

Biostatistiques – Licence 2

BIO2006L

Marc Bailly-Bechet

Université Claude Bernard Lyon I – France

`marc.bailly-bechet@univ-lyon1.fr`

Table des matières

- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité
- 3 Statistiques descriptives, estimation et intervalles de confiance
- 4 Tests d'hypothèses
- 5 Tests du χ^2
- 6 ANOVA 1
- 7 ANOVA 2
- 8 Analyse bivariable et corrélation
- 9 Modèle et régression linéaire
- 10 Décomposition de variance et test de linéarité

Organisation globale du module

- $\frac{2}{3}$ de biostatistiques, avec 9 CM et 14 TD de 1h30, en début de semestre
- $\frac{1}{3}$ de bioinformatique, avec 3 CM de 1h30 et 5 TP de 3h00, en fin de semestre

L'UE étant en CCI, il n'y a aucune seconde session, ni pour les biostatistiques, ni pour la bioinformatique.

Notation et examens

Biostatistiques (67%) :

- 2 contrôles (CC) au cours du semestre, soit 33% de la note finale au total.
- 1 examen en fin de semestre : 34% de la note finale.

Bioinformatique (33%) :

- 1 rapport de TP : 8% de la note finale
- 1 examen de TP sur machine en fin de semestre : 25% de la note finale.

Des rattrapages seront organisés si besoin en fin de semestre pour les éventuelles absences **justifiées**.

Comment réussir en stats ?

Comment réussir en stats ?

~~En étant naturellement super fort, c'est dans les gènes~~

En travaillant au moins une heure par semaine sur le module, en dehors des cours et TD

Comment bien travailler en stats ?

- En assistant aux cours et aux TD
- En préparant les exercices de TD chez vous
- En interagissant avec les enseignants et les autres étudiants
- En employant des livres de statistiques pour s'entraîner, à la BU :
 - Probabilités et statistiques (Couty/Debord/Fredon) 570.15 COU (Rappels de cours succints et beaucoup d'exercices corrigés)
 - Biostatistique Licence PCEM PCEP (Beauscart), coll. Omniscience 570.15 VAL
 - Biostatistique pour les sciences de la vie et de la santé par Triola, Marc M. – Triola, Mario F 570.15 TRI

Contacts

Pour les interactions, plusieurs possibilités :

- Le forum du cours sur Spiral-Connect
- Par email :

Cours `marc.bailly-bechet@univ-lyon1.fr`

TD voir TOMUSS

Organisation `marie-claude.venner@univ-lyon1.fr`

- De vive voix : Bat. Mendel, Mezzanine, bureau "Lacroix/Mary".

Pourquoi faire des statistiques en biologie ?

Pourquoi faire des statistiques en biologie ?

- Variabilité** : Une expérience en biologie donne rarement un résultat tranché ou parfaitement reproductible.
- Quantité** : Les nouvelles technologies biologiques permettent de recueillir des quantités faramineuses de données.

Les statistiques vues de loin

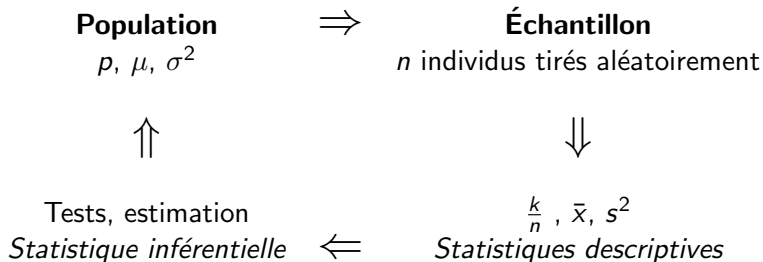
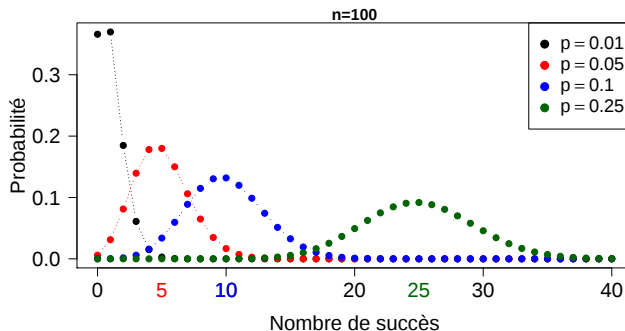


Table des matières

- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité**
- 3 Statistiques descriptives, estimation et intervalles de confiance
- 4 Tests d'hypothèses
- 5 Tests du χ^2
- 6 ANOVA 1
- 7 ANOVA 2
- 8 Analyse bivariable et corrélation
- 9 Modèle et régression linéaire
- 10 Décomposition de variance et test de linéarité

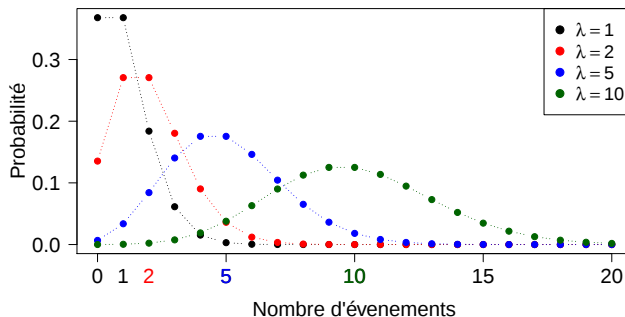
Loi binomiale

La loi binomiale est la loi de probabilité décrivant le *nombre* de réussites parmi un ensemble de tirages aléatoires et indépendants. Elle se note $\mathcal{B}(n, p)$ avec n le nombre de tirages et p la probabilité de réussite à chaque tirage.



Loi de Poisson

La loi de Poisson (de Siméon Denis Poisson, 1781-1840) est la loi de probabilité décrivant le *nombre d'événements* aléatoires et indépendants arrivant dans le même intervalle de temps ou d'espace. Elle se note $\mathcal{P}(\lambda)$ avec λ l'espérance et la variance de la loi.

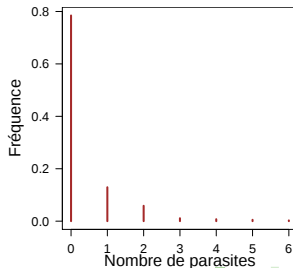
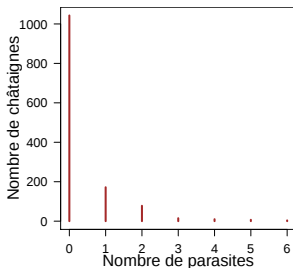


Représentation graphique d'une variable discrète : diagramme en bâtons

La distribution de données discrètes se représente à l'aide d'un diagramme en bâtons. Ici, on s'intéresse à un parasite de la châtaigne, le balanin (*Curculio elephas*).

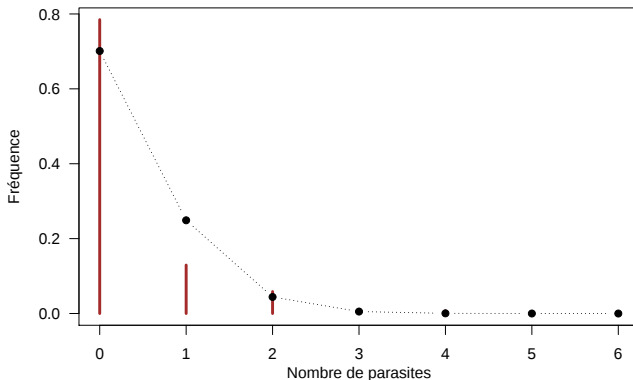
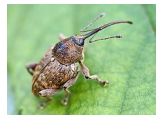


Nb. de parasites x_i	0	1	2	3	4	5	6 et plus
Nombre de fruits n_i ayant x_i parasites	1043	172	78	15	10	7	4
Fréquence $f_i = \frac{n_i}{\sum_i n_i}$	0.785	0.129	0.059	0.011	0.007	0.005	0.004

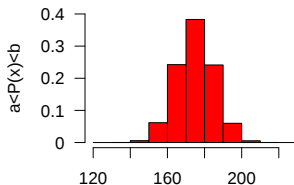
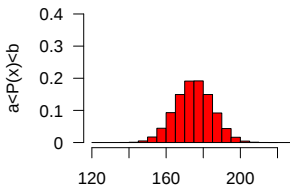
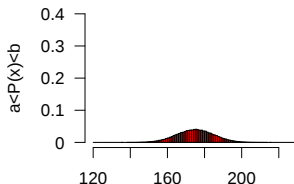
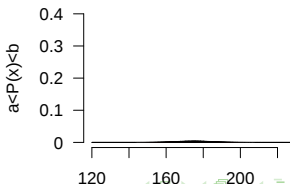


Une variable qui semble suivre une loi de Poisson

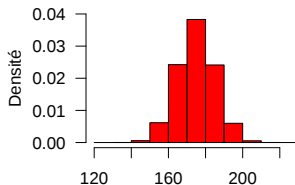
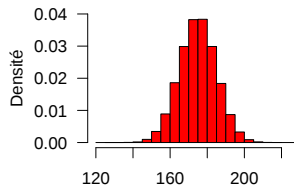
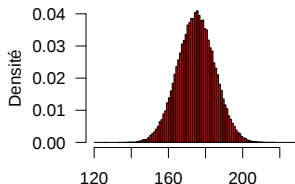
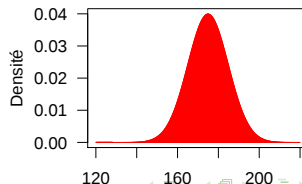
Les données de parasitisme du balanin suivent une décroissance exponentielle, et on peut supposer que les parasites sont indépendants dans chaque châtaigne. On trace sur le diagramme en bâtons une loi de Poisson :



