de la loi normale centrée réduite est

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad x \in \mathbb{R}.$$



- Si des données suivent approximativement une loi Normale, alors l'histogramme des données centrées réduites doit ressembler à la courbe précédente.
- Pour standardiser les données $x_1, \ldots, x_n$

$$\frac{x_i - \overline{x}}{\sigma}, \qquad i = 1, \ldots, n$$

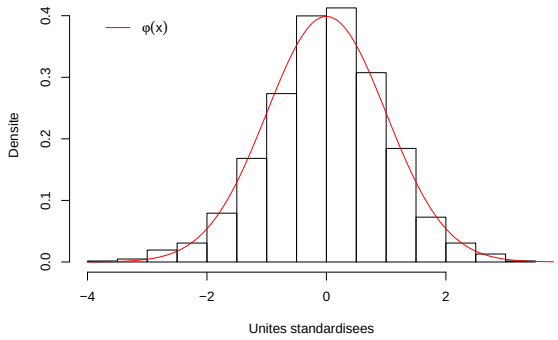


Figure 7: *Comparaison de l'histogramme des poids à la naissance (standardisés) et de la densité de la loi normale centrée réduite.*

- Les données semblent donc être bien représentées par une loi Normale.
- Si tel est le cas, on peut alors utiliser cette loi pour répondre à certaines questions.

Exemple : Calcul de $\Pr[\textsf{Poids} \leq 138]$

- La fonction de répartition de la loi normale centrée réduite corresponds à l'aire sous la courbe φ jusqu'à un point z

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx.$$

- Cela corresponds à la probabilité d'être inférieur à z sous cette loi normale centrée réduite

Pour notre exemple, on a donc

$$\Pr[\textsf{Poids} \leq 138] = \Pr\left[\frac{\textsf{Poids} - \overline{x}}{\sigma} \leq \frac{138 - \overline{x}}{\sigma}\right] \approx \Phi(1) = 0.84.$$

En calculant cette probabilité à partir des données on trouve 0.85 mais ne permettrait pas de faire les calculs en dehors du domaine observé.

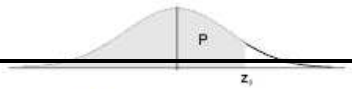
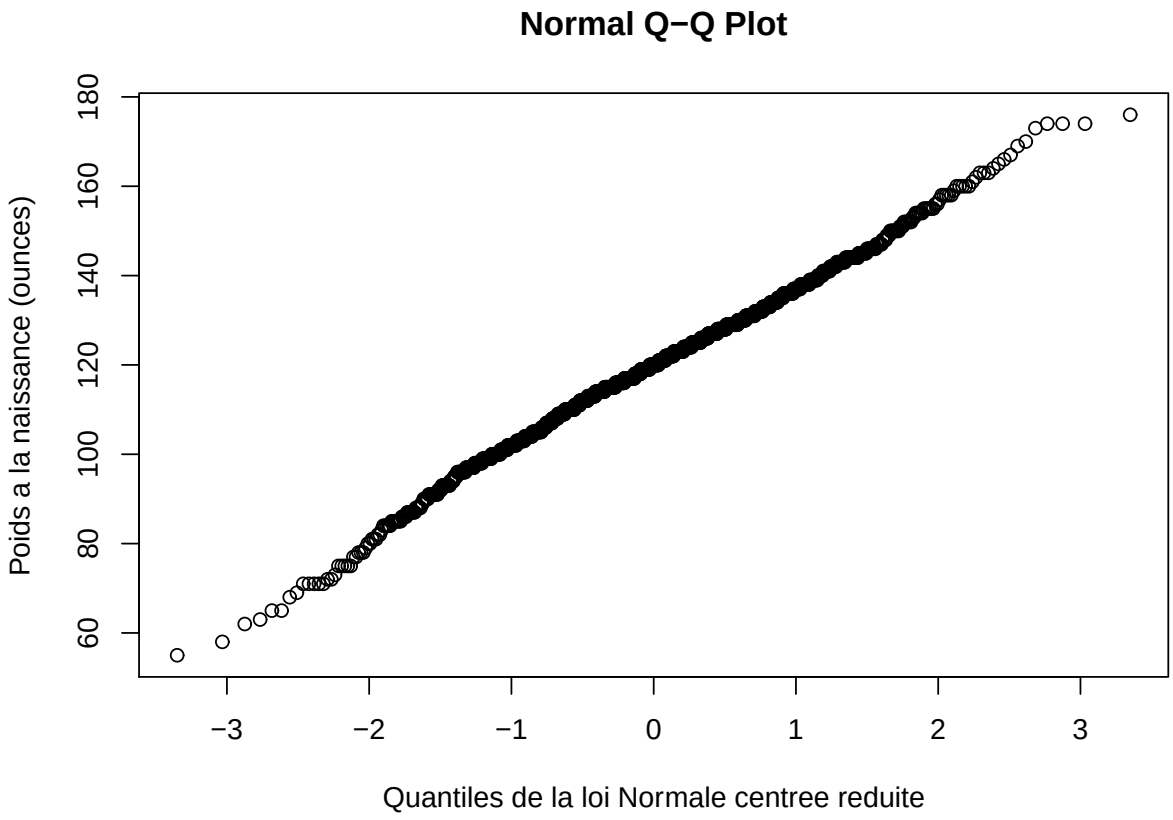


TABLE C.1. Cumulative normal distribution—values of P corresponding to z_p for the standard normal curve.

z_p	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.719	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8707	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

0.4 QQ-plot

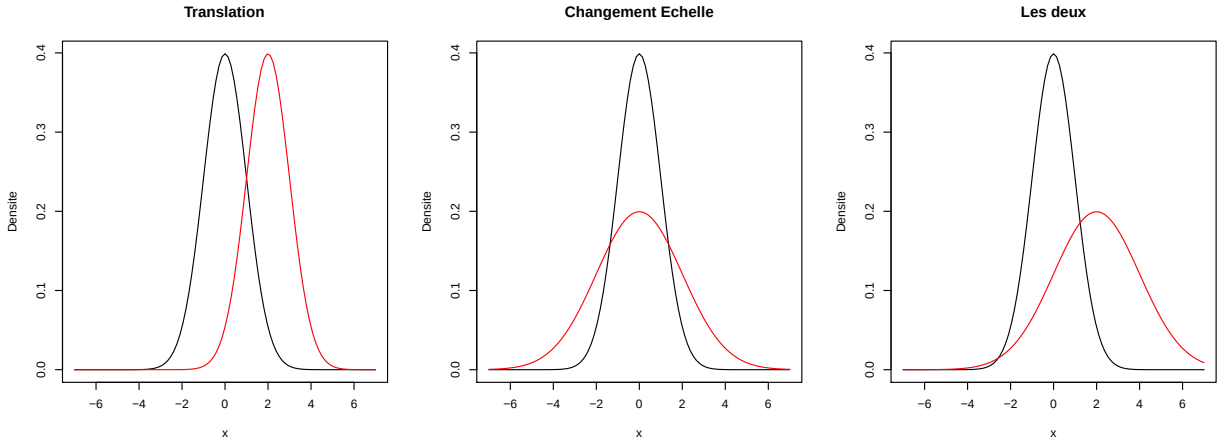
QQ-plot



- □ C'est quoi ? Une représentation graphique comparant des observations.
- □ Utilité :
  - (a) Vérifier si les données suivent une loi particulière
  - (b) Vérifier si deux jeux de données ont la même loi
- □ Construction :
  - Classer les observations par ordre croissant, i.e., $x_{1:n}, \ldots, x_{n:n}$
  - A ces points on associe les probabilités $p_i = i/(n+1)$, $i = 1, \ldots, n$.
  - (a) Représenter les points $\{x_{i:n}, Q_{p_i}\}_{i=1,\ldots,n}$, par exemple pour la loi normale centrée réduite $Q_{p_i} = \Phi^{-1}(p_i)$.
  - (b) Classer le deuxième jeu de données par ordre croissant et représenter les points $\{x_{i:n}, y_{i:n}\}_{i=1,\ldots,n}$.

Interprétation : QQ-plot (loi normale) des poids à la naissance

- Si les observations étaient $N(0,1)$ alors le nuage de points se concentrerait autour de la droite $y = x$
- Ici le nuage de points se concentre autour d'une droite mais pas $y = x$, disons $y = ax + b$.
- Si $b \neq 0 \Longrightarrow$ Translation
- Si $a \neq 1 \Longrightarrow$ Changement d'échelle (variabilité)
- Enfin si le nuage de points n'est pas "linéaire" cela indique que les deux distributions ont des formes différentes.



QQ-plot du poids des mères fumeuse / non-fumeuse

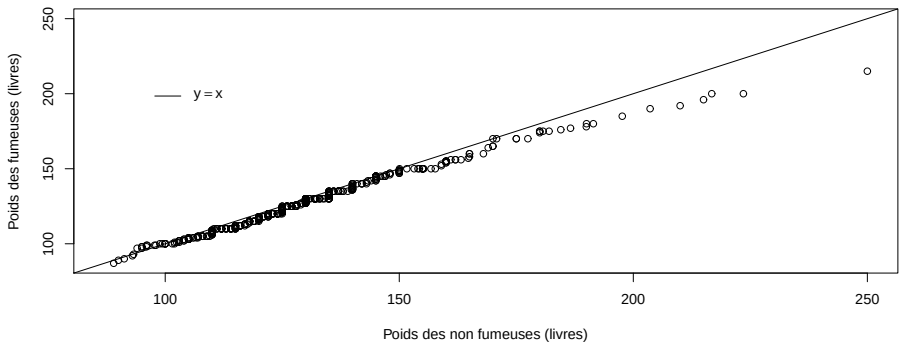
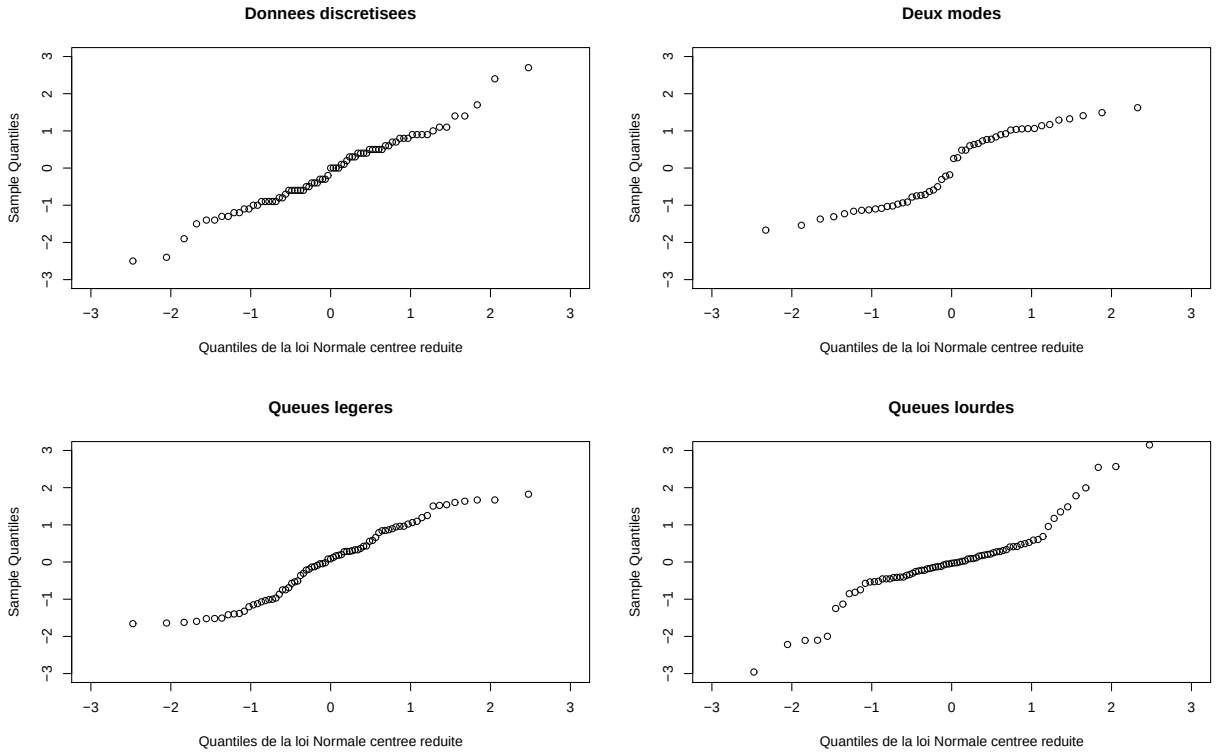


Figure 8: *QQ-plot du poids des mères (livres) selon le statut fumeur ou non.*

- Le nuage de points semble être linéaire la plupart du temps (sauf à l'extrémité droite du graphique)
  - Ceci indique que les mères fumant ont tendance à peser moins que les non fumeuses
  - Le "décrochage" à droite indique que "les plus lourdes" non fumeuses pèsent plus que "les plus lourdes" fumeuses

Quelques QQ-plot (loi normale) pathologiques



Ce que nous avons vu

- ☐ Histogramme : moyen graphique de représenter la distribution des observations
- ☐ Résumés numériques : moyenne, médiane, quantile, écart-type, distance inter-quartile
- ☐ Boxplot : moyen graphique synthétique de représenter la distribution des observations
- ☐ Loi Normale
- ☐ QQ-plot : par rapport à une loi, pour deux jeux de données

Qui joue aux jeux vidéo ?
Échantillonnage aléatoire simple

TUESDAY, MARCH 3, 1998 — San Francisco Chronicle

Computing Computers' Effect

Critics ask if there's too much technology in the classroom

By John Wildermuth

When Bill Gates, founder of Microsoft, cruised through the Bay Area in January, he took time out to visit some of his company's youngest customers.

The billionaire software developer spent part of the morning at Cesar Chavez Academy in East Palo Alto, watching a giggling bunch of third-graders work on computer art programs and search for information on the Internet.

"I need to see what you're doing with the computers and how we can make it better," Gates told the twenty 8- and 9-year-olds. But some people are asking the question, better for whom? Better for the students whose parents desperately want them to get the education that best prepares them for the future, or better for the school districts, politicians and technology companies carrying on a wild love affair with a high-tech vision of education's future?

President Clinton says in a speech that he wants to see "the day when computers are as much a part of a classroom as blackboards." California legislators plan to commit about $460 million over the next four years to get technology into every one of the state's 1,400 high schools. Schools throughout the state are spending millions of dollars to wire classrooms, buy computers and show children from kindergarten on up how to surf the Net. ...

Making computers part of the daily classroom activity is one of the toughest parts of the new technology, said Rhonda Neagle, technology coordinator for the New Haven Unified School District in Union City. ...

"What we're asking teachers to do is change the way they teach," she said. "With only six (student) computers in a room, teachers need to provide more group work and maybe use a different setup for their course."

0.5 Problématique

À Berkeley (Université de Californie), environ 300 étudiants suivent un cours de statistiques 1 chaque année.

- ☐ Quelle proportion d'étudiants ont joué aux jeux vidéos la semaine précédent les examens ?
- ☐ En moyenne, combien de temps un étudiant a-t-il joué ?
- ☐ En automne 1997, ils étaient 314. Il est impensable de faire une étude sur ces 314 étudiants, trop fastidieux, coûteux.
- ☐ Mais il paraît assez intuitif de se concentrer sur un sous groupe plus raisonnable.

Plusieurs questions nous viennent alors à l'esprit

- ☐ Comment construire ce "sous groupe" ?
- ☐ Quelle taille faut il lui donner ?
- ☐ Que peut on dire sur le temps moyen des 300 étudiants à partir du temps moyen du sous groupe ?

0.6 Théorie

Dans ce cours nous allons introduire l'échantillonnage aléatoire simple.

- ☐ C'est une méthode probabiliste permettant de choisir les étudiants constituant le sous groupe.
- ☐ Les méthodes probabilistes sont importantes puisqu'elles permettent, en contrôlant le hasard, de connaître les relations entre le sous groupe et le groupe entier.

Vocabulaire : Population

- La population est le groupe que l'on veut étudier, e.g., les 314 étudiants
- La population est constituée d'unités, e.g., 1 unité = 1 étudiant
- Taille de la population notée N est le nombre d'unités dans la population, e.g., $N = 314$
- Les variables sont des informations particulières données pour chaque membre de la population, e.g., temps de jeux
- Un paramètre est un résumé des variables sur toute la population, e.g. le temps moyen de jeux la semaine précédent l'examen

Vocabulaire : Échantillon

- L'échantillon sont les unités choisies pour faire notre étude
- La taille de l'échantillon notée n est le nombre d'unités présentes dans l'échantillon
- Une statistique est un résumé numérique des variables calculé à partir de l'échantillon.

Tout échantillon est construit à partir d'une *règle de décision*. Pour notre étude, les étudiants ont été "numérotés" de 1 à 314 et un ordinateur a choisi 91 numéros entre 1 et 314 successivement, i.e., aucun étudiant n'a pu être sélectionné deux fois. De plus tout au long du processus de sélection chaque numéro disponible avait la même probabilité d'être choisi.

Ceci constitue un échantillonage aléatoire simple.

Échantillonage aléatoire simple

- C'est une méthode pour affecter une probabilité à chaque échantillon de taille n extrait d'une population de taille N.
- Pour notre étude, $n = 91$ et $N = 314$.
- Combien y a-t-il d'échantillons possible ?

$$\underbrace{314}_{\text{1er étudiant}} \times \underbrace{313}_{\text{2ème étudiant}} \times \cdots \times \underbrace{224}_{\text{91ème étudiant}}$$

- Mais ceci tient compte de l'ordre de sélection, i.e., qui a été choisi en premier, second, ...
- Pour notre étude, peu nous importe l'ordre de sélection donc le nombre d'échantillon est en fait

$$\frac{314 \times 313 \times \cdots \times 224}{91 \times 90 \times \cdots \times 1}$$

- Ce nombre s'écrit $\binom{314}{91}$ ou encore C_{314}^{91}.
- De manière générale c'est donc $\binom{N}{n}$ ou C_N^n.

La règle de décision défini par l'échantillonnage aléatoire simple impose que chacun de ces échantillons ait la même chance d'être sélectionné.

Chaque échantillon a donc une probabilité $1/\binom{N}{n}$ d'être sélectionné.

Les individus de la population ayant été numéroté de 1 à 314, on a donc

$$\Pr[\text{n}^\circ\ 1 \text{ sélectionné en premier}] = 1/314$$
$$\Pr[\text{n}^\circ\ 1 \text{ sélectionné en deuxième}] = ???$$
$$\Pr[\text{n}^\circ\ 1 \text{ est dans l'échantillon}] = 91/314$$

Toutefois il y a de la dépendance dans le processus de sélection

$$\Pr[\text{n}^\circ\ 1 \text{ et n}^\circ\ 2 \text{ sont dans l'échantillon}] = \frac{91 \times 90}{314 \times 313} \neq \frac{91}{314} \times \frac{91}{314}$$

Loi de probabilité des numéros sélectionné

- Posons $I(1)$ le premier numéro pris au hasard parmis $1, \ldots, N$.
- De même $I(2)$ sera le deuxième numéro, $\ldots$, $I(n)$ le dernier numéro sélectionné.
- Les $I(1), \ldots, I(n)$ sont des variables aléatoires (discrètes).

On a donc

$$\Pr[I(1) = 1] = \frac{1}{N}$$
$$\Pr[I(1) = 1, I(2) = 2] = \frac{1}{N \times (N-1)}$$

et de manière générale pour $1 \leq j_1 \neq j_2 \neq \cdots \neq j_n \leq N$, on a

$$\Pr[I(1) = j_1, I(2) = j_2, \ldots, I(n) = j_n] = \frac{1}{N \times (N-1) \times \cdots \times (N-n+1)}$$

Que nous dit l'horrible équation précédente ?

Rien de plus que l'échantillonnage aléatoire simple pose une structure aléatoire sur l'échantillon.

- Des échantillons différents auront des valeurs différentes pour leurs variables (temps de jeux $\neq$) et donc des statistiques différentes (temps moyen de jeux $\neq$).
- Autrement dit, une statistique admet une loi de probabilité liée à la procédure d'échantillonnage.
- C'est donc une variable aléatoire !

Faire une illustration sur R

0.7 Moyenne empirique

Moyenne empirique

Notons $x_1, \ldots, x_N$ le nombre d'heures de jeux pour l'étudiant numéro $1, \ldots, N$. Notre étude s'intéresse au paramètre

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Comme nous n'avons pas accès aux N étudiants, il paraît assez intuitif de considérer

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} x_{I(j)}$$

On l'appelle la moyenne empirique. C'est une statistique (elle est aléatoire puisque les $I(j)$ le sont). On dira aussi que $\overline{X}$ est un estimateur de μ puisque c'est une statistique estimant le paramètre μ.

Espérance de la moyenne empirique

- Calculons l'espérance du temps de jeux pour le premier individu sélectionné.

$$\mathbb{E}[X_{I(1)}] = \sum_{j=1}^{N} x_j \Pr[I(1) = j] = \sum_{j=1}^{N} x_j \frac{1}{N} = \frac{1}{N} \sum_{j=1}^{N} x_j = \mu.$$

De même puisque tous les individus sont équiprobables, on a

$$\mathbb{E}[x_{I(j)}] = \mu, \qquad j = 1, \ldots, n.$$

L'espérance de $\overline{X}$ est alors

$$\mathbb{E}[\overline{X}] = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[X_{I(j)}] = \mu.$$

On dit alors que notre estimateur $\overline{X}$ est sans biais ou non biaisé.

Variance de la moyenne empirique

- Le transparent précédent nous dit qu'en moyenne $\overline{X}$ notre estimateur est exact.
- $\overline{X}$ varie t il beaucoup autour de μ, i.e., on veut connaître sa variance

$$\mathrm{Var}[\overline{X}] = \mathbb{E}\left[\left(\overline{X} - \mathbb{E}[\overline{X}]\right)^2\right]$$

- Par analogie on définit également la variance sur la population (c'est donc un paramètre)

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{N} (x_j - \mu)^2$$

- On peut montrer que $\mathrm{Var}[\overline{X}] = \frac{1}{n}\sigma^2 \frac{N-n}{N-1}$, et donc que l'écart type de l'estimateur (on parle alors d'erreur standard) est

$$\mathrm{SE}(\overline{X}) := \sqrt{\mathrm{Var}[\overline{X}]} = \frac{1}{\sqrt{n}} \sigma \sqrt{\frac{N-n}{N-1}}$$

Pour votre culture

- Le terme $k = (N-n)/(N-1)$ s'appelle le facteur de correction en population finie.
- Il vient facilement que ce facteur de correction vaut approximativement

$$k = 1 - \frac{n-1}{N-1} \approx 1 - \frac{n}{N}$$

- Ainsi lorsque le rapport n/N est très petit, ce qui est souvent le cas en statistique, le facteur de correction est proche de 1 et

$$\mathrm{SE}[\overline{X}] = \sqrt{\frac{k}{n}}\sigma \approx \frac{\sigma}{\sqrt{n}}.$$

- Ainsi k est souvent ignoré mais pour notre étude nous ne devons pas puisque

$$k = (314-91)/(314-1) = 0.71.$$

Variance empirique

- Le problème avec $\mathrm{SE}(\overline{X}) = \frac{1}{\sqrt{n}}\sigma\sqrt{\frac{N-n}{N-1}}$, c'est que σ doit être connu
- Ce n'est généralement pas le cas. Après tout on ne connaissait pas μ alors pourquoi connaître σ^2...
- Un estimateur de σ^2 est

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{I(j)} - \overline{X}\right)^2$$

- Et puisque

$$\mathrm{Var}[\overline{X}] = \frac{1}{n}\sigma^2\frac{N-n}{N-1},$$

un estimateur de cette quantité est donc

$$\frac{s^2}{n}\frac{N-n}{N-1}.$$

- En fait cet estimateur de $\mathrm{Var}[\overline{X}]$ n'est pas entièrement satisfaisant puisqu'on peut montrer que

$$\mathbb{E}\left[\frac{s^2}{n}\frac{N-n}{N-1}\right] = \frac{N}{N-1}\mathrm{Var}[\overline{X}],$$

il est donc biaisé.

- Un estimateur sans biais de $\mathrm{Var}[\overline{X}]$ est alors

$$\frac{s^2}{n}\frac{N-n}{N}$$

- Remarque : Pour une population de taille raisonnable, ces deux estimateurs ne diffèrent que très peu.

0.8 Proportions

Proportions

- Parfois le paramètre d'intérêt est une proportion
- Pour notre étude cela pourrait être la proportion des étudiant qui ont joué aux jeux vidéos la semaine précédent l'examen
- Dans de tels cas on introduit une variable binaire valant 0 ou 1. Par exemple

$$x_1 = \begin{cases} 1, & \text{si l'étudiant numéro 1 a joué} \\ 0, & \text{si l'étudiant numéro 1 n'a pas joué} \end{cases}$$

- Ainsi $\tau = \sum_{j=1}^{N} x_j$ est le nombre d'étudiants ayant joué aux jeux vidéo la semaine précédent l'examen
- De même $\pi = \sum_{j=1}^{N} x_j / N$ est la proportion de tels étudiants

- D'après les transparents précédents, $\overline{X}$ est un estimateur sans biais de π
- et donc $N\overline{X}$ est un estimateur sans biais de τ.
- Dans ce contexte particulier, les résultats précédents sont toujours valide mais se simplifient un peu.
- En effet

$$\begin{aligned} \sigma^2 &= \frac{1}{N}\sum_{j=1}^{N}(x_j - \pi)^2 = \frac{1}{N}\sum_{j}\left(x_j^2 - 2\pi x_j + \pi^2\right) \\ &= \pi - 2\pi^2 + \pi^2 = \pi(1-\pi) \end{aligned}$$

- Et un estimateur de $\mathrm{Var}[\overline{X}]$ est

$$\frac{\overline{X}(1-\overline{X})}{n-1}\frac{N-n}{N}$$

0.9 Théorème Central Limite et Intervalles de confiance

Théorème central limite

Version orale :

> Si la taille de l'échantillon est grande, alors la distribution de la moyenne empirique suit approximativement une loi normale.