Probabilité absolue

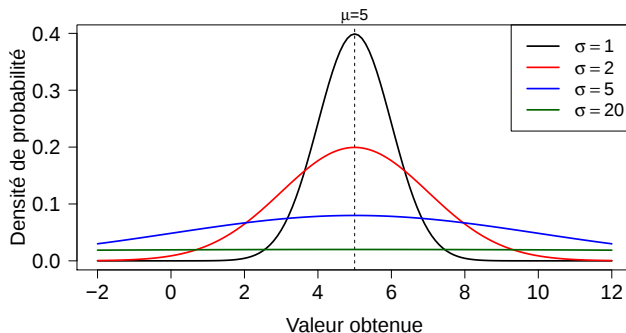
Pas de 10 cm**Pas de 5 cm****Pas de 1 cm****Pas de 0.1 cm**

Densité de probabilité

Pas de 10 cm**Pas de 5 cm****Pas de 1 cm****Limite continue**

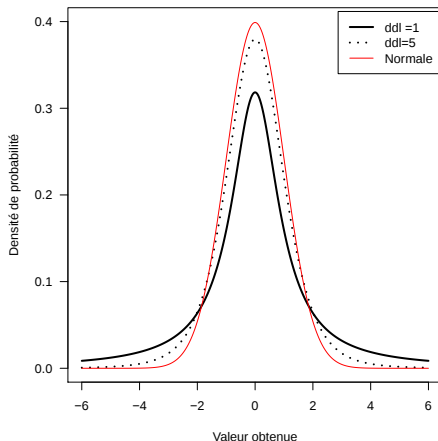
Loi normale

La loi normale est la loi de probabilité des variables aléatoires continues dépendantes d'un grand nombre de causes indépendantes et additives. Elle se note $\mathcal{N}(\mu, \sigma)$ avec μ l'espérance de la loi et σ l'écart-type.



Loi de Student

C'est une variante de la loi normale, plus étalée sur les ailes. Ici on la compare à une loi normale centrée réduite.



Représentation graphique d'une variable continue : histogramme

On s'intéresse au mélange de mangues issu de $k = 4$ différentes récoltes pour faire du jus. Chaque type de mangue a un taux de glucose différent et relativement imprécis.



Concentration (g.L^{-1})	Moyenne	Nb de mangues
X	x_j^*	n_j
$[135, 165[$	150	17
$[165, 180[$	172.5	23
$[180, 195[$	187.5	14
$[195, 225[$	210	8

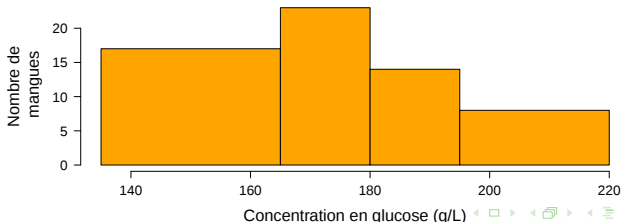


Table des matières

- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité
- 3 Statistiques descriptives, estimation et intervalles de confiance**
- 4 Tests d'hypothèses
- 5 Tests du χ^2
- 6 ANOVA 1
- 7 ANOVA 2
- 8 Analyse bivariable et corrélation
- 9 Modèle et régression linéaire
- 10 Décomposition de variance et test de linéarité

Moyenne **observée** sur des données groupées

On veut la moyenne du taux de glucose dans le mélange final de nos 4 types de mangues :



Concentration (g.L ⁻¹)	Moyenne	Nb de mangues
X	x_j^*	n_j
[135, 165[	150	17
[165, 180[	172.5	23
[180, 195[	187.5	14
[195, 225[	210	8

$$\bar{x} = \frac{1}{62} (150 \times 17 + 172.5 \times 23 + \dots) = \frac{10822.5}{62} = 174.56 \text{ g.L}^{-1}$$

Variance et écart-type **observés**, données groupées

La variance sur des données groupées se calcule ainsi :

Concentration (g.L ⁻¹)	Moyenne	Nb de mangues
X	x_j^*	n_j
[135, 165[	150	17
[165, 180[	172.5	23
[180, 195[	187.5	14
[195, 225[	210	8



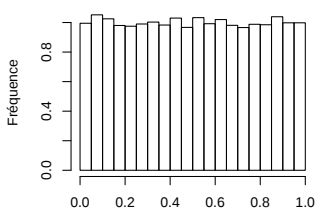
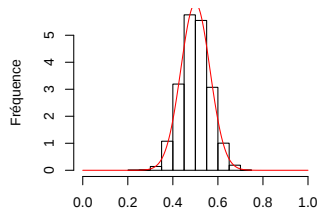
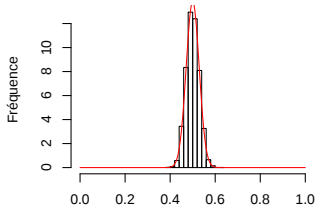
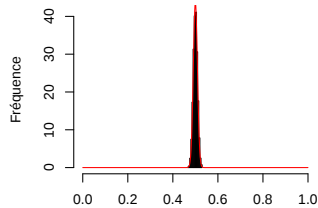
$$\bar{x} = 174.56 \text{ g.L}^{-1}$$

$$s^2 = \frac{1}{62} (17 \times 150^2 + 23 \times 172.5^2 + \dots) - 174.56^2$$

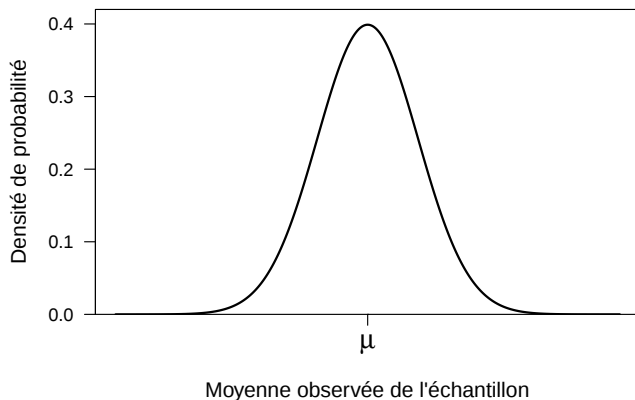
$$= 365.60$$

$$s = \sqrt{365.60} = 19.12 \text{ g.L}^{-1}$$

Loi de la moyenne de n v.a., n grand

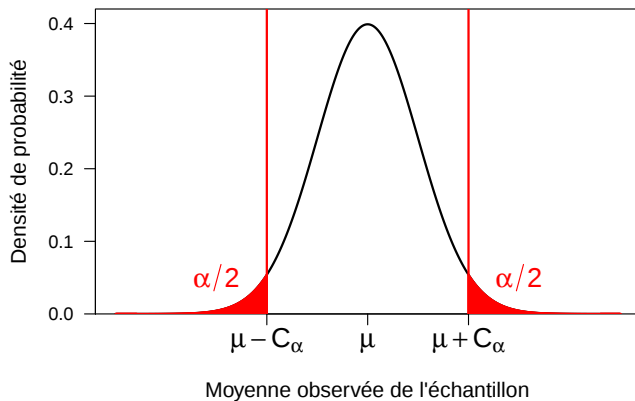
 $n=1$  $n=20$  $n=100$  $n=1000$

Distribution d'échantillonnage d'une moyenne observée



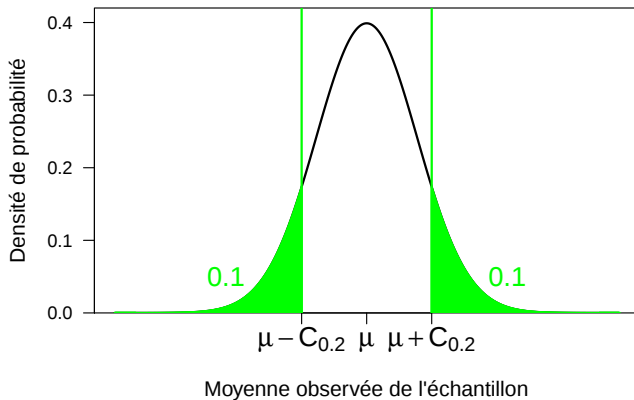
Quantiles de la loi normale

$$P(\mu - C_\alpha < \bar{x} < \mu + C_\alpha) = 1 - \alpha$$



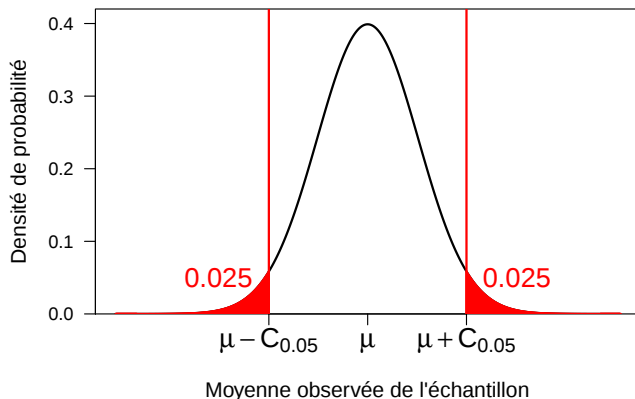
Quantiles de la loi normale, $\alpha = 0.20$

$$P(\mu - C_{0.20} < \bar{x} < \mu + C_{0.20}) = 0.80$$



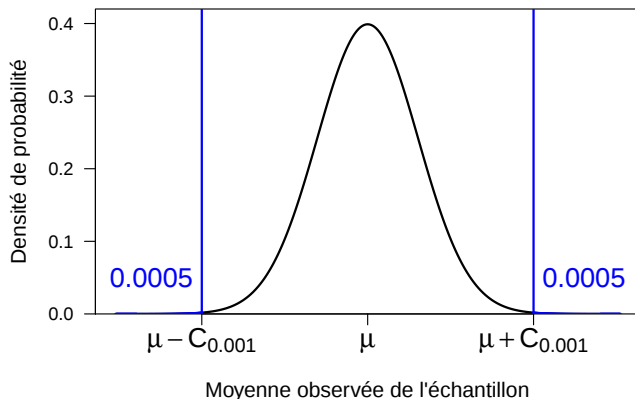
Quantiles de la loi normale, $\alpha = 0.05$