Version formelle

Théorème 1 (Théorème central limite). *Soient $X_1, X_2, \ldots$ des variables aléatoire indépendantes et de même loi de moyenne $\mu < \infty$ et de variance $\sigma^2 < \infty$, alors*

$$Z = \sqrt{n}\frac{\overline{X} - \mu}{\sigma}$$

converge (en loi) vers une loi normale centrée réduite, notée $N(0,1)$.

"Oui mais vous nous avez dit que pour un échantillonnage aléatoire simple, les $x_{I(j)}$, $j = 1, \ldots, n$, étaient dépendants !!!"

- C'est une très bonne remarque !
- Toutefois si le rapport n/N est petit, la dépendance est alors tellement faible que le TCL reste valide.
- Le TCL est donc un outil souvent applicable et très puissant.
- Très puissant puisqu'il ne suppose pas connaître la loi des X_i (juste que l'espérance et la variance existent).

Intervalles de confiance

- Le TCL peut être utilisé afin d'obtenir des intervalles de confiance.
- Par exemple un intervalle de confiance à 68% est

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}}, \overline{X} + \frac{\sigma}{\sqrt{n}}\right]$$

- De même un intervalle de confiance à 95% est

$$\left[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

- Et plus généralement un intervalle de confiance à $(1-\alpha)\%$

$$\left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right], \qquad z_{1-\alpha/2} \text{ lu dans la table } N(0,1)$$

- En pratique σ n'est pas connu et remplacé par s défini avant

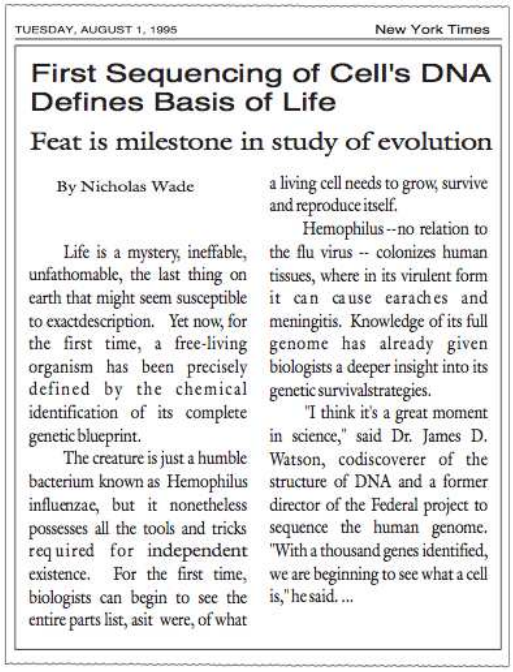
Niveaux de confiance

- Que signifie le terme niveau de confiance à 95% ?
- Si l'on considère plusieurs échantillons $\overline{X}$ prendra des valeurs différentes
- De même on aura donc des intervalles de confiance différents
- Alors la moyenne sur la population μ appartiendra à ces intervalles de confiance dans 95% des cas

Ce que nous avons vu

- Vocabulaire statistique : échantillon, population, unité...
- Moyenne empirique
- Variance empirique
- TCL
- Intervalles de confiance

Motifs dans l'ADN
Estimations et tests



TUESDAY, AUGUST 1, 1995 New York Times

First Sequencing of Cell's DNA Defines Basis of Life

Feat is milestone in study of evolution

By Nicholas Wade

Life is a mystery, ineffable, unfathomable, the last thing on earth that might seem susceptible to exactdescription. Yet now, for the first time, a free-living organism has been precisely defined by the chemical identification of its complete genetic blueprint.

The creature is just a humble bacterium known as Hemophilus influenzae, but it nonetheless possesses all the tools and tricks required for independent existence. For the first time, biologists can begin to see the entire parts list, asit were, of what a living cell needs to grow, survive and reproduce itself.

Hemophilus--no relation to the flu virus -- colonizes human tissues, where in its virulent form it can cause earaches and meningitis. Knowledge of its full genome has already given biologists a deeper insight into its genetic survivalstrategies.

"I think it's a great moment in science," said Dr. James D. Watson, codiscoverer of the structure of DNA and a former director of the Federal project to sequence the human genome. "With a thousand genes identified, we are beginning to see what a cell is," he said. ...

0.10 Problématique

Problématique

- ☐ Le cytomegalovirus humain (CMV) est un virus dangereux pour les personnes immunodéficientes
- ☐ Pour combattre le virus, des chercheurs s'intéressent à sa manière de se répliquer
- ☐ En particulier à un endroit particulier de son ADN, l'origine, qui contient l'information pour sa reproduction
- ☐ L'ADN est formé de seulement 4 lettres (A, C, G, T), une séquence d'ADN contient donc de nombreux motifs.
- ☐ Certains motifs peuvent indiquer une position importante comme l'origine.
- ☐ Un palindrome complémentaire est un de ces motifs pour lequel une séquence lue de droite à gauche correspond au complémentaire de la séquence lue normalement

$$\text{GGGCATGCCC}, \qquad \text{Comp(A)} = \text{C}, \quad \text{Comp(G)} = \text{T}$$

- ☐ Le CMV appartient à la famille des herpès virus comme l'herpès simplex et le virus d'Epstein–Barr
- ☐ Ces deux autres virus marquent l'origine par des palindromes complémentaires

  HS par un palindrome très long (144 lettres)
  EB par un amas de courts palindromes

- ☐ Certains biologistes conjecturent que son origine est marquée par un amas de palindromes
- ☐ Pour localiser l'origine on pourrait découper l'ADN en segments et tester quels sont les segments pouvant se répliquer
- ☐ Ceci est valide mais très coûteux et long

- Une approche statistique permettant d'identifier des amas anormaux de palindromes permettrait d'affiner les zones de recherche
- Le séquençage du CMV a été fait en 1990
- En 1991 des chercheurs ont répertoriés la position de différents motifs
- Le plus long palindrome pour le CMV fait 18 paires de bases et contient 296 palindromes ayant entre 10 et 18 paires de bases
- Nos données sont donc les positions de ces palindromes

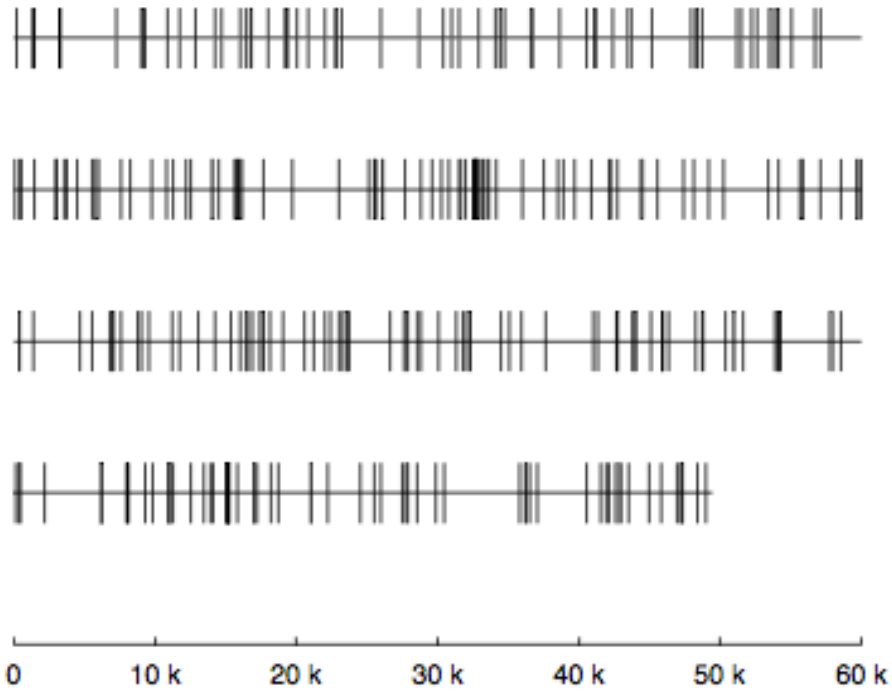


Figure 9: *Positions des 296 palindromes du CMV [Chee et al., 1990].*

0.11 Théorie

Processus de Poisson

- Le processus de Poisson (homogène) est un modèle probabiliste pour l'occurrence de phénomènes aléatoires, e.g., arrivée dans une file d'attente, positions des étoiles dans le ciel...
- Prenons l'exemple de la file d'attente. Il est raisonnable de supposer que
  - chaque personne agit indépendemment
  - pour une personne donnée et pour un laps de temps très court, il est peu probable que cette personne entre dans la file
  - mais puisqu'il y a beaucoup de personnes, il est probable que certains d'entre eux joignent la file durant ce laps de temps

- Le processus de Poisson est un modèle naturel pour modéliser des points (ou temps) distribués complètement au hasard sans régularité apparente
- Ce processus suppose plusieurs hypothèses
  - Le taux d'apparition, noté λ, des points ne dépend pas de l'espace/du temps (homogénéité)
  - Les nombres de points appartenant à deux régions disjointes sont indépendants
  - Les points ne peuvent pas se superposer

Utilité du processus de Poisson

Puisque selon ce modèle les positions des palindromes apparaissent sans aucune régularité, un écart des observations à ce modèle nous indiquera une région où le nombre de palindromes est anormalement élevé.

Il s'agira donc ici

- d'ajuster notre modèle à nos données

$$\Longrightarrow \text{Estimation de paramètres}$$

- de voir si le modèle "colle" aux données

$$\Longrightarrow \text{Test d'adéquation à une loi}$$

Loi de Poisson

- Que peut on bien faire avec des points répartis au hasard ? Compter non ?
- Le processus de Poisson de taux λ suppose que la probabilité qu'il y ait k points dans un intervalle de longueur 1 est

$$\Pr[N = k] = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k \in \mathbb{N}$$

- Une variable aléatoire N ayant cette loi suit une loi de Poisson de paramètre λ notée $Poiss(\lambda)$.
- Le paramètre λ est un taux et représente le nombre de points moyen par unité de longueur
- C'est aussi l'espérance de la loi de Poisson, i.e., $\mathbb{E}[N] = \lambda$ où $N \sim Poiss(\lambda)$
- Enfin si on s'intéressait à un intervalle de taille L alors le nombre de points suivrait une loi $Poiss(\lambda L)$ (homogénéité)

Estimation de paramètres

- Pour notre étude, nous devons ajuster notre modèle (le processus de Poisson) à nos données (les positions des palindromes)
- En statistique on parle alors d'estimation de paramètres (d'une loi)
- Dans ce cours nous allons voir deux méthodes différentes

1. La méthode des moments :

   On parle alors de l'estimateur des moments

2. La méthode du maximum de vraisemblance :

   On parle alors de l'estimateur du maximum de vraisemblance

La méthode des moments

- L'idée de la méthode des moments est très simple

  "On égalise les *moments* de l'échantillon (empirique) à ceux de la population (théorique)"

- Mais c'est quoi un moment ?
- Le k-ième moment (par rapport à l'origine), $k \in \mathbb{N}_*$,
  - Population (théorique) : $\mathbb{E}[X^k]$
  - Échantillon (empirique) : $\frac{1}{n}\sum_{i=1}^{n} X_i^k$

Exemple 1. Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Poiss(\lambda)$, i.e., suivant indépendemment une loi de Poisson. Estimer λ.

Exemple 2. Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, estimer μ et σ^2.

La méthode du maximum de vraisemblance

Définition 1. Soient $x_1, \ldots, x_n$ des données supposées être une réalisation d'un échantillon aléatoire $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x, \theta)$, i.e., de densité f alors la vraisemblance pour θ est

$$L(\theta) = f(x_1; \theta) \times f(x_2; \theta) \times \cdots \times f(x_n; \theta)$$

Définition 2. L'estimateur du maximum de vraisemblance $\hat{\theta}_{ML}$ d'un paramètre θ est celui qui, parmi tous les θ possibles, donne à l'échantillon obtenu la plus grande vraisemblance d'être obtenu, i.e.

$$L(\hat{\theta}_{ML}) \geq L(\theta), \qquad \text{pour tout } \theta$$

Étapes de la méthode

On se facilite grandement les calculs en maximisant $\ell(\theta) := \ln L(\theta)$. Les étapes pour trouver $\hat{\theta}_{ML}$ sont

1. Calculer $L(\theta)$
2. Poser $\ell(\theta) = \ln L(\theta)$ (c'est la log-vraisemblance)
3. Trouver $\hat{\theta}_{ML}$ tel que $\{d\ell/d\theta\}(\hat{\theta}_{ML}) = 0$
4. (Vérifier qu'il s'agit bien d'un maximum)

Exemple 3. Supposons que $x_1, \ldots, x_n$ soient des réalisations indépendantes d'une loi exponentielle, i.e., de densité

$$f(x; \lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0, \quad \lambda > 0.$$

Trouvez $\hat{\lambda}_{ML}$.

Exemple 4. Pareil mais avec une loi $Poiss(\lambda)$.

Propriétés de $\hat{\theta}_{ML}$

- Lorsque les observations sont i.i.d., i.e., de même loi et indépendantes, $\hat{\theta}_{ML}$ a souvent de très bonnes propriétés
- En effet sous des hypothèses de régularités, on a pour n grand

$$\hat{\theta}_{ML} \dot{\sim} N\left(\theta, \frac{1}{nI(\theta)}\right), \qquad I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2}\right]$$

Exemple 5. Trouvez la loi de $\hat{\lambda}_{ML}$ pour l'exemple de la loi exponentielle.

Test d'adéquation à une loi

- En stat, on supposera souvent que les observations sont des réalisations indépendantes d'une certaine loi, e.g., loi de Poisson.
- Cela ne veut bien sûr pas dire que les observations ont exactement cette loi mais que cette loi reproduit bien l'aléa de nos observations
- Pour notre étude, nous aimerions que le processus de Poisson représente bien nos données globalement
- Si tel est le cas, nous pourrions alors détecter une concentration anormale de palindromes dans une région particulière
- Vérifier si un modèle probabiliste représente bien nos données s'appelle un text d'adéquation à une loi
- Dans cette partie nous allons voir le test d'adéquation du χ^2

- Ici nos données sont le nombre de palindromes dans des segments disjoints (4000 paires de bases)

7	1	5	3	8	6	1	4	5	3
6	2	5	8	2	9	6	4	9	4
1	7	7	14	4	4	4	3	5	5
3	6	5	3	9	9	4	5	6	1
7	6	7	5	3	4	4	8	11	5
3	6	3	1	4	8	6			

Table 4: *Nombre de palindromes dans les 57 premiers segments disjoints de l'ADN du CMV. Nombre total = 294.*

- Cette représentation des données n'est pas très pratique
- On va donc les formater d'une manière plus pratique pour notre test

Nombre de palindromes	Nombre d'intervalles observés	attendus
0–2	7	
3	8	
4	10	
5	9	
6	8	
7	5	
8	4	
≥ 9	6	
Total	57	

Table 5: *Distribution du nombre de palindrômes dans les 57 premiers segments.*

- Le nombre d'intervalles attendus correspond au nombre attendu sous notre modèle, i.e., la loi de Poisson

- Par exemple le nombre attendu d'intervalles ayant 3 palindromes est

$$\underbrace{57}_{n} \times \underbrace{\Pr[N=3]}_{\text{proba d'être dans la classe}} = 57 \times \frac{\lambda^3}{3!} e^{-\lambda}$$

- Il faut donc connaître λ
- La méthode des moments (ou du max de vrais.) nous dit que

$$\hat{\lambda} = \overline{X} = 294/57 = 5.16$$

- Ainsi l'effectif attendu pour avoir 3 palindromes est

$$57 \times \Pr[N=3] = 57 \times \frac{5.16^3}{3!} e^{-5.16} = 7.5$$

- On fait de même pour les autres lignes.

Nombre de palindromes	Nombre d'intervalles observés	attendus
0–2	7	6.4
3	8	7.5
4	10	9.7
5	9	10.0
6	8	8.6
7	5	6.3
8	4	4.1
≥ 9	6	4.5
Total	57	57

Table 6: *Distribution du nombre de palindrômes dans les 57 premiers segments.*

- La statistique pour notre test du χ^2 est alors

$$T_{\text{obs}} = \frac{(7-6.4)^2}{6.4} + \frac{(8-7.5)^2}{7.5} + \frac{(10-9.7)^2}{9.7} + \frac{(9-10)^2}{10} + \frac{(8-8.6)^2}{8.6} + \frac{(5-6.3)^2}{6.3} + \frac{(4-4.1)^2}{4.1} + \frac{(6-4.5)^2}{4.5} = 1.0$$

- Notez que T_{obs} est une sorte de "distance" entre nos observations et notre modèle
- Intuitivement, si cette "distance" est petite notre modèle colle aux données, sinon c'est un mauvais modèle
- Question : A partir de quelle distance peut on dire que le modèle est bon ?
- La théorie nous dit que si notre modèle est le bon, alors T_{obs} suit une loi du chi-deux à six degrés de liberté notée χ^2_6
- Pour conclure sur l'adéquation de notre modèle,

$$\Pr[\chi^2_6 > T_{\text{obs}}] = 0.98,$$

  ainsi nous avons de forte chance d'avoir une "distance" supérieure à 1.0.
- On peut conclure que le processus de Poisson représente bien la position des palindromes.

TABLE C.2. Percentiles of the Chisquare distribution—values of c_p corresponding to P.

df	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.1}$	$\chi^2_{.9}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	0.000039	0.00016	0.00098	0.0039	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.61	5.99	7.38	9.21	10.60
3	0.0717	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.57	95.70	100.62	140.23	146.57	152.21	158.95	163.65

Présentation du test plus formelle

- On calcule les effectifs observés pour chaque classe, notés eff. obs$_j$, $j = 1, \ldots, K$ (K nombre de classes)
- On calcule les effectifs théoriques/attendus pour chaque classe

$$\text{eff. théo.}_j = n \Pr[\text{être dans la classe } j]$$

- On calcule la statistique

$$T_{\text{obs}} = \sum_{j=1}^{K} \frac{(\text{eff. obs}_j - \text{eff. théo}_j)^2}{\text{eff. théo}_j}$$

- Si notre modèle est vrai, alors

$$T_{\text{obs}} \sim \chi^2_{K-1-p}, \qquad \text{où } K \text{ nb. de classes, } p \text{ nb. de paramètres estimés}$$

- On calcule la p-valeur

$$p\text{-valeur} = \Pr[\chi^2_{K-1-p} > T_{\text{obs}}]$$

Vocabulaire des tests statistiques

- Un test statistique (d'adéquation ou autre) "compare" toujours 2 hypothèses et vérifie laquelle des deux est la plus vraisemblable à partir des données
- Plus formellement un test statistique s'écrit

$$H_0 : \underbrace{\text{hypothèse nulle}}_{\text{proc. de Poisson}} \quad \text{contre} \quad H_1 : \underbrace{\text{hypothèse alternative}}_{\text{pas proc. de Poisson}}$$

- Une statistique de test T dont la loi est connue sous H_0
- Un niveau de confiance $1 - \alpha$, typiquement $\alpha = 10\%$ ou 5%.
- Une zone de rejet

$$\left\{ x \in \mathbb{R} : \Pr_{H_0}[T > x] < \alpha \right\},$$

pour laquelle on décide en faveur de H_1 si T_{obs} appartient à la zone de rejet.

Les erreurs dans un test statistique

- Un test statistique est toujours associé à deux types d'erreurs
- L'erreur de première espèce

$$\alpha = \Pr_{H_0}[\text{décider en faveur de } H_1]$$

- et l'erreur de seconde espèce

$$\beta = \Pr_{H_1}[\text{décider en faveur de } H_0]$$

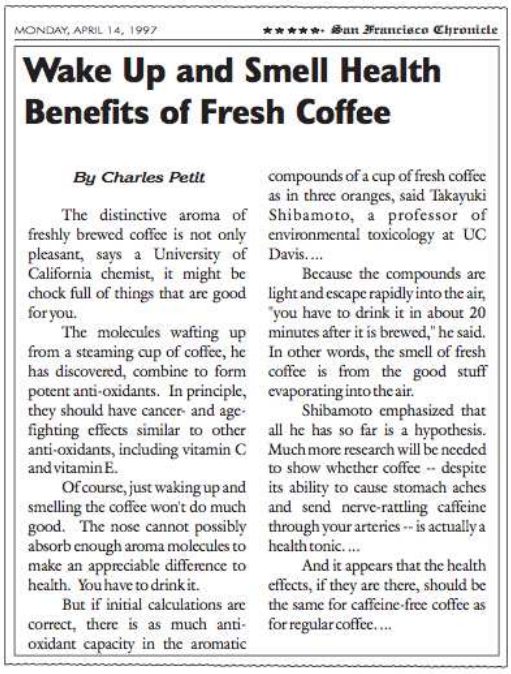
- Pour la plupart des tests, c'est l'utilisateur qui décide de α. Typiquement $\alpha = 10\%$ ou $\alpha = 5\%$.
- Pour α fixé, β est alors déterminée — bien que pas toujours connue explicitement
- A plusieurs tests on préferera le test le plus puissant, i.e., pour α fixé celui qui maximise la puissance

$$1 - \beta = \Pr_{H_1}[\text{décider en faveur de } H_1]$$

Ce que nous avons vu

- Le processus de Poisson homogène
- La loi de Poisson
- Méthode des moments, maximum de vraisemblance
- Test d'adéquation du χ^2
- Vocabulaire des tests statistiques

Saurez-vous faire la différence ?
Plans d'expérience



MONDAY, APRIL 14, 1997 ★★★★★ San Francisco Chronicle

Wake Up and Smell Health Benefits of Fresh Coffee

By Charles Petit

The distinctive aroma of freshly brewed coffee is not only pleasant, says a University of California chemist, it might be chock full of things that are good for you.

The molecules wafting up from a steaming cup of coffee, he has discovered, combine to form potent anti-oxidants. In principle, they should have cancer- and age-fighting effects similar to other anti-oxidants, including vitamin C and vitamin E.

Of course, just waking up and smelling the coffee won't do much good. The nose cannot possibly absorb enough aroma molecules to make an appreciable difference to health. You have to drink it.

But if initial calculations are correct, there is as much anti-oxidant capacity in the aromatic compounds of a cup of fresh coffee as in three oranges, said Takayuki Shibamoto, a professor of environmental toxicology at UC Davis....

Because the compounds are light and escape rapidly into the air, "you have to drink it in about 20 minutes after it is brewed," he said. In other words, the smell of fresh coffee is from the good stuff evaporating into the air.

Shibamoto emphasized that all he has so far is a hypothesis. Much more research will be needed to show whether coffee -- despite its ability to cause stomach aches and send nerve-rattling caffeine through your arteries -- is actually a health tonic....

And it appears that the health effects, if they are there, should be the same for caffeine-free coffee as for regular coffee....

0.12 Problématique

Problématique

- Ronald A. Fisher (1890–1962) a énormément contribué à la statistique moderne
- Une de ses études a été motivé par une lady anglaise prétendant qu'elle pouvait faire la différence entre un thé auquel le lait était ajouté avant ou après le thé.

- Nous allons voir étudier la capacité à différencier le café normal et décaféiné

Les données

- Nos données sont juste le nombre de tasses de cafés identifiées à raison comme "café normal"
- Ces données peuvent donc se mettre sous la forme d'un tableau 2×2

		Café servi		
		Normal	Décaféiné	
Testeur dit	Normal	a	b	$a+b$
	Décaféiné	c	d	$c+d$
		$a+c$	$b+d$	n

- Évidemment le nombre total de tasses servies est $n = a + b + c + d$
- Nous voyons qu'il y a $a + c$ cafés normaux de servis et que le testeur dit qu'il y en a $a + b$.
- Si le sujet sait faire la différence, alors $b \approx 0$ et $c \approx 0$
- Nous allons voir comment tester si l'individu a de telles capacités

0.13 Théorie

Loi hypergéométrique

- Supposons que $n = 8$ tasses sont servies, 4 tasses de chaque sorte
- Le testeur est informé de ce plan d'expérience
- Cela implique donc des contraintes sur notre tableau

$$a + b = a + c = c + d = b + d = 4$$

- En conséquence, la connaissance d'une seule case nous permet de remplir le tableau entièrement, i.e., si a est connu

$$b = 4 - a, \qquad c = 4 - a, \qquad d = a$$

- Nous allons maintenant introduire l'aléatoire

Modèle probabliste

- Considérons les hypothèses suivante

H_0 : le testeur ne sait pas faire la différence

H_1 : le testeur sait faire la différence

- Il y a donc

$\binom{8}{4} = 70$ façons de choisir 4 cafés normaux parmi 8

- Sous H_0 chacune de ses classifications à la même probabilité $1/\binom{8}{4}$
- Mais une seule est la vraie qui a pour probabilité $1/70$

Test exact de Fisher

- Soit N le nombre de cafés normaux mal classifiés
- Alors

$$\Pr[N = a] = \frac{\binom{4}{a}\binom{4}{4-a}}{\binom{8}{4}}, \qquad a = 0, \ldots, 4,$$

cette loi est connue sous le nom de loi hypergéométrique

Table 7: *Probabilités selon la loi hypergéométrique de mal classer a cafés comme normaux.*