$$P(\mu - C_{0.05} < \bar{x} < \mu + C_{0.05}) = 0.95$$



Quantiles de la loi normale, $\alpha = 0.001$

$$P(\mu - C_{0.001} < \bar{X} < \mu + C_{0.001}) = 0.999$$



Quantiles de la loi normale centrée réduite

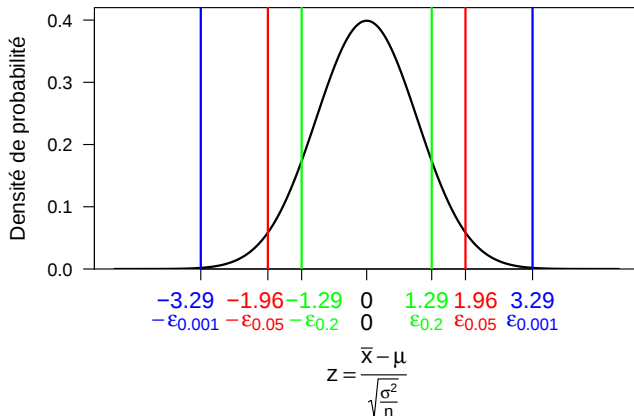


Table des écarts réduits

α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	∞	2.576	2.326	2.170	2.054	1.960	1.881	1.812	1.751	1.695
0.10	1.645	1.598	1.555	1.514	1.476	1.440	1.405	1.372	1.341	1.311
0.20	1.282	1.254	1.227	1.200	1.175	1.150	1.126	1.103	1.080	1.058
0.30	1.036	1.015	0.994	0.974	0.954	0.935	0.915	0.896	0.878	0.860
0.40	0.842	0.824	0.806	0.789	0.772	0.755	0.739	0.722	0.706	0.690
0.50	0.674	0.659	0.643	0.628	0.613	0.598	0.583	0.568	0.553	0.539
0.60	0.524	0.510	0.496	0.482	0.468	0.454	0.440	0.426	0.412	0.399
0.70	0.385	0.372	0.358	0.345	0.332	0.319	0.305	0.292	0.279	0.266
0.80	0.253	0.240	0.228	0.215	0.202	0.189	0.176	0.164	0.151	0.138
0.90	0.126	0.113	0.100	0.088	0.075	0.063	0.050	0.038	0.025	0.013

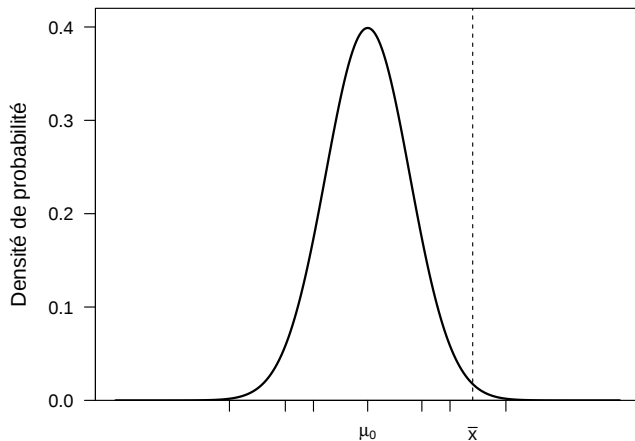
La probabilité α s'obtient par addition des nombres écrits en marge.

Exemple : pour $\alpha = 0.00 + 0.05 = 0.05$, $\varepsilon = 1.960$.

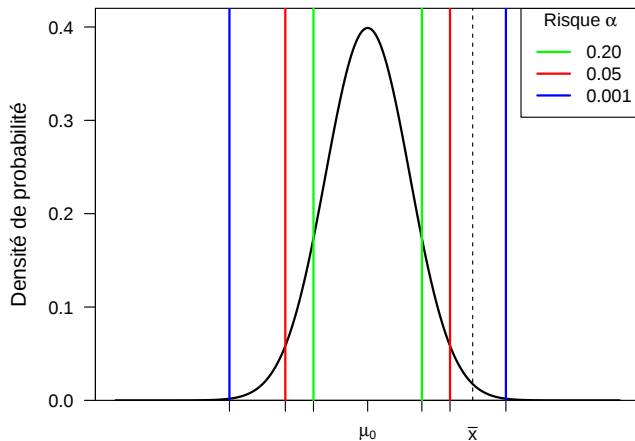
Table des matières

- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité
- 3 Statistiques descriptives, estimation et intervalles de confiance
- 4 Tests d'hypothèses**
- 5 Tests du χ^2
- 6 ANOVA 1
- 7 ANOVA 2
- 8 Analyse bivariable et corrélation
- 9 Modèle et régression linéaire
- 10 Décomposition de variance et test de linéarité

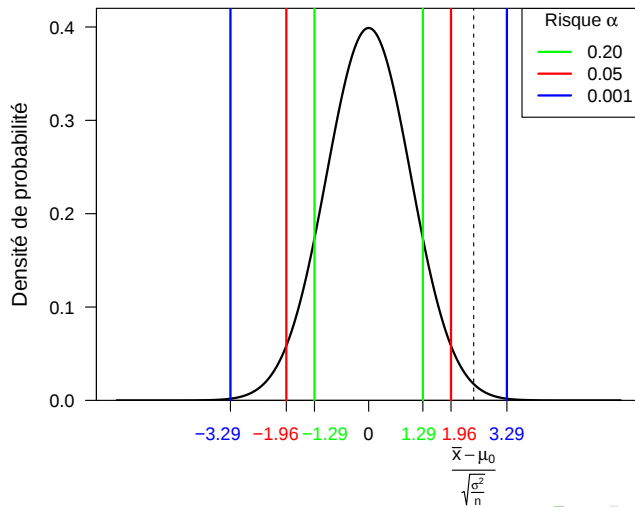
Distribution d'échantillonnage et moyenne observée



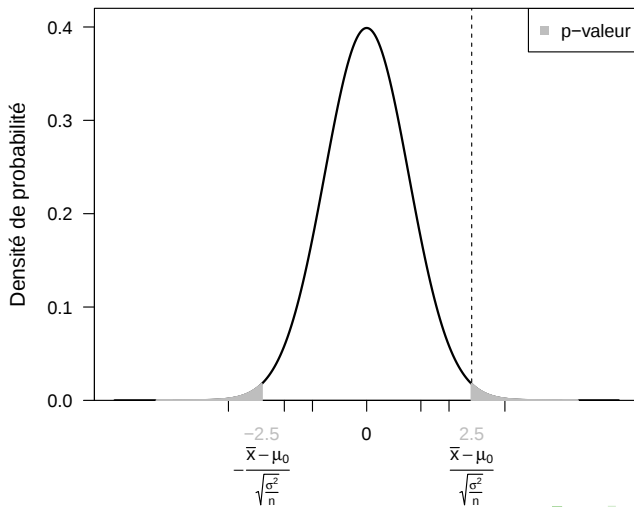
Distribution d'échantillonnage et moyenne observée



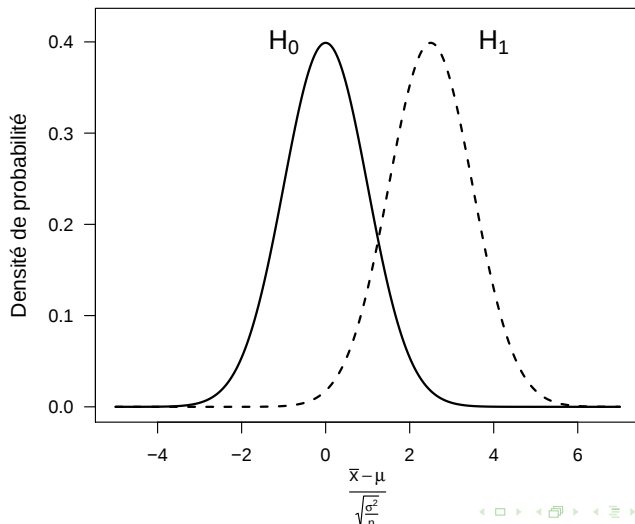
Distribution d'échantillonnage centrée réduite



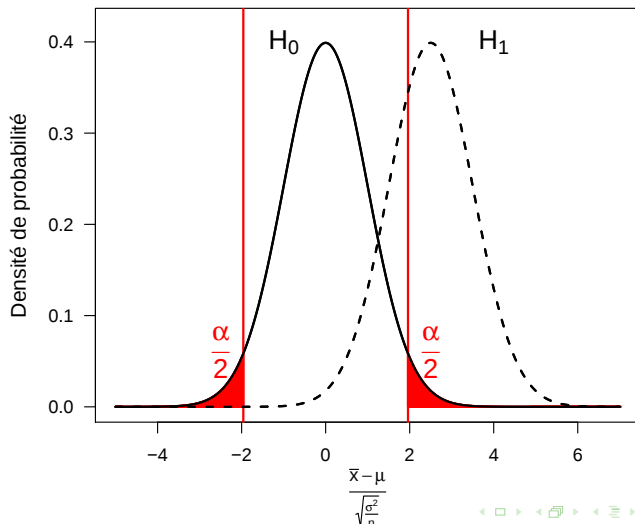
Le monstre sacré : la p -valeur



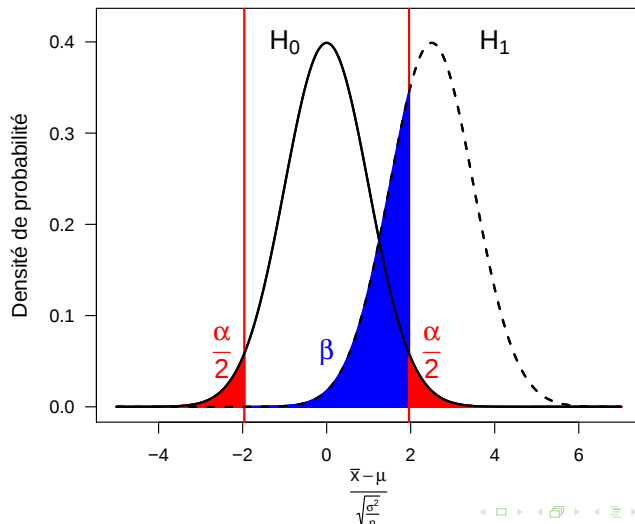
Risque de deuxième espèce



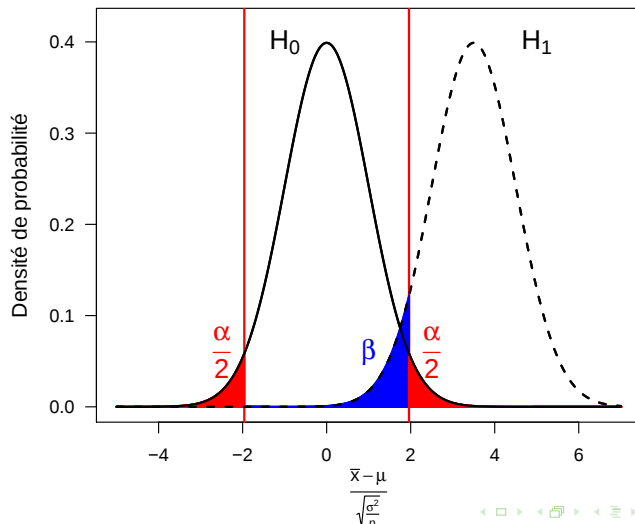
Risque de deuxième espèce



Risque de deuxième espèce

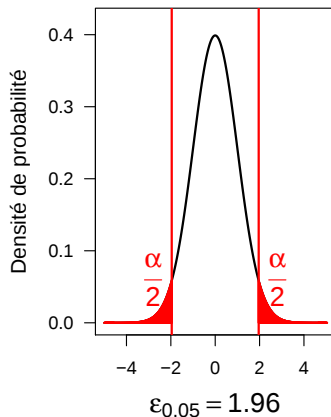


β dépend d'une réalité inconnue

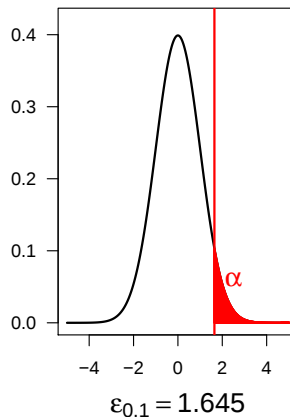


Test unilatéral, $\alpha = 5\%$

$$H_1 : \mu \neq \mu_0$$



$$H_1 : \mu > \mu_0$$



Temps d'hibernation et influence géographique

On dispose des temps d'hibernation de différents individus, dans un échantillon des Alpes ($n_A = 6$) et un des Pyrénées ($n_P = 5$).



Alpes	195	184	166	185	149	151
Pyrénées	224	186	188	213	169	

Temps d'hibernation et influence géographique

On dispose des temps d'hibernation de différents individus, dans un échantillon des Alpes ($n_A = 6$) et un des Pyrénées ($n_P = 5$).



Alpes	195	184	166	185	149	151
Pyrénées	224	186	188	213	169	

Test de Fisher de comparaison des variances :

$$\bar{x}_A = 171.67, \hat{\sigma}_A^2 = 369.47$$

$$\bar{x}_P = 196, \hat{\sigma}_P^2 = 491.5$$

$$F_{obs} = 1.33 < F_{0.05}^{4,5} = 5.19 \Rightarrow \text{conservation de } H_0 : \sigma_A^2 = \sigma_P^2$$

p -valeur : 0.75

Variance commune : $\hat{\sigma}^2 = 423.70$

Temps d'hibernation et influence géographique

On dispose des temps d'hibernation de différents individus, dans un échantillon des Alpes ($n_A = 6$) et un des Pyrénées ($n_P = 5$).



Alpes	195	184	166	185	149	151
Pyrénées	224	186	188	213	169	

Test de Student de comparaison des moyennes :

$$\bar{x}_A = 171.67, \bar{x}_P = 196 \quad \hat{\sigma}^2 = 423.70$$

$$t_{obs} = 1.95 < t_{0.05}^9 = 2.262 \Rightarrow \text{conservation de } H_0 : \mu_A = \mu_P$$

p -valeur : 0.08

On note que la p -valeur permet une discussion critique du résultat.

Table des matières

- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité
- 3 Statistiques descriptives, estimation et intervalles de confiance
- 4 Tests d'hypothèses
- 5 Tests du χ^2**
- 6 ANOVA 1
- 7 ANOVA 2
- 8 Analyse bivariable et corrélation
- 9 Modèle et régression linéaire
- 10 Décomposition de variance et test de linéarité

Les gros pandas vivent-ils en altitude ?

On mesure le poids de $N = 200$ pandas vivant à différentes altitudes.



Poids (kgs)	70-80	80-90	90-100	100-110	110-120
Altitude (m)					
1500-2000	28	5			
2000-2500	13	29	24	12	
2500-3000	6	12	21	12	5
3000-3500		2	10	16	5

Les catégories sociales se mélangent-elles ?

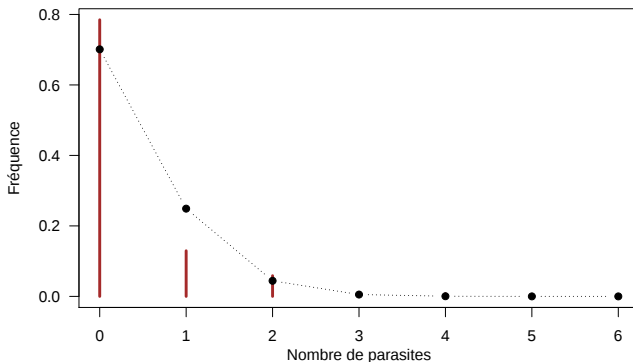
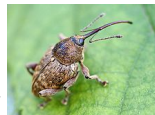
On note la catégorie sociale du mari et de la femme chez des couples pris au hasard.



	Homme	Cadre	Technicien	Employé
Femme				
Cadre		158	53	18
Technicienne		201	309	113
Employée		159	323	348

Les balanins suivent-ils vraiment une loi de Poisson ?

Nb. de parasites x_i	0	1	2	3	4	5	6+
Nombre de fruits n_i ayant x_i parasites	1043	172	78	15	10	7	4
Fréquence $f_i = \frac{n_i}{\sum_i n_i}$	0.785	0.129	0.059	0.011	0.007	0.005	0.004



Les balanins suivent-ils vraiment une loi de Poisson ?

$$k = 7 \quad n = \sum_{j=1}^k n_j = 1329$$

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k n_j x_j = \frac{1}{1329} (1043 \times 0 + 172 \times 1 + \dots)$$

= 0.36 parasites par fruit



Effectif théorique de la classe 0 :

$$n \times \mathcal{P}(\lambda = 0.36, X = 0) = 1329 \times \frac{0.36^0 e^{-0.36}}{0!} = 1329 \times 0.697 = 927.21$$

Nb. de parasites x_j	0	1	2	3	4	5	6 et plus
Nombre de fruits n_j ayant x_j parasites	1043	172	78	15	10	7	4
Nombre de fruits <i>théorique</i> ayant n_j parasites	927.21	333.79	60.08	7.21	0.64	0.05	0.02

On regroupe les classes, puis on calcule les χ^2 partiels et le χ^2 global.

$$\chi^2_{\{0\}} = \frac{(1043 - 927.21)^2}{927.21} = 14.45$$



Nb. de parasites x_j	0	1	2	3 et plus	Somme
Nombre de fruits n_j <i>observés</i> ayant x_j parasites	1043	172	78	36	1329
Nombre de fruits <i>théorique</i> ayant n_j parasites	927.21	333.79	60.08	7.92	1329
χ^2 partiel	14.45	78.42	5.34	99.56	197.78

$$\chi^2_{0.05, 4-1-1 \text{ ddl}} = 5.991 < 197.78 \Rightarrow H_0 \text{ rejetée et } H_1 \text{ acceptée.}$$
[illegible]

Les mesures de labos différents sont-elles comparables ?

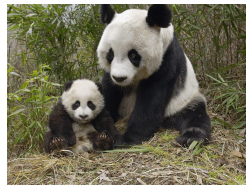
Trois labos dans 3 pays effectuent des comptages sur des échantillons bactériens issus de sols :



Souches Laboratoire	A	B	C	D
Lab. Belgique	146	38	57	4
Lab. Japon	89	17	26	0
Lab. Uruguay	101	27	40	3

Les gros pandas vivent-ils en altitude ?

On mesure le poids de $N = 200$ pandas vivant à différentes altitudes.

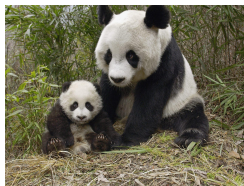


Poids (kgs) Altitude (m)	70-80	80-90	90-100	100-110	110-120	Somme
1500-2000	28	5				33
2000-2500	13	29	24	12		78
2500-3000	6	12	21	12	5	56
3000-3500		2	10	16	5	33
Somme	47	48	55	40	10	200

Le poids et l'altitude sont liés : règle de Bergmann

On calcule les effectifs théoriques. Par exemple pour la première case :

$$t_{11} = \frac{33 \times 47}{200} = 7.755$$



Poids (kgs) Altitude (m)	70-80	80-90	90-100	100-110	110-120	Somme
1500-2000	7.755	7.920	9.075	6.600	1.650	33
2000-2500	18.330	18.720	21.450	15.600	3.900	78
2500-3000	13.160	13.440	15.400	11.200	2.800	56
3000-3500	7.755	7.920	9.075	6.600	1.650	33
Somme	47	48	55	40	10	200

$$\chi_{obs}^2 = \frac{(28 - 7.755)^2}{7.755} + \frac{(5 - 7.920)^2}{7.920} + \frac{(0 - 9.075)^2}{9.075} + \dots = 123.81 > \chi_{0.05, 4 \times 3}^2 = 21.03$$

p -valeur : 1.10^{-20}

Le poids et l'altitude sont liés : règle de Bergmann

Attention si les effectifs théoriques dans certaines cases sont inférieurs à 5 !