Nombre d'erreurs	0	1	2	3	4
Probabilité	$1/70$	$16/70$	$36/70$	$16/70$	$1/70$

- Ce tableau nous donne donc les p-valeurs exactes que le testeur n'a pas ces capacités
- Si le testeur ne fait aucune erreur p-valeur $= 1/70 \approx 0.014$ et s'il en fait une

$$p\text{-valeur} = \Pr[N \leq 1] = \frac{1}{70} + \frac{16}{70} \approx 0.24$$

- Pour ce cas particulier, on ne rejetterai pas H_0 si le sujet fait au plus 1 erreurs — pour les niveaux de confiance usuels de 90% ou 95%.
- En revanche s'il fait aucune erreur on rejetterai H_0 au profit de H_1 : il sait faire la différence
- Ce test est connu sous le nom du text exact de Fisher

Remarque. Notez qu'il y a qu'un nombre fini de classifications possibles. En conséquence il y a un nombre fini de p-valeurs et les régions critiques pour les niveau 90% et 95% peuvent ne pas exister

Un deuxième plan d'expérience

- Pour tester notre sujet sur nos 8 tasses de café, on aurait pu procéder autrement
- Avant le test, on jette une pièce. Si pile, on sert un café normal ; sinon un décaféiné. Le sujet ne voyant rien de tout cela bien sûr.
- Pour ce plan d'expérience il y a maintenant

$$2^8 = 256 \quad \text{classifications possibles}$$

- Sous H_0 chaque classification a pour proba 1 / 256 et une seule est la bonne
- Notons B (resp. C) est la variable aléatoire représentant le nombre de café normaux (resp. décaf) classés à tort comme décaf (resp. normaux)
- La probabilité de faire $b + c$ erreurs est alors

$$\Pr[N = b + c] = \frac{\binom{8}{b+c}}{256}, \qquad b + c = 0, \ldots, 8$$

Un troisième plan d'expérience

- Une autre approche consisterai à servir les cafés par paires, i.e., un normal + un décaféiné
- Comme pour le premier plan, nous n'avons besoin de connaître uniquement le nombre c des cafés normaux classés comme décaféinés
- Puisqu'on ne s'occupe que des cafés classés comme normaux, il y a

$$2^4 = 16 \quad \text{classifications possibles}$$

- Sous H_0 chacune de ces classifications ont la même proba $1/16$ et une seule est la vraie
- La probabilité de faire c erreurs est alors

$$\Pr[N = c] = \frac{\binom{4}{c}}{16}, \qquad c = 0, \ldots, 4$$

- Les plans d'expériences rendent plus ou moins difficile de faire un sans faute
- Sur notre exemple nous avons

Plan d'expérience n°	1	2	3
$\Pr[N = 0]$	0.014	0.004	0.06

- Choisir un bon plan d'expérience est une branche à part entière en statistique — et que nous ne verrons pas

Test approché de Fisher : z-test

- Si n est grand, il est difficile voire impossible de calculer les p-valeurs
- On a alors recours au test approché de Fisher
- L'idée consiste à remplacer la loi hypergéométrique par son approximation selon la loi normale
- Notons par A, B, C et D les variables aléatoires liées à notre tableau de comptage
- Pour le premier plan d'expérience, A suit une loi hypergéométrique et on peut montrer que (admis)

$$\mathbb{E}[A] = \frac{(a+b)(a+c)}{n}$$
$$\mathrm{Var}[A] = \frac{(a+b)(a+c)(b+d)(c+d)}{n \times n \times (n-1)}$$

- Il convient donc de centrer et réduire a

$$z = \frac{a - \mathbb{E}[A]}{SD(A)} \approx \frac{a - \frac{(a+b)(a+c)}{n}}{\sqrt{\frac{(a+b)(a+c)(b+d)(c+d)}{n^3}}} = \frac{\sqrt{n}(ad-bc)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

- La statistique z suit approximativement une loi normale centrée réduite et les p-valeurs sont donc déduites de cette loi
- La loi normale étant continue et la loi hypergéométrique discrète, on utilisera souvent la correction $a + 0.5$ dans l'expression de z

Tableaux de contingence

- Un tableau 2×2 classant les sujets selon une variable binaire est appelé un tableau de contingence
- Pour la leçon 2 sur les jeux vidéos on a

		Sexe		
		Homme	Femme	
Aime jouer	Oui	43	26	69
	Non	8	12	20
		51	38	89

- Supposons que nous voulions tester si le sexe a une influence sur le fait d'aimer jouer, i.e., en termes statistiques

H_0 : aimer jouer et le sexe sont indépendants

H_1 : aimer jouer et le sexe ne sont pas indépendants

Test d'indépendance du χ^2

- Vous vous rappelez du test d'adéquation du χ^2 non ?

$$\sum_{j=1}^{K} \frac{(\text{eff. obs}_j - \text{eff. théo}_j)^2}{\text{eff. théo}_j}, \qquad K \text{ nb. de classes}$$

- Pour nous on a donc $K = 4$
- Si le sexe est indépendant du fait d'aimer jouer alors

$$\pi_A = \alpha\beta, \quad \pi_B = \alpha(1-\beta), \quad \pi_C = (1-\alpha)\beta, \quad \pi_D = (1-\alpha)(1-\beta),$$

où α est la proba. de choisir un étudiant qui aime jouer et β la proba. de choisir un étudiant homme.

- Les effectifs théoriques sont alors

$$\mathbb{E}[A] = n\alpha\beta, \quad \mathbb{E}[B] = n\alpha(1-\beta), \quad \mathbb{E}[C] = n(1-\alpha)\beta, \quad \mathbb{E}[D] = n(1-\alpha)(1-\beta)$$

- α et β étant inconnu, il paraît naturel de les estimer par leur proportions empiriques
- On a alors pour A

$$\mathbb{E}[A] \approx n\frac{a+b}{n}\frac{a+c}{n} = 89 \times \frac{69}{89} \times \frac{51}{89} = 39.5$$

- Pour les autres effectifs théoriques on trouve 11.5, 29.5 et 8.5
- La statistique de test vaut alors

$$\frac{(43-39.5)^2}{39.5} + \frac{(26-29.5)^2}{29.5} + \frac{(8-11.5)^2}{11.5} + \frac{(12-8.5)^2}{8.5} = 3.2$$

- La p-valeur est alors $\Pr[\chi^2_{4-1-2} > 3.2] = 0.08$
- Au niveau 5% on ne rejette donc pas H_0 et le sexe est indépendant du fait d'aimer jouer ou non.

z-test sur deux échantillons (comparaison de proportions)

- Supposons maintenant que les étudiants hommes et femmes ont été échantillonnés indépendamment
- Alors on a

$$A \sim Bin(51, \gamma_A), \quad C = 51 - A, \quad B \sim Bin(38, \gamma_B), \quad D = 38 - B,$$

où γ_A (resp. γ_B) est la proba. d'aimer jouer chez une homme (resp. femme)
- Une autre formulation de notre question de base serait alors

$$H_0 : \gamma_A = \gamma_B = \gamma \qquad \text{contre} \qquad H_1 : \gamma_A \neq \gamma_B$$

- On estime facilement ces deux paramètres par leurs versions empiriques

$$\hat{\gamma}_A = \frac{a}{a+c} = \frac{43}{51}, \qquad \hat{\gamma}_B = \frac{b}{b+d} = \frac{26}{38}$$

- Sous H_0, $\frac{A}{a+c} - \frac{B}{b+d}$ a pour espérance 0 et variance

$$\gamma(1-\gamma)\left(\frac{1}{a+c} + \frac{1}{b+d}\right)$$

- Ainsi la statistique

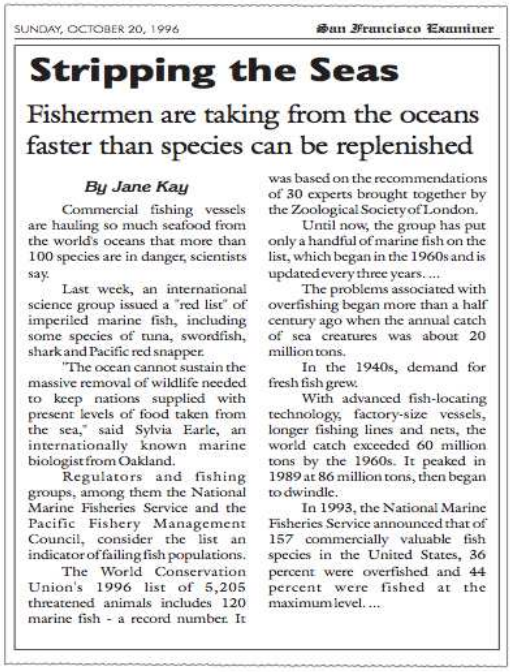
$$z = \frac{\frac{A}{a+c} - \frac{B}{b+d}}{\sqrt{\gamma(1-\gamma)\left(\frac{1}{a+c} + \frac{1}{b+d}\right)}} \dot{\sim} N(0,1)$$

- γ est estimé naturellement par $\frac{a+b}{n}$, de sorte que $z_{\text{obs}} = 1.8$ et la p-valeur vaut $\Pr[|Z| > 1.8] = 0.08$
- Remarquons que $\sqrt{3.2} = 1.8$ ce qui est normal puisque $z^2 \sim \chi^2_1$
- Ce test se confond avec le test d'indépendance du χ^2.

Ce que nous avons vu

- ☐ Loi hypergéométrique
- ☐ Test exact de Fisher
- ☐ z-test
- ☐ Tableaux de contingence
- ☐ Test d'indépendance du χ^2
- ☐ z-test sur deux échantillons

Courbe de croissance du crabe dormeur
Modèle linéaire simple



SUNDAY, OCTOBER 20, 1996 San Francisco Examiner

Stripping the Seas

Fishermen are taking from the oceans faster than species can be replenished

By Jane Kay

Commercial fishing vessels are hauling so much seafood from the world's oceans that more than 100 species are in danger, scientists say.

Last week, an international science group issued a "red list" of imperiled marine fish, including some species of tuna, swordfish, shark and Pacific red snapper.

"The ocean cannot sustain the massive removal of wildlife needed to keep nations supplied with present levels of food taken from the sea," said Sylvia Earle, an internationally known marine biologist from Oakland.

Regulators and fishing groups, among them the National Marine Fisheries Service and the Pacific Fishery Management Council, consider the list an indicator of failing fish populations.

The World Conservation Union's 1996 list of 5,205 threatened animals includes 120 marine fish - a record number. It was based on the recommendations of 30 experts brought together by the Zoological Society of London.

Until now, the group has put only a handful of marine fish on the list, which began in the 1960s and is updated every three years. ...

The problems associated with overfishing began more than a half century ago when the annual catch of sea creatures was about 20 million tons.

In the 1940s, demand for fresh fish grew.

With advanced fish-locating technology, factory-size vessels, longer fishing lines and nets, the world catch exceeded 60 million tons by the 1960s. It peaked in 1989 at 86 million tons, then began to dwindle.

In 1993, the National Marine Fisheries Service announced that of 157 commercially valuable fish species in the United States, 36 percent were overfished and 44 percent were fished at the maximum level. ...