Ici on peut regrouper facilement les deux dernières colonnes.



Poids (kgs) Altitude (m)	70-80	80-90	90-100	100-120	Somme
1500-2000	7.755	7.920	9.075	8.250	33
2000-2500	18.330	18.720	21.450	19.500	78
2500-3000	13.160	13.440	15.400	14.000	56
3000-3500	7.755	7.920	9.075	8.250	33
Somme	47	48	55	50	200

$$\chi_{obs}^2 = 120.34 > \chi_{0.05, 3 \times 3}^2 = 16.92$$

p -valeur : 1.10^{-21}

Table des matières

- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité
- 3 Statistiques descriptives, estimation et intervalles de confiance
- 4 Tests d'hypothèses
- 5 Tests du χ^2
- 6 ANOVA 1**
- 7 ANOVA 2
- 8 Analyse bivariable et corrélation
- 9 Modèle et régression linéaire
- 10 Décomposition de variance et test de linéarité

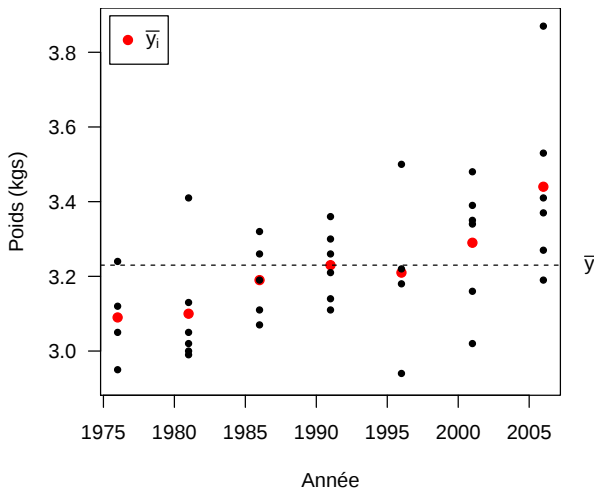
Evolution du poids des marmottes

On dispose des poids de différents individus, au total $N = 37$, pour chaque année.



Année	1976	1981	1986	1991	1996	2001	2006
Poids (kgs)	2.95	2.99	3.07	3.11	2.94	3.34	3.87
	3.24	3.00	3.26	3.26	3.18	3.02	3.41
	3.12	3.41	3.19	3.30	3.50	3.16	3.27
	3.05	3.02	3.32	3.14	3.22	3.48	3.37
		3.05	3.11	3.21		3.39	3.19
		3.13		3.36		3.35	3.53

Représentation graphique



Calculs intermédiaires

$$SCE_{tot} = (2.95 - 3.23)^2 + (3.24 - 3.23)^2 + \dots + (2.99 - 3.23)^2 + \dots$$

$$SCE_{inter} = 4 \times (3.09 - 3.23)^2 + 6 \times (3.1 - 3.23)^2 + \dots$$

$$SCE_{intra} = SCE_{tot} - SCE_{inter}.$$



Année	1976	1981	1986	1991	1996	2001	2006
Poids (kgs)	2.95	2.99	3.07	3.11	2.94	3.34	3.87
	3.24	3.00	3.26	3.26	3.18	3.02	3.41
	3.12	3.41	3.19	3.30	3.50	3.16	3.27
	3.05	3.02	3.32	3.14	3.22	3.48	3.37
		3.05	3.11	3.21		3.39	3.19
		3.13		3.36		3.35	3.53
$\bar{y}_i$ (kgs)	3.09	3.1	3.19	3.23	3.21	3.29	3.44
n_i	4	6	5	6	4	6	6

$$\bar{y} = 3.23 \text{ kgs}$$

Analyse par ANOVA 1



Variabilité	ddl	SCE	CM	F_{obs}	$F_{seuil}(\alpha = 0.05)$
Totale	36	1.327			
Intra	30	0.851	0.028		
Inter	6	0.476	0.079	2.821	2.42

On peut donc conclure, au risque de 5%, que le facteur temps a bien un effet sur le poids des marmottes.

Analyse par ANOVA 1 sous 

Variabilité	ddl	SCE	CM	F_{obs}	$F_{seuil}(\alpha = 0.05)$
Totale	36	1.327			
Intra	30	0.851	0.028		
Inter	6	0.476	0.079	2.821	2.42

```

              Df Sum Sq Mean Sq F value Pr(>F)
inter          6  0.4756  0.07927   2.794 0.0279 *
Residuals     30  0.8510  0.02837
---

```

Signif. codes: 0

p -valeur : 0.0279

Les variances étaient-elles égales ?



La taille des échantillons étant différente, on procède à un test de Bartlett sous  :

Bartlett test of homogeneity of variances

data: mamval and inter

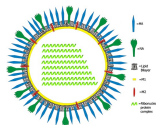
Bartlett's K-squared = 6.0374, df = 6, p-value = 0.419

Table des matières

- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité
- 3 Statistiques descriptives, estimation et intervalles de confiance
- 4 Tests d'hypothèses
- 5 Tests du χ^2
- 6 ANOVA 1
- 7 ANOVA 2**
- 8 Analyse bivariable et corrélation
- 9 Modèle et régression linéaire
- 10 Décomposition de variance et test de linéarité

Réplication du virus de la grippe

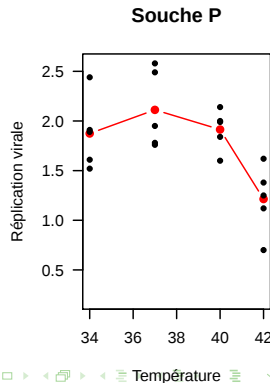
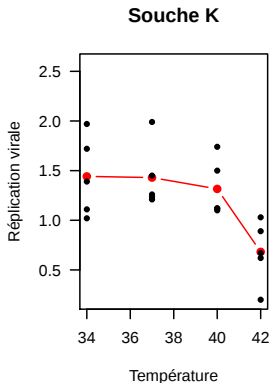
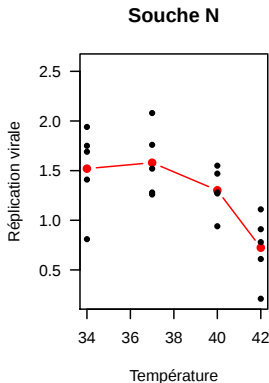
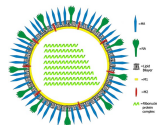
On mesure la vitesse de réplication (unités arbitraires) du virus de la grippe dans différentes conditions, pour différentes souches.



Souche	Température							
	34°C		37°C		40°C		44°C	
N	1.75	1.94	1.28	1.52	1.47	1.27	0.78	0.21
	1.41	0.81	2.08	1.26	1.55	0.94	1.11	0.61
	1.69		1.76		1.28		0.91	
K	1.02	1.97	1.26	1.21	1.74	1.12	0.89	0.67
	1.72	1.11	1.24	1.45	1.12	1.10	1.03	0.62
	1.39		1.99		1.50		0.20	
P	1.91	2.44	1.78	2.58	1.99	1.84	1.25	1.62
	1.52	1.61	1.76	1.95	2.14	1.60	1.38	1.12
	1.89		2.49		2.00		0.70	

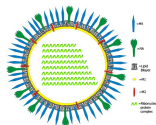
Représentation graphique

On représente souvent les moyennes sur le même graphe – ici ils sont séparés pour une meilleure lecture.



Réplication du virus de la grippe

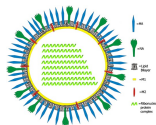
On calcule toutes les **moyennes** : par case, en ligne et en colonne.



	Température								
Souche	34°C		37°C		40°C		44°C		Moyennes
N	1.75	1.94	1.28	1.52	1.47	1.27	0.78	0.21	1.28
	1.41	0.81	2.08	1.26	1.55	0.94	1.11	0.61	
	1.69	1.52	1.76	1.58	1.28	1.30	0.91	0.72	
K	1.02	1.97	1.26	1.21	1.74	1.12	0.89	0.67	1.22
	1.72	1.11	1.24	1.45	1.12	1.10	1.03	0.62	
	1.39	1.44	1.99	1.43	1.50	1.31	0.20	0.68	
P	1.91	2.44	1.78	2.58	1.99	1.84	1.25	1.62	1.78
	1.52	1.61	1.76	1.95	2.14	1.60	1.38	1.12	
	1.89	1.87	2.49	2.11	2.00	1.91	0.70	1.21	
Moyennes	1.61		1.71		1.51		0.87		

Réplication du virus de la grippe : ANOVA2

Après application des formules, on obtient :



	ddl	Somme des carrés	Carrés moyens	F	F _{seuil}
Température	3	6.40	2.13	18.52	2.84
Souche	2	3.78	1.89	16.38	3.23
Interaction	6	0.14	0.02	0.2	2.34
Résidus	48	5.53	0.12		

table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Temperature	3	6.395	2.1318	18.519	4.05e-08	***
Souche	2	3.772	1.8861	16.384	3.77e-06	***
Temperature:Souche	6	0.138	0.0229	0.199	0.975	
Residuals	48	5.526	0.1151			

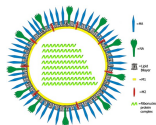
Signif. codes: 0

Conclusions statistiques

- $F_{Interaction} \leq F_{6,48}^{0.05}$: Pas d'interaction, le taux de réplication *varie de la même manière* en fonction de la température pour les différentes souches.
- $F_{Temperature} > F_{3,48}^{0.05}$: le taux de réplication moyen dépend de la température.
- $F_{Souche} > F_{2,48}^{0.05}$: le taux de réplication moyen dépend de la souche.

Réplication du virus de la grippe : deuxième cas

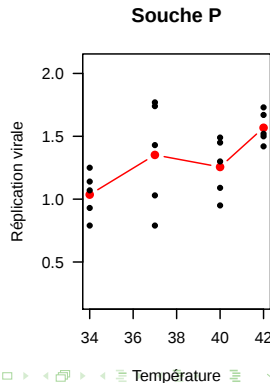
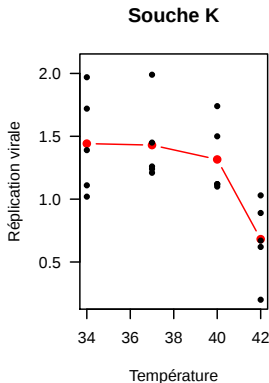
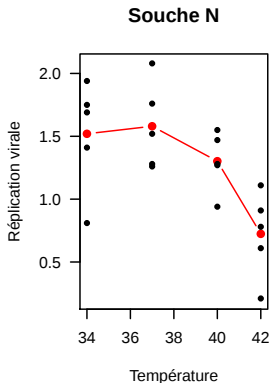
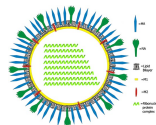
Par rapport au premier cas, seule la souche P a changé.



Souche	Température							
	34°C		37°C		40°C		44°C	
N	1.75	1.94	1.28	1.52	1.47	1.27	0.78	0.21
	1.41	0.81	2.08	1.26	1.55	0.94	1.11	0.61
	1.69		1.76		1.28		0.91	
K	1.02	1.97	1.26	1.21	1.74	1.12	0.89	0.67
	1.72	1.11	1.24	1.45	1.12	1.10	1.03	0.62
	1.39		1.99		1.50		0.20	
P	1.07	0.93	1.77	0.79	1.49	1.30	1.5	1.42
	0.79	1.25	1.74	1.43	1.45	1.09	1.67	1.52
	1.14		1.03		0.95		1.73	

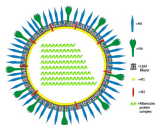
Représentation graphique : deuxième cas

On représente souvent les moyennes sur le même graphe – ici ils sont séparés pour une meilleure lecture.



Calculs intermédiaires : deuxième cas

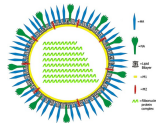
On calcule toutes les **moyennes** : par case, en ligne et en colonne.



	Température								
Souche	34°C		37°C		40°C		42°C		Moyennes
N	1.75	1.94	1.28	1.52	1.47	1.27	0.78	0.21	1.28
	1.41	0.81	2.08	1.26	1.55	0.94	1.11	0.61	
	1.69	1.52	1.76	1.58	1.28	1.30	0.91	0.72	
K	1.02	1.97	1.26	1.21	1.74	1.12	0.89	0.67	1.22
	1.72	1.11	1.24	1.45	1.12	1.10	1.03	0.62	
	1.39	1.44	1.99	1.43	1.50	1.31	0.20	0.68	
P	1.07	0.93	1.77	0.79	1.49	1.30	1.5	1.42	1.30
	0.79	1.25	1.74	1.43	1.45	1.09	1.67	1.52	
	1.14	1.03	1.03	1.35	0.95	1.25	1.73	1.57	
Moyennes	1.33		1.45		1.29		0.99		

Réplication du virus de la grippe : ANOVA2, deuxième cas

Après application des formules, on obtient :



	ddl	Somme des carrés	Carrés moyens	F	F _{seuil}
Température	3	1.74	0.58	5.66	2.84
Souche	2	0.08	0.04	0.39	3.23
Interaction	6	3.24	0.54	5.27	2.34
Résidus	48	4.91	0.10		

```

Df Sum Sq Mean Sq F value Pr(>F)
Temperature      3   1.738   0.5793    5.660 0.002106 **
Souche           2   0.079   0.0396    0.387 0.681497
Temperature:Souche 6   3.239   0.5398    5.274 0.000305 ***
Residuals       48   4.913   0.1023

```

Signif. codes: 0

Conclusions statistiques : deuxième cas

- $F_{Interaction} > F_{6,48}^{0.05}$: le taux de réplication **ne varie pas de la même manière** en fonction de la température pour les différentes souches.

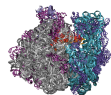
Donc on ne peut pas conclure sur les effets de la température ou de la souche, puisque ceux-ci ne sont pas systématiques.

Table des matières

- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité
- 3 Statistiques descriptives, estimation et intervalles de confiance
- 4 Tests d'hypothèses
- 5 Tests du χ^2
- 6 ANOVA 1
- 7 ANOVA 2
- 8 Analyse bivariable et corrélation**
- 9 Modèle et régression linéaire
- 10 Décomposition de variance et test de linéarité

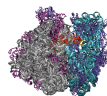
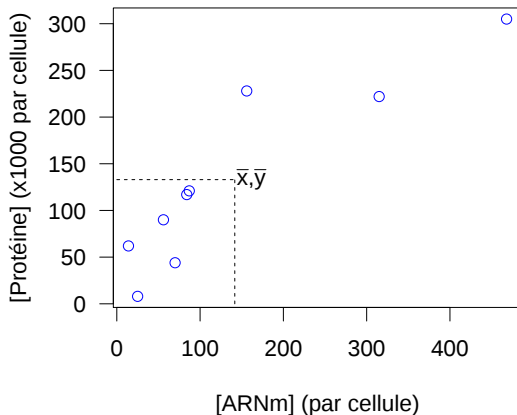
Existe-t-il un lien quantitatif entre concentration d'ARNm et de protéine ?

On étudie dans plusieurs tissus la concentration en ARNm et en protéine pour un gène donné.



ARNm (/cellule)	Protéine ($\times 1000$ /cellule)
87	121
156	228
56	90
70	44
14	62
25	8
468	305
84	117
315	222

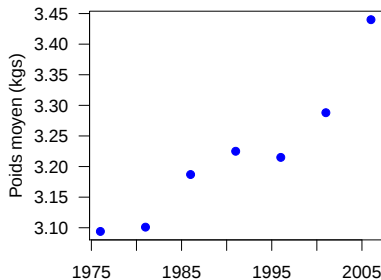
Représentation graphique



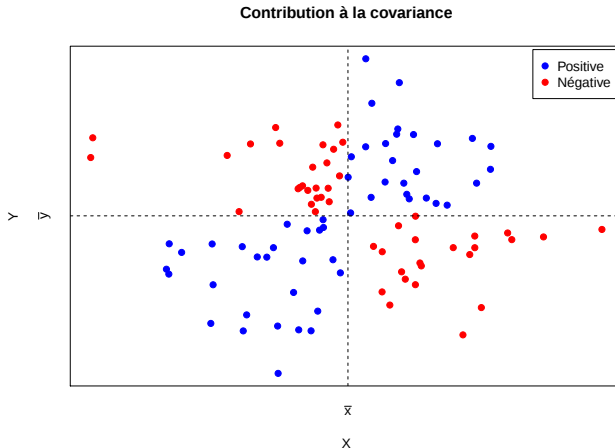
La variable contrôlée est toujours placée en abscisse!



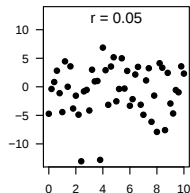
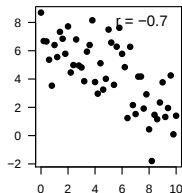
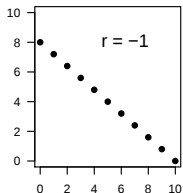
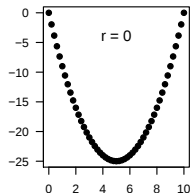
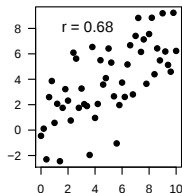
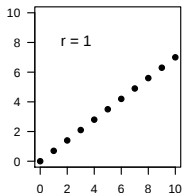
Année	Poids moyen (kgs)
1976	3.09
1981	3.10
1986	3.19
1991	3.23
1996	3.21
2001	3.29
2006	3.44



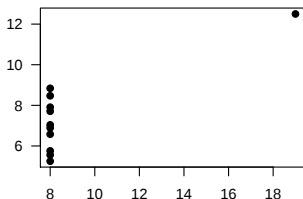
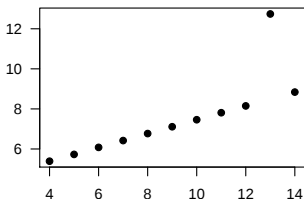
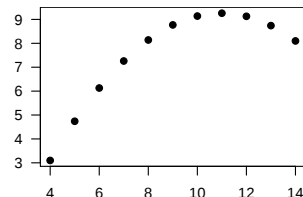
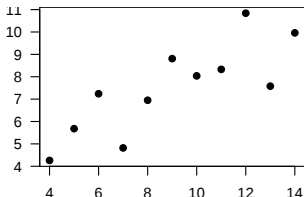
Les différentes contributions à la covariance



Interprétation graphique de r



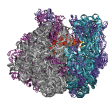
Le quartet d'Anscombe



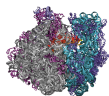
Pour ces 4
graphes :
 $r = 0.82$

Calculs intermédiaires

Variables	ARNm X (/cellule)	Protéine Y (x1000/cellule)
	87	121
	156	228
	56	90
	70	44
	14	62
	25	8
	468	305
	84	117
	315	222
<i>N</i>	9	9
Moyennes	141.667	133
Variances	20604.667	8596.222
Écarts-types	143.543	92.716



Résultats finaux



$$s_{XY} = 12022.889$$

$$r_{XY} = \frac{12022.889}{143.543 \times 92.716} = 0.903, r_{XY}^2 = 0.815$$

$$t_{obs} = \frac{0.903\sqrt{7}}{\sqrt{1-0.815}} = 5.561$$

La valeur seuil pour $N = 7$ ddl et un risque $\alpha = 0.05$ vaut 2.365 :

$t_{obs} > t_{seuil}$, on rejette H_0 et on accepte H_1 .