0.14 Problématique

Problématique

- ☐ Aux USA les crabes dormeurs sont péchés sur la côte ouest de décembre à juin
- ☐ Chaque année presque tous les crabes mâles adultes sont péchés
- ☐ Les femelles sont relâchées afin préserver la ressource
- ☐ Afin de réduire les fluctuations du nombre annuel de crabes péchés, il a été demandé de pouvoir pécher les femelles
- ☐ Se pose donc la question du "gabarit" des carapaces des femelles indiquant la relâche ou la capture
- ☐ Il s'agit donc de modéliser la courbe de croissance des carapaces des femelles

Les données

- Tailles avant et après mue sur 472 crabes dormeurs femelles
- Données mixtes issues de laboratoires et de capture-recapture
- Les données issues de capture-recapture ont été obtenue "en marquant" 12 000 crabes
- Afin d'obtenir à nouveau ces crabes auprès des pêcheurs, une loterie avec un prix de 500$ a été effectué

Avant mue	113.6	118.1	142.3	125.1	98.2	119.5	116.2
Après mue	127.7	133.2	154.8	142.5	120.0	134.1	133.8
Accroissement	14.1	15.1	17.4	21.8	14.6	17.6	
Source	0	0	1	1	1	1	1

Table 8: *Partie du tableau de données des 472 tailles (mm) des femelles crabes. Source : 0 si laboratoire, 1 sinon.*

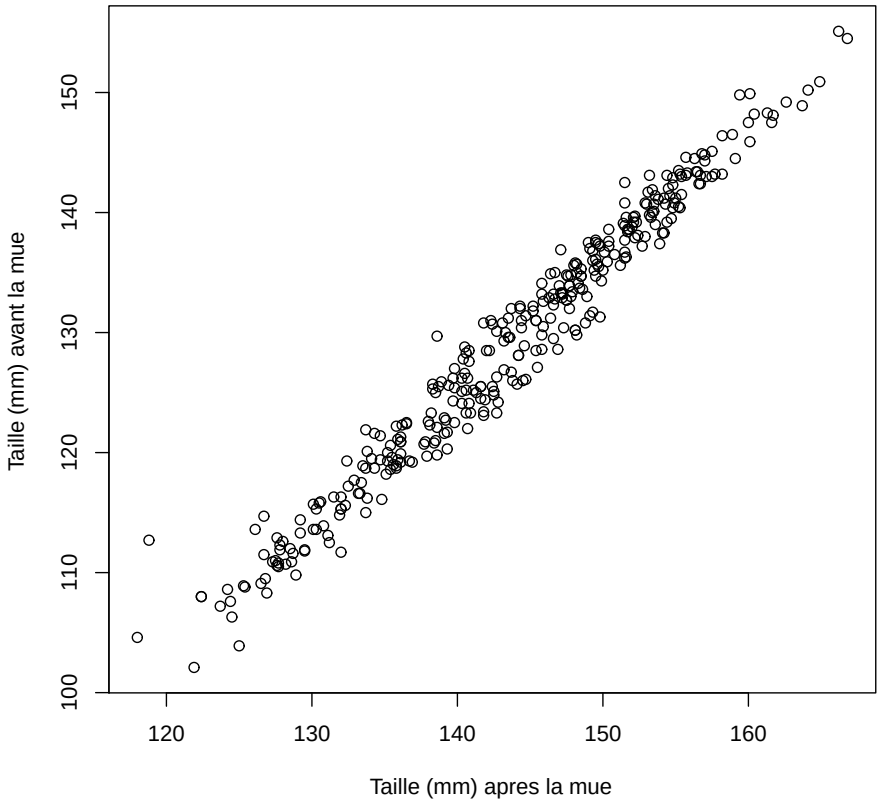


Figure 10: *Taille des carapaces des femelles crabes après et avant la mue.*

0.15 Théorie

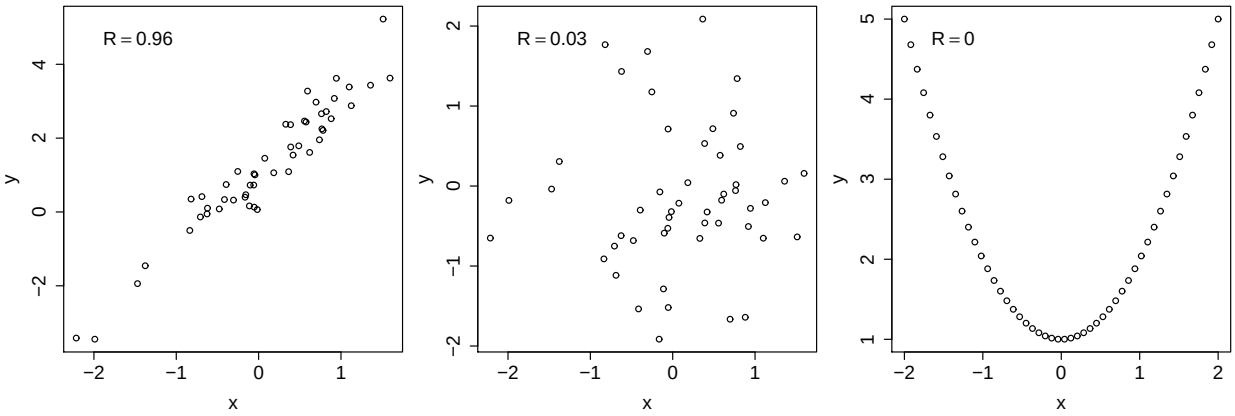
Coefficient de corrélation

- La figure précédente nous montre qu'il y a une forte relation linéaire entre la taille avant et après la mue, i.e., les points s'amassent autour d'une droite
- Le coefficient de corrélation (linéaire) mesure la force de cette relation
- Soient $(x_1, y_1), \ldots, (x_n, y_n)$ les couples des tailles après et avant la mue
- Le coefficient de corrélation est donné par

$$R = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \overline{x}}{SD(x)} \times \frac{y_i - \overline{y}}{SD(y)}, \qquad \text{Crabes: } R = 0.98$$

- C'est une mesure sans unité (on a standardisé les x_i et y_i) et qui varie entre -1 et 1.
- Lorsque $R = 1$ ou $R = -1$ les points sont parfaitement alignés sur une droite dont la pente est du signe de R

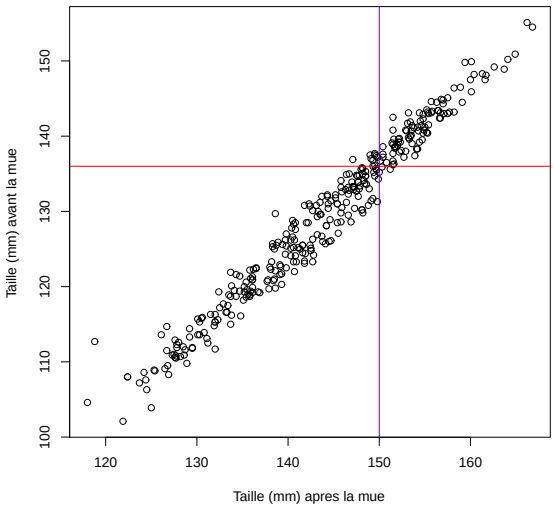
Une mesure de la dépendance linéaire seulement !

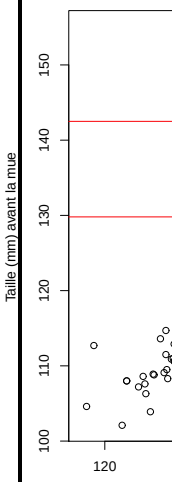


- Le premier nuage de points montre une forte dépendance linéaire. L'utilisation de R est justifiée
- Le deuxième nuage de points montre aucune réelle dépendance entre x et y. L'utilisation de R est justifiée
- Le troisième montre une dépendance parfaite mais non linéaire entre x et y. L'utilisation de R est maladroite
- Il est donc toujours bon de visualiser le nuage de points

Vers le modèle linéaire

A partir de la figure, on voit qu'un crabe ayant une taille après la mue de 150mm a une taille avant la mue d'environ 136mm En prenant cette fois des tailles entre 147.5 et 152.5, on trouve 69 crabes. Leur taille avant mue moyenne est 150 et l'écart-type 2.8 En faisant pareil pour 8 classes de tailles





- En cherchant à prédire la taille avant mue d'un crabe de taille 150 mm après mue, il paraît raisonnable d'énoncer la taille 136 En cherchant à prédire la taille avant mue d'un crabe de taille 150 mm après mue, il paraît raisonnable d'énoncer la taille 136 associée à une erreur de 2.8 En cherchant à prédire la taille avant mue d'un crabe de taille après mue quelconque, on serait tenter de tracer une droite passant par les $*$ et d'avoir une "enveloppe d'erreur"
- Cette droite est elle la meilleure possible ?

Quelques notations / rappel

- Moyennes empiriques $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Variances empiriques

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

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \overline{y^2} - \overline{y}^2$$

- Écart-types empiriques $SD(x) = \sqrt{V(x)}$, $SD(y) = \sqrt{V(y)}$
- Covariance empirique

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x}\,\overline{y},$$

  i.e., la moyenne des produits moins le produit des moyennes

Méthode des moindres carrés

- Sous certaines hypothèses la méthode des moindres carrés donne la meilleure prédiction possible
- Elle consiste à trouver a et b minimisant

$$\sum_{i=1}^{n}(y_i - ax_i - b)^2, \quad \text{i.e., } \sum(\text{taille}_{\text{avant mue}} - a\text{taille}_{\text{après mue}} - b)^2$$

- L'estimateur des moindres carrés, i.e., la solution au problème ci-dessus, est

$$\hat{a} = \frac{S_{xy}}{V(x)}, \qquad \hat{b} = \overline{y} - \hat{a}\overline{x}$$

- La droite de régression ou droite des moindres carrés est alors

$$\hat{y} = \hat{a}x + \hat{b},$$

i.e., $\hat{y}$ est la meilleure prédiction de y sachant x.

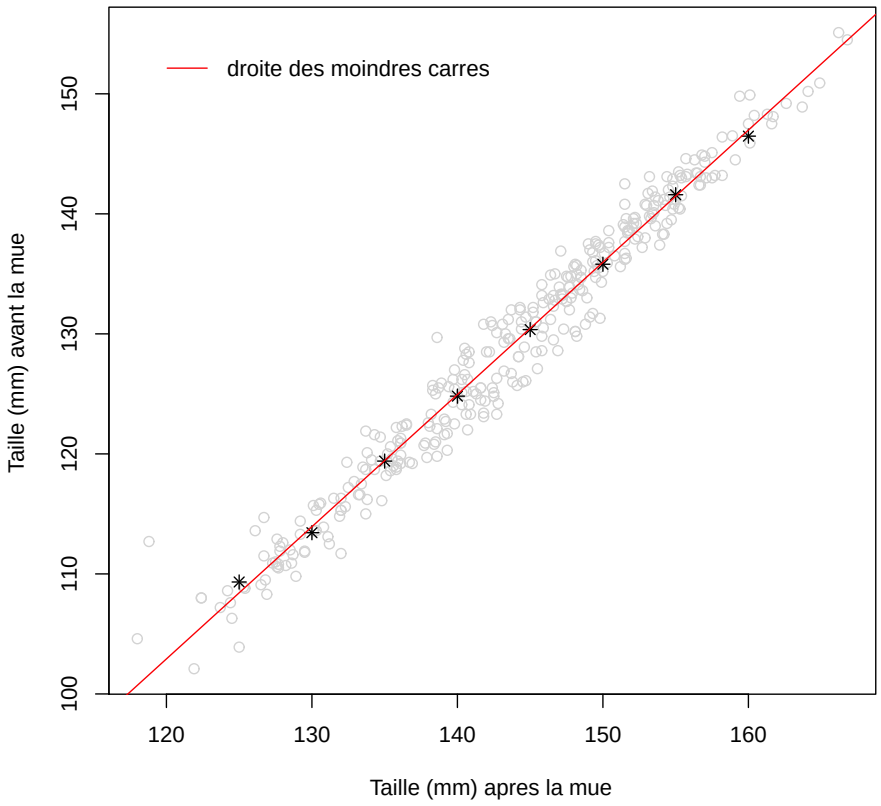


Figure 11: *Droite des moindres carrés pour les tailles des crabes — $\hat{a} = 1.1, \hat{b} = -29$. Les tailles moyennes selon 8 classes sont représentées par le symbole $*$.*

Le modèle linéaire simple

- N'étant pas naïf, nous savons bien que la droite $\hat{y} = \hat{a}x + \hat{b}$ n'est pas parfaite
- Il est donc souvent utile d'imposer une loi de probabilité sur les erreurs
- Le modèle linéaire simple suppose que les erreurs, notées ε_i, sont gaussiennes, i.e.,

$$Y_i = ax_i + b + \varepsilon_i, \qquad i = 1, \ldots, n,$$

  où $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ et a et b sont les paramètres de régression à estimer.
- Puisque les ε_i sont des variables aléatoires, la réponse Y_i l'est également mais les observations y_i non
- Hypothèse importante d'homoscédasticité : la variance des erreurs est constante
- Attention les erreurs, qui sont des variables aléatoires, ne correspondent pas aux résidus r_i, qui sont juste des nombres.

Utilité des hypothèses supplémentaires

- On appelle résidus

$$r_i = y_i - \hat{a}x_i - \hat{b} = y_i - \hat{y}_i$$

- Attention les résidus ne sont pas les erreurs ε_i
- La variance des erreurs σ^2 est estimée sans biais par la variance résiduelle

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n-2}, \qquad \text{car 2 paramètres de régression}$$

- Les estimateurs $\hat{a}$ et $\hat{b}$ suivent une loi normale

$$\hat{a} \sim N\left(a, \frac{\sigma^2}{nV(x)}\right), \qquad \hat{b} \sim N\left(b, \frac{\sigma^2 \overline{x^2}}{nV(x)}\right)$$

- En pratique on remplacera σ^2 qui est inconnue par son estimation $\hat{\sigma}^2$

- Il est également possible de connaître la distribution de $\hat{y}_* = \hat{a}x_* + \hat{b}$, pour un x_* donné.

$$\hat{y}_* \sim N\left\{ax_* + b, \sigma^2\left(1 + \frac{1}{n} + \frac{(\overline{x} - x_*)^2}{V(x)}\right)\right\}$$

- Comme précédemment, on remplacera σ^2 qui est inconnue par son estimation $\hat{\sigma}^2$.
- Les résultats précédents permettent d'obtenir des intervalles de confiances et de faire des test d'hypothèses.

Tests d'hypothèses

- Il est souvent utile de tester si un des paramètres de la droite de régression vaut une valeur particulière, typiquement 0
- Par exemple pour tester

$$H_0 : a = a_0 \qquad \text{contre} \qquad H_1 : a \neq a_0$$

on va utiliser la statistique de test

$$T = \frac{\hat{a} - a_0}{SE(\hat{a})} \sim t_{n-2}, \qquad \text{sous } H_0$$

- Exemple : Pour $a_0 = 0$, on a $T_{\text{obs}} = (1.1 - 0)/0.011 = 100$ et puisque $\Pr[|t_{342-2}| > |T_{\text{obs}}|] \approx 0$, on rejette donc H_0.
- La même méthode s'applique bien entendu pour b.

Intervalles de confiances

- En utilisant le même résultat sur la distribution de T, on peut obtenir des intervalles de confiance
- Par exemple un intervalle de confiance à 95% pour a est

$$[\hat{a} - t_{n-2,0.975} \times SE(\hat{a}), \hat{a} + t_{n-2,0.975} \times SE(\hat{a})]$$

- Pour la pente de notre exemple nous obtenons

$$[1.1 - 1.96 \times 0.011, 1.1 + 1.96 \times 0.011] = [1.08; 1.12]$$

- Alors que pour l'ordonnée à l'origine

$$[-29 - 1.96 \times 1.6, -29 + 1.96 \times 1.6] = [-32; -26]$$

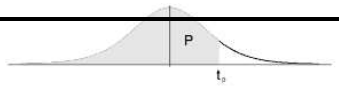


TABLE C.3. Percentiles of the t distribution—values of t_p corresponding to P.

df	$t_{.60}$	$t_{.70}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
∞	0.253	0.524	0.842	1.282	1.645	1.96	2.326	2.576

Pour voir si vous avez tout retenu

Voici une sortie de *R* sur notre exemple

```
> fit <- lm(pre.size ~ post.size) ## Ajuste le modêle linéaire
> summary(fit)                    ## Affiche plein d'informations

Call:
lm(formula = pre.size ~ post.size)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6233 -1.3044  0.1231  1.3016 11.1038

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.26843    1.58114  -18.51   <2e-16 ***
post.size     1.10155    0.01098  100.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.998 on 340 degrees of freedom
Multiple R-squared: 0.9673,Adjusted R-squared: 0.9672
F-statistic: 1.007e+04 on 1 and 340 DF,  p-value: < 2.2e-16
```

Propriétés de la droite de régression

- La droite des moindres carrés passe par le point $(\overline{x}, \overline{y})$
- $\sum_{i=1}^{n} r_i = 0$
- $\sum_{i=1}^{n} x_i r_i = 0$
- $\sum_{i=1}^{n} \hat{y}_i r_i = 0$

Ainsi

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 = \ldots = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} r_i^2$$

conduisant à la décomposition de la somme des carrés totale

$$SC_{\text{Total}} = SC_{\text{R}} + SC_{\text{E}}$$

en une partie due à la régression et une partie due à l'erreur.

- Pour notre exemple, $SC_{\text{Total}} = SC_{\text{R}} + SC_{\text{E}} = 40192 + 1357$.

Coefficient de détermination

- Nous venons de voir la décomposition de la somme des carrés totale

$$SC_{\text{Total}} = SC_{\text{R}} + SC_{\text{E}}$$

- La proportion de la variation totale expliquée par le modèle

$$R^2 = \frac{SC_{\text{R}}}{SC_{\text{Total}}} = \frac{SC_{\text{Total}} - SC_{\text{E}}}{SC_{\text{Total}}}$$

  est appelé coefficient de détermination; $0 \leq R^2 \leq 1$.
- Si
  - $R^2 \approx 1$, alors $y_i \approx \hat{y}_i$ et donc $r_i \approx 0$: le modèle explique les données presque parfaitement
  - $R^2 \approx 0$, alors $\hat{a} \approx 0$ et x n'explique presque rien de la variation totale.

- Pour notre exemple, nous avons donc

$$R^2 = \frac{40192}{41549} = 0.96$$

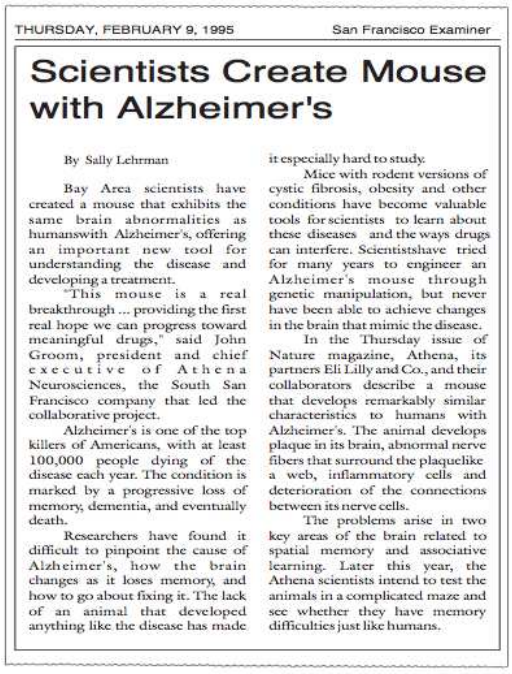
- Le modèle explique donc presque toute la variation

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.6233 -1.3044  0.1231  1.3016 11.1038