La concentration en protéine dans une cellule est linéairement reliée à la concentration dans les ARNm correspondants.

Sous , tout est plus simple

```
> cov(dat$ARNm, dat$Protéine)  
[1] 13525.75
```

Sous , tout est plus simple

```
> cov(dat$ARNm, dat$Protéine)
```

```
[1] 13525.75
```

```
> cov(dat$ARNm, dat$Protéine)*8/9
```

```
[1] 12022.89
```

```
> cor(dat$ARNm, dat$Protéine)
```

```
[1] 0.9033836
```

```
> cor.test(dat$ARNm, dat$Protéine)
```

Pearson's product-moment correlation

data: dat\$ARNm and dat\$Protéine

t = 5.5736, df = 7, p-value = 0.0008389

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.5980901 0.9797174

sample estimates:

cor

0.9033836

La vengeance des pandas alpins

On mesure le poids de $N = 200$ pandas vivant à différentes altitudes.



Poids (kgs) Altitude (m)	70-80	80-90	90-100	100-110	110-120
1500-2000	28	5			
2000-2500	13	29	24	12	
2500-3000	6	12	21	12	5
3000-3500		2	10	16	5

Test du coefficient de corrélation linéaire sur des données groupées

En notant X l'altitude et Y le poids, les calculs donnent (vérifiez que vous arrivez à le retrouver!) :

$$\bar{x} = 2472.5 \quad s_X^2 = 226743.75 \quad \bar{y} = 90.9 \quad s_Y^2 = 141.19$$

$$s_{XY} = 3587.34 \quad r = 0.634 \quad r^2 = 0.402$$

$$t_{obs} = \frac{0.634\sqrt{198}}{\sqrt{1 - 0.402}} = 11.536$$

Pour 198 ddl et un risque $\alpha = 0.05$, $t_{seuil} = 1.96$, on rejette H_0 et on accepte H_1 : il existe une relation linéaire entre le poids des pandas et l'altitude à laquelle ils vivent.

Test du coefficient de corrélation linéaire sur des données groupées avec

```
> ### X et Y ont été mis en mémoire  
> mean(X)
```

```
[1] 2472.5
```

```
> mean(Y)
```

```
[1] 90.9
```

```
> cor.test(X,Y)
```

Pearson's product-moment correlation

data: X and Y

t = 11.536, df = 198, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.5430291 0.7102682

sample estimates:

cor

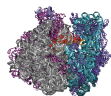
0.634004

Table des matières

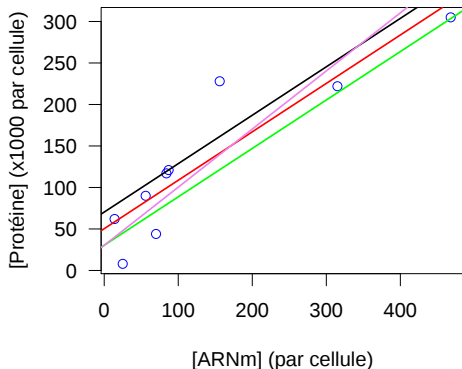
- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité
- 3 Statistiques descriptives, estimation et intervalles de confiance
- 4 Tests d'hypothèses
- 5 Tests du χ^2
- 6 ANOVA 1
- 7 ANOVA 2
- 8 Analyse bivariable et corrélation
- 9 Modèle et régression linéaire**
- 10 Décomposition de variance et test de linéarité

Quelle droite entre [ARNm] et [protéine] ?

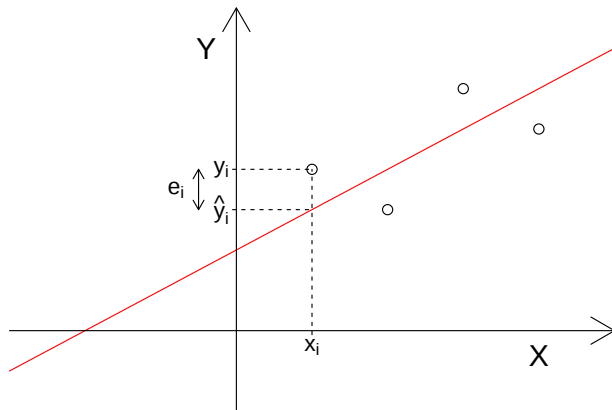
On veut maintenant modéliser la concentration en ARNm et en protéine pour un gène donné.



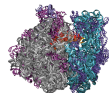
ARNm (/cellule)	Protéine (x1000/cellule)
87	121
156	228
56	90
70	44
14	62
25	8
468	305
84	117
315	222



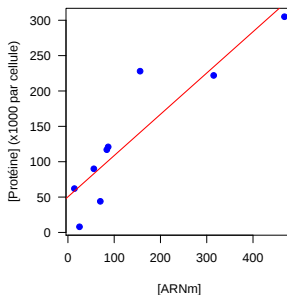
Notion d'écart à la valeur prédite



"N'utilise pas tes yeux, Luke. Utilise... la régression"



ARNm X (/cellule)	Protéine Y (x1000/cellule)
87	121
156	228
56	90
70	44
14	62
25	8
468	305
84	117
315	222



$$n = 9, \bar{x} = 141.667, \bar{y} = 133$$

$$\hat{a} = 0.584, \hat{b} = 50.266$$

$$s_X^2 = 20604.667, s_{XY} = 12022.889$$

$$Y = 0.584X + 50.266$$

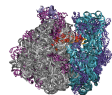
Sous ?

On peut faire tous les calculs sous  :

```
> ARNm<-c(87,156,56,70,14,25,468,84,315)
> Proteine<-c(121,228,90,44,62,8,305,117,222)
> lm(Proteine~ARNm)
```

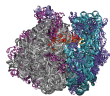
```
Call:
lm(formula = Proteine ~ ARNm)
```

```
Coefficients:
(Intercept)          ARNm
    50.3370         0.5835
```



L'attaque des résidus

[ARNm]	[Protéine] (x1000)	Estimation	Erreur
x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
87	121	101.074	19.926
156	228	141.37	86.63
56	90	82.97	7.03
70	44	91.146	-47.146
14	62	58.442	3.558
25	8	64.866	-56.866
468	305	323.578	-18.578
84	117	99.322	17.678
315	222	234.226	-12.226

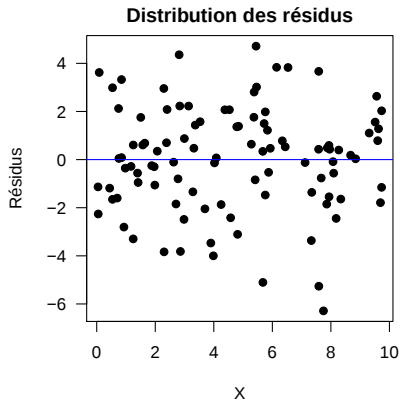
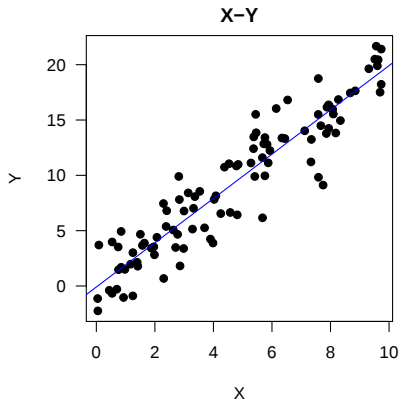


On vérifie :

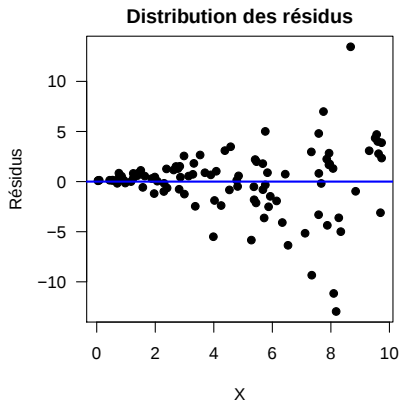
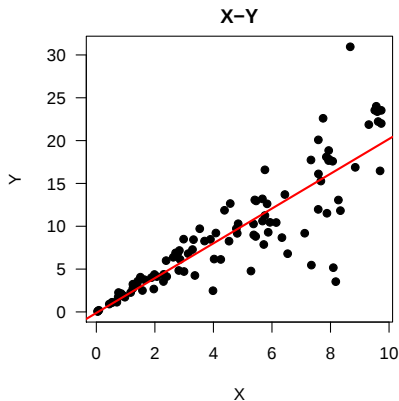
$$\sum_{i=1}^n e_i = 0$$

(ici
 $\sum_{i=1}^n e_i = 0.006$ à
 cause des erreurs
 d'arrondis)

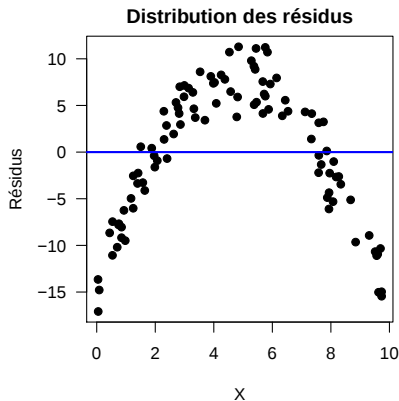
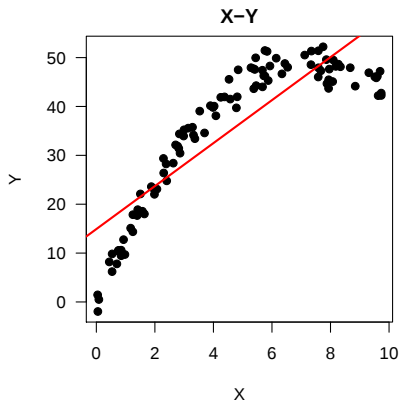
La contre-attaque des résidus