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.26843    1.58114  -18.51   <2e-16 ***
post.size     1.10155    0.01098  100.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.998 on 340 degrees of freedom
Multiple R-squared: 0.9673,Adjusted R-squared: 0.9672
F-statistic: 1.007e+04 on 1 and 340 DF,  p-value: < 2.2e-16
```


Ce que nous avons vu

- Coefficient de corrélation
- Méthode des moindres carrés
- Modèle linéaire simple
- Tests d'hypothèses sur a, b
- Intervalles de confiances sur a, b
- Coefficient de détermination

Trisomie 21 chez les souris
Analyse de la variance



THURSDAY, FEBRUARY 9, 1995 San Francisco Examiner

Scientists Create Mouse with Alzheimer's

By Sally Lehrman

Bay Area scientists have created a mouse that exhibits the same brain abnormalities as humanswith Alzheimer's, offering an important new tool for understanding the disease and developing a treatment.

"This mouse is a real breakthrough ... providing the first real hope we can progress toward meaningful drugs," said John Groom, president and chief executive of Athena Neurosciences, the South San Francisco company that led the collaborative project.

Alzheimer's is one of the top killers of Americans, with at least 100,000 people dying of the disease each year. The condition is marked by a progressive loss of memory, dementia, and eventually death.

Researchers have found it difficult to pinpoint the cause of Alzheimer's, how the brain changes as it loses memory, and how to go about fixing it. The lack of an animal that developed anything like the disease has made it especially hard to study.

Mice with rodent versions of cystic fibrosis, obesity and other conditions have become valuable tools for scientists to learn about these diseases and the ways drugs can interfere. Scientistshave tried for many years to engineer an Alzheimer's mouse through genetic manipulation, but never have been able to achieve changes in the brain that mimic the disease.

In the Thursday issue of Nature magazine, Athena, its partners Eli Lilly and Co., and their collaborators describe a mouse that develops remarkably similar characteristics to humans with Alzheimer's. The animal develops plaque in its brain, abnormal nerve fibers that surround the plaquelike a web, inflammatory cells and deterioration of the connections between its nerve cells.

The problems arise in two key areas of the brain related to spatial memory and associative learning. Later this year, the Athena scientists intend to test the animals in a complicated maze and see whether they have memory difficulties just like humans.

0.16 Problématique

Problématique

- ☐ La trisomie 21 est un syndrome congénital survenant lorsqu'un enfant reçoit un chromosome 21 supplémentaire de ses parents
- ☐ Aux USA, 250000 personnes sont atteintes de ce syndrome
- ☐ En 1980, on a découvert que seuls les gênes en "bas" du chromosome 21 étaient à l'origine de ce syndrome
- ☐ Les scientifiques travaillent encore afin de mieux localiser le(s) gêne(s) responsable(s)
- ☐ Dans ce but des portions du chromosome 21 humain sont ajouter à l'ADN de souris de laboratoire
- ☐ Si la souris transgénique montre les symptômes alors la portion d'ADN contient le(s) gêne(s) responsable(s)

- ☐ Afin de déterminer si une souris est atteinte, elle est soumis à des tests (visuels) d'apprentissage, d'intelligence
- ☐ Malheureusement 500 des souris de laboratoires sont aveugles
- ☐ Pour ces dernières seule leur masse est connue
- ☐ Se pose donc la question de savoir si le poids peut aider à mieux cerner le(s) gêne(s) responsable(s)

Données

- Le centre du génome humain au laboratoire Lawrence Berkeley a construit un panel de souris transgéniques
- Chacune de ces souris "souches" contient une des quatre parties du chromosome 21 humain
- La deuxième génération est issue de la reproduction des souris "souches" et d'autres souris non transgénique
- Ils ont procédé de même pour les générations suivantes, de sorte que la généalogie de chaque souris est parfaitement connue

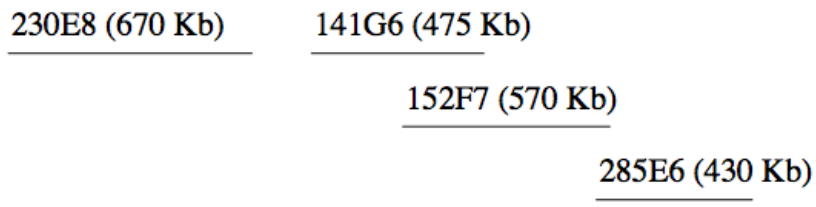


Figure 12: *Localisation des quatres fragment d'ADN du chromosome 21.*

Données

ADN	C	C	C	C	A	A	A	A	A
Lignée	50	50	50	50	4	4	28	28	28
Transgénique	1	0	0	1	1	1	0	1	0
Sexe	1	1	0	0	1	1	1	1	1
Age	113	113	112	112	119	119	115	115	115
Poids	31.6	30.1	23.1	26.3	31.2	28.4	28.1	30.1	29.1
Cage	1	1	5	5	7	7	9	9	10

Table 9: *Parties des observations issues de 532 souris transgéniques. Transgénique : Oui (1). Sexe : Mâle (1), Age (jours), Poids (g), Cage : numéro de la cage.*

- Pour la suite on notera
  - A le fragment 141G6
  - B le fragment 152F7
  - C le fragment 230E8
  - D le fragment 285E6

0.17 Théorie

La moyenne empirique vue par les moindres carrés

- Soient $y_1, \ldots, y_n$ le poids de nos souris
- Il est facile de voir que la moyenne empirique $\overline{y}$ minimise la fonction

$$f(\beta) = \sum_{i=1}^{n} (y_i - \beta)^2$$

- En effet en dérivant par rapport à β il vient

$$f'(\beta) = 0 \iff -2\sum_{i=1}^{n}(y_i - \beta) = 0 \iff n\beta = \sum_{i=1}^{n} y_i$$
$$\iff \beta = \overline{y}$$

- Il existe un lien entre la moyenne empirique est les moindres carrés

- Si nous avons plusieurs groupes, par exemple A, B, C et D, on pourrait être intéressé à estimer la moyenne pour chaque groupe
- Considérons deux groupes pour commencer. Transgénique (T) et non transgénique (NT)
- On introduit alors des variables binaires $e_1, \ldots, e_n$ valant 1 si la souris est transgénique et 0 sinon
- L'estimation par les moindres carrés de

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 e_i)^2$$

est $\hat{\beta}_0 = \overline{y}_{\mathrm{NT}}$ et $\hat{\beta}_1 = \overline{y}_{\mathrm{T}} - \overline{y}_{\mathrm{NT}}$, où

$$\overline{y}_{\mathrm{T}} = \frac{1}{n_{\mathrm{T}}} \sum_{(\mathrm{T})} y_i, \qquad \overline{y}_{\mathrm{NT}} = \frac{1}{n_{\mathrm{NT}}} \sum_{(\mathrm{NT})} y_i$$

Pourquoi c'est ça !#?#?

- Les souris sont soit transgénique soit non transgénique mais certainement pas les deux à la fois !
- Ainsi

$$\begin{aligned} f(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 e_i)^2 \\ &= \sum_{(\mathrm{NT})} (y_i - \beta_0)^2 + \sum_{(\mathrm{T})} (y_i - \beta_0 - \beta_1)^2 \end{aligned}$$

- Chaque somme se fait sur un groupe disjoint de souris
- Puisqu'un carré est positif, minimiser $f(\beta_0, \beta_1)$ revient à minimiser chacune des sommes
- Cela revient donc à faire deux moindres carrés et donc

$$\hat{\beta}_0 = \overline{y}_{\mathrm{NT}}, \qquad \hat{\beta}_0 + \hat{\beta}_1 = \overline{y}_{\mathrm{T}}.$$

Généralisation

- Revenons maintenant à nos 4 groupes A, B, C, D
- On introduit alors 4 variable binaires e_A, e_B, e_C et e_D

$$e_{A,i} = \begin{cases} 1, & \text{si la } i\text{-ième souris a le fragment A} \\ 0, & \text{sinon} \end{cases}$$

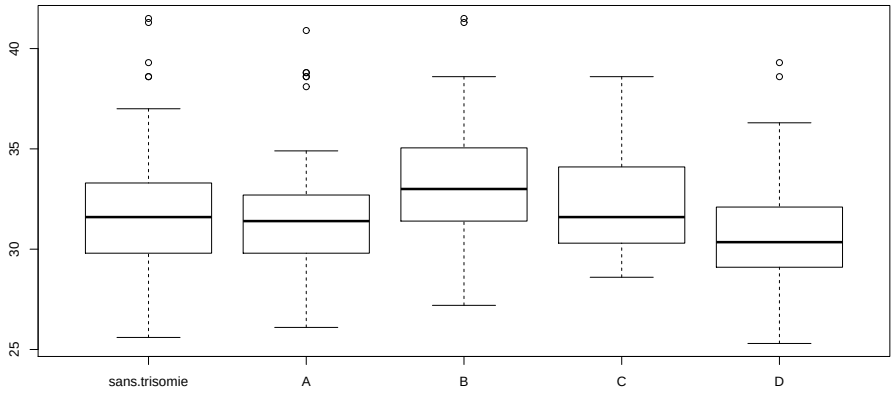
- On minimise alors par rapport à $(\beta_0, \beta_A, \beta_B, \beta_C, \beta_D)$

$$\begin{aligned} &\sum_{i=1}^{n} (y_i - \beta_0 - \beta_A e_{A,i} - \beta_B e_{B,i} - \beta_C e_{C,i} - \beta_D e_{D,i})^2 \\ &= \sum_{\text{(sans trisomie)}} (y_i - \beta_0)^2 + \sum_{\text{(A)}} (y_i - \beta_0 - \beta_A)^2 + \sum_{\text{(B)}} (y_i - \beta_0 - \beta_B)^2 + \\ &\quad \sum_{\text{(C)}} (y_i - \beta_0 - \beta_C)^2 + \sum_{\text{(D)}} (y_i - \beta_0 - \beta_D)^2 \end{aligned}$$

- La solution des moindres carrés à ce problème est donc

$$\hat{\beta}_0 = \overline{y}_0, \quad \hat{\beta}_A = \overline{y}_A - \overline{y}_0, \quad \hat{\beta}_B = \overline{y}_B - \overline{y}_0, \quad \hat{\beta}_C = \overline{y}_C - \overline{y}_0 \quad \hat{\beta}_D = \overline{y}_D - \overline{y}_0$$

- Lorsque nous avons beaucoup de groupes, il est souvent pertinent de faire un boxplot pour chaque groupe afin d'avoir une première idée



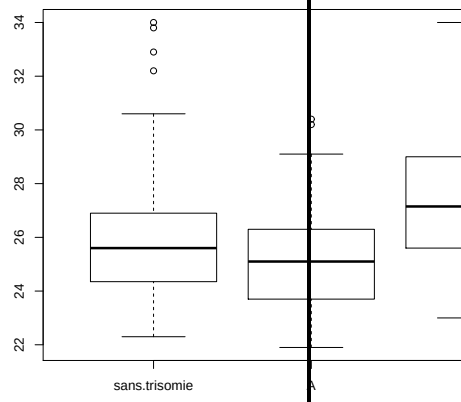


Figure 13: *Boxplot du poids des souris mâles selon leur catégorie.*

Figure 14: *Boxplot du poids des souris femelles selon leur catégorie.*

Modèle pour la moyenne

- Nous venons juste de voir que les moyennes sur chaque groupe pouvait être calculées par la méthode des moindres carrés
- En fait nous avons obtenu des estimations du modèle suivant

$$\mathbb{E}[Y_i] = \begin{cases} \beta_0, & \text{sans trisomie} \\ \beta_0 + \beta_A, & \text{si le fragment A est présent} \\ \beta_0 + \beta_B, & \text{si le fragment B est présent} \\ \beta_0 + \beta_C, & \text{si le fragment C est présent} \\ \beta_0 + \beta_D, & \text{si le fragment D est présent} \end{cases}$$

- Ce modèle a l'avantage de s'interpréter facilement
- Par exemple, β_A représente la différence entre la moyenne des poids transgénique A et des souris sans trisomie.

t-test le retour

- Typiquement on serait intéresser à savoir si le fragment A n'a pas d'influence
- En version statistique cela s'écrit donc

$$H_0 : \beta_A = 0 \qquad \text{contre} \qquad H_1 : \beta_A \neq 0$$

- La statistique pour ce test est $T = \hat{\beta}_A / SE(\hat{\beta}_A)$
- Et sous H_0 et quelques hypothèses de régularités,

$$T \sim t_{n-5}, \qquad \text{car 5 groupes/paramètres}$$

- On rejettera donc H_0 dès lors que la p-valeur

$$\Pr[|t_{n-5}| > |T_{\text{obs}}|]$$

sera petite, e.g., inférieure à 0.10 ou 0.05

Somme des carrés

- Pour tous ajustements par moindre carrés on a la décomposition

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SC_{\text{Total}}}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SC_{\text{E}}}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{SC_{\text{R}}}$$

- Pour notre cas particulier, cette expression se simplifie puisque

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}\sum_{g=1}^{5}(y_i-\overline{y}_g)^2$$
$$\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2=\sum_{g=1}^{5}n_g(y_i-\overline{y}_g)^2,$$

avec $\overline{y}_g=n_g^{-1}\sum_{(g)}y_i$.

ANOVA (ANalysis Of VAriance)

Table 10: *Tableau ANOVA.*

	ddl	Somme des carrés	Carré Moyen	F-statistique
Groupe	$5-1$	$\sum_{i=1}^{5}n_g(\overline{y}_g-\overline{y})^2$	$\frac{\sum_{i=1}^{5}n_g(\overline{y}_g-\overline{y})^2}{5-1}$	$\frac{\sum_{i=1}^{5}n_g(\overline{y}_g-\overline{y})^2/(5-1)}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-5)}$
Résidu	$n-5$	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-5}$	—
Total	$n-1$	$\sum_{i=1}^{n}(y_i-\overline{y})^2$		

Théorème 2. *Si les Y_i sont non corrélés, $\mathrm{Var}[Y_i]=\sigma^2$ et les moyennes $\overline{Y}_g$, $g=1,\ldots,G$ suivent (approximativement) une loi Normale. Alors*

$$\frac{\sum_{i=1}^{G}n_g(\overline{y}_g-\overline{y})^2/(G-1)}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-G)}\sim F_{G-1,n-G},$$

où F_{ν_1,ν_2} représente une loi de Fisher à ν_1 et ν_2 degrés de liberté.

La table de Fisher

TABLE C.4. Percentiles of the F distribution—values of F_{90}, for k degrees of freedom in the numerator and m degrees of freedom in the denominator.

$m\backslash k$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	39.86	49.5	53.59	55.83	57.24	58.2	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.2	5.18	5.18	5.17	5.16	5.15	5.14
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.9	3.87	3.84	3.83	3.82	3.8	3.79	3.78
5	4.06	3.78	3.62	3.52	3.45	3.4	3.37	3.34	3.32	3.3	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.9	2.87	2.84	2.82	2.8	2.78	2.76	2.74
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.7	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.5	2.46	2.42	2.4	2.38	2.36	2.34	2.32
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.3	2.28	2.25	2.23	2.21	2.18
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.2	2.18	2.16	2.13	2.11	2.08
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.3	2.27	2.25	2.21	2.17	2.12	2.1	2.08	2.05	2.03	2.00
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.1	2.06	2.04	2.01	1.99	1.96	1.93
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.2	2.16	2.14	2.1	2.05	2.01	1.98	1.96	1.93	1.9	1.88
14	3.1	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.1	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83
15	3.07	2.7	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.9	1.87	1.85	1.82	1.79
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75
17	3.03	2.64	2.44	2.31	2.22	2.15	2.1	2.06	2.03	2.0	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72
18	3.01	2.62	2.42	2.29	2.2	2.13	2.08	2.04	2.0	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69
19	2.99	2.61	2.4	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.7	1.67
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.0	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.9	1.86	1.81	1.76	1.73	1.7	1.67	1.64	1.6
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.8	1.74	1.72	1.69	1.66	1.62	1.59
24	2.93	2.54	2.33	2.19	2.1	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.7	1.67	1.64	1.61	1.57
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54
27	2.9	2.51	2.3	2.17	2.07	2.0	1.95	1.91	1.87	1.85	1.8	1.75	1.7	1.67	1.64	1.6	1.57	1.53
28	2.89	2.5	2.29	2.16	2.06	2.0	1.94	1.9	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52
29	2.89	2.5	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.5
40	2.84	2.44	2.23	2.09	2.0	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.6	1.54	1.51	1.48	1.44	1.4	1.35
120	2.75	2.35	2.13	1.99	1.9	1.82	1.77	1.72	1.68	1.65	1.6	1.55	1.48	1.45	1.41	1.37	1.32	1.26

$m\backslash k$	1	2
1	161.45	199.5
2	18.51	19.0
3	10.13	9.55
4	7.71	6.94
5	6.61	5.79
6	5.99	5.14
7	5.59	4.74
8	5.32	4.46
9	5.12	4.26
10	4.96	4.1
11	4.84	3.98
12	4.75	3.89
13	4.67	3.81
14	4.6	3.74
15	4.54	3.68
16	4.49	3.63
17	4.45	3.59
18	4.41	3.55
19	4.38	3.52
20	4.35	3.49
21	4.32	3.47
22	4.3	3.44
23	4.28	3.42
24	4.26	3.4
25	4.24	3.39
26	4.23	3.37
27	4.21	3.35
28	4.2	3.34
29	4.18	3.33
30	4.17	3.32
40	4.08	3.23
60	4.0	3.15
120	3.92	3.07

Utilité de l'ANOVA

- Permet de savoir si notre modèle est bon globalement
- Plus formellement, le test s'écrit

$$H_0 : \beta_A = \beta_B = \beta_C = \beta_D = 0$$
$$H_1 : \text{au moins un des } \beta_A, \beta_B, \beta_C, \beta_D \text{ est non nul}$$

- L'ordonnée à l'origine, i.e., β_0, n'est pas concernée
- Ne dit pas quel coefficient est nul, faire un t-test pour cela
- Idéalement on souhaite rejeter H_0, ce qui dans notre cas implique que le poids moyen dépend d'au moins un des groupes

```
>anova(fitMiceMale)
Analysis of Variance Table

Response: weight2
           Df  Sum Sq Mean Sq F value    Pr(>F)
DNA2        4  191.31  47.828  6.1443 9.725e-05 ***
Residuals 260 2023.88   7.784
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
>summary(fitMiceMale)
Call:
lm(formula = weight2 ~ DNA2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3276 -1.8276 -0.3103  1.4535  9.5724

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.9276     0.2723 117.262   <2e-16 ***
DNA2A         0.1708     0.4424   0.386   0.6998
DNA2B         1.2826     0.5232   2.452   0.0149 *
DNA2C         0.5224     0.7938   0.658   0.5111
DNA2D        -1.6811     0.5051  -3.328   0.0010 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.79 on 260 degrees of freedom
Multiple R-squared: 0.08636,Adjusted R-squared: 0.07231
F-statistic: 6.144 on 4 and 260 DF,  p-value: 9.725e-05
```


L'ANOVA en général

- L'ANOVA ne s'applique pas qu'aux moyennes selon modalité(s) mais aux modèles linéaires en général
- Le principe reste exactement le même

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_{p,i} + \varepsilon_i, \qquad i = 1, \ldots, n$$

- L'ANOVA test ici

$$H_0 \colon \beta_1 = \ldots = \beta_p = 0, \quad \text{contre} \quad H_1 \colon \exists i \in \{1, \ldots, p\}, \beta_i \neq 0$$

- Et la statistique suivra alors une $F_{p-1,n-p}$

Retour sur la courbe de croissance des crabes dormeurs

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.6233 -1.3044  0.1231  1.3016 11.1038

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.26843    1.58114  -18.51   <2e-16 ***
post.size     1.10155    0.01098  100.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.998 on 340 degrees of freedom
Multiple R-squared: 0.9673,Adjusted R-squared: 0.9672
F-statistic: 1.007e+04 on 1 and 340 DF,  p-value: < 2.2e-16
```

- ☐ Bien entendu ici la variable explicative "taille après mue" est utile
- ☐ Vous devez à present tout connaître sur cette sortie numérique, hormis `Adjusted R-squared` que nous ne verrons pas

Ce que nous avons vu

- ☐ Lien entre moyenne empirique et moindre carrés
- ☐ Modèle pour la moyenne selon catégorie
- ☐ Retour sur le t-test
- ☐ ANOVA et loi de Fisher