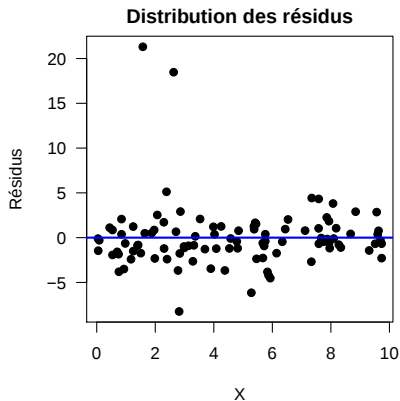
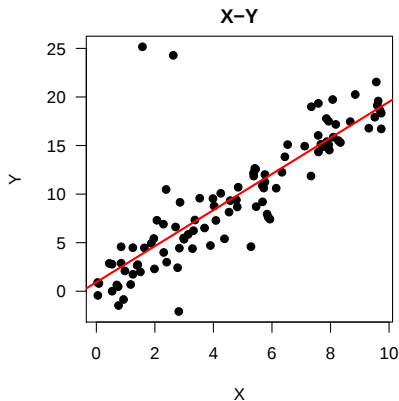
Les résidus ninjas hétéroscédastiques



Le sacrifice des résidus autocorrélés



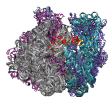
Les résidus extrêmes : renaissance



En pratique : intervalles de confiance

ARNm X (/cellule)	Protéine Y (x1000/cellule)
87	121
156	228
56	90
70	44
14	62
25	8
468	305
84	117
315	222

$$n = 9, \bar{x} = 141.667, \bar{y} = 133$$



$$s_X^2 = 20604.667, s_{XY} = 12022.889$$

$$\hat{\sigma}_R^2 = \frac{N}{N-2} \sum_{i=1}^N e_i^2 = 2032.5$$

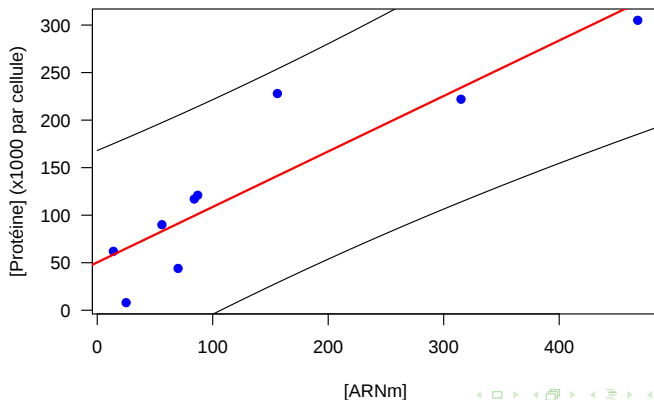
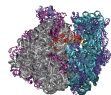
$$\hat{a} = 0.584, \hat{b} = 50.266$$

$$t_{(9-2),0.05} = 2.365$$

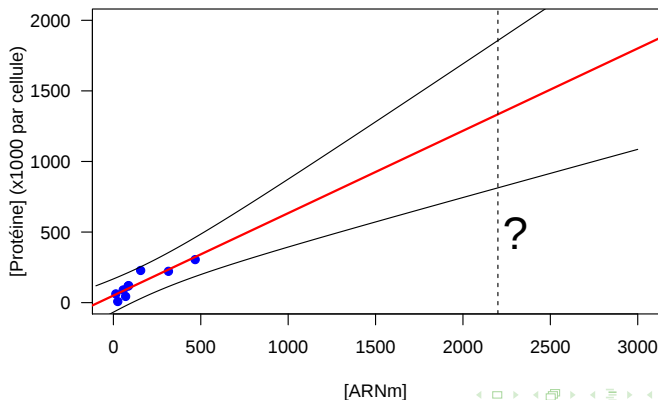
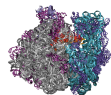
$$IC_{0.05} a : 0.336 < a < 0.832$$

$$IC_{0.05} b : 14.725 < b < 85.807$$

Prédictions pour les valeurs individuelles

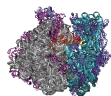


Prédictions à grande échelle ?



Test de la pente

On veut vérifier si notre gène d'intérêt a la même relation [ARNm]-[protéine] qu'un ensemble de gènes de référence, pour lesquels on a $\gamma = 0.42$.



Test de la pente :

$$t_{obs} = \left| \frac{0.584 - 0.42}{\sqrt{\frac{2032.5}{9 \times 20604.667}}} \right|$$
$$t_{obs} = 1.567, \quad t_{7,0.05} = 2.365$$

$t_{obs} < t_{7,0.05}$: On conserve H_0 au risque β inconnu, $a = \gamma$.

Table des matières

- 1 Présentation du cours de biostatistiques et bioinformatique
- 2 Variables aléatoires et lois de probabilité
- 3 Statistiques descriptives, estimation et intervalles de confiance
- 4 Tests d'hypothèses
- 5 Tests du χ^2
- 6 ANOVA 1
- 7 ANOVA 2
- 8 Analyse bivariable et corrélation
- 9 Modèle et régression linéaire
- 10 Décomposition de variance et test de linéarité**

Le retour des marmottes du Jedi

Quel test effectuer ? ANOVA1 ou test de corrélation linéaire ? Lequel est le plus général ?

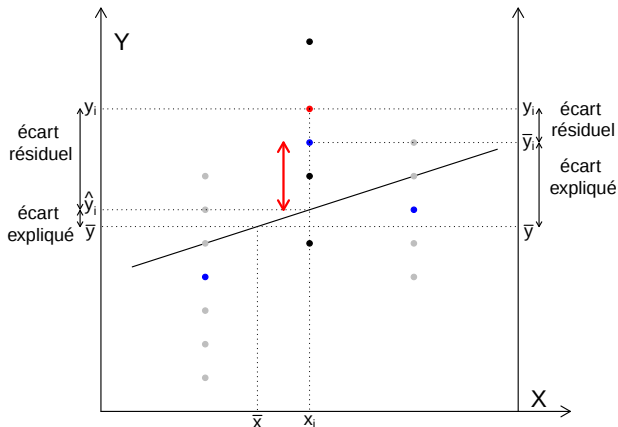


Année	1976	1981	1986	1991	1996	2001	2006
Poids (kgs)	2.95	2.99	3.07	3.11	2.94	3.34	3.87
	3.24	3.00	3.26	3.26	3.18	3.02	3.41
	3.12	3.41	3.19	3.30	3.50	3.16	3.27
	3.05	3.02	3.32	3.14	3.22	3.48	3.37
		3.05	3.11	3.21		3.39	3.19
		3.13		3.36		3.35	3.53

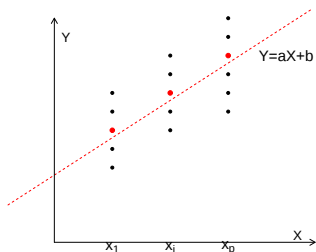
Graphiquement

Modèle linéaire

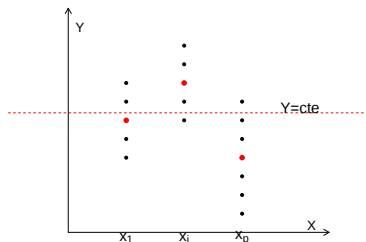
Modèle ANOVA 1



Exemple d'interprétation graphique

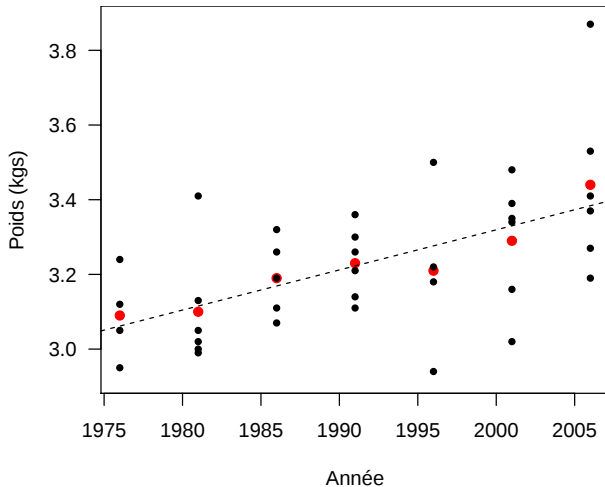


Les 2 tests sont significatifs, $\eta^2 \simeq R^2$:
Relation fortement linéaire.



Le test de l'ANOVA 1 est significatif,
mais pas le test de linéarité, $\eta^2 \gg R^2$:
Relation non linéaire.

L'Empire des marmottes contre-attaque



L'ultime sacrifice des marmottes fantômes



Analyse par ANOVA 1.

Variabilité	ddl	SCE	CM	F_{obs}	$F_{seuil}(\alpha = 0.05)$
Totale	36	1.327			
Intra	30	0.851	0.028		
Inter	6	0.476	0.079	2.821	2.42

On peut donc conclure, au risque de 5%, que le facteur temps a bien un effet sur le poids des marmottes.

Les marmottes des étoiles : nouvelle génération



Analyse de linéarité.

Variabilité	ddl	SCE	CM	F_{obs}	$F_{seuil}(\alpha = 0.05)$
Inter (résiduelle)	30	0.851	0.028		
Expliquée	1	0.423	0.423		
Ecart à la régression	5	0.053	0.011	0.393	2.53

F_{obs} est inférieur à 1 (et donc à fortiori à F_{seuil}), on doit donc conserver H_0 et accepter la linéarité du poids des marmottes avec le temps.

En guise de conclusion générale. . .

Vous devriez maintenant savoir :

- Quand et comment faire les tests correspondants aux situations biologiques classiques,
- Comment interpréter les sorties d'un langage statistique comme  dans le cadre de tests,
- Qu'on peut faire des statistiques en manipulant des concepts et pas des nombres.

Si ces cours vous ont intéressé, rendez-vous en L3, et pourquoi pas en Bioinformatique, Statistiques et Modélisation (BISM) ?

May the stats be with you